HIV-1 and Human Genetic Variation

Paul J McLaren[1,2] and Jacques Fellay[3,4,5†]

[1]National HIV and Retrovirology Laboratory at JC Wilt Infectious Diseases Research Centre, National Microbiology Laboratories, Public Health Agency of Canada, Winnipeg, Canada

[2]Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Canada

[3]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[4]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[5]Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[†]E-Mail: jacques.fellay@epfl.ch

Abstract

Over the past four decades, research on the natural history of Human Immunodeficiency Virus (HIV) infection has described how HIV wreaks havoc on human immunity and causes Acquired Immunodeficiency Syndrome (AIDS). HIV host genomic research, which aims to understand how human genetic variation affects our response to HIV infection, has progressed from early candidate gene studies to recent multi-omic efforts, benefiting from spectacular advances in sequencing technology and data science. In addition to invading cells and co-opting the host machinery for replication, HIV also stably integrates into our own genome. The study of the complex interactions between the human and retroviral genomes has improved our understanding of pathogenic mechanisms and suggested novel preventive and therapeutic approaches against HIV infection.

[H1] Introduction

HIV-1 is the human retrovirus responsible for the HIV/AIDS **[G]** pandemic, which has claimed more than 30 million lives over the past four decades. HIV infection continues to be a major global public health issue, with currently around 40 million people living with HIV (PLWH). Lifelong antiretroviral therapy (ART) has transformed the disease into a manageable chronic health condition. When available, ART enables PLWH to lead long and heatlhy lives, but there is still no effective vaccine and no cure.

Early in the pandemic, it became clear that the risk of HIV acquisition is highly variable across humans. Socioeconomic and behavioural factors played a central role in this variability, with some risk groups, such as intravenous drug users (IVDU) and men having sex with men (MSM), being disproportionately affected[1]. Still, even among the most highly-exposed individuals, a fraction remained HIV-negative[2,3]. Similarly, important differences in the natural course of HIV infection (such as time from infection to AIDS diagnosis and occurrence of opportunistic infections or malignancies) were only partially explained by known variables such as age and comorbidities[4]. Taken together, these clinical and epidemiological observations suggested a role for additional factors in the modulation of the individual response to HIV, including inherited variation in the genes and pathways involved in the retroviral life cycle and in innate or adaptative immunity against the infection.

HIV enters its main target cell, the CD4+ T lymphocyte, by binding to its receptor CD4 and to the co-receptor C-C chemokine receptor type 5 **[G]** (CCR5)[5]. This binding event triggers fusion of the viral and human cell membranes, initiating a complex intracellular life cycle that will lead to the production of new viruses (***Figure 1***). The natural immune response against HIV infection relies mostly on CD8+ T cells, also called cytotoxic T lymphocytes **[G]** (CTL). Upon primary infection, intense HIV replication results in very high plasma viral load, measured as copies of the HIV RNA genome per ml of plasma, which is then partly controlled by the specific CD8+ T cell response. The very diverse Human Leukocyte Antigen **[G]** (HLA) class I molecules have a central role in this immune response, by presenting small viral fragments, called epitopes **[G]**, at the surface of infected cells. Recognition of these epitopes by CTL leads to the elimination of HIV-infected cells. A more efficient immune response is linked to lower viral load during the chronic phase of an untreated infection and to slower disease progression, though it is unable to eliminate the virus[6].

As a retrovirus, HIV can be described as a genomic pathogen. Indeed, it not only uses the molecular machinery of the infected cell for replication and dissemination, but it also has the remarkable capacity to integrate a DNA copy of its RNA genome into a host cell chromosome. By becoming part of the human genome, HIV can persist in long-term cellular reservoirs for decades, making it extremely challenging to develop therapeutic strategies resulting in complete eradication[7].

To better fight HIV infection, we must once again consider the old Delphic maxim: "know thyself". Because HIV is an expert at hijacking human cells and immunity, we have no choice but to improve our understanding of our inner machinery, starting with the most fundamental layer of biological information, the human genome. The exploration of human diversity at the DNA level, long hampered by technological limitations, was fueled by the development of new and more powerful tools over the past decades[8]. Thanks to progress in our understanding of human genetic diversity, in genotyping and sequencing technology, as well as in bioinformatics and data science, it became possible to search for genetic factors that modulate the individual response to HIV, including resistance and susceptibility to infection and natural history of the disease in PLWH[9].

In this Review, we first present an overview of the technological and conceptual developments that fueled HIV host genomic research. Then we describe the major genetic factors modulating the natural history of HIV infection – in the HLA class I region and the *CCR5* locus. Next, we highlight the recent convergence of human and HIV genomics, which allows longitudinal analyses of host-pathogen genetic interactions. Finally, we explain how genomic knowledge is poised to have a positive impact on people living with HIV (PLWH), notably through pharmacogenomic interventions and stratification of care based on polygenic risk scores **[G]** before discussing short and long-term perspectives for translational research and clinical applications of human genomics in the HIV field.

[H1] Discovering host genetic variants

The search for human genetic differences that have an impact on HIV-related outcomes was first motivated by clinical observations – the striking variability in individual trajectories of patients in the absence of treatment. It was further propelled by a desire to uncover fundamental physiopathological mechanisms, which can be uncovered by the careful exploration of genomic variants and their impact on host and viral molecular processes.

[H2] Candidate gene studies. In the candidate gene approach, population-level associations are sought between HIV-related phenotypes and specific genetic variants in genes that have been selected based on previous biological knowledge or functional work. The selected variants are typically typed using targeted genotyping assays or Sanger sequencing of the region of interest. This framework was first applied to HIV host genetics in the early 1990's, in analyses of allelic variants in genes known or suspected to play a role in HIV pathogenesis or antiretroviral immune response. Therefore, genetic associations were reported in two broad categories: genes coding for proteins involved in the HIV life cycle (such as *PPIA*[10] and *TSG101*[11]); and immune-related genes encoding molecules implicated in innate and adaptive immune pathways (such as *MBL2*[12] and *TLR9*[13]) as well as in specific antiretroviral defense mechanisms (such as *APOBEC3G*[14] and

TRIM5[15]). Dozens of genes were tested in multiple cohorts. Unfortunately, as has been the case in the broader field of human genetics, most reported associations turned out to be false positives, notably owing to the small sizes of the studied cohorts, population stratification **[G]** and lack of correction for multiple testing. Replication attempts in larger cohorts, where these factors could be better controlled, showed no association for the vast majority of variants[16–18]. In fact, only two major discoveries remain from the candidate gene era: the protective effect of a homozygous 32 bp deletion in *CCR5* (CCR5Δ32) against HIV acquisition[19–21]; and the modulating effect of HLA alleles on HIV progression **[G]** for which early studies, largely in MSM of European ancestry in the United States, noted a strong impact of the *HLA-B* alleles B*57 and B*27 on delaying time to AIDS onset[22].

[H2] Genome-wide association studies. Advances in genotyping and sequencing technologies progressively transformed human genetic analyses during the first decade of this century. In particular, the commercial availability of genome-wide genotyping arrays marked the beginning of the era of genome-wide association studies (GWAS). The principle of a GWAS is to simultaneously test very large numbers of genetic variants throughout the genome for potential associations with a phenotype of interest[23]. This truly agnostic approach finally allowed for a more comprehensive exploration of the human genome. To date, most GWAS have been based on genotyping of single nucleotide polymorphisms (SNPs) followed by imputation, a process that leverages the linkage disequilibrium property of the human genome to statistically infer the genotypes that are not directly measured. This approach allows near-comprehensive testing of common variants (that is, variants with a minor allele frequency of >1%) in most human populations[24].

The first GWAS of any infectious disease focused on the level of detectable viral genetic material in the blood of untreated, chronically infected individuals during the period of HIV latency **[G]**[25]. This phenotype, known as setpoint viral load **[G]** (spVL), was selected because of its relative ease of measurement and its known correlation to rate of progression to AIDS[26] and transmission potential[27]. The spectrum of interrogated variants was limited by early DNA genotyping arrays, yet genome-wide significant associations were identified in the HLA class I region, the most polymorphic locus in the human genome, known to have a crucial role in the modulation of T cell immunity (see ‘HLA variation in HIV control’, below). These findings were soon validated and expanded by other GWAS, performed in independent cohorts, which demonstrated that the genetic architecture **[G]** of HIV spVL is comparable between the general population of PLWH[16,28–30] and a particular group of individuals able to maintain low viral loads for prolonged periods of time in the absence of antiretroviral therapy (ART), the so-called HIV controllers **[G]** [17,31]. The absence of specific genetic factors explaining the HIV controller phenotype was a disappointment in terms of potential therapeutic development. It is, however, consistent with what has been found for many complex human traits and diseases –that individuals at the extremes of the phenotypic distribution are more likely to carry multiple

common variants with weak effects rather than rare, high-impact variants[32]. Beyond genotyping, a single exome sequencing study has been published so far in the HIV field[33], which also indicated that rare coding variants with large effect sizes are unlikely to make a major contribution to host control of HIV infection.

GWAS were less successful in the search for determinants of HIV resistance, with no definitive evidence found of human genetic polymorphims conferring altered susceptibility to HIV, apart from *CCR5* variation[18,34]. However, recent genome sequencing studies of extreme exposure phenotypes[35] have shown promising associations in *CD101,* a gene encoding an immunoglobulin superfamily member implicated in regulatory T cell function[36], and *UBE2V1*, which encodes a ubiquitin-conjugating enzyme involved in pro-inflammatory cytokine expression[37,38] that associates with the HIV restriction factor **[G]** *TRIM5-*α[38]. Although both *CD101* and *UBE2V1* are plausible candidates, further functional studies are required to validate their role in HIV susceptibility. Finally, analyses of GWAS data provide evidence for residual heritability owing to additive genetic effects beyond *CCR5*[18] and genetic overlap with behavioral and socioeconomic traits[39]. These results suggest larger genomic studies of HIV acquisition may identify additional loci that impact susceptibility and warrant further investigation, potentially in large biobanks.

A number of intrinsic limitations means that it remains difficut to investigate the genetic mechanisms potentially involved in HIV resistance. For example, sample sizes are usually small (in the tens or hundreds) because studies need to be performed on highly-exposed, yet uninfected, individuals, such as patients with hemophilia exposed to HIV through contaminated blood products[34], sex workers in hyper-endemic areas[40], or serodiscordant couples (stable heterosexual couples with one partner HIV-infected and the other partner HIV-seronegative at enrollment)[29]. Frailty (or survival) bias is a limitation in cross-sectional studies of HIV cohorts with long-term follow-up, as these cohorts are enriched for genetic factors protecting against HIV disease progression. Another limitation is misclassification bias in studies comparing the genomes of HIV-infected patients to unselected controls from the general population, in which most individuals are in fact susceptible to HIV infection[18]. The identification of additional genetic determinants of individual susceptibility to HIV infection will require increased sample sizes (ideally in the thousands), as well as the use of sequencing approaches to characterize the rare functional variants that are not interrogated in studies based on genotyping arrays.

[H1] HLA variation in HIV control

[H2] The HLA locus in infectious diseases. The human major histocompatability complex (MHC) located on chromosome 6 is one of the most genetically diverse loci in the genome[41]. The extended MHC occupies ~7.6Mb of the human genome[42] and encodes more than 400 genes, many of which are key mediators of the innate and adaptive immune responses. Within this locus, alleles at the classical class I (*HLA-A*, *-B*, *-C*) and

class II (*HLA-DR*, *-DQ*, *-DP*) genes have been associated with numerous autoimmune, inflammatory and infectious diseases (reviewed in [43,44]) with recent preprints demonstrating extensive disease associations in large biobanks from multiple populations[45,46]. In the context of infectious disease, class I HLA proteins present endogenous peptides on the surface of infected cells for recognition by CTLs, triggering development of an adaptive response. As discussed below, variability in epitope specificity of HLA proteins and expression levels of class I HLA alleles has a dramatic impact on progression of HIV disease.

[H2] Effects of amino acid variabilty encoded by HLA alleles. An individual's genotype at the class I *HLA* genes has been consistently demonstrated to be the major host genetic determinant of HIV spVL and rate of disease progression across geographic contexts and ancestries[17,22,47–50]. This observation was put in the genome-wide context by the first GWASs of HIV spVL[25] and HIV controllers[17] that exclusively identified SNPs in strong linkage disequilibrium with classical *HLA-B* alleles. Although array-based techniques for genotyping of DNA samples do not allow for direct resolution of classical HLA alleles, computational methods leveraging linkage disequilibrium structure between SNPs and sequence-based HLA types in reference populations allow for accurate imputation of classical HLA types from GWAS data[51]. Application of this technique to a sample of >6,000 PLWH of European ancestry underscored the dramatic effect of *HLA-B*57:01* on reducing viral load, which was, on average, ~0.8log10 RNA copies/ml lower in individuals carrying this allele [52]. This study also demonstrated strong associations at multiple other classical class I HLA alleles that had a range of effects, from decreasing spVL (B*57:01, B*27:05, B*13:02, B*14:02, C*06:02, C*08:02, C*12:02) to increasing spVL (B*07:02, B*08:01, C*07:01. C*07:02, C*04:01.

To better understand how functional variation in HLA class I proteins can impact HIV spVL, recent studies have tested variable amino acid positions within these proteins to fine-map the classical allele associations. In a GWAS performed by the International HIV Controllers study, this technique was applied to demonstrate that previously identified associations between HIV control and classical HLA alleles such as B*57:01, could be explained by variability across a small number of amino acid positions within the HLA-B protein[17]. The strongest effect was observed at position 97 of the protein, which accommodates six alternative amino acids, including valine, which is unique to B*57 haplotypes. A recent preprint describing the comprehensive analysis of the impact of HLA amino acid polymorphisms on spVL in a multi-ethnic sample of >12,000 PLWH identified 3 amino acid positions in HLA-B (positions 67, 97 and 156) and one in HLA-A (position 77) as independently associating with spVL[53]. The positions within HLA-B map to classical HLA alleles known to impact spVL, whereas the HLA-A position suggests that HLA-A functions independently of HLA-B. Interestingly, all four positions are located in the peptide binding groove of the respective HLA protein, supporting the hypothesis that epitope presentation is key for natural suppression of HIV replication.

Furthermore, there was no substantial evidence that the effects of these polymorphic positions differed across ancestry groups, suggesting biological relevance across global contexts.

Several mechanisms of action have been proposed to explain why different alleles of the same HLA gene have differential effects on HIV progression. Studies of epitope specificity have shown that certain protective alleles, including B*57:01 (which uniquely carries valine at position 97), and B*27:05 (which carries the protective cystine and asparagine residues at positions 67 and 97 respectively), drive compensatory mutations in the HIV genome leading to reduced viral fitness[54–56]. In addition to differential epitope specificity, the CTL effector function induced by epitope presentation has been implicated in HIV control, with CTLs in carriers of some protective HLA alleles exhibiting enhanced proliferative capacity and more polyfunctional responses[57–59].

[H2] Within-host diversity and epitope presentation. In addition to the impact of specific HLA class I alleles on HIV progression at the population level, the within-host diversity of HLA alleles may be important at the individual level. An early study looking at the impact of allele combinations revealed that maximum heterozygosity at HLA class I genes (that is, individuals carrying two different alleles at all three class I genes) associated with a reduced time to AIDS[47]. This observation was supported by a GWAS that showed that individuals carrying different HLA alleles at each class I gene had significantly lower viral load than homozygous individuals, even after accounting for the additive effect at each allele[60]. This heterozygote advantage likely comes from the ability to present multiple HIV epitopes, supporting the hypothesis that breadth of presentation is beneficial in preventing HIV progression. To further test this hypothesis, a recent *in silico* study used novel algorithms to predict the binding affinity of all possible 9-mer peptides in the HIV proteome to HLA proteins encoded by the different class I alleles[61]. Coupling these predicted affinities to clinical and genetic data demonstrated that spVL was negatively correlated with the breadth of the peptide repertoire bound by an individual's HLA protein isoforms. Moreover, HLA-B isoforms had the largest predicted breadth of epitope recognition and conferred the strongest reduction of viral load (***Figure 2A***). However, quantity of epitopes alone is unlikely to fully explain the protective capacity of an individual's HLA alleles, as subsets of epitopes that are uniquely presented by protective HLA isoforms explained more of the observed variance in spVL than the entire predicted set. This observation is further supported by an *in silico* and functional study that demonstrated that HIV epitopes that encode structurally important residues are preferentially targeted by protective HLA isoforms and associate with elite control of replication[62]. Thus, the quantity and quality of HIV epitopes presented by combinations of HLA isoforms are the key drivers of spVL.

[H2] Non-classical effects of HLA variation. In addition to the classical effects of *HLA* genes on peptide presentation, several studies have suggested that non-classical effects may play a part in limiting HIV replication *in vivo*. In particular, the variable expression levels of classical *HLA-C* alleles has been linked to HIV control, with those expressed at high levels conferring protection against disease progression[63]. This effect has been observed across ancestries and has been linked to the absence of a variable microRNA-148a (miR-148a) binding site in the 3′ untranslated region of *HLA-C* [64]. The proposed model suggests that mRNA from alleles lacking the miR-148a binding site escape suppression by miR-148a; as a consequence, proteins encoded by these alleles and loaded with HIV epitopes are expressed at higher levels on infected cells, allowing for greater rates of detection by CTLs[64] (***Figure 2B***). Similarly, proteins encoded by *HLA-A* alleles are also expressed at variable levels on the cell surface[65]. However, in contrast to *HLA-C*, *HLA-A* alleles expressed at high levels associate with poorer control of viral replication and with faster disease progression[66]. A combination of genetic and functional studies indicated that increased *HLA-A* expression levels correlated with higher viremia in a combined cohort of more than 9,000 people living with HIV from sub-Saharan Africa and the United States. It was proposed that this effect may be the result of enhanced production of the HLA class I signal peptide that regulates *HLA-E* expression, a hypothesis that was supported by a correlation between *HLA-A* expression and *HLE-A* expression among 58 healthy donors tested[66]. HLA-E is a ligand for Natural Killer Group Protein 2A (NKG2A) and their interaction results in strong inhibition of NK cell degranulation ***Figure 2C***). Thus, the enhanced production of the HLA class I signal peptide in individuals carrying highly expressing *HLA-A* alleles may lead to enhanced inhibition of immune responses in infected individuals, resulting in poorer clinical outcomes.

Finally, it has also been observed that the combination of HLA genotype and expression of particular killer cell immunoglobulin-like receptor **[G]** (KIR) proteins variably modulated HIV disease course[67]. The KIR proteins are a highly variable set of cell surface receptors expressed on natural killer (NK) cells (and some T cells) that, when engaged by their cognate receptors, either activate or inhibit NK cell-mediated killing (recently reviewed in[68]). In particular, the combination of the activating *KIR3DS1* allele with a set of *HLA-B* alleles that carry isoleucine in the the Bw4 epitope (Bw4-I80), is highly associated with HIV control[69]. Taken together, these results demonstrate the complex interplay between epitope presentation, HLA protein expression and NK inhibition.

[H1] CCR5 variation in HIV infection

[H2] CCR5Δ32 and resistance against HIV infection. Perhaps the most highly touted example of human genetic variability restricting infectious diseases is the observation that individuals carrying two copies of a loss-of-function variant in the gene encoding the cell receptor CCR5 are highly resistant to infection by HIV. CCR5 is a chemokine receptor expressed on the surface of multiple subsets of monocytes and lymphocytes,

including CD4+ T cells, the major HIV target cells **[G]**. At the earliest stages of infection, the HIV envelope protein gp120 binds CD4 and CCR5 on the cell surface resulting in fusion of the viral and host cell membranes and release of the viral genome into the target cell. The discovery that individuals who carry homozygous loss-of-function alleles at *CCR5* are resistant to infection was first made in a group of MSM that were multiply exposed to the virus but remained uninfected[19]. It was determined that these men all shared a 32 base pair deletion in the *CCR5* gene (the *CCR5Δ32* allele) that leads to the production of a non-functional protein, and the absence of functional CCR5 on the cell surface prevents HIV entering target cells (Figure 3a). The *CCR5Δ32* allele is observed at ~10% frequency in individuals of European ancestry (homozygosity occurs at a frequency of 1%), at a reduced frequency in southern Europeans compared to those in the north[70] and is not observed at an appreciable frequency in other continental populations. Compound heterozygotes (that is, individuals carrying one copy of *CCR5Δ32* and a second loss-of-function *CCR5* variant) are also resistant to infection, although these individuals are exceedingly rare[71].

The observation that individuals lacking CCR5 expression are resistant to HIV infection directly led to the development of the antiviral drug Maraviroc, a CCR5 antagonist[72], and also to the world's first ethically fraught attempt at human embryo engineering[73]. Perhaps most interestingly, bone marrow transplants between *CCR5Δ32* homozygous donors and HIV-infected recipients have resulted in the only two confirmed cases of long-term HIV cure[74,75]. Although promising, this effect has been difficult to replicate in engineered autologous stem cell models[76] and is unlikely to be scalable to the level necessary to stem the pandemic. Additionally, the protection is not absolute, as several confirmed cases of infection in *CCR5Δ32* homozygotes have been reported (reviewed in[77]), presumably by viruses that utilize the minor co-receptor CXCR4 or by dual-tropic viruses.

[H2] Associations between CCR5 variation and spontaneous HIV control. In addition to the impact of homozygosity on preventing infection, individuals with a single *CCR5Δ32* copy exhibit lower spVL and delayed disease progression compared to those with two functional copies[19,20,78], likely because the reduced levels of CCR5 protein on the cell surface lower the efficiency of HIV entry into target cells (***Figure 3a)***. The *CCR5* locus was also identified in GWAS, first in a study of ~2,500 people living with HIV in Europe[16] and then in an expanded set of 6,300 individuals from across the globe[52]. However, the *CCR5Δ32* allele was not directly assayed on the genotyping platforms used in these studies, thus only proxy SNPs were identified. In a combined analysis of GWAS data and direct *CCR5Δ32* genotyping, it was observed that the *CCR5Δ32* allele was not the most strongly associated variant in the region, suggesting that multiple independent genetic effects occur at this locus. Conditional analysis accounting for the effect of CCR5Δ32 showed that an additional marker, rs1015164, was also strongly associated with spVL. Functional analysis of this variant

showed that it regulates expression of an antisense long noncoding RNA (lncRNA) called *CCR5-AS*, which overlaps the *CCR5* gene[79]. This study further showed that increased expression of *CCR5-AS* resulted in increased *CCR5* expression because CCR5-AS interfered with Raly-mediated degradation **[G]** of CCR5 mRNA. Moreover, knockdown of *CCR5-AS* reduced the susceptibility of CD4+ T cells to HIV-1 infection *ex vivo* (***Figure 3b***). These results demonstrate that the clinical course of untreated HIV infection is directly influenced by the innate level of CCR5 expression within the infected individual. Whether additional functional polymorphisms in *CCR5* have similar effects remains an open question.

[H1] Host:pathogen genetic variation

[H2] Pathogen sequence variation as an indicator of host genetic pressure. Most studies performed so far in the field of host genetics focused on clinically defined outcomes, such as susceptibility to infection or disease progression. However, intermediate phenotypes have been shown to be very valuable in identifying subtle genetic association signals that are not always detectable using more complex clinical outcomes. A particularly promising intermediate phenotype, unique by its nature to infectious diseases, is variation in the pathogen genome (***Figure 4***). HIV is a highly variable virus that establishes a life-long infection. Therefore, it represents an ideal model to search for the potential effects of intra-host selective pressure on a human pathogen. While some of the variants observed in the HIV genomic sequence are present in the transmitted/founder virus **[G]**, another fraction is acquired during the course of the disease, resulting at least partially from selective pressure exerted by the host response to infection. Signs of host-driven selection are clearly visible in the HIV genome. In particular, specific variants have been described in key viral epitopes presented by HLA class I molecules and targeted by CTL responses[80]. Mutations have also been reported in regions targeted by KIR, suggesting escape from immune pressure by NK cells[81,82]. A non-negligible fraction of the HIV-1 genome (~12%) is under positive selection, but only about half of the positively selected sites map to canonical CD8+ T cell epitopes[83], indicating that additional host factors could be driving evolution in non-epitope sites.

[H2] Genome-to-genome studies. Computational approaches developed over the past decade have allowed more comprehensive analyses of the reciprocal genetic signals resulting from the host-pathogen interaction[84,85]. Joint analyses of human and HIV sequence variation start with the generation of large-scale genomic data from paired samples. The retroviral genome can be isolated and sequenced either as native RNA during replicative infection or as proviral DNA, integrated into the host genome, during latent infection. Human genomic information can be obtained using genotyping or sequencing technology. The principle of genome-to-genome (G2G) studies **[G]** is then to perform a systematic search for associations between human genetic polymorphisms and viral sequence variants, at the nucleotide or amino acid levels. Because of the very large number of models run in parallel – one GWAS for each viral variant – this approach requires stringent

correction for multiple testing. By mapping all interacting loci, G2G studies have the potential to uncover the most important genes and pathways involved in specific responses to infectious agents, thereby revealing novel diagnostic or therapeutic targets. In addition to identifying the sites of genetic interplay between virus and host, this study design makes it possible to estimate the biological consequences of such interactions and to estimate the relative impact of human and viral genetic variation on phenotypic outcomes, by assessing associations between human-driven escape at viral sites and a quantitative clinical phenotype. In spite of these promises, it must be acknowledged that G2G studies have not led, as of today, to the identification of novel HIV restriction factors in the human genome[86]. Future studies will require larger sample sizes to increase power, but also more diversity with a strong focus on the inclusion of PLWH of non-European ancestries.

Nevertheless, studies based on the combined analysis of host and pathogen genomic variation have already demonstrated their potential in other infections. In particular, the use of a similar study design in chronic hepatitis C virus (HCV) infection highlighted the evolutionary pressure exerted by both innate (interferon lambda) and acquired (HLA class II) immune defense mechanisms[87,88]. The intra-host evolution of DNA viruses can also be investigated using a G2G approach, as shown in a recent study that revealed several associations between human and Epstein-Barr virus (EBV) sequence variation in immunosuppressed PLWH[89].

[H1] HIV precision medicine

[H2] Expanding antiretroviral therapy to eradicate HIV. With the now accepted knowledge that PLWH who do not have detectable plasma viral loads cannot transmit the virus to others[90], the United Nations joint Programme on HIV/AIDS (UNAIDS) set an ambitious 90-90-90 target[91], where 90% of infected people know their status, 90% of those are on antiviral therapy and 90% of those are suppressing the virus below the level of detection. This aspirational treatment target would practically mean, given currently available technologies, that more than 34 million people would be on lifelong chemotherapy. Although this treatment as prevention **[G]** approach would undoubtedly result in decreases in transmission and dramatic increases in life expectancy for the HIV infected population, it also requires a deeper understanding of how human genetic variation relates to variability in drug toxicity and response to long-term therapy.

[H2] HIV pharmacogenetics. In addition to affecting HIV disease progression in untreated individuals, human genetic variability has also been implicated in modifying response to treatment. A major achievement in the fight against HIV has been the development of multiple, effective therapeutics that target several stages of the viral life cycle. These include; entry inhibitors, which prevent binding of the viral spike protein gp120 to host cell receptors and fusion of the virus with host cell membranes; nucleoside and non-nucleoside reverse transcriptase inhibitors, which prevent reverse transcription of the viral RNA genome into DNA; integrase inhibitors, which prevent integration of the viral DNA product into the host genome; and protease inhibitors, which prevent cleavage of viral polyproteins into their functional sub-units (***Figure 5***). For several classes of

anti-HIV therapy, human genetic variability is known to influence response to the drug, which in some cases leads to severe adverse events and treatment discontinuation[92]. Paradoxically, the *HLA-B* allele B*57:01, most notably associated with control of infection, also predisposes carriers to a severe hypersensitivity reaction to the nucleoside reverse transcriptase inhibitor Abacavir[93,94]; high specificity binding of Abacavir alters the binding pocket of HLA-B*57:01 and triggers reactivity to self-peptides[95]. Similarly, variants in genes encoding the drug-metabolizing enzymes CYP2B6[96–98], CYP2A6[99], CYP2C9[100], CYP2C19[100], CYP3A[101] and ABCC2[101] have all been associated with slow metabolization kinetics of their cognate drugs (***Table 1***), in some cases leading to drug accumulation in the brain, psychiatric complications and treatment stoppage[99]. The frequency of many of these polymorphisms varies depending on ancestral background, leading to reduced drug tolerance and therefore efficacy in some populations. For example, the *6 allele of *CYP2B6* (rs3745274), which results in slow metabolism of Efavirenz (EFV) and Nevirapine, two non-nucleoside reverse transcriptase inhibitors recommended for first-line use by the World Health Organization until recently, has a ~two-fold higher frequency in some African populations compared to Europeans[102]. This increased frequency and resulting adverse events led to thousands of cases of treatment discontinuation in Zimbabwe when the nation adopted a single-pill EFV-containing regimen[103]. This example highlights the need to not only tailor the therapy to the individual, but in some cases to the population as well. Newer generations of HIV therapies, such as integrase inhibitors and advanced nucleoside reverse transcriptase inhibitors, have more favorable pharmacokinetic and safety profiles[104]. However, the effects of long-term treatment with these drugs and any potential interactions with human genetic variability remain to be understood.

[H2] Complex trait genomics in HIV medicine. In addition to direct interactions between host genotype and drug metabolism, patients on long-term HIV therapy also experience early onset of several chronic diseases including cardiovascular disease[105–107], metabolic syndrome[108], kidney disease[109,110] and liver fibrosis[111]. These conditions are all known to have high heritability in the HIV uninfected population, and genetic risk factors for Type 2 diabetes[112] and cardiovascular disease[113] have been shown to be enhanced in PLWH on therapy. Recently, there has been a push to develop polygenic risk scores (PRS) in the general population. These scores, built by summing the additive effects of dozens to thousands of genetic variants within an individual, have been shown to have strong predictive ability for multiple metabolic, inflammatory, tumoural and cardiovascular conditions[32]. Investigations of PRS in the specific context of HIV infected individuals receiving long-term antiretroviral therapy have just begun, with the recent demonstrations that prediction of chronic kidney disease can be improved through the addition of a PRS to the known clinical and pharmacological risk factors[114,115], and that a PRS can be useful to stratify PLWH at a high risk of cardiometabolic diseases who may benefit from preventive therapies[116]. An important caveat is that PRS are not necessarily transferable across ancestral groups and, as in all areas of genomics, attention should be paid to enhancing diversity and ensuring equity in precision medicine approaches.

[H1] Conclusion and future perspectives

Host genomic studies have advanced our understanding of HIV biology in several important ways. Firstly, the demonstration of the dominant impact of HLA variation on HIV progression in the context of the whole genome reinforced the need to focus on T cell responses in vaccine design. Moreover, the ability to accurately infer HLA allele types and protein-level variability from genotyping array data, an approach first piloted in HIV genomic studies, has greatly increased our understanding of how amino acid variability in HLA molecules contributes to multiple medically important traits. Secondly, dense genotyping and large sample sizes enabled the discovery of multiple, independent signals in the *CCR5* locus, which provided a deeper understanding of how expression of *CCR5* is regulated and how it modulates HIV infection beyond the known impact of the CCR5Δ32 allele. Finally, amassing genome-wide data for large cohorts of PLWH has enabled the validity of previous candidate gene associations to be assessed, providing a new standard for identifying novel loci of HIV restriction.

In recent years, there have been several barriers to further advancing our understanding of how host genomics affects HIV susceptibility and progression. Firstly, current studies have predominantly included individuals of European ancestry, mirroring the lack of diversity in genomics in general[117], which is particularly problematic because the vast majority of PLWH are non-White. The example of the population-specific *CCR5Δ32* allele further highlights the need to stretch beyond European cohorts to determine if other population-specific effects may exist. Attaining the large sample sizes required for genomic discovery in non-European populations will require substantial investment of resources and building of capacity in low- and middle-income countries. Furthermore, understanding the potential function of genetic variants identified in diverse samples will require a shift towards inclusivity across genomics databases[118]. Secondly, with improvements in HIV care and broad adoption of test and treat strategies, the focus of host genomics studies has necessarily shifted away from natural history of infection phenotypes to intermediate phenotypes, pharmacogenomics of long-term therapy, comorbidities or vaccine response. Thirdly, understanding other classes of genetic variation that are not well captured by genotyping arrays, for example, diversity of KIR alleles and T cell receptor usage, the other partner in the HLA interaction, should be investigated to better understand how genetic variation in key innate and adaptive immune genes impact disease outcomes. However, capturing these types of variation requires in-depth sequencing to resolve genetic diversity and, in the case of T cell receptor variation, targeted immune assays to capture the relevant cells. Progress on computational methods for inferring variation at some complex loci from genotyping array data[51,119] or next-generation sequence data[120–122] will greatly aid these efforts.

The full translational potential of host genomics discovery in HIV has yet to be realized. Although the association between HLA allele type, epitope binding and HIV control have been well established, this knowledge has yet to be translated into an effective preventative or therapeutic vaccine. As mentioned above,

treatment of PLWH with CCR5-deficient cells has shown potential as an HIV cure, but several technological improvements in autologous cell editing will be required before it becomes a scalable strategy. In addition to targeting host genes for editing, *in vitro* studies have also shown that it is feasible to directly target and excise the integrated proviral genome[123,124]. Although an extremely promising strategy, delivery of the necessary machinery to latently infected cells remains a challenge.

The host genomics approach established in HIV research has since been applied to several other infectious diseases, including those posing substantial threats to human health, such as hepatitis C virus[125,126], tuberculosis[127], malaria[128] and even SARS-CoV2[129], among others. These studies have time and again uncovered novel therapeutic targets and mechanisms to identify the individuals that are most vulnerable to specific infections. As the world struggles with a novel pandemic-causing RNA virus, the lessons we can learn from how the human genome contributes to variability in outcome have never been more important.

References

1. Friedland, G. H. & Klein, R. S. Transmission of the human immunodeficiency virus. *N. Engl. J. Med.* **317**, 1125–35 (1987).
2. Kulkarni, P. S., Butera, S. T. & Duerr, A. C. Resistance to HIV-1 infection: lessons learned from studies of highly exposed persistently seronegative (HEPS) individuals. *AIDS Rev* **5**, 87–103 (2003).
3. Horton, R. E., McLaren, P. J., Fowke, K., Kimani, J. & Ball, T. B. Cohorts for the study of HIV-1-exposed but uninfected individuals: benefits and limitations. *J. Infect. Dis.* **202 Suppl**, S377–S381 (2010).
4. Sabin, C. A. & Lundgren, J. D. The natural history of HIV infection. *Curr. Opin. HIV AIDS* **8**, 311–7 (2013).
5. Chen, B. Molecular Mechanism of HIV-1 Entry. *Trends Microbiol.* **27**, 878–891 (2019).
6. Letvin, N. L. & Walker, B. D. Immunopathogenesis and immunotherapy in AIDS virus infections. *Nat. Med.* **9**, 861–6 (2003).
7. Dufour, C., Gantner, P., Fromentin, R. & Chomont, N. The multifaceted nature of HIV latency. *J. Clin. Invest.* **130**, 3381–3390 (2020).
8. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
9. Naranbhai, V. & Carrington, M. Host genetic variation and HIV disease: from mapping to mechanism. *Immunogenetics* **69**, 489–498 (2017).
10. An, P. *et al.* Regulatory polymorphisms in the cyclophilin A gene, PPIA, accelerate progression to AIDS. *PLoS Pathog* **3**, e88 (2007).
11. Bashirova, A. A. *et al.* Consistent effects of TSG101 genetic variability on multiple outcomes of exposure to human immunodeficiency virus type 1. *J. Virol.* **80**, 6757–63 (2006).
12. Boniotto, M. *et al.* Polymorphisms in the MBL2 promoter correlated with risk of HIV-1 vertical transmission and AIDS progression. *Genes Immun* **1**, 346–348 (2000).
13. Bochud, P.-Y. *et al.* Polymorphisms in Toll-like receptor 9 influence the clinical course of HIV-1 infection. *AIDS* **21**, 441–6 (2007).
14. An, P. *et al.* APOBEC3G genetic variants and their influence on the progression to AIDS. *J. Virol.* **78**, 11070–6 (2004).
15. Javanbakht, H. *et al.* Effects of human TRIM5alpha polymorphisms on antiretroviral function and susceptibility to human immunodeficiency virus infection. *Virology* **354**, 15–27 (2006).
16. Fellay, J. *et al.* Common genetic variation and the control of HIV-1 in humans. *PLoS Genet* **5**, e1000791 (2009).
17. International HIV Controllers Study *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science (80-.).* **330**, 1551–1557 (2010).

Commented [DC1]: Whether a study is 'the first' to show something is often open to debate, so we try to avoid such phrasing. I have made some edits below, but please feel free to edit as you think suitable. For example, instead of saying the 'first', you could state is it 'among the first' or 'among the earliest'?

18. McLaren, P. J. *et al.* Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathog.* 9, e1003515 (2013).

This GWAS of HIV-1 acquisition showed no evidence for association outside of CCR5 including lack of replication of 22 loci previously claimed to impact acquisition

19. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE . *Science (80-.).* 273, 1856–1862 (1996).

We believe this study to be the first description of host genetic resistance to an infectious disease observed in people homozygous for a loss of function mutation in CCR5 (previously CKR5)

20. Huang, Y. *et al.* The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nat. Med.* **2**, 1240–3 (1996).

21. Liu, R. *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86**, 367–377 (1996).

22. Kaslow, R. A. *et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med* **2**, 405–411 (1996).

23. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

24. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).

25. Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science (80-.).* 317, 944–947 (2007).

In what is considered to be the first genome-wide association study of an infectious disease phenotype, the authors demonstrate the large impact of genetic variation in the HLA region on determining HIV setpoint viral load

26. Mellors, J. W. *et al.* Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion. *Ann. Intern. Med.* **122**, 573–579 (1995).

27. Quinn, T. C. *et al.* Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *N. Engl. J. Med.* **342**, 921–9 (2000).

28. Pelak, K. *et al.* Host Determinants of HIV-1 Control in African Americans. *J. Infect. Dis.* **201**, 1141–1149 (2010).

29. Lingappa, J. R. *et al.* Genomewide association study for determinants of HIV-1 acquisition and viral set point in HIV-1 serodiscordant couples with quantified virus exposure. *PLoS One* **6**, e28632 (2011).

30. Luo, M. *et al.* A genetic polymorphism of FREM1 is associated with resistance against HIV infection in the Pumwani Sex Worker Cohort. *J. Virol.* **86**, 11899–11905 (2012).

31. Mclaren, P. J. *et al.* Fine-mapping classical HLA variation associated with durable host control of HIV-1 infection in african americans. *Hum. Mol. Genet.* **21**, 4334–4347 (2012).
32. Khera, A. V *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
33. McLaren, P. J. *et al.* Evaluating the Impact of Functional Genetic Variation on HIV-1 Control. *J. Infect. Dis.* **216**, 1063–1069 (2017).
34. Lane, J. *et al.* A genome-wide association study of resistance to HIV infection in highly exposed uninfected individuals with hemophilia A. *Hum. Mol. Genet.* **22**, 1903–1910 (2013).
35. Mackelprang, R. D. *et al.* Whole genome sequencing of extreme phenotypes identifies variants in CD101 and UBE2V1 associated with increased risk of sexually acquired HIV-1. *PLoS Pathog.* **13**, e1006703 (2017).
36. Bouloc, A., Bagot, M., Delaire, S., Bensussan, A. & Boumsell, L. Triggering CD101 molecule on human cutaneous dendritic cells inhibits T cell proliferation via IL-10 production. *Eur. J. Immunol.* **30**, 3132–3139 (2000).
37. Xia, Z.-P. *et al.* Direct activation of protein kinases by unanchored polyubiquitin chains. *Nature* **461**, 114–119 (2009).
38. Pertel, T. *et al.* TRIM5 is an innate immune sensor for the retrovirus capsid lattice. *Nature* **472**, 361–365 (2011).
39. Powell, T. R. *et al.* The behavioral, cellular and immune mediators of HIV-1 acquisition: New insights from population genetics. *Sci. Rep.* **10**, 3304 (2020).
40. Petrovski, S. *et al.* Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population. *AIDS* **25**, 513–518 (2011).
41. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).
42. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).
43. de Bakker, P. I. W. & Raychaudhuri, S. Interrogating the major histocompatibility complex with high-throughput genomics. *Hum. Mol. Genet.* **21**, 1–8 (2012).
44. Kennedy, A. E., Ozbek, U. & Dorak, M. T. What has GWAS done for HLA and disease associations? *Int. J. Immunogenet.* **44**, 195–211 (2017).
45. Venkataraman, G. R. *et al.* Pervasive additive and non-additive effects within the HLA region contribute to disease risk in the UK Biobank. *bioRxiv* 2020.05.28.119669 (2020). doi:10.1101/2020.05.28.119669
46. Sakaue, S. *et al.* A global atlas of genetic associations of 220 deep phenotypes. *medRxiv* 2020.10.23.20213652 (2020). doi:10.1101/2020.10.23.20213652
47. Carrington, M. *et al.* HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage.

Science (80-.). **283**, 1748–1752 (1999).

48. Migueles, S. A. *et al.* HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2709–2714 (2000).

49. Gao, X. *et al.* Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N. Engl. J. Med.* **344**, 1668–1675 (2001).

50. Kiepiela, P. *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432**, 769–775 (2004).

51. Jia, X. *et al.* Imputing amino Acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).

52. McLaren, P. J. *et al.* Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc. Natl. Acad. Sci. U. S. A.* 1–6 (2015). doi:10.1073/pnas.1514867112

53. **Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. *medRxiv* 2020.07.16.20155606 (2020). doi:10.1101/2020.07.16.20155606**

This paper presents a framework for HLA allele imputation and association testing from genome-wide SNP data and presents a detailed fine-mapping of HLA functional variation in ~12,000 people living with HIV

54. Kelleher, A. D. *et al.* Clustered Mutations in HIV-1 Gag Are Consistently Required for Escape from Hla-B27–Restricted Cytotoxic T Lymphocyte Responses. *J. Exp. Med.* **193**, 375–386 (2001).

55. Martinez-Picado, J. *et al.* Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. *J. Virol.* **80**, 3617–3623 (2006).

56. Schneidewind, A. *et al.* Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. *J. Virol.* **81**, 12382–12393 (2007).

57. Migueles, S. A. *et al.* HIV-specific CD8+ T cell proliferation is coupled to perforin expression and is maintained in nonprogressors. *Nat. Immunol.* **3**, 1061–1068 (2002).

58. Horton, H. *et al.* Preservation of T Cell Proliferation Restricted by Protective HLA Alleles Is Critical for Immune Control of HIV-1 Infection. *J. Immunol.* **177**, 7406–7415 (2006).

59. Almeida, J. R. *et al.* Superior control of HIV-1 replication by CD8+ T cells is reflected by their avidity, polyfunctionality, and clonal turnover. *J. Exp. Med.* **204**, 2473–2485 (2007).

60. Arora, J. *et al.* HLA Heterozygote Advantage against HIV-1 Is Driven by Quantitative and Qualitative Differences in HLA Allele-Specific Peptide Presentation. *Mol. Biol. Evol.* **37**, 639–650 (2020).

61. Arora, J. *et al.* HIV peptidome-wide association study reveals patient-specific epitope repertoires

associated with HIV control. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 944–949 (2019).

62. **Gaiha, G. D. *et al.* Structural topology defines protective CD8+ T cell epitopes in the HIV proteome. *Science* 364, 480–484 (2019).**

Using network analysis to quantify the structural importance of amino acids in HIV proteins, the authors show that mutations in epitopes presented by protective class I HLA alleles disproportionately impair viral replication

63. Apps, R. *et al.* Influence of HLA-C expression level on HIV control. *Science (80-.).* 340, 87–91 (2013).

This paper demonstrates that HLA-C expression level rather than epitope presentation is the key mediator of its impact on host control of HIV replication

64. Kulkarni, S. *et al.* Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature* **472**, 495–498 (2011).

65. Ramsuran, V. *et al.* Epigenetic regulation of differential HLA-A allelic expression levels. *Hum. Mol. Genet.* **24**, 4268–75 (2015).

66. Ramsuran, V. *et al.* Elevated HLA-A expression impairs HIV control through inhibition of NKG2A-expressing cells. *Science* 359, 86–90 (2018).

Furthering work on non-classical HLA effects in HIV control, this paper demonstrates how expression level of the HLA-A signal peptide can regulate the HLA-E:NKG2A interaction modulating spVL

67. Martin, M. P. & Carrington, M. Immunogenetics of HIV disease. *Immunol. Rev.* **254**, 245–264 (2013).

68. Pende, D. *et al.* Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. *Front. Immunol.* **10**, 1179 (2019).

69. Martin, M. P. *et al.* Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet* **31**, 429–434 (2002).

70. Novembre, J., Galvani, A. P. & Slatkin, M. The geographic spread of the CCR5 Δ32 HIV-resistance allele. *PLoS Biol.* **3**, 1954–1962 (2005).

71. Hladik, F. *et al.* Combined effect of CCR5-Delta32 heterozygosity and the CCR5 promoter polymorphism -2459 A/G on CCR5 expression and resistance to human immunodeficiency virus type 1 transmission. *J. Virol.* **79**, 11677–84 (2005).

72. Fätkenheuer, G. *et al.* Efficacy of short-term monotherapy with maraviroc, a new CCR5 antagonist, in patients infected with HIV-1. *Nat. Med.* **11**, 1170–2 (2005).

73. Wang, H. & Yang, H. Gene-edited babies: What went wrong and what could go wrong. *PLoS Biol.* **17**, e3000224 (2019).

74. Hutter, G. *et al.* Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N Engl J Med* 360, 692–698 (2009).

This is the first description of functional cure of HIV infection in a patient who received a stem cell transplant from a donor homozygous for the CCR5Δ32 polymorphism

75. Gupta, R. K. *et al.* Evidence for HIV-1 cure after CCR5Δ32/Δ32 allogeneic haemopoietic stem-cell transplantation 30 months post analytical treatment interruption: a case report. *lancet. HIV* **7**, e340–e347 (2020).
76. Tebas, P. *et al.* Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. *N. Engl. J. Med.* **370**, 901–910 (2014).
77. Smoleń-Dzirba, J. *et al.* HIV-1 Infection in Persons Homozygous for CCR5-Δ32 Allele: The Next Case and the Review. *AIDS Rev.* **19**, 219–230 (2017).
78. Michael, N. L. *et al.* The role of viral phenotype and CCR-5 gene defects in HIV-1 transmission and disease progression. *Nat. Med.* **3**, 338–40 (1997).
79. Kulkarni, S. *et al.* CCR5AS lncRNA variation differentially regulates CCR5, influencing HIV disease outcome. *Nat. Immunol.* **20**, 824–834 (2019).
80. Carlson, J. M., Le, A. Q., Shahid, A. & Brumme, Z. L. HIV-1 adaptation to HLA: a window into virus-host immune interactions. *Trends Microbiol.* **23**, 212–24 (2015).
81. Alter, G. *et al.* HIV-1 adaptation to NK-cell-mediated immune pressure. *Nature* **476**, 96–100 (2011).
82. Hölzemer, A. *et al.* Selection of an HLA-C*03:04-Restricted HIV-1 p24 Gag Sequence Variant Is Associated with Viral Escape from KIR2DL3+ Natural Killer Cells: Data from an Observational Cohort in South Africa. *PLoS Med.* **12**, e1001900; discussion e1001900 (2015).
83. Snoeck, J., Fellay, J., Bartha, I., Douek, D. C. & Telenti, A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology* **8**, 87 (2011).
84. **Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* 2, e01123 (2013).**

This paper proposes the G2G method and demonstrates that viral genetic variation can be a more powerful phenotype than clinical markers for host-pathogen genomic studies

85. Palmer, D. S. *et al.* Mapping the drivers of within-host pathogen evolution using massive data sets. *Nat. Commun.* **10**, 3017 (2019).
86. Fellay, J. & Pedergnana, V. Exploring the interactions between the human and viral genomes. *Hum. Genet.* **139**, 777–781 (2020).
87. **Ansari, M. A. *et al.* Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat. Genet.* 49, 666–673 (2017).**

Applying the G2G approach to Hepatitis C Virus, the authors demonstrate that it can be used to detect viral evolution owing to both innate and adaptive host immune pressure

88. Chaturvedi, N. *et al.* Adaptation of hepatitis C virus to interferon lambda polymorphism across multiple viral genotypes. *Elife* **8**, (2019).

89. Rüeger, S. *et al.* The influence of human genetic variation on Epstein-Barr virus sequence diversity. *Sci. Rep.* **11**, 4586 (2021).
90. Eisinger, R. W., Dieffenbach, C. W. & Fauci, A. S. HIV Viral Load and Transmissibility of HIV Infection: Undetectable Equals Untransmittable. *JAMA* **321**, 451–452 (2019).
91. UNAIDS. *90-90-90 An ambitious treatment target to help end the AIDS epidemic*. (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2014).
92. Lubomirov, R. *et al.* Association of pharmacogenetic markers with premature discontinuation of first-line anti-HIV therapy: An observational cohort study. *J. Infect. Dis.* **203**, 246–257 (2011).
93. Mallal, S. *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727–732 (2002).
94. Hetherington, S. *et al.* Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet* **359**, 1121–1122 (2002).
95. Norcross, M. A. *et al.* Abacavir induces loading of novel self-peptides into HLA-B*57: 01: an autoimmune model for HLA-associated drug hypersensitivity. *AIDS* **26**, F21-9 (2012).
96. Rotger, M. *et al.* Predictive value of known and novel alleles of CYP2B6 for efavirenz plasma concentrations in HIV-infected individuals. *Clin. Pharmacol. Ther.* **81**, 557–566 (2007).
97. Arab-Alameddine, M. *et al.* Pharmacogenetics-based population pharmacokinetic analysis of efavirenz in HIV-1-infected individuals. *Clin. Pharmacol. Ther.* **85**, 485–494 (2009).
98. Haas, D. W. *et al.* Pharmacogenetics of efavirenz and central nervous system side effects: an Adult AIDS Clinical Trials Group study. *AIDS* **18**, 2391–2400 (2004).
99. di Iulio, J. *et al.* In vivo analysis of efavirenz metabolism in individuals with impaired CYP2A6 function. *Pharmacogenet. Genomics* **19**, 300–309 (2009).
100. Lubomirov, R. *et al.* Pharmacogenetics-based population pharmacokinetic analysis of etravirine in HIV-1 infected individuals. *Pharmacogenet. Genomics* **23**, 9–18 (2013).
101. Lubomirov, R. *et al.* ADME pharmacogenetics: investigation of the pharmacokinetics of the antiretroviral agent lopinavir coformulated with ritonavir. *Pharmacogenet. Genomics* **20**, 217–230 (2010).
102. Zanger, U. M. & Klein, K. Pharmacogenetics of cytochrome P450 2B6 (CYP2B6): advances on polymorphisms, mechanisms, and clinical relevance. *Front. Genet.* **4**, (2013).
103. Nordling, L. How the genomics revolution could finally help Africa. *Nature* **544**, 20–22 (2017).
104. Kanters, S. *et al.* Comparative efficacy and safety of first-line antiretroviral therapy for the treatment of HIV infection: a systematic review and network meta-analysis. *Lancet HIV* **3**, e510–e520 (2016).
105. Currier, J. S. *et al.* Coronary heart disease in HIV-infected individuals. *J. Acquir. Immune Defic. Syndr.* **33**, 506–12 (2003).
106. Obel, N. *et al.* Ischemic heart disease in HIV-infected and HIV-uninfected individuals: a population-

based cohort study. *Clin. Infect. Dis.* **44**, 1625–31 (2007).

107. Lang, S. *et al.* Increased risk of myocardial infarction in HIV-infected patients in France, relative to the general population. *AIDS* **24**, 1228–30 (2010).

108. Périard, D. *et al.* Atherogenic dyslipidemia in HIV-infected individuals treated with protease inhibitors. The Swiss HIV Cohort Study. *Circulation* **100**, 700–5 (1999).

109. Nadkarni, G. N. *et al.* The burden of dialysis-requiring acute kidney injury among hospitalized adults with HIV infection: a nationwide inpatient sample analysis. *AIDS* **29**, 1061–6 (2015).

110. Scherzer, R. *et al.* Association of tenofovir exposure with kidney disease risk in HIV infection. *AIDS* **26**, 867–75 (2012).

111. Li, Y. *et al.* Combination Antiretroviral Therapy Is Associated With Reduction in Liver Fibrosis Scores in HIV-1-Infected Subjects. *Medicine (Baltimore).* **95**, e2660 (2016).

112. Rotger, M. *et al.* Impact of single nucleotide polymorphisms and of clinical risk factors on new-onset diabetes mellitus in HIV-infected individuals. *Clin. Infect. Dis.* **51**, 1090–1098 (2010).

113. Rotger, M. *et al.* Contribution of Genetic Background, Traditional Risk Factors, and HIV-Related Factors to Coronary Artery Disease Events in HIV-Positive Persons. *Clin. Infect. Dis.* **57**, 112–121 (2013).

114. Dietrich, L. G. *et al.* Contribution of Genetic Background and Data Collection on Adverse Events of Anti-human Immunodeficiency Virus (HIV) Drugs (D:A:D) Clinical Risk Score to Chronic Kidney Disease in Swiss HIV-infected Persons With Normal Baseline Estimated Glomerular Filtrat. *Clin. Infect. Dis.* **70**, 890–897 (2020).

115. Dietrich, L. G. *et al.* Rapid Progression of Kidney Dysfunction in Swiss People Living with HIV: Contribution of Polygenic Risk Score and D:A:D Clinical Risk Score. *J. Infect. Dis.* (2020). doi:10.1093/infdis/jiaa695

116. Chang, H. *et al.* Genetic architecture of cardiometabolic risks in people living with HIV. *BMC Med.* **18**, 288 (2020).

117. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).

118. Rotman, D. DNA databases are too white. *MIT Technology Review* (2021).

119. Vukcevic, D. *et al.* Imputation of KIR Types from SNP Variation Data. *Am. J. Hum. Genet.* **97**, 593–607 (2015).

120. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).

121. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

122. Roe, D. & Kuang, R. Accurate and Efficient KIR Gene and Haplotype Inference From Genome Sequencing Reads With Novel K-mer Signatures. *Front. Immunol.* **11**, 583013 (2020).

123. Sarkar, I., Hauber, I., Hauber, J. & Buchholz, F. HIV-1 proviral DNA excision using an evolved recombinase. *Science* **316**, 1912–5 (2007).
124. Hu, W. *et al.* RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11461–6 (2014).
125. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399–401 (2009).
126. Thomas, D. L. *et al.* Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature* **461**, 798–801 (2009).
127. Thye, T. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* **42**, 739–741 (2010).
128. Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature* **526**, 253–257 (2015).
129. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, (2020).
130. Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet.* **14**, e1007309 (2018).

Author Contributions

The authors contributed equally to all aspects of the article

Competing interests

The authors declare no competing interests.

744 **Table 1: Genetic variants known to affect the pharmacokinetics of anti-HIV drugs**
745

Drug	Class	Gene	Variant	Effect on patient
Abacavir	NRTI	*HLA-B*	*57:01 (rs2395029)	Hypersensitivity reaction
Atazanavir	PI	*UGT1A1*	rs8175347 (*28,*37)	Predisposition to hyperbilirubinemia
Efavirenz	NNRTI	*CYP2A6*	rs1801272 (*2), rs5031016 (*7,*10,*19), rs28399433 (*9)	Predisposition to high drug plasma levels and treatment discontinuation
Efavirenz & Nevirapine	NNRTI	*CYP2B6*	rs3745274 (*6), rs12721655 (*8,13), rs35303484 (*11), rs36060847 (*12), rs35773040 (*14), rs35979566 (*15), rs28399499 (*16,18)	Predisposition to high drug plasma levels
Etravirine	NNRTI	*CYP2C9, CYP2C19*	rs1057910 (*3), rs4424285 (*2)	Predisposition to high drug plasma levels
Lopinavir	PI	*ABCC2, CYP3A*,	rs717620(T), rs6945984(C)	Predisposition to high drug plasma levels
Lopinavir	PI	*SLCO1B1*	rs4149056(*5), rs17329885(*4)	Predisposition to low drug plasma levels

746 NRTI, nucleoside reverse transcriptase inhibitor; PI, protease inhibitor; NNRTI, non-nucleoside reverse
747 transcriptase inhibitor
748
749

Figure legends

Figure 1. **Schematic representation of the HIV life cycle and the HLA-mediated host response**. The viral envelope glycoprotein gp120 binds CD4 and CCR5 on the surface of target cells triggering fusion of the viral and host cell membranes; host genomic studies have implicated genetic variants in *CCR5* (blue background) listed as modifiers of infectivity. Reverse transcription of the single-stranded RNA genome into double-stranded (ds) DNA occurs using proteins carried by the infecting virion. Viral dsDNA is trafficked to the nucleus where it is integrated into the human genome. Transcription of viral dsDNA results in viral gene expression and genome replication. Viral mRNA is translated into polyproteins which are cleaved by the viral protease. Functional proteins assemble with copies of the viral genome at the cell membrane and mature virions bud from the surface. In parallel, as part of the immune response, viral proteins are digested by the host proteasome and processed through tapasin I and II (orange rectangles) into the golgi where the epitopes are loaded in the HLA class I molecules. The peptide-loaded HLA protein is trafficked to the cell surface and presented to CTLs. Variability in epitope presentation by HLA-B alleles, such as the protective allele B*57:01, and in expressino levels of HLA-C and HLA-A alleles modify repsone to infection and spVL.

Figure 2. **Classical and non-classical effects of HLA class I on HIV suppression**. A) HIV infected cells expressing protective HLA-B alleles tend to present a more diverse and more structurally conserved set of HIV epitopes compared to non-protective alleles. Interactions with protective alleles tend to produce a more polyfunctional cytotoxic T-lymphocyte response. B) HLA-C protein isoforms vary broadly in their level of expression on the surface of infected cells. *HLA-C* alleles that do not have a binding site for microRNA-148a (miRNA-148a) in the 3' untranslated region of their mRNAs escape suppression and present more peptide on the cell surface than alleles with a miRNA-148a binding site, resulting in initiation of stronger CTL responses C) Different HLA-A alleles express different amounts of HLA-A signal peptide, which positively correlates with HLA-E peptide expression. HLA-E interacts with the NKG2A receptor on the surface of natural killer (NK) cells and, when highly expressed, inhibits killing of infected cells.

Figure 3. **CCR5 expression modifies HIV progression** A) A 32-basepair deletion in CCR5 (CCR5Δ32) results in reduced expression of CCR5 on the surface of target cells. Heterozygous individuals exhibit reduced CCR5 expression, lower setpoint viral loads (spVLs) and slower disease progression. Individuals carrying two defective copies of the CCR5 gene show no surface expression and are highly resistant to HIV infection. Additionally, a single nucleotide polymorphism downstream of CCR5 (rs1015164) affects cell surface levels of CCR5. In homozygous reference (A/A) and heterozygous (A/G) individuals respectively, surface expression of CCR5 is normal, whereas G/G homozygous individuals, have lower CCR5 surface expression and lower spVLs. B) A single nucleotide polymorphism downstream of CCR5 (rs1015164) regulates the expression of

Commented [DC2]: Please check figure legends carefully against the figures, especially for descriptions of colours/shapes/symbols.

Commented [JF3R2]: Ok

Commented [DC4]: Please amend to reflect preferred way of indicating these variants (coloured text cannot be used).

Commented [JF5R4]: To be modified in the figure, as suggested

Commented [DC6]: Mention the variants listed in the figure.

an antisense RNA termed CCR5-AS. When CCR5-AS (orange) is expressed at high or intermediate levels in homozygous reference (A/A) and heterozygous (A/G) individuals respectively, CCR5 mRNA (blue) is protected from Raly- mediated degradation and results in normal levels of surface expression. In G/G homozygous individuals, CCR5-AS expression is diminished and CCR5 mRNA is degraded resulting in lower surface expression at the cellular level and lower set point viral load overall.

Figure 4. **Detecting genomic signatures of host-pathogen interactions in matched host:virus samples.** First, genetic variants in the host (human) and pathogen (viral) genomes are identified from genome-wide genotyping or sequencing data and catalogued. A genome-wide search for associations between human polymorphisms and viral variants is then performed, which needs to consider the risk of systematic signal inflation owing to population stratification. On the human side, this can be addressed by methods that infer genetic ancestry, such as principal component analysis followed by the inclusion of the top principal components as covariates, or by the use of mixed models that incorporate the full covariance structure of the study population[130]. On the pathogen side, various phylogenetic-based and model-based approaches have been proposed[85]. Significant associations, after correction for multiple testing, reveal the loci involved in host-pathogen genomic conflicts. The image showing viral-host interactions is adapted with permission from Bartha et al., 2013. (ref 84)

Figure 5. **Antiretroviral drugs target multiple stages of the HIV life cycle**. Commonly used antiretroviral drugs target receptor binding, reverse transcription, integration and protease cleavage. Genetic variation in several human genes (bold text) have been shown to modify drug metabolism and contribute to adverse drug reactions (detailed in Table 1).

Commented [DC7]: Amend as necessary to match figure.

Commented [JF8R7]: To be modified on the figure, as indicated

Glossary

C-C CHEMOKINE RECEPTOR TYPE 5

(CCR5) A beta-chemokine receptor that is involved in lymphocyte trafficking. In combination with CD4, CCR5 is the major host cell receptor for HIV and interacts with gp120 in the viral envelope to promote cell entry and infection.

CYTOTOXIC T LYMPHOCYTE

(CTL) A cytotoxic T cell (also called CD8+ T cell) is an effector T lymphocyte that specifically kills target cells that express an appropriate peptide:MHC class I complex recognized by its T cell receptor.

EPITOPES

Parts of an antigen that make contact with a particular antibody or T cell receptor and is thus capable of stimulating an immune response.

GENETIC ARCHITECTURE

Underlying genetic basis of a given trait, in terms of variant number, effect size, allele frequency and interactions.

GENOME-TO-GENOME STUDIES

Methods that test the hypothesis that host genetic variability causes pathogen genetic variability and can indicate novel host restriction factors.

HIV/AIDS

Human Immunodeficiency Virus (HIV) is the causative agent of acquired mmunodeficiency syndrome (AIDS), which is a state of severe immune deficiency defined as an HIV infection with either a CD4+ T cell count < 200 cells per μL or the occurrence of a specific AIDS-defining illness.

HIV CONTROLLERS

A group of people living with HIV whose plasma HIV RNA load is spontaneously maintained at very low levels for several years (usually at least 3 to 5 years) in the absence of antiretroviral therapy.

HIV LATENCY

The long-term persistence of HIV in an integrated, but transcriptionally inactive, form in the host genome. Because latent HIV resides in memory T cells, it persists indefinitely even in patients on suppressive antiretroviral therapy. This latent reservoir is a major barrier to curing HIV infection.

HIV PROGRESSION

The natural disease course of HIV infection in untreated individuals, characterized by an acute phase, a chronic phase and development of AIDS. The rate of HIV progression varies dramatically in the infected population.

HIV TARGET CELLS

The cells primarily infected by HIV, namely CD4+ T cells and macrophages, both of which are key components of a healthy immune system.

HUMAN LEUKOCYTE ANTIGEN

(HLA) A protein, encoded by one of a group of *HLA* genes, that presents antigens that train the adaptive immune response. *HLA* genes are highly variable and allelic variants encode proteins that are differentially able to present antigens based on the amino acid sequences in the peptide binding grooves.

KILLER IMMUNOGLOBULIN-LIKE RECEPTORS

(KIR) A family of highly polymorphic activating and inhibitory receptors that serve as key regulators of human natural killer (NK) cell function.

POLYGENIC RISK SCORES

(PRS) Statistics that are calculated by enumerating the number of risk alleles associated with a particular phenotype (often weighted by their population-level effect sizes) that are present in a single individual and comparing the individual's score to the distribution of risk scores in the population.

POPULATION STRATIFICATION

Presence of systematic differences in allele frequencies between population subgroups owing to systematic differences in ancestry.

RALY-MEDIATED DEGRADATION

A mechanism in which the Raly protein binds to the 3' UTR of an mRNA to promote its degradation.

RESTRICTION FACTOR

A host cellular protein that participates in antiviral defense by interfering with specific steps of the viral replication cycle.

SETPOINT VIRAL LOAD

(SPVL) Mean log viral load (HIV RNA copies per ml of plasma) measured in an HIV-infected individual during the chronic phase of infection. The setpoint viral load varies substantially within a population and correlates with disease progression.

TRANSMITTED/FOUNDER VIRUS

The single viral variant that is responsible for a new infection after being transferred from an infected individual to an uninfected individual. Sexually transmitted HIV infections are typically established from a single transmitted/founder virus.

TREATMENT AS PREVENTION

Strategy to prevent the sexual transmission of HIV through the prescription of HIV medication. It is a highly effective option for preventing HIV transmission. People living with HIV who take antiretroviral drugs and maintain an undetectable viral load have effectively no risk of sexually transmitting the virus to their HIV-negative partners.

ToC blurb

McLaren and Fellay review our current understanding of the effects of human genetic variation on HIV infection and disease progression, and how this knowledge is contributing to preventative and therapeutic approaches.

Commented [DC9]: This blurb will appear on the online homepage and in the ToC of the print issue. Please check for accuracy. Maximum 40 words.

Commented [JF10R9]: Ok perfect