Host genetics and infectious disease: new tools, insights and translational opportunities

Andrew J. Kwok1, Alex Mentzer1,2 and Julian C. Knight1

Abstract | Understanding how human genetics influence infectious disease susceptibility offers the opportunity for new insights into pathogenesis, potential drug targets, risk stratification, response to therapy and vaccination. As new infectious diseases continue to emerge, together with growing levels of antimicrobial resistance and an increasing awareness of substantial differences between populations in genetic associations, the need for such work is expanding. In this Review, we illustrate how our understanding of the host–pathogen relationship is advancing through holistic approaches, describing current strategies to investigate the role of host genetic variation in established and emerging infections, including COVID-19, the need for wider application to diverse global populations mirroring the burden of disease, the impact of pathogen and vector genetic diversity and a broad array of immune and inflammation phenotypes that can be mapped as traits in health and disease. Insights from study of inborn errors of immunity and multi-omics profiling together with developments in analytical methods are further advancing our knowledge of this important area.

Disease syndromes caused by infectious agents have occurred throughout the history of modern humans1. As a result of our continued interactions with pathogens, our genomes have been shaped through processes of co-evolution, with pathogen-imposed selection pressures leading to selection signatures in ancient and modern human genomes2. As one of several illustrative examples, genetic diversity involving human red blood cell structure and function is being impacted by an evolutionary arms race with malaria that is reciprocally seen in the parasite genome3. In the era of modern medicine with antimicrobial drugs and powerful organ support systems, we are curbing the selection pressure exerted through the traditional high mortality associated with many of these infectious agents in resource-rich settings, but the worldwide health-care burden of infectious diseases remains substantial1. This is exemplified by the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)4,5 and the rapid evolution into a global COVID-19 pandemic requiring coordinated global responses. Additionally, the ongoing and dynamic challenges of managing malaria, tuberculosis (TB) and human immunodeficiency virus (HIV) infection alongside the constant threat of global influenza and other outbreaks such as Ebola virus disease, all particularly afflicting underserved populations, remind us of the importance of advancing infectious disease medicine.

A seminal study of adoptees in the 1980s reported increased risk of death from infectious disease in children whose biological parents succumbed to an infectious disease7, highlighting the significance of human genetic background in susceptibility to infections. As genetic technology has developed, mirroring other disease syndromes, a raft of genetic loci influencing susceptibility to infectious diseases have been discovered through genome-wide association studies (GWAS), including for HIV infection, TB, hepatitis and malaria8 (Table 1). Such studies aim to identify common variants on the basis of a polygenic model of complex multifactorial traits. However, it has become clear that, although human genetics play a role in disease susceptibility, in contrast to other syndromes, GWAS of case–control design may not be the most powerful method to tease out complex host–pathogen relationships. Alongside developments in genotyping technologies enabling GWAS, the in-depth understanding of single-gene disorders facilitated by next-generation sequencing technologies has enabled us to better understand the functional mechanisms of host defence to infection through experiments of nature9 resulting in loss of function (LOF) or, in increasingly recognized instances, gain of function (GOF) of discrete genes (Fig. 1). Here, rare mutations of variable penetrance result in single-gene inborn errors of immunity and lead to life-threatening infectious
| Reported trait/phenotype | Population                        | Sample size | Replication cohort? | SNP            | Purported gene/implicated gene            | P              | Odds ratio | Refs |
|--------------------------|-----------------------------------|-------------|---------------------|----------------|-------------------------------------------|----------------|------------|------|
| HIV infection            |                                   |             |                     |                |                                           |                |            |      |
| Set point viral load     | European                          | 2,362       | Yes                 | rs9264942      | HLA-C                                     | $5.9 \times 10^{-32}$ | NA         | 164,165 |
|                          | European                          | 2,362       | Yes                 | rs2395029      | HLA-B*5701, HCP5                          | $4.5 \times 10^{-35}$ | NA         | 164,165 |
| Set point viral load     | European                          | 1,712       | Yes                 | rs9264942      | HLA-C                                     | $2.8 \times 10^{-35}$ (univariate regression model) | 2.9         | 166     |
|                          | European                          | 1,712       | Yes                 | rs4418214      | MICA                                      | $1.4 \times 10^{-34}$ (univariate regression model) | 4.4         | 166     |
|                          | African                           | 1,233       | No                  | rs2523608      | HLA-B                                     | $8.9 \times 10^{-20}$ (univariate regression model) | 2.6         | 166     |
|                          | African and Hispanics             | 1,233,677   | No                  | rs2523590      | HLA-B                                     | $1.7 \times 10^{-13}$, $8.3 \times 10^{-8}$ (univariate regression model) | 2.4         | 166     |
| Susceptibility           | African Americans and European Americans | 3,136 | Yes                 | rs4878712      | FRMPD1                                    | $4 \times 10^{-7}$ | NA         | 167     |
| Viral load               | Chinese                           | 538         | No                  | rs2442719      | HLA-B                                     | $7.85 \times 10^{-7}$ | NA         | 168     |
| HIV-1 acquisition        | European                          | 13,581      | No                  | NA             | CCR5,532                                  | $5 \times 10^{-6}$ | 0.2        | 169     |
| Viral load               | European                          | 6,315       | No                  | rs59440261     | HCP5, MICA                                | $2.0 \times 10^{-83}$ | NA         | 170     |
|                          | European                          | 6,315       | No                  | rs1015164      | CCRL2                                     | $1.5 \times 10^{-19}$ | NA         | 170     |
| Black, Hispanic, Asian, | European                          | 2,440       | No                  | rs57989216     | AL671883.3                                | $2.6 \times 10^{-13}$ | NA         | 170     |
| Early-onset tuberculosis| African                           | 3,176       | No                  | rs2057178      | AGMO                                      | $2.63 \times 10^{-4}$ | 0.77       | 170     |
| Early progression to    | Peruvian (mixed ancestry)         | 4,002       | No                  | rs73226617     | ATP1B3                                    | $3.93 \times 10^{-4}$ | 1.18       | 171     |
| active tuberculosis      | East Asian                        | 2,053       | Yes                 | rs12437118     | ESRBB                                     | $1.72 \times 10^{-31}$ | 1.277      | 171     |
| Susceptibility           | East Asian                        | 2,053       | Yes                 | rs6114027      | TGM6                                      | $2.37 \times 10^{-13}$ | 1.339      | 171     |
|                          | East Asian (Chinese)              | 2,509       | Yes                 | rs4240897      | MFN2                                      | $1.41 \times 10^{-31}$ | 0.79        | 171     |
| Resistance               | African (Ghanaian)                | 3,176       | Yes                 | rs2057178      | WT1                                       | $2.63 \times 10^{-4}$ | 0.77        | 171     |
| Early-onset tuberculosis| Middle Eastern (Moroccan)         | 558         | Yes                 | rs916943       | AGMO                                      | $2 \times 10^{-6}$ | 2.73        | 171     |
| Malaria                  | African                           | 7,908       | Yes                 | rs10900585     | ATP2B4                                    | $6.1 \times 10^{-9}$ | 0.65        | 177     |
| Severe Plasmodium       | African                           | 7,908       | Yes                 | rs2334880      | MARVELD3                                  | $3.9 \times 10^{-8}$ | 1.24        | 177     |
| falciparum malaria      | African                           | 7,908       | Yes                 | rs817619       | ABO                                       | $4.0 \times 10^{-23}$ | 1.73        | 177     |
|                          | African                           | 7,908       | Yes                 | rs372091       | HBB                                       | $1.1 \times 10^{-14}$ | 0.44        | 177     |
|                          | African                           | 25,904      | Yes                 | rs184895969    | Glycophorin genes                         | $9.5 \times 10^{-13}$ | 0.67        | 44      |
| Leprosy                  | East Asian (Chinese)              | 24,330      | Yes                 | rs2221593      | BATF3                                     | $3.09 \times 10^{-8}$ | 1.15        | 178     |
| Susceptibility           | East Asian (Chinese)              | 24,330      | Yes                 | rs77061563     | CIITA                                     | $6.23 \times 10^{-25}$ | 0.84        | 178     |
|                          | East Asian (Chinese)              | 24,330      | Yes                 | rs637343       | CCDC88B                                   | $8.84 \times 10^{-12}$ | 1.24        | 178     |
|                          | East Asian (Chinese)              | 23,766      | Yes                 | rs6807915      | SYN2                                      | $1.94 \times 10^{-8}$ | 0.89        | 95      |
|                          | East Asian (Chinese)              | 23,766      | Yes                 | rs4720118      | BBS9                                      | $3.85 \times 10^{-10}$ | 1.16        | 95      |
|                          | East Asian (Chinese)              | 23,766      | Yes                 | rs55894533     | CTSB                                      | $5.07 \times 10^{-10}$ | 1.15        | 95      |
|                          | East Asian (Chinese)              | 23,766      | Yes                 | rs10100465     | MED30                                     | $2.85 \times 10^{-10}$ | 0.85        | 95      |
disease (Table 2). More recently, the development and availability of multi-omics experimental techniques promise to deliver similar mechanistic understanding of the functional importance of genome-level variation at genomic loci implicated in disease pathogenesis.

This Review aims to summarize the insights gained from the wide range of -omics approaches being used to understand genetic susceptibility to infectious disease. We begin by discussing recent findings and limitations of GWAS applied to infectious diseases, in particular the need for application in diverse global populations. We focus on lessons learnt from different infections and on a region of the genome consistently linked to infectious disease susceptibility: the major histocompatibility (MHC) locus encoding the vital immune human leukocyte antigen (HLA) genes (Box 1). We then proceed to discuss novel approaches, many of which involve large-scale sequencing, and highlight the importance of monogenic architecture leading to infectious disease susceptibility, although only selected examples will be covered as this topic was reviewed comprehensively recently. This is followed by exploration of the potential of future integrated analyses of human–pathogen variation and application of multi-omics platforms. We postulate how these findings may inform our understanding of the genetic determinants of infectious disease susceptibility, how this knowledge may in turn improve our understanding of individual immune and vaccine responses, and how this might be directed more universally for the benefit of human health. Finally, we discuss strategic approaches to dissecting host genetic factors in COVID-19 as a case study in how such work is being implemented as part of global efforts to address this emerging public health threat.

Common-variant associations: HLA and beyond

For major infectious diseases, including HIV-1 infection, TB, malaria, leprosy and viral hepatitis, genetic evidence has been accrued primarily through GWAS, aiming to identify disease-associated common variants in populations on the basis of a polygenic model of complex multifactorial traits. Recent work involving these traits has highlighted the progress and challenges that continue to be associated with mapping and functionally interrogating genetic associations through GWAS.

Genetic factors are hypothesized to contribute to the observed heterogeneity in response to infectious disease, and twin studies are an important source of information to estimate heritability while recognizing the risks of overestimation. It is also important to note

| Reported trait/phenotype | Population | Sample size | Replication cohort? | SNP | Purported gene/implicated gene | P | Odds ratio | Refs |
|--------------------------|------------|-------------|---------------------|-----|--------------------------------|---|-----------|------|
| **Hepatitis B**          |            |             |                     |     |                                |   |           |      |
| Chronic HBV infection    | East Asian | 1,888       | Yes                 | rs7453920 | HLA-DQB2                     | 4.93 × 10⁻³⁷ | 0.53 | 179 |
|                         | East Asian | 1,888       | Yes                 | rs3130542 | HLA-C                        | 9.49 × 10⁻³⁴ | 1.33 | 179 |
|                         | East Asian | 1,888       | Yes                 | rs4821116 | UBE2L3                       | 1.71 × 10⁻³² | 0.82 | 179 |
| Persistent HBV infection | East Asian | 2,308       | Yes                 | rs7000921 | INTS10                       | 3.2 × 10⁻¹² | 0.78 | 180 |
| **Hepatitis C**          |            |             |                     |     |                                |   |           |      |
| Chronic HCV infection    | East Asian (Japanese) | 36,112 | Yes             | rs9275572 | HLA-DQB1                    | 3.59 × 10⁻⁵⁶ | 0.79 | 181 |
|                         | Mixed      | 2,401       | Yes                 | rs12979860 | IL-28B                       | 2.17 × 10⁻⁵⁰ | 0.45 | 50  |
|                         | Mixed      | 2,401       | Yes                 | rs4273729 | HLA-DQB1                    | 1.71 × 10⁻³⁸ | 0.59 | 50  |
| Spontaneous clearance of HCV | Mixed   | 4,423       | Yes                 | rs74597329 | IFNL4–IFNL3                  | 5.99 × 10⁻³⁰ | 2.14 | 51  |
|                         | Mixed      | 4,423       | Yes                 | rs2647006 | HLA-DQB1, HLA-DQA2           | 1.15 × 10⁻¹² | 1.71 | 51  |
|                         | Mixed      | 4,423       | Yes                 | rs1754257 | GPR158                       | 1.8 × 10⁻⁷  | 1.06 | 51  |
| **Dengue fever**         |            |             |                     |     |                                |   |           |      |
| Dengue shock syndrome    | Vietnamese | 8,697       | Yes                 | rs3132468 | MICB                         | 4.4 × 10⁻¹¹ | 1.34 | 182 |
|                         | Vietnamese | 8,697       | Yes                 | rs3765524 | PLCE1                        | 3.1 × 10⁻¹⁰ | 0.8  | 182 |
| **Non-typhoidal Salmonella infection** |            |             |                     |     |                                |   |           |      |
| Non-typhoidal Salmonella bacteraemia | African   | 3,591       | Yes                 | rs13390936 | STAT4                        | 8.62 × 10⁻¹⁰ | 7.61 | 125 |
| **Typhoid**              |            |             |                     |     |                                |   |           |      |
| Enteric fever            | Asian (Vietnamese and Nepalese) | 4,243 | Yes             | rs7765379 | HLA-DRB1–HLA-DQB1               | 2.29 × 10⁻¹³ | 0.22 | 183 |
| **Rheumatic heart disease (after Streptococcus pyogenes infection)** |            |             |                     |     |                                |   |           |      |
| Susceptibility           | Oceanian (Polynesians, South Asians and mixed) | 2,852 | No               | rs11846409 | IGHV4-61*02                  | 3.6 × 10⁻⁹  | 1.43 | 67  |

Data sourced from the US National Human Genome Research Institute–European Bioinformatics Institute GWAS Catalog. The table includes most recent studies (2015 onwards), as well as previous definitive studies in the field of infectious disease. HBV, hepatitis B virus; HCV, hepatitis C virus; NA, not available; SNP, single-nucleotide polymorphism.
**a PRR signalling**

Viral RNA → CARD → CTD Helicase domain → TRIM25 → MAVS → Mitochondrion → TRIF/TLR → RLR → Cytoplasm → TRAF6 → TBK1 → IKKα/IKKβ → NF-κB

Susceptibility to HSV encephalitis → IRF3 → Type I interferons

Susceptibility to HSV encephalitis → IRF7 → Type III interferons

Severe influenza

**b RIPK signalling in immune pathways**

TLR4 + MyD88 + TRIF/ RIKK1 → TRAM + TAB1/ TAB2/ TAK1 → MAPK activation

Susceptibility to HSV encephalitis

RAK1 → IRF9 → STAT1/2 → ISGs

MAPK activation

NF-κB activation

MAPK → Apoptosis + necroptosis

RIPK3

**c Interferon pathways**

Type I interferons

CMV viraemia and Streptococcus viridans bacteraemia

Disseminated strain measles, HSV-6

* Gain of function: fungal, bacterial, viral (HSV) infections

* Loss of function: mycobacterial infections

Type II interferons

**d Antigen presentation pathways**

Viral antigen (intracellular) → Proteasome → TAP1/2 → Endoplasmic reticulum → MHC class I

Bacterial respiratory tract infection, bronchectasis

CD8+ T cell

TCR

Extracellular antigen

CD4+ T cell

TCR

Antigen

Endosome

CLIP (invariant chain) → MHC class II

MHC class II gene expression: CIITA, RFX5, RFXAP, RFXANK → recurrent severe infections

Mutations in genes encoding transcription factors regulating MHC class II gene expression: CIITA, RFX5, RFXAP, RFXANK → Mendelian susceptibility to mycobacterial disease

SPPL2A — required to cleave CLIP → Mendelian susceptibility to mycobacterial disease
that while ongoing work investigating rare and structural
genomic variants may help further address the ‘missing heritability’
conundrum — the disparity between overall
trait heritability and observed effects through GWAS
— this has been reconciled mostly by the recognition
that genetic predisposition in most common traits is spread
and shared across large numbers of variants
(in the thousands), each with a typically modest effect[5].
Understanding of the genetic architecture of disease
predisposition continues to develop, with an omnigenic
model for complex traits proposing genetic contribution
from both core genes with strong genetic evidence and
very many networked peripheral genes of very low effect
size involving regulatory circuits[12,13].

Of particular interest in infectious disease GWAS has
been the role of genetic variation at the locus encoding
HLA genes in determining the individual response to
infection, given the role of encoded MHC molecules in
binding and presenting antigens and thus determining
which antigens are ‘seen’ by the immune system (BOX 1).
However, the extreme polymorphism of this region, high
gene density, complex linkage disequilibrium and regions
of homology make fine mapping genetic associations
involving HLA very challenging[14].

Across the genome, a substantial majority of genetic
variants identified by GWAS are located in non-coding
regions of the genome, requiring complementary func-
tional genomic annotations to link such variants to mod-
ulation of individual genes or pathways. This is key to
understanding the functional basis of such associations,
in particular the specific modulated genes or pathways
that may represent potential drug targets. Knowledge of
network connectivity substantially enhances the ability
to then identify nodal genes that may be the best ther-
apeutic targets[15]. Prior genetic evidence for targets sub-
stantially increases the success rate in late-stage clinical
trials[16]. As well as supporting the development of new
drugs, it provides important opportunities to support
repurposing existing drugs where a genetic associa-
tion with a target is shared across traits, indicating new
indications for such approved drugs[17].

Genetics can also enable analytical approaches such as
Mendelian randomization that leverage genetic varia-
tion as instrumental variables to establish causal relations-
ships between non-genetic (modifiable) risk factors and
disease[18]. A further example of translational relevance is
the opportunity for genetic evidence to potentially strat-
ify risk; for example, to identify high-risk individuals for
whom early or pre-emptive intervention may be appro-
priate. Aggregated genetic risk through polygenic risk
scores and consideration of large-effect rare variants are
gaining traction as useful clinical tools[19], with reports,
for example, of application to dengue fever[20].

HIV-1 infection. The two genomic regions most consist-
ently associated with HIV infection outcomes are the
MHC[21] and the CCR5 locus (CCR5 is a coreceptor for the
virus) [22–24]. Arguably, HIV-1 infection is the disease with
the greatest insights into HLA–disease association
pathogenesis, including seminal early advances support-
ing the hypothesis of ‘heterozygous advantage’, with
individuals with greater heterozygosity of class I alleles
(HLA-A, HLA-B and HLA-C) having delayed progres-
sion to AIDS[25]. A more recent discovery has demonstrated
that differential expression of HLA-A contributes to
disease progression through HLA-E–mediated effects,
implicating the interaction between HLA-E and the
NKG2A natural killer cell inhibitory receptor as a ther-
apapeutic target[26]. Both HLA-B and HLA-C have been
consistently implicated in HIV control, but strong
coinheritance or linkage disequilibrium in the MHC
region has made disentangling the contributions of
each difficult.

The repeated identification of HLA associations in
HIV-1 GWAS has contributed, in part, to a catalogue of
work focusing on the killer immunoglobulin-like
receptor (KIR) genes that encode a family of highly poly-
orphic activating and inhibitory receptors expressed
mainly by natural killer cells that interact with HLA.
There has been a surprising lack of KIR gene associa-
tions in GWAS to date, which may be due to a multitude
of reasons, including substantial diversification of KIR
gene haplotypes, incomplete knowledge of the nature
and diversity of such haplotypes across populations, and
poor coverage of current genotyping arrays for this
region, making it difficult to pinpoint specific genetic
variants across populations. Thus, this region could
represent an archetypal complex region that we have
yet to accurately tag in even the most comprehensive of
modern GWAS. Moreover, the effects of KIR gene vari-
ants may be detected only in the presence of their spe-
cific HLA ligand. As evidence for this effect, KIR3DL1
haplotypes expressed highly in combination with specific
HLA-B alleles (Bw4) were found to slow disease pro-
gression when compared with a low-expression haplotype

Linkage disequilibrium
The non-random association of alleles at different loci in
a population, often owing to physical proximity of the loci
on a chromosome.

Heterozygous advantage
Where heterozygotes have fitness advantages over
homozygotes.

Haplotypes
Sets of genetic variants that
tend to be inherited together.

Allotypes
A group of alleles that encode
a unique protein sequence.
This table lists inborn errors of immunity with genes and associated infectious disease phenotypes indicated. The classification for inborn errors of immunity is based on the part of the immune system affected. PIDs are associated with overt immunological abnormalities and have high/complete penetrance; they are often associated with other clinical phenotypes, such as autoimmunity, autoinflammation, cancer and allergy. Examples of other monogenic inborn errors of immunity are shown, which typically disrupt host defence against one or a few infectious agents, differentiated by penetrance. AD, autosomal dominant; AR, autosomal recessive; IFN-γ, interferon-γ; TLR, Toll-like receptor. "High-penetrance/complete-penetrance inborn errors of immunity (Mendelian infections)" are usually familial. "Low-penetrance inborn errors of immunity are typically sporadic and rare."
Population structure
Also referred to as population stratification. The presence of differences in allele frequencies between subpopulations in a population.

Expression quantitative trait locus (eQTL)
A genomic region that is associated with a proportion of variation in mRNA expression levels.

Balancing selection
A selection process that maintains genetic diversity.

or lack of the Bw4 allele. By contrast, homozygosity for the CCR5Δ32 mutation has been the most robustly confirmed genetic background for protection against HIV, and antisense long non-coding RNA CCR5AS, which affects CCR5 mRNA stability and cell surface expression, has recently been shown to influence HIV infection susceptibility. Furthermore, knowledge of the CCR5 pathway has led to the successful development of a class of antiviral agents targeting the entry of HIV-1 into target host cells.

**Tuberculosis.** Despite some of the earliest evidence for TB susceptibility being driven by genetic heterogeneity, a consistent genetic signal of association in any one locus has yet to be consistently demonstrated. The largest study to date was undertaken in a combined sequencing and genotyping effort including 8,162 cases of combined presentations of TB in the Icelandic population. This study identified variants in the MHC class II region as being associated with TB susceptibility, perhaps emphasizing the need to undertake GWAS in highly homogeneous populations when tackling ancient pathogens such as *Mycobacterium tuberculosis*. By contrast, other GWAS applied to TB of varied organ involvement in diverse populations have demonstrated association of pulmonary TB with variants in ESRRB and TGM6 in a Chinese population and as ASAPI in a Russian population, although the biological basis by which these genes mediate an effect in TB is not always clear. For example, TB mouse models with CRISPR–Cas9 deletion of TGM6 lead to increased cytokine transcripts in the lungs, greater tissue damage and higher bacterial burden, but whether the cytokine levels are directly due to altered TGM6 levels is yet to be established.

The lack of consistent signals in the TB GWAS highlight the enormous challenges in undertaking multinational studies, where the observations may be attributed to differences in populations because of population structure either at the host level or the pathogen level or may be owing to technical differences, such as the genotyping array used. A meta-analysis of these studies that might help explain the observed incompatibilities and may inform future research in the area is eagerly awaited.

**Leprosy.** By contrast, many genes have been identified through GWAS for susceptibility to leprosy due to the related species *Mycobacterium leprae* that unsurprisingly implicate the immune system, including, most notably, the HLA-DR region. Whereas a GWAS in an Indian cohort found HLA-DR, HLA-DQ and TLR1 (encoding Toll-like receptor 1) to be associated with susceptibility to leprosy, a further GWAS of Chinese patients failed to replicate the TLR1 association despite adequate power, implicating population-specific effects. This is again in line with the hypothesis of the genetic variation responsible for variability in the infectious disease trait (human genetic architecture) being subjected to selection pressure from pathogens, as different populations would have been exposed to various infective agents. The large size of the case–control cohorts has enabled the identification and downstream functional characterization of some of the non-HLA loci; for example, for NOD2, where the single-nucleotide polymorphism (SNP) rs1981760 is associated with differential gene expression by being an expression quantitative trait locus (eQTL) in neutrophils. The SNP acts by affecting binding of the transcription factor STAT3 and influencing neutrophil interferon-β (IFNβ) responses, an effect that has been also reported in TB.

**Box 1 | The MHC and HLA genes**

The major histocompatibility complex (MHC) is a locus on chromosome band 6p21, which in humans includes the classical human leukocyte antigen (HLA) genes that encode a group of proteins responsible for presenting antigens to T cells, the MHC molecules. Split into three regions (classes I, II and III), the locus includes genes for well-characterized proteins involved in antigen processing and presentation as well as a diverse range of other immune functions. The classical and highly polymorphic HLA-A, HLA-B and HLA-C genes and non-classical and less polymorphic HLA-E, HLA-F and HLA-G genes are located in the class I region. The class II region includes HLA-DR, HLA-DP, HLA-DQ and HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB2, HLA-DRB3, HLA-DRB4 and HLA-DRB5 genes. Genes involved in other aspects of immunity are found in the class III region, such as those for complement components. Although the primary function of both class I and class II molecules is to bind and present peptides to T cells, class I molecules also interact with killer immunoglobulin-like receptors (KIRs) and leukocyte immunoglobulin-like receptor (LILRs) on natural killer cells and monocytes, respectively.

This intimate relationship with immune function coincides with the fact that the MHC is the genomic locus with the most associations with disease. However, the high polymorphism, homology and gene density and the strong linkage disequilibrium in the locus make the contributions and causal mechanisms of individual loci difficult to determine. The concept of ‘polymorphic frozen blocks’ arose from the substantial linkage disequilibrium and may be explained by genetic mechanisms — for example, insufficient time for recombination owing to isolated populations in recent history/lower meiotic pairing of highly variable genomic regions — or it may result from preservation of coordinated sets of alleles given the likely epistasis of the genes.

The extensive diversity of class I and class II alleles is likely to reflect selection pressures from infectious diseases, with both heterozygote advantage and frequency-dependent selection having been proposed to lead to population HLA diversity. However, one interesting aspect of HLA function that provides an alternative explanation is the interaction between maternal KIRs and fetal HLA-C molecules, which influence reproductive fitness, which may drive HLA diversity.

**Malaria.** This parasitic infection is perhaps the most famous example of host genetics influencing infectious disease susceptibility, with haemoglobinopathies — such as sickle haemoglobin (HbS), thalassaemia and glucose 6-phosphate dehydrogenase (G6PD) deficiency — long being associated with protection against malaria. As one example of an underlying biological mechanism, haemoglobin oxidation products in HbS affect the ability of the *Plasmodium falciparum* parasite to hijack red blood cell cytoskeletal remodelling. Balancing selection, illustrated by the heterozygote advantage conferred by HbS, provides an important mechanism for maintaining advantageous genetic diversity in human populations.

Malaria is an example where modern GWAS have confirmed several loci previously identified largely by candidate gene or protein studies, albeit still not most of them; most notably, despite many early observations of associations with HLA, GWAS projects have not found equivalent evidence. The Malaria Genomic Epidemiology Network (MalariaGEN) has successfully identified or confirmed genetic loci associated...
with severe *Plasmodium falciparum* malaria across Africa, Asia and Australasia, including *HBB* (ß-globin gene), *ABO* (gene for glycosyltransferase enzyme that determines ABO blood groups), *ATPB284* (gene for a calcium transporter on the erythrocyte membrane), *G6PD* (activity levels of the enzyme affect oxidative stress in red blood cells) and *CD40LG* (a key gene in B cell immune responses). A notable finding was that among both different phenotypes and different populations, the heterogeneity of effect of risk alleles was substantial, possibly explaining the inability to replicate some of the previous genetic findings, together with pathogen diversity. The consortium further identified a locus neighbouring glycophorin genes (encoding erythrocyte cell surface receptors that allow *Plasmodium falciparum* invasion) to be implicated in malaria susceptibility. This finding highlights the possibility that loci conferring resistance to malaria involve polymorphisms with features of ancient balancing selection based on haplotype sharing between humans and chimpanzees, in keeping with the hypothesis of a protracted evolutionary arms race between host and pathogen leading to the host and pathogen genetic diversity that we see today. The link between glycophorin genes and malaria resistance was further elucidated in a study using next-generation sequencing data and reference haplotypes that identified a complex structural rearrangement involving GYPA and GYPB gene copy number that reduced the risk of severe malaria by 40% and that explained the previous association signal.

**Viral hepatitis.** Whereas three-quarters of patients with acute hepatitis C virus (HCV) infections eventually reach chronicity, the remaining quarter clear the virus spontaneously. As a key example of a strong GWAS signal not related to HLA with major translational impact, polymorphisms in the gene encoding IFNλ3 (*IFNL3*; formerly known as *IL28B*) have been associated with resolution of HCV infection and with differences in response to HCV drug treatment (viral clearance following treatment with pegylated interferon and ribavirin) in individuals of both European and African ancestry, enabling development of a precision medicine approach to therapy (Fig. 2). HLA class II associations with spontaneous HCV clearance have been reported by GWAS. In a separate GWAS recently of African, European and Hispanic populations, HLA class II signals were replicated, while loci around the IFNA genes and *GPR158* were also implicated in spontaneous viral clearance.

**Novel approaches and technologies**

**Human population diversity.** Lessons learnt in infectious disease susceptibility GWAS to date highlight several principles for increasing the chances of success. A key emerging area is the extent of population-specific associations genome-wide (TABLE 1). This in turn emphasizes the relative paucity of GWAS undertaken in diverse worldwide populations, many of which have the highest burden of infectious disease. Greater efforts will be needed to focus GWAS and other genetic studies on such populations, especially where ethnic background may be associated with differential risk. It is thus imperative that genotyping arrays are designed for worldwide populations and that the availability of large imputation reference panels from diverse populations enables effective statistical inference of additional non-genotyped SNPs (and hence substantially increasing the informativeness of GWAS at no additional cost). New technologies such as those based on low-coverage whole-genome sequencing may provide an alternative cost-effective approach for genotyping common and rarer genetic variants.

As is apparent from the studies reported so far herein, infections can affect populations limited by geography and local ecology, and it is therefore crucial that genetic associations are tested for in non-European cohorts. However, such analyses come with technical challenges...
owing to the need to adjust the data for potential confounding by differences in population structure or relatedness that may cause an excess of type I errors. These aspects have been appreciated for some time, and many methods have been developed to facilitate such adjustments. One method is principal component analysis (a mathematical method to summarize the main sources of variance in data) to model ancestry differences and account for differing allele frequencies in different populations. A further development is in statistical modelling using linear mixed model algorithms that can account for population structure and cryptic relatedness simultaneously.

The Population Architecture Using Genomics and Epidemiology (PAGE) Consortium has demonstrated both the power and the importance in accounting for multi-ethnic and diverse populations in various non-infectious disease and physical trait settings; a GWAS conducted by the consortium demonstrated that in different populations, the effect size of implicated loci differed, exemplifying the need to account for population structure. Any initiatives planning to develop personalized medicine in infectious diseases based on GWAS will therefore have to depend on the availability of genetic studies that are commensurately precise in a population context and that include an appreciation of the genomic complexity of regions such as the MHC, KIR and immunoglobulin heavy chain (IGH) regions.

**Leveraging self-report.** The maturation of methods for undertaking GWAS has been paralleled by improvements in methods for recruiting and testing associations. A key study from the personal genetic provider 23andMe used self-report to assess the influence of host genetic background on susceptibility to a range of common infectious diseases of low mortality (for example, common cold, rubella and cold sores). Including more than 200,000 individuals of European descent, this study reported many genetic loci related to the immune system and embryonic development, including genes both within and outside the MHC. Despite the large number of associations identified, the use of surveys requiring retrospective recall to identify cases will require follow-up replication and validation. A similar method was used for whooping cough, where self-report of the characteristic cough in childhood was associated with variation across the MHC.

**Phenome-wide association studies.** Another development in recent years has been to establish the phenome-wide association study (PheWAS) approach. This has been facilitated by the use of electronic health records for clinical phenotyping. Instead of taking a single disease phenotype and scanning genome-wide as in GWAS, this approach takes a number of genetic loci (usually based on a priori biological knowledge) and looks for associations with any of a large number of phenotypes. Although revealing for many immune-related diseases, it has had limited use thus far in the field of infectious diseases. A PheWAS on HLA predictably highlighted autoimmune diseases, but found only a relatively small number of associations with infectious diseases. Given the known key role of HLA genes in infectious diseases, it is likely that this was due to a lack of statistical power or challenges in defining cases from controls using hospital data.

Nevertheless, one example of a PheWAS in infectious disease genetics identified an association between the rs3211783 SNP in the factor X gene, encoding an enzyme of the coagulation cascade, with infection-related phenotypes. Recent evidence for the role of factor X in antibacterial immunity has emerged, and more studies will be needed to confirm this role, but the study provides an example of what may come in the field of infectious disease genetics through future PheWAS.

**Genome sequencing and large-scale bioresources.** Standard GWAS and imputation methods are unable to test the effects of rare, large-effect variants, as these variants are not covered by genotyping arrays and are not common enough in haplotype references. To overcome this challenge, direct whole-genome sequencing (WGS) or whole-exome sequencing (WES) approaches have been used. In one example, WES was first performed to identify rare alleles that are relevant in West Nile neuroinvasive disease, enabling subsequent imputation of the rare alleles into a primary cohort. The findings were then extended to larger cohorts for genome-wide association analysis. Loci revealed as being important included HERC5, an interferon-stimulated gene, and an intergenic region between CD83 and JARID2, which included a site for STAT5A transcription factor binding. With very few other such examples in the literature for any trait, it is becoming increasingly clear that the successful application of these methods will require hundreds to millions of individuals.

To this end, several population-level bioresources and personal genetic services such as the UK Biobank, the China Kadoorie Biobank and the 100,000 Genomes Project — which are large repositories of samples and data — are enabling the use of novel approaches and enhancing the quality of data analysis using existing methods. In light of early discoveries afforded by the availability of genetic data from these studies, it will be exciting to see how they may provide further insights into infectious disease as they mature.

**Long-read sequencing technologies and highly polymorphic loci.** A GWAS of rheumatic heart disease in Pacific Island populations recently identified a signal in the IGH locus. This disease, which occurs following *Streptococcus pyogenes* infection, used to have a more widespread worldwide prevalence and is now rather limited to underserved populations across the Pacific. This intriguing finding in the IGH region has long been suspected for many diseases but has been infrequently observed. Studying the IGH region has been challenging owing to its high level of complexity (including substantial copy number variation) and the sparse coverage of this locus on genotyping arrays. However, with the advent of new long-read sequencing and other technologies, it is likely that we will better capture the extent of IGH diversity and possible disease associations, similarly to what has been achieved for regions such as the MHC and KIR regions.
New insights from inborn errors of immunity

Another key route to understanding infectious disease susceptibility is through the study of inborn errors of immunity that disrupt host defence against infection. In contrast to more common variants with small to moderate effect sizes identified by GWAS, rare mutations (of variable penetrance) typically result in a large phenotypic effect with substantial infections for the individual that may involve multiple pathogens. Such analysis has been highly informative in revealing mechanisms of infectious disease pathogenesis (FIG. 1; TABLE 2). Intersection with GWAS is seen: for example, single-gene defects in complement components result in increased susceptibility to meningococcal disease, while GWAS identified SNPs within complement factor H (CFH) and CFH-related protein 3, consolidating evidence for the role of complement in host defence against meningococcal infection. Application of WGS and WES has now identified more than 400 genetic causes of inborn errors of immunity, many of which cause primary immunodeficiencies (PIDs), including genes of known and novel mediators of innate and adaptive immunity (TABLE 2) mutated in various manners leading to the full range of mechanisms for Mendelian inheritance patterns. One notable advancement in this area has been the recognition of disease-causing genes that do not necessarily lead to overt immunological abnormalities, with differences in penetrance and specificities in clinical infection phenotype among other features (TABLE 2).

The link between certain immune pathways with specific clinical infection phenotypes has been well documented. For example, the TLR1/2/6–Toll–interleukin-1 receptor domain-containing adaptor protein (TIRAP)–nuclear factor-kB (NF-kB) pathway influences invasive pneumococcal disease, whereas the TLR3–interferon pathway affects susceptibility to herpes simplex virus encephalitis (FIG. 1a), whereas, in cases of life-threatening influenza, WES identified compound heterozygous mutations in IRF7, which reduces both type I and type III interferon production. Influenza virus pneumonia may also arise owing to rare inborn errors of immunity involving IRF9 and TLR3 deficiencies, whereas homozygous CCR5-null and FUT2-null alleles protect against CCR5-tropic HIV and norovirus, respectively.

Mendelian susceptibility to mycobacterial disease can arise owing to mutations in genes encoding proteins involved in the production of, or response to, the cytokine IFNγ. Such mutations have important treatment implications; for example, the deficiency of IFNγ signalling can be ameliorated by therapeutic supplementation with IFNγ or haematopoietic stem cell transplantation, dependent on the degree of functional pathway responsiveness. WES has revealed a novel cause of Mendelian susceptibility to mycobacterial disease in a case of homozygous splice-site mutations in SPPL2A (required to cleave CD74, the invariant chain of HLA class II); the mutation is proposed to disrupt functional polarization of T cells owing to altered priming by dendritic cells after mycobacterial antigen presentation.

One hurdle in monogenic disorders is incomplete penetrance. Polymorphism of TIRAP is known to influence susceptibility to invasive pneumococcal disease, TB, malaria and other bacterial infections. In a study of a family with TIRAP deficiency, seven of eight individuals with the genetic defect did not have the severe staphylococcal infection that the index case did. A possible explanation was the lack of antibodies to lipoteichoic acid found in the index patient compared with other family members that may allow monocytes in the unaffected family members to respond to lipoteichoic acid stimulation of TL2 despite deficiency in TIRAP.
**Integrating host and pathogen genetic variation**

Although the impact of host genetic variation on susceptibility to infection is often viewed separately from that of pathogen genetics, it is highly likely that a major limitation in identifying novel genetic signals in GWAS applied to common infections will be attributable to heterogeneity at the level of the infectious agent. Approaches attempting to simultaneously interrogate host and pathogen transcriptomic responses (dual RNA sequencing) have been developed and applied to infectious diseases \(^{98-100}\); this joint characterization has also occurred at the genetic level to account for variability in both the host genome and the pathogen genome. In a Dutch cohort, a GWAS of both the host and the pathogen in pneumococcal meningitis allowed delineation of the contributions of host genetic background to the susceptibility and severity of disease versus the effects of pneumococcal genetic variation on invasive potential but not disease severity \(^{101}\). The study revealed an intronic SNP in a host ubiquitin-converting enzyme gene, *UBE2U*, to be significantly associated, with possible interaction with other host genes, including *PGM1* (encoding phosphoglucomutase 1) and *ROR1* (encoding a tyrosine-protein kinase transmembrane receptor).

Similarly, a joint genetic study of HCV infection with human genome-wide genotyping and genome sequencing of HCV genotype 3 infections led to more biological insights, including understanding of reciprocal effects (for example, of human genetic variation in driving viral genome polymorphism) \(^{102}\). Rather than association tests using disease status, this study examined the associations between the host genetic background and variable sites of the viral proteome. Host HLA alleles were found to be associated with HCV amino acids leaving multiple `footprints` in the HCV genome; and, in addition, a specific interaction between the HCV NSSA protein and human *IFNL4* (encoding IFNλ4) genotype was found to affect viral load.

Simultaneous evaluation of host and pathogen genetics was also used in HIV datasets to reveal the influence of HLA genes. By considering how a virus mutates throughout the course of its infection, the relevance of the host genetic background can be considered. A recent study analysed a large number of existing HIV datasets, using host HLA class I genotypes and pathogen reverse transcriptase and protease sequences and modelling the processes of within-host evolution of the virus \(^{103}\). Known allele–epitope combinations (for example, escape at codons 173/195 of reverse transcriptase when paired with the *HLA-B*/*51* allele) were replicated, but many epitopes highlighted to be HLA associated were not previously reported. Deeper analysis of viral sequence subtypes showed that differential selection pressures between the same HLA allele (*HLA-B*/*58*) and the different viral subtypes occurred at a reverse transcriptase codon that differed between the two subtypes. *HLA-B*/*58* selection pressure may have consolidated the divergence in viral sequence subtypes.

The selective pressure of other pathogens on human populations has been delineated by genome-wide analyses of natural selection. For example, for cholera, targets of positive selection were identified in a population from the Ganges River Delta, supporting strong selective pressure by the pathogen on innate immune pathways, including NF-κB and inflammasome signalling \(^{104}\). Studies of dual-pronged interrogation of host and infectious agent genetics in malaria have yet to be conducted at a large scale, but there have been projects to characterize heterogeneity in both *Plasmodium falciparum*, *Plasmodium knowlesi* and *Plasmodium malariae* and the malarial vector *Anopheles gambiense* at genetic and gene expression level \(^{105-107}\). These studies will be informative for drug resistance and vector control; for example, one study highlighted the need to collect vector genomic data to establish which strategies lead to insecticide resistance \(^{107}\).

Several major challenges remain for studies considering both host and pathogen genetic variation. Most notably, the multiple testing burden means that substantially larger sample sizes will be needed in future studies, although an important analytical strategy is dimensionality reduction, which takes advantage of genome-wide correlation between variants \(^{106}\). Additionally, understanding the impact of complex spatio-temporal and ecological scales requires careful acquisition of comprehensive metadata that can be used with genomic data \(^{106}\).

**Genetics of immune function phenotypes**

In the evaluation of any disease, accounting for phenotype heterogeneity is of paramount importance. The use of various -omics technologies is a strategy to address this (Fig. 3). For example, transcriptomic analysis in
sepsis has been able to stratify patients into different disease endotypes regardless of the source of infection. Crucially, these endotypes have been shown to have interaction effects with steroid treatment, with evidence that patients of a relatively immunocompetent endotype have increased mortality on being given steroids. This finding highlights the need for accurate phenotyping in future clinical trials where the effect of treatment may be diluted by heterogeneity in patient cohorts, and such improved phenotyping will also enable more successful application of GWAS.

Thus, aside from clinical disease phenotypes, innovative approaches to examining infectious disease genetics have investigated intermediate phenotypes that may be relevant to the disease. For example, differential blood cell types are under substantial genetic influence. Furthermore, levels of immunoglobulin are also genetically influenced not only by genetic variants in IGH genes and related immunoglobulin genes and the MHC but also by RUNX3 variants altering isoform shifting and FCGR2B variants impacting IgG binding to its receptor. These approaches are now also being translated in the context of antigen specificity, including the role of genetic variation in agent-specific antibody responses, with a role for MHC identified in various responses, including anti-Epstein–Barr virus and anti-rubella virus IgG levels. It will be interesting to determine what findings these markers of previous exposure may contribute in the context of acute infection risk in the future.

In a different attempt to examine the role of genetic variants in controlling the human immune system, Roederer et al. performed detailed immunophenotyping in 669 female twins. The top 151 heritable immune traits from extensive flow cytometry were analysed. For example, FCGR2A variants were associated with multiple immune phenotypes such as expression levels of various T cell markers such as CD27 or CD161, and FCGR2A has been associated with HIV infection progression.

The use of microbial stimuli to produce intermediate phenotypes has also led to interesting insights. One study used 528 lymphoblastoid cell lines and subjected them to challenges with eight different kinds of pathogens, selected carefully for their high health-care burden and the wide range of host responses elicited against them. Genome-wide association tests revealed 17 significant SNPs; for example, a SNP in ZBTB20 (which functions as a transcriptional repressor) was associated with both increased salmonella-induced pyroptosis and chlamydia replication. A PheWAS approach was further used with clinical phenotypes from the eMERGE dataset, which highlighted the same SNP, this time relevant in viral hepatitis.

Ultimately, the link from gene to biological function requires characterization of candidate loci for mechanistic understanding of genetic contributions. The vast majority of loci found in GWAS reside in non-coding regions, implicating gene regulation in disease pathogenesis. Much effort in recent years has therefore been focused on understanding the genetic basis of various forms of molecular traits (molecular quantitative trait locus (QTL) mapping). Such associations are often highly context specific, dependent, for example, on cell type and activation state, with effects specific to exposure to particular pathogens, and dependent on the disease state; for example, in patients with sepsis due to community acquired pneumonia. This characterization of molecular traits can inform GWAS interpretation; for example, in a GWAS of non-typoidal Salmonella bacteraemia, a SNP acting as an eQTL for STAT4 was found, with its effect reported to be a reduction in IFNγ production by natural killer cells. Genetic modulators of gene expression may also operate at the level of alternative splicing, including after stimulation by influenza A virus, with evidence that genetic regulation of isoform use was mostly distinct from that of gene expression. Immune cell-specific protein QTLs have also been demonstrated, including effects on expression levels of HLA-DRB1 in specific cell types that may be relevant in diseases such as sepsis, where reduced HLA-DR expression is known to correlate with poor outcome. A further study mapped cytokine production to genetic variants in response to microbial stimuli, identifying QTLs involving pathways containing pattern recognition receptors, cytokine and complement inhibitors, and the kallikrein system.

The maturation of technologies has seen multi-omics approaches applied to increasing numbers of diseases to identify and link genomic alterations to biological mechanisms. The infectious disease field has seen recent examples emerge: the combination of the plasma proteome, metabolome and lipidome and peripheral blood mononuclear cell transcriptome in analysing Ebola virus disease pathogenesis is one such example. Macrophages and neutrophils were found to be particularly relevant cell types. Although this study itself did not tie the findings to the host genetic background, it is Nevertheless an important example of the use of multi-omics functional characterization, and it is likely that we will soon be seeing the linking of multi-omics data back to genetic background.

**Genetics of responses to vaccinations**

As vast amounts of data continue to accrue on the relationship between the human host, its immune responses and infectious agents, the question as to how these data are best used is timely. Arguably, vaccination has offered one of the most effective public health efforts of modern medicine and is the most cost-effective opportunity to substantially reduce the incidence of many important infections, such as malaria, HIV infection and TB. Thus, given the large amount of evidence of human genetics influencing both non-specific and specific immune responses and susceptibility to disease, it is curious that there is still a relative lack of investigations looking at the effect of host genetics on vaccine response, especially as the study of responses against conserved epitopes offers a means to control for pathogen diversity. Here, we review the state of the field in terms of the genetics of response to vaccination and its potential utility.

The vaccine targeting hepatitis B virus (HBV) remains the most studied, with many genetic associations identified in the MHC region, and there is
controversy over whether it is HLA-DR, HLA-DP or class III MHC (involving the C4A gene) that plays independent or linked roles\(^{13-17}\). It is clear, however, that while HLA-DRB1*03 and HLA-DRB1*07 are linked to lower antibody responses, the opposite is true of HLA-DRB1*01, HLA-DRB1*13 and HLA-DRB1*15 (REFS\(^{134,135,136}\)), and the associations between HLA-DP and vaccine response against HBV are robust, as they are for other clinical phenotypes associated with the infection, particularly in Asian populations.

Rubella vaccination has seen similar HLA associations, although evidence is again conflicting. The HLA-DPB1 locus has been associated with rubella vaccination antibody responses\(^{139,140}\), but SNPs in the region have been associated in opposite directions, highlighting the need for further replication and functional studies to identify plausible biological mechanisms. Increased antibody responses to measles vaccination has been associated with the HLA-DQA1*0201 allele\(^{141}\).

Although a twin study has shown that the contribution of HLA genes is relatively small compared with non-HLA genes in antibody-inducing vaccines\(^{142}\), there has been relatively little robust and replicated evidence of what the non-HLA genes may be. A recent two-stage GWAS looking at data from more than 3,600 children revealed both HLA and non-HLA associations on studying a combination of capsular group C meningococcal, Haemophilus influenzae type b and tetanus toxoid vaccines\(^{143}\). This study found SNPs in the locus coding for signal-regulatory proteins SIRPA, SIRPB and SIRPG were associated with greater antibody titres in group C meningococcal vaccines. Notably, the association was present only for serum bactericidal antibodies (functional antibodies as assessed by rabbit/human complement assays, and not with total meningitis C-specific IgG), emphasizing the need for caution about the agent-specific readout used. The study also identified four HLA class II alleles associated with tetanus toxoid vaccine IgG concentrations, with the lead SNP being an eQTL for HLA-DRB1 and HLA-DRB5 (REF\(^{144}\)).

Inborn errors of immunity have been reported to occasionally result in life-threatening disease following administration of live attenuated vaccines, for example, live poliovirus vaccine (vaccine-associated paralytic polio in patients with agammaglobulinemia\(^{145}\)), whereas defects in interferon immunity may result in severe illness following administration of the yellow fever or MMR vaccine (measles strain) (IFNAR1 deficiencies, respectively) (FIG. 1c).

Overall, current work highlights that new cohorts and mechanistic studies to further explore the effect of host genetics on vaccine response are needed to address this important area.

**Genetic strategies for a new disease: COVID-19**

Emerging infections such as SARS-CoV-2 pose a major threat to human health, and the early observation of striking heterogeneity in clinical presentations and outcomes following infection has highlighted the ongoing need to apply genetic and genomic approaches to understand drivers of individual susceptibility, severity and outcomes to this disease. Only a small minority of those infected with SARS-CoV-2 develop severe disease, with recognized risk factors including older age, male sex and co-morbidities, including obesity\(^{150}\). Genetic factors are hypothesized to contribute to the observed heterogeneity in response, and early reports such as from TwinsUK indicate relatively high heritability estimates for self-reported COVID-19 symptoms such as anosmia\(^{151}\). Genetic approaches provide the opportunity to gain novel insights into disease pathogenesis through the identification of specific disease associations involving particular genes or pathways, to define novel drug targets or to develop personalized medicine approaches with early intervention or therapy tailored to the individual. The opportunity to identify repurposing opportunities for approved medications based on shared genetic associations across traits is a particularly important consideration in a pandemic situation such as COVID-19, where, while not based on genetic evidence, the RECOVERY trial has demonstrated the therapeutic utility of existing drugs such as dexamethasone for treatment of severe disease\(^{152}\).

The international response to answering the question of whether human genetics alters susceptibility to or outcome from SARS-CoV-2 infection is an exemplar of modern genomic collaboration. The establishment of large appropriately phenotyped cohorts for GWAS in a pandemic situation has been enabled by prospective collection through existing or hibernating studies, rapid deployment of new studies and leveraging existing population biobank studies with large numbers of already genotyped individuals. Monthly data releases from the UK Biobank of COVID-19 results together with outcome and phenotypic information early in the pandemic highlight the opportunity provided by such population biobanks, while application of meta-analysis methods across cohorts will maximize power for genetic discovery. International collaborative efforts are key to facilitate genetic studies in infectious diseases such as COVID-19. For example, the COVID-19 Host Genetics Initiative\(^{153}\) was rapidly established to support sharing of relevant tools and data with an emphasis on GWAS. The COVID Human Genetic Effort\(^{154}\), by contrast, is focused on identifying monogenic cases with rare, highly penetrant mutations through analysis of young patients (less than 50 years of age) who were previously well and developed life-threatening disease, as well as those naturally resistant despite repeated exposure. Integrative analysis, considering both host and viral genetic variation, will likely be highly informative, although early reports suggest that viral genetic variation did not significantly affect outcomes in COVID-19 (REF\(^{155}\)).

An early exemplar of GWAS during the COVID-19 pandemic is a study by the Severe Covid-19 GWAS Group, who conducted a GWAS of 1,980 patients with severe disease (hospitalized with respiratory failure requiring oxygenation or mechanical ventilation) in Italian and Spanish epicentres of the pandemic versus population controls. The European meta-analysis revealed genome-wide significant associations at 3p21.31 and 9q34.2 (REF\(^{156}\)). The limitations of self-reported data notwithstanding, a large GWAS of more than 1.05 million
individually from the 23andMe research platform replicated the associations at these two loci for disease severity, while the GenOMICC investigators recruited patients with COVID-19 from intensive care units only and robustly reproduced the 3p21.31 signal in addition to finding associations at other loci, including ones with genes of well-known antiviral function such as 12q24.13 (OAS1, OAS2 and OAS3) and 21q22.1 (IFNAR2).

Despite the important achievements, these results exemplify current challenges in GWAS biology. First, the 3p21.31 association, which was also supported by preliminary results from the COVID-19 Host Genetics Initiative, spans multiple genes, including SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6 and XCR1, but will require careful functional genomic and other annotations to establish the causal genes underlying the genomic signal, as non-coding variants may modulate gene expression and other regulatory mechanisms at a distance. Intriguingly, there is evidence that the risk haplotype, spanning ~50 kb at 3p21.31, was inherited from Neanderthals and shows substantial variation in frequency between populations, being carried by ~50% of people in South Asia, while being almost absent in East Asia, suggesting it may have been affected by selection in the past. Second, the association at 9q34.2 coincides with the ABO blood group locus, and increased risk with blood group A was shown, but the mechanism for this association remains unclear and will require further study especially as this locus is frequently a site of type I error resulting from population stratification. Despite the apparent lack of significant signal of association at the locus encoding HLA genes in these initial GWAS of COVID-19, there is considerable interest in the importance of such variation in the response to infection, with evidence, for example, of differing capacity between HLA alleles in presentation of highly conserved SARS-CoV-2 peptides to immune cells.

There is also early evidence from the COVID Human Genetic Effort to support the role of rare variants in the risk of severe disease on the basis of a study of 659 patients (0.1–99 years of age) hospitalized with critical disease due to COVID-19 and requiring mechanical ventilation or organ support in intensive care units. Following genome and exome sequencing, the study authors tested the hypothesis that inborn errors of TLR3- and IRF7-dependent type I interferon immunity contribute to critical disease risk, analysing 13 genomic regions comprising loci previously shown to be mutated in critical influenza pneumonia and connected loci mutated in patients with other viral illnesses. Significant enrichment of rare variants predicted to result in LOF was found at these loci relative to people with asymptomatic or benign infection, with experimental validation including demonstration of inborn errors at eight loci in up to 23 patients (3.5%) of different ages (17–77 years) and population ancestries. The importance of type I interferons in protective immunity against SARS-CoV-2 was underlined by an accompanying article showing evidence for a high level of neutralizing autoantibodies to type I interferons in other patients with critical COVID-19, together highlighting the opportunities for therapeutic intervention based on screening at-risk individuals and developing targeted interventions. Indeed, in the GenOMICC study described above, beyond uncovering key potential genetic contributions to COVID-19 severity, the group further used Mendelian randomization to demonstrate a link between low IFNAR2 expression and high TYK2 expression and life-threatening disease, the former further evidencing the importance of type I interferons in COVID-19 pathogenesis. Altogether, the emerging insights from this fast-moving field of research exemplify the complex nature of the genetic architecture of susceptibility to infectious disease that may have relevance not only for the agent inflicting the largest pandemic of our generation but also potentially for a range of other infections afflicting our public health worldwide.

Conclusions and perspectives

A huge wealth of data is becoming available that confirms the influence of our host genetic variation on how we become susceptible to and respond to diseases caused by infectious agents. Although traditional case–control GWAS have yielded many insights, a large amount of information is also coming from other methods, such as intermediate phenotype mapping and multi-omics approaches. Issues of heterogeneity in host and pathogen genetic diversity and difficulty in arriving at consensus for disease case definition both hamper power in GWAS approaches. However, as recent studies emphasize, these problems are not insurmountable with appropriate power and will be aided by greater precision in defining disease phenotypes and application in a range of populations.

The challenge now comes with bringing all of these findings and approaches together to translate them into a clinical benefit. Moving forwards, it will be critical to consider context specificity. The cellular studies and vaccine responses emphasize the importance of considering pathogen-specific immune responses in the context of both time and appropriate sample type. The extent and nature of pathogen diversity is a key context, with large gaps remaining in our understanding. Our ability to leverage genetic information will also improve as we further appreciate the overlaps between monogenic syndromes with a high infection risk and the more common infectious diseases with similar pathogens. All of these opportunities are likely to come as data from cohorts such as the UK Biobank and the China Kadoorie Biobank mature.

The clinical benefit of the large body of genetic work has already been demonstrated through associations translated to clinical implementation, such as abacavir hypersensitivity eliminated by screening for HLA-B*5701 and HCV treatment influenced by IL28B. We should be inspired by these lessons learnt as we move forwards into an era where large-scale genomics may help predict our risks of disease or facilitate future vaccine development and deployment through an understanding of genetic diversity at both the individual level and the population level to enable more tailored application in a precision medicine approach that maximizes effectiveness for a given person or population group.

Published online 4 December 2020
1. Mühlemann, B. et al. Ancient hepatitis B viruses from the Bronze Age to the medieval period. Nature 557, 418–425 (2018).
2. Quintana-Murci, L. Human immunity through the lens of evolutionary genetics. Cell 177, 184–199 (2019).
3. Karlsson, E. K., Kwiatkowski, D. P. & Sabaté, P. C. Natural selection and infectious disease in human populations. Nat. Rev. Genet. 15, 379–393 (2014).
4. Stanaway, J. D. et al. Global, regional, and national burden of bacterial sepsis in children in 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. Lancet 392, 1923–1994 (2018).
5. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269 (2020).
6. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China. Nature 579, 272–735 (2020).
7. Sørensen, T. I. A., Nielsen, G. G., Andersen, P. K. & Jernigan, R. B. An expanded view of complex traits: from polygenic to omnigenic. Nature Rev. Genet. 21, 47–58 (2020).
8. Chapman, S. J. & Hill, A. V. S. Human genetic epidemiology. Annu. Rev. Genom. Evol. 1, 13–40 (2010).
9. Claussnitzer, M. et al. A brief history of human disease genetics. Nat. Rev. Genet. 19, 461–472 (2018).
10. Young, A. I. Solving the missing heritability problem. PLoS Genet. 15, e1008222 (2019).
11. Clausonstorff, M. et al. A brief history of human disease genetics. Nat. Rev. Genet. 19, 175–188 (2018).
12. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: toward a synthesis of the germ and genetic theories. Annu. Rev. Pathol. Mech. Dis. https://doi.org/10.1146/annurev-pathmol-021920-091016 (2020).
13. Yang, Y. I. Solving the missing heritability problem. PLoS Genet. 15, e1008222 (2019).
14. Nabirotchkin, S. et al. Next-generation drug repurposing using human genetics and network biology. Project shapes medical genetics in Africa. Nature 517, 527–532 (2015).
15. Ge, D. et al. Genetic variation in IL28B predicts hepatitis C virus-induced liver cirrhosis: a meta-analysis. Nat. Genet. 43, 105–109 (2011).
16. Key, F. M., Teixeira, J. C., de Filippo, C. & Andrés, A. M. The role of structural variation of red blood cell invasion receptors. Nature 523, 1140–1152 (2015).
17. Thomas, D. W. et al. HLA variation and disease. Nat. Rev. Immunol. 5, 661–666 (2005).
18. Ramsuran, V. et al. Elevated HLA-A expression impairs HIV control through inhibition of NK2CA-expressing cells. Science 359, 86–90 (2018).
19. Deng, H. K. et al. Identification of a major co-receptor for primary isolates of H1N1. Nature 381, 405–411 (1996).
20. Leffler, E. M. et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 340, 1578–1582 (2013).
21. Thomas, D. W. et al. HLA diversity through structural variation of red blood cell invasion receptors. Science 356, 1140–1152 (2017).
22. Thomas, D. W. et al. HLA diversity through structural variation of red blood cell invasion receptors. Science 356, 1140–1152 (2017).
23. Key, F. M., Teixeira, J. C., de Filippo, C. & Andrés, A. M. The role of structural variation of red blood cell invasion receptors. Nature 523, 1140–1152 (2015).
24. Thomas, D. W. et al. HLA diversity through structural variation of red blood cell invasion receptors. Nature 523, 1140–1152 (2015).
25. Thomas, D. W. et al. HLA diversity through structural variation of red blood cell invasion receptors. Nature 523, 1140–1152 (2015).
26. Duggal. P. et al. Genome-wide association study of spontaneous resolution of hepatitis C virus infection from multiple cohorts. Am. Intern. Med. 158, 235–245 (2013).
27. Westra, J. D. et al. Multi-ancestry genome-wide association study of spontaneous clearance of hepatitis C virus. Gastroenterology 156, 1469–1475 (2019).
28. Duggal, P. et al. Genome-wide association study of spontaneous resolution of hepatitis C virus infection from multiple cohorts. Am. Intern. Med. 158, 235–245 (2013).
29. Westra, J. D. et al. Multi-ancestry genome-wide association study of spontaneous clearance of hepatitis C virus. Gastroenterology 156, 1469–1475 (2019).
30. Westra, J. D. et al. Multi-ancestry genome-wide association study of spontaneous clearance of hepatitis C virus. Gastroenterology 156, 1469–1475 (2019).
31. Ellison, M. A. et al. Next-generation whole genome sequencing improves imputation quality and detection of rare variant associations in admixed and Hispanic latino populations. PLoS Genet. 15, e1008500 (2019).
32. Ramsuran, V. et al. Elevated HLA-A expression impairs HIV control through inhibition of NK2CA-expressing cells. Science 359, 86–90 (2018).
33. Deng, H. K. et al. Identification of a major co-receptor for primary isolates of H1N1. Nature 381, 405–411 (1996).
34. Leffler, E. M. et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 340, 1578–1582 (2013).
35. Thomas, D. W. et al. HLA diversity through structural variation of red blood cell invasion receptors. Science 356, 1140–1152 (2017).
36. Thomas, D. W. et al. HLA diversity through structural variation of red blood cell invasion receptors. Nature 523, 1140–1152 (2015).
37. Thomas, D. W. et al. HLA diversity through structural variation of red blood cell invasion receptors. Nature 523, 1140–1152 (2015).
38. Jones, C. R., Flower, B. F., Barber, E., Simmons, B. & Cooke, G. S. Treatment optimisation for hepatitis C in the era of direct-acting antivirals. Liver Int. 36, 338–345 (2016).
39. Lewis, L. A. & Ram, S. Meningococcal disease and the complement system. Virulence 5, 88–126 (2014).
40. Schneider, G. et al. Hematological manifestations of diabetes and dysregulation: pathophysiological consequences, modern analysis, and clinical management. Mol. Immunol. 114, 299–311 (2019).
153. Lambert, N. D. et al. Polymorphisms in HLA-DPB1 Are associated with differences in rubella virus-specific humoral immunity after vaccination. *J. Infect. Dis.* 211, 898–905 (2015).

154. Ovsyannikova, I. G., Pankratz, V. S., Larrabee, B. R., Jacobson, R. M. & Poland, G. A. HLA genotypes and rubella vaccine immune response: additional evidence. *Virology* 472, 159–164 (2015).

155. Jacobson, R. M., Ovsyannikova, I. G., Vierkant, R. A., Shane Pankratz, V. & Poland, G. A. Human leukocyte antigen associations with humoral and cellular immunity following a second dose of measles-containing vaccine: persistence, dampening, and extinction of associations found after a first dose. *Sci. Rep.* 11, 17052 (2021).

156. Newport, J. M. et al. Genetic regulation of immune responses to vaccines in early life. *Genes Immun.* 5, 122–129 (2004).

157. O’Connor, D. et al. Common genetic variations associated with the persistence of immunity following childhood immunization. *Cell Rep.* 27, 2541–2555.e4 (2019).

This article shows how GWAS can be applied to study persistence of immunity following childhood vaccination.

158. Zeller, T. et al. Genetics and beyond — the transcriptome of human monocytes and disease susceptibility. *Hum. Genet.* 137, 2057–2070 (2019).

This article shows how rare IFNAR1 deficiency can lead to life-threatening complications following vaccination with live attenuated viruses.

159. Duncan, C. J. A. et al. Human IFNAR2 deficiency: lessons for antiviral immunity. *Sci. Transl. Med.* 7, 307ra145 (2015).

160. Burns, C. et al. A novel presentation of homozygous loss-of-function STAT1 mutations in an infant with hyperinflammation — a case report and review of the literature. *J. Allergy Clin. Immunol. Pract.* 4, 777–779 (2016).

161. Moens, L. et al. A novel kindred with inherited STAT2 deficiency and severe viral illness. *J. Allergy Clin. Immunol.* 139, 165–170 (2017).

162. Docherty, A. B. et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* 369, m1985 (2020).

163. Williams, F. M. et al. Self-reported symptoms of covid-19 19 and relationships most predictive of SARS-CoV-2 infection, are heritable. Preprint at medRxiv https://doi.org/10.1101/2020.04.22.20083422 (2020).

164. RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with covid-19 — preliminary report. *N. Engl. J. Med.* 385, 224–232 (2020).

165. 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. *Eur. J. Hum. Genet.* 28, 715–718 (2020).

166. Casanova, J. L. et al. A global effort to define the human genetics of protective immunity to SARS-CoV-2 infection. *Cell* 181, 1194–1199 (2020).

167. Zhou, X. et al. Host factors related to the clinical outcome of COVID-19. *Nature* https://doi.org/10.1038/s41586-020-2355-0 (2020).

168. Severe Covid-19 GWAS Group. Genomewide association study of severe covid-19 with respiratory failure. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMoa201456 (2020).

This study illustrates the early successful application of GWAS in the COVID-19 pandemic, identifying the 3p21.31 locus and a potential role for the ACE2 receptor and angiotension system.

169. Shelton, J. F. et al. Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility. Preprint at medRxiv https://doi.org/10.1101/2020.04.20.20088518 (2020).

170. Paro-Castineira, E. et al. Genetic mechanisms of critical illness in Covid-19. Preprint at medRxiv https://doi.org/10.1101/2020.04.29.20090048 (2020).

171. Smemo, S. et al. Obesity-association variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375 (2014).

172. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* https://doi.org/10.1038/s41586-020-2819-3 (2020).

173. Greenberg, A. et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* 370, re3780 (2021).

174. Bastard, P. et al. Auto-antibodies against type I IFNs in patients with life-threatening COVID-19. *Science* 368, eabd4585 (2020).

175. Palladino, T. A. et al. A genome-wide association study of major determinants for host control of HIV-1. *Science* 317, 946–947 (2007).

176. Korbel, J. O. et al. Polymorphisms in HLA-DPB1 genetic and non-genetic associations with COVID-19. *Immunol. Rev.* 2020. e1003515 (2020).

177. Hyundai, J. & Winternitz, J. Advances in the evolutionary genomics of people with complement deficiencies and patients who have undergone spleenectomy. *Clin. Microbiol. Rev.* 23, 740–780 (2010).

178. Tange, T. & XLP, S. Clinical features and molecular etiology due to mutations in SH2D1A encoding SAP in Japanese. *J. Clin. Immunol.* 34, 772–779 (2014).

179. Nahum, A. Chronic mucocutaneous candidiasis: a spectrum of genetic disorders. *LymphoSign J.* https://doi.org/10.14785/lymphosgn.2017-0011 (2017).

180. Lanterriere, F. et al. Deep dermatophytosis and inherited CARD9 deficiency. *N. Engl. J. Med.* 369, 1704–1714 (2013).

181. Parham, P. & Moffett, A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat. Rev. Immunol.* 13, 133–144 (2013).

182. Dawkins, R. et al. Genomics of the major histocompatibility complex: haplotypes, duplication, retroversions and disease. *Immunol. Rev.* 167, 275–304 (1999).

183. Travisولد, J. & Kout, J. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics. Hum. Genet.* 14, 301–323 (2013).

184. Parham, P. & Ohta, T. Population biology of antigen presentation by MHC class I molecules. *Science* 272, 67–74 (1996).

185. Hedrick, P. W. Pathogen resistance and genetic variation at MHC loci. *Evolution 56*, 1902–1908 (2002).

186. Radwan, J., Babik, W., Kaufman, J., Lenz, T. L. & Winternitz, J. Advances in the evolutionary understanding of MHC polymorphism. *Trends Genet.* 36, 298–311 (2020).

Acknowledgements Funding for this work was received from the UK National Institute for Health Research Oxford Biomedical Research Centre (to J.K.C. and A.M.), a Wellcome Trust Investigator Award (204969/Z/16/Z to J.J.K.C.), the Chinese Academy of Medical Sciences Innovation Fund for Medical Science (grant number 2018IM2-002 to J.J.C.), the Croucher Foundation (to A.J.K.) and a Wellcome Trust Grant (200532/Z/09/Z) to core facilities at the Wellcome Centre for Human Genetics.

Author contributions The authors contributed to all aspects of the article.

Competing interests The authors declare no competing interests.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**RELATED LINKS**

COVID-19 Host Genetics Initiative: https://www.covid19hg.org/

COVID Human Genetic Efforts: https://www.covid19gcg.org

**UK Biobank of COVID-19**: https://www.ukbiobank.ac.uk/

© Springer Nature Limited 2020