Dissecting human population variation in single-cell responses to SARS-CoV-2

Yann Aquino

SORBONNE UNIVERSITÉ

INSTITUT PASTEUR

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Bioinformatique et biologie des systèmes

École doctorale 515 | Complexité du vivant

Dissecting human population variation in single-cell immune responses to viral infection

Présentée par

Yann AQUINO

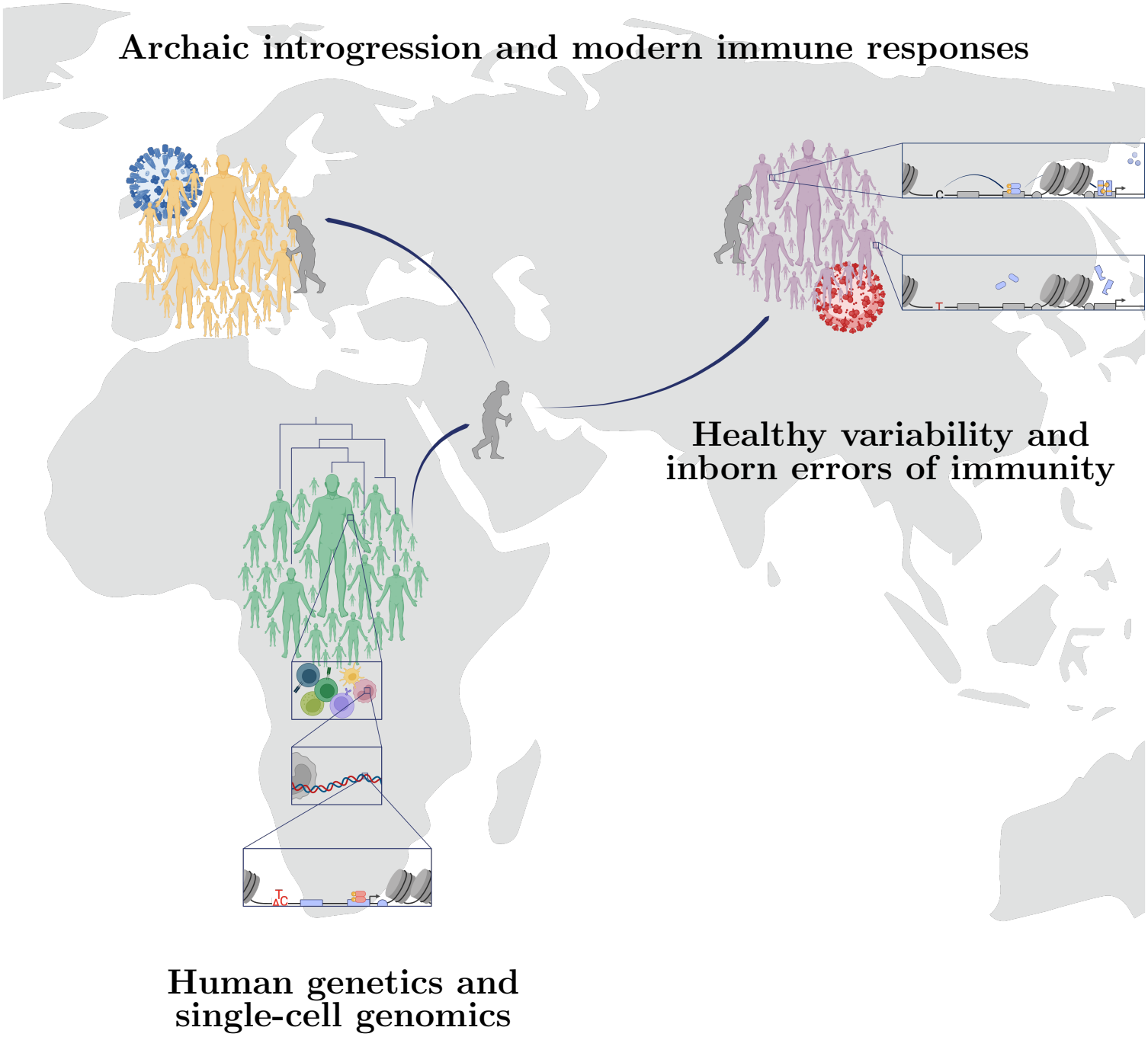
pour obtenir le grade de

DOCTEUR DE SORBONNE UNIVERSITÉ

le 15 décembre 2023

devant le jury composé de

M^me^	Monique VAN DER WIJST	Rapporteure
M.	Luis Bruno BARREIRO	Rapporteur
M^me^	Sarah KIM-HELLMUTH	Examinatrice
M^me^	Laure SÉGUREL	Examinatrice
M.	Gilles FISCHER	Président du jury
M.	Lluís QUINTANA-MURCI	Directeur de thèse
M.	Maxime ROTIVAL	Membre invité


Archaic introgression and modern immune responses
Healthy variability and
inborn errors of immunity
Human genetics and
single-cell genomics

Para el Nonno, por una vida dedicada a la enseñanza.

Preface

In the Human Evolutionary Genetics Unit of Institut Pasteur, I worked in a diverse environment. The multidisciplinary nature of the team—marrying population genetics methods with functional genomics approaches—is reflected in the variety of questions adressed by my thesis project. Why do healthy humans respond differently to infection by the same virus? How does the immune response vary across populations and cell types? What is the role of genetics in this variability? Where do the genetic predictors of immune variability come from?

Hoping to channel this complexity, I wrote this manuscript in three Parts. First, I introduce the main concepts underlying my contributions to the field of population single-cell genomics. Second, I present said contributions in their published state, along with a few lines describing the context in which they were produced. Finally, I discuss how my work relates to other pieces of knowledge published in the field, as well as some perspectives on possible future developments.

Human genetics and single-cell genomics. Genetic diversity is the unifying concept of this multidimensional project; each Chapter in the first Part spans a question around it. Chapter 1 is centered on how information on human genetic diversity is used to map the genetic bases of complex traits, with a particular focus on disease phenotypes. I start with an introduction to some of the main data bases of human genetic diversity of the 21st century—built in the wake of the Human Genome Project—that are essential for most genomic analysis tasks today. In particular, I then focus on how these resources are used to map the genetic bases of complex traits genome-wide, and how molecular endophenotypes can easen causal inference of genetic effects on phenotype. Finally, I discuss how single-cell assays of molecular endophenotypes across different layers of gene expression regulation are helpful to disentangle the context-dependency of these effects, and thus maximize the chance to detect bona fide causal links between genotype and phenotype. ■

Archaic introgression and modern immune responses. Next, Chapter 2 explores how extant human genetic diversity can inform on the impact of ancient events in the shared human evolutionary history that shaped present-day immune responses. I begin with a high-level overview of the early evolution of the ancestors of anatomically modern humans, including their expansion out of Africa, and their interactions with 'archaic' human forms in Eurasia. I then review some of the main methods used to detect archaic genetic material in modern human genomes, as well as pieces of evidence that show these exchanges helped modern humans adapt to new pathogenic environments during the colonization of Eurasia, Oceania and The Americas. Finally, I focus on the role of viral pathogens as drivers of human evolution, and argue the importance of characterizing the evolutionary forces that shaped the genetic architecture of infectious disease risk in order to better defend against future viral outbreaks. ■

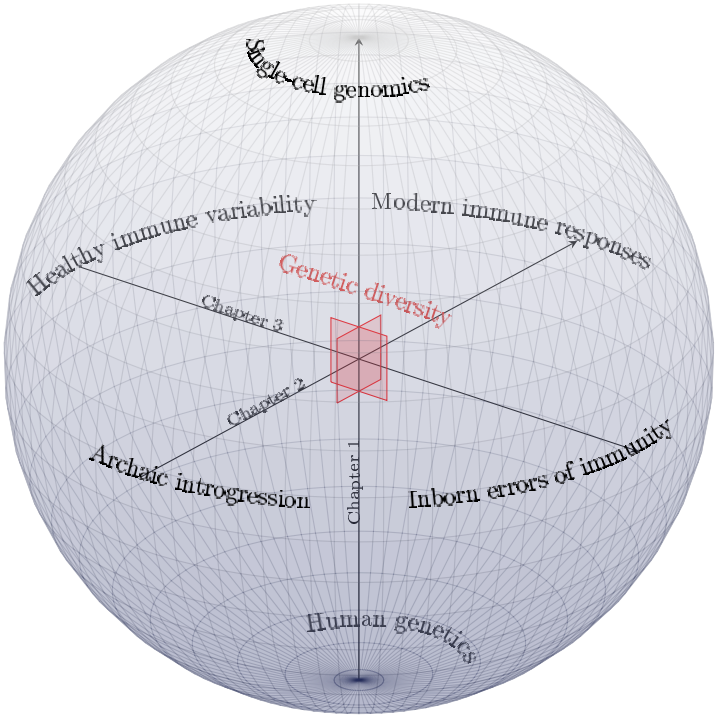
Healthy variability and inborn errors of immunity. Finally, Chapter 3 expands beyond the realm of healthy human genetic diversity and into inborn errors of immunity. I start with a brief presentation of the innate immune response to viruses in the peripheral blood, from the point of view of gene regulatory networks wired to sense pathogens and mount an appropriate transcriptional response. Next, I go over the genetic and nongenetic drivers of immune response variability across

healthy individuals and in the context of viral infection, before moving on to the case of individuals with genetic deficiencies of antiviral immunity. In this last Section, I review how the study of errors of antiviral immunity in rare individuals can yield insights into the predictors of variability in infectious disease risk, including common immunological factors with potential population-level impacts on disease susceptibility and the response to vaccination. ■

My intention is to introduce all of these concepts and ideas in a way that reflects my own intuition of the underlying biology. Although I aimed to be as extensive as possible, so as to place each method employed in Aquino et al. (2023) in its rightful context, I do not pretend to provide a comprehensive review of the genomicist's toolkit. For that, I strongly recommend the 'Handbook of statistical genomics' by Balding et al. (2019).

Given the context in which my project started in early 2020, all of my discussion is strongly influenced by the coronavirus outbreak of 2019 and its associated disease. In fact, the whole project itself was shaped by the pandemic, the need for a comprehensive assessment of the drivers of immune variability in the response to viral infection, and the unprecented amounts of data and resources pooled to answer this need.

All in all, my work is inscribed within the larger paradigm of precision medicine. With the advancement of technology and human knowledge, a medicine tailored to each individual, their innate and acquired features and their environment, is within reach. This implies a strong evolutionary understanding of the genetic bases of disease risk, and of the relative contributions of genetic and nongenetic predictors of its variability across individuals.



Acknowledgments

This manuscript crystallises over three years of lessons, hard work, blood [a], sweat [b] and tears [c]. In what follows, I intend to render proper tribute to each and every person who contributed to make these some of the most fulfilling and growth-inducing years of my life. I may not cite everyone by name, but please know that I am immensely grateful to all who helped carry this project to completion; our interactions shaped me into the scientist I am today.

To my jurors. I thank the jury of expert and busy scientists who took the time to evaluate and comment my work: Dr. Monique van der Wijst and Dr. Luis B. Barreiro, your kind and thoughtful assessments of my manuscript touched me greatly—they are a source of inspiration and motivation for my future research endeavours. I relish in advance of the discussion we will have on December 15$^{\text{th}}$, alongside Dr. Sarah Kim-Hellmuth, Dr. Laure Ségurel and Dr. Gilles Fischer, whom I also warmly thank for their time and for sharing their expertise with me. ■

To my collaborators. I thank all of my collaborators for their contributions to my research work, as well as the people who allowed me to contribute to their projects. In particular, I thank Dr. Sarah Merkling, Dr. Giovanna Barba-Spaeth and Dr. Stefano Pietropaoli for teaching me the basics of viral culture and work in a biosafety level 3 environment. I also thank Dr. Milena Hasan and Dr. Valentina Libri for their training on single-cell technologies. I thank Dr. Darragh Duffy and his team for their work, and more generally all members of the Milieu Intérieur Consortium for the insightful discussions we had. I thank Dr. Mickaël Ménager and Dr. Camille de Cevins for their help disentangling cell-type identities from single-cell transcriptomic data. Finally, I thank Dr. Danyel Lee, Dr. Shen-Ying Zhang and the Human Genetics of Infectious Diseases team led Dr. Laurent Abel and Dr. Jean-Laurent Casanova for their input on my work, as well as for including me on a highly relevant and high-profile piece of scientific work. ■

To my teammates. I am also thankful for the group of amazing researchers and people with whom I had the pleasure to work for the past three years. You have given me too much to fit in a few lines, but I will try. Aurélie, Lisa and Christine, thank you for walking me through the first stressful months at the bench, thank you for your kindness. Guillaume and Etienne, thank you for sharing your knowledge and experience on all things computational. Maguelonne and Jérémy, thank you for introducing me to the Pasteurian night-life, and for keeping our team's fashion standards high. Oğuzhan and Jan, thank you for your sweetness, your friendliness and your party-readiness. Sara, thank you for the stories, for the business opportunities and for the Wordle puzzles. Dang and Sebastián, thank you for trying your best to beat me at Mario Kart. Axel, thank you for the cocktails and the guitar concerts. Jacob, thank you for your lessons on Swedish folk songs. Gastón, thank you for the great house parties. Javier and Lara, thank you for your kindness and cheer.

a. Ask Maxime—he knows.
b. It gets hot inside P3 Nocard in the summertime.
c. See *a.*

To my dearly departed (they are fine, they just left) officemates: Mary, thank you for embodying the spirit of the self-proclaimed 'mean' office, thank you for the guidance, thank you for the sass, thank you for the laughs, thank you for the great tacos; Gaspard, thank you for the eye-opening philosophical musings, thank you for the light-hearted yet insightful discussions.

Above everything, thank you all for your kindness and for the good times we spent together, inside and outside Pasteur. After these three years, several questions remain, yet one thing is clear: no aspiring doctoral candidate could wish or hope for a better team than the Human Evolutionary Genetics Unit of Institut Pasteur, ascertainment bias be damned. ■

To my mentors. Lluís, applying for a Master's intern position in your team was arguably one of the easiest choices I have ever made. As soon as I read your subject proposal—the first of a long list—I knew yours was the right project for me; I did not really need to read the others. Nearly four years later, I could not be happier about where that choice led me. Joining your team, I expected to learn everything about human immune diversity and adaptation to viruses; yet, little did I know at the time that I was also going to learn a great deal about European monarchies and their dirty laundry through your spontaneous podcasts. A happy surprise. Jokes aside, thank you for taking a chance on me. Thank you for the lessons. Thank you for your trust and the opportunities to represent the team internationally through talks and collaborations. Thank you for your steady direction, which has brought me academic recognition beyond my wildest expectations.

Maxime, thank you for teaching me your language of $\widehat{\beta}$, σ^2, ε and other such unholy things. Thank you for your unwavering patience and for the generosity with which you shared your time and knowledge with me. Thank you for supervising my work and for helping me correct my errors, and find interesting biological insights amid the dreams you broke. Thank you for knowing when to be my supervisor, and when to be a friend. ■

À mes amis. Je remercie tous les amis—brestois ou magistériens, vous vous reconnaîtrez—qui m'ont accompagné durant ces trois années et qui, avec leur bonne humeur et leur gentillesse, ont rendu ma vie de thésard plus agréable. J'ai toutefois une pensée toute particulière pour Marie, Gaël et Paul, qui ont été davantage présents dans cette partie de ma vie—peut-être parce que thésards eux-mêmes—et qui m'ont en plus apporté leur expeRtise. ■

À ma famille. Je voudrais remercier ma tante Colette, mon oncle Christian et ma cousine Sonia, qui m'ont accueilli quand je suis arrivé en France, il y a de ça dix ans. Sans la stabilité que vous m'avez apporté ces premiers mois—pourtant décisifs—loin de chez moi, et sans votre soutien inconditionnel ensuite, je ne serais sûrement pas là aujourd'hui. Je ne peux que regretter que Tonton Christian ne soit pas parmi nous aujourd'hui pour fêter comme il aimait le faire, mais je pense fort à lui. Du fond du cœur, merci.

À mon autre famille, celle que j'ai trouvé à Douarnenez. Patricia, Estelle, Joffrey, Anne-Marie, Jean-Ronan, Anne-Yvonne, Alain, Samuel et Corentin, merci de m'avoir ouvert grand les bras et de m'avoir accepté comme l'un des vôtres. Je chéris les moments passés à vos côtés, qui ont souvent allégé les périodes stressantes de ma longue vie d'étudiant. ■

À mes parents. Je remercie ma mère pour son amour inconditionnel, pour m'avoir inculqué l'importance de la discipline et pour la nationalité française, sans quoi je n'aurais certainement pas fait mes études en France. Je te sais fière de mes travaux, et tu peux l'être de toi-même aussi; pour ces raisons, mes réussites sont aussi les tiennes.

A mi padre le agradezco su amor, su presencia y apoyo constante, y el haberme enseñado la importancia del trabajo. Sin ustedes no habría llegado hasta aquí. ■

À Ophélie. À ma complice d'aventure. Rien de tout cela n'aurait été possible sans ta compagnie, sans ton soutien ni tes conseils, sans tes blagues ni nos fou-rires, sans l'amour qu'on partage. Rien de tout cela n'a d'importance si je ne peux le partager avec toi. Tu es mon repère, tu es ma force. Je t'aime. ■

List of Figures

List of Boxes

List of Acronyms

1KG 1000 Genomes (Project Consortium)

ADP Adenosine diphosphate

AI Artificial intelligence

AMD Age-related macular degeneration

AMH Anatomically modern humans

ARG Ancestral recombination graph

scATAC-seq Single-cell assays for transposase-accessible-chromatin sequencing

ATP Adenosine triphosphate

BBJ BioBank Japan (Project)

BMI Body mass index

CD Cluster of differentiation

CD/CV Common disease/common variant

CEPH Centre d'Étude du Polymorphisme Humain

CITE-seq Cellular indexing of transcriptome and epitopes by sequencing

CMV Cytomegalovirus

COVID-19 Coronavirus disease 2019

CPM Counts per million

CRE *Cis*-regulatory element

CRE-PLS *Cis*-regulatory element with promoter-like signature

CRE-ELS *Cis*-regulatory element with enhancer-like signature

CRF Conditional random field

CTCF CCCTC-binding factor

DAMP Damage-associated molecular pattern

DC Dendritic cell

DL Deep learning

pDC Plasmacytoid DC

DE Differential expression; differentially expressed

DHS DNase hypersensitive site

DNA Deoxyribonucleic acid

cDNA Complementary DNA

mtDNA Mitochondrial DNA

EBI European Bioinformatics Institute

EMRA Effector memory re-expressing CD45RA (T cells)

ENCODE Encyclopedia of DNA elements

FBM Feature-barcode matrix

FC Fold-change

FE Fold-enrichment

GCTA Genome-wide complex trait analysis

GLM Generalized linear model

GRC Genome Reference Consortium

GREML Genome-based restricted maximum likelihood

GRN Gene regulatory network

GTEx Genotype Tissue Expression (Consortium)

GWAS Genome-wide association study

HGDP Human Genome Diversity Project

HGP Human Genome Project

HGSVC Human Genome Structural Variation Consortium

HIV Human immunodeficiency virus

HPRC Human Pangenome Reference Consortium

HVG Highly variable gene

IAV Influenza A virus

IBD Inflammatory bowel disease; identical-by-descent (of alleles)

IEI Inborn error of immunity

IFN Interferon

IL Interleukin

ILS Incomplete lineage sorting

IRF IFN regulatory factor

ISG IFN-stimulated gene

LD Linkage disequilibrium

MAF Minor allele frequency

MAIT Mucosal-associated invariant T (cells)

MERS-CoV 'Middle East respiratory syndrome' coronavirus

MIS-C Multisystem inflammatory syndrome in children

MRE Multiregional evolution, (Model of)

MXB Mexican Biobank

NB Negative binomial

NCBI National Center for Biotechnology Information

NGS Next-generation sequencing

NIH National Institutes of Health

NK Natural killer (cell)

NLP Natural language processing

NMF Non-negative matrix factorization

OAS 2'-5'-oligoadenylate synthetase

OOA Out-of-Africa

OR Odds ratio

PAMP Pathogen-associated molecular pattern

PBMC Peripheral blood mononuclear cell

PCA Principal components analysis

PP Posterior probability

PRR Pattern-recognition receptor

PRS Polygenic risk score

QC Quality control

QTL Quantitative trait locus

caQTL Chromatin accessibility QTL

coeQTL Co-expression QTL

eQTL Expression QTL

ieQTL Interaction expression QTL

molQTL Molecular QTL

pQTL Protein QTL

sQTL Splicing QTL

RNA Ribonucleic acid

mRNA Messenger RNA

scRNA-seq Single-cell RNA-sequencing

SARS-CoV 'Severe acute respiratory syndrome' coronavirus

SGDP Simons Genome Diversity Project

SNP Single nucleotide polymorphism

SLE Systemic lupus erythematosus

TMRCA Time to most recent common ancestor

t-SNE t-distributed stochastic neighbor embedding

T2T Telomere-to-Telomere (Consortium)

TF Transcription factor

TFBS TF binding site

TLR Toll-like receptor

TRE *Trans*-regulatory element

TSS Transcriptional start site

TWAS Transcriptome-wide association study

UKB United Kingdom Biobank

UMAP Uniform manifold approximation and projection

UMI Unique molecular indentifier

UTR Untranslated region

VIP Virus-interacting protein

WGS Whole-genome sequencing

WHO World Health Organization

WTCCC Wellcome Trust Case Control Consortium

Table of contents

Introduction

The long shared history of humans and viruses. Diseases caused by pathogenic microbes have been a leading cause of mortality for hundreds of thousands of years of human evolution (Casanova and Abel, 2005). In particular, viruses are known to have played an especially important role in shaping the genetic basis of present-day human immune responses (Enard et al., 2016). As increased human activity accelerates the rate of viral spillover (Jones et al., 2013), characterizing the relative contributions of the genetic, immunological, environmental and microbial predictors of immune variability (Casanova and Abel, 2013), as well as the impact of evolutionary forces on the genetic architecture of infectious disease risk (Quintana-Murci, 2019; Sella and Barton, 2019) is fundamental to better prepare against future outbreaks. ■

The extent of human immune variability in the response to viruses was strikingly illustrated by the ongoing 'coronavirus disease 2019' (COVID-19) pandemic, sparked by the 2019 outbreak of a novel coronavirus strain inducing severe acute respiratory syndrome (SARS-CoV-2). Indeed, infection by SARS-CoV-2 is characterized by a wide range of possible outcomes, from asymptomatic cases, to COVID-19 patients requiring intensive care, and even death. Although recent estimates place the rate of life-threatening COVID-19 at 2% to 4% of patients and the SARS-CoV-2 infection fatality ratio at around 0.5% to 1% of cases[d] (O'Driscoll et al., 2021; Bollyky et al., 2022), the World Health Organization (2020a) estimates that nearly 7 million lives have been lost to the disease, reflecting the prevalence of silent SARS-CoV-2 infection: Sah et al. (2021) estimated the rate of asymptomatic cases at around 35% through the meta-analysis of over 350 studies.

In fact, differences in COVID-19 susceptibility might explain the velocity of the global spread of SARS-CoV-2. During the first year of the pandemic, over 80 million people were reportedly infected by SARS-CoV-2, despite widespread travel restrictions and social distancing measures. In contrast, basic epidemiological interventions were sufficient to contain the first SARS-CoV-1 outbreak after 8 months, limiting the death toll to approximately 8 thousand lives in 2002. Given that both coronaviruses feature similar aerosol and surface stabilities (van Doremalen et al., 2020), the differences in transmission of SARS-CoV-1 and 2 could be explained by the latter's highly infectious presymptomatic phase and high rate of asymptomatic cases (He et al., 2020). Hence, it is paramount to disentangle the genetic, immunological and environmental predictors of variability in the response to SARS-CoV-2, so as to more accurately describe variability in COVID-19 courses across healthy individuals and populations worldwide.

Several large-scale genome-wide association studies have contributed to map the common genetic predictors of susceptibility to SARS-CoV-2 infection and severe COVID-19 forms. In particular, the COVID-19 Host Genetics Initiative (2020, 2021, 2022, 2023) has greatly participated to this endeavour by proposing unified definitions for susceptibility and severity, as well as concentrating efforts and resources from research institutes world-wide. These and other studies (Ellinghaus et al., 2020; Pairo-Castineira et al., 2021; Shelton et al., 2021; Kousathanas et al., 2022; Horowitz et al., 2022) have unveiled several immune-relevant genomic regions associated to higher COVID-19 risk

d. For reference, the global case fatality ratio of tuberculosis—the largest infectious killer prior to the COVID-19 pandemic—was estimated at 14% in 2019 by the World Health Organization (2019).

(Shelton et al., 2021; Kousathanas et al., 2022), but also variants that protect against severe forms of the disease (Ellinghaus et al., 2020). Yet, assessing the effects on immune phenotypes that link these genetic factors to disease risk requires complementary functional genomic data.

In particular, transcriptomic data are commonly used to establish intermediate links between genotype and gene expression endophenotypes, which can help interpret the genetic association with an organismal phenotype (Lappalainen et al., 2013) like COVID-19 risk. For instance, Kousathanas et al. (2022) show evidence of genetic effects on COVID-19 severity that are mediated by genetically controlled changes in the expression of the *CCR9* chemokine locus, located in the genomic region most strongly associated to the risk of severe COVID-19 (Ellinghaus et al., 2020). However, while such results may point to real causal associations between genotype and phenotype, their biological relevance must be assessed with caution, as measured genetic effects on immune phenotypes are known to be strongly dependent on the nature (Barreiro et al., 2012) and length (Kim-Hellmuth et al., 2017) of stimulation, as well as cell-type identity (Kim-Hellmuth et al., 2020).

From this view, assays that allow to characterize the transcriptome at single-cell resolution are particularly useful to disentangle the context-dependency of genetic effects. Yet, although several single-cell transcriptomic studies of the response to SARS-CoV-2 have provided valuable insights into the ætiology of COVID-19 (Lee et al., 2020; Wilk et al., 2020; Ren et al., 2021; Stephenson et al., 2021), a systematic assessment of the effects of common genetic variants on the response to SARS-CoV-2 across immune cell types and human populations—likely to improve the interpretability of the wealth of genome-wide associations revealed by the COVID-19 Host Genetics Initiative (2020, 2021, 2022, 2023) and others—was lacking. Hence, we at the Human Evolutionary Genetics Unit of Institut Pasteur set out to characterize the boundaries of natural variability in the immune response to SARS-CoV-2 across healthy and diverse individuals, map its genetic basis, and describe the evolutionary forces that shaped it (Aquino et al., 2023).

Our evolutionary perspective is essential because demographic events in human evolutionary history, as well as events of adaptation driven by natural selection are known to have shaped the genetic architectures of complex traits—such as COVID-19 risk—across human populations worldwide (Sella and Barton, 2019; Uricchio, 2020). Specifically, while recent population expansions can lead to private variants with large effects on phenotype in different populations (Lohmueller, 2014), negative selection on large-effect deleterious variants could explain why the bulk of complex trait heritability is distributed across many variants with small effects (O'Connor et al., 2019). Hence, approaches that are agnostic to these evolutionary processes are ill-equipped to explain disparities in heritable COVID-19 risks across populations (Shelton et al., 2021).

Moreover, it has been shown that some of the genetic variants associated to COVID-19 severity were introduced into the genomes of anatomically modern humans following admixture with archaic hominin forms, and have been brought to high frequencies in different human populations by natural selection. For example, Zeberg and Pääbo (2020, 2021) show that the haplotype spanning the *CCR9* locus associated to COVID-19 severity is of Neandertal origin, while another Neandertal haplotype spanning the *OAS1-3* locus in around 30% of Eurasian genomes has been shown to be associated to a 22% reduction in the risk of severe forms in COVID-19 patients (Ellinghaus et al., 2020), as well as to harbor variants displaying strong signals of selection (Sams et al., 2016).

More generally, viruses have been singled out as drivers of human adaptation to changing local environments (Enard et al., 2016; Souilmi et al., 2021), including via such events of adaptive archaic introgression (Enard and Petrov, 2018, 2020). These results highlight the pervasiveness of pathogen-related selective pressures during human evolutionary history, and how they have contributed to shape present-day human genetic diversity (Barreiro and Quintana-Murci, 2010; Quintana-Murci, 2019; Rotival and Quintana-Murci, 2020). Together, the impact of evolutionary forces on the genetic

architectures of complex disease, the evidence of Neandertal influences on modern human adaptation and the role of viruses as drivers of natural selection, support the relevance of evolutionary genetics approaches in the dissection of genetic effects on COVID-19 susceptibility and severity.

Yet, common genetic variation only explains a fraction of the inter-individual variability in COVID-19 risks. Indeed, since relatively early on in the pandemic, it was shown that male sex (Takahashi et al., 2020) and advanced age (O'Driscoll et al., 2021) were the largest predictors of severe COVID-19 forms. Remarkably, these studies also highlighted links between these host intrinsic factors and differences in the proportions of specific immune cell types, that could in turn explain variation in disease courses. For instance, Takahashi et al. (2020) associated severe COVID-19 risk to poor $CD8^+$ T cell responses in biological males, but high myeloid cytokine plasma concentrations in females. Notably, it was later shown that changes in immune cell type proportions could also translate the effects of extrinsic environmental exposures linked to COVID-19 risk. For example, latent infection by cytomegalovirus (CMV) is associated to increased frequencies of cytotoxic $CD8^+$ T cell subsets (Patin et al., 2018), as well as to the risk of severe COVID-19 forms, even in patients under 60 years old (Weber et al., 2022).

These results underline a more general trend in immune responses to SARS-CoV-2. Namely, while efficient cytotoxic immune responses protect against severe COVID-19, exacerbated myeloid activity can steer the course of the disease towards critical forms. In particular, severe COVID-19 has been associated to changes in cytotoxic lymphoid T and natural killer (NK) cell compartments, as well as T and NK cell exhaustion (Diao et al., 2020; Xu et al., 2020; Wilk et al., 2020; Stephenson et al., 2021; Lee et al., 2020). On the other hand, changes in the monocyte compartment can trigger inflammatory cytokine 'storms' that characterize dysregulated immune responses in the plasma of severe COVID-19 patients (Ren et al., 2021; Stephenson et al., 2021).

The current literature suggests that the exacerbated inflammatory features of severe COVID-19 cases stem from problems in the interferon-mediated regulation of the response to SARS-CoV-2, due to changes in cellular composition (Lee et al., 2020; Wilk et al., 2020; Ren et al., 2021; Stephenson et al., 2021), but also to the joint impact of low-effect common variants genome-wide (COVID-19 Host Genetics Initiative, 2020, 2021, 2022, 2023). Moreover, the study of rare and strongly deleterious inborn errors of interferon immunity (Ciancanelli et al., 2015; Hernández et al., 2018; Lim et al., 2019; Zhang et al., 2020) and their auto-immune phenocopies (Bastard et al., 2020, 2021a,b) has also shed light on the biology of COVID-19 risk disparities (Zhang et al., 2022), in particular through the discovery interferon-neutralizing auto-antibodies that phenocopy the inborn deficiency, but are much more prevalent in the general population, can explain up to 20% of life-threatening COVID-19 cases among the elderly, but also variable responses to vaccination (Bastard et al., 2020, 2021a,b). Hence, even though inborn errors of immunity are rare by definition, their study offers major opportunities to dissect the gene regulatory networks underlying disease traits, which can lead to insights with population-level impacts (Casanova and Abel, 2022).

In summary, the data generated in response to the COVID-19 pandemic present unparalleled opportunities to learn about the host intrinsic and extrinsic predictors of immune variability in the response to viruses in health and disease. These insights have been used to inform and adapt publich health policy during the ongoing pandemic (O'Driscoll et al., 2021), but they should also be leveraged towards the establishment of a precision medicine—tailored to each population and its environment—in preparation for the pandemics to come (Jones et al., 2013). Towards this goal, systematic assessments of the respective contributions of the environmental and genetic predictors of immune variability across human populations and immune cell types are essential, and should be performed through the lens of evolutionary genetics (Quintana-Murci, 2019) to account for the impact of evolutionary forces on the genetic architecture of complex diseases.

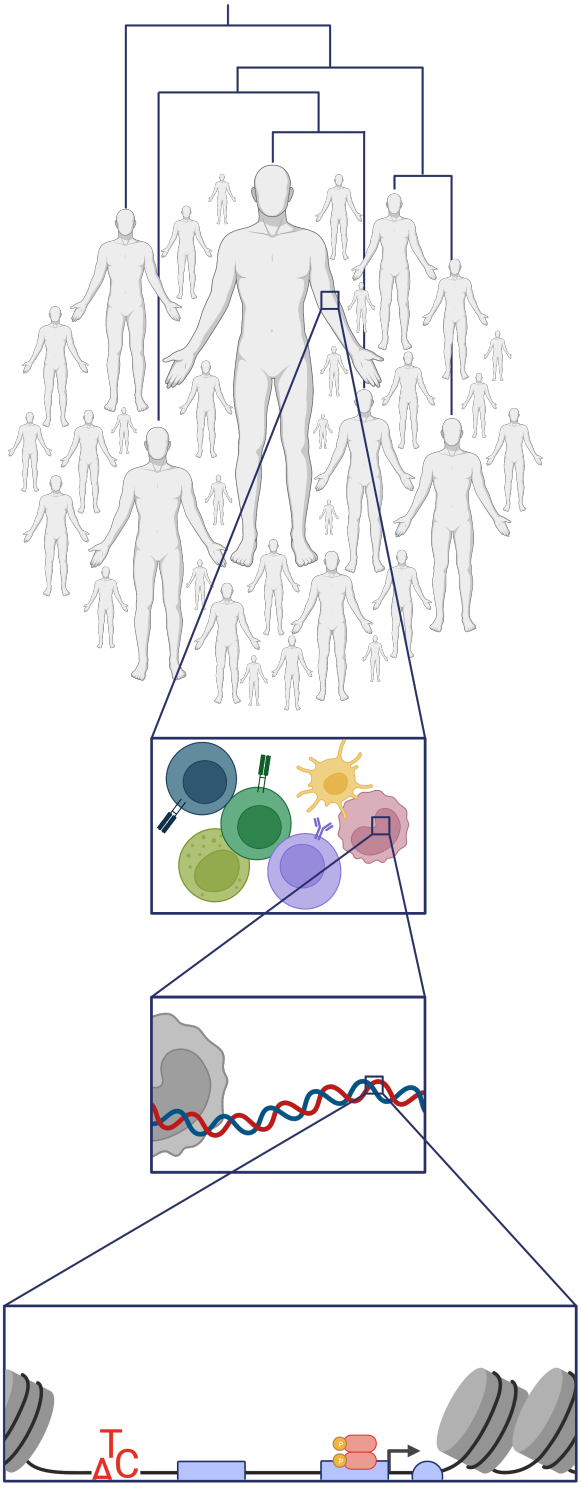
Part I

State of the art

1 Human genetics and single-cell genomics

'The most significant thing about the nucleic acids is that we don't know what they do'

– James Watson, according to Francis Crick (1958)



1.1 From individual genomes to the human pangenome

The 20$^{\text{th}}$ century was marked by several milestone discoveries in genetics that ultimately led to the first observation of the sequence of the human genome. This series of discoveries was sparked by the identification of chromosomes and DNA as the respective cellular and molecular media of genetic information. On this basis, the elucidation of DNA's double-helix structure provided further insights into the biological mechanisms behind heredity, thus spurring several successful efforts to ascertain genomic sequences throughout the tree of life. Yet, the length and complexity of the human genome precluded its sequencing. Even then, in the pre-genomic era of the 20$^{\text{th}}$ century, it was clear that establishing a complete human genome sequence would require a collective and concerted effort of unprecented magnitude in the life sciences. However, the potential benefits to be drawn from a completely sequenced human genome were also abundantly clear. Hence, the Human Genome Project (HGP) was formally launched in October 1990.

The turn of the 21$^{\text{st}}$ century saw the culmination of the HGP. For over a decade, hundreds of researchers across the world toiled towards a common goal: assembling a sequence of the nuclear human genome. The first draft of the assembly was published in 2001 (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Two years later—and precisely two decades ago—the draft was augmented to reach 99.99% accuracy across 99% of euchromatic bases [a], yielding the first human genome sequence (International Human Genome Sequencing Consortium, 2004).

1.1.1 Features and limitations of a composite human genome reference

The completion of the HGP catalyzed our understanding of the mechanisms behind the flow of genetic information (International Human Genome Sequencing Consortium, 2001). Among other key insights, it provided the first concrete observations of the distribution of coding genomic features, as well as regulatory ones (e.g., CpG islands). It also enabled the first measures of genomic quantities such as mutation and recombination rates, and allowed researchers to conduct genome-wide comparative and phylogenetic studies between humans and other species with sequenced genomes.

One important characteristic of this assembly is that it does not represent the genome of a given individual. Instead, it is a mosaic sequence pieced together from fragments of the genomes of several individuals (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). As such, it provided one of the first comprehensive assessments of genetic variation in humans, with around 1.4 million detected single nucleotide polymorphisms (SNPs) (International Human Genome Sequencing Consortium, 2001).

All in all, the HGP revolutionized the study of human genetics and paved the way towards genome-wide studies in humans. For example, it helped the identification of disease genes and drug targets by combining the *in silico* and *in vitro* screening of relevant candidates. Thus, access to a complete human genome sequence reduced the time to pinpoint putative causal genetic loci from years (i.e., with genetic positional cloning) to months (International Human Genome Sequencing Consortium, 2001).

1.1.2 Assessing human genetic diversity with next-generation sequencing methods

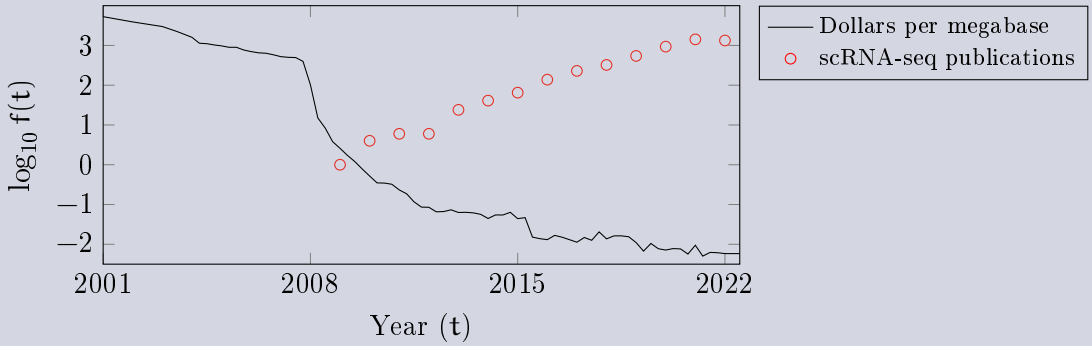
Following the final release of the HGP's reference genome, the International HapMap Consortium set out to build a 'haplotype map' of the human genome, leveraging genetic data from over two hundred individuals originating from four different populations (International HapMap Consortium, 2005). The HapMap Consortium used array-based methods to infer genotypes at a given set of polymorphic loci, thus providing an extension to the HGP's reference by focusing on sequences

a. Excluding regions of heterochromatin at all centromeres, acrocentric chromosome arms, the distal half of chromosome Y and secondary constrictions immediately adjacent to centromeres.

that vary across individuals. These results allowed to confirm the generality of many observations made on the human genome assembly, such as the distribution of 'hotspots' of recombination, and provided a basis for approaches that bypass resequencing by leveraging linkage disequilibrium (LD) patterns—that is, genotype associations between variant loci—to impute information. However, the authors also concluded that whole-genome sequencing (WGS) remained nonetheless critical to gain a finer view of LD structure in the human genome, as well as to identify rare and structural variation across individuals and populations (International HapMap Consortium, 2005).

Box 1 | Evolution of sequencing costs through the years. The entire field of genomics relies on our ability to sequence nucleic acid polymers such as DNA and RNA. Hence, the development of methods to address novel questions through genomics is tightly linked to the evolution of sequencing performance. As faster and cheaper sequencing technologies appear, new genomics methods become accessible.

As shown by the black line in the Figure below, the decrease in sequencing costs over the last two decades has been more than exponential. According to the National Human Genome Research Institute (NHGRI) (Wetterstrand, 2022), sequencing a megabase of DNA cost around five thousand dollars near the time of the initial release from the International Human Genome Sequencing Consortium (2001). In 2008, the NHGRI's sequencing centers shifted from Sanger sequencing to next-generation sequencing (NGS) technologies, leading to a steep drop in sequencing costs. As newer NGS options appeared, cost decreases accelerated, until around 2015 when they plateaued at their present value of approximately a cent of a dollar per megabase of DNA.



The red dots in the Figure show the number of published research pieces applying single-cell RNA-sequencing (scRNA-seq; § 1.3.1, page 26) and reported in PubMed each year, showcasing how the drop in sequencing costs transformed genomics research by enabling the development of novel assays (Tang et al., 2009).

The advent of next-generation sequencing (NGS) methods—pioneered by 454 Life Sciences' pyrosequencing platforms, followed by Illumina's solutions—enabled large-scale efforts aiming to build cohorts of sequenced genomes from different human populations world-wide (Box 1). In this context, the 2010 study by the 1000 Genomes (1KG) Project Consortium (1000 Genomes Project Consortium et al., 2010) is specially important, as it underlined the first large-scale effort to characterize human genetic variation through WGS. After this pilot study, the following installment of the 1KG Project resulted in a description of human genetic variation from over 2.5 thousand donors sampled across 26 populations world-wide (1000 Genomes Project Consortium et al., 2015). Since then, the cohort has been expanded with the addition of nearly 700 related donors (Byrska-Bishop et al., 2022).

Extensive catalogs of human genetic diversity like the 1KG Project are essential for dissecting the genetic and nongenetic drivers of phenotypic variation across individuals from different ancestries. Yet, some researchers have raised concerns about potential reductionism entailing from the use of such cohorts (Mathieson and Scally, 2020; Coop, 2022). Humans rarely live in discrete and perennial clusters, and the set of factors that link individuals through ancestry within these groups are highly complex. It has long been established that the lion's share of human genetic diversity lies within these groups, with between-population differences explaining only as much as 15% of genetic variation (Lewontin, 1972). In the context of genetic modelling, 'genetic populations' become approximations; as in any statistical modelling framework, approximations are bound to lead to bias. Thus, some of these authors argue for precise—albeit not concise—labeling of human groups with different genetic, environmental and social backgrounds, so as to avoid confounded comparisons.

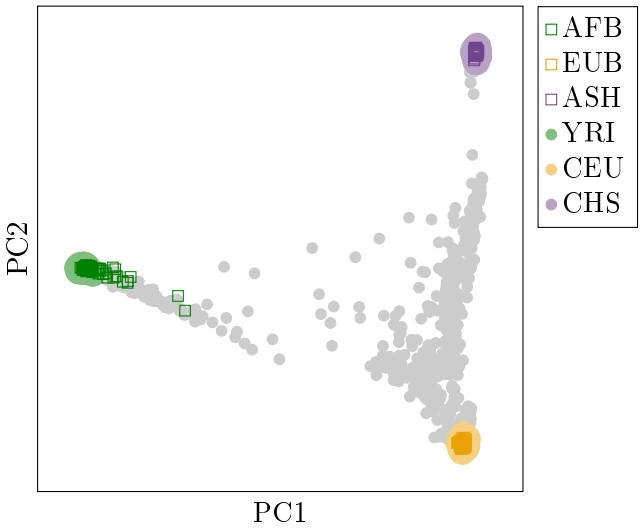


Figure 1.1 | Principal components (PC) analysis of genotypes from diverse cohorts. The square points represent the genotypes of 160 individuals of Central African (AFB) or West European (EUB) descent from the EvoImmunoPop cohort (Quach et al., 2016), as well as an additional 71 individuals of East Asian descent (ASH) (Aquino et al., 2023). The circular colored points represent the genotypes of individuals in each of the 'Yoruba in Ibadan' (YRI), 'Utah residents with Northern and Western European ancestry from the CEPH (Centre d'Étude du Polymorphisme Humain) collection' (CEU) and 'Han Chinese South, China' (CHS) panels of the 1000 Genomes Project Consortium's cohort. The circular gray points represent the genotypes of individuals in all other panels.

For illustration, Figure 1.1 shows the two dimensions that maximize genetic distances between around 3 thousand donors of different ancestries, as determined by principal components analysis (PCA) of genotyping data across over 370 thousand SNP loci. Most of these genotypes, represented by circular points, belong to individuals from the latest release of the 1KG Project Consortium's reference panels. The square points show the genotypes of individuals of Central African (AFB), West European (EUB) or East Asian (ASH) descent from two unrelated cohorts. Because they appear genetically closer to individuals from the 'Yoruba in Ibadan' (YRI) group than to any other panel, it is tempting to assign AFB individuals an 'African ancestry' label. However, it would be wrong to assume that the YRI panel alone can stand as a comprehensive representation of human diversity in Africa. Thus, according to Coop (2022), such statements about 'genetic ancestry groups' should be replaced with factual descriptions of observed genetic similarity. In the present example, AFB individuals would be more accurately described as being 'most genetically similar, according to Euclidean distances on the first two principal components of available genotyping data, to the individuals in the YRI panel from the 1KG Project Consortium's cohort'.

All in all, the partitioning of human genetic diversity is a non-trivial issue, and still the subject of lively discussion. When using 'genetic population' constructs for the purposes of study design, researchers should favor preciseness over conciseness in their descriptions of these groups. Given the societal undertones that are often attached to scientific studies of population differences, precise descriptions of the comparisons being carried out are essential to limit the mainstream spread of misinformation. Furthermore, a fair assessment of the generality of the findings of such studies is only possible if comparisons are well characterized.

Another related problem that limits the generality of observations from these catalogs of genetic diversity is that virtually all published genome-wide analyses in humans are based on a single reference. Over the years since its first publication, the human reference genome has been updated to newer versions. Yet, two limitations remain. First, a substantial portion of the reference is still not sequenced: 6.7% of the primary chromosome scaffolds in the Genome Reference Consortium's (GRC) most recent release of the human genome, GRCh38.p14 (Liao et al., 2023). Second, even though it was partly pieced together from the genomes of different individuals, the composite sequence produced by the HGP paints an incomplete picture of human genetic variation. Because of the relationships that link genetic variants to one another—expressed through LD—a true depiction of human genetic diversity requires integrating information from several genomes.

To tackle the first of these challenges, the Telomere-to-Telomere (T2T) Consortium leveraged novel long-read NGS methods to sequence all molecules in the human nuclear genome from one end to the other, resulting in a gapless assembly of all autosomes and the X gonosome (Nurk et al., 2022). The T2T-CHM13 haplotype is the most extensive reference of the human genome ever generated, and an absolutely remarkable achievement in the history of human genetics. By filling-in missing and simulated sequences in the prior references, T2T-CHM13 lifts the veil on these complex genomic regions, enabling their study and uncovering previously unknown genetic polymorphisms. However, even this comprehensive resource is unable to accurately represent human genetic diversity.

In fact, no individual genome sequence can appropriately capture this variability. Both the GRCh38 and the T2T-CHM13 references are characterized by the slanted representation of particular human groups with respect to the frequency with which they occur world-wide. Specifically, while European-origin individuals make up around 16% of the global population, well over 90% of the sequences in T2T-CHM13 have a predicted European origin (Nurk et al., 2022). This slant translates a systemic bias in human genomics: as of 2018, 52% of studies seeking the genetic basis of complex traits, including diseases, were performed in European populations. Overall, 78% of donors recruited for these studies were of European descent (Sirugo et al., 2019).

European-based references are likely to bias discovery towards genetic variants that are frequent in European genomes, or whose effects appear in a European background. Yet, there is no guarantee as to the generalizability of these results to other populations that live in different environments. Furthermore, such studies are underpowered to find loci with small effects in Europeans, but that may considerably impact phenotype in other genetic and environmental backgrounds. Hence, the use of individual genome references induces biases both scientifical and ethical. These studies output angled results, and they do not benefit all individuals equally.

Addressing these biases requires a new paradigm in human genomics: one which entails an overhaul of the topology of the reference itself. Namely, while a reference genome is a linear sequence of nucleotide monomers, a so-called 'pangenome' is a graph that integrates information from multiple genomes into a unique reference.

In pangenome graphs, nodes are sub-sequences of nucleotides and edges show the possible ways in which these components can be concatenated to compose an entire genome. The different haplotypes underlying the pangenome can be recovered by 'walking along' the graph. Thus, the pangenome

becomes an apt framework to express complex genomic variation. In fact, pangenomes have been used for decades to represent the genetic diversity in prokaryotic clades, and to distinguish sequences that are shared by all or most individuals from those that are private to some.

Efforts to build the pangenomes of species with longer and more intricate genomes have only recently become feasible, owing in part to the advent of long-read NGS methods such as those used by the T2T Consortium. These long reads are especially useful to resolve complex assemblies, covering regions rich in repeated sequences, for example. Recently, the first eukaryotic pangenomes were built for major crop species, such as soybean, wheat and corn; the construction of pangenomes to help the breeding of domestic animals and cattle is also being explored.

1.1.3 Features and limitations of a fully representative human pangenome

Earlier this year, the Human Pangenome Reference Consortium (HPRC) officially released a first draft of the human pangenome, composed of fully-phased haplotypes from 47 genetically diverse individuals (Liao et al., 2023). Among other findings, this new reference provides a more complete view of structural variants (i.e., insertion/deletion events longer that 51 base pairs) and copy-number variants, two types of genomic variation that are notoriously hard to capture with short-read sequencing methods. In particular, the HPRC reports that 71% of copy-number-variant genes with respect to the GRCh38 reference were private to a single haplotype (Liao et al., 2023). This illustrates well the relevance of a reference that integrates information from multiple genomes.

Moreover, the advantages of the human pangenome are not limited to structural and copy-number variant detection. After aligning the short-read WGS data from the extended 1KG Project Consortium's cohort (Byrska-Bishop et al., 2022) to their reference, Liao et al. (2023) were able to detect over 60 thousand new variants in each sample. Importantly, some of these novel variants were found in medically relevant, but previously unresolved, genomic regions. Thus, the human pangenome opens new genomic windows in which to look for disease-associated loci and putative therapeutic targets. Furthermore, it is also likely to have a substantial positive impact on transcriptomic studies, by improving mRNA-derived read mapping in these previously unresolved windows.

The HPRC aims to increase the number of genomes in their reference up to 350 by next year. Although the potential benefits of this new framework of reference have been very clearly outlined by Liao et al. (2023), and have borne fruit in other species, they have yet to be translated into practical human genetics research. As is the case with most paradigm shifts, incorporating the human pangenome into everyday practice will require many satellite changes to well-established workflows and analytical pipelines. For example, although there is a growing consensus around pangenome graphs, the field is not yet entirely set on the best way to visualize this new reference. More generally, almost all aspects of human pangenome construction, analysis and representation are active research areas (Sirén and Paten, 2022; Hickey et al., 2023; Garrison et al., 2023). Some authors believe that a full transition will not be complete before a decade (Eisenstein, 2023).

1.2 Genome-wide associations and causal effect inference

The HGP revealed the complexity of the human genomic landscape and provided a reference on which to locate genetic elements. Yet, actually mapping the positions of loci associated to traits of interest requires associating particular expressions of phenotype to combinations of genotype. In this context, catalogs of human genetic variation—such as the HapMap Consortium's genotyping data, the 1KG Project Consortium's WGS data or the human pangenome—are very useful to dissect the sources of phenotypic variability across individuals. Today, the diversity captured by these resources is widely availed of to establish statistical links between genotype and disease-associated traits through genome-wide association studies (GWASs).

1.2.1 Genome-wide association studies to link genotype and disease traits

Previous to the appeareance of GWASs, genome-wide approaches to identify variants associated to disease susceptibility were mainly limited to family-based linkage mapping or candidate gene studies. These methods seek genetic regions that are linked to disease traits within pedigrees more often than would be expected by chance, and that are thus likely to contain causal variants of the disease (Hirschhorn and Daly, 2005). However, while linkage mapping is powerful to detect genetic associations with monogenic 'Mendelian' disorders (Gusella et al., 1983; White et al., 1985), it generally falls short when used to search for variants associated to complex diseases. For example, while linkage mapping successfully revealed variants in the *NOD2* locus associated to the susceptibility to Crohn's disease (i.e., a component of inflammatory bowel disease, IBD) (Hugot et al., 2001; Ogura et al., 2001), it was later estimated that these variants jointly explain but a fraction of the excess risk of IBD relative to siblings, suggesting unidentified genetic factors (Daly and Rioux, 2004).

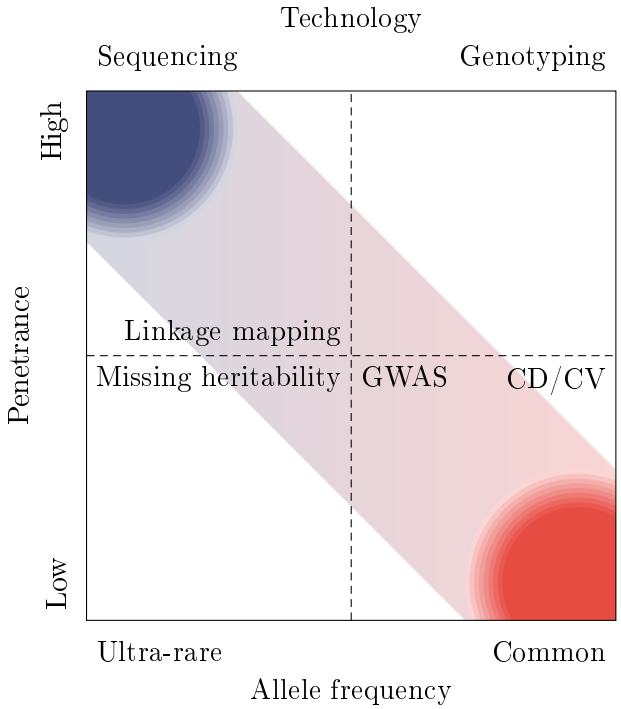


Figure 1.2 | The genetic architecture of Mendelian and complex diseases. The 'common disease/common variant' (CD/CV) hypothesis predicts that the genetic component of most frequent diseases is made up of relatively common low-effect genetic variants. Genome-wide association studies (GWASs) are most useful to identify these variants. In contrast, linkage mapping and candidate gene studies are most useful to find relatively rare albeit highly penetrant disease-associated variants. Yet, variants identified by both of these approaches are unable to completely account for the heritability of most diseases. A part of this 'missing heritability' may be explained by rare variants not included in most commercial genotyping arrays, by an over-estimation of trait heritability due to gene-by-environment interactions in twin studies, or by types of variants that are under-represented in current catalogs of genetic variation, such as structural variants. Adapted from Manolio et al. (2009).

The loss of power of linkage mapping in complex contexts can be explained by differences in the genetic architectures of monogenic and complex diseases. Genetic variants behind Mendelian diseases are by definition highly penetrant and often rare, likely due to natural selection purging them from the population. Thus, Mendelian-disease-causing variants are expected to be tightly linked with disease status within each family. In contrast, diseases with complex architectures can be explained by several genetic and nongenetic factors. Although the joint effect of these genetic factors on phenotype can be substantial, each individual variant is expected to have a marginally low effect. Thus, even though these variants are associated to disease traits, they can be relatively common in the population.

To illustrate the differences between the genetic architectures of complex and Mendelian diseases, Figure 1.2 shows a simplified schematic representation of the relationship between population allele frequencies and the magnitude of allele effects on disease traits. Together, the frequency and effect size of a risk allele determine its contribution to the total heritability (Appendix A, page 190) of a disease trait (Appendix D, page 205) (§ 2.3.3, page 50). In this context, Figure 1.2 also illustrates the 'common disease/common variant' (CD/CV) hypothesis, which posits that most frequent disorders are at least partly due to the joint impact of many low-effect genetic variants found in over 1 to 5% of individuals in a population; although deleterious, these alleles would have risen to such high frequencies following the rapid expansion of humans from small founder populations around 15 to 18 thousand years ago (Reich and Lander, 2001).

On the basis of the CD/CV hypothesis, and owing to the development of dense genotyping arrays and of resources of human genetic diversity like the HapMap Consortium's, GWASs emerged as powerful tools to find the genetic determinants of complex diseases.

1.2.2 Statistical models behind genotype-phenotype-association testing

In their most basic implementation, GWASs are performed using linear models of the form

$$\mathbf{Y} = \alpha_{\mathbf{Y}} + \beta_{\mathbf{Y}\mathfrak{i}}\mathbf{G}_{\mathfrak{i}} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{Y}$ is a vector of quantitative trait values across individuals, vector $\mathbf{G}_{\mathfrak{i}}$ encodes genotypes at polymorphic locus $\mathfrak{i}$ along the genome, $\alpha_{\mathbf{Y}}$ is an intercept, $\beta_{\mathbf{Y}\mathfrak{i}}$ captures the effect of genotype on phenotype and $\boldsymbol{\varepsilon}$ is a term for independent and identically distributed noise. The significance of the genotype-phenotype association is tested under the null hypothesis $\mathcal{H}_0 : \beta_{\mathbf{Y}\mathfrak{i}} = 0$. Rejection of the null pinpoints variant $\mathfrak{i}$ as a quantitative trait locus (QTL) at family-wise type I error rate α.

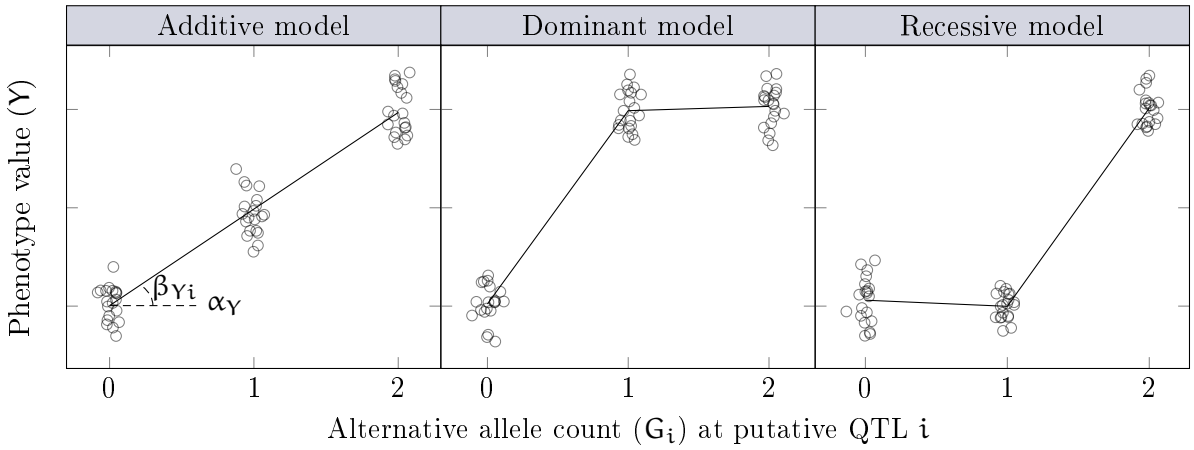


Figure 1.3 | Additive, dominant and recessive models of genotype-phenotype association. The dots show values of phenotype Y for different levels of genotype G_i, encoding the number of alternative alleles found at locus i in each individual, for three models that reflect different functional relationships between the alternative and reference alleles. The lines connect the mean values of Y conditional on G_i in each model. Under the additive model, α_Y is equal to the mean of Y across homozygotes for the reference allele, and β_{Yi} is equal to the increase in Y for each copy of the alternative allele.

The encoding of genotypes in G_i depends on the functional relationship between the alleles at putative QTL $\mathfrak{i}$. Most commonly, an additive model is used in which $\beta_{\mathbf{Y}\mathfrak{i}}$ captures the change in Y for each copy of the alternative allele. Alternatively, dominant or recessive models can be used. Figure 1.3 illustrates these different models using a simulated set of observations of phenotype Y across 60 individuals homogeneously stratified across levels of genotype G_i.

1.2.3 Features and limitations of genome-wide association studies

The first GWAS ever published used a case-control study design to screen for genetic variants associated to age-related macular degeneration (AMD) in a cohort of 96 cases and 50 controls (Klein et al., 2005). Ironically, although the GWAS approach was devised to find frequent variants with small effects on phenotype, Klein et al. (2005) identified several large-effect and common variants associated to AMD that are not predicted by the CD/CV hypothesis (Fig. 1.2).

Picking up on this budding trend, the Wellcome Trust Case Control Consortium (WTCCC) used a cohort of approximately 14 thousand cases and 3 thousand controls to identify genetic variants associated to the susceptibility to seven major common diseases, including IBD, type I and type II diabetes (The Wellcome Trust Case Control Consortium, 2007), and bipolar disorder. By working with thousands of cases and controls, and across different diseases, the authors of the WTCCC aimed to provide a general framework of recommendations for performing GWASs. In line with the CD/CV hypothesis, this landmark study identified several novel associations between each disease and common moderate-effect variants (i.e., study-wide minor allele frequency over 1%).

Missing heritability. Since then, thousands of GWASs have identified hundreds of thousands of associations between genetic variants and diverse traits. Yet, the effect sizes of most of these variants on disease risk are limited—with estimated odds-ratios (OR) between 1.2 and 1.5—and they cannot fully explain the segregation patterns of the corresponding diseases in pedigrees [b] (Manolio et al., 2009). Taken together, these observations point to missing components of complex disease heritability that elude detection by GWASs or linkage mapping studies (Fig. 1.2).

One possible explanation of the 'missing heritability' conundrum is an over-estimation of trait heritability due to unadjusted gene-by-environment interactions in family-based studies. The height phenotype has been of historical interest in this context (Appendix A, page 190). Although the genetic component of height has been acknowledged for well over a century (Galton, 1886), interest in the environmental predictors of adult height is only relatively recent (Jelenkovic et al., 2016). Importantly, neglecting the effect of a shared environment on variation in height—or any other heritable trait—among relatives can inflate the estimated heritability of the trait.

Another potential source of heritability may be missed through the unmeasured effect of rare variants—with a population minor allele frequency (MAF) below 1%—that are not included in the design of most genotyping arrays commonly used in GWASs. Indeed, owing to the LD structure of the human genome, and depending on each particular population, most SNP-related common (MAF $> 5\%$) haplotypes can be characterized by genotyping approximately 500 thousand variants (International HapMap Consortium, 2005). Hence, most commercial arrays are limited to these sets of 'tag' SNPs, which do not include rare variants. Importantly, however, rare variants not detected by these methods could have substantial effects on disease risk and at least partially explain missing heritability. In this context, the advent of NGS methods (Box 1) enabled WGS efforts such as those led by the 1KG Project Consortium, to capture rare genetic variability.

Structural variants represent another possible source of heritability that is currently neglected. As mentioned previously, this type of variation is known to be under-represented in the reference genomes currently used by the community, but advances such as the T2T-CHM13 haplotype and the human pangenome, that leverage long-read NGS methods and are able to capture this diversity, are closing the gap. Recently, long-read NGS methods were used by Ebert et al. (2021) from the Human Genome Structural Variation Consortium (HGSVC) to assemble haplotypes from diverse individuals and detected over 100 thousand new structural variants.

b. Except for some particularly well-studied traits, like height, for which saturated maps of common genetic variants that contribute to heritability have been built using heritability estimation methods (Yang et al., 2010, 2011) able to leverage genetic information across millions of unrelated individuals (Yengo et al., 2018, 2022).

The human pangenome will also allow to tackle another possible explanation of missing heritability: the biased representation of some human populations in genetic data bases. Indeed, while most GWASs performed to date have used cohorts of mainly European ancestry, it is known that individuals of African ancestry carry the largest share of human genetic diversity. Thus, it is expected that including such under-represented groups with diverse genetic backgrounds will better our grasp on human genetic diversity. In fact, using WGS data from over 400 individuals from different African ethnic groups, Choudhury et al. (2020) uncovered millions of novel genetic variants. Importantly, all of these nominally healthy individuals carried at least one variant anotated as 'pathogenic' in the National Center for Biotechnology Information's ClinVar data base, highlighting how genetic effects observed in a predominantly European cohort may not translate to other populations. ■

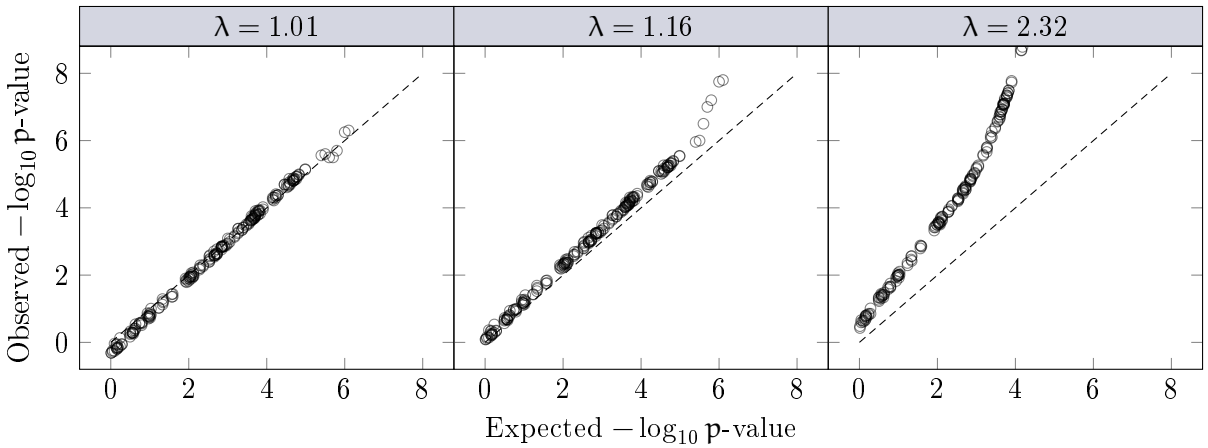


Figure 1.4 | Quantile-quantile plots to visualize genomic inflation of positive association tests. The x and y axes respectively show a measure of the expected and observed statistical significance of associations between a phenotype and the genotype at variant loci genome-wide. The dashed line is the subspace in which these two metrics are equal. Early deviations from this line reflect genomic inflation of positive associations.

Multiple-testing burden. Modern GWASs involve tests for association between a phenotype and upwards of hundreds of thousands of variant loci. Therefore, multiple-test correction of raw association p-values is essential. Most often, the general Bonferroni correction is used to control the family-wise type I error rate α. Assuming that the human genome carries around 1 million independent haplotypes, the widely accepted 'genome-wide significance' threshold for GWASs in humans is set to $5 \cdot 10^{-8}$ for an initial $\alpha = 0.05$. However, this may lead to an overly conservative correction if the independence assumption does not hold (Risch and Merikangas, 1996).

A complementary method to control the rate of false positive associations is provided by the genomic inflation factor λ (Devlin and Roeder, 1999). Briefly, λ quantifies the ratio between the median strength of the observed genotype-phenotype associations and the median strength expected if all variant loci were unlinked. Therefore, a strongly positive λ reflects bias induced by confounding factors such as genotype associations omitted in the model. ■

Population stratification. Although Equation (1.1) is very useful to estimate expected mean values of phenotype in a population, conditional on disease-status and genotype, other differences between these cases and controls might confound the association if left unaccounted for in the model. Thus, the basic linear model is often complemented with a matrix $\mathbf{Z}$ of covariates and associated vector $\boldsymbol{\gamma}$ of coefficients that capture the effects of known confounders,

$$\mathbf{Y} = \alpha_Y + \beta_{Yi}\mathbf{G}_i + \mathbf{Z}^T\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{1.2}$$

Commonly adjusted-for covariates include age and sex, as well as metrics of genetic similarity between individuals. Indeed, population stratification within the cohorts of cases and controls may also lead to spurious associations between genotype and phenotype.

For example, if population stratification exists such that populations A and B are respectively over-represented among cases and controls, then frequent variants in population A may falsely appear as associated to disease risk. In this context, principal component coordinates such as those shown in Figure 1.1 are frequently used as a proxy for genetic distance between individuals to account for population stratification in the model.

Because unadjusted population stratification creates artificial associations between genotype and phenotype, it also leads to genomic inflation. Using genome-wide association results simulated under varying degrees of population stratification, Figure 1.4 illustrates how so-called 'quantile-quantile' or 'Q-Q' plots that relate the observed and expected statistical significance of the associations are useful tools to visualize genomic inflation. ■

Phenotype definition. Mapping QTLs as per Equation (1.2) requires encoding the trait of interest as a quantitative variable $\mathbf{Y}$. For some phenotypes, such as measures of disease risk, this is relatively straightforward. Other phenotypes, however, need to be transformed or approximated in order to fit the GWAS framework; the final definition of the trait can translate prior beliefs from the authors into biases of the study.

For example, in their attempt to map the genetics of same-sex sexual behavior, Ganna et al. (2019) labelled as 'non-heterosexual' all participants who had ever had a same-sex partner. Yet, reducing this complex social behavior—and with it, the spectrum of human sexuality—down to a binary variable is likely to yield a biased answer to the original question. Indeed, such a simple criterion can confound associations by grouping together individuals with widely different sexual behaviors—including predominantly heterosexual ones. ■

Correlation-causation phallacy. When mapping the genetic basis of a trait through GWAS, the hope is to find causal links between genotype and phenotype. Under the CD/CV hypothesis, many common genetic variants with small marginal effects are expected to participate to variation in a given complex trait. Although the impact of each single variant is weak, their joint effect on phenotype can be substantial. From this point of view, the polygenic risk score (PRS) has been proposed as a tool to summarize these effects into a single composite metric of the genetic contribution to complex phenotypic variation (Wray et al., 2007).

For example, schizophrenia is a complex and severe psychiatric disorder with a largely known genetic component; its heritability has been estimated at up to 80% based on twin studies (Cardno and Gottesman, 2000; Sullivan et al., 2003). The first GWAS on schizophrenia was published in 2009, using a cohort of over 3 thousand cases and more than 3 thousand controls of European ancestry (The International Schizophrenia Consortium, 2009). This landmark study uncovered thousands of associations genome-wide, albeit with very few strong effects. The authors then proposed different PRSs for the innate risk of developing schizophrenia, as linear combinations of several tens of thousands of SNPs weighted by their estimated impact on disease. Although the authors' rationale was sound—pooling information across many variants to overcome limited power to detect small effects—their metric was only able to explain around 3% of variation in case status in their sample.

Around that time, advances in NGS methods spurred a decrease in sequencing costs (Box 1) that ultimately enabled larger studies with increased power to detect small effects (Zheutlin and Ross, 2018). In 2014, a schizophrenia GWAS used a cohort of around 150 thousand mostly European donors to derive a new PRS explaining around 7% of variance in case status (Pantelis et al., 2014).

Yet, even though increasingly larger data sets can improve the predictive power of PRSs for complex traits, cryptic population stratification can complicate the establishment of causal links between genotype and phenotype. For instance, Curtis (2018) found stark population differences in the schizophrenia PRS proposed by Pantelis et al. (2014), using multi-ancestry genotyping data from the HapMap Consortium (International HapMap 3 Consortium et al., 2010). Interestingly, the average difference in PRS between HapMap Europeans and Africans was around ten-fold larger than the mean difference between European schizophrenia cases and controls, suggesting that the PRS may be picking up on ancestry-related correlations, possibly because it was derived from a mainly European cohort.

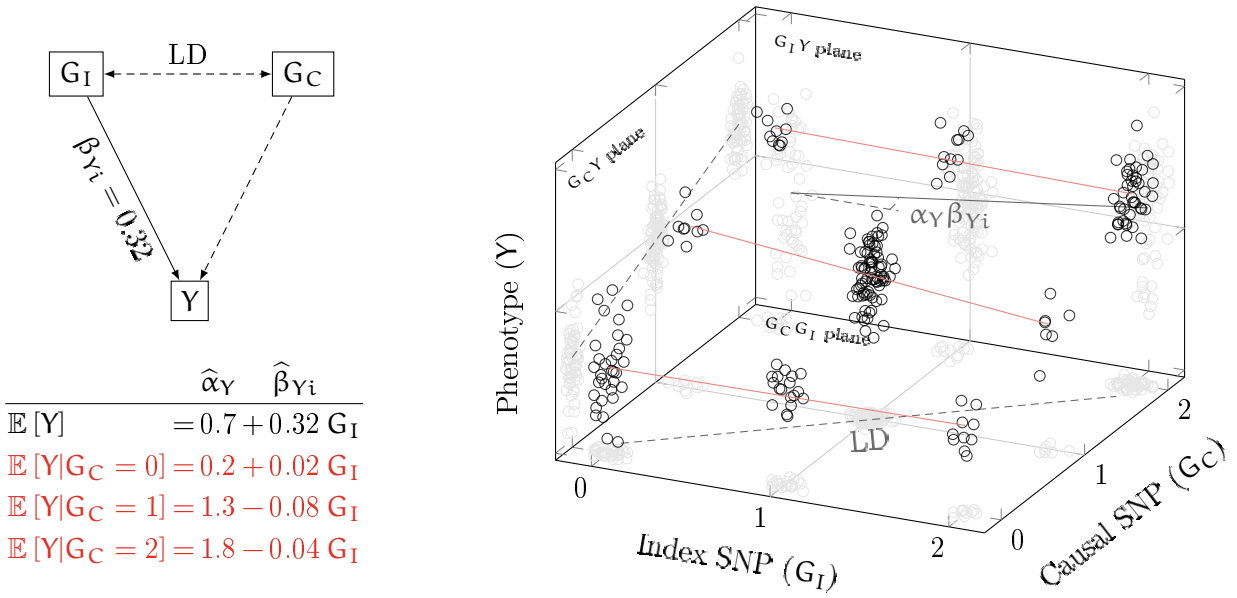


Figure 1.5 | Conditional independence between phenotype and index genotypes. In a genome-wide association study (GWAS), an association between a phenotype Y and the genotype G_I at an index single nucleotide polymorphism (SNP) locus may appear, even when there is no real causal relationship between the two, through the unmeasured effect of a causal SNP G_C that is associated to G_I through linkage disequilibrium (LD). On the left-hand side graph, the solid line represents the association between phenotype and index SNP estimated by GWAS; the dashed lines show the associations masked to the study when G_C is not known. The right-hand side scatter plot shows 200 observations across the three variables simulated under a model in which Y and G_I are independent, but both are linked to G_C. The grey dots on the G_IY, G_CY and $G_C G_I$ planes show the joint distribution of observations along the corresponding pair of variables. The regression line of Y on G_I is shown in black; the equation is shown under the left-hand side graph. The regression lines of Y on G_I, conditional on G_C, are shown in red; the corresponding equations are shown under the left-hand side graph.

The interpretation of GWAS results is also complicated by confounding genotype correlations. Because of LD, GWAS hits highlight haplotypes—not particular SNPs—as putatively causal of phenotype differences. The highlighted haplotype may well carry one or more bona fide causal variants, but these might not include the 'index' SNP that tags the haplotype. Yet, the causal variant may act as a confounder, and create a statistical association between phenotype and genotype at the 'tag' SNP, even if the two are conditionally independent.

For example, Figure 1.5 illustrates a fairly simple case of confounded association between an index SNP G_I and a phenotype Y, created by the effect of a real causal variant G_C in LD with G_I, and using simulated data across 200 observations of the three variables with covariance matrix

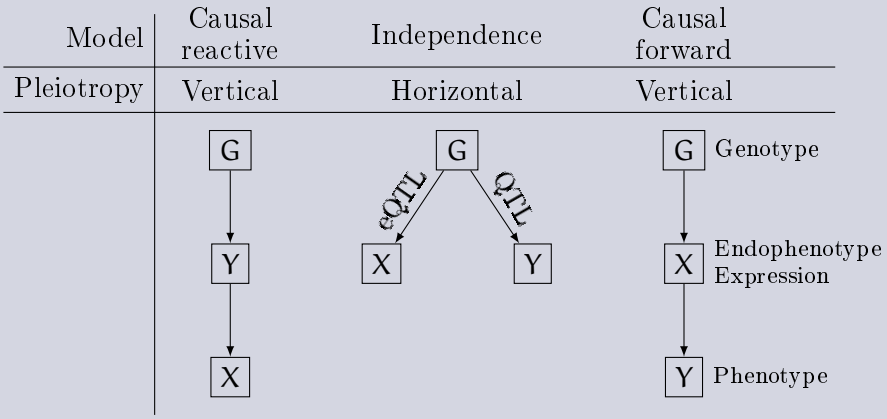
$$\begin{matrix} G_I & G_C & Y \end{matrix} \\ \begin{bmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0.6 \\ 0 & 0.6 & 1 \end{bmatrix} \begin{matrix} G_I \\ G_C \\ Y \end{matrix} . \tag{1.3}$$

The graph on the left shows the relationships between these three variables. The black solid line indicates the effect estimated in a GWAS; the dashed lines indicate latent effects. Interestingly, although Y was explicitly simulated to be independent of G_I, β_{Yi} is significantly larger than zero (Student's two-sided t-test $p = 1.5 \times 10^{-7}$). The slope is shown on the $G_I Y$ plane of the scatter plot on the right. The unmeasured causal effect of G_C on Y—which artificially links G_I to Y through the LD between G_C and G_I—appears on the $G_C Y$ plane. When G_C is revealed to the model, and the strength of the association between G_I and Y is estimated conditional on the genotype at G_C, these coefficients become indistinguishable from zero at a type I error rate of 5% (Student's two-sided t-test $p > 0.13$). The corresponding regression lines are shown in red on the scatter plot.

Box 2 | Genotype, endophenotype and phenotype. The ultimate goal of a genome-wide association study (GWAS) is to map the genetic basis of a complex trait. Yet, causal inference from GWAS results is complicated by several factors, including their associative nature, as well as linkage disequilibrium (LD) between nearby SNPs.

In a GWAS, quantitative trait loci (QTLs) are mapped by regressing the value Y of the focal phenotype on the genotype G at each tested variant locus, as per Equation (1.2). Thus, QTL results are associative; they do not imply any causality between G and Y. Furthermore, true causal links may be obscured by LD in the region.

Whenever changes in phenotype Y can be explained by genetically-controlled changes in a gene expression trait X, functional data can be used to improve the interpretability of GWAS results. In this context, variants associated to the endophenotype X are 'expression' (e) QTLs. A genetic variant that is a QTL of phenotype Y and an eQTL of endophenotype X becomes a likely causal candidate.



Crick's dogma directs the general flow of organic matter and genetic information downstream from DNA (Crick, 1970). On this basis, the association between the genotype G, the gene expression endophenotype X and the focal phenotype Y can take three forms, illustrated in the Figure above.

Each topology is associated to a type of pleiotropy. If phenotype and endophenotype are independent, but an association appears through the confounding effect of genotype G acting on both traits, the variant displays horizontal pleiotropy. Otherwise, vertical pleiotropy arises when a genetically-controlled change in one trait triggers a change in the other. When changes in the expression endophenotype mediate the genetic effects on phenotype, forward causality can be inferred by merging GWAS and eQTL data.

All in all, Figure 1.5 illustrates why even when the variables in Equation (1.2) are properly defined, population stratification is aptly adjusted for, and the false positive rate is controlled, a significant correlation between genotype and phenotype does not necessarily imply that the tested variant has any causal effect. ■

The limitations outlined above, including the confounding effects of nearby SNPs, weaken the power of the GWAS approach to identify causal variants underlying complex traits. As human groups with different genetic backgrounds are characterized by distinct LD patterns, building more diverse cohorts can improve the resolution of the mapping and the interpretability of GWAS results. That is, associations identified in genomes from different populations can be intersected so as to trim putatively causal haplotypes and zero in on the bona fide causal variant.

More generally, genetic 'fine-mapping' methods are used to disentangle the confounding patterns of LD between nearby SNPs, so as to find causal variants—assuming at least one exists—in a genomic region associated to a trait (Schaid et al., 2018). Several approaches to fine-mapping are available, but the common basic idea is to prioritize SNPs based (i) on the strength of their association with the trait and (ii) on the strength of the LD they share with the top-associated 'peak' SNP. Thus, the signal in a genomic region can be refined into linked SNP sets, each acting independently on a trait. These components can then be ornamented with functional information—such as annotations from data bases or gene expression data—to infer which variants are more likely to be causal.

Adding functional data on 'intermediate' endophenotypes associated to genetics and to the focal phenotype can also help infer causal genetic effects (Lappalainen et al., 2013). Even though GWAS loci linked to disease are over-represented among coding sequences—relative to their frequency in genotyping arrays—the fact remains that only a minor fraction of them fall on readily interpretable regions: relatively soon after the GWAS boom, Hindorff et al. (2009) estimated that more than 80% of SNP QTLs discovered across over 150 studies were located in intergenic or intronic regions of the genome. Moreover, common disease-associated variants have been reported to be concentrated in regions that regulate gene expression (Maurano et al., 2012). In this context, gene expression (e) QTLs are commonly used as an aid to the interpretation of GWAS results (Box 2).

1.2.4 Molecular endophenotypes to aid quantitative trait locus interpretation

The heritability of gene expression in humans has long been established using twin models (Powell et al., 2012; Grundberg et al., 2012; Wright et al., 2014), other family-based approaches (Dixon et al., 2007) and population-based methods among unrelated individuals (Price et al., 2011; Lloyd-Jones et al., 2017) (Appendix A, page 190). For instance, Wright et al. (2014) leveraged the patterns of genetic similarity among 13 hundred pairs of monozygotic and dizygotic twins to estimate a mean narrow-sense heritability of around 14% across more than 18 thousand transcripts expressed in peripheral blood, and mapped close to 7 thousand independent eQTLs.

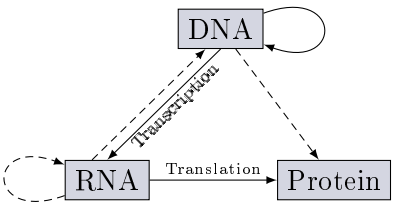


Figure 1.6 | The central dogma of molecular biology. Francis Crick's early interpretation of the flow of genetic information from DNA to protein involved a 'general' stream present in all cells, which could be complemented by 'special' tributaries in particular contexts. The transfer of information from DNA to (messenger) RNA is called 'transcription'; information is then transferred from RNA to protein through 'translation'. Solid and dashed arrows represent general and special transfers, respectively. Adapted from Crick (1970).

In fact, genetic effects on gene expression are expected under Crick's central dogma of molecular biology (Crick, 1958, 1970). In its barest meaning, the dogma is a directed graph of the general flow of matter and information from DNA to messenger (m) RNA through transcription, and from mRNA to protein through translation, as illustrated in Figure 1.6. It follows that changes in one node or edge of the graph can affect the transfer of information to other nodes downstream from it. For instance, variation in the DNA sequence of a gene promoter can change the rate at which mRNA molecules are transcribed by modifying the affinity of a transcription factor (TF) for the altered nucleotide motif. In turn, this may lead to inter-individual differences in the concentration of the protein encoded by that gene.

From this molecular perspective, an eQTL is a genetic variant associated to the abundance of mRNA molecules transcribed from an 'eGene' in a cell or tissue. Depending on their position relative to their target eGene, eQTLs are divided in two classes. In general, proximal eQTLs found on the same chromosome and within a megabase of the transcriptional start site (TSS) of their eGene are termed '*cis*-eQTLs'. Distal eQTLs found outside this window—possibly even on another DNA molecule than the eGene—are called '*trans*-eQTLs'.

Alternative splicing is another genetically-controlled feature of transcription that is often used as an endophenotype to improve the interpretability of GWAS QTLs. From this perspective, splicing (s) QTLs are genetic variants associated to alternative exon usage during the maturation of mRNA from 'sGenes'. The same classification criteria based on the position of eQTLs apply to sQTLs. Together, eQTLs and sQTLs are items of a longer list of molecular (mol) QTLs. Other examples of molQTLs include methylation (me), protein (p) and chromatin accessibility (ca) QTLs.

Mapping eQTLs is relatively straightforward, using linear models analogous to Equation (1.2),

$$\mathbf{X} = \alpha_X + \beta_{Xi}\mathbf{G}_i + \mathbf{Z}^\top\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1.4}$$

where $\mathbf{X}$ contains values of the gene expression endophenotype, α_X is equal to the mean of $\mathbf{X}$ across homozygotes for the reference allele, β_{Xi} captures the effect of putative eQTL i on the endophenotype, $\boldsymbol{\gamma}$ captures the effects of the covariates in $\mathbf{Z}$, and $\boldsymbol{\varepsilon}$ is independent and normally distributed noise. In contrast, sQTL mapping needs resolving isoform-specific expression, and is complicated by inherent correlations in mRNA isoform levels: higher usage of an exon set implies lower expression of other isoforms (Garrido-Martín et al., 2021; Yamaguchi et al., 2022).

Lessons learned from atlases of tissue-specific gene expression regulation

Early estimates of the heritability of gene expression revealed a variable impact of genetics across tissues (Price et al., 2011; Powell et al., 2012; Grundberg et al., 2012). For instance, Price et al. (2011) estimated that *cis* regulation could explain 37% of heritability of gene expression in blood, versus only 24% in adipose tissue. While these studies helped grasp the genetic bases of endophenotypes underlying complex traits, their span was limited to easily accessible tissues like blood and skin. In order to overcome this limitation, and fully exploit the wealth of knowledge produced by GWASs, the National Institutes of Health (NIH) conceived the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013).

One of the main motivators behind the GTEx project was to build a data resource in order to facilitate the study of gene expression and genetic variation across human tissues. Over the years since its conception, the repository has been updated several times (The GTEx Consortium, 2015, 2017, 2020). The latest release of the GTEx atlas extends across over 49 tissues sampled post-mortem from 838 individuals, and is an established reference in the human genomics community.

Molecular QTL effects are pervasive. Estimates from the current version of the GTEx atlas testify to the widespread impact of genetics on transcription (The GTEx Consortium, 2020). Across all tissues, 95% and 67% of protein-coding genes are respectively eGenes and sGenes of at least one variant; non-eGenes in a given tissue are enriched in genes not expressed in that context.

Focusing on genetic effects on mRNA abundance, 43% of the common variants (MAF $\geqslant$ 1%) catalogued in GTEx are *cis*-eQTLs in at least one tissue. In line with previous reports (Yang et al., 2017), there is an overlap between the genetic bases of short and long-range regulation of gene expression in GTEx: the top *trans*-eQTLs discovered in each tissue are around six times more likely to also be *cis*-eQTLs, relative to variants that are not *trans*-eQTLs. Furthermore, between 20% to 80% of *trans*-eGene effects are mediated by *cis*-eQTLs across several tissues. ■

GWAS QTLs are enriched in eQTLs. These results also emphasize the utility of mapping eQTLs in the quest for the genetic basis of complex traits (The GTEx Consortium, 2020). While a *cis*-eQTL was discovered in 43% of variants tested for association with gene expression, this percentage increased to 63% of associations in the GWAS catalog curated by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EBI) (Sollis et al., 2023), yielding a 1.46-fold enrichment (FE) for *cis*-eQTLs among QTLs. Interestingly, *trans*-eQTLs were more clearly over-represented in the NHGRI-EBI GWAS catalog, with a 6.97-FE.

Prior work from the GTEx Consortium and others had already highlighted a link between eQTLs and GWAS loci in disease-relevant contexts (Dimas et al., 2009; Westra et al., 2013; The GTEx Consortium, 2015). For instance, Westra et al. (2013) show that the rs4917014 variant linked to systematic lupus erythematosus (SLE) in the NHGRI-EBI GWAS catalog is a *trans*-eQTL of multiple biomarkers of SLE, such as *C1QB*, that encodes a protein of the complement system. ■

GWAS QTL effects are weaker than eQTL effects. Francis Crick advocated the role of proteins as the functional units behind complex biological phenomena (Crick, 1958). In line with the central dogma (Fig. 1.6), genetic variants that cause variation in complex traits act through changes in protein activity, whether they lead to missense or nonsense changes in the aminoacid chain, changes in codon usage or changes in protein abundance. However, the layers of post-transcriptional regulation buffer protein activities against variation downstream from transcription. Hence, eQTL effect sizes are expected to be larger than those of GWAS QTL effects (Battle et al., 2015). In line with this expectation, eGene expression was at least doubled by the alternative allele of 22% of *cis*-eQTLs in average across all tissues (The GTEx Consortium, 2020). In contrast, most GWAS variants yield a small increase in relative risk of at most 1.5-fold (Manolio et al., 2009).

In fact, Mostafavi et al. (2022) suggest that GWAS and eQTL mapping studies are biased to detect different types of variants: while GWAS QTLs are involved in complex regulatory schemes and have measurable effects on organismal traits, eQTLs are associated to simpler regulation of gene expression, albeit with stronger effects. ■

Molecular QTLs are enriched in functional genomic elements. The data in GTEx become all the more powerful when compounded with annotations from the 'Encyclopedia of DNA elements' (ENCODE; Appendix B, page 193). In fact, both eQTLs and sQTLs are enriched in functional genomic elements. Notably, while sQTLs are enriched almost exclusively among transcribed regions, eQTLs are also over-represented among annotated co-transcriptional regulators in noncoding regions (The GTEx Consortium, 2020). In particular, while *cis*-sQTLs are most strongly enriched among splice sites (FE $\approx$ 256), *cis*-eQTLs are mainly found in untranslated regions (UTRs) and promoter sequences [c] (FE $\approx$ 4). Similar to previous observations, there is little overlap between *cis*-eQTL and *cis*-sQTL effects in GTEx (Lappalainen et al., 2013; Li et al., 2016).

c. *Cis*-eQTLs also show FE $\approx$ 32 among splice sites, but these ultimately amount to very few variants.

In contrast to short-range regulators of gene expression, *trans*-eQTLs appear over-represented among annotations linked to pre-transcriptional and post-transcriptional regulatory steps. For instance, they show a 4-FE in motifs for the CCCTC-binding factor (CTCF) that creates bundles of closed chromatin by tethering to far-apart loci in DNA (Appendix B, page 193).

Overall, the collaborative resources output by the GTEx and ENCODE Consortia enable an integrative description of the impact of variation at functional genomic elements on gene expression. For example, the rs9896202 SNP—reported as a *cis*-eQTL of *CBX8* in lung tissue—lies just a few base pairs upstream of its eGene. Interestingly, this locus is annotated in ENCODE as a binding motif for TF EGR1. These data, added to the cross-tissue correlation (Spearman's $\rho = -0.69$) observed between the expression of *CBX8* and *EGR1*, suggest that the *cis*-eQTL effect of rs9896202 could reflect a disruption of the binding motif for EGR1 (The GTEx Consortium, 2020). ■

The proximal genetic control of gene expression is characterized by two general properties that reflect the biased representation of *cis*-eQTLs among certain functional genomic elements. First, regulation is often extremely local: most *cis*-eQTLs are found within—or close to—the body of the corresponding eGene. Second, short-range control is mostly private: while some eGenes are regulated by several *cis*-eQTLs, these are a minority. Both of these observations are illustrated in Figure 1.7, using experimental data from *cis*-eQTLs mapped in a 100-kilobase window around genes expressed in resting myeloid cells from human peripheral blood.

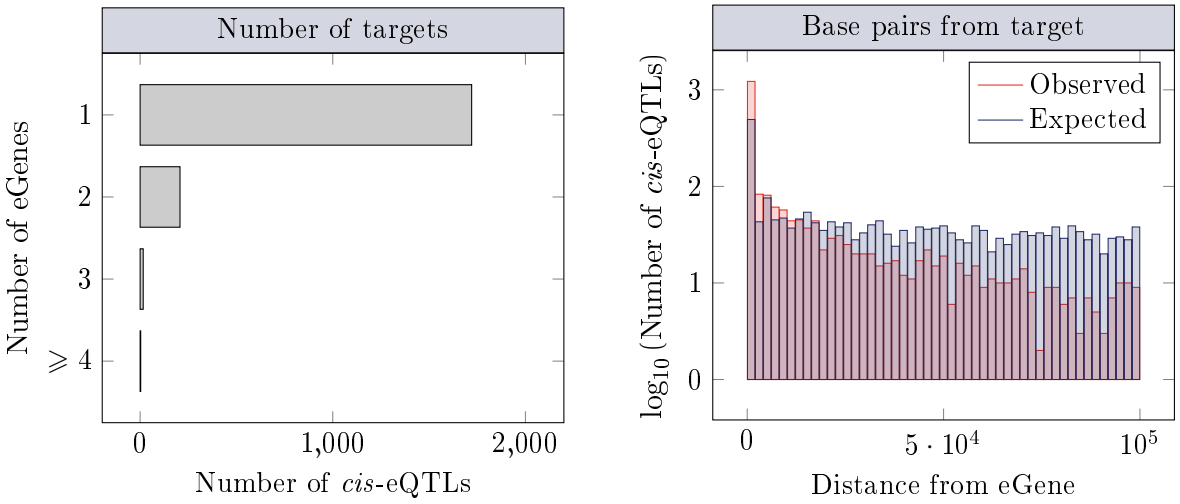


Figure 1.7 | Quantitative genomic features of *cis*-expression quantitative trait loci. Both histograms illustrate genomic features of the genetic basis of gene expression, using *cis*-expression quantitative trait loci (eQTLs) discovered ($|\beta_X| \geqslant 0.2$, Benjamini-Hochberg-adjusted Student's two-sided t-test $p < 0.01$) in a 100-kilobase region around genes expressed in resting monocytes as an example (Aquino et al., 2023). The plot on the left concerns the absolute number of target (i.e., eGenes) associated to each *cis*-eQTL. The plot on the right focuses on the number of the base pairs separating *cis*-eQTLs from the body of each eGene. Red bars show results for genome-wide significant *cis*-eQTLs in resting monocytes; blue bars show results for the same variants, but mapped on permuted gene expression data, thus reflecting spurious associations.

In contrast, *trans*-eQTLs are often associated to the expression of multiple genes spread far and wide across the genome. One way in which a *trans*-eQTL can arise is by changing the recognition of nucleotide motifs by proteins. On the one hand, these changes can affect DNA loci associated to pre-transcriptional regulation of gene expression. For example, variation at CTCF binding sites can create *trans*-eQTLs (The GTEx Consortium, 2020; Võsa et al., 2021) by bringing chromatin segments on different chromosomes into contact (Delaneau et al., 2019). On the other hand, variation in sequences that encode TF DNA-binding domains—or proteins that regulate TF activity, like cell signalling receptors and kinases—can also affect the expression of genes spread far apart across the genome, but acting in the same transcriptional program wired in a gene regulatory network (GRN; § 3.1.2, page 57).

For instance, Quach et al. (2016) report the rs5743618 missense variant in the locus of Toll-like receptor (TLR) 1, that affects the expression of over 400 genes in myeloid cells stimulated with a synthetic mimic of bacterial lipopeptides. TLR1 is an immune cell surface receptor that recognizes patterns associated to bacterial and fungal pathogens; its activation triggers GRNs essential for the antibacterial response. Accordingly, the targets of the rs5743618 *trans*-eQTL are enriched in genes involved in the response to bacterial infection, including mediators of inflammation such as *CCL5* and *IL10*. Further attesting to the biological relevance of these results, rs5743618 appears associated to immune-related disorders like allergies and asthma in the NHGRI-EBI GWAS catalog.

These observations, added to the insights garnered from the GTEx and ENCODE resources, highlight the importance of considering both coding and noncoding regions when assessing the genetic bases of complex diseases (Hindorff et al., 2009; The GTEx Consortium, 2020; ENCODE Project Consortium, 2020). Mapping eQTLs in regulatory regions is paramount in order to gain a mechanistic understanding that bridges the gap between the measured GWAS QTL and the observed phenotype (Cookson et al., 2009).

1.2.5 Inferring causal links between genotype and phenotype

Considering regulatory associations becomes particularly important when the likely causal gene underlying a GWAS trait is located far away from the eQTL that controls its expression. The pitfalls of this scenario are especially well illustrated by the associations at the 'fat mass and obesity associated' (*FTO*) locus. Shortly after the dawn of the GWAS era, several studies of body mass index (BMI) and obesity-related disorders reported links with variants in the introns of the *FTO* gene (Dina et al., 2007; Frayling et al., 2007; Scuteri et al., 2007). Yet, in spite of the reproducible associations, no study at the time was able to establish a functional link between increased BMI and the FTO protein (Klöting et al., 2008; Grunnet et al., 2009).

Several years later, Smemo et al. (2014) proposed that the functional link between obesity and *FTO* could actually be mediated by the Iroquois homeobox protein 3 (IRX3) TF. First, the authors reported direct contact between the intronic stretch of *FTO* associated to obesity and the promoter of *IRX3* in the human genome, suggesting that the former could act as a regulatory factor of the latter. Accordingly, the obesity-related variants in *FTO* were then discovered as eQTLs of *IRX3* in human brain tissue. Supporting a functional link between BMI and the activity of IRX3 in the brain, weight losses were recorded in transgenic mice with deficient *Irx3* expression in the hypothalamus. Together, these data suggest that *FTO* is a regulator of *IRX3*, which would then be the true predictor of polygenic obesity.

Although the exact molecular mechanisms and cellular contexts in which *FTO* and/or *IRX3* can affect BMI are still not fully understood, recent work points towards a role for macrophage-expressed *IRX3* in metabolic inflammation leading to obesity (Yao et al., 2021).

All in all, the *FTO*/*IRX3* association is a good example of how functional data can often help disentangle the effect of variants in non-coding regions. Figure 1.8 illustrates this in a schematical way, using QTL and eQTL data simulated across 30 observations. Although the strongest QTL signal lies in the body of gene B, and may thus appear as a tempting causal agent, the true simulated QTL locus i is located in a noncoding stretch of DNA between genes A and B. In this context, adding gene expression data simplifies the picture in two ways. First, it indicates that the strongest QTL signal actually has no effect on the transcription of gene B, while in turn the true causal locus i is the strongest eQTL of eGene B. Second, expression data indicates that gene A is not a target of this variant. Hence, endophenotype data can help gain a better understanding of the molecular mechanisms behind the QTL association, as well as narrow down the list of candidate genes to study (Umans et al., 2021).

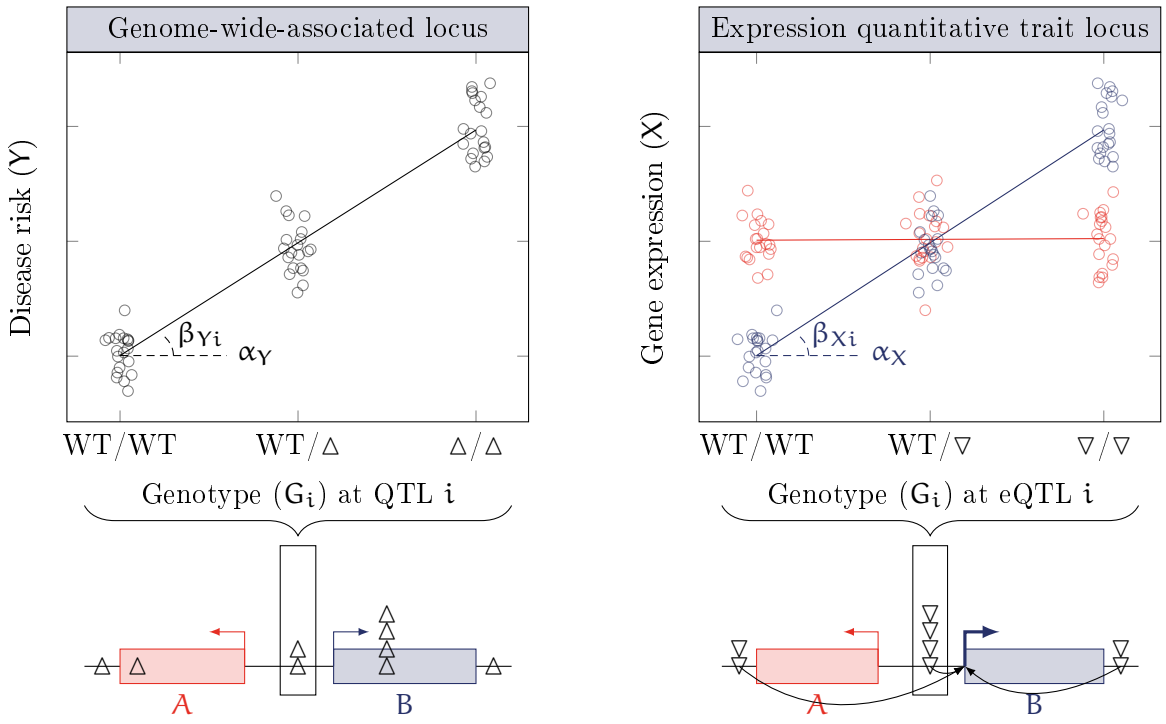


Figure 1.8 | Refining genome-wide associations with expression traits. Genome-wide association studies (GWASs) are commonly used to map quantitative trait loci (QTLs) associated to disease risk (Y) by comparing the genotype (G_i) at a given locus i across cohorts of cases and controls. Due to linkage disequilibrium and/or cryptic confounders, several associations may be found in a given genomic window, even though there is a single causal variant. Furthermore, the intepretation of GWAS results is complicated when hits fall on non-coding regions. Mapping expressed quantitative trait loci (eQTLs) may help interpret GWAS results and refine the list of putative causal variants by highlighting those that are also associated to gene expression (X). In this example, although the strongest QTL association falls on the body of gene B, the true causal QTL i lies in a noncoding region between genes A and B. Adding expression data clears the picture up by highlighting locus i as the strongest eQTL for gene B, but not for gene A. Upward and downward pointing triangles quantify the strength of the QTL and eQTL associations, respectively. The black line represents the association at putative QTL i; the red and blue lines give the slope of the putative eQTLs for genes A and B, respectively. Adapted from Umans et al. (2021).

Figure 1.8 also illustrates a case of 'colocalization', where a variant displays horizontal pleiotropy through simultaneous association with two traits (Box 2). In this case, the QTL of trait Y is also an eQTL of the gene expression endophenotype X. Importantly, colocalization assumes no causal link between X and Y, only a shared ætiology to changes in both.

Several approaches to colocalization analyses have been described in recent years (Nica et al., 2010; Hormozdiari et al., 2014; Giambartolomei et al., 2014; Zhu et al., 2016; Ongen et al., 2017; Wen et al., 2017). In particular, Giambartolomei et al. (2014) devised a Bayesian method to quantify the posterior probability of colocalization for two traits X and Y from summary association statistics. Briefly, the method iteratively tests five hypotheses for each assayed genetic variant. The first one ($\mathcal{H}_0$) is that there is no association between the variant and either trait. The second and third hypotheses ($\mathcal{H}_1$, $\mathcal{H}_2$) posit that the variant is associated to trait X but not Y and vice versa. The fourth hypothesis ($\mathcal{H}_3$) is that genetic associations exist for both traits, but they are mediated by different variants [d]. Finally, the fifth hypothesis ($\mathcal{H}_4$) is that the same variant is associated to both traits. Thus, a high posterior probability for $\mathcal{H}_4$ ($PP_{\mathcal{H}4}$) is good evidence for colocalization.

When applied in the context of QTL and molQTL mapping—where trait Y is the focal phenotype and trait X is a molecular endophenotype—colocalization analyses can provide strong evidence for regulatory genetic effects on phenotype (Fig. 1.8). For example, the rs9896202 *cis*-eQTL linked to expression of *CBX8*—a biomarker of tumorigenesis—is a presumed disruptor of EGR1 binding that shares a $PP_{\mathcal{H}4}$ of 68% with breast cancer risk (The GTEx Consortium, 2020; Shi et al., 2021).

d. Rejection of this hypothesis does not exclude cases where two variants in perfect LD are linked to each trait.

More generally, researchers from the GTEx Consortium assessed colocalization between proximal molQTLs and over 5 thousand GWAS loci spanning 87 complex traits (The GTEx Consortium, 2020). While 23% of these QTLs colocalize with *cis*-sQTLs, this proportion rises up to 43% for *cis*-eQTLs, further emphasizing the overlap between the genetic bases of complex traits and gene expression, and the relevance of molQTL mapping when seeking the former (Cookson et al., 2009).

Yet, even convincing evidence of colocalization between a QTL and an eQTL is insufficient to sustain claims of regulatory causality. That is, colocalization methods are unable to distinguish cases of horizontal pleiotropy, where traits X and Y are not directly related but share a common genetic ætiology, from cases of vertical pleiotropy, where the effect of genotype on trait Y is mediated by trait X (Box 2). Only in the latter case can regulatory causal effects be inferred.

The transcriptome-wide association study (TWAS) framework has recently been proposed as an approach to infer vertical pleiotropy between genotype, phenotype and an expression endophenotype (Gamazon et al., 2015; Park et al., 2017; Barbeira et al., 2018). Although several TWAS methods exist, the common basic principle is to impute genetically regulated expression (GReX) values for each assayed gene based on reference genotyping and transcriptome data sets, and then test whether imputed gene expression values are significantly associated to genetic variation and the focal phenotype in the cohort of interest.

In the TWAS implementation by Gamazon et al. (2015)—for a given GWAS trait Y and an expression endophenotype X, and for each gene g in the transcriptome—the method starts by training an elastic net regression model to estimate the effect on the expression of g of genetic variants in a fixed-width window around it. The model is of the form

$$\mathbf{X} = \alpha_X + \mathbf{G}^T\boldsymbol{\beta}_X + \mathbf{Z}^T\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1.5}$$

where $\mathbf{X}$ is the expression of gene g across n individuals in the reference transcriptome data set, α_X is an intercept, $\mathbf{G}_{n\times m}$ encodes the reference genotypes of m variants around gene g, $\boldsymbol{\beta}_X$ captures the respective contributions of these variants on $\mathbf{X}$, $\boldsymbol{\gamma}$ captures the effects of the covariates in $\mathbf{Z}$ and $\boldsymbol{\varepsilon}$ is independent and normally distributed noise. Then, access to a full GWAS data set is required to impute GReX values $\mathbf{X^R}$ from the estimated weights $\boldsymbol{\beta}_X$ and the observed genotypes $\mathbf{G^R}$ in the cohort of interest,

$$\mathbf{X^R} = \alpha_{X^R} + \mathbf{G^R}^T\boldsymbol{\beta}_X. \tag{1.6}$$

Finally, the extent to which genetically controlled changes in the expression of gene g can explain the GWAS trait is measured by correlating $\mathbf{X^R}$ and $\mathbf{Y}$ (Gamazon et al., 2015).

More recently, Barbeira et al. (2018) developed a version of this method able to leverage GWAS and eQTL-mapping summary statistics, bypassing the need for complete data sets. For a given gene g, the association between $\mathbf{X^R}$ and Y is approximated as a weighted sum W of a 'QTL' component of GWAS Z-scores for the set of m variants around g, scaled by an 'eQTL' component representing their contribution to variance in the expression of g,

$$W \approx \sum_{i=1}^{m} \overbrace{\beta_{Xi}\frac{\widehat{\sigma}_i}{\widehat{\sigma}_g}}^{\text{eQTL}} \underbrace{\frac{\widehat{\beta}_{Yi}}{se\left(\widehat{\beta}_{Yi}\right)}}_{\text{QTL}}, \tag{1.7}$$

where β_{Xi} is the weight of variant i on the prediction of the expression of gene g, $\widehat{\sigma}_i$ and $\widehat{\sigma}_g$ are the respective estimated variances of the genetic variant and the predicted expression of g, and $\widehat{\beta}_{Yi}$ and $se\left(\widehat{\beta}_{Yi}\right)$ are the effect size estimated by the GWAS for variant i and its standard error.

The colocalization and TWAS approaches are complementary; when properly combined, they can yield fine mechanistic insights into the processes underlying GWAS discoveries. For instance, the latest GTEx release reports the rs1775555 SNP as a *cis*-eQTL of *GATA3* and a *trans*-eQTL of *MSTN*. *GATA3* encodes a TF involved in many aspects of immune regulation and *MSTN* encodes a secreted ligand of transforming growth factor (TGF) β, a major immune cytokine (Wan, 2014; Wang et al., 2018). Both associations colocalize in fibroblasts with a $PP_{\mathcal{H}4}$ over 99%, and TWAS results suggest ($p = 2.1 \times 10^{-22}$) that the *trans*-eQTL effect of rs1775555 on *MSTN* is mediated by its *cis*-eQTL effect on *GATA3* (The GTEx Consortium, 2020). Interestingly, both genetic associations colocalize with a $PP_{\mathcal{H}4}$ over 97% with multiple immune traits, such as asthma and eczema.

More broadly, around 90% of gene-trait associations with strong evidence of colocalization also show significant TWAS results, in median across the more than 5 thousand GWAS loci considered in that work (The GTEx Consortium, 2020). However, caution must be taken when interpreting the significance of TWAS results for eGenes with few eQTLs. The sum W in Equation (1.7) is maximal when (i) the strongest eQTLs are also the strongest QTLs, and if (ii) the direction of eQTL and QTL effects are consistent with mediation by gene expression. Indeed, if there is no association between eQTL effects on g and QTL effects on Y, the signs in Equation (1.7) are expected to cancel out, leading to a lower overall W value for the gene-trait association. If several variants are associated to the expression of gene g, such coordinated effects on gene expression and the GWAS trait may be robust evidence for genetic effects on phenotype mediated by gene expression changes. However, if the expression of g is controlled by only one variant, a significant gene-trait association may result from spurious association between the eQTL and QTL effects.

1.2.6 Conditioning on context-dependency to improve mapping

Given the pervasiveness of genetic effects on phenotype, it is paramount to disentangle the regulatory associations to disease in the human genome. The incomplete overlap between QTLs and eQTLs could at least partially be due to a lack of power to map these variants. Yet, even though the most recent GTEx atlas more than doubled the number of samples relative to its previous installment, yielding more powerful studies to discover molQTLs and their targets, the proportion of eQTLs that colocalize with disease risk increased only modestly (The GTEx Consortium, 2017, 2020). Hence, a purely incremental approach does not appear as a viable line of action towards this goal (Umans et al., 2021; Mostafavi et al., 2022) (§ 2.3.3, page 50).

Other technical factors that could explain missing QTL and eQTL links include batch effects, confounding factors like population stratification (§ 1.2.3, page 10), and the so-called 'winner's curse': a general statistical effect that can explain lack of replicability in GWAS settings. Briefly, the curse appears when the measured association of the peak SNP in a discovery cohort is stochastically stronger than its ground-truth value. Due to this randomness, it is unlikely that the same association will again be the strongest in another cohort (Bigdeli et al., 2016).

Differences in biological context can also explain why some regulatory associations are missed. For instance, some links between genetics and gene expression are only revealed after a certain stimulus (Barreiro et al., 2012). Such 'response' (r) eQTLs differ from standard eQTLs in the sense that they are not associated to absolute mRNA abundances, but to the change in transcript counts following stimulation. This type of conditional genetic effect is widespread in actors of the immune system (Kim-Hellmuth et al., 2017; Zhernakova et al., 2017). For example, Figure 1.9 shows a reQTL associated to a change in the expression of *MMP1*, but specifically after stimulation by 'severe acute respiratory syndrome' coronavirus 2 (SARS-CoV-2; COV) ($\beta_{Xi}^{COV} = 1.47$, $p = 2 \times 10^{-16}$), and not by the influenza A virus (IAV) ($\beta_{Xi}^{IAV} = 0.04$, $p = 0.4$). Interestingly, the matrix metalloproteinase 1 product of *MMP1* has been pointed out as a marker of the severity of the 'coronavirus disease 2019' (COVID-19) triggered by SARS-CoV-2 infection (Syed et al., 2021).

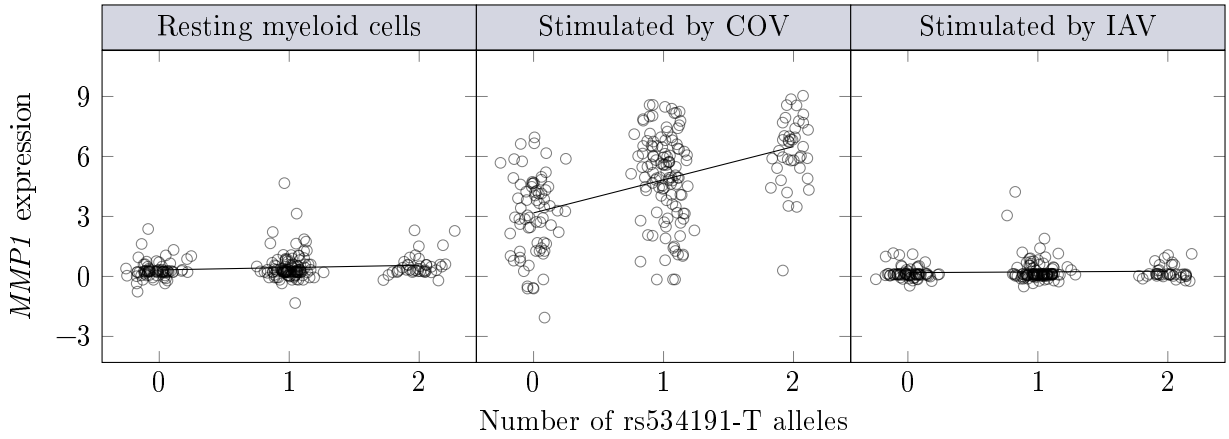


Figure 1.9 | A stimulus-dependent response expression quantitative trait locus. The y axis shows the relative abundance (i.e., counts per million) of mRNA molecules transcribed from the *matrix metalloproteinase 1* gene (*MMP1*) in myeloid cells from each of 222 individuals at the basal state, or following six hours of stimulation by 'severe acute respiratory syndrome' coronavirus 2 (SARS-CoV-2; COV) or the influenza A virus (IAV). The x axis stratifies gene expression observations according to the genotype at SNP rs534191 carried by each individual. Adapted from Aquino et al. (2023).

Moreover, regulatory activity in response to a stimulus can be largely dependent on the period of stimulation. For instance, across 417 reQTLs mapped in monocytes exposed to three different stimuli and assayed at two time points, Kim-Hellmuth et al. (2017) estimate that while 13% to 51% were stimulation-specific at a given time point, 32% to 64% were specific to a given period of stimulation, suggesting a highly dynamic genetic control of immune gene expression.

The impact of genetics on molecular endophenotypes may also be conditioned by tissue-dependent regulation. Humans are knit from a multitude of different tissues, each with different structural and functional properties. Although the heritability of gene expression across these systems varies, the functions carried by each tissue at least partially depend on specific GRNs (Price et al., 2011; Powell et al., 2012; Grundberg et al., 2012; The GTEx Consortium, 2017, 2020; Yao et al., 2020).

Here again, the GTEx atlas provides an optimal setting to assess the tissue-dependency of genetic effects on molecular endophenotypes (The GTEx Consortium, 2020). Overall, the correlation patterns of proximal and distal eQTLs, as well as *cis*-sQTLs, recapitulate the known patterns of similarity between the 49 tissues in GTEx. Interestingly, while genetic control across seemingly different tissues, such as breast and uterine tissue, appears relatively similar (Spearman's $\rho > 0.8$), the blood is an outlier with distinct regulation patterns. Focusing on the variants themselves, proximal control of expression and splicing is either very specific or widely shared across tissues: the majority of *cis*-eQTLs and *cis*-sQTLs are either detected in 5 GTEx tissues or less, or found in more than 45 of them. In turn, over 70% of *trans*-eGenes are detected in 5 tissues or less.

In summary, studies that seek links between genotype, endophenotype and phenotype need to assay relevant contexts in which regulation is active. Active regulation may require prior stimulation (Barreiro et al., 2012; Fairfax et al., 2014; Kim-Hellmuth et al., 2017) (Fig. 1.9), but it may also vary across tissues. Indeed, it has been shown that tissues that are clearly linked to a given trait—such as liver tissue for cholesterol levels or brain tissue for the risk of Alzheimer's disease—are enriched in high expression levels of eGenes, as well as strong effects of corresponding *cis*-eQTLs, that also colocalize with the trait-associated GWAS loci in the relevant tissue (The GTEx Consortium, 2020).

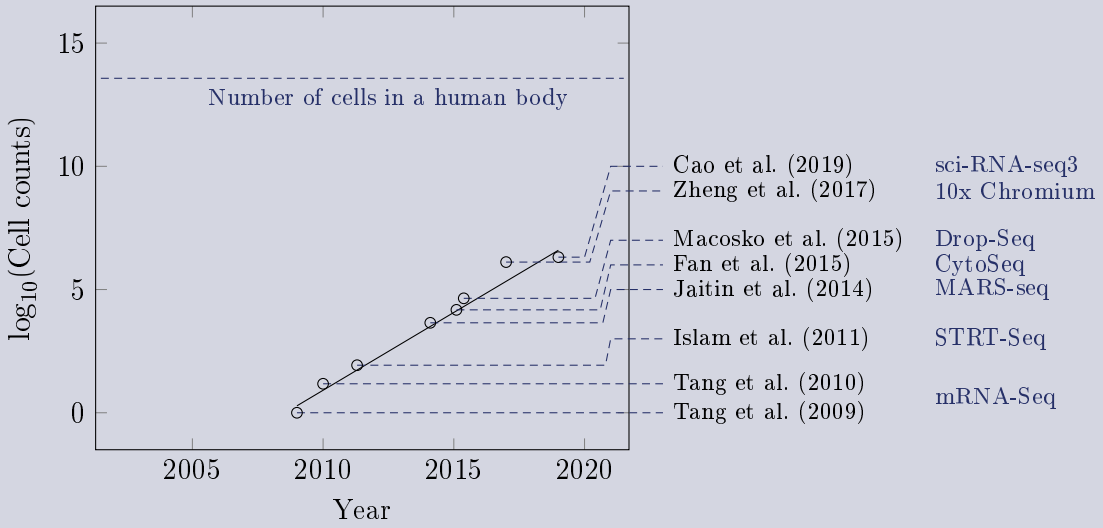
Tissues are themselves composed of many different types of cells embedded in an extra-cellular matrix. Each subset of cells fulfills a particular role, such that complex tissular functions arise from the coordinated action of different cell types. Hence, eQTLs mapped through bulk measurements of mRNA abundance capture a mixture of heterogeneous GRNs that obscures cell-type specific regulatory mechanisms in each tissue, thereby limiting the functional interpretability of association results (van der Wijst et al., 2018a; Kim-Hellmuth et al., 2020).

Box 3 | Evolution of single-cell sequencing before 2020. The field of single-cell sequencing was born from the development of next-generation methods for high-throughput sequencing of nucleic acids. Tang et al. (2009) were the first to describe a method for sequencing complementary (c) DNA retro-transcribed from the mRNA molecules contained in a single cell (mRNA-Seq). A year later, the same group showed that their method could be used to characterize cell-resolved transcriptomes across around ten cells in parallel (Tang et al., 2010). From this princeps work, further developments in sequencing technologies enabled the characterization of ever larger mixtures of single cells.

The Figure below, adapted from Kharchenko (2021), highlights recent developments that led to order-of-magnitude jumps in single-cell sequencing capacities. A more comprehensive description of earlier developments is provided by Svensson et al. (2018).

Islam et al. (2011) proposed a method for single-cell tagged reverse transcription followed by sequencing (STRT-Seq) and used it to map the transcriptome of around a hundred single cells. In contrast to mRNA-Seq, that was based on sequencing by oligonucleotide ligation and detection (SOLiD), STRT-Seq leveraged Illumina's sequencing-by-synthesis solution, yielding a greater throughput. A couple of years later, Jaitin et al. (2014) also leveraged Illumina sequencing to analyse around 5 thousand cells using massively parallel single-cell RNA-sequecing (MARS-seq).

The next jump in cell sample size came with the introduction of CytoSeq, a plate-based approach using microwell arrays and combinatorial indexing beads, capable of assaying several thousands of single cells (Fan et al., 2015). That same year, Macosko et al. (2015) described the first droplet-based single-cell RNA-sequencing method (Drop-Seq), increasing throughput to just shy of 50 thousand cells.



In 2017, 10x Genomics commercialized droplet-based sequencing through their Chromium Controller, and showed that their product could be used to reliably map the transcriptome of millions of cells (Zheng et al., 2017). Today, 10x Genomics' solutions are among the most popular single-cell sequencing technologies. Although other frameworks are available that can yield higher cell counts at a lower cost through combinatorial designs, they are not used as widely yet (Cao et al., 2019).

In fact, an eQTL allele can have opposing effects on the expression of the same eGene in different cell types. For example, in line with previous reports by Fairfax et al. (2012), Yazar et al. (2022) associate the rs4987360-G allele to a decrease in the expression of *SELL* in classical monocytes, but an increase in *SELL* mRNA abundance in naïve B cells. The *SELL* gene encodes the homing receptor CD62L, essential for immune cell trafficking between the blood and the lymph nodes. Through rs4987360, the innate and adaptive branches of the immune system (§ 3.1, page 55) can then weigh in on the overall expression levels of CD62L to ensure proper leukocyte recirculation: a key feature of the system.

One way to skirt this heterogeneity is by sorting cells beforehand, so as to measure mRNA abundance in purified cell types (Barreiro et al., 2012; Fairfax et al., 2012, 2014; Nédélec et al., 2016; Quach et al., 2016; Ishigaki et al., 2017). Alternatively, computational methods that deconvolve cell type mixtures based on bulk expression data have also been proposed (Venet et al., 2001; Westra et al., 2015; Aran et al., 2017). Using one of these deconvolution methods, researchers from the GTEx Consortium recorded a strong correlation between the measures of pairwise similarity based on tissue cellular composition and proximal regulation of gene expression (Aran et al., 2017; The GTEx Consortium, 2020). This suggests that the cross-tissue correlations in gene regulatory activity described in GTEx could arise simply from tissues sharing similar cell type proportions, not from the activation of common GRNs.

In silico cell-type deconvolution also revealed widespread of cellular composition variability across GTEx samples (The GTEx Consortium, 2020). Researchers leveraged this variability to map cell-type interaction (i) proximal molQTLs in seven tissues including liver, skin and whole blood. This revealed over a thousand neutrophil-specific ieQTLs in whole blood, a fraction of which had not been detected in bulk analyses.

Kim-Hellmuth et al. (2020) further generalized cell type deconvolution across 35 GTEx tissues, and mapped ieQTL and isQTL variants affecting the expression of over 3 thousand protein-coding genes in a cell-type dependent manner. Interestingly, eGenes of eQTLs found in a single tissue are enriched ieQTL targets, relative to eGenes shared across more than one tissue, suggesting that ieQTLs participate to the tissue-specificity of gene expression regulation. This is in line with cross-tissue patterns of genetic control of expression emerging from similarities in cellular composition.

Similar to standard eQTLs, ieQTLs show a 1.3-FE in GWAS loci, in median across the 87 previously considered traits (Kim-Hellmuth et al., 2020; The GTEx Consortium, 2020). For over a third of the 1,370 eGenes that showed signs of proximal genetic control by a GWAS locus, only the cell-type resolved ieQTL—not the fine-mapped standard eQTL—showed strong evidence of colocalization ($PP_{\mathcal{H}4} > 0.5$) with the GWAS trait. Overall, these results highlight the importance of considering cell-type specific regulation when mapping the genetic bases of complex traits, and emphasize the need for population-level QTL mapping studies at single-cell scale in this endeavour (Kim-Hellmuth et al., 2020) .

1.3 Genomic features at single-cell resolution

In vitro cell-sorting technologies and in silico deconvolution methods are useful to map the genetic bases of complex traits at cell-type resolution (Venet et al., 2001; Barreiro et al., 2012; Fairfax et al., 2012, 2014; Lee et al., 2014; Çalışkan et al., 2015; Westra et al., 2015; Nédélec et al., 2016; Quach et al., 2016; Aran et al., 2017; Ishigaki et al., 2017; Kim-Hellmuth et al., 2017, 2020; Piasecka et al., 2018; Schmiedel et al., 2018; Ye et al., 2018). However, these approaches are respectively limited by the availability of markers to tag specific cell types and their ability to capture rare subsets in bulk expression data (van der Wijst et al., 2018a).

1.3.1 Quantification of transcript abundance at single-cell scale

The development of NGS technologies revolutionized many aspects of biology (Box 1). In particular, it spurred the development of methods to characterize transcript abundances at single-cell resolution directly from a mixture of cells. Tang et al. (2009) first described single-cell RNA-sequencing (scRNA-seq) of transcripts contained in a developing mouse blastomere. The authors repurposed a previously described method for amplification of complementary (c) DNA, that they characterized through sequencing by oligonucleotide ligation and detection (Shendure et al., 2005; Kurimoto et al., 2006, 2007).

Where previous microarray-based transcriptomic technologies required micrograms of tissue sample, this new method allowed to capture a larger portion of the transcriptome—over 5 thousand more genes—from a single cell (Tang et al., 2009). In the following decade, several other single-cell sequencing approaches were developed to assay increasing numbers of single cells (Box 3).

Most modern scRNA-seq workflows apply droplet-based methods that use microfluidic networks to encapsulate single cells in an oil emulsion (Heumos et al., 2023). Each nanoliter-volume drop in the emulsion should contain a cell, a support carrying copies of a nucleotide 'barcode' sequence that identifies the droplet, and all the reagents needed to transform the mRNA molecules it contains into sequenceable cDNA fragments. After retro-transcription and ligation of the barcodes to the resulting cDNA molecules, the emulsion is broken up to create a pooled cDNA sequencing library. Each sequenced read in the library can then be retraced back in silico to the droplet whence it came thanks to its barcode (Macosko et al., 2015; Zheng et al., 2017). Before pooling, each transcript is also assigned a unique molecular identifier (UMI), so as to limit the impact of biased cDNA synthesis and polymerase chain reaction amplification during library preparation (Islam et al., 2014).

After barcode demultiplexing, the set of reads coming from each droplet is aligned to a common reference so as to characterize expressed genomic features. This information is then summarized in a feature-barcode matrix (FBM) that contains the number of UMIs assigned to each mapped transcript counted in each droplet in the assay. The FBM is the starting point of most pipelines for pre-processing of raw scRNA-seq data.

Yet, the analysis of scRNA-seq data is a rapidly evolving domain. By analyzing trends in the 'scRNA-tools' data base (Zappia et al., 2018), Zappia and Theis (2021) recorded a supralinear increase in the number of catalogued methods since 2016. As of 2021, over a thousand tools were available for different aspects of scRNA-seq data analysis. For example, there are more than a hundred different ways to normalize scRNA-seq count data. This reflects the striking breadth of biological questions that single-cell transcriptomics can address, but it also shows that many aspects of scRNA-seq data analysis are still under active development.

Although there is still much debate around many common analysis tasks—such as the relevance of nonlinear dimensionality reduction for visualization—there is also growing consensus in the single-cell genomics community for some of the central aspects of scRNA-seq data analysis. Recently, Heumos et al. (2023) reviewed these processing and analysis steps in detail, providing a consolidated set of expert recommendations in single-cell data analysis across multiple modalities today. In the end, the best suited tool for each job often ultimately depends on the data themselves.

Single-cell assumptions and barcode filtering steps

In principle, scRNA-seq reads associated to a valid barcode should originate from an oil droplet containing a single live cell (Macosko et al., 2015; Zheng et al., 2017). Violating this assumption is likely to bias or blur results from downstream analysis tasks (Heumos et al., 2023). Hence, quality control (QC) steps that filter barcodes associated to empty droplets or low-quality cells, and 'doublet' detection methods that remove barcodes associated to more than one cell are essential.

Most scRNA-seq data analysis pipelines use three core QC metrics: the number of mapped reads and the number of genes detected in each barcode, as well as the fraction of reads that map to the mitochondrial genome. The first two metrics give an indication of the amount of information carried by each barcode; less informative barcodes with low read counts and/or mapped features will only burden analyses and lead to noisier results. The third metric is often used to filter out dying or low-quality cells. All in all, it is important to consider all three metrics jointly and to interpret their distributions in light of the biology of the assayed cells. For instance, muscle cells can be associated to several thousands of reads with a relatively high mitochondrial fraction due to a highly metabolically active state (Mercer et al., 2011; Kuppe et al., 2021).

Several doublet detection methods have been described for single-cell sequencing data so far (Kang et al., 2018; Germain et al., 2021; Xi and Li, 2021; Neavin et al., 2022). Because doublet barcodes are expected to share defining features—such as a higher number of mapped reads—that distinguish them from 'singlets', many of these methods work by simulating transcriptional profiles from artificial doublets and comparing them to those of observed barcodes. In single-cell assays of tissues composed of widely different cell types with well described markers, heterotypic doublets that mix together distinct cell types can also be detected through aberrant expression of markers.

In single-cell eQTL mapping studies that use pooled designs, homotypic doublets of cells from different individuals are particularly dangerous, as they can lead to confounded genotype effects. Kang et al. (2018) propose a suite of tools that use genetic variation across individuals to disentangle singlets from doublets based the genetic variants detected from scRNA-seq data. These methods can thus be used to demultiplex pooled samples and trace barcoded cells back to each donor.

Yet, these genotype-based demultiplexing methods are unable to detect doublets of cells coming from the same individual. Such 'cryptic' doublets can be found by approaches that use the patterns of transcriptional similarity between singlets and doublets. A genetic singlet that clusters together with doublets is more likely to be a doublet. Because each approach uses and provides different sources of information, it is recommended to base doublet filtering decisions on the results from several methods at once (Heumos et al., 2023).

A statistical description of transcript counts in single cells

The filtered FBM is an array of whole numbers $\boldsymbol{A} \in \mathbb{W}^{m \times n}$. Each matrix item A_{gc} gives the number of unique molecules from transcript $g \in \{1, \cdots, m\}$ counted in cell $c \in \{1, \cdots, n\}$. Importantly, because mRNA capture and cDNA synthesis are imperfect, A_{gc} represents a fraction of the true total number of transcript g molecules contained in cell c. Thus, scRNA-seq data are compositional in nature (Quinn et al., 2018).

That is, the observed count for a given transcript g is a function

$$A_{gc} = f(T_{gc}) \tag{1.8}$$

of the true underlying mRNA abundance T_{gc} in cell h, where f translates variation due to mRNA capture efficiency, amplification bias, and other technical confounders like sequencing depth across transcripts and cells (Stegle et al., 2015; Lun et al., 2016b).

This also means that the 'library size', or the sum of transcript counts assigned to barcode c,

$$L_c = \sum_{i=1}^{m} \boldsymbol{A}_{\cdot c}, \tag{1.9}$$

is an artifact of the underlying mRNA abundance in the live cell that constrains the observed abundances A_{gc} as proportions of an arbitrary sum (Quinn et al., 2018). Because library size

depends on technical factors, absolute differences in abundance A_{gc} are not informative, but relative abundances normalized for L_c are. The simplest approach would be to simply divide A_{gc} by L_c, but this does not resolve the compositional character of transcriptome data (Quinn et al., 2018).

Several methods have been proposed to tackle this problem in bulk RNA-seq data. Robinson and Oshlack (2010) first described the 'trimmed mean of M-values' as an effectively normalized metric for differential expression (DE) RNA-seq analyses. Their method uses a trimmed and weighted mean modeled from a subset of transcripts—assuming that most transcripts are not DE between samples—as a library 'size factor' to normalize the data. Along the same lines, Anders and Huber (2010) proposed normalizing the data against a gene expression median estimated across transcripts. Both of these methods work well for bulk RNA-seq data, but the sparsity of the FBM complicates their application to scRNA-seq data (Box 4).

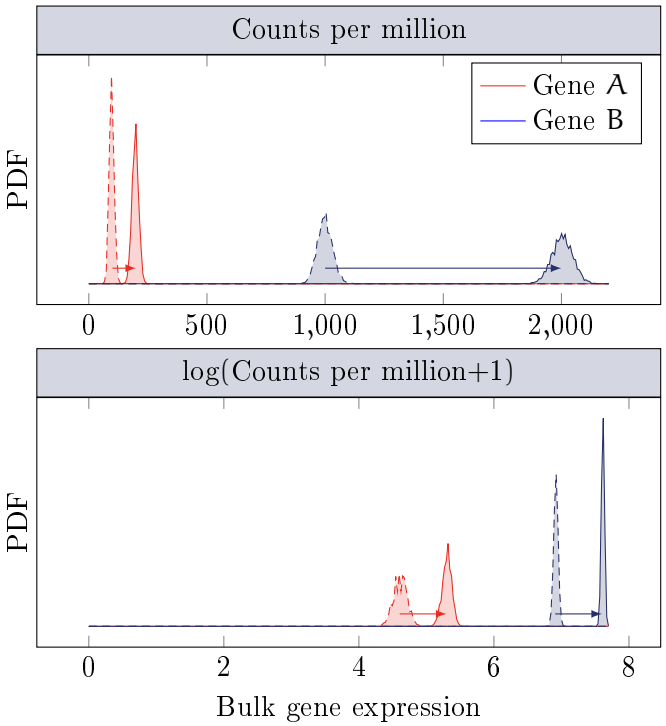


Figure 1.10 | Absolute and relative differences in gene expression. Probability density function (PDF) of 200 raw and log-transformed simulated expression values for two genes, before and after stimulation. Dashed lines show non-stimulated gene expression. Transcript abundances were modeled as purely Poissonian processes.

The shifted logarithm transformation is a popular choice for scRNA-seq data normalization. Lun et al. (2016a) describe a 'log-normalization' that computes relative transcript abundances as

$$X_{gc} = \log\left(\frac{A_{gc}}{h(L_c)} + s\right), \tag{1.10}$$

where h is a function that sums library sizes across cells with similar count depths and deconvolves the pooled values into cell-specific size factors, and $s > 0$ is a 'pseudocount' that ensures the transformation is defined for $A_{gc} \geqslant 0$. Aggregating library sizes across multiple cells helps dampen the problematic effects of low and null values (Lun et al., 2016a).

The log-transformation itself has other added benefits. First, it translates absolute transcript count differences into log-fold changes in gene expression. That is, changes in gene expression are interpreted relative to the overall expression level of each gene. For instance, Figure 1.10 illustrates a simulated scenario in which two genes A and B show a two-fold increase in bulk expression following a given stimulation. In the upper panel, expression is given in terms of 'counts per million' (CPM), a linear transformation of the data that rescales raw counts by a factor of 10^{-6}. Although the

magnitude of DE of both genes is the same, the absolute difference in the expression of the highly expressed gene B dominates the DE signal of lowly expressed gene A. Considering log-fold changes reveals the similar patterns of DE between both genes, as shown in the lower panel.

The log-transformation is also useful because it brings the distribution of transformed UMI counts closer to a Gaussian, and thus closer to the assumptions made by several downstream methods involving statistical inference through classical linear models. However, this is not the case when null expression values are frequent. For example, Figure 1.11 shows the distribution of CPM values before and after log-transformation for the *CSTH* gene in CD14$^+$ monocytes. Due to the prevalence of zeroes in the FBM (Box 4), both distributions feature a spike at zero. Several methods have been proposed to correct this bimodality through statistical modeling (Fan et al., 2016a; Stuart et al., 2019), but a simpler approach is to filter out very lowly expressed genes beforehand, so as to reduce the sparsity of the FBM and bring the distribution of log-transformed expression closer to normal.

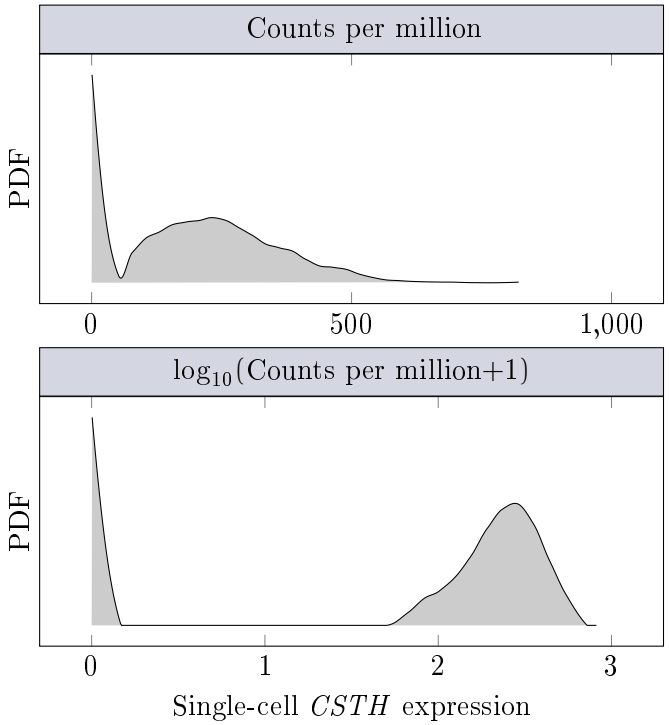


Figure 1.11 | Log-normalization and normality. Probability density function (PDF) of raw and log-transformed count-per-million expression values of *CSTH* in single CD14$^+$ monocytes. Adapted from Kharchenko (2021), using a single-cell RNA-sequencing data set of around 10 thousand peripheral blood mononuclear cells sampled from a healthy donor, made available by 10x Genomics.

In this sense, the log-transformation also has variance-stabilizing properties. That is, taking the logarithm modifies the distribution of gene expression values such that its mean and variance are less related [e]. Variance stabilization is critical for scRNA-seq data, as they are characterized by a clear mean-variance relationship that can confound downstream analyses; yet, log-normalization is not sufficient to eliminate it [f]. Using a reference scRNA-seq data set of around 10 thousand peripheral mononuclear blood cells made available by 10x Genomics, Figure 1.12 illustrates how genes with the highest log-normalized expression values also display the largest variance in expression.

e. Briefly, if gene expression values X follow a negative binomial distribution with mean μ and overdispersion ϕ (Appendix B, page 193; Box 4), the mean-variance relationship will be quadratic $\mathbb{V}[X] = \mu + \phi\mu^2$ (Ahlmann-Eltze and Huber, 2023). The log-transformation changes the multiplicative variation into additive expression noise.

f. Although Ahlmann-Eltze and Huber (2023) argue that log-transformation followed by principal components analysis is at least as effective as other more sophisticated scRNA-seq data pre-processing approaches.

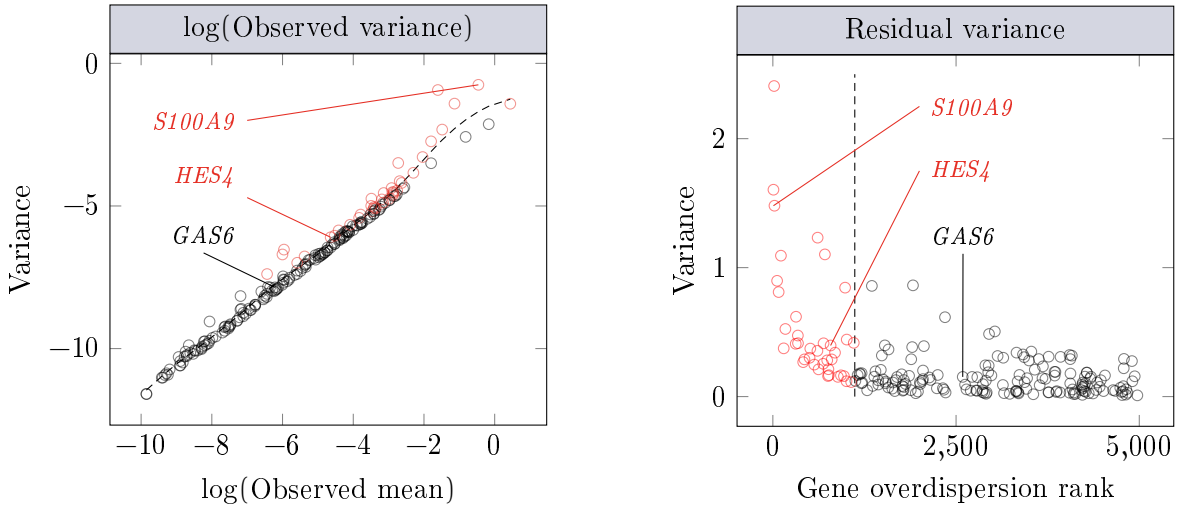


Figure 1.12 | Mean and variance of single-cell gene expression data. Single-cell RNA-sequencing data are characterized by a clear mean-variance relationship, through which genes with the highest expression levels are also the most variable. This expression variance has a biological and a technical component. Assuming that the technical component is Poisson-distributed, the expected technical noise can be modeled given mean expression values observed genome-wide. At each expression level, genes with a variance above this expectation—shown by the dashed line in the left panel—are considered biologically meaningful. 'Highly variable genes' with the largest residual biological variance components are shown in red in both panels. For example, *S100A9* is marker used to disnguish myeloid cells from other immune subsets. Adapted from Kharchenko (2021), using a single-cell RNA-sequencing data set of around 10 thousand peripheral blood mononuclear cells sampled from a healthy donor, made available by 10x Genomics.

Other methods are thus needed to correct the mean-variance relationship in scRNA-seq data, by decomposing it into its biological and technical components. In the absence of biological drivers of gene expression variation, the number k out of N randomly sampled transcripts that maps to a given gene g follows a binomial distribution parameterized $\mathcal{B}\left(N, \frac{k}{N}\right)$. As N grows to infinity, $\mathcal{B}\left(N, \frac{k}{N}\right)$ tends to a Poisson $\mathcal{P}(k)$. Thus, if the proportion of reads coming from gene g is constant across cells, the the number of transcripts sampled at random from a large number of transcripts that belong to g is expected to be Poissonian.

Hence, Lun et al. (2016b) propose to dissect the biological and technical components of gene expression variation by simulating Poisson-distributed data—assuming technical noise is close to Poissonian–given observed gene expression mean values, to derive the mean-variance trend expected in the absence of biological variation. Excess variance in gene expression relative to this null expectation is assigned to biological component of gene expression variance. Genes with the largest biological variance components are tagged as interesting 'highly variable genes' (HVGs; Fig. 1.12).

Efficient variance stabilization and HVG selection are paramount for downstream analyses. The latest single-cell transcriptomic technologies can produce UMI counts for millions of cells across thousands of genes in a single experimental run (Box 3). Hence, most modern scRNA-seq data sets live in enormously multidimensional spaces. For computations to remain tractable, it is necessary to reduce the dimensionality of FBMs. Filtering HVGs is one way to accomplish this, removing features that carry mostly technical noise, or very little biological information, from the matrix.

Linear algebra methods like principal components analysis (PCA) and non-negative matrix factorization (NMF) are also commonly used to reduce the dimensionality of scRNA-seq data into a smaller set of linear combinations of expressed features that best summarize the covariances in the data (Shao and Höfer, 2017; Zhu et al., 2017). The need for dimensionality reduction highlights the importance of variance stabilization. For instance, if the mean-variance relationship remains, PCA will be driven by predominantly technical variation across highly variable genes, overlooking global correlation patterns across genes with lower expression levels (Fig. 1.10) (Svensson, 2020). Performing PCA on a set of biologically relevant HVGs makes it more likely that observations are projected on a subspace of the FBM defined by interesting and informative transcripts.

Size-factor-based methods make up one of two large families of approaches to scRNA-seq data normalization (Vallejos et al., 2015; Lun et al., 2016a). The other is composed of methods based on probabilistic models of transcript counts (Kharchenko et al., 2014; Grün et al., 2014; Hafemeister and Satija, 2019). For example, Hafemeister and Satija (2019) propose using Pearson residuals from a negative-binomial (NB; Box 4) generalized linear model (GLM) of UMI counts as normalized gene expression values. More precisely, they model the number of UMIs assigned to transcript $\mathbf{g}$ across all cells with a GLM with NB-distributed error and log-link function,

$$\log\left(\mathbb{E}\left[\boldsymbol{A}_{\mathbf{g}\cdot}\right]\right) = \alpha + \beta_{\mathcal{A}\mathbf{g}} \log_{10}\left(\mathbf{L}\right), \tag{1.11}$$

where $\mathbf{L}$ is the vector of library sizes across all cells. To avoid overfitting the model to the expression profile of each gene, model parameters are regularized through kernel regression across genes using a Gaussian kernel. Pearson residuals are then computed from observed expression values $\boldsymbol{A}_{\mathbf{gc}}$, given the values expected under the regularized NB model. The authors convincingly show that their method accounts for the effect of count depth $\mathbf{L}_{\mathbf{c}}$ on the measured expression of each gene across a wide range of expression levels. In contrast, by using a one-size-fits-all correction factor across all transcripts expressed in each cell, log-normalization does not efficiently adjust the effect of count depth on the measured expression of highly active genes.

Box 4 | The sparsity of single-cell transcriptomic data. Gene expression is a naturally stochastic process (Novick and Weiner, 1957; Ko et al., 1990; Raj and van Oudenaarden, 2008; Eldar and Elowitz, 2010; Weidemann et al., 2023). Owing to this biological noise, but also to technical factors like inefficient transcript capture from the small stock of mRNA molecules present in each single cell, single-cell RNA-sequencing (scRNA-seq) data are characterized by a low signal-to-noise ratio relative to bulk RNA-seq. Collectively, these factors contribute to 'dropout' events, in which a gene that is robustly expressed in some cells is not detected in another subset of similar cells (Kharchenko et al., 2014). Due to the prevalence of these false null values, zeroes in scRNA-seq data are considered non-informative.

Stochastic gene expression is generally modeled with a 'negative binomial' (NB) distribution resulting from a mixture of Gamma and Poisson processes (Appendix B, page 193). Briefly, the NB distribution differs from a Poisson in that its dispersion differs from its mean μ, allowing for an additional overdispersion parameter ϕ. If ϕ is allowed to vary across genes, most zeroes in biological data can be explained under the NB (Svensson, 2020).

In their description of single-cell differential expression, Kharchenko et al. (2014) were the first to explicitly account for an excess of null values in scRNA-seq data through a mixture of an NB 'signal' component in which detected transcript abundance correlates with its true underlying abundance, and a Poisson 'dropout' component in which the signal is not detected. Since then, other probabilistic models for 'zero-inflated' scRNA-seq data have been proposed (Pierson and Yau, 2015; Finak et al., 2015; Risso et al., 2018).

However, recent work suggests that the prevalence of null values in droplet-based scRNA-seq data is accurately modeled by the NB distribution alone and is as expected from count data, such that 'zero-inflation' is likely to arise from biological variability (Vieth et al., 2017; Svensson, 2020).

As is often the case in scRNA-seq data analysis tasks, the best choice of normalization method depends on features of the input data, as well as on the intended downstream analyses. For instance, while log-normalization has been suggested to improve the performance of dimensionality reduction (Booeshaghi et al., 2022), Pearson residuals lead to selection of HVGs with a greater biological variance component (Lause et al., 2021). However, the method described by Hafemeister and Satija (2019) becomes intractable on data sets of over a million cells. Ahlmann-Eltze and Huber (2023) argue that log-transformation followed by principal components analysis is generally at least as effective as other more sophisticated approaches.

Batch effect correction and data set integration

The recent exponential increase in the capacity of scRNA-seq to assay single cells (Box 3) sparked large scale efforts to extensively characterize single-cell transcriptomic variation across human tissues and individuals (Regev et al., 2017; van der Wijst et al., 2020; Tabula Sapiens Consortium, 2022). Drawing biological insight from such valuable compendia requires merging information from different data sets—across distinct experimental settings, cohorts, tissues and conditions—to eliminate technical noise while preserving biologically meaningful variability. Where normalization methods focus on dampening the effects of technical confounders—chiefly cell count depth—within a data set, batch correction is concerned with technical variation between data sets.

Here again, the diversity of proposed batch-correction methods is wide, ranging from methods that leverage linear models (Butler et al., 2018; Korsunsky et al., 2019; Hie et al., 2019) to deep learning approaches (Lopez et al., 2018; Lotfollahi et al., 2019; Xu et al., 2021). While complex correction tasks often require the latter machine learning tools, linear-embedding models perform well on simpler tasks (Luecken et al., 2022; Heumos et al., 2023).

In particular, the linear method described by Korsunsky et al. (2019) is the top choice when batch structure is known and clear. Given raw PC coordinates and a set of batch variables, the tool progressively learns—through iterations of clustering and projection—a linear function specific to each cell, and outputs a set of adjusted PC-like dimensions such that cells project into clusters of mixed batch levels. These batch-corrected dimensions can then be used as input for clustering or for nonlinear embedding methods, such as uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) or t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008).

Recent years have also seen an explosion in single-cell assays of genomic features other than gene expression. For instance, single-cell assays for transposase-accessible chromatin sequencing (scATAC-seq) (Buenrostro et al., 2015), cellular indexing of transcriptome and epitopes by sequencing (CITE-seq) (Stoeckius et al., 2017) and single-cell reduced-representation bisulfite sequencing (scRRBS) for methylome analyses (Guo et al., 2013) were all introduced within the last decade. Layering different sources of information from the genome of a single cell can lead to finer biological insight (Argelaguet et al., 2021) (§ III, page 137). Yet, although some methods like CITE-seq can produce more than one data modality from the same experiment, different modalities are most often assayed independently. Hence, tools are needed to integrate these data sets together.

Single-cell transcriptomic data sets are commonly represented as graphs in which each cell is connected to its k-nearest neighbors on the gene expression manifold (Islam et al., 2011). The graph representation provides numerous computational advantages (Kharchenko, 2021). For example, UMAP and t-SNE rely on it to approximate the underlying gene expression space. Although both methods can produce visually attractive summaries of complex data sets, their intepretatibility remains shallow. Most implementations of the algorithms carry a significant stochastic component, and both methods can produce distorted approximations that do not accurately represent the gene expression manifold (Chari and Pachter, 2023).

Nearest-neighbor graph representations are also useful for integration tasks, when paired measurements are available between data sets. For example, Hao et al. (2021) used a weighted nearest-neighbor approach to project peripheral blood mononuclear cell barcodes from different assays of chromatin accessibility, gene expression and surface protein expression on a common reduced-dimension space. By transferring information across modalities, the authors were then able to extensively characterize the innate immune response to vaccination against the human immunodeficiency virus. When paired observations are not available, other methods using artificial intelligence can be used (Gayoso et al., 2021; Cao and Gao, 2022; Ashuach et al., 2023).

1.3.2 Mapping molecular quantitative trait loci in single cells

As previously mentioned, the eQTL framework appears as a viable solution to interpret GWAS results because most genome-wide associations fall in non-coding regions with regulatory potential (Hindorff et al., 2009; Maurano et al., 2012) (§ 1.2.4, page 15). Furthermore, it has been shown that genes around GWAS loci are more likely to be eGenes, and that the most significant GWAS hits are more likely to also be eQTLs (Nicolae et al., 2010). It is thus common to assume that most GWAS hits can be explained through genetically controlled changes in transcript abundances. Yet, across different studies, only between 5% and 40% of GWAS QTLs colocalize with eQTLs (Giambartolomei et al., 2014, 2018; Hormozdiari et al., 2016; Chun et al., 2017).

Connally et al. (2022) argue that while failure to colocalize is not enough to dismiss the mediator potential of transcriptional variability, most GWAS hits are actually explained by genetic variants that regulate the expression of nearby genes, but whose effects cannot be captured by classical bulk eQTL studies. The authors instead point to a 'missing regulation' layer that could be recovered through more detailed models of context-dependent gene regulation (§ 1.2.6, page 22).

A prime example of such extended models of gene regulation is showcased in recent work by Oelen et al. (2022), who mapped cell-type and stimulus-dependent eQTLs in human peripheral blood mononuclear cells (PBMCs; § 3.1.1, page 55) exposed to *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa* or *Candida albicans* for different periods. Overall, their results reflect the widespread impact of cellular and environmental context on the genetic control of the transcriptional responses to pathogens. In particular, this genetic control is primarily impacted by cell-type-dependent factors, rather than by differences across stimuli or time points.

To tease apart the downstream effects of these different layers of regulation, Oelen et al. (2022) leveraged their scRNA-seq data set to build gene 'co-expression' (coe) networks in each context (van der Wijst et al., 2018b; van Dam et al., 2018), and mapped coeQTLs that affect how eGene expression levels correlate across cells (van der Wijst et al., 2018a; Li et al., 2023). In other words, Oelen et al. (2022) used genetically-controlled transcriptional relationships to approximate the gene regulatory networks (GRNs; § 3.1.2, page 57) underlying the response to stimulation by particular pathogens in distinct PBMC types. For instance, the authors highlight the rs12230244 coeQTL of *CLEC12A*—a gene that encodes a C-type lectin-like receptor important for the regulation of inflammatory responses. Pathway analysis of the set of genes co-expressed with *CLEC12A* in each condition revealed an enrichment in genes regulated by interferon (IFN) at the early time point. Additionally, the set of genes for which the rs12230244-T allele was associated to a more positive co-expression relationship are enriched in binding sites for IFN regulatory factors (IRFs). Together, these results suggest a GRN in which stimulation by pathogens triggers an IRF-mediated IFN response, which leads to increased IRF activation and subsequent ISG induction. In this context, co-expression loss could be explained by the rs12230244-C allele disrupting an IRF binding site in regulatory regions of *CLEC12A* in the response to infection. The biological relevance of this locus is highlighted by the interaction between *CLEC12A* expression, a PRS for SLE—a disease characterized by increased IFN activity—and the genotype at rs12230244 (Oelen et al., 2022).

The coeQTL framework is especially interesting because it allows to reliably infer the chain of regulatory events underlying complex traits in specific contexts using data from relatively few individuals and conditions (van Dam et al., 2018; Li et al., 2023). However, Oelen et al. (2022) also note that coeQTL mapping does not enable causal inference of genetic effects (§ 1.2.5, page 19), and highlight the importance of integrating scRNA-seq data sets with other modalities of single-cell data to improve the interpretability of genome-wide associations with complex traits by mapping the genetic bases of molecular endophenotypes other than gene expression (Argelaguet et al., 2021).

Previous studies in bulk have shown how assays of different layers of gene expression regulation can be integrated in synergy to reach deeper insight. For instance, by coupling eQTL and caQTL mapping in human macrophages exposed to IFN-γ and/or *Salmonella enterica*, Alasoo et al. (2018) revealed pervasive regulatory 'priming' of genetically controlled transcriptional immune responses to stimulation. In over half of all colocalized caQTL-reQTL pairs in each condition, the caQTL was already detected in resting cells. In contrast, only around 10% to 25% of response caQTLs in each condition colocalized with a resting eQTL. Hence, integrating different modalities of gene expression regulation data allows to track the flow of genetic information from genotype to phenotype.

More recently, Aygün et al. (2023) mapped eQTLs and caQTLs in developing human brain cells and tested different models of causality (Box 2) between chromatin accessibility and gene expression, in order to map the genetic factors underlying complex neurological traits. In neuronal progenitors, the expression of 168 genes was significantly mediated by genetically-controlled changes in chromatin accessibility. In particular, the indel rs10717382 SNP was shown to regulate the expression of *SLC26A7* through its effect on the accessibility of an upstream locus by modifying the binding affinity of transcriptional repressor NKX2-2. Notably, the eQTL signal at rs10717382 also colocalized with the GWAS signal at the rs57117164 SNP linked to structural variation in a brain region associated to emotional recognition. Thus, these results suggest a causal path through which genetic regulation of *SLC26A7* in differentiating neurons could explain inter-individual differences in brain structure, underlying the risk of emotional disorders.

All in all, the key to retrieving the 'missing regulation' brought up by Connally et al. (2022) is likely to lie in multimodal single-cell data. Yet, although several methods exist to characterize molecular endophenotypes other than gene expression at single-cell resolution (Guo et al., 2013; Buenrostro et al., 2015; Stoeckius et al., 2017), the repertoire of single-cell molQTL mapping studies is only slowly diversifying (Benaglio et al., 2023). In fact, there is still not a single piece of published work coupling single-cell molQTL mapping across more than one modality (Cuomo et al., 2023).

Next-generation single-cell expression quantitative trait locus mapping methods

Most single-cell eQTL mapping frameworks applied to date proceed in three steps. First, single-cell transcriptional profiles are clustered on the basis of nearest-neighbor graphs that approximate the underlying highly multidimensional gene expression manifold. The clusters are then annotated on the basis of DE marker genes, or using automated annotation approaches (Clarke et al., 2021). Finally, UMI counts are aggregated—most often as a sum or average—into 'pseudobulk' values by cell type and donor to create context-specific vectors of expression $\mathbf{X}$ for each gene, that can be plugged into models like Equation (1.4) for downstream analysis steps.

Pseudobulking brings scRNA-seq data closer to the assumptions of such linear models: there is a single observation per sample and context, and the phenotype is close to normally distributed across samples (Cuomo et al., 2023). Figure 1.13 illustrates this using pseudobulk and single-cell *MMP1* expression data from the same set of SARS-CoV-2-stimulated myeloid cells. While the distribution of bulk expression values is approximately Gaussian, single cell expression values are discrete in nature and thus better modelled using zero-inflated Poisson or negative binomial distributions (Box 4; Appendix B, page 193).

While pseudobulk measurements dampen the noisiness, sparsity and discrete nature of scRNA-seq data (Box 4), they also create other challenges. For instance, the uncertainty around each pseudobulk estimate depends on the number of cells in the aggregate, which can vary widely between contexts and donors, although this problem can be easily countered by filtering out samples with a cell count below a given threshold. The aggregation also limits the eQTL mapping to the variants that are associated to changes in average gene expression values across cells from different individuals. Yet, 'dispersion' eQTLs associated to instability in cell-to-cell gene expression—but not necessarily mean changes—may also explain disease (Sarkar et al., 2019; Cuomo et al., 2023).

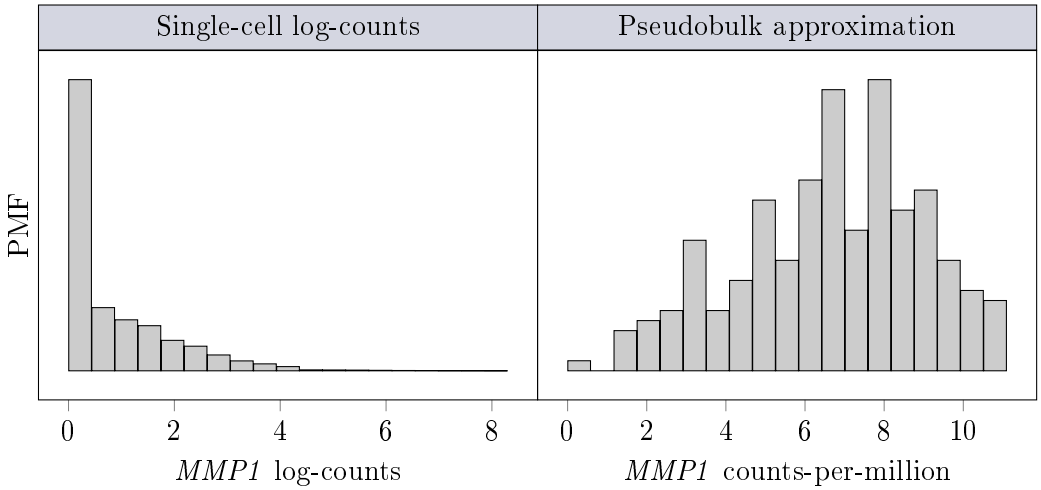


Figure 1.13 | Single-cell and pseudobulk gene expression. Probability mass function (PMF) of log-transformed pseudobulk and single-cell *MMP1* expression values in the same set of SARS-CoV-2 stimulated myeloid cells (Aquino et al., 2023). For the single-cell expression values, unique molecular identifier (UMI) counts were log-normalized following the method proposed by Lun et al. (2016a).

Finally, it has also been suggested that pseudobulk eQTL mapping altogether defeats the purpose of scRNA-seq by obscuring the heterogeneity contained within clusters of cells and by discretizing naturally continuous processes (Cuomo et al., 2023). Although the concept of bona fide single-cell eQTL mapping is not recent (Wills et al., 2013), a new generation of methods is emerging that holds the promise to more fully exploit the diversity of cell states captured by single-cell genomic data (Nathan et al., 2022; Cuomo et al., 2022).

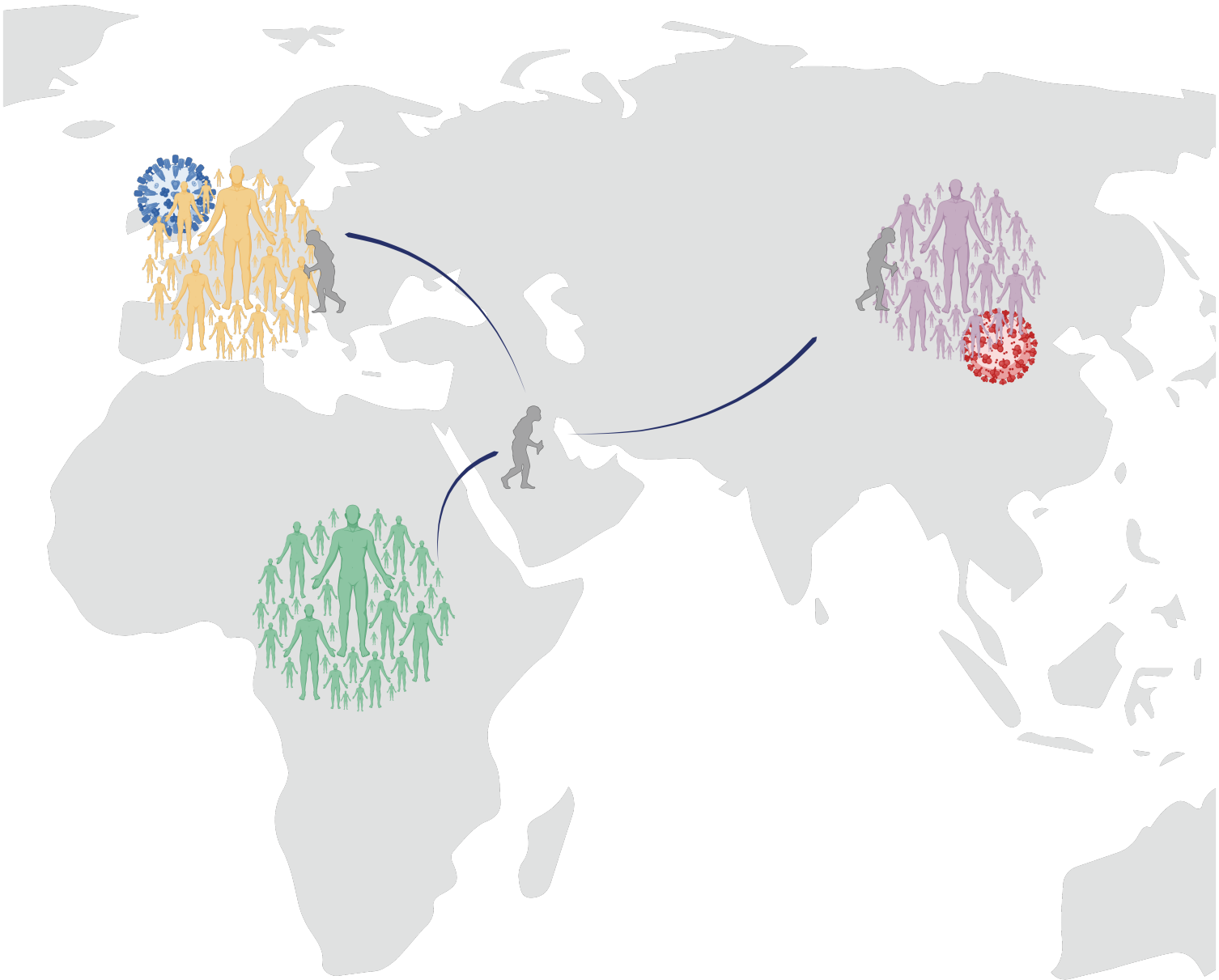
For instance, by enabling the mapping of sceQTLs along continuous trajectories, these novel methods may uncover dynamic genetic regulation mechanisms arising during activation of immune cells by pathogens, or interferon-mediated signalling in disease contexts. For example, through pseudobulk eQTL mapping in PBMCs from SLE patients and healthy controls, Perez et al. (2022) revealed dynamic genetic regulation of *SLFN5* expression, conditional on type I IFN activation—a hallmark of SLE—across eight immune cell types. Applied in a context like this, next-generation sceQTL mapping could potentially reveal other disease-relevant cell states, by characterizing the dynamic genetic regulation of gene expression at greater resolution.

However promising, these new approaches are currently limited by modeling considerations and computational tractability (Cuomo et al., 2023). For example, the assumption of independent observations in linear modelling (Eq. (1.4)) is violated by the nested structure of cell-level and individual-level observations, and more complex models are needed to account for these effects. Although pseudobulk approaches remain essential for several tasks in single-cell genomics, as the body of theoretical work (Nathan et al., 2022; Cuomo et al., 2022) and computational capabilities grow (Gewirtz et al., 2022), it is likely that the field will shift towards next-generation scmolQTL mapping methods to fully exploit single-cell genomic data and characterize the genetic bases of complex traits at unprecedented resolution.

2 Archaic introgression and modern immune responses

'People who carry this Neandertal risk variant for COVID-19 have a reduced risk of becoming infected with HIV'

– Svante Pääbo, during his Nobel Prize lecture (2022)



2.1 Anatomically modern humans and our sister species

Genome-wide association studies of complex phenotypes and molecular endophenotypes have proven useful to unveil gene regulatory networks underlying human disease (§ 1.2, page 7). Yet, these methods can only paint a static picture of immune diversity among extant humans. To fully understand how immune response differences emerged during human evolution, it is essential to characterize the evolutionary forces that shaped present-day human genetic diversity, and how these forces have acted on regulatory loci controlling the immune response (Barreiro and Quintana-Murci, 2010; Fan et al., 2016b; Nielsen et al., 2017; Quintana-Murci, 2019). In the words of Theodosius Dobzhansky (1973), 'Nothing in biology makes sense except in the light of evolution'.

2.1.1 Genetic and archæological evidence of human origins in Africa

Charles Darwin (1871) was among the first to propose an African origin of humans, based on evolutionary relationships inferred between anatomically modern humans (AMHs) and great apes endemic to Africa (Huxley, 1863). Since then, excerpts from the East African fossil record have shed more light on the evolution of the genus *Homo* (Leakey and Leakey, 1978; Wood, 1991; Bromage et al., 1995; Kimbel et al., 1997). Among the best-described early exponents of the genus, *Homo habilis* and *Homo erectus* are thought to have roamed the African continent around two million to two hundred thousand years ago (Spoor et al., 2007; Herries et al., 2020). Relative to their common *Australopithecus* ancestor, both species had larger cranial capacities and modern limb proportions, although the features of *Homo erectus* were markedly closer to AMH features.

Another interesting feature of *Homo erectus* fossils is their distribution across the globe. In fact, the first *Homo erectus* specimens were unearthed in Asia, sparking a debate about the true geographical origin of the species (Delson, 1985; Wood, 1991; Walker and Leakey, 1993; Gabunia et al., 2001). The common view today is that *Homo erectus* appeared in Africa and then expanded into Eurasia, where it evolved into other species of *Homo* (Asfaw et al., 2016).

The mode of descent linking AMHs to *Homo erectus* has been another subject of intense debate. The model of multiregional evolution (MRE) proposed that AMH populations evolved independently from the separate groups of *Homo erectus* that spread into Eurasia, sharing limited gene flow (Wolpoff et al., 1984). In contrast, the 'Out-of-Africa' (OOA) hypothesis draws upon evidence for a much more recent African common ancestor of AMHs that expanded later and replaced non-African populations of archaic hominins (Box 5) (Lewin, 1987). The MRE model was mainly supported by archæological data reflecting a continuity in the fossil records, which was at odds with the replacement predicted by the OOA model. However, archæological data alone could not inform on whether the continuity was also genetic or solely due to cultural diffusion.

Later into the 20$^{\text{th}}$ century, genetic studies of mitochondrial (mt) DNA provided strong support for the OOA hypothesis. The maternally inherited mtDNA molecule is an interesting tool because its sequence is relatively short and it evolves rapidly, allowing genealogical inference even among closely related populations (Brown et al., 1979; Giles et al., 1980). In particular, through analyses of mtDNA sequences from 147 diverse individuals, Cann et al. (1987) concluded to the presence of a common ancestor of all AMH mtDNA in Africa around 200 thousand years ago. A few years later, Quintana-Murci et al. (1999) leveraged analyses of mtDNA to propose a coastal exit route from East Africa into the Arabian Peninsula. However, mtDNA also left several unanswered questions about the origins of AMHs and their dispersal out of Africa (Nielsen et al., 2017). Because mtDNA is not recombinant, the whole molecule is akin to a single genetic variant. Hence, phylogenetic trees drawn from mtDNA sequences are blind to the bulk of genomic signatures of human evolution. Here again, the development of genotyping and sequencing tools was key to address these questions by unlocking genome-wide assessments of variation across populations (Box 1, page 4).

The theory of isolation-by-distance and the 'stepping stone' model of population structure both predict an increase in genetic differentiation between populations located further apart (Malécot, 1948; Kimura and Weiss, 1964). Ramachandran et al. (2005) showed these expected patterns across populations world-wide, through the analysis of over 700 variant loci across more than a thousand individuals in the Human Genome Diversity Project-Centre d'Étude du Polymorphisme Humain (HGDP-CEPH) cohort (Cann et al., 2002; Rosenberg et al., 2002). The authors then proposed a model for human expansion out of Africa as a series of 'founder effects' originating from a single starting point in East Africa. Each founder effect is equivalent to a population bottleneck, during which only a subset of the population of migrants—and thus only a subset of the alleles segregating in the population—move on to colonize other locations. In line with this model, Ramachandran et al. (2005) also reported a decay in genetic diversity proportional to the geographical distance separating each population from Addis Ababa, Ethiopia.

Jakobsson et al. (2008) later described increased levels of linkage disequilibrium (LD) in non-African populations world-wide, proportional to their geographical distance from Africa, in line with the model proposed by Ramachandran et al. (2005). However, recent work provides strong evidence against the assumption of a single African AMH origin (Ragsdale et al., 2023), in line with previous reports of complex ancestral population structure in Africa (Tishkoff et al., 2009).

Box 5 | Hominids and hominins. The term 'hominid' designates the set of all extant and extinct great apes, including gorillas, chimpanzees and humans. 'Hominins' are the subset of hominids that descend from the last common ancestor of chimpanzees and humans, including species with 'archaic' features relative to anatomically modern humans. Today, *Homo sapiens* is the only extant human species, but fossil records harbor a wide diversity of archaic hominins, including Neanderthal and Denisova.

The complexity of ancestral human population structure is reflected in the African fossil record. Recently, several skull and mandible fragments bearing AMH cranial and facial features were discovered at the Jebel Ihroud site in Morocco (Hublin et al., 2017). Dated at around 300 thousand years ago, the Jebel Ihroud fossils are the oldest human remains discovered to date. Before Jebel Ihroud, human remains with AMH features had been unearthed at the Herto and Omo Kibish sites—respectively dated to around 160 and 200 thousand years ago—in Ethiopia (White et al., 2003; McDougall et al., 2005). Taken together, these pieces of fossil evidence support the OOA hypothesis of AMH origin, and highlight the complex pan-African evolutionary history of humans.

2.1.2 Retracing the recent modern human expansion out of Africa

Although the studies led by Ramachandran et al. (2005) and Jakobsson et al. (2008) revealed genomic signatures predicted by the OOA model, the genetic and fossil evidence available at the time did not suffice to resolve many features of the human expansion into Eurasia, Oceania and The Americas. For instance, while Mirazón Lahr and Foley (1994) proposed multiple dispersal routes, White et al. (2003) proposed a single exit path through Ethiopia, based on their fossil discoveries in the Herto site. The timing of these events was also contended, with estimates ranging from around 40 up to 170 thousand years ago (Groucutt et al., 2015).

In 2016, a trilogy of genetic studies on nearly 800 high-quality genomes—from geographically diverse and typically understudied populations—provided novel insights of the peopling of Eurasia and Oceania (Malaspinas et al., 2016; Mallick et al., 2016; Pagani et al., 2016). Interestingly, the three studies reached different conclusions. In line with multiple waves of dispersal from Africa,

Pagani et al. (2016) stated an early migration into Australasia around 120 thousand years ago, followed by other dispersals. The two other studies concluded to a single exit event followed by different branchings. As previously proposed by Rasmussen et al. (2011), Malaspinas et al. (2016) proposed a separation between the colonizers of mainland Eurasia and Australasia. In contrast, Mallick et al. (2016) suggested a separation between West and East Eurasians, who then went on to people Oceania. In line with the reports by Malaspinas et al. (2016), Choin et al. (2021) estimated the separation between the ancestors of Eurasians and those of Papua New Guineans to around 58 thousand years ago.

All in all, there is still much to learn about AMH origins. Even though decades of research have produced archæological and genomic results predicted by the OOA hypothesis (Quintana-Murci et al., 1999; White et al., 2003; Ramachandran et al., 2005; Jakobsson et al., 2008; Hublin et al., 2017; Malaspinas et al., 2016), much uncertainty remains around the finer details of early human evolution. The broad model accepted today is that humans with anatomically modern features appeared in Africa around 300 thousand years ago, and that the most significant successful migration out of the continent happened around 60 thousand years ago.

2.1.3 Encounters with other human species in Eurasia

It is also widely accepted that *Homo sapiens* was not the first human species to venture out of Africa. Fossil and bone evidence reveals the diversity of non-African human species that likely descended from the previous *Homo erectus* expansions included in the OOA model (King, 1864; Schoetensack, 1908; Bermúdez de Castro et al., 1997; Krause et al., 2010). When AMHs arrived, Eurasia was already inhabited by at least two groups of archaic hominins: Neandertals to the west (Stringer and Hublin, 1999; Krause et al., 2007) and Denisovans to the east (Krause et al., 2010).

The type Neandertal specimen was unearthed in the eponymous valley in Germany in 1856 (King, 1864). In 1997, a team of researchers led by Swedish geneticist Svante Pääbo—pioneers in the field of palæogenetics, or the study of 'ancient' DNA—sequenced segments of mtDNA found on that specimen (Krings et al., 1997). They estimated the divergence time between the AMH and Neandertal mitochondrial lineages to around 550 to 690 thousand years ago. A decade later, Pääbo's group published a complete sequence of the Neandertal mitochondrial genome, and refined the divergence time estimation to around 660 thousand years ago (Green et al., 2008). Shortly after, the same group drew the first draft of a complete Neandertal nuclear genome from three individuals found in Vindija Cave, Croatia (Green et al., 2010).

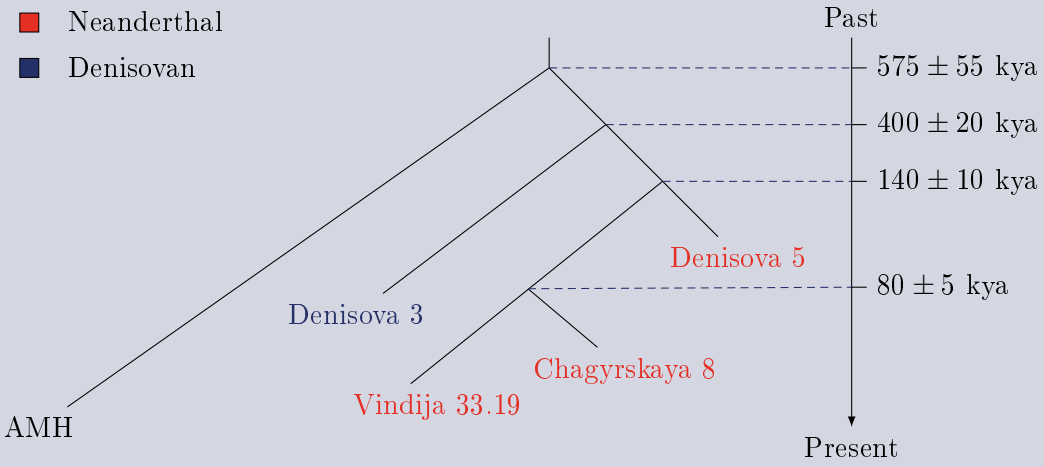
Because human nuclear genomes are mosaics of recombining DNA segments, the genome-wide comparisons led by Green et al. (2010) shed much more light on the evolutionary relationships linking AMHs to Neandertals, relative to previous single-marker analyses of mtDNA. In particular, the authors pointed out genetic variants present at high frequencies across present-day AMHs but absent in Neandertals, which showed signals suggestive of natural selection and could thus have been important in early human evolution. For instance, a negatively selected variant in the locus of *RUNX2* could contribute to differences in bone structure between Neandertals and AMHs (Green et al., 2010). On a more fundamental level, Green et al. (2010) also showed proof-of-concept for the study of full archaic hominin genomes sequenced from ancient DNA contained in bones old of several hundred thousand years.

Since 2010, three Neandertal (Prüfer et al., 2014, 2017; Mafessoni et al., 2020) and one Denisovan (Meyer et al., 2012) high-quality genome sequences have been obtained (Box 6). All but one of these genomes have been sequenced from bones found in caves of the Altai mountain range. The Denisova Cave has been particularly fruitful: Denisova 3 is the type specimen of the Denisovan group (Reich et al., 2010; Krause et al., 2010; Meyer et al., 2012), Denisova 5 provided the DNA

for the first Neandertal high-quality genome sequence (Prüfer et al., 2014)—often referred to as the 'Altai' Neandertal—and Denisova 11 was shown to descend directly from interbreeding between a Neandertal female and a Denisovan male (Slon et al., 2018).

More recently, Mafessoni et al. (2020) obtained a high-quality genome sequence from Neandertal bones found in the Chagyrskaya Cave in the Altai region. Among other interesting results, the authors report that their Chagyrskaya 8 individual lived in a population related to the Neandertal mother of Denisova 11.

Box 6 | High-quality archaic hominin genomes. As the ancestors of anatomically modern humans (AMHs) expanded out of Africa, they met populations of Neandertals and Denisovans that had inhabited Eurasia for hundreds of thousands of years. Recent advances in genomics have enabled the high-coverage sequencing of three Neandertal (Prüfer et al., 2014, 2017; Mafessoni et al., 2020) and one Denisovan (Meyer et al., 2012) genomes. The Figure below gives a schematic representation of the evolutionary relationships between these individuals and AMHs. The split times between AMHs and archaic hominins (Nielsen et al., 2017), and within archaic hominin individuals (Mafessoni et al., 2020) are rounded estimates expressed in 'thousands of years ago' (kya). The clustering of leaves in the tree reflects the true patterns of genetic similarity estimated between these individuals, but the branch lengths are not to scale.



Vindija 33.19 belongs to a female Neandertal who lived around 50 kya (Prüfer et al., 2017). It is the most commonly used Neandertal reference genome in studies of archaic introgression. Relative to the older Denisova 5 Neandertal (Prüfer et al., 2014), Vindija 33.19 shares more alleles with non-African AMHs (Prüfer et al., 2017). Furthermore, almost all Neandertal segments carried by AMH genomes are closer to Vindija 33.19 than to Denisova 5 (Prüfer et al., 2017).

Overall, by comparing ancient and modern DNA sequences, palæogenomic studies furnished key insights into the relationships among Neandertals and Denisovans, but also anatomically modern humans and their closest evolutionary relatives. In particular, they provided strong evidence to support previous claims of interbreeding between these groups, resulting in the 'archaic introgression' of segments of DNA from Neandertals and Denisovans into the genomes of anatomically modern humans (Rotival and Quintana-Murci, 2020).

2.2 Signals of archaic introgression across genomic and geographical regions

Neanderthals appear in the fossil record from approximately 400 thousand years ago (Stringer and Hublin, 1999; Hublin, 2009; Meyer et al., 2016) to around 40 thousand years ago (Finlayson et al., 2006; Higham et al., 2014), ranging from England and Spain (Stringer and Hublin, 1999; Arsuaga et al., 2014) to Central Asia (Krause et al., 2007) and as far south as the Middle East (Valladas et al., 1987). The geographical and temporal overlap between the distributions of AMHs and Neandertals raises the possibility of interbreeding between the two species.

Early contentions against AMH-Neandertal admixture were based on anatomical (Bräuer et al., 2006; Bailey et al., 2009) and genetic (Currat and Excoffier, 2004) observations. In particular, it was advanced that Neandertals could not have contributed to the genetic make-up of AMHs because variation in Neandertal mtDNA was not included in the pool of present-day human genetic diversity (Krings et al., 1997; Serre et al., 2004; Orlando et al., 2006).

These views were challenged by the first genome-wide analysis of Neandertal nuclear DNA. Namely, Green et al. (2010) estimated around 2% of Neandertal ancestry in the genomes of Eurasian AMHs, who shared more alleles with Neandertal than AMHs from sub-Saharan Africa, suggesting at least one interbreeding event in the non-African AMH lineage. Sankararaman et al. (2012) then suggested admixture between Neandertal and AMHs to have taken place somewhere in the Middle East around 47 to 65 thousand years ago, based on archæological evidence and analyses of LD in present-day AMH populations. A few years later, fossil evidence from Manot Cave in Israel confirmed that AMHs were in the right place at the right time to interbreed with Neandertal, around 50 to 60 thousand years ago (Hershkovitz et al., 2015).

More recently, human remains from Peştera cu Oase in Romania have revealed direct links between Neandertals and AMHs (Fu et al., 2015). Namely, the Oase 1 mandible was shown to have belonged to an individual with a direct Neandertal ancestor four to six generations before them. Also, genome-wide analyses from the remains of three individuals found in Bacho Kiro Cave, Bulgaria, revealed recent contributions from Neandertal ancestors six to seven generations in the past (Hajdinjak et al., 2021).

The picture that emerges from these and other results reviewed elsewhere (Lalueza-Fox, 2021) is one of assimilation of archaic hominin groups into populations of modern humans. When they expanded out of Africa, AMHs encountered Neandertal and Denisovan populations spread across the Eurasian continent, with whom they admixed. These species of archaic hominins went extinct shortly after the arrival of anatomically modern humans, but their legacy remains in the genomes of present-day humans. Importantly, these archaic genetic variants are not inert: through effects on the regulation of gene expression (Silvert et al., 2019), they contribute to the phenotypic diversity across present-day humans, including immune differences in the response to pathogens (Deschamps et al., 2016; Sams et al., 2016; Quach et al., 2016; Zeberg and Pääbo, 2020, 2021).

2.2.1 Detecting events of archaic introgression in modern human genomes

The initial assessment of Neandertal contribution to modern non-African genomes by Green et al. (2010) based on the draft Neandertal genome sequence was later refined on the basis of the high-coverage Neandertal genome obtained from Denisova 5 (Box 6). Prüfer et al. (2014) estimated the proportion of Neandertal ancestry in modern genomes to vary between 1.48% and 1.96% in Europeans, and from 1.64% to 2.14% in East Asians and Native Americans. Prüfer et al. (2017) then showed that virtually all the Neandertal-origin nucleotide sequences found in modern genomes were closer to the high-coverage Vindija 33.19 genome sequence than to Denisova 5. Based on this new reference, the proportion of Neandertal ancestry in modern genomes was estimated between 1.8% and 2.4% in Western Eurasia, and 2.3% to 2.6% in East Asia (Prüfer et al., 2017).

Although the status of Denisovans as a separate species of hominin is still debated, studies of Denisovan and Neandertal genetic diversity suggest independent population histories for these two groups of archaic hominins (Reich et al., 2010; Meyer et al., 2012). These studies and others have also highlighted differences in the Denisovan contribution to present-day human genomes. Although there are signs of gene flow from Denisovans into AMH populations in South and East Asia (Skoglund and Jakobsson, 2011; Qin and Stoneking, 2015; Vernot et al., 2016; Browning et al., 2018), the highest proportions of Denisovan ancestry are found in Oceania (Malaspinas et al., 2016; Choin et al., 2021), reaching up to 5% in some populations (Reich et al., 2010).

The varying patterns of archaic ancestry across modern populations world-wide may inform on the number of 'pulses' of gene flow from Neandertals and Denisovans into AMH genomes. For example, Browning et al. (2018) explain the higher rates of Denisovan introgression in the genomes of East Asians relative to South Asians through two separate pulses of archaic gene flow. The authors suggest a first event of introgression into the genome of the common ancestor of South and East Asian populations, followed by another pulse after the separation of the South and East Asian lineages. The first Denisovan component—found in modern South and East Asian genomes—appears to have descended from a population distantly related to the Altai Denisovans (Box 6). By contrast, the second Denisovan component is specific to East Asian genomes and shows strong genetical similarity with Denisova 3 (Browning et al., 2018).

Even though Denisova 3 is the only reference genome available to date, Browning et al. (2018) were able to find the South Asian Denisovan component using an S*-statistic that does not rely on an archaic reference genome. In general, S*-statistics rely on the expectation that (i) the time to the most recent common ancestor (TMRCA) between archaic introgressed and non-introgressed haplotypes in a modern genome is longer than the TMRCA between non-introgressed haplotypes, and that (ii) due to the relatively recent timing of archaic admixture, introgressed segments are less likely to be broken down by recombination, relative to older sequences that have remained polymorphic over long time periods due to incomplete lineage sorting (ILS). Thus, archaic variants stand out in long divergent haplotypes bound by strong LD; assuming that AMHs and archaic hominins interbred around 60 thousand years ago, the expected length of archaic segments is around 50 kilobases (Browning et al., 2018).

S*-statistics are designed to capture the set of variants in LD within a fixed-width—generally 50-kilobases wide—sliding genomic window that maximizes a pairwise scoring function of genotype distances. In the first implementation by Vernot and Akey (2014), variants j and $j+1$ are scored

$$S(j, j+1) = \begin{cases} -\infty, & d(j, j+1) > 5, \\ -10^4, & d(j, j+1) \in \{1, \cdots, 5\}, \\ 5000 + bp(j, j+1), & d(j, j+1) = 0, \\ 0, & j = \max(J), \end{cases} \tag{2.1}$$

where $bp(j, j+1)$ is the number of base pairs separating the two variant loci and

$$d(j, j+1) = \sum_i |GT(i, j) - GT(i, j+1)|, \tag{2.2}$$

measures the distance between their genotypes in individual i, and $GT(i, \cdot) = \{0, 1, 2\}$. For each subset J of the whole set of putative archaic introgressed variants V_i in a given genomic window in individual i, the value of the statistic is then computed as

$$S(J) = \sum_{j=1}^{\max(J)-1} S(j, j+1). \tag{2.3}$$

Finally, the value of S^* in the genomic region for individual i is assigned as the maximum value obtained across all subsets J in V_i,

$$S^*_i = \max_{J \in V_i} S(J). \tag{2.4}$$

Although useful, this scoring function scales poorly; Vernot and Akey (2014) applied it iteratively on subsets of twenty individuals (Balding et al., 2019).

Browning et al. (2018) implemented a new scoring function that follows the same basic principle, but allows to take local mutation and recombination rates into account and scales better over larger sample sizes (Balding et al., 2019). The statistical significance of high S' values is assessed against a null S' distribution obtained through simulations of demographic models with no archaic interbreeding. Hence, the error rates of S'-statistic tests depend on the availability of well-calibrated demographic models, as well as precise estimates of mutation and recombination rates. Computing S' also requires defining a 'target' population in which putative archaic segments segregate and a 'reference' population—usually an African reference panel—expected not to have interbred with archaic hominins. The reference population is useful to prevent confounded introgression signals due to ILS of a variant shared between the target and the reference populations.

Using their implementation of the S^*-statistic, Browning et al. (2018) detected around 1.4 gigabases of archaic introgressed DNA segments across 19 panels in the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) (§ 1.1.2, page 3). Across individuals in each panel, Puerto Ricans had the lowest proportion of archaic ancestry, at around 0.80%, and Han Chinese the highest, at around 1.23% (Browning et al., 2018).

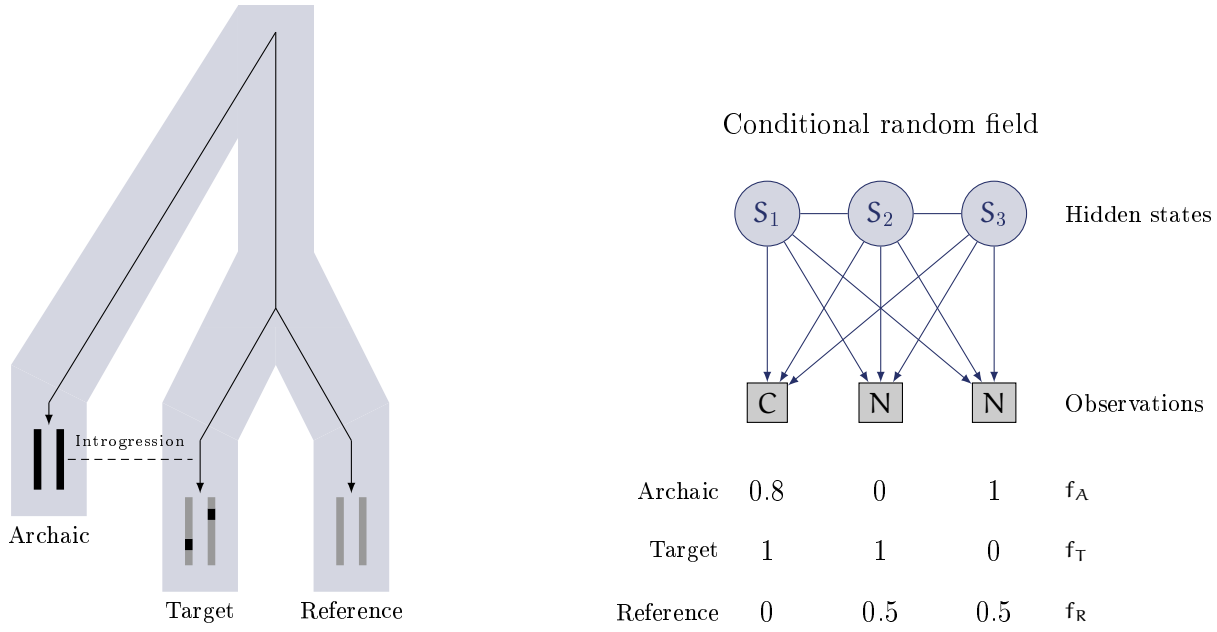


Figure 2.1 | Conditional random fields to detect archaic introgression. The schematic phylogenetic tree on the left illustrates the expectations shared by methods to detect archaic introgression. First, the time to the most recent common ancestor (TMRCA) between archaic introgressed and non-introgressed haplotypes in a modern genome should be longer than the TMRCA between two non-introgressed haplotypes. Second, these methods rely on the presence of a 'reference' population for which admixture with archaic hominins is not expected. The right panel shows a simplified conditional random field (CRF) model that aims to infer the 'archaic' or 'modern' hidden state of three variants S_1, S_2 and S_3, from observed states that are 'consistent' (C) or 'inconsistent' (N) with archaic introgression on the basis of derived allele frequencies in the archaic, target and reference populations. Horizontal lines represent 'transition' functions that capture linkage disequilibrium (LD) between SNPs; vertical and diagonal lines are 'emission' functions that capture the relationships between states. Adapted from Balding et al. (2019).

Besides summary S*-statistics, the other large family of methods to detect archaic introgression is composed of formal probabilistic models that explicitly use information from archaic genomes. For instance, Sankararaman et al. (2014) suggested using a conditional random field (CRF) model to infer the hidden 'archaic' or 'modern' state of single nucleotide polymorphisms (SNPs) along the genome, by comparing the states observed in a target (T), reference (R) and archaic (A) population. Briefly, the state of a SNP is defined as 'consistent' (C) or 'inconsistent' (N) with archaic introgression depending on the derived allele frequencies f. in these populations. If

$$f_T - f_R = 1 \cap |f_A - f_T| < 1, \tag{2.5}$$

the allele is ancestral in the test population, absent from the reference and present in the archaic genome; its state is consistent with archaic introgression. In contrast, if

$$0 < f_R < 1 \cap |f_A - f_T| = 1, \tag{2.6}$$

the allele is present in the reference population, and either absent from the target or absent from the archaic population; its state is inconsistent with archaic introgression (Balding et al., 2019).

As illustrated in Figure 2.1, CRF models involve two types of functions. The horizontal lines in the graph represent the 'transition' functions used to capture LD between SNPs. The vertical and diagonal lines represent 'emission' functions that capture relationships between observed and hidden states. In contrast to hidden Markov models (Prüfer et al., 2014), inference of each hidden state in CRF models is not based solely on the current observed SNP state, but also depends on states observed before and after in the sequence (Balding et al., 2019).

Sankararaman et al. (2016) applied the CRF framework using different emission functions to (i) model the allelic patterns at each SNP and to (ii) recover signals of Neandertal and Denisovan ancestries across multiple SNPs (Balding et al., 2019). Across 257 high-quality genomes from the Simons Genome Diversity Project (Mallick et al., 2016), they detected a total of 257 megabases of Denisovan-origin DNA in Oceanian populations and 673 megabases of Neandertal-origin DNA in non-African populations. Interestingly, Sankararaman et al. (2016) also found that introgressed segments from Denisovans into Oceanian genomes were on average longer than those introgressed from Neandertal, suggesting that these populations interbred with Denisovans more recently in evolutionary history. More recently, Choin et al. (2021) shed light on the complex genetic interactions between Oceanians and highly structured groups of archaic hominins, leveraging both CRF models and the S*-statistic implementation from Browning et al. (2018).

Regarding gene flow from Neandertals into non-Africans, there are still many open questions. Since the earliest estimates of archaic ancestry, modern East Asian genomes have been attributed a higher Neandertal proportion relative to Europeans (Prüfer et al., 2014, 2017). More recently, Browning et al. (2018) estimated Neandertal ancestry to be around 30% higher in East Asian panels from the 1000 Genomes Project, relative to Europeans. The difference could be explained through a single admixture pulse in the ancestor of Europeans and East Asians around 40 to 50 thousand years ago (Sankararaman et al., 2012; Hershkovitz et al., 2015; Moorjani et al., 2016), followed by dilution of Neandertal ancestry in Europeans through subsequent migration of unadmixed individuals out of Africa (Browning et al., 2018). However, it could also be that the purge of Neandertal alleles from East Asian genomes through natural selection was less efficient due to smaller population sizes (Keinan et al., 2007; Sankararaman et al., 2014) or longer generation times (Coll Macià et al., 2021).

Other authors have proposed more complex histories of Neandertal-AMH admixture outside Africa, using models that incorporate multiple pulses of archaic introgression (Vernot et al., 2016; Villanea and Schraiber, 2019). In particular, Villanea and Schraiber (2019) state that dilution of

Neandertal ancestry in European AMH populations by influx of non-admixed individuals is not sufficient to explain its current distribution in Eurasia, and propose a model in which an initial pulse of Neandertal genomic material was supplemented by other admixture events. In line with previous reports supporting small effective sizes across Neandertal populations (Prüfer et al., 2014; Castellano et al., 2014), Villanea and Schraiber (2019) further suggest that the Neandertals that admixed with AMHs were closely related and lived in a restricted geographical region; the present-day differences in Neandertal-ancestry proportions across Eurasian populations would reflect the length of time during which their ancestors lived in that region, in contact with Neandertals. Yet, Villanea and Schraiber (2019) also state that their model cannot explain why East Asian genomes in particular have such a high Neandertal contribution.

Recently, Iasi et al. (2021) suggested that simple-pulse models such as used by Sankararaman et al. (2012) and Moorjani et al. (2016) are not adapted to date admixture with Neandertals because they assume that gene flow happened in a narrow time frame. Furthermore, whereas the different Denisovan genetic components in present-day human genomes helped infer multiple gene flow pulses from Denisovans (Browning et al., 2018; Choin et al., 2021), multiple-pulse Neandertal admixture inference is complicated because practically all Neandertal segments in modern genomes are most similar to Vindija 33.19 (Prüfer et al., 2017) (Box 6). In this context, Iasi et al. (2021) propose an extended pulse model of Neandertal admixture, and report that modern genome data from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) is compatible with gene flow spread over hundreds of generations and centered around 49 thousand years ago. The authors also discuss the importance of collecting more ancient DNA data from Neandertals living around the time of admixture to refine the time frame of gene flow into modern human genomes, and ascertain the number of pulses and their length.

2.2.2 Archaic introgression and the adaption to new environments

Whether it descends from Neandertals or Denisovans, archaic ancestry is generally deleterious in modern genomic backgrounds (Harris and Nielsen, 2016; Juric et al., 2016; Petr et al., 2019; Rotival and Quintana-Murci, 2020). For instance, Simonti et al. (2016) report robust associations between the risk of depression and Neandertal-origin variants picked up genome-wide by S*-statistic tests. While some argue that deleteriousness could arise from epistatic interactions between archaic and modern genetic variants (Sankararaman et al., 2014, 2016), others suggest that deleterious variants could have survived in Neandertal genomes because of inefficient natural selection (Harris and Nielsen, 2016) due to low effective population sizes (Prüfer et al., 2014; Castellano et al., 2014) (Appendix C, page 199).

In any case, the non-uniform distribution of Neandertal and Denisovan-origin segments in modern genomes suggests that archaic introgressed DNA evolved under strong purifying selection (Sankararaman et al., 2014, 2016; Vernot et al., 2016). Through simulations, Harris and Nielsen (2016) estimated that—if deleterious mutations have an additive effect on fitness—strong selection against Neandertal-AMH hybrids could bring the Neandertal admixture proportion from 10% to between 2% and 3% in just around 20 generations, at which point all AMH individuals would have a similar fraction of archaic ancestry, only at different genomic loci.

Using CRF models, Sankararaman et al. (2016) uncovered 18 regions longer than 10 megabases and with a Neandertal ancestry proportion below 0.001 in non-African genomes, and 63 such regions depleted in Denisovan ancestry in the genomes of present-day Oceanians. Overall, these archaic ancestry 'deserts' were found preferentially in conserved and gene-enriched regions of the modern human genome, consistent with a role for negative selection purifying the genome from archaic variants with a deleterious effect on phenotype (§ 2.1, page 37).

However, there is also overwhelming evidence for archaic variants that were preserved in the genomes of AMHs—and even sometimes brought to relatively high frequencies—by natural selection, owing to their beneficial effects on phenotype (Sankararaman et al., 2014, 2016; Sams et al., 2016; Quach et al., 2016; Gittelman et al., 2016; Racimo et al., 2017; Browning et al., 2018; Rotival and Quintana-Murci, 2020; Zeberg and Pääbo, 2021).

According to early estimates from samples of African, Eurasian and Oceanian populations, only around 20% of high-frequency archaic haplotypes—that is, falling beyond the 99$^{\text{th}}$ percentile of the empirical allele frequency distribution—contain protein-coding variants, suggesting that most functional archaic variants were selected for their regulatory effects on gene expression (Gittelman et al., 2016). For example, Sams et al. (2016) revealed robust signals of positive selection on a Neandertal haplotype spanning the *OAS1-3* locus—coding for 2'-5' oligoadenylate synthetase (OAS) sensors of viral nucleic acids—in the genomes of non-Africans (Mendez et al., 2013), and bearing expression quantitative trait loci (eQTLs; § 1.2.4, page 15) for *OAS1*, *OAS2* and *OAS3*, as well as splicing QTLs for *OAS1* and *OAS2*. The authors also highlight the rs10774671 splice variant of *OAS1*, which is common in African and non-African genomes, but found almost strictly in Neandertal-like haplotypes in non-Africans, suggesting that this variant could have been reintroduced into AMH genomes by archaic introgression following its loss during the OOA migration (Sams et al., 2016).

Namely, Sams et al. (2016) describe a likely case of adaptive archaic introgression, through which admixture with archaic hominins helped AMHs recover a part of genetic diversity lost through the OOA bottleneck, and adapt to new environmental pressures during their expansion from Africa (Fan et al., 2016b; Gittelman et al., 2016; Racimo et al., 2017).

Other known examples of adaptive archaic introgression include Neandertal variants that affect skin pigmentation (Dannemann and Kelso, 2017), hair structure (Sankararaman et al., 2014) and lipid metabolism (Racimo et al., 2017), but also immune functions through genes like *CCR9*, *CXCR6* (Gittelman et al., 2016; Browning et al., 2018), *TLR1* (Deschamps et al., 2016; Dannemann et al., 2016), and the aforementioned *OAS1-3* (Mendez et al., 2013; Sams et al., 2016). In turn, variants introgressed from Denisovans seem to primarily affect immune traits, through genes like *OAS1* (Mendez et al., 2012), *STAT2*, *IRF4* and *TNFAIP3* (Browning et al., 2018; Choin et al., 2021).

2.3 Viral pathogens as drivers of human evolution

Beyond their contribution to adaptive introgression, pathogens have played a pervasive role in human evolution (Barreiro and Quintana-Murci, 2010; Quintana-Murci and Clark, 2013; Fan et al., 2016b). In particular, selective pressures imposed by pathogenes have contributed to shape present day human genetic diversity through adaptive allele frequency changes at immune-related loci (Haldane, 1949; Karlsson et al., 2014; Quintana-Murci, 2019) (Appendix C, page 199).

Depending on whether alleles are advantageous or detrimental for a particular phenotype, two types of directional natural selection are defined. While negative selection tends to decrease the frequency of deleterious alleles, alleles that enable a better adaption to the local environment will tend to increase in frequency under positive selection. Several evolutionary models of positive selection exist, including classic selective 'sweeps' in which a novel and markedly advantageous allele takes over the population, selection on 'standing' pre-existing variation that becomes advantageous after a change in the environment, and polygenic adaptation through the joint effect of several loci.

2.3.1 Evolutionary relevance of human interactions with viruses

In general, genes that accomplish essential functions do not tolerate variation well—as any variant that inactivates the gene will lead to a stark loss in fitness—and thus evolve under strong negative selection (Quintana-Murci, 2019). Interestingly, Deschamps et al. (2016) showed that genes

involved in innate immunity (§ 3.1, page 55) evolve under stronger negative selection relative to randomly sampled protein-coding genes. Several essential mediators of antiviral immunity have also been shown to evolve under strong purifying selection, including *TLR3*, *TLR7*, *TLR8* and *TLR9* (Barreiro et al., 2009), *STAT1* (Deschamps et al., 2016) and interferon γ (Manry et al., 2011).

Likewise, Enard et al. (2016) report that human proteins that physically interact with viruses also evolve under stronger purifying selection relative to those that do not interact with viruses, further emphasizing the importance of interactions with viral pathogens during human evolution. Remarkably, the authors show that even virus-interacting proteins (VIPs) with no clear antiviral role, but which are involved in core cellular functions like transcription, are enriched in strong signals of adaptation. Overall, using genomic data from over 20 mammal species, Enard et al. (2016) estimate that around 30% of adaptive amino acid changes in the deeply conserved human proteome are due to viral pressures, highlighting viruses as major drivers of human adaptation.

Enard and Petrov (2020) further conforted this view through analyses of modern genomes from 26 panels from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) (§ 1.1.2, page 3). In particular, the approximately 45 hundred VIP loci in the human genome appear to be strongly enriched in signals of selective sweeps across these diverse populations (Box 7). Furthermore, the authors specify that these recent sweeps seem to have been driven preferentially by interactions with RNA viruses, but not DNA viruses.

This is especially interesting because RNA viruses have been implicated in most of the major known zoonoses in recent human history, including the West African Ebola virus epidemic of 2013, the 1981 human immunodeficiency virus (HIV) outbreak, as well as influenza A virus (IAV) pandemics (Kreuder Johnson et al., 2015; Geoghegan et al., 2017), and zoonotic diseases account for more than 60% of human infectious diseases (Taylor et al., 2001). In contrast, the majority of DNA viral pathogens would have spilled over to humans deeper in human evolutionary history (Geoghegan et al., 2017), and are generally less virulent today.

Taken together, these results pinpoint RNA viruses as major drivers of human adaptation (Enard et al., 2016; Enard and Petrov, 2018, 2020), and emphasize the relevance of evolutionary genetics approaches to study natural selection in understanding immune trait differences across present-day human populations (Barreiro and Quintana-Murci, 2010; Quintana-Murci, 2019).

2.3.2 Detecting evolutionary events of human adaption to respiratory viruses

The importance of these questions was most recently highlighted by the ongoing 'coronavirus disease 2019' (COVID-19) pandemic. COVID-19 characterizes symptomatic infection by a novel strain of coronavirus—a family of RNA viruses—triggering a 'severe acute respiratory syndrome' (SARS-CoV-2). Recent analyses point to several likely SARS-CoV-2 spill-over events in a section of the Huanan Seafood Wholesale Market dealing with live wild animals (Pekar et al., 2022).

The most recent estimates from the World Health Organization place the global COVID-19 death toll at nearly 7 million lives (World Health Organization, 2020a). Yet, the risk of severe and potentially lethal COVID-19 forms is not uniformly distributed across human populations. For instance, even when cultural and socio-economic factors are duly accounted for, African-American individuals in the United States are almost twice as likely to require hospitalization following a positive COVID-19 test, relative to European-Americans (Shelton et al., 2021). These and other results reviewed elsewhere (§ 3.2.2, page 62) highlight a role for genetic predictors of population differences in COVID-19 risk (COVID-19 Host Genetics Initiative, 2020, 2021, 2022, 2023).

It is also known that some of these genetic factors came into AMH genomes through archaic introgression. For instance, Zeberg and Pääbo (2020) report that the strongest genetic association with COVID-19 hospitalization risk known to date (Ellinghaus et al., 2020; Kousathanas et al., 2022)

is driven by a Neandertal haplotype around 50 kilobases in length and overlapping immune-relevant genes like the aforementioned *CCR9* and *CXCR6*. This haplotype is present in the genomes of around 50% of South Asians and around 16% of Europeans [a], and is associated to a 1.6 odds-ratio of requiring hospitalization relative to healthy controls (Zeberg and Pääbo, 2020).

Zeberg and Pääbo (2021) also describe a 75-kilobase Neandertal haplotype in the *OAS1-3* cluster previously highlighted by Mendez et al. (2013) and Sams et al. (2016), that is present in around 30% of Eurasian genomes and is associated to a 22% reduction in risk of developing severe COVID-19 forms relative to reported COVID-19 cases.

Although a systematic evaluation of the impacts of archaic introgression and natural selection on population disparities in COVID-19 risk was lacking at the time (Aquino et al., 2023), previous studies of VIP loci had already revealed the widespread role of Neandertal introgression in AMH adaptation to viral pressures outside Africa (Enard and Petrov, 2018, 2020).

The poison-antidote model of adaptive archaic introgression

Enard and Petrov (2018) report an enrichment of VIP loci in long and frequent Neandertal segments identified in Eurasian genomes by Sankararaman et al. (2014), as well as in previously identified targets of adaptive archaic introgression (Gittelman et al., 2016; Racimo et al., 2017; Jagoda et al., 2018), relative to non-VIP loci. Interestingly, the enrichment is even stronger when focusing on RNA-VIPs in the genomes of Europeans. Out of the 11 RNA viruses considered by Enard and Petrov (2018), IAV and HIV had the highest counts of identified VIP loci—1,500 and 1,171 respectively. These loci showed strong enrichments in frequent and long archaic introgressed segments in European genomes, particularly when considering virus-specific loci in genomic regions with high recombination rates, consistent with the model of adaptive archaic introgression.

Box 7 | Virus-interacting proteins in the genetics toolkit. Virus-interacting protein (VIP) loci are useful tools to study human adaptation to past viral pressures. Of the 5,291 VIP loci considered in Souilmi et al. (2021), 36% are 'high-confidence' hits ascertained through manual curation of the virology literature that report low-throughput molecular interaction methods, such as yeast two-hybrid assays (Enard et al., 2016; Enard and Petrov, 2018). The remaining 64% of VIP loci were identified using high-throughput assays, including mass-spectrometry-based methods.

VIP loci are enriched in natural selection targets, reflected in their lower average rate of nonsynonymous mutations relative to non-VIP loci (Enard et al., 2016). They are also enriched in regions of the genome dense in coding, regulatory and conserved elements (Enard and Petrov, 2018). In line with their evolutionary constraint (Luisi et al., 2015), VIPs are involved in central hubs of protein-protein interactions more often than non-VIPs (Dyer et al., 2008; Halehalli and Nagarajaram, 2015), reflecting their functional relevance. Yet, it has also been shown that under positive directional selection, VIP loci support molecular evolution rates up to three times higher than non-VIP loci (Uricchio et al., 2019). Finally, because phylogenetically close viruses interact with similar sets of host proteins, present-day VIP data can be used to infer interactions between human ancestors and ancient viruses (Enard et al., 2016).

a. The archaic rs35044562-A haplotype has a frequency of $f = 0.081$ in the 'Utah residents with Northern and Western European ancestry' (CEU) reference panel of the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2022), and is thus expected to be carried by $f^2 + 2f(1-f) = 15.5\%$ of individuals with similar ancestry.

Enard and Petrov (2018) place these observations in line with a 'poison-antidote' model in which interactions between Neandertal and AMHs led to an exchange of pathogens between the two human species, but also to gene flow through which each group received genetic factors adaptively selected to withstand infection by those pathogens. Following introgression, the advantageous Neandertal haplotypes swept over AMH genomes, driven by positive directional selection. Overall, Enard and Petrov (2018) estimate that around 30% of all long and frequent—100-kilobase or longer and present in at least 15% of the population—Neandertal introgressed segments in Eurasian genomes may have been positively selected soon after admixture.

In a general sense, the poison-antidote model of Neandertal-AMH admixture provides a unified framework to explain how archaic introgression can accelerate human adaptation to viral pressures, encompassing previously described instances of adaptive archaic introgression at immune-relevant loci (Sams et al., 2016; Deschamps et al., 2016; Quach et al., 2016). From another viewpoint, it also emphasizes the role of the viral pathogens themselves as drivers of archaic introgression and human adaptation through VIP loci (Enard and Petrov, 2018, 2020).

Remnants of an ancient coronavirus epidemic in anatomically modern human genomes

In the context of the COVID-19 pandemic, Souilmi et al. (2021) applied the VIP framework (Box 7) to retrace specific events of adaptation to coronaviruses in the evolutionary histories of human populations world-wide. More specifically, the authors focused evolutionary genetics approaches on the subset of 420 coronavirus (CoV) VIP loci in the genomes of hundreds of diverse individuals across the 26 panels of the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015). Over 300 of these CoV-VIPs were shown to interact with SARS-CoV-2 by mass-spectrometry assays (Gordon et al., 2020), including its main entry receptor: the angiotensin-converting enzyme (ACE) 2 (Zhou et al., 2020a). The remaining proteins are reported interactants of other coronaviruses linked to previous respiratory syndrome epidemics, such as the previous severe acute respiratory syndrome epidemic of 2002 (SARS-CoV-1) and the 2012 outbreak in the Middle East (MERS-CoV).

Using haplotype-based methods (Appendix C, page 199), Souilmi et al. (2021) found strong enrichments in signals of selective sweeps at CoV-VIP loci, but only in the genomes of East Asian populations. Remarkably, the enrichment was found to be private to coronaviruses when compared to VIPs for 17 other viruses. Based on these observations, the authors suggest that adaptation could have been driven by hypothetical past epidemics caused by ancient coronaviruses—or related viruses that interact similarly with host proteins—pressuring the ancestors of modern East Asians.

To test this hypothesis, Souilmi et al. (2021) dated the putative adaptation events using novel methods enabling the approximation of full-likelihood models to detect natural selection (Speidel et al., 2019; Stern et al., 2019), based on ancestral recombination graphs (Appendix C, page 199). The authors report 42 CoV-VIPs that could have become the targets of natural selection between 21 and 27 thousand years ago. Relative to randomly sampled genes matched for statistical evidence of selective sweeps, the clustering of CoV-VIP selection events in this particular period is very unlikely due to chance (peak significance $\mathrm{p} = 1.1 \times 10^{-4}$). Thus, these results are in line with the emergence of a coronavirus-related viral pressure in East Asia around 25 thousand years ago. This time frame is particularly interesting, because it coincides with the appeareance of the ancestor of the Sarbecovirus—a subgenre of β-coronavirus that contains SARS-CoV-1 and 2—around 23 thousand years ago (Ghafari et al., 2021).

In support of the biological relevance of these results, 50% of the 42 putatively selected CoV-VIPs are implicated in biological pathways related to viral infection, versus 29% for the whole set of 420 CoV-VIPs in the human proteome. Furthermore, some of these loci have been associated to inter-individual differences in COVID-19 susceptibility and severity, and others include known molecular targets of drugs against COVID-19 (Souilmi et al., 2021).

2.3.3 Human evolutionary history and precision medicine

The COVID-19 pandemic has placed evolutionary genetics in the limelight as a framework to understand differences in infectious disease risk between modern-day individuals. All humans alive today descend from the survivors of past environmental pressures that contributed to shape human genetic diversity into its present state. Hence, modern human genomes can be exploited through evolutionary genetics approaches to shed light on the genetic architecture of complex disease traits, and thus inform on current diease risk disparities across populations world-wide (Barreiro and Quintana-Murci, 2010; Quintana-Murci, 2019; Sella and Barton, 2019).

Evolutionary genetics and the genome-wide association study framework

In fact, evolutionary genetics theory is intimately related to the genome-wide association study (GWAS) framework (§ 1.2, page 7) used to map the genetic bases of complex traits, including disease risk (Sella and Barton, 2019). Briefly, because the contribution of any genetic variant to the heritability of a trait is determined by its effect on the trait—measurable through a GWAS—and the frequency of the effect allele, a full interpretation of GWAS results implies characterizing the evolutionary forces that shaped current allele frequency patterns (Appendix D, page 205). Indeed, there is evidence that past demographic and natural selection events in human evolutionary history have shaped the genetic architecture of common diseases in ways that affect GWAS success rates, as well as the proportion of trait heritability explained by these studies (Uricchio, 2020).

In this context, several hypotheses have been proposed to explain the part of 'missing heritability' (§ 1.2.3, page 10) between family-based study and GWAS estimates (Manolio et al., 2009), including overestimation of heritability by the former (Zuk et al., 2012; Ruby et al., 2018), lack of GWAS power to detect the low effects of the multitude of variants simultaneously affecting highly polygenic traits [b] (Yang et al., 2011) and a significant contribution of rare variation with large effects to complex trait heritability (Marouli et al., 2017).

Importantly, each of these hypotheses implies a statement about the evolution of complex traits (Uricchio, 2020). First, an overestimation of trait heritability by family-based genetic association studies may be symptomatic of incorrect assumptions in the model of additive genetic inheritance, whereby the distribution of a trait in a population could evolve through interactions between alleles and/or with their environment (Jelenkovic et al., 2016). Second, high polygenicity suggests a large mutational target of genomic predictors of complex traits, in line with the 'infinitesimal' model of polygenic inheritance proposed by Fisher (1918) (Appendix A, page 190), and the more recent 'omnigenic' hypothesis (Boyle et al., 2017). Third, the biased contribution of rare variants to heritability may reflect recent episodes of population expansion followed by negative selection purifying the genome from large-effect deleterious variants (Lohmueller, 2014).

Through simulations under various population genetics models, Lohmueller (2014) shows that recent population expansions (Keinan and Clark, 2012; Gao and Keinan, 2014) can lead to higher proportions of rare nonsynonymous SNPs—relative to non-expanded populations—with potentially deleterious effects. Moreover, if the impact of these mutations on reproductive fitness correlates with their effect on complex traits, the part of additive genetic variance in the trait explained by rare variation, as well as the number of variants that contribute to the trait, will also be greater in expanded populations; yet, the corresponding decrease in per-site heritability will lower the power to detect these contributions through a GWAS (Lohmueller, 2014). Thus, even if the heritability of a given trait is constant across populations, the power to detect its additive genetic component may vary depending on the recent demographic history of each population, which should be taken into account in GWAS designs (Lohmueller, 2014; Mathieson, 2021).

b. Although saturated maps of polygenic traits (Yengo et al., 2022) have been built using more recent approaches (Yang et al., 2010, 2011) to estimate heritability across millions of unrelated individuals (Appendix A, page 190).

In line with past work by John B. S. Haldane (1937) on the effect of variation on fitness, Lohmueller (2014) also reports that alleles that appear in the population during the simulated expansion have a lower mean effect on phenotype, relative to alleles that appeared in non-expanded populations during the same time frame, consistent with a role for negative selection purifying the genome from strongly deleterious alleles. Importantly, by purging variants with large effects and sizeable contributions to the heritability of a given trait, negative selection may lead to increased polygenicity for that trait (O'Connor et al., 2019). The idea is that negative selection reduces the contribution to heritability of large-effect variants by decreasing their frequency (Appendix D, page 205), but weak-effect variants that escape negative selection are allowed to become common in the population (Eyre-Walker, 2010; Shi et al., 2016; Zeng et al., 2018). Thus, a large number of such variants with individually weak effects on phenotype will be required to reach similar heritability. O'Connor et al. (2019) refer to this phenomenon as the 'flattening' of the heritability signal.

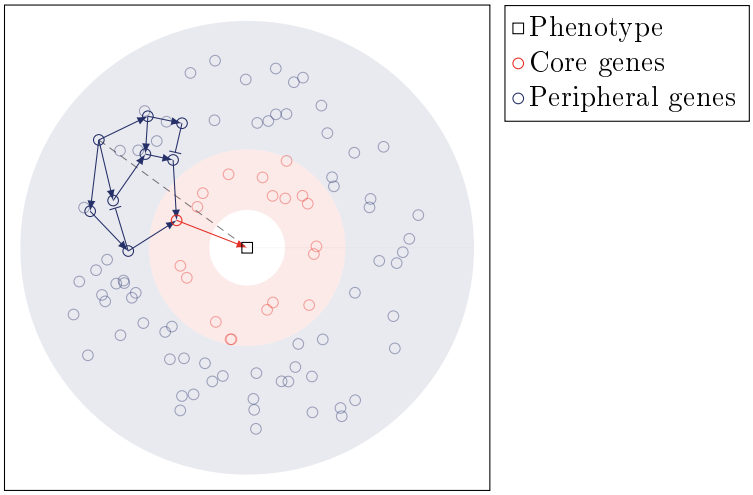


Figure 2.2 | A schematic of the model of omnigenic inheritance. The model of omnigenic inheritance assumes that the gene regulatory network (GRN) underlying a given trait is sufficiently interconnected so that all genes expressed in a given context are susceptible to regulate each other (Boyle et al., 2017; Liu et al., 2019). The omnigenic GRN is composed of several layers of genes: while 'core' genes affect the phenotype directly, the indirect effect of 'peripheral' genes must be mediated by other genes, potentially located in other peripheral layers. The solid lines in this toy example highlight the causal path between a peripheral gene and the phenotype. The dashed line shows the relationship as seen through the lens of a genome-wide association study. Adapted from Liu et al. (2019).

One suprising feature of polygenic disease traits is that the weak-effect variants that jointly explain most heritability do not seem to cluster around genes involved in pathways that are likely biologically relevant for the trait (Boyle et al., 2017). In fact, the contribution of each chromosome to trait heritability seems to be roughly proportional to its length (Visscher et al., 2006; Shi et al., 2016), consistent with a uniform distribution of polygenic trait predictors along the genome (Boyle et al., 2017). For instance, Loh et al. (2015) estimate that more than 70% of megabase-length genomic windows contribute to the heritability of schizophrenia, a very highly polygenic trait (Shi et al., 2016). Together, these results highlight the need for a new classification of the genetic predictors of disease traits that goes beyond plausibly relevant biological pathways.

Assuming that the gene regulatory network (GRN; § 3.1.2, page 57) underlying each complex trait is sufficiently dense so that all genes expressed in a given context regulate each other, the omnigenic model of inheritance proposes an alternative clustering of trait-associated genes and variants, based on their contributions to trait heritability rather than by biological pathways (Boyle et al., 2017). In its original formulation, the omnigenic model is represented as a GRN with multiple layers. Where the expression of genes in the 'core' layer directly affects the expected value of the trait in a population, the indirect effect of 'peripheral' genes on phenotype is mediated through regulation of core genes (Boyle et al., 2017; Liu et al., 2019).

For illustration, Figure 2.2 shows a schematic of the omnigenic model, highlighting the causal path between a gene in the outermost peripheral layer and variation in a phenotype. Figure 2.2 only highlights a single path, but the idea is that the expression of core genes can be affected—ever so minutely—by a large number of peripheral genes, which have a joint substantial effect on the distribution of the trait in the population (Liu et al., 2019). These individual peripheral effects can impact gene regulatory functions like transcription, protein degradation and post-translational modifications. Importantly, in contrast to 'sensory' GRNs—such as the one presented in Figure 3.2 (§ 3.1.2, page 57)—the regulatory role of extracellular receptors is neglected in the omnigenic GRN, so that receptor-encoding genes can also be classified as core genes (Liu et al., 2019).

This new classification has been shown to easen interpretation of GWAS hits around core genes of polygenic traits for which the underlying biology is well characterized (Sinnott-Armstrong et al., 2021). Yet, because few genes are expected to affect phenotype directly, the omnigenic model predicts that the overall heritability explained by core variants will be limited, and most of the variability in any particular trait will be explained by peripheral genes with regulatory function (Boyle et al., 2017). This is in line with the prevalence of GWAS hits in non-coding regions of the genome (Hindorff et al., 2009) and the large overlap between non-coding GWAS hits and regions of active chromatin (Maurano et al., 2012; Sella and Barton, 2019).

Remarkably, the omnigenic framework also provides insights into the limited overlap observed between GWAS QTL and eQTL variants. While some authors have suggested that the missing link between GWAS and eQTL mapping may reside in the context-specificity (§ 1.2.6, page 22) of eQTL effects (Barreiro et al., 2012; Fairfax et al., 2012, 2014; Lee et al., 2014; Çalışkan et al., 2015; Westra et al., 2015; Nédélec et al., 2016; Quach et al., 2016; Aran et al., 2017; Ishigaki et al., 2017; Kim-Hellmuth et al., 2017, 2020; Piasecka et al., 2018; Schmiedel et al., 2018; Ye et al., 2018; Fairfax et al., 2014; Umans et al., 2021; Yazar et al., 2022), others have shown that the contribution of currently mapped context-specific eQTLs remains modest (Connally et al., 2022). Mostafavi et al. (2022) propose a complementary explanation for this 'missing regulation' by showing that GWAS hits often fall near transcription factor genes, genes with complex regulatory landscapes—and thus expected to make a significant contribution to heritability in the omnigenic framework—and genes under strong selective constraints, while eQTLs are predominantly found near promoters of genes with simpler regulatory schemes that do not seem to be under selection. That is, GWAS QTL and eQTL mapping studies would be fundamentally biased towards discovering different types of variants—those with discernible functional effects on organismal traits in the former case, and those that significantly affect gene expression in the latter—largely due to the effects of natural selection on the genetic architecture of complex traits (O'Connor et al., 2019; Mostafavi et al., 2022).

Altogether, these and other results reviewed elsewhere (Sella and Barton, 2019; Uricchio, 2020) highlight the importance of evolutionary genetics approaches in understanding the current genetic architecture of polygenic traits, as well as the variability in GWAS success rates across complex diseases and human populations. These insights are key to achieve accurate and powerful GWAS designs in the quest to establish a precision medicine, as population-specific genetic architectures limit the ability to predict disease risk across populations world-wide (Martin et al., 2017, 2019).

Evolutionary genetics, polygenic risk scores and precision medicine

Precision medicine is an emerging paradigm in healthcare that seeks to tailor prevention and treatment strategies to the innate and acquired features of each individual through the integration of genetic, environmental and lifestyle data. A major point of interest in this framework is to develop scores able to predict disease risk across individuals and populations based solely on genetic information (Khera et al., 2018). Briefly, the standard polygenic risk score (PRS) for a given complex disease is a linear combination of the genotypes at independent GWAS loci associated to the disease

in a population, weighted by the strength of the genotype-phenotype link measured by the GWAS (Choi et al., 2020). Hence, if a PRS is applied to a 'target' population that is genetically distant from the 'training' population in which the GWAS was performed, differences in LD structure and allele frequency distributions can affect the translability of the estimates, and generally decrease the accuracy of the PRS (Martin et al., 2017, 2019). These considerations are especially important given the Eurocentrism of current human genomics research (Sirugo et al., 2019) (§ 1.1.1, page 3).

Using population genetics models of different demographic and adaptive scenarios, Durvasula and Lohmueller (2021) show how private evolutionary histories can create population differences in the genetic architecture of complex traits, which ultimately lead to PRS accuracy losses proportional to the genetic distance between populations. Across 37 complex traits in the United Kingdom Biobank (UKB) (Ollier et al., 2005), the authors estimate that European-specific variants contribute up to around 80% of the heritability of traits associated to mutations—across a wide range of minor allele frequencies—with deleterious effects on fitness (Lohmueller, 2014), suggesting that for some traits, PRS accuracy losses may stem from low sharing of causal variants across populations.

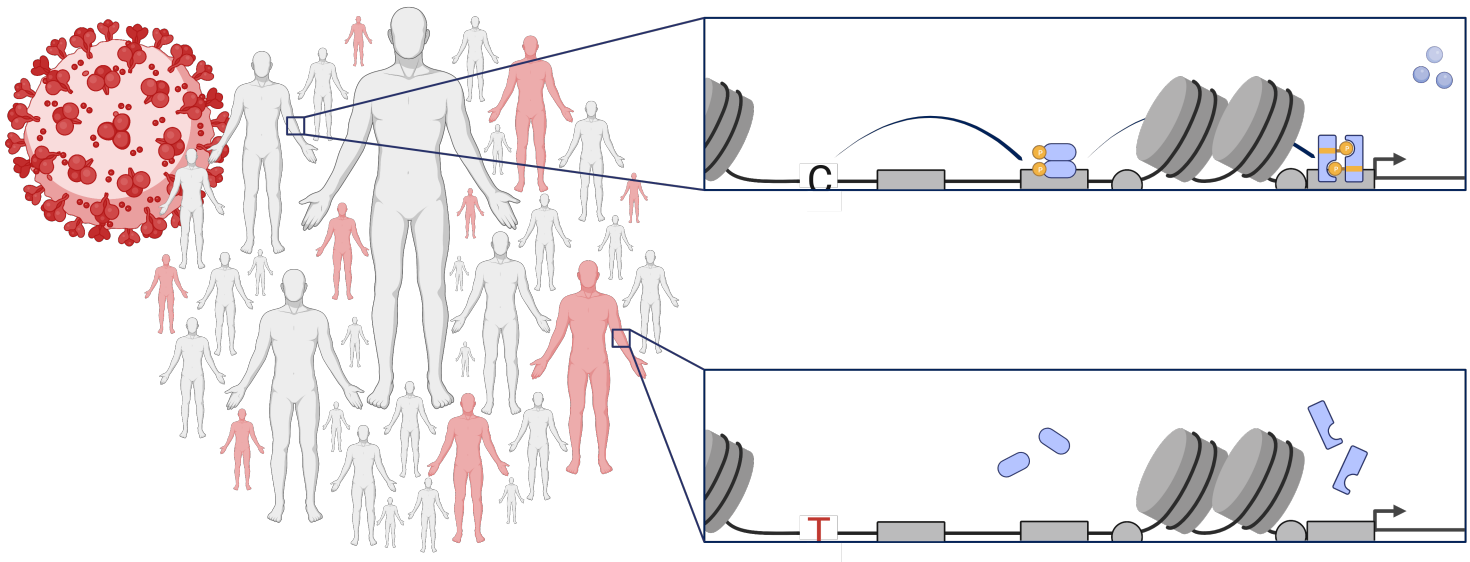
To further characterize the low PRS transferability, Mathieson (2021) proposes an extension of the omnigenic model (Fig. 2.2) that includes environmental factors as nodes in the network. In this case, the ultimate effect of a peripheral gene on phenotype is the result of its interactions with other genetic variants in a given environment. Importantly, while the wiring of the omnigenic network underlying a complex trait is likely to vary across individuals and populations, the GWAS framework is blind to this complexity (Fig. 2.2). This is not a problem when a GWAS is performed on individuals from the same population—with little genetic and environmental substructure—such that node values are drawn from the same distribution and the expected effect of the variant is the same (Mathieson, 2021). However, comparison of genetic effects across populations may be confounded by changes in the peripheral layers of the network, even if the direct effect of the core genes on the phenotype is the same (Mathieson, 2021). Hence, the 'omni-environmental' model may explain why, while many GWAS loci replicate across populations, effect sizes do not correlate well.

For instance, through genome-wide analyses of over a hundred traits across 179 thousand donors of East Asian origin from the BioBank Japan (BBJ) Project (Nagai et al., 2017), Sakaue et al. (2021) replicated the direction of over 90% of 2,305 associations reported in UKB, although the Pearson's correlation coefficient across all effect size estimates was only $\mathrm{r} = 0.11$. These and other results reviewed by Mathieson (2021) suggest that, alongside population differences in LD and allele frequency patterns, variability in genetic effect sizes—potentially driven by variation in genetic and environmental interactions in the omnigenic network—can indeed explain low PRS transferability.

These predictions of the extended omnigenic model carry important implications for future PRS applications (Mathieson, 2021). Specifically, if the variability across populations of estimated genetic effect sizes linked to a complex trait is predominantly driven by underlying differences in the omni-environmental network—rather than by population differences in LD structure and allele frequencies—this will limit the efficacy of statistical approaches that aim to improve PRS transferability through fine-mapping (Weissbrod et al., 2022) or incorporation of local ancestry (Wang et al., 2023a). While integration of data from the target population in the training set can yield 'multi-ethnic' PRSs that translate better (Márquez-Luna et al., 2017) when the confounding interactions are genetic, variation in environmental factors across populations sets an upper bound on PRS transferability (Mathieson, 2021). For some traits and in some contexts, the only way to produce an accurate PRS could be to train it on data from the target population, or from a genetically close population in a similar environment (Mathieson, 2021). Thus, in the pursuit of a precision medicine that is accessible to all, it is essential to consider the evolutionary forces that shape the genetic architectures of complex diseases across human populations world-wide, as well as to diversify data bases of genetic associations to disease (Zhou et al., 2022; Wang et al., 2023b).

3 Healthy variability and inborn errors of immunity

'No man is an island, entire of itself;
every man is a piece of the continent,
a part of the main'
– John Donne, while battling the presumably infectious disease that claimed his life (1624)



3.1 The first hours of the immune response to viruses

Any assessment of variability requires (i) observing a variable metric and (ii) defining the entities across which this metric fluctuates. In the study of variation in human immune responses to viral infection, both of these conditions are fulfilled by core features of the immune system. Indeed, the capacity of the human immune system to respond to infection implies the definition of each individual as a discrete and cohesive unit, distinct from the pathogen that triggers the response. That is, the response to infection is controlled by the cellular and molecular components of the immune system, which are also paramount in defining biological individuality: in distinguishing 'Self' from 'Not-self' (Burnet, 1969; Pradeu and Carosella, 2006).

The different responses mounted by the immune system to protect Self from Not-self have been historically divided in two classes (Owen et al., 2013). On the one hand, 'innate' immunity comprises the physical boundaries of the organism, as well as elements that provide short-term and non-specific protection. Its deep conservation—from drosophila to humans—attests to its essentiality (Hoffmann et al., 1999). On the other hand, 'adaptive' immunity is aquired through repeated encounters with pathogens; the immune system learns to provide long-term and specific protection against them. Although the distinction is not as clear-cut in truth—for example, innate immune elements can control adaptive immunity (Janeway, 1989)—this schematic view of immunity remains useful when dealing with the complexity of the human immune system.

Because innate immunity is 'the first line of defense' against infection, and less dependent on the nature of a given stimulus than its adaptive counterpart, it appears as the more appropriate aspect of immunity to focus studies of inter-individual and population immune variation on.

The innate boundaries of Self are organized in multiple layers—both at the macroscopical and microscopical scales—to confer protection from a wide range of external and internal agents. From an anatomical point of view, the first barriers to infection by external agents are the epitheliæ and mucosæ (e.g., nasal and intestinal) that cover the bodily surfaces in direct contact with the environment. Behind the frontline, a wide variety of immune cells patrol the blood and the tissues to maintain homeostasis (Owen et al., 2013).

3.1.1 Peripheral blood mononuclear cell responses to viral stimulation

Hematopoiesis yields a very heterogeneous set of immune cell types from a common stem cell state. The first step in this process of differentiation separates the myeloid and lymphoid lineages of immune cells (Owen et al., 2013). The myeloid lineage is mainly divided into granulocytes and phagocytes, such as the various monocyte, macrophage and dendritic cell (DC) subsets. Lymphoid cells include B, $CD4^+$ T and $CD8^+$ T lymphocyte populations, as well as natural killer (NK) cells. While the activity of adaptive cells depends on the activity of the RAG1-RAG2 recombinase, innate cells do not undergo somatic recombination (Patin et al., 2018)

Myeloid cells are most commonly linked to innate immunity because they are the most apt to detect incoming pathogens, and coordinate an appropriate response from lymphoid cells and other myeloid cell types (Shi and Pamer, 2011). From this perspective, the architecture of the innate immune response follows a two-tier schematic of myeloid sensors and lymphoid effectors.

Although particular types of immune cells can have very specific functions—such as the role of eosinophil granulocytes in combatting parasitic worms—the coordinated action of multiple cell types is a key feature of an effective immune response. Hence, efficient cell-cell communication is paramount. In general, cells in close proximity to each other—for example, battling pathogens in the same focus of infection—can interact via juxtacrine signalling through membrane-bound ligands and receptors. For cells that are further apart, small and soluble cytokine proteins mediate paracrine signalling. In fact, cytokines can even establish long-range endocrine signalling between cells in different anatomical structures via the blood.

For instance, interferons (IFNs) are cytokines of major relevance in infectious disease. Since their discovery almost 70 years ago, many types of IFNs have been described (Isaacs and Lindenmann, 1957; Borden et al., 2007). Among these, the type I IFN-α and IFN-β are master coordinators of the immune response against viruses, through induction of the expression of IFN-stimulated genes (ISGs). In turn, ISG products participate to combat the pathogen by interfering with viral replication, enhancing the activity of myeloid and lymphoid subsets, and recruiting immune cells from the circulation into the focus of infection (McNab et al., 2015; Bourdon et al., 2020).

In this context, blood plays a crucial role in transporting immune cells and molecules through the body. In fact, because it is also a readily accessible tissue, studies of human immunity often rely on samples of peripheral blood. In particular, peripheral blood mononuclear cells (PBMCs) are widely used to model the immune response. PBMC types include major lymphoid populations of B, $CD4^+$ T, $CD8^+$ T and NK cells, but also monocytes and DCs.

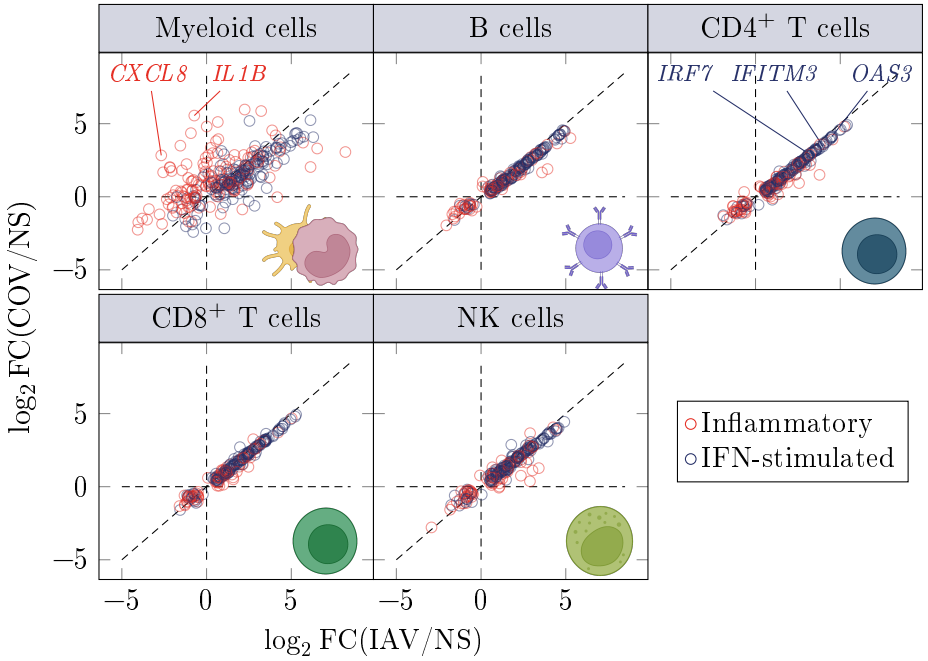


Figure 3.1 | Peripheral blood mononuclear cell responses to viral stimulation. The x and y axes respectively show the $\log_2$-fold-change (FC) in gene expression after six hours of stimulation by influenza A virus (IAV) or SARS-CoV-2 (COV), relative to the non-stimulated (NS) condition, for genes of hallmark inflammatory or interferon (IFN)-stimulated antiviral pathways, acros five types of peripheral blood mononuclear cells (Aquino et al., 2023). Genes not annotated to these pathways are not shown, as well as genes whose change in expression was below the minimum effect size of interest ($|\log_2 \mathrm{FC}| \leqslant 0.5$) or not significantly different from zero at a 1% false-discovery rate (Student's two-sided t-test adjusted $p > 0.01$) .

Importantly, although the PBMC subset does not entirely capture the diversity of immune populations in whole blood, it contains all the actors required for a complete transcriptional immune response. For example, Figure 3.1 shows the $\log_2$-fold change in gene expression following six hours of stimulation by 'severe acute respiratory syndrome' coronavirus 2 (SARS-CoV-2; COV) or influenza A virus (IAV) in five types of PBMCs, for genes associated to canonical inflammatory and IFN-stimulated responses. Notably, while the ISG response to both viruses is highly correlated across all cell types, myeloid cells display a SARS-CoV-2-specific component driven by the expression of well-known pro-inflammatory cytokines like interleukin (IL)-1β and IL-8, respectively encoded by *IL1B* and *CXCL8*. This is especially interesting because both cytokines have been pointed out as biomarkers of the severity of the 'coronavirus disease 19' (COVID-19) triggered by SARS-CoV-2 infection (Lee et al., 2020; Li et al., 2021). Hence, the *in vitro* PBMC model is able to capture the transcriptional immune responses to both viruses (Lee et al., 2020), including complex features such as the myeloid inflammatory response to SARS-CoV-2 described *in vivo*.

3.1.2 Gene regulatory networks of the immune response

When the barriers of Self are breached, invading microbes are sensed by cells—monocytes, granulocytes and DCs, but also epithelial cells—bearing pattern-recognition receptors (PRRs) that bind pathogen-associated molecular patterns (PAMPs), as well as damage-associated molecular patterns (DAMPs) on debris (Owen et al., 2013). Different PRRs are found in different cellular compartments, where they recognize different motifs. For example, Toll-like receptors (TLRs) 1 and 6 ornament the cell membrane, where they recognize bacterial and fungal PAMPs. In contrast, TLR3 and TLR8 are found in endosomes, where they are most likely to encounter PAMPs linked to intra-cellular pathogens, like viral nucleic acids (Takeda et al., 2003).

Upon binding of a PAMP or DAMP, PRR activation sparks cytoplasmic signalling cascades that transmit and amplify the external signal in order to mount an appropriate response by regulating protein activity. Information is transduced from one protein to the next through post-translational modifications catalyzed by enzymes. For instance, in protein phosphorylation—the most common post-translational modification in eukaryotes—protein kinases use adenosine triphosphate (ATP) as a substrate to covalently bind a phosphate group to particular aminoacid residues. The transfer of chemical energy primes the phosphorylated protein into a different 'active' or 'inactive' form.

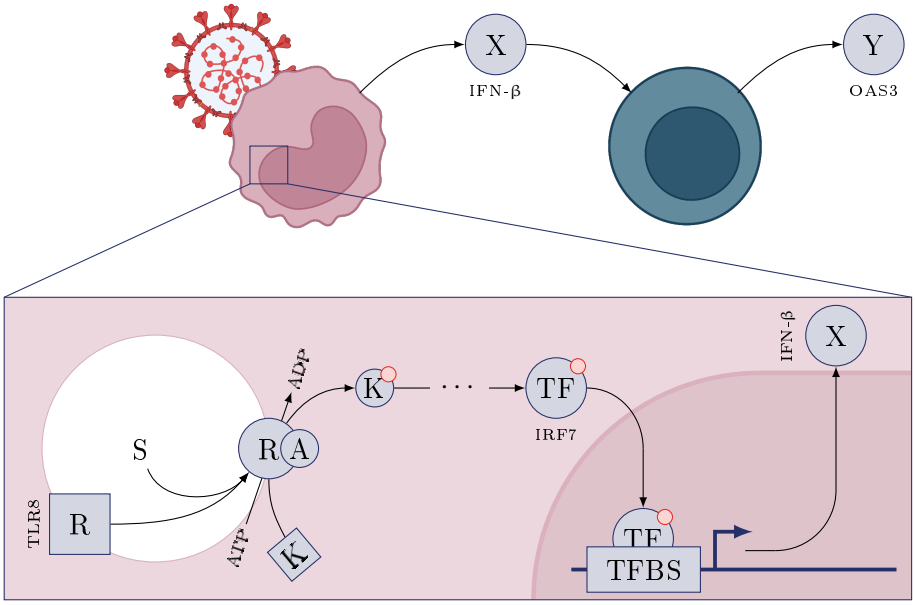


Figure 3.2 | Two tiers of immune gene regulatory networks. Monocytes can sense viral single-stranded RNA through endosomal Toll-like receptor (TLR) 8. Upon detection of the viral stimulus (S), TLR8 changes into an active conformation, able to recruit adaptor (A) protein MyD88. This sparks a cascade of phosphorylation events that results in the activation of several transcription factors (TFs). Among them, interferon regulatory factor (IRF) 7 induces the expression of several antiviral genes (X), including those that encode type I interferon (IFN) α and β. IRF7 induces expression through recruitment of the transcriptional machinery to the proximity of its targets, by binding particular transcription factor binding site (TFBS) nucleotide motifs. The IFN-α and IFN-β secreted by monocytes can then signal other cells to activate gene regulatory networks (GRNs), to induce expression of interferon-stimulated genes like *OAS3*, for example. ATP, adenosine triphosphate; ADP, adenosine diphosphate.

Signalling cascades can also affect protein activity by regulating gene expression through the action of one or more *trans*-regulatory factors (TRFs). In general, TRFs play on the probability that the transcriptional machinery is recruited around a *cis*-regulatory element (CRE) locus at which they bind DNA (Appendix B, page 193). In particular, transcription factors (TFs) recognize and tether to specific TF binding site (TFBS) nucleotide motifs. TFBSs may be found on promoter sequences directly upstream of transcription start sites, but they may also lay several kilobases away from the genes whose expression they regulate. These distal loci are commonly referred to as 'enhancers' of transcription. Promoters and enhancers are other examples of CREs.

Relationships between TFs and genes are not private: a single TF can regulate the expression of several genes and vice versa. The upshot of this regulatory promiscuity is that TF activity can integrate complex signals from the environment into the wiring of gene regulatory networks (GRNs) to ellicit nuanced cellular responses. Through the binding of TFBSs in the promoters or enhancers of several different genes, a relatively small set of TFs can activate intricate gene expression programs in response to changes in the environment of the cell, such as the appereance of a pathogen.

The archetypal GRN is composed of an environmental input S that is detected by a receptor R which, through the action of one or more adaptor proteins A, triggers a cytoplasmic signalling cascade ending with the activation of a TF that regulates the output of one or more targets X. For example, Figure 3.2 illustrates a simplified GRN of the myeloid response to viral single-stranded RNA. Upon sensing of viral single-stranded RNA in the endosomes, TLR8 sparks a series of phosphorylation events that ultimately results in the activation of several major immune-relevant TFs, including interferon regulatory factor (IRF) 7. Phosphorylated IRF7 is translocated into the nucleus where it regulates the rate of expression of type I IFNs and pro-inflammatory cytokines. In turn, these secreted mediators can go on to activate GRNs in other cells, for example resulting in the induction of ISGs like *IFITM3*, *OAS3*, or *IRF7* itself (Schneider et al., 2014).

In fact, the correlation patterns observed in Figure 3.1 are in line with a tiered architecture of the innate immune response to viruses. High correlations between the inflammatory and ISG responses to both viruses in lymphoid cell types (Pearson's $\mathrm{r} > 0.96$, $\mathrm{p} < 2.2 \times 10^{-16}$) could result from these cells reacting to IFN molecules in the medium rather than to the viral particles themselves. In contrast, monocytes and DCs are better suited to detect the viruses, which could explain the heterogeneity of viral responses in myeloid cells (Pearson's $\mathrm{r} = 0.64$, $\mathrm{p} < 2.2 \times 10^{-16}$).

The impact of genetic variation in gene regulatory networks

All GRN interactions between CREs and TRFs involve chemical reactions. The rates at which these reactions occur depend on the concentrations of the interacting partners, as well as on the affinity between them, which itself depends on their physico-chemical properties. For example, the probability that a TF tethers to a TFBS motif depends on the properties of the aminoacid sequence in the DNA-binding domain of the TF. This sequence is itself encoded in the gene that produces the TF. Thus, genetic variation at the TFBS CRE, but also in the *trans*-regulatory element (TRE) that encodes the TF, can lead to variability in the expression of genes targeted by these regulators of transcription (Flynn et al., 2022) (Appendix B, page 193).

Genetically-controlled differences in gene expression can be detected through the mapping of expression quantitative trait loci (eQTLs). As previously mentioned (§ 1.2.4, page 15), two types of eQTLs are distinguished depending on the number of base pairs separating each variant from its target 'eGene'. On the one hand, *cis*-eQTLs are associated to the expression of genes within a megabase-width window around them, on the same chromosome. On the other hand, *trans*-eQTLs regulate the expression of eGenes further away, possibly on another chromosome.

From this view, while variants of TFBS motifs in promoter sequences can yield *cis*-eQTL signals, a *trans*-eQTL associated to the expression of multiple genes in a coordinated program may arise from genetic variation in a TF gene, or in other TREs encoding upstream actors of GRNs. For example, Piasecka et al. (2018) report the rs4833095 *trans*-eQTL in the *TLR1/6/10* locus as a master regulator of the expression of over a hundred genes—including inflammatory cytokines *IL1B*, *IL6* and *IL12B*—in whole blood, following stimulation by *Escherichia coli*.

Genetic variants associated to immune differences between individuals provide the substrate for natural selection to adapt immune transcriptional programs to pathogen-related pressures (Quintana-Murci, 2019) (§ 2.3, page 46). For example, Barreiro et al. (2009) revealed signatures of selection at most of the known TLR loci in the human genome. The authors highlight the essentiality to

host defense of intracellular sensors of nucleic acids—TLR3, TLR7, TLR8 and TLR9—that evolved under the strongest purifying selection. In contrast, cell-surface-bound TLR1, TLR2, TLR4, TLR5, TLR6 and TLR10 display more relaxed selective constraints, probably reflecting the overlap in their respective sensor roles (Takeda et al., 2003; Quintana-Murci, 2019).

Importantly, the redundancy in surface TLR roles creates opportunities for natural selection: Barreiro et al. (2009) also report signals of positive selection at the *TLR1/6/10* locus known to harbor master regulators of immune responses (Quach et al., 2016; Piasecka et al., 2018), but only in the genomes of non-Africans. This illustrates how past local adaptation to pathogens contributed to shape the patterns of modern human genetic diversity across the globe (Appendix C, page 199).

3.2 Predictors of immune variability across healthy individuals

The role of pathogens (Fumagalli et al., 2011), and viruses in particular (Enard and Petrov, 2018, 2020), as drivers of human adaptation through natural selection is strongly supported by a large body of literature (Barreiro and Quintana-Murci, 2010; Karlsson et al., 2014; Fan et al., 2016b; Quintana-Murci, 2019) (§ 2.3, page 46). In this context, recent studies of human immune variability have focused on the genetic basis of transcriptional variability in the response to viral infection (Nédélec et al., 2016; Quach et al., 2016; Randolph et al., 2021). In particular, Randolph et al. (2021) used single-cell RNA-sequencing (scRNA-seq; §1.3.1, page 26) on PBMCs sampled from 90 individuals of African and European descent and exposed IAV for 6 hours, revealing a wide network of cell-type specific genetic ancestry effects on the early immune response. Overall, the authors estimate that across all genes differentially expressed between African and European-descent individuals, and that show evidence of local genetic control, *cis*-eQTLs explain over 50% of the variance in expression differences associated to genetic ancestry.

However, Randolph et al. (2021) also point out that only around half of the genes with genetic-ancestry-associated expression levels show evidence of local genetic control, suggesting that other predictors of population gene expression differences—including unmapped *cis*-eQTLs, *trans*-eQTLs and environmental factors—covary with genetic ancestry and impact innate immune parameters.

3.2.1 Genetic and nongenetic drivers of natural immune variability

Combining DNA genotyping with flow cytometry in whole blood samples from a cohort of a thousand individuals (Thomas et al., 2015) stratified across biological sex and age groups spanning five decades of life—from 20 to 70 years old—Patin et al. (2018) identified smoking, age, sex and latent infection with cytomegalovirus (CMV; Box 8) as the main non-genetic drivers in immune parameters across nominally healthy individuals.

Focusing on non-genetic intrinsic factors, Patin et al. (2018) report an increase in the proportion of $CD16^+$ monocytes with age, which might contribute to the establishment of low-grade, chronic and sterile inflammatory states associated to ageing (Franceschi et al., 2018). The authors also report a decrease in the proportions of naïve $CD4^+$ and $CD8^+$ T cells, which respectively fall at a rate of 1.6% and 3.6% per year. Regardless of age, female sex is associated to lower counts of activated NK cells, but a higher number of 'mucosal-associated invariant' T (MAIT) cells. Finally, CMV seropositivity was associated to a 12.5-fold increase in $CD4^+$ 'effector memory re-expressing CD45RA' (EMRA) T cells, and a 4.6-fold increase of $CD8^+$ EMRA T cells (Patin et al., 2018).

Out of 39 lifestyle and demographic variables—including past infections and vaccinations, as well as educational attainment—smoking status was the only non-genetic extrinsic factor significantly associated to changes in immune parameters (Patin et al., 2018). Namely, active smoking was associated to an approximately 40% increase in active and memory regulatory T cell (T_{reg}) proportions, as well as a decrease in the number of NK and other innate lymphoid cell subsets, $\gamma\delta$ T cells and MAIT cells.

Patin et al. (2018) also performed genome-wide association studies (GWASs; § 1.2, page 7) of 166 cell-type specific immune phenotypes across over 5 million single nucleotide polymorphisms (SNPs) genotyped for all individuals in the cohort, to find the common genetic basis of natural immune variability. At a conservative genome-wide significance threshold of $\mathrm{p} < 10^{-10}$, the authors identified 14 independent QTLs associated to 42 traits, including measurements of 36 immune protein markers. Around 80% of these measurements were associated to protein (p) QTLs within 50 kilobases of the gene encoding the corresponding protein. Interestingly, 5 out of the 9 pQTLs were also mapped as eQTLs of nearby genes, including cases in which the same QTL was associated to the abundance of mRNA and protein product of a given gene. For example, the rs2223286 SNP that is annotated in the Genotype-Tissue Expression atlas (The GTEx Consortium, 2015, 2017, 2020) as an eQTL for the *Selectin L* (*SELL*) gene in whole blood ($\mathrm{p} = 3.2 \times 10^{-9}$) was detected as a pQTL of CD62L in eosinophil granulocytes ($\mathrm{p} = 1.6 \times 10^{-35}$) by Patin et al. (2018).

Remarkably, 80% of genome-wide associations with immunophenotypes were detected in innate immune cells—including monocytes, DCs and NK cells—while only 47% of all traits where measured in innate cells. Moreover, 33% of the genetic associations with immunophenotypes of adaptive cells were found in naïve B and T subsets, while these cells represent less than 10% of the total adaptive compartment (Patin et al., 2018). To systematically assess the contributions of genetic and non-genetic factors to innate and adaptive immunophenotype variance, Patin et al. (2018) used linear models including age, sex, CMV serostatus, smoking status and all significant QTLs of each trait. Across all immune traits, the proportion of variance explained by genetics is 66% larger for innate immunophenotypes relative to adaptive ones. In contrast, non-genetic factors explain a fraction of innate immune trait variance that is 46% smaller than for adaptive immunophenotypes.

Taken together, these results suggest that the parameters of innate immune cells and naïve adaptive subsets are preferentially genetically controlled, while mature and memory adaptive cell states are more dependent on the non-genetic and environmental factors included in life history of each individual (Patin et al., 2018). Among non-genetic factors, smoking status showed the strongest difference in predictive potential of innate and adaptive immune traits, followed by age, CMV serostatus and sex (Patin et al., 2018).

Sex biases on the genetic regulation of gene expression

Working on the same cohort (Thomas et al., 2015), Piasecka et al. (2018) set out to characterize the effects of age, sex, immune cell composition and genetics on the expression of 560 immune genes in response to various microbial pathogens, including *Escherichia coli* and *Staphylococcus aureus* bacteria, *Candida albicans* fungi and a live strain of IAV. Overall, the authors found that while age and sex had widespread albeit moderate effects on the transcriptional immune response, eQTL effects were stronger but targeted specific gene sets. Yet, across all stimulation conditions, the respective contributions of age, sex and genetics were dwarved by the impact of cellular composition, as measured by the global proportion of leukocytes in whole blood (Piasecka et al., 2018).

For instance, *cis*-eQTLs explained an average 10.3% of variance in the expression of 132 immune genes in response to IAV, whereas sex and age affected the expression of all 560 immune genes each, but only explained around 1% and 2% of expression variance, respectively. Leukocyte proportions also affected the expression of all tested genes, explaining around 16% of expression variance.

Interestingly, although age and sex both drive variation in immune cell proportions across healthy individuals (Patin et al., 2018), and cellular composition can impact immune gene expression read-outs (Perez et al., 2022), Piasecka et al. (2018) report that age and sex directly impacted the expression of 85% and 76% of genes tested at the basal state, respectively. For some of these genes, direct age and sex-effects were coupled to expression changes mediated by cellular composition. For example, and age-relate decrease in $CD8^+$ T cell proportions mediated expression changes on 44%

of immune genes indirectly affected by age. Likewise, indirect sex-effects mediated by a decrease in the proportion of $CD4^+$ T cells in males impacted the expression of 26% of genes indirectly affected by sex, while a decrease in $CD14^+$ monocyte proportions in females mediated changes in the expression of 21% immune genes under indirect sex-effects.

Finally, although interactions between age, sex and genetics have been described in the context of complex human diseases (Yao et al., 2014), Piasecka et al. (2018) report very limited evidence of such conditional eQTL effects, suggesting that the genetic control of genes involved in immune-related pathways is somewhat independent of age and sex.

To decipher the biology behind sex-related differences in gene expression and regulation across human tissues, Oliva et al. (2020) used the latest data release from the Genotype-Tissue Expression (GTEx; § 1.2.4, page 16) Consortium (The GTEx Consortium, 2020). Overall, the authors report widespread transcriptional differences between sexes—affecting around 13 thousand genes or 37% of the human transcriptome across tissues—albeit with small effects, with a median fold-change of 1.04. Furthermore, sex-effects on gene expression were found to be strongly dependent on tissue, as 18% of sex-biased genes were only so in one tissue, suggesting tissue-specific regulation mechanisms (Oliva et al., 2020). Nonetheless, for 76% of sex-biased genes in more than one tissue, the effects of sex on gene expression were largely consistent across tissues.

In line with these and previous observations (The GTEx Consortium, 2020), sex-biased genetic regulation of gene expression was estimated as strongly tissue-specific. Out of the 369 sex-biased *cis*-eQTLs associated to the expression of 366 eGenes, 70% were specific to breast tissue, and only one variant was detected in two tissues—rs41309559 on the X chromosome, associated to the expression of *ASB9* in skeletal ($\mathfrak{p} = 6.2 \times 10^{-4}$) and cardiac ($\mathfrak{p} = 4.9 \times 10^{-6}$) muscle.

Remarkably, the 261 sex-biased eQTLs in breast tissue were enriched for cell-type interacting *cis*-eQTLs (Kim-Hellmuth et al., 2020) whose effects are conditioned by the proportions of different cell types (§ 1.2.6, page 22). In particular, 42% of sex-biased eQTL signals—including the strongest unadjusted signal—were lost when adjusting the models for estimated epithelial cell abundances in the breast (Oliva et al., 2020). Moreover, mediation analyses revealed that 23% of sex-biased eQTL effects were significantly mediated by the abundance of epithelial cells. Taken together, these results suggest that a large fraction of sex-effects on the genetic regulation of gene expression could be explained by cell-type specific eQTLs in cells found at different frequencies across sexes (Oliva et al., 2020), once more highlighting the importance of high-resolution context definition in eQTL mapping studies (§ 1.2.6, page 22).

Age biases on the genetic regulation of gene expression

Yamamoto et al. (2022) also used GTEx data (The GTEx Consortium, 2020) to assess the impacts of aging on the genetic regulation of gene expression across human tissues. In line with previous observations (Viñuela et al., 2018; Balliu et al., 2019), the authors observed a general loss in eQTL predictive power across tissues with age, which they associated to increased variability in gene expression among older individuals. Gene expression heterogeneity associated to aging could in turn be explained by the stronger cumulative impact of environmental and lifestyle factors, as well as by relaxed selective constraints on the variance of gene expression (Medawar, 1946, 1952).

In line with these observations, the narrow-sense heritability ($\mathfrak{h}^2$; Appendix A, page 190) of gene expression—estimated as the contribution to genetically regulated expression variance (Eq. (1.6), page 21) of SNPs around each gene—generally decreases with age (Yamamoto et al., 2022). However, Yamamoto et al. (2022) also found age-effects on gene expression regulation to be strongly tissue-dependent. While the average $\mathfrak{h}^2$ of gene expression varied from 2.9% to 5.7% across tissues, the proportion of gene expression variance explained by age was as low as 0.04% in the pancreas, and as high as 7.9% in whole blood in average across expressed genes.

3.2.2 Genetic and nongenetic drivers of immune variability in response to viruses

Through their contributions to the natural variation of immune parameters, age and sex, but also genetic and environmental factors can explain variability in host immune responses to pathogens (Patin et al., 2018; Piasecka et al., 2018; Randolph et al., 2021). In particular, pathogenic viruses have been shown to ellicit especially heterogeneous responses (Piasecka et al., 2018), in addition to having been strong drivers of adaptation during human evolutionary history (§ 2.3, page 46).

The extent of human immune variability in the response to viruses was strikingly illustrated by the ongoing 'coronavirus disease 2019' (COVID-19) pandemic, sparked by the 2019 outbreak of a novel coronavirus strain inducing severe acute respiratory syndrome (SARS-CoV-2). In just under three years, SARS-CoV-2 infection and COVID-19-related complications claimed the lives of around six million people world-wide (Wang et al., 2022); yet, SARS-CoV-2 infection is also characterized by a high rate of asymptomatic infection at approximately 35% of cases (Sah et al., 2021). Moreover, COVID-19 courses and outcomes are highly variable, ranging from light cold-like symptoms to death (O'Driscoll et al., 2021; Pei et al., 2021).

Age and sex were clearly identified as the strongest predictors of variability in COVID-19 courses relatively early on during the pandemic (Takahashi et al., 2020; O'Driscoll et al., 2021). Through analyses of genetic and lifestyle data across over a million individuals from diverse ancestries, Shelton et al. (2021) estimated that the odds of testing positive for SARS-CoV-2 infection were 1.2 times higher for self-reported male participants relative to females. Moreover, males were also found to be more likely to require hospitalization (χ^2-test $\mathrm{p} = 4.3 \times 10^{-8}$). Irrespective of self-reported sex, hospitalization rates also increased steadily from around 4% in individuals below the age of 30, to around 30% in individuals over 80 years old (Shelton et al., 2021). In general, although the risk of death following infection by SARS-CoV-2 is always higher in males than in females, the magnitude of this difference increases with age (O'Driscoll et al., 2021).

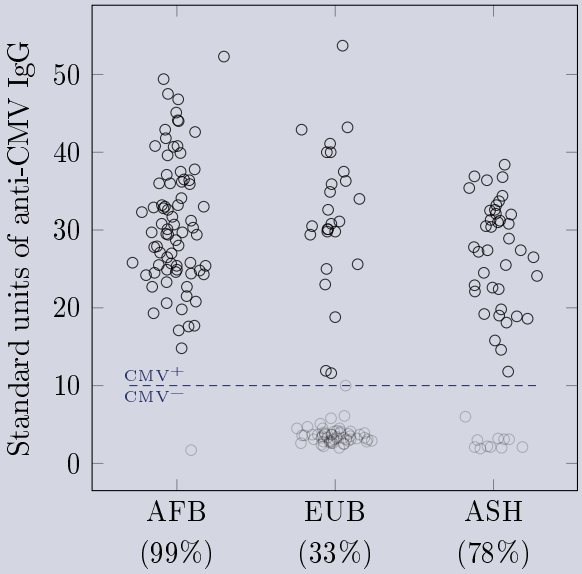
Interestingly, it was shown that some of these effects on COVID-19 risk were mediated by changes in immune cell proportions. For example, Takahashi et al. (2020) report that while lymphoid T cells and non-classical monocytes are more strongly induced in biological females and males respectively, high myeloid cytokine plasma levels are associated to COVID-19 severity exclusively in females, and poor $CD8^+$ T cell responses are likely to lead to severe COVID-19 in males only. Notably, while older male COVID-19 patients showed lower activated $CD8^+$ T cell proportions relative to younger male cases, this relationship with age was not observed in females.

Changes in cellular composition associated to COVID-19 risk are also interesting because they can mediate effects from the environment. As previously mentioned (§ 3.2.1, page 59), and in line with previous observations in the NK cell compartment (Gumá et al., 2004), Patin et al. (2018) describe an increase in $CD8^+$ EMRA T cells in CMV-seropositive individuals (Box 8). In line with these observations, infection by CMV has recently been linked to worsening COVID-19 symptoms in younger patients under 60 years old, and in the absence of co-morbidities (Weber et al., 2022).

Other host intrinsic factors were also quickly associated to higher COVID-19 risks. Using a multivariate logistic regression model of the probability of hospitalization among SARS-CoV-2-infected individuals—including age, sex, mean household income, educational attainment, body mass index (BMI), ancestry and various comorbidities as covariates—Shelton et al. (2021) identified obesity as the most significant predictor of COVID-19-related hospitalization risk; regardless of age and sex, individuals with a BMI over 30 are twice as likely to require hospitalization relative to other cases. The authors also estimated that African-Americans were 82% more likely to require hospitalization following a positive SARS-CoV-2 test relative to European-Americans, suggesting a genetic basis to COVID-19 risk disparities, although between-group differences in vaccination rates could also contribute to this picture.

Through genome-wide association studies (GWASs; § 1.2, page 7) of COVID-19 diagnosis and severity phenotypes (Box 9), Shelton et al. (2021) strengthened the basis of evidence for previously reported links (Ellinghaus et al., 2020; Zhao et al., 2021) between COVID-19 risk and the ABO blood group—especially relevant given the implication of dysregulated blood clotting (Levi et al., 2020) in severe COVID-19 forms—as well as a gene-rich region in chromosome 3 (Ellinghaus et al., 2020) spanning immune-relevant genes like *XCR1* and *SLC6A20*, but also *CCR9* and *CXCR6* that have been proposed as targets of archaic adaptive introgression (§ 2.2.2, page 45)

Box 8 | Cytomegalovirus seropositivity across populations. Cytomegalovirus (CMV) is a herpesvirus associated to lifelong latent infection in humans (Cannon et al., 2010). CMV seroprevalence tends to be higher in females than in males, and to increase with age. Another remarkable feature of CMV infection rates is that their geographical distribution is associated to by-country income levels: mean CMV seroprevalence is higher in developing countries in Africa, Asia and South America, than in relatively developed countries of Northern and Western Europe (Cannon et al., 2010).



The Figure shows CMV serologies ascertained via enzyme-linked immunosorbent assays for 209 individuals of Central African (AFB, $n = 78$), West European (EUB, $n = 80$) or East Asian (ASH, $n = 51$) origin (Aquino et al., 2023). Standard units of anti-CMV immunoglobulin (Ig) G abundances are shown. The dashed line represents the detection threshold used to identify seropositive donors. Observations above this threshold are shown in higher opacity; seroprevalence estimates in each group are shown in parentheses.

Although CMV infection was long thought to have no discernible consequences on human health, recent studies have linked it to several cardiovascular problems and worse prognoses following viral infection (Weber et al., 2022).

As members of the COVID-19 Host Genetics Initiative (2020, 2021, 2022, 2023), Kousathanas et al. (2022) used whole-genome sequencing data across over 7 thousand critically ill COVID-19 patients—presenting acute lung injury and respiratory failure—and more than 48 thousand controls (Box 9, 'A2' analysis type) of diverse origins to perform a GWAS of COVID-19 severity.

The work by Kousathanas et al. (2022) is especially valuable because the authors followed up on the genome-wide associations with several statistical and functional genomic approaches to infer causal links between genotype and severe COVID-19 risk (§ 1.2, page 7). For instance, the authors used Bayesian statistical fine-mapping (Wang et al., 2020) to refine the associations in the 3p21.31 gene cluster (Ellinghaus et al., 2020; Shelton et al., 2021) as two independent signals falling on the 5' untranslated region of *SLC6A20*, as well as on introns of *LZTFL1* and non-coding regions downstream of it.

Box 9 | 'Coronavirus disease 2019' phenotype definitions. The heterogeneity in 'coronavirus disease 2019' (COVID-19) outcomes highlights the need for accurately defining focal phenotypes (§ 1.2.3, page 12) when conducting genome-wide association studies of COVID-19 susceptibility and severity. Indeed, even nuanced differences across phenotype definitions can lead to apparently conflicting results or false associations (Roberts et al., 2022).

For example, Shelton et al. (2021) define one 'diagnosis' phenotype between cases that reported a positive COVID-19 test and controls that tested negative, and four 'severity' phenotypes between increasingly dire COVID-19 cases—from hospitalization to respiratory support—and controls that did not report a COVID-19 diagnosis. The authors justify these definitions stating that data collection took place early in the pandemic, when most of the general population was likely still unexposed to the virus.

	Sufficient reporting			Severity degree		
Subtype	Self	Physician	Laboratory	Hospital	ICU or death	Control
A2	☐	☐	☑	☑	☑	Non-case
B1	☐	☐	☑	☑	☐	Non-hospitalized
B2	☐	☐	☑	☑	☐	Non-case
C2	☑	☑	☑	☐	☐	Non-case

The COVID-19 Host Genetics Initiative (2020, 2021, 2022, 2023) proposes another set of consolidated phenotypes. The Table above summarizes the four main types of analyses used today. Briefly, analysis type 'A2' concerns very severe COVID-19 cases confirmed by laboratory tests and requiring hospitalization in an intensive care unit (ICU) or leading to death, compared to all individuals who are not cases. In contrast, analysis subtype 'B2' focuses on less severe cases that require hospitalization but not admission into an ICU, compared to the general population, or to controls that did not require hospitalization in the 21 days following a positive test, for subtype 'B1'. Finally, analysis type 'C2' considers COVID-19 cases confirmed by a laboratory or physician, or self-reported in a questionnaire, compared to the general population.

Using eQTL data from the GTEx Consortium (The GTEx Consortium, 2020), Kousathanas et al. (2022) then performed transcriptome-wide association studies (§ 1.2.5, page 19) in whole blood and lung tissue. Among other interesting hits, genetic effects on COVID-19 severity were significantly mediated by increased expression of *CCR9* in whole blood, and increased expression of *MUC1*—encoding a component of the mucus lining airway epitheliæ—in both blood and lung tissue. Moreover, high colocalization probabilities ($PP_{\mathcal{H}4}$) were estimated between COVID-19 risk loci and GETx eQTLs around these genes: rs73064425 for *CCR9* in blood ($PP_{\mathcal{H}4} > 0.8$), and rs41264915 for *MUC1* in blood ($PP_{\mathcal{H}4} > 0.8$) and the lung ($PP_{\mathcal{H}4} > 0.5$) (§ 1.2.5, page 19).

Kousathanas et al. (2022) also unveiled new associations between COVID-19 severity and the genotype at immune-relevant loci, including mediators of IFN signalling like *IFNA10*, *IFNAR2*, *TYK2*, *IL10RB* and *PLSCR1*, as well as genes involved in the differentiation of lymphoid and myeloid cells (§ 3.1.1, page 55), such as *BCL11A*, *TAC4* and *CSF2*. Remarkably, *CSF2* encodes the granulocyte-macrophage colony stimulating factor (GM-CSF), which has been proposed as a candidate therapeutic target against COVID-19 (Lang et al., 2020).

The latest installment of the COVID-19 Host Genetics Initiative (2023) provides a comprehensive summary of the known genetic bases of COVID-19 susceptibility and severity, and how they affect different actors of the cellular immune responses to SARS-CoV-2. Across two severity and one susceptibility phenotypes (Box 9, analysis types 'A2', 'B2' and 'C2'), 51 independent loci reached genome-wide significance and were annotated to a gene based on distance or functional scores. Out of these, 15 genes were linked to biological pathways involved in viral entry into cells, mucus production by epithelial cells and type I IFN signalling (COVID-19 Host Genetics Initiative, 2023).

For example, variants around the loci encoding the ACE2 (rs190509934, $\mathrm{p} = 1.5 \times 10^{-18}$) and SLC6A20 (rs73062389, $\mathrm{p} = 1.3 \times 10^{-102}$) receptors—both important mediators of SARS-CoV-2 entry—are associated to COVID-19 susceptibility. Moreover, the rs9305744 variant of *TMPRSS2*, which encodes a type II transmembrane serine protease needed for the cleavage of Spike proteins prior to SARS-CoV-2 entry, is associated to COVID-19 severity ($\mathrm{p} < 2.5 \times 10^{-7}$).

Regarding IFN-mediated immunity, variants associated to COVID-19 severity were found near the loci of TYK2 (rs34536443, $\mathrm{p} < 1.7 \times 10^{-21}$) and JAK1 (rs11208552, $\mathrm{p} < 2.2 \times 10^{-9}$). Downstream in the GRN, the rs10066378 SNP near the *IRF1* TF locus was also associated to disease severity ($\mathrm{p} = 2.7 \times 10^{-9}$), as were variants near the locus of IFN-α (rs28368148, $\mathrm{p} < 7.1 \times 10^{-7}$) and its receptor *IFNAR2* (rs78143111, $\mathrm{p} < 2.8 \times 10^{-16}$). Finally, the rs10774679 SNP near the *OAS1-3* ISG locus is also associated to COVID-19 severity ($\mathrm{p} < 7.9 \times 10^{-11}$), is an eQTL of *OAS1*, *OAS2* and *OAS3* in GTEx data ($\mathrm{p} < 6.7 \times 10^{-12}$), and is carried by a haplotype likely introgressed from Neandertal into modern human genomes (Zeberg and Pääbo, 2021).

All in all, these results highlight genetic COVID-19 risk factors predominantly strewn along the GRN (§ 3.1.2, page 57) of type I IFN signalling, as well as the repercussion of adaptation events in shared human evolutionary history that shaped present-day genetic diversity (§ 2, page 36). Yet, they do not provide a comprehensive view of how these impacts vary across human populations and immune cell types (Aquino et al., 2023).

The hallmarks of severe viral infection and the importance of interferons

The relevance of type I IFN responses was also highlighted by single-cell studies (Lee et al., 2020; Wilk et al., 2020; Ren et al., 2021; Stephenson et al., 2021) of the transcriptional immune response to SARS-CoV-2 (§ 1.3, page 25). In particular, Lee et al. (2020) compared the transcriptional profiles of PBMCs sampled from patients with mild or severe COVID-19, or patients hospitalized following infection by IAV, allowing derivation of characteristic features of responses to SARS-CoV-2—relative to another respiratory RNA virus—at cell-type resolution.

Across all PBMC types, cells from COVID-19 patients expressed stronger inflammatory transcriptional programs—mainly driven by tumor necrosis factor (TNF) and IL-1β—relative to cells from patients with severe influenza (Lee et al., 2020). Inflammation was further exacerbated in patients with severe COVID-19 due to a dysregulated type I IFN response emanating from the monocyte compartment, that was not observed in mild COVID-19 cases (Lee et al., 2020).

The biological relevance of these results is supported by bulk transcriptomic observations on lung tissue from deceased COVID-19 patients (Blanco-Melo et al., 2020). Specifically, the lung transcriptional signatures of lethal COVID-19 cases were enriched in genes upregulated by type I

IFN-signalling and TNF/IL-1β-driven inflammation (Lee et al., 2020). Moreover, genes upregulated by monocytes from severe relative to mild COVID-19 cases were enriched in genes differentially expressed in lethal COVID-19 and healthy lung biopsies (Blanco-Melo et al., 2020; Lee et al., 2020).

The picture that emerges from these and other studies (Hadjadj et al., 2020; Ren et al., 2021; Stephenson et al., 2021) is that early dysregulation of type I IFN signalling upon infection by SARS-CoV-2 can propagate along the GRNs of the immune response, resulting in lower viral clearance and unhinged inflammatory reactions involving TNF, IL-1β, IL-6 and IL-8, among other cytokines. These so-called cytokine 'storms' are mainly mediated by myeloid cells, and lead to tissular infiltration of immune cells and production of more inflammatory mediators. Ultimately, this auto-amplifying process leads to lung tissue damage and COVID-19 pneumonia.

Such immune dysregulation can be explained through intrinsic (Takahashi et al., 2020; O'Driscoll et al., 2021), environmental (Patin et al., 2018; Weber et al., 2022) and lifestyle factors (Patin et al., 2018) that can alter immune cellular composition, as well as by common genetic variation (Ellinghaus et al., 2020; COVID-19 Host Genetics Initiative, 2020, 2021, 2022, 2023; Shelton et al., 2021; Kousathanas et al., 2022).

Importantly, rare and Mendelian inborn errors of immunity (IEIs), as well as their immune-mediated phenocopies, have also been associated to higher COVID-19 risks (Zhang et al., 2020; Bastard et al., 2020). Through comparison of 13 genes involved in TLR3-dependent induction or IRF7-dependent amplification (Fig. 3.2) of type I IFN signalling—two pathways known to harbor IEI risk factors of severe influenza—across over 600 COVID-19 patients with severe pneumonia and more than 500 individuals with asymptomatic SARS-CoV-2 infection or mild COVID-19 symptoms, Zhang et al. (2020) identified 24 IEIs at 8 loci—*TLR3*, *TICAM1*, *UNC93B1*, *TBK1*, *IRF3*, *IRF7*, *IFNAR1* and *IFNAR2*—in 23 severe COVID-19 patients aged 17 to 77 years.

Remarkably, Zhang et al. (2020) also report an absence of type I and type III IFN production in plasmacytoid dendritic cells (pDCs)—an immune subset known for its ability to rapidly secrete great quantities of IFN-α upon viral infection (Siegal et al., 1999; Cella et al., 1999)—from an IRF7-deficient patient exposed to SARS-CoV-2 in vitro. In line with this observation, IFN-α was undetectable in the serum of 10 out of the 23 severe COVID-19 patients carrying IEIs, including 5 IRF7-deficient individuals (Zhang et al., 2020). Together, these results highlight the clinical relevance of type I IFNs and IFN-producing cell types like pDCs in the context of COVID-19.

Further emphasizing the importance of type I IFNs, Bastard et al. (2020) found neutralizing immunoglobulin G auto-antibodies targeting IFN-α and/or IFN-ω in the sera of 101 out of 987 patients with severe COVID-19 pneumonia, but only in 4 out of 1,227 healthy individuals sampled before the pandemic. Notably, while the 101 severe COVID-19 cases varied widely in age—from 25 to 87 years old—more than half were over 65 years old. Also, 94% of these individuals were male, representing a significantly larger fraction than males among life-threatening COVID-19 cases without auto-antibodies (75%, Fisher's exact $\mathrm{p} = 2.5 \times 10^{-16}$).

The authors associate the presence of anti-IFN-α2 auto-antibodies in the serum to an inability to block SARS-CoV-2 replication in vitro, and show that auto-antibodies were already present in the sera from two of the COVID-19 patients sampled pre-pandemic, suggesting that auto-antibodies are a cause of critical COVID-19 rather than a consequence (Bastard et al., 2020).

Importantly, no auto-antibodies were detected in the 23 severe COVID-19 patients described by Zhang et al. (2020), suggesting that although IEIs of type I IFN signalling and auto-antibodies against type I IFNs have similar effects on the immune response to SARS-CoV-2, they act through different and independent pathways. Together, these two risk factors jointly explain up to 20% of life-threatening cases of COVID-19 pneumonia in patients over 70 years old (Bastard et al., 2020, 2021a; Zhang et al., 2020; Casanova and Abel, 2022).

3.3 Genetic susceptibility to infectious diseases and the infection enigma

Infectious pathogens have been major drivers of adaptation throughout human evolutionary history (§ 2.3, page 46). It is estimated that for approximately 200 thousand years, and up until the late 19$^{\text{th}}$ century, around half of human children under the age of 15 years died to infection-induced fever (Casanova and Abel, 2005). In 19$^{\text{th}}$-century England, and outside of any major epidemics, infectious diseases were the reported cause of around 60% of deaths (Casanova and Abel, 2005). Today, with modern hygiene practices, vaccines and drugs, widely lethal infectious diseases—such as tuberculosis, the acquired immunodeficiency syndrome or COVID-19—have become the exception rather than the rule (Casanova and Abel, 2021, 2022).

Yet, in rare cases even generally harmless pathogens can cause life-threatening disease in young individuals with few risk factors (Ciancanelli et al., 2015; Hernández et al., 2018; Lim et al., 2019). By contrast, some individuals are known to display persistent lifelong resistance to infection by even the deadliest of pathogens (Cobat et al., 2009). Casanova and Abel (2013, 2020) call this reflection of human immune variability the 'infection enigma'.

3.3.1 Synthetic theory of immune variability in infectious disease

Casanova and Abel (2013) propose a synthesis of four theories to explain the widespread immune variability revealed by infectious diseases. First, the germ theory of infectious diseases—championed by Louis Pasteur (1862) and Robert Koch (1882), among others—attributes immune variability to the microbial pathogens themselves. In contrast, the ecological theory considers variation due to other environmental factors, excluding the causal infectious agent itself, but potentially including co-infection by other pathogens and remnants of immune responses against previously encountered pathogens (Sun and Metzger, 2008). Third, the immunological theory focuses on somatic variation acquired through adaptive responses. Finally, the genetic theory looks for inborn predictors of innate and adaptive immunity encoded in the DNA (Casanova and Abel, 2005). In the synthetic theory of infectious disease, all of these host-extrinsic and intrinsic factors—microbial, environmental, immunological and genetic—interact together, giving rise to a wide range of clinical phenotypes. The relevance of jointly considering the different predictors of infectious disease risk has most recently been emphasized by the COVID-19 pandemic.

The microbial component of COVID-19 is reflected in the variable pathogenicity observed across different SARS-CoV-2 variants of concern (VOCs; Box 10). Indeed, since the reporting of the first VOC eight months into the pandemic, all subsequent VOCs from the B.1 SARS-CoV-2 lineage have been associated to increased hospitalization rates following infection—relative to the ancestral Wuhan-1 lineage—likely due to mutations in the Spike viral entry protein (Tegally et al., 2021; Funk et al., 2021; Markov et al., 2023). Although the currently circulating 'Omicron' VOC is associated to decreased hospitalization rates relative to the previous 'Delta' one, it seems to be linked to higher susceptibility to infection, as Omicron reinfection rates are much higher (Volz et al., 2021).

Besides virus-centric factors, changes in host immune cellular composition—particularly in the myeloid and cytotoxic lymphoid compartments—have been associated to increased risks of life-threatening disease (Diao et al., 2020; Lee et al., 2020; Xu et al., 2020; Wilk et al., 2020; Ren et al., 2021; Stephenson et al., 2021). Interestingly, the variability in abundance of some of these immune subsets, such as terminally differentiated CD8^{+} EMRA T cells, has been associated to latent infection by CMV (Box 8) (Patin et al., 2018), which has recently been associated to COVID-19 severity in relatively young patients below 60 years old. Together, these results reflect the ecological and immunological components of COVID-19 risk, whereby previous infection by another virus alters immune parameters in a way that predisposes to severe outcomes of SARS-CoV-2 infection.

From a genetic perspective, common and rare genetic variants have been associated to inter-individual differences in COVID-19 susceptibility and severity (§ 3.2.2, page 62). In this context, the study of immune responses in different IEI contexts has proven its worth in unveiling biological mechanisms underlying variation in infectious disease risks (Casanova and Abel, 2021, 2022; Zhang et al., 2022). A clear illustration of this potential is given by the discovery of an IRF7 deficiency in a severe influenza patient (Ciancanelli et al., 2015), which led to the identification of other genetic risk factors of severe IAV infection (Hernández et al., 2018; Lim et al., 2019), as well as genetic and auto-immune predictors of a substantial part of variability in COVID-19 severity (Zhang et al., 2020; Bastard et al., 2020, 2021a) (§ 3.2.2, page 62).

Box 10 | Coronavirus variants of concern. In the context of the 'coronavirus disease 2019' (COVID-19) pandemic triggered by the 2019 'severe acute respiratory syndrome' coronavirus 2 (SARS-CoV-2) outbreak, the World Health Organization (WHO)—together with other international health institutions—keeps track of emerging SARS-CoV-2 'variants of concern' (VOCs) for global health. In the terms of the WHO, a VOC is a SARS-CoV-2 variant with 'genetic changes that are predicted or known to affect virus characteristics' an their impact on human health, has had 'a growth advantage over other circulating variants in more than one WHO region' and entails a 'detrimental change in clinical disease severity', affects 'the ability of health systems to provide care to patients' or causes a 'significant decrease in the effectiveness of available vaccines' (World Health Organization, 2020b).

All former and currently circulating VOCs belong to the B.1 SARS-CoV-2 lineage characterized by an aspartate (D) to glycine (G) substitution in the 614^{th} aminoacid of the Spike protein (D614G), relative to the ancestral Wuhan-1 lineage. The D614G mutation alone is associated to increased transmissibility, higher viral load and lower patient age, but its effects on infection severity are uncertain (Volz et al., 2021). However, VOCs carrying D614G alongside other mutations that do increase virulence can lead to increased hospitalization rates (Markov et al., 2023). For instance, the 'Alpha' VOC (B.1.1.7) identified in the United Kingdom in September 2020, and responsible for the second major wave of COVID-19 cases, carried 16 Spike mutations other than D614G (Funk et al., 2021). Later, the 'Delta' VOC (B.1.617.2) identified in India in October 2020 sparked the third major wave of the pandemic. The currently circulating 'Omicron' VOC lineages (BA.1, BA.2, BA.3, BA.4, BA.5) carry over 30 Spike mutations and are associated to lower hospitalization rates than Delta, but much higher reinfection rates (Pulliam et al., 2022).

Interestingly, the mirroring between factors of inborn susceptibility to severe IAV and SARS-CoV-2 infection suggest a shared basis of genetic susceptibility to viral infection, and highlight the relevance of comparing SARS-CoV-2 to other respiratory viruses to derive specific patterns of COVID-19 severity. In support of this view, Figure 3.3 shows response (r) eQTL effects (§ 1.2.6, page 22) estimated in PBMCs exposed to SARS-CoV-2 (COV) or IAV for six hours. While most reQTLs have a similar effect on gene expression in lymphoid responses to both viruses, myeloid cells show much more heterogeneous genetic control. In particular, the strongest SARS-CoV-2 reQTL ($\beta_{\text{Xi}}^{\text{COV}} = 1.47$, $\text{p} = 2 \times 10^{-16}$; $\beta_{\text{Xi}}^{\text{IAV}} = 0.04$, $\text{p} = 0.4$) affects the expression of *MMP1*—encoding matrix metalloproteinase 1 (Syed et al., 2021), a marker of COVID-19 severity—and is specific to the response to SARS-CoV-2 in myeloid cells (Fig. 1.9, page 23).

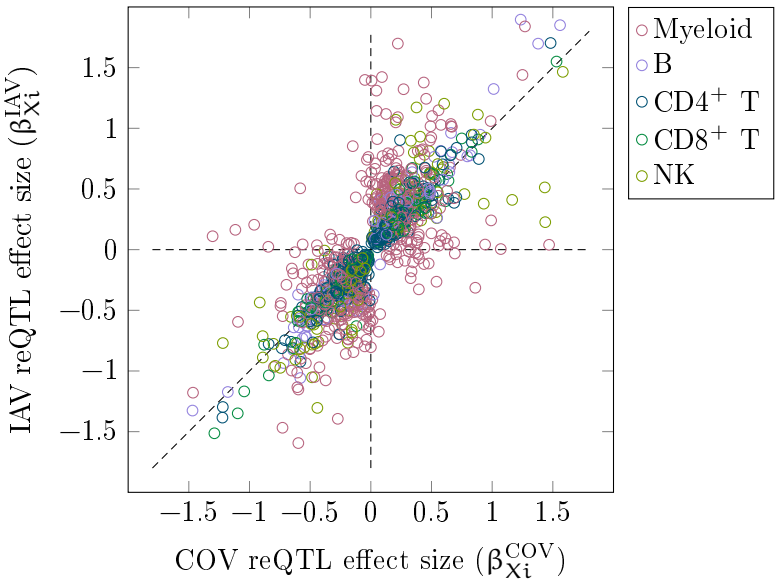


Figure 3.3 | Shared genetic basis of transcritional response to respiratory viruses. Comparison of response expression quantitative trait locus (reQTL) effect sizes (β_{Xi}) between peripheral blood mononuclear cells (PBMCs) stimulated with 'severe acute respiratory syndrome' coronavirus 2 (SARS-CoV-2; COV) or the influenza A virus (IAV). Each dot represents a specific reQTL effect, colored according to the PBMC type in which it was estimated. Adapted from Aquino et al. (2023).

Comparing the responses to SARS-CoV-2 and IAV in the context of IEIs can also highlight the relative importance of specific cell types in the response to one virus or the other. For example, while pDCs have been shown to be essential for the IRF7 and TLR7-mediated (Zhang et al., 2020; Asano et al., 2021) production of critical type I IFNs in the reponse to SARS-CoV-2, the pDC response from UNC93B1-deficient individuals—insensitive to TLR7 stimulation—infected by seasonal IAV seems to be unaffected (Ciancanelli et al., 2015).

3.3.2 Inborn errors of immunity and precision medicine

In summary, the wealth of immunological and clinical data generated in response to the COVID-19 pandemic has propelled multiple efforts to understand the various predictors of human immune variability in the response to viruses. In this setting, data from IEI patients has provided very valuable insights into the pathogenicity of SARS-CoV-2 infection. In particular, the identification of IEIs of IFN-mediated immunity ultimately led to the discovery of auto-immune predictors of up to 20% of life-threatening COVID-19 cases (Zhang et al., 2020; Bastard et al., 2020, 2021a). In hindsight, it is likely that other IEIs will be discovered in the 80% of severe cases not currently explained by genetic deficiencies or its immunological counterparts (Zhang et al., 2022). Moreover, given the evidence for a common genetic basis of susceptibility to severe viral infection (Fig. 3.3), these findings will also likely be relevant for other diseases (Casanova and Abel, 2021).

The discovery of anti-IFN auto-antibodies in a COVID-19 context—and in light of the synthetic theory of infectious diseases—also illustrates how studying rare IEI genetic determinants of life-threatening disease at the individual level can unveil more common immunological predictors of severe disease courses with potential population-level impacts. Within a sample of over 30 thousand individuals, Bastard et al. (2021a) estimated the proportion of neutralizing auto-antibodies against IFN-α and/or IFN-ω in the general population at around 1.1% in individuals below 70 years old, 4.4% in individuals between 70 and 79 years old and 7.1% in individuals between 80 and 85 years old. The increase in prevalence with age suggests an acquired environmental component in the ætiology of anti-IFN auto-antibodies, potentially including previous infections by other viruses or previous immune-related disorders (Panem et al., 1982; Bello-Rivero et al., 2004; Gupta et al., 2016).

Together, these results highlight the importance of testing for anti-IFN auto-antibodies, particularly in the sera of elderly individuals or patients with immune disorders, as their presence has several direct clinical implications, including severe infectious disease courses and adverse reactions to vaccination (Bastard et al., 2021a,b).

In this sense, the synthetic theory of infectious diseases proposed by Casanova and Abel (2013) is inscribed in the larger paradigm of precision medicine. Any accurate prediction of disease risk requires a comprehensive assessment of the predictors of immune response variability—including its microbial, environmental, genetic and nongenetic components—and in light of potential evolutionary determinants (Dobzhansky, 1973; Quintana-Murci, 2019) (§ 2.3.3, page 50). Such descriptions are essential for the development of a medicine adapted to the great human diversity, and able to efficiently counter current and future pathogenic threats (Jones et al., 2013).

Part II

Contributions to the field

Single-cell and bulk RNA-sequencing reveal differences in monocyte susceptibility to influenza A virus infection between Africans and Europeans

Immune variability from a myeloid perspective. The outcome of viral infection varies widely across human individuals. This variability has been the subject of long-standing interest at the Human Evolutionary Genetics (HEG) Unit of Institut Pasteur, Paris. In particular, HEG-ites focus on the environmental, genetic and evolutionary drivers of population immune response variation.

In this context, Quach et al. (2016) used RNA-sequencing (RNA-seq) data to characterize the myeloid transcriptional immune response to a series of stimuli—including a live influenza A virus (IAV) strain—across 200 healthy donors of African (AFB, $\mathfrak{n} = 100$) and European (EUB, $\mathfrak{n} = 100$) self-reported ancestry from the EvoImmunoPop (EIP) cohort. With this landmark study, the HEG group greatly contributed to consolidate the notion that adaptive archaic introgression could explain present-day population differences in the response to viral challenge, specifically through changes in the expression of immune-relevant genes in myeloid cells (§ 2, page 36).

Myeloid cells are both sensors and effectors of the immune system; their action is paramount for an effective response to pathogens (§ 3.1, page 55). Moreover, monocytes are particularly important for the antiviral response, owing to their capacity to both produce type I interferons (IFNs), and respond to them via the expression of IFN-stimulated genes (ISGs). Hence, monocytes are a relevant model in which to study inter-individual variability in the peripheral innate immune response to viruses (§ 3.2, page 59).

The complexity of the circulating monocyte subset can be decomposed along the expression gradient of two cluster of differentiation (CD) markers. Upon infection, classical $CD14^{++}CD16^{-}$, intermediate $CD14^{++}CD16^{+}$ and non-classical $CD14^{+}CD16^{++}$ monocytes are recruited from the blood into the tissular focus of infection, where they participate to combat the invading pathogen. Each of these subtypes of monocytes has its own set of features and is associated to different roles.

To account for the cellular heterogeneity in the monocyte lineage in our description of the myeloid drivers of immune variation, we performed single-cell (sc) RNA-seq (§ 1.3, page 25) on primary monocytes from four AFB and four EUB donors of the EIP cohort, stimulated *ex vivo* with IAV, and chosen among extremely low or high responders identified in Quach et al. (2016). Because of monocytes' pivotal role as immune sensors, we focused on the early dynamics of the response, recovering samples at 0, 2, 4, 6 and 8 hours post-stimulation.

We first used the scRNA-seq data to define the different circulating monocyte subsets and assess their susceptibility to productive infection by IAV. On this basis, we defined subsets of groups of 'unexposed', 'bystander' and 'infected' $CD16^{-}$ or $CD16^{-}$ cells in order to refine the characterization of transcriptional states in each sample. We were thus able to explain differences in basal activation across samples through variation in the proportion of monocytes susceptible to infection. Finally, we used these results, as well as flow cytometry data from Quach et al. (2016), to hypothesize about myeloid drivers of population variation in immune responses to viral infection. ■

frontiers in Immunology

HYPOTHESIS AND THEORY
published: 29 November 2021
doi: 10.3389/fimmu.2021.768189

Single-Cell and Bulk RNA-Sequencing Reveal Differences in Monocyte Susceptibility to Influenza A Virus Infection Between Africans and Europeans

Mary B. O'Neill[1], Hélène Quach[2], Julien Pothlichet[3], Yann Aquino[1,4], Aurélie Bisiaux[1], Nora Zidane[5], Matthieu Deschamps[1], Valentina Libri[6], Milena Hasan[6], Shen-Ying Zhang[7,8,9], Qian Zhang[7,8,9], Daniela Matuozzo[8,9], Aurélie Cobat[8,9], Laurent Abel[7,8,9], Jean-Laurent Casanova[7,8,9,10], Nadia Naffakh[11], Maxime Rotival[1†] and Lluis Quintana-Murci[1,12†]*

[1] Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, Centre National de la Recherche Scientifique (CNRS), Paris, France, [2] Muséum National d'Histoire Naturelle, UMR7206, Centre National de la Recherche Scientifique (CNRS), Université de Paris, Paris, France, [3] DIACCURATE, Paris, France, [4] Sorbonne Université, Collège doctoral, Paris, France, [5] Biodiversity and Epidemiology of Bacterial Pathogens Unit, Institut Pasteur, Paris, France, [6] Cytometry and Biomarkers UTechS, Institut Pasteur, Paris, France, [7] St. Giles Laboratory of Human Genetics of Infectious Diseases, The Rockefeller University, New York, NY, United States, [8] Laboratory of Human Genetics of Infectious Diseases, Necker Hospital for Sick Children, INSERM UMR 1163, Necker Hospital for Sick Children, Paris, France, [9] Imagine Institute, Paris University, Paris, France, [10] Howard Hughes Medical Institute, New York, NY, United States, [11] RNA Biology of Influenza Virus Unit, Institut Pasteur, Paris, France, [12] Chair of Human Genomics and Evolution, Collège de France, Paris, France

There is considerable inter-individual and inter-population variability in response to viruses. The potential of monocytes to elicit type-I interferon responses has attracted attention to their role in viral infections. Here, we use single-cell RNA-sequencing to characterize the role of cellular heterogeneity in human variation of monocyte responses to influenza A virus (IAV) exposure. We show widespread inter-individual variability in the percentage of IAV-infected monocytes. Notably, individuals with high cellular susceptibility to IAV are characterized by a lower activation at basal state of an IRF/STAT-induced transcriptional network, which includes antiviral genes such as *IFITM3*, *MX1* and *OAS3*. Upon IAV challenge, we find that cells escaping viral infection display increased mRNA expression of type-I interferon stimulated genes and decreased expression of ribosomal genes, relative to both infected cells and those never exposed to IAV. We also uncover a stronger resistance of $CD16^+$ monocytes to IAV infection, together with $CD16^+$-specific mRNA expression of *IL6* and *TNF* in response to IAV. Finally, using flow cytometry and bulk RNA-sequencing across 200 individuals of African and European ancestry, we observe a higher number of $CD16^+$ monocytes and lower susceptibility to IAV infection among monocytes from individuals of African-descent. Based on these data, we hypothesize that higher basal monocyte activation, driven by environmental factors and/or weak-effect genetic variants, underlies the lower cellular susceptibility to IAV

Edited by:
Petter Brodin,
Science for Life Laboratory (SciLifeLab), Sweden

Reviewed by:
Karthik Shekhar,
University of California, Berkeley, United States
Trine Hyrup Mogensen,
Aarhus University, Denmark

***Correspondence:**
Lluis Quintana-Murci
quintana@pasteur.fr

[†]These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Systems Immunology, a section of the journal Frontiers in Immunology

Received: 15 September 2021
Accepted: 27 October 2021
Published: 29 November 2021

Citation:
O'Neill MB, Quach H, Pothlichet J, Aquino Y, Bisiaux A, Zidane N, Deschamps M, Libri V, Hasan M, Zhang S-Y, Zhang Q, Matuozzo D, Cobat A, Abel L, Casanova J-L, Naffakh N, Rotival M and Quintana-Murci L (2021) Single-Cell and Bulk RNA Sequencing Reveal Differences in Monocyte Susceptibility to Influenza A Virus Infection Between Africans and Europeans. Front. Immunol. 12:768189. doi: 10.3389/fimmu.2021.768189

infection of individuals of African ancestry relative to those of European ancestry. Further studies are now required to investigate how such cellular differences in IAV susceptibility translate into population differences in clinical outcomes and susceptibility to severe influenza.

Keywords: Monocytes, single-cell 'omics, transcriptomics, ancestry, population, influenza virus

INTRODUCTION

Respiratory viruses with pandemic potential pose enormous health and economic impacts on the human population. In the last century, we have witnessed outbreaks of several coronaviruses, including SARS-CoV-2, SARS-CoV-1 and MERS, and a number of avian and swine influenza A viruses (IAV). A particularly harrowing and shared feature of these pandemics are the sudden deaths of otherwise healthy individuals (1). A hyperinflammatory state characterized by high levels of inflammatory cytokines, often referred to as a 'cytokine storm' (2, 3), has emerged as a hallmark of these severe viral infections. While still controversial, there is increasing evidence to suggest that the mononuclear phagocyte system is an important immunological determinant of this phenotype (4–6). Upon viral infection, sentinel cells such as lung-resident macrophages trigger complex signaling cascades that recruit leukocytes to the site of infection, among them monocytes. These infiltrating monocytes differentiate into monocyte-derived dendritic cells or macrophages, enabling viral clearance through the induction of the adaptive response, and help replenish the pool of tissue-resident alveolar macrophages (4, 7).

In humans, circulating monocytes are divided into classical (~80%), intermediate (~15%), and nonclassical (~5%) subsets, based on surface receptor expression of the cluster-determinant antigens CD14 and CD16 (8). While nonclassical monocytes ($CD14^{+}CD16^{++}$) are long-lived and 'patrol' healthy tissues through long-range crawling on the endothelium, classical ($CD14^{++}CD16^{-}$) and intermediate ($CD14^{++}CD16^{+}$) monocytes are recruited to the lung in response to viral infection, where they secrete inflammatory cytokines and chemokines, as well as type I interferons (IFNs) (7, 9–11). In most individuals, recruited cells help clear infection despite being susceptible to infection themselves (12, 13); yet, in some individuals, a dysfunctional immune response occurs resulting in widespread lung inflammation. Whether monocyte subsets behave differently upon viral exposure, and how direct viral sensing and exposure to secreted cytokines shape monocyte activation and differentiation are not well understood.

Variation in blood composition and cellular proportions have been shown to be one of the main factors underlying transcriptional variation in immune genes across individuals (14), with these proportions being influenced by both genetic and non-heritable factors (15–17). Recently, we characterized the genetic architecture of transcriptional responses of primary monocytes from 200 individuals of African and European ancestry to *ex vivo* challenge with viral stimuli (18). In this model, where we were able to control for viral determinants of disease (i.e. dose and strain), we reported marked inter- and intra-population differences in transcriptional responses to IAV. While our analyses revealed numerous *cis*-expression quantitative trait loci (18), genetic variants could only account for a small fraction of expression variation, in line with other studies (14, 19).

Here, we implemented single-cell RNA-sequencing (scRNA-seq) on human primary monocytes exposed to IAV to investigate (i) the effects of direct viral infection versus activation by exposure to secreted cytokines, (ii) the subset-specific responses of monocytes to viral challenge, and (iii) the extent of inter-individual and between-population variation in the proportions of monocyte subsets and the degree of monocyte susceptibility to IAV infection. Our study reveals a profound reprogramming of monocyte transcriptomes upon viral infection and shows a proinflammatory role of $CD16^{+}$ monocytes following IAV challenge. Furthermore, it highlights that African-ancestry individuals are characterized by both a higher frequency of $CD16^{+}$ monocytes and a generally lower susceptibility of their monocytes to IAV infection. Based on these results, we propose that population differences in the composition of circulating monocytes and their susceptibility to infection may contribute to the higher severity of IAV infections reported among African-ancestry individuals.

RESULTS

Using scRNA-Seq to Investigate Cellular Heterogeneity

To investigate the role of cellular heterogeneity in driving immune variability across individuals, we performed a time-course experiment where we monitored the $CD14^{+}$ fraction of peripheral blood mononuclear cells (PBMCs) from eight donors, both in the presence and absence of viral challenge. To maximize inter-individual variability, we chose individuals from two distinct ancestries whose cells demonstrated extreme responses to viral stimuli in a previous bulk RNA-seq experiment (18). Droplet-based scRNA-seq was performed on monocytes from all eight donors immediately before infection initiation (T_0), as well as at 2 (T_2), 4 (T_4), 6 (T_6), and 8 (T_8) hours post challenge with A/USSR/90/1977(H1N1) at a multiplicity of infection (MOI) equal to 1 (IAV-challenged) and mock infection (non-infected). To mitigate batch effects, we pooled IAV-challenged and non-infected cells from distinct donors in each library, assigning cells to their condition in silico *via* genetic barcoding (20). After stringent quality control where we removed low-quality, dying, and contaminants of the $CD14^{+}$ monocyte isolation, our final dataset contained 88,559 high-quality cells, among which we

predicted >99% monocyte purity at T_0 (**Figure 1A; Supplementary Figures 1** and **2**). At later time points, a substantial fraction of non-infected cells (up to 70% at T_8) were predicted to be macrophage-like, indicating monocyte differentiation over the course of the experiment. For clarity, we refer to cells as monocytes at T_0 and as monocyte-derived cells from T_2-T_8.

Stable *FCGR3A* Expression Distinguishes Monocyte Subsets Over Time

We next sought to characterize each cell by its mRNA expression of the canonical monocyte markers, CD14 and CD16, given that much of the structure in our data was associated with *FCGR3A* (aka *CD16*) mRNA expression. In droplet-based scRNA-seq, encapsulation of ambient mRNAs emanating from dying cells can occur during library preparation leading to spurious mRNA detection (21). We thus used a statistical framework to test whether *CD14* and *CD16* were expressed at a level significantly higher than expected when accounting for potential contamination from the ambient pool (*Methods*). Despite having been positively selected for the CD14 antigen, only 32.4% of monocytes significantly expressed *CD14* at T_0; this percentage further decreased at later time points and remained <15% across all time points and conditions (average 6.4% s.d.: 5.0%, **Supplementary Figures 3A–C**). On the other hand, 12.1% of monocytes significantly expressed *FCGR3A* (*CD16*) (referred to as *CD16*$^+$) at T_0, this marker proving much more stable across conditions and time points (9.3% of *CD16*$^+$ cells on average, s.d.: 1.8%, **Figure 1B** and **Supplementary Figures 3D–F**). While we deciphered classical, intermediate, and nonclassical monocytes subsets at T_0 (**Supplementary Note 1; Supplementary Figure 4** and **Supplementary Data 1**), we focus on the simpler distinction of *CD16*$^-$ and *CD16*$^+$ subsets given that positive-selection for monocytes does not capture the entire nonclassical population and that we were unable to distinguish the intermediate and nonclassical subsets after T_0.

Functional Features of Monocyte Subsets Are Conserved Upon Manipulation

To assess how transcriptional profiles of *CD16*$^-$ and *CD16*$^+$ monocytes and their derived-cells differ, we focused on the 5,681 genes expressed with a normalized $\log_2$ count > 0.1 in at least one condition, time point, and subset (**Supplementary Data 2A**). We found that the $\log_2$ fold change ($\log_2$FC) in gene expression between *CD16*$^{+/-}$ subsets remained relatively stable over the course of the experiment (Pearson r between time points >0.42

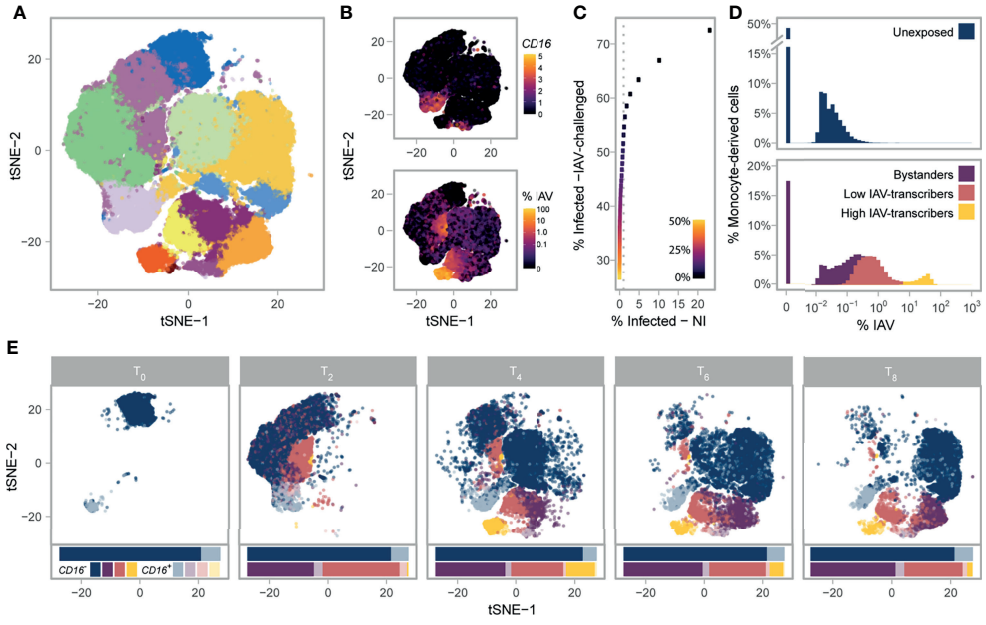


FIGURE 1 | Single-cell RNA-sequencing of 88,559 monocytes and their derived cells. **(A)** Post-QC tSNE colored by unsupervised graph-based clusters. **(B)** Post-QC tSNE colored by *FCGR3A* (*CD16*) $\log_2$ normalized counts (top), or percentage of viral mRNAs (bottom). **(C)** Determination of the maximum contamination fraction by ambient RNA. The number of non-infected cells deemed to significantly express IAV transcripts (presumed false positives) versus the number of IAV-challenged cells deemed to significantly express IAV transcripts across a range of maximum contamination fractions from 1-50% (color bar). Dotted grey line is drawn at 1% on the x-axis. A maximum contamination fraction of 10% results in 1% of non-infected cells being classified as infected (false positive proxy), and half of IAV-challenged cells showing evidence of viral transcription. **(D)** Distribution of counts of viral origin across all donors, from T_2 to T_8. Cells are shown separately for non-infected (top) and IAV-challenged (bottom) conditions. Fill color reflects the cell state assignments. Note that the threshold used to define infected cells is dependent on the number of viral mRNAs in the ambient pool, and varies across libraries. **(E)** Post-QC tSNE stratified by time point. For each time point, cells are colored according to their *CD16*$^{+/-}$ status (see key) and their assigned cell state (same as depicted in D). For each condition and time point, stacked bar charts below the tSNE represent the relative proportions of the various cell states and subsets. IAV, Influenza A virus; NI, non-infected; tSNE, t-distributed Stochastic Neighbor Embedding.

and >0.52 for the non-infected and IAV-challenged conditions respectively, p-values<$2.2x10^{-16}$; **Supplementary Figure 5A**), and differentially expressed genes between $CD16^{+/-}$ subsets were largely the same across conditions (Pearson r = 0.92, p-value < $2.2x10^{-16}$; **Supplementary Figure 5B**). We thus searched for genes that were consistently differentially expressed between $CD16^{+}$ and $CD16^{-}$ cells across all time points (including T_0), conditions, and donors. We identified 266 genes over-expressed ($log_2FC>0.2$, FDR<1%) in $CD16^{+}$ cells relative to $CD16^{-}$ cells, and 389 genes that showed the opposite pattern, and performed a GO-term enrichment analysis on these genes (**Supplementary Data 2B**). Consistent with previous reports (22–24), $CD16^{-}$ subsets were characterized by high expression of several proinflammatory *S100 Calcium Binding Proteins* (*S100A12*, *S100A9*, and *S100A8*), contributing to a sizable GO-term enrichment in the defense response to fungus pathway (GO:0050832: OR=41.3, FDR=$4.9x10^{-4}$), while $CD16^{+}$ subsets were characterized by high expression of Fc-gamma receptor signaling pathway genes (GO:0038096: OR=8.7, FDR=$6.2x10^{-6}$). Notably, $CD16^{+}$ subsets over-expressed several type I IFN stimulated genes (ISGs) relative to $CD16^{-}$ subsets (e.g. GO:0071357: OR=5.3, FDR=$2.6x10^{-3}$), including the well-known viral restriction factors *IFITM3* and *OAS1*. Collectively, these results demonstrate *CD16* is a reliable marker at the mRNA level and that $CD16^{+/-}$ monocyte subsets maintain functional differences upon manipulation.

scRNA-Seq Highlights Heterogeneity in Monocyte Susceptibility and Viral Transcription

Using the presence of IAV transcripts as a proxy for infection (**Figure 1B**), we next sought to distinguish cells that were successfully infected from those that were not. Among monocyte-derived cells that were exposed to IAV, we found that 50.3% expressed IAV transcripts above ambient levels when allowing up to 10% of mRNAs to come from the ambient pool. In contrast, less than 1% of non-infected cells showed evidence of viral transcription, supporting the validity of the threshold used to detect IAV expressing cells (**Figure 1C**). We deemed cells with statistical evidence for expression of IAV transcripts from the IAV-challenged condition as 'infected', while the remaining cells from this condition were considered as 'bystanders', as these either did not come into contact with the virus or were able to fully repress viral mRNA transcription. When comparing the percentage of infected cells between subsets, we noticed that $CD16^{+}$ cells were slightly less likely to be infected than $CD16^{-}$ cells (42.3% sd: 4.0% for $CD16^{+}$ relative to 49.4% sd: 5.4% for $CD16^{-}$, generalized linear model with $CD16^{+/-}$ status, donor, and time point as covariates, p-value=0.006), possibly related to the higher expression of ISGs observed in this subset (**Supplementary Data 2A, B**). We further confirmed experimentally that intermediate and nonclassical ($CD16^{++}$) monocytes display increased resistance to IAV challenge by monitoring intracellular IAV nucleoprotein by flow cytometry in PBMCs challenged with another H1N1 strain (**Supplementary Figure 6**).

We observed that the proportions of viral mRNAs among infected cells were bimodally distributed and largely varied between the clusters identified in our unsupervised analysis (**Figure 1D**). We used a Gaussian mixture model to locate the two modes of the distribution and further sub-classify infected cells into those with lower IAV mRNA levels (<1-6%) and those with higher IAV mRNA levels (6-83%); while viral mRNA levels are dictated by both the rate of transcription and degradation, for simplicity we refer to these infected cell states as 'low IAV-transcribers' and 'high IAV-transcribers', respectively. The proportions of infected cells among individuals remained largely unchanged over the course of the experiment; however, high IAV-transcribers were virtually absent at 2h (<2% of infected cells), peaked to ~36% of IAV-infected cells at 4h, and decreased to 8.5% by 8h, suggesting that high-IAV transcribers represent a transient state of IAV-infection preceding IAV-induced apoptosis (**Figure 1E**). These results reveal profound heterogeneity in monocyte susceptibility and subsequent viral transcription upon IAV challenge.

Interplay of Cytokine and Ribosome Networks Drive Cell States Upon Infection

To characterize host transcriptional responses over time, we next subsampled each subset ($CD16^{-}$/$CD16^{+}$), cell state (unexposed, bystander, infected), and time point in our scRNA-seq data to a uniform number of cells to avoid biases emanating from differences in sample sizes. Limited by the number of $CD16^{+}$ high IAV-transcribing cells, we randomly sampled 100 cells from each subgroup, while ensuring representation of all donors. We then focused on the 6,669 host genes with average log_2 normalized count >0.1 in at least one subgroup (**Supplementary Data 3A**). Overall, $CD16^{-}$ and $CD16^{+}$ subsets behaved similarly upon stimulation with changes in gene expression between cell states being strongly correlated among subsets (Pearson r=0.83-0.95, p-values<$2.2x10^{-16}$; **Supplementary Figure 7**). GO term enrichment analyses of shared responses (FDR<1% & $log_2FC>0.2$ in same direction in both subsets) uncovered several functional categories interacting to shape the activation state of cells (**Figure 2A**; **Supplementary Data 3B**). Both bystander and infected cells showed increased mRNA expression of genes involved in antigen processing and presentation *via* class I MHC (GO:0019885, OR=53.7, FDR=$2.0x10^{-6}$) and ISGs (GO:0034340, OR=14.8, FDR=$3.3x10^{-20}$). Yet, bystander cells showed increased mRNA expression of ISGs and defense response to virus pathways relative to infected cells (GO:0034340, OR=13.4, FDR=$4.4x10^{-7}$; GO:0051607, OR=9.0, FDR=$2.1x10^{-7}$), while infected cells displayed higher mRNA expression of mitochondrial (GO:0005743, OR=4.7, FDR=$3.3x10^{-3}$) and ribosomal genes (GO:0005840, OR=117, FDR=$1.0x10^{-78}$). Notably, type-I IFN genes themselves tended to be preferentially expressed by infected cells (e.g. log_2 normalized count at 6h for *IFNB1* ~0.12/0.29 in $CD16^{-}$ and $CD16^{+}$ subsets, respectively, vs <0.01 for bystander cells of both subsets), although this difference was only barely significant in our setting (FDR=0.03), likely due to the highly transient nature of IFN expression.

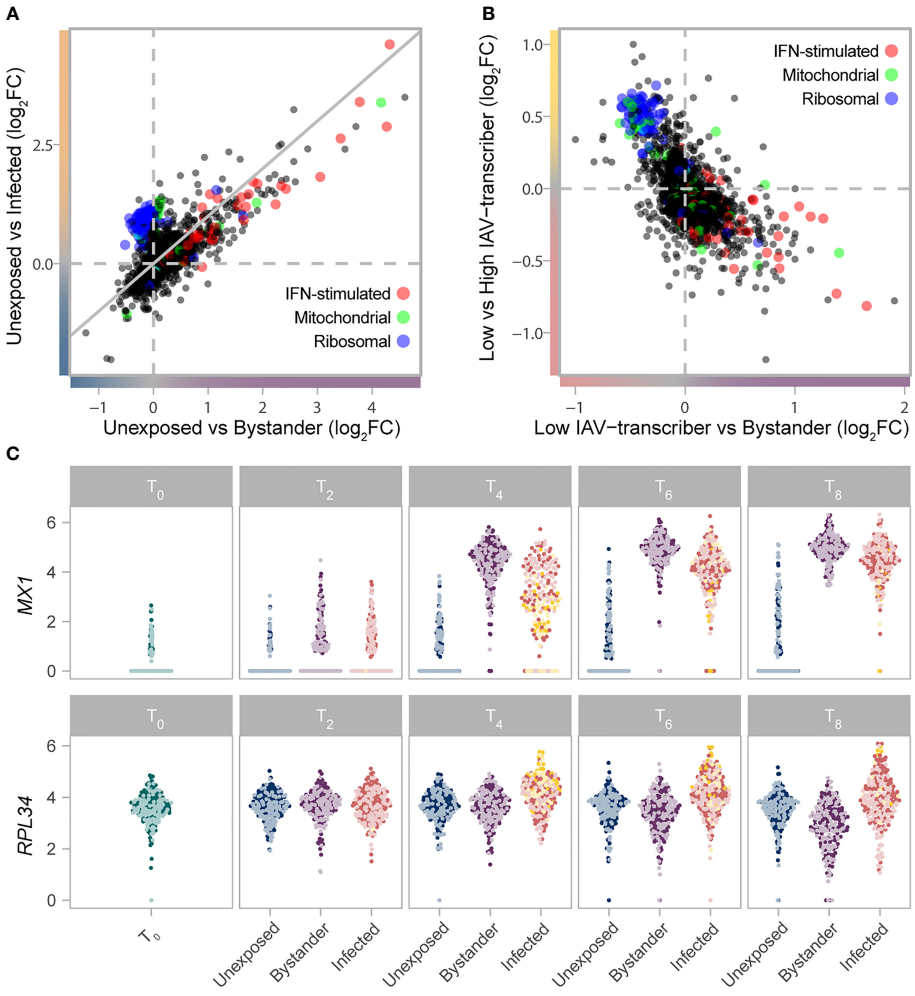


FIGURE 2 | Gradient of mRNA expression from ribosomal and IFN-stimulated genes separates bystander and infected cells. **(A)** Transcriptional responses of cells upon IAV challenge (T_2-T_8) highlight the interplay between IFN-stimulated (GO:0034340), ribosomal (GO:0005840), and mitochondrial (GO:0005743) genes. The log_2FC change in gene expression between unexposed and bystander cells is plotted on the x-axis, while the log_2FC change in gene expression between unexposed and infected cells is plotted on the y-axis. Values are plotted based on a meta-analysis across time points and subsets, of a subsampled dataset with balanced representation of all donors. **(B)** The interplay between IFN-stimulated (GO:0034340), ribosomal (GO:0005840), and mitochondrial (GO:0005743) genes among cells exposed to IAV. The log_2FC change in gene expression between low IAV-transcribing infected and bystander cells is plotted on the x-axis, while the log_2FC change in gene expression between low IAV-transcribing infected and high IAV-transcribing infected cells is plotted on the y-axis. Values are plotted based on a meta-analysis across monocyte subsets at T_4. **(C)** mRNA expression levels of representative IFN-stimulated (*MX1*) and ribosomal (*RPL34*) genes across the subsampled dataset. Colors reflect the cell state and subset assignment depicted in **Figures 1D, E**. IFN, Interferon; FC, fold change.

Among infected cells, ribosomal genes showed higher activity among high IAV-transcribing cells relative to low IAV-transcribing cells (**Figure 2B**, comparison only made at T_4 due to sample size constraints, e.g. GO:0019083, OR=137, FDR=$6.1x10^{-65}$). This observation is consistent with the notion that the expression of viral proteins is dependent on cellular ribosomes, with recent data suggesting that IAVs do not induce a global shut-off of cellular translation but rather a reshaping of the translation landscape (25–27). Likewise, among bystander cells, numerous ribosomal genes were downregulated at later time points relative to unexposed cells (**Figures 2A, C**; GO:0019083, OR=5.3, FDR=$4.2x10^{-6}$), suggesting that repression of ribosomal subunits plays an active role in limiting viral replication. Collectively, these results suggest that expression of ISGs and ribosomal genes interact to shape cell states upon IAV challenge.

Increased IRF and STAT Activity Drives Stronger Antiviral Response

Despite qualitatively similar responses to infection between $CD16^-/CD16^+$ subsets (**Supplementary Figure 7**), we hypothesized that subtle differences in the intensity of such responses might contribute to the increased resistance of *CD16*$^+$ cells to infection. We thus performed an interaction test on the subsampled scRNA-seq data, and searched for genes for which transcriptional response upon IAV challenge differed between *CD16*$^-$ and *CD16*$^+$ subsets in either infected and/or bystander

cells (**Supplementary Figures 7A, B**; **Supplementary Data 3A**). At FDR≤1%, we identified a total of 335 such genes, of which 98 differed between subsets only in bystander cells, 144 only in infected cells, and 93 in both. Hierarchical clustering highlighted eight major patterns of transcriptional responses (modules) among the 335 genes, several of which were associated with specific biological functions (**Figure 3A; Supplementary Data 3C, D**). Notably, module 1 (green) was enriched for genes in the antiviral response pathway (GO:0051607, OR=23.2, FDR=5.43×10^{-7}) and displayed a stronger response in infected *CD16*$^+$ cells relative to *CD16*$^-$ infected cells. Of additional interest was the transient *CD16*$^+$-specific transcription of the inflammatory cytokine genes *IL6* and *TNF*, following viral challenge (**Figure 3B**). We also found that several genes involved in the regulation and production of IL-6 and TNFα were over-expressed in *CD16*$^+$ subsets at all time points and conditions (**Supplementary Data 2B**), but only see active transcription of the cytokines upon viral exposure. These results reveal the strong antiviral and inflammatory potential of *CD16*$^+$ relative to *CD16*$^-$ monocytes in response to viral infection (28).

We next sought to characterize the regulatory architecture underlying the 335 genes whose transcriptional response to IAV challenge differed between monocyte subsets. Using SCENIC (29), we identified 113 high-confidence gene regulatory networks, or 'regulons', which were active in non-infected and/or IAV-challenged cells, each composed of a transcription factor (TF) and a set of predicted targets (genes). We used these 113 regulons to search for an enrichment/depletion of TF targets among the eight modules of genes displaying subset-specific response to infection (**Supplementary Data 3E**). Among modules associated with an increased expression in cells exposed to IAV (modules 1-5), we observed a widespread over-representation of targets of IFN regulatory factors (IRFs) and signal transducing and activators of transcription (STATs) (**Figure 3C**), reinforcing the central role of the IFN response upon IAV challenge. Interestingly, several of these factors displayed subset-specific activity themselves in response to IAV (*IRF1/2/7* and *STAT1/2/3*, FDR<1%), mirroring the expression patterns of module 1 (Pearson r>0.92). These results collectively highlight a *CD16*$^+$-specific inflammatory response upon IAV challenge and suggest stronger activation of IRF and STAT transcription factors as a driver of the increased antiviral response observed in *CD16*$^+$ cells upon IAV infection.

Basal Activation Differences Correlate With Monocyte Susceptibility

To explore the degree of inter-individual variation upon viral challenge, we next quantified IAV transcripts in the monocyte-derived cells of each individual, and created pseudo-bulk estimates by averaging the percent of viral mRNAs per-cell across all cells from each donor at each time point (**Figure 4A**). While viral mRNAs peaked at the same time for all individuals, we observed extensive variation in the levels of viral mRNAs and percentages of infected cells across individuals (**Figure 4B**). To identify specific genes that might underlie infection potential, we focused on the 4,589 genes that were expressed at >0.1 $\log_2$ normalized counts in at least one canonical monocyte subset at T_0. We identified a total of 3,131 genes that differed among our eight donors in either classical, intermediate, and/or nonclassical monocyte subsets (Kruskal-Wallis Rank Test, FDR=1%; **Supplementary Data 4A**). Within each subset, focusing on genes that significantly differed between donors, we

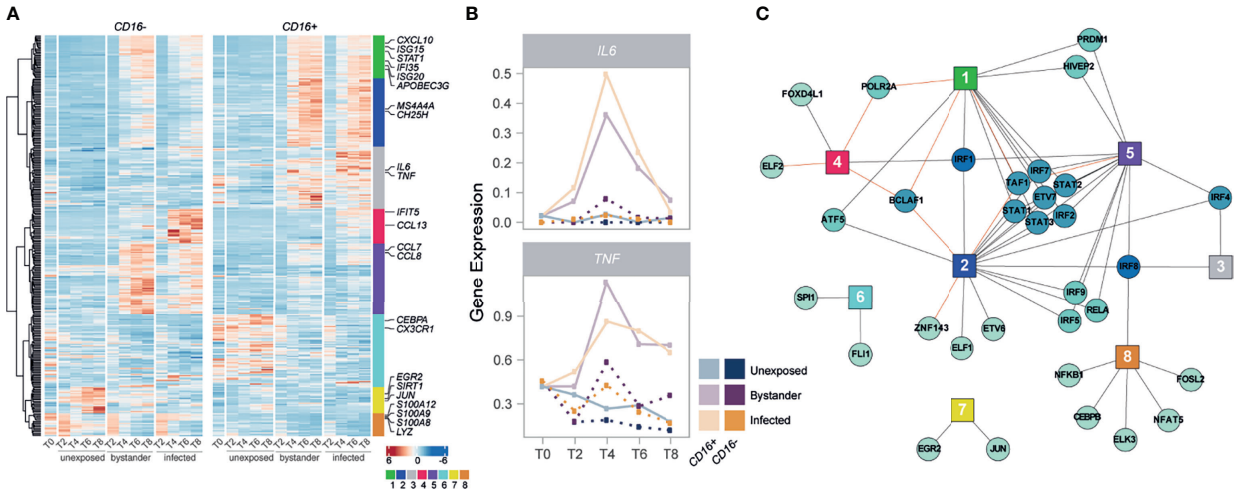


FIGURE 3 | IRFs and STATs have a central role in the subset-specific responses to IAV infection. **(A)** Heatmap of scaled gene expression from 335 genes displaying a subset-specific response to infection challenge. Genes are grouped into 8 modules based on hierarchical clustering of their expression patterns. Representative genes from each module are labelled. **(B)** Mean expression over time of *IL6* and *TNF*, across the different monocyte subsets and cell states. **(C)** Network of transcription factors (round nodes) associated with each gene expression module (square nodes). Transcription factor nodes are colored according to the number of modules they are associated with. Black lines represent enrichments of the module in TF targets, while red lines represent depletions.

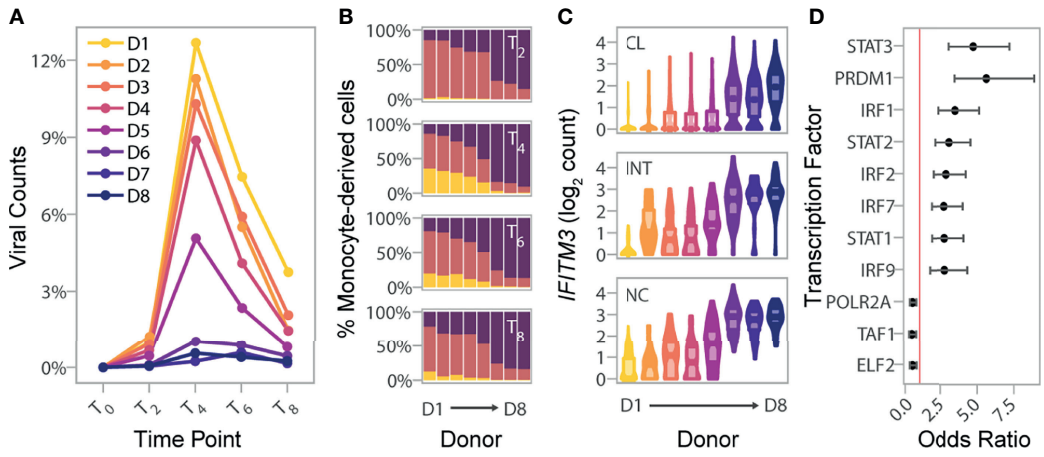


FIGURE 4 | Basal IRF/STAT-induced transcriptional network underlies inter-individual differences in monocyte susceptibility and IAV levels. **(A)** Pseudo-bulk estimates of the percentage of counts of viral origin in IAV-challenged condition (T_2-T_8). Donors are colored based on the rank of these pseudo-bulk estimates at the peak of viral transcription, T_4, from that with highest observed viral mRNA level (D1) to that of the lowest (D8). **(B)** Proportions of cell states from the IAV-challenged condition at T_2, T_4, T_6, and T_8, in the eight donors. X-axis is ordered by decreasing viral mRNA levels found at T_4 (D1-D8). **(C)** Log normalized expression values of *IFITM3* across all cells, stratified by canonical monocyte subsets, and separated by donor. Colors reflect the different donors depicted in **(A)** For each donor and monocyte subset, the violin plots show the full distribution of *IFITM3* expression across individual cells and boxplots highlight the median and interquartile range. **(D)** Enrichment of SCENIC-predicted targets among the 118 genes whose basal expression at T_0 correlates with the percentage of infected cells at later time points (odds ratio and 95% confidence interval). Red line designates an odds ratio equal to 1. Only TFs significantly enriched among the 118 candidate genes are shown (FDR < 0.05). CL, classical; INT, intermediate; NC, nonclassical.

searched for those for which mean expression at basal state was correlated with the percentage of infected cells at T_4 among our eight donors. Despite our limited sample size, we found that cellular susceptibility was strongly correlated with basal expression of the well-known host viral restriction factor *IFITM3*. Although it reached significance only in nonclassical monocytes (FDR~1%), the association remained strong in other subsets (p-value< 4.1×10^{-4}; **Figure 4C**).

We next relaxed our search to all genes for which basal expression showed nominal correlation (p-value<0.01) with the percentage of infected cells at T_4. Depending on the monocyte subset, between 3.6 to 8.3% of genes matched these criteria, resulting in a set of 118 genes displaying correlation with monocyte susceptibility in at least one subset. These 118 genes were collectively enriched for several related biological processes such as defense response to virus (GO:0051607, OR=15.3, FDR=9.2×10^{-19}) and ISGs (GO:0034340, OR=19.6 FDR=8.4×10^{-15}) (**Supplementary Data 4B**). Among genes contributing to this enrichment, we found additional antiviral genes such as *OAS3*, and *MX1*, as well as the critical TF, *IRF7*, involved in the severity of IAV-infection both in mice and humans (30–32). Finally, overlap with the TF targets identified by SCENIC revealed strong enrichments of several IRFs and STATs among the 118 genes, including IRF7, as well as STAT1, STAT2 and IRF9 that form the tripartite IFN-stimulated gene factor 3 (ISGF3) (**Figure 4D**; **Supplementary Data 4C**). Together, our results provide evidence that the basal mRNA expression of genes related to IFN-induced and antiviral responses are indicative of the proportion of cells that will become infected in the first cycle of IAV infection.

African-Ancestry Monocytes Are More Resistant to Infection

Lastly, we wondered how our findings of inter-individual variation might extrapolate to the population level. In a previous study (18), we challenged the primary monocytes from 200 Belgian individuals of African (AFB) and European (EUB) ancestry with the same IAV strain and MOI used in the present study, and performed bulk RNA-seq at 6 hours post infection (hpi). While basal (T_0) expression profiles were not collected, flow cytometry labelling of CD14 and CD16 was performed on the $CD14^+$-selected monocytes for the majority of donors. Interestingly, AFB individuals had higher proportions of $CD16^+$ cells than EUB individuals (**Figure 5A**; **Supplementary Figure 8**). In light of our findings that $CD16^+$ cells are more resistant to IAV infection, we hypothesized that this might translate to lower infection rates among AFB monocytes relative to EUB monocytes.

To test this hypothesis, we mapped the bulk RNA-seq profiles collected 6hpi challenge with IAV for the 200 individuals to a combined human-IAV reference. Excluding 1 sample with low quality RNAs, we found that 0.02-13.5% of RNA-seq reads from each sample were of viral origin (**Figure 5B**). Reassuringly, these percentages correlated with IAV mRNA levels estimated from the single-cell experiment across all time points for the eight donors used in the present study (Pearson r>0.84, p-values<8.9×10^{-3}), with the strongest correlation being observed at the peak of viral transcription (T_4) (Pearson r=0.97, p-value=5.1×10^{-5}). These observations indicate that *ex vivo* cellular susceptibility is highly reproducible among individuals, even across different experimental protocols and technologies.

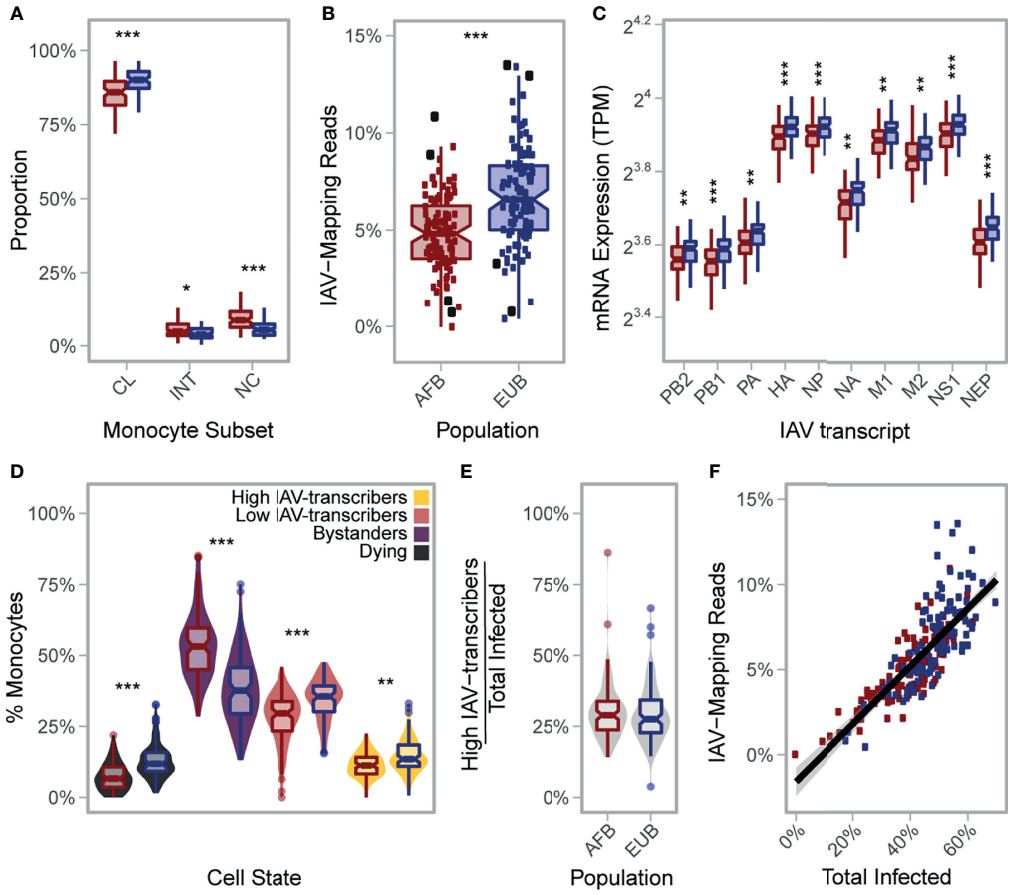


FIGURE 5 | African-ancestry individuals display increased number of *CD16*$^{+}$ cells and lower susceptibility to IAV infection. **(A)** Variation in the number of classical (CD14^{++}CD16^{-}), intermediate (CD14^{++}CD16^{+}), and nonclassical (CD14^{+}CD16^{++}) monocytes across African- and European- ancestry individuals following CD14^{+} selection from PBMCs (n_{AFB} = 89, n_{EUB} = 85). Colors reflect population (AFB in red and EUB in blue). All three subsets are significantly different between populations (p-value<0.01). **(B)** Inter- and intra-population variation in the percentage of RNA-seq reads mapping to the IAV genome (n_{AFB} = 100, n_{EUB} = 99). Colors reflect population (AFB in red and EUB in blue). The percentage of RNA-seq reads mapping to the IAV genome is significantly higher in European-ancestry individuals relative to African-ancestry individuals (p-value = 5.3×10^{-8}). Donors used in the scRNA-seq experiment (n_{AFB} = 4, n_{EUB} = 4) are designated with enlarged black squares. **(C)** Inter- and intra-population variation in viral mRNA expression at 6hpi (n_{AFB} = 100, n_{EUB} = 99). Expression levels for each of the 10 primary transcripts of IAV are plotted. Colors reflect population (AFB in red and EUB in blue). All IAV transcripts are significantly higher in European-ancestry individuals on average (p-value<0.001). **(D)** Estimated distribution of the percentage of cells from each cell state in the bulk RNA-seq data (n_{AFB} = 100, n_{EUB} = 99). Fill colors reflect cell state assignments, while outlines of boxplots reflect population (AFB in red and EUB in blue). **(E)** Distribution of the percentage of high IAV-transcribers among infected cells, stratified by population. One individual with no infected cell was excluded (n_{AFB} = 99, n_{EUB} = 99). **(F)** Percentage of RNA-seq reads of viral origin as a function of the estimated proportion of infected cells (n_{AFB} = 100, n_{EUB} = 99), colored by population (AFB in red and EUB in blue). **(A, C)** Outlier points are not displayed. AFB, African-ancestry individuals from Belgium; EUB, European-ancestry individuals from Belgium; TPM, transcripts per million; IAV, influenza A virus; MFI, mean fluorescent intensity; CL, classical; INT, intermediate; NC, nonclassical. *p-value < 0.01; **p-value < 0.001; ***p-value < 0.0001.

Among the 199 bulk profiles, AFB and EUB samples presented overlapping but significantly shifted distributions of total IAV-mapping reads (**Figure 5B**, 4.9% vs. 6.8% of reads, respectively, Wilcoxon p-value=5.3×10^{-8}), and of each of the 10 primary viral transcripts (**Figure 5C**, Wilcoxon p-values<5.5×10^{-4}).

Using the transcriptional profiles obtained from the scRNA-seq data at T_6, we estimated the proportion of reads coming from each inferred cell state in these bulk RNA-seq profiles (**Figures 5D, E; Supplementary Note 2** and **Supplementary Figure 9A**). We found that, on average, AFB monocytes were more resistant to IAV infection than EUB monocytes (39.2% vs. 48.9% infected, respectively, Wilcoxon p-value=5.3×10^{-10}). Differences in the estimated percentage of infected cells alone explained 63% of the inter-individual variability in viral mRNA levels (**Figure 5F**), and was sufficient to account for the observed difference in viral mRNA levels between AFB and EUB individuals (p-value=0.16 after adjusting on infected cells, compared to p-value=5.3×10^{-8} without adjustment). Nonetheless, variation in the percentage of high/low IAV-transcribers among infected cells accounted for an additional 19% of variance in viral mRNA expression (**Supplementary Note 2** and **Supplementary Figure 9B**). Finally, the ratio of

CD16^{+}/*CD16*$^{-}$ cells negatively correlated with the percentage of infected cells, albeit weakly (-0.27, *p*-value=0.0165 adjusted on population). Altogether, these results show that population differences in viral mRNA levels are primarily driven by the overall proportion of cells that will ultimately become infected, with only a fraction of the differences being attributable to the different proportions of *CD16*$^{+/-}$ subsets observed in individuals of African and European ancestry.

DISCUSSION AND HYPOTHESIS

We performed scRNA-seq on primary monocytes, before and after *ex vivo* IAV challenge, to assess transcriptional differences between monocytes infected by IAV (i.e. infected) versus those activated only by exposure to secreted cytokines (i.e. bystanders), and to identify subset-specific responses of monocytes to viral challenge. We found that bystander cells display increased mRNA expression of ISGs relative to infected cells; yet, we additionally observed both an induction of ribosomal gene mRNA expression in IAV-transcribing cells and a down regulation of these genes in bystander cells at later time points. While the former is likely induced by the virus to enhance mRNA translation (33), the repression of ribosomal expression observed in bystander cells may reflect a host mechanism to contain infection by shutting down the translational machinery of neighboring cells, and we speculate that this may hold true across other cells types and constitute a general cellular defense mechanism against viral infections. Interestingly, the interplay of ribosomal and ISG expression also distinguished infected cells into two distinct states (high and low IAV-transcribers), providing an explanation for the high cell-to-cell variation in IAV replication observed among circulating monocytes, which has also been documented in other cell types and during natural infection (34–41). Notably, type-I IFN genes themselves tended to be preferentially expressed by infected cells in a highly transient manner, suggesting a potential role of monocyte infection in the triggering of the type I IFN response among bystander cells.

While these patterns were generally shared across *CD16*$^{-}$ and *CD16*$^{+}$ subsets, we found *CD16*$^{+}$ cells to be slightly more resistant to infection. This is likely attributable to their higher absolute expression of some ISGs relative to *CD16*$^{-}$ cells (independent of viral exposure), as well as their more robust upregulation of antiviral genes upon IAV challenge, which we found to be driven by stronger activity of IRF transcription factors. Interestingly, *CD16*$^{+}$ cells displayed transient mRNA expression of *IL6* and *TNF* upon viral exposure (both infected and bystander cells), two cytokines that have been widely implicated in cytokine storms (5). Collectively, these findings highlight the opposing roles of ISG and ribosomal gene mRNA expression on viral transcription, and reveal the stronger antiviral and pro-inflammatory potential of *CD16*$^{+}$ monocyte subsets.

At the population level, we found that the ratio of *CD16*$^{+}$/*CD16*$^{-}$ at basal state was predictive of the percentage of monocytes that were susceptible to IAV infection, and observed that African-ancestry individuals, from our sample, harbored more *CD16*$^{+}$ monocytes on average than European-ancestry individuals residing in the same city (Ghent, Belgium), consistent with previous observations (42). Independently of monocyte subset proportions, we identified that individuals presenting lower monocyte susceptibility to IAV had a higher basal activation of an IRF/STAT-driven antiviral program. These findings suggest that the fate of a monocyte hinges upon its basal activation state, and that the infection potential differs both within an individuals' monocyte population, in part based on the differentiation status of the cell (i.e. *CD16*-positivity), but also between individuals, where a *CD16*$^{-}$ cell from one individual may have a higher antiviral state than a *CD16*$^{+}$ cell from another individual.

Our finding of a decreased ability of IAV to infect and replicate in monocytes from individuals of African-ancestry was recently replicated in an independent cohort of American individuals with varying levels of African and European ancestry whose PBMCs were challenged with the 2009 pandemic H1N1 strain (43). While the cause of these population differences remains to be determined, we did not find evidence that strong-effect genetic factors, nor evidence of past exposure to H1N1, could explain such an association with the viral replication phenotype. Nevertheless, the observed inter- and intra-population differences are noteworthy in and of themselves, and may reflect the influence of both weak-effect genetic loci, and non-heritable factors, such as stress, nutrition or lifestyle, on transcriptional variation of immune genes (14, 15). Future studies are needed to determine if such population differences hold true across other cell types, such as lung epithelial cells.

Given our finding that *CD16*$^{+}$ subsets are the main drivers of inflammatory cytokine gene expression such as *IL6* and *TNF*, and that African-ancestry individuals harbor a larger fraction of these subsets, it is tangible to conceive that monocyte subset composition prior to infection may influence disease outcome. A lower percentage of infected monocytes could also contribute to a faster disease progression, as we find that infected monocytes continue to express antigen-presenting genes. Thus, a higher number of infected cells could lead to a stronger activation of the adaptive immune system. In support of these hypotheses, patients with severe influenza and COVID-19 harbor higher proportions of intermediate monocytes in peripheral blood than patients with mild disease (44, 45), and African Americans are more often hospitalized than other self-defined ethnic groups by both influenza (46, 47) and COVID-19 (48, 49), even when adjusting for age and various social factors such as poverty and vaccination status. In light of these observations, we hypothesize that the higher percentage of *CD16*$^{+}$ monocyte observed among African-ancestry individuals may, in conjunction with a stronger basal activation of their monocytes, contribute to poor infectious outcomes. Further studies are now needed to formally establish the clinical relevance of monocyte heterogeneity in the context of viral infections, IAV in particular, and determine its potential use as a biomarker.

LIMITATIONS OF STUDY

In this Hypothesis and Theory article, we analyze single-cell transcriptional heterogeneity of circulating monocytes before and after *ex vivo* IAV challenge, and propose that differences in basal monocyte activation underlie population disparities in cellular susceptibility to IAV. We acknowledge that our study is based on positively-selected monocytes isolated from PBMCs, and that infection of this cell population *in vivo* would be expected to take place in the lung, after an initial infection has been established. Future studies should investigate how these findings translate to other cell populations - including lung epithelial cells and resident monocyte and macrophage populations - and whether they influence clinical outcomes. This could be achieved, for instance, by using single cell techniques to measure how nasal epithelial cells from healthy patients of various ancestries differ in their expression of viral RNAs and proteins upon IAV challenge. Additionally, sputum extract could be collected from mild and severe influenza patients of both ancestries, to compare the single cell transcriptome of lung epithelial cells and resident macrophage populations, both across ancestries and in relation with disease severity. Another caveat of the study is the lack of detailed lifestyle observations in the cohort used, precluding us from examining in further detail the influence of non-genetic factors. Further studies are now needed to evaluate how non-genetic factors, such as social status, chronic stress levels (and the induced physiological response), previous exposures to pathogens or even the microbiome, could contribute to shape basal monocyte activation and prime the innate immune response to viral infections.

METHODS

Experimental Model and Subjects

All individuals from this study were part of the EVOIMMUNOPOP cohort, which has been previously described (18). Human blood was obtained from healthy volunteers who gave informed consent, and the PBMC fraction was isolated and frozen. In brief, 200 healthy male donors living in Belgium of self-reported African descent (AFB) or European descent (EUB) were recruited. Inclusion was restricted to nominally healthy individuals between 19 and 50 years of age at the time of sample collection. The majority of our African-descent individuals originated from West Central Africa, with >90% of our sample being born in either Cameroon or Congo. Serological testing was performed for all donors to exclude those with serological signs of past or ongoing infection with human immunodeficiency virus (HIV), hepatitis B virus (HBV) or hepatitis C virus (HCV).

Single-Cell Analyses and RNA-Sequencing

For eight selected donors [4 individuals from each ancestry, selected from extremes of the first principal component of gene expression in our previous study of monocyte response to IAV challenge (18)], 100×10^6 PBMCs were thawed, washed twice and resuspended in complete medium: pre-warmed RPMI-1640 Glutamax medium, supplemented with 10% FCS and 1% penicillin/streptomycin (Cat#15140-122, Life Technologies). Monocytes were then positively selected with magnetic CD14 microbeads, according to the manufacturer's instructions (Cat#130-050-201, Miltenyi Biotec). The number of monocytes was determined with the Countless2 automated cell counter system (Cat#AMQAX1000, ThermoFisher Scientific) in the presence of trypan blue. For each donor, monocytes were seeded at 0.5×10^6 monocytes per well on 24-well NUNC plates in 500 µL of complete media and allowed to rest for one hour at 37°C under 5% CO_2. Five-hundred microliters of complete media (non-infected) or A/USSR/90/1977(H1N1) at a concentration of 1×10^6 pfu/mL in complete media (IAV-challenged, MOI=1) were added to each sample. Following one hour of staging at 4°C, plates were centrifuged at 1300 rpm for 10 minutes at 4°C, media was removed by pipette, and each well was washed with 1mL complete media. The spin was repeated, media removed by pipette, and samples were resuspended in 1mL pre-warmed complete media before being transferred to an incubator at 37°C under 5% CO_2 to initiate infection (T_0).

At each time point (T_0, T_2, T_4, T_6, and T_8), samples were mixed by pipetting and transferred to Eppendorf tubes. Wells were washed with 300uL of PBS + 0.04% BSA and transferred to the same tubes. Collection tubes were centrifuged at 1300 rpm for 10 minutes, media was removed and replaced with 1mL PBS + 0.04% BSA and an aliquot of 10µL was taken to count each sample on a Countless2 automated cell counter system, before repeating the centrifugations. Individual samples were adjusted to 2×10^6 live cells/mL.

Samples were multiplexed for running on the 10X Chromium (Cat#120223 & 1000074, 10X Genomics) by mixing equal proportions from 6-8 samples in a manner that balanced conditions and allowed us to assess for batch effects across lanes (**Supplementary Table 1**). Multiplexed samples were counted with the Countless2 automated cell counter system and adjusted to target recovery of 10,000 cells per reaction of the Chromium Single Cell 3' Reagent Kits v3 (Cat#1000092 & 1000078, 10X Genomics) assuming a recovery rate of 50%. GEM Generation & Barcoding, Post GEM-RT Cleanup & cDNA Amplification, and 3' Gene Expression Library Construction were performed as per manufacturer's instructions (50). All 13 libraries were mixed prior to sequencing across 13 different lanes from an Illumina HiSeq X (28bp barcode + 91bp insert – target 400 M reads pairs per lane), leading to a total of 5.3 billon reads.

Sample Genotyping

Genotyping data [accession EGAS00001001895] were obtained for all 200 individuals from the EvoImmunoPop cohort based on both Illumina HumanOmni5-Quad BeadChips and whole-exome sequencing with the Nextera Rapid Capture Expanded Exome kit (18). The 3,782,260 SNPs obtained after stringent quality control were then used for imputation, based on the 1,000 Genomes Project imputation reference panel (Phase 1 v3.2010/11/23) (51), leading to a final set of 19,619,457 high-quality SNPs, of which 7,766,248 SNPs had a MAF ≥5% in our cohort.

Processing of scRNA-Seq Data

Basic pre-processing of the sequencing data was performed with CellRanger v3.0.2 (52), including the *mkfastq*, *count*, and *aggr* commands. Default parameters and our combined human-IAV reference were used, and batch correction was disabled in the *aggr* command. Cell-containing droplets (n=132,130) were traced back to individual donors using two independent methods, Demuxlet and SoupOrCell, which capitalize on genetic variation in the sequencing reads (20, 53). Barcodes with ambiguous and/or non-concordant calls between the two programs were used to establish suitable QC metrics. We found that barcodes deemed as *doublets* (i.e. the droplet contained two or more cells originating from different donors) were more likely to be nearest-neighbors in a *knn*-graph with other *doublets* than assigned *singlets*. We used this feature to identify droplets presumed to contain two or more cells originating from the same donor; barcodes with > 5 *doublets* as nearest-neighbors were excluded from further analysis (**Supplementary Figures 1A, B**). Additionally, droplets containing low-quality cells (i.e. damaged, dying) were excluded using the following thresholds: <1500 total counts, <500 genes, or >50% mitochondrial gene content (**Supplementary Figure 1C**). This QC resulted in 96,386 single cells.

Transcriptomes (i.e. counts) were adjusted for the presence of ambient RNA with SoupX, (https://github.com/constantAmateur/SoupX, accessed November 28, 2019) (21), using estimated contamination fractions (per 10X library) from SoupOrCell (53). SoupX-adjusted counts were normalized using pool-based size factors followed by deconvolution as implemented in the scran R package (54). Feature selection was performed by (i) constructing a mean-variance trend in the log-counts and retaining genes found to exhibit more variation than expected assuming Poisson-distributed technical noise, as implemented in the makeTechTrend and TrendVar functions from package scran (54), and (ii) selecting genes expressed in at least 25 cells (n=22,603). The first 10 PCs of the data were retained for data visualization and clustering analyses. Graph-based clustering was performed by building the shared nearest-neighbor graph with the buildSNNGraph function from scran (54) using a series of k values, and cell clusters were defined with the igraph Walktrap algorithm (55). Similar clustering results were obtained based on the *knn*-graphs generated using k=25, 50, 75, and 100, and k=25 was used for all downstream analyses (**Supplementary Figure 2A**). Cell types were predicted using SingleR and the built-in BlueprintEncodeData reference (56). Based on the clustering and cell-type predictions, we removed cells belonging to clusters associated with lymphoid cell types or low QC metrics from downstream analyses (**Supplementary Figures 2B, C**).

Accounting for Ambient RNA Contamination in scRNA-Seq Data and Assigning Cell States

Droplet-based scRNA-seq methods capture ambient mRNAs present in the cell suspension in addition to cell specific mRNAs. To estimate which cells in our experiment were genuinely expressing mRNAs for *CD14*, *FCGR3A (CD16)*, and those originating from the virus, we implemented a two-step strategy utilizing the estimateNonExpressionCells function of the SoupX package (21). This function estimates whether each cell contains significantly more counts of a provided gene-set than would be expected under a Poisson model, given the estimated ambient RNA from its library of origin and the maximum contamination fraction. First, we used the viral genes to estimate the true maximum contamination fraction, based on the assumption that cells from the non-infected state should only contain viral reads from ambient mRNA captured in their droplets. To do so, we modified the estimateNonExpressionCells function to return p-values, and performed the test on each of our 13 libraries with a range of maximum contamination values from 1-50% (step of 1%) using the viral genes. We then computed FDR adjusted p-values for each maximum contamination value on the 88,559 high-quality, single monocytes. The number of non-simulated cells deemed to significantly express IAV transcripts (FDR<0.01) was used as a proxy for false positives. In examining the relationship between this number and the number of IAV-challenged cells found to significantly express viral transcripts at FDR<0.01 (**Figure 1C**), we found that a maximum contamination fraction of 10% resulted in a 1% false positive rate (defined as the percentage of non-infected cells from T_2-T_8 that were deemed to significantly express IAV transcripts). This parameter value was then used to correct for contamination from ambient for all genes considered (*CD14*, *FCGR3A* and IAV transcripts).

Assigning Cell States and Investigating Sources of Variability in IAV Levels

We used a maximum contamination fraction of 10% to test for significant expression of IAV transcripts in each cell (**Figures 1C–E**). IAV-challenged cells that contained a significant amount of IAV transcripts were considered as infected, while the others were deemed bystanders. To distinguish low from high IAV-transcribing cells, a Gaussian mixture model was fitted to the total percentage of viral mRNAs per cell across all infected cells, using the normalmixEM function from mixtools R package with k=2 (57). Each cell was assigned to the cluster with the highest posterior probability, and the cluster of cells with higher IAV content was annotated as high IAV-transcribing.

Characterizing Monocyte Subsets and Transcriptional Profiles From scRNA-Seq Data

Principal components analysis of 6,601 cells at T_0 was used to order monocytes along a differentiation axis separating $CD14^+$ cells from $CD16^+$ cells. We then computed the average percentage of classical and nonclassical monocytes obtained by flow cytometry across the eight donors, weighting each individual by the number of high-quality cells in the scRNA-seq data at T_0. Based on these percentages (87.1% for classical and 7.6% for nonclassical), we annotated the monocytes on each side of the differentiation axis as classical and nonclassical, respectively, with the remaining 5.3% of monocytes being annotated as intermediates. Validity of our approach was

confirmed by correlating the proportion of monocytes assigned to each subset across the eight donors, with the percentage of classical, intermediate and nonclassical monocytes estimated by flow cytometry.

Differential expression between subsets was assessed for the 4,859 genes expressed at a normalized $\log_2$ count > 0.1 in any of the 3 subsets. Specifically, Wilcoxon rank tests were implemented in the scran package (54), using the findMarkers function and blocking on donor. We considered genes to be differentially expressed (DE) between monocyte subsets when gene expression was significant at an FDR≤1% and $\log_2$FC>0.2. The 848 genes that differed between classical (CL) and nonclassical (NC) monocyte subsets were classified according to their behavior in intermediate monocytes (INT). They were either deemed 'similar to classical' (DE between INT and NC, but not between INT and CL), 'similar to nonclassical' (DE between INT and CL, but not between INT and NC), or 'intermediate' (all other cases).

At later time points, comparisons between $CD16^+$ and $CD16^-$ monocytes subsets were done based on 5,681 genes expressed with a normalized $\log_2$ count > 0.1 on average in either subset, in at least one condition and time point. For each subset, $\log_2$ fold change in gene expression relative to T_0 were correlated across times points. Differential expression between $CD16^+$ and $CD16^-$ cells was assessed with findMarkers (54), based on Wilcoxon rank tests and blocking on donors, time points and condition. Again, an FDR≤1% and $\log_2$FC>0.2 were required to define differentially expressed genes. To assess how $CD16^{+/-}$ status alters the infection of monocytes by IAV, we used logistic regression to model bystander/infected status as a function of $CD16^{+/-}$ status, while adjusting on donor, and time point (as factors).

Characterizing Subset-Specific Responses to IAV Challenge

To allow comparison between responses of $CD16^+$ and $CD16^-$ monocytes, 100 cells were subsampled from each subset and cell state, and at each time point. When subsampling, we ensured balanced representation of all donors across each monocyte subset and cell state, by using sampling weights that were inversely proportional to each donor representation in the original dataset. After sampling, a total of 6,669 genes with normalized $\log_2$ counts >0.1 on average in at least one group (cell-state x subset x time point) was selected for further analyses. For each monocyte subset, differences in expression between cell states (unexposed, bystander, infected) as well as between high- and low-IAV transcribing infected cells were performed using the findMarkers function from the scran package (54) and blocked on time point. For each comparison, genes were considered to be differentially expressed between cell states when gene expression was significant at an FDR=1% (Wilcoxon rank tests) and the $\log_2$ fold change was > 0.2. In addition, for each comparison between cell states, we tested for differences in response between subsets using a linear model of the form:

$$Expr_i \sim State_i + subset_i + State_i \cdot subset_i \quad (1)$$

where $Expr_i$ is the expression of the gene being tested in cell i, $State_i$ is an indicator variable that distinguishes the two cell states being compared (e.g. unexposed and bystander), and $subset_i$ is an indicator variable that reflects the $CD16^{+/-}$ status of cell i. The 335 genes with significant interactions at a 1% FDR (for unexposed-bystander and unexposed-infected comparisons) were clustered using the hclust R function with method 'Ward.D2'. DynamicTreeCut algorithm (58) was used to identify eight major patterns of response to IAV.

Transcription Factor Enrichment Analyses

To estimate Transcription Factor (TF) activity and define TF-targets relationships, we ran the R SCENIC pipeline (29) on the expression matrix (pre-normalization) on a random subsample of 4800 cells (100 cells from each donor at each time point and each condition, pre-exclusion of dying and contaminant cells) with default parameters. For each gene, motif-enrichment was considered for either *cis*-regulatory regions located <10kb from the TSS (distal regulatory elements), or between 500 bp upstream and 100 bp downstream of the promoter (proximal regulatory elements). To do so, motif-enrichment scores for all human genes (hg38 build, refseq_r80), were retrieved from https://resources.aertslab.org/cistarget and used as input for the Rcistarget package (29).

Sets of high-confidence targets for the 113 TFs whose activity could be quantified by SCENIC were then extracted and used for enrichment analysis. For each gene module, TF enrichment was assessed using a Fisher's exact test with the 6,669 expressed genes as background (**Supplementary Data 3**). Resulting p-values were adjusted using a global Benjamini-Hochberg correction for all eight modules and 113 TFs.

For each TF, with its targets enriched among one of the eight modules, TF activity inferred by SCENIC was used to test for subset-specific activity using a linear model of the form:

$$TF_i \sim State_i + subset_i + State_i \cdot subset_i \quad (2)$$

where TF_i is the activity of the TF being tested in cell i, $State_i$ is an indicator variable that distinguishes the two cell states being compared (e.g. unexposed and bystander), and $subset_i$ is an indicator variable that reflects the $CD16^{+/-}$ status of cell i. Average TF activity was then computed for each cell state, subset and time point, and correlated with gene expression of the associated module, to assess the link between TF activation and the TF-target enriched modules.

Association of the Outcome of IAV Infection With Basal Gene Expression

For each of the three monocyte subsets detected at basal state, a Kruskall-wallis test was used to search for genes whose expression levels significantly differ across donors. Within each monocyte subset, we then computed the average expression of each gene for all eight donors and correlated it with the percentage of infected cells at 4hpi. Genes that differed in expression between donors (FDR≤1%), and passed a nominal p-value threshold of 0.01 for association with IAV levels in any of the 3 subsets, were selected for downstream enrichment analyses.

For genes nominally correlated with viral mRNA levels, TF enrichment was assessed as previously using a Fisher's exact test with all 4,859 genes expressed at T_0 as background (**Supplementary Data 4**), and Benjamini-Hochberg correction for all 113 TFs was applied.

Gene Ontology Enrichment Analyses

All Gene Ontology (GO) enrichment analyses were performed with the GOSeq package using default settings (59). Background gene sets consisted of all genes that had average log-normalized expression values > 0.1 in at least one of the groupings being examined, and are described in the text. Only enrichments significant at FDR≤5% are reported.

Pseudo-Bulk Estimates From scRNA-Seq Data

Pseudo bulk estimates of IAV mRNA levels were computed by measuring, for each donor and time point, the mean percentage of reads of viral origin across all cells from the sample. At each time point, we then used a Pearson's correlation test to compare pseudo-bulk estimates for the 8 donors with IAV mRNA levels obtained in bulk data at 6hpi.

Monocyte Subset Characterization of EVOIMMUNOPOP Samples *via* Flow Cytometry

For 174 of the 200 EVOIMMUNOPOP donors, proportions of classical, intermediate and nonclassical monocytes were determined based on a fraction of 10^5 $CD14^+$ positively-selected monocytes, stained according to the manufacturer's instructions, with fluorescent APC-conjugated anti-CD14 and PE-conjugated anti-CD16 antibodies (Cat#130-091-243 and Cat #130-091-245, respectively, Miltenyi Biotec). Samples were then analyzed on a MACSQuant Analyzer 10 benchtop flow cytometer (Miltenyi Biotec).

Quantification of Canonical Monocyte Subsets in EVOIMMUNOPOP Samples

FlowJo v10.6.1 software (60) was used with the gating strategy depicted in **Supplementary Figure 8** to quantify monocyte subsets for 174 EVOIMMUNOPOP donors. Population-level differences in proportion of canonical monocyte subsets were assessed using Wilcoxon Rank tests. Correlation of the ratio of $CD16^+$ to $CD16^-$ cells with IAV mRNA levels was assessed using a linear model of the form

$$IAV \sim ratio + Pop, \quad (3)$$

where '*IAV*' are IAV mRNA levels, '*ratio*' is the percentage of $CD16^+$ monocytes (nonclassical+intermediates) divided by the percentage of $CD16^-$ monocytes (classical), and '*Pop*' is and indicator variable separating AFB from EUB individuals.

Analysis of Bulk RNA-Seq Profiles From the EVOIMMUNOPOP Cohort

A combined human-IAV reference was generated by concatenation of the primary human genome assembly (GRCh38) with the 8 segments of the human influenza A virus (IAV) A/USSR/90/1977(H1N1) genome (accession numbers CY010372-CY010379). Comprehensive human gene annotation was obtained from GENCODE (release 27) and merged with the 12 known transcripts of A/USSR/90/1977 (H1N1). RNA-seq reads (FASTQs) for all 970 samples that passed quality control in our previous study (18) [accession EGAS00001001895] were mapped to the combined reference with the STAR aligner (v.2.5.0a) (61) and assessed for quality with QualiMap 'bamqc' and 'rnaseq' (62, 63). Expression of viral mRNAs was measured as the percentage of uniquely-mapped reads aligning to the IAV genome. Reassuringly, the mean percentage of RNA-seq reads among samples from the IAV-challenged condition was 5.86% versus <0.01% in the other four conditions. Comparison of the percentage of IAV reads between populations was done using a Wilcoxon rank test. StringTie (v.1.3.3) (64) was used to quantify expression levels in transcripts per million mapped reads (TPM) for each annotated transcript. Gene expression data were filtered to remove genes with little evidence of activation (mean zTPM score < -3) (65) in any of the 5 conditions, and their quality was checked by principal component analysis (PCA). As GC content, 5′/3′ bias, date of the experiment and library batch were previously determined to be the strongest confounding factors on transcript expression (18), we corrected the data for these factors. First, we adjusted the data for GC content and 5′/3′ bias using linear models. Then, we imputed missing values by k-nearest neighbor imputation and adjusted for experiment date and library batch by sequentially running ComBat (66) for each batch effect, with condition and population as covariates. After batch effect correction, only IAV-stimulated samples were kept for downstream analyses.

Cell States Deconvolution From Bulk RNA Sequencing

To assess the percentage of total transcripts that originate from each cell state across the 199 IAV-challenged samples, we pooled cells from T_6 into 3 groups, based on their assigned cell-state (bystander, infected: high and low IAV-transcribing) and to which we added a 4^{th} group containing all singlets that were either (i) assigned to cluster numbers 3, 8, 10, and 11 (dying cells) or (ii) discarded based on their high mitochondrial content or low read counts (dead cells). We then estimated pseudo-bulk profiles for each group by summing UMIs across all cells and computing the number of UMIs associated to each gene per million of sequenced UMIs. TPM profiles obtained from bulk data were then normalized to improve comparison with pseudo-bulk. Specifically, we first computed a global pseudo-bulk profile of the entire single cell dataset as the average of the pseudo bulk profiles from the 4 cell states (bystander, infected: high and low IAV-transcribing, or dying/dead), weighted by the percentage of UMIs they contribute to the overall pool of cells. To account for the difference in how gene expression is quantified between the two methods (3' end counts for scRNA-seq and full-length gene coverage for bulk RNA-seq), we computed for each gene i a normalization factor s_i given by

$$s_i = \log(\overline{TPM_i}) - \log(PB_i) \quad (4)$$

where PB_i is the number of UMI per million for gene i in the global pseudo-bulk profile, and $\overline{TPM_i}$ is the average expression of the gene i in the 199 IAV-stimulated samples from the bulk RNA-seq data. For each gene, s_i was then subtracted from the log transformed TPM to yield normalized TPM profiles. We next applied DeconRNAseq (67) to the normalized log TPM profiles from all individuals, using the log-transformed pseudo bulk profiles from the 4 cell states as a basis for deconvolution. Quality of the deconvolution was assessed using leave-one-out cross validation, based on the eight individuals for whom we had scRNA-seq data. Specifically, for each of these eight individuals, bulk mRNAs were decomposed using pseudo-bulk profiles recomputed based on the seven other individuals. The resulting proportions were then compared with the percentage of UMIs that originate in each cell-state in the scRNA-seq to assess the quality of the deconvolution. Excluding IAV genes from bulk transcriptomic profiles prior to preforming the deconvolution had virtually no impact on the estimated proportions (Pearson r > 0.98 with proportions estimated without excluding IAV genes), confirming that our estimates were not driven by IAV expression alone. Comparisons between populations were performed using Wilcoxon rank tests.

The effect of the percentage of infected cells and percentage of high IAV-transcribing cells among infected cells on the total IAV mRNA levels were assessed by modelling

$$IAV \sim INF + POP \quad (5)$$

And

$$IAV \sim HI + POP \quad (6)$$

where IAV are the IAV mRNA levels across the 199 bulk mRNA samples, INF and HI are respectively the percentage of infected cells and the percentage of high IAV-transcribing cells among infected cells that we estimated from the deconvolution, and POP is a factor variable reflecting the population (EUB or AFB). The fraction η of population differences attributable to difference in rate of infection was estimating by comparing model (5) with model (7) below

$$IAV \sim POP \quad (7)$$

and computing $\eta = 100 \times (1 - \frac{\beta_5}{\beta_7})$, where β_i is the effect of population on IAV levels in model (i). To assess how the contribution of the percentage of high IAV transcribing cells to total IAV mRNA levels differed between populations, we used a linear model of the form

$$IAV \sim HI + POP + HI : POP \quad (8)$$

and tested for significant effect of the interaction term $HI : POP$ on IAV mRNA levels.

Flow Cytometry Analysis of Monocyte Susceptibility to IAV Infection

Frozen PBMCs from 8 individuals included in the EVOIMMUNOPOP cohort were thawed and allowed to rest overnight at 37°C, 5% CO_2 in 25cm^2 flasks. Cells were then seeded at 2×10^6/ml in untreated 96-well plates in RPMI-1640 GlutaMAX supplemented with 10% FCS in the presence of A/PR/8/34 (H1N1) (Charles River Laboratories) at a MOI=1 or media alone for 6h at 37°C, 5% CO_2. At the end of the incubation, cells were washed in FACS buffer (1X PBS supplemented with 2% FCS and 1mM EDTA) and stained with the LIVE/DEAD fixable violet dead cell stain kit (Cat#L34955, Life Technologies) and human Fc block for 15 min at 4°C, protected from light. Cells were then washed and stained with a mix of 6 surface antibodies for 20 min at 4°C, protected from light (anti-human CD19 BV510 Cat#562947, CD3 APC Cat#561811, CD16 PerCp-Cy5.5 Cat#560717, CD69 PE-Cy7 Cat#557745 from BD Biosciences, anti-human CD14 Cat#301806 from Biolegend, anti-human CD56 APC-Vio770 Cat#130-114-739 from Miltenyi Biotech). After centrifugation at 300g for 5 min, cells were incubated with the Fixation/Permeabilization solution from the Cytofix/Cytoperm kit (BD Biosciences) for 15min at 4°C, followed by intracellular staining with FITC conjugated anti-NP (Cat#MA1-7322, ThermoFisher Scientific) in BD Perm/Wash buffer (1X) for 30min at 4°C. Cells were washed and acquired using a MACSQuant (Miltenyi Biotec), and data were analyzed with FlowJo v10 with the gating strategy depicted in **Supplementary Figure 6A**.

DATA AVAILABILITY STATEMENT

The bulk RNA-seq data used in this study are available at the European Genome Phenome archive under accession number [EGAS00001001895]. The single cell RNA-seq data generated during this study are available at the European Genome Phenome archive under accession number [EGAS00001005000]. Code generated as part of this study is available on Github (https://github.com/h-e-g/PopDiff_MonocyteIAV).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Board of Institut Pasteur (EVOIMMUNO POP-281297) and the relevant French authorities (CPP, CCITRS, and CNIL). All experimental methods were conducted in accordance with the Declaration of Helsinki principles. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MO'N, MR, and LQ-M conceived and designed the study. MO'N, HQ, JP, YA, AB, NZ, and MD conducted the experiments at the Human Evolutionary Genetics Unit. MO'N, MR, YA, and DM developed computational methods and performed bioinformatic analyses. JP, VL, MH, S-ZY, QZ, AC,

LA, J-LC, and NN provided resources, expertise and feedback. MR and LQ-M supervised the study. LQ-M secured funding. MO'N, MR and LQ-M wrote the manuscript with input from all authors. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Institut Pasteur, the Collège de France, the French Government's Investissement d'Avenir program, Laboratoires d'Excellence "Integrative Biology of Emerging Infectious Diseases" (ANR-10- LABX-62-IBEID) and "Milieu Intérieur" (ANR-10-LABX-69-01), the Fondation de France (n°00106080), and the Fondation pour la Recherche Médicale (Equipe FRM DEQ20180339214). MO'N was supported by a European Molecular Biology Organization long-term fellowship (ALTF 229-2017).

ACKNOWLEDGMENTS

We wish to thank Tzachi Hagai and John Marioni for input and feedback, and Sylvie van Der Werf and Vincent Enouf for providing resources and protocols relating to influenza viruses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2021.768189/full#supplementary-material

REFERENCES

1. Ryabkova VA, Churilov LP, Shoenfeld Y. Influenza Infection, SARS, MERS and COVID-19: Cytokine Storm - The Common Denominator and the Lessons to be Learned. *Clin Immunol (Orlando Fla)* (2021) 223:108652. doi: 10.1016/j.clim.2020.108652
2. Krammer F, Smith GJD, Fouchier RAM, Peiris M, Kedzierska K, Doherty PC, et al. Influenza. *Nat Rev Dis Primers* (2018) 4(1):3. doi: 10.1038/s41572-018-0002-y
3. Zhang Q, Bastard P, Bolze A, Jouanguy E, Zhang SY, Effort CHG, et al. Life-Threatening COVID-19: Defective Interferons Unleash Excessive Inflammation. *Med (N Y)* (2020) 1(1):14–20. doi: 10.1016/j.medj.2020.12.001
4. Stegelmeier AA, van Vloten JP, Mould RC, Klafuric EM, Minott JA, Wootton SK, et al. Myeloid Cells During Viral Infections and Inflammation. *Viruses* (2019) 11(2):168. doi: 10.3390/v11020168
5. Fajgenbaum DC, June CH. Cytokine Storm. *N Engl J Med* (2020) 383 (23):2255–73. doi: 10.1056/NEJMra2026131
6. Guo C, Li B, Ma H, Wang X, Cai P, Yu Q, et al. Single-Cell Analysis of Two Severe COVID-19 Patients Reveals a Monocyte-Associated and Tocilizumab-Responding Cytokine Storm. *Nat Commun* (2020) 11(1):3924. doi: 10.1038/s41467-020-17834-w
7. Alon R, Sportiello M, Kozlovski S, Kumar A, Reilly EC, Zarbock A, et al. Leukocyte Trafficking to the Lungs and Beyond: Lessons From Influenza for COVID-19. *Nat Rev Immunol* (2021) 21(1):49–64. doi: 10.1038/s41577-020-00470-2
8. Ziegler-Heitbrock L, Ancuta P, Crowe S, Dalod M, Grau V, Hart DN, et al. Nomenclature of Monocytes and Dendritic Cells in Blood. *Blood* (2010) 116 (16):e74–80. doi: 10.1182/blood-2010-02-258558
9. Geissmann F, Jung S, Littman DR. Blood Monocytes Consist of Two Principal Subsets With Distinct Migratory Properties. *Immunity* (2003) 19(1):71–82. doi: 10.1016/S1074-7613(03)00174-2
10. Auffray C, Fogg D, Garfa M, Elain G, Join-Lambert O, Kayal S, et al. Monitoring of Blood Vessels and Tissues by a Population of Monocytes With Patrolling Behavior. *Science* (2007) 317(5838):666–70. doi: 10.1126/science.1142883
11. Ziegler-Heitbrock L. The CD14+ CD16+ Blood Monocytes: Their Role in Infection and Inflammation. *J Leukoc Biol* (2007) 81(3):584–92. doi: 10.1189/jlb.0806510
12. Hoeve MA, Nash AA, Jackson D, Randall RE, Dransfield I. Influenza Virus A Infection of Human Monocyte and Macrophage Subpopulations Reveals Increased Susceptibility Associated With Cell Differentiation. *PloS One* (2012) 7(1):e29443. doi: 10.1371/journal.pone.0029443
13. Hou W, Gibbs JS, Lu X, Brooke CB, Roy D, Modlin RL, et al. Viral Infection Triggers Rapid Differentiation of Human Blood Monocytes Into Dendritic Cells. *Blood* (2012) 119(13):3128–31. doi: 10.1182/blood-2011-09-379479
14. Piasecka B, Duffy D, Urrutia A, Quach H, Patin E, Posseme C, et al. Distinctive Roles of Age, Sex, and Genetics in Shaping Transcriptional Variation of Human Immune Responses to Microbial Challenges. *Proc Natl Acad Sci USA* (2018) 115(3):E488–97. doi: 10.1073/pnas.1714765115
15. Brodin P, Jojic V, Gao T, Bhattacharya S, Angel CJ, Furman D, et al. Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* (2015) 160(1-2):37–47. doi: 10.1016/j.cell.2014.12.020
16. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* (2016) 167(5):1415–29.e19. doi: 10.1016/j.cell.2016.10.042
17. Patin E, Hasan M, Bergstedt J, Rouilly V, Libri V, Urrutia A, et al. Natural Variation in the Parameters of Innate Immune Cells Is Preferentially Driven by Genetic Factors. *Nat Immunol* (2018) 19(3):302–14. doi: 10.1038/s41590-018-0049-7
18. Quach H, Rotival M, Pothlichet J, Loh YE, Dannemann M, Zidane N, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* (2016) 167(3):643–56.e17. doi: 10.1016/j.cell.2016.09.024
19. Ouwens KG, Jansen R, Nivard MG, van Dongen J, Frieser MJ, Hottenga JJ, et al. A Characterization of Cis- and Trans-Heritability of RNA-Seq-Based Gene Expression. *Eur J Hum Genet* (2020) 28(2):253–63. doi: 10.1038/s41431-019-0511-5
20. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed Droplet Single-Cell RNA-Sequencing Using Natural Genetic Variation. *Nat Biotechnol* (2018) 36(1):89–94. doi: 10.1038/nbt.4042
21. Young MD, Behjati S. SoupX Removes Ambient RNA Contamination From Droplet-Based Single-Cell RNA Sequencing Data. *Gigascience* (2020) 9(12): giaa151. doi: 10.1093/gigascience/giaa151
22. Wong KL, Tai JJ, Wong WC, Han H, Sem X, Yeap WH, et al. Gene Expression Profiling Reveals the Defining Features of the Classical, Intermediate, and Nonclassical Human Monocyte Subsets. *Blood* (2011) 118(5):e16–31. doi: 10.1182/blood-2010-12-326355
23. Segura V, Valero ML, Cantero L, Muñoz J, Zarzuela E, García F, et al. In-Depth Proteomic Characterization of Classical and Non-Classical Monocyte Subsets. *Proteomes* (2018) 6(1):8. doi: 10.3390/proteomes6010008
24. Schmidl C, Renner K, Peter K, Eder R, Lassmann T, Balwierz PJ, et al. Transcription and Enhancer Profiling in Human Monocyte Subsets. *Blood* (2014) 123(17):e90–9. doi: 10.1182/blood-2013-02-484188
25. Bercovich-Kinori A, Tai J, Gelbart IA, Shitrit A, Ben-Moshe S, Drori Y, et al. A Systematic View on Influenza Induced Host Shutoff. *eLife* (2016) 5:e18311. doi: 10.7554/eLife.18311
26. Machkovech HM, Bloom JD, Subramaniam AR. Comprehensive Profiling of Translation Initiation in Influenza Virus Infected Cells. *PloS Pathog* (2019) 15 (1):e1007518. doi: 10.1371/journal.ppat.1007518
27. Li S. Regulation of Ribosomal Proteins on Viral Infection. *Cells* (2019) 8 (5):508. doi: 10.3390/cells8050508

28. Cros J, Cagnard N, Woollard K, Patey N, Zhang SY, Senechal B, et al. Human CD14dim Monocytes Patrol and Sense Nucleic Acids and Viruses *via* TLR7 and TLR8 Receptors. *Immunity* (2010) 33(3):375–86. doi: 10.1016/j.immuni.2010.08.012
29. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nat Methods* (2017) 14(11):1083–6. doi: 10.1038/nmeth.4463
30. Allen EK, Randolph AG, Bhangale T, Dogra P, Ohlson M, Oshansky CM, et al. SNP-Mediated Disruption of CTCF Binding at the IFITM3 Promoter Is Associated With Risk of Severe Influenza in Humans. *Nat Med* (2017) 23(8):975. doi: 10.1038/nm.4370
31. Zhang Q. Human Genetics of Life-Threatening Influenza Pneumonitis. *Hum Genet* (2020) 139(6-7):941–8. doi: 10.1007/s00439-019-02108-3
32. Ciancanelli MJ, Huang SX, Luthra P, Garner H, Itan Y, Volpi S, et al. Infectious Disease. Life-Threatening Influenza and Impaired Interferon Amplification in Human IRF7 Deficiency. *Science* (2015) 348(6233):448–53. doi: 10.1126/science.aaa1578
33. Panthu B, Terrier O, Carron C, Traversier A, Corbin A, Balvay L, et al. The NS1 Protein From Influenza Virus Stimulates Translation Initiation by Enhancing Ribosome Recruitment to mRNAs. *J Mol Biol* (2017) 429(21):3334–52. doi: 10.1016/j.jmb.2017.04.007
34. Wang C, Forst CV, Chou TW, Geber A, Wang M, Hamou W, et al. Cell-To-Cell Variation in Defective Virus Expression and Effects on Host Responses During Influenza Virus Infection. *mBio* (2020) 11(1):e02880–19. doi: 10.1128/mBio.02880-19
35. Steuerman Y, Cohen M, Peshes-Yaloz N, Valadarsky L, Cohn O, David E, et al. Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing. *Cell Syst* (2018) 6(6):679–91.e4. doi: 10.1016/j.cels.2018.05.008
36. Russell AB, Trapnell C, Bloom JD. Extreme Heterogeneity of Influenza Virus Infection in Single Cells. *eLife* (2018) 7:e32303. doi: 10.7554/eLife.32303
37. Sun J, Vera JC, Drnevich J, Lin YT, Ke R, Brooke CB. Single Cell Heterogeneity in Influenza A Virus Gene Expression Shapes the Innate Antiviral Response to Infection. *PloS Pathog* (2020) 16(7):e1008671. doi: 10.1371/journal.ppat.1008671
38. Ramos I, Smith G, Ruf-Zamojski F, Martínez-Romero C, Fribourg M, Carbajal EA, et al. Innate Immune Response to Influenza Virus at Single-Cell Resolution in Human Epithelial Cells Revealed Paracrine Induction of Interferon Lambda 1. *J Virol* (2019) 93(20):e00559–19. doi: 10.1128/JVI.00559-19
39. Russell AB, Elshina E, Kowalsky JR, Te Velthuis AJW, Bloom JD. Single-Cell Virus Sequencing of Influenza Infections That Trigger Innate Immunity. *J Virol* (2019) 93(14):e00500–19. doi: 10.1128/JVI.00500-19
40. Kudo E, Song E, Yockey LJ, Rakib T, Wong PW, Homer RJ, et al. Low Ambient Humidity Impairs Barrier Function and Innate Resistance Against Influenza Infection. *Proc Natl Acad Sci USA* (2019) 116(22):10905–10. doi: 10.1073/pnas.1902840116
41. Cao Y, Guo Z, Vangala P, Donnard E, Liu P, McDonel P, et al. Single-Cell Analysis of Upper Airway Cells Reveals Host-Viral Dynamics in Influenza Infected Adults. *bioRxiv* (2020) 2020.04.15.042978. doi: 10.1101/2020.04.15.042978
42. Appleby LJ, Nausch N, Midzi N, Mduluza T, Allen JE, Mutapi F. Sources of Heterogeneity in Human Monocyte Subsets. *Immunol Lett* (2013) 152(1):32–41. doi: 10.1016/j.imlet.2013.03.004
43. Randolph H, Mu Z, Fiege J, Thielen B, Grenier J, Cobb M, et al. Single-Cell RNA-Sequencing Reveals Pervasive But Highly Cell Type-Specific Genetic Ancestry Effects on the Response to Viral Infection. *bioRxiv* (2020), 2020.12.21.423830. doi: 10.1101/2020.12.21.423830
44. Cole SL, Dunning J, Kok WL, Benam KH, Benlahrech A, Repapi E, et al. M1-Like Monocytes Are a Major Immunological Determinant of Severity in Previously Healthy Adults With Life-Threatening Influenza. *JCI Insight* (2017) 2(7):e91868. doi: 10.1172/jci.insight.91868
45. Zhou YH, Fu B, Zheng X, Wang D, Zhao C, Qi Y, et al. Pathogenic T Cells and Inflammatory Monocytes Incite Inflammatory Storm in Severe COVID-19 Patients. *Natl Sci Rev* (2020) 7(6):998–1002. doi: 10.1093/nsr/nwaa041
46. Chandrasekhar R, Sloan C, Mitchel E, Ndi D, Alden N, Thomas A, et al. Social Determinants of Influenza Hospitalization in the United States. *Influenza Other Respir Viruses* (2017) 11(6):479–88. doi: 10.1111/irv.12483
47. Hadler JL, Yousey-Hindes K, Perez A, Anderson EJ, Bargsten M, Bohm SR, et al. Influenza-Related Hospitalizations and Poverty Levels - United States, 2010-2012. *MMWR Morb Mortal Wkly Rep* (2016) 65(5):101–5. doi: 10.15585/mmwr.mm6505a1
48. Shelton JF, Shastri AJ, Ye C, Weldon CH, Filshtein-Somnez T, Coker D, et al. Trans-Ethnic Analysis Reveals Genetic and non-Genetic Associations With COVID-19 Susceptibility and Severity. *medRxiv* (2020) 2020.09.04.20188318. doi: 10.1101/2020.09.04.20188318
49. Center for Disease Control and Prevention. *Risk for COVID-19 Infection, Hospitalization, and Death By Race/Ethnicity* (2021). Available at: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html (Accessed August 30, 2021).
50. 10xgenomics. *Chromium Single Cell 3′ Reagent Kits User Guide (V3 Chemistry)* (2020). Available at: https://support.10xgenomics.com/permalink/OAXXHQmeWa60IWK66uqSo (Accessed August 30, 2021).
51. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An Integrated Map of Genetic Variation From 1,092 Human Genomes. *Nature* (2012) 491(7422):56–65. doi: 10.1038/nature11632
52. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nat Commun* (2017) 8:14049. doi: 10.1038/ncomms14049
53. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: Robust Clustering of Single-Cell RNA-Seq Data by Genotype Without Reference Genotypes. *Nat Methods* (2020) 17(6):615–20. doi: 10.1038/s41592-020-0820-1
54. Lun AT, McCarthy DJ, Marioni JC. A Step-by-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data With Bioconductor. *F1000Res* (2016) 5:2122. doi: 10.12688/f1000research.9501.2
55. P Pons and M Latapy eds. *Computing Communities in Large Networks Using Random Walks. Computer and Information Sciences - ISCIS 2005* Vol. 2005. Berlin, Heidelberg: Springer Berlin Heidelberg (2005).
56. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage. *Nat Immunol* (2019) 20(2):163–72. doi: 10.1038/s41590-018-0276-y
57. Benaglia T, Chauveau D, Hunter DR, Young DS. Mixtools: An R Package for Analyzing Mixture Models. *J Stat Software* (2010) 1(6):2009. doi: 10.18637/jss.v032.i0
58. Langfelder P, Zhang B, Horvath S. Defining Clusters From a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R. *Bioinformatics* (2008) 24(5):719–20. doi: 10.1093/bioinformatics/btm563
59. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias. *Genome Biol* (2010) 11(2):R14. doi: 10.1186/gb-2010-11-2-r14
60. *FlowJo™ Software, Version 10.6.1*. Ashland, OR: Becton, Dickinson and Company (2019).
61. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* (2013) 29(1):15–21. doi: 10.1093/bioinformatics/bts635
62. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: Evaluating Next-Generation Sequencing Alignment Data. *Bioinformatics* (2012) 28(20):2678–9. doi: 10.1093/bioinformatics/bts503
63. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data. *Bioinformatics* (2016) 32(2):292–4. doi: 10.1093/bioinformatics/btv566
64. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie Enables Improved Reconstruction of a Transcriptome From RNA-Seq Reads. *Nat Biotechnol* (2015) 33(3):290–5. doi: 10.1038/nbt.3122
65. Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR. Finding the Active Genes in Deep RNA-Seq Gene Expression Studies. *BMC Genomics* (2013) 14:778. doi: 10.1186/1471-2164-14-778
66. Johnson WE, Li C, Rabinovic A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics (Oxford England)* (2007) 8(1):118–27. doi: 10.1093/biostatistics/kxj037
67. Gong T, Szustakowski JD. DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq

Data. *Bioinformatics* (2013) 29(8):1083–5. doi: 10.1093/bioinformatics/btt090

Conflict of Interest: JP was employed by DIACCURATE.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Dissecting human population variation in single-cell responses to SARS-CoV-2

Environmental and genetic drivers of immune variability. The year 2019 was marked by the outbreak of a novel coronavirus strain responsible of severe acute respiratory syndrome (SARS-CoV-2) in humans. The ensuing 'coronavirus disease 2019' (COVID-19) pandemic highlighted the importance of dissecting the variability of outcomes to viral infection (§ 3.2.2, page 62).

Advanced age and male sex quicky appeared as the two main predictors of severe COVID-19 risk (Takahashi et al., 2020; O'Driscoll et al., 2021), which was also linked to changes in immunological parameters (Lee et al., 2020; Bastard et al., 2020; Hadjadj et al., 2020), as well as genetic variation (Shelton et al., 2021; Kousathanas et al., 2022). Together, these factors may explain why COVID-19 risks are not uniformly distributed across individuals and populations with different genetic and environmental backgrounds.

From an evolutionary point of view, there is increasing evidence to suggest that viral pathogens have exerted one of the strongest selective pressures on our genomes (Enard and Petrov, 2018). For example, strong genetic adaptation starting around 25 thousand years ago—in coincidence with the appereance in East Asia of the parental clade of coronaviruses—appears to have targeted multiple genes encoding proteins that interact with coronaviruses, such as *ACE2*, in East Asian populations only (Souilmi et al., 2021) (§ 2.3, page 46).

Furthermore, segments introgressed from Neandertal into Europeans and overlapping immune relevant genes, such as *OAS1* and *CCR9*, have been associated to both an increased and a decreased risk of severe COVID-19 (Zeberg and Pääbo, 2020, 2021). Yet, prior to this study we did not know how these events affected our immune responses to SARS-CoV-2 (§ 2, page 36) , nor how these responses vary across human populations (§ 3.2, page 59) and immune cell types (§ 3.1, page 55).

To address these questions, we at the HEG Unit proceeded in four steps. First, we sought to define the boundaries of natural variability in the immune response to viral infection by performing scRNA-seq on peripheral blood mononuclear cells (PBMCs) from over 200 healthy donors across three human groups with different genetic and environmental backgrounds (i.e., $n = 80$ AFB and $n = 80$ EUB from the EIP cohort, plus $n = 62$ other donors of East Asian origin), stimulated with either SARS-CoV-2 or IAV. Next, we tested for associations between gene expression variation and genetic variability, through the mapping of expressed quantitative trait loci (eQTLs; § 1.2.4, page 15), to delineate the genetic bases of transcriptional response variability in a cell type-specific manner. Third, we tested these eQTLs for signals of natural selection and Neanderthal introgression to assess their evolutionary origins. Finally, we assessed the overlap between these genetic factors and the variants that have been associated to COVID-19 risk genome-wide (§ 1.2.1, page 8).

Overall, we provide an extensive assessment of genetic, nongenetic and evolutionary predictors of variability in immune responses to viral infection across healthy individuals and populations, including plausible causal links between genotype, gene expression endophenotypes and COVID-19 phenotypes at cell-type resolution. ■

Article

Dissecting human population variation in single-cell responses to SARS-CoV-2

https://doi.org/10.1038/s41586-023-06422-9

Received: 10 November 2022

Accepted: 11 July 2023

Published online: 09 August 2023

Open access

Check for updates

Yann Aquino[1,2,27], Aurélie Bisiaux[1,27], Zhi Li[1,27], Mary O'Neill[1,27], Javier Mendoza-Revilla[1], Sarah Hélène Merkling[3], Gaspard Kerner[1], Milena Hasan[4], Valentina Libri[4], Vincent Bondet[5], Nikaïa Smith[5], Camille de Cevins[6], Mickaël Ménager[6,7], Francesca Luca[8,9,10], Roger Pique-Regi[8,9], Giovanna Barba-Spaeth[11], Stefano Pietropaoli[11], Olivier Schwartz[12], Geert Leroux-Roels[13], Cheuk-Kwong Lee[14], Kathy Leung[15,16], Joseph T. Wu[15,16], Malik Peiris[17,18,19], Roberto Bruzzone[18,19], Laurent Abel[20,21,22], Jean-Laurent Casanova[20,21,22,23,24], Sophie A. Valkenburg[18,25], Darragh Duffy[5,19], Etienne Patin[1], Maxime Rotival[1,28] ✉ & Lluis Quintana-Murci[1,26,28] ✉

Humans display substantial interindividual clinical variability after SARS-CoV-2 infection[1–3], the genetic and immunological basis of which has begun to be deciphered[4]. However, the extent and drivers of population differences in immune responses to SARS-CoV-2 remain unclear. Here we report single-cell RNA-sequencing data for peripheral blood mononuclear cells—from 222 healthy donors of diverse ancestries—that were stimulated with SARS-CoV-2 or influenza A virus. We show that SARS-CoV-2 induces weaker, but more heterogeneous, interferon-stimulated gene activity compared with influenza A virus, and a unique pro-inflammatory signature in myeloid cells. Transcriptional responses to viruses display marked population differences, primarily driven by changes in cell abundance including increased lymphoid differentiation associated with latent cytomegalovirus infection. Expression quantitative trait loci and mediation analyses reveal a broad effect of cell composition on population disparities in immune responses, with genetic variants exerting a strong effect on specific loci. Furthermore, we show that natural selection has increased population differences in immune responses, particularly for variants associated with SARS-CoV-2 response in East Asians, and document the cellular and molecular mechanisms by which Neanderthal introgression has altered immune functions, such as the response of myeloid cells to viruses. Finally, colocalization and transcriptome-wide association analyses reveal an overlap between the genetic basis of immune responses to SARS-CoV-2 and COVID-19 severity, providing insights into the factors contributing to current disparities in COVID-19 risk.

A notable feature of the COVID-19 pandemic is the substantial clinical variation among individuals infected with SARS-CoV-2, ranging from asymptomatic infection to fatal disease[1–3]. Risk factors include advanced age[1] as well as male sex[5], comorbidities[6] and host genetics[4,7,8]. Furthermore, variation in innate immunity[9–11]—including inborn errors or neutralizing auto-antibodies against type I interferons[12–14]—contribute to variation in clinical outcome, and epidemiological and genetic data suggest differences between populations[6,7,15,16]. This, together with reports of ancestry-related differences in transcriptional responses to immune challenges[17–19], calls for investigations of the magnitude and drivers of variation in immune responses to SARS-CoV-2 across populations worldwide.

Pathogen-imposed selection pressures have been paramount during human evolution[20]. Human adaptation to RNA viruses, through selective sweeps or archaic admixture, has been identified as a source of population genetic differentiation[18,21,22] and adaptation signals have been reported at coronavirus-interacting proteins in East Asians[23,24]. There is also evidence for links between archaic introgression and immunity[25], with Neanderthal haplotypes associated with COVID-19 severity[26,27]. However, the effects of natural selection and archaic admixture on immune responses to SARS-CoV-2 remain to be investigated.

We addressed these questions by exposing peripheral blood mononuclear cells (PBMCs) from individuals of Central African, West European and East Asian descent to SARS-CoV-2 and, for comparison, to influenza A virus (IAV). By combining single-cell RNA-sequencing (scRNA-seq) with quantitative and population genetics approaches, we delineate environmental and genetic drivers of population differences in immune responses to SARS-CoV-2.

Single-cell responses to RNA viruses

We characterized transcriptional responses to SARS-CoV-2 and IAV by performing scRNA-seq analysis of PBMCs from 222 SARS-CoV-2-naive donors originating from three geographical locations (Central Africa,

A list of affiliations appears at the end of the paper.

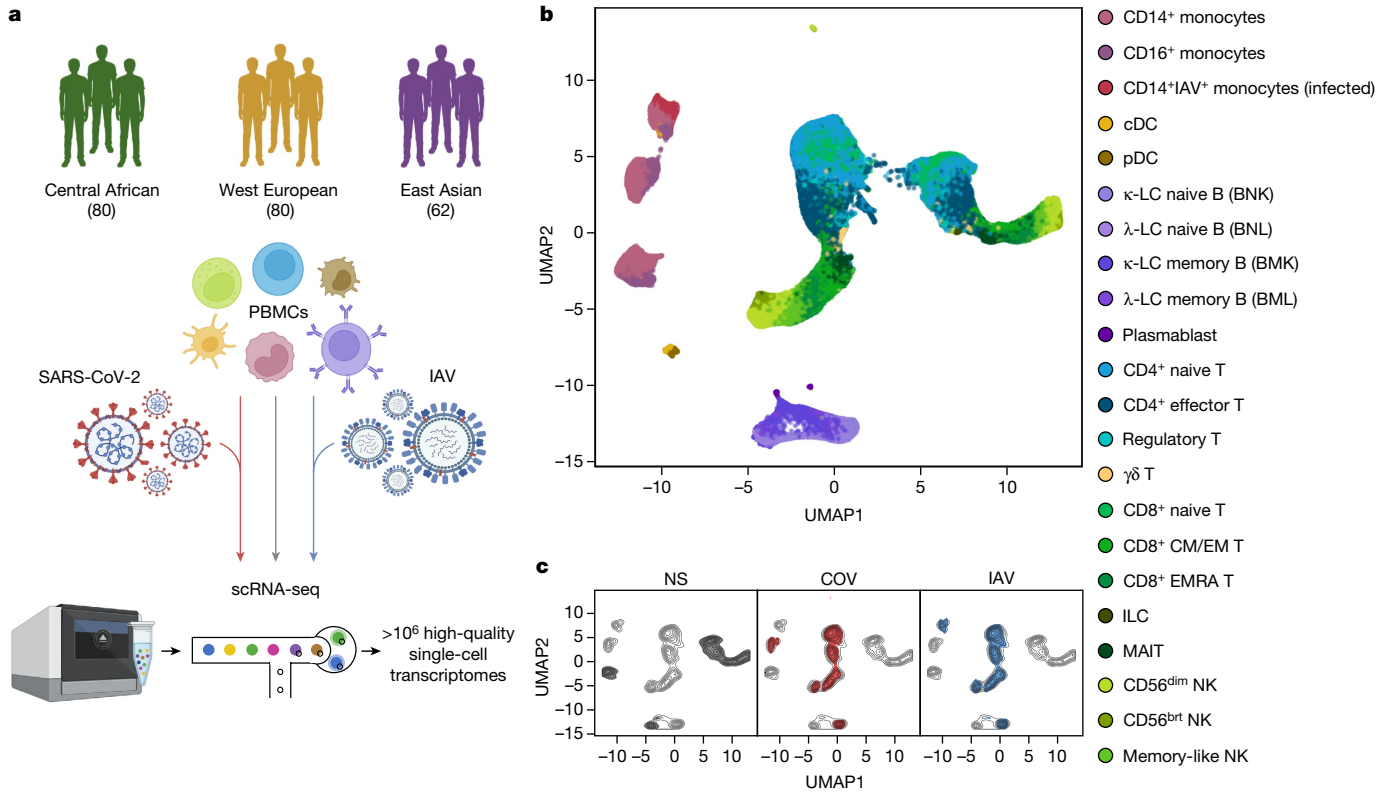


Fig. 1 | Population single-cell responses to SARS-CoV-2 and IAV. a, The study design. The diagram was created using BioRender. **b**,**c**, Uniform manifold approximation and projection (UMAP) embedding of 1,047,824 PBMCs: resting (non-stimulated; NS) or stimulated with SARS-CoV-2 (COV) or IAV for 6 h. **b**, The colours indicate the 22 cell types inferred. **c**, The distribution of cells in the NS, COV and IAV conditions on UMAP coordinates. The contour plot indicates the overall density of cells, and the coloured areas delineate regions of high cell density in each condition (NS (grey), COV (red) and IAV (blue)).

$n = 80$ male; West Europe, $n = 80$ male; East Asia, $n = 36$ female and 26 male) and with different genetic ancestries (Supplementary Fig. 1 and Supplementary Table 1). PBMCs were treated for 6 h (Supplementary Note 1, Supplementary Fig. 2 and Supplementary Table 2) with a mock-control (non-stimulated), SARS-CoV-2 (ancestral strain, BetaCoV/France/GE1973/2020) or IAV (H1N1/PR/8/1934). We captured over 1 million high-quality single-cell transcriptomes (Fig. 1a, Supplementary Fig. 3 and Supplementary Table 3a). By combining transcriptome-based clusters with cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq; Methods), we defined 22 cell types across myeloid, B, CD4$^+$ T, CD8$^+$ T and natural killer (NK) immune lineages (Fig. 1b, Supplementary Fig. 4 and Supplementary Table 3b–d). After virus exposure, most cell types showed moderate changes in abundance, with the strongest changes observed in the myeloid lineage after IAV treatment (Supplementary Note 2 and Supplementary Table 3e).

After adjusting for technical factors (Methods and Supplementary Fig. 5), we found that lineage identity was the main driver of gene expression variation (around 32%), followed by virus exposure (around 27%) (Fig. 1b,c). Both viruses induced a strong transcriptional response, with 2,914 genes upregulated (false-discovery rate (FDR) < 0.01, $\log_2$[FC] > 0.5; out of 12,655 with detectable expression; Supplementary Table 3f). These responses were highly correlated across lineages and featured a strong induction of interferon-stimulated genes (ISGs) (Extended Data Fig. 1a). However, myeloid responses were markedly heterogeneous, with SARS-CoV-2 inducing a transcriptional network enriched in inflammatory-response genes (Gene Ontology (GO): 0006954; fold-enrichment (FE) = 3.4, FDR < 4.9×10^{-8}; Supplementary Table 3g). For example, *IL1A*, *IL1B* and *CXCL8* were highly and specifically upregulated in response to SARS-CoV-2 ($\log_2$[FC] > 2.8, FDR < 2.3×10^{-36}), consistent with in vitro and in vivo studies[28,29].

To assess interindividual variability in the response to viruses, we summarized each individual's response as a function of their mean ISG expression (Supplementary Table 3h). SARS-CoV-2 induced more variable ISG activity than IAV across lineages[30], with myeloid cells displaying the strongest differences (Levene test, $P < 6.2 \times 10^{-6}$; Extended Data Fig. 1b). We determined the contributions of the various interferons (IFNs) to variation of ISG activity using single-molecule arrays (SIMOA) to quantify the levels of secreted IFNα, IFNβ and IFNγ. In the SARS-CoV-2 condition, IFNα accounted for up to 57% of ISG variability (Extended Data Fig. 2a,b), consistent with its determinant role in COVID-19 pathogenesis[13]. *IFNA1-21* transcripts were mostly produced by infected CD14$^+$ monocytes and plasmacytoid dendritic cells (pDCs) after IAV stimulation, whereas pDCs were the only important source of *IFNA1-21* after SARS-CoV-2 stimulation (that is, producing 88% of transcripts; Extended Data Fig. 2c). *IFNA1-21* expression by pDCs was weaker after stimulation with SARS-CoV-2 ($\log_2$[FC] = 6.4 versus 12.5 for IAV, Wilcoxon's rank-sum test, $P = 1.2 \times 10^{-16}$). Nevertheless, patterns of interindividual variability for ISG activity were notably similar after virus treatment ($r = 0.60$, Pearson's $P < 1.2 \times 10^{-22}$; Extended Data Fig. 2d), indicating that the IFN-driven response is largely shared between SARS-CoV-2 and IAV.

Cellular heterogeneity across populations

We assessed how immune responses differ across populations by comparing male individuals of African and European ancestries, who were sampled in a single recruitment effort thereby mitigating potential batch effects (Methods). As East Asian donors were recruited independently and present distinct demographic characteristics (Supplementary Table 1), they were excluded from cross-population comparisons. Focusing on cellular proportions, we detected marked population

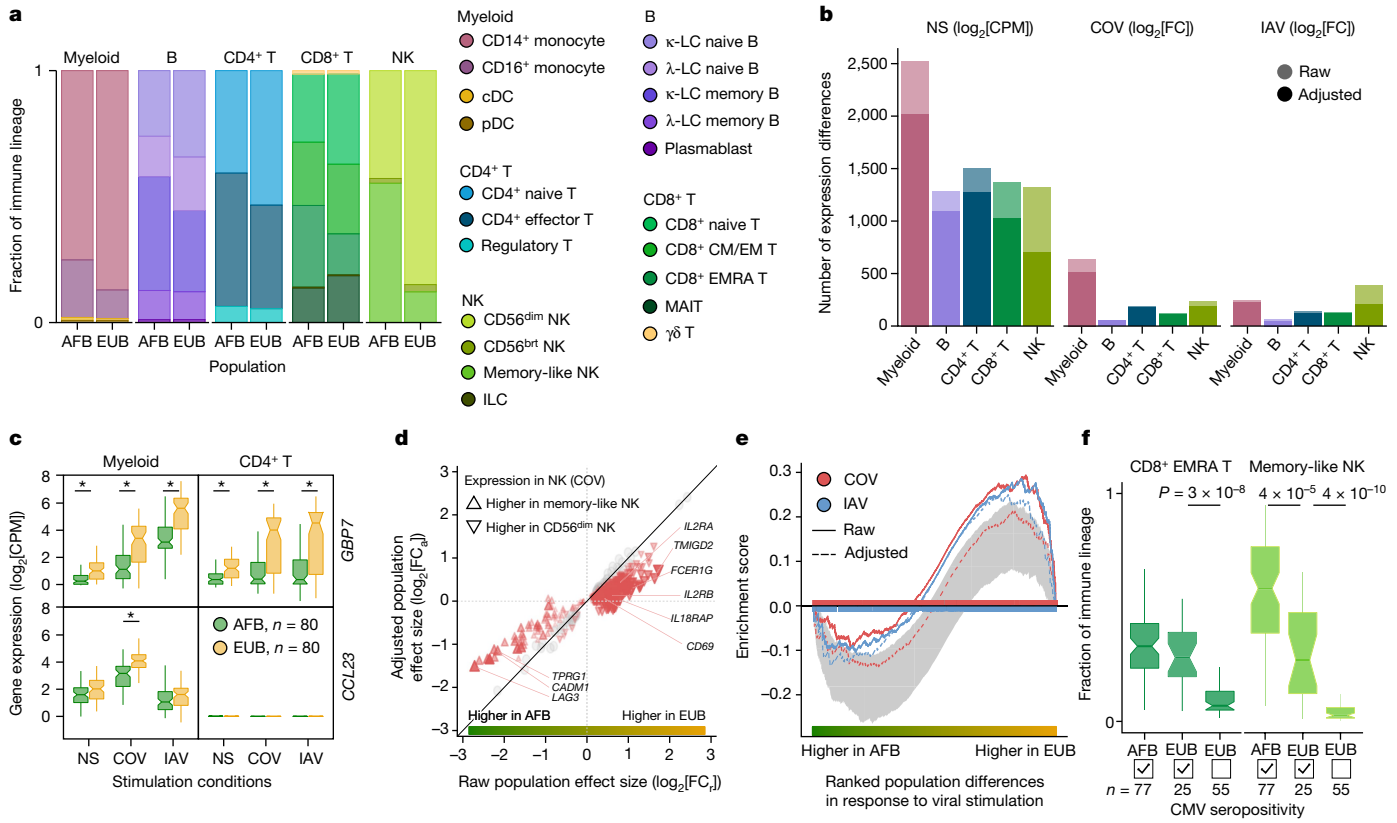


Fig. 2 | Cellular composition affects the transcriptional responses to viral stimuli. a, Cell type proportions within each immune lineage in Africans (AFB) and Europeans (EUB). brt, bright. **b**, The number of genes differentially expressed between the African and European groups in the basal state (NS) or in response to SARS-CoV-2 (COV) or IAV, in each immune lineage. Numbers are provided before and after adjustment for cellular composition. **c**, Examples of popDRGs, either shared across cell types and viruses (*GBP7*) or specific to SARS-CoV-2-stimulated myeloid cells (*CCL23*). Statistical analysis was performed using two-sided Student's *t*-tests with adjustment using the Benjamini–Hochberg method; $^*P < 0.001$. Exact *P* values are provided in Supplementary Table 4b. **d**, The effect of adjusting for cellular composition on genes differentially expressed between populations after exposure to SARS-CoV-2. Adjustment reduces raw population fold-changes (FC_a versus FC_r) in the expression of genes that are differentially expressed between memory-like NK cells and $CD56^{dim}$ NK cells (red triangles; genes with similar expression are shown in grey). **e**, The effect of adjusting for cellular composition on population differences in the response to viral stimulation for genes involved in the positive regulation of cell migration (GO:0030335) in the NK lineage. For each stimulus, gene set enrichment analysis enrichment curves are shown before and after adjusting on the basis of cellular composition. Grey shades indicate the 95% expected range for the enrichment curve when gene labels are permuted at random. **f**, The distribution of $CD8^+$ EMRA T and memory-like NK cell frequencies in Africans and Europeans according to $CMV^{+/-}$ serostatus. *P* values ($P < 0.01$) calculated using Wilcoxon's two-sided rank-sum tests are shown. For **c** and **f**, the centre line shows the median; the notches show the 95% confidence intervals (CIs) of the median; the box limits show the upper and lower quartiles; and the whiskers show 1.5× interquartile range. The number (*n*) of independent biological samples is indicated where relevant.

differences in lineage composition, particularly for NK cells (Fig. 2a and Supplementary Table 4a). A subset identified as memory-like NK cells[31] constituted 55.2% of the NK compartment in African-descent individuals, but only 12.2% in Europeans (Wilcoxon's rank-sum test, $P < 1.3 \times 10^{-18}$; Extended Data Fig. 3a,b and Supplementary Fig. 6). African donors also presented higher proportions of $CD16^+$ monocytes[32] and memory lymphocyte subsets, such as memory B cells, effector $CD4^+$ T cells and effector memory re-expressing CD45RA (EMRA) $CD8^+$ T cells (Wilcoxon's rank-sum test, $P < 4.7 \times 10^{-3}$).

Across lineages, we found 3,389 genes displaying population differences in expression in the basal state (popDEGs; FDR < 0.01, $|\log_2[\text{fold change (FC)}]| > 0.2$) and 898 and 652 displaying differential responses between populations (popDRGs; FDR < 0.01, $|\log_2[\text{FC}]| > 0.2$) after stimulation with SARS-CoV-2 and IAV, respectively (Fig. 2b and Supplementary Table 4b,c). popDRGs included key immunity regulators, such as the IFN-responsive *GBP7* and the gene coding for the macrophage inflammatory protein MIP-3, *CCL23*, both of which were more strongly upregulated in Europeans (Fig. 2c). The *GBP7* response was common to both viruses and all lineages ($\log_2[\text{FC}] > 0.88$, Student's *t*-test, adjusted *P* (P_{adj}) $< 1.4 \times 10^{-3}$), but that of *CCL23* was specific to SARS-CoV-2-stimulated myeloid cells ($\log_2[\text{FC}] = 0.72$, Student's *t*-test, $P_{adj} = 5.3 \times 10^{-4}$). We estimated that population differences in cellular composition accounted for 15–47% of popDEGs and for 7–46% of popDRGs, with the strongest impact on NK cells (Fig. 2b,d and Extended Data Fig. 3c). Variation in cellular composition mediated pathway-level differences in response to viral stimulation between populations (Supplementary Table 4d). For example, in virus-stimulated NK cells, genes involved in the promotion of cell migration, such as *CSF1* or *CXCL10*, were more strongly induced in Europeans (normalized enrichment score > 1.5, gene set enrichment analysis, $P_{adj} < 0.009$). However, the loss of this signal after adjustment for cellular composition (Fig. 2e) indicates that fine-scale cellular heterogeneity drives population differences in immune responses to SARS-CoV-2.

Repercussions of CMV infection

We next investigated the sources of population differences in cellular composition. We found no strong genetic effects on cellular proportions (Supplementary Note 3 and Supplementary Table 4e), suggesting a predominantly environmental origin to such population differences. As latent cytomegalovirus (CMV) infection alters cellular proportions[33–35] and its prevalence varies across populations[36], we determined

the CMV$^{+/-}$ serostatus of the samples. All but one of the African-descent individuals were CMV^{+} (99%), versus 31% of Europeans, and CMV^{+} was associated with higher proportions of memory-like NK and CD8^{+} EMRA T cells in Europeans (Fig. 2f and Extended Data Fig. 3d). Using mediation analysis, we estimated that CMV serostatus accounts for up to 73% of the differences in the proportion of these cell types between Africans and Europeans; these differences substantially impact the transcriptional response to SARS-CoV-2 (Supplementary Table 4f,g, Supplementary Notes 4 and 5 and Supplementary Fig. 7). However, other than its effects on cellular composition, CMV^{+} had a limited direct effect on SARS-CoV-2 responses, with only one gene presenting significant expression differences in response to this virus (*ERICH3* in CD8^{+} T cells, $\log_2$FC = 1.7, FDR = 0.007; Supplementary Table 4h). These findings highlight how differing environmental exposures, such as CMV infection, may lead to population differences in the responses to SARS-CoV-2 through changes in the lymphoid composition.

Genetic basis of the leukocyte response

To assess the effects of human genetic variants on transcriptional variation, we mapped expression quantitative trait loci (eQTLs) jointly in all three populations, focusing on *cis*-regulatory variants. At an FDR of 1%, we identified 1,866–4,323 independent eQTLs per lineage, affecting 5,198 genes (Fig. 3a and Supplementary Table 5a). Among the 9,150 eQTLs detected, 11% were ancestry specific (n = 973; Supplementary Note 6), underscoring the importance of including diverse ancestries in genomics research. Increasing the resolution to 22 cell types revealed an additional 3,603 eQTLs (Extended Data Fig. 4a,b and Supplementary Table 5b). We found that 79% of eQTLs were replicated ($P < 0.01$) in at least three cell types, but only 22% were common to all lineages. In total, 812 eQTLs were cell-type-specific, around 45% of which were detected in myeloid cells (Extended Data Fig. 4b), including a pDC-specific eQTL (rs114273142) at *MIR155HG*—hosting a micro RNA that promotes sensitivity to type I IFNs[37] (Extended Data Fig. 4c and Supplementary Note 7). Broadly, eQTL effect sizes were more correlated across ontogenetically related cell types (mean correlation within and between lineages of r = 0.60 and 0.47, Wilcoxon's rank-sum test, $P = 6.2 \times 10^{-6}$; Extended Data Fig. 4d).

Focusing on variants that altered responses to viral stimuli (reQTLs), we identified 1,505 reQTLs affecting 1,213 genes (Supplementary Table 5c,d). Supporting the replicability of the results, our IAV reQTLs are enriched in genes that are reported to contain IAV-specific eQTLs[19] (OR > 3.2, Fisher's exact test, $P < 9.4 \times 10^{-4}$), with more than 98% of replicated eQTLs affecting expression in the same direction (Supplementary Note 8, Supplementary Fig. 8 and Supplementary Table 5e). The correlation of reQTL effect sizes across ontogenetically related cell types was weaker than for eQTLs (r = 0.36 and 0.50, respectively, Wilcoxon's rank-sum test, $P < 5.6 \times 10^{-13}$; Extended Data Fig. 4d). Furthermore, the proportion of virus-dependent reQTLs differed across cell types. In lymphoid cells, only 7.7% of reQTLs differed in effect size between viruses (interaction $P < 0.01$; Fig. 3b,c), whereas 49% of myeloid reQTLs were virus dependent (interaction $P < 0.01$), with 46 and 185 reQTLs displaying specific, stronger effects after SARS-CoV-2 and IAV stimulation, respectively. The strongest SARS-CoV-2 reQTL (rs534191, Student's t-test, $P = 1.96 \times 10^{-16}$ (SARS-CoV-2) and $P = 0.05$ (IAV); Fig. 3d) was identified in myeloid cells at *MMP1*, encoding a biomarker of COVID-19 severity[38]. These analyses reveal that the effects of virus-induced reQTLs are cell-type dependent and highlight the virus specificity of the genetic basis of the myeloid response.

Ancestry effects on immune responses

To evaluate the contribution of genetic variation to population differences in immune responses, we focused on popDEGs and popDRGs. We found that 11–24% of the genes expressed in each lineage had at least one eQTL, but this proportion increased up to 56% and 60% for popDEGs and popDRGs that were not explained by cellular heterogeneity, respectively (Fisher's exact test, $P < 1.4 \times 10^{-6}$; Fig. 3e and Extended Data Fig. 5a). The popDEGs and popDRGs displaying the largest population differences were more likely to be under genetic control and associated with large-effect (r)eQTLs (Extended Data Fig. 5b–d). We used mediation analysis to assess, for each gene, immune lineage and virus, the fraction of population differences explained by genetics (that is, the most significant eQTL) or cellular heterogeneity (Supplementary Table 6 and Supplementary Note 9). Cellular composition had a broad effect on population differences in gene expression and viral responses (explaining 16–62% of differences per lineage and virus, with the strongest effect in NK cells), whereas genetics had a weaker effect (explaining 13–35% of population differences; Fig. 3f and Extended Data Fig. 5e). However, genetics had strong effects on a gene subset (141–433 genes per lineage) for which they accounted for 32–58% of population differences. For example, 81–100% of the differences in *GBP7* expression between Africans and Europeans were explained by a single variant displaying strong population differentiation (rs1142888, derived allele frequency (DAF) = 0.13 and 0.53 in Africans and Europeans, respectively, fixation index (F_{ST}) = 0.26, $|\beta_{eQTL}| > 1.7$ across lineages after stimulation). Thus, population variation in immune responses is driven largely by cellular heterogeneity, but genetic variants with marked allele frequency variation contribute to population differences at specific loci.

Natural selection and SARS-CoV-2 responses

To investigate the contribution of natural selection to population differences in immune responses, we first searched for overlaps between (r)eQTLs and genome-wide signals of local adaptation, measured by the population branch statistic (PBS)[39]. We identified 1,616 eQTLs (1,215 genes) and 180 reQTLs (166 genes) displaying strong population differentiation (empirical $P < 0.01$), 90 of which were ancestry specific (Supplementary Table 7a and Supplementary Note 6). Among genes harbouring putatively adaptive (r)eQTLs, we found key players in IFN-mediated antiviral immunity, such as *DHX58* and *TRIM14* in Africans, *ISG20*, *IFIT5*, *BST2* and *IFITM2-3* in Europeans, and *IFI44L* and *IFITM2* in East Asians.

We then used CLUES[40] to identify rapid changes in (r)eQTL frequency over the last 2,000 generations (that is, 56,000 years) in each population (Supplementary Fig. 9 and Supplementary Table 7b). We found signals of rapid adaptation (maximum $|Z| > 3$) targeting the same (*IFITM2*, *IFIT5*) or different (*ISG20*, *IFITM3*, *TRIM14*) eQTLs at highly differentiated genes, suggesting repeated adaptations targeting IFN-mediated antiviral immunity (Supplementary Note 10, Supplementary Table 7c and Supplementary Fig. 10). We determined whether selection had altered gene expression in specific cell types or in response to SARS-CoV-2 or IAV by testing for increased population differentiation (PBS) at (r)eQTLs within each cell type, relative to random single-nucleotide polymorphisms (SNPs) matched for allele frequency, linkage disequilibrium (LD) and distance to the nearest gene. In the basal state, eQTLs were more strongly differentiated in Europeans, the strongest signal observed for γδ T cells (Extended Data Fig. 6a). Among popDEGs for which genetics mediates more than 50% of the differences between Africans and Europeans, 34% presented signals of rapid adaptation in Europeans (versus 21% in Africans, Fisher's exact test, $P = 7.7 \times 10^{-6}$). For example, population differences at *GBP7* have been driven by a frequency increase, over the last 782–1,272 generations, of the rs1142888-G allele in Europeans (maximum $|Z| > 4.3$, Extended Data Fig. 6b).

Focusing on responses to viruses, SARS-CoV-2 reQTLs displayed increased population differentiation in East Asians (FE = 1.24, one-sided resampling, $P < 2 \times 10^{-4}$; Extended Data Fig. 6c) and were enriched in East-Asian-specific variants (OR > 4.2, Fisher's exact test, $P < 2.3 \times 10^{-6}$; Supplementary Note 6 and Supplementary Table 7d). Furthermore, among SARS-CoV-2-specific reQTLs, 28 reQTLs (5.3%) displayed

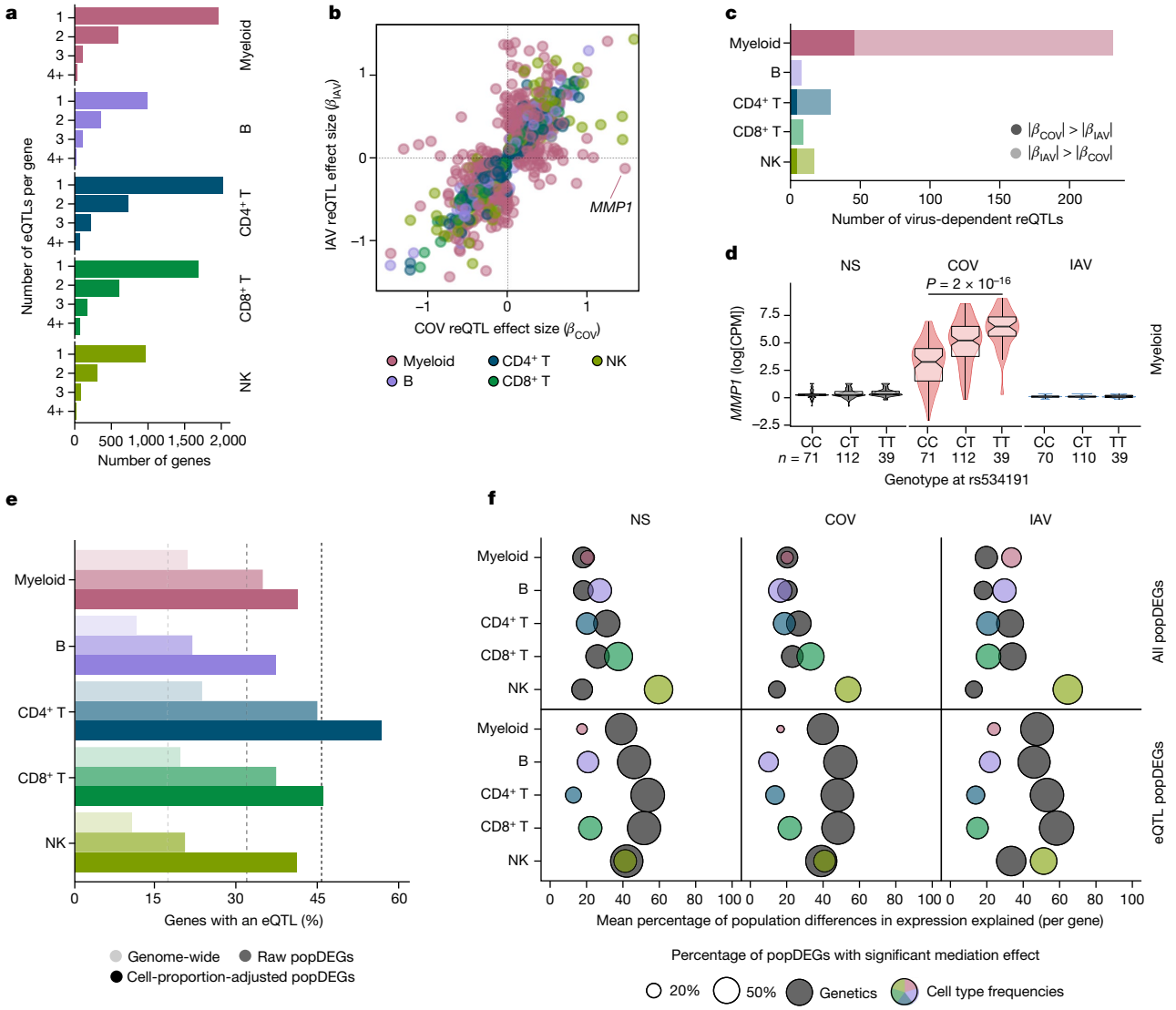


Fig. 3 | Genetic basis of immune responses to RNA viruses. a, The number of eQTLs detected per gene within each immune lineage. **b**, Comparison of reQTL effect sizes (β) between SARS-CoV-2- and IAV-stimulated cells. Each dot represents a specific reQTL (that is, SNP, gene and lineage) and its colour indicates the lineage in which it was detected. **c**, The number of virus-dependent reQTLs (two-sided Student's *t*-test nominal interaction, $P < 0.01$) in each immune lineage, coloured according to the lineage and the stimulus for which the reQTL has the largest effect size. **d**, Example of a SARS-CoV-2-specific reQTL at *MMP1*. *P* values ($P < 0.01$) calculated using Student's two-sided *t*-tests are shown. The centre line shows the median; the notches show the 95% CIs of the median; the box limits show the upper and lower quartiles; the whiskers show 1.5× interquartile range; and the points show outliers. **e**, Enrichment in eQTLs among genes that are differentially expressed between populations (popDEGs). For each immune lineage, the bars indicate the percentage of genes with a significant eQTL, at the genome-wide scale and among popDEGs, before or after adjustment for cellular composition. **f**, For each lineage and stimulus, the *x* axis indicates the mean contribution of either genetics (that is, the most significant eQTL per gene in each lineage and stimulus) or cellular composition to population differences in expression, across all popDEGs (top) or popDEGs associated with an eQTL (bottom). The size of the dots reflects the percentage of genes with a significant mediated effect at an FDR of 1% (Supplementary Table 6). The number (*n*) of independent biological samples is indicated where relevant.

signals of adaptation in East Asians starting 770–970 generations ago (around 25,000 years)—a timeframe associated with genetic adaptation at SARS-CoV-2-interacting proteins[23] (OR relative to other populations = 2.6, Fisher's exact test, $P = 7.3 \times 10^{-4}$; Fig. 4a and Extended Data Fig. 7a–c). An example is the immune mediator *LILRB1*, which has a SARS-CoV-2-specific reQTL (rs4806787) in pDCs (Extended Data Fig. 7d). However, the selection events making the largest contribution to the differentiation of SARS-CoV-2 responses in East Asia (top 5% PBS) began before this period (more than 970 generations ago, OR = 1.94, Fisher's exact test, $P = 0.019$; Fig. 4b). For example, the rs1028396-T allele (80% frequency in East Asia versus 16–25% elsewhere), associated with a weaker response of *SIRPA* to SARS-CoV-2 in CD14⁺ monocytes, presents a selection signal beginning more than 45,000 years ago (Fig. 4b and Extended Data Fig. 7e). SIRPα inhibits infection by endocytic viruses, including SARS-CoV-2[41]. These results suggest recurrent genetic adaptation targeting antiviral immunity over the last 50,000 years, contributing to present-day population differences in immune responses to SARS-CoV-2.

Neanderthal heritage on immune functions

We investigated the effects of Neanderthal introgression on immune responses to viruses by defining 100,345 'archaic' variants (aSNPs) and testing for biased eQTL representation among aSNPs relative to

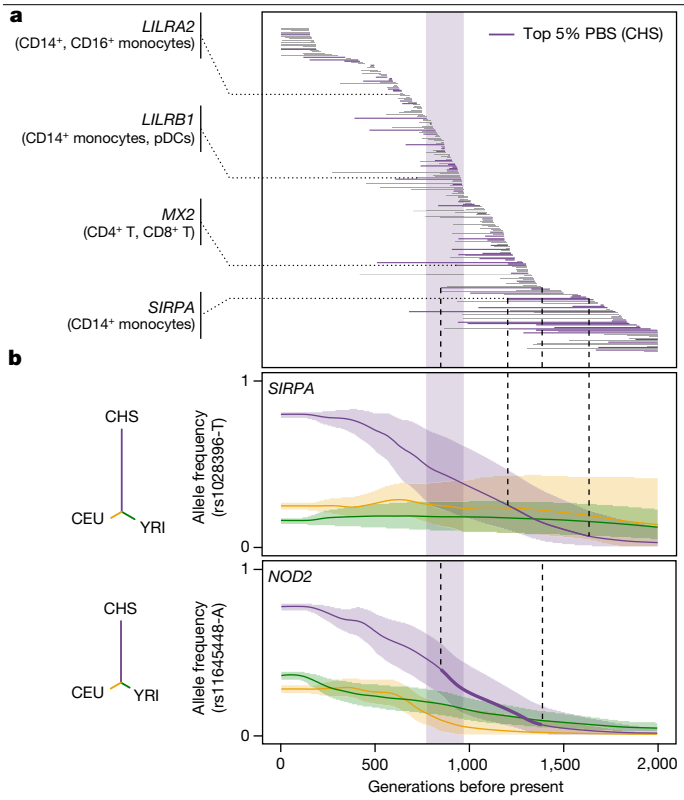


Fig. 4 | Natural selection effects on population differentiation of immune responses. a, Estimated periods of selection, over the past 2,000 generations, for 245 SARS-CoV-2 reQTLs with significant signals of rapid adaptation in East Asians (CHS) (maximum $|Z| > 3$). Each horizontal line represents a variant, sorted in descending order of time to onset of selection. The area shaded in purple highlights the period (770–970 generations ago) associated with genetic adaptation at host coronavirus-interacting proteins in East Asians[23]. Several immunity-related genes are highlighted. **b**, Allele frequency trajectories of two SARS-CoV-2 reQTLs (rs1028396 at *SIRPA* and rs11645448 at *NOD2*) in Africans (YRI, green), Europeans (CEU, yellow) and East Asians (CHS, purple). The full lines indicate the maximum a posteriori estimate of allele frequency at each epoch and shaded areas indicate the 95% CIs. The dendrograms show the estimated unrooted population phylogeny for each eQTL based on PBS (that is, the branch length between each pair of populations is proportional to $-\log_{10}[1-F_{ST}]$).

random, matched SNPs (Methods). We found that archaic haplotypes were 1.4–1.5 times more likely to alter gene expression in the basal state (one-sided permutation test, $P = 3 \times 10^{-4}$) and after stimulation with SARS-CoV-2 or IAV (one-sided permutation test, $P = 9 \times 10^{-4}$ and 3×10^{-3}, respectively) in Europeans, and this trend was only marginally significant in East Asians after viral stimulation (FE > 1.2, one-sided permutation test, $P < 2 \times 10^{-2}$; Extended Data Fig. 8a and Supplementary Table 8a–c). Enrichment was strongest in SARS-CoV-2-stimulated CD16$^+$ monocytes from Europeans, suggesting that archaic haplotypes altering myeloid responses have been preferentially retained in their genomes. Archaic haplotypes with eQTLs are generally present at higher frequencies compared with archaic haplotypes without eQTLs (Δf(introgressed allele) >3.2%, Student's *t*-test, $P_{adj} < 8 \times 10^{-3}$; Extended Data Fig. 8b and Supplementary Table 8d,e), even after adjustment for minor allele frequency (MAF) to ensure similar power for eQTL detection, supporting the adaptive nature of Neanderthal regulatory alleles.

To characterize the functional consequences of archaic introgression at the cell-type level, we focused on introgressed eQTLs for which the archaic allele was found at its highest frequency in Eurasians (that is, 5% most frequent). These included known adaptively introgressed variants at *OAS1-3* or *PNMA1* in Europeans and *TLR1*, *FANCA* or *IL10RA* in East Asians[18,42–46], for which we delineated the cellular and molecular effects (Extended Data Figs. 8c and 9a and Supplementary Table 8f). Yet, we identified previously unreported signals of Neanderthal introgression affecting immunity phenotypes. For example, an introgressed reQTL (rs58964929-A, 38% of Europeans versus 22% of East Asians) decreases *UBE2F* responses to SARS-CoV-2 and IAV in monocytes (Extended Data Fig. 9b). UBE2F is involved in neddylation, a post-translational modification that is required for the nuclear translocation of IRF7 by myeloid cells after RNA virus infection and, therefore, for the induction of type I IFN responses[47]. Likewise, an introgressed eQTL (rs11119346-T, 43% in East Asians versus less than 3% in Europeans) downregulates *TRAF3IP3*—a negative regulator of the cytosolic RNA-induced IFN response[48]—in IAV-infected monocytes, thereby favouring IFN release after viral infection (Extended Data Fig. 9c,d). We also identified a 35.5 kb Neanderthal haplotype reaching 61% frequency in East Asians (versus 24% in Europeans, tagged by rs9520848-C allele) that is associated with higher basal expression of the cytokine gene *TNFSF13B* by MAIT cells (Extended Data Fig. 9e,f). Collectively, these results reveal how archaic introgression has altered immune functions in present-day Eurasians at the molecular and cellular level.

Contribution of eQTLs to COVID-19 risk

We investigated the contributions of genetic variants altering responses to SARS-CoV-2 ex vivo to COVID-19 risk in vivo by determining whether (r)eQTLs were more strongly associated with COVID-19 GWAS hits[8] than random, matched SNPs (Methods). We observed an enrichment in eQTLs at loci associated with susceptibility (reported cases) and severity (hospitalized or critical cases) (FE = 4.1 and FE > 3.8, respectively, one-sided resampling, $P < 10^{-4}$), and a specific enrichment in reQTLs at severity loci (FE > 3.7, one-sided resampling, $P < 3 \times 10^{-3}$; Fig. 5a). This trend was observed across most cell lineages (Extended Data Fig. 10a). Colocalization analyses identified 40 genes at which there was a high probability of (r)eQTL colocalization with COVID-19 hits (posterior probability that both traits are linked to the same SNP (PP_{H4}) > 0.8) and transcriptome-wide association studies (TWASs) linked predicted gene expression with COVID-19 risk for 30 of these genes ($FDR_{TWAS} < 0.01$; Supplementary Table 9a). These included direct regulators of innate immunity, such as *IFNAR2* in non-stimulated CD4$^+$ T cells, *IRF1* in non-stimulated NK and CD8$^+$ T cells, *OAS1* in lymphoid cells stimulated with SARS-CoV-2 and IAV, and *OAS3* in SARS-CoV-2-exposed CD16$^+$ monocytes (Fig. 5b and Extended Data Fig. 10b,c). These results support a contribution of immunity-related (r)eQTLs to COVID-19 risk.

Focusing on the evolutionary factors affecting COVID-19 risk, we identified 20 eQTLs that (1) colocalized with COVID-19 hits ($PP_{H4} > 0.8$) and (2) presented positive selection signals (top 1% PBS, $n = 13$ eQTLs) or evidence of archaic introgression ($n = 7$ eQTLs), 14 of which regulate genes of which the expression is correlated with COVID-19 susceptibility and/or severity ($FDR_{TWAS} < 0.01$) (Fig. 6). For example, two variants in high LD at *DR1* (rs569414 and rs1559828, $r^2 > 0.73$) displayed extremely high levels of population differentiation, probably due to selection outside Africa (DAF = 0.13 in Africa versus higher than 0.62 in Eurasia; Extended Data Fig. 10d). DR1 suppresses type I IFN responses[49] and the selected alleles, which decrease COVID-19 severity, reduce *DR1* expression in most immune cells (Fig. 6). Likewise, an approximately 39 kb Neanderthal haplotype, spanning the *MUC20* locus in Eurasians, contains the rs2177336-T allele that increases *MUC20* expression in SARS-CoV-2-stimulated cells, particularly for CD4$^+$ T cells, and decreases COVID-19 susceptibility (Fig. 6). Together, these results reveal how past selection or Neanderthal introgression have impacted immune responses that contribute to present-day disparities in COVID-19 risk.

Discussion

Here we show that cell type composition is a major driver of population differences in immune responses to SARS-CoV-2. The higher proportions

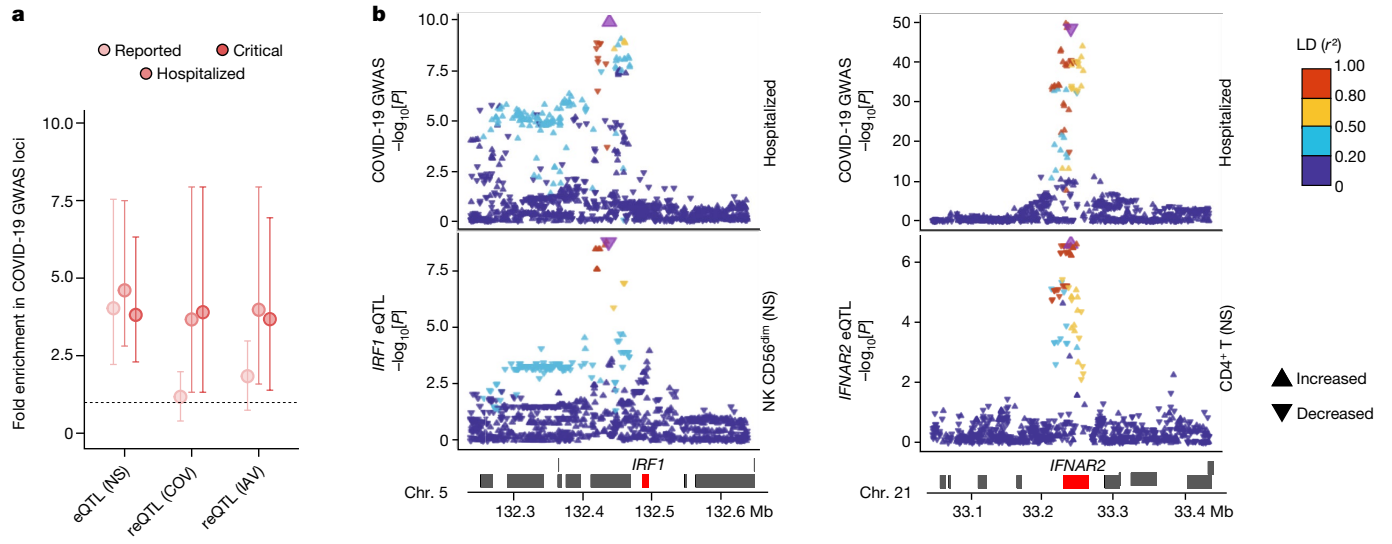


Fig. 5 | eQTLs and reQTLs contribute to COVID-19 risk. a, Enrichment in GWAS loci associated with COVID-19 susceptibility and severity at eQTLs and reQTLs. Data are the mean and 2.5th–97.5th percentiles (95% CIs) of fold enrichments observed over $n = 10{,}000$ resamplings. **b**, Colocalization of *IRF1* and *IFNAR2* eQTLs with COVID-19 severity loci. Top, the $-\log_{10}[P]$ profiles (two-sided Student's *t*-tests) for association with COVID-19-related hospitalization. Bottom, the $-\log_{10}[P]$ profiles for association with expression in non-stimulated CD56^{dim} NK cells (*IRF1*) and CD4$^+$ T cells (*IFNAR2*). The colour code reflects the degree of LD (r^2) with the consensus SNP identified by colocalization analyses (purple). For each SNP, the direction of the arrow indicates the direction of the effect. Chr., chromosome.

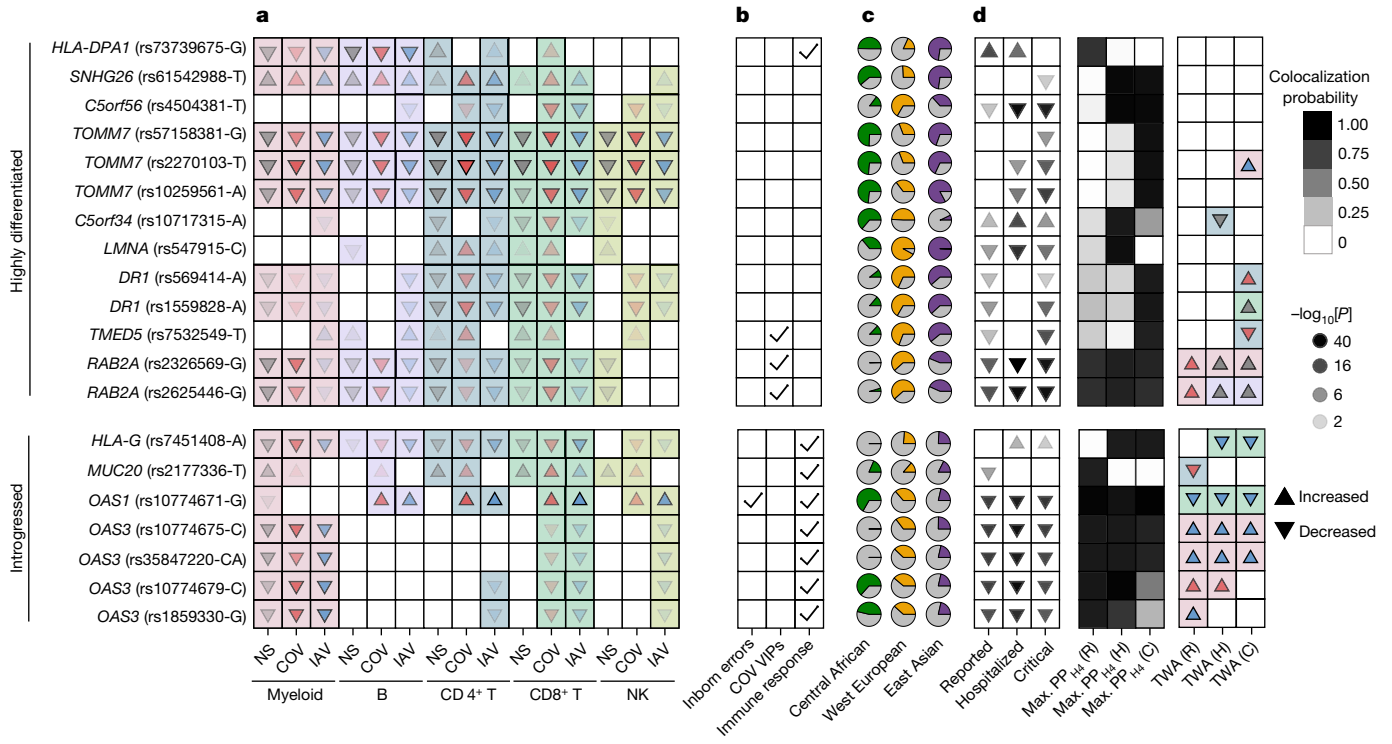


Fig. 6 | Adaptation and archaic introgression at COVID-19-associated (r) eQTLs. a–**d**, Features of (r)eQTLs colocalizing with COVID-19 risk loci ($PP_{H4} > 0.8$) and presenting either strong population differentiation (top 1% PBS genome-wide) or evidence of Neanderthal introgression. **a**, Effects of the target allele on gene expression across immune lineages and stimulation conditions. **b**, Clinical and functional annotations of associated genes. **c**, Present-day population frequencies of the target allele. **d**, The effects of the target allele on COVID-19 risk (infection, hospitalization and critical state), colocalization probability and the lineage and condition in which gene expression most likely affects COVID-19 risk as detected by transcriptome-wide association (TWA) analyses. For expression or COVID-19 associations, the arrows indicate increases/decreases in expression or disease risk with each copy of the target allele, and the opacity reflects the strength of association (two-sided Student's *t*-test $-\log_{10}[P]$). For the TWA analysis, the arrows indicate the effect of an increase in gene expression on the risk of COVID-19. In **a** and **d**, the arrow colours indicate stimulation conditions (non-stimulated (grey), SARS-CoV-2-stimulated (red), IAV-stimulated (blue)) and the background colour indicates the lineage (myeloid (pink), B (purple), CD4$^+$ T (blue), CD8$^+$ T (green), NK (light green)). For each eQTL, the target allele is defined as (1) the derived allele for highly differentiated eQTLs or (2) the allele that segregates with the archaic haplotype for introgressed eQTLs. When the ancestral state is unknown, the minor allele is used as a proxy for the derived allele. Note that, in some cases (for example, *OAS1*), the introgressed allele can be present in Africa, which is attributed to the reintroduction in Eurasia of an ancient allele by Neanderthals[46]. C, critical; H, hospitalized; R, reported.

of memory cells in lymphoid lineages from individuals of African descent, along with their association with CMV infection, highlight how previous environmental exposures can contribute to population disparities in cellular activation states. Neglecting socioenvironmental factors that covary with ancestry may therefore inflate the estimated effects of genetic ancestry on phenotypic variation. One such factor is CMV, affecting leukocyte responses to SARS-CoV-2, but the impact of other exposures on population variation in immune responses remains to be determined. Common genetic variants can also contribute to immune response variation, but their effects primarily apply to a subset of genes showing strong population differentiation. This is illustrated by the rs1142888-G allele, which accounts for the greater than 2.8-fold higher levels of *GBP7* expression in response to viral stimulation in Europeans compared with in Africans. The higher frequency of this allele in Europe probably results from selection occurring 21,900–35,600 years ago. GBP7 facilitates IAV replication by suppressing innate immunity[50], but also regulates host defence to intracellular bacteria such as *Listeria monocytogenes* and *Mycobacterium tuberculosis*[51], providing a plausible mechanism for positive selection at this locus.

This study also shows that natural selection and Neanderthal introgression contributed to differentiate present-day immune responses to SARS-CoV-2. We found traces of selection targeting SARS-CoV-2-specific reQTLs around 25,000 years ago in the ancestors of East Asians, coinciding with the proposed timing of an epidemic that affected the evolution of host coronavirus-interacting proteins[23,24]. However, there is little overlap between alleles selected during this period and variants underlying COVID-19 risk, suggesting changes in the genetic basis of infectious diseases over time, possibly due to the evolution of viruses themselves. Nevertheless, we identified cases (for example, *DR1*, *OAS1-3*, *TOMM7*, *MUC20*) in which selection or archaic introgression contributed to changes in both SARS-CoV-2 immune responses and COVID-19 outcome. Studies based on ancestry-aware polygenic risk scores from cross-population GWAS will be required to establish a formal link between past adaptation and present-day population differences in COVID-19 risk.

Finally, the genetic dissection of variation in transcriptional responses to SARS-CoV-2 provides mechanistic insights into the effects of alleles that are associated with COVID-19 risk. Variants of *IRF1*, *IFNAR2* and *DR1* associated with lower COVID-19 severity increase type I IFN signalling in lymphoid cells by upregulating *IRF1* and *IFNAR2* or downregulating *DR1*, attesting to the importance of efficient IFN signalling for a favourable clinical outcome[4,12–14]. Another example is *MUC20*, at which we identified a Neanderthal-introgressed eQTL that increases *MUC20* expression in SARS-CoV-2-stimulated $CD4^+$ T cells and decreases COVID-19 susceptibility. Given the role of mucins in forming a barrier against infection in the respiratory tract, the high *MUC20* expression in ciliated epithelial cells from the bronchus[52] and the detection of the *MUC20* eQTL in pulmonary tissue (Supplementary Note 11), we suggest that the greater resistance to infection conferred by the Neanderthal haplotype may result from a similar effect on *MUC20* expression in the respiratory tract.

We note two main limitations of our results. First, our samples mostly originate from male individuals, so the impact of sex on immune variation was not addressed. Sex has a widespread yet moderate effect on both transcriptional responses to microbial threats[53] and the genetic regulation of gene expression[54], supporting the transferability of our main conclusions. Nonetheless, examining sex-balanced cohorts will enable the characterization of possible sex-specific differences at the population scale. Second, given the sample sizes and cell counts needed to accurately define population variation in immune activity, we focused on a single system (PBMCs) and selected viral strains. Although PBMCs constitute a valuable model to characterize peripheral immune activation by SARS-CoV-2[9,10], they provide an incomplete representation of the pulmonary epithelium—the primary infection site for respiratory viruses. However, we found that 38% of the eQTLs identified in this study are also detected in lung tissue[55], rising to 72% for eQTLs shared across immune lineages (Supplementary Note 11 and Supplementary Table 9b). Further studies are needed to examine the transferability of our findings to other cell types and to investigate how diverse viral strains affect the dynamics of host responses to SARS-CoV-2.

Overall, our results highlight the value of single-cell approaches in capturing the full diversity of peripheral immune responses to RNA viruses, particularly SARS-CoV-2, and provide insights into environmental, genetic and evolutionary drivers of immune response variation across individuals and populations.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06422-9.

1. O'Driscoll, M. et al. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140–145 (2021).
2. Pei, S., Yamana, T. K., Kandula, S., Galanti, M. & Shaman, J. Burden and characteristics of COVID-19 in the United States during 2020. *Nature* **598**, 338–341 (2021).
3. Sah, P. et al. Asymptomatic SARS-CoV-2 infection: a systematic review and meta-analysis. *Proc. Natl Acad. Sci. USA* **118**, e2109229118 (2021).
4. Casanova, J. L. & Abel, L. From rare disorders of immunity to common determinants of infection: following the mechanistic thread. *Cell* **185**, 3086–3103 (2022).
5. Takahashi, T. et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* **588**, 315–320 (2020).
6. Navaratnam, A. V., Gray, W. K., Day, J., Wendon, J. & Briggs, T. W. R. Patient factors and temporal trends associated with COVID-19 in-hospital mortality in England: an observational study using administrative data. *Lancet Respir. Med.* **9**, 397–406 (2021).
7. Kousathanas, A. et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* **607**, 97–103 (2022).
8. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).
9. Stephenson, E. et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
10. Wilk, A. J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
11. Carvalho, T., Krammer, F. & Iwasaki, A. The first 12 months of COVID-19: a timeline of immunological insights. *Nat. Rev. Immunol.* **21**, 245–256 (2021).
12. Zhang, Q. et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, eabd4570 (2020).
13. Bastard, P. et al. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**, eabd4585 (2020).
14. Manry, J. et al. The risk of COVID-19 death is much greater and age dependent with type I IFN autoantibodies. *Proc. Natl Acad. Sci. USA* **119**, e2200413119 (2022).
15. Shelton, J. F. et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat. Genet.* **53**, 801–808 (2021).
16. Bennett, T. D. et al. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. *JAMA Netw. Open* **4**, e2116901 (2021).
17. Nedelec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669 (2016).
18. Quach, H. et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643–656 (2016).
19. Randolph, H. E. et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* **374**, 1127–1133 (2021).
20. Quintana-Murci, L. Human immunology through the lens of evolutionary genetics. *Cell* **177**, 184–199 (2019).
21. Enard, D. & Petrov, D. A. Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. *Cell* **175**, 360–371 (2018).
22. Enard, D. & Petrov, D. A. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philos Trans. R. Soc. Lond. B* **375**, 20190575 (2020).
23. Souilmi, Y. et al. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *Curr. Biol.* **31**, 3504–3514 (2021).
24. Wang, W. & Han, G. Z. Ancient adaptative evolution of ACE2 in East Asians. *Genome Biol. Evol.* **13**, evab173 (2021).
25. Kerner, G., Patin, E. & Quintana-Murci, L. New insights into human immunity from ancient genomics. *Curr. Opin. Immunol.* **72**, 116–125 (2021).
26. Zeberg, H. & Paabo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).
27. Zeberg, H. & Paabo, S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc. Natl Acad. Sci. USA* **118**, e2026309118 (2021).
28. Lee, J. S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci. Immunol.* **5**, eabd1554 (2020).

29. Leon, J. et al. A virus-specific monocyte inflammatory phenotype is induced by SARS-CoV-2 at the immune-epithelial interface. *Proc. Natl Acad. Sci. USA* **119**, e2116853118 (2022).
30. Hadjadj, J. et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718–724 (2020).
31. Ram, D. R. et al. Tracking KLRC2 (NKG2C)$^{+}$ memory-like NK cells in SIV^{+} and rhCMV^{+} rhesus macaques. *PLoS Pathog.* **14**, e1007104 (2018).
32. O'Neill, M. B. et al. Single-cell and bulk RNA-sequencing reveal differences in monocyte susceptibility to influenza A virus infection between Africans and Europeans. *Front. Immunol.* **12**, 768189 (2021).
33. Bigley, A. B., Spielmann, G., Agha, N., O'Connor, D. P. & Simpson, R. J. Dichotomous effects of latent CMV infection on the phenotype and functional properties of CD8^{+} T-cells and NK-cells. *Cell Immunol.* **300**, 26–32 (2016).
34. Guma, M. et al. Imprint of human cytomegalovirus infection on the NK cell receptor repertoire. *Blood* **104**, 3664–3671 (2004).
35. Patin, E. et al. Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nat. Immunol.* **19**, 302–314 (2018).
36. Zuhair, M. et al. Estimation of the worldwide seroprevalence of cytomegalovirus: a systematic review and meta-analysis. *Rev. Med. Virol.* **29**, e2034 (2019).
37. Wang, P. et al. Inducible microRNA-155 feedback promotes type I IFN signaling in antiviral innate immunity by targeting suppressor of cytokine signaling 1. *J. Immunol.* **185**, 6226–6233 (2010).
38. Syed, F. et al. Excessive matrix metalloproteinase-1 and hyperactivation of endothelial cells occurred in COVID-19 patients and were associated with the severity of COVID-19. *J. Infect. Dis.* **224**, 60–69 (2021).
39. Yi, X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
40. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet.* **15**, e1008384 (2019).
41. Sarute, N. et al. Signal-regulatory protein alpha is an anti-viral entry factor targeting viruses using endocytic pathways. *PLoS Pathog.* **17**, e1009662 (2021).
42. Deschamps, M. et al. Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.* **98**, 5–21 (2016).
43. Gittelman, R. M. et al. Archaic hominin admixture facilitated adaptation to out-of-Africa environments. *Curr. Biol.* **26**, 3375–3382 (2016).
44. Racimo, F., Marnetto, D. & Huerta-Sanchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
45. Choin, J. et al. Genomic insights into population history and biological adaptation in Oceania. *Nature* **592**, 583–589 (2021).
46. Huffman, J. E. et al. Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19. *Nat. Genet.* **54**, 125–127 (2022).
47. Zhao, M. et al. Myeloid neddylation targets IRF7 and promotes host innate immunity against RNA viruses. *PLoS Pathog.* **17**, e1009901 (2021).
48. Deng, M. et al. TRAF3IP3 negatively regulates cytosolic RNA induced anti-viral signaling by promoting TBK1 K48 ubiquitination. *Nat. Commun.* **11**, 2193 (2020).
49. Hsu, S. F., Su, W. C., Jeng, K. S. & Lai, M. M. A host susceptibility gene, DR1, facilitates influenza A virus replication by suppressing host innate immunity and enhancing viral RNA replication. *J. Virol.* **89**, 3671–3682 (2015).
50. Feng, M. et al. Inducible guanylate-binding protein 7 facilitates influenza A virus replication by suppressing innate immunity via NF-κB and JAK-STAT signaling pathways. *J. Virol.* **95**, e02038-20 (2021).
51. Kim, B. H. et al. A family of IFN-γ-inducible 65-kD GTPases protects against bacterial infection. *Science* **332**, 717–721 (2011).
52. Papatheodorou, I. et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83 (2020).
53. Piasecka, B. et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc. Natl Acad. Sci. USA* **115**, E488–E497 (2018).
54. Oliva, M. et al. The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).
55. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

[1]Human Evolutionary Genetics Unit, Institut Pasteur, Université Paris Cité, CNRS UMR2000, Paris, France. [2]Collège Doctoral, Sorbonne Université, Paris, France. [3]Insect-Virus Interactions Unit, Institut Pasteur, Université Paris Cité, CNRS UMR2000, Paris, France. [4]Cytometry and Biomarkers UTechS, Institut Pasteur, Université Paris Cité, Paris, France. [5]Translational Immunology Unit, Institut Pasteur, Université Paris Cité, Paris, France. [6]Université Paris Cité, Imagine Institute, Laboratory of Inflammatory Responses and Transcriptomic Networks in Diseases, Atip-Avenir Team, INSERM UMR1163, Paris, France. [7]Labtech Single-Cell@Imagine, Imagine Institute, INSERM UMR1163, Paris, France. [8]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA. [9]Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, USA. [10]Department of Biology, University of Rome Tor Vergata, Rome, Italy. [11]Structural Virology Unit, Institut Pasteur, Université Paris Cité, CNRS UMR3569, Paris, France. [12]Virus and Immunity Unit, Institut Pasteur, Université Paris Cité, CNRS UMR3569, Paris, France. [13]Ghent University and University Hospital, Ghent, Belgium. [14]Hong Kong Red Cross Blood Transfusion Service, Hospital Authority, Hong Kong SAR, China. [15]WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China. [16]Laboratory of Data Discovery for Health (D24H), Hong Kong Science Park, Hong Kong SAR, China. [17]Division of Public Health Laboratory Sciences, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China. [18]HKU-Pasteur Research Pole, School of Public Health, The University of Hong Kong, Hong Kong SAR, China. [19]Centre for Immunology and Infection, Hong Kong Science Park, Hong Kong SAR, China. [20]St Giles Laboratory of Human Genetics of Infectious Diseases, The Rockefeller University, New York, NY, USA. [21]Laboratory of Human Genetics of Infectious Diseases, INSERM UMR1163, Necker Hospital for Sick Children, Paris, France. [22]Université Paris Cité, Imagine Institute, Paris, France. [23]Department of Pediatrics, Necker Hospital for Sick Children, Paris, France. [24]Howard Hughes Medical Institute, New York, NY, USA. [25]Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia. [26]Chair Human Genomics and Evolution, Collège de France, Paris, France. [27]These authors contributed equally: Yann Aquino, Aurélie Bisiaux, Zhi Li, Mary O'Neill. [28]These authors jointly supervised this work: Maxime Rotival, Lluis Quintana-Murci. ✉e-mail: maxime.rotival@pasteur.fr; quintana@pasteur.fr

Methods

Sample collection

The individuals of self-reported African (AFB) and European (EUB) descent studied are part of the EVOIMMUNOPOP cohort[18]. In brief, 390 healthy male donors (188 AFB and 202 EUB) were recruited between 2012 and 2013 in Ghent (Belgium), thus, before the COVID-19 pandemic. Blood was obtained from the healthy volunteers, and the PBMC fraction was isolated and frozen. Inclusion in the current study was restricted to 80 nominally healthy individuals of each ancestry, aged between 19 and 50 years at the time of sample collection. Donors of African descent originated from West Central Africa, with >90% being born in either Cameroon or the Democratic Republic of Congo. For this study, 71 additional individuals of East Asian descent (ASH) were included, of whom 62 were retained after quality control (see the 'scRNA-seq library preparation and data processing' section). ASH individuals were recruited at the School of Public Health, University of Hong Kong, and were included in a community-based sero-epidemiological COVID-19 study (research protocol number JTW 2020.02). Inclusion for the study described here was restricted to nominally healthy ASH individuals (30 men and 41 women) aged between 19 and 65 years of age and seronegative for SARS-CoV-2. Samples were collected at the Red Cross Blood Transfusion Service (Hong Kong) where the PBMC fraction was isolated and frozen. Target sample sizes were determined to ensure >80% power for the detection of eQTLs with R2 higher than 0.2, at a $P < 5 \times 10^{-9}$ threshold.

In this study, we refer to individuals of Central African (AFB), West European (EUB) and East Asian (ASH) ancestries to describe individuals who are genetically similar (that is, lowest F_{ST} values) to populations from West-Central Africa, Western Europe and East Asia, using the 1000 Genomes (1KG) Project[56] data as a reference (Supplementary Fig. 1a). Notably, the AFB, EUB and ASH samples present no evidence of recent genetic admixture with populations originating from another continent, besides two AFB donors who respectively present 22% of Near Eastern- and 25% of European-ancestries. Such a moderate level of admixture in fewer than 1% of individuals is unlikely to have any significant impact on the results presented.

All of the samples were collected after written informed consent was obtained from the donors, and the study was approved by the ethics committee of Ghent University (B670201214647), the Institutional Review Board of the University of Hong Kong (UW 20-132), and the relevant French authorities (CPP, CCITRS and CNIL). This study was also monitored by the Ethics Board of Institut Pasteur (EVOIMMUNOPOP-281297).

Genome-wide DNA genotyping

The AFB and EUB individuals were previously genotyped at 4,301,332 SNPs, using the Omni5 Quad BeadChip (Illumina) with processing as previously described[18]. The additional 71 ASH donors were genotyped separately at 4,327,108 SNPs using the Infinium Omni5-4 v.1.2 BeadChip (Illumina). We updated SNP identifiers based on Illumina annotation files (https://support.illumina.com/content/dam/illumina-support/documents/downloads/productfiles/humanomni5-4/v1-2/infinium-omni5-4-v1-2-a1-b144-rsids.zip) and called the genotypes of all ASH individuals jointly on GenomeStudio (v.2011.1; https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html). We then removed SNPs with (1) no 'rs' identifiers or with no assigned chromosome or genomic position (n = 14,637); (2) duplicated identifiers (n = 5,059); or (3) a call rate of <95% (n = 10,622). We then used the 1KG Project Phase 3 data[56] as a reference for merging the ASH genotyping data with that of AFB and EUB individuals and detecting SNPs misaligned between the three genotype datasets. Before merging, we removed SNPs that (1) were absent from either the Omni5 or 1KG datasets (n = 469,535); (2) were transversions (n = 138,410); (3) had incompatible alleles between datasets, before and after allele flipping (n = 1,250); and (4) had allele frequency differences of more than 20% between the AFB and Luhya from Webuye, Kenya (LWK) and Yoruba from Ibadan, Nigeria (YRI), or between the EUB and Utah residents with Northern and Western European ancestries (CEU) and British individuals from England and Scotland (GBR), or between the ASH and Southern Han Chinese (CHS) (n = 777). Once the data had been merged, we performed principal component analysis (PCA) using PLINK (v.1.9)[57] and ensured that the three study populations (that is, AFB, EUB and ASH) overlapped with the corresponding 1KG populations, to exclude batch effects between genotyping platforms (Supplementary Fig. 1a). The final genotyping dataset included 3,723,840 SNPs.

Haplotype phasing and imputation

After merging genotypes from AFB, EUB and ASH donors, we filtered genotypes for duplicates with bcftools norm --rm-dup all (v.1.16)[58] and lifted all genotypes over to the human genome assembly GRCh38 with GATK's (v.4.1.2.0) LiftoverVcf using the RECOVER_SWAPPED_ALT_REF=TRUE option[59]. We then filtered out duplicated variants again before phasing genotypes with SHAPEIT4 (v.4.2.1)[60] and imputing missing variants with Beagle5.1 (v.18May20.d20)[61], treating each chromosome separately. For both phasing and imputation, we used the genotypes of 2,504 unrelated individuals from the 1KG Project Phase 3 data as a reference (downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502 and lifted over to GRCh38) and downloaded genetic maps from the GitHub pages of the associated software (that is, SHAPEIT4 for phasing and *Beagle5.1* for imputation). A third round of duplicate filtering was performed after phasing and before imputation using Beagle5.1 (v.18May20.d20)[61]. Phasing was performed using the setting --pbwt-depth=8 and imputation was performed assuming an effective population size (N_e) of 20,000. The quality of imputation was assessed by cross-validation; specifically, we performed 100 independent rounds of imputation excluding 1% of the variants and compared the imputed allelic dosage with the observed genotypes for these variants (Supplementary Fig. 1b,c). The results obtained confirmed that imputation quality was satisfactory, with 98% of common variants (that is, MAF > 5%) having an $r^2 > 0.8$ for the correlation between observed and imputed genotypes (>95% concordance for 96% of common variants). After imputation, variants with a MAF < 1% or with a low predicted quality of imputation (that is, DR2 < 0.9) were excluded, yielding a final dataset of 13,691,029 SNPs for downstream analyses.

Viruses used in this study

To evaluate population differences in the immune responses to SARS-CoV-2, we chose the viral strain that circulated in France at the time of our experiments (Autumn 2020). This reference strain (BetaCoV/France/GE1973/2020) was supplied by the National Reference Centre for Respiratory Viruses hosted by Institut Pasteur and headed by S. van der Werf. The human sample from which the strain was isolated was provided by L. Andreoletti from the Robert Debré Hospital. To characterize the distinctive features of SARS-CoV-2 responses, we included in our study the IAV as a reference comparison of another respiratory RNA virus. Specifically, we chose the PR8 strain (IAV, PR/8, H1N1/1934) based on our previous experience with this virus, its availability in the laboratory and its ability to trigger IFN responses in healthy human donors[53,62]. The PR8 strain used was purchased from Charles River Laboratories (3X051116) and provided in ready-to-use aliquots that were stored at −80 °C.

SARS-CoV-2 stock production

To produce SARS-CoV-2, we used African green monkey kidney Vero E6 cells that were tested for mycoplasma contamination and maintained at 37 °C in 5% CO_2 in Dulbecco's minimum essential medium (DMEM) (Sigma-Aldrich) supplemented with 10% fetal bovine serum (FBS, Dutscher) and 1% penicillin–streptomycin (Gibco, Thermo Fisher Scientific). Vero E6 cells were plated at 80% confluence in 150 cm^2 flasks

and infected with SARS-CoV-2 at a multiplicity of infection (MOI) of 0.01 in DMEM supplemented with 2% FBS and 1% penicillin–streptomycin. After 1 h, the inoculum was removed and replaced with DMEM supplemented with 10% FBS and 1% penicillin–streptomycin, and cells were incubated for 72 h at 37 °C in 5% CO_2. The cell culture supernatant was collected and centrifuged for 10 min at 3,000 rpm to remove cellular debris, and polyethylene glycol (PEG; PEG8000, Sigma-Aldrich) precipitation was performed to concentrate the viral suspension. In brief, 1 l of viral stock was incubated with 250 ml of 40% PEG solution (that is, 8% PEG final) overnight at 4 °C. The suspension was centrifuged at 10,000*g* for 30 min at 4 °C and the resulting pellet was resuspended in 100 ml of RPMI medium (Gibco, Thermo Fisher Scientific) supplemented with 10% FBS (hereafter R10) and viral aliquots were stored at −80 °C. SARS-CoV-2 viral titres were determined using a focus-forming unit assay as previously described[63]. In brief, Vero E6 cells were plated in a 96-well plate with 2×10^4 cells per well. The cellular monolayer was infected with serial dilutions (1:10) of viral stock and overlaid with a semi-solid 1.5% carboxymethylcellulose (Sigma-Aldrich) and 1× MEM medium for 36 h at 37 °C. Cells were then fixed with 4% paraformaldehyde (Sigma-Aldrich), and permeabilized with 1× phosphate-buffered saline, 0.5% Triton X-100 (Sigma-Aldrich). Infectious foci were stained with a human anti-SARS-CoV-2 spike antibody (H2-162, Hugo Mouquet's laboratory, Institut Pasteur) and the corresponding HRP-conjugated secondary antibody (Sigma-Aldrich). Foci were stained using a 3,3′-diaminobenzidine staining solution (DAB, Sigma-Aldrich) and counted using the BioSpot suite of the C.T.L. ImmunoSpot S6 Image Analyzer.

In vitro peripheral blood mononuclear cell stimulation

We performed scRNA-seq analysis of SARS-CoV-2-, IAV- and mock-stimulated (referred to as the non-stimulated condition) PBMCs from healthy donors (80 AFB, 80 EUB and 71 ASH) in 16 experimental runs. We first performed a kinetic experiment (run 1) on samples from 4 AFB and 4 EUB individuals stimulated for 0, 6 and 24 h to validate our in vitro model across different timepoints (Supplementary Note 1, Supplementary Fig. 2 and Supplementary Table 2). The 6 h timepoint was identified as the optimal timepoint for the analysis (Supplementary Note 1). We then processed the rest of the cohort, over runs 2 to 15. Finally, we reprocessed some samples (run 16) to assess the technical variability in our setting and to increase in silico cell counts (see the 'scRNA-seq library preparation and data processing' section). Ancestry-related batch effects were minimized by scheduling sample processing to ensure a balanced distribution of AFB, EUB and ASH donors within each run. Donors used for feach run were randomly selected within each population.

For each run, cryopreserved PBMCs were thawed in a 37 °C water bath, transferred to 25 ml of R10 medium (that is, RPMI 1640 supplemented with 10% heat-inactivated FBS) at 37 °C, and centrifuged at 300*g* for 10 min at room temperature. Cells were counted, resuspended at 2×10^6 cells per ml in warm R10 in 25 cm^2 flasks, and rested overnight (that is, 14 h) at 37 °C. The next morning, PBMCs were washed and resuspended at a density of 3.3×10^6 cells per ml in R10; 120 µl of a suspension containing 4×10^5 cells from each sample was then plated in a 96-well untreated plate (Greiner Bio-One) for each of the three sets of stimulation conditions. We added 80 µl of either R10 (non-stimulated), SARS-CoV-2 or IAV stock (corresponding to 4×10^5 focus-forming units diluted in R10) to the cells, so as to achieve a multiplicity of infection (MOI) of 1 and an optimal PBMC concentration of 2×10^6 cells per ml. Cells were incubated at 37 °C for 0, 6 or 24 h for the kinetic experiment (run 1), and for 6 h for all subsequent runs (runs 2 to 16), in a biosafety level 3 (BSL-3) facility at Institut Pasteur, Paris. The plates were then centrifuged at 300*g* for 10 min and supernatants were stored at −20 °C until use (see 'Supernatant cytokine assays' section). All of the samples from the same run were resuspended in Dulbecco's phosphate-buffered saline (Gibco), supplemented with 0.04% bovine serum albumin (BSA, Miltenyi Biotec), and multiplexed in eight pools according to a pre-established study design (Supplementary Fig. 3a and Supplementary Table 3a). The cells from each pool were counted using the Cell Countess II automated cell counter (Thermo Fisher Scientific) and the cell density was adjusted to 1,000 viable cells per µl of 0.04% BSA in phosphate-buffered saline. When performing stimulations, researchers were blinded to the genotype and environmental exposures of the individual.

scRNA-seq library preparation and data processing

We generated scRNA-seq cDNA libraries using the Chromium Controller (10x Genomics) according to the manufacturer's instructions for the Chromium Single Cell 3′ Library and Gel Bead Kits (v.3.1). Library quality and concentration were assessed using the Agilent 2100 Bioanalyzer and a Qubit fluorometer (Thermo Fisher Scientific). The final products were processed for high-throughput sequencing on a HiSeqX platform (Illumina).

Paired-end sequencing reads from each of the 133 scRNA-seq cDNA libraries (13 libraries from the kinetic experiment and 120 from the population-level study) were independently mapped onto the concatenated human (GRCh38), SARS-CoV-2 (hCoV-19/France/GE1973/2020) and IAV (A/Puerto Rico/8/1934(H1N1)) genome sequences using the STARsolo aligner (v.2.7.8a)[64] (Supplementary Fig. 3b). We obtained a mean of 10,785 cell-containing droplets per library, and each droplet was assigned to its sample of origin with Demuxlet (v.0.1)[65], based on the genotyping data available for each individual. Singlet/doublet calls were compared with the output of Freemuxlet (v.0.1)[65] to ensure good agreement (Supplementary Fig. 3c–e). We loaded feature-barcode matrices for all cell-containing droplets identified as singlets by Demuxlet in each scRNA-seq library onto a SingleCellExperiment (v.1.14.1) object[66]. Data from barcodes associated with low-quality or dying cells were removed with a hard threshold-based filtering strategy based on three metrics: cells with fewer than 1,500 total unique molecular identifier (UMI) counts, 500 detected features or a mitochondrial gene content exceeding 20% were removed from each sequencing library (Supplementary Fig. 3f). We also discarded samples from nine ASH donors from whom fewer than 500 cells were obtained in at least one condition (Supplementary Fig. 3g).

We then log-normalized raw UMI counts with a unit pseudocount and library size factors (that is, the number of reads associated with each barcode) were calculated with quickClusters and computeSumFactors from the scran package (v.1.20.1)[67]. We then calculated the mean and variance of log-transformed counts for each gene and broke the variance down into a biological and a technical component with the fitTrendPoisson and modelGeneVarByPoisson functions of scran. This approach assumes that technical noise is Poisson-distributed and simulates Poisson-distributed data to derive the mean-variance relationship expected in the absence of biological variation. Excess variance relative to the null hypothesis is considered to correspond to the biological variance. We retained only those genes for which the biological variance component was positive with an FDR below 1%. We used this filtered feature set and the technical variance component modelled from the data to run PCA with denoisePCA from scran, thus discarding later components more likely to capture technical noise. Doublets (that is, barcodes assigned to cells from different individuals captured in the same droplet) are likely to be in close neighbourhoods when projected onto a subspace of the data of lower dimensionality. We therefore used a *k*-nearest neighbours approach to discard cryptic doublets (that is, barcodes associated to different cells from the same individual captured in the same droplet). Barcodes identified as singlets by Demuxlet but having over 5 out of 25 doublet nearest neighbours in the PCA space were reassigned as doublets and excluded from further analyses (Supplementary Fig. 3h).

After data preprocessing, we performed a second round of UMI count normalization, feature selection and dimensionality reduction on the cleaned data to avoid bias due to the presence of low-quality cells and

cryptic doublets. Differences in sequencing depth were equalized between batches (that is, sequencing libraries) using multiBatchNorm from batchelor (v.1.8.1) to scale library size factors according to the ratio of mean counts between batches[68] (Supplementary Fig. 3i). We accounted for the different mean-variance trends in each batch, by applying modelGeneVarByPoisson separately for each sequencing library, and then combining the results for all batches with combineVar from scran[67]. We then bound all 133 separate preprocessed feature-barcode matrices into a single merged SingleCellExperiment object, log-normalized UMI counts according to the scaled size factors and selected genes with mean log-expression values over 0.01 or a biological variance compartment exceeding 0.001 (Supplementary Fig. 3j). On the basis of this set of highly variable genes and the variance decomposition, we then performed PCA on the whole dataset using denoisePCA, and then used Harmony (v.0.1.0) on the PCs to adjust for library effects[69].

Clustering and cell type assignment

We performed cluster-based cell type identification in each stimulation condition, according to a four-step procedure. We first performed low-resolution (res. parameter = 0.8) shared nearest-neighbour graph-based ($k = 25$) clustering using FindClusters from Seurat (v.4.1.1) with assignment to one of three meta-clusters (that is, myeloid, B lymphoid and T/NK lymphoid) on the basis of the transcriptional profiles of the cells for canonical markers (for example, *CD3E-F*, *CD14*, *FCGR3A*, *MS4A1*) (Supplementary Fig. 4a,b). We next performed a second round of clustering at higher resolution (res. parameter = 3) within each meta-cluster and stimulation condition (Supplementary Fig. 4c). We systematically tested for differential expression between each cluster and the other clusters of the same meta-cluster and stimulation condition. This made it possible to define unbiased markers ($|\log_2 FC| \neq 0$, FDR < 0.01) for each cluster (Supplementary Fig. 4d). We then used these expression profiles of these genes to assign each cluster manually to one of 22 different cell types (Supplementary Fig. 4e), which, for some analyses, were collapsed into five major immune lineages. This step was performed in parallel by three investigators to consolidate consensus assignments. We also used cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) data, generated for a subset of cells (2% of the whole dataset), to validate our assignments and redefine clusters presenting ambiguous transcriptional profiles (for example, memory-like NK cells; Supplementary Fig. 4f).

By calling cell types from high-resolution, homogeneous clusters, assigned independently for each lineage and stimulation condition (that is, non-stimulated, SARS-CoV-2, and IAV), we were able to preserve much of the diversity in our dataset, while avoiding potential confounding effects due to the stimulation conditions. However, some clusters were characterized by markers associated with different cell types. Most of these clusters corresponded to mixtures of similar cell types (for example, the expression of *CD3E*, *CD8A*, *NKG7* and *CD16* suggested a mixture of cytotoxic $CD8^+$ T and NK cells) and were consistent with the known cell hierarchy. Other, less frequent clusters expressed a combination of markers usually associated with lineages originating from different progenitors (for example, *CD3E* and *CD19*, associated with T and B lymphocytes, respectively). These clusters were considered to be incoherent and were discarded. In the fourth and final step of our procedure, we used linear discriminant analysis to resolve within each condition the mixtures that were consistent with the established cell hierarchy, to obtain a final cell assignment (Supplementary Fig. 4g,h). For clusters of mixed identity AB, we built a training dataset from 10,000 observations sampled from the set of cells called as A or B, preserving the corresponding frequencies of these cells in the whole dataset. We then used a model trained on these data to predict the specific cellular identities within the mixed cluster.

Cell abundance from each immune lineage/cell type was compared between non-stimulated and SARS-CoV-2-/IAV-stimulated conditions, using Wilcoxon's signed-rank test matching on the individual. FDR was calculated across all conditions and lineages with the Benjamini–Hochberg procedure (p.adjust function with the 'fdr' method). Viral stimulation had a moderate effect on the estimated cell proportions and, although significant differences were detected, the total number of cells per cell type was generally consistent across conditions (Supplementary Note 2 and Supplementary Table 3e).

Cellular indexing of transcriptomes and epitopes by sequencing

To confirm the identity of specific cell types expressing ambiguous markers at the RNA level, during the last experimental run (run 16), half the cells from each experimental condition were used to perform CITE-seq, according to the manufacturer's instructions (10x Genomics). PBMCs were washed, resuspended in chilled 1% BSA in phosphate-buffered saline and incubated with human TruStain FcX blocking solution (BioLegend) for 10 min at 4 °C. Cells were then stained with a cocktail of TotalSeq-B antibodies (BioLegend) previously centrifuged at 14,000*g* for 10 min (Supplementary Table 3b). The cells were incubated for 30 min at 4 °C in the dark and were then washed three times. Cell density was then adjusted to 1,000 viable cells per µl in 1% BSA in phosphate-buffered saline. We generated scRNA-seq libraries and cell protein libraries (L131–L134) with the Chromium Single Cell 3′ Reagent Kit (v.3.1), using the Feature Barcoding technology for Cell Surface Proteins (10x Genomics).

Supernatant cytokine assays

Before protein analysis, sample supernatants were treated in the BSL-3 facility to inactivate the viruses, according to a published protocol for SARS-CoV[70], which we validated for SARS-CoV-2. In brief, all of the samples were treated with 1% (v/v) Triton X-100 (Sigma-Aldrich) for 2 h at room temperature, which effectively inactivated both SARS-CoV-2 and IAV. The protein concentration was then determined with a commercial Luminex multi-analyte assay (Biotechne, R&D Systems) and the SIMOA Homebrew assay (Quanterix). For the Luminex assay, we used the XL Performance Kit according to the manufacturer's instructions, and proteins were determined using the Bioplex 200 (Bio-Rad) system. Furthermore, IFNα, IFNγ (duplex) and IFNβ (single-plex) protein concentrations were quantified in SIMOA digital ELISA tests developed as Quanterix Homebrews according to the manufacturer's instructions (https://portal.quanterix.com/). The SIMOA IFNα assay was developed with two autoantibodies specific for IFNα isolated and cloned (Evitria) from two patients with autoimmune polyglandular syndrome type 1 (also known as autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy)[71] and covered by patent application WO2013/098419. These antibodies can be used for the quantification of all IFNα subtypes with a similar sensitivity. The 8H1 antibody clone was used to coat paramagnetic beads at a concentration of 0.3 mg ml^{-1} for use as a capture antibody. The 12H5 antibody was biotinylated (biotin/antibody ratio = 30:1) and used as the detector antibody, at a concentration of 0.3 µg ml^{-1}. The SBG enzyme for detection was used at a concentration of 150 pM. Recombinant IFNα17/αI (PBL Assay Science) was used as calibrator. For the IFNγ assay, the MD-1 antibody clone (BioLegend) was used to coat paramagnetic beads at a concentration of 0.3 mg ml^{-1} for use as a capture antibody. The MAB285 antibody clone (R&D Systems) was biotinylated (biotin/antibody ratio = 40:1) and used as the detector antibody at a concentration of 0.3 µg ml^{-1}. The SBG enzyme used for detection was used at a concentration of 150 pM. Recombinant IFNγ protein (PBL Assay Science) was used as a calibrator. For the IFNβ assay, the 710322-9 IgG1, kappa, mouse monoclonal antibody (PBL Assay Science) was used to coat paramagnetic beads at a concentration of 0.3 mg ml^{-1}, for use as a capture antibody. The 710323-9 IgG1 kappa mouse monoclonal antibody was biotinylated (biotin/antibody ratio = 40:1) and used as the detector antibody at a concentration of 1 µg ml^{-1}. The SBG enzyme for detection was used at a concentration of 50 pM. Recombinant IFNβ protein (PBL Assay Science) was used

as a calibrator. The limit of detection of these assays was 0.8 fg ml^{-1} for IFNα, 20 fg ml^{-1} for IFNγ and 0.2 pg ml^{-1} for IFNβ, considering the dilution factor of 10.

Flow cytometry

Frozen PBMCs from three AFB (CMV$^+$) and six EUB (three CMV$^+$, three CMV$^-$) donors were thawed and allowed to rest overnight, as previously described. For each donor, 10^6 cells were resuspended in phosphate-buffered saline supplemented with 2% FBS and incubated with human Fc blocking solution (BD Biosciences) for 10 min at 4 °C. Cells were then stained with the following antibodies for 30 min at 4 °C: CD3 VioGreen (BW264/56, Miltenyi Biotec), CD14 V500 (M5E2, BD Biosciences), CD57 Pacific Blue (HNK-1, BioLegend), NKp46 PE (9E2/NKp46, BD Biosciences), CD16 PerCP-Cy5.5 (3G8, BD Biosciences), CD56 APC-Vio770 (REA196, Miltenyi Biotec), NKG2A FITC (REA110, Miltenyi Biotec) and NKG2C APC (REA205, Miltenyi Biotec). The cells were then washed and acquired on the MACSQuant cytometer (Miltenyi Biotec), and the data were analysed using FlowJo (v.10.7.1)[72].

Quantification of batch effects and replicability

Once all the samples had been processed, we used the kBET metric (v.0.99.6)[73] to assess the intensity of batch effects and to quantify the relative effects of technical and biological variation on cell clustering. This made it possible to confirm that the variation across libraries, and across experimental runs, remained limited relative to the variation across individuals or across conditions (Supplementary Fig. 5a). We used technical replicates to assess the replicability of our observations across independent stimulations. Agreement was good between the cell proportions and the ISG activity scores inferred across independent runs ($r > 0.82$, $P < 7.6 \times 10^{-13}$) (Supplementary Fig. 5b,c).

Pseudobulk estimation, normalization and batch correction

Individual variation in gene expression was quantified at two resolutions: five major immune lineages and 22 cell types. We aggregated raw UMI counts from all high-quality single-cell transcriptomes ($n = 1{,}047{,}824$) into bulk expression estimates by summing gene expression values across all cells assigned to the same lineage/cell type and sample (that is, individual and stimulation conditions) using the aggregateAcrossCells function of scuttle (v.1.2.1)[74]. We then normalized the raw aggregated UMI counts by library size, generating 3,330 lineage-wise (222 donors × 3 sets of conditions × 5 lineages) and 14,652 cell type-wise (666 samples × 22 cell types) pseudobulk counts-per-million (CPM) values, for all genes in our dataset. CPM values were then $\log_2$-transformed, with an offset of 1 to prevent non-finite values and to stabilize variation for weakly expressed genes. Genes with a mean CPM < 1 across all conditions and lineages/cell types were considered to be non-expressed and were discarded from further analyses, leading to a final set of 12,667 genes at the lineage level (12,672 genes when increasing granularity to 22 cell types), including 12 viral transcripts. To quantify the experimental variation induced by the experimental run, library preparation and sequencing, and to remove unwanted batch effects, we first used the lmer function of the lme4 package (v.1.1-27.1)[75] to fit a linear model of the following form in each stimulation condition and for each lineage/cell type:

$$\log(1 + \mathrm{CPM}_i) = \alpha + \mathrm{IID}_i + \mathrm{LIB}_i + \mathrm{RUN}_i + \mathrm{FLOW}_i + \varepsilon_i \quad (1)$$

where CPM_i is the gene expression in sample i (that is, one replicate of a given individual and set of experimental conditions); α is the intercept; $\mathrm{IID}_i \sim \mathcal{N}(0, \sigma^2_{\mathrm{IND}})$ captures the effect of the corresponding individual on gene expression; $\mathrm{LIB}_i \sim \mathcal{N}(0, \sigma^2_{\mathrm{LIB}})$ captures the effect of 10x Genomics library preparation; $\mathrm{RUN}_i \sim \mathcal{N}(0, \sigma^2_{\mathrm{RUN}})$ captures the effect of the experimental run; $\mathrm{FLOW}_i \sim \mathcal{N}(0, \sigma^2_{\mathrm{Flowcell}})$ captures the effect of the sequencing flow cell; and ε_i captures residual variation between samples. We then subtracted the estimated value of the library, experimental run and flow cell effects (as provided by the ranef function) from the transformed CPMs of each sample, to obtain batch-corrected CPM values. Finally, we averaged the batch-corrected CPM values obtained across different replicates for the same individual and set of stimulation conditions, to obtain final estimates of gene expression.

For each cell type and stimulation condition, an inverse-normal rank-transformation was applied to the $\log_2$[CPM] of each gene, before testing for differences in gene expression between populations and mapping eQTL. Within each lineage and set of stimulation conditions, we ranked, for each gene, the pseudobulk expression values of all individuals, assigning ranks at random for ties, and replaced each observation with the corresponding quantile from a normal distribution with the same mean and s.d. as the original expression data. This inverse-normal rank-transformation rendered downstream analyses robust to zero-inflation in the data and outlier values, while maintaining the rank-transformed values on the same scale as the original data.

Variance explained by lineage identity and viral exposure

We used CAR scores[76] to quantify the fraction of gene expression variance that is explained by variation across immune lineages and stimulation conditions. First, we built per-gene linear models regressing pseudobulk expression levels on two sets of dummy variables, encoding both lineage identity and stimulation condition. Specifically, we used a model of the form:

$$\mathrm{Expr}_{ils} = \alpha + \sum_{l=2}^{5} \beta_l I_{\{\mathrm{lineage}=l\}} + \sum_{l=2}^{5} \sum_{s=2,3} \gamma_{ls} I_{\{\mathrm{lineage}=l \text{ and } \mathrm{stim}=s\}} + \varepsilon_{ils} \quad (2)$$

Where Expr_{ils} is the log-transformed expression of the target gene for donor i, in lineage l and in the condition of stimulation s; α is the intercept measuring the mean expression of the reference lineage (CD4$^+$ T cells) in the non-stimulated state; β_l are parameters that capture the mean difference (log-fold change) in expression between lineage l and the reference lineage; I is an indicator variable equal to 1 when the subscript condition is met, and 0 otherwise; γ_{ls} are parameters that capture the mean log-fold change in expression of lineage l in response to stimulus s; and ε_{ils} are normally distributed residuals. We then ran the carscore function from care R package (v.1.1.11)[76] on each model, setting $\lambda = 0$ (that is, no shrinkage), to obtain the CAR score associated with each parameter. In brief, care decorrelates a set of predictors using a Mahalanobis whitening transformation and computes CAR scores as marginal correlations between these decorrelated predictors and the response variable. This enables direct estimation of the contribution of each predictor to the variance of the response variable as the square of its CAR score. The variance explained by cellular identity (lineage) and stimulation is then computed as:

$$\mathrm{Var}_{\mathrm{lineage}} = \sum_{l=2}^{5} \mathrm{CAR}(\beta_l)^2 \text{ and } \mathrm{Var}_{\mathrm{stim}} = \sum_{l=2}^{5} \sum_{s=1,2} \mathrm{CAR}(\gamma_{ls})^2 \quad (3)$$

ISG activity calculation

ISGs strongly respond to both viruses across all lineages/cell types. We therefore evaluated each donor's ISG expression level in the basal state or after stimulation with either SARS-CoV-2 or IAV by constructing an ISG activity score. For the human genes in our filtered gene set ($n = 12{,}655$), we defined as ISGs ($n = 174$) those genes included in the union of GSEA's hallmark (https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=H) IFNα response and IFNγ response gene sets, but excluded those from the inflammatory response set. We then used AddModuleScore from Seurat (v.4.1.1)[77] to measure ISG activity as the mean pseudobulk expression level of ISGs in each sample minus the mean expression for one hundred randomly selected non-ISGs matched for mean magnitude of expression. In all analyses, ISG activity

scores were adjusted for cell mortality of the sample by fitting a model of the form:

$$\mathrm{ISG}_i = \alpha + \beta_p \,\mathrm{Population}_i + \beta_m \,\mathrm{CellMortality}_i + \varepsilon_i \quad (4)$$

and subtracting the effect of cell mortality from the raw ISG scores. In this model, ISG_i denotes the ISG activity score of individual i; α is the intercept, $\mathrm{Population}_i$ and $\mathrm{CellMortality}_i$ are variables reflecting the donor's population and the cell mortality of the sample; β_p and β_m are parameters capturing the effect of the population and cell mortality on ISG activity; and ε_i are normally distributed residuals. The difference in variance of ISG activity between SARS-CoV-2 and IAV was assessed using Levene's test. For comparisons with SIMOA-estimated IFN levels, the carscore function from the care R package[76] was used to model ISG activity as a function of levels of IFNα, IFNβ and IFNγ, adjusting for population, age, sex and cell mortality. The percentage of ISG variance attributable to each IFN (α, β, or γ) was estimated as the square of the resulting CAR scores.

Testing for differences in lineage/cell type abundance between populations

We compared immune cell abundance between donors of African and European ancestries by contrasting the average number and percentage of cells assigned to each lineage/cell type between donors from both populations. To assess the statistical significance of population differences in cell type frequency, we first corrected cellular frequencies for the confounding effects of age, cell mortality and total number of cells in each sample (that is, donor × condition) using a linear model of the form

$$\begin{aligned}\mathrm{CellularFrequency}_{cis} = {} & \alpha + \beta_p \,\mathrm{Population}_i + \beta_a \,\mathrm{Age}_i \\ & + \beta_m \,\mathrm{CellMortality}_i + \beta_c \,\mathrm{NCells}_{is} + \varepsilon_i\end{aligned} \quad (5)$$

and subtracting the effect of these three covariates from the raw cell frequencies. In this model, $\mathrm{CellularFrequency}_{cis}$ denotes the frequency/number of the lineage/cell type c under consideration in individual i and condition s; α is the intercept; $\mathrm{Population}_i$, Age_i, $\mathrm{CellMortality}_i$ and NCells_{is} are variables reflecting the donor's age and population, the cell mortality of the sample and the total number of cells recovered in condition s; β_p, β_a, β_m and β_c are parameters capturing the effect of these covariates on cellular composition; and ε_i are normally distributed residuals. The adjusted cell frequencies were then compared between populations using Wilcoxon's rank-sum tests.

Mapping the genetic determinants of immune cell composition

We performed genome-wide association studies of the proportions of each immune lineage/cell type in the different stimulation conditions. In brief, we used PLINK (v.1.9)[78] to estimate at each locus the additive effect of each copy of the alternate allele on two quantitative traits: (1) the number of cells of each lineage relative to total number of cells in the sample and (2) the number of cells of each cell type relative to the lineage under consideration. In total, we performed 79 GWASs: one for each of the 27 immune classes (5 lineages and 22 cell types), in each of the 3 experimental conditions (except for the IAV-infected $\mathrm{CD14}^+$ monocytes, which are only present in the IAV condition). In each GWAS, we modelled cell type frequencies across individuals as

$$\begin{aligned}\mathrm{CellularFrequency}_{cis} = {} & \alpha + \beta \,\mathrm{SNP}_i + \beta_p \,\mathrm{Population}_i + \beta_s \,\mathrm{Sex}_i + \beta_a \,\mathrm{Age}_i \\ & + \beta_m \,\mathrm{CellMortality}_i + \beta_c \,\mathrm{NCells}_{is} + \varepsilon_i\end{aligned} \quad (6)$$

where the $\mathrm{CellularFrequency}_{cis}$ is the rank-transformed percentage of lineage/cell type c in the sample (that is, among cells from donor i in condition s); $\mathrm{SNP}i$ is the number of alternative alleles of donor i for the target SNP; $\mathrm{Population}_i$, Sex_i and Age_i are variables reflecting the donor's characteristics (population of origin, genetic sex and age); $\mathrm{CellMortality}_i$ and NCells_{is} are variables reflecting technical parameters (respectively, the percentage of dead cells after thawing the cryopreserved PBMCs and the count of high-quality cells in the sample); β, β_p, β_s, β_a, β_m and β_c are parameters capturing the effect of these variables on cellular composition; and ε_i are normally distributed residuals. For each SNP, we used Bonferroni correction to adjust for the number cell types and the condition tested and considered $P_{\mathrm{adj}} < 5 \times 10^{-8}$ as genome-wide significant. Winner's curse-adjusted Z-score and R^2 were computed using FDR inverse quantile transformation[79].

Effect of CMV infection on cell composition

We determined the CMV serostatus of AFB ($n = 78$), EUB ($n = 80$) and ASH ($n = 49$) donors with a human anti-IgG CMV ELISA kit (Abcam) on plasma samples, according to the manufacturer's instructions. We assessed the contribution of CMV infection to differences in cellular composition between Africans and Europeans using mediation analysis. Specifically, we used the mediate function of the mediation package of R (v.4.5.0)[80] to model the frequency of each cell type, as a function of population, CMV serostatus and covariates:

$$\mathrm{CellularFrequency}_i = \alpha + \beta \,\mathrm{CMV}_i + \delta \, I_i^{\mathrm{EUB}} + Z_i^T \cdot \boldsymbol{\gamma} + \varepsilon_i \quad (7)$$

$$\mathrm{logit}(\mathrm{Prob}(\mathrm{CMV}_i = 1)) = \alpha' + \delta' I_i^{\mathrm{EUB}} + Z_i^T \cdot \boldsymbol{\gamma}' \quad (8)$$

where $\mathrm{CellularFrequency}_i$ corresponds to the basal state frequency of the cell type under consideration; α and α' are two intercepts; β is the effect of the CMV serostatus (CMV_i) on cellular proportions; δ and δ' are the (direct) effect of population (captured through the indicator variable I_i^{EUB}) on cell type frequency and CMV serostatus; $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ capture the confounding effect of covariates (that is, age and cell mortality) on both gene expression and CMV serostatus; and ε_i are normally distributed residuals. Under this model, we implicitly assumed that the effect of CMV serostatus is the same across populations. Although this assumption cannot be tested due to the lack of CMV^- individuals in the African group, we used an interaction test to evaluate whether the effect of CMV serostatus on cell composition is similar between European and East Asian donors (Supplementary Note 4 and Supplementary Fig. 7). To do so, we defined the following model, with the same notations as before

$$\mathrm{CellularFrequency}_i = \alpha + \beta \,\mathrm{CMV}_i + \delta \, I_i^{\mathrm{ASH}} + \theta \,\mathrm{CMV}_i I_i^{\mathrm{ASH}} + Z_i^T \cdot \boldsymbol{\gamma} + \varepsilon_i \quad (9)$$

and performed a Student's t-test for the null hypothesis that the effect of CMV is the same in Europeans and East Asians ($\mathcal{H}_0 : \theta = 0$).

Modelling population effects on the variation of gene expression

To estimate population effects on gene expression while mitigating any potential batch effect relating to sample processing, we first focused exclusively on AFB and EUB individuals, as all these individuals were recruited during the same sampling campaign and their PBMCs were processed at the same time, with the same experimental procedure[18]. For each immune lineage, cell type, stimulation condition and gene, we then built a separate linear model of the form:

$$\mathrm{Expr}_i = \alpha + \beta_r I_i^{\mathrm{EUB}} + Z_i^T \cdot \boldsymbol{\gamma} + \varepsilon_i \quad (10)$$

where Expr_i is the rank-transformed gene expression (log-normalized CPM) for individual i in the lineage/cell type and condition under consideration; I_i^{EUB} is an indicator variable equal to 1 for European-ancestries individuals and 0 otherwise; and Z_i represents the set of core covariates of the sample that includes the individual's age and cellular mortality (that is, the proportion of dying cells in each thawed vial, as a proxy of sample quality). Moreover, ε_i are the normally distributed residuals and α,β_r,γ are the fitted parameters of the models. In particular, α is the intercept, β_r indicates the $\log_2$-transformed fold change difference in

expression between individuals of European and African ancestries, and γ captures the effects of the set of core covariates on gene expression.

We reasoned that differences in the variance of gene expression between populations might inflate the number of false positives. We therefore used the vcovHC function of sandwich (v.2.5-1)[81] with the Type='HC3' option to compute sandwich estimators of variance that are robust to residual heteroskedasticity. We estimated the β_r coefficients and their standard error with the coeftest function of lmtest (v.0.9-40)[82]. The FDR was calculated across all conditions and lineages using the Benjamini–Hochberg procedure (p.adjust function with the 'fdr' method). Genes with an FDR < 1% and $|\beta_r| > 0.2$ were considered to be differentially expressed between populations (that is, 'raw' popDEGs). We adjusted for cellular composition within each lineage L by introducing into model (10) a set of variables $(F_{j\cdot})_{j\in L}$ encoding the frequency in the PBMC fraction of each cell type j comprising the lineage (for example, naive, effector and regulatory subsets of $CD4^+$ T cells).

$$\mathrm{Expr}_i = \alpha' + \beta_a I_i^{\mathrm{EUB}} + Z_i^T \cdot \boldsymbol{\gamma}' + \sum_{j\in L} \delta_j F_{j,i} + \varepsilon_i \quad (11)$$

The notation is as above, with α, β_a, γ' the fitted parameters of the model. In this model, δ_j is the effect on gene expression of a 1% increase in cell type j and β_a indicates the cell composition-adjusted $\log_2$-transformed fold change in the difference in expression between AFB and EUB individuals. The significance of β_a was calculated as described above, with a sandwich estimator of variance and the coeftest function. FDR was calculated across all conditions and lineages to yield a set of "cell-composition-adjusted" popDEGs. We assessed the impact of cellular composition on differences in gene expression between populations, by defining Student's t-test statistic $T_{\Delta\beta}$ as follows:

$$T_{\Delta\beta} = \frac{\hat{\beta}_a - \hat{\beta}_r}{\mathrm{Var}(\hat{\beta}_a - \hat{\beta}_r)} = \frac{\hat{\beta}_a - \hat{\beta}_r}{\hat{s}_a^2 + \hat{s}_r^2 - 2\rho\hat{s}_a\hat{s}_r} \quad (12)$$

where $\widehat{\beta_r}$ and $\widehat{\beta_a}$ are the raw and cell-composition-adjusted differences in expression between populations; $\hat{s}_r$ and $\hat{s}_a$ are the estimated standard error of $\widehat{\beta_r}$ and $\widehat{\beta_a}$, respectively; and ρ is the observed correlation in permuted data between the $\widehat{\beta_r}$ and $\widehat{\beta_a}$ statistics. Under the null hypothesis that population differences are not affected by cellular composition, $T_{\Delta\beta}$ should follow an approximate Gaussian distribution with mean 0 and variance 1, enabling the definition of a P value $P_{\Delta\beta}$. We then considered the set of raw popDEGs that (1) were not significant after adjustment (FDR > 1% or $|\beta_a| < 0.2$) and (2) displayed significant differences between the raw and adjusted effect sizes ($|T_{\Delta\beta}| > 1.96$) imputable to the effect of cellular composition.

For the assessment of population differences in response to viral stimuli (that is, popDRGs), we used the same approach, but with the replacement of log-normalized counts with the log-fold change difference in expression between the stimulation conditions for each of the two viruses and non-stimulated conditions.

Pathway enrichment analyses

We performed functional assessments of the effects of cellular composition variability on differences in gene expression between donors in the basal state and in response to each virus, using the fgsea R package (v.1.18.1)[83] and the default options. This made it possible to perform a gene set enrichment analysis with population differences in each lineage ranked by the magnitude of the effect of ancestry on the expression or response of the gene before (β_r) and after (β_a) adjustment for differences in cellular composition.

Fine mapping of eQTL

For eQTL mapping, we used variants with MAF > 5% in at least one of the three populations considered, resulting in a set of 10,711,657 SNPs, of which 4,164,060 were located <100 kb from a gene. We used MatrixEQTL (v.2.3)[84] to map eQTLs in a 100 kb region around each gene and obtain estimates of eQTL effect sizes and their standard error. eQTL mapping was performed separately for each immune lineage/cell type and condition, based on rank-transformed gene expression values. eQTL analyses were performed adjusting for population, age, chromosomal sex, cell composition (within each lineage), as well as cell mortality and total number of cells in the sample, and a data-driven number of surrogate variables included to capture unknown confounders and remove unwanted variability. Specifically, for each immune lineage/cell type and condition, surrogate variables were obtained using the sva function from the sva R package (v.3.40.0)[85] with option method='two-steps', providing all other covariates as known confounders (mod argument). The number of surrogate variables to use in each lineage/cell type and condition was determined automatically based on the results from num.sv function with method='be'[85].

For each gene, immune lineage/cell type and stimulation condition, Z-values (that is, the effect size of each eQTL divided by the standard error of effect size) were then calculated, and the fine mapping of eQTLs was performed using SuSiE (v.0.11.42)[86] (susie_rss function of the susieR R package), with a default value of up to 10 independent eQTLs per gene. Imputed genotype dosages were extracted in a 100 kb window around each gene and regressed against the population of origin (that is, AFB, EUB or ASH). Genes with fewer than 50 SNPs in the selected window were discarded from the analysis. Pairwise correlations between the population-adjusted dosages were then assessed, to define the genotype correlation matrix to be used for the fine mapping of eQTLs. In rare cases (<0.1% of tested gene × condition combinations), the susie_rss function did not converge, even when the number of iterations was increased to $>10^6$. These runs were therefore discarded, and the associated eQTLs were assigned a null Z-score during FDR computation (see below). For each eQTL, the index SNP was defined as the SNP with the highest posterior inclusion probability (that is, the α parameter in the output of SuSiE) for that eQTL, and the 95% credible interval was obtained as the minimal set of SNPs S such that $\alpha_s > 0.01$ for all $s \in S$ and $\sum_{s\in S} \alpha_s > 0.95$. Only eQTLs with a log-Bayes factor (lbf) > 3 were considered for further analyses.

For each lineage and set of stimulation conditions, each eQTL identified by SuSiE was assigned an eQTL evidence score, defined as the absolute Z-value of association between the eQTL index SNP and the associated gene. We then used a pooled permutation strategy to define the genome-wide number of significant eQTLs (that is, eQTL × gene combinations) expected under the null hypothesis, for different thresholds of the eQTL evidence score. We repeated the eQTL mapping procedure on the dataset after randomly permuting genotype labels within each population. We then counted, for each possible evidence score threshold T, the number of eQTLs identified in the observed and permuted data. Finally, we retained as a significance threshold the lowest threshold giving several significant eQTLs in the permuted data (false positives) of less than 1% the number of eQTLs identified in the observed data (false positives + true positives).

Aggregation of eQTLs across cell types and stimulation conditions

The eQTL index SNP may differ between cellular states (immune lineage/cell type and stimulation condition), even in the presence of a single causal variant. It is therefore necessary to aggregate eQTLs to ensure that the same locus is tagged by a single variant across cellular states. To this end, we applied the following procedure, for each gene: (1) let C_g be the set of cellular states where a significant eQTL was detected for gene g, and S_g be the associated list of eQTLs (that is, cellular state × index SNP). We aim to define a minimal set of SNPs, M_g, that overlaps the 95% credible intervals of all significant eQTLs in S_g. (2) For each SNP s in a 100-kb window around each gene, compute the expected number of cellular states in which the SNP has a causal effect on gene expression $E[N_c(s)]$ as:

$$E[N_c(s)] = \sum_{j \in C_g} \mathrm{PP}_{sj} \quad (13)$$

where PP_{sj} is the posterior probability that SNP s has a causal effect on the expression of gene g in the cellular state j (cell type × condition). (3) Find the SNP s that maximizes $E[N_c(s)]$, and add it to M_g. (4) Remove from S_g all eQTLs for which the 95% credible interval contains SNP s. (5) Repeat steps 1–3 until S_g is empty.

At the end of this procedure, M_g provides the list of independent eQTL index SNPs (referred to as eSNPs) for gene g, for which we extracted summary statistics across all cellular states.

Mapping of response eQTLs

For the mapping of response eQTLs (reQTLs), we repeated the same procedure as for the mapping of eQTLs, using the rank-transformed $\log_2$-fold change as input rather than gene expression. This included reQTL mapping using MatrixEQTL[84], fine mapping with SuSiE[86], permutation-based FDR computation, and aggregation of reQTL across immune lineages, cell types and stimulation conditions. Surrogate variables were computed directly from $\log_2$-transformed fold changes. For IAV-infected monocytes (detected only in the IAV condition), fold changes were computed relative to $CD14^+$ monocytes in the non-stimulated condition. This produced a list of independent reQTL index SNPs M', like that obtained for eQTLs, for which we extract summary statistics across all cellular states.

Sharing of eQTLs across cell types and stimulation conditions

After extracting the set M of independent eSNPs across all genes, we defined cell-type-specific eQTLs as eQTLs significant genome-wide in a single cell type. We accounted for the possibility that some eQTLs may be missed in specific cell types due to a lack of power by introducing a second definition of eQTL sharing based on nominal P values. Specifically, we considered an eQTL to be cell type-specific at a nominal significance if, and only if, it was significant genome-wide in a single cell type and its nominal P value of association was greater than 0.01 in all other cell types. For each pair of cell types, the correlation of eQTL effect sizes was calculated on the set of all eQTLs passing the nominal significance criterion (Student's t-test, $P < 0.01$) in at least one of the two cell types. To understand how the effect of genetics on immune response varies between SARS-CoV-2 and IAV, we defined an interaction statistic that enabled us to test for differences in reQTL effect size between the two viruses. Specifically, within each immune lineage/cell type, we defined:

$$T_{\mathrm{int}} = \frac{\widehat{\beta}_{\mathrm{IAV}} - \widehat{\beta}_{\mathrm{COV}}}{\mathrm{Var}(\widehat{\beta}_{\mathrm{IAV}} - \widehat{\beta}_{\mathrm{COV}})} = \frac{\widehat{\beta}_{\mathrm{IAV}} - \widehat{\beta}_{\mathrm{COV}}}{\hat{s}^2_{\mathrm{IAV}} + \hat{s}^2_{\mathrm{COV}}} \quad (14)$$

When the reQTL effect size is identical between the two viruses, we expect T_{int} to be normally distributed around 0 with variance 1, allowing to derive an interaction P value. We therefore defined as virus-dependent reQTLs those with a nominal interaction $P < 0.01$ and as virus-specific reQTLs those that passed a nominal P value threshold of 0.01 in only one of the two stimulation conditions.

Comparison of eQTLs and eGenes across studies

To assess the replicability of the eQTLs detected in our study, we compared the eGenes that we identified with those reported in three single-cell studies of resting and stimulated PBMCs[19,87,88]. For each study, we first reassigned each reported cell type to one of the five major lineages that we identified. We then retrieved, for each lineage/condition, the union of all genes with a reported eQTL in at least one of the cell types associated to that lineage, using the following thresholds: ref. 87, $P < 10^{-5}$; ref. 88, FDR < 0.05; ref. 19, lfsr < 0.1. The resulting gene sets were considered as eGenes for each lineage/condition. Enrichments of eQTLs in previously identified eGenes were tested for each study and lineage separately, using a Fisher's exact test to assess whether genes reported to contain an eQTL in a given lineage were more likely to present an eQTL in the same lineage in our study. For each comparison, we used as background sets all genes tested for eQTLs in both studies. When the set of tested genes was not reported in the study, as reported previously[87], we used the union of eQTL genes across all cell types, as a proxy for the set of tested genes. For reQTLs, we compared, for each lineage, genes with a reQTL after IAV stimulation in our study with genes with an eQTL at lfsr < 0.1 specifically after IAV stimulation (but not in non-stimulated cells) in ref. 19. Finally, we compared the direction of effect at shared eQTLs between our study (FDR < 0.01) and that of ref. 19 (lfsr < 0.1), focusing on the eQTL index SNP reported by the latter and assessing the percentage of eQTLs with concordant direction of effect in our data.

To assess the extent to which our findings in PBMCs replicate in the lung, we downloaded Genotype-Tissue Expression (GTEx) lung eQTL data[55] from the eQTL catalogue (uniformly processed summary statistics; http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/GTEx/ge/GTEx_ge_lung.all.tsv.gz). We next considered the index SNP from each eQTL, focusing on the subset of genes with median TPM > 10 in the lung and eQTLs with MAF > 5% in the GTEx dataset. We considered any eQTL with (1) $P < 0.01$ in the lung and (2) the same effect direction between lung and the lineage/cell type/condition in which it is the most significant in our study as replicated. As a comparison, we evaluated the amount of eQTLs that would be replicated when selecting SNPs at random, matching for MAF in GTEx (bins of 5%) and the distance between the eQTL index SNPs and the nearest gene (that is, bins of 0–1, 1–5, 5–10, 10–20, 20–50 and 50–100 kb), and computed the fold-enrichment in replicated eQTLs as the ratio between the observed and expected number of replicated eQTLs.

Mediation analyses

For all popDEGs and popDRGs, we evaluated the proportion of the difference in expression or response to viral stimulation between populations attributable to either genetic factors (that is, eQTLs) or cellular composition, using the mediate function of the mediation R package (v.4.5.0)[80]. Mediation analysis made it possible to separate the differences in expression/response between populations that were mediated by genetics (that is, differences in allele frequency of a given eQTL between populations, ζ_g), or cellular composition (that is, difference in cell type proportions between populations, ζ_c) from those occurring independently of the eQTL/cell type considered (independent or direct effect δ). It was then possible to estimate the respective proportion of population differences mediated by genetics τ_g and cellular composition τ_c as $\tau_c = \frac{\zeta_c}{\zeta_c + \zeta_g + \delta}$ and $\tau_g = \frac{\zeta_g}{\zeta_g + \zeta_c + \delta}$, with $\zeta_c + \zeta_g + \delta$ corresponding to the total differences in expression/response between populations. For each popDEG and popDRG, we focused on either (1) the most strongly associated SNP in a 100 kb window around the gene, regardless of the presence or absence of a significant (r)eQTL, or (2) the cell type differing most strongly between populations in each lineage (that is, $CD16^+$ monocytes in the myeloid lineage, κ-light-chain-expressing memory cells in the B cell lineage, effector cells in $CD4^+$ T cell lineage, $CD8^+$ EMRA cells in the $CD8^+$ T cell lineage and memory cells in the NK cell lineage). For each popDEG and potential mediator M (that is, eQTL SNP or cell subtype proportion), we then ran mediate with the following models:

$$\mathrm{Expr}_i = \alpha + \beta\, M_i + \delta\, I_i^{\mathrm{EUB}} + Z_i^T \cdot \boldsymbol{\gamma} + \varepsilon_i \quad (15)$$

$$M_i = \alpha' + \delta'\, I_i^{\mathrm{EUB}} + Z_i^T \cdot \boldsymbol{\gamma}' + \varepsilon'_i \quad (16)$$

where Expr_i corresponds to normalized expression values in the cell type/condition under consideration; α and α' are two intercepts; β is

the effect of the mediator M_i on gene expression; δ and δ' are the (direct) effect of population (captured through the indicator variable I_i^{EUB}) on gene expression and on the mediator; γ and γ' capture the confounding effect of covariates (that is, age and cell mortality) on both gene expression and the mediator; and ε_i and ε_i' are normally distributed residuals. For popDRGs, we used the same approach, replacing normalized gene expression values with the $\log_2$-transformed fold change in gene expression between the stimulated and unstimulated states.

Detection of signals of natural selection

We avoided SNP ascertainment bias by performing natural selection analyses with high-coverage sequencing data from the 1KG Project[89]. We downloaded the GRCh38 phased genotype files from the New York Genome Center FTP server and calculated the pairwise F_{ST} (ref. 90) between our three study populations (AFB, EUB or ASH) and all 1KG populations to identify the 1KG populations that were the most genetically similar to our study populations. All selection and introgression analyses (see the 'Archaic introgression analyses' section) were based on the Yoruba from Ibadan, Nigeria (YRI), Utah residents with Northern and Western European ancestries (CEU) and Southern Han Chinese (CHS) populations, as these 1KG populations had the lowest F_{ST} values with our three study groups. We filtered the data to include only autosomal biallelic SNPs and insertions/deletions (indels), and removed sites that were invariant (that is, monomorphic) across the three populations. We identified loci presenting signals of positive selection (local adaptation) with the PBS[39], based on the Reynold's F_{ST} estimator[90] between pairs of populations. PBS values were calculated for the YRI, CEU, and CHS populations separately, with the other two populations used as the control and outgroup. For each population, genome-wide PBS values were then ranked, and variants with PBS values within the top 1% were considered to be putative targets of selection. For annotation of the selected eQTLs, the ancestral and derived states at each site were inferred from six-way EPO multiple alignments for six primate species (available from http://ftp.ensembl.org/pub/release-71/emf/ensembl-compara/epo_6_primate/), and the effect size was reported for the derived allele. For sites without an ancestral/derived state in the EPO alignment, the effect of the allele with the lowest frequency worldwide was reported.

We assessed the extent to which different sets of eQTLs displayed signals of local adaptation in permutation-based enrichment analyses. For each population, we compared the mean PBS values at (r)eQTLs for each set of cell type/stimulation condition with the mean PBS values obtained for 10,000 sets of randomly resampled sites. Resampled sites were matched with eQTLs for MAF (mean MAF across the three populations, bins of 0.01), LD scores (quintiles) and distance to the nearest gene (bins of 0–1, 1–5, 10–20, 20–50 and >100 kb). For each population and set of eQTLs, we defined the fold-enrichment (FE) in positive selection as the ratio of observed/expected values for mean PBS and extracted the mean and 95% confidence interval of this ratio across all resamplings. One-sided resampling *P* values were calculated as the number of resamplings with a FE > 1 divided by the total number of resamplings. Resampling *P* values were then adjusted for multiple testing by the Benjamini–Hochberg method.

Detecting and dating episodes of local adaptation

We inferred the frequency trajectories of all eQTLs and reQTLs during the past 2,000 generations (that is, 56,000 years before the present, with a generation time of 28 years), systematically by using CLUES (commit no. 7371b86, 27 May 2021)[40]. We first used Relate (v.1.1.8)[91] on each population separately to reconstruct tree-like ancestral recombination graphs (ARGs) around each SNP in the genome and to estimate effective population sizes across time on the basis of the rate of coalescence events over the inferred ARGs. Using CLUES[40], we then estimated at each eQTL or reQTL, the most likely allele frequency trajectories for each sampled ARG and averaged these trajectories across all possible ARGs.

We then analysed changes in inferred allele frequencies over time to identify selection events characterized by a rapid change in allele frequency (Supplementary Fig. 9a). We considered the posterior mean of allele frequency at each generation and smoothed the inferred allele frequency trajectories by loess regression (with span = 0.1) to ensure progressive changes in allele frequency over time and to minimize the artifacts induced by the inference process. Finally, for each variant and in each population, we calculated the change in allele frequency f at each generation as the difference in the smoothed allele frequency between two consecutive generations:

$$\dot{f}(t) = \frac{\mathrm{d}f}{\mathrm{d}t}(t) = f(t+1) - f(t) \tag{17}$$

Under an assumption of neutrality, the count of a particular allele at generation $t+1$ is the result of a Bernoulli trial parameterized $\mathcal{B}(N, f)$, where N is the size of the haploid population. The variance of allele frequencies at generation $t+1$ is therefore greater for alleles present at higher frequencies in generation t,

$$V[f] = \frac{f(1-f)}{N}. \tag{18}$$

We adjusted for this by scaling the change in allele frequency $\dot{f}$ by a normalizing factor dependent on the allele frequency at generation t, such that the variance of the normalized change in allele frequency $\dot{g}$ was the same across all variants,

$$\dot{g} = \frac{\dot{f}}{\sqrt{f(1-f)}} \tag{19}$$

Finally, at each generation, we divided the normalized change in allele frequency $\dot{g}$ by its s.d. across all eQTLs and reQTLs, to calculate a Z-score for detecting alleles for which the normalized change in allele frequency exceeded genome-wide expectations,

$$Z = \frac{\dot{g}}{\mathrm{s.d.}(\dot{g})} \tag{20}$$

For each variant and generation, we then considered an absolute Z-score > 3 to constitute evidence of selection and we inferred the onset of selection of a variant as the first generation in which $|Z| > 3$.

Simulations, power and type I error estimates

We assessed the ability of our approach to detect (and date) events of natural selection correctly from the trajectories of allele frequencies by using simulations with SLiM (v.4.0.1)[92] under various selection scenarios. Simulations were performed under a Wright–Fisher model for a single mutation occurring around 5,000 generations ago, at a frequency varying from $f_{min} = \frac{1}{N}$ to $f_{max} = 1 - \frac{1}{N}$ in steps of 1%, where N is the simulated population size. We allowed population size to vary over time according to published estimates[91] for the YRI, CEU and CHS populations (Supplementary Fig. 9b). We then performed simulations both under an assumption of neutrality (1,000 simulations for each starting frequency) and assuming a 200-generation-long episode of selection (100 simulations for each starting frequency and selection scenario). Selection episodes were simulated with an onset of selection 1,000, 2,000, 3,000 or 4,000 generations ago, and with a selection coefficient ranging from 0.01 to 0.05 (Supplementary Fig. 9c). We saved computation time by performing a tenfold scaling in line with SLiM recommendations. For each selected scenario and variant, simulated allele frequencies were retrieved every ten generations, and smoothed using loess regression with a span of 0.1. We then calculated

normalized differences in smoothed allele frequencies for each simulated variant and scaled these differences at each generation, on the basis of their s.d. among neutral variants, to obtain *Z*-scores. For each selection scenario, we focused on the centre of the selection interval and determined the type I error and power for various thresholds of absolute *Z*-scores varying from 0 to 6. We found that a threshold of 3 yielded both a low type I error (<0.2% false positives) and a satisfactory power for detecting selection events (Supplementary Fig. 9c). Finally, at each generation, we estimated the percentage of simulations, under an assumption of neutrality or a particular selection scenario, for which the absolute *Z*-score exceeded a threshold of 3. We found that significant *Z*-scores were equally rare at each generation under the assumption of neutrality, but that selected variants presented a clear and localized enrichment in significant *Z*-scores for intervals in which we simulated selection (Supplementary Fig. 9d).

Archaic introgression analyses

For the definition of regions of the modern human genome of archaic ancestry (Neanderthal or Denisovan), we downloaded the VCFs from the high-coverage Neanderthal Vindija[93] and Denisovan Altai[94] genomes (human genome assembly GRCh37; http://cdna.eva.mpg.de/neandertal/Vindija/) and applied the corresponding genome masks (FilterBed files). We then removed sites within segmental duplications and lifted over the genomic coordinates to the GRCh38 assembly with CrossMap (v.0.6.3)[95]. We used two statistics to identify introgressed regions in the CEU and CHS populations: (1) conditional random fields (CRF)[96,97], which uses reference archaic and outgroup genomes to identify introgressed haplotypes; and (2) the *S′* method[98], which identifies stretches of probably introgressed alleles without requiring the definition of an archaic reference genome.

For CRF-based calling, we phased the data with SHAPEIT4 (v.4.2.1)[60], using the recommended parameters for sequence data, and focused on biallelic SNPs for which the ancestral/derived state was unambiguously defined. We then performed two independent runs of CRF to detect haplotypes inherited from Neanderthal or Denisova. For Neanderthal-introgressed haplotypes, we used the Vindija Neanderthal genome as the archaic reference and YRI individuals merged with the Altai Denisovan genome as the outgroup. For Denisovan-introgressed haplotypes, we used the Altai Denisovan genome as the archaic reference panel and YRI individuals merged with the Vindija Neanderthal genome as the outgroup. The results from the two independent CRF runs were analysed jointly, and we retained alleles with a marginal posterior probability $P_{\text{Neanderthal}} \geq 0.9$ and $P_{\text{Denisova}} < 0.5$ as Neanderthal-introgressed haplotypes and those containing alleles with $P_{\text{Denisova}} \geq 0.9$ and $P_{\text{Neanderthal}} < 0.5$ as Denisovan-introgressed haplotypes. For the *S′*-based calling of introgressed regions, we considered all biallelic SNPs with an allele frequency of <1% in the YRI population to be Eurasian-specific alleles. We then ran the Sprime (v.07Dec18.5e2; https://github.com/browning-lab/sprime) separately for the CEU and CHS populations to identify and score putatively introgressed regions containing a high density of Eurasian-specific alleles. Putatively introgressed regions with a *S′* score of >150,000 were considered to be introgressed. This cut-off score has been shown to provide a good trade-off between power and accuracy based on simulations of introgression under realistic demographic scenarios[98]. For both calling methods (that is, CRF and *S′*), we used the recombination map from the 1KG Project Phase 3 data release[56].

After the calling of introgressed regions throughout the genome for each population, we defined SNPs of putative archaic origin (archaic SNPs, aSNPs) as those (1) located in an introgressed region defined by either the CRF or *S′* method; (2) with one of their alleles being rare or absent (MAF < 1%) in the YRI population, but present in the Vindija Neanderthal or Denisovan Altai genomes; and (3) in high LD ($r^2 > 0.8$) with at least two other aSNPs and, to exclude incomplete lineage sorting, comprising an LD block of >10 kb. This yielded a set of 100,345 high-confidence aSNPs (Supplementary Table 8a). We further categorized aSNPs as of Neanderthal origin, Denisovan origin or shared origin according to their presence/absence in the Vindija Neanderthal and Denisovan Altai genomes. Finally, we considered any site that was in high LD with at least one aSNP in the same population in which introgression was detected to be introgressed, and classified introgressed haplotypes as of Neanderthal origin, Denisovan origin or shared origin according to the most frequent origin of aSNPs in the haplotype. For introgressed SNPs, we defined the introgressed allele as (1) the allele rare or absent from individuals of African ancestries if the SNP was an aSNP; and (2) for non-aSNPs, the allele most frequently segregating with the introgressed allele of linked aSNPs. In each population, introgressed alleles with a frequency in the top 1% for introgressed alleles genome-wide were considered to present evidence of adaptive introgression.

The enrichment of introgressed haplotypes in eQTLs or reQTLs was assessed separately for each population (CEU and CHS), first by stimulation condition and then by cell type within each condition. To avoid biases related to an increased power for the detection of eQTLs that segregate at higher frequencies in European genomes, (that is, $n_{\text{EUB}} = 80$ and $n_{\text{ASH}} = 62$), we considered a set of $n = 10{,}276$ eQTLs mapped on a downsampled dataset composed of the same number of individuals from each population (EUB and ASH) within each cell type and condition. This downsampled set of eQTLs was highly concordant with the original eQTL mapping (that is, >95% sharing at the lineage level). Within each cell type/stimulation condition, we considered the set of all (r)eQTLs for which the index SNP displayed at least a marginal association (Student's *t*-test, $P < 0.01$) with gene expression. For each population and (r)eQTL set, we then grouped (r)eQTLs in high LD ($r^2 > 0.8$), retaining a single representative per group, and counted the total number of (r)eQTLs for which the index SNP was in LD ($r^2 > 0.8$) with an aSNP (that is, introgressed eQTLs). We then used PLINK (v.1.9) --indep-pairwise (with a 500 kb window, 1 kb step, an r^2 threshold of 0.8, and a MAF > 5%)[57] to define tag-SNPs for each population, and we determined the expected number of introgressed SNPs by resampling tag-SNPs at random with the same distribution for MAF, LD scores and distance to the nearest gene. We performed 10,000 resamplings for each (r)eQTL set and population. One-sided resampling-based *P* values were calculated as the frequency at which the number of introgressed SNPs among resampled SNPs exceeded the number of introgressed SNPs among (r)eQTLs. Resampling-based *P* values were then adjusted for multiple testing using the Benjamini–Hochberg method.

We searched for signals of adaptive introgression by determining whether introgressed haplotypes that altered gene expression were introgressed at a higher frequency than introgressed haplotypes with no effect on gene expression. For each stimulation cell type/condition, we focused on the set of introgressed eQTLs segregating with a MAF > 5% in each population (retaining a single representative per LD group) and compared the frequency of the introgressed allele with that of introgressed tag-SNPs genome-wide. We modelled $r_{(\text{Freq})}$, the (rank-transformed) frequency of introgressed tag-SNPs according to the presence/absence of a linked eQTL (I_{eQTL}), and the mean MAF of the SNP across the three populations (giving a higher power for eQTL detection).

$$r_{(\text{Freq})} \approx \alpha + \beta\, I_{\text{eQTL}} + \gamma\, \overline{\text{MAF}} \qquad (21)$$

where I_{eQTL} is an indicator variable equal to 1 if the SNP is in LD with an eQTL ($r^2 > 0.8$) and 0 otherwise; $\overline{MAF}$ is the mean MAF calculated separately for each population; α is the intercept of the model; β measures the difference in rank $r_{(\text{Freq})}$ between eQTLs and non eQTLs; and γ is a nuisance parameter capturing the relationship between $\overline{MAF}$ and $r_{(\text{Freq})}$. Under this model, the difference in frequency between eQTLs and non-eQTLs can be tested directly in a Student's *t*-test with $\mathcal{H}_0 : \beta = 0$.

Enrichment in COVID-19-associated loci and colocalization analyses

We downloaded summary statistics from the COVID-19 Host Genetics Initiative (release 7: https://www.covid19hg.org/results/r7)[8] for three GWASs: (1) A2—very severe respiratory cases of confirmed COVID-19 versus the general population; (2) B2—hospitalized COVID-19 cases versus the general population; (3) C2—confirmed COVID-19 cases versus the general population. We assessed the enrichment in eQTLs and reQTLs of COVID-19 susceptibility/severity loci by considering, for each eQTL/reQTL, the A2, B2 or C2 *P* values of the index SNP and calculating the percentage of eQTLs/reQTLs with a significant GWAS *P* value of 10^{-4}. This percentage was then compared to that obtained for the resampled set of SNPs, matched for distance to the nearest gene (bins of 0–1, 1–5, 5–10, 10–20, 20–50 and 50–100 kb) and MAF (1% MAF bins). We performed 10,000 resamplings for each set of eQTLs/reQTLs tested. The use of different *P*-value thresholds for COVID-19-associated hits (10^{-3} to 10^{-5}) yielded similar results. Note that, despite the strong overlap (OR > 200, Fisher's test, $P = 4.2 \times 10^{-40}$) between loci associated with susceptibility (C2) and severity (A2 or B2)[8], 81 out of 105 COVID-19 associated eQTLs (at nominal $P < 10^{-4}$) are associated specifically with either susceptibility ($n = 19$) or severity ($n = 62$), supporting the relevance of considering these traits separately in our analysis.

To identify specific eQTLs/reQTLs colocalized with GWAS hits, we first considered all (r)eQTLs for which the index SNPs were located within 100 kb of a SNP associated with COVID-19 susceptibility/severity ($P < 10^{-5}$). For each immune lineage/cell type, and condition for which the eQTL/reQTL reached genome-wide significance, we next extracted all SNPs in a 500 kb window around the index SNP for which summary statistics were available for both the eQTLs/reQTLs and COVID-19 GWAS phenotypes (A2, B2 and C2) and performed a colocalization test using the coloc.signals function of the coloc (v.5.1.0) R package. We set a prior probability for colocalization p_{12} of 10^{-5} (that is, the recommended default value). Any pair of (r)eQTL/COVID-19 phenotypes with a posterior probability for colocalization $PP_{H4} > 0.8$ was considered to display significant colocalization.

Transcriptome-wide association tests

Using the summary statistics from the COVID-19 Host Genetics Initiative[8], we applied the S-PrediXcan framework (v.0.6.11)[99] to leverage our genotype-expression dataset and identify associations between genotypes and COVID-19 traits that could be mediated by the regulation of gene expression. These analyses were conducted separately in each of the 5 lineages or 22 cell types and the 3 experimental conditions of our setting.

To perform these transcriptome-wide association tests (across the 12,655 human genes of our final dataset), we proceeded in two steps. First, we used the pseudobulk gene expression levels detected in each cell type/lineage and condition, together with the associated genotypes, from each of the 222 donors to build reference transcriptome datasets. We then trained elastic net regression models on these references to estimate the effect on gene expression of each SNP within a 100 kb window around each gene. These models were of the form:

$$\mathrm{Expr}_{ij} = \alpha + X_{ij}^T \cdot \boldsymbol{w}_j + Z_i^T \cdot \boldsymbol{\gamma} + \varepsilon_i \quad (22)$$

where Expr_{ij} is the rank-transformed expression (log-normalized CPM) of gene *j* for individual *i* in the lineage/cell type and condition under consideration; α is an intercept; X_{ij} are the genotypes of common variants in a 100 kb window around gene *j*; Z_i represents the set of core covariates of the sample that includes the individual's age and population of origin, the cellular mortality of the sample and the frequency in the PBMC fraction of each cell type *k* comprising the lineage. Moreover, $\mathbf{w}_j$ and $\boldsymbol{\gamma}$ are parameter vectors capturing the effect of genotypes and covariates, and the ε_i are normally distributed residuals.

We followed the S-PrediXcan pipeline[99] using the regression coefficients w_j as weights to predict the association between the genetically controlled expression (GReX) of each gene *j* (given by $\mathrm{GReX}_{ij} = \alpha + X_{ij}^T \cdot \boldsymbol{w}_j$) and the trait of interest. Specifically, we combined these weights with SNP covariances calculated from our data to approximate *Z*-scores of association with COVID-19-trait as

$$Z_j^{\mathrm{TWA}} \approx \sum_l w_{lj} \frac{\widehat{\sigma}_l}{\widehat{\sigma}_j} \frac{\widehat{\beta}_l}{\mathrm{s.e.}(\widehat{\beta}_l)} \quad (23)$$

where the Z_j^{TWA} statistic measures the association between gene *j*'s GReX and each of the three COVID-19 traits; w_{lj} is the weight of SNP *l* in the prediction of gene *j*'s expression, $\widehat{\sigma}_l$ and $\widehat{\sigma}_j$ are, respectively, the estimated variances of the SNP and the predicted gene expression, and $\widehat{\beta}_l$ and $\mathrm{s.e.}(\widehat{\beta}_l)$ are the effect size estimated by each GWAS for SNP *l* and its standard error, respectively.

Statistical analyses

Unless explicitly specified, all statistical tests are two-sided and based on measurements from independent samples.

Inclusion and ethics

The current research project builds on samples collected in Ghent (Belgium) and Hong-Kong SAR (China) and has been conducted in collaboration with local researchers. Roles and responsibilities were agreed among collaborators ahead of the research. Research conducted in this study is relevant to local participants and has been reviewed by local ethics committees (committee of Ghent University, Belgium, B670201214647; Institutional Review Board of the University of Hong Kong, UW 20-132), and the relevant French authorities (CPP, CCITRS and CNIL). This study was also monitored by the Ethics Board of Institut Pasteur (EVOIMMUNOPOP-281297). All manipulations of live viruses were performed in a high-security BSL-3 environment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The scRNA-seq data generated and analysed in this study have been deposited in the Institut Pasteur data repository, OWEY, which is available online (https://doi.org/10.48802/owey.e4qn-9190). The genome-wide genotyping data generated or used in this study have been deposited in OWEY and can be accessed online (https://doi.org/10.48802/owey.pyk2-5w22). In accordance with the General Data Protection Regulation (GDPR) in force in the European Union, the aforementioned data can be accessed only from the institutional data repository after authorization by the relevant Data Access Committee (DAC). The DAC ensures that data access and use is authorized for academic research relating to the variability of the human immune response, as defined in the informed consent signed by research participants. COVID-19 GWAS summary statistics used in the present study can be downloaded from https://www.covid19hg.org/results/r7. Human (1000G data, low (phase 3) and high (NYGC) coverage), archaic (Vindija and Denisova) and ancestral (EPO6) genomes used can be retrieved from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502 (1000G phase 3), https://www.internationalgenome.org/data-portal/data-collection/30x-grch38 (1000G high coverage), http://cdna.eva.mpg.de/neandertal/Vindija/ (archaic) and http://ftp.ensembl.org/pub/release-71/emf/ensembl-compara/epo_6_primate/ (EPO6). Uniformly processed summary statistics from GTEx lung tissue were downloaded from http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/ (GTEx/lung/ge/all: study_id: QTS000015, dataset_id: QTD000271, file: QTD000271.all.tsv.gz). Source data are provided with this paper.

Inborn errors of OAS–RNase L in SARS-CoV-2–related multisystem inflammatory syndrome in children

Pathological drivers of immune variability. The widespread variability in COVID-19 courses is a clear illustration of the 'infection enigma' coined by Casanova and Abel (2013, 2020): while SARS-CoV-2 infection is generally asymptomatic, it can lead to lethal pneumonia in some cases. Previous work from Casanova and Abel's groups at The Rockefeller University and Institut Imagine revealed inborn errors of immunity (IEIs) and auto-antibodies targeting type I IFN signalling, which can jointly explain up to 20% of severe COVID-19 pneumonia cases in patients over 70 years old (Zhang et al., 2020; Bastard et al., 2020, 2021a) (§ 3.3, page 67).

In this context, Lee et al. (2022) hypothesized that IEIs could also contribute to some cases of multisystem inflammatory syndrome in children (MIS-C), another severe phenotype related to SARS-CoV-2 infection featuring generalized inflammation across various organs. Importantly, though exacerbated monocyte activation and clonal expansion of some T cell subsets are key features of MIS-C, the cellular and molecular causes of this syndrome were still unknown at the time.

Through whole-genome or whole-exome sequencing of 558 COVID-19 patients with MIS-C, Lee et al. (2022) found IEIs in the 2'-5'-oligoadenylate synthetase (OAS)-RNase L pathway in five children between 4 months and 14 years old. *OAS1*, *OAS2* and *OAS3* are ISGs that encode sensors of viral nucleic acids, and are thus key players of the IFN-mediated antiviral response. Upon sensing of viral doubled-stranded RNA, the OAS enzymes produce 2'-5'-linked oligoadenylates, which activate the RNase L enzyme that catalyzes RNA degradation.

After functionally validating the OAS-RNase L loss-of-functions in these patients, Lee et al. (2022) identified myeloid cells as the highest transcribers of *OAS1*, *OAS2* and *OAS3* and *RNASEL* using bulk transcriptomics, and characterized exaggerated inflammatory cytokine responses to double-stranded RNA in monocyte and macrophage cell lines with OAS-RNase L knock-outs.

To fully characterize the innate response to a live virus strain across multiple immune cell types at single-cell resolution, I performed scRNA-seq on PBMC samples from four OAS-RNase-L-deficient MIS-C patients and from healthy controls, exposed to SARS-CoV-2 or non-infectious medium for six hours. Overall, I detected a widespread effect of OAS-RNase L deficiency on the transcriptional response to SARS-CoV-2 across all PBMC types—significantly affecting the response of around 48% to 94% of differentially expressed genes in each cell type—with myeloid cells being the most affected. Remarkably, myeloid responses to SARS-CoV-2 included a private inflammatory component driven by *IL1B* and *CCL3*, and were significantly stronger in cells from MIS-C patients. These differences were functionally translated into stronger inflammatory responses, tumor necrosis factor production and IFN production, as well as stronger IFN-mediated responses, respectively in OAS-RNase-L-deficient myeloid and $CD4^+$ T cells.

Together, these results support a role for exaggerated myeloid-driven inflammation—owing to a dysregulation of inflammatory cytokine production due to impaired RNase L activity—in the pathogenesis of MIS-C in children with inborn OAS-RNase L deficiencies (§ 3.3.1, page 67). ■

RESEARCH ARTICLE

CORONAVIRUS

Inborn errors of OAS–RNase L in SARS-CoV-2–related multisystem inflammatory syndrome in children

Danyel Lee[1,2,3], Jérémie Le Pen[4]†, Ahmad Yatim[1]†, Beihua Dong[5]†, Yann Aquino[6,7]†, Masato Ogishi[1]†, Rémi Pescarmona[8]†, Estelle Talouarn[2,3]†, Darawan Rinchai[1]‡, Peng Zhang[1]‡, Magali Perret[8]‡, Zhiyong Liu[1], Iolanda Jordan[9,10,11,12,13], Sefika Elmas Bozdemir[14], Gulsum Iclal Bayhan[15], Camille Beaufils[16], Lucy Bizien[2,3], Aurelie Bisiaux[6], Weite Lei[1], Milena Hasan[17], Jie Chen[1], Christina Gaughan[5], Abhishek Asthana[5], Valentina Libri[17], Joseph M. Luna[4,18], Fabrice Jaffré[19], H.-Heinrich Hoffmann[4], Eleftherios Michailidis[4,20], Marion Moreews[21], Yoann Seeleuthner[2,3], Kaya Bilguvar[22,23], Shrikant Mane[24], Carlos Flores[25,26,27], Yu Zhang[29,30], Andrés A. Arias[1,31,32], Rasheed Bailey[1], Agatha Schlüter[33], Baptiste Milisavljevic[1], Benedetta Bigio[1], Tom Le Voyer[2,3], Marie Materna[2,3], Adrian Gervais[2,3], Marcela Moncada-Velez[1], Francesca Pala[29], Tomi Lazarov[34], Romain Levy[2,3], Anna-Lena Neehus[2,3], Jérémie Rosain[2,3], Jessica Peel[1], Yi-Hao Chan[1], Marie-Paule Morin[16], Rosa Maria Pino-Ramirez[35], Serkan Belkaya[36], Lazaro Lorenzo[1], Jordi Anton[12,37,38], Selket Delafontaine[39], Julie Toubiana[40,41], Fanny Bajolle[42], Victoria Fumadó[10,12,43,44], Marta L. DeDiego[45], Nadhira Fidouh[46], Flore Rozenberg[47], Jordi Pérez-Tur[48,49,50], Shuibing Chen[19], Todd Evans[19], Frédéric Geissmann[34], Pierre Lebon[51], Susan R. Weiss[52], Damien Bonnet[42], Xavier Duval[53,54,55,56], CoV-Contact Cohort§, COVID Human Genetic Effort¶, Qiang Pan-Hammarström[57], Anna M. Planas[58,59], Isabelle Meyts[60], Filomeen Haerynck[61], Aurora Pujol[62,63], Vanessa Sancho-Shimizu[64,65], Clifford L. Dalgard[66,67], Jacinta Bustamante[1,2,3,68], Anne Puel[1,2,3], Stéphanie Boisson-Dupuis[1,2,3], Bertrand Boisson[1,2,3], Tom Maniatis[69], Qian Zhang[1,2,3], Paul Bastard[1,2,3,70], Luigi Notarangelo[29], Vivien Béziat[1,2,3], Rebeca Perez de Diego[71,72], Carlos Rodriguez-Gallego[28,73], Helen C. Su[29,30], Richard P. Lifton[24,74], Emmanuelle Jouanguy[1,2,3], Aurélie Cobat[1,2,3]#, Laia Alsina[10,12,38,75]#, Sevgi Keles[76]#, Elie Haddad[77]#, Laurent Abel[1,2,3]**, Alexandre Belot[21,78]**, Lluis Quintana-Murci[6,79]**, Charles M. Rice[4]**, Robert H. Silverman[5]**, Shen-Ying Zhang[1,2,3]*††, Jean-Laurent Casanova[1,2,3,80,81]††

Multisystem inflammatory syndrome in children (MIS-C) is a rare and severe condition that follows benign COVID-19. We report autosomal recessive deficiencies of *OAS1*, *OAS2*, or *RNASEL* in five unrelated children with MIS-C. The cytosolic double-stranded RNA (dsRNA)–sensing OAS1 and OAS2 generate 2′-5′-linked oligoadenylates (2-5A) that activate the single-stranded RNA–degrading ribonuclease L (RNase L). Monocytic cell lines and primary myeloid cells with OAS1, OAS2, or RNase L deficiencies produce excessive amounts of inflammatory cytokines upon dsRNA or severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) stimulation. Exogenous 2-5A suppresses cytokine production in OAS1-deficient but not RNase L–deficient cells. Cytokine production in RNase L–deficient cells is impaired by MDA5 or RIG-I deficiency and abolished by mitochondrial antiviral-signaling protein (MAVS) deficiency. Recessive OAS–RNase L deficiencies in these patients unleash the production of SARS-CoV-2–triggered, MAVS-mediated inflammatory cytokines by mononuclear phagocytes, thereby underlying MIS-C.

Interindividual clinical variability in the course of primary infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is immense in unvaccinated individuals (*1*–*4*). We have shown that inborn errors of type I interferon (IFN) immunity and their phenocopies—autoantibodies against type I IFNs—collectively underlie at least 15% of cases of critical COVID-19 pneumonia in unvaccinated patients (*5*–*9*). Common genetic variants act as more modest risk factors (*10*–*13*). Children were initially thought to be rarely affected by COVID-19, as only 0.001 to 0.005% of infected children had critical pneumonia (*2*). However, another severe SARS-CoV-2–related phenotype, multisystem inflammatory syndrome in children (MIS-C), occurs predominantly in children, typically 4 weeks after infection (*14*–*16*). Its prevalence is estimated at ~1 per 10,000 infected children (*17*–*19*). Children with MIS-C do not suffer from hypoxemic pneumonia and typically display no detectable viral infection of the upper respiratory tract at disease onset. However, most MIS-C cases test positive for anti–SARS-CoV-2 antibodies, and almost all cases have a history of exposure to SARS-CoV-2 (*17*, *20*). Initial reports of MIS-C described it as an atypical form of Kawasaki disease (KD) (*16*, *21*–*25*), as its clinical features include fever, rash, abdominal pain, myocarditis, lymphadenopathy, coronary aneurysm, and elevated biological markers of acute inflammation.

The elevated markers frequently detected in MIS-C patients suggest that inflammation occurs in various organs (*21*, *22*, *26*–*36*). These markers include surrogates of cardiovascular endothelial injury [e.g., troponin and B-type natriuretic peptide (BNP)] and gastrointestinal epithelial injury [e.g., lipopolysaccharide (LPS)–binding protein (LBP) and soluble CD14] (*36*). Various leukocyte subsets are also affected. Sustained monocyte activation has been consistently reported as a key immunological feature of MIS-C, with high levels of proinflammatory markers, including ferritin, interleukin-1 receptor antagonist (IL-1RA), IL-6, IL-10, IL-18, monocyte chemoattractant protein 1 (MCP1, or CCL2), and tumor necrosis factor (TNF) (*21*, *22*, *26*–*36*). In addition, the levels of biomarkers related to type II IFN (IFN-γ) signaling, which are not necessarily specific to monocyte activation, often increase during the early phase of disease (*22*, *31*–*36*). An immunological phenotype specific to MIS-C, observed in ~75% of patients, is the polyclonal expansion of $CD4^+$ and $CD8^+$ T cells bearing the Vβ21.3 segment (*32*, *34*, *36*–*38*). In this multitude of molecular, cellular, and clinical abnormalities, the root cause of MIS-C remains unknown (*39*). We hypothesized that monogenic inborn errors of immunity (IEIs) to SARS-CoV-2 may underlie MIS-C in some children and that the identification of these inborn errors may clarify the molecular, cellular, and immunological basis of disease (*15*, *40*).

Results

Identification of homozygous rare predicted loss-of-function variants of OAS1 *or* RNASEL *in two MIS-C patients*

We performed whole-exome or whole-genome sequencing for 558 patients with MIS-C from the international COVID Human Genetic Effort (CHGE) cohort (https://www.covidhge.com/) (fig. S1). We first searched for homozygous or hemizygous rare predicted loss-of-function (pLOF) variants with a high degree of confidence in human genes with a gene damage index of <13.83 (*41*). We then restricted the list to genes involved in host response to viruses (Gene Ontology term "response to virus," GO:0009615). We identified two unrelated patients homozygous for stop-gain variants, of *OAS1* in one patient (P1, p.R47*) and *RNASEL* in the other (P5, p.E265*) (Fig. 1A, fig. S2A, and Table 1). OAS1 (2′-5′-oligoadenylate synthetase 1) is one of the four members of the OAS family (OAS1, OAS2, OAS3, and the catalytically inactive OASL). These proteins are type I IFN–inducible cytosolic proteins that produce 2′-5′-linked oligoadenylates (2-5A) upon binding to double-stranded RNA (dsRNA). The 2-5A, in turn, induce the dimerization and activation of the latent endoribonuclease RNase L, which degrades single-stranded RNA (ssRNA) of viral or human origin (*42*, *43*). No homozygous variants fulfilling these criteria were identified in any of the 1288 subjects with asymptomatic or mild SARS-CoV-2 infection (SARS-CoV-2–infected controls) in the CHGE database (fig. S1). MIS-C patients therefore display significant

enrichment (P = 0.013) in homozygous pLOF variants of the *OAS1* and *RNASEL* genes, suggesting that these loci are probably relevant to MIS-C pathogenesis. Moreover, although OAS1, OAS2, OAS3, and RNase L are expressed in various cell types in mice and humans, their levels are particularly high in myeloid cells, including monocytes and macrophages (*44*–*46*). Thus, autosomal recessive (AR) deficiencies of the OAS–RNase L pathway may underlie MIS-C by impairing the restriction of viral replication and/or enhancing the virus-triggered inflammatory response in monocytes, macrophages, dendritic cells, or other cell types.

Identification of biallelic rare experimentally deleterious variants of OAS1, OAS2, *or* RNASEL *in five MIS-C patients*

OAS1, *OAS2*, *OAS3*, and *RNASEL* have consensus negative selection (CoNeS) scores for negative selection of 2.25, 0.79, 1.46, and 0.66, respectively, consistent with findings for known monogenic IEIs with an AR mode of inheritance (*47*). We therefore extended our search to all homozygous or potential compound-heterozygous nonsynonymous or essential splicing site variants with a minor allele frequency (MAF) of <0.01 at these four loci in our MIS-C cohort. We identified a total of 12 unrelated patients and 16 different variants of *OAS1*, *OAS2*, *OAS3*, and *RNASEL* (Table 1). To study the expression and function of these 16 variants in vitro, we first analyzed RNase L–mediated ribosomal RNA (rRNA) degradation after the cotransfection of RNase L–deficient HeLa M cells with the corresponding *OAS1*, *OAS2*, *OAS3*, or *RNASEL* cDNAs (*48*–*51*) (Fig. 1, B to D, and fig. S2, B and C). The p.R47* OAS1 (homozygous in P1) mutant protein was not produced and was LOF (Fig. 1B and fig. S2D). The three mutant OAS2 proteins detected (p.R535Q, p.Q258L, and p.V290I) were produced in normal amounts, but p.R535Q (homozygous in P2 and P3) had minimal activity, and p.Q258L and p.V290I (both found in P4) had lower levels of activity than the wild-type (WT) protein (Fig. 1, A and C). All the *OAS3* variants were produced in normal amounts, and all but one (p.R932Q, found in the heterozygous state in one patient) of these variants had normal levels of activity (fig. S2C). The *RNASEL* p.E265* variant (homozygous in P5) was expressed as a truncated protein and was LOF (Fig. 1D and fig. S2E), whereas the p.I264V variant was neutral in expression and function (Fig. 1D). We also quantified the function of the OAS1 and OAS2 mutants in a fluorescence resonance energy transfer (FRET) assay, which confirmed that P1's *OAS1* variant was LOF and that the *OAS2* variants of P2, P3, and P4 were hypomorphic (21 to 43%, 32 to 76%, and 36 to 75% of WT OAS2 activity for p.Q258L, p.V290I, and p.R535Q, respectively) (Fig. 1, E and F). Thus, we identified five unrelated MIS-C patients homozygous or compound heterozygous for rare and deleterious alleles of three of the four genes controlling the OAS–RNase L pathway (Fig. 1A and fig. S2A). The patients' genotypes were confirmed by Sanger sequencing and familial segregation. Their clinical and immunological features were consistent with those previously reported for other MIS-C patients (*21*, *22*, *26*–*36*, *52*) (Fig. 1, G to I, and Table 2).

Enrichment in homozygous deleterious OAS1, OAS2, *and* RNASEL *variants in MIS-C patients*

We found no homozygous rare (MAF < 0.01) deleterious variants of the three genes in the

[1]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA. [2]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Paris, France. [3]Paris City University, Imagine Institute, Paris, France. [4]Laboratory of Virology and Infectious Disease, The Rockefeller University, New York, NY, USA. [5]Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [6]Human Evolutionary Genetics Unit, Institut Pasteur, Paris City University, CNRS UMR 2000, Paris, France. [7]Doctoral College, Sorbonne University, Paris, France. [8]Laboratory of Immunology, Lyon Sud Hospital, Lyon, France. [9]Pediatric Intensive Care Department, Hospital Sant Joan de Déu, Barcelona, Spain. [10]Kids Corona Platform, Barcelona, Spain. [11]Center for Biomedical Network Research on Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain. [12]Department of Surgery and Surgical Specializations, Faculty of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain. [13]Respiratory and Immunological Dysfunction in Pediatric Critically Ill Patients, Institute of Recerca Sant Joan de Déu, Barcelona, Spain. [14]Bursa City Hospital, Bursa, Turkey. [15]Ankara City Hospital, Yildirim Beyazit University, Ankara, Turkey. [16]Immunology and Rheumatology Division, Department of Pediatrics, University of Montreal, CHU Sainte-Justine, Montreal, QC, Canada. [17]Center for Translational Research, Institut Pasteur, Paris City University, Paris, France. [18]Department of Biochemistry and Center for RNA Science and Therapeutics, Case Western Reserve University, Cleveland, OH, USA. [19]Department of Surgery, Weill Cornell Medical College, New York, NY, USA. [20]Department of Pediatrics, School of Medicine, Emory University, Atlanta, GA, USA. [21]International Center of Infectiology Research (CIRI), University of Lyon, INSERM U1111, Claude Bernard University, Lyon 1, CNRS, UMR5308, ENS of Lyon, Lyon, France. [22]Departments of Neurosurgery and Genetics and Yale Center for Genome Analysis, Yale School of Medicine, New Haven, CT, USA. [23]Department of Medical Genetics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey. [24]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA. [25]Research Unit, Nuestra Señora de la Candelaria University Hospital, Santa Cruz de Tenerife, Spain. [26]Genomics Division, Institute of Technology and Renewable Energies (ITER), Granadilla de Abona, Spain. [27]CIBERES, ISCIII, Madrid, Spain. [28]Department of Clinical Sciences, University Fernando Pessoa Canarias, Las Palmas de Gran Canaria, Spain. [29]Laboratory of Clinical Immunology and Microbiology, Division of Intramural Research, NIAID, NIH, Bethesda, MD, USA. [30]NIAID Clinical Genomics Program, NIH, Laboratory of Clinical Immunology and Microbiology, Division of Intramural Research, NIAID, NIH, Bethesda, MD, USA. [31]Primary Immunodeficiencies Group, University of Antioquia (UdeA), Medellin, Colombia. [32]School of Microbiology, University of Antioquia (UdeA), Medellin, Colombia. [33]Neurometabolic Diseases Laboratory, IDIBELL–Hospital Duran I Reynals, CIBERER U759, ISIiii, Madrid, Spain. [34]Immunology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [35]Pediatrics Department, Hospital Sant Joan de Déu, Barcelona, Spain. [36]Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey. [37]Pediatric Rheumatology Division, Hospital Sant Joan de Déu, Barcelona, Spain. [38]Study Group for Immune Dysfunction Diseases in Children (GEMDIP), Institute of Recerca Sant Joan de Déu, Barcelona, Spain. [39]Department of Pediatrics, University Hospitals Leuven, Leuven, Belgium. [40]Department of General Pediatrics and Pediatric Infectious Diseases, Necker Hospital for Sick Children, Assistance Publique–Hôpitaux de Paris (AP-HP), Paris City University, Paris, France. [41]Biodiversity and Epidemiology of Bacterial Pathogens, Pasteur Institute, Paris, France. [42]Department of Pediatric Cardiology, Necker Hospital for Sick Children, AP-HP, Paris City University, Paris, France. [43]Pediatrics Infectious Diseases Division, Hospital Sant Joan de Déu, Barcelona, Spain. [44]Infectious Diseases and Microbiome, Institute of Recerca Sant Joan de Déu, Barcelona, Spain. [45]Department of Molecular and Cellular Biology, National Center for Biotechnology (CNB-CSIC), Madrid, Spain. [46]Laboratory of Virology, Bichat–Claude Bernard Hospital, Paris, France. [47]Laboratory of Virology, AP-HP, Cochin Hospital, Paris, France. [48]Molecular Genetics Unit, Institute of Biomedicine of Valencia (IBV-CSIC), Valencia, Spain. [49]CIBERNED, ISCIII, Madrid, Spain. [50]Joint Research Unit in Neurology and Molecular Genetics, Institut of Investigation Sanitaria La Fe, Valencia, Spain. [51]Medical School, Paris City University, Paris, France. [52]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [53]Bichat–Claude Bernard Hospital, Paris, France. [54]University Paris Diderot, Paris 7, UFR of Médecine-Bichat, Paris, France. [55]IAME, INSERM, UMRS1137, Paris City University, Paris, France. [56]Infectious and Tropical Diseases Department, AP-HP, Bichat–Claude Bernard Hospital, Paris, France. [57]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden. [58]Department of Neuroscience and Experimental Therapeutics, Institute for Biomedical Research of Barcelona (IIBB), Spanish National Research Council (CSIC), Barcelona, Spain. [59]Institute for Biomedical Investigations August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [60]Department of Pediatrics, University Hospitals Leuven and Laboratory for Inborn Errors of Immunity, KU Leuven, Leuven, Belgium. [61]Primary Immunodeficiency Research Laboratory, Center for Primary Immunodeficiency Ghent, Ghent University Hospital, Ghent, Belgium. [62]Neurometabolic Diseases Laboratory, IDIBELL–Hospital Duran I Reynals; and Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. [63]CIBERER U759, ISCiii, Madrid, Spain. [64]Department of Paediatric Infectious Diseases and Virology, Imperial College London, London, UK. [65]Centre for Paediatrics and Child Health, Faculty of Medicine, Imperial College London, London, UK. [66]The American Genome Center, Collaborative Health Initiative Research Program, Uniformed Services University of the Health Sciences, Bethesda, MD, USA. [67]Department of Anatomy, Physiology, and Genetics, Uniformed Services University of the Health Sciences, Bethesda, MD, USA. [68]Study Center for Primary Immunodeficiencies, Necker Hospital for Sick Children, AP-HP, Paris, France. [69]New York Genome Center, New York, NY, USA. [70]Pediatric Hematology-Immunology and Rheumatology Unit, Necker Hospital for Sick Children, AP-HP, Paris, France. [71]Laboratory of Immunogenetics of Human Diseases, Innate Immunity Group, IdiPAZ Institute for Health Research, La Paz Hospital, Madrid, Spain. [72]Interdepartmental Group of Immunodeficiencies, Madrid, Spain. [73]Department of Immunology, University Hospital of Gran Canaria Dr. Negrín, Canarian Health System, Las Palmas de Gran Canaria, Spain. [74]Laboratory of Human Genetics and Genomics, The Rockefeller University, New York, NY, USA. [75]Clinical Immunology and Primary Immunodeficiencies Unit, Pediatric Allergy and Clinical Immunology Department, Hospital Sant Joan de Déu, Barcelona, Spain. [76]Necmettin Erbakan University, Konya, Turkey. [77]Department of Pediatrics, Department of Microbiology, Immunology and Infectious Diseases, University of Montreal and Immunology and Rheumatology Division, CHU Sainte-Justine, Montreal, QC, Canada. [78]National Reference Center for Rheumatic, Autoimmune and Systemic Diseases in Children (RAISE), Pediatric Nephrology, Rheumatology, Dermatology Unit, Hospital of Mother and Child, Hospices Civils of Lyon, Lyon, France. [79]Human Genomics and Evolution, Collège de France, Paris, France. [80]Department of Pediatrics, Necker Hospital for Sick Children, Paris, France. [81]Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA.
***Corresponding author. Email: shzh289@rockefeller.edu** †These authors contributed equally to this work. ‡These authors contributed equally to this work.
§A full list of CoV-Contact Cohort collaborators and their affiliations is provided at the end of the paper. ¶A full list of COVID Human Genetic Effort collaborators and their affiliations is provided at the end of the paper. #These authors contributed equally to this work. **These authors contributed equally to this work.
††These authors contributed equally to this work.

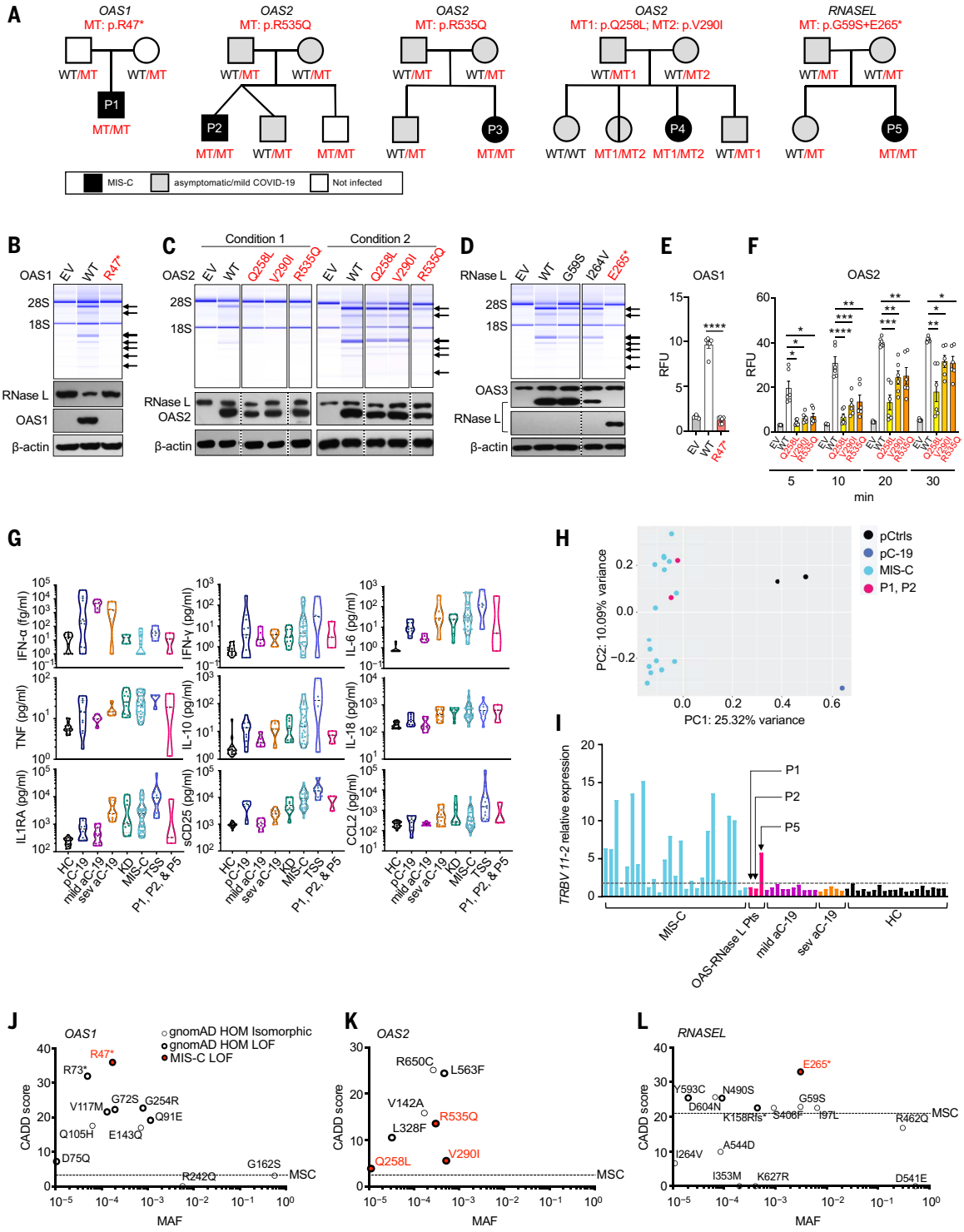


Fig. 1. Biallelic *OAS1*, *OAS2*, and *RNASEL* variants in patients with MIS-C. (**A**) Family pedigrees with allele segregation. Mutant, "MT" in red; wild-type, "WT" in black. (**B** to **D**) Functional assays for WT and mutant OAS1 (B), OAS2 (C), and RNase L (D). Variants for which homozygotes or compound heterozygotes were present in our MIS-C cohort were tested. (Upper panels) RNase L–mediated cleavage of rRNA in a cell-free system based on transfected HeLa M cells. (Lower panels) Immunoblots of the indicated proteins. EV, empty vector. Arrows indicate degraded rRNA species. OAS2 variants (C) were tested under two different sets of conditions (see methods). The results shown in (B) to (D) are representative of three independent experiments. (**E** and **F**) FRET assay of 2-5A synthesized in response to poly(I:C) stimulation by WT and MT OAS1 (E) or OAS2 (F). RFU, relative fluorescence units. The data shown are the means ± SEM of six biological replicates. Statistical analysis was performed as described in the methods. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. (**G**) Concentrations of various cytokines in plasma samples from OAS–RNase L–deficient patients during MIS-C (P1, P2, and P5); comparison with those of healthy controls (HC), pediatric (pC-19) or adult COVID-19 pneumonia (aC-19) patients, typical Kawasaki disease patients (KD), other MIS-C patients with no known genetic etiology (MIS-C), and patients with toxic shock syndrome (TSS). (**H**) PCA of gene expression quantified by whole-blood bulk RNA-seq for P1 and

P2 during MIS-C relative to pediatric controls (pCtrls), previously published MIS-C patients, and a pediatric patient with mild COVID-19 (pC-19). (**I**) Relative levels of *TRBV 11-2* (encoding Vβ21.3) RNA in blood samples from P1, P2, and P5 during MIS-C, relative to other MIS-C patients, adults with mild or severe COVID-19 (mild aC-19, sev aC-19), and healthy controls. (**J** to **L**) CADD-MAF graph of *OAS1* (J), *OAS2* (K), and *RNASEL* (L) variants for which homozygotes are reported in gnomAD and/or found in our MIS-C cohort. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

Table 1. Homozygous or potentially compound-heterozygous rare nonsynonymous variants of the *OAS* and *RNASEL* genes in MIS-C patients. Homozygous or potentially compound-heterozygous nonsynonymous variants with a minor allele frequency (MAF) < 0.01 (gnomAD) found in our cohort of MIS-C patients. CADD_Phred, combined annotation-dependent depletion Phred score; Exp function, experimental function of each variant as tested in the RNase L–dependent rRNA degradation assay (OAS1, OAS2, RNase L) and FRET assay (OAS1, OAS2); Hom, homozygous; Het, heterozygous.

Gene	Nucleotide change	Amino acid change	Zygosity	MAF (gnomAD)	CADD_Phred	Exp function
OAS1	c.139C>T	p.Arg47* (R47*)	Hom	0.00017327	36	LOF
OAS2	c.1604G>A	p.Arg535Gln (R535Q)	Hom	0.00028695	13.58	Hypomorph
OAS2	c.773A>T	p.Gln258Leu (Q258L)	Het	–	3.888	Hypomorph
OAS2	c.868G>A	p.Val290Ile (V290I)	Het	0.0005153	5.585	Hypomorph
OAS3	c.145G>A	p.Ala49Thr (A49T)	Het	0.00243639	9.48	Isomorph
OAS3	c.1475G>A	p.Arg492His (R492H)	Het	0.0054987	9.95	Isomorph
OAS3	c.1703G>A	p.Arg568Lys (R568K)	Het	0.00104951	0.472	Isomorph
OAS3	c.2795G>A	p.Arg932Gln (R932Q)	Het	0.0094	23.2	LOF
OAS3	c.3089A>G	p.Gln1030Arg (Q1030R)	Het	–	23.9	Isomorph
OAS3	c.1586A>G	p.Gln529Arg (Q529R)	Het	0.00000401	5.85	Isomorph
OAS3	c.792C>A	p.His264Gln (H264Q)	Het	0.001001261	0.924	Isomorph
OAS3	c.442C>T	p.Pro148Ser (P148S)	Het	0.000036	22.9	Isomorph
OAS3	c.3259G>A	p.Val1087Met (V1087M)	Het	0.003936537	22.5	Isomorph
RNASEL	c.790A>G	p.Ile264Val (I264V)	Hom	0.00000401	6.597	Isomorph
RNASEL	c.793G>T	p.Glu265* (E265*)†	Hom	0.0031	33	LOF
RNASEL	c.175G>A	p.Gly59Ser (G59S)†	Hom	0.0031	22.9	Isomorph

†*RNASEL* variants p.E265* and p.G59S were in complete linkage disequilibrium (https://www.internationalgenome.org), forming a haplotype.

1288 SARS-CoV-2–infected controls or in a control cohort of 334 patients under the age of 21 years with asymptomatic or mild infection or COVID-19 pneumonia (fig. S1). Thus, there was a significant enrichment in such homozygotes among MIS-C patients relative to infected controls (P = 0.001) or controls under 21 years old (P = 0.046), suggesting that AR deficiencies of three genes of the OAS–RNase L pathway (*OAS1*, *OAS2*, and *RNASEL*) specifically underlie MIS-C. We further assessed the probability of AR deficiencies of these three gene products being causal for MIS-C by evaluating the expression and function of all nonsynonymous variants of *OAS1*, *OAS2*, and *RNASEL* for which homozygotes were reported in the Genome Aggregation Database (gnomAD, v2.1.1 and v3.1.1, 28 variants in total) in our RNase L–mediated rRNA degradation assay (fig. S2, F to H, and table S1). In total, 13 *OAS1*, *OAS2*, or *RNASEL* variants were deleterious and present in the homozygous state in 19 individuals in the gnomAD database (Fig. 1, J to L). The estimated cumulative frequency of homozygous carriers of LOF variants at the three loci was ~0.00013 [95% confidence interval (CI): 7.2×10^{-5} to 20×10^{-5}] in the general population. The rarity of AR OAS–RNase L deficiencies in the general population is therefore consistent with that of MIS-C. Moreover, the enrichment in these deficiencies observed in MIS-C patients relative to the individuals included in gnomAD was highly significant ($P = 2 \times 10^{-6}$). These findings suggest that AR deficiencies of OAS1, OAS2, and RNase L are genetic etiologies of MIS-C.

The expression pattern for the OAS–RNase L pathway implicates mononuclear phagocytes

We studied the basal expression of *OAS1*, *OAS2*, *OAS3*, and *RNASEL* in cells from different tissues. Consistent with data from public databases (*44*), our in-house human cell RNA sequencing (RNA-seq) and reverse transcription–quantitative polymerase chain reaction (RT-qPCR) data showed that myeloid blood cells had higher basal mRNA levels for the four genes than did the tissue-resident cells tested (Fig. 2, A and B). In all cell types studied, both type I and type II IFN treatments up-regulated the levels of mRNA for *OAS1*, *OAS2*, and *OAS3*, whereas the levels of *RNASEL* mRNA were not influenced by these IFNs (fig. S3A). Previous studies reported a relationship between cell type–dependent activation of the OAS–RNase L pathway and basal levels of expression in mice (*45*, *46*). MIS-C occurs 3 to 6 weeks after SARS-CoV-2 infection, but the virus and/or viral proteins may still be detectable in nonrespiratory tissues, such as the intestine or heart, at disease onset in some patients (*32*, *34*, *37*). In addition, $CD4^+$ and $CD8^+$ T cells carrying Vβ21.3 expand, which implies a superantigen-like viral driver of MIS-C (*32*, *34*, *36*–*38*) and suggests that the virus or its antigens persist. Thus, AR deficiencies of the OAS–RNase L pathway may underlie MIS-C by impairing SARS-CoV-2 restriction and/or enhancing virus-triggered inflammatory responses in monocytes and other mononuclear phagocytes.

OAS–RNase L deficiencies have no impact on SARS-CoV-2 replication in A549 epithelial cells and fibroblasts

Previous studies have shown that the overproduction of exogenous OAS1 can result in the restriction of SARS-CoV-2 replication in A549 lung epithelial cells in the absence of exogenous type I IFN (*53*, *54*). However, the five OAS–RNase L–deficient patients had MIS-C without pneumonia. We assessed SARS-CoV-2 replication in A549 cells rendered permissive to SARS-CoV-2 by the stable expression of angiotensin-converting enzyme 2 (ACE2) and transmembrane protease serine 2 (TMPRSS2), which facilitates viral entry. Knockout (KO) of OAS1 or OAS2 did not increase the proportion of SARS-CoV-2–infected cells at 24 or 48 hours relative to that for the parental WT A549 cells, regardless of the presence or absence

Table 2. Demographic and clinical information for MIS-C patients biallelic for deleterious variants of the OAS–RNase L pathway. IEI, inborn error of immunity; SCV2, SARS-CoV-2; IVIG, intravenous immunoglobulins; ND, not determined; CRP, C-reactive protein; sCD25, soluble IL-2Rα.

Patient	P1	P2	P3	P4	P5
IEI (inheritance mode)	*OAS1* (AR)	*OAS2* (AR)	*OAS2* (AR)	*OAS2* (AR)	*RNASEL* (AR)
Age at MIS-C diagnosis	3 months	3 years	14 years	9 years	4 years
Sex	Male	Male	Female	Female	Female
Ethnicity	Filipino	Spanish	Turkish	Turkish	French Canadian
Resident country	Spain	Spain	Turkey	Turkey	Canada
SCV2 virology	Nasal swab PCR (−); blood PCR (−); blood anti-SCV2 IgG (+); blood antigen N (−)	Nasal swab PCR (−); blood PCR (−); blood anti-SCV2 IgG (+); blood antigen N (−)	Nasal swab PCR (−); blood PCR (ND); blood total anti-SCV2 (+); blood antigen N (ND)	Nasal swab PCR (−); blood PCR (ND); blood anti-SCV2 IgM and IgG (+); blood antigen N (ND)	Nasal swab PCR (−); blood PCR (−); blood anti-SCV2 IgG (+); blood antigen N (−)
Hemogram	Normal	Normal	Normal	Normal	Normal
Increased markers of multiorgan inflammation	CRP, ferritin, pro-BNP, GM-CSF, IL-1RA, MCP1, sCD25, IL-18, TNF	CRP, ferritin, pro-BNP, MCP1, sCD25, IL-1RA, IL-18, TNF	CRP, ferritin, troponin	Ferritin, troponin, pro-BNP	sCD25
TRBV 11-2 expansion	(−)	(−)	ND	ND	(+)
Clinical presentation	Kawasaki-like disease: fever, gastrointestinal symptoms, hepatosplenomegaly, aseptic meningitis with neurological symptoms (irritability), peripheral edema, lymphadenopathy, bilateral coronary aneurysm (Z score +8, +8.7), possible cerebral arterial aneurysm	Kawasaki disease: fever, rash, bilateral eyelid edema and erythema, conjunctival hyperemia	Kawasaki-like disease: fever, rash, bilateral nonpurulent conjunctivitis, strawberry tongue, abdominal pain, vomiting, dyspnea, mild mitral insufficiency. One and a half months prior, the patient had fever, headache, and sore throat when her mother had COVID-19. The patient developed oligoarticular juvenile idiopathic arthritis 5 months after MIS-C.	Fever, vomiting, coughing, myocarditis, left ventricular failure, pulmonary edema with paracardiac infiltration, polyneuropathy	Kawasaki disease: fever, rash, erythema and edema of the feet, anterior uveitis, cervical lymphadenopathy
Treatment	IVIG, aspirin, corticosteroids, anticoagulation therapy	IVIG, aspirin	IVIG, methylprednisolone, heparin	IVIG, pulse steroid, anakinra, mechanical ventilation	IVIG
Outcome	Recovery	Recovery	Recovery, but with persistent arthralgia in both knees 1.5 years after MIS-C	Recovery	Recovery

of exogenous IFN-α2b (Fig. 2, C and D, and fig. S3B). Only RNase L KO cells resulted in a mild increase in susceptibility to SARS-CoV-2 relative to WT cells in the absence of IFN-α2b, consistent with previous findings (*55*). We also used patient-specific SV40-transduced human dermal fibroblasts (SV40-fibroblasts) stably expressing ACE2 as a surrogate cell type for studying the impact of OAS–RNase L deficiencies on tissue-resident cell-intrinsic defenses against SARS-CoV-2 (*5*). Consistent with the lack of pneumonia in these patients, no increase in SARS-CoV-2 susceptibility was observed in any of the fibroblasts with *OAS1* (from P1), *OAS2* (P3 and P4), or *RNASEL* (P5) mutations up to 72 hours after infection in the presence or absence of exogenous IFN-α2b, despite the complete loss of expression of OAS1 or RNase L in the cells of P1 and P5, respectively (Fig. 2, E and F). This contrasted with the increased susceptibility reported for fibroblasts from a patient with AR complete IFNAR1 deficiency (*56*) and critical COVID-19 pneumonia.

OAS–RNase L deficiencies have no impact on SARS-CoV-2 replication in THP-1 cells

Only abortive SARS-CoV-2 infection has been reported in human mononuclear phagocytes, including monocytes and macrophages, which express very little to no ACE2 (*57*–*59*). However, basal *Oas* and *Rnasel* expression levels have previously been correlated with murine coronavirus or vesicular stomatitis virus (VSV) restriction in mouse macrophages (*60*). We tested the hypothesis that deficiencies of OAS–RNase L might result in productive SARS-CoV-2 infection in mononuclear phagocytes by assessing the replication of SARS-CoV-2. Unlike WT A549 cells stably transduced with ACE2 and TMPRSS2, in which SARS-CoV-2 can be detected 24 hours after infection, no SARS-CoV-2 was detected in THP-1–derived macrophages (*61*), whether parental or with a KO of OAS1, OAS2, or RNase L (Fig. 2, G and H, and fig. S3C). Thus, no myeloid

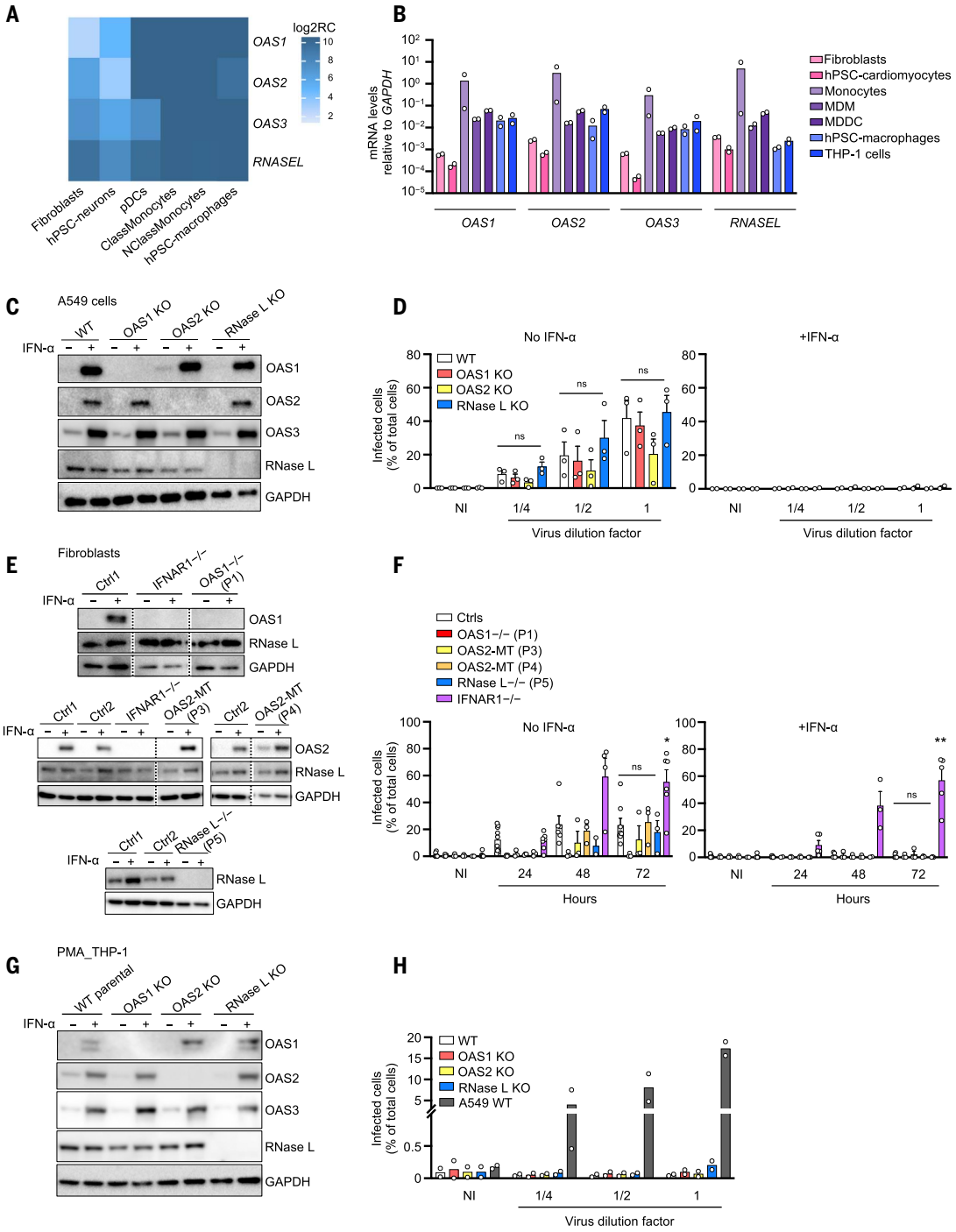


Fig. 2. Expression pattern of the OAS–RNase L pathway genes and their role in SARS-CoV-2 restriction. (**A** and **B**) Relative *OAS1*, *OAS2*, *OAS3*, and *RNASEL* mRNA levels measured by bulk RNA-seq (A) or RT-qPCR (B), in various cell types. hPSC, human pluripotent stem cell; ClassMonocytes, classical monocytes; NClassMonocytes, nonclassical monocytes; MDM, monocyte-derived macrophages; MDDC, monocyte-derived dendritic cells; Log2RC, $\log_2$ read count. (**C** and **D**) Immunoblot of the indicated proteins (C) and immunofluorescence (IF) of SARS-CoV-2 nucleocapsid (N) protein (D) in A549+ACE2/TMPRSS2 cells with and without knockout (KO) of OAS1, OAS2, or RNase L. IF analysis for N protein was performed 24 hours after infection with various dilutions of SARS-CoV-2. Dilution factors of $^1/_4$, $^1/_2$, and 1 correspond to MOI values of 0.0002, 0.0005, and 0.001, respectively. GAPDH, glyceraldehyde-3-phosphate dehydrogenase; NI, noninfected. (**E** and **F**) Immunoblot of the indicated proteins (E) and IF analysis for the SARS-CoV-2 N protein (F) in SV40-fibroblasts +ACE2 from healthy controls (Ctrl1 and Ctrl2), patients with *OAS-RNASEL* mutations (P1, P3, P4, and P5), and a previously reported patient with complete IFNAR1 deficiency (*IFNAR1*$^{-/-}$). IF analysis for N protein was performed at various time points after infection at a MOI of 0.08. (**G** and **H**) Immunoblot of the indicated proteins (G) and IF analysis for the SARS-CoV-2 N protein (H) in THP-1 cells with and without KO of OAS1, OAS2, or RNase L. IF analyses for N protein were performed in PMA-primed THP-1 cells 24 hours after infection with various dilutions of SARS-CoV-2. Dilution factors of $^1/_4$, $^1/_2$, and 1 correspond to MOI values of 0.012, 0.025, and 0.05, respectively. WT A549+ACE2/TMPRSS2 cells were included as a positive control for SARS-CoV-2 infection. The data points are means ± SEM from three [(D) and (F)] or means from two [(B) and (H)] independent experiments with three to six technical replicates per experiment. Statistical analyses were performed as described in the methods. ns, not significant; *P < 0.05, **P < 0.01.

SARS-CoV-2 replication was detected in the presence or absence of deficiencies of the OAS–RNase L pathway, at least in this cellular model of mononuclear phagocytes (*60*).

OAS–RNase L deficiencies result in an exaggerated inflammatory response to intracellular dsRNA in THP-1 cells

Sustained monocyte activation has repeatedly been reported to be a key immunological feature of MIS-C (*22*, *31*–*36*). We studied the impact of OAS–RNase L deficiencies on cellular responses to intracellular (cytosolic) or extracellular (endosomal) stimulation with dsRNA in THP-1 cells. Consistent with a previous study (*62*), THP-1 cells and THP-1–derived macrophages with a KO for OAS1, OAS2, or RNase L displayed enhanced activation, as demonstrated by their higher levels of IFN-λ1, IFN-β, IL-1β, IL-6, CXCL9, CXCL10, and TNF secretion 24 hours after stimulation with various doses of intracellular polyinosinic:polycytidylic acid [poly(I:C)] (Fig. 3A and fig. S4A), as well as higher mRNA induction for *IL6* and *CXCL9* 8 hours after stimulation (fig. S4, B and C). Cell viability was similar to that of WT THP-1 cells after intracellular poly(I:C) stimulation (fig. S4D). Small hairpin RNA–mediated knockdown (KDn) of the expression of *OAS1*, *OAS2*, and *RNASEL* in THP-1 cells confirmed these findings (fig. S4E). The transduction of THP-1 cells with a KO of the corresponding gene with the WT cDNA of *OAS1*, *OAS2*, or *RNASEL*, respectively, resulted in cytokine secretion levels similar to those observed in parental cells, whereas transduction with mutant cDNAs corresponding to the patients' variants had no such effect (*OAS1* variant of P1 and *RNASEL* variant of P5) or a lesser effect (*OAS2* variants of P2, P3, and P4) (Fig. 3B and fig. S5, A to C). Thus, OAS–RNase L deficiencies result in exaggerated inflammatory responses to intracellular dsRNA stimulation in THP-1 cells. Enhanced responses may also occur in the mononuclear phagocytes of our patients, underlying MIS-C.

The inflammatory response to intracellular dsRNA in THP-1 cells is MAVS dependent

Intracellular dsRNA is known to stimulate the RIG-I/MDA5–MAVS pathway (RIG-I, retinoic acid–inducible gene I; MDA5, melanoma differentiation-associated protein 5; MAVS, mitochondrial antiviral-signaling protein), inducing type I IFNs and other cytokines in various cell types (*63*), in addition to the OAS–RNase L pathway (*42*, *64*). Indeed, unlike WT THP-1 cells, MAVS KO THP-1 cells did not respond to intracellular poly(I:C) stimulation, and *RNASEL* gene KDn did not result in enhanced activation (Fig. 3C and fig. S5, D and E), confirming that the response to poly(I:C) is dependent on MAVS-mediated signaling in these cells. The enhancement of the intracellular poly(I:C) response after *RNASEL* KDn was partially attenuated in RIG-I or MDA5 KO THP-1 cells (Fig. 3C and fig. S5, D and E), suggesting that both dsRNA sensors may be involved. Another dsRNA agonist that specifically activates RIG-I, 5′ triphosphate double-stranded RNA (5′ppp-dsRNA), induced enhanced responses in RNase L KO THP-1 cells similar to those seen with poly(I:C) (Fig. 3D). By contrast, the activation of other sensing pathways, including the extracellular ssRNA-sensing toll-like receptor 7 (TLR7) and TLR8 pathways (R848), the TLR4 pathway (LPS), and the intracellular DNA agonist-sensing DAI pathway (ISD), resulted in responses in RNase L KO or KDn THP-1 cells that were similar to those of the parental WT cells (Fig. 3D and fig. S5F). Thus, the exaggerated inflammatory responses to cytosolic dsRNA observed in THP-1 cells deficient for OAS–RNase L appear to require RIG-I/MDA5 sensing and MAVS activation.

Activation of the OAS–RNase L pathway can suppress inflammatory responses in THP-1 cells

Intracellular dsRNA stimulates both the RIG-I/MDA5–MAVS and OAS–RNase L pathways (*42*, *63*, *64*). We therefore investigated whether the dsRNA-sensing MAVS-dependent signaling pathway was itself hyperactivated as a result of OAS–RNase L deficiency. After intracellular poly(I:C) stimulation, interferon regulatory factor 3 (IRF3) and nuclear factor κB (NF-κB) phosphorylation levels were similar in RNase L KO and WT THP-1 cells (Fig. 3E). Thus, the molecular mechanisms by which OAS–RNase L deficiency results in an exaggerated inflammatory response appears to involve an impairment of RNase L activation resulting in a lack of host RNA transcriptional and/or translational inhibition (*65*–*68*), rather than a hyperactivation of the MAVS-dependent pathways. Consistent with this hypothesis, treatment with exogenous 2-5A, which is normally generated by OASs upon dsRNA sensing and activates RNase L (*42*, *43*), rescued the inflammatory phenotype in OAS1 KO THP-1 cells after intracellular poly(I:C) stimulation (Fig. 3F). By contrast, dephosphorylated 2-5A, which is unable to activate RNase L (*69*, *70*), had no such effect (fig. S5G). Moreover, exogenous 2-5A treatment decreased the response to TLR7/8 activation in WT THP-1 cells (Fig. 3G). Treatment with 2-5A had a much weaker effect or even no suppressive effect in RNase L KDn or KO THP-1 cells (Fig. 3F and fig. S5, G and H). Thus, the exaggerated inflammatory response in OAS–RNase L–deficient mononuclear cells appears to result from the activation of the MAVS-dependent pathway (but not of other nucleic acid–sensing pathways) and an impairment of RNase L activation by OAS1- or OAS2-derived 2-5A after dsRNA sensing. This imbalance creates a phenotype that is probably a consequence of an impairment of the posttranscriptional activities of RNase L (*65*–*68*).

OAS–RNase L deficiencies result in an exaggerated inflammatory response to SARS-CoV-2 in THP-1 cells

We investigated whether OAS–RNase L deficiencies resulted in exaggerated inflammatory responses to SARS-CoV-2 in mononuclear phagocytes. Bulk RNA-seq on THP-1 cells with KO of OAS1, OAS2, or RNase L stimulated with intracellular poly(I:C) or SARS-CoV-2 revealed transcriptomic profiles different from those of the parental cells (Fig. 4, A and B, and fig. S6A). Gene set enrichment analysis (GSEA) against Hallmark gene sets (*71*) revealed an enrichment in genes relating to inflammatory responses and IFN-γ signaling in OAS–RNase L–deficient cells, showing that these cells displayed an exacerbated inflammatory response not only to synthetic dsRNA but also to SARS-CoV-2 (Fig. 4, C and D). Moreover, RNase L KO THP-1 cells had higher levels of IL-6 and CXCL10 secretion than WT cells when cocultured with SARS-CoV-2–infected Vero cells, which support SARS-CoV-2 replication (*72*, *73*) (Fig. 4E and fig. S6, B and C). Bulk RNA-seq further confirmed this observation at the transcriptome level (Fig. 4F and fig. S6D), revealing an enrichment in the expression of genes relating to inflammatory responses and IFN-α signaling in RNase L KO cells relative to WT cells (Fig. 4G). In addition, transfection with total RNA from SARS-CoV-2–infected Vero cells, but not from uninfected Vero cells, also induced enhanced responses in RNase L KO THP-1 cells relative to parental WT cells, with an enrichment in genes relating to inflammatory responses and IFN-γ signaling (Fig. 4H and fig. S6E). These findings suggest that OAS–RNase L deficiency results in excessive inflammatory responses in mononuclear phagocytes following both abortive SARS-CoV-2 infection and coculture with SARS-CoV-2–replicating cell types. This is likely due to defective activation of the OAS-RNase L pathway following the engulfment of the virus or infection-related by-products, leading to the release of dsRNA into the cytosol (*73*).

OAS–RNase L deficiencies result in an enhanced inflammatory response to intracellular dsRNA in primary mononuclear cells

We then studied the impact of OAS–RNase L deficiencies on the response to intracellular poly(I:C) stimulation in human peripheral blood mononuclear cells (PBMCs). Routine blood cell counts and immunotyping for the five patients revealed no significant abnormalities in blood leukocyte subsets, a result confirmed by deep immunophenotyping by mass cytometry [cytometry by time of flight (CyTOF)] (fig. S7A and table S2). After intracellular poly(I:C) stimulation, PBMCs from

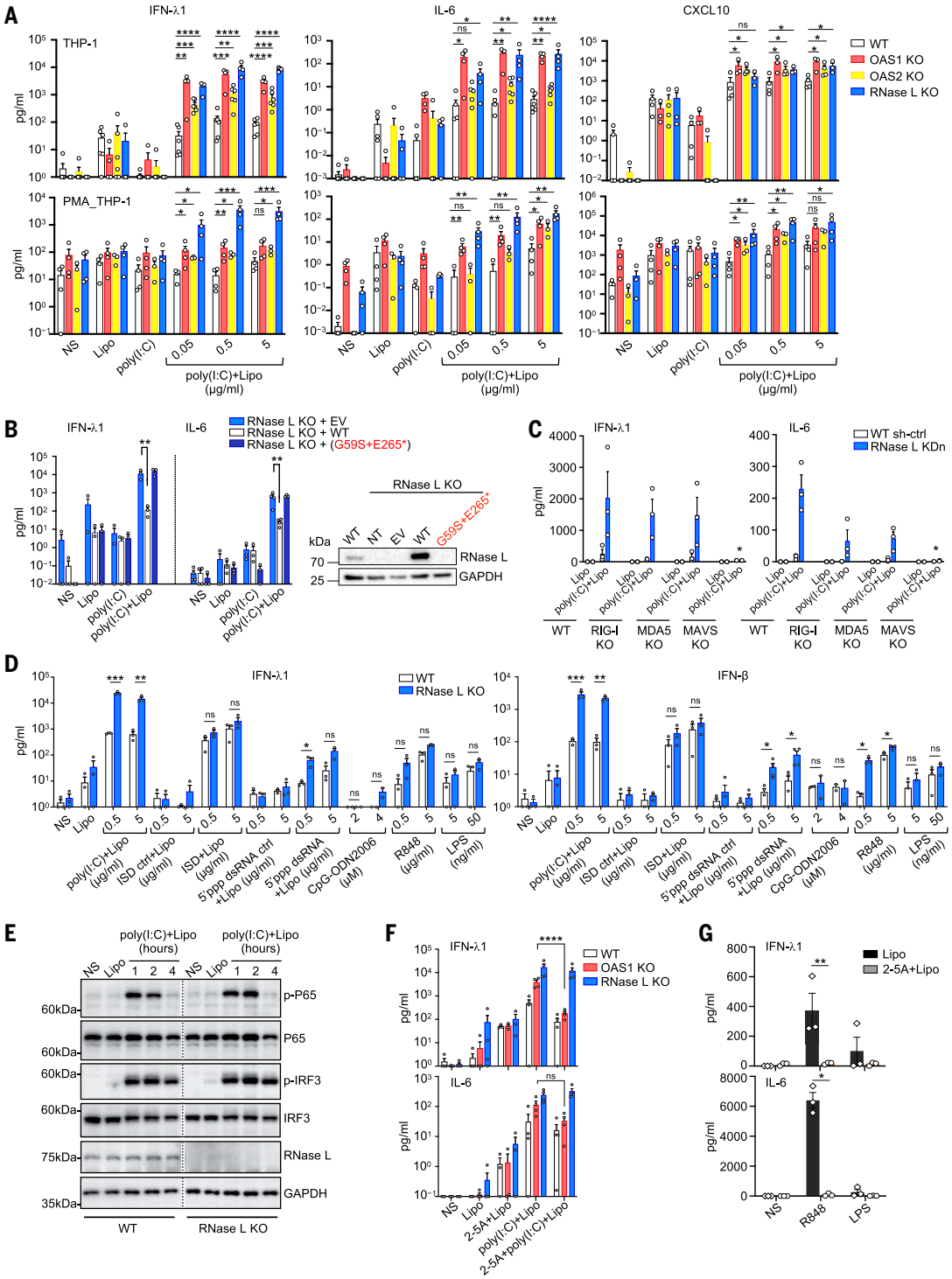


Fig. 3. Exaggerated inflammatory responses of OAS–RNase L-deficient THP-1 cells. (**A**) Concentrations of various cytokines in the supernatant of OAS1 KO, OAS2 KO, RNase L KO, or parental THP-1 cells (upper panels) or PMA-primed THP-1 cells (lower panels) treated as indicated for 24 hours. (**B**) IFN-λ1 and IL-6 concentrations in the supernatant of RNase L KO THP-1 cells transduced with the WT or P5's variant *RNASEL* cDNA, or empty vector (EV), and treated as indicated for 24 hours. On the right, RNase L protein levels, as assessed by immunoblotting. NT, not transfected. (**C**) IFN-λ1 and IL-6 concentrations in the supernatant of parental, RIG-I KO, MDA5 KO, or MAVS KO THP-1 cells with or without (WT sh-ctrl) RNase L knockdown (KDn), treated as indicated for 24 hours. (**D**) IFN-λ1 and IFN-β concentrations in the supernatant of parental or RNase L KO THP-1 cells, treated as indicated for 24 hours. (**E**) Immunoblot of phosphorylated P65 and IRF3 in parental and RNase L KO THP-1 cells treated as indicated. The results shown are representative of two independent experiments. (**F**) IFN-λ1 and IL-6 concentrations in the supernatant of parental, OAS1 KO, or RNase L KO THP-1 cells treated as indicated for 24 hours. (**G**) IFN-λ1 and IL-6 concentrations in WT THP-1 cells treated as indicated for 24 hours. In (A) to (D), (F), and (G), the data points are means ± SEM from three to five independent experiments with one to two technical replicates per experiment. Statistical analysis was performed as described in the methods. ns, not significant; *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. NS, nonstimulated; Lipo, lipofectamine only; poly(I:C), extracellularly added poly(I:C); poly(I:C)+Lipo, intracellular poly(I:C) in the presence of lipofectamine; 2-5A+Lipo, intracellular 2-5A in the presence of lipofectamine; 2-5A+poly(I:C)+Lipo, intracellular poly(I:C) in addition to intracellular 2-5A.

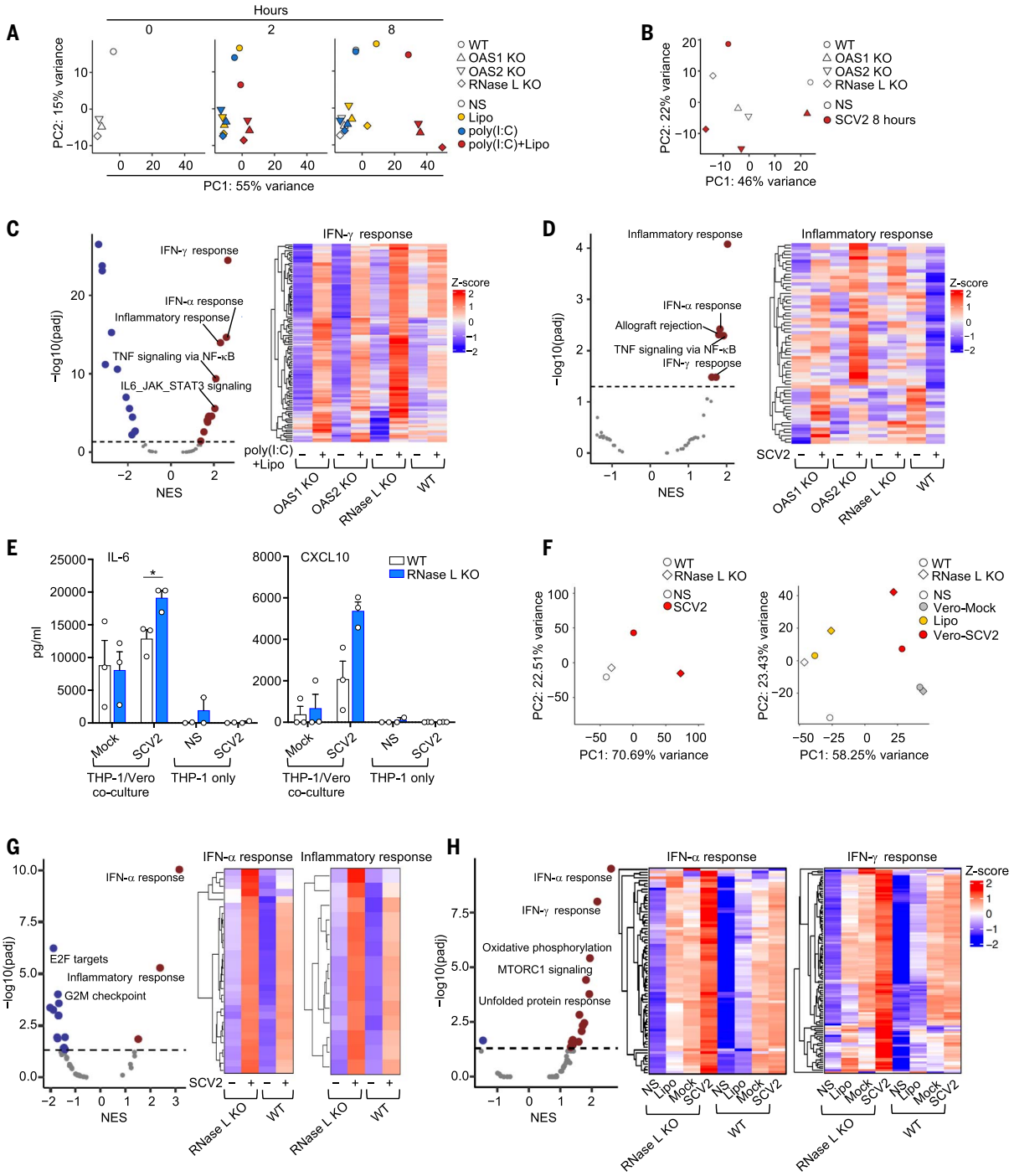


Fig. 4. Exaggerated inflammatory responses to SARS-CoV-2 of OAS–RNase L–deficient THP-1 cells. (**A** and **B**) PCA of RNA-seq–quantified gene expression for OAS1 KO, OAS2 KO, RNase L KO, and parental (WT) THP-1 cells left nonstimulated (NS), treated as indicated for 2 or 8 hours (A), or stimulated with SARS-CoV-2 (SCV2) at a MOI of 0.01 for 8 hours (B). (**C** and **D**) Differential expression analysis (DEA) and gene set enrichment analysis (GSEA) for genes induced by 8 hours of intracellular poly(I:C) stimulation (C) or by 8 hours of SCV2 stimulation (D). The OAS1 KO, OAS2 KO, and RNase L KO THP-1 cells were compared with parental (WT) THP-1 cells. Volcano plots show immune system–related pathways. NES, normalized enrichment score. Heatmaps show gene expression for the "IFN-γ response" (C) or "inflammatory response" (D) Hallmark gene sets. (**E**) IL-6 and CXCL10 concentrations in the supernatant of parental or RNase L KO THP-1 cells treated as indicated for 24 hours. The data points are means ± SEM from three independent experiments with three technical replicates per experiment. Statistical analysis was performed as described in the methods. *$P < 0.05$. (**F**) PCA of RNA-seq–quantified gene expression, for RNase L KO and parental THP-1 cells cocultured with Vero cells with or without SCV2 infection for 24 hours (left) or transfected for 8 hours with RNA from Vero cells with or without SCV2-infection (right). (**G** and **H**) DEA and GSEA for genes induced in RNase L KO THP-1 cells, compared with parental THP-1 cells after 24 hours of coculture with SCV2-infected or mock-infected Vero cells (G), or after 8 hours of transfection with RNA from SCV2-infected or mock-infected Vero cells (H). Volcano plots show immune system–related pathways. Heatmaps show gene expression for the indicated Hallmark gene sets. Heatmaps represent Z-score–scaled $\log_2$ read counts per million. NS, nonstimulated; Lipo, lipofectamine; SCV2, SARS-CoV-2.

P2 (OAS2 deficient), P3 (OAS2 deficient), and P5 (RNase L deficient) secreted larger amounts of the inflammatory cytokines studied than cells from healthy controls (Fig. 5A and fig. S7B). This enhanced inflammatory response to intracellular poly(I:C) stimulation was monocyte dependent, as the depletion of monocytes from the PBMCs of healthy controls strongly decreased this response (fig. S7C). Moreover, the shRNA-mediated KDn of *OAS1*, *OAS2*, or *RNASEL* in monocyte-derived dendritic cells (MDDCs) from healthy controls resulted in an enhanced inflammatory response to intracellular poly(I:C) stimulation, as shown by the higher levels of inflammatory cytokines, including IFN-λ1, IL-6, TNF, and IL-12, than were observed with WT parental cells (Fig. 5B). Thus, deficiencies of the OAS–RNase L pathway also result in exaggerated inflammatory responses to intracellular dsRNA stimulation in primary mononuclear phagocytes, or at least in monocytes and MDDCs.

Enhanced myeloid cell activation by SARS-CoV-2 in patient PBMCs

We studied the impact of OAS–RNase L deficiencies on the responses of the various PBMC populations to SARS-CoV-2 by performing single-cell RNA sequencing (scRNA-seq) on PBMCs from P1 (OAS1), P2 (OAS2), P3 (OAS2), and P5 (RNase L) and comparing the results with those for healthy controls. Regardless of genotype, 6 hours of stimulation with SARS-CoV-2 induced a strong immune response across all five major immune cell types including myeloid, B, $CD4^+$ T, $CD8^+$ T, and natural killer (NK) cells (Fig. 5C), with 1301 unique differentially expressed genes (DEGs) (data S1). OAS–RNase L deficiency significantly changed the response of 48 to 94% of the DEGs in each lineage, with myeloid cells being the most affected. Cellular responses were generally stronger in the OAS–RNase L–deficient patients and were essentially limited to the IFN-α and IFN-γ response pathways. Myeloid cell responses were characterized by a distinct proinflammatory component, such as *IL1B* and *CCL3*, that was stronger in OAS–RNase L–deficient cells (Fig. 5D and data S2). We then calculated pseudobulk estimates by cell type. Consistent with the single-cell observations, genes strongly up-regulated by SARS-CoV-2 in OAS–RNase L–deficient myeloid cells were enriched in types I and II IFN signature genes and TNF signature genes, whereas those strongly up-regulated in $CD4^+$ T cells were enriched in type I IFN signature genes (Fig. 5E). Thus, there is an exaggerated inflammatory response to intracellular dsRNA or extracellular SARS-CoV-2 stimulation in primary monocytes and other mononuclear phagocytes with deficiencies of the OAS–RNase L pathway cultured alone or with other PBMC populations. This provides a plausible pathogenic mechanism for MIS-C, in which this condition is driven by the exacerbated activation of mononuclear phagocytes. This hypothesis is also supported by scRNA-seq on PBMCs from P5 (RNase L deficient) collected during MIS-C and the convalescence period. Enhanced expression levels were observed for IFN-α, IFN-γ, or TNF signature genes in monocytes, myeloid dendritic cells (mDCs), B lymphocytes, plasmacytoid dendritic cells (pDCs), and activated T cells of P5 relative to healthy pediatric controls (Fig. 5, F and G, and fig. S8, A to D). Quantitatively inferred cell–cell communications (*74*) revealed that MIS-C in the RNase L–deficient patient was probably driven by a signal from hyperactivated monocytes and mDCs directed at $CD8^+$ αβ T cells (Fig. 5, H and I, and fig. S8, E to G). This situation differs from that observed in patients with COVID-19 pneumonia without MIS-C but is similar to reports for previously described MIS-C patients (fig. S9) (*33*, *34*, *36*), identifying exaggerated myeloid cell activation due to OAS–RNase L deficiency as the core driver of the immunological and clinical phenotypes of MIS-C in our patients.

Discussion

We report AR deficiencies of OAS1, OAS2, and RNase L as genetic etiologies of MIS-C in five unrelated children, corresponding to ~1% of the international cohort of patients studied. OAS–RNase L–deficient monocytic cell lines, monocyte-derived dendritic cells modeling patient genotypes, and primary monocytes from patients displayed excessive inflammatory responses to intracellular dsRNA, SARS-CoV-2, SARS-CoV-2–infected cells, and their RNA, providing a plausible mechanism for MIS-C. In these patients, MIS-C may result primarily from an excessive response of monocytes and other mononuclear phagocytes to SARS-CoV-2 dsRNA intermediates or by-products, followed by the presentation of a viral superantigen to T cells, resulting in the activation and expansion of $V\beta21.3^+$ $CD4^+$ and $CD8^+$ T cells. The molecular basis of the exacerbated inflammatory response to SARS-CoV-2 due to OAS–RNase L deficiency in mononuclear phagocytes involves an impairment of the activation of RNase L by the dsRNA-sensing molecules OAS1 and OAS2, probably resulting in defective post-transcriptional RNase L activity (*67*, *68*) and the unchecked RIG-I/MDA5–MAVS–mediated production of inflammatory cytokines. Alternative molecular mechanisms cannot be excluded (*64*, *75*). The SARS-CoV-2–related RNA products that trigger phagocyte activation, the viral superantigen(s) that activate T cells, and the human leukocyte antigen (HLA) restriction elements all remain to be discovered. Our findings also do not exclude the possibility that AR OAS–RNase L deficiency additionally affects antiviral responses in cells of other tissues injured during MIS-C, such as cardiomyocytes, enterocytes, and endothelial cells. The role of this pathway in T cells themselves merits further investigation. MIS-C in other patients may result from IEIs that may or may not be related to the OAS–RNase L pathway. Our findings also suggest that other forms of Kawasaki disease may be caused by other virus-specific IEIs in other patients (*15*).

The notion that the OAS–RNase L pathway is essential for antiviral immunity in mononuclear phagocytic cells was first proposed nearly 40 years ago (*60*). Intriguingly, the OAS–RNase L pathway is apparently dispensable for protective immunity to SARS-CoV-2 in the respiratory tract. None of the five MIS-C patients had a pulmonary phenotype, and no viral replication was detectable in the upper respiratory tract of any of the five children at the onset of MIS-C. Nevertheless, genome-wide association studies have suggested that common variants in the vicinity of *OAS1* may be weakly associated with COVID-19 severity (*10*, *11*, *53*, *76*–*79*). Our finding that the human OAS–RNase L pathway is crucial for regulation of the mononuclear phagocyte response to SARS-CoV-2, but not for SARS-CoV-2 restriction in the respiratory tract, suggests that the main protective action of this pathway is mediated by the control of phagocyte-driven systemic inflammation at a later stage of disease rather than viral restriction in the respiratory tract early on. These findings are also consistent with the discovery of germline gain-of-function *OAS1* mutations in humans with an autoinflammatory syndrome involving myeloid cells (*80*, *81*).

The five patients, now aged 1 to 15 years, are normally resistant to diseases caused by other common viruses. Since the discovery of the OAS–RNase L pathway in the 1970s (*65*, *82*, *83*), this pathway has been one of the most intensively studied type I IFN–inducible pathways (*42*, *84*). Biochemically, the three OASs have different subcellular distributions and different dsRNA optima for activation, they synthesize 2-5A of different lengths (*42*, *85*), and they appear to have antiviral activity against different viruses (*86*–*88*). The only well-established function of 2-5A is the activation of RNase L (*66*), and any of the three OASs appears to be sufficient for the biochemical activation of RNase L in human cells in vitro. RNase L has been shown to have antiviral activity against certain viruses (dengue virus and Sindbis virus), but not others (Zika virus), in murine and human cells in vitro (*85*, *89*). In vivo RNase L deficiency in mice drives susceptibility to various viruses (e.g., encephalomyocarditis virus, coxsackievirus B4, murine coronavirus, etc.) (*45*, *85*). Our data suggest that human OAS1, OAS2, and RNase L are each essential for the correct regulation of immunity to SARS-CoV-2 but are otherwise largely redundant in natural conditions of infection. It is also clear that the RNase L–dependent functions of OAS1 and

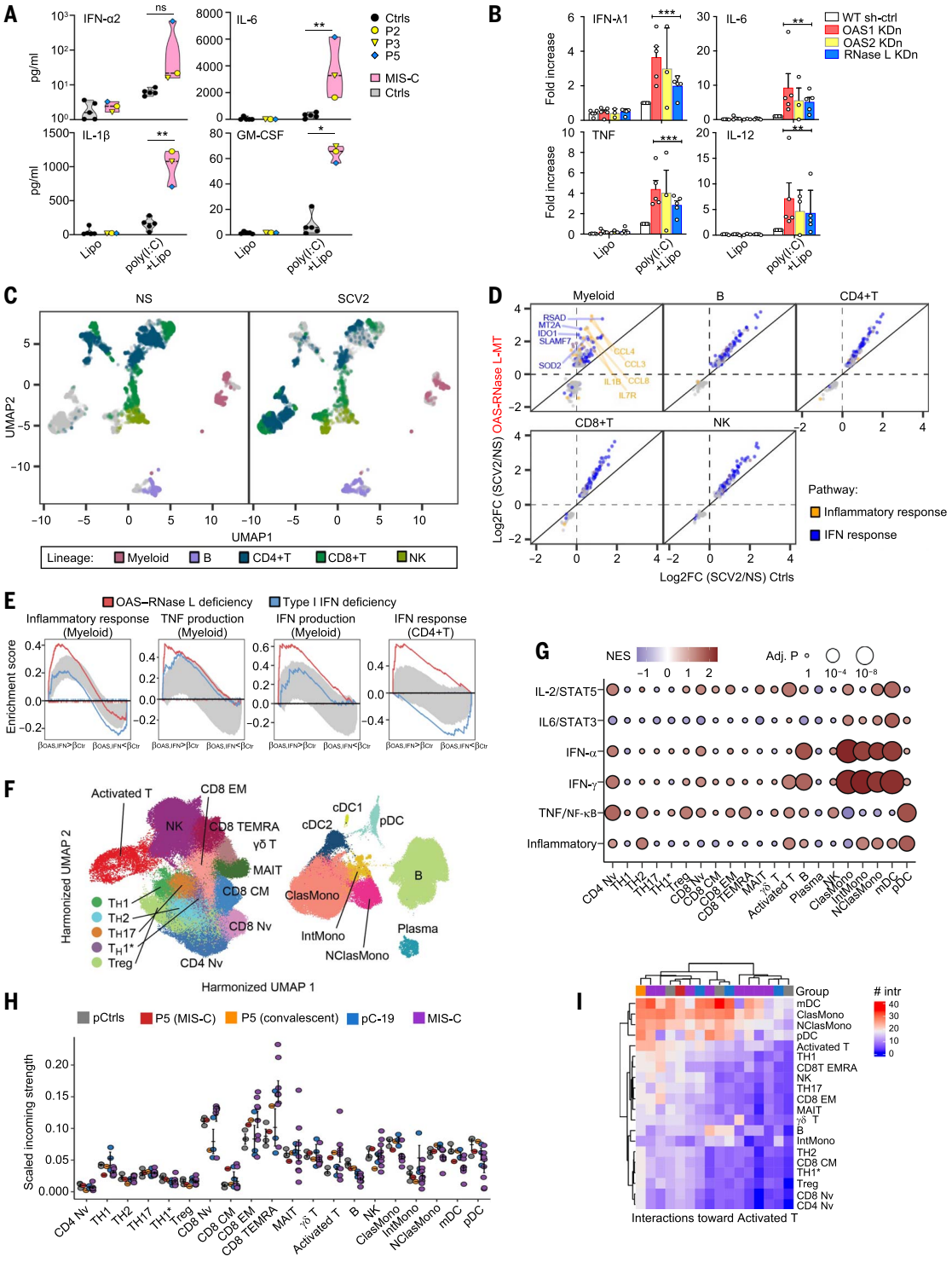


Fig. 5. Exaggerated myeloid cell activation in response to SARS-CoV-2 underlies MIS-C. (**A**) Concentrations of cytokines in the supernatant of PBMCs from OAS–RNase L–deficient patients (grouped in the pink violin zone) and three healthy pediatric and two healthy adult controls (Ctrls; gray violin zone). The data points are means of biological duplicates. (**B**) Fold-increase in the concentrations of cytokines in the supernatant of MDDCs with KDn of OAS1, OAS2, or RNase L, or transduced with control shRNA (WT sh-ctrl). The fold-change is expressed relative to the values for poly(I:C)+lipo-stimulated WT sh-ctrl cells. Data shown are means ± SEM from three independent experiments, with one to two technical replicates per experiment. For (A) and (B), statistical analysis was performed as described in the methods. NS, nonstimulated. ns, not significant; *P < 0.05, **P < 0.01, ***P < 0.001. (**C** to **E**) scRNA-seq of PBMCs from OAS–RNase L–deficient patients (OAS–RNase L-MT) or healthy controls after 6 hours of incubation with SARS-CoV-2 (SCV2) or mock infection (NS). (C) Uniform manifold approximation and projection (UMAP) of single PBMC transcriptomes. (D) Cell type–specific transcriptional responses. Genes passing the FDR < 0.01 and |log2FC| > 0.5 thresholds are shown. (E) GSEA of SCV2-induced genes across immune-related Hallmark gene sets. PBMCs from three patients with type I IFN pathway deficiency are controls for defective type I IFN responses. Gray zone highlights the expected enrichment scores under the null hypothesis (95% CI calculated over 100 randomized genes). (**F** to **I**) scRNA-seq of PBMCs from P5 and from healthy controls. A published dataset for pediatric patients with acute SARS-CoV-2 infection (pC-19) and MIS-C was also integrated. (F) UMAP of clustering analysis. (G) Pseudobulk differential expression analysis with GSEA. P5 (convalescent phase) was compared with local pediatric controls (pCtrls). Immune-related pathways are shown. [(H) and (I)] Intercellular communication analysis with CellChat. (H) Incoming signal strength and (I) the number of interactions for representative cell subsets.

OAS2 are crucial for the regulation of immunity to SARS-CoV-2 within the same cells, as the genetic deficiency of any of these three components results in the same immunological and clinical phenotype, namely MIS-C.

Materials and methods

Patients

We enrolled an international cohort of 558 MIS-C patients (aged 3 months to 19 years, 60.4% boys and 39.6% girls) originating from Europe, Africa, Asia, and America and living in 16 different countries. All patients met the WHO diagnostic criteria for MIS-C (*52*). We focus here on five of these patients (P1 to P5). Written informed consent was obtained

in the country of residence of each patient, in accordance with local regulations and with institutional review board (IRB) approval. Experiments were conducted in the United States and in France, in accordance with local regulations and with the approval of the IRB of the Rockefeller University and the Institut National de la Santé et de la Recherche Médicale, respectively. Approval was obtained from the French Ethics Committee (Comité de Protection des Personnes), the French National Agency for Medicine and Health Product Safety, the Institut National de la Santé et de la Recherche Médicale in Paris, France (protocol no. C10-13), and the Rockefeller University Institutional Review Board in New York, USA (protocol no. JCA-0700). For patients sequenced by National Institute of Allergy and Infectious Diseases (NIAID) through the American Genome Center (TAGC) other than the five patients described in this paper, written informed consent was obtained in the country of residence of each patient, in accordance with local regulations and with IRB approval: Ethics Committee of the Fondazione IRCCS Policlinico San Matteo, Pavia, Italy (protocol 20200037677); Comitato Etico Interaziendale A.O.U. Città della Salute e della Scienza di Torino, Turin, Italy (protocol 00282/2020); and IRB at Children's Hospital of Philadelphia (protocol 18-014863).

The five patients with MIS-C and AR deficiencies of the OAS–RNase L pathway—two boys and three girls—ranged in age from 3 months to 14 years at the time of diagnosis and all fulfilled the WHO criteria for MIS-C (Table 2) (*52*). They originated from the Philippines (P1), Spain (P2), Turkey (P3 and P4), and Canada (of French descent) (P5) and lived in Spain, Turkey, and Canada. P1 (*OAS1* mutation) (*29*), P3 (*OAS2*), and P4 (*OAS2*) had a severe course of MIS-C, with coronary aneurysm, myocarditis, and polyneuropathy, respectively. P2 (*OAS2*) and P5 (*RNASEL*) had a milder course of MIS-C, with a typical Kawasaki disease presentation. None of these patients presented any clinical or radiological evidence of pneumonia. Cytokine profiling of serum obtained from P1, P2, and P5 during MIS-C revealed high levels of IFN-γ, soluble CD25, IL-18, IL-1RA, and MCP1 (CCL2) (Fig. 1G), consistent with previously published immune profiles of MIS-C and in contrast to those for pulmonary COVID-19 (*21*). Bulk mRNA sequencing (RNA-seq) of whole-blood RNA from P1 and P2 collected during the MIS-C phase revealed transcriptomic signatures clearly different from those of healthy controls and a pediatric case of acute COVID-19 pneumonia, but similar to those of previously reported MIS-C patients (Fig. 1H) (*33*). T cell receptor Vβ repertoire analysis confirmed the expansion of *TRBV 11-2* (encoding Vβ21.3) in one of the three MIS-C–phase samples available (P5, with AR RNase L deficiency) (Fig. 1I). The clinical and immunological features of the five patients were, therefore, consistent with those previously reported for other MIS-C patients (*21*, *22*, *26*–*36*).

Whole-exome, whole-genome, and Sanger sequencing

Genomic DNA was extracted from whole blood. Whole-exome sequencing (WES) or whole-genome sequencing (WGS) was performed at several sequencing centers, including the Genomics Core Facility of the Imagine Institute (Paris, France), the Yale Center for Genome Analysis (USA), the New York Genome Center (NY, USA), the American Genome Center (TAGC, Uniformed Services University of the Health Sciences, Bethesda, USA), and the Genomics Division–Institute of Technology and Renewable Energies (ITER) of the Canarian Health System sequencing hub (Canary Islands, Spain). More technical details are provided in the supplementary materials. For the Sanger sequencing of *OAS1*, *OAS2*, and *RNASEL* variants, the relevant regions of *OAS1*, *OAS2*, and *RNASEL* were amplified by PCR, purified by ultracentrifugation through Sephadex G-50 Superfine resin (Amersham-Pharmacia-Biotech), and sequenced with the Big Dye Terminator Cycle Sequencing Kit on an ABI Prism 3700 apparatus (Applied Biosystems).

Whole-exome sequencing data analysis

We performed an enrichment analysis focusing on the three candidate genes in our cohort of 558 MIS-C patients and 1288 children and adults with asymptomatic or paucisymptomatic SARS-CoV-2 infection (controls). We considered variants that were predicted to be loss-of-function or missense, with a highest population MAF < 0.01, not included in segmental duplication regions (gnomAD v2.1.1). We considered genes corresponding to the Gene Ontology term "response to virus" (GO:0009615), with a gene damage index of <13.83 (*41*), corresponding to the 90% least-damaged genes. We searched for all homozygous variants in MIS-C patients, SARS-CoV-2–infected controls, and the gnomAD database. We compared the proportions of patients and controls carrying experimentally confirmed deleterious homozygous variants by means of a logistic regression model, accounting for the ethnic heterogeneity of the cohorts by including the first five principal components of the principal components analysis (PCA), and for data heterogeneity (WGS and WES with various kits and calling processes) by including the two first PCs of a PCA on individual sequence-quality parameters, as previously described (*9*). The PCA for ethnic heterogeneity was performed with PLINK (v1.9) on WES and WGS data, with the 1000 Genomes Project phase 3 public database as a reference, using >15,000 exonic variants with a MAF > 0.01 and a call rate > 0.99. The PCA for data heterogeneity was performed with the R FactoMineR package and the following individual sequence quality parameters calculated with bcftools stats: number of alleles, number of ALT alleles, number of heterozygous variants, Ts/Tv ratio, number of indels, mean depth of coverage, number of singletons, and number of missing genotypes. We also compared the frequency of experimentally confirmed deleterious homozygous variants of the three genes between our MIS-C cohort and gnomAD using Fisher's exact test.

Cell culture

Primary cultures of human fibroblasts were established from skin biopsy specimens from patients or healthy controls. They were transformed with an SV40 vector, as previously described (*56*), to create immortalized SV40-fibroblast cell lines. SV40-fibroblasts, human embryonic kidney 293T (HEK293T) cells, and A549 cells were cultured in Dulbecco's modified essential medium (DMEM; GIBCO) with 10% fetal bovine serum (FBS) (GIBCO). THP-1 cells were cultured in RPMI 1640 medium (GIBCO) with 10% FBS. For the generation of phorbol-12-myristate-13-acetate (PMA)–primed THP-1–derived macrophages, THP-1 cells were incubated with 50 ng/ml of PMA for 48 hours then left without PMA overnight before stimulation. PBMCs were cultured in RPMI 1640 medium (GIBCO) with 10% FBS. For intracellular poly(I:C) or SARS-CoV-2 stimulation of the PBMCs, blood samples were obtained from the OAS–RNase L–deficient patients 2 months to 1 year after acute-phase MIS-C and from five healthy controls with (two pediatric controls and one adult control) or without (one pediatric control and one adult control) prior asymptomatic or mild SARS-CoV-2 infection ~6 months before sample collection. For the differentiation of monocyte-derived dendritic cells, monocytes were isolated from PBMCs with the Pan Monocyte Isolation kit (Miltenyi Biotec) and cultured with 50 ng/ml of recombinant human granulocyte-macrophage colony-stimulating factor (GM-CSF; PeproTech) and 20 ng/ml of recombinant human IL-13 (PeproTech) for 7 days before cell stimulation experiments.

Plasmids

For overexpression studies in HEK293T cells, WT cDNAs for *OAS1* and *RNASEL* in a pCMV6 backbone were purchased from Origene. For rRNA degradation assays, human *OAS1* (GenBank accession no. BC071981.1), *OAS2* (GenBank accession no. BC049215.1), *OAS3* (GenBank accession no. BC113746), and *RNASEL* (GenBank accession no. L10381.1) cDNAs were inserted into p3X-FLAG-CMV-10 (Sigma) as previously described (*75*, *88*). Patient-specific variants or variants from the gnomAD database were

generated by site-directed mutagenesis PCR with the Super Pfx DNA Polymerase (CWbio). For stable lentivirus-mediated transduction with *ACE2* and *RNASEL*, cDNAs for WT and patient-specific *ACE2* or *RNASEL* variants were inserted into pTRIP-SFFV-CD271-P2A, a modified pTRIP-SFFV-mtagBFP-2A (Addgene 102585) in which mtagBFP is replaced with CD271, with InFusion (Takara Bio), according to the manufacturer's instructions. We used the XhoI and BamHI restriction sites. For stable lentivirus-mediated transduction with *OAS1* and *OAS2*, cDNAs for WT and patient-specific *OAS1* or *OAS2* variants were inserted into a modified pSCRPSY vector (KT368137.1) with a PaqCI cutting site expressing blue fluorescent protein (BFP). The PaqCI site was used for cDNA insertion with InFusion. We checked the entire sequences of the *OAS1*, *OAS2*, *OAS3*, and *RNASEL* cDNAs in the plasmids by Sanger sequencing.

Cell-free system assays of OAS and RNase L activity

Assays for OAS and RNase L activity were performed with a modified cell-free system assay based on HeLa M cells (*49*, *50*). The HeLa M cells were cultured in DMEM with 10% FBS, and their identity was confirmed by the presence of short tandem repeat loci with a 94.12% match to HeLa cells (ATCC CCL2, Genetica, Burlington, NC). We previously reported that HeLa M cells have no RNase L expression (*51*). Cells were plated in 24-well dishes (6×10^4 cells per well) with empty vector (p3X-FLAG-CMV-10) or vector containing WT or mutant human *OAS1* (GenBank accession no. BC071981.1), *OAS2* (GenBank accession no. BC049215.1), *OAS3* (GenBank accession no. BC113746), or *RNASEL* (GenBank accession no. L10381.1) cDNAs. HeLa M cells were cotransfected with cDNAs in the presence of Lipofectamine 2000 for 20 hours. Conditions were optimized for each type of enzyme assayed. RNase L assays were performed on cells cotransfected with 300 ng of WT or mutant *RNASEL* cDNA and 100 ng of WT *OAS3* cDNA. *OAS1* assays were performed with 300 ng of *OAS1* cDNA and 100 ng of *RNASEL* cDNA. OAS2 assays were performed with 300 ng (condition 1) or 600 ng (condition 2) of *OAS2* cDNA and 100 ng of *RNASEL* cDNA, and *OAS3* assays were performed with 300 ng of *OAS3* cDNA and 100 ng of *RNASEL* cDNA. The lysis-activation-reaction (LAR) buffer contained 0.1% (by volume) Nonidet P-40, 50 mM Tris-HCl pH 7.5, 0.15 M NaCl, 2 mM EDTA, 10 mM $MgCl_2$, 2 mM ATP, 400 U/ml of RNaseOUT (Thermo Fisher Scientific), and 2.5 μg/ml of poly(I):poly(C) (Millipore catalog no. 528906). LAR buffer (75 μl) was added to each well of cells on ice and the contents of the wells were then transferred to tubes on ice. The lysates were then incubated at 30°C for 30 min, except in OAS2 assays, for which lysates were incubated at 37°C (condition 1) or 30°C (condition 2) for 40 and 50 min, respectively. Total RNA was isolated with RLT buffer supplemented with guanidinium isothiocyanate and the EZ-10 Spin Columns Total RNA Minipreps Super kit (BIO BASIC). RNA was separated on RNA chips with an Agilent Bioanalyzer 2000, from which images and RNA integrity numbers (RINs) were obtained. For immunoblots, aliquots of the lysates (10 μg of protein) were separated by SDS–polyacrylamide gel electrophoresis (SDS-PAGE) in a 7% acrylamide gel. Immunoblots were probed with a monoclonal antibody against the Flag epitope or β-actin (Sigma-Aldrich).

FRET-based OAS enzyme assays

FRET assays of the amount of 2-5A synthesized by WT and mutant isoforms of OAS1 or OAS2 were performed with lysates of transfected HeLa M cells (*90*). Cells were plated in 24-well dishes (6×10^4 cells per well), cultured for 24 hours and transfected for 20 hours with Lipofectamine 2000 transfection reagent (Thermo Fisher Scientific) and 0.5 μg empty vector (p3X-FLAG-CMV-10), or 500 ng of vector containing WT or mutant OAS1 or OAS2. Cells were washed with cold PBS and then lysed with 100 μl of LAR buffer [containing ATP and poly(I:C)] per well on ice. The lysates were transferred to tubes on ice and incubated at 30°C for 50 min before heating at 95°C for 10 min (to stop the reaction and denature proteins) and vortexing twice. The lysates were centrifuged at 12,000*g* for 10 min. The supernatants were then collected and diluted 10-fold in H_2O. Diluted samples (2 μl) were added to 45 μl of cleavage buffer (25 mM Tris-HCl, pH 7.4, 0.1 M KCl, 10 mM $MgCl_2$, 50 μM ATP pH 7.4, and 7 mM β-mercaptoethanol) containing 40 nM RNase L and 135 nM FRET probe in 96-well plates. The probe used was a 36-nucleotide synthetic oligoribonucleotide probe with multiple RNase L cleavage sites, a fluorophore (6-FAM or 6-carboxyfluorescein) at the 5′ terminus, and the black hole quencher-1 (BHQ1) at the 3′ terminus (IDT, Inc.) (*90*). FRET assays were performed at room temperature, every 5 min, for 30 min. Fluorescence was measured in relative fluorescence units (RFU), with excitation at 485 nm and emission at 535 nm, with a Varioskan LUX multimode microplate reader and Skanit version 6.0.1 software (Thermo Fisher Scientific). There were six biological replicates for each treatment group. Standard curves were plotted in triplicate with 0.1 to 30 nM ppp5′A2′p5′A2′p5′A (trimer 2-5A) synthesized with isolated OAS1 and purified by high-performance liquid chromatography (HPLC) (*70*).

Cytokine quantification in plasma samples

Cytokine quantification in plasma samples was performed as previously described (*32*). Briefly, whole blood was sampled into EDTA tubes. The plasma concentrations of IFN-γ, IL-1RA, IL-10, IL-18, IL-6, MCP-1, soluble CD25, and TNF were then determined with Simpleplex technology and an ELLA instrument (Protein Simple) according to the manufacturer's instructions. Plasma IFN-α concentrations were determined with a single-molecule array (Simoa) on an HD-1 Analyzer (Quanterix) with a commercial kit for IFN-α2 quantification (Quanterix). Blood samples from P1, P2, and P5 were obtained on days 7, 4, and 9 after symptom onset, respectively.

TRBV 11-2 relative expression levels

Whole blood was collected into PAXgene (BD Biosciences) or Tempus (Thermo Fisher Scientific) blood RNA tubes or EDTA tubes. RNA was extracted with the corresponding RNA extraction kits or with the Maxwell 16 LEV Blood RNA kit and a Maxwell extractor (Promega) and quantified by spectrometry (Nanovue). For P5, RNA was extracted from sorted T cells with the RNeasy Plus microkit (Qiagen). Relative expression levels were determined for *TRBV 11-2* with nCounter analysis technology (NanoString Technologies), by calculating *TRBV 11-2* mRNA levels relative to other *TRBV* mRNA levels and normalizing against the median value for the healthy volunteer group. Blood samples from P1, P2, and P5 were obtained on days 7, 4, and 9 after symptom onset, respectively.

Immunoblots

Total protein extracts were prepared by lysing cells in NP40 lysis buffer (150 mM NaCl, 50 mM Tris pH 8.0, and 1.0% NP40) supplemented with cOmplete Protease Inhibitor cocktail (Roche, Mannheim, Germany). Equal amounts of protein from each sample were subjected to SDS-PAGE, and the proteins were blotted onto polyvinylidene difluoride membranes (Bio-Rad). The membranes were then probed with the desired primary antibody followed by the appropriate secondary antibody. Primary antibodies against the following targets were used: Flag tag (Sigma-Aldrich, cat: F1804), human OAS1 (Cell Signaling, cat: 14498), OAS2 (Proteintech, cat: 19279-1-AP), RNase L (Cell Signaling, cat: 27281), RIG-I (Cell Signaling, cat: 3743), MDA5 (Cell Signaling, cat: 5321), MAVS (Cell Signaling, cat: 3993), phospho-IRF3 (Cell Signaling, cat: 4947), total IRF3 (Cell Signaling, cat: 11904), phospho-p65 (Cell Signaling, cat: 3033), and total p65 (Santa Cruz, cat: sc-372). Membranes were probed with a horseradish peroxidase (HRP)–conjugated antibody against GAPDH (Proteintech, cat: HRP-60004), as a protein loading control. Antibody binding was detected by enhanced chemiluminescence (Thermo Fisher Scientific).

RT-qPCR

Total RNA was extracted from THP-1 cells and various other cell types with the Quick-RNA MicroPrep kit (Zymo Research). RNA was

reverse-transcribed with random hexamers and the Superscript III first-strand cDNA synthesis system (Invitrogen). Quantitative real-time PCR was then performed with the TaqMan universal PCR master mix (Applied Biosystems). For gene expression assays, TaqMan probes for *OAS1*, *OAS2*, *OAS3*, *RNASEL*, *IL6*, and *CXCL9* were used (Thermo Fisher Scientific). We used β-glucuronidase (*GUSB*) for normalization (Applied Biosystems). The results were analyzed with the ΔCt or ΔΔCt method. For SARS-CoV-2 genomic RNA quantification, RNA was extracted from 3×10^5 THP-1 cells infected with SARS-CoV-2 for 24 hours. Cells were washed three times with PBS and lysed for RNA extraction. Equal amounts of total RNA were reverse-transcribed with random hexamers and the Superscript III first-strand cDNA synthesis kit (Invitrogen). Equal amounts of cDNA were used for the qPCR reaction. Primers and probes for the N gene (N2 region), the RNA-dependent RNA polymerase (RdRP) gene, and their respective standards were purchased from IDT technologies. All qPCR reactions were analyzed with the QuantStudio 3 system.

Gene knockout

OAS1 knockout THP-1 cells and the parental WT cells were kindly provided by W.-B. Lee (*62*). The THP-1 cells with knockouts for RIG-I, MDA5, and MAVS were purchased from Invivogen. A549 KO cells were kindly provided by S. Weiss (*55*). For the generation of OAS2 and RNase L KO THP-1 cells, a set of three single-guide RNAs for *OAS2* or *RNASEL* (Synthego) were combined with TrueCut Cas9 protein v2 (Invitrogen) and used for the nucleofection of the cells with Cell Line Nucleofection kit V (Lonza) and AMAXA Nucleofector 2b (Lonza), according to the manufacturer's instructions. The cells were cultured for several days and then plated at clonal density in 96-well plates and amplified. Genomic DNA was extracted from multiple clones, and genomic regions of ~450 bp around the *OAS2* or *RNASEL* single guide RNAs were subjected to Sanger sequencing. The absence of the protein was confirmed by immunoblotting. The loss of RNase L activity in RNase L KO THP-1 cells was confirmed in an rRNA degradation assay. The sequences of the guide RNAs for OAS2 and RNase L knockouts were 5′-AGCUGAGAGCAAUGGGAAAU-3′, 5′-UCAGACACUGAUCGACGAGA-3′, and 5′-UGCACCAGGGGGAACUGUUC-3′ (*OAS2*); and 5′-GCAGUGGAGAAGAAGCACUU-3′, 5′-GCAGGUGGCAUUUACCGUCA-3′, and 5′-UUUGACCUUACCAUACACAG-3′ (*RNASEL*). The sequencing primers were 5′-CAGTTTCAGTTTCCTGGCTCTGG-3′ and 5′-GCACATAATAGGCACCCAGCAC-3′ for *OAS2* and 5′-CTCTGTTGCCAGAGAATCCCAATTTAC-3′, 5′-CAATCGCTGCGAGGATAAAAGG-3′, 5′-GAGCGTGAAGCTGCTGAAAC-3′, and 5′-TGTACTGGCTCCACGTTTG-3′ for *RNASEL*.

Gene knockdown

The shRNA-mediated silencing experiments were performed with GIPZ (Horizon Discovery) lentiviral vectors encoding microRNA-adapted shRNAs targeting the open reading frame of *OAS1* (catalog nos. 200201641 and 200293786), *OAS2* (200260991 and 200255637), and *RNASEL* (200226261 and 200226578), or a nonsilencing control shRNA (RHS4346). Lentiviral particles encoding shRNA were generated by the transient transfection of HEK293T cells with lentiviral GIPZ vectors and a mixture of packaging plasmids with X-tremeGENE 9 transfection reagent, used according to the manufacturer's instructions. Briefly, HEK293T cells at 80 to 90% confluence in a six-well plate were transfected with 1.5 μg of the lentiviral vector GIPZ, 1 μg of the packaging plasmid (psPAX2, Addgene), and 0.5 μg of the envelope plasmid (pMD2G, Addgene). The medium was changed the following day, and the virus-containing supernatant was collected 48 hours after transfection, passed through a filter with 0.45-μm pores, and used directly for cell transduction or stored at −80°C.

For the transduction of THP-1 cells, the cells were incubated with supernatants containing the lentiviral particles. The medium was replaced with fresh medium the following day, and puromycin was added 3 days after transduction, to a final concentration of 2 μg/ml. Protein production was analyzed by immunoblotting after 4 days of selection. All the experiments were performed between days 7 and 14 after transduction.

For shRNA-mediated knockdown experiments in primary monocyte-derived dendritic cells (MDDCs), a high transduction efficiency (>60% GFP^+ cells) was achieved by cotransduction with shRNA-encoding lentiviral particles and virion-like particles (VLPs) carrying the SIV viral protein Vpx (VLP-Vpx). Vpx suppresses the SAMHD1-mediated restriction of lentiviral reverse transcription in myeloid cells. VLP-Vpx were produced by transfecting HEK293T cells with 1.5 μg of the packaging vector SIV3+ (derived from SIVmac251) and 0.5 μg of the envelope plasmid pMD2G with XtremeGENE9. Monocytes were isolated from PBMCs from healthy donors by negative selection with the Pan Monocyte Isolation Kit (Miltenyi Biotec). Freshly purified monocytes were transduced with shRNA-encoding lentiviral particles and VLP-Vpx in the presence of protamine (8 μg/ml). Transduced cells were allowed to differentiate into MDDCs in the presence of recombinant human GM-CSF (10 ng/ml) and IL-4 (25 ng/ml) for 5 days.

Lentiviral transduction

HEK293T cells were dispensed into a six-well plate at a density of 8×10^5 cells per well. The next day, cells were transfected with pCMV-VSV-G (0.2 μg), pHXB2-env (0.2 μg; NIH-AIDS Reagent Program; 1069), psPAX2 (1 μg; Addgene plasmid no. 12260), and either pTRIP-SFFV-CD271-P2A empty vector or encoding the protein of interest (1.6 μg) in Opti-MEM (Gibco; 300 μl) containing X-tremeGENE 9 (Sigma Aldrich; 10 μl), according to the manufacturer's instructions. After 6 hours, the medium was replaced with 3 ml of fresh culture medium, and the cells were incubated for a further 24 hours for lentiviral particle production. The viral supernatant was collected and passed through a syringe filter with 0.2-μm pores (Pall) to remove debris. Protamine sulfate (Sigma; 10 μg/ml) was added to the supernatant, which was then used immediately or stored at −80°C until use.

For the transduction of THP-1 cells with *OAS1*, *OAS2*, or *RNASEL*, the corresponding gene KO THP-1 cells were dispensed into a 12-well plate at a density of 1×10^6 cells per well, in 500 μl of culture medium per well. Viral supernatant was then added (500 μl per well) the next day. For the transduction of SV40-fibroblasts with ACE2, healthy control or patient-specific SV40-fibroblasts were used to seed six-well plates at a density of 5×10^5 cells per well. Viral supernatant was added (500 μl per well) the next day. The cells were then incubated for a further 48 hours at 37°C. Transduction efficiency was evaluated by surface staining for CD271 (Miltenyi Biotec) for the pTRIP vector, or by flow cytometry to evaluate BFP expression levels for the pSCRPSY vector. MACS column separation was performed with selection beads for CD271-positive cells (Miltenyi Biotec) if the proportion of CD271-positive cells was <80%. Cells transduced with the pSCRPSY vector were selected with puromycin or by flow cytometry. Protein production was subsequently validated by immunoblotting.

SARS-CoV-2 infection

The SARS-CoV-2 NYC isolate was obtained from the saliva of a deidentified patient on 28 July 2020. The sequence of the virus is publicly available (GenBank OM345241). The virus isolate was initially amplified in Caco-2 cells (passage 1, or P#1 stock). For the generation of P#2 and P#3 working stocks, Caco-2 cells were infected with the P#1 and P#2 viruses, respectively, at a multiplicity of infection (MOI) of 0.05 plaque-forming units (PFU)/cell and incubated for 6 and 7 days, respectively, at 37°C. The virus-containing supernatant was then harvested, clarified by centrifugation (3000*g* for 10 min), and filtered through a disposable vacuum filter system with 0.22-μm pores. The P#3 stock used in this study had a titer of 3.4×10^6 PFU/ml determined on Vero E6 cells with a 1% methylcellulose overlay, as previously described (*72*).

A549 + ACE2/TMPSS2 cells, human SV40-fibroblasts + ACE2, or THP-1 cells were used to seed 96-well plates at a density of 1.5×10^4 cells per well, 4×10^3 cells per well, and 1×10^5

cells per well, respectively, in the presence or absence of IFN-α2b at a concentration of 1000 IU/ml. The cells were infected with SARS-CoV-2 24 hours later by directly adding 10 μl of virus stock at various dilutions to the wells (final volume: 110 μl). Cells were infected for 24, 48, or 72 hours. The cells were fixed with neutral buffered formalin at a final concentration of 10% and stained for SARS-CoV-2 with an anti-N antibody (catalog no. GTX135357; GeneTex). An Alexa Fluor 488- or Alexa Fluor 647-conjugated secondary antibody (Invitrogen) was used. Plates were imaged with an ImageXpress micro XL and analyzed with MetaXpress (Molecular Devices).

Cell stimulation

THP-1 cells were used to coat a 96-well plate at a density of 1×10^5 cells per 100 μl of culture medium. For stimulations of PBMCs and MDDCs, we used 1×10^5 cells and 5×10^5 cells per 100 μl of culture medium, respectively. The cells were stimulated with the indicated stimulus at the specified concentrations, with or without lipofectamine 2000 (Invitrogen), according to the manufacturer's instructions. Poly(I:C), 5′ppp-dsRNA, 5′ppp-dsRNA control, ISD, ISD control, R848, CPG-ODN2006, and LPS were purchased from Invivogen. For exogenous 2′5′-linked oligoadenylate (2-5A) or dephosphorylated 2-5A, we used 20 μM of 2-5A for transfection in the presence of lipofectamine simultaneously with the other stimuli [poly(I:C), R848, or LPS]. Dephosphorylated 2-5A (A2′p5′A2′p5′A) was prepared by treating 2-5A with shrimp alkaline phosphatase (SAP) (Thermo Fisher Science) to remove the 5′-triphosphoryl group from 2-5A, rendering it unable to activate RNase L (*69*, *70*). The dephosphorylation reaction mixture contained 5 mM 2-5A incubated with five units of SAP at 37°C for 1 hour, according to the manufacturer's protocol. Samples were denatured by incubation at 95°C for 5 min. Supernatants containing dephosphorylated 2′,5′-A3 were removed after centrifugation at 18,000*g* for 15 min at 4°C. Dephosphorylated 2-5A was then validated by HPLC and FRET assays for RNase L activity. After cell stimulation, the cells or supernatants were harvested, and their cytokine mRNA and protein levels were assessed by RT-qPCR and with a multiplex bead assay (BioLegend), respectively.

Detection of secreted cytokines in a multiplex bead assay

The harvested supernatants of stimulated THP-1 cells, PBMCs, and other types of cells were prepared and used for the LEGENDplex multiplex bead assay (BioLegend), according to the manufacturer's instructions. Samples were analyzed by flow cytometry on an Attune NxT flow cytometer, according to the manufacturer's instructions. Data were analyzed with LEGENDplex Cloud-based Data Analysis Software.

Luciferase assay

THP-1 cells expressing an ISRE-lucia luciferase reporter gene were purchased from Invivogen (THP1-Dual). Cells were stimulated according to the conditions specified above. The supernatant was collected and used for the luciferase assay in accordance with the manufacturer's instructions.

Coculture of THP-1 and SARS-CoV-2–infected cells

Vero cells were plated in a six-well plate and infected at a MOI of 0.05 (as determined by plaque assay on Vero E6 cells) for a total of 48 hours. The supernatant of the infected cells was carefully removed, and the infected cells were then transferred to fresh THP-1 culture medium. A fixed volume of the resulting cell suspension was then dispensed onto WT or RNase L KO THP-1 cells plated in a 96-well plate at a density of 1×10^5 cells in 100 μl. THP-1 cells stimulated with SARS-CoV-2 only were stimulated in parallel for 24 hours. THP-1 cells were stimulated for a total of 24 hours before collection of the supernatant for cytokine determinations and cells for total RNA extraction.

Transfection of THP-1 cells with RNA from SARS-CoV-2–infected cells

Total RNA was extracted from mock-infected Vero cells or Vero cells infected with SARS-CoV-2 at a MOI of 0.05 for a total of 72 hours. THP-1 cells were transfected with 2 μg/ml of total RNA extract for 8 hours. THP-1 cells were then collected for total RNA extraction.

Deep immunophenotyping by mass cytometry (CyTOF)

CyTOF was performed on whole blood with the Maxpar Direct Immune Profiling Assay (Fluidigm), according to the manufacturer's instructions, as previously described (*7*). Cells were frozen at −80°C after overnight staining to eliminate dead cells, and acquisition was performed on a Helios machine (Fluidigm). The antibodies used for staining are listed in table S3. All the samples were processed within 24 hours of sampling. Data analysis was performed with OMIQ software.

Bulk RNA sequencing (RNA-seq)

Total RNA was extracted from THP-1 cells or sorted blood cell populations. Cells were left untreated or were stimulated with poly(I:C) in the presence of lipofectamine or infected with SARS-CoV-2. RNA was extracted with the Quick-RNA MicroPrep kit (Zymo Research) or the RNeasy Micro Kit (Qiagen) and treated with DNase I (Zymo Research and Qiagen) to remove residual genomic DNA. RNA-seq libraries were prepared with the Illumina RiboZero TruSeq Stranded Total RNA Library Prep Kit (Illumina) and sequenced on the Illumina NovaSeq platform in the 100 nucleotide, paired-end configuration. Each library was sequenced twice.

The RNA-seq FASTQ files were first inspected with fastqc to ensure that the raw data were of high quality. The sequencing reads of each FASTQ file were then aligned with the GENCODE human reference genome GRCh37.p13 with STAR aligner v2.6 and the alignment quality of each BAM file was evaluated with RSeQC. Reads were quantified with featureCounts v1.6.0 to generate gene-level feature counts from the read alignment, based on GENCODE GRCh37.p13 gene annotation. The gene-level feature counts were then normalized and $\log_2$-transformed with DESeq2, to obtain gene expression values for all genes and all samples. Differential gene expression analyses were conducted by contrasting the intracellular poly(I:C)-stimulated samples or the SARS-CoV-2–infected samples with the nonstimulated samples. For each gene expression analysis, we performed trimmed mean of M values (TMM) normalization and gene-wise generalized linear model regression by edgeR, and the genes displaying significant differential expression were selected according to the following criteria: FDR ≤ 0.05 and |log2(FoldChange)| ≥ 1. Differential gene expression was plotted as a heatmap with ComplexHeatmap, and genes and samples were clustered according to complete linkage and the Euclidean distances of gene expression values. GSEA was conducted with the fgsea package, by projecting the ranking of fold-change in expression onto the Hallmark gene sets (*71*).

Single-cell RNA sequencing of PBMCs

We performed scRNA-seq on SARS-CoV-2– and mock-stimulated PBMCs sampled from four individuals with inborn errors of the OAS–RNaseL pathway (P1 with OAS1 deficiency, P2 and P3 with OAS2 deficiency, P5 with RNase L deficiency), three individuals with inborn errors of type I IFN immunity, and eight healthy donors—one pediatric control and one adult control with a history of past asymptomatic SARS-CoV-2 infection, and two pediatric controls and four adult controls with no history of prior SARS-CoV-2 infection. The cryopreserved PBMCs were thawed, stimulated, and processed for scRNA-seq. Across all samples, we captured 46,157 high-quality single-cell transcriptomes that were classified into five major immune cell lineages: myeloid, B, $CD4^+$ T, $CD8^+$ T, and NK cells. The data were then analyzed as described in detail in the supplementary materials.

We also performed scRNA-seq on cryopreserved PBMCs from P5 (RNase L–deficient, aged 4 years) sampled during the acute (9 days after MIS-C onset) and convalescent (~1 month after onset) phases, together with cells from one healthy adult and two pediatric controls.

We compared the data obtained with a previously published dataset for patients with pediatric acute SARS-CoV-2 infection or MIS-C (*33*). Clustering analysis showed lower levels of monocytes and type 1 and type 2 conventional dendritic cells (cDCs) in these patients and an expansion of the activated T cell population strongly expressing *MKI67* (Fig. 5F and fig. S8, A and B). Other subsets were largely unaffected. Pseudobulk differential expression analysis was performed at the single-cell level for monocytes, mDCs, B lymphocytes, plasmacytoid dendritic cells (pDCs), and activated T cells. Bulk RNA-seq was performed on sorted nonclassical monocytes and pDCs to further confirm the scRNA-seq findings. We also quantitatively inferred cell–cell communications with CellChat (*74*) to identify the signal-outgoing and the signal-receiving cell subsets. The data generated during this study were analyzed in an integrative manner with historical controls from the laboratory (one pediatric and seven adult controls), publicly available control PBMC datasets downloaded from the 10X Genomics web portal (https://support.10xgenomics.com/single-cell-gene-expression/datasets), and a previously published dataset for patients with acute SARS-CoV-2 infection and MIS-C (GEO accession: GSE167029), as described in detail in the supplementary materials. In addition, two other previously published sets of scRNA-seq data for pediatric healthy controls and children with acute SARS-CoV-2 infection or MIS-C (GSE166489) (*91*) were used for an independent cohort analysis.

Statistical analysis

For experiments performed in vitro, quantitative data were obtained for cells carrying the different mutations and control cells, or cells treated with different stimuli, from at least three biological replicates. For each biological replicate, up to six technical replicates were performed and averaged for downstream analysis. Cytokine determinations were log-transformed after subtracting the limit of detection for the experiment concerned. Mean quantitative values were compared between cells carrying the various mutations and control cells or cells treated with different stimuli in unequal-variance *t* tests. Where relevant, statistical test results are indicated in the corresponding figures (ns, not significant; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$).

REFERENCES AND NOTES

1. A. T. Levin *et al*., Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *Eur. J. Epidemiol.* **35**, 1123–1138 (2020). doi: 10.1007/s10654-020-00698-1; pmid: 33289900
2. M. O'Driscoll *et al*., Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140–145 (2021). doi: 10.1038/s41586-020-2918-0; pmid: 33137809
3. K. Bhaskaran *et al*., Factors associated with deaths due to COVID-19 versus other causes: Population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. *Lancet Reg. Health Eur.* **6**, 100109 (2021). doi: 10.1016/j.lanepe.2021.100109; pmid: 33997835
4. E. J. Williamson *et al*., Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020). doi: 10.1038/s41586-020-2521-4; pmid: 32640463
5. Q. Zhang *et al*., Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, eabd4570 (2020). doi: 10.1126/science.abd4570; pmid: 32972995
6. P. Bastard *et al*., Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**, eabd4585 (2020). doi: 10.1126/science.abd4585; pmid: 32972996
7. T. Asano *et al*., X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. *Sci. Immunol.* **6**, eabl4348 (2021). doi: 10.1126/sciimmunol.abl4348; pmid: 34413140
8. P. Bastard *et al*., Autoantibodies neutralizing type I IFNs are present in ~4% of uninfected individuals over 70 years old and account for ~20% of COVID-19 deaths. *Sci. Immunol.* **6**, eabl4340 (2021). doi: 10.1126/sciimmunol.abl4340; pmid: 34413139
9. Q. Zhang *et al*., Recessive inborn errors of type I IFN immunity in children with COVID-19 pneumonia. *J. Exp. Med.* **219**, e20220131 (2022). doi: 10.1084/jem.20220131; pmid: 35708626
10. E. Pairo-Castineira *et al*., Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021). doi: 10.1038/s41586-020-03065-y; pmid: 33307546
11. H. Zeberg, S. Pääbo, The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020). doi: 10.1038/s41586-020-2818-3; pmid: 32998156
12. A. Kousathanas *et al*., Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* **607**, 97–103 (2022). doi: 10.1038/s41586-022-04576-6; pmid: 35255492
13. COVID-19 Host Genetics Initiative, Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021). doi: 10.1038/s41586-021-03767-x; pmid: 34237774
14. S. B. Morris *et al*., Case series of multisystem inflammatory syndrome in adults associated with SARS-CoV-2 infection — United Kingdom and United States, March–August 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 1450–1456 (2020). doi: 10.15585/mmwr.mm6940e1; pmid: 33031361
15. V. Sancho-Shimizu *et al*., SARS-CoV-2–related MIS-C: A key to the viral and genetic causes of Kawasaki disease? *J. Exp. Med.* **218**, e20210446 (2021). doi: 10.1084/jem.20210446; pmid: 33904890
16. E. Whittaker *et al*., Clinical characteristics of 58 children with a pediatric inflammatory multisystem syndrome temporally associated with SARS-CoV-2. *JAMA* **324**, 259–269 (2020). doi: 10.1001/jama.2020.10369; pmid: 32511692
17. M. Ahmed *et al*., Multisystem inflammatory syndrome in children: A systematic review. *EClinicalMedicine* **26**, 100527 (2020). doi: 10.1016/j.eclinm.2020.100527; pmid: 32923992
18. E. M. Dufort *et al*., Multisystem inflammatory syndrome in children in New York State. *N. Engl. J. Med.* **383**, 347–358 (2020). doi: 10.1056/NEJMoa2021756; pmid: 32598830
19. A. B. Payne *et al*., Incidence of multisystem inflammatory syndrome in children among US persons infected with SARS-CoV-2. *JAMA Netw. Open* **4**, e2116420 (2021). doi: 10.1001/jamanetworkopen.2021.16420; pmid: 34110391
20. L. R. Feldstein *et al*., Characteristics and outcomes of US children and adolescents with multisystem inflammatory syndrome in children (MIS-C) compared with severe acute COVID-19. *JAMA* **325**, 1074–1087 (2021). doi: 10.1001/jama.2021.2091; pmid: 33625505
21. M. J. Carter *et al*., Peripheral immunophenotypes in children with multisystem inflammatory syndrome associated with SARS-CoV-2 infection. *Nat. Med.* **26**, 1701–1707 (2020). doi: 10.1038/s41591-020-1054-6; pmid: 32812012
22. C. R. Consiglio *et al*., The immunology of multisystem inflammatory syndrome in children with COVID-19. *Cell* **183**, 968–981.e7 (2020). doi: 10.1016/j.cell.2020.09.016; pmid: 32966765
23. L. Hoste, R. Van Paemel, F. Haerynck, Multisystem inflammatory syndrome in children related to COVID-19: A systematic review. *Eur. J. Pediatr.* **180**, 2019–2034 (2021). doi: 10.1007/s00431-021-03993-5; pmid: 33599835
24. J. Toubiana *et al*., Distinctive features of Kawasaki disease following SARS-CoV-2 infection: a controlled study in Paris, France. *J. Clin. Immunol.* **41**, 526–535 (2021). doi: 10.1007/s10875-020-00941-0; pmid: 33394320
25. B. Cherqaoui, I. Koné-Paut, H. Yager, F. L. Bourgeois, M. Piram, Delineating phenotypes of Kawasaki disease and SARS-CoV-2-related inflammatory multisystem syndrome: A French study and literature review. *Rheumatology* **60**, 4530–4537 (2021). doi: 10.1093/rheumatology/keab026; pmid: 33493353
26. C. N. Gruber *et al*., Mapping systemic inflammation and antibody responses in multisystem inflammatory syndrome in children (MIS-C). *Cell* **183**, 982–995.e14 (2020). doi: 10.1016/j.cell.2020.09.034; pmid: 32991843
27. C. Diorio *et al*., Multisystem inflammatory syndrome in children and COVID-19 are distinct presentations of SARS-CoV-2. *J. Clin. Invest.* **130**, 5967–5975 (2020). doi: 10.1172/JCI140970; pmid: 32730233
28. P. Y. Lee *et al*., Distinct clinical and immunological features of SARS-CoV-2-induced multisystem inflammatory syndrome in children. *J. Clin. Invest.* **130**, 5942–5950 (2020). doi: 10.1172/JCI141113; pmid: 32701511
29. A. Esteve-Sole *et al*., Similarities and differences between the immunopathogenesis of COVID-19-related pediatric multisystem inflammatory syndrome and Kawasaki disease. *J. Clin. Invest.* **131**, e144554 (2021). doi: 10.1172/JCI144554; pmid: 33497356
30. H. Bukulmez, Current understanding of multisystem inflammatory syndrome (MIS-C) following COVID-19 and its distinction from Kawasaki disease. *Curr. Rheumatol. Rep.* **23**, 58 (2021). doi: 10.1007/s11926-021-01028-4; pmid: 34216296
31. L. A. Vella *et al*., Deep immune profiling of MIS-C demonstrates marked but transient immune activation compared to adult and pediatric COVID-19. *Sci. Immunol.* **6**, eabf7570 (2021). doi: 10.1126/sciimmunol.abf7570; pmid: 33653907
32. M. Moreews *et al*., Polyclonal expansion of TCR Vβ 21.3$^+$ CD4$^+$ and CD8$^+$ T cells is a hallmark of multisystem inflammatory syndrome in children. *Sci. Immunol.* **6**, eabh1516 (2021). doi: 10.1126/sciimmunol.abh1516; pmid: 34035116
33. C. de Cevins *et al*., A monocyte/dendritic cell molecular signature of SARS-CoV-2-related multisystem inflammatory syndrome in children with severe myocarditis. *Med (N Y)* **2**, 1072–1092.e7 (2021). doi: 10.1016/j.medj.2021.08.002; pmid: 34414385
34. A. Ramaswamy *et al*., Immune dysregulation and autoreactivity correlate with disease severity in SARS-CoV-2-associated multisystem inflammatory syndrome in children. *Immunity* **54**, 1083–1095.e7 (2021). doi: 10.1016/j.immuni.2021.04.003; pmid: 33891889
35. L. A. Vella, A. H. Rowley, Current insights into the pathophysiology of multisystem inflammatory syndrome in children. *Curr. Pediatr. Rep.* **9**, 83–92 (2021). doi: 10.1007/s40124-021-00257-6; pmid: 34692237
36. K. Sacco *et al*., Immunopathological signatures in multisystem inflammatory syndrome in children and pediatric COVID-19. *Nat. Med.* **28**, 1050–1062 (2022). doi: 10.1038/s41591-022-01724-3; pmid: 35177862
37. R. A. Porritt *et al*., HLA class I–associated expansion of TRBV11-2 T cells in multisystem inflammatory syndrome in children. *J. Clin. Invest.* **131**, e146614 (2021). doi: 10.1172/JCI146614; pmid: 33705359
38. L. Hoste *et al*., TIM3$^+$*TRBV11-2* T cells and IFNγ signature in patrolling monocytes and CD16$^+$ NK cells delineate MIS-C. *J. Exp. Med.* **219**, e20211381 (2022). doi: 10.1084/jem.20211381; pmid: 34914824
39. J. L. Casanova, L. Abel, Mechanisms of viral inflammation and disease in humans. *Science* **374**, 1080–1086 (2021). doi: 10.1126/science.abj7965; pmid: 34822298
40. S. Y. Zhang, Q. Zhang, J. L. Casanova, H. C. Su; COVID Team, Severe COVID-19 in the young and healthy: Monogenic inborn errors of immunity? *Nat. Rev. Immunol.* **20**, 455–456 (2020). doi: 10.1038/s41577-020-0373-7; pmid: 32555547
41. Y. Itan *et al*., The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13615–13620 (2015). doi: 10.1073/pnas.1518646112; pmid: 26483451
42. S. L. Schwartz, G. L. Conn, RNA regulation of the antiviral protein 2′-5′-oligoadenylate synthetase. *Wiley Interdiscip. Rev. RNA* **10**, e1534 (2019). doi: 10.1002/wrna.1534; pmid: 30989826
43. B. Dong, R. H. Silverman, 2-5A-dependent RNase molecules dimerize during activation by 2-5A. *J. Biol. Chem.* **270**, 4133–4137 (1995). doi: 10.1074/jbc.270.8.4133; pmid: 7876164
44. M. Uhlén *et al*., Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). doi: 10.1126/science.1260419; pmid: 25613900
45. L. Zhao *et al*., Cell-type-specific activation of the oligoadenylate synthetase–RNase L pathway by a murine coronavirus. *J. Virol.* **87**, 8408–8418 (2013). doi: 10.1128/JVI.00769-13; pmid: 23698313

46. S. Banerjee, A. Chakrabarti, B. K. Jha, S. R. Weiss, R. H. Silverman, Cell-type-specific effects of RNase L on viral induction of beta interferon. *mBio* **5**, e00856-14 (2014). doi: 10.1128/mBio.00856-14; pmid: 24570368
47. F. Rapaport *et al*., Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2001248118 (2021). doi: 10.1073/pnas.2001248118; pmid: 33408250
48. B. Dong, M. Niwa, P. Walter, R. H. Silverman, Basis for regulated RNA cleavage by functional analysis of RNase L and Ire1p. *RNA* **7**, 361–373 (2001). doi: 10.1017/S1355838201002230; pmid: 11333017
49. R. H. Silverman, J. J. Skehel, T. C. James, D. H. Wreschner, I. M. Kerr, rRNA cleavage as an index of ppp(A2'p)nA activity in interferon-treated encephalomyocarditis virus-infected cells. *J. Virol.* **46**, 1051–1055 (1983). doi: 10.1128/jvi.46.3.1051-1055.1983; pmid: 6190010
50. D. H. Wreschner, T. C. James, R. H. Silverman, I. M. Kerr, Ribosomal RNA cleavage, nuclease activation and 2-5A (ppp(A2'p)$_n$A) in interferon-treated cells. *Nucleic Acids Res.* **9**, 1571–1581 (1981). doi: 10.1093/nar/9.7.1571; pmid: 6164990
51. Y. Xiang *et al*., Effects of RNase L mutations associated with prostate cancer on apoptosis induced by 2',5'-oligoadenylates. *Cancer Res.* **63**, 6795–6801 (2003). pmid: 14583476
52. L. A. Henderson *et al*., American College of Rheumatology clinical guidance for multisystem inflammatory syndrome in children associated with SARS-CoV-2 and hyperinflammation in pediatric COVID-19: version 3. *Arthritis Rheumatol.* **74**, e1–e20 (2022). doi: 10.1002/art.42062; pmid: 35118829
53. A. Wickenhagen *et al*., A prenylated dsRNA sensor protects against severe COVID-19. *Science* **374**, eabj3624 (2021). doi: 10.1126/science.abj3624; pmid: 34581622
54. O. Danziger, R. S. Patel, E. J. DeGrace, M. R. Rosen, B. R. Rosenberg, Inducible CRISPR activation screen for interferon-stimulated genes identifies OAS1 as a SARS-CoV-2 restriction factor. *PLOS Pathog.* **18**, e1010464 (2022). doi: 10.1371/journal.ppat.1010464; pmid: 35421191
55. Y. Li *et al*., SARS-CoV-2 induces double-stranded RNA-mediated innate immune responses in respiratory epithelial-derived cells and cardiomyocytes. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022643118 (2021). doi: 10.1073/pnas.2022643118; pmid: 33811184
56. P. Bastard *et al*., Herpes simplex encephalitis in a patient with a distinctive form of inherited IFNAR1 deficiency. *J. Clin. Invest.* **131**, e139980 (2021). doi: 10.1172/JCI139980; pmid: 32960813
57. X. Song *et al*., Little to no expression of angiotensin-converting enzyme-2 on most human peripheral blood immune cells but highly expressed on tissue macrophages. *Cytometry A* 10.1002/cyto.a.24285 (2020). doi: 10.1002/cyto.a.24285; pmid: 33280254
58. T. S. Rodrigues *et al*., Inflammasomes are activated in response to SARS-CoV-2 infection and are associated with COVID-19 severity in patients. *J. Exp. Med.* **218**, e20201707 (2021). doi: 10.1084/jem.20201707; pmid: 33231615
59. J. Zheng *et al*., Severe acute respiratory syndrome coronavirus 2–induced immune activation and death of monocyte-derived human macrophages and dendritic cells. *J. Infect. Dis.* **223**, 785–795 (2021). doi: 10.1093/infdis/jiaa753; pmid: 33277988
60. I. Gresser, F. Vignaux, F. Belardelli, M. G. Tovey, M. T. Maunoury, Injection of mice with antibody to mouse interferon alpha/beta decreases the level of 2'-5' oligoadenylate synthetase in peritoneal macrophages. *J. Virol.* **53**, 221–227 (1985). doi: 10.1128/jvi.53.1.221-227.1985; pmid: 2981340
61. W. Chanput, J. J. Mes, H. J. Wichers, THP-1 cell line: An in vitro cell model for immune modulation approach. *Int. Immunopharmacol.* **23**, 37–45 (2014). doi: 10.1016/j.intimp.2014.08.002; pmid: 25130606
62. W. B. Lee *et al*., OAS1 and OAS3 negatively regulate the expression of chemokines and interferon-responsive genes in human macrophages. *BMB Rep.* **52**, 133–138 (2019). doi: 10.5483/BMBRep.2019.52.2.129; pmid: 30078389
63. T. Kawai, S. Akira, Innate immune recognition of viral infection. *Nat. Immunol.* **7**, 131–137 (2006). doi: 10.1038/ni1303; pmid: 16424890
64. A. Chakrabarti, B. K. Jha, R. H. Silverman, New insights into the role of RNase L in innate immunity. *J. Interferon Cytokine Res.* **31**, 49–57 (2011). doi: 10.1089/jir.2010.0120; pmid: 21190483
65. A. G. Hovanessian, J. Wood, E. Meurs, L. Montagnier, Increased nuclease activity in cells treated with pppA2'p5'A2'p5' A. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 3261–3265 (1979). doi: 10.1073/pnas.76.7.3261; pmid: 114998
66. A. Zhou, B. A. Hassel, R. H. Silverman, Expression cloning of 2-5A-dependent RNAase: A uniquely regulated mediator of interferon action. *Cell* **72**, 753–765 (1993). doi: 10.1016/0092-8674(93)90403-D; pmid: 7680958
67. K. Malathi *et al*., A transcriptional signaling pathway in the IFN system mediated by 2'-5'-oligoadenylate activation of RNase L. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14533–14538 (2005). doi: 10.1073/pnas.0507551102; pmid: 16203993
68. J. M. Burke, S. L. Moon, T. Matheny, R. Parker, RNase L reprograms translation by widespread mRNA turnover escaped by antiviral mRNAs. *Mol. Cell* **75**, 1203–1217.e5 (2019). doi: 10.1016/j.molcel.2019.07.029; pmid: 31494035
69. M. Knight *et al*., Radioimmune, radiobinding and HPLC analysis of 2-5A and related oligonucleotides from intact cells. *Nature* **288**, 189–192 (1980). doi: 10.1038/288189a0; pmid: 6159552
70. A. Asthana, C. Gaughan, B. Dong, S. R. Weiss, R. H. Silverman, Specificity and mechanism of coronavirus, rotavirus, and mammalian two-histidine phosphoesterases that antagonize antiviral innate immunity. *mBio* **12**, e0178121 (2021). doi: 10.1128/mBio.01781-21; pmid: 34372695
71. A. Liberzon *et al*., The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015). doi: 10.1016/j.cels.2015.12.004; pmid: 26771021
72. E. J. Mendoza, K. Manguiat, H. Wood, M. Drebot, Two detailed plaque assay protocols for the quantification of infectious SARS-CoV-2. *Curr. Protoc. Microbiol.* **57**, ecpmc105 (2020). doi: 10.1002/cpmc.105; pmid: 32475066
73. A. C. G. Salina *et al*., Efferocytosis of SARS-CoV-2-infected dying cells impairs macrophage anti-inflammatory functions and clearance of apoptotic cells. *eLife* **11**, e74443 (2022). doi: 10.7554/eLife.74443; pmid: 35666101
74. S. Jin *et al*., Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021). doi: 10.1038/s41467-021-21246-9; pmid: 33597522
75. A. Chakrabarti *et al*., RNase L activates the NLRP3 inflammasome during viral infections. *Cell Host Microbe* **17**, 466–477 (2015). doi: 10.1016/j.chom.2015.02.010; pmid: 25816776
76. N. Magusali *et al*., A genetic link between risk for Alzheimer's disease and severe COVID-19 outcomes via the *OAS1* gene. *Brain* **144**, 3727–3741 (2021). doi: 10.1093/brain/awab337; pmid: 34619763
77. S. Zhou *et al*., A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667 (2021). doi: 10.1038/s41591-021-01281-1; pmid: 33633408
78. H. Zeberg, S. Pääbo, A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2026309118 (2021). doi: 10.1073/pnas.2026309118; pmid: 33593941
79. J. E. Huffman *et al*., Multi-ancestry fine mapping implicates *OAS1* splicing in risk of severe COVID-19. *Nat. Genet.* **54**, 125–127 (2022). doi: 10.1038/s41588-021-00996-8; pmid: 35027740
80. K. Cho *et al*., Heterozygous mutations in *OAS1* cause infantile-onset pulmonary alveolar proteinosis with hypogammaglobulinemia. *Am. J. Hum. Genet.* **102**, 480–486 (2018). doi: 10.1016/j.ajhg.2018.01.019; pmid: 29455859
81. T. Magg *et al*., Heterozygous *OAS1* gain-of-function variants cause an autoinflammatory immunodeficiency. *Sci. Immunol.* **6**, eabf9564 (2021). doi: 10.1126/sciimmunol.abf9564; pmid: 34145065
82. A. G. Hovanessian, R. E. Brown, I. M. Kerr, Synthesis of low molecular weight inhibitor of protein synthesis with enzyme from interferon-treated cells. *Nature* **268**, 537–540 (1977). doi: 10.1038/268537a0; pmid: 560630
83. I. M. Kerr, R. E. Brown, pppA2'p5'A2'p5'A: an inhibitor of protein synthesis synthesized with an enzyme fraction from interferon-treated cells. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 256–260 (1978). doi: 10.1073/pnas.75.1.256; pmid: 272640
84. M. Drappier, T. Michiels, Inhibition of the OAS/RNase L pathway by viruses. *Curr. Opin. Virol.* **15**, 19–26 (2015). doi: 10.1016/j.coviro.2015.07.002; pmid: 26231767
85. R. H. Silverman, Viral encounters with 2',5'-oligoadenylate synthetase and RNase L during the interferon antiviral response. *J. Virol.* **81**, 12720–12729 (2007). doi: 10.1128/JVI.01471-07; pmid: 17804500
86. R. J. Lin *et al*., Distinct antiviral roles for human 2',5'-oligoadenylate synthetase family members against dengue virus infection. *J. Immunol.* **183**, 8035–8043 (2009). doi: 10.4049/jimmunol.0902728; pmid: 19923450
87. Y. C. Kwon, J. I. Kang, S. B. Hwang, B. Y. Ahn, The ribonuclease L-dependent antiviral roles of human 2',5'-oligoadenylate synthetase family members against hepatitis C virus. *FEBS Lett.* **587**, 156–164 (2013). doi: 10.1016/j.febslet.2012.11.010; pmid: 23196181
88. Y. Li *et al*., Activation of RNase L is dependent on OAS3 expression during infection with diverse human viruses. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2241–2246 (2016). doi: 10.1073/pnas.1519657113; pmid: 26858407
89. J. N. Whelan, Y. Li, R. H. Silverman, S. R. Weiss, Zika virus production is resistant to RNase L antiviral activity. *J. Virol.* **93**, e00313-19 (2019). doi: 10.1128/JVI.00313-19; pmid: 31142667
90. C. S. Thakur, Z. Xu, Z. Wang, Z. Novince, R. H. Silverman, A convenient and sensitive fluorescence resonance energy transfer assay for RNase L and 2',5' oligoadenylates. *Methods Mol. Med.* **116**, 103–113 (2005). doi: 10.1385/1-59259-939-7:103; pmid: 16000857
91. K. Sacco *et al*, Multiomics approach identifies novel age-, time- and treatment-related immunopathological signatures in MIS-C and pediatric COVID-19, version 1.0.0, Zenodo (2021); https://doi.org/10.5281/zenodo.5524378.

ACKNOWLEDGMENTS

We thank the patients and their families for participating in our research. We thank all members of both branches of the Laboratory of Human Genetics of Infectious Diseases for discussions and technical and administrative support. We thank A. W. Ashbrook for managing the Rice laboratory BSL3 facility. We thank the Memorial Sloan Kettering Cancer Center for help isolating and sequencing the NY SARS-CoV-2 isolate. We thank R. Padgett and G. Stark at the Cleveland Clinic for discussions. We also thank A. Codina and C. Jou from the Biobanc de l'Hospital Infantil Sant Joan de Déu per a la Investigació, which is integrated into the Spanish Biobank Network of ISCIII, for sample and data procurement. **Funding:** The Laboratory of Human Genetics of Infectious Diseases is supported by the Howard Hughes Medical Institute, the Rockefeller University, the St. Giles Foundation, the National Institutes of Health (NIH) (R01AI088364 and R21AI160576), the National Center for Advancing Translational Sciences (NCATS), NIH Clinical and Translational Science Award (CTSA) program (UL1TR001866), the Yale Center for Mendelian Genomics and the GSP Coordinating Center funded by the National Human Genome Research Institute (NHGRI) (UM1HG006504 and U24HG008956), the Yale High-Performance Computing Center (S10OD018521), the Fisher Center for Alzheimer's Research Foundation, the Meyer Foundation, the JBP Foundation, the French National Research Agency (ANR) under the "Investments for the Future" program (ANR-10-IAHU-01), the Integrative Biology of Emerging Infectious Diseases Laboratory of Excellence (ANR-10-LABX-62-IBEID), the French Foundation for Medical Research (FRM) (EQU201903007798), the ANR GenMISC (ANR-21-COVR-039), the ANRS-COV05, ANR GENVIR (ANR-20-CE93-003) and ANR AABIFNCOV (ANR-20-CO11-0001) projects, the ANR-RHU program (ANR-21-RHUS-08), the European Union's Horizon 2020 research and innovation program under grant agreement 824110 (EASI-genomics), the HORIZON-HLTH-2021-DISEASE-04 program under grant agreement 01057100 (UNDINE), the ANR-RHU Program ANR-21-RHUS-08 (COVIFERON), the Square Foundation, Grandir – Fonds de solidarité pour l'enfance, the Fondation du Souffle, the SCOR Corporate Foundation for Science, the French Ministry of Higher Education, Research, and Innovation (MESRI-COVID-19), Institut National de la Santé et de la Recherche Médicale (INSERM), and Paris Cité University. We acknowledge support from the National Institute of Allergy and Infectious Diseases (NIAID) of the NIH under award R01AI104887 to R.H.S. and S.R.W. The Laboratory of Human Evolutionary Genetics (Institut Pasteur) is supported by the Institut Pasteur, the Collège de France, the French Government's Investissement d'Avenir program, Laboratoires d'Excellence "Integrative Biology of Emerging Infectious Diseases" (ANR-10-LABX-62-IBEID) and "Milieu Intérieur" (ANR-10-LABX-69-01), the Fondation de France (no. 00106080), the FRM (Equipe FRM DEQ20180339214 team), and the ANR COVID-19-POPCELL (ANR-21-CO14-0003-01). A.Puj. is supported by ACCI20-759 CIBERER, EasiGenomics H2020 Marató TV3 COVID 2021-31-33, the HORIZON-HLTH-2021-ID: 101057100 (UNDINE), the Horizon 2020 program under grant no. 824110 (EasiGenomics grant no. COVID-19/PID12342), and the CERCA Program/Generalitat de Catalunya. The Canarian Health System sequencing hub was funded by the Instituto de Salud Carlos III (COV20_01333 and COV20_01334), the Spanish Ministry of Science and Innovation (RTC-2017-6471-1; AEI/FEDER, UE), Fundación MAPFRE Guanarteme (OA21/131), and Cabildo Insular de Tenerife (CGIEU0000219140 and "Apuestas científicas del ITER para colaborar

en la lucha contra la COVID-19"). The CoV-Contact Cohort was funded by the French Ministry of Health and the European Commission (RECOVER project). Our studies are also funded by the Ministry of Health of the Czech Republic Conceptual Development of Research Organization (FNBr, 65269705) and ANID COVID0999 funding in Chile. G. Novelli and A. Novelli are supported by Regione Lazio (Research Group Projects 2020) No. A0375-2020-36663, GecoBiomark. A.M.P., M.L.D., and J.P.-T. are supported by the Inmungen-CoV2 project of CSIC. This work was supported in part by the Intramural Research Program of the NIAID, NIH. The research work of A.M.P, M.L.D., and J.P.-T. was funded by the European Commission –NextGenerationEU (Regulation EU 2020/2094), through CSIC's Global Health Platform (PTI Salud Global). I.M. is a senior clinical investigator at FWO Vlaanderen supported by a VIB GC PID grant, by FWO grants G0B5120N (DADA2) and G0E8420N, and by the Jeffrey Modell Foundation. I.M. holds an ERC-StG MORE2ADA2 grant and is also supported by ERN-RITA. A.Y. is supported by fellowships from the European Academy of Dermatology and Venereology and the Swiss National Science Foundation and by an Early Career Award from the Thrasher Research Fund. Y.-H.C. is supported by an A*STAR International Fellowship (AIF). M.O. was supported by the David Rockefeller Graduate Program, the New York Hideyo Noguchi Memorial Society (HNMS), the Funai Foundation for Information Technology (FFIT), the Honjo International Scholarship Foundation (HISF), and the National Cancer Institute (NCI) F99 Award (F99CA274708). A.A.A. was supported by Ministerio de Ciencia Tecnología e Innovación MINCIENCIAS, Colombia (111584467551/CT 415-2020). D.L. is supported by a fellowship from the FRM for medical residents and fellows. E.H. received funding from the Bank of Montreal Chair of Pediatric Immunology, Foundation of CHU Sainte-Justine, CIHR grants PCC-466901 and MM1-181123, and a Canadian Pediatric Society IMPACT study. Q.P.-H. received funding from the European Union's Horizon 2020 research and innovation program (ATAC, 101003650), the Swedish Research Council, and the Knut and Alice Wallenberg Foundation. Work in the Laboratory of Virology and Infectious Disease was supported by NIH grants P01AI138398-S1, 2U19AI111825, R01AI091707-10S1, and R01AI161444; a George Mason University Fast Grant; the G. Harold and Leila Y. Mathers Charitable Foundation; the Meyer Foundation; and the Bawd Foundation. R.P.L. is on the board of directors of both Roche and the Roche subsidiary Genentech. J.L.P. was supported by a Francois Wallace Monahan Postdoctoral Fellowship at the Rockefeller University and by a European Molecular Biology Organization Long-Term Fellowship (ALTF 380-2018). **Author contributions:** D.L., J.L.P., A.Y., B.D., Y.A., M.O., R.P., M.P., Z.L., L.B., A.Bi., W.L., M.H., J.C., C.G., A.A., V.L., J.M.L., F.J., H.-H.H., E.M., M.Mo., K.B., S.M., C.F., Y.Z., A.A.A., R.B., A.S., T.L.V., M.Ma., A.G., M.M.-V., F.P., T.L., R.L., A.-L.N., J.R., J.P., Y.-H.C., M.-P.M., R.M.P.-R., S.B., L.L., M.L.D., N.F., F.R., J.P.-T., S.C., T.E., F.G., P.L., S.R.W., A.M.P., C.L.D., J.B., A.Pue., S.B.-D., B.Bo., T.M., Q.Z., L.N., V.B., R.P.L., E.J., A.Be., L.Q.-M., C.M.R., R.H.S., S.-Y.Z., and J.-L.C. performed or supervised experiments, generated and analyzed data, and contributed to the manuscript by providing figures and tables. E.T., D.R., P.Z., Y.S., B.M., B.Bi., A.C., and L.Ab. performed computational analysis of data. I.J., S.E.B., G.I.B., C.B., J.A., S.D., J.T., F.B., V.F., D.B., X.D., Q.P.-H., I.M., F.H., A.Puj., V.S.-S., P.B., R.P.d.D., C.R.-G., H.C.S., L.Al., S.K., and E.H. evaluated and recruited patients to COVID and/or control cohorts of patients. D.L, S.-Y.Z., and J.-L.C. wrote the manuscript. S.-Y.Z. and J.-L.C. conceptualized and supervised the project. All authors edited the manuscript. **Competing interests:** E.H. received honoraria from CSL-Behring, Takeda, and Octapharma. R.H.S. is a consultant to Laronde, Inc., and Inception Therapeutics, Inc. H.C.S. is also affiliated with the Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania. R.H.S. is also affiliated with the Departments of Biochemistry, Molecular Biology and Microbiology, and Molecular Medicine, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, Ohio; the Departments of Biological, Geological, and Environmental Sciences and Chemistry, Cleveland State University, Cleveland, Ohio; and Ohio School of Biological Sciences, Kent State University, Kent, Ohio. The other authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. The materials and reagents used are commercially available and nonproprietary, with the exception of SARS-CoV-2 working stock and the gene-KO or patient-specific cell lines generated from this study. The SARS-CoV-2 working stock is available from C.M.R. under a material transfer agreement (MTA) with the Rockefeller University. The cell lines generated from this study are available from S.-Y.Z. and J.-L.C. upon request under MTAs from the Rockefeller University and the Imagine Institute. Patient-specific cellular materials from patients enrolled at NIAID are available from H.C.S. under a MTA with the NIH, provided that the request fulfills all articles listed in a MTA with the originating institute where the materials were collected. WGS data for patients sequenced by NIAID through TAGC were deposited under database of Genotypes and Phenotypes (dbGaP) accession number phs002245. Other genomic sequences of the patients reported in this paper are available from the authors upon request under a data transfer agreement. The raw RNA-seq data generated from this study are deposited in the NCBI database under the NCBI-SRA project PRJNA898284.

CoV-Contact Cohort

Loubna Alavoine[1], Sylvie Behillil[2], Charles Burdet[3], Charlotte Charpentier[3,4], Aline Dechanet[5], Diane Descamps[3,6], Xavier Duval[1,3,7], Jean-Luc Ecobichon[1], Vincent Enouf[8], Wahiba Frezouls[1], Nadhira Houhou[5], Ouifiya Kafif[5], Jonathan Lehacaut[1], Sophie Letrou[1], Bruno Lina[9], Jean-Christophe Lucet[10], Pauline Manchon[5], Mariama Nouroudine[1], Valentine Piquard[5], Caroline Quintin[1], Michael Thy[11], Sarah Tubiana[1], Sylvie van der Werf[8], Valérie Vignali[1], Benoit Visseaux[3,10], Yazdan Yazdanpanah[3,10], Abir Chahine[12], Nawal Waucquier[12], Maria-Claire Migaud[12], Dominique Deplanque[12], Félix Djossou[13], Mayka Mergeay-Fabre[14], Aude Lucarelli[15], Magalie Demar[13], Léa Bruneau[16], Patrick Gérardin[17], Adrien Maillot[16], Christine Payet[18], Bruno Lavielle[19], Fabrice Laine[19], Christophe Paris[19], Mireille Desille-Dugast[19], Julie Fouchard[19], Denis Malvy[20], Duc Nguyen[20], Thierry Pistone[20], Pauline Perreau[20], Valérie Gissot[21], Carole Le Goas[21], Samatha Montagne[22], Lucie Richard[23], Catherine Chirouze[24], Kévin Bouiller[24], Maxime Desmarets[25], Alexandre Meunier[26], Benjamin Lefèvre[27], Hélène Jeulin[28], Karine Legrand[29], Sandra Lomazzi[30], Bernard Tardy[31], Amandine Gagneux-Brunon[32], Frédérique Bertholon[33], Elisabeth Botelho-Nevers[32], Kouakam Christelle[34], Leturque Nicolas[34], Layidé Roufai[34], Karine Amat[35], Sandrine Couffin-Cadiergues[34], Hélène Espérou[36], Samia Hendou[34]

[1]Centre d'Investigation Clinique, INSERM CIC 1425, Hôpital Bichat Claude Bernard, AP-HP, Paris, France. [2]Institut Pasteur, Paris, France. [3]Université de Paris, IAME, INSERM U1137, Paris, France, Hôpital Bichat Claude Bernard, AP-HP, Paris, France. [4]Service de Virologie, Université de Paris, INSERM, IAME, UMR 1137, AP-HP, Hôpital Bichat Claude Bernard, F-75018 Paris, France. [5]Hôpital Bichat Claude Bernard, AP-HP, Paris, France. [6]IAME INSERM U1140, Hôpital Bichat Claude Bernard, AP-HP, Paris, France. [7]Centre d'Investigation Clinique, INSERM CIC 1425, AP-HP, IAME, Paris University, Paris, France. [8]Institut Pasteur, U3569 CNRS, Université de Paris, Paris, France. [9]Virpath Laboratory, International Center of Research in Infectiology, Lyon University, INSERM U1111, CNRS U5308, ENS, UCBL, Lyon, France. [10]IAME INSERM U1138, Hôpital Bichat Claude Bernard, AP-HP, Paris, France. [11]Center for Clinical Investigation, AP-HP, Hôpital Bichat Claude Bernard, Paris, France. [12]Centre d'Investigation Clinique, INSERM CIC 1403, Centre Hospitalo Universitaire de Lille, Lille, France. [13]Service des maladies infectieuses, Centre Hospitalo Universitaire de Cayenne, Guyane, France. [14]Centre d'Investigation Clinique, INSERM CIC 1424, Centre Hospitalier de Cayenne, Cayenne, Guyane Française. [15]Service Hôpital de jour Adulte, Centre Hospitalier de Cayenne, Guyane, France. [16]Centre d'Investigation Clinique, INSERM CIC 1410, Centre Hospitalo universitaire de la Réunion, La Réunion, France. [17]Centre d'Investigation Clinique, INSERM CIC 1410, CHU Reunion, Saint-Pierre, Reunion Island. [18]Centre d'Investigation Clinique, INSERM CIC 1410, Centre de Ressources Biologiques, Centre Hospitalo universitaire de la Réunion, La Réunion, France. [19]Centre d'Investigation Clinique, INSERM CIC 1414, Centre Hospitalo universitaire de Rennes, Rennes, France. [20]Service des maladies infectieuses, Centre Hospitalo universitaire de Bordeaux, Bordeaux, France. [21]Centre d'Investigation Clinique, INSERM CIC 1415, CHRU Tours, Tours, France. [22]CRBT, Centre Hospitalo universitaire de Tours, Tours, France. [23]Pole de Biologie Médicale, Centre Hospitalo universitaire de Tours, Tours, France. [24]Service des maladies infectieuses, Centre Hospitalo universitaire de Besançon, Besançon, France. [25]Service des maladies infectieuses, Centre d'investigation clinique, INSERM CIC1431, Centre Hospitalier Universitaire de Besançon, Besançon, France. [26]Centre de Ressources Biologiques - Filière Microbiologique de Besançon, Centre Hospitalier Universitaire, Besançon, France. [27]Université de Lorraine, CHRU-Nancy and APEMAC, Infectious and Tropical Diseases, Nancy, France. [28]Laboratoire de Virologie, CHRU de Nancy Brabois, Vandoeuvre-lès-Nancy, France. [29]INSERM CIC-EC 1433, Centre Hospitalo universitaire de Nancy, Nancy, France. [30]Centre de ressources Biologiques, Centre Hospitalo universitaire de Nancy, Nancy, France. [31]Centre d'Investigation Clinique, INSERM CIC 1408, Centre Hospitalo universitaire de Saint-Étienne, Saint-Étienne, France. [32]Service des maladies infectieuses, Centre Hospitalo universitaire de Saint-Étienne, Saint-Étienne, France. [33]Service des maladies infectieuses, CRB42-BTK, Centre Hospitalo Universitaire de Saint-Étienne, Saint-Étienne, France. [34]Pole Recherche Clinique, INSERM, Paris, France. [35]IMEA Fondation Léon M'Ba, Paris, France. [36]INSERM Pôle Recherche Clinique, Paris, France.

COVID Human Genetic Effort

Laurent Abel[1], Hassan Abolhassani[2], Sergio Aguilera-Albesa[3], Alessandro Aiuti[4], Ozge Metin Akcan[5], Nihal Akcay[6], Gulsum Alkan[7], Suzan A. Alkhater[8], Luis Miguel Allende[9], Yosunkaya Alper[5], Naima Amenzoui[10], Mark S. Anderson[11], Lisa Arkin[12], Melodie Aubart[13], Iryna Avramenko[14], Şehnaz Aydemir[15], Zeynep Gökçe Gayretli Aydin[16], Caner Aytekin[17], Gökhan Aytekin[18], Selma Erol Aytekin[5], Silvia Yumi Bando[19], Kathie Beland[20], Serkan Belkaya[21], Catherine M. Biggs[22], Agurtzane Bilbao Aburto[23], Geraldine Blanchard-Rohner[24], Daniel Blázquez-Gamero[9], Marketa Bloomfield[25], Dusan Bogunovic[26], Anastasia Bondarenko[27], Alessandro Borghesi[28], Amed Aziz Bousfiha[29], Oksana Boyarchuk[30], Petter Brodin[31], Yenan Bryceson[32], Giorgia Bucciol[33], Valeria Calcaterra[34], Giorgio Casari[4], Andre Cavalcanti[35], Jale Bengi Celik[36], George P. Chrousos[37], Roger Colobran[38], Antonio Condino-Neto[39], Francesca Conti[40], Megan Cooper[41], Taner Coskuner[42], Cyril Cyrus[43], Enza D'Auria[44], Selket Delafontaine[45], Beth A. Drolet[12], Burcu Bursal Duramaz[46], Loubna El Zein[47], Marwa H. Elnagdy[48], Melike Emiroglu[7], Emine Hafize Erdeniz[49], Marianna Fabi[50], Hagit Baris Feldman[51], Jacques Fellay[52], Filip Fencl[25], Filippos Filippatos[53], Julie Freiss[54], Jiri Fremuth[55], Alenka Gagro[56], Blanca Garcia-Solis[57], Gianluca Vergine[58], Rafaela González-Montelongo[59], Yahya Gul[60], Belgin Gülhan[61], Sara Sebnem Kilic Gultekin[62], Marta Gut[63], Rabih Halwani[64], Lennart Hammarström[65], Nevin Hatipoğlu[66], James Heath[67], Sarah E. Henrickson[68], Elisa Hernandez-Brito[69], Ilse Hoffman[70], Levi Hoste[71], Elena Hsieh[72], Antonio Íñigo-Campos[59], Yuval Itan[73], Petr Jabandziev[74], Bahar Kandemir[60], Saliha Kanık-Yüksek[61], Hasan Kapakli[75], Adem Karbuz[76], Ozgur Kasapcopur[77], Robin Kechiche[78], Yasemin Kendir Demirkol[79], Omer Kilic[80], Stella Kim Hansen[81], Adam Klocperk[25], Yu-Lung Lau[82], Jan Leb[25], José M. Lorenzo-Salazar[59], Carrie L. Lucas[83], Majistor Maglorius[84], Laura Marque[85], Yeray Novoa Medina[86], Abián Montesdeoca Melián[87], Alexios-Fotios A. Mentis[37], Michele T. Pato[81], Athanasios Michos[53], Joshua D. Milner[88], Trine H. Mogensen[89], Adrián Muñoz-Barrera[59], Serdar Nepesov[90], João Farela Neves[91], Ashley Ng[12], Lisa F. P. Ng[92], Antonio Novelli[93], Giuseppe Novelli[94], Fatma Nur Oz[95], J. Gonzalo Ocejo-Viñals[96], Satoshi Okada[97], Zerrin Orbak[98], Ahmet Osman Kilic[60], Hind Ouair[29], Şadiye Kübra Tüter Öz[7], Tayfun Özçelik[99], Esra Akyüz Özkan[49], Aslınur Özkaya Parlakay[100], Carlos N. Pato[80], Estela Paz-Artal[9], Simon Pelham[101], Isabelle Pellier[54], Quentin Philippot[84], Laura Planas-Serra[102], Samira Plassart[103], Petra Pokorna[104], Meltem Polat[95], Cecilia Poli[105], Carolina Prando[106], Laurent Renia[107], Jacques G. Rivière[108], Agustí Rodríguez-Palmero[109], Lucie Roussel[110], Luis A. Rubio-Rodriguez[59], Moro Salifu[81], Lumir Sasek[55], Laura Sasia[111], Anna Scherbina[112], Erica Schmitt[41], Anna Sediva[25], Esra Sevketoglu[113], Katerina Slaba[74], Ondrej Slaby[114], Ali Sobh[115], Jordi Solé-Violán[116], Pere Soler-Palacin[108], Lien De Somer[117], Betül Sözeri[42], András N. Spaan[118], Yuriy Stepanovskiy[27], Stuart G. Tangye[119], Gonul Tanir[95], Elizabeth-Barbara Tatsi[53], Christian W. Thorball[120], Selda Hancerli Torun[121], Stuart Turvey[22], Ahmad Abou Tayoun[122], Sathishkumar Ramaswamy[123], Mohammed J. Uddin[124], Emel Uyar[61], Juan Valencia-Ramos[125], Ana Maria Van Den Rym[57], Hulya Vatansev[60], Martín Castillo de Vera[126], François Vermeulen[33], Donald C. Vinh[110], Alla Volokha[27], Horst von Bernuth[127], Carine Wouters[33], Aysun Yahşi[61], Volkan Yarar[75], Osman Yesilbas[128], Mehmet Yıldız[77], Mayana Zatz[129], Pawel Zawadzki[130], Gianvincenzo Zuccotti[131], Shen-Ying Zhang[132], Jean-Laurent Casanova[133]

[1]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, Paris, France. [2]Department of Biosciences and Nutrition, Karolinska Institutet, SE14183, Huddinge, Sweden.

[3]Pediatrics Department, Navarra Health Service Hospital, Pamplona, Spain. [4]Pediatric Immunohematology, San Raffaele Hospital, Salute San Raffaele University, Italy. [5]Necmettin Erbakan University, Konya, Turkey. [6]Bakirkoy Dr. Sadi Konuk Research and Training Hospital, Pediatric Intensive Care Unit, Istanbul, Turkey. [7]Division of Pediatric Infectious Diseases, Department of Pediatrics, Selcuk University Faculty of Medicine, Konya, Turkey. [8]College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia; Department of Pediatrics, King Fahad Hospital of the University, Al-Khobar, Saudi Arabia. [9]Department of Pediatrics, Hospital Universitario 12 de Octubre, Madrid, Spain. [10]Children Infectious and Clinical Immunology Department, Abderrahim Harouchi Hospital, Faculty of Medicine and Pharmacy, Averroes University Hospital, Hassan 2 University, Casablanca, Morocco. [11]Diabetes Center, University of California, San Francisco, CA, USA. [12]University of Wisconsin School of Medicine, Madison, WI, USA. [13]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163; Pediatric Neurology Department, Necker-Enfants Malades Hospital, AP-HP, Paris, France. [14]Department of Propedeutics of Pediatrics and Medical Genetics, Danylo Halytsky Lviv National Medical University, Lviv, Ukraine. [15]Dr. Ali Kemal Belviranli State Hospital, Konya Turkey. [16]Department of Pediatrics, Division of Pediatric Infectious Disease, Faculty of Medicine, Karadeniz Technical University, Trabzon, Turkey. [17]Department of Pediatric Immunology, Dr. Sami Ulus Maternity and Children's Health and Diseases Training and Research Hospital, Ankara, Turkey. [18]Konya City Hospital, Konya, Turkey. [19]Laboratory of Pediatric Genomics, Faculty of Medicine, University of Sao Paulo, Sao Paulo, Brazil. [20]CHU Sainte-Justine, Montreal, QC, Canada. [21]Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey. [22]Department of Pediatrics, University of British Columbia, Vancouver, BC, Canada; BC Children's Hospital Research Institute, Vancouver, BC, Canada. [23]Servicio de Pediatría, Hospital Universitario Cruces, Spain. [24]Unit of Immunology and Vaccinology, Division of General Pediatrics, Department of Pediatrics, Gynecology and Obstetrics, Geneva University Hospitals, University of Geneva, Geneva, Switzerland. [25]Department of Pediatrics, 2nd Faculty of Medicine, Charles University in Prague and Motol University Hospital, Prague, Czech Republic. [26]Center for Inborn Errors of Immunity, Precision Immunology Institute, Mindich Child Health and Development Institute, Department of Microbiology, Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [27]Pediatric Infectious Disease and Pediatric Immunology Department, Shupyk National Healthcare University of Ukraine, Kyiv, Ukraine. [28]Neonatal Intensive Care Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; Fellay Lab, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [29]Clinical Immunology Unit, Casablanca Children's Hospital, Ibn Rochd Medical School, King Hassan II University, Casablanca, Morocco. [30]Department of Children's Diseases and Pediatric Surgery, I. Horbachevsky Ternopil National Medical University, Ukraine. [31]Science for Life Laboratory, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden. [32]Center for Hematology and Regenerative Medicine, Department of Medicine, Karolinska Institute, Stockholm, Sweden. [33]Department of Pediatrics, University Hospitals Leuven, Belgium. [34]Department of Pediatrics, Buzzi Children's Hospital, Milan, Italy; Department of Internal Medicine, University of Pavia, Pavia, Italy. [35]Department of Pediatrics, Clinical Hospital of Federal University of Pemambuco, Recife, Brazil. [36]Department of Anesthesiology and Reanimation, Selcuk University Faculty of Medicine, Konya, Turkey. [37]University Research Institute of Maternal and Child Health and Precision Medicine, National and Kapodistrian University of Athens, Athens, Greece. [38]Immunology Division, Genetics Department, Vall d'Hebron Research Institute, Vall d'Hebron Barcelona Hospital Campus, Universitat Autònoma de Barcelona, Barcelona, Spain. [39]Department of Immunology, Institute of Biomedical Sciences, University of Sao Paulo, Sao Paulo, Brazil. [40]Pediatric Unit-IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy. [41]Division of Rheumatology and Immunology, Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, USA. [42]Division of Pediatric Rheumatology, Umraniye Training and Research Hospital, University of Health Sciences, Istanbul, Turkey. [43]Department of Biochemistry, College of Medicine, Imam Abdulrahman Bin Faisal University, Saudi Arabia. [44]Department of Pediatrics, Buzzi Children's Hospital, Milan, Italy. [45]Department of Pediatrics, University Hospitals Leuven, Laboratory for Inborn Errors of Immunity, KU Leuven, Leuven, Belgium. [46]University of Health Sciences, Kanuni Sultan Suleyman Training and Research Hospital, Istanbul, Turkey. [47]Lebanese University, Faculty of Sciences I, Biology Department, Rafic Hariri Campus, Beirut, Lebanon. [48]Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Mansoura University, Mansoura, Egypt. [49]Ondokuz Mayis University, Samsun, Turkey. [50]Pediatric Emergency Unit, Scientific Institute for Research and Healthcare (IRCCS), Sant'Orsola Hospital, Bologna, Italy. [51]The Genetics Institute, Tel Aviv Sourasky Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [52]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; Precision Medicine Unit, Biomedical Data Sciences Center, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland; Swiss Institute of Bioinformatics, Lausanne, Switzerland. [53]First Department of Pediatrics, National and Kapodistrian University of Athens, Athens, Greece; University Research Institute of Maternal and Child Health and Precision Medicine, National and Kapodistrian University of Athens, Athens, Greece. [54]Unité d'Hématologie-Immunologie-Oncologie Pédiatrique, Centre Hospitalier Universitaire d'Angers, Angers, France. [55]Department of Pediatrics - PICU, Faculty of Medicine in Pilsen, Charles University in Prague, Czech Republic. [56]Department of Pediatrics, University of Zagreb School of Medicine, Children's Hospital Zagreb, Zagreb, Josip Juraj Strossmayer University of Osijek, Medical Faculty Osijek, Osijek, Croatia. [57]Laboratory of Immunogenetics of Human Diseases, IdiPAZ Institute for Health Research, La Paz Hospital, Madrid, Spain. [58]Unità Operativa Complessa Pediatria, Ospedale degli Infermi di Rimini, Rimini, Italy. [59]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain. [60]Necmettin Erbakan University, Konya, Turkey. [61]Ankara City Hospital, Ankara, Turkey. [62]Pediatric Immunology-Rheumatology Division, Uludag University Medical Faculty, Department of Pediatrics, Bursa, Turkey. [63]CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [64]College of Medicine, University of Shajah, Shajah, UAE. [65]Division of Clinical Immunology, Department of Laboratory Medicine, Karolinska University Hospital Huddinge, Stockholm, Sweden. [66]Pediatric Infectious Diseases Unit, Bakirkoy Dr. Sadi Konuk Training and Research Hospital, University of Health Sciences, Istanbul, Turkey. [67]Institute of Systems Biology, Seattle, WA, USA. [68]Children's Hospital of Philadelphia, Division of Allergy Immunology, Philadelphia, PA, USA; Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [69]Department of Immunology, Hospital Universitario de Gran Canaria Dr. Negrín, Canarian Health System, Las Palmas de Gran Canaria, Spain. [70]Department of Pediatrics, University Hospitals Leuven, Belgium. [71]Primary Immunodeficiency Research Lab, Center for Primary Immunodeficiency Ghent, Ghent University Hospital, Ghent, Belgium. [72]Department of Pediatrics, Section of Allergy and Immunology, Department of Immunology and Microbiology, University of Colorado Anschutz Medical Campus, Children's Hospital Colorado, Aurora, CO, USA. [73]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [74]Department of Pediatrics, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic. [75]Balikesir City Hospital, Balikesir, Turkey. [76]Prof. Dr. Cemil Taşcıoglu City Hospital, Istanbul, Turkey. [77]Department of Pediatric Rheumatology, Cerrahpasa Medical School, Istanbul University-Cerrahpasa, Istanbul, Turkey. [78]Service de Rhumatologie Pediatrique, CHU Bicetre, France. [79]Department of Pediatric Genetics, Health Sciences University, Umraniye Education and Research Hospital, Istanbul, Turkey. [80]Eskişehir Osmangazi University, Faculty of Medicine, Clinic of Pediatric Infectious Diseases, Eskişehir, Turkey. [81]Institute for Genomic Health, SUNY Downstate, Health Science University, Brooklyn, NY, USA. [82]Department of Paediatrics and Adolescent Medicine, LKS Faculty of Medicine, the University of Hong Kong, Hong Kong. [83]Department of Immunobiology, Yale University School of Medicine, New Haven, CT, USA. [84]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children; Paris Descartes University, Imagine Institute, Paris, France. [85]Immunodeficiencies and Infectious Diseases Unit, Pediatric Department, Centro Materno-Infantil do Norte, Centro Hospitalar Universitario do Porto, Porto, Portugal. [86]Department of Pediatrics, Complejo Hospitalario Universitario Insular-Materno Infantil, Canarian Health System, Las Palmas de Gran Canaria, Spain. [87]Guanarteme Health Care Center, Canarian Health System, Las Palmas de Gran Canaria, Spain. [88]Department of Pediatrics, Columbia University Irving Medical Center, New York, NY, USA. [89]Department of Biomedicine, Aarhus University, Aarhus, Denmark. [90]Pediatric Allergy and Immunology Unit, Istanbul Medipol Üniversity, Istanbul, Turkey. [91]Primary Immunodeficiencies Unit, Hospital Dona Estefânia, CHULC, EPE; CEDOC, Center for Chronic Diseases, Lisbon, Portugal. [92]A*STAR Infectious Disease Labs, Agency for Science, Technology and Research, Singapore. [93]Translational Cytogenomics Research Unit, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy. [94]Department of Biomedicine and Prevention, Tor Vergata University of Rome, Italy. [95]Department of Pediatric Infectious Disease, SBU Ankara Dr. Sami Ulus Maternity Child Health and Diseases Training and Research Hospital, Ankara, Turkey. [96]Department of Immunology, Hospital Marques de Valdecilla, Santander, Spain. [97]Department of Pediatrics, Hiroshima University Graduate School of Biomedical and Health Sciences, Hiroshima, Japan. [98]Department of Pediatrics, Ataturk University, Erzurum, Turkey. [99]Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey. [100]Yildirim Beyazit University, Ankara City Hospital, Ankara, Turkey. [101]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA. [102]Neurometabolic Diseases Laboratory, IDIBELL–Hospital Duran I Reynals; CIBERER U759, ISIiii Madrid, Spain. [103]Service de Rhumatologie pédiatrique, Hôpital Femme-Mère-Enfant, Groupement Hospitalier Est - Bâtiment "Pinel", Bron, France. [104]Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [105]Immunogenetics and Translational Immunology Program, Instituto de Ciencias e Innovación en Medicina (ICIM), School of Medicine, Clínica Alemana-Universidad del Desarrollo, Santiago de Chile, Chile. [106]Faculdades Pequeno Príncipe, Instituto de Pesquisa Pelé Pequeno Príncipe, Curitiba, Brazil. [107]A*STAR Infectious Disease Labs, Agency for Science, Technology and Research, Singapore; Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore; School of Biological Sciences, Nanyang. Technological University, Singapore. [108]Pediatric Infectious Diseases and Immunodeficiencies Unit, Vall d'Hebron Research Institute, Vall d'Hebron Barcelona Hospital Campus, Universitat Autònoma de Barcelona, Barcelona, Spain. [109]Palmero Pediatrics Department, University Hospital Germans Trias i Pujol, Badalona, Barcelona, Spain. [110]Department of Medicine, Division of Infectious Diseases, McGill University Health Centre, Montréal, QC, Canada; Infectious Disease Susceptibility Program, Research Institute, McGill University Health Centre, Montréal, QC, Canada. [111]Hospital Infantil Municipal de Córdoba, Córdoba, Argentina. [112]Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Moscow, Russia. [113]Pediatric Intensive Care Unit, Bakirkoy Dr. Sadi Konuk Training and Research Hospital, University of Health Sciences, Istanbul, Turkey. [114]Department of Biology UKB Kamenice, Masaryk University / Faculty of Medicine, Brno, Czech Republic. [115]Department of Pediatrics, Mansoura University Children's Hospital, Faculty of Medicine, Mansoura University, Mansoura, Egypt. [116]Critical Care Unit, Hospital Universitario de Gran Canaria Dr. Negrín, Canarian Health System, Las Palmas de Gran Canaria, Spain; Universidad Fernando Pessoa, Canarias, Spain; CIBER de Enfermedades Respiratorias. Instituto de Salud Carlos III, Madrid, Spain. [117]Department of Pediatrics, University Hospitals Leuven, Leuven, Belgium. [118]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA; Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, Netherlands. [119]Garvan Institute of Medical Research, Darlinghurst, NSW, Australia; St Vincent's Clinical School, Faculty of Medicine, UNSW Sydney, Sydney, NSW, Australia. [120]Precision Medicine Unit, Biomedical Data Sciences Center, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland. [121]Department of Pediatric Infectious Disease, Faculty of Medicine, Istanbul University, Istanbul, Turkey. [122]Genomics Center, Al Jalila Children's Specialty Hospital, Dubai, UAE; Center for Genomic Discovery, Mohammed Bin Rashid University, Dubai, UAE. [123]Genomics Center, Al Jalila Children's Specialty Hospital, Dubai, UAE. [124]College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE; Cellular Intelligence (Ci) Lab, GenomeArc Inc., Toronto, ON, Canada. [125]Pediatrics Department, Division Pediatric Intensive Care Unit, Hospital Universitario de Burgos, Burgos, Spain. [126]Doctoral Health Care Center, Canarian Health System, Las Palmas de Gran Canaria, Spain. [127]Department of Pediatric Pneumology, Immunology and Intensive Care, Charité

Universitätsmedizin, Berlin University Hospital Center, Berlin, Germany; Labor Berlin GmbH, Department of Immunology, Berlin, Germany; Berlin Institutes of Health (BIH), Berlin-Brandenburg Center for Regenerative Therapies, Berlin, Germany. [128]Department of Pediatrics, Division of Pediatric Critical Care Medicine, Faculty of Medicine, Karadeniz Technical University, Trabzon, Turkey. [129]Biosciences Institute, University of São Paulo, São Paulo, Brazil. [130]MNM Diagnostics, Poznań, Poland. [131]Department of Pediatrics, Buzzi Children's Hospital, Milan, Italy; Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy. [132]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, Paris, France; Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA. [133]Necker Hospital for Sick Children and INSERM, Paris, France; The Rockefeller University and Howard Hughes Medical Institute, New York, NY, USA.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo3627
Materials and Methods
Figs. S1 to S9
Tables S1 to S3
References (*92*–*100*)
MDAR Reproducibility Checklist
Data S1 and S2

View/request a protocol for this paper from *Bio-protocol*.

Submitted 29 January 2022; resubmitted 16 August 2022
Accepted 14 December 2022
Published online 20 December 2022
10.1126/science.abo3627

Part III

Discussion and perspectives

As a graduate student of Institut Pasteur's Human Evolutionary Genetics (HEG) Unit, I was directly involved in three single-cell genomics projects aimed at disentangling variability in the antiviral immune response across healthy individuals, or patients with inborn errors of immunity affecting the response to viruses. Across these projects, I collaborated with a diverse array of researchers—from virologists and evolutionary geneticists to clinicians—both at the bench and on the computational side. After three years of work, all of these collaborations have bore their fruit, strengthening our understanding of the bases of human immune variability in health and disease.

Myeloid cellular predictors of the susceptibility to viral infection

During my first months in the team, I contributed to the work led by Mary O'Neill et al. (2021), assessing the cellular determinants of susceptibility to infection by the influenza A virus (IAV) in myeloid cells. This is a milestone in the HEG Unit because it represents the first application of single-cell RNA-sequencing (scRNA-seq) by the team. While the experiments had already been performed when I joined the Unit in January 2020, I contributed to setting up and testing the scRNA-seq data analysis pipelines that were used in this study, and that have served as a basis of reference for subsequent 'single-cell' work (Lee et al., 2022; Aquino et al., 2023).

Through analyses of scRNA-seq data, we characterized the response of monocyte subsets to IAV infection over time and linked differences in monocyte infection rates to variability in the induction of interferon (IFN responses: infection escapees feature higher IFN-stimulated gene (ISG) expression relative to infected cells. We showed that low IAV-transcribers can be identified early after infection and display higher basal ISG expression, leading us to hypothesize a role for basal ISG activation in driving inter-individual variation in myeloid susceptibility to IAV infection. Finally, we used these results to characterize the drivers of population variation of myeloid responses to viral infection and highlight differences in cellular susceptibility to IAV infection between individuals of African and European origin, in line with reports by Randolph et al. (2021).

Although we did not have a large enough sample to make reasonably powered population-level inferences in O'Neill et al. (2021), the suggestion that population differences in susceptibility to viral infection could be driven by variation in the basal proportion of myeloid cells amenable to infection is especially interesting. Knowing that individuals of African origin tend to have higher proportions of $CD16^+$ monocytes (Quach et al., 2016), the biological relevance of this hypothesis was emphasized by epidemiological and clinical data collected during the 'coronavirus disease 2019' (COVID-19) pandemic, linking African ancestry to increased COVID-19 severity (Shelton et al., 2021), and COVID-19 severity to increased $CD16^+$ monocyte proportions (Zhou et al., 2020b).

Natural variability of single-cell transcriptional immune responses to viruses

As the COVID-19 pandemic kicked into high gear over the year 2020, I and other colleagues from the HEG Unit led a collaborative effort—spread across several research institutes worldwide, but centered around Institut Pasteur's biosafety level 3 facilities and our team's growing expertise in single-cell genomics—to dissentangle the genetic, nongenetic and evolutionary drivers of variability in the susceptibility to and the severity of 'severe acute respiratory syndrome' coronavirus 2 (SARS-CoV-2) infection. Namely, Aquino et al. (2023) represents the first systematic assessment of the genetic and nongenetic basis of immune variability in the response to SARS-CoV-2 across humans populations at single-cell resolution. For instance, while Randolph et al. (2021) had previously revealed ancestry-related differences in single peripheral blood mononuclear cell (PBMC) transcriptional responses to viral infection across African-American and European-American healthy donors, the authors focused on IAV as an infectious model. Also, other studies of transcriptional immune responses had identified hallmarks of SARS-CoV-2 responses, including impaired IFN induction

(Hadjadj et al., 2020) and an exacerbated myeloid inflammatory component (Leon et al., 2022), but were not performed at single-cell resolution. Finally, while previous single-cell studies in severe COVID-19 contexts had characterized transcriptional responses across several PBMC types at high resolution (Wilk et al., 2020) and compared the responses to SARS-CoV-2 and IAV (Lee et al., 2020), their relatively small samples precluded population-level inferences.

Availing of our large sample size, we dissected the respective contributions of cellular composition and genetic factors to population variation in transcriptional immune responses to SARS-CoV-2 and IAV across over 20 immune cell types sampled from diverse and healthy individuals. Importantly, we found that variation in relative immune cell abundances explains most gene expression differences between human populations, while genetic effects are generally stronger albeit limited to targeted gene subsets in each cell type.

Differences in cellular composition can also explain differing COVID-19 courses. For instance, previous single-cell descriptions of transcriptional immune responses in the peripheral blood of COVID-19 patients have linked severe outcomes of SARS-CoV-2 infection to an increased abundance of 'exhausted' natural killer (NK) cells expressing *LAG3*, *PDCD1* and *HAVCR2* (Wilk et al., 2020). We show that variation between individuals of African and European ancestry in the abundance of memory-like NK cells (Ram et al., 2018) with similar transcriptional profiles can be largely explained by latent infection by cytomegalovirus (CMV), which has been recently pointed out as a marker of COVID-19 severity, even in relatively young individuals (Weber et al., 2022).

These results suggest that differences in immune cell composition could partially explain the reported disparities in severe COVID-19 risk between individuals of African and European ancestry (Shelton et al., 2021) by mediating differential exposure to CMV in Central Africa and West Europe; these people respond to viral infection differently because they live in different places and are exposed to different pathogens. More generally, our results also highlight the relevance of scRNA-seq to accurately characterize cell type composition, which serves as a useful proxy to model differential environmental exposures in population-level studies of immune variation.

Yet, it is certain that genetic variation also plays a role in severe COVID-19 risk disparities worldwide, as attested by the wealth of genome-wide association results produced through several large collaborative efforts spurred by the pandemic (COVID-19 Host Genetics Initiative, 2020, 2021, 2022, 2023; Ellinghaus et al., 2020; Pairo-Castineira et al., 2021; Shelton et al., 2021; Kousathanas et al., 2022; Horowitz et al., 2022). Through context-specific colocalization and transcriptome-wide association tests, our work builds on this knowledge base to inform on the likely cell type and condition in which this genetic control takes place, thus inching towards causal inference of genetic effects on COVID-19 susceptibility and severity. For example, we show data suggesting that the effect of the likely Neandertal-origin rs10774679 quantitative trait locus (QTL) on COVID-19 hospitalization risk could be mediated by its effect on the expression of *OAS3* in $CD16^{+}$ monocytes exposed to SARS-CoV-2.

Another distinctive feature of our approach is the use of evolutionary genetics methods to retrace the evolutionary history of the QTLs associated to gene expression (e) or changes in expression in response (r) to viral stimulation. More specifically, we tested eQTL and reQTL variants for patterns of allele frequency differentiation—either across human populations or through time—that could result from natural selection, as well as for evidence of introgression from archaic hominin genomes.

Assessing the antiviral response as a substrate for natural selection, our most remarkable finding is perhaps the fact that selection has preferentially affected reQTL variants that control the reponse to SARS-CoV-2, but not IAV, specifically in genomes of East Asian descent. Motivated by previous reports (Souilmi et al., 2021) of local adaptation to coronavirus-related pressures in East Asia around 25 thousand years ago—in coincidence with the appeareance of the ancestors of SARS-CoV-2 in

the region—we built on our results by estimating the time frame within the last 56 thousand years during which allele frequencies at these loci changed more rapidly than expected by chance. Mirroring the reports by Souilmi et al. (2021), we found that rapid allele frequency changes at SARS-CoV-2-specific reQTLs between 21 to 27 thousand years ago were 2.6 times more likely to have happened in East Asian genomes, relative to Central African and West European genomes.

We also found striking evidence supporting a widespread effect of archaic introgression on gene expression in the genomes of Eurasians, where Neandertal haplotypes are up to 1.5 times more likely to harbor eQTLs compared to random matched variants, and regulatory Neandertal haplotypes are more frequent than archaic introgressed segments without eQTLs. Furthermore, we report novel signals of archaic introgression, including the rs58964929 reQTL of *UBE2F*—which encodes an important protein for the nuclear translocation of IFN regulatory factor (IRF) 7 in myeloid cells stimulated with RNA viruses—in $CD14^+$ monocytes exposed to SARS-CoV-2 and IAV.

Overall, the results in Aquino et al. (2023) draw a panorama of genetic factors with complex evolutionary histories that contribute to present-day disparities in COVID-19 risk through their effects on gene expression across different immune cell types. However, in line with our results from O'Neill et al. (2021), we also emphasize the role of cellular composition differences—which can mediate the effects of different environmental exposures, like latent viral infection—on human population differences in the transcriptional immune response to viruses among healthy individuals.

Inborn errors of immunity to map the genetic basis of susceptibility to infection

Natural immune variability across healthy individuals and populations is only one side of a larger picture. As extensively discussed by Casanova and Abel (2005, 2013, 2020, 2021, 2022), data from patients carrying inborn errors of immunity (IEIs) are useful to gain insight into the so-called 'infection enigma', whereby some individuals can resist infection by even the deadliest pathogens while others are especially susceptible to severe infection by generally innocuous microbes.

In the context of the COVID-19 pandemic, and through collaborations with Casanova and Abel's research groups at Institut Imagine and The Rockefeller University in New York, I assessed the impact of IEIs in the 2'-5'-oligoadenylate synthetase (OAS)-RNase L pathway—a major component of the IFN-mediated antiviral immune response—on multisystem inflammatory syndrome in children (MIS-C), a severe phenotype associated to SARS-CoV-2 infection.

Knowing that inborn errors of IFN-mediated immunity explain increased susceptibility to severe infection by IAV (Ciancanelli et al., 2015; Hernández et al., 2018; Lim et al., 2019) and SARS-CoV-2 (Zhang et al., 2020, 2022), Lee et al. (2022) screened MIS-C patients in search of IEIs and found several deficiencies touching different levels of the OAS-RNase L pathway. Aware of the HEG Unit's expertise in single-cell genomics and our ability to work with live SARS-CoV-2, the authors then reached out to us in order to dig further into the cellular and molecular determinants of immune-mediated MIS-C in OAS-RNase-L-deficient children.

Through analyses of scRNA-seq data from PBMCs sampled from four pediatric MIS-C patients and stimulated with SARS-CoV-2, I identified a clear inflammatory signature emanating from OAS-RNase-L-deficient myeloid cells, and associated to stronger induction of tumor necrosis factor and IFN-mediated responses by other cell types—such as $CD4^+$ T cells—downstream in the peripheral blood immune gene regulatory network.

Together with the results from other transcriptomic and biochemical assays presented in Lee et al. (2022), my findings contribute to refining the ætiology of MIS-C by supporting a role for exacerbated myeloid inflammatory responses to SARS-CoV-2 in the generalized inflammatory state observed in these patients. The myeloid responses have downstream effects on the expansion of particular subsets of $CD4^+$ and $CD8^+$ T cells, as well as upregulation of lymphoid IFN responses,

which also participate to the maintenance of inflammation. More specifically, we propose that myeloid-driven inflammation stems from a dysregulation of RNase-L-mediated post-transcriptional control of inflammatory cytokines (Malathi et al., 2005; Burke et al., 2019), although others have associated RNase L to production of interleukin (IL) 1β by the NOD-like-receptor-P3 (NLRP3) inflammasome in mice (Chakrabarti et al., 2015).

Interestingly, and in contrast to other IFN IEIs (Ciancanelli et al., 2015; Hernández et al., 2018; Lim et al., 2019; Zhang et al., 2020, 2022), none of the OAS-RNase-L-deficient MIS-C patients seem particularly sensitive to severe outcomes of infection by SARS-CoV-2 or other respiratory viruses. Yet, several common variants in the *OAS1-3* locus have been associated to differences in severe COVID-19 risk (Pairo-Castineira et al., 2021; Zeberg and Pääbo, 2021). For example, through colocalization and Mendelian randomization methods applied to protein (p) QTL data from the COVID-19 Host Genetics Initiative (2020), Zhou et al. (2021) linked the Neandertal adaptive *OAS1* splice variant rs10774671 described by Sams et al. (2016) to reduced risks of severe COVID-19 in individuals of European descent. In Aquino et al. (2023), we suggest that the protective effect of the rs10774671-G allele is mediated by its effect on *OAS1* expression in lymphoid cells.

Taken together, these results suggest that the effects of genetic variation in the OAS-RNase L pathway are mainly mediated by the establishment of systemic inflammation by immune cells, rather than by an effect on SARS-CoV-2 replication in the lung at earlier stages of infection, in line with recent reports of myeloid-driven auto-inflammatory syndromes in patients with *OAS1* mutations (Magg et al., 2021).

Key takeaways. My work in O'Neill et al. (2021), Lee et al. (2022) and Aquino et al. (2023) showcases the utility of single-cell genomics methods to dissect context-specific genetic effects on antiviral immunity in healthy and IEI backgrounds in an evolutionary framework. From the bulk of what I learned working in the HEG Unit, I consider three messages to be of particular importance.

The first observation is that a substantial portion of differences in the transcriptional immune response to viral infection across Africans and Europeans is driven by variation in immune cellular composition, and not by innate differences between the two human groups. This is most clearly presented in Aquino et al. (2023), where we were able to go a step further, and link variation in the memory-like NK subset—showing the starkest frequency differences—to increased exposure to CMV in Africa, but it was already suggested by observations in the myeloid compartment since the work we did in O'Neill et al. (2021).

The second observation is the similarity in the genetic bases of the transcriptional responses to IAV and SARS-CoV-2 that we uncovered in Aquino et al. (2023). On the one hand, I find this particularly interesting because it suggests the existence of a set of common gene regulatory networks underlying the innate immune response to respiratory RNA viruses. On the other, this gives me hope that the bulk of what we have learned about the predictors of severe COVID-19 risk will translate well to other viral infectious diseases.

The third observation is that across human groups with different genetic and environmental backgrounds, natural selection seems to have targeted different different components of the genetic basis of the antiviral response, so as to reach local adaptive equilibria between strong antiviral immunity and controlled responses to infection, rather than directional selection favoring increased or decreased antiviral immunity in specific human populations.

Altogether, my work reflects the importance of thinking about the causality of the genetic predictors of human immune variability from an evolutionary perspective, as natural selection and archaic introgression have both contributed to present-day variation in human immune responses, but also because evolution has shaped the genetic architecture of complex traits, including infectious disease risks (Sella and Barton, 2019; Uricchio, 2020). ■

Besides the knowledge we produced, another major contribution from the work in Aquino et al. (2023) is the high-quality multimodal single-cell transcriptomic data set—combining scRNA-seq with cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) for a subset of cells—we produced across over a million single immune cells.

In our hands, CITE-seq data was particularly useful to disentangle NK and CD8$^+$ T cell subsets with highly similar, highly cytotoxic effector profiles. Specifically, assigning the 'NK' identity from scRNA-seq data alone was troublesome because—at the transcriptional level—putative NK cells expressed several subunits of CD3, a canonical marker of T lymphocytes. Adding CITE-seq data for surface immune protein markers enabled us to distinguish NK cells from CD8$^+$ T along another dimension. As shown in Figure 4.4, while both cell types express similar levels of *CD3E* messenger (m) RNA, median antibody-derived tags (ADT) against the CD3 protein are close to null in NK cells and much higher in CD8$^+$ T cells, consistent with their known marker protein profiles.

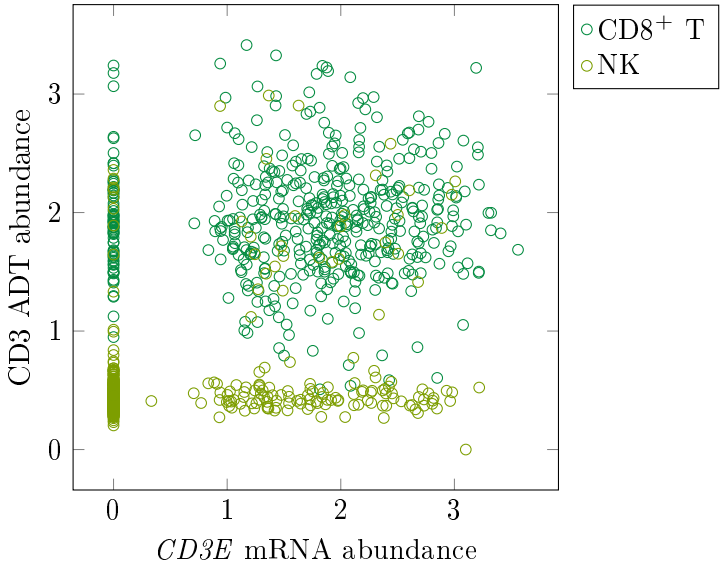


Figure 4.4 | Cell-type classification with multimodal single-cell data. Single-cell RNA-sequencing data can be coupled to other data modalities to easen cell-type identification. The x and y axes respectively show mRNA abundances for *CD3E* and antibody-derived tag (ADT) abundances for CD3—a canonical marker of T cells—across a thousand single cells identified as natural killer (NK) or CD8$^+$ T cells. While *CD3E* messenger (m) RNA abundance alone does not suffice to resolve cell identity, the two groups can be distinguished along the CD3 ADT dimension. Adapted from Aquino et al. (2023).

We used CITE-seq ad hoc to reinforce cell-type characterization in Aquino et al. (2023), but work from other authors has shown the potential of CITE-seq measurements to derive complex insights behind severe COVID-19 risk. As previously mentioned, up to 20% of life-threatening COVID-19 cases in older patients (Casanova and Abel, 2022) can be explained by inborn errors of IFN-mediated immunity (Zhang et al., 2020, 2022) or corresponding auto-immune phenocopies due to auto-antibodies targeting type I IFNs (Bastard et al., 2020, 2021a). In this context, van der Wijst et al. (2021) used CITE-seq to retrace the dynamic effects of anti-IFN-α2 auto-antibodies on PBMC responses from COVID-19 patients across two weeks of hospitalization. In line with other observations (Wilk et al., 2020; Ren et al., 2021; Stephenson et al., 2021), the authors report an increased abundance of CD14$^+$ monocytes and plasmablasts, and a decreased abundance of cytotoxic immune subsets, in the blood of critical COVID-19 cases (van der Wijst et al., 2021). Interestingly, the changes in CD14$^+$ monocyte and CD8$^+$ T cell abundance were even more stark in critical COVID-19 patients carrying anti-IFN-α2 auto-antibodies. van der Wijst et al. (2021) also report a progressive increase in CD14$^+$ and CD16$^+$ monocyte frequencies in COVID-19 patients since the date of onset of symptoms, as well as the date of hospitalization.

At the transcriptional level, van der Wijst et al. (2021) associated COVID-19 severity and the presence of anti-IFN-α2 auto-antibodies to impaired IFN responses. Specifically, in the myeloid compartment of critical COVID-19 patients, IFN-stimulated gene (ISG) expression was lower than in less severe COVID-19 cases on the day of hospitalization, and remained low throughout the two-week record. Moreover, plasmacytoid dendritic cells from critical COVID-19 patients with auto-antibodies expressed lower ISG levels, as compared to critical cases without auto-antibodies.

Availing of the other dimension of CITE-seq, van der Wijst et al. (2021) then linked impaired ISG responses to changes in surface protein expression in PBMCs. In particular, the authors highlight the increased expression of leukocyte-associated immunoglobulin-like receptor (LAIR) 1 on the surface of $CD14^+$ monocytes from COVID-19 patients with lower ISG expression. Remarkably, auto-antibodies against LAIR1 have been reported in the plasma of patients with severe COVID-19, but not moderate or light cases (Wang et al., 2021). Hence, through analyses of multimodal single-cell genomics data, van der Wijst et al. (2021) point to a novel surface protein biomarker of severe COVID-19 associated to impaired transcriptional IFN responses in specific immune cell types, which could actually play a role in disease progression, as attested by the presence of anti-LAIR1 auto-antibodies in severe cases (Wang et al., 2021).

CITE-seq and scRNA-seq are only two of the many methods available in the single-cell genomics toolkit (Heumos et al., 2023). As single-cell library preparation (Rosenberg et al., 2018) and sequencing costs continue to drop, higher multimodality and throughputs will become accessible, leading to larger data sets across more layers of gene expression regulation. By filling in the blanks between genotype and phenotype, these new methods will enable more robust causal inference of genetic effects (Heumos et al., 2023; Cuomo et al., 2023).

Latent regulatory predictors of immune variation

Where CITE-seq and scRNA-seq respectively assay gene expression at the protein and transcript level, single-cell assays for transposase-accessible-chromatin sequencing (scATAC-seq) allow to profile open chromatin regions (OCRs) in single nuclei (Buenrostro et al., 2015). Chromatin accessibility data are an especially useful tool to unravel the complex context-dependent regulatory grammar of gene expression. First, accessibility at any given OCR results from the integration of complex genetic (Degner et al., 2012; Benaglio et al., 2023) and epigenetic (Kundaje et al., 2015) signals—such as histone marks and DNA methylation—which can translate effects from environmental exposures (Bergstedt et al., 2022). Second, OCRs in different human tissues contain different sets of active *cis*-regulatory genomic elements like promoters, enhancers and transcription factor (TF) binding site (TFBS) motifs (ENCODE Project Consortium, 2012, 2020; Zhang et al., 2021), reflecting the tissue-specificity of gene expression regulation (Kim-Hellmuth et al., 2017). Finally, because chromatin accessibility is a necessary condition for transcription, coupling scATAC-seq and scRNA-seq data is clearly relevant from a biological standpoint.

Context-dependent chromatin accessibility data also come in handy to interpret genome-wide associations between genotype and disease phenotypes. Across over 5 thousand single nucleotide polymoprhism (SNP) loci in noncoding genomic regions associated to more than 6 hundred disease traits, Maurano et al. (2012) report that around 77% of genome-wide significant associations fall in an OCR or are correlated to accessibility in that region. Moreover, Degner et al. (2012) estimate that 55% of strong eQTLs mapped across 70 human cell lines (Pickrell et al., 2010) are also associated to chromatin accessibility at nearby OCRs. Together, these results reveal a wide overlap between the genetic bases of gene expression regulation at different levels along Crick's dogma, and highlight the relevance of pairing caQTL and eQTL mapping to improve the interpretation of genome-wide association results.

The pervasiveness of caQTLs along the human genome (Degner et al., 2012; Maurano et al., 2012) may reflect variation in epigenetic mark patterns—such as histone-tail modifications or DNA methylation—across human individuals (Waszak et al., 2015; Bergstedt et al., 2022) and populations (Carja et al., 2017; Husquin et al., 2018), which correlate with differential OCR accessibility and changes in gene expression (Luo et al., 2018). In particular, Husquin et al. (2018) estimate that around 70% of population DNA methylation differences in monocytes sampled from 156 healthy individuals of African and European origin can be explained by genetic variation around CpG sites. Interestingly, the authors also linked differences in DNA methylation to differences in the expression of 230 genes in response to Toll-like receptor activation or stimulation by a live IAV strain (Quach et al., 2016). Moreover, Alasoo et al. (2018) have shown direct links between genetically controlled chromatin accessibility variation and the magnitude of the transcriptional response to *Salmonella enterica* bacteria in human myeloid cells. Together, these results suggest a role for genetically controlled OCR accessibility variation in population differences in infectious disease risk.

There is yet to be a published assessment of the contribution of caQTLs to disease risk disparities across human populations and immune cell types (Benaglio et al., 2023). However, Aracena et al. (2022) show that a fraction of population differences in the myeloid transcriptional response to IAV infection between individuals of African and European ancestry (Quach et al., 2016; Randolph et al., 2021) can be explained by an epigenetic basis made of factors that affect chromatin accessibility. More specifically, the authors performed bulk eQTL, caQTL, histone-mark (hm) QTL and DNA-methylation (me) QTL mapping on monocyte-derived macrophages sampled from 35 healthy individuals, before and after 24 hours of exposure to IAV.

Relative to eQTLs, chromatin QTLs were most strongly enriched in population-differential traits, in line with a stronger genetic control of population differences in gene expression regulation, rather than gene expression itself. Aracena et al. (2022) also report higher chromatin accessibility around genes involved in inflammatory pathways in African-origin macrophages, that could explain why antiviral immune responses from African-origin individuals tend to display a stronger inflammatory component (Quach et al., 2016). Interestingly, most of these signals were lost when regressing out the effects of the strongest QTLs, suggesting an important genetic contribution to the observed population differences. Overall, the observations by Aracena et al. (2022) reflect the need for a systematic evaluation of the latent regulatory predictors of immune variation across human populations and immune cell types. To this end, we performed scATAC-seq on non-stimulated PBMC samples from the cohort of Central African (AFB), West European (EUB) and East Asian (ASH) healthy donors reported in Aquino et al. (2023).

Focusing first on AFB and EUB individuals sampled during the same recruitment event (Quach et al., 2016), we recovered over 200 thousand high-quality chromatin accessibility profiles across 21 different PBMC types. As shown in Figure 4.5, the estimates of cellular composition in each population—evaluated as the relative contribution of each immune cell type to the make-up of its corresponding lineage—taken from scATAC-seq data are highly similar to those reported in Aquino et al. (2023) from scRNA-seq data (Pearson's $\mathrm{r} = 0.98$, $\mathrm{p} < 2.2 \times 10^{-16}$). In line with our hypothesis of latent regulatory predictors underlying population differences in the transcriptional immune response to viruses, this suggests that the immune cellular components that explain most immune gene expression differences between healthy individuals of African and European descent also have distinct chromatin accessibility profiles, which can translate differences in environmental exposures. For example, population differences in memory-like NK cell abundances can be largely explained by latent cytomegalovirus (CMV) infection (Aquino et al., 2023). With the scATAC-seq data, we will be able to dig deeper into these differences, so as to add new context-dependent regulatory nodes to the putative causal chain linking exposure to CMV and COVID-19 severity.

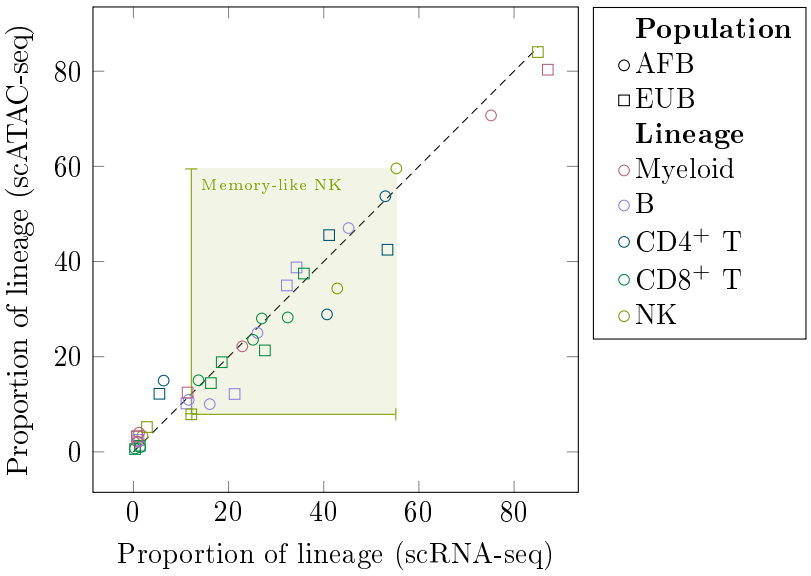


Figure 4.5 | Cellular composition inference from single-cell data. For each of five major immune lineages, each dot shows the proportion of the lineage made up by each of the cell types it contains, estimated from single-cell RNA-sequencing (scRNA-seq) or single-cell assays for transposase-accessible-chromatin sequencing (scATAC-seq) data, in median across peripheral blood mononuclear cells sampled from 80 individuals of Central African (AFB) or 80 individuals West European (EUB) origin (Aquino et al., 2023). Each dot represents a cell type, but is colored according to the lineage it belongs to. The population from which each estimate was made is shown by different dot shapes. The largest cellular composition difference between AFB and EUB is highlighted.

These scATAC-seq data will also allow us to look for the genetic determinants of gene expression regulation differences that may also underlie current disparities in COVID-19 risk. The single-cell resolution of our methods is of particular importance here, as it has been shown that besides genetic, immune cellular composition is the largest driver of inter-individual epigenetic marks associated to chromatin accessibility, such as DNA methylation (Bergstedt et al., 2022).

Across all different immune cell types, we mapped 104,775 unique caQTLs nominally associated (Student's t-test $p < 10^{-5}$) to the accessibility at 29,934 OCRs. Out of all unique OCRs under putative genetic control, 67.6% are annotated as 'distal' intergenic or intronic regions, 21.9% fall on annotated promoter regions and 10.5% fall on exons, in line with what is expected from a typical ATAC-seq experiment (Yan et al., 2020). For example, the rs11080327-A allele is associated to increased chromatin accessibility in B cells ($p = 5.6 \times 10^{-8}$) at an OCR in the first intron of *SLFN5* annotated as a TFBS for STAT1 (ENCODE Project Consortium, 2020), and is remarkable because *SLFN5* is a well-known ISG involved in antiviral responses, and rs11080327-A has been associated to stronger *SLFN5* expression in response to type I IFN (Perez et al., 2022) and IAV (Schott et al., 2022) in B cells. In particular, we also associated rs11080327-A to a stronger *SLFN5* response in B cells exposed to SARS-CoV-2 and IAV ($p < 9.9 \times 10^{-12}$) in Aquino et al. (2023).

Although preliminary, these observations highlight the relevance of our caQTL mapping, as well as the relevance of coupling scATAC-seq and scRNA-seq data from paired samples to disentangle the genetic basis of inter-individual variability in the genetic basis of the antiviral immune response. Knowing the genetic variants associated to putative regulatory activity in different immune cell types and human populations, we will be able to test them for signals of archaic introgression and natural selection, so as to retrace their evolutionary history and impact on molecular endophenotypes.

Molecular QTL mapping studies are useful to infer causal links between genotype and phenotype; yet, they can only paint a static portrait of these associations based on extant human diversity. In contrast, the environmental factors—including pathogens, diet and cultural practices—that shape human genetic diversity are dynamic. Temporal shifts in selective pressures can lead to evolutionary 'mismatch', when a previously advantageous allele becomes deleterious following a change in the environment. Evolutionary insights are useful to reconcile such cases of apparent maladaptation.

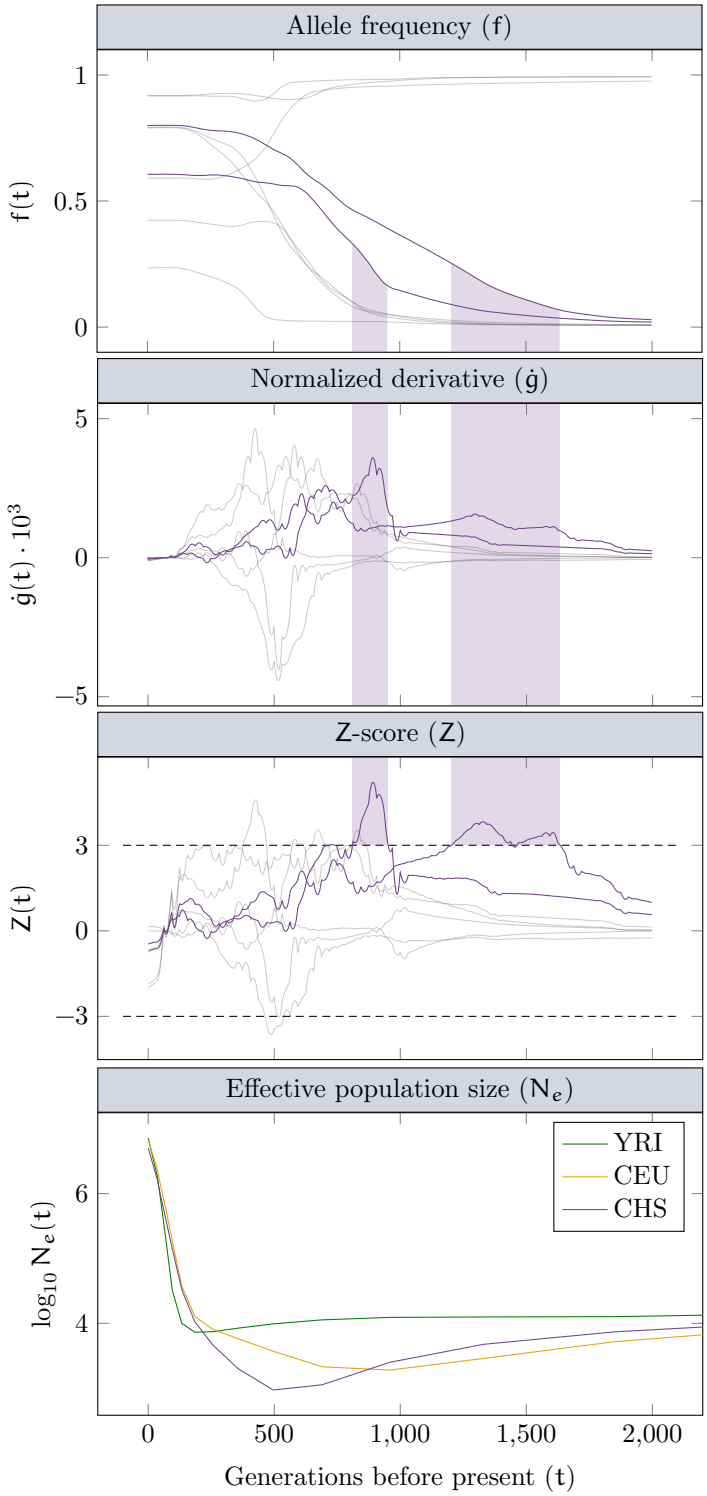


Figure 4.6 | **Timing natural selection signals from allele frequency trajectories.** The top panel shows allele frequency trajectories of two quantitative trait loci (QTLs)–rs4806787 and rs1028396–respectively affecting the expression (e) of *LILRB1* in plasmacytoid dendritic cells and *SIRPA* in CD14^{+} monocytes in response (r) to 'severe acute respiratory syndrome' coronavirus 2, inferred across the past 2 thousand generations using data from the 'Han Chinese South' (CHS) reference panel from the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2022). The panel below shows per-generation changes of derived allele frequencies, normalized for allele frequency. The third panel shows Z-scores calculated as the normalized derivative, scaled at each generation by the standard deviation of derivatives across all eQTLs. Periods of selection are estimated as the range, in generations, over which the rate of change in the frequency of each allele deviates significantly from expectations under the hypothesis of neutrality (i.e., $|Z| > 3|$). Across the first three panels, the corresponding allele frequency trajectories, first derivatives and Z-scores for five random SNPs sampled from the set of all (r)eQTLs detected in this study are shown in gray. The purple transparent intervals indicate the inferred timing of onset of selection at the two highlighted loci. The bottom-most panel shows effective population sizes estimated by Speidel et al. (2019) using data from the CHS, 'Yoruba in Ibadan' (YRI) and 'Utah residents with Northern and Western European ancestry' (CEU) reference panels. Adapted from Aquino et al. (2023).

Detection of natural selection signals from inferred allele frequency trajectories

An example of pleiotropic trade-off is illustrated by the G1 and G2 variants of the gene encoding apolipoprotein L-1 (*APOL1*). Alleles of both variants have been associated to chronic kidney disorders in present-day African American individuals (Genovese et al., 2010). Yet, G1 is found at high frequencies in African genomes—borne by 38% of the individuals in the 'Yoruba in Ibadan' panel of the 1000 Genomes Project Consortium (1000 Genomes Project Consortium et al., 2010), but absent from European, Japanese or Chinese panels (Genovese et al., 2010). Consistent with this stark population differentiation, haplotype-based methods support the hypothesis that G1 rose to high frequency in Africa through positive selection (Genovese et al., 2010).

Both G1 and G2 confer protection against infection by *Trypanosoma brucei rhodesiense*—the parasite reponsible for 'African sleeping sickness'—providing a likely target for natural selection to increase the frequency of G1 in genomes of African descent (Genovese et al., 2010). However, its advantageous, protective effect against sleeping sickness is lost for African Americans that live outside the endemic range of the causal parasite. Taken together, these results emphasize the importance of considering the genetic make-up of each individual in the context of their local environment, and in light of their evolutionary roots.

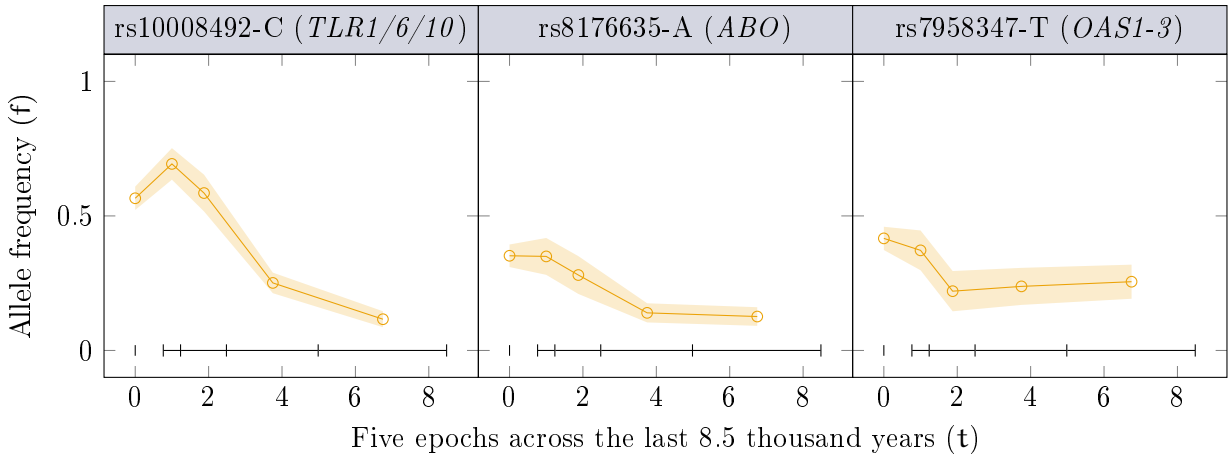


Figure 4.7 | **Observed allele frequency trajectories through ancient genomes.** For the three variants in immune-relevant loci, each dot shows the frequency f of the derived allele estimated from the set of ancient DNA samples available in each of five time transects spanning from 8.5 thousand years ago (kya) to the present. Confidence intervals were computed as $f \pm 1.96 \cdot \sqrt{\frac{f(1-f)}{n(t)}}$, where $n(t)$ is the number of samples at each epoch. Neolithic period, 8.5 to 5 kya, $n = 729$; Bronze Age, 5 to 2.5 kya, $n = 893$; Iron Age, 2.5 to 1.25 kya, $n = 319$; Middle Ages, 1.25 to 0.75 kya, $n = 435$. Adapted from Kerner et al. (2023b).

In Aquino et al. (2023), we used two different methods to detect signatures of natural selection. On the one hand, we used a summary population branch statistic (PBS) that captures extreme patterns of allele frequency differentiation (Yi et al., 2010) between reference African, European and East Asian panels from the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2022). On the other, we used a method to approximate the full likelihood that selection acted on a given variant, based on hidden allele frequency trajectories inferred from ancestral recombination graphs of genetic variation in each reference panel (Stern et al., 2019; Speidel et al., 2019).

From the inferred allele frequency trajectories at each (r)eQTL, we defined periods of time in the last 2 thousand generations—that is, 56 thousand years—during which allele frequencies changed more rapidly than expected under random genetic drift using an intuitive and highly scalable approach. Briefly, we considered the posterior mean $f(t)$ of allele frequency at each generation t—smoothed by loess regression to ensure progressive changes and minimize artifacts induced from

the inference process—and computed its first derivative as

$$\dot{f}(t) = \frac{df}{dt}(t) = f(t+1) - f(t) \tag{4.1}$$

to get the rate of allele frequency change across generations. Under neutrality, allele frequencies follow a binomial distribution parameterized $\mathcal{B}(N, f)$, where N is the size of the haploid population (Appendix C, page 199). The variance of allele frequency changes across generations is thus a function of the frequency of the allele at each generation,

$$\mathbb{V}\left[\dot{f}\right] = \frac{f(1-f)}{N}. \tag{4.2}$$

To account for this, we scaled the allele frequency derivative $\dot{f}(t)$ by a function of allele frequency, such that the variance of normalized allele frequency differences,

$$\dot{g} = \frac{\dot{f}}{\sqrt{f(1-f)}} \tag{4.3}$$

was the same across all variants. Finally, to avoid confounding effects from generalized frequency changes across several alleles due to evolutionary forces other than natural selection, we computed a Z-score of generation-wise allele frequency change as

$$Z = \frac{\dot{g}}{\widehat{\sigma}(\dot{g})}, \tag{4.4}$$

where $\widehat{\sigma}(\dot{g})$ is the estimated standard deviation of scaled and normalized allele frequency changes. We infer rapid changes in allele frequency for each variant and generation where the absolute Z-score is greater than 3. This process is illustrated in Figure 4.6.

Using random SARS-CoV-2 reQTLs, Figure 4.6 shows that normalizing the derivative of the allele frequency trajectory and scaling it as a Z-score aptly adjusts the rate of allele frequency change to account for periods during which allele frequencies varied more widely, possibly under the effect of other evolutionary forces like changes in effective population sizes. However, it also showcases a limitation of our method: if allele frequencies evolve at a near-constant rate, it is only able to estimate the onset of natural selection, not the full length of the episode of adaptation. Furthermore, our inference of natural selection can only be as good as the reconstruction of the allele frequency trajectory. Through simulated evolutionary scenarios across different times of onset of selection and selection cofficients, we showed that the heuristic $|Z| > 3$ threshold yields a reasonably powered method to detect events of local adaptation (Aquino et al., 2023). For instance, the probability to detect an episode of selection starting around a thousand generations ago with selection coefficient $s = 0.05$ is over 0.9.

With this new method, we detected rapid changes in allele frequencies concentrated around 25 thousand years ago at 245 SARS-CoV-2-specific reQTLs, and specifically in genomes of East Asian descent, in line with previous reports by Souilmi et al. (2021) of an ancient coronavirus-related epidemic driving local adaptation at virus-interacting protein (VIP) loci in East Asia around the same time frame. Yet, we also found that adaptation events in the top 5% PBS values—that is, contributing the most to present-day allele frequency differentiation patterns—in East Asia started over 27 thousand years ago (Aquino et al., 2023). Among even more extreme outliers in the top 1% PBS values, we report three variants that are associated to the risk of COVID-19 through their effects on expression at two coronavirus VIP loci—*RAB2A* and *TMED5*—in different PBMC types. In a context of accelerated viral spillover (Jones et al., 2013), these signals of past adaptation can highlight candidates for treatment or prevention of viral diseases (Souilmi et al., 2021).

Detection of natural selection signals from observed allele frequency trajectories

Both the PBS (Yi et al., 2010) and the approximate full likelihood (Stern et al., 2019) methods infer signals of natural selection from extant human genetic diversity. While their foundations are grounded on a solid basis of evolutionary genetics theory—built by Fisher (1918), Haldane (1937, 1949), Wright (1965), Griffiths and Marjoram (1996, 1997), Krone and Neuhauser (1997), Coop and Griffiths (2004), and many others reviewed by Hartl and Clark (2018)—these methods are blind to the true evolution of allele frequencies across time.

Recently, the advent of next-generation DNA sequencing methods and the improvement of techniques to extract DNA from fossilized remains have enabled a new framework to retrace allele frequency trajectories through direct observations from 'ancient' DNA (Kerner et al., 2023a). For instance, Kerner et al. (2023b) used 2,376 ancient and 503 modern human genomes to characterize episodes of local adaptation to pathogenic pressures in Europe during the last 10 thousand years, since the Neolithic period and through the Bronze, Iron and Middle Ages, up until today.

From the pool of ancient DNA samples available in each time transect, the authors estimated derived-allele frequencies at over a million SNPs, and used these observed allele frequency trajectories to infer the strength and the time of onset of selection at each SNP. Kerner et al. (2023b) then defined 89 non-consecutive loci that concentrated the strongest signals of selection, and found these candidate targets of natural selection to be enriched in genes with immune-relevant functions.

Kerner et al. (2023b) report that most episodes of directional selection at these 89 loci started after the beginning of the Bronze Age: a period characterized by population expansions in Europe. Larger populations are associated to more efficient natural selection, and may also result in stronger pathogenic pressures due to increased promiscuity and mobility, which can potentiate epidemic outbreaks (Kerner et al., 2023b). Notably, the authors also link these adaptive changes in allele frequency to a decreased polygenic risk to infectious diseases, but an increased risk to autoimmune and inflammatory diseases among present-day humans. For example, Figure 4.7 shows allele frequency trajectories estimated by Kerner et al. (2023b) at three immune-relevant loci, including *TLR1/6/10* and *OAS1-3*, that were targeted by positive selection after the Bronze Age.

Altogether, the work by Kerner et al. (2023b) emphasizes the usefulness of ancient DNA in the framework of evolutionary medicine, to better understand human evolutionary adaptation to different contexts and environments (Marciniak and Perry, 2017; Perry, 2021), and thus draw a more accurate picture of the genetic bases of modern complex diseases (Kerner et al., 2023a).

Yet, ancient DNA is not without limitations. For instance, although some ancient DNA samples have been sequenced using unbiased 'shotgun sequencing' methods, around 70% of available ancient genome-wide data have been produced using the same array (Mathieson et al., 2015), which is biased towards certain types of sequences (Ávila-Arcos et al., 2023). Also, ancient DNA only allows to look as far back as there are available samples. To date, 9,695 DNA samples older than 500 years have been published, but 98% are younger than 10 thousand years (Kerner et al., 2023a). Another important limitation imposed by the availability of ancient DNA is that almost 60% of samples available today come from remains found on the European continent (Kerner et al., 2023a; Ávila-Arcos et al., 2023). This is in part due to the better conservation of DNA in temperate climates—relative to African or South American regions, which jointly contribute less than 5% of ancient DNA samples—but it also reflects a more general trend in human genomics research (Sirugo et al., 2019). Namely, while individuals of European descent represent only a minor fraction of the global human population, they are grossly over-represented in ancient (Kerner et al., 2023a) and modern (Sirugo et al., 2019) DNA data bases. This has important ethical implications (Popejoy and Fullerton, 2016), but it also limits the generality of inferences drawn from these data.

In particular, the lack of diversity in current genetic data bases limits the capacity of a polygenic risk score (PRS) to predict complex disease risks across human populations. Briefly, the standard PRS is a metric in the precision medicine toolkit that estimates individual-specific risks of developing a disease trait, as a linear combination of genome-wide association study (GWAS) effect size estimates, weighted by the genotype of each individual. However, if the discovery cohort on which the GWAS is performed is genetically distant from the target cohort on which the PRS is applied, differences in the structure of linkage disequilibrium and allele frequency distributions can affect the translability of the estimates, and generally decrease PRS accuracy (Martin et al., 2017, 2019). Therefore, the current European focus of human genomics research limits these approaches.

In this context, several statistical approaches have been proposed to improve PRS transferability. In fact, even a relatively modest diversification of discovery cohorts through 'multi-ancestry' GWASs on diverse groups of individuals, or meta-analysis of 'single-ancestry' GWAS results across different populations, has been shown to improve PRS accuracy across populations (Ruan et al., 2022; Wang et al., 2023a). Another interesting perspective is the development of weighted PRS approaches that leverage recently admixed populations—whom are often neglected by genetic studies—to consider the effects of different local ancestries while maintaining relatively homogeneous environmental exposures (Wang et al., 2023a). However, for traits for which most of the population variation in GWAS estimates is driven by environmental factors, the only solution to obtain an accurate PRS is to perform the GWAS directly on the target population, or one with very close genetic and environmental backgrounds (Mathieson, 2021).

Overall, the development of a precision medicine able to tailor healthcare to the genetic and nongenetic determinants of disease risk in each individual relies on the establishment of data bases and 'biobanks' that accurately represent human diversity. For example, Sohail et al. (2023) from the Centro de Ciencias Genómicas de la Universidad Nacional Autónoma de México just published the Mexican Biobank (MXB) of genome-wide genotypes and complex phenotypes—including metabolic and socioeconomic traits—across more than 6 thousand individuals from different cultural regions in Mexico. Across several metabolic traits including blood glucose and cholesterol levels, the authors show that dividing the data set into 'training' and 'testing' partitions—to perform a GWAS and test the associated PRS, respectively—yields more accurate predictions, compared to those obtained from a PRS derived from a United Kingdom Biobank (UKB) GWAS data set, despite the much lower sample size (Sohail et al., 2023). However, Sohail et al. (2023) also highlight the need to build larger samples, so as to make more accurate predictions.

Another obstacle in the way to a widely translatable PRS for certain complex traits is the diversity of phenotype definitions. Briefly, computing a PRS on merged GWAS results from studies with different trait definitions is likely to yield a score underpowered to predict either definition well. In this context, the Global Biobank Meta-analysis Initiative (GBMI) (Zhou et al., 2022) is a powerful network of international collaboration and data sharing across 23 biobanks—including UKB but not MXB yet—across the globe, that aims to improve overall PRS accuracy by diversifying the resources used in human genomics research and building larger samples, as well as setting unified definitions for GWAS phenotypes and covariates (Wang et al., 2023b).

Besides the GBMI, other world-class organizations have recognized the need to diversify data bases in the establishment of a precision medicine. Specifically, through its expertise in single-cell genomics and its continued efforts to include understudied human groups in their work, the HEG Unit is part of the Chan-Zuckerberg Initiative to generate an African-ancestry immune cell atlas at single-cell resolution, in order to tackle the paucity of data across ethnically diverse groups of African individuals and assess immune gene regulation in more diverse contexts.

In particular, the contribution of the HEG Unit to the African Immune Cell Atlas Project is centered around using scRNA-seq and scATAC-seq data to characterize the environmentally driven epigenetic variation behind immune differences between individuals of African origin with a similar genetic background, but living in very distinct rural and urban settings. Deciphering how the environment can affect the genetic regulation of chromatin accessibility and transcription in an immune context is essential towards an improved understanding of the parameters that drive immune variability across healthy individuals worldwide, and how the risk for infectious and non-infectious diseases is transmitted across generations.

Gene-by-environment interactions are a central component of the model of 'omni-environmental' inheritance proposed by Mathieson (2021) as an extension to the original omnigenic model (Boyle et al., 2017; Liu et al., 2019). In this network, environmental and genetic factors interact with each other extensively to form the basis of a complex trait, but the GWAS framework is blind to these interactions, and can only measure the overall effect of observed genotypes on observed phenotypes (Boyle et al., 2017; Liu et al., 2019). Molecular QTL mapping has improved the capacity for causal inference by shedding light on some regulatory nodes and edges in the network. However, the effectiveness of these methods to uncover the genetic bases of complex diseases is limited by their focus on common genetic variation, the 'flattening' of heritability signals across weak-effect variants by negative selection against strongly deleterious variants (O'Connor et al., 2019), the large sample sizes required to detect these weak effects and the context-dependency of gene expression regulation, among other factors reviewed by Umans et al. (2021).

Large language models of genomic data and nucleotide transformers

Following recent technological and algorithmical developments, artificial intelligence (AI) has quickly emerged as a promising tool to disentangle the complexity of omni-environmental networks and map the nodes and edges leading from measurable genotypes to observable phenotypes (Dias and Torkamani, 2019). More specifically, 'deep learning' (DL) methods differ from classical 'machine learning' AI algorithms in that they do not need human feedback to adapt their learning. That is, DL algorithms are self-learning systems able to extract patterns from highly complex data structures across upwards of millions of data points, which makes them especially interesting for the analysis of complex genomic and clinical data (Dias and Torkamani, 2019; Topol, 2023).

The first successful attempts to predict gene expression from mammalian genomic sequences used convolutional neural network (CNN) architectures to infer regulatory activity based on chromatin mark data, and then predict the effect of genetic variants on expression (Zhou et al., 2018; Kelley et al., 2018; Kelley, 2020; Agarwal and Shendure, 2020) (Appendix E, page 206). While these methods set a new state of the art in the application of DL algorithms to genomic data, they were limited by the local scope of convolutional filters, which only allowed to consider regulatory interactions within 20 to 40 kilobases; yet, regulatory interactions between enhancers and repressors of gene expression are known to span far greater distances (Gasperini et al., 2020; Avsec et al., 2021).

In this context, Avsec et al. (2021) developed a new DL model mixing CNN layers with components of the 'transformer' architecture (Appendix E, page 206). Briefly, transformers are a class of DL models used to process large sequential input data—such as natural language text (Vaswani et al., 2017; Brown et al., 2020) but also DNA sequences (Ji et al., 2021)—and predict an output sequence. Relative to previous recurrent neural network (RNN) architectures that analyse sequence data in a step-wise manner, transformers are imbued with additional 'attention' mechanisms that allow them to process arbitrarily long sequences by learning which components of the input are really informative—regardless of their position in the sequence, and without having to process the entire sequence beforehand—and should be attended to, so as to accurately predict the output

(Vaswani et al., 2017). For example, this allows to predict the activity at a transcription start site based on the sequence-inferred activity of enhancers downstream and upstream in the genomic sequence (Avsec et al., 2021).

Avsec et al. (2021) report that their transformer can capture regulatory interactions between genes and context-dependent regulatory elements up to 100 kilobases apart in human and mouse genomes, and use this information to predict tissue-specific gene expression with higher accuracy than CNN-based methods. Furthermore, the authors show that the model gives more attention to experimentally validated enhancers, suggesting that transformer attention levels can be used to prioritize gene-enhancer pairs at least as well as other state-of-the-art methods (Fulco et al., 2019; Avsec et al., 2021). The transformer was also shown to provide more accurate eQTL effect predictions—relative to a CNN-based DL model (Kelley, 2020)—across 47 out of 48 tissues from the Genotype-Tissue Expression Consortium (The GTEx Consortium, 2020), for high-confidence fine-mapped variants (Wang et al., 2020) in regions with simple LD patterns (Avsec et al., 2021). Hence, the transformer is a useful tool for predicting regulatory activity and the effect of noncoding variants on gene expression: two key tasks towards disentangling omni-environmental network effects.

However, it has more recently been argued that, while the transformer trained by Avsec et al. (2021) can predict the impact of short-range regulatory regions on gene expression well, it struggles with the impact of long-range enhancers (Karollus et al., 2023). That is, although its receptive field spans up to 200 kilobases, the transformer underestimates the effect of distal regulatory elements, and its predictions are driven by stronger proximal effects less than 100 kilobases away. This may reflect the underlying biology, but it also may be due to the model artificially down-weighting noisier distal effects, or to its inability to account for the tridimensional chromatin conformations that bring distant genomic elements into contact (Avsec et al., 2021; Karollus et al., 2023). In line with previous trends in natural language processing (Vaswani et al., 2017; Kaplan et al., 2020; Rae et al., 2021), Karollus et al. (2023) suggest that the improvement in prediction accuracy described by Avsec et al. (2021) could be solely due to the sheer increase in estimated parameter counts relative to previous DL models, rather than the larger receptive field of their transformer. Karollus et al. (2023) also propose that training the model on chromatin-contact data and/or gene expression and epigenetic data across multiple cell types and species could help build a transformer that is actually able to exploit a large receptive field.

From a similar perspective, Dalla-Torre et al. (2023) propose a collection of transformers—ranging in size from 50 million to 2.5 billion parameters—and trained on over 3 thousand human genome sequences and 850 genomes from other species, ranging from bacteria to mammals. In line with the suggestions by Karollus et al. (2023), the authors report that—at a constant parameter count—increased intra-species and inter-species variability increases prediction accuracy. Interestingly, training on genomic data from several species also increases prediction accuracy of models applied to human data, suggesting that the algorithm learned to pick up on conserved and important functional sequences during training (Dalla-Torre et al., 2023). Dalla-Torre et al. (2023) also validate the suggestion that increasing parameter counts leads to improved accuracy on biological data, but show that their models can be fine-tuned to a subset of parameters allowing to maintain similar performance at a fraction of the computational cost. Yet, although these models have a larger attention span than previous transformers (Ji et al., 2021), at 12 kilobases it is still much smaller than the one proposed by Avsec et al. (2021).

All in all, nucleotide transformers have proven their worth in predicting functional regulatory activity along the human genome and the context-dependent effects of this regulation on gene expression (Ji et al., 2021; Avsec et al., 2021; Dalla-Torre et al., 2023). However, their capacity to accurately predict individual-specific molecular endophenotypes from DNA sequence data alone is

currently limited (Karollus et al., 2023) by the difficulty of simultaneously estimating the impact of long-range and short-range regulatory activity (Nguyen et al., 2023). With growing technological improvements and availability of multimodal data sets, it is likely that more DL models will be developed and applied towards dissecting the genetic and environmental interactions underlying complex traits (Boyle et al., 2017; Liu et al., 2019; Mathieson, 2021; Dalla-Torre et al., 2023).

Yet, both artificial and organic intelligence have a role to play in this endeavour. Although performing GWAS and molecular QTL mapping studies ad infinitum is not a viable approach to uncover all regulatory interactions (Umans et al., 2021; Mostafavi et al., 2022), these frameworks have provided several key insights into the general structure of the omni-environmental networks underlying complex disease traits (Dimas et al., 2009; Westra et al., 2013; van der Wijst et al., 2018b, 2020; The GTEx Consortium, 2020; COVID-19 Host Genetics Initiative, 2023), and how their genetic architecture has been shaped by evolution (Lohmueller, 2014; O'Connor et al., 2019; Sella and Barton, 2019; Uricchio, 2020). As the field of DL genomics develops, the data produced with these tools can be used to train stronger models able to predict genotype-phenotype effects more accurately (Karollus et al., 2023). More importantly, these insights will continue to be essential for scientists to interpret the biological relevance of the machine's predictions.

Bibliography

1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Vikram Agarwal and Jay Shendure. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Reports*, 31(7), 2020.

Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, pages 1–8, 2023.

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

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:1–9, 2010.

Yann Aquino, Aurélie Bisiaux, Zhi Li, Mary O'Neill, Javier Mendoza-Revilla, Sarah Hélène Merkling, Gaspard Kerner, Milena Hasan, Valentina Libri, Vincent Bondet, et al. Dissecting human population variation in single-cell responses to SARS-CoV-2. *Nature*, 621(7977):120–128, 2023.

Katherine A. Aracena, Yen-Lung Lin, Kaixuan Luo, Alain Pacis, Saideep Gona, Zepeng Mu, Vania Yotova, Renata Sindeaux, Albena Pramatarova, Marie-Michelle Simon, et al. Epigenetic variation impacts ancestry-associated differences in the transcriptional response to influenza infection. *bioRxiv*, pages 2022–05, 2022.

Dvir Aran, Zicheng Hu, and Atul J. Butte. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*, 18:1–14, 2017.

Ricard Argelaguet, Anna S. E. Cuomo, Oliver Stegle, and John C. Marioni. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39(10):1202–1215, 2021.

Juan Luis Arsuaga, Ignacio Martínez, Lee J. Arnold, Arantza Aranburu, Ana Gracia-Téllez, Warren D. Sharp, Rolf M. Quam, Christophe Falguères, Ana Pantoja-Pérez, James Bischoff, et al. Neandertal roots: Cranial and chronological evidence from Sima de los Huesos. *Science*, 344 (6190):1358–1363, 2014.

Takaki Asano, Bertrand Boisson, Fanny Onodi, Daniela Matuozzo, Marcela Moncada-Vélez, Majistor Raj Luxman Maglorius Renkilaraj, Peng Zhang, Laurent Meertens, Alexandre Bolze, Marie

Materna, et al. X-linked recessive TLR7 deficiency in ˜ 1% of men under 60 years old with life-threatening COVID-19. *Science Immunology*, 6(62):eabl4348, 2021.

Berhane Asfaw, W. Henry Gilbert, Yonas Beyene, William K. Hart, Paul R. Renne, Giday WoldeGabriel, Elisabeth S. Vrba, and Tim D. White. Remains of *Homo erectus* from Bouri, Middle Awash, Ethiopia. In *Human Evolution Source Book*, pages 305–312. Routledge, 2016.

Tal Ashuach, Mariano I. Gabitto, Rohan V. Koodli, Giuseppe-Antonio Saldi, Michael I. Jordan, and Nir Yosef. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, pages 1–10, 2023.

María C Ávila-Arcos, Maanasa Raghavan, and Carina Schlebusch. Going local with ancient DNA: A review of human histories from regional perspectives. *Science*, 382(6666):53–58, 2023.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021.

Nil Aygün, Dan Liang, Wesley L. Crouse, Gregory R. Keele, Michael I. Love, and Jason L. Stein. Inferring cell-type-specific causal gene regulatory networks during human neurogenesis. *Genome Biology*, 24(1):1–25, 2023.

Shara E. Bailey, Timothy D. Weaver, and Jean-Jacques Hublin. Who made the Aurignacian and other early Upper Paleolithic industries? *Journal of Human Evolution*, 57(1):11–26, 2009.

David J. Balding, Ida Moltke, and John C. Marioni. *Handbook of statistical genomics*. John Wiley and Sons, fourth edition, 2019.

Brunilda Balliu, Matthew Durrant, Olivia de Goede, Nathan Abell, Xin Li, Boxiang Liu, Michael J. Gloudemans, Naomi L. Cook, Kevin S. Smith, David A. Knowles, et al. Genetic regulation of gene expression and splicing during a 10-year period of human aging. *Genome Biology*, 20(1): 1–16, 2019.

Alvaro N. Barbeira, Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, Eric S. Torstenson, Kaanan P. Shah, Tzintzuni Garcia, Todd L. Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, 9(1):1825, 2018.

Luis B. Barreiro and Lluís Quintana-Murci. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics*, 11(1):17–30, 2010.

Luis B. Barreiro, Meriem Ben-Ali, Hélène Quach, Guillaume Laval, Etienne Patin, Joseph K. Pickrell, Christiane Bouchier, Magali Tichit, Olivier Neyrolles, Brigitte Gicquel, et al. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genetics*, 5(7):e1000562, 2009.

Luis B. Barreiro, Ludovic Tailleux, Athma A. Pai, Brigitte Gicquel, John C. Marioni, and Yoav Gilad. Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *Proceedings of the National Academy of Sciences*, 109(4):1204–1209, 2012.

Nicholas H. Barton, Alison M. Etheridge, and Amandine Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50–73, 2017.

Paul Bastard, Lindsey B. Rosen, Qian Zhang, Eleftherios Michailidis, Hans-Heinrich Hoffmann, Yu Zhang, Karim Dorgham, Quentin Philippot, Jérémie Rosain, Vivien Béziat, et al. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science*, 370(6515): eabd4585, 2020.

Paul Bastard, Adrian Gervais, Tom Le Voyer, Jérémie Rosain, Quentin Philippot, Jérémy Manry, Eleftherios Michailidis, Hans-Heinrich Hoffmann, Shohei Eto, Marina Garcia-Prat, et al. Autoantibodies neutralizing type I IFNs are present in ˜ 4% of uninfected individuals over 70 years old and account for˜ 20% of COVID-19 deaths. *Science Immunology*, 6(62):eabl4340, 2021a.

Paul Bastard, Eleftherios Michailidis, Hans-Heinrich Hoffmann, Marwa Chbihi, Tom Le Voyer, Jérémie Rosain, Quentin Philippot, Yoann Seeleuthner, Adrian Gervais, Marie Materna, et al. Auto-antibodies to type I IFNs can underlie adverse reactions to yellow fever live attenuated vaccine. *Journal of Experimental Medicine*, 218(4):e20202486, 2021b.

Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of transcript variability in single mammalian cells. *Cell*, 163(7):1596–1610, 2015.

Alexis Battle, Zia Khan, Sidney H Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, 2015.

Iraldo Bello-Rivero, Majel Cervantes, Yeny Torres, Joel Ferrero, Eulises Rodrıguez, Jesús Pérez, Idrian Garcıa, Gisou Dıaz, and Pedro López-Saura. Characterization of the immunoreactivity of anti-interferon alpha antibodies in myasthenia gravis patients. *Journal of Autoimmunity*, 23(1): 63–73, 2004.

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

Jacob Bergstedt, Sadoune Ait Kaci Azzou, Kristin Tsuo, Anthony Jaquaniello, Alejandra Urrutia, Maxime Rotival, David T. S. Lin, Julia L. MacIsaac, Michael S. Kobor, Matthew L. Albert, et al. The immune factors driving DNA methylation variation in human blood. *Nature Communications*, 13(1):5895, 2022.

José María Bermúdez de Castro, Juan Luis Arsuaga, Eudald Carbonell, Antonio Rosas, I. Martınez, and Marina Mosquera. A hominid from the Lower Pleistocene of Atapuerca, Spain: possible ancestor to Neandertals and modern humans. *Science*, 276(5317):1392–1395, 1997.

Bradley E. Bernstein, John A. Stamatoyannopoulos, Joseph F. Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A. Marra, Arthur L. Beaudet, Joseph R. Ecker, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28 (10):1045–1048, 2010.

T. Bernard Bigdeli, Donghyung Lee, Bradley Todd Webb, Brien P. Riley, Vladimir I. Vladimirov, Ayman H. Fanous, Kenneth S. Kendler, and Silviu-Alin Bacanu. A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. *Bioinformatics*, 32(17):2598–2603, 2016.

Daniel Blanco-Melo, Benjamin E. Nilsson-Payant, Wen-Chun Liu, Skyler Uhl, Daisy Hoagland, Rasmus Møller, Tristan X Jordan, Kohei Oishi, Maryline Panis, David Sachs, et al. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, 181(5):1036–1045, 2020.

Thomas J. Bollyky, Erin N. Hulland, Ryan M. Barber, James K. Collins, Samantha Kiernan, Mark Moses, David M. Pigott, Robert C. Reiner Jr., Reed J. D. Sorensen, Cristiana Abbafati, et al. Pandemic preparedness and COVID-19: an exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021. *The Lancet*, 399(10334):1489–1512, 2022.

A Sina Booeshaghi, Ingileif B Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. *bioRxiv*, pages 2022–05, 2022.

Ernest C. Borden, Ganes C. Sen, Gilles Uze, Robert H. Silverman, Richard M. Ransohoff, Graham R. Foster, and George R. Stark. Interferons at age 50: past, current and future impact on biomedicine. *Nature Reviews Drug Discovery*, 6(12):975–990, 2007.

Marie Bourdon, Caroline Manet, and Xavier Montagutelli. Host genetic susceptibility to viral infections: the role of type I interferon induction. *Genes and Immunity*, 21(6-8):365–379, 2020.

Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

G. Bräuer, H. Broeg, and C. B. Stringer. Earliest Upper Paleolithic crania from Mladeč, Czech Republic, and the question of Neanderthal-modern continuity: metrical evidence from the frontofacial region. *Neanderthals Revisited: New Approaches and Perspectives*, pages 269–279, 2006.

Timothy G. Bromage, Friedemann Schrenk, and Frans W. Zonneveld. Paleoanthropology of the Malawi Rift: an early hominid mandible from the Chiwondo Beds, northern Malawi. *Journal of Human Evolution*, 28(1):71–108, 1995.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Wesley M. Brown, Matthew George Jr., and Allan C. Wilson. Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences*, 76(4):1967–1971, 1979.

Sharon R. Browning, Brian L. Browning, Ying Zhou, Serena Tucci, and Joshua M. Akey. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell*, 173(1):53–61, 2018.

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

James M. Burke, Stephanie L. Moon, Tyler Matheny, and Roy Parker. RNase L reprograms translation by widespread mRNA turnover escaped by antiviral mRNAs. *Molecular Cell*, 75(6):1203–1217, 2019.

F. Macfarlane Burnet. *Self and Not-self: Cellular Immunology*. Melbourne University Press, Carlton, Vic., 1969.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18):3426–3440, 2022.

Long Cai, Chiraj K. Dalal, and Michael B. Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455(7212):485–490, 2008.

Minal Çalışkan, Samuel W. Baker, Yoav Gilad, and Carole Ober. Host genetic variation influences gene expression response to rhinovirus infection. *PLoS Genetics*, 11(4):e1005111, 2015.

Howard M. Cann, Claudia de Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F. Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

Rebecca L. Cann, Mark Stoneking, and Allan C/ Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(6099):31–36, 1987.

Michael J. Cannon, D. Scott Schmid, and Terri B. Hyde. Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Reviews in Medical Virology*, 20(4): 202–213, 2010.

Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

Alastair G. Cardno and Irving I. Gottesman. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *American Journal of Medical Genetics*, 97(1):12–17, 2000.

Oana Carja, Julia L. MacIsaac, Sarah M. Mah, Brenna M. Henn, Michael S. Kobor, Marcus W. Feldman, and Hunter B. Fraser. Worldwide patterns of human epigenetic variation. *Nature Ecology and Evolution*, 1(10):1577–1583, 2017.

Jean-Laurent Casanova and Laurent Abel. Inborn errors of immunity to infection: the rule rather than the exception. *The Journal of Experimental Medicine*, 202(2):197–201, 2005.

Jean-Laurent Casanova and Laurent Abel. The genetic theory of infectious diseases: a brief history and selected illustrations. *Annual Review of Genomics and Human Genetics*, 14:215–243, 2013.

Jean-Laurent Casanova and Laurent Abel. The human genetic determinism of life-threatening infectious diseases: genetic heterogeneity and physiological homogeneity?, 2020.

Jean-Laurent Casanova and Laurent Abel. Mechanisms of viral inflammation and disease in humans. *Science*, 374(6571):1080–1086, 2021.

Jean-Laurent Casanova and Laurent Abel. From rare disorders of immunity to common determinants of infection: Following the mechanistic thread. *Cell*, 185(17):3086–3103, 2022.

Sergi Castellano, Genís Parra, Federico A. Sánchez-Quinto, Fernando Racimo, Martin Kuhlwilm, Martin Kircher, Susanna Sawyer, Qiaomei Fu, Anja Heinze, Birgit Nickel, et al. Patterns of coding variation in the complete exomes of three Neandertals. *Proceedings of the National Academy of Sciences*, 111(18):6666–6671, 2014.

Marina Cella, David Jarrossay, Fabio Facchetti, Olga Alebardi, Hideo Nakajima, Antonio Lanzavecchia, and Marco Colonna. Plasmacytoid monocytes migrate to inflamed lymph nodes and produce large amounts of type i interferon. *Nature Medicine*, 5(8):919–923, 1999.

Arindam Chakrabarti, Shuvojit Banerjee, Luigi Franchi, Yueh-Ming Loo, Michael Gale, Gabriel Núñez, and Robert H. Silverman. RNase L activates the NLRP3 inflammasome during viral infections. *Cell Host and Microbe*, 17(4):466–477, 2015.

Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Computational Biology*, 19(8):1–20, 08 2023.

Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F. O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15(9):2759–2772, 2020.

Jérémy Choin, Javier Mendoza-Revilla, Lara R. Arauna, Sebastián Cuadros-Espinoza, Olivier Cassar, Maximilian Larena, Albert Min-Shan Ko, Christine Harmant, Romain Laurent, Paul Verdu, et al. Genomic insights into population history and biological adaptation in Oceania. *Nature*, 592(7855):583–589, 2021.

Ananyo Choudhury, Shaun Aron, Laura R Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik Bensellak, Gordon Wells, Judit Kumuthini, Daniel Shriner, Yasmina J Fakim, et al. High-depth African genomes inform human migration and health. *Nature*, 586(7831):741–748, 2020.

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

Michael J. Ciancanelli, Sarah X. L. Huang, Priya Luthra, Hannah Garner, Yuval Itan, Stefano Volpi, Fabien G. Lafaille, Céline Trouillet, Mirco Schmolke, Randy A. Albrecht, et al. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. *Science*, 348(6233): 448–453, 2015.
First reported case of genetic susceptibility to influenza pneumonia.

Zoe A. Clarke, Tallulah S. Andrews, Jawairia Atif, Delaram Pouyabahar, Brendan T. Innes, Sonya A. MacParland, and Gary D. Bader. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature Protocols*, 16(6):2749–2764, 2021.

Aurélie Cobat, Caroline J. Gallant, Leah Simkin, Gillian F. Black, Kim Stanley, Jane Hughes, T. Mark Doherty, Willem A. Hanekom, Brian Eley, Jean-Philippe Jaïs, et al. Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. *Journal of Experimental Medicine*, 206(12):2583–2591, 2009.

Moisès Coll Macià, Laurits Skov, Benjamin Marco Peter, and Mikkel Heide Schierup. Different historical generation intervals in human populations inferred from Neanderthal fragment lengths and mutation signatures. *Nature Communications*, 12(1):5317, 2021.

Noah J. Connally, Sumaiya Nazeen, Daniel Lee, Huwenbo Shi, John Stamatoyannopoulos, Sung Chun, Chris Cotsapas, Christopher A. Cassa, and Shamil R. Sunyaev. The missing link between genetic association and regulatory function. *eLife*, 11:e74970, 2022.

William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194, 2009.

Graham Coop. Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics. *aRxiv*, 2022.

Graham Coop and Robert C. Griffiths. Ancestral inference on gene trees under selection. *Theoretical Population Biology*, 66(3):219–232, 2004.

COVID-19 Host Genetics Initiative. The COVID-19 host genetics initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics*, 28(6):715–718, 2020.

COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature*, 600(7889):472–477, 2021.

COVID-19 Host Genetics Initiative. A first update on mapping the human genetic architecture of COVID-19. *Nature*, 608(7921):E1–E10, 2022.

COVID-19 Host Genetics Initiative. A second update on mapping the human genetic architecture of COVID-19. *Nature*, 621(7977):E7–E26, 2023.

Francis H. Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 8, 1958.

Francis H. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

Anna S. E. Cuomo, Tobias Heinen, Danai Vagiaki, Danilo Horta, John C. Marioni, and Oliver Stegle. CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Molecular Systems Biology*, 18(8):e10663, 2022.

Anna S. E. Cuomo, Aparna Nathan, Soumya Raychaudhuri, Daniel G. MacArthur, and Joseph E. Powell. Single-cell genomics meets human genetics. *Nature Reviews Genetics*, pages 1–15, 2023.

Mathias Currat and Laurent Excoffier. Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biology*, 2(12):e421, 2004.

David Curtis. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatric Genetics*, 28(5):85–89, 2018.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, et al. The Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pages 2023–01, 2023.

Mark J. Daly and John D. Rioux. New approaches to gene hunting in IBD. *Inflammatory Bowel Diseases*, 10(3):312–317, 2004.

Michael Dannemann and Janet Kelso. The contribution of Neanderthals to phenotypic variation in modern humans. *The American Journal of Human Genetics*, 101(4):578–589, 2017.

Michael Dannemann, Aida M. Andrés, and Janet Kelso. Introgression of Neandertal- and Denisovan-like haplotypes contributes to adaptive variation in human Toll-like receptors. *The American Journal of Human Genetics*, 98(1):22–33, 2016.

Charles Darwin. *The descent of man, and selection in relation to sex.* John Murray, Albemarle Street, 1871.

Jacob F. Degner, Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E. Crawford, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

Olivier Delaneau, M. Zazhytska, Christelle Borel, G. Giannuzzi, Guillaume Rey, Cédric Howald, S. Kumar, Halit Ongen, Konstantin Popadin, D. Marbach, et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, 364(6439):eaat8266, 2019.

Eric Delson. Palæoanthropology: Palæobiology and age of African *Homo erectus*. *Nature*, 316 (6031):762–763, 1985.

Matthieu Deschamps, Guillaume Laval, Maud Fagny, Yuval Itan, Laurent Abel, Jean-Laurent Casanova, Etienne Patin, and Lluís Quintana-Murci. 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, 2016.

Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4): 997–1004, 1999.

Bo Diao, Chenhui Wang, Yingjun Tan, Xiewan Chen, Ying Liu, Lifen Ning, Li Chen, Min Li, Yueping Liu, Gang Wang, et al. Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). *Frontiers in Immunology*, page 827, 2020.

Raquel Dias and Ali Torkamani. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, 11(1):1–12, 2019.

Antigone S. Dimas, Samuel Deutsch, Barbara E. Stranger, Stephen B. Montgomery, Christelle Borel, Homa Attar-Cohen, Catherine Ingle, Claude Beazley, Maria Gutierrez Arcelus, Magdalena Sekowska, et al. Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science*, 325(5945):1246–1250, 2009.

Christian Dina, David Meyre, Sophie Gallina, Emmanuelle Durand, Antje Körner, Peter Jacobson, Lena MS Carlsson, Wieland Kiess, Vincent Vatin, Cécile Lecœ ur, et al. Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nature Genetics*, 39(6):724–726, 2007.

Anna L. Dixon, Liming Liang, Miriam F. Moffatt, Wei Chen, Simon Heath, Kenny C. C. Wong, Jenny Taylor, Edward Burnett, Ivo Gut, Martin Farrall, et al. A genome-wide association study of global gene expression. *Nature Genetics*, 39(10):1202–1207, 2007.

Theodosius Dobzhansky. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35:125–129, 1973.

Arun Durvasula and Kirk E. Lohmueller. Negative selection on complex traits limits phenotype prediction accuracy between populations. *The American Journal of Human Genetics*, 108(4): 620–631, 2021.

Matthew D. Dyer, T. M. Murali, and Bruno W. Sobral. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathogens*, 4(2):e32, 2008.

Peter Ebert, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372 (6537):eabf7117, 2021.

Michael Eisenstein. Every base everywhere all at once: Pangenomics comes of age. *Nature*, 616 (7957):618–620, 2023.

Avigdor Eldar and Michael B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467 (7312):167–173, 2010.

David Ellinghaus, Frauke Degenhardt, Luis Bujanda, Maria Buti, Agustín Albillos, Pietro Invernizzi, Javier Fernández, Daniele Prati, Guido Baselli, Rosanna Asselta, Marit M. Grimsrud, et al. Genome-wide association study of severe Covid-19 with respiratory failure. *New England Journal of Medicine*, 383(16):1522–1534, 2020.

David Enard and Dmitri A. Petrov. Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. *Cell*, 175(2):360–371, 2018.

David Enard and Dmitri A. Petrov. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 375(1812):20190575, 2020.

David Enard, Le Cai, Carina Gwennap, and Dmitri A. Petrov. Viruses are a dominant driver of protein adaptation in mammals. *eLife*, 5:e12469, 2016.

ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57, 2012.

ENCODE Project Consortium. Expanded encyclopædias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

Adam Eyre-Walker. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(suppl_1): 1752–1756, 2010.

Benjamin P. Fairfax, Seiko Makino, Jayachandran Radhakrishnan, Katharine Plant, Stephen Leslie, Alexander Dilthey, Peter Ellis, Cordelia Langford, Fredrik O. Vannberg, and Julian C. Knight. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nature Genetics*, 44(5):502–510, 2012.

Benjamin P. Fairfax, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, Chris McGee, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343(6175): 1246949, 2014.

H. Christina Fan, Glenn K. Fu, and Stephen P. A. Fodor. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222):1258367, 2015.

Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C. Yung, Joseph L. Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, 13(3):241–244, 2016a.

Shaohua Fan, Matthew E. B. Hansen, Yancy Lo, and Sarah A. Tishkoff. Going global by adapting local: A review of recent human adaptation. *Science*, 354(6308):54–59, 2016b.

Justin C. Fay and Chung-I Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3): 1405–1413, 2000.

E. A. Feingold, P. J. Good, M. S. Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, F. S. Collins, T. R. Gingeras, D. Kampa, E. A. Sekinger, et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):1–13, 2015.

Clive Finlayson, Francisco Giles Pacheco, Joaquín Rodríguez-Vidal, Darren A. Fa, José María Gutierrez López, Antonio Santiago Pérez, Geraldine Finlayson, Ethel Allue, Javier Baena Preysler, Isabel Cáceres, et al. Late survival of Neanderthals at the southernmost extreme of Europe. *Nature*, 443(7113):850–853, 2006.

Ronald A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1918.

Ronald A. Fisher. The genetical theory of natural selection, 1930.

Elise D. Flynn, Athena L. Tsu, Silva Kasela, Sarah Kim-Hellmuth, Francois Aguet, Kristin G. Ardlie, Harmen J. Bussemaker, Pejman Mohammadi, and Tuuli Lappalainen. Transcription factor regulation of eQTL activity across individuals and tissues. *PLoS Genetics*, 18(1):e1009719, 2022.

Claudio Franceschi, Paolo Garagnani, Paolo Parini, Cristina Giuliani, and Aurelia Santoro. Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nature Reviews Endocrinology*, 14(10):576–590, 2018.

Timothy M. Frayling, Nicholas J. Timpson, Michael N. Weedon, Eleftheria Zeggini, Rachel M. Freathy, Cecilia M. Lindgren, John R. B. Perry, Katherine S. Elliott, Hana Lango, Nigel W. Rayner, et al. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894, 2007.

Qiaomei Fu, Mateja Hajdinjak, Oana Teodora Moldovan, Silviu Constantin, Swapan Mallick, Pontus Skoglund, Nick Patterson, Nadin Rohland, Iosif Lazaridis, Birgit Nickel, et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, 524(7564):216–219, 2015.

Yun-Xin Fu and Wen-Hsiung Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3): 693–709, 1993.

Charles P. Fulco, Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, Rockwell Anyoha, Benjamin R. Doughty, Tejal A. Patwardhan, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. *Nature Genetics*, 51(12):1664–1669, 2019.

Matteo Fumagalli, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admettla, Linda Pattini, and Rasmus Nielsen. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, 7(11):e1002355, 2011.

Tjede Funk, Anastasia Pharris, Gianfranco Spiteri, Nick Bundle, Angeliki Melidou, Michael Carr, Gabriel González, Alejandro García-Leon, Fiona Crispie, Lois O'Connor, et al. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Eurosurveillance*, 26(16):2100348, 2021.

L. Gabunia, A. Vekua, and D. Lordkipanidze. New human fossils from Dmanisi, eastern Georgia. *Archæology, Ethnology and Anthropology of Eurasia*, 2(6):128–139, 2001.

Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Francis Galton. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279):135–145, 1889.

Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, GTEx Consortium, Dan L. Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

Andrea Ganna, Karin J. H. Verweij, Michel G. Nivard, Robert Maier, Robbee Wedow, Alexander S. Busch, Abdel Abdellaoui, Shengru Guo, J. Fah Sathirapongsasuti, 23andMe Research Team 16, et al. Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science*, 365(6456):eaat7693, 2019.

Feng Gao and Alon Keinan. High burden of private mutations due to explosive human population growth and purifying selection. *BMC Genomics*, 15(4):1–7, 2014.

Diego Garrido-Martín, Beatrice Borsari, Miquel Calvo, Ferran Reverter, and Roderic Guigó. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nature Communications*, 12(1):727, 2021.

Erik Garrison, Andrea Guarracino, Simon Heumos, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, Sebastian Vorbrugg, Santiago Marco-Sola, Christian Kubica, et al. Building pangenome graphs. *bioRxiv*, pages 2023–04, 2023.

Molly Gasperini, Jacob M. Tome, and Jay Shendure. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, 21(5):292–310, 2020.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, 2021.

Giulio Genovese, David J. Friedman, Michael D. Ross, Laurence Lecordier, Pierrick Uzureau, Barry I. Freedman, Donald W. Bowden, Carl D. Langefeld, Taras K. Oleksyk, Andrea L. Uscinski Knob, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, 329(5993):841–845, 2010.

Jemma L. Geoghegan, Sebastián Duchêne, and Edward C. Holmes. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathogens*, 13(2):e1006215, 2017.

Pierre-Luc Germain, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D. Robinson. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research*, 10, 2021.

Ariel D. H. Gewirtz, F. William Townes, and Barbara E. Engelhardt. Expression QTLs in single-cell sequencing data. *bioRxiv*, pages 2022–08, 2022.

Mahan Ghafari, Peter Simmonds, Oliver G. Pybus, and Aris Katzourakis. Prisoner of War dynamics explains the time-dependent pattern of substitution rates in viruses. *bioRxiv*, pages 2021–02, 2021.

Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10(5):e1004383, 2014.

Claudia Giambartolomei, Jimmy Zhenli Liu, Wen Zhang, Mads Hauberg, Huwenbo Shi, James Boocock, Joe Pickrell, Andrew E. Jaffe, CommonMind Consortium, Bogdan Pasaniuc, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, 34(15):2538–2545, 2018.

Richard E. Giles, Hugues Blanc, Howard M. Cann, and Douglas C. Wallace. Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences*, 77(11):6715–6719, 1980.

Rachel M. Gittelman, Joshua G. Schraiber, Benjamin Vernot, Carmen Mikacenic, Mark M. Wurfel, and Joshua M. Akey. Archaic hominin admixture facilitated adaptation to out-of-Africa environments. *Current Biology*, 26(24):3375–3382, 2016.

David E. Gordon, Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M White, Matthew J. O'Meara, Veronica V. Rezelj, Jeffrey Z. Guo, Danielle L. Swaney, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(7816): 459–468, 2020.

Richard E. Green, Anna-Sapfo Malaspinas, Johannes Krause, Adrian W. Briggs, Philip L. F. Johnson, Caroline Uhler, Matthias Meyer, Jeffrey M. Good, Tomislav Maricic, Udo Stenzel, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–426, 2008.

Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, et al. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, 2010.

T. Ryan Gregory. Animal genome size database. `http://www.genomesize.com`, 2002. Accessed: August 23rd, 2023.

Robert C. Griffiths and Paul Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.

Robert C. Griffiths and Paul Marjoram. An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257, 1997.

Huw S. Groucutt, Michael D. Petraglia, Geoff Bailey, Eleanor M. L. Scerri, Ash Parton, Laine Clark-Balzan, Richard P. Jennings, Laura Lewis, James Blinkhorn, Nick A. Drake, et al. Rethinking the dispersal of *Homo sapiens* out of Africa. *Evolutionary Anthropology*, 24(4):149–164, 2015.

Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, 2014.

Elin Grundberg, Kerrin S. Small, Åsa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping *cis*-and *trans*-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, 2012.

Louise G. Grunnet, Emma Nilsson, Charlotte Ling, Torben Hansen, Oluf Pedersen, Leif Groop, Allan Vaag, and Pernille Poulsen. Regulation and function of *FTO* mRNA expression in human skeletal muscle and subcutaneous adipose tissue. *Diabetes*, 58(10):2402–2408, 2009.

Mónica Gumá, Ana Angulo, Carlos Vilches, Natalia Gómez-Lozano, Núria Malats, and Miguel López-Botet. Imprint of human cytomegalovirus infection on the NK cell receptor repertoire. *Blood*, 104(12):3664–3671, 2004.

Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23(12):2126–2135, 2013.

Sarthak Gupta, Ioanna P. Tatouli, Lindsey B. Rosen, Sarfaraz Hasni, Ilias Alevizos, Zerai G. Manna, Juan Rivera, Chao Jiang, Richard M. Siegel, Steven M. Holland, et al. Distinct functions of autoantibodies against interferon in systemic lupus erythematosus: a comprehensive analysis of anticytokine autoantibodies in common rheumatic diseases. *Arthritis and Rheumatology*, 68(7): 1677–1687, 2016.

James F. Gusella, Nancy S. Wexler, P. Michael Conneally, Susan L. Naylor, Mary Anne Anderson, Rudolph E. Tanzi, Paul C. Watkins, Kathleen Ottina, Margaret R. Wallace, Alan Y. Sakaguchi, et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940): 234–238, 1983.
First instance of a disease gene mapped by studying genetic linkage within families.

Jérôme Hadjadj, Nader Yatim, Laura Barnabei, Aurélien Corneau, Jeremy Boussier, Nikaïa Smith, Hélène Péré, Bruno Charbit, Vincent Bondet, Camille Chenevier-Gobeaux, et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science*, 369(6504): 718–724, 2020.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, 2019.

Mateja Hajdinjak, Fabrizio Mafessoni, Laurits Skov, Benjamin Vernot, Alexander Hübner, Qiaomei Fu, Elena Essel, Sarah Nagel, Birgit Nickel, Julia Richter, et al. Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry. *Nature*, 592(7853):253–257, 2021.

John B. S. Haldane. The effect of variation of fitness. *The American Naturalist*, 71(735):337–349, 1937.

John B. S. Haldane. Disease and evolution. *La Ricerca Scientifica*, 19:68–76, 1949.

Rachita Ramachandra Halehalli and Hampapathalu Adimurthy Nagarajaram. Molecular principles of human virus protein–protein interactions. *Bioinformatics*, 31(7):1025–1033, 2015.

Maike M. K. Hansen, Ravi V. Desai, Michael L. Simpson, and Leor S. Weinberger. Cytoplasmic amplification of transcriptional noise generates substantial cell-to-cell variability. *Cell Systems*, 7 (4):384–397, 2018.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Kelley Harris and Rasmus Nielsen. The genetic cost of Neanderthal introgression. *Genetics*, 203(2): 881–891, 2016.

Daniel L. Hartl and Andrew G. Clark. *Principles of population genetics*. Sinauer Associates, Inc. Publishers, fourth edition, 2018.

Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26(5):672–675, 2020.

Nicholas Hernández, Isabelle Melki, Huie Jing, Tanwir Habib, Susie S. Y. Huang, Jeffrey Danielson, Tomasz Kula, Scott Drutman, Serkan Belkaya, Vimel Rattina, et al. Life-threatening influenza pneumonitis in a child with inherited IRF9 deficiency. *Journal of Experimental Medicine*, 215 (10):2567–2585, 2018.

Andy I. R. Herries, Jesse M. Martin, A. B. Leece, Justin W. Adams, Giovanni Boschian, Renaud Joannes-Boyau, Tara R. Edwards, Tom Mallett, Jason Massey, Ashleigh Murszewski, et al. Contemporaneity of *Australopithecus*, *Paranthropus*, and early *Homo erectus* in South Africa. *Science*, 368(6486):eaaw7293, 2020.

Israel Hershkovitz, Ofer Marder, Avner Ayalon, Miryam Bar-Matthews, Gal Yasur, Elisabetta Boaretto, Valentina Caracuta, Bridget Alex, Amos Frumkin, Mae Goder-Goldberger, et al. Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans. *Nature*, 520(7546):216–219, 2015.

Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, pages 1–23, 2023.

Glenn Hickey, Jean Monlong, Jana Ebler, Adam M Novak, Jordan M Eizenga, Yan Gao, Tobias Marschall, Heng Li, and Benedict Paten. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, pages 1–11, 2023.

Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6):685–691, 2019.

Tom Higham, Katerina Douka, Rachel Wood, Christopher Bronk Ramsey, Fiona Brock, Laura Basell, Marta Camps, Alvaro Arrizabalaga, Javier Baena, Cecillio Barroso-Ruíz, et al. The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, 512(7514):306–309, 2014.

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

Joel N. Hirschhorn and Mark J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

Jules A. Hoffmann, Fotis C. Kafatos, Charles A. Janeway Jr., and R. A. B. Ezekowitz. Phylogenetic perspectives in innate immunity. *Science*, 284(5418):1313–1318, 1999.

Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 610–611, 2014.

Farhad Hormozdiari, Martijn van de Bunt, Ayellet V. Segre, Xiao Li, Jong Wha J. Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics*, 99 (6):1245–1260, 2016.

Julie E. Horowitz, Jack A. Kosmicki, Amy Damask, Deepika Sharma, Genevieve H. L. Roberts, Anne E. Justice, Nilanjana Banerjee, Marie V. Coignet, Ashish Yadav, Joseph B. Leader, et al. Genome-wide analysis provides genetic evidence that *ACE2* influences COVID-19 risk and yields risk scores associated with severe disease. *Nature Genetics*, 54(4):382–392, 2022.

Jean-Jacques Hublin. The origin of Neandertals. *Proceedings of the National Academy of Sciences*, 106(38):16022–16027, 2009.

Jean-Jacques Hublin, Abdelouahed Ben-Ncer, Shara E. Bailey, Sarah E. Freidline, Simon Neubauer, Matthew M. Skinner, Inga Bergmann, Adeline Le Cabec, Stefano Benazzi, Katerina Harvati, et al. New fossils from Jebel irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, 546 (7657):289–292, 2017.

Jean-Pierre Hugot, Mathias Chamaillard, Habib Zouali, Suzanne Lesage, Jean-Pierre Cézard, Jacques Belaiche, Sven Almer, Curt Tysk, Colm A O'Morain, Miquel Gassull, et al. Association of *nod2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411 (6837):599–603, 2001.

Lucas T. Husquin, Maxime Rotival, Maud Fagny, Hélène Quach, Nora Zidane, Lisa M. McEwen, Julia L. MacIsaac, Michael S. Kobor, Hugues Aschard, Etienne Patin, et al. Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. *Genome Biology*, 19(1):1–17, 2018.

Thomas H. Huxley. *Evidence as to Man's Place in Nature*. Williams and Norgate, 1863.

Leonardo N. M. Iasi, Harald Ringbauer, and Benjamin M. Peter. An extended admixture pulse model reveals the limitations to Human–Neandertal introgression dating. *Molecular Biology and Evolution*, 38(11):5156–5174, 2021.

International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063): 1299–1320, 2005.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
First draft of the human genome using hierarchical shotgun sequencing.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.
Although the euchromatic sequence of the human genome was declared finished in April 2003, the associated paper was published several months later.

Alick Isaacs and Jean Lindenmann. Virus interference: The interferon. *Proceedings of the Royal Society of London. Series B-Biological Sciences*, 147(927):258–267, 1957.

Kazuyoshi Ishigaki, Yuta Kochi, Akari Suzuki, Yumi Tsuchida, Haruka Tsuchiya, Shuji Sumitomo, Kensuke Yamaguchi, Yasuo Nagafuchi, Shinichiro Nakachi, Rika Kato, et al. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nature Genetics*, 49(7):1120–1125, 2017.

Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–1167, 2011.

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, 2014.

Evelyn Jagoda, Daniel J. Lawson, Jeffrey D. Wall, David Lambert, Craig Muller, Michael Westaway, Matthew Leavesley, Terence D. Capellini, Marta Mirazón Lahr, Pascale Gerbault, et al. Disentangling immediate adaptive introgression from selection on standing introgressed variation in humans. *Molecular Biology and Evolution*, 35(3):623–630, 2018.

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343 (6172):776–779, 2014.

Mattias Jakobsson, Sonja W. Scholz, Paul Scheet, J. Raphael Gibbs, Jenna M. VanLiere, Hon-Chung Fung, Zachary A. Szpiech, James H. Degnan, Kai Wang, Rita Guerreiro, et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003, 2008.

Charles A. Janeway. Approaching the asymptote? evolution and revolution in immunology. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 54, pages 1–13. Cold Spring Harbor Laboratory Press, 1989.

Aline Jelenkovic, Reijo Sund, Yoon-Mi Hur, Yoshie Yokoyama, Jacob B. Hjelmborg, Sören Möller, Chika Honda, Patrik K. E. Magnusson, Nancy L. Pedersen, Syuichi Ooki, et al. Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. *Scientific Reports*, 6(1):28496, 2016.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Bryony A. Jones, Delia Grace, Richard Kock, Silvia Alonso, Jonathan Rushton, Mohammed Y. Said, Declan McKeever, Florence Mutua, Jarrah Young, John McDermott, et al. Zoonosis emergence linked to agricultural intensification and environmental change. *Proceedings of the National Academy of Sciences*, 110(21):8399–8404, 2013.

Ivan Juric, Simon Aeschbacher, and Graham Coop. The strength of selection against Neanderthal introgression. *PLoS Genetics*, 12(11):e1006340, 2016.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1): 89–94, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv*, 2020.

Elinor K. Karlsson, Dominic P. Kwiatkowski, and Pardis C. Sabeti. Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, 15(6):379–393, 2014.

Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1):1–29, 2023.

Alon Keinan and Andrew G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.

Alon Keinan, James C. Mullikin, Nick Patterson, and David E. Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics*, 39(10):1251–1255, 2007.

David R. Kelley. Cross-species regulatory sequence activity prediction. *PLoS Computational Biology*, 16(7):e1008050, 2020.

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018.

Gaspard Kerner, Jeremy Choin, and Lluís Quintana-Murci. Ancient DNA as a tool for medical research. *Nature Medicine*, 29(5):1048–1051, 2023a.

Gaspard Kerner, Anna-Lena Neehus, Quentin Philippot, Jonathan Bohlen, Darawan Rinchai, Nacim Kerrouche, Anne Puel, Shen-Ying Zhang, Stéphanie Boisson-Dupuis, Laurent Abel, et al. Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe. *Cell Genomics*, 3(2), 2023b.

Peter V. Kharchenko. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, 18(7):723–732, 2021.

Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

Amit V. Khera, Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, 2018.

Yuseob Kim and Wolfgang Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, 2002.

Sarah Kim-Hellmuth, Matthias Bechheim, Benno Pütz, Pejman Mohammadi, Yohann Nédélec, Nicholas Giangreco, Jessica Becker, Vera Kaiser, Nadine Fricker, Esther Beier, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nature Communications*, 8(1):266, 2017.

Sarah Kim-Hellmuth, FranAguet, Meritxell Oliva, Manuel Muñoz Aguirre, Silva Kasela, Valentin Wucher, Stephane E. Castel, Andrew R. Hamel, Ana Viñuela, Amy L. Roberts, Serghei Mangul, Xiaoquan Wen, Alvaro N Barbeira, Diego Garrido-Martín, Brian B. Nadel, Yuxin Zou, Rodrigo Bonazzola, Jie Quan, Andrew Brown, Ángel Martínez-Pérez, José Manuel Soria, The GTEx Consortium, Gad Getz, Emmanouil T. Dermitzakis, Kerrin S. Small, Matthew Stephens, Hualin S. Xi, Hae Kyung Im, Roderic Guigó, Ayellet V Segrè, Barbara E. Stranger, Kristin G. Ardlie, and Tuuli Lappalainen. Cell type–specific genetic regulation of gene expression across human tissues. *Science*, 369(6509):eaaz8528, 2020.

William H. Kimbel, Donald C. Johanson, and Yoel Rak. Systematic assessment of a maxilla of *Homo* from Hadar, Ethiopia. *American Journal of Physical Anthropology*, 103(2):235–262, 1997.

Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

Motoo Kimura and George H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561, 1964.

William King. The reputed fossil man of the Neanderthal. In *Quarterly Journal of Science*, pages 88–97. 1864.

Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005. **First genome-wide association study ever published.**

N. Klöting, D. Schleinitz, K. Ruschke, J. Berndt, A. Fasshauer, M .and Tönjes, M. R. Schön, P. Kovacs, M. Stumvoll, and M. Blüher. Inverse relationship between obesity and *FTO* gene expression in visceral adipose tissue in humans. *Diabetologia*, 51:641–647, 2008.

M. S. Ko, Hiromitsu Nakauchi, and Naomi Takahashi. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *The EMBO journal*, 9(9):2835–2842, 1990.

Robert Koch. Die ätiologie der Tuberkulose. *Berliner Klinische Wochenschrift*, 19:221–230, 1882.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019.

Athanasios Kousathanas, Erola Pairo-Castineira, Konrad Rawlik, Alex Stuckey, Christopher A Odhams, Susan Walker, Clark D Russell, Tomas Malinauskas, Yang Wu, Jonathan Millar, et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature*, 607(7917): 97–103, 2022.

Johannes Krause, Ludovic Orlando, David Serre, Bence Viola, Kay Prüfer, Michael P. Richards, Jean-Jacques Hublin, Catherine Hänni, Anatoly P. Derevianko, and Svante Pääbo. Neanderthals in central Asia and Siberia. *Nature*, 449(7164):902–904, 2007.

Johannes Krause, Qiaomei Fu, Jeffrey M. Good, Bence Viola, Michael V. Shunkov, Anatoli P. Derevianko, and Svante Pääbo. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464(7290):894–897, 2010.

Christine Kreuder Johnson, Peta L. Hitchens, Tierra Smiley Evans, Tracey Goldstein, Kate Thomas, Andrew Clements, Damien O. Joly, Nathan D. Wolfe, Peter Daszak, William B. Karesh, et al. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific Reports*, 5(1):14830, 2015.

Matthias Krings, Anne Stone, Ralf W. Schmitz, Heike Krainitzki, Mark Stoneking, and Svante Pääbo. Neandertal DNA sequences and the origin of modern humans. *Cell*, 90(1):19–30, 1997.

Stephen M. Krone and Claudia Neuhauser. Ancestral processes with selection. *Theoretical Population Biology*, 51(3):210–237, 1997.

Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

Christoph Kuppe, Mahmoud M. Ibrahim, Jennifer Kranz, Xiaoting Zhang, Susanne Ziegler, Javier Perales-Patón, Jitske Jansen, Katharina C. Reimer, James R. Smith, Ross Dobie, et al. Decoding myofibroblast origins in human kidney fibrosis. *Nature*, 589(7841):281–286, 2021.

Kazuki Kurimoto, Yukihiro Yabuta, Yasuhide Ohinata, Yukiko Ono, Kenichiro D. Uno, Rikuhiro G. Yamada, Hiroki R. Ueda, and Mitinori Saitou. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Research*, 34 (5):e42–e42, 2006.

Kazuki Kurimoto, Yukihiro Yabuta, Yasuhide Ohinata, and Mitinori Saitou. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nature Protocols*, 2(3):739–752, 2007.

Carles Lalueza-Fox. Neanderthal assimilation? *Nature Ecology and Evolution*, 5(6):711–712, 2021.

Frederick M. Lang, Kevin M.-C. Lee, John R. Teijaro, Burkhard Becher, and John A. Hamilton. Gm-csf-based treatments in covid-19: reconciling opposing therapeutic approaches. *Nature Reviews Immunology*, 20(8):507–514, 2020.

Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar González-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, 22(1):1–20, 2021.

Meave G. Leakey and Richard E. Leakey. The fossil hominids and an introduction to their context, 1968-1974. *Koobi Fora Research Project*, 1, 1978.

Danyel Lee, Jérémie Le Pen, Ahmad Yatim, Beihua Dong, Yann Aquino, Masato Ogishi, Rémi Pescarmona, Estelle Talouarn, Darawan Rinchai, Peng Zhang, et al. Inborn errors of OAS–RNase L in SARS-CoV-2–related multisystem inflammatory syndrome in children. *Science*, 379 (6632):eabo3627, 2022.

Jeong Seok Lee, Seongwan Park, Hye Won Jeong, Jin Young Ahn, Seong Jin Choi, Hoyoung Lee, Baekgyu Choi, Su Kyung Nam, Moa Sa, Ji-Soo Kwon, et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Science Immunology*, 5(49):eabd1554, 2020.

Mark N. Lee, Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, Portia I. Chipendo, F. Ann Ran, Kamil Slowikowski, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343(6175): 1246980, 2014.

Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

I. Leitch, E. Johnston, J. Pellicer, O. Hidalgo, and M. D. Bennett. Angiosperm DNA C-values database (release 9.0). `http://data.kew.org/cvalues`, 2019. Accessed: August 23rd, 2023.

Juliette Leon, Daniel A. Michelson, Judith Olejnik, Kaitavjeet Chowdhary, Hyung Suk Oh, Adam J. Hume, Silvia Galván-Peña, Yangyang Zhu, Felicia Chen, Brinda Vijaykumar, et al. A virus-specific monocyte inflammatory phenotype is induced by SARS-CoV-2 at the immune–epithelial interface. *Proceedings of the National Academy of Sciences*, 119(1):e2116853118, 2022.

Marcel Levi, Jecko Thachil, Toshiaki Iba, and Jerrold H Levy. Coagulation abnormalities and thrombosis in patients with COVID-19. *The Lancet Hæmatology*, 7(6):e438–e440, 2020.

Roger Lewin. Africa: Cradle of modern humans. *Science*, 237(4820):1292–1295, 1987.

Richard C. Lewontin. The apportionment of human diversity. *Evolutionary Biology: Volume 6*, pages 381–398, 1972.

Richard C. Lewontin and Jesse Krakauer. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1):175–195, 1973.

Lili Li, Jie Li, Meiling Gao, Huimin Fan, Yanan Wang, Xin Xu, Chunfeng Chen, Junxiao Liu, Jocelyn Kim, Roghiyh Aliyari, et al. Interleukin-8 as a biomarker for disease prognosis of coronavirus disease-2019 patients. *Frontiers in Immunology*, 11:602395, 2021.

Shuang Li, Katharina T. Schmid, Dylan H. de Vries, Maryna Korshevniuk, Corinna Losert, Roy Oelen, Irene V. van Blokland, Hilde E. Groot, Morris A. Swertz, Pim van der Harst, et al. Identification of genetic variants that impact gene co-expression relationships using large-scale single-cell data. *Genome Biology*, 24(1):1–37, 2023.

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

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K Lucas, Jean Monlong, Haley J. Abel, et al. A draft human pangenome reference. *Nature*, 617(7960):312–324, 2023.
First draft of the human pangenome.

Hye Kyung Lim, Sarah X. L. Huang, Jie Chen, Gaspard Kerner, Olivier Gilliaux, Paul Bastard, Kerry Dobbs, Nicholas Hernandez, Nicolas Goudin, Mary L. Hasek, et al. Severe influenza pneumonitis in children with inherited TLR3 deficiency. *Journal of Experimental Medicine*, 216 (9):2038–2056, 2019.

Xuanyao Liu, Yang I. Li, and Jonathan K. Pritchard. Trans effects on gene expression can drive omnigenic inheritance. *Cell*, 177(4):1022–1034, 2019.

Luke R. Lloyd-Jones, Alexander Holloway, Allan McRae, Jian Yang, Kerrin Small, Jing Zhao, Biao Zeng, Andrew Bakshi, Andres Metspalu, Manolis Dermitzakis, et al. The genetic architecture of gene expression in peripheral blood. *The American Journal of Human Genetics*, 100(2):228–237, 2017.

Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K. Finucane, Brendan K. Bulik-Sullivan, Samuela J. Pollack, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Teresa R. de Candia, Sang Hong Lee, Naomi R. Wray, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):1385–1392, 2015.

Kirk E. Lohmueller. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genetics*, 10(5):e1004379, 2014.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F. Müller, Daniel C. Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

Pierre Luisi, David Alvarez-Ponce, Marc Pybus, Mario A Fares, Jaume Bertranpetit, and Hafid Laayouni. Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. *Genome Biology and Evolution*, 7(4):1141–1154, 2015.

Aaron T. L. Lun, Karsten Bach, and John C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):1–14, 2016a.

Aaron T. L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5, 2016b.

Chongyuan Luo, Petra Hajkova, and Joseph R. Ecker. Dynamic DNA methylation: In the right place at the right time. *Science*, 361(6409):1336–1340, 2018.

Stuart MacGregor, Belinda K. Cornes, Nicholas G. Martin, and Peter M. Visscher. Bias, precision and heritability of self-reported and clinically measured height in australian twins. *Human Genetics*, 120:571–580, 2006.

Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Fabrizio Mafessoni, Steffi Grote, Cesare de Filippo, Viviane Slon, Kseniya A. Kolobova, Bence Viola, Sergey V. Markin, Manjusha Chintalapati, Stephane Peyrégne, Laurits Skov, et al. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences*, 117(26):15132–15136, 2020.

Thomas Magg, Tsubasa Okano, Lars M. Koenig, Daniel F. R. Boehmer, Samantha L. Schwartz, Kento Inoue, Jennifer Heimall, Francesco Licciardi, Julia Ley-Zaporozhan, Ronald M. Ferdman, et al. Heterozygous OAS1 gain-of-function variants cause an autoinflammatory immunodeficiency. *Science Immunology*, 6(60):eabf9564, 2021.

Anna-Sapfo Malaspinas, Michael C. Westaway, Craig Muller, Vitor C. Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Y. Cheng, Jacob E. Crawford, et al. A genomic history of Aboriginal Australia. *Nature*, 538(7624):207–214, 2016.

Krishnamurthy Malathi, Jayashree M. Paranjape, Elena Bulanova, Minsub Shim, Jeanna M. Guenther-Johnson, Pieter W. Faber, Thomas E. Eling, Bryan R. G. Williams, and Robert H. Silverman. A transcriptional signaling pathway in the IFN system mediated by 2'-5'-oligoadenylate activation of RNase L. *Proceedings of the National Academy of Sciences*, 102(41):14533–14538, 2005.

Gustave Malécot. *The mathematics of heredity*. Masson et Cie., Paris, 1948.

Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

Jérémy Manry, Guillaume Laval, Etienne Patin, Simona Fornarino, Yuval Itan, Matteo Fumagalli, Manuela Sironi, Magali Tichit, Christiane Bouchier, Jean-Laurent Casanova, et al. Evolutionary genetic dissection of human interferons. *Journal of Experimental Medicine*, 208(13):2747–2759, 2011.

Stephanie Marciniak and George H. Perry. Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics*, 18(11):659–674, 2017.

Peter V. Markov, Mahan Ghafari, Martin Beer, Katrina Lythgoe, Peter Simmonds, Nikolaos I. Stilianakis, and Aris Katzourakis. The evolution of SARS-CoV-2. *Nature Reviews Microbiology*, 21(6):361–379, 2023.

Eirini Marouli, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R. Wood, Troels R. Kjaer, Rebecca S. Fine, Yingchang Lu, Claudia Schurmann, Heather M. Highland, et al. Rare and low-frequency coding variants alter human adult height. *Nature*, 542(7640):186–190, 2017.

Carla Márquez-Luna, Po-Ru Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Alkes L. Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology*, 41(8):811–823, 2017.

Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, 2019.

Iain Mathieson. The omnigenic model and polygenic prediction of complex traits. *The American Journal of Human Genetics*, 108(9):1558–1563, 2021.

Iain Mathieson and Aylwyn Scally. What is ancestry? *PLoS Genetics*, 16(3):e1008624, 2020.

Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.

Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

Ian McDougall, Francis H. Brown, and John G. Fleagle. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433(7027):733–736, 2005.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018.

Finlay McNab, Katrin Mayer-Barber, Alan Sher, Andreas Wack, and Anne O'Garra. Type i interferons in infectious disease. *Nature Reviews Immunology*, 15(2):87–103, 2015.

Peter Brian Medawar. Old age and natural death. *Modern Quarterly*, 2:30–49, 1946.

Peter Brian Medawar. An unsolved problem of biology. 1952.

Fernando L. Mendez, Joseph C. Watkins, and Michael F. Hammer. Global genetic variation at *OAS1* provides evidence of archaic admixture in Melanesian populations. *Molecular Biology and Evolution*, 29(6):1513–1520, 2012.

Fernando L. Mendez, Joseph C. Watkins, and Michael F. Hammer. Neandertal origin of genetic variation at the cluster of OAS immunity genes. *Molecular Biology and evolution*, 30(4):798–801, 2013.

Tim R. Mercer, Shane Neph, Marcel E. Dinger, Joanna Crawford, Martin A. Smith, Anne-Marie J. Shearwood, Eric Haugen, Cameron P. Bracken, Oliver Rackham, John A. Stamatoyannopoulos, et al. The human mitochondrial transcriptome. *Cell*, 146(4):645–658, 2011.

Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–226, 2012.

Matthias Meyer, Juan-Luis Arsuaga, Cesare de Filippo, Sarah Nagel, Ayinuer Aximu-Petri, Birgit Nickel, Ignacio Martínez, Ana Gracia, José María Bermúdez de Castro, Eudald Carbonell, et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature*, 531 (7595):504–507, 2016.

Marta Mirazón Lahr and Robert A. Foley. Multiple dispersals and modern human origins. *Evolutionary Anthropology*, 3(2):48–60, 1994.

Dennis O. Mook-Kanamori, Catharina E. M. van Beijsterveldt, Eric A. P. Steegers, Yurii S. Aulchenko, Hein Raat, Albert Hofman, Paul H. Eilers, Dorret I. Boomsma, and Vincent W. V. Jaddoe. Heritability estimates of body size in fetal life and early childhood. *PLoS One*, 7(7): e39901, 2012.

Priya Moorjani, Sriram Sankararaman, Qiaomei Fu, Molly Przeworski, Nick Patterson, and David E. Reich. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proceedings of the National Academy of Sciences*, 113(20): 5652–5657, 2016.

Hakhamanesh Mostafavi, Jeffrey P. Spence, Sahin Naqvi, and Jonathan K. Pritchard. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. *bioRxiv*, pages 2022–05, 2022.

Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

Akiko Nagai, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, Akiko Tamakoshi, Zentaro Yamagata, Taisei Mushiroda, et al. Overview of the BioBank Japan Project: study design and profile. *Journal of Epidemiology*, 27(Supplement III):S2–S8, 2017.

Luca Nanni, Stefano Ceri, and Colin Logie. Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. *Genome Biology*, 21(1):1–25, 2020.

Aparna Nathan, Samira Asgari, Kazuyoshi Ishigaki, Cristian Valencia, Tiffany Amariuta, Yang Luo, Jessica I. Beynor, Yuriy Baglaenko, Sara Suliman, Alkes L. Price, et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*, 606(7912):120–128, 2022.

Drew Neavin, Anne Senabouth, Jimmy Tsz Hang Lee, Aida Ripoll, sc-eQTLGen Consortium, Lude Franke, Shyam Prabhakar, Chun Jimmie Ye, Davis J. McCarthy, Marta Melé, et al. Demuxafy: Improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *bioRxiv*, pages 2022–03, 2022.

Yohann Nédélec, Joaquín Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, Andrew Freiman, Aaron J. Sams, Steven Hebert, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*, 167 (3):657–669, 2016.

Andrew Ng. Neural networks and deep learning. `https://www.coursera.org/learn/neural-networks-deep-learning`, 2021a. Accessed in February 2023.

Andrew Ng. Improving deep neural networks. `https://www.coursera.org/learn/deep-neural-networks`, 2021b. Accessed in February 2023.

Andrew Ng. Structuring machine learning projects. `https://www.coursera.org/learn/machine-learning-projects`, 2021c. Accessed in February 2023.

Andrew Ng. Convolutional neural networks. `https://www.coursera.org/learn/convolutional-neural-networks`, 2021d. Accessed in February 2023.

Andrew Ng. Sequence models. `https://www.coursera.org/learn/nlp-sequence-models`, 2021e. Accessed in February 2023.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv*, 2023.

Alexandra C. Nica, Stephen B. Montgomery, Antigone S. Dimas, Barbara E. Stranger, Claude Beazley, Inês Barroso, and Emmanouil T. Dermitzakis. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, 6(4): e1000895, 2010.

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

Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J. Hubisz, Andrew G. Clark, and Carlos Bustamante. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11): 1566–1575, 2005.

Rasmus Nielsen, Joshua M. Akey, Mattias Jakobsson, Jonathan K. Pritchard, Sarah Tishkoff, and Eske Willerslev. Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310, 2017.

Aaron Novick and Milton Weiner. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences*, 43(7):553–566, 1957.

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.
Publication of the telomere-to-telomere CHM13 assembly of the human genome.

Luke J. O'Connor, Armin P. Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L. Price. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, 2019.

Roy Oelen, Dylan H. de Vries, Harm Brugge, M. Grace Gordon, Martijn Vochteloo, single-cell eQTLGen consortium, BIOS Consortium, Chun J. Ye, Harm-Jan Westra, Lude Franke, and Monique G. P. van der Wijst. Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nature Communications*, 13(1):3267, 2022.

Yasunori Ogura, Denise K Bonen, Naohiro Inohara, Dan L Nicolae, Felicia F Chen, Richard Ramos, Heidi Britton, Thomas Moran, Reda Karaliuskas, Richard H Duerr, et al. A frameshift mutation in nod2 associated with susceptibility to crohn's disease. *Nature*, 411(6837):603–606, 2001.

Meritxell Oliva, Manuel Muñoz-Aguirre, Sarah Kim-Hellmuth, Valentin Wucher, Ariel DH Gewirtz, Daniel J Cotter, Princy Parsana, Silva Kasela, Brunilda Balliu, Ana Viñuela, et al. The impact of sex on gene expression across human tissues. *Science*, 369(6509):eaba3066, 2020.

William Ollier, Tim Sprosen, and Tim Peakman. UK Biobank: from concept to reality. 2005.

Mary B. O'Neill, Hélène Quach, Julien Pothlichet, Yann Aquino, Aurélie Bisiaux, Nora Zidane, Matthieu Deschamps, Valentina Libri, Milena Hasan, Shen-Ying Zhang, et al. Single-cell and bulk RNA-sequencing reveal differences in monocyte susceptibility to influenza A virus infection between Africans and Europeans. *Frontiers in Immunology*, 12:768189, 2021.

Halit Ongen, Andrew A. Brown, Olivier Delaneau, Nikolaos I. Panousis, Alexandra C. Nica, GTEx Consortium, and Emmanouil T. Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nature Genetics*, 49(12):1676–1683, 2017.

Ludovic Orlando, Pierre Darlu, Michel Toussaint, Dominique Bonjean, Marcel Otte, and Catherine Hänni. Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. *Current Biology*, 16(11):R400–R402, 2006.

Judith A. Owen, Jenni Punt, Sharon A. Stranford, and Patricia P. Jones. *Kuby Immunology*. WH Freeman New York, seventh edition, 2013.

Megan O'Driscoll, Gabriel Ribeiro Dos Santos, Lin Wang, Derek A. T. Cummings, Andrew S. Azman, Juliette Paireau, Arnaud Fontanet, Simon Cauchemez, and Henrik Salje. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, 590(7844):140–145, 2021.

Luca Pagani, Daniel John Lawson, Evelyn Jagoda, Alexander Mörseburg, Anders Eriksson, Mario Mitt, Florian Clemente, Georgi Hudjashov, Michael DeGiorgio, Lauri Saag, et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538(7624):238–242, 2016.

Erola Pairo-Castineira, Sara Clohisey, Lucija Klaric, Andrew D. Bretherick, Konrad Rawlik, Dorota Pasko, Susan Walker, Nick Parkinson, Max Head Fourman, Clark D. Russell, et al. Genetic mechanisms of critical illness in COVID-19. *Nature*, 591(7848):92–98, 2021.

Sandra Panem, Irene J. Check, Dorothy Henriksen, and Jan Vilček. Antibodies to alpha-interferon in a patient with systemic lupus erythematosus. *Journal of Immunology*, 129(1):1–3, 1982.

Christos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietiläinen, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

Yongjin Park, Abhishek Sarkar, Kunal Bhutani, and Manolis Kellis. Multi-tissue polygenic models for transcriptome-wide association studies. *bioRxiv*, page 107623, 2017.

Louis Pasteur. *Mémoire sur les corpuscules organisés qui existent dans l'atmosphère: examen de la doctrine des générations spontanées*. Annales de chimie et de physique, 1862.

Etienne Patin, Milena Hasan, Jacob Bergstedt, Vincent Rouilly, Valentina Libri, Alejandra Urrutia, Cécile Alanio, Petar Scepanovic, Christian Hammer, Friederike Jönsson, et al. Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nature Immunology*, 19(3):302–314, 2018.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

Karl Pearson and Alice Lee. On the laws of inheritance in man: Inheritance of physical characters. *Biometrika*, 2(4):357–462, 1903.

Sen Pei, Teresa K Yamana, Sasikiran Kandula, Marta Galanti, and Jeffrey Shaman. Burden and characteristics of COVID-19 in the United States during 2020. *Nature*, 598(7880):338–341, 2021.

Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena Mariana Malpica Serrano, Alexander Crits-Christoph, Nathaniel L. Matteson, et al. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science*, 377(6609):960–966, 2022.

Richard K. Perez, M. Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C. Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, et al. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589): eabf1970, 2022.

George H. Perry. Evolutionary medicine, 2021.

Martin Petr, Svante Pääbo, Janet Kelso, and Benjamin Vernot. Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences*, 116(5):1639–1644, 2019.

Barbara Piasecka, Darragh Duffy, Alejandra Urrutia, Hélène Quach, Etienne Patin, Céline Posseme, Jacob Bergstedt, Bruno Charbit, Vincent Rouilly, Cameron R. MacPherson, et al. 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–E497, 2018.

Joseph K. Pickrell, 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. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010.

Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):1–10, 2015.

Alice B. Popejoy and Stephanie M. Fullerton. Genomics is failing on diversity. *Nature*, 538(7624): 161–164, 2016.

Joseph E. Powell, Anjali K. Henders, Allan F. McRae, Margaret J. Wright, Nicholas G. Martin, Emmanouil T. Dermitzakis, Grant W. Montgomery, and Peter M. Visscher. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Research*, 22(3):456–466, 2012.

Thomas Pradeu and Edgardo D. Carosella. The self model and the conception of biological identity in immunology. *Biology and Philosophy*, 21:235–252, 2006.

Alkes L. Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics*, 7(2):e1001317, 2011.

Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, 2014.

Kay Prüfer, Cesare se Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot, Laurits Skov, Pinghsun Hsieh, Stéphane Peyrégne, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, 358(6363):655–658, 2017.

Juliet R. C. Pulliam, Cari van Schalkwyk, Nevashan Govender, Anne von Gottberg, Cheryl Cohen, Michelle J. Groome, Jonathan Dushoff, Koleka Mlisana, and Harry Moultrie. Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science*, 376 (6593):eabn4947, 2022.

Pengfei Qin and Mark Stoneking. Denisovan ancestry in East Eurasian and native American populations. *Molecular Biology and Evolution*, 32(10):2665–2674, 2015.

Hélène Quach, Maxime Rotival, Julien Pothlichet, Yong-Hwee Eddie Loh, Michael Dannemann, Nora Zidane, Guillaume Laval, Etienne Patin, Christine Harmant, Marie Lopez, et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. *Cell*, 167(3):643–656, 2016.

Thomas P. Quinn, Ionas Erb, Mark F. Richardson, and Tamsyn M. Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, 2018.

Lluís Quintana-Murci. Human immunology through the lens of evolutionary genetics. *Cell*, 177(1): 184–199, 2019.

Lluís Quintana-Murci and Andrew G. Clark. Population genetic tools for dissecting innate immunity in humans. *Nature Reviews Immunology*, 13(4):280–293, 2013.

Lluís Quintana-Murci, Ornella Semino, Hans-Jürgen Bandelt, Giuseppe Passarino, Ken McElreavey, and A. Silvana Santachiara-Benerecetti. Genetic evidence of an early exit of homo sapiens sapiens from Africa through eastern Africa. *Nature Genetics*, 23(4):437–441, 1999.

Fernando Racimo, Davide Marnetto, and Emilia Huerta-Sánchez. Signatures of archaic adaptive introgression in present-day human populations. *Molecular Biology and Evolution*, 34(2):296–317, 2017.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: methods, analysis and insights from training gopher. *arXiv*, 2021.

Aaron P. Ragsdale, Elizabeth G. Weaver, Timothy D.and Atkinson, Eileen G. Hoal, Marlo Möller, Brenna M. Henn, and Simon Gravel. A weakly structured stem for human origins in Africa. *Nature*, pages 1–9, 2023.

Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006.

Daniel R. Ram, Cordelia Manickam, Brady Hueber, Hannah L. Itell, Sallie R. Permar, Valerie Varner, and R. Keith Reeves. Tracking KLRC2 (NKG2C)+ memory-like NK cells in SIV+ and rhCMV+ rhesus macaques. *PLoS Pathogens*, 14(5):e1007104, 2018.

Sohini Ramachandran, Omkar Deshpande, Charles C. Roseman, Noah A. Rosenberg, Marcus W. Feldman, and L. Luca Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, 102(44):15942–15947, 2005.

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

Morten Rasmussen, Xiaosen Guo, Yong Wang, Kirk E. Lohmueller, Simon Rasmussen, Anders Albrechtsen, Line Skotte, Stinus Lindgreen, Mait Metspalu, Thibaut Jombart, et al. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, 334(6052):94–98, 2011.

Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *eLife*, 6:e27041, 2017.

David E. Reich and Eric S. Lander. On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502–510, 2001.

David E. Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, 2010.

Xianwen Ren, Wen Wen, Xiaoying Fan, Wenhong Hou, Bin Su, Pengfei Cai, Jiesheng Li, Yang Liu, Fei Tang, Fan Zhang, et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 184(7):1895–1913, 2021.

Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):284, 2018.

Genevieve H. L. Roberts, Raghavendran Partha, Brooke Rhead, Spencer C. Knight, Danny S. Park, Marie V. Coignet, Miao Zhang, Nathan Berkowitz, David A. Turrisini, Michael Gaddis, et al. Expanded COVID-19 phenotype definitions reveal distinct patterns of genetic association and protective effects. *Nature Genetics*, 54(4):374–381, 2022.

Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):1–9, 2010.

Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385): 176–182, 2018.

Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.

Maxime Rotival and Lluís Quintana-Murci. Functional consequences of archaic introgression and their impact on fitness. *Genome Biology*, 21(1):1–4, 2020.

Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Lin He, Akira Sawa, Alicia R. Martin, et al. Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, 54(5):573–580, 2022.

J. Graham Ruby, Kevin M. Wright, Kristin A. Rand, Amir Kermany, Keith Noto, Don Curtis, Neal Varner, Daniel Garrigan, Dmitri Slinkov, Ilya Dorfman, et al. Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics*, 210(3):1109–1124, 2018.

Pardis C. Sabeti, David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, Jill V. Platko, Nick J. Patterson, Gavin J. McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

Pratha Sah, Meagan C. Fitzpatrick, Charlotte F. Zimmer, Elaheh Abdollahi, Lyndon Juden-Kelly, Seyed M. Moghadas, Burton H. Singer, and Alison P. Galvani. Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, 118(34):e2109229118, 2021.

Saori Sakaue, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba, Akira Narita, Takahiro Konuma, Kenichi Yamamoto, Masato Akiyama, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics*, 53(10):1415–1424, 2021.

Aaron J. Sams, Anne Dumaine, Yohann Nédélec, Vania Yotova, Carolina Alfieri, Jerome E. Tanner, Philipp W. Messer, and Luis B. Barreiro. Adaptively introgressed neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biology*, 17(1): 1–15, 2016.

Grant Sanderson. Three-blue one-brown (3B1B). `https://www.youtube.com/@3blue1brown`, 2020. Accessed between January 2020 and October 2022.

Sriram Sankararaman, Nick Patterson, Heng Li, Svante Pääbo, and David E. Reich. The date of interbreeding between Neandertals and modern humans. *PLoS Genetics*, 2012.

Sriram Sankararaman, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David E. Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357, 2014.

Sriram Sankararaman, Swapan Mallick, Nick Patterson, and David E. Reich. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology*, 26(9): 1241–1247, 2016.

Abhishek K. Sarkar, Po-Yuan Tung, John D. Blischak, Jonathan E. Burnett, Yang I. Li, Matthew Stephens, and Yoav Gilad. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genetics*, 15(4):e1008045, 2019.

Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

Benjamin J. Schmiedel, Divya Singh, Ariel Madrigal, Alan G. Valdovino-Gonzalez, Brandie M. White, Jose Zapardiel-Gonzalo, Brendan Ha, Gokmen Altay, Jason A. Greenbaum, Graham McVicker, et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell*, 175(6):1701–1715, 2018.

William M. Schneider, Meike Dittmann Chevillotte, and Charles M. Rice. Interferon-stimulated genes: a complex web of host defenses. *Annual Review of Immunology*, 32:513–545, 2014.

Otto Schoetensack. *Der unterkiefer des Homo heidelbergensis aus den Sanden von Mauer bei Heidelberg*. Wilhelm Engelmann, 1908.

Benjamin H. Schott, Liuyang Wang, Xinyu Zhu, Alfred T. Harding, Emily R. Ko, Jeffrey S. Bourgeois, Erica J. Washington, Thomas W. Burke, Jack Anderson, Emma Bergstrom, et al. Single-cell genome-wide association reveals that a nonsynonymous variant in ERAP1 confers increased susceptibility to influenza virus. *Cell genomics*, 2(11), 2022.

Angelo Scuteri, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, et al. Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLoS Genetics*, 3(7):e115, 2007.

Guy Sella and Nicholas H. Barton. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annual Review of Genomics and Human Genetics*, 20:461–493, 2019.

David Serre, André Langaney, Mario Chech, Maria Teschler-Nicola, Maja Paunovic, Philippe Mennecier, Michael Hofreiter, Göran Possnert, and Svante Pääbo. No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biology*, 2(3):e57, 2004.

Chunxuan Shao and Thomas Höfer. Robust classification of single-cell transcriptome data by non-negative matrix factorization. *Bioinformatics*, 33(2):235–242, 2017.

Janie F Shelton, Anjali J Shastri, Chelsea Ye, Catherine H Weldon, Teresa Filshtein-Sonmez, Daniella Coker, Antony Symons, Jorge Esparza-Gordillo, Stella Aslibekyan, et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nature Genetics*, 53(6):801–808, 2021.

Jay Shendure, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, 2005.

Chao Shi and Eric G. Pamer. Monocyte recruitment during infection and inflammation. *Nature Reviews Immunology*, 11(11):762–774, 2011.

Dongjie Shi, Lei Ao, Hua Yu, Yongzhi Xia, Juan Li, Wenjie Zhong, and Haijian Xia. Chromobox homolog 8 (CBX8) in human tumor carcinogenesis and prognosis: a pancancer analysis using multiple databases. *Frontiers in Genetics*, 12:745277, 2021.

Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99 (1):139–153, 2016.

Frederick P. Siegal, Norimitsu Kadowaki, Michael Shodell, Patricia A. Fitzgerald-Bocarsly, Kokila Shah, Stephen Ho, Svetlana Antonenko, and Yong-Jun Liu. The nature of the principal type 1 interferon-producing cells in human blood. *Science*, 284(5421):1835–1837, 1999.

Martin Silvert, Lluís Quintana-Murci, and Maxime Rotival. Impact and evolutionary determinants of Neanderthal introgression on transcriptional and post-transcriptional regulation. *The American Journal of Human Genetics*, 104(6):1241–1250, 2019.

Corinne N. Simonti, Benjamin Vernot, Lisa Bastarache, Erwin Bottinger, David S. Carrell, Rex L. Chisholm, David R. Crosslin, Scott J. Hebbring, Gail P. Jarvik, Iftikhar J. Kullo, et al. The phenotypic legacy of admixture between modern humans and Neandertals. *Science*, 351(6274): 737–741, 2016.

Nasa Sinnott-Armstrong, Sahin Naqvi, Manuel Rivas, and Jonathan K. Pritchard. Gwas of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife*, 10:e58615, 2021.

Jouni Sirén and Benedict Paten. GBZ file format for pangenome graphs. *Bioinformatics*, 38(22): 5012–5018, 2022.

Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019.

Pontus Skoglund and Mattias Jakobsson. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences*, 108(45):18301–18306, 2011.

Viviane Slon, Fabrizio Mafessoni, Benjamin Vernot, Cesare de Filippo, Steffi Grote, Bence Viola, Mateja Hajdinjak, Stéphane Peyrégne, Sarah Nagel, Samantha Brown, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, 561(7721):113–116, 2018.

Scott Smemo, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L. Credidio, Débora R. Sobreira, Nora F. Wasserman, et al. Obesity-associated variants within *fto* form long-range functional connections with *irx3*. *Nature*, 507 (7492):371–375, 2014.

Mashaal Sohail, María Palma-Martínez, Amanda Y Chong, Consuelo D. Quinto-Cortés, Carmina Barberena-Jonas, Santiago G. Medina-Muñoz, Aaron Ragsdale, Guadalupe Delgado-Sánchez, Luis Pablo Cruz-Hervert, Leticia Ferreyra-Reyes, et al. Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature*, 2023.

Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

Yassine Souilmi, M. Elise Lauterbur, Ray Tobler, Christian D. Huber, Angad S. Johar, Shayli Varasteh Moradi, Wayne A. Johnston, Nevan J. Krogan, Kirill Alexandrov, and David Enard. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *Current Biology*, 31(16):3504–3514, 2021.

Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019.

Fred Spoor, Meave G. Leakey, Patrick N. Gathogo, Frank H. Brown, Susan C. Antón, Ian McDougall, Christopher Kiarie, Frederick K. Manthi, and Louise N. Leakey. Implications of new early *Homo* fossils from Ileret, east of Lake Turkana, Kenya. *Nature*, 448(7154):688–691, 2007.

Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

Emily Stephenson, Gary Reynolds, Rachel A. Botting, Fernando J. Calero-Nieto, Michael D. Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B. Worlock, Masahiro

Yoshida, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, 27(5):904–916, 2021.

Aaron J. Stern, Peter R. Wilton, and Rasmus Nielsen. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, 15 (9):e1008384, 2019.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

Chris B. Stringer and Jean-Jacques Hublin. New age estimates for the Swanscombe hominid, and their significance for human evolution. *Journal of Human Evolution*, 6(37):873–877, 1999.

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

Patrick F. Sullivan, Kenneth S. Kendler, and Michael C. Neale. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*, 60(12):1187–1192, 2003.

Keer Sun and Dennis W. Metzger. Inhibition of pulmonary antibacterial defense by interferon-γ during recovery from influenza infection. *Nature Medicine*, 14(5):558–564, 2008.

Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, 2018.

Fahim Syed, Wei Li, Ryan F. Relich, Patrick M. Russell, Shanxiang Zhang, Michelle K. Zimmerman, and Qigui Yu. Excessive matrix metalloproteinase-1 and hyperactivation of endothelial cells occurred in COVID-19 patients and were associated with the severity of COVID-19. *The Journal of Infectious Diseases*, 224(1):60–69, 2021.

Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.

Takehiro Takahashi, Mallory K. Ellingson, Patrick Wong, Benjamin Israelow, Carolina Lucas, Jon Klein, Julio Silva, Tianyang Mao, Ji Eun Oh, Maria Tokuyama, et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature*, 588(7837):315–320, 2020.

Kiyoshi Takeda, Tsuneyasu Kaisho, and Shizuo Akira. Toll-like receptors. *Annual Review of Immunology*, 21(1):335–376, 2003.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.
Seminal paper describing the single-cell RNA-sequencing method.

Fuchou Tang, Catalin Barbacioru, Siqin Bao, Caroline Lee, Ellen Nordman, Xiaohui Wang, Kaiqin Lao, and M Azim Surani. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell*, 6(5):468–478, 2010.

Louise H. Taylor, Sophia M. Latham, and Mark E. J. Woolhouse. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1411):983–989, 2001.

Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Sureshnee Pillay, Emmanuel James San, Nokukhanya Msomi, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854):438–443, 2021.

Mukund Thattai and Alexander van Oudenaarden. Stochastic gene expression in fluctuating environments. *Genetics*, 167(1):523–530, 2004.

The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

The GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675): 204–213, 2017.

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

The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

Charles A. Thomas. The genetic organization of chromosomes. *Annual Review of Genetics*, 5(1): 237–256, 1971.

Stéphanie Thomas, Vincent Rouilly, Etienne Patin, Cécile Alanio, Annick Dubois, Cécile Delval, Louis-Guillaume Marquier, Nicolas Fauchoux, Seloua Sayegrih, Muriel Vray, et al. The Milieu Intérieur study—an integrative approach for study of human immunological variance. *Clinical immunology*, 157(2):277–293, 2015.

Sarah A. Tishkoff, Floyd A. Reed, Françoise R. Friedlaender, Christopher Ehret, Alessia Ranciaro, Alain Froment, Jibril B. Hirbo, Agnes A. Awomoyi, Jean-Marie Bodo, Ogobara Doumbo, et al. The genetic structure and history of Africans and African Americans. *Science*, 324(5930):1035–1044, 2009.

Eric J. Topol. As artificial intelligence goes multimodal, medical applications multiply, 2023.

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

Lawrence H. Uricchio. Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. *Human Genetics*, 139(1):5–21, 2020.

Lawrence H. Uricchio, Dmitri A. Petrov, and David Enard. Exploiting selection at linked sites to infer the rate and strength of adaptation. *Nature Ecology and Evolution*, 3(6):977–984, 2019.

Heléne Valladas, Jean-Louis Joron, Georges Valladas, Baruch Arensburg, Ofer Bar-Yosef, Anna Belfer-Cohen, Paul Goldberg, Henri Laville, Lilliane Meignen, Yoel Rak, et al. Thermoluminescence dates for the Neanderthal burial site at Kebara in Israel. *Nature*, 330(6144):159–160, 1987.

Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology*, 11(6):e1004333, 2015.

Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Monique G. P. van der Wijst, Harm Brugge, Dylan H. de Vries, Patrick Deelen, Morris A. Swertz, LifeLines Cohort Study, BIOS Consortium, and Lude Franke. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics*, 50(4):493–497, 2018a.

Monique G. P. van der Wijst, Dylan H. de Vries, Harm Brugge, Harm-Jan Westra, and Lude Franke. An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Medicine*, 10(1):1–15, 2018b.

Monique G. P. van der Wijst, Dylan H. de Vries, Hilde E. Groot, Gosia Trynka, Chung-Chau Hon, Marc-Jan Bonder, Oliver Stegle, M. C. Nawijn, Youssef Idaghdour, Pim van der Harst, et al. The single-cell eQTLGen consortium. *eLife*, 9:e52155, 2020.

Monique G. P. van der Wijst, Sara E. Vazquez, George C. Hartoularos, Paul Bastard, Tianna Grant, Raymund Bueno, David S. Lee, John R. Greenland, Yang Sun, Richard Perez, et al. Type I interferon autoantibodies are associated with systemic immune alterations in patients with COVID-19. *Science Translational Medicine*, 13(612):eabh2624, 2021.

Neeltje van Doremalen, Trenton Bushmaker, Dylan H. Morris, Myndi G. Holbrook, Amandine Gamble, Brandi N. Williamson, Azaibi Tamin, Jennifer L. Harcourt, Natalie J. Thornburg, Susan I. Gerber, et al. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *New England Journal of Medicine*, 382(16):1564–1567, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

David Venet, F. Pecasse, Carine Maenhaut, and Hugues Bersini. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(suppl_1):S279–S287, 2001.

Craig J. Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. doi: 10.1126/science.1058040. **First draft of the human genome using whole-genome shotgun sequencing**.

Benjamin Vernot and Joshua M. Akey. Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, 343(6174):1017–1021, 2014.

Benjamin Vernot, Serena Tucci, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf, Rachel M. Gittelman, Michael Dannemann, Steffi Grote, Rajiv C. McCoy, Heather Norton, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*, 352(6282): 235–239, 2016.

Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.

Fernando A. Villanea and Joshua G. Schraiber. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nature Ecology and Evolution*, 3(1):39–44, 2019.

Ana Viñuela, Andrew A. Brown, Alfonso Buil, Pei-Chien Tsai, Matthew N. Davies, Jordana T. Bell, Emmanouil T. Dermitzakis, Timothy D. Spector, and Kerrin S. Small. Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Human Molecular Genetics*, 27(4):732–741, 2018.

Peter M. Visscher, Sarah E. Medland, Manuel A. R. Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2(3):e41, 2006.

Erik Volz, Verity Hill, John T. McCrone, Anna Price, David Jorgensen, Áine O'Toole, Joel Southgate, Robert Johnson, Ben Jackson, Fabricia F. Nascimento, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell*, 184(1):64–75, 2021.

Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, et al. 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, 2021.

Ha My T. Vy and Yuseob Kim. A composite-likelihood method for detecting incomplete selective sweep from population genomic data. *Genetics*, 200(2):633–649, 2015.

Alan Walker and Richard E. Leakey. *The Nariokotome Homo erectus skeleton*. Harvard University Press, 1993.

Yisong Y. Wan. *GATA3*: a master of many trades in immune regulation. *Trends in Immunology*, 35(6):233–242, 2014.

Chao Wang, Yan-Ling Chen, Wan-Ping Bian, Shao-Lin Xie, Ge-Le Qi, Li Liu, Phyllis R. Strauss, Ji-Xing Zou, and De-Sheng Pei. Deletion of *mstna* and *mstnb* impairs the immune system and affects growth performance in zebrafish. *Fish and Shellfish Immunology*, 72:572–580, 2018.

Eric Y. Wang, Tianyang Mao, Jon Klein, Yile Dai, John D. Huck, Jillian R. Jaycox, Feimei Liu, Ting Zhou, Benjamin Israelow, Patrick Wong, et al. Diverse functional autoantibodies in patients with COVID-19. *Nature*, 595(7866):283–288, 2021.

Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B*, 82(5):1273–1300, 2020.

Haidong Wang, Katherine R. Paulson, Spencer A. Pease, Stefanie Watson, Haley Comfort, Peng Zheng, Aleksandr Y. Aravkin, Catherine Bisignano, Ryan M. Barber, Tahiya Alam, et al. Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020-21. *The Lancet*, 399(10334):1513–1536, 2022.

Ying Wang, Masahiro Kanai, Taotao Tan, Mireille Kamariza, Kristin Tsuo, Kai Yuan, Wei Zhou, Yukinori Okada, The BioBank Japan Project, Hailiang Huang, Patrick Turley, Elizabeth G. Atkinson, and Alicia R. Martin. Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genomics*, 2023a.

Ying Wang, Shinichi Namba, Esteban Lopera, Sini Kerminen, Kristin Tsuo, Kristi Läll, Masahiro Kanai, Wei Zhou, Kuan-Han Wu, Marie-Julie Favé, et al. Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genomics*, 3(1), 2023b.

Sebastian M. Waszak, Olivier Delaneau, Andreas R. Gschwind, Helena Kilpinen, Sunil K. Raghav, Robert M. Witwicki, Andrea Orioli, Michael Wiederkehr, Nikolaos I. Panousis, Alisa Yurovsky, et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell*, 162(5):1039–1050, 2015.

Simone Weber, Victoria Kehl, Johanna Erber, Karolin I. Wagner, Ana-Marija Jetzlsperger, Teresa Burrell, Kilian Schober, Philipp Schommers, Max Augustin, Claudia S. Crowell, et al. CMV seropositivity is a potential novel risk factor for severe COVID-19 in non-geriatric patients. *PLoS One*, 17(5):e0268530, 2022.

Douglas E. Weidemann, James Holehouse, Abhyudai Singh, Ramon Grima, and Silke Hauf. The minimal intrinsic stochasticity of constitutively expressed eukaryotic genes is sub-Poissonian. *Science Advances*, 9(32):eadh5138, 2023.

Omer Weissbrod, Masahiro Kanai, Huwenbo Shi, Steven Gazal, Wouter J. Peyrot, Amit V. Khera, Yukinori Okada, Alicia R. Martin, Hilary K. Finucane, et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nature Genetics*, 54(4):450–458, 2022.

Xiaoquan Wen, Roger Pique-Regi, and Francesca Luca. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genetics*, 13(3):e1006646, 2017.

Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, Benjamin P. Fairfax, Katharina Schramm, Joseph E. Powell, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 2013.

Harm-Jan Westra, Danny Arends, Tõnu Esko, Marjolein J. Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen, Hanieh Yaghootkar, Benjamin P. Fairfax, Anand Kumar Andiappan, et al. Cell specific eQTL analysis without sorting cells. *PLoS Genetics*, 11(5):e1005223, 2015.

K.A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). `http://www.genome.gov/sequencingcostsdata`, 2022. Accessed: June 6th, 2023.

Ray White, Scott Woodward, Mark Leppert, Peter O'Connell, Mark Hoff, John Herbst, Jean-Marc Lalouel, Michael Dean, and George Vande Woude. A closely linked genetic marker for cystic fibrosis. *Nature*, 318(6044):382–384, 1985.

Tim D. White, Berhane Asfaw, David DeGusta, Henry Gilbert, Gary D. Richards, Gen Suwa, and F. Clark Howell. Pleistocene Homo sapiens from middle Awash, Ethiopia. *Nature*, 423(6941): 742–747, 2003.

Aaron J. Wilk, Arjun Rustagi, Nancy Q. Zhao, Jonasel Roque, Giovanny J. Martínez-Colón, Julia L. McKechnie, Geoffrey T. Ivison, Thanmayi Ranganath, Rosemary Vergara, Taylor Hollis, et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine*, 26(7):1070–1076, 2020.

Quin F. Wills, Kenneth J. Livak, Alex J. Tipping, Tariq Enver, Andrew J. Goldson, Darren W. Sexton, and Chris Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*, 31(8):748–752, 2013.

Milford H. Wolpoff, Xinzhi Wu, and Alan G. Thorne. Modern *Homo sapiens* origins: a general theory of hominid evolution involving the fossil evidence from East Asia. *The origins of modern humans: a world survey of the fossil evidence*, 6:411–483, 1984.

Andrew R. Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H. Pers, Stefan Gustafsson, Audrey Y. Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46 (11):1173–1186, 2014.

Bernard A. Wood. Hominid cranial remains. *Koobi Fora Research Project*, 4, 1991.

World Health Organization. Global tuberculosis report 2019, 2019.

World Health Organization. 'Coronavirus Disease 2019' Dashboard. `https://covid19.who.int`, 2020a. Accessed: September 27th, 2023.

World Health Organization. Tracking SARS-CoV-2 variants. `https://www.who.int/activities/tracking-SARS-CoV-2-variants`, 2020b. Accessed: September 27th, 2023.

Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17(10):1520–1528, 2007.

Fred A. Wright, Patrick F. Sullivan, Andrew I. Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, et al. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, 46(5):430–437, 2014.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

Sewall Wright. The interpretation of population structure by f-statistics with special regard to systems of mating. *Evolution*, pages 395–420, 1965.

Nan Miles Xi and Jingyi Jessica Li. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Systems*, 12(2):176–194, 2021.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021.

Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, Lei Huang, Chao Zhang, Shuhong Liu, Peng Zhao, Hongxia Liu, Li Zhu, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, 8(4):420–422, 2020.

Kensuke Yamaguchi, Kazuyoshi Ishigaki, Akari Suzuki, Yumi Tsuchida, Haruka Tsuchiya, Shuji Sumitomo, Yasuo Nagafuchi, Fuyuki Miya, Tatsuhiko Tsunoda, Hirofumi Shoda, et al. Splicing qtl analysis focusing on coding sequences reveals mechanisms for disease susceptibility loci. *Nature Communications*, 13(1):4659, 2022.

Ryo Yamamoto, Ryan Chung, Juan Manuel Vázquez, Huanjie Sheng, Philippa L. Steinberg, Nilah M. Ioannidis, and Peter H. Sudmant. Tissue-specific impacts of aging and genetics on gene expression patterns in humans. *Nature Communications*, 13(1):5803, 2022.

Feng Yan, David R. Powell, David J. Curtis, and Nicholas C. Wong. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, 21:1–16, 2020.

Fan Yang, Jiebiao Wang, Brandon L. Pierce, Lin S. Chen, François Aguet, Kristin G. Ardlie, Beryl B. Cummings, Ellen T. Gelfand, Gad Getz, Kane Hadley, et al. Identifying *cis*-mediators for *trans*-eQTLs across many human tissues using genomic mediation analysis. *Genome Research*, 27(11):1859–1871, 2017.

Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

Chen Yao, Roby Joehanes, Andrew D. Johnson, Tianxiao Huan, Tonu Esko, Saixia Ying, Jane E. Freedman, Joanne Murabito, Kathryn L. Lunetta, Andres Metspalu, et al. Sex-and age-interacting eQTLs in human complex diseases. *Human Molecular Genetics*, 23(7):1947–1956, 2014.

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

Jingfei Yao, Dongmei Wu, Chunyan Zhang, Ting Yan, Yiheng Zhao, Hongyu Shen, Kaili Xue, Xun Huang, Zihao Wang, and Yifu Qiu. Macrophage IRX3 promotes diet-induced obesity and metabolic inflammation. *Nature Immunology*, 22(10):1268–1279, 2021.

Seyhan Yazar, José Alquicira-Hernández, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R. P. Taylor, Linda Clarke, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376(6589): eabf3041, 2022.

Chun Jimmie Ye, Jenny Chen, Alexandra-Chloé Villani, Rachel E. Gate, Meena Subramaniam, Tushar Bhangale, Mark N. Lee, Towfique Raj, Raktima Raychowdhury, Weibo Li, et al. Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of *ERAP2* transcripts under balancing selection. *Genome Research*, 28(12):1812–1825, 2018.

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

Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, et al. A saturated map of common genetic variants associated with human height. *Nature*, 610(7933):704–712, 2022.

Xin Yi, Yu Liang, Emilia Huerta-Sánchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, Xun Xu, Hui Jiang, Nicolas Vinckenbosch, Thorfinn Sand Korneliussen, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–78, 2010.

Luke Zappia and Fabian J. Theis. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biology*, 22:1–18, 2021.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Computational Biology*, 14(6):e1006245, 2018.

Hugo Zeberg and Svante Pääbo. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*, 587(7835):610–612, 2020.

Hugo Zeberg and Svante Pääbo. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proceedings of the National Academy of Sciences*, 118(9): e2026309118, 2021.

Jian Zeng, Ronald De Vlaming, Yang Wu, Matthew R. Robinson, Luke R. Lloyd-Jones, Loic Yengo, Chloe X. Yap, Angli Xue, Julia Sidorenko, Allan F. McRae, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics*, 50(5):746–753, 2018.

Daniel Zenklusen, Daniel R. Larson, and Robert H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural and Molecular Biology*, 15(12):1263–1271, 2008.

Kai Zhang, James D. Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B. Poirion, Yunjiang Qiu, Yang E. Li, Kyle J. Gaulton, Allen Wang, et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell*, 184(24):5985–6001, 2021.

Qian Zhang, Paul Bastard, Zhiyong Liu, Jérémie Le Pen, Marcela Moncada-Velez, Jie Chen, Masato Ogishi, Ira K. D. Sabli, Stephanie Hodeib, Cecilia Korol, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science*, 370(6515):eabd4570, 2020.

Qian Zhang, Paul Bastard, Aurélie Cobat, and Jean-Laurent Casanova. Human genetic and immunological determinants of critical COVID-19 pneumonia. *Nature*, 603(7902):587–598, 2022.

Jiao Zhao, Yan Yang, Hanping Huang, Dong Li, Dongfeng Gu, Xiangfeng Lu, Zheng Zhang, Lei Liu, Ting Liu, Yukun Liu, et al. Relationship between the ABO blood group and the coronavirus disease 2019 (COVID-19) susceptibility. *Clinical Infectious Diseases*, 73(2):328–331, 2021.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017.

Daria V. Zhernakova, Patrick Deelen, Martijn Vermaat, Maarten van Iterson, Michiel van Galen, Wibowo Arindrarto, Peter van't Hof, Hailiang Mei, Freerk van Dijk, Harm-Jan Westra, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*, 49(1):139–145, 2017.

Amanda B. Zheutlin and David A. Ross. Polygenic risk scores: what are they good for? *Biological Psychiatry*, 83(11):e51–e53, 2018.

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, 2018.

Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273, 2020a.

Sirui Zhou, Guillaume Butler-Laporte, Tomoko Nakanishi, David R. Morrison, Jonathan Afilalo, Marc Afilalo, Laetitia Laurent, Maik Pietzner, Nicola Kerrison, Kaiqiong Zhao, et al. A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nature Medicine*, 27(4):659–667, 2021.

Wei Zhou, Masahiro Kanai, Kuan-Han H. Wu, Humaira Rasheed, Kristin Tsuo, Jibril B. Hirbo, Ying Wang, Arjun Bhattacharya, Huiling Zhao, Shinichi Namba, et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2(10), 2022.

Yonggang Zhou, Binqing Fu, Xiaohu Zheng, Dongsheng Wang, Changcheng Zhao, Yingjie Qi, Rui Sun, Zhigang Tian, Xiaoling Xu, and Haiming Wei. Pathogenic T-cells and inflammatory monocytes incite inflammatory storms in severe COVID-19 patients. *National Science Review*, 7 (6):998–1002, 2020b.

Lan Zhu and Carlos D. Bustamante. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics*, 170(3):1411–1421, 2005.

Xun Zhu, Travers Ching, Xinghua Pan, Sherman M. Weissman, and Lana Garmire. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ*, 5:e2888, 2017.

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

Or Zuk, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

Appendix

A The partitioning of phenotypic variance

Decomposing the innate and acquired contributors of phenotypic variation is a central theme of genetics. Because it is an easily accessible, albeit complex phenotype, much work has focused on the 'nature versus nurture' of human height. In fact, the field of quantitative genetics bloomed from studies of height at the turn of the 20$^{\text{th}}$ century. Interestingly, advances in this domain were tightly linked to the development of core concepts in modern statistics, such as regression and variance.

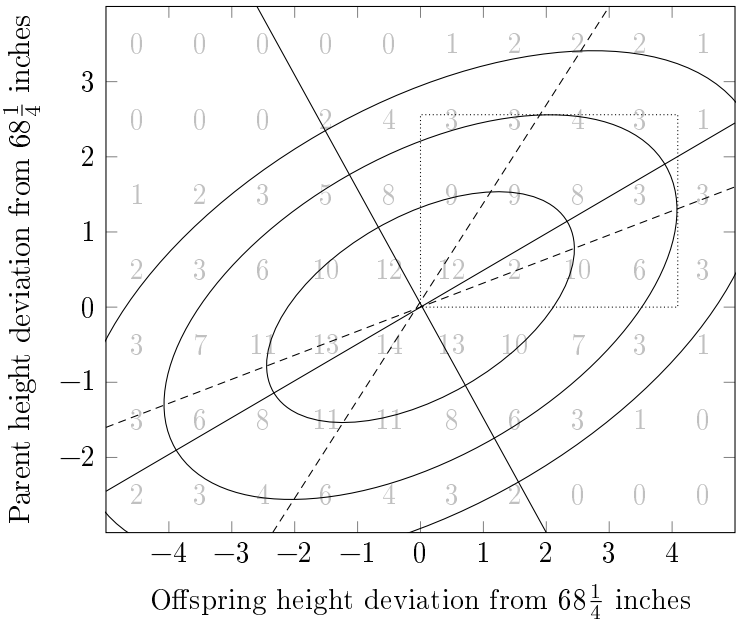


Figure A.1 | The first graphically explicit bivariate relationship. Francis Galton produced the first known scatter plot showing an explicit relationship between two variables. The scattered numbers show the joint distribution of deviations in the heights of offspring and their parents from a reference value of $68\frac{1}{4}$ inches. The concentric ellipsoids show the isodensity contours at three magnitudes of height deviation. The solid lines are the major and minor axes of the ellipsoids and correspond to the principal components of the data. The tangent lines of the middle ellipsoid are shown by the dotted lines. The dashed lines link the center of the data to the tangential points of the ellipsoids. Adapted from Galton (1886).

In 1885, during his presidential address to the Anthropological Section of the British Association, Francis Galton produced data showing that a part of variation in height across individuals could be explained by differences in the height of their parents (Galton, 1886; Jelenkovic et al., 2016). These data were presented in the form of the first ever scatter plot showing an explicit relationship between two variables, as illustrated in Figure A.1.

Geometrical insights from this graph laid the bases for the development of correlation and regression by Galton and others (Galton, 1889; Pearson, 1895; Lee Rodgers and Nicewander, 1988). Based on Galton's theorization of the 'co-relation' between two variables, Karl Pearson developed the mathematical formula for his eponymous 'product-moment correlation coefficient'. Several years later, Pearson and Lee (1903) built on Galton's data to show height correlations among relatives, strengthening the evidence for a genetic basis of height.

In 1918, Ronald Fisher formalized the statistical principles behind modern quantitative genetics by showing that Mendelian inheritance could lead to continuous variation in quantitative traits (Fisher, 1918). In particular, Fisher estimated the 'heritability' of height as a measure of the proportion of phenotypic variation attributable to genetics, and showed that continuous variation in height could be explained through the joint effect of many loci, in line with his 'infinitesimal' model of polygenic inheritance.

Fisher (1918) modeled the value of continuous quantitative trait z in an individual as the sum of a genotypic component G and an environmental component E,

$$Y = G + E. \tag{A.1}$$

In this context, G can be viewed as the average phenotypic value $\overline{Y}$ summarized over all possible environments an individual is likely to cross. Because each diploid parent transmits a haploid genome down to their offspring, only a fraction of value G is effectively transmitted. Fisher called this subcomponent the 'additive' genetic value A. In contrast, the 'non-additive' genetic subcomponent D contains the effects of other alleles within loci (e.g., dominance) and between loci (i.e., epistasis).

From Equation (A.1), and overlooking complex epistatic effects, the phenotypic value of an individual can be expressed as a linear combination of genetic and environmental components,

$$Y = \alpha_Y + (A + D) + E, \tag{A.2}$$

where α_Y is the mean phenotypic value across all individuals in the population. Then, the expected phenotypic value of an individual across a universe of environments can be calculated from the respective maternal and paternal additive genetic values $A_{\female}$ and $A_{\male}$,

$$\mathbb{E}[Y] = \alpha_Y + \frac{A_{\female} + A_{\male}}{2}. \tag{A.3}$$

The ideas introduced by Fisher (1918) were seminal in quantitative genetics, but they also preluded his own advances in statistics. In particular, Fisher (1918) defined the 'variance' of a random variable as the square σ^2 of its standard deviation.

Because additive genetic effects A are average effects estimated from the regression of phenotype Z on genotype, they are by construction independent from other effects in the model. Hence, the variance of the phenotypic value of an invidual is the sum of the variances of its components,

$$\sigma^2_Y = \sigma^2_A + \sigma^2_D + \sigma^2_E. \tag{A.4}$$

Such partitioning of variance foreshadowed Fisher's development of the analysis of variance (ANOVA).

Fisher's variance decomposition is also basal to his definition of heritability. From this view, broad-sense heritability H^2 is defined as the ratio between the variance of the genetic component of phenotype—which may include dominance and epistatic effects—in an individual over the total variance of the phenotype,

$$H^2 = \frac{\sigma^2_A + \sigma^2_D}{\sigma^2_Y}. \tag{A.5}$$

In contrast, narrow-sense heritability h^2 considers only the part of phenotypic variance explained by the additive genetic component,

$$h^2 = \frac{\sigma^2_A}{\sigma^2_Y}. \tag{A.6}$$

Heritability is a central concept of quantitative and evolutionary genetics. For instance, the non-null heritability of a trait is an essential condition for directional selection. Even if individuals with a particular value of the trait perform better than the rest within a generation, if its genetic basis is negligeable or if it is not passed to the next generation, selection will not target the trait.

Narrow-sense heritability was traditionally estimated by so-called 'twin-studies' that leverage the ressemblance between genetically identical monozygotic twin pairs on the one hand, and on the other dizygotic twin pairs that share half of their alleles identical-by-descent (IBD) in expectation. For example, MacGregor et al. (2006) estimated the narrow-sense heritability of height at approximately 80% using a cohort of over 800 monozygotic and dizygotic twin pairs, respectively.

The proportion of alleles shared IBD by parents and their offspring can also be leveraged to estimate trait heritability. For example, Mook-Kanamori et al. (2012) used parent-offspring trios to estimate the heritability of body height—using femur length as a proxy—during fetal life and childhood. Interestingly, the authors estimated increased heritability from mid-pregnancy to infancy.

More recent methods allow to estimate heritability from population-level genotyping data. In particular, the 'genome-based restricted maximum likelihood' (GREML) tool in the 'genome-wide complex trait analysis' (GCTA) kit is a statistical model that can estimate the narrow-sense heritability of a trait given genotypes at a set of single nucleotide polymorphism (SNP) loci (Yang et al., 2010, 2011). Using an early version of this method, Yang et al. (2010) found that around 300 thousand common SNP variants—with a minor allele frequency (MAF) of at least 1%—could explain between 40% to 50% of height variation among around 4 thousand unrelated individuals. The authors reasoned that the deficit relative to the $\mathfrak{h}^2 \approx 0.8$ estimated by family-based studies could be explained by incomplete linkage disequilibrium between the SNPs considered in the study and true causal SNPs associated to height.

The following years saw several attempts to bridge this gap, with ever increasing sample sizes, culminating in a genome-wide association study of height across over 5 million donors of diverse ancestries (Lango Allen et al., 2010; Wood et al., 2014; Yengo et al., 2018, 2022). With as many individuals, Yengo et al. (2022) managed to build a saturated map of common genetic variants (MAF $\geqslant 1\%$) associated with human height. The authors report that only 12,111 SNPs—grouped in around 7 thousand distinct genomic windows covering 21% of the human genome—are enough to explain almost all common SNP-based heritability in Europeans.

B The encyclopædia of genomic and epigenomic elements

Ontogeny describes the process through which multicellular organisms develop from single cells. Multicellular organisms are complex systems of organs knit from different tissues composed of coordinated subsets of cells specialized to accomplish particular functions. The genome in each original single cell contains all the information needed to build all the cell types and tissues in the fully developed organism. Thus, the lack of correlation between eukaryotic genome sizes and the number of tissues composing each organism was deemed paradoxical by some (Thomas, 1971).

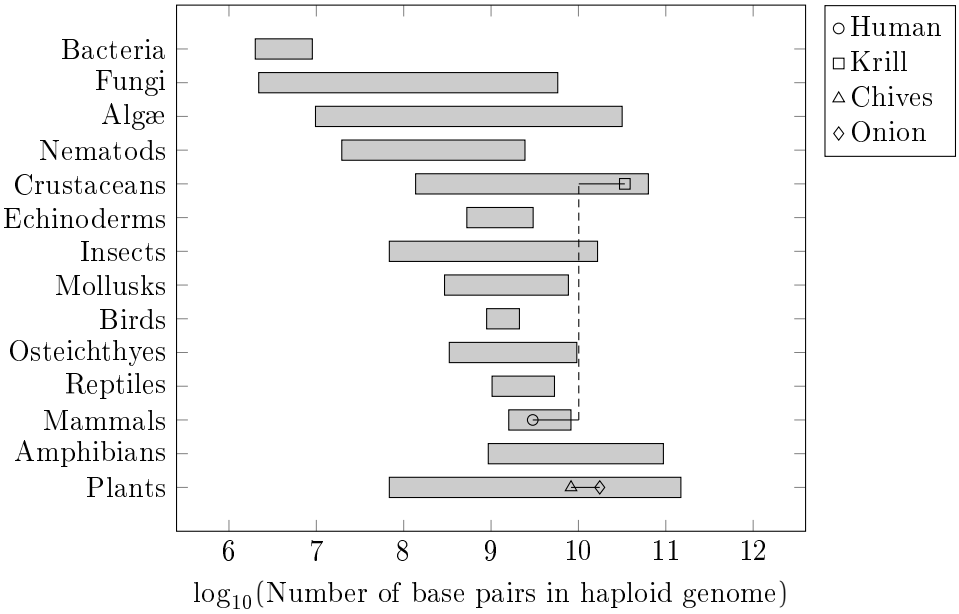


Figure B.1 | Organism complexity and genome size. Range of recorded genome sizes, expressed in number of base pairs, across different clades of the phylogenetic tree of life. Some species are highlighted to illustrate that organism complexity—at least as measured by the number of different cell types and tissues—does not depend directly on genome size. Animal genome sizes recovered from Gregory (2002); Plant, fungal and bacterial genomes recovered from Leitch et al. (2019).

Figure B.1 shows the range of recorded genome sizes—expressed in numbers of base pairs in the DNA sequence—for various clades across the phylogenetic tree. Although there is no concrete measure of multidimensional biological complexity, it could be argued that the human body is much more complex than that of krill, based on the number of different cell types and tissues that compose each organism. Yet, the genome of krill—a centimeter-long crustacean—is around ten-fold larger than the human genome. Perhaps more convincingly, there is a two-fold difference in the length of the haploid genomes of the common onion and of chives, two arguably very similar species.

The 'C-value' paradox coined by Thomas (1971) can be explained by regulatory DNA sequences. It is now known that only around 1% to 2% of the human genome is composed of protein-coding genes (ENCODE Project Consortium, 2020). Most of the functional sequences in the human genome have a regulatory role, controlling the rate at which different proteins are expressed in different contexts, thus allowing cells bearing the same genome to adapt to changing environments and differentiate into distinct cell types.

In the wake of the Human Genome Project (§ 1.1, page 3), the 'Encyclopedia of DNA elements' (ENCODE) Consortium set out to annotate the newly minted human genome sequence, so as to more fully exploit the information contained within, including in regulatory DNA (International Human Genome Sequencing Consortium, 2004; Feingold et al., 2004).

After a pilot study focusing on approximately 1% of the genome, the second phase of the ENCODE Project leveraged next-generation sequencing (Box 1, page 4) methods to inventorize functional DNA elements genome-wide (ENCODE Project Consortium, 2007, 2012). ENCODE 2 uncovered novel functional sequences, such as DNA methylation motifs, in previously uncharted stretches of non-coding DNA. In total, the authors assigned a regulatory role to around 80% of the human genome. In 2020, the ENCODE Consortium published an expanded version of the encyclopædia compiling results from thousands of different assays of DNA and RNA features across hundreds of independent human tissue and cell samples (ENCODE Project Consortium, 2020).

The ENCODE Project revealed a very hetegeneous set of *cis*-regulatory elements (CREs) in human regulatory DNA, including gene promoters, 'enhancers' of transcription, transcription factor (TF) binding sites (TFBSs) and DNA methylation sites. The unifying feature that binds these regulators together is that they affect the probability that the transcriptional machinery binds DNA through changes in chromatin conformation. Hence, candidate (c) CREs are commonly detected using assays of chromatin accessibility. In particular, sensitivity to DNase I—an enzyme that 'cuts' naked DNA—is commonly used to find cCREs (ENCODE Project Consortium, 2020).

DNase I hypersensitive sites (DHSs) are then annotated based on data from other assays of DNA and RNA features to classify cCREs (ENCODE Project Consortium, 2020). For instance, chromatin immunoprecipitation followed by sequencing (ChIP-seq) is used to find TFBS motifs, as well as motifs bound by the CCCTC-binding factor (CTCF), a transcriptional repressor that creates bundles of closed chromatin. ChIP-seq is also used to characterize histone marks associated to chromatin accessibility, such as trimethylation of histone 3 on lysine 4 (H3K4me3) or acetylation of histone 3 on lysine 27 (H3K27ac). Assays for transposase-accessible chromatin followed by sequencing (ATAC-seq) also enable the definition of 'peaks' of open chromatin. Other assays are used to map RNA-binding protein motifs and transcription start sites (TSSs), as well as characterize levels of gene expression and DNA methylation.

The 'ground-level' annotations defined by all these different assays are then integrated to classify cCREs and define chromatin states (ENCODE Project Consortium, 2020). ENCODE distinguishes eight different types of active cCREs depending on whether they are bound by CTCF or ornamented with particular histone marks. For example, 'cCREs with promoter-like signatures' (cCRE-PLSs) are DHSs found within 200 base pairs of a TSS and displaying a high signal of H3K4me3. In contrast, 'cCREs with enhancer-like signatures' (cCRE-ELSs) have low H3K4me3 scores when found within 200 base pairs of a TSS, but they are characterized by a marked H3K27ac signature. cCRE-ELSs found within 2 kilobases of a TSS are tagged as 'proximal' (cCRE-pELSs); the rest are called 'distal' (cCRE-dELSs).

Not all regulatory elements of transcription are written in the DNA sequence. Histone marks like H3K4me3 and H3K27ac are examples of 'epigenetic' factors able to carry information across cell generations somewhat independently of DNA sequence. The transfer of epigenetic information during mitosis is key for cells to differentiate into distinct subsets with minutely specialized functions, even though they carry the same genome. The 'epigenome' of a cell is composed of all the structural and functional addenda—such as histone marks, but also chromatin regulators and noncoding RNAs—that supplement genetic information. Characterizing the human epigenome across different tissues is thus essential to fully understand the impacts of genotype on phenotype in disease-relevant contexts.

In 2008, the National Institutes of Health (NIH) started the Roadmap Epigenomics Program with the aim to map epigenome references across human tissues and cell lines (Bernstein et al., 2010). Seven years later, the Roadmap Epigenomics published over 100 novel epigenomes characterized across five core histone marks associated with promoters, enhancers, transcribed regions and heterochromatin regions, as well as levels of chromatin accessibility, DNA methylation and gene expression (Kundaje et al., 2015). Today, the data produced by Roadmap Epigenomics is hosted by the ENCODE Consortium, along with their own genomic and epigenomic references.

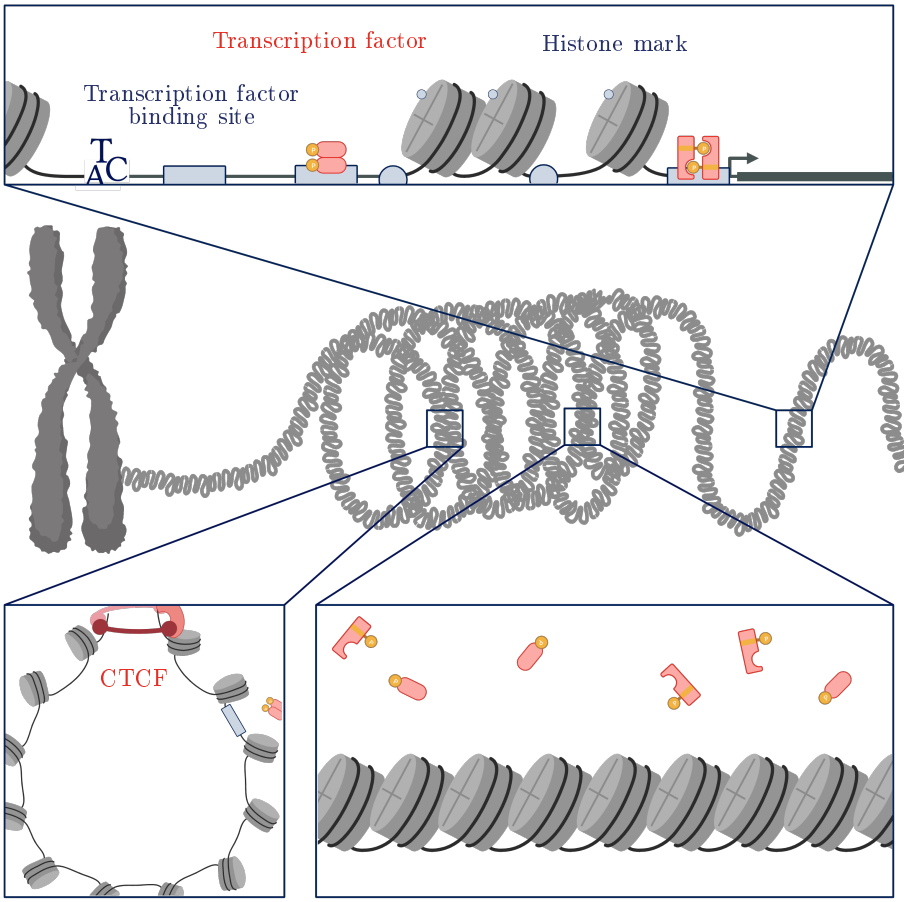


Figure B.2 | The regulatory grammar of gene expression. Mechanisms of transcriptional gene expression regulation play on the probability that the transcriptional machinery binds DNA near coding sequences. In general, chromatin accessibility is regulated by *trans*-regulatory factors (TRFs) that recognize *cis*-regulatory elements (CREs) on the chromatin or directly in the DNA sequence. For example, CTCF is a TRF that affects transcription by creating loops of chromatin through recognition of CCCTC-motif CREs on DNA. Genetic variation around CREs or *trans*-regulatory elements (TREs) that encode TRFs can give rise to expression quantitative trait loci.

The genomic and epigenomic elements strewn along DNA define the regulatory grammar of gene expression regulation. Through interactions with *trans*-regulatory factors (TRFs), these CREs control chromatin accessibility to adapt transcription to changes in the cellular environment. Some CRE-TRF interactions are schematically illustrated in Figure B.2. For example, CTCF is a protein TRF that affects transcription by looping chromatin—in coordinated action with other protein partners—at CCCTC-motif CREs that it recognizes in the DNA sequence (Nanni et al., 2020).

Figure B.2 also illustrates how genetic variation at or around CREs and TREs can give rise to expression quantitative trait loci (eQTLs; § 1.2.4, page 15). In fact, Degner et al. (2012) report that around 16% of QTLs associated to chromatin accessibility at DHSs in human lymphoblastoid cells are also classified as eQTLs in the same model. Conversely, up to 55% of eQTLs mapped in these cell lines are also DHS QTLs. Furthermore, Maurano et al. (2012) estimate that around 77% of GWAS SNPs in non-coding regions fall in or are correlated with DHS loci. Together, these results emphasize the impact of gene expression regulation on complex traits.

At the scale of the tissue, gene expression regulation through CRE-TRF interactions appears as a deterministic process. The binding of a TRF on a CRE alters chromatin accessibility in such a way that the transcription rates of the regulated genes are changed. Yet, at the cellular scale, gene expression is an intrinsically stochastic process, owing to the small numbers of molecules involved in CRE-TRF interactions. In a diploid cell, each particular CRE locus is present in at most two copies; whether a TRF diffuses close enough to recognize and bind a CRE depends on random Brownian motion (Novick and Weiner, 1957; Ko et al., 1990; Raj and van Oudenaarden, 2008).

When the number of available molecules becomes limiting for chemical reaction, the infrequency of interactions introduces random variation into the process. This stochasticity is masked in bulk studies of gene expression, thanks to the statistics of large numbers that average out the noise. However, at the single-cell level, protein expression can vary randomly even among genetically identical cells in experimentally controlled environments (Ko et al., 1990).

Eukaryotic gene expression noise can be conceptually divided in three components (Eldar and Elowitz, 2010). The first component of noise involves co-transcriptional mechanisms that affect messenger (m) RNA production. Importantly, not all genes are equally noisy. While some genes experience a constitutive low level of noise, other genes are under higher noise regimes.

One of the contributing factors to higher noise is transcriptional 'bursting'. The rate at which mRNA is transcribed from these genes is not uniform; transcription rates alternate between 'On' and 'Off' states dictated by CRE activity (Raj et al., 2006). During these bursts, multiple mRNA molecules are transcribed from active genes, and then translated into protein. In contrast, the promoters of low-noise regime genes do not toggle between 'On' and 'Off' states.

The second component of gene expression noise involves molecular mechanisms downstream from transcription. For example, slow nuclear export of mature mRNA may buffer random fluctuations in transcription (Battich et al., 2015). Likewise, protein accumulation in the cytoplasm—protein lifetimes usually last longer than the period between transcriptional bursts—may average out variation due to bursty expression (Eldar and Elowitz, 2010). In contrast, complex pathways of mRNA decay that alternate between translation and degradation-competent states lay introduce more random variation in mRNA concentrations (Hansen et al., 2018).

Finally, expression noise can propagate among genes involved in the same gene regulatory network (GRN; § 3.1.2, page 57). For instance, randomness in the expression of a TF can introduce noisy variation in the expression of its target genes (Eldar and Elowitz, 2010). Relative to the 'intrinsic' noise introduced during transcription, the 'extrinsic' noise that propagates along GRNs can affect multiple genes at once.

The simplest model of intrinstic transcriptional noise represents constitutive gene expression as a Poissonian process (Raj and van Oudenaarden, 2008). The abundance m of transcripts from a given gene is modelled as a birth-death process described by synthesis and degradation rates μ and γ, respectively. The expected number of mRNA copies is the ratio of synthesis to degration, and equal to the Poisson rate parameter λ,

$$\frac{\mu}{\gamma} = \lambda = \mathbb{E}[m] = \mathbb{V}[m]. \tag{B.1}$$

This model assumes constant and independent birth-death rates μ and γ. If the transcript is very abundant, its rate of change with time can be approximated as a deterministic process (Munsky et al., 2012), with a first-order differential equation

$$\frac{dm}{dt} = \mu - \lambda m, \quad m \gg 1. \tag{B.2}$$

When m is not particularly large, Equation (B.2) must be rewritten in a probabilistic framework.

From Equation (B.2), the probability of producing one copy during a short time interval dt is equal to $\mu\, dt$, and the probability of degrading one transcript is equal to $\gamma\, m\, dt$.

For the reformulation to be valid, there must be an equilibrium state in which the probability of seeing m copies of a transcript and producing another one is equal to the probability of seeing $m+1$ copies and having one degrade. Hence,

$$P(m)\,\mu = P(m+1)\,\gamma\,(m+1) \tag{B.3}$$

for any m, which is only possible if m follows a Poisson distribution (Munsky et al., 2012).

While the Poisson approximation describes the expression of less noisy genes well, it is not a good fit for expression data from genes that experience bursty transcription (Zenklusen et al., 2008; Weidemann et al., 2023). To incorporate more complex gene expression regulation schemes, more complex probabilistic models are needed. In particular, the two-state model of bursty transcription introduces two other parameters, k_{On} and k_{Off}, that represent the rates at which promoter activity transitions from an inactive to an active state and vice versa (Munsky et al., 2012).

Conceptually, promoter activity is used as a proxy of chromatin accessibility (Fig. B.2). The transition rate parameters k_{On} and k_{Off} are used to model the probability of the promoter being in the 'On' state as a Beta-distributed variable,

$$\begin{aligned} P(\mathrm{On}) &\sim \mathcal{B}\left(\widehat{k}_{On}, \widehat{k}_{Off}\right), \\ \widehat{k}_{On} &= \frac{k_{On}}{\gamma}, \\ \widehat{k}_{Off} &= \frac{k_{Off}}{\gamma}. \end{aligned} \tag{B.4}$$

When genes are transcribed in short but intense bursts, $\widehat{k}_{Off}$ is much larger than $\widehat{k}_{On}$ and much greater than one, and the distribution of $P(\mathrm{On})$ converges to a Gamma

$$P(\mathrm{On}) \sim \mathcal{G}\left(\widehat{k}_{On}, \widehat{k}_{Off}\right), \quad \widehat{k}_{Off} \gg \widehat{k}_{On} \gg 1. \tag{B.5}$$

Thus, under a bursty transcriptional regime, the otherwise constant expected number of mRNA copies in a cell becomes a random variable

$$m \sim \mathcal{NB}\left(\lambda; \widehat{k}_{On}, \widehat{k}_{Off}\right) \tag{B.6}$$

distributed according to a Gamma-Poisson mixture (Raj et al., 2006). Another name for this mixture is the 'negative binomial' distribution.

The probability $P(\mathrm{On})$ integrates information from upstream regulators of transcription, such as DNA methylation, histone marks and pioneer TFs. By modifying chromatin accessibility, these factors can adapt the size and frequency of transcriptional bursts in response to changes in the cellular environment, like hormones, cytokines or pathogens (Raj and van Oudenaarden, 2008). As mentioned previously, this variation can propagate further along the GRN and introduce extrinsic noise into the expression of downstream genes, possibly in other cells (Eldar and Elowitz, 2010).

For example, Figure B.3 illustrates a simulated case in which transcripts from gene g are synthesized ten times faster than they are degraded. Although $\frac{\mu}{\gamma} = 10$ mRNA molecules are expected in each cell, the real observed number of transcript copies also depends on the chromatin context in which gene g lies. This information is integrated into the model through the Gamma-distributed probability that its promoter toggles into the 'On' state. In contexts where chromatin around gene g is more likely to be open, $P(\mathrm{On})$ is higher and so is the expected number of transcripts.

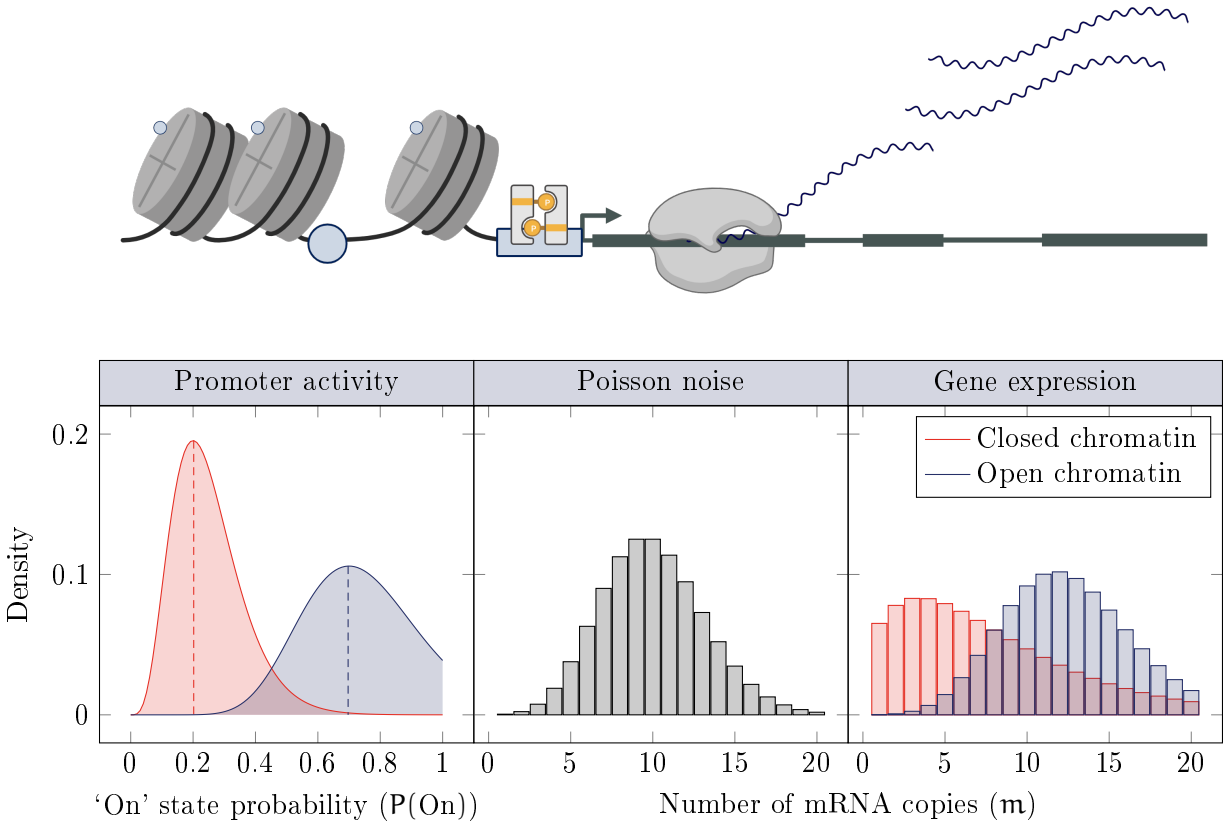


Figure B.3 | **Two-state model of gene expression regulation.** Gene expression is a stochastic process. For some genes, the uncertainty around the expected number m of transcripts, given constant and uncorrelated mRNA synthesis and degradation rates, is well described as a Poisson process. For other genes with more complex regulation schemes, the two-state model of expression regulation introduces a variable that captures the probability P(On) that the gene promoter is active. Thus, observed transcript counts are distributed as negative binomial random variables. In this example, two instances of P(On) were simulated as Gamma processes to model promoter activity in an open chromatin context and a closed chromatin context. Basal gene expression noise was modeled with a Poisson rate parameter $\lambda = 10$. Simulated gene expression values are distributed as a Gamma-Poisson mixture—otherwise known as the 'negative binomial' distribution—with P(On) equal to the maximum density estimate in each chromatin context, given by the dashed lines.

Importantly, noise can be a feature of gene expression rather than a bug (Eldar and Elowitz, 2010). Indeed, gene expression noise can be used to coordinate transcriptional responses in large GRNs (Cai et al., 2008; van Dam et al., 2018) and can provide a substrate for adaptation by enlarging the range of phenotypes that result from a given genotype (Thattai and van Oudenaarden, 2004).

C The local adaptation to environmental pressures

Evolution manifests itself through changes in allele frequencies through time. In this context, the Wright-Fisher model describes allele frequency changes across non-overlapping generations, under the simplest mode of genetic inheritance in a panmictic population of N individuals evolving under neutrality (Kimura, 1968) and in the absence of mutation (Fisher, 1930; Wright, 1931). If locus g has two alleles A and a, and X_t is the number of A alleles in the population at generation t, the model is a discrete-time Markov chain that describes the evolution of X_t across generations.

At each generation t, the pool of alleles for generation $t+1$ is randomly sampled with replacement as a binomial process. The transition matrix of the Markov chain contains the probabilities $(P_t)_{ij}$ that X_t changes from i copies to j copies after a generation. The items $(P_t)_{ij}$ are given by the probability mass function of the binomial distribution,

$$(P_t)_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}. \tag{C.1}$$

That is, the number X_{t+1} of A copies at the next generation, knowing that there are $X_t = i$ copies in the present generation, follows a binomial distribution with size $2N$ and success probability equal to the frequency $f_A(t)$ of A at generation t,

$$(X_{t+1}|X_t = i) \sim \mathcal{B}\left(2N, f_A(t)\right), \quad f_A(t) = \frac{i}{2N}. \tag{C.2}$$

From the law of total expectation and the properties of the binomial distribution,

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X|Y]\right], \qquad \mathbb{E}[X \sim \mathcal{B}(n,p)] = np, \tag{C.3}$$

the expected state of the Markov chain X_t at any point in time is equal to the initial number of A copies in the first generation,

$$\mathbb{E}[X_{t+1}] = \mathbb{E}\left[\mathbb{E}[X_{t+1}|X_t]\right] = \mathbb{E}\left[2N\frac{i}{2N}\right] = \mathbb{E}[i] = \mathbb{E}[X_t] = \cdots = 2N\, f_A(0). \tag{C.4}$$

Then, if the difference in A allele frequency from generation t to generation $t+1$ is

$$\Delta f_A = f_A(t+1) - f_A(t) = \frac{j}{2N} - \frac{i}{2N} = \frac{X_{t+1}}{2N} - \frac{X_t}{2N}, \tag{C.5}$$

from the properties of the binomial distribution,

$$\mathbb{E}[\Delta f_A] = 0 \tag{C.6}$$

and

$$\mathbb{V}[\Delta f_A] = \frac{\frac{X_t}{2N}\left(1 - \frac{X_t}{2N}\right)}{N} \tag{C.7}$$

While Equation (C.6) reflects neutrality, Equation (C.7) shows that significant allele frequency variation will only happen over time scales of N generations; yet, if the population is small, allele frequencies can drastically change through random 'drift', even though they are not expected to under the model. The probability that an allele is fixed or lost from the population also depends on its initial frequency. Figure C.1 illustrates these key features of the neutral Wright-Fisher model.

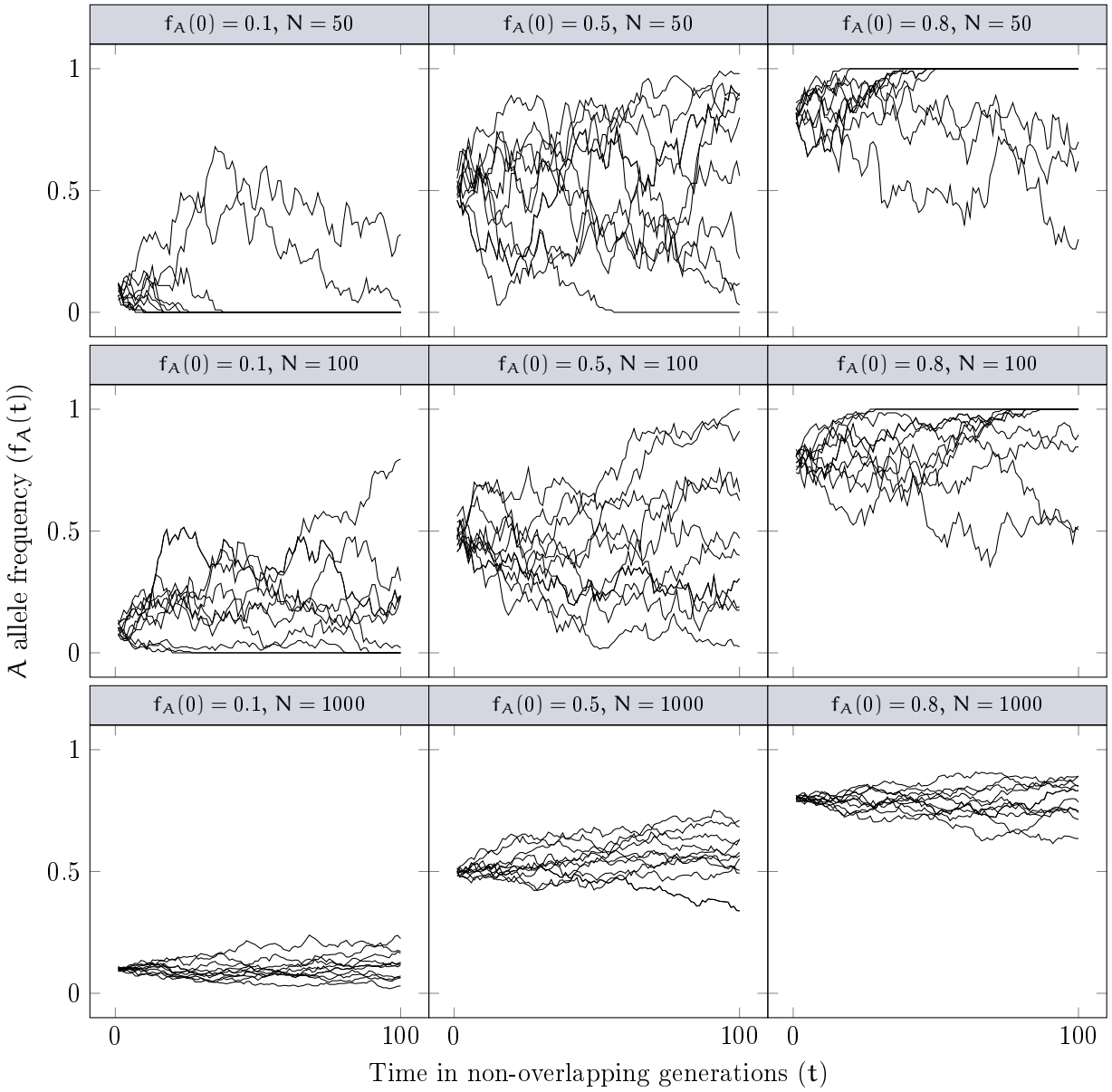


Figure C.1 | The neutral Wright-Fisher model. The neutral Wright-Fisher model describes the evolution of allele frequencies $\frac{X_t}{2N}$ across non-overlapping generations in a population of size N as a random binomial process. Two key features of the model are that allele frequencies are not expected to change from one generation to the next, but significant allele frequency changes can happen randomly, even over a relatively short periods, if N is small. In each panel, the lines show the frequencies of ten independent alleles evolving under the neutral Wright-Fisher model for 100 generations in a population of size N and from a initial frequency $\frac{X_0}{2N}$.

The biological assumptions of the neutral Wright-Fisher model are in fact not very realistic. For instance, the probability of mating is not uniformly distributed across all possible pairs of individuals in most real diploid populations. Also, the sampling of alleles at each generation is not entirely random, due to the linkage disequilibrium (LD) in human genomes. Furthermore, evolutionary forces like mutation, demographic events and natural selection can make allele frequencies deviate from what is expected under this model.

Natural selection arises when inter-individual differences in fitness—phenotype differences that alter survival probability or reproductive success—are at least partially driven by heritable genetic factors (Balding et al., 2019). If locus g is such a factor, the trajectory of allele A (Fig. C.1) will depend on the relative fitnesses of the genotypes AA, Aa and Aa. If the fitnesses associated to alleles A and a are η_A and η_a, respectively, the relative genotype fitnesses can be expressed as $\eta_A^2 : \eta_A\eta_a : \eta_a^2$, or as their associated selection coefficients $s_{aa} = 1-\eta_a^2 = 0$, $s_{Aa} = 1-\eta_A\eta_a$ and $s_{AA} = 1-\eta_A^2$, if $\eta_a^2 = 1$. The frequency trajectory of the selected allele A can then be traced using

$$\mathbb{E}\left[f_A(t+1)|f_A = f_A(t)\right] = f_A\frac{1+s_{Aa}(1-f_A)+s_{AA}f_A}{1+2s_{Aa}f_A(1-f_A)+s_{AA}f_A^2} \neq f_A(0), \tag{C.8}$$

and is not expected to be equal to the starting allele frequency, as in the neutral Wright-Fisher model in Equation (C.4). Natural selection can thus drive the stable adaption of populations to changes in their local environment.

Hence, uncovering the genomic signatures of selection is key to understanding differences in phenotype across populations. Although emerging approaches and growing data sets allow to draw observed allele frequency trajectories using ancient DNA (Kerner et al., 2023a), most commonly used methods to detect natural selection use present-day human genetic diversity to infer changes in allele frequency. Some of these methods use summary statistics to pick up on alterations of the site frequency spectrum (SFS) (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000), characteristic haplotype structures (Sabeti et al., 2002) or patterns of population differentiation (Lewontin and Krakauer, 1973; Yi et al., 2010), while others are based on probability models and likelihood functions (Kim and Stephan, 2002; Nielsen et al., 2005; Stern et al., 2019).

Yet, the amount of stochasticity surrounding allele frequency trajectories in the human genome makes computation of full-likelihood models intractable. In particular, because the human genome is recombinant, several chromosomal lineages are possible given a unique sample of allele frequencies at any variant locus. That is, each recombination event uncouples the ancestry of a DNA segment from the ancestry of the individual that carries it, spreading it over two different chromosomal lineages. The ancestry of a recombinant DNA segment cannot be described with a tree topology anymore, it must be described as an ancestral recombination graph (ARG) that integrates all the possible different trees resulting from recombination (Griffiths and Marjoram, 1996, 1997). For example, Figure C.2 illustrates the joint ARG inferred from a sample of six chromosomal lineages with four recorded recombination events. Depending on where cross-overs happen, different genealogies arise.

Using full-likelihood models to detect natural selection is implies considering all possible marginal trees embedded in an ARG conditional on a given allele frequency trajectory, which makes analytical solutions of these models computationally intractable (Coop and Griffiths, 2004).

To skirt this limitation, Kim and Stephan (2002) developed a method that leverages the ARG framework to test the statistical significance of skews in the SFS and local reductions of variations predicted by the 'genetic hitchhiking' theory of selective sweeps. Briefly, the authors propose a likelihood ratio test to compare models without selection ($\mathcal{H}_0 : s = 0$) to models with selective sweeps of different magnitude at various genomic loci. The overall likelihood of each model is calculated as a composite of individual likelihood functions computed at each position in the tested sequence. Since then, several other authors have proposed extended versions of composite likelihood methods (Nielsen et al., 2005; Zhu and Bustamante, 2005; Vy and Kim, 2015).

More recently, Speidel et al. (2019) developed a scalable method to reconstruct ARGs around single nucleotide polymorphisms (SNPs) genome-wide, which enables more powerful analyses of natural selection by providing a computationally tractable algorithm to infer population sizes, split times and mutation rates using thousands of samples.

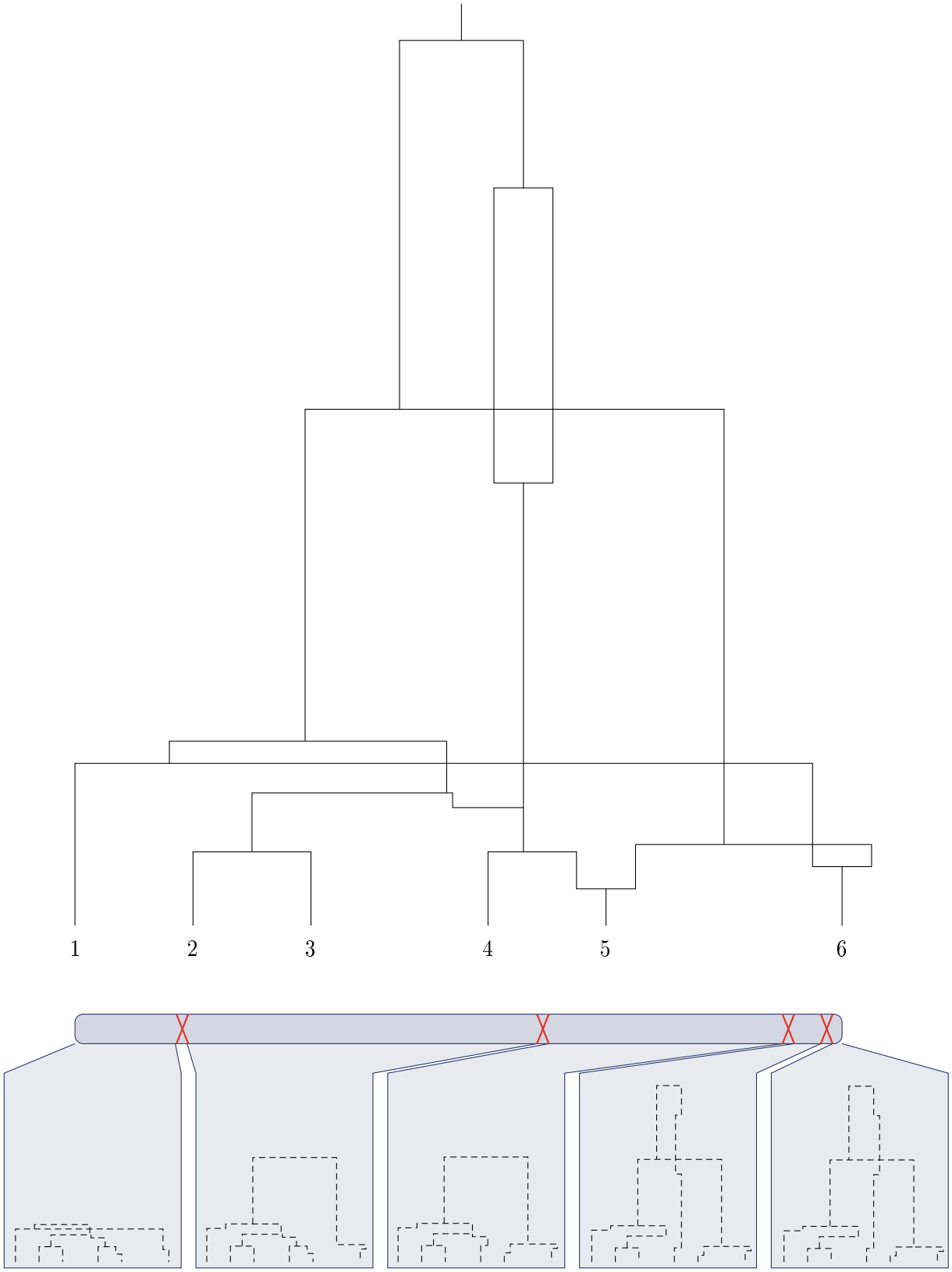


Figure C.2 | **The ancestral recombination graph.** An ancestral recombination graph (ARG) that integrates all possible ancestry topologies in a DNA segment with four random recombination events. The blue horizontal rectangle represents a stretch of chromosome. Four crossing-overs, resulting from recombination events and shown as red crosses, were simulated randomly along the length of the chromosomal segment. The five marginal trees embedded in the ARG are shown below; each one is associated to its corresponding segment of recombinant DNA. Adapted from Balding et al. (2019), Chapter 5.

Briefly, the method takes phased haplotype information across SNPs and individuals—as well as a map that specifies recombination rates along the human genome, and the sequence of a reference human ancestor genome to define ancestral and derived allele states—and outputs the series of marginal trees embedded in the ARG inferred at each SNP position, given a fixed mutation rate and effective population size (Speidel et al., 2019).

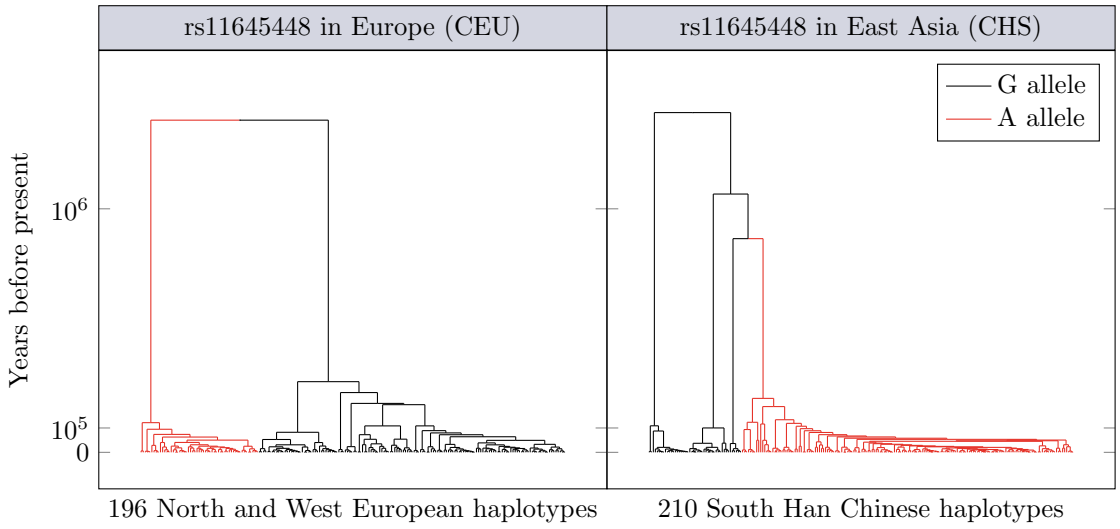


Figure C.3 | Marginal trees in the ancestral recombination graph. Marginal trees around the rs11645448 SNP embedded in the ancestral recombination graph (ARG) inferred using Relate (Speidel et al., 2019) on genotype data from 98 individuals in the 'Utah residents with Northern and Western European ancestry' (CEU) panel of the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2022) and 105 individuals in the 'Han Chinese South' (CHS) panel, with a fixed mutation rate $\mu = 1.25 \times 10^{-8}$ and effective population size $N_e = 3 \times 10^4$. Each leaf in each tree represents a haplotype. The branches carrying the derived allele at this variant are shown in red.

For example, Figure C.3 shows marginal trees around the rs11645448 SNP embedded in ARGs inferred using phased genotyping data from a European and an East Asian panel from the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2022). Interestingly, rs11645448 has been pointed out as an expression quantitative trait locus (eQTL; § 1.2.4, page 15) of *NOD2* in $CD16^+$ monocytes exposed to 'severe acute respiratory syndrome' coronavirus (SARS-CoV) 2 (Aquino et al., 2023). In Figure C.3, the red branches in each tree indicate the lineages carrying the derived A allele at SNP rs11645448. Remarkably, this allele seems to have rapidly expanded, but only in the genomes of East Asians, suggesting an episode of local adaptation.

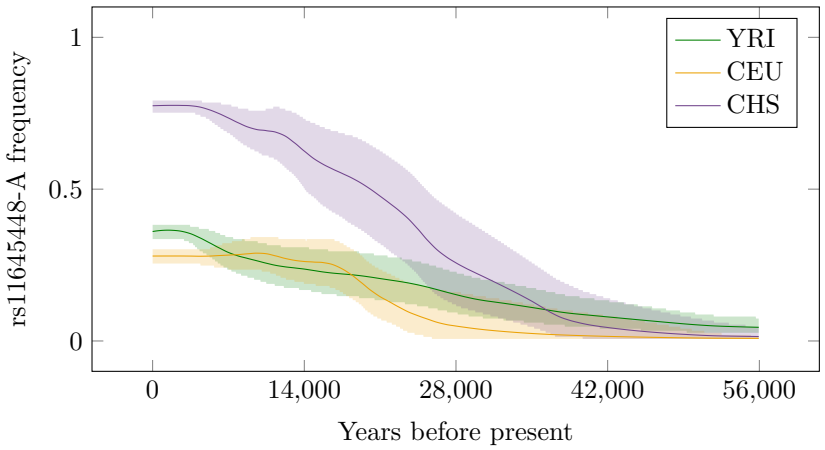


Figure C.4 | Allele frequency trajectories inferred from ancestral recombination graphs. Frequency trajectories of the derived allele at the rs11645448 expression quantitative trait locus of *NOD2* in Africa and Eurasia across the last 56,000 years, or 2,000 generations with a 28-year generation time. The 'Yoruba in Ibadan' (YRI), 'Utah residents with Northern and Western European ancestry' (CEU) and 'Han Chinese South' (CHS) reference panels from the 1000 Genomes Project Consortium (Byrska-Bishop et al., 2022) were used as proxies of genetic diversity in Africa, Europe and East Asia, respectively. Adapted from Aquino et al. (2023).

Importantly, because the method proposed by Speidel et al. (2019) integrates information on effective population size variation through inferred coalescence rates, the trees in Figure C.3 are calibrated against the effects of evolutionary demographic processes—such as so-called population 'bottlenecks' and expansions—that can confound signals of natural selection in the SFS.

These genealogies can then be used as input for other evolutionary genetics analyses. For instance, Stern et al. (2019) propose a method based on ARGs inferred from modern genetic data to approximate the full likelihood that selection acts on a given variant locus, and simultaneously infer the frequency trajectories of its alleles. Given marginal trees sampled from the ARG around the site of interest, the algorithm then runs a hidden Markov model on each tree to infer hidden allele frequency states based on coalescence rates observed at discrete points in time. The transition probabilities of allele frequencies depend on the selection coefficient s (Eq. C.8), which is ultimately the parameter of interest. By marginalizing out the latent allele frequency trajectory, the probability of each local tree—represented by the vector C containing the number of ancestral and derived lineages at each epoch—conditional on the selection coefficient can be recovered as

$$P(C|s) = \sum_{x \in \chi} P(C|f_A = x)P(f_A = x|s) \tag{C.9}$$

where χ is the space of all possible trajectories. The full likelihood function of s at this locus is then approximated through importance sampling over all local trees (Stern et al., 2019).

Figure C.4 plots trajectory frequencies for the putatively adaptive rs11645448-A allele (Fig. C.3) across the last 56 thousand years in Africa, Europe and East Asia, inferred using the method developed by Stern et al. (2019) and based on local genealogies modeled using the method by Speidel et al. (2019). In line with what is shown in Figure C.3, the rs11645448-A allele appears to have rapidly increased in frequency over the last 40 to 15 thousand years in the genomes of East Asians, but not in European genomes, nor in Africa.

D The evolutionary forces behind heritability

The genome-wide association study (GWAS) framework is a powerful tool to map the genetic bases of complex traits (§ 1.2, page 7). Its foundations are deeply rooted in evolutionary genetics theory (Reich and Lander, 2001; Sella and Barton, 2019). For instance, the validity of the GWAS approach in disease contexts relies on the truth of the 'common disease/common variant' (CD/CV) hypothesis that the genetic basis of most common diseases is composed of several low-effect variants carried by 1 to 5% of individuals in the population (Fig. 1.2, page 8). From an evolutionary genetics perspective, Reich and Lander (2001) provide support for the CD/CV hypothesis—as well as a plausible explanation for the pervasiveness of frequent disease-associated alleles—through a relatively simple demographic model incorporating a rapid expansion from founder populations around 15 to 18 thousand years ago.

In a sense, GWAS results express how the heritability of a trait is distributed along the genome (Sella and Barton, 2019) (Appendix A, page 190). Assuming additive genetic effects on trait Y,

$$\mathbf{Y} = \alpha_Y + \beta_{Yi}\mathbf{G}_i + \boldsymbol{\varepsilon}, \tag{D.1}$$

the contribution of any given locus i to the total heritability of Y depends (i) on the effect β_{Yi} of the genotype $\mathbf{G}_i$ at locus i on the trait and (ii) on the frequency f_i of the effect allele,

$$\begin{aligned}\mathbb{V}[\mathbf{Y}] &= 0 + \beta_{Yi}^2\mathbb{V}[\mathbf{G}_i] + \mathbb{V}[\boldsymbol{\varepsilon}] \\ &= 2\beta_{Yi}^2 \cdot f_i(1-f_i) + \sigma^2,\end{aligned} \tag{D.2}$$

where $\mathbb{V}[\mathbf{G}_i] = f_i(1-f_i)$ is given by the binomial sampling of alleles in the population from the previous generation (Appendix C, page 199) and $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2$ is given by the properties of the classical linear model in Equation (D.1).

If Y is a polygenic trait, its narrow-sense heritability is given by the ratio between the sum of additive genetic components across all L independent variants that contribute to variability in Y, and the total variance of Y,

$$h^2 = \sum_{i=1}^{L} \frac{2\beta_{Yi}^2 \cdot f_i(1-f_i)}{\mathbb{V}[\mathbf{Y}]} = \sum_{i=1}^{L} R_i^2, \tag{D.3}$$

where R_i^2 is the coefficient of determination of locus i. Under the additive infinitesimal model (Fisher, 1918; Barton et al., 2017), L is assumed to be very large and each β_{Yi} accordingly weak.

Importantly, the relationship between function and heritability reflected in GWAS results must be interpreted in light of the evolutionary forces that shaped allele frequencies f_i into their current states, and thus affect the contribution of each locus to heritability (Sella and Barton, 2019).

E The transformer architecture and genomic data

In the era of big data, there is a great interest in developing artificial intelligence (AI) methods that can be trained to detect intricate patterns in complex data sets—including genomic (Dias and Torkamani, 2019) and clinical ones (Topol, 2023)—so as to derive insights and predict outcomes. In this context, 'deep learning' (DL) methods differ from classical 'machine learning' approaches in that their learning process requires no explicit human intervention (Ng, 2021a,b,c,d,e).

From a very broad perspective, DL algorithms draw increasingly derived features from raw data through stacked 'neural network' (NN) layers. Each node in the NN is a linear combination of inputs from nodes in the previous layer; the goal of the DL algorithm is to optimize these functions so as to minimize the overall prediction error of the model during training. In AI terms—and in analogy to a biological, organic NN—each 'neuron' in an artificial NN layer is a linear combination of the 'activation' values of neurons in the previous layer. Thus, the model learns the optimal set of weights for each neuron that minimize the cost of error given training input and output data, and are likely to yield the most accurate predictions when it is applied to other data sets. The depth of the learning process depends on the number of layers in the NN: a DL model is characterized by a large number of 'hidden' NN layers.

For illustration, Figure E.1 shows a shallow two-layer feedforward NN (FNN) with a single hidden layer. In this densely connected neural network, each neuron in layer l takes all inputs from the previous layer, and outputs an activation that is passed on as input to the neurons in the following layer. For instance, the activation of the second neuron in the hidden layer of the NN in Figure E.1 is computed as

$$a_2^{[1](i)} = g\left(z_2^{[1](i)}\right) = g\left(\boldsymbol{w_2^{[1]}}^{\mathsf{T}} \boldsymbol{a}^{[0](i)} + b_2^{[1]}\right), \tag{E.1}$$

where $\boldsymbol{a}^{[0](i)}$ is the vector of activations of the input layer, $\boldsymbol{w_2^{[1]}}$ is the weight parameter vector and $b_2^{[1]}$ is the bias parameter associated to this neuron, and g is an 'activation function' that transforms the linear combination $z_2^{[1](i)}$ into the activation $a_2^{[1](i)}$.

More generally, in an L-layer FNN trained across m input-output examples, the jl-th neuron of layer l in training example i is computed as

$$a_{jl}^{[1](i)} = g\left(z_{jl}^{[l](i)}\right) = g\left(\boldsymbol{w_{jl}^{[1]}}^{\mathsf{T}} \boldsymbol{a}^{[l-1](i)} + b_{jl}^{[1]}\right), \tag{E.2}$$

where $jl = \{1, \cdots, n^{[l]}\}$, $n^{[l]}$ is the number of neurons in layer l, $l = \{1, \cdots, L\}$ and $i = \{1, \cdots, m\}$. In practice though, neuron activations are not computed serially one-by-one: all of these vectors can be stacked together into matrices, so that the activation of a layer—across $n^{[l]}$ neurons and m training examples—is computed in parallel as

$$\boldsymbol{A}^{[l]} = g\left(\boldsymbol{Z}^{[l]}\right) = g\left(\boldsymbol{W}^{[l]}\boldsymbol{A}^{[l-1]} + \boldsymbol{b}^{[l]}\right). \tag{E.3}$$

where $\boldsymbol{W}^{[l]} \in \mathbb{R}^{n^{[l]} \times n^{[l-1]}}$, $\boldsymbol{A}^{[l-1]} \in \mathbb{R}^{n^{[l-1]} \times m}$ and $\boldsymbol{b}^{[l]} \in \mathbb{R}^{n^{[l]} \times 1}$ (Fig. E.1).

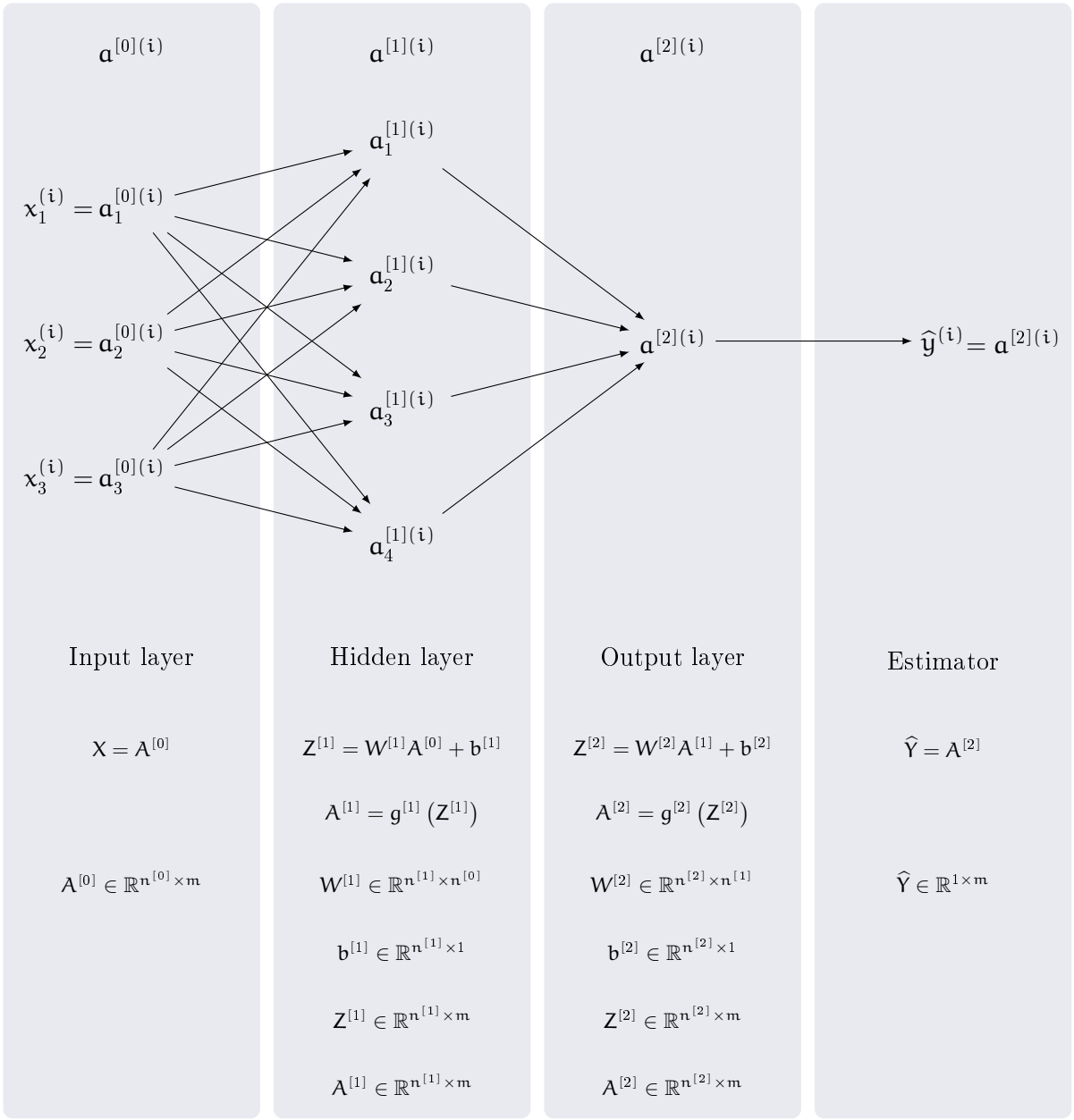


Figure E.1 | A shallow feedforward neural network. A feedforward neural net is a directed acyclic graph that links input data to output data through a series of interconnected layers of neurons. Each neuron in layer l and training example i outputs an activation $a^{[l](i)}$ that is used as input by neurons in layer $l+1$. Thus, the activations $A^{[l+1]}$ result from the linear combination $Z^{[l+1]}$ of activations $A^{[l]}$, transformed by activation function g. The estimation of the model given each set of input data is equal to the activation of the output layer.

The choice of the activation function g depends on the task at hand. For instance, if the FNN in Figure E.1 was used to solve a binary classification problem, a sigmoid activation would be appropriate to estimate the probability that example i belongs to the focal class,

$$a^{[2](i)} = g\left(z^{[2](i)}\right) = \frac{1}{1 + e^{-z^{[2](i)}}} = \frac{1}{1 + e^{-w^{[2]\top} a^{[1](i)} + b^{[2]}}} = \hat{y}^{(i)}. \tag{E.4}$$

Each neuron in the FNN would be equivalent to a logistic regression unit, which takes inputs from the previous layers and outputs a probability. The overall output of the FNN for training example i, $a^{[2](i)} = \hat{y}^{(i)}$, would be the estimated probability that it belongs to the focal class. This can then be compared to the actual label of training example i, so as to compute a 'cost function' of the parameters of the FNN. As previously mentioned, the goal during training is to find the set of weight and bias parameters—across all neurons and training examples—that minimize the cost function, through rounds of 'forward' and 'backward' progagation along the network.

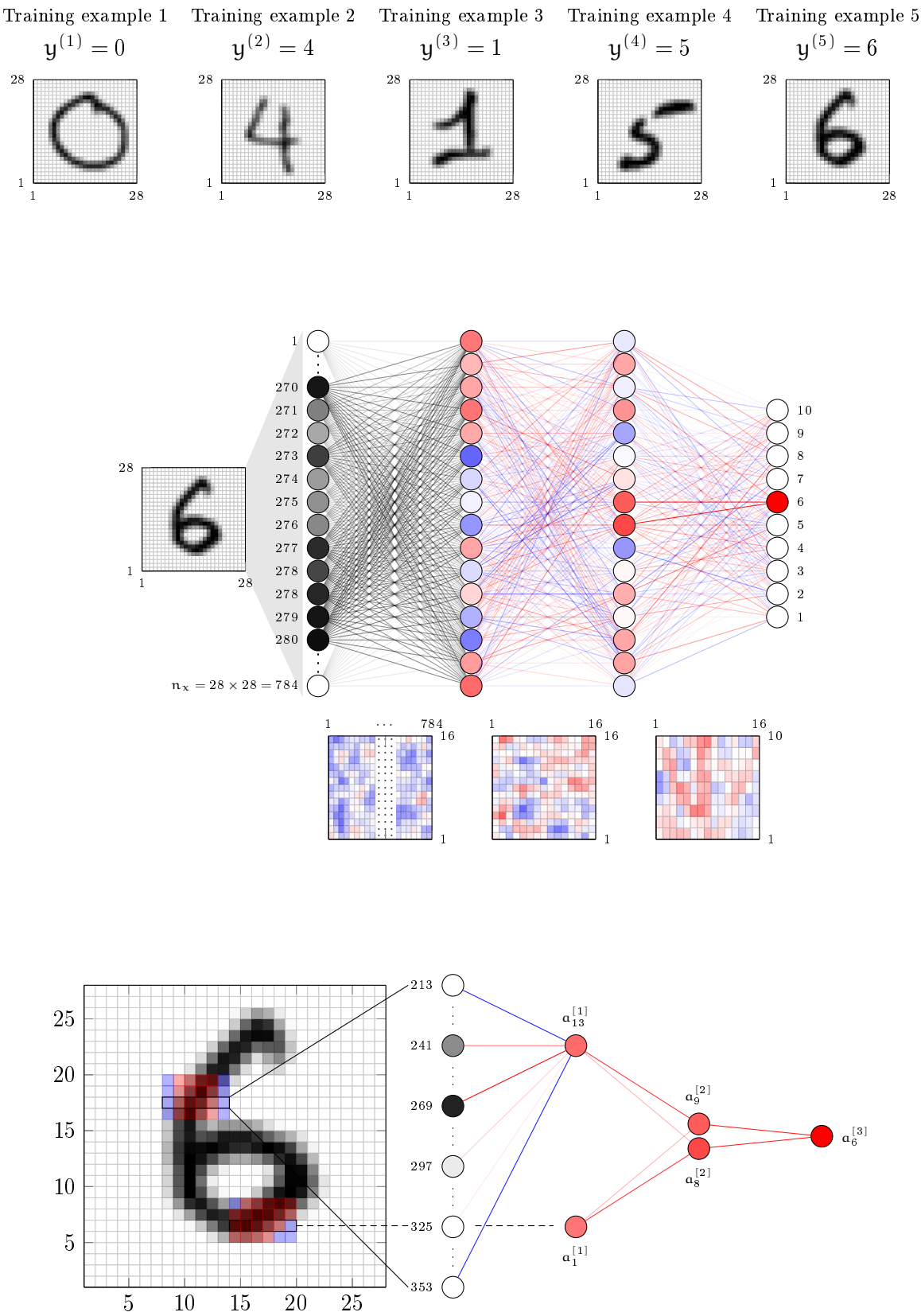


Figure E.2 | **Training a feedforward neural network.** Feedforward neural networks can be trained to recognize hand-written digits. The top panel shows five training examples of labeled digits encoded as 28×28 matrices of grayscale pixel values. The middel panel then focuses on a trained network with two hidden layers. The matrix is unfolded into a vector of length $28 \times 28 = 784$ that is used as input data $a^{[0](5)}$. The weight matrix associated to each pair of successive layer is shown below. The bottom panel focuses on two regions of the training example to illustrate how weights can be tuned during training to detect particular metafeatures in the input data, and integrate them to predict the correct outcome.

More generally, the sigmoid is an example of 'ridge' activation functions that transform a linear combination of inputs into a nonlinear activation. Other examples include the rectified linear unit (ReLU), the Heaviside and the hyperbolic tangent functions. The common property that makes them useful in an NN context is that they allow to introduce nonlinearities into the data as it progresses along the NN, so that each layer is able to extract increasingly complex features as compositions of simpler features detected in previous layers. This is meant to be analogous to how in a biological NN, the activation of some neurons induces the activation of another subset of neurons, integrating information to ultimately mount a coordinated response. In this context, training an NN comes down to downweighting the combinations of neuron activations that lead to incorrect outcomes given labeled input data (Sanderson, 2020).

For illustration, Figure E.2 shows an example of a shallow FNN with two hidden layers used to recognize hand-written digits. Each digit is encoded as a 28×28 matrix of grayscale pixel values. A commonly held view is that while different neurons in the first layer of the network may learn—by assigning the appropriate weights to the corresponding pixels in the input data—to recognize pixel patterns associated to vertical edges and loops, these insights could then be used by the second layer of neurons to learn that the digit '6' is composed of a loop with a vertical edge on top. Although this does not accurately reflect the function of an FNN, the basic intuition is true: the network learns to derive feature combinations from the input data through successive layers of abstraction.

In reality, no particular neuron in the network has any predefined function or role: neurons acquire their 'tasks' during training. The performance of each neuron in accomplishing its task depends on a set of weight and bias parameters. If neurons perform their tasks well, the FNN will tend to have a parameter set that minimizes the cost function during training. For example, a neuron well-trained to detect vertical edges in a particular region of the matrix may acquire positive weights associated to pixels in the region of interest, but negative weights at the bounds of the region, such that its weighted sum will be maximal if there is a vertical edge, but will decrease if there are also pixels associated to a horizontal edge, for example (Fig. E.2). In turn, the bias parameter is useful to control the rate at which neurons fire. For example, it can be set such that neurons that cover regions of high pixel density, but where the distribution of pixels is rather homogeneous and noninformative, do not fire as often.

This simple and specific example illustrates the general mechanism behind more complex DL algorithms. The wheels and cogs of an NN are made from core linear algebra tools coupled to nonlinear activation functions, so that the algorithm is able to capture nonlinear relationships and patterns in the data by estimating weight and bias parameters associated to neural activation. The real 'learning' happens through rounds of forward and backward propagation along the NN, that progressively update parameters from their randomly initialized values to the values that minimize the prediction error of the algorithm. The simple FNN can be ornamented with other network motifs to quicken the learning process or prevent model overfitting, but this general principle remains valid.

Other NN architectures exist apart from the FNN, that are useful for other tasks. In particular, the convolutional (C) NN is especially useful for image processing, while the recurrent (R) NN is mostly used for natural language processing (NLP) applications. Briefly, the CNN is a special type of FNN architecture characterized by the presence of 'convolutional' and 'pooling' hidden layers. As illustrated in Figure E.3, the basic idea is that—through successive convolution and pooling—these layers learn to reduce the dimensionality of complex input data to its most informative features, which can then be presented in a fully-connected layer. Briefly, convolutional layers scan the input data—or the ouput of previous convolutions—step-wise with a convolution filter or 'kernel' that extracts informative features, which are then aggregated in the pooling layers, so as to reduce the number of parameters to estimate.

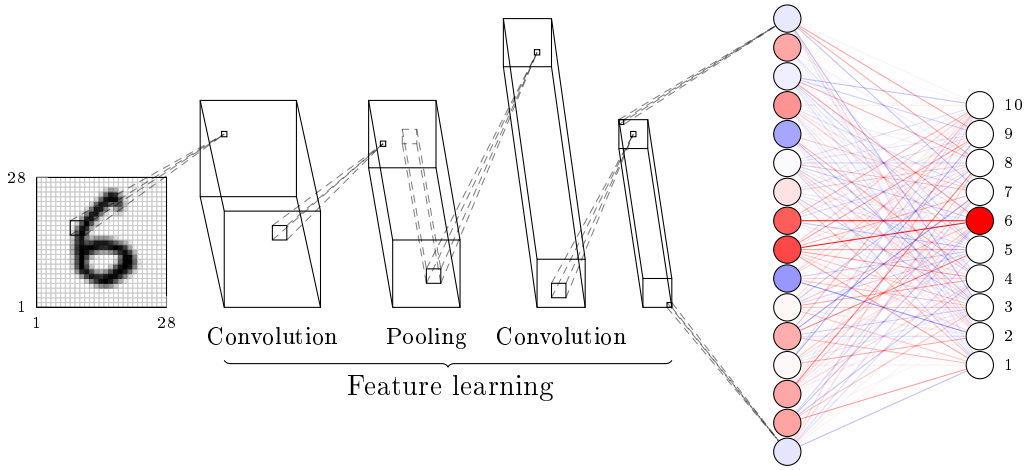


Figure E.3 | A schematic convolutional neural network. Convolutional neural networks are a type of feed-forward neural networks characterized by the presence of 'convolutional' and 'pooling' layers. The basic idea is that, across these layers, the network learns to detect the most informative features of multidimensional data, so as to reduce its dimensionality into a shorter vector, that is used as input for a classical fully-connected layer. During convolution, the input is scanned step-wise by a convolutional filter, or 'kernel', of a given size to produce each ouput of the convolutional layer. During pooling, input values are passed to an aggregating function—like the sum or the average—so as to reduce the dimensionality of the input.

Relative to a non-convolutional FNN (Fig. E.2), this framework allows to handle highly multidimensional data more efficiently, by first learning its most informative features, and then using these as inputs to a fully-connected FNN (Fig. E.3). This is also the basis for 'encoder-decoder' type architectures that analyse the input data through convolutions to produce an lower-dimensional encoding, that is then decoded through transposed convolution to produce an output of same dimension as the input, but driven by its most informative features. For instance, encoder-decoder architectures are commonly used to denoise low-resolution images into higher resolution.

Encoder-decoder network motifs are also commonly used within RNN architectures in sequence-to-sequence NLP tasks, like neural machine translation. As illustrated in Figure E.4, the RNN architecture differs from the FNN and CNN in that it is cyclic, as any neuron can take its own output as input, which makes them especially appropriate to process variable-length sequence data. More precisely, an RNN is built as instances of the same recurrent neural motif progressing through moments of a sequence, rather than as a stack of layers of neurons (Figs. E.2 and E.3).

In its most simple implementation, the basic recurrent unit (BRU) accomplishes two successive operations. First, the activation at moment t is computed as a linear combination of the input $x^{<t>}$ to the BRU and its previous activation $a^{<t-1>}$,

$$a^{<t>} = g_a\left(\boldsymbol{w}_{aa}a^{<t-1>} + \boldsymbol{w}_{ax}x^{<t>} + b_a\right), \tag{E.5}$$

where g_a is an activation function, $\boldsymbol{w}_{aa}$ is the vector of activation weights that determine how each of the activations in the previous instance affects the activation of the current instance, $\boldsymbol{w}_{ax}$ is the vector of input weights that connects the elements of the input to the corresponding activation in the current instance, and b_a is the activation bias parameter. Second, the hidden activation state is used to compute the output of the BRU,

$$\hat{y}^{<t>} = g_y\left(\boldsymbol{w}_{ya}a^{<t>} + b_y\right), \tag{E.6}$$

where g_y is an output function, and $\boldsymbol{w}_{ya}$ and b_y are output weight and bias parameters, respectively. Importantly, because the BRU is always the same, all the input, activation and output parameters are constant: the activation of BRU at the t-th moment depends only on the input sequence and the previous activation.

Although the BRU can be ornamented with 'memory' and 'gating' functions to add functionality to the RNN, the general principle illustrated in Equations (E.5), (E.6) and Figure E.4 remains the same. In an encoder-decoder setup with an RNN, the whole input sequence must be fed into the encoder part of the network to produce a hidden activation state, which serves as the 'context' used by the decoder to produce the output sequence. However, one major drawback of naive RNN encoder-decoder architectures is that the context is strongly influenced by the last units of the input sequence, while the context of the first units is diluted. This limitation can be adressed by adding 'attention' modules that preserve the local context of each part of the input.

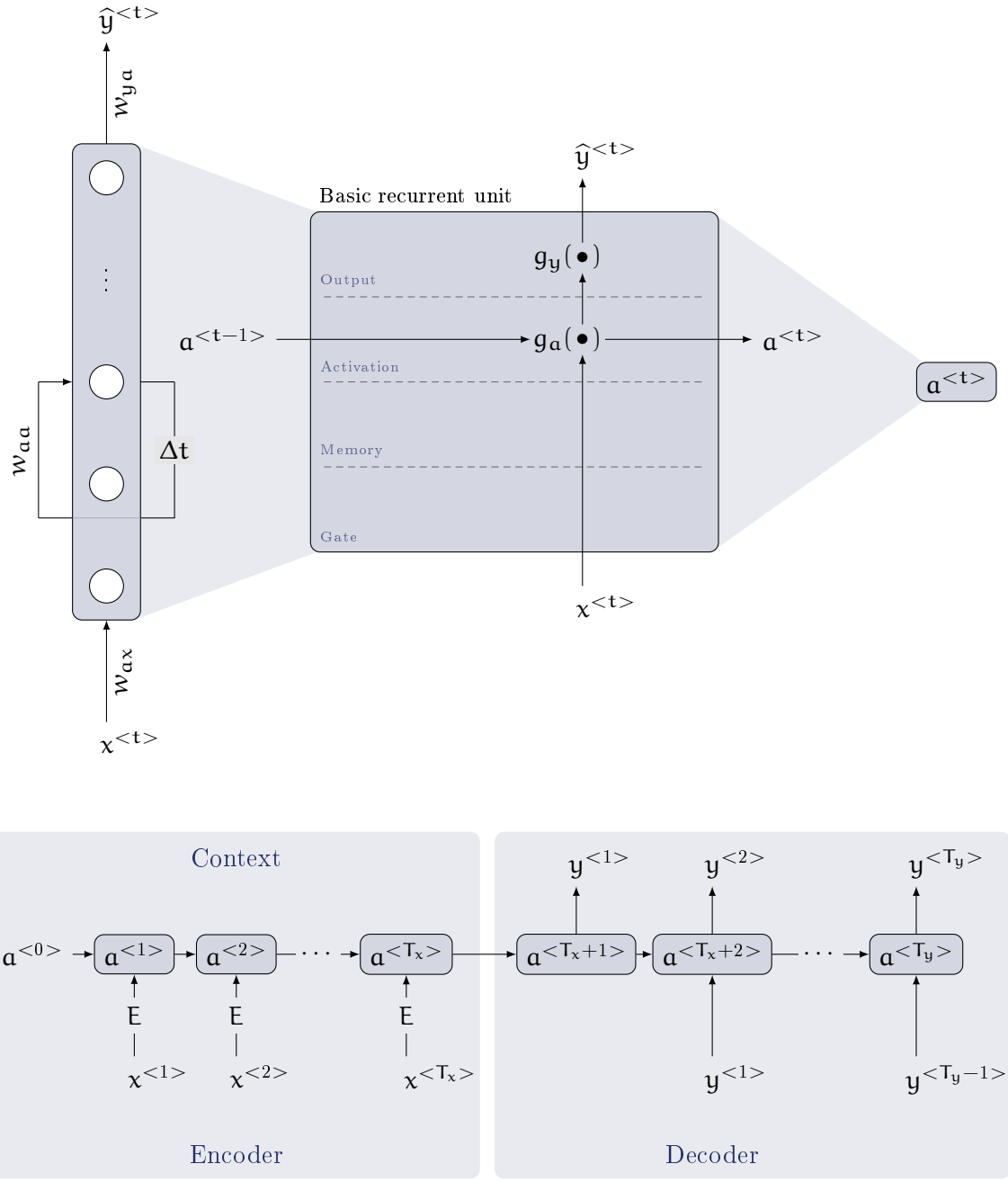


Figure E.4 | An encoder-decoder recurrent neural network architecture. Recurrent neural network (RNN) architectures are very different from convolutional and non-convolutional feedforward neural networks. Instead of being built as superposed layers of neurons, an RNN is represented as the progression of a unique recurrent motif through moments of a sequence. While the inputs $x^{<t>}$, activations $a^{<t>}$ and outputs $\hat{y}^{<t>}$ of the recurrent unit depend on the moment of the sequence, the weight and bias parameters are always the same. The RNN is commonly used in encoder-decoder type architectures, where the input sequence is encoded as a hidden embedding of informative features, which is fed to the decoder portion of the network, and used as 'context' to produce the output sequence.

In particular, 'transformers' are a class of NN architecture imbued with attention mechanisms that have greatly increased in popularity in recent years. In contrast to the RNN topology, transformers are able to process the entire input at once, and learn to focus only on its informative components. Thus, they can in theory be used to process arbitrarily lengthy input sequences.

Remarkably, transformers also leverage encoder-decoder layers. Each encoder layer consists of two major components: a self-attention mechanism and an FNN. The self-attention mechanism accepts input encodings from the previous encoder layer and weights their relevance to each other to generate output encodings. The FNN then further processes each output encoding individually. These output encodings are then passed to the next encoder layer as its input, as well as to the decoder layers. Importantly: the encoder is bidirectional: attention can be placed on components of the sequence before and after the current unit.

Each decoder layer consists of three major components: a self-attention mechanism, an attention mechanism over the encodings, and an FNN. Thus, the decoder functions in a similar fashion to the encoder, but with an additional attention mechanism which draws relevant information from the encodings generated by the encoders. This mechanism can also be called the *encoder-decoder attention*.