

The genetic architecture of HIV-1 virulence

François Blanquart ^{1,*}, Chris Wymant ^{2,3}, Matthew Hall ^{2,3}, Robert Power^{3,4}, Tanya Golubchik ⁵, Astrid Gall⁶, Mariateresa de Cesare^{2,3,7}, George Macintyre-Cockett^{2,3,7}, Margreet Bakker^{8,9}, Daniela Bezemer^{10,11}, Migle Gabrielaite ¹², Swee Hoe Ong ¹³, Michelle Kendall¹⁴, Rafael Sauter³, Norbert Bannert¹⁵, Jacques Fellay ^{16,17}, M. Kate Grabowski¹⁸, Barbara Günsheimer-Bartmeyer¹⁹, Huldrych F. Günthard^{20,21}, Pia Kivelä²², Roger D. Kouyos^{20,21}, Oliver Laeyendecker ²³, Rasmus L. Marvig ²⁴, Karolin Meixenberger ²⁵, Laurence Meyer²⁶, Ard van Sighem^{10,11}, David Bonsall^{2,3,7}, Marc van der Valk^{10,11}, Ben Berkhout²⁷, Paul Kellam ^{28,29}, Marion Cornelissen³⁰, Peter Reiss ³¹, Christophe Fraser^{2,3,*}

¹Center for Interdisciplinary Research in Biology, CNRS, Collège de France, PSL Research University, 11 place Marcelin Berthelot, Paris 75231, France

²Pandemic Sciences Institute, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford OX3 7LF, United Kingdom

³Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford OX3 7LF, United Kingdom

⁴St Edmund Hall, University of Oxford, Queen's Ln, Oxford OX1 4AR, United Kingdom

⁵Sydney Infectious Diseases Institute, Faculty of Medicine and Health, University of Sydney; Biomedical Building, 1 Central Avenue, Eveleigh, NSW 2015, Australia

⁶EMBO, Meyerhofstrasse 1, 69117 Heidelberg, Germany

⁷Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Roosevelt Dr, Headington, Oxford OX3 7BN, United Kingdom

⁸Amsterdam University Medical Centers, University of Amsterdam, Medical Microbiology and Infection Prevention, Laboratory of Experimental Virology, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands

⁹Amsterdam institute for Immunology and Infectious diseases, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands

¹⁰Stichting HIV Monitoring, Tafelbergweg 51, 1105 BD Amsterdam, The Netherlands

¹¹Department of Infectious Diseases, Amsterdam Institute for Immunology and Infectious Diseases, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands

¹²Institute of Biotechnology, Life Sciences Center, Vilnius University, Sauletekio al. 7, 10257 Vilnius, Lithuania

¹³Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

¹⁴Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom

¹⁵Department of Infectious Diseases, Unit 18, Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany

¹⁶School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Station 19, 1015 Lausanne, Switzerland

¹⁷Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Rue du Bugnon 46, 1011 Lausanne, Switzerland

¹⁸Department of Pathology, Johns Hopkins School of Medicine, 446 Carnegie Building, 600 N Wolfe St, 21287 Baltimore MD, USA

¹⁹Department of Infectious Disease Epidemiology, Unit 34 HIV/AIDS, STI and Blood-borne infections, Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany

²⁰Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Raemistrasse 100, 8091 Zurich, Switzerland

²¹Institute of Medical Virology, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

²²Department of Infectious Diseases, Helsinki University Hospital, 00029 HUS, Helsinki, Finland

²³Division of Intramural Research, National Institute of Allergy and Infectious Diseases, 855 North Wolfe St, 21209 Baltimore, MD, USA

²⁴Department of Genomic Medicine, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

²⁵Department of Infectious Diseases, Unit 18 Sexually transmitted bacterial Pathogens (STI) and HIV, Robert Koch Institute, Seestrasse 10, 13353 Berlin, Germany

²⁶Inserm CESP U1018, APHP Bicetre Hospital, Paris Saclay University, 82 rue du general Leclerc, 94276 le Kremlin-Bicetre, France

²⁷Laboratory of Experimental Virology, Department of Medical Microbiology, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 15, 1105 AZ, Amsterdam, The Netherlands

²⁸Department of Infectious Disease, South Kensington Campus, Imperial College London, London, SW7 2AZ, United Kingdom

²⁹RQ Biotechnology Ltd, London, Scale Space, 58 Wood Lane, London, W12 7RZ, United Kingdom

³⁰Medical Microbiology and Infection Prevention, Laboratory of Clinical Virology, Amsterdam University Medical Centers, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands

³¹Amsterdam University Medical Center, University of Amsterdam, Global Health, Amsterdam institute for Immunology and Infectious diseases and Amsterdam Institute for Global Health and Development, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

*Corresponding authors. François Blanquart, Center for Interdisciplinary Research in Biology, CNRS, Collège de France, PSL Research University, 11 place Marcelin Berthelot, Paris 75231, France. E-mail: francois.blanquart@college-de-france.fr; Christophe Fraser, Pandemic Sciences Institute and Big Data Institute, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford OX3 7LF, United Kingdom. E-mail: christophe.fraser@ndm.ox.ac.uk.

Abstract

The virulence of Human Immunodeficiency Virus-1 (HIV-1) is partly determined by viral genetic variation. Finding individual genetic variants affecting virulence is important for our understanding of HIV pathogenesis and evolution of virulence; however, very few have been identified. To this end, within the “Bridging the Evolution and Epidemiology of HIV in Europe” (BEEHIVE) collaboration, we produced whole-genome HIV sequence data for 2294 seroconverters from European countries for a genome-wide association study (GWAS). We considered two phenotypes: (i) set-point viral load (SPVL), the approximately stable viral load from 6 to 24 months after infection, and (ii) the rate of CD4 cell count decline. We developed a GWAS method that corrects for population structure with random

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

effects, accounts for two or more alleles at each locus, and tests for the effect of multiple genetic variants including single-nucleotide polymorphisms (SNPs), k-mers, insertions and deletions, within-host variant frequency, the number of rare point mutations, and drug resistance. We confirmed with this new approach that viral genomes explained 26% [95% CI 17%–35%] of the variance in SPVL, while they explained only 0.9% [0.0%–2.1%] of the variance in the rate of CD4 cell count decline. After correction for multiple testing, among all tested variants, only two significantly explained SPVL: an epitope mutation allowing escape from the host HLA-B*57 allele and lowering SPVL by $-0.26 \log_{10}$ copies/ml and an epitope mutation allowing escape from the host HLA-B*35 allele and increasing SPVL by $+0.22 \log_{10}$ copies/ml. We attempted to replicate these two large effects in two additional independent datasets together encompassing 2445 seroconverters, with mixed results. Overall, the inferred effects of all SNPs and amino-acid variants weakly correlated (R^2 ranging from 0.08 to 0.87%, P -values from 0.001 to 0.32) between our main dataset and these two additional datasets. Lastly, a lasso regression of phenotypes on genetic variants confirmed the heritability of SPVL and explained up to 6% of variance in SPVL in cross-validation datasets. These findings suggest that HIV SPVL is determined by viral genomes through HLA escape variants with potentially large, host-dependent effects that may not always be detected at the population level and many other variants with effects too weak to reach genome-wide significance in our GWAS.

Keywords: HIV-1; virulence; genome-wide association study; polygenic trait; CTL escape

Introduction

An estimated 39 million individuals live with HIV as of 2022 (Joint United Nations Programme on HIV/AIDS (UNAIDS)). Untreated HIV infection leads to deaths within years, with a large variability in time from infection to death. Although antiretroviral therapy prevents onwards transmission and extends life expectancy to approximately normal duration, 24% of individuals living with HIV are still untreated, and in 2022, 1.3 million new infections occurred, and 630 000 individuals died of AIDS (Joint United Nations Programme on HIV/AIDS (UNAIDS)). HIV infection is a chronic infection with a large mutation rate, enabling the rapid within-host evolution of drug resistance (Wei et al. 1995) and cytotoxic T lymphocyte (CTL) escape (Goulder et al. 2001) mutations. At the between-host level, virulence may also evolve over decades (Herbeck et al. 2012). HIV virulence is difficult to measure directly but can be indirectly estimated by the set-point viral load (SPVL, the approximately stable value of viremia in early untreated infection), the rate of decline in host CD4 cell counts, or the rate of progression to AIDS. All of these are highly variable between individuals. Several host factors affecting this variability are known, including sex, age, and the human genotype. Specifically, variability at human leukocyte antigen (HLA) loci [including both HLA types and individual single-nucleotide polymorphisms (SNPs) (Fellay et al. 2007, 2009)] and other mutations like the deletion $\Delta 32$ in the CCR5 receptor gene (McDermott et al. 1998) are associated with variability in viral load or disease progression. However, until recently, the contribution of the viral genome to virulence—the viral heritability—was less clear. This is quantified through the heritability of virulence: the proportion of variability in virulence that is due to variability in viral genetics. The heritability of SPVL was initially estimated to be around 30% from regression of SPVL in donor–recipient pairs (reviewed in Fraser et al. 2014). Subsequent studies based on viral genomes suggested either larger (around 50%) (Alizon et al. 2010) or much smaller (6%) (Hodcroft et al. 2014) heritability. The most recent studies, using improved methods and viral phylogenies inferred from new genomic datasets, found an SPVL heritability of 20%–30% consistent with donor–transmission pairs (Bachmann et al. 2017, Blanquart et al. 2017, Bertels et al. 2018, Mitov and Stadler 2018, Bastide et al. 2020) and a somewhat lower heritability for the slope of CD4 cell count decline estimated at 11% or 17% (Blanquart et al. 2017, Bertels et al. 2018).

The mechanisms of pathogenesis, the mode of evolution of virulence traits, and the evolutionary forces acting on these traits could all be better explained with improved knowledge of the genetics of virulence. At a broad genetic level, subtype D was found to be associated with faster disease progression than subtype A (Kaleebu et al. 2001, 2002), subtype B with faster disease

progression than subtypes A and C (Touloumi et al. 2013), and various recombinant forms have been associated with faster disease progression (Palm et al. 2014, Kouri et al. 2015). A highly virulent lineage associated with high SPVL and faster CD4 cell count decline was recently discovered (Wymant et al. 2022). At a finer genetic level, a deletion in the *nef* gene leads to slower disease progression (Learnmont et al. 1999), and viral variants that switch from using the CCR5 co-receptor to the CXCR4 co-receptor earlier in infection may lead to faster disease progression (Asjö et al. 1986, Koot et al. 1993, Ghosn et al. 2017). The effect of the coreceptor switch may underlie the difference in virulence between subtypes (Taylor et al. 2008). Lastly, SNPs in the viral genome conferring escape from the CTL response or resistance to antiretroviral therapy may be costly, as some of them rapidly revert to the wild type form upon transmission (Leslie et al. 2004, Yang et al. 2015). These SNPs are usually strongly associated with host factors such as HLA type or with antiretroviral treatment, making it difficult to measure their effect on viral load.

Several genome-wide association studies (GWASs) have been conducted to decipher more precisely the determinants of HIV virulence. Using paired human and viral genomic data with viral loads from 1071 individuals, Bartha et al. did not find any viral mutation significantly associated with viral load, though they reported strong associations between human SNPs in the HLA region and certain HIV amino acid variants (Bartha et al. 2013, 2017). They reported that a larger fraction of the variance in viral load was explained by viral variants than by host HLA SNPs. In another GWAS on the viral load of 2122 individuals living with HIV, Gabrielaite et al. reported four viral amino acid variants significantly associated with viral load after Bonferroni correction, with modest effects ranging from -0.084 to $0.097 \log_{10}$ copies/ml, and confirmed the very strong associations between several human SNPs and HIV amino acid variants (Gabrielaite et al. 2021). These two studies were conducted on diverse cohorts, only the viral load was examined, and the measure of this phenotype was not standardized across cohorts.

More standardized phenotypes and a large sample size could improve the power to detect specific HIV lineages and unravel novel variants and mechanisms explaining variability in virulence. Within the BEEHIVE collaboration, we assembled a large dataset of HIV seroconverters in Europe (from cohorts in Belgium, Finland, France, Germany, Netherlands, Switzerland, and the UK) and Uganda and used whole HIV genomes, SPVL re-measured with a standardized set of assays, and CD4 cell counts, to conduct a GWAS of viral genetic variants on virulence phenotypes. A subset of these data has been used before to quantify the heritability of SPVL (Blanquart et al. 2017).

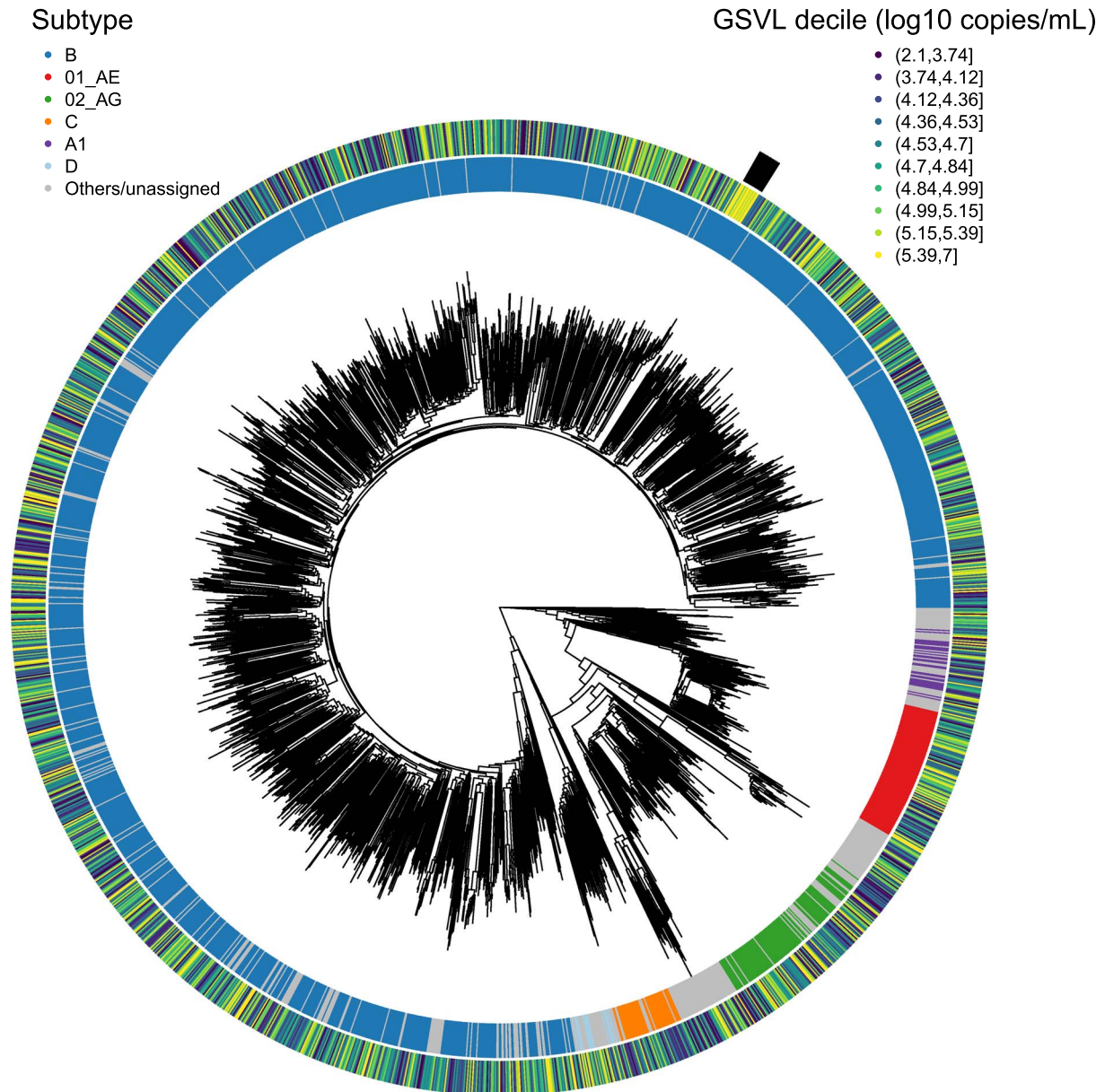


Figure 1. Phylogenetic tree of HIV-1 sequences used in the main analysis, with subtype (determined by COMET) indicated by colour in the inner ring, and ‘gold standard’ viral load (GSVL) categorized into deciles indicated by colour in the outer ring. The black rectangle marks the hypervirulent variant. 33 outlier sequences in terms of root-to-tip distance were removed for clarity.

Results

Our main dataset comprised data from 2294 European HIV seroconverters matching the inclusion criteria, with whole-genome HIV sequences from samples obtained early in infection (before antiretroviral treatment) and available phenotypes (Fig. 1). Seroconverters are individuals for whom the date of HIV infection is approximately known, as they had a negative test result in the year before the first positive HIV test or other evidence of recent infection. The four phenotypes were: (i) ‘gold standard’ viral load (GSVL), a single viral load for each individual re-measured for this study with a standardized set of assays, and measured for the same sample used for viral genome sequencing; (ii) adjusted set-point viral load (SPVL_a), the mean of all (log) viral loads an individual had measured in the defined time window 6–24 months postinfection, adjusted for various host factors; (iii) adjusted normalized SPVL (‘SPVL_{an}’), same as (ii) but where the distribution of viral loads across timepoints and individuals is normalized; and (iv) the rate of CD4 cell count decline (‘CD4 slope’).

We developed a linear mixed-model GWAS method to test for several types of HIV genetic variants (including loci with more than two alleles) and appropriately correct for population structure. Population structure is the relatedness between viral genomes in the sample. This structure may generate noncausal associations between genetic variants and the phenotype of interest by confounding. As part of the linear mixed model, a likelihood ratio test gives a single *P*-value for an association between variation at a locus and the phenotype, underlying all *P*-values quoted hereafter. While prior studies tested the effect of HIV amino acid variants on phenotypes (Bartha *et al.* 2013, Gabrielaite *et al.* 2021), we tested a more comprehensive list of types of variants: SNPs, amino acid variants, k-mers (substrings of consecutive nucleotides in the viral genome sequence), insertions and deletions (‘length’ variants), within-host frequencies, the fraction of the viral genome composed of rare point mutations, and predicted transmitted drug resistance. Our method also estimates the heritability of the phenotypes (the fraction of phenotypic variance

explained by genetic factors). In a second step, we attempted to replicate our findings on two additional datasets of 323 and 2122 HIV seroconverters. See Material and Methods, Section 4.4, for more details.

The main genome-wide association study identifies four ‘hits’ associated with viral load

From the GWAS, we initially identified in total 23 hits (genetic variation significantly associated with phenotype, with P -value < 0.05 after multiple testing correction) associated with at least one of the three viral load phenotypes or the CD4 slope (Fig. 2). Moreover, sequences harbouring a greater number of rare point mutations (defined as those point mutations at frequency lower than 5% in our sample) had lower viral load (Supplementary Figure 1). We found no evidence of an association between viral load and predicted transmitted drug resistance.

The power calculation showed that only four of the hits, all associated with one of the three viral load phenotypes (not CD4 slope), as well as the burden of rare point mutations, could possibly be tested in our additional data (the power to replicate these was greater than 0). The others were not replicable because the variants did not appear in the additional sequence data. Most ($N=17$) of the nonreplicable hits were within-host frequency variants and length variants associated with viral load. These variants could be artefacts caused by amplification or sequencing errors that are associated with viral load; alternatively, the sequencing method used for the additional BEEHIVE dataset [veSEQ-HIV (Bonsall et al. 2020)] could be less able to detect within-host frequency variants than the method used for the main dataset (Wymant et al. 2022). The last two nonreplicable hits were a 2-mer and an amino acid variant associated with CD4 slope. The alternative variants for these hits were found in fewer than 10 additional sequences, and we therefore did not attempt to replicate the association observed for these variants. In the following, we therefore focus on the four replicable k-mer hits, all associated with viral load.

The four hits were at positions 1413, 1514, 6570, and 9008 with respect to the 9719-bp-long HXB2 reference (Supplementary Table 1, Fig. 2A).

The first hit in order of position along the genome, at position 1413–1416, is a synonymous 4-mer variant. A change from the most common 4-mer AGCT to the variants AGCC or GGCT was found to be associated with changes in GSVL by -0.68 or $-0.19 \log_{10}$ copies/ml, respectively (P -value 1.2×10^{-6}). The variation in this 4-mer overlapping the amino acids 208 and 209 in *gag* (coding for the p24 capsid protein) does not alter the two resulting amino acids (glutamic acid-alanine, GAA GCT in the reference).

The second hit, at position 1514, is a nonsynonymous variant. A change from the most common base C to an A (GagT242N) or to a G (GagT242S) was found to be associated with changes in GSVL of -0.26 or $-0.039 \log_{10}$ copies/ml, respectively (P -value 2.1×10^{-8}). This is an epitope mutation (Leslie et al. 2004) allowing escape of the host HLA-B*57 allele associated with suppression of viremia (Kaslow et al. 1996). The T242N B*57 escape mutant found in HLA-B*57-positive hosts reverts upon transmission to HLA-B*57-negative hosts (Leslie et al. 2004).

The third hit, at position 6570–6574, is a nonsynonymous 5-mer variant. A change from the most common 4-mer CTAAG to the variants CTACA, CTGAA, TTAAA, or TTGAA was associated with changes in GSVL of -0.15 , -0.056 , 0.022 , or -0.87 , respectively (P -value 7.4×10^{-5}). The variation in this k-mer overlapping the amino acids 116 and 117 in *gp120* alters the two resulting amino acids (leucine-lysine, CTA AAG in the reference)

to leucine–glutamine, leucine–lysine, leucine–lysine, and leucine–lysine, respectively, in the variants. This hit was significant only for GSVL, and we predicted very little power to replicate it (power 0.06).

The fourth hit, at position 9008, is a nonsynonymous variant. A change from the most common base G to A (NefR71K) or C (NefR71T) was associated with changes in GSVL of $+0.22$ or $+0.11 \log_{10}$ copies/ml, respectively (P -value 1.4×10^{-5}). The variant NefR71T escapes the host HLA-B*35:01 allele (Ueno et al. 2007, 2008).

The epidemiological and genomic variant effect sizes on GSVL in the re-optimized model with all significant k-mer variants are presented on Table 1 (see Supplementary Tables 2–4 for other phenotypes).

We could also use our linear mixed model to estimate the heritability of the three viral load phenotypes estimated here as 0.26 [95% CI, 0.17 – 0.35], 0.25 [0.15 – 0.34], and 0.24 [0.15 – 0.32] (Table 2). Although different measurements contribute to GSVL and SPVL, they are highly correlated (Pearson coefficient of correlation $\rho=0.84$). Both SPVL measures are very highly correlated as they differ only by the normalization step ($\rho=0.99$). The CD4 slope had lower heritability, at 0.009 [0 – 0.021]. These results are in line with other recent analyses (Blanquart et al. 2017, Bertels et al. 2018, Mitov and Stadler 2018, Bastide et al. 2020). Moreover, using an alternative description of genetic effects based on a principal component analysis as in Wymant et al. (2022), we found that the first 100 principal components explained 15% of the variance in GSVL. The principal component corresponding to the recently detected hypervirulent variant ‘VB’ (Wymant et al. 2022) alone explained 1.4% of variance in GSVL.

We then checked whether SNPs allowing escape to host CTLs or mutations conferring resistance to antiretroviral drugs were more likely to be significantly associated with GSVL. In spite of the two clearest hits being known epitope mutations, we found overall no associations between these functions and the significance of their inferred effect on GSVL (Fig. 2C). The lack of enrichment for drug resistance mutations is consistent with the absence of association in the GWAS between any virulence phenotype and the predicted drug resistance.

Replication in two independent datasets

We attempted to replicate our four hits in two additional datasets, with mixed results. One additional dataset consisted of the 598 last seroconverters included in the BEEHIVE study, with HIV sequences and viral loads generated subsequently and, for purely logistical reasons, with different methods. The other was published HIV amino-acid sequence and viral load data from 2122 seroconverters enrolled in the INSIGHT START study (Gabrielaite et al. 2021).

The four hits were not replicated in the BEEHIVE additional study ($N=323$). For the first hit, at position 1413–1416, only the alternative allele GGCT could be tested and associated with a reduction in GSVL of $-0.098 \log_{10}$ copies/ml (P -value .4). The second hit, the variant C1514A (GagT242N), was associated with an increase in GSVL of $+0.026 \log_{10}$ copies/ml (resp. reduction $-0.014 \log_{10}$ copies/ml for C1514G) (P -value = .98). For the third hit, at position 6570–6574, the alleles present in the additional dataset were not the same as for the main dataset (the resulting low power to replicate, 0.06, was anticipated by the power analysis) and did not have a significant effect on GSVL (P = .59). The fourth hit, the variant G9008A (NefR71K), was found to be associated with a decrease in GSVL of $-0.21 \log_{10}$ copies/ml (P -value = .082), an effect opposite to that of the discovery data. The fact that the

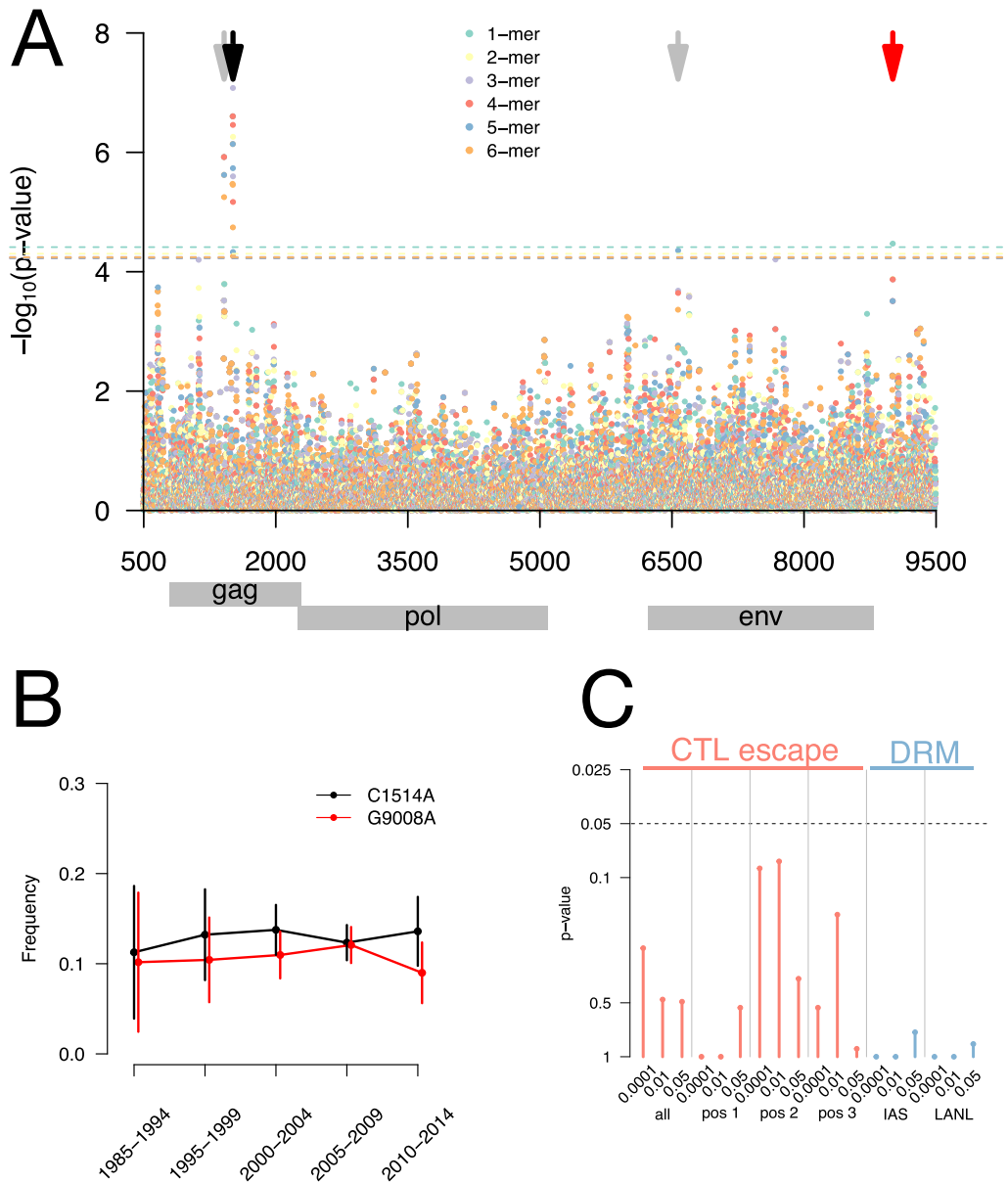


Figure 2. (A) Manhattan plot for the GWAS of the ‘gold standard’ viral load (GSVL) phenotype against all k-mer variants, from 1-mer (SNP) to 6-mer. The x-axis shows the position in base pairs along the HIV genome. The y-axis shows the negative \log_{10} -transformed P-value. The dashed horizontal lines show the Bonferroni-corrected threshold for significance at the 0.05 level. Four variants (at positions 1413, 1514, 6570, and 9008, marked with arrows) exceeded the threshold. Two of them are known CTL escape mutations (black and red arrows). (B) Frequency of the two CTL escape mutations over time in the main data, with vertical bars indicating the 95% binomial confidence intervals. (C) Barplot of the P-values of the enrichment analysis, in which we tested the association between significance of each position in the GWAS (defined at thresholds 0.0001, 0.01, and 0.05) and the CTL escape (red) or drug resistance mutations (blue) phenotypic properties of SNPs. The dashed line is $P = .05$ after correction for multiple testing.

first and third hits were not associated with a known biological function and that the variation at these loci was neutral (first hit) or largely different in the additional dataset (third hit) suggest that these were probably false positive.

In the main dataset, we also tested for an association between the fraction of the genome composed of rare point mutations (mutations present in our sample population at a frequency < 5%) and the phenotypes. Although having a greater number of rare point mutations was strongly associated with low viral load in the main data (Supplementary Figure 1), this finding was not replicated either (Supplementary Information).

With the amino-acid sequences of the INSIGHT START study, only the variants C1514A (GagT242N) and G9008A (NefR71K) could be tested, and only the effect of C1514A was replicated.

Associations were previously inferred with the PLINK2 software (Gabrielaitė et al. 2021), with the following covariates: age, sex assigned at birth, and the first four principal components of the genetic data from both human and viral populations to account for population structure. Nef71K was associated with a decrease in viral load of $-0.019 \log_{10}$ copies/ml change in viral load ($P = .36$), while Gag 242 N was associated with a decrease in viral load of $-0.065 \log_{10}$ copies/ml ($P = .00061$). Of note, none of the four variants significantly associated with viral load in the INSIGHT START study (Pol980E (integrase), Tat53R, Rev7D, and Env571W) were significant in our GWAS.

Accordingly, the phenotypes in the additional BEEHIVE dataset were poorly predicted by a polygenic score built from the hits (significance at <0.05) (Supplementary Table 5). A polygenic score

Table 1. Fixed effects estimated in the final re-optimized model for GSVL, and 95% confidence intervals. We kept only the k-mer variants in the linear model. The reference factors were female; age category 0–19; pegivirus absent; hepatitis absent; and sample type unknown. The k-mers are denoted by the start and end of the position in the alignment and the allele numbers. The 3-mers at positions 1124–1126 and 7676–7678 could not be replicated and are therefore not mentioned in detail in the main text

| Fixed effect | Estimate [95% CI] |
|-------------------------------------|----------------------------------|
| Intercept | 4.7 [4.47; 4.93] |
| Male | 0.147 [0.0534; 0.241] |
| Age category 20–39 | 0.13 [−0.0685; 0.329] |
| 40–59 | 0.218 [0.0143; 0.422] |
| 60+ | 0.206 [−0.0688; 0.481] |
| Unknown | 0.837 [−0.00167; 1.68] |
| Pegivirus present | −0.125 [−0.206; −0.0439] |
| Hepatitis present | −0.0309 [−0.195; 0.133] |
| Time to sample (days) | −0.000404 [−0.000599; −0.000209] |
| Sample type plasma | −0.0611 [−0.137; 0.0144] |
| Sample type serum | 0.041 [−0.102; 0.184] |
| kmer_1 124_1 126 (Ref: AGA) | |
| kmer_1 124_1 126_1 (AAA) | −0.0749 [−0.274; 0.124] |
| kmer_1 124_1 126_2 (AAC) | −0.126 [−0.296; 0.0442] |
| kmer_1 124_1 126_3 (AG) | −0.38 [−0.791; 0.031] |
| kmer_1 124_1 126_4 (AGC) | −0.176 [−0.256; −0.0963] |
| kmer_1 124_1 126_5 (AGG) | −0.235 [−0.435; −0.0351] |
| kmer_1 124_1 126_6 (GCC) | 0.131 [−0.233; 0.495] |
| kmer_1 124_1 126_7 (GGA) | −0.0269 [−0.367; 0.313] |
| kmer_1 124_1 126_8 (GGC) | −0.717 [−1.1; −0.33] |
| kmer_1 124_1 126_9 (pooled) | 0.00334 [−0.186; 0.192] |
| kmer_1 413_1 416 (Ref: AGCT) | |
| kmer_1 413_1 416_1 (AGCC) | −0.836 [−1.24; −0.43] |
| kmer_1 413_1 416_2 (GGCT) | −0.106 [−0.195; −0.017] |
| kmer_1 413_1 416_3 (pooled) | −0.307 [−0.594; −0.0204] |
| kmer_1 514_1 514 (Ref: C) | |
| kmer_1 514_1 514_1 (A) | −0.223 [−0.311; −0.135] |
| kmer_1 514_1 514_2 (G) | 0.0446 [−0.177; 0.266] |
| kmer_7 676_7 678 (Ref: TAA) | |
| kmer_7 676_7 678_1 (CAA) | 0.0744 [−0.0397; 0.189] |
| kmer_7 676_7 678_2 (TAG) | −0.196 [−0.463; 0.0714] |
| kmer_7 676_7 678_3 (pooled) | −0.43 [−0.929; 0.0692] |
| kmer_9 008_9 008 (Ref: G) | |
| kmer_9 008_9 008_1 (A) | 0.188 [0.101; 0.275] |
| kmer_9 008_9 008_2 (C) | 0.142 [−0.132; 0.416] |

Table 2. Heritability of each phenotype

| Phenotype | Heritability [95% CI] |
|--------------------------|-----------------------|
| GSVL | 0.26 [0.17–0.35] |
| SPVL adjusted | 0.25 [0.15–0.34] |
| SPVL adjusted normalized | 0.24 [0.15–0.32] |
| CD4 slope | 0.009 [0.−0.021] |

built from an expanded list of significant variants (significance <0.2) did not perform well either.

Finally, we investigated whether the effect of all variants estimated in the main versus additional datasets was positively correlated. We conducted this analysis focusing on SNPs for the comparison with additional BEEHIVE data and for amino-acid variants for the comparison with INSIGHT START data. We consistently found weak positive correlations between effects with varying degrees and levels of significance across the three viral load phenotypes and CD4 slopes (Table 3). These correlations substantiate that the inferred effects, although too weak to reach genome-wide significance, represent genuine weak biological effects of mutations on phenotypes.

To interpret further our findings, we conducted a GWAS on simulated data varying the number of loci in an attempt to reproduce the poor performance of the polygenic score and the positive association in effect sizes (Material and Methods, Section 4.6). When the cumulative effects of loci (beyond the correction for population structure) explains about 10% heritability, and when there are 200 and 500 loci explaining this heritability, it is not an unexpected outcome to find no success of the polygenic score (P -value > .05) but a weak correlation between effect sizes in discovery and additional datasets (P -value < .05) (Supplementary Figure 2). Thus, the polygenic architecture and small effect sizes may be the main explanation behind our findings.

Alternative genome-wide association study based on lasso regression

As the mixed success of our main GWAS may be explained by the small heritability beyond the population structure, and the polygenic architecture of the phenotypes, we last developed an alternative GWAS based on lasso regression on SNPs and k-mer variants (Material and Methods, Section 4.7). Lasso regression does not explicitly correct for population structure, instead directly capturing the variants best explaining the phenotype of

Table 3. Regression coefficients β of effect sizes in additional versus main dataset, over all SNPs tested. The left columns show the BEEHIVE additional dataset; the right columns show the INSIGHT START additional dataset. The INSIGHT START data did not include the phenotype CD4 slope. The *P*-values correspond to two-sided *t*-tests. The dash symbol indicates “not applicable”.

| Phenotype | β (BEEHIVE additional) | <i>P</i> | <i>R</i> ² | β (INSIGHT START) | <i>P</i> | <i>R</i> ² |
|--------------------------|------------------------------|----------|-----------------------|--------------------------|----------|-----------------------|
| GSVL | 0.100 [0.0119; 0.189] | .0262 | 0.18% | 0.0130 [−0.0128; 0.0389] | .323 | 0.08% |
| SPVL adjusted | 0.0796 [−0.0321; 0.191] | .163 | 0.08% | 0.0321 [0.0129; 0.0513] | .00106 | 0.87% |
| SPVL adjusted normalized | 0.109 [−0.00922; 0.228] | .0707 | 0.14% | 0.0288 [0.0103; 0.0474] | .00233 | 0.75% |
| CD4 slope | 0.0917 [−0.0197; 0.203] | .106 | 0.12% | – | – | – |

interest (Lees *et al.* 2018). This alternative approach confirmed the high heritability of viral load traits and low heritability of CD4 slopes in our dataset. The heritability was quantified by the coefficient of determination (R^2) for the predicted versus true phenotype in the training data (a random subset of 80% of individuals in the BEEHIVE main dataset). This coefficient was 23% for viral load phenotypes (Fig. 4A) and 2% for the CD4 slope. The lasso regression led to a better prediction of the phenotype in the cross-validation sample (a random subset of 20% of individuals in the BEEHIVE main dataset) ($R^2 = 3.6\%–6.0\%$ for viral load, Fig. 4B). The lower R^2 than for the training suggests contributions of false positives, possibly compounded with additional heterogeneity brought about by population-dependent effects or epistasis and varying patterns of linkage disequilibrium. The prediction was very poor for the BEEHIVE additional dataset ($R^2 = 0.49\%–2.4\%$, Fig. 4B), although the correlation was still often significant (Fig. 4C). This suggests that the heterogeneity in measure of phenotype, sequencing method, and represented countries in the additional BEEHIVE data (Section 4.1.4), further reduced the portability of the polygenic score. Lastly, the lasso regression captured on average 260–280 loci explaining viral load phenotypes, and only 12 explaining CD4 slope (Fig. 4D). The variants often captured in the lasso models on viral load included the two hits identified in the main GWAS and dozens of other variants (Supplementary Tables 6–9). Of note, in this approach, the CTL escape mutations were over-represented: while 23% of sites of the HXB2 reference are annotated as CTL escape mutations, about half the variants captured by the lasso included CTL escape sites.

Discussion

Using whole-genome HIV sequences associated with detailed phenotypic data in an international cohort collaboration of HIV seroconverters in Europe, we conducted a GWAS searching for viral genetic determinants of virulence traits. We confirm with a larger dataset and distinct methods the substantial heritability of viral load phenotype, at $h^2 = 0.26$ [95% CI, 0.17–0.35], and the lower heritability of the CD4 slope, at $h^2 = 0.009$ [0–0.021]. Heritability of the CD4 slope was previously estimated at around 0.11, with a wide confidence interval also encompassing 0. The viral heritability of the CD4 slope has only been measured in BEEHIVE and in one independent dataset [at 17% (Bertels *et al.* 2018)], and further work on this phenotype would be needed. Previous estimates of heritability were done either by correlating the viral loads within a small number of identified transmission pairs or for larger datasets using models of phenotype evolution on a phylogenetic tree. Our linear mixed model directly links the phenotype with variation in the viral genomes, using a different set of assumptions on the expected covariances between phenotypes of related individuals; arriving at a similar estimated

heritability adds confidence to previous work. We also report a set of negative findings: no detected effect of within-host minor frequencies, of insertions and deletions, no burden of rare point mutations, and no effect of predicted transmitted drug resistance, on any of the viral phenotypes.

However, we found two CTL escape mutations: the C1514A (GagT242N) mutation associated with a reduction in GSVL of $-0.26 \log_{10}$ copies/ml, and the G9008A mutation (NefR71K) associated with an increase in GSVL of $+0.22 \log_{10}$ copies/ml in our main dataset ($N = 2294$ seroconverters). The attempted replication of their effects gave mixed results. They were not replicated in our additional BEEHIVE dataset ($N = 323$ seroconverters), but the effect of the GagT242N mutation was replicated in the INSIGHT START additional dataset. To further investigate the effect of these two CTL escape mutations, we estimated their association with viral load using a third small additional dataset comprising the subset of individuals from the Swiss HIV Cohort Study studied in Bartha *et al.* (2013), excluding those who were already included in the BEEHIVE study ($N = 119$). The inferred effects of the two variants were aligned with those inferred from our GWAS but did not reach significance in this small dataset (Supplementary Table 10). In spite of mixed results from the additional datasets, we have further indications from the literature that these hits have a genuine biological effect. The variant C1514A (GagT242N) was previously identified because it was associated with suppression of viremia (Kaslow *et al.* 1996). It is strongly associated with the host HLA allele HLA-B*57 and reverts upon transmission to a non-HLA-B*57 host, suggesting N242 confers a fitness cost (Leslie *et al.* 2004). This association with host HLA alleles was replicated in two recent GWASs (Bartha *et al.* 2013, Gabrielaite *et al.* 2021), as was the negative effect on viral load (Fig. 3, right panel). Specifically, in these two recent GWASs, this variant was identified because it was strongly associated with HLA alleles. The overall negative effect on viral load was detected but not significant after correction for multiple testing in these studies. However, focusing on the effect of this SNP stratified by host HLA type revealed that the SNP was associated with lower viral load in the vast majority of individuals without B*57:03 or B*57:01 HLA allele and higher viral load in the individuals with B*57:03 or B*57:01 HLA alleles (Gabrielaite *et al.* 2021). This further confirms the cost of this mutation in non-B*57 hosts. The effect of the second hit (the G9008A mutation or NefR71K) on viral load is less documented. It was also previously identified as strongly associated with specific host HLA alleles (Bartha *et al.* 2013; Gabrielaite *et al.* 2021); in one previous study, it had a positive association with viral load, again not significant at the genome-wide level (Bartha *et al.* 2013). In the other (the INSIGHT START additional dataset, Fig. 3), the effect was weak and opposite.

Our mixed results regarding the replication of hits and the portability of the polygenic score can be attributed to several

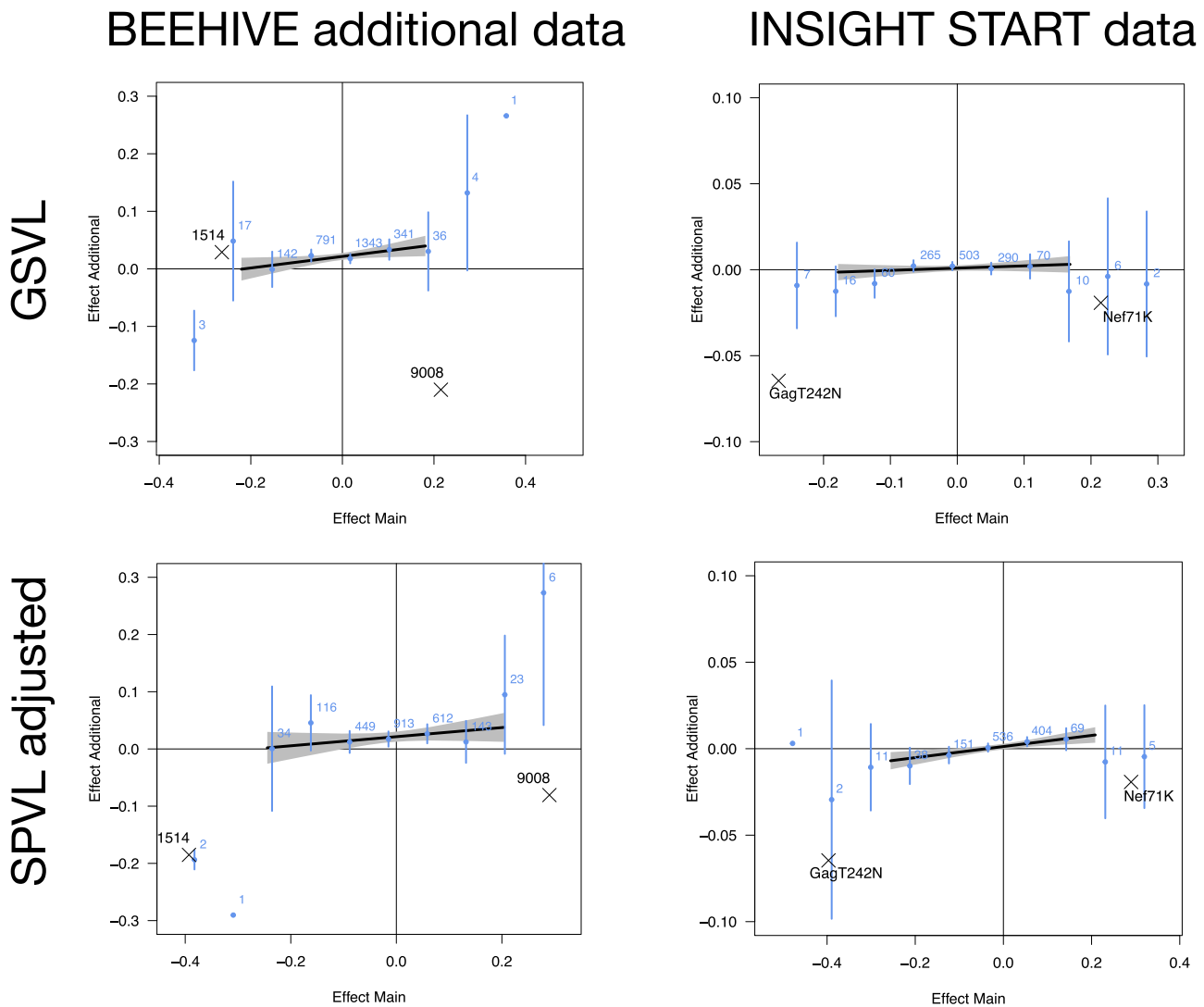


Figure 3. Effect sizes on GSVL (top panels) and SPVL adjusted (bottom panels), in main versus BEEHIVE (left panels) and INSIGHT START (right panels) additional datasets. The line and shaded area show the regression coefficient with 95% interval. The blue points and segments show the means and standard errors of the effect sizes in additional data, in 10 evenly spaced intervals. The numbers in blue show the number of effects represented by each point. The two SNP hits are shown as crosses (positions 1514 and 9008 in BEEHIVE, GagT242N and Nef71K in INSIGHT START).

factors. First, the statistical power to replicate the effects of the two CTL escape mutations in the additional BEEHIVE dataset was limited, ranging from only 45%–50% depending on the viral load phenotype analysed. Moreover, the impact of these CTL escape mutations on viral load may be influenced by the host's HLA alleles (Gabrielaite et al. 2021). Since the additional dataset comprised individuals from different countries than the main dataset (Supplementary Table 11)—and given that HLA allele frequencies vary geographically—this likely further reduced our power to replicate the findings. Consistent with the failure to replicate these hits, the polygenic score constructed from them showed poor predictive performance. Nevertheless, effect sizes of all variants in the discovery and additional datasets showed a weak correlation. These observations were compatible with a model where the amount of heritability that is not already explained by the population structure is around 10% and the trait is polygenic (200–500 loci). To better understand the limitations affecting replicability and portability, we applied a complementary approach using lasso regression. The polygenic score constructed from lasso regression performed well on the training

dataset (confirming the heritability of viral load) and less well on the cross-validation dataset—composed of 20% of the main data—again likely due to false positives and/or overfitting. Its performance further declined on the additional BEEHIVE data, especially for the GSVL phenotype. This decline can be explained by data heterogeneity, including differences in sequencing methods (see Section 4.1.4) and phenotype measurements for GSVL. Biological factors also likely contributed, such as population-specific effects and variations in linkage disequilibrium across populations (Peterson et al. 2019). Additional studies would be needed to establish our results; however, the set-point viral load is now harder or impossible to measure thanks to widespread testing and rapid initiation of treatment after diagnosis. Our study might remain one of the largest studies of the genetic architecture of virulence in the years to come.

Our study had some limitations. First, we did not have access to the host genotype. Indeed, the main ambition of our project was to identify viral variants affecting virulence independently of the host genotype. It turned out that the two significant hits were CTL escape mutations, with effects potentially strongly dependent on

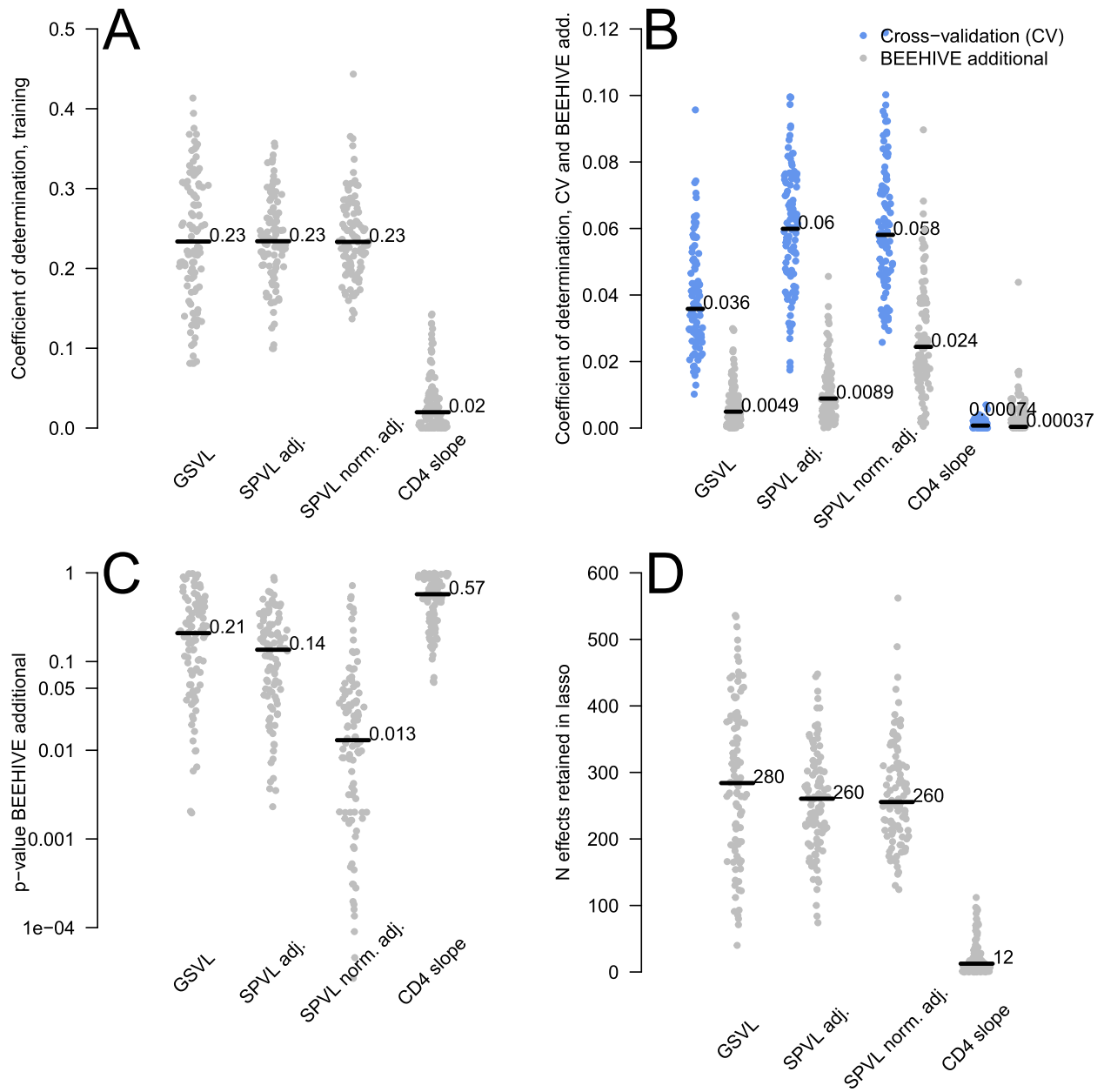


Figure 4. Results from the GWAS based on lasso regression. (A) The coefficient of determination for the predicted versus true phenotype in training data, for each phenotype. Each point shows results from one of the 100 random splits of the main BEEHIVE data (representing 80% of main data). The line shows the median, and the number shows the value of this median. (B) The coefficients of determination for the predicted versus true phenotype in cross-validation data (blue, 20% of main data) and in the additional BEEHIVE dataset (grey). (C) Level of significance of the correlation between predicted versus true phenotype in the additional BEEHIVE data. (D) Number of variants captured in the final model for each phenotype.

host genetics. This could explain our inability to replicate our findings in additional datasets. However, even in the face of this strong host dependence, these two variants had a significant effect on viral load on average in the population, which had not been detected in previous studies. Second, in our main GWAS, we could only detect variants not strongly associated with the HIV population structure. This considerably limits the power to discover variants but is necessary to establish a potential causal effect of a variant on the phenotype of interest. For example, our main dataset comprises the smaller dataset in which we previously identified a hypervirulent strain (increasing viral load and accelerating CD4 decline) (Wymant *et al.* 2022). In the main GWAS, we did not discover any significant SNP characterizing this hypervirulent strain because any lineage effect was included

in the random effect, correcting for population structure. The effect of the hypervirulent strain contributed to the measured heritability. The lasso GWAS, in contrast, captures both causal variants and variants associated with the population structure, which limits the causal interpretation.

What are the implications of our discoveries for the evolution of HIV virulence? Within hosts, CTL escape mutations may evolve and carry a cost that reduces replication. When transmitted to other hosts with different HLA types, these viruses with costly mutations have a lower SPVL (van Dorp *et al.* 2014). Between hosts, viral load may evolve to an optimal value under a trade-off between transmissibility and the duration of the asymptomatic phase of infection (Fraser *et al.* 2007). Under which of these two evolutionary forces do the variants we discovered evolve? On the

one hand, the evolution of the two CTL escape mutations with a significant effect on viral load is certainly driven by within-host selection. The stable frequency of these two mutations is compatible with predictions from models of evolution of such mutations with nonzero reversion rates (Fryer et al. 2010). The C1514A (GagT242N) escape mutation results in a negative average effect on viral load because the corresponding host HLA allele is rare and reversion may be slow enough that this costly mutation is still found mainly in individuals ‘not’ carrying the HLA allele (Ueno et al. 2008). The average positive impact of G9008A (NefR71K), on the contrary, could be explained if this mutation is found mainly in hosts carrying the matching allele, which would happen if reversion is fast. The other CTL escape mutations captured by the lasso GWAS could evolve under similar forces. In addition, the many mutations contributing to viral load heritability could evolve under the transmission–virulence trade-off. It has been argued that within-host selection of escape mutations is the dominant mode of HIV evolution, while the trade-off at the between-host level is irrelevant (van Dorp et al. 2014). However, the model inspiring this claim did not allow the evolution of mutations solely affecting viral load but only of CTL escape/reversion mutations. We find significant heritability of viral load, and many variants identified by the lasso GWAS are not CTL escape mutations. This would suggest that many mutations of small effect influence viral load independently of a role in immune escape and could drive the viral load to an optimal level resulting from the trade-off. For example, a model with hundreds of deleterious mutations reducing viral load enables evolution to the optimum predicted by the trade-off (Longley et al. 2025). In support of the trade-off hypothesis, the rate of branching at the tips of the phylogenetic tree of HIV subtype B (a subset of the present data), a proxy for viral fitness, was largest for viral sequences associated with an intermediate viral load (Zhao et al. 2022). A plausible dynamical model of HIV virulence evolution would include both CTL escape mutations and mutations impacting replicative capacity, evolving under the action of selection within the host (in particular but not exclusively by the immune system), at the transmission bottleneck (Carlson et al. 2014), and under the transmission–virulence trade-off.

Future GWAS of HIV virulence could focus on larger datasets that include both viral phenotypes and genomes and host HLA information to uncover more gene-by-gene (GxG) interactions. This would provide further insights into the interplay between HIV antigens and the host immune system. While our estimate of heritability of viral load at 20%–30% seems robust, attempting to resolve the problem of ‘missing heritability’ would identify individual variants with very small effects on viral load and may not offer valuable mechanistic insights. It may be more interesting to develop methods that use both GWAS results and longitudinal genomic datasets to infer the adaptive evolution of polygenic traits like viral load.

In summary, this is the largest and most detailed GWAS of HIV virulence phenotypes to date, compared with several large independent datasets. We confirm the substantial heritability and the polygenic nature of viral load, thanks to new evidence from mixed linear models and from comparisons of the inferred effects of variants across independent datasets. We confirm two CTL escape mutations significantly associated with viral load, with effects that are probably host-population dependent. The genetic architecture of viral load challenges our ability to identify mechanisms of pathogenesis, except for CTL escape variants. However, our findings imply that HIV-1 virulence is a quantitative trait with large genetic variation and can readily evolve. It remains largely

Table 4. Number of HIV seroconverters included in the BEEHIVE study, by country

| Country | N |
|-------------|------|
| Belgium | 54 |
| Finland | 31 |
| France | 418 |
| Germany | 357 |
| Netherlands | 786 |
| Switzerland | 1052 |
| Uganda | 209 |
| UK | 797 |
| Total | 3704 |

open to investigation (Rindler et al. 2022) how this trait will evolve when the time from infection to treatment is reduced, thanks to ART, and may be highly heterogenous in the host population because of disparities in access to ART.

Material and methods

Epidemiological data from European and Ugandan cohorts

Inclusion criteria

We selected HIV seroconverters from the Antwerp cohort in Belgium, the Swiss HIV Cohort Study in Switzerland (Scherrer et al. 2022), a cohort in Finland (Kivelä et al. 2005), the ANRS PRIMO Cohort in France (Harzic et al. 2002), the HIV-1 Seroconverter Study in Germany (Poggensee et al. 2007), the ATHENA cohort in the Netherlands (Dieleman et al. 2002), the Rakai Community Cohort Study in Uganda (Chang et al. 2016), and the UK Register of HIV seroconverters in the United Kingdom (Committee 1996)—all part of the BEEHIVE (‘Bridging the Evolution and Epidemiology of HIV in Europe’) collaboration (Table 4).

These were selected as the set of all individuals meeting the study’s inclusion criteria:

- (i) Individuals were seroconverters (i.e. the first positive test was less than 1 year after the last negative test or the individual presented with evidence of recent infection—laboratory evidence or seroconversion illness), ensuring the date of infection was well estimated.
- (ii) No antiretroviral therapy (ART) was taken in the first 6 months following the first positive test.
- (iii) At least one viral load or one sample from which viral load can be determined was taken between 0 and 24 months following the first positive test and before the start of ART.
- (iv) At least one sample of at least 500 μ l of frozen EDTA plasma or serum was taken between 0 and 24 months following the first positive test while antiretroviral therapy-naive. This was the sample on which ‘GSVL’ was measured.

All participating individuals consented to this study. All studies within the cohorts were approved by in-country institutional review boards. The overall BEEHIVE study, which only accessed anonymized data, was approved by the ethics panel of the European Research Council.

Number of HIV seroconverters per participating country

We initially identified 4253 putative HIV seroconverters from the European cohorts. Among these, 3704 seroconverters meeting the criteria remained, with a blood sample taken in the time

Table 5. Number of HIV seroconverters retained for the main GWAS and the additional analysis. The total number of individuals with enough genomic positions determined in the viral sequence alignment across all four phenotypes is 2294 for the main dataset and 323 for the additional dataset

| Phenotype | N | Main BEEHIVE dataset | | Additional BEEHIVE dataset | |
|--------------------------|------|----------------------|--------------------|----------------------------|--------------------|
| | | N in alignment | N enough positions | N in additional alignment | N enough positions |
| GSVL | 3016 | 2427 | 2249 | 434 | 322 |
| SPVL adjusted | 2960 | 2436 | 2247 | 348 | 252 |
| SPVL adjusted normalized | 2960 | 2436 | 2247 | 348 | 252 |
| CD4 slope | 2378 | 1967 | 1847 | 282 | 211 |

Table 6. Characteristics of the individuals in the main dataset. ‘1stQu’ abbreviates the first quartile and ‘3rdQu’ the third quartile

| GSVL | Year sampled | Country sampled | Transmission mode | Sex | Age at infection | Ethnicity | Subtype |
|-------------|--------------|------------------------|---------------------|-------------------|------------------|--------------------|---------------------------|
| Min: 2 | Min: 1985 | Belgium: 48 (2%) | HEAM/TRANSF: 4 (0%) | Female: 321 (14%) | Min: 17 | BLACK: 35 (2%) | 01_AE: 106 (5%) |
| 1stQu: 4.26 | 1stQu: 2003 | Finland: 30 (1%) | HET: 505 (22%) | Male: 1973 (86%) | 1stQu: 28 | OTHUNK: 1792 (78%) | 02_AG: 104 (5%) |
| Median: 4.7 | Median: 2007 | France: 340 (15%) | HET/IDU: 44 (2%) | | Median: 34 | WHITE: 467 (20%) | A: 134 (6%) |
| 3rdQu: 5.07 | 3rdQu: 2009 | Germany: 299 (13%) | IDU: 115 (5%) | | 3rdQu: 41 | | B: 1765 (77%) |
| Max: 7 | Max: 2015 | Netherlands: 435 (19%) | MSM: 1583 (69%) | | Max: 78 | | C: 46 (2%) |
| | | Switzerland: 939 (41%) | MSM/IDU: 1 (0%) | | | | D: 74 (3%) |
| | | Uganda: 39 (2%) | OTHUNK: 42 (2%) | | | | Other/unknown: 65 (3%) |
| | | UK: 164 (7%) | | | | | |

window 0–24 months after the first positive test, and therapy-naïve. Table 4 shows the breakdown by cohort.

Number of HIV seroconverters included in the genome-wide association study

Only a subset of the 3704 HIV seroconverters in BEEHIVE are included in the GWAS, shown in Tables 5 and 6.

First, we included only individuals with some sequence data and with no suspected contamination. Contamination—the incorrect inclusion of some sequencing output from one sample into the sequencing output of a different sample—was identified from examination of the sequences. Specifically, we excluded all amplicon regions with a median coverage lower than 200 or containing more than 20% of ambiguous nucleotides (‘N’). Additionally, 17 sequences that might have suffered from amplicon-based contamination were removed. Such contamination was first detected by visual inspection of the sequences, looking for diversity patterns similar to that generated by recombination with a breakpoint at an amplicon boundary, then confirmed with the Recombination Detection Program (Martin et al. 2010). Second, we included only individuals with single infections. Dual infection—having acquired two distinct viral strains—has previously been reported to be associated with increased viral load (van der Kuyl and Cornelissen 2007, Redd et al. 2013, Janes et al. 2015); 4.9% of individuals in our dataset were classified as dual from the reads using the software phyloscanner (Wymant et al. 2018b) and excluded from this analysis. Thirdly, additional inclusion criteria varied according to our separate analyses of the different phenotypes:

- (i) Gold standard viral load (GSVL): for analyses of this phenotype, we included all individuals with a nonmissing GSVL, with the exception of individuals whose GSVL was at the lower detection limit of the assay ($2 \log_{10}$ copies/ml) when

the set-point viral load was greater than $3 \log_{10}$ copies/ml for the same individual. The rationale for this was to eliminate low GSVL values due to a failure of the viral load assay.

- (ii) *Adjusted* set-point viral load: for analyses of this phenotype, we included all individuals with nonmissing set-point viral load.
- (iii) *Adjusted and normalized* set-point viral load: for analyses of this phenotype, we included all individuals with nonmissing set-point viral load.
- (iv) The slope of CD4 cell count decay: for analyses of this phenotype, we included all individuals with nonmissing CD4 cell count decay. For this phenotype, we considered only CD4 counts (not CD4 percentages) and measured a slope only when the individual had measurements from at least three different dates.

Following these steps of data selection, we obtained four overlapping subsets of individuals with data on the phenotype of interest and the covariates (column ‘N’ in Table 5).

We then considered the intersection of each of these four subsets with the HIV sequences present in the main genome alignment (column ‘N in alignment’ in Table 5). We further removed individuals with more than half of the nucleotide positions in the HIV sequence missing (column ‘N enough positions’ in Table 5). This was to avoid spurious high relatedness between individuals with many nucleotide positions in the HIV sequence missing, because such positions were set to the population mean when controlling for population structure in the GWAS (see below). These four subsets of the data together encompass 2294 individuals.

Number and characteristics of HIV seroconverters included in the BEEHIVE additional data

To replicate our results in an independent dataset, we sequenced the viral genomes from 598 additional HIV seroconverters. These

were the last individuals and associated samples included in the project. We applied the same steps of data selection as for the main data, eventually leading to 211–322 HIV genomes with associated phenotype in the additional data, depending on the phenotype (Table 5). For logistical reasons, the sequencing pipeline had to be restarted in a new laboratory over the course of the project. In consequence, the additional dataset differs in the sequencing technology, and has a distinct composition in terms of country of origin of the samples (Table 6, Supplementary Table 11). More precisely, we compared systematically the viral load, year sampled, country, transmission mode, sex assigned at birth, age, ethnicity and viral subtype in the main and additional datasets. The GSVL was $+0.19 \log_{10}$ copies/ml higher in the additional dataset (t-test, $N=2249$ and 323 , $P = 8 \times 10^{-5}$). Years sampled were 1.4 later in the additional dataset (Mann-Whitney U test, $P = 8 \times 10^{-7}$). We found no significant difference for sex and age category (chi-square test, P -values 0.972 and 0.64) and highly significant differences in country, transmission mode, ethnicity, and viral subtype (chi-square test, all P -values $< .001$). Notably, the countries Netherlands, Uganda, and the UK, and subtypes A, B, and D were over-represented in the additional dataset.

Sequence data

Sequencing and *de novo* assembly of contigs

For the main dataset, viral RNA was extracted from plasma or serum samples manually (Cornelissen et al. 2018). Extracted RNA was reverse-transcribed and amplified using a set of universal primers, then fragmented and sequenced using Illumina MiSeq or HiSeq 2500 technology generating paired-end reads (each pair recording the sequence at either end of a fragment of nucleic acid), which varied in length between 100 and 300 bp (Gall et al. 2012). IVA (Hunt et al. 2015) was used for *de novo* assembly of the reads into a set of contigs for each sample.

For the BEEHIVE additional dataset, RNA extraction, reverse transcription, amplification, fragmentation, and sequencing were performed using the laboratory component of the veSEQ-HIV method (Bonsall et al. 2020). SPAdes (Bankevich et al. 2012) was used for *de novo* assembly of the reads into a set of contigs for each sample.

Whole-genome reconstruction, within-host diversity, alignment

For each sample, we processed the reads and contigs with shiver (Wymant, Blanquart, et al. 2018). In summary, a custom reference for mapping was built for each sample using the contigs, filling gaps with a set of reference whole-genome sequences (Foley et al. 2018) for any genomic regions not covered by the contigs. Reads were mapped to these custom references using SMALT (Ponstingl and Ning 2010). At each nucleotide position, shiver recorded the number of mapped reads supporting each of the different bases, insertions, or deletions present here (capturing within-host minority variants) and also which one of these was most common (i.e. calling the consensus base). To reduce the low-level noise prevalent throughout next-generation sequencing, we masked (marked as missing) positions where the number of mapped reads was below a threshold. We set that threshold as 5X for the additional data, 30X for the main MiSeq data, and 300X for the main HiSeq data, following inspection of the extent to which the mean similarity between mapped reads and the mapping reference decreased as the depth of coverage decreased.

Translation, subtyping, phylogenetics

We used an in-house pipeline to translate the nucleotide sequence to an amino acid sequence for each gene (Golubchik 2018).

We subtyped each sequence using COMET (Struck et al. 2014). We corrected the assignment of subtypes CRF12_BF, CRF42_BF, or CRF17_BF to subtype B as described previously (Blanquart et al. 2017).

The phylogenetic tree (Fig. 1) was inferred by maximum likelihood with IQ-TREE, using the GTR + I + R model of substitution (Minh et al. 2020). The root of the tree was placed out of the clades defining the major subtypes and so as to maximize the R^2 of the root-to-tip distance versus date correlation.

Phenotypes

Outcome variables ('phenotypes') in the genome-wide association study

We performed a GWAS on each of the four phenotypes:

- (i) The 'gold standard' viral load (GSVL). This is the viral load re-measured using a standardized set of assays within the BEEHIVE study, on a single blood sample taken less than 24 months after the first positive HIV test and before the start of ART. If viral load had been previously measured with one of three assays (COBAS AmpliPrep/COBAS TaqMan HIV-1 Test, v2.0 from Roche; Abbott RealTime HIV-1 Assay from Abbott; Quantiplex HIV-1 RNA Assay, version 3.0 from Chiron Diagnostics, Emeryville, CA), on the same visit when the sample used to determine the viral sequence was taken, we did not repeat the assay. Otherwise, viral loads were repeated with COBAS AmpliPrep/COBAS TaqMan HIV-1 Test, v2.0 on the same sample used to determine the viral sequence. We defined the GSVL as \log_{10} of the single viral load (in copies per ml) measured in this way.
- (ii) The *adjusted* set-point viral load. The set-point viral load (SPVL) was obtained from the series of \log_{10} viral loads (in copies per ml) previously measured on the same individuals, in the time window between 6 and 24 months after the first positive HIV test and when the individual was therapy-naive. Viral loads were measured with a variety of assays and types of samples. We adjusted for these sources of heterogeneity as follows. For each individual, a linear model was fitted to the viral load measures of all individuals, as a function of a fixed effect indicating whether the measure is from the focal individual or not, a fixed effect adjusting for sample type, and a random effect adjusting for assay type. The adjusted set-point viral load was the inferred fixed effect of the indicator variable for each individual. This phenotype was then standardized to have mean 0 and standard deviation 1. We additionally recorded the number of viral load measures entering in the calculations of each SPVL to estimate the measurement error.
- (iii) The *adjusted and normalized* set-point viral load. This phenotype was computed as the adjusted set-point viral load, except that before adjustment, the distribution of all \log_{10} viral loads was transformed to obtain a standard normal distribution (of mean 0 and standard deviation 1). We defined this additional phenotype to test if heritability would be higher if the trait is transformed to better conform to the distribution resulting from the mixed model (normal distribution).
- (iv) The CD4 slope (the rate of CD4 cell count decline) quantifies the rate of an individual's disease progression. We selected

individuals who had at least three CD4 cell count measures between the date of the first positive HIV test and the date that ART was first prescribed. For each individual independently, we fitted a linear model by ordinary least-squares regression describing the decline in CD4 cell count over time. We recorded the slope of this relationship, as well as the standard error of the slope, to estimate the measurement error of this phenotype.

Phenotypes in the BEEHIVE additional dataset

The phenotypes in the BEEHIVE additional dataset were defined exactly as the phenotypes in the main dataset, with one exception. For 193 individuals, a GSVL measurement was not available, but a 'sequence-derived' viral load (arising from the quantitativity of the amplification in the protocol; see [Bonsall et al. 2020](#) for more details) was available. This was derived from the same sample used for sequencing (as were the GSVL measurements), and we used this instead.

Associating virulence phenotypes and genotypes Adjustment for covariates

For the adjusted SPVL (whether normalized or not) and CD4 slope phenotypes, we adjusted for the following covariates: sex, age category, presence/absence of pegivirus, and presence/absence of hepatitis. For the GSVL phenotype, we adjusted for those same covariates as well as time to sample (a continuous variable) and sample type.

In more detail, these covariates were:

- Sex: 1973 males and 322 females
- Age category (age at first positive test): $N = 47$ in [0, 20), 1550 in [20, 40), 640 in [40, 60), 53 in [60, 80) years old, and 4 unknown.
- Presence ($N = 327$) or absence ($N = 1967$) of pegivirus ([Stapleton 2022](#)). We assigned all reads of a sample to a taxon using Kraken ([Wood and Salzberg 2014](#)). A sample was set as 'pegivirus present' if it included at least one read assigned to pegivirus, and as 'pegivirus absent' otherwise.
- Presence ($N = 69$) or absence ($N = 2225$) of hepatitis. Hepatitis presence was detected as for pegivirus.
- Time to sample: the time from the date of the first positive test to the date of the sample, a continuous variable ranging from 0 to 730 days.
- Sample type: the type of the 'BEEHIVE' sample, used for sequencing and for the GSVL: 182 serum, 1755 plasma, and 357 other or unknown.

We did not adjust for the mode of transmission as a potential factor because it was very correlated with the phylogeny, as some genetic clusters correspond to specific modes of transmission. Thus, including mode of transmission as a covariate would bias downwards our estimate of heritability. In our cohort, mode of transmission was not associated with the CD4 slope, but it was associated with viral load: specifically, MSM have $+0.13$ to $+0.22$ (depending on the viral phenotype) higher \log_{10} viral load compared to individuals with heterosexual transmission.

Structure of the linear mixed model

We developed a custom GWAS pipeline designed for HIV, which allows testing for the effect of variants with more than two alleles at a given locus (unlike in human GWAS, notably), corrects for the population structure, and accounts for variable phenotypic measurement error (Supplementary Information). It rests on a linear mixed-effects statistical model that is applied in turn to

each locus. The fixed effects include the effect of the locus of interest and the adjustment for the covariates mentioned above in [Section 4.1](#). The random effects include the adjustment for population structure and the error variance. We chose the random effect to describe population structure after initial model comparisons showed that they capture more genetic variance and result in less inflation than description based on adding principal components as fixed effects in the linear model.

The random effects describing population structure are modelled as an independent, normally distributed effect on the phenotype from each allele at each locus ([Lippert 2013](#)). The error variance is adjusted for the varying level of precision on the phenotypic measure of different individuals. Although population structure can be characterized by the matrix of distances between genotypes via a phylogeny, we used here a matrix of similarities determined directly from the genomic sequences. This has the advantage of avoiding assumptions on the absence of recombination implied by the use of a phylogeny to describe evolutionary relationships. Furthermore, we did not include subtype as a covariate because it is already included in our control for population structure.

After optimizing the parameters describing the random effects in a first step, the statistical significance of each variant, in turn, is computed in a second step by comparing the linear models with and without the variant using a likelihood ratio test. We corrected for multiple testing with a Bonferroni correction based on the effective number of tests ([Supplementary materials](#)).

The heritability is calculated based on the optimized model of the first step (without variant), by estimating with resampling the fraction of variance in phenotype explained by the population structure (genetic effect). The 95% confidence intervals on heritability were estimated using the inverse of the Hessian estimated at the maximum likelihood (ML) parameters, estimating the variance-covariance matrix of the error. The lower and upper bounds on heritability were then computed by randomly resampling the random effects from the multivariate normal distribution of error estimated from the Hessian, computing the heritability in each random sample, and finally computing the 2.5% and 97.5% quantiles of heritability in the resulting distribution (parametric bootstrap).

The 95% confidence intervals on fixed effects were estimated using the standard deviation of error of the linear mixed model. This was approximated as the sum of two components. The first (the largest) is the standard deviation of error in the standard linear model (without random effects). The second is the standard deviation of ML fixed effects, across 200 replicates where we randomly resampled the random effects from the multivariate normal distribution estimated from the Hessian.

Enrichment analysis

To investigate whether positions with a significant effect on phenotypes are enriched in variants affecting escape to the host immune system or drug resistance, we conducted an enrichment analysis. We did this analysis only for the GSVL phenotype and for SNP variants. We analysed the association between a position having a significant effect, and the SNP variant at this position being a CTL escape mutation, or a drug resistance mutation. The significance of the association was computed by Fisher's exact test on the contingency table. We did so for SNP significance thresholds $P = .0001$, $P = .01$, $P = .05$. For CTL escape mutations, we looked at all positions and the first, second, and third positions in codons. We used the classification of CTL escape mutations inferred from associations between HLA types and HIV-1 subtype B sequence

polymorphisms (Carlson *et al.* 2012). For drug resistance mutations, we looked both into the classification of the International Antiviral Society (IAS) (Wensing *et al.* 2015) and the Los Alamos National Laboratory HIV Database ('HIV Databases' n.d.).

Selection of significant variants and inference of effect size

Short list of significant variants

The GWAS conducted for all combinations of genetic variant and phenotype resulted in a list of variants with a significant effect at the 0.05 level. For each phenotype, we reduced this list of variants by removing those that were redundant (e.g. SNP and 2-mer variants can describe variation at the same position in slightly different ways), favouring the shortest variant and nucleotide over amino acid variants. For the burden of rare mutations, we selected the frequency threshold that was found to be associated with the lowest *P*-value. We finally re-optimized the linear model with all short-listed variants included as covariates and decomposed the variance in phenotype explained by each epidemiological and genetic covariate. We constructed the linear model to estimate the polygenic score using the estimated effect sizes associated with epidemiological covariates and genetic variants in this final linear model. We also tested a polygenic score constructed from the similar re-optimized linear model when the *P*-value threshold was 0.2 (instead of 0.05).

Checking the correction for population structure

As an indicator of sufficient correction for population structure, we drew quantile–quantile plots of the distribution of *P*-values against the theoretical distribution under the null hypothesis that none of the variants have a genuine effect on phenotype (*P*-value is uniform in [0, 1]). We also summarized the similarity of the two distributions by computing the 'genomic inflation factor' (or lambda value), obtained by dividing the median value of the observed chi-squared statistic by the median expected chi-squared statistic expected under the null hypothesis (Power *et al.* 2017). This inflation factor should be 1 in the absence of inflation and should, in principle, not exceed 1.05. For SNPs, the inflation factor was 0.98, 0.99, 1.00 for the three viral load phenotypes, indicating slight conservatism, and 1.04 for the CD4 slope, indicating slight inflation. For other variants (length variants, amino acid variants, within-host frequency variants), it was between 0.91 and 1.31.

Power to replicate the findings in the additional BEEHIVE dataset

We computed the power to replicate the effects of each of the short-listed variants ('hits') in the additional dataset. For simplicity, we computed power based on simulating a simpler linear model that includes only the effect of the focal variant. We built pseudo-datasets where the effect sizes were as inferred from the GWAS, and the error variance was the phenotypic variance. We computed the true-positive rate of the focal variant as the fraction of simulations that identified the variant as having a significant effect at the 0.05 level (without correction for multiple testing). When the focal variant tested consisted of a reference allele and a single alternative allele, we used a one-sided *t*-test (with an effect in the same direction as observed in the main analysis) to improve power.

Analysis of the additional dataset

For the additional BEEHIVE dataset, we used the GWAS pipeline described above. We first specifically tested only the effects

of the hits identified in the main dataset that we had some power to replicate (Section 2.2). In the GWAS on the additional data, we did not apply a correction of the *P*-value to account for multiple testing. Second, we correlated the effect of each SNP in the main GWAS with that in the additional data, for the SNPs that have inferred effect in both datasets, using linear regression (Section 2.2, and Table 3). Third, we used the re-optimized linear model from the main GWAS (5.1) to construct a polygenic score that we correlated with the true phenotype (Supplementary Table 5, Section 2.2).

Analyses of the INSIGHT START dataset

We also compared our findings to those of the GWAS already conducted on the INSIGHT START data, analysed with the PLINK2 software (Cc *et al.* 2015, Gabrielaite *et al.* 2021). Details of the cohort, the viral sequencing and bioinformatic techniques, the measurement of phenotypes, and the GWAS method were reported in Gabrielaite *et al.* (2021). Briefly, these data represent individuals living with HIV spanning five continents and 23 countries. Compared to our BEEHIVE dataset (Table 6), individuals in this cohort had similar age distribution (median 34, interquartile range 29–45 years old), sex distribution (female–male 17%–83%), mode of transmission (2% IDU, 29% HET, 68% MSM) (Table 1 in Gabrielaite *et al.* 2021). However, subtype B was less frequent (64% of sequences) and viral load was lower (4.29 log₁₀ copies/ml, interquartile range [3.81–4.71]) in INSIGHT START data than in BEEHIVE. HIV-1 viral load was associated with HIV amino-acid variants. The following covariates were adjusted for in the linear model: age, sex assigned at birth, and the first four principal components of the genetic data from both human and viral populations to account for population structures. They identified four amino-acid variants significantly associated with viral load at the 0.05 level after Bonferroni correction: Pol980E (integrase), Tat53R, Rev7D, and Env571W (gp41) (the reference being HXB2). First, we identified the association of Nef71K and T242N variants with viral load in this previous analysis (Section 2.2). Next, we matched all amino-acid variants tested in INSIGHT START with all amino-acid variants tested in BEEHIVE, resulting in 1230 amino-acid variants with inferred associations with viral load in both INSIGHT START and BEEHIVE (Section 2.2). We correlated these associations across the two datasets with linear regression (Table 3). For the BEEHIVE main dataset, we did so for associations inferred from each phenotype in turn: GSVL, SPVL adjusted, and SPVL adjusted normalized.

Simulation model to interpret of the failure to replicate our findings in additional datasets

To support our interpretation of the nonreplication of hits and the absence of portability of polygenic score, we ran a GWAS on simulated data. In simulations, we assumed all covariates, as well as population structure, are properly adjusted and can be omitted from these simulations. We modelled 2250 viral genomes in the discovery GWAS and 322 viral genomes in the replication GWAS, as for our main GWAS on GSVL. We modelled 5530 polymorphic loci, corresponding to the number of SNP variants tested. Among these 5530 loci, we modelled that from 10 to 500 loci have a genuine effect on the phenotype. The effects of these loci contribute additional heritability on top of the heritability associated with the population structure modelled by the random effect. For each locus, we drew the frequency of the variant in a beta distribution mimicking the true frequency of variants in the dataset. The effect size of each locus was drawn in a normal distribution with standard deviation $1/\sqrt{n_{\text{loci}}}$, ensuring the total

genetic variance explained by the loci remains constant across different number of loci tested. The environmental and error components of the phenotype were normally distributed with a standard deviation of 0.7. This value was tuned to ensure that the heritability explained by the loci is around 10% (between 10% and 11% across simulations).

For each of the 100 simulated datasets, we ran a GWAS with a linear model at each of the 5530 positions to infer the effect and significance of the locus. We then selected a short list of hits—positions significant at the 0.05 level (with a Bonferroni correction corresponding to the 1291 effective tests). From these hits, we predicted a polygenic score in the replication data and correlated that polygenic score to the simulated phenotype of the replication data. Additionally, we re-inferred effects in the replication data and correlated all discovery effects with replication effects (similarly to Fig. 3). We examined the *P*-values of these two correlations to determine under what conditions the polygenic score does not succeed (*P*-value > .05) and effect sizes are associated in discovery and additional datasets (*P*-value < .05).

Constructing a polygenic score with lasso regression

The absence of replicability of hits and the low portability of the polygenic score could mainly be explained by the small effects of individual loci once population structure has been accounted for. Thus, we attempted to develop an alternative GWAS and build polygenic scores that include together both the population structure and effects of individual loci, using lasso regression. Lasso regression is a penalized regression model with an L1 regularization—a term penalizing the log-likelihood proportionally to the sum of absolute values of parameters. Lasso regression has been successfully used for pathogen GWAS (Lees et al. 2018). It does not explicitly control for population structure but instead identifies a set of variants that best explain the phenotype of interest. We fitted lasso regression models to explain our four phenotypes from SNPs and kmer variants, retaining as potential predictors only those variants where the minority form is represented in at least 10 individuals, as in the main GWAS. We split the data in a training set representing 80% of individuals and a cross-validation set representing 20% of individuals to test the performance of the algorithm. We did so for 100 random splits of the BEEHIVE main dataset. Furthermore, we tested the performance of the algorithm on the BEEHIVE additional dataset. Our measure of performance of the predictor was the R^2 . To account for potential heterogeneity between training and additional dataset that greatly affected downwards the R^2 , we used the R^2 of the linear regression between predictor (resulting from the linear model selected by the lasso regression) and true phenotype.

Acknowledgements

We thank Jan Albert and Luca Ferretti for helpful comments and Christian Thorball for his help preparing the data for the additional dataset from the Swiss HIV Cohort Study. We thank people living with HIV of all cohorts across Europe for contributing samples to respective biobanks, allowing this work, and physicians, study nurses, and the datacentres for the high quality of data. Mígla Gabrielaite and Rasmus L. Marvig thank the International Network for Strategic Initiatives in Global HIV Trials (INSIGHT).

Supplementary data

Supplementary data are available at *VEVOLU Journal* online.

Conflict of interest: H F. G., outside this study, reports grants from the Swiss National Science Foundation, Swiss HIV Cohort

Study, a subcontract to a Bill and Melinda Gates Foundation grant, Gilead and ViiV (unrestricted research grants), and the Yvonne Jacob Foundation, all paid to his institution; personal fees as an advisor/consultant for Merck, ViiV, Johnson and Johnson, GSK, Janssen and Novartis and Gilead, and data and safety monitoring board remuneration from Merck, all outside the submitted work. M. v. d. V. has received research grants and fees for participation in advisory boards from Gilead, MSD and ViiV all paid to his institution. P. R. through his institution unrelated to the current work has received independent scientific grant support from Gilead Sciences, Merck & Co and ViiV Healthcare, and has served on scientific advisory boards for Gilead Sciences, ViiV Healthcare, and Merck & Co, honoraria for which were all paid to his institution.

Funding

This work was supported by the European Research Council Advanced Grant (grant number PBDR-339251) to C.F.

The Swiss HIV Cohort Study was supported by the Swiss National Science Foundation grant # 33FI-0_229621 and H.F.G. by the Yvonne Jacob foundation.

Additional support was provided by the Division of Intramural Research, NIAID, NIH.

M.G. is funded by HORIZON-MSCA-2023-PF-01 (Grant agreement No: 101151221).

T.G. is supported by an Investigator Grant (GNT2025445) from the National Health and Medical Research Council, Australia (NHMRC).

Data availability

The code underlying the analyses and the minimal dataset necessary to reproduce the figures are available on https://github.com/BDI-pathogens/BEEHIVE_GWAS

References

- Alizon S, von Wyl V, Stadler T et al. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog* 2010;**6**:e1001123. Public Library of Science.
- Asjö B, Morfeldt-Månson L, Albert J et al. Replicative capacity of human immunodeficiency virus from patients with varying severity of HIV infection. *Lancet (London, England)* 1986;**2**:660–2.
- Bachmann N, Turk T, Kadelka C et al. Parent-offspring regression to estimate the heritability of an HIV-1 trait in a realistic setup. *Retrovirology* 2017;**14**:33. <https://doi.org/10.1186/s12977-017-0356-3>
- Bankevich A, Nurk S, Antipov D et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77. <https://doi.org/10.1089/cmb.2012.0021>
- Bartha I, Carlson JM, Brumme CJ et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *elife* 2013;**2**:e01123. eLife Sciences Publications Limited.
- Bartha I, McLaren PJ, Brumme C et al. Estimating the respective contributions of human and viral genetic variation to HIV control. *PLoS Comput Biol* 2017;**13**:e1005339. Public Library of Science San Francisco, CA USA.
- Bastide P, Ho LST, Baele G et al. Efficient bayesian inference of general gaussian models on large phylogenetic trees. *The Annals of Applied Statistics* 2020;**15**:971–97.

- Bertels F, Marzel A, Leventhal G et al. Dissecting HIV virulence: Heritability of setpoint viral load, CD4+ T-cell decline, and parasite pathogenicity. *Mol Biol Evol* 2018;**35**:27–37. Oxford University Press.
- Blanquart F, Wymant C, Cornelissen M et al. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biol* 2017;**15**:e2001855. Public Library of Science San Francisco, CA USA.
- Bonsall D, Golubchik T, de Cesare M et al. A comprehensive genomics solution for HIV surveillance and clinical monitoring in low-income settings. *J Clin Microbiol* 2020;**58**:e00382–20. American Society for Microbiology. <https://doi.org/10.1128/jcm.00382-20>
- Carlson JM, Brumme CJ, Martin E et al. Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *J Virol* 2012;**86**:13202–16. <https://doi.org/10.1128/JVI.01998-12>
- Carlson JM, Schaefer M, Monaco DC et al. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science (New York, NY)* 2014;**345**:1254031. <https://doi.org/10.1126/science.1254031>
- Chang CC, Chow CC, Tellier LCAM et al. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 2015;**4**:s13742-015-0047-8. <https://doi.org/10.1186/s13742-015-0047-8>
- Chang LW, Grabowski MK, Ssekubugu R et al. Heterogeneity of the HIV epidemic in agrarian, trading, and fishing communities in Rakai, Uganda: An observational epidemiological study. *Lancet HIV* 2016;**3**:e388–96Elsevier. [https://doi.org/10.1016/S2352-3018\(16\)30034-0](https://doi.org/10.1016/S2352-3018(16)30034-0)
- Cornelissen M, Gall A, van der Kuyl A et al. Workup of human blood samples for deep sequencing of HIV-1 genomes. *Methods Mol Biol* 2018;**1746**:55–61. Viral Metagenomics. Springer.
- Dieleman JP, Jambroes M, Gyssens IC et al. Determinants of recurrent toxicity-driven switches of highly active antiretroviral therapy. The ATHENA cohort. *AIDS* 2002;**16**:737–45.
- van Dorp CH, van Boven M, De Boer RJ. Immuno-epidemiological modeling of HIV-1 predicts high heritability of the set-point viral load, while selection for CTL escape dominates virulence evolution. *PLoS Comput Biol* 2014;**10**:e1003899. Public Library of Science.
- Fellay J, Ge D, Shianna KV et al. Common genetic variation and the control of HIV-1 in humans. *PLoS Genet* 2009;**5**:e1000791. Public Library of Science.
- Fellay J, Shianna KV, Ge D et al. A whole-genome association study of major determinants for host control of HIV-1. *Science* 2007;**317**:944–7. American Association for the Advancement of Science.
- Foley BT, Korber BTM, Leitner TK et al. *HIV Sequence Compendium* 2018. Los Alamos, NM (United States): Los Alamos National Lab (LANL), 2018.
- Fraser C, Hollingsworth TD, Chapman R et al. Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. *Proc Natl Acad Sci* 2007;**104**:17441–6. National Acad Sciences.
- Fraser C, Lythgoe K, Leventhal GE et al. Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective. *Science* 2014;**343**:1243727. American Association for the Advancement of Science.
- Fryer HR, Frater J, Duda A et al. Modelling the evolution and spread of HIV immune escape mutants. *PLoS Pathog* 2010;**6**:e1001196. Public Library of Science San Francisco, USA.
- Gabrielaite M, Bennedbæk M, Zucco AG et al. Human immunotypes impose selection on viral genotypes through viral epitope specificity. *J Infect Dis* 2021;**224**:2053–63.
- Gall A, Ferns B, Morris C et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J Clin Microbiol* 2012;**50**:3838–44. Am Soc Microbiol.
- Ghosn J, Bayan T, Meixenberger K et al. CD4 T cell decline following HIV seroconversion in individuals with and without CXCR4-tropic virus. *J Antimicrob Chemother* 2017;**72**:2862–8. <https://doi.org/10.1093/jac/dkx247>
- Goulder PJ, Brander C, Tang Y et al. Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature* 2001;**412**:334–8. Nature Publishing Group.
- Harzic M, Pellegrin I, Deveau C et al. Genotypic drug resistance during HIV-1 primary infection in France (1996–1999): Frequency and response to treatment. *AIDS* 2002;**16**:793–6.
- Herbeck JT, Müller V, Maust BS et al. Is the virulence of HIV changing? A meta-analysis of trends in prognostic markers of HIV disease progression and transmission. *AIDS* 2012;**26**:193–205. <https://doi.org/10.1097/QAD.0b013e32834db418>
- HIV Databases (n.d.). Retrieved June 29, 2023, from <<https://www.hiv.lanl.gov/content/index>>
- Hodcroft E, Hadfield JD, Fearnhill E et al. The contribution of viral genotype to plasma viral set-point in HIV infection. *PLoS Pathog* 2014;**10**:e1004112. Public Library of Science.
- Hunt M, Gall A, Ong SH et al. IVA: Accurate de novo assembly of RNA virus genomes. *Bioinformatics* 2015;**31**:2374–6. Oxford University Press.
- Janes H, Herbeck JT, Tovanabutra S et al. HIV-1 infections with multiple founders are associated with higher viral loads than infections with single founders. *Nat Med* 2015;**21**:1139–41. Nature Publishing Group. <https://doi.org/10.1038/nm.3932>
- Joint United Nations Programme on HIV/AIDS (UNAIDS). *The Path that Ends AIDS: 2023 UNAIDS Global AIDS Update*. UN: The United Nations, 2023.
- Kaleebu P, French N, Mahe C et al. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. *J Infect Dis* 2002;**185**:1244–50. <https://doi.org/10.1086/340130>
- Kaleebu P, Ross A, Morgan D et al. Relationship between HIV-1 env subtypes A and D and disease progression in a rural Ugandan cohort. *AIDS* 2001;**15**:293–9. LWV.
- Kaslow RA, Carrington M, Apple R et al. Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med* 1996;**2**:405–11. <https://doi.org/10.1038/nm0496-405>
- Kivelä PS, Krol A, Salminen MO et al. High plasma HIV load in the CRF01-AE outbreak among injecting drug users in Finland. *Scand J Infect Dis* 2005;**37**:276–83.
- Koot M, Keet IP, Vos AH et al. Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. *Ann Intern Med* 1993;**118**:681–8. American College of Physicians.
- Kouri V, Khouri R, Alemán Y et al. CRF19_cpx is an evolutionary fit HIV-1 variant strongly associated with rapid progression to AIDS in Cuba. *EBioMedicine* 2015;**2**:244–54. Elsevier.
- van der Kuyl AC, Cornelissen M. Identifying HIV-1 dual infections. *Retrovirology* 2007;**4**:67. <https://doi.org/10.1186/1742-4690-4-67>
- Learmont JC, Geczy AF, Mills J et al. Immunologic and virologic status after 14 to 18 years of infection with an attenuated strain of HIV-1—A report from the Sydney blood Bank cohort. *N Engl J Med* 1999;**340**:1715–22. Mass Medical Soc.
- Lees JA, Galardini M, Bentley SD et al. Pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* 2018;**34**:4310–2. <https://doi.org/10.1093/bioinformatics/bty539>

- Leslie A, Pfafferott K, Chetty P et al. HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med* 2004;**10**: 282–9. Nature Publishing Group.
- Lippert C. *Linear Mixed Models for Genome-Wide Association Studies*. Eberhard Karls Universität Tübingen, 2013.
- Longley H, Fraser C, Wymant C et al. Attenuation of HIV severity by slightly deleterious mutations can explain the long-term trajectory of virulence evolution. *bioRxiv* 2025. <https://doi.org/10.1101/2025.05.12.653435>
- Martin DP, Lemey P, Lott M et al. RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* 2010;**26**: 2462–3. <https://doi.org/10.1093/bioinformatics/btq467>
- McDermott DH, Zimmerman PA, Guignard F et al. CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS cohort study (MACS). *Lancet* 1998;**352**:866–70. Elsevier.
- Minh BQ, Schmidt HA, Chernomor O et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;**37**:1530–4. <https://doi.org/10.1093/molbev/msaa015>
- Mitov V, Stadler T. A practical guide to estimating the heritability of pathogen traits. *Mol Biol Evol* 2018;**35**:756–72. Oxford University Press.
- Palm AA, Esbjörnsson J, Månsson F et al. Faster progression to AIDS and AIDS-related death among seroincident individuals infected with recombinant HIV-1 A3/CRF02_AG compared with sub-subtype A3. *J Infect Dis* 2014;**209**:721–8. Oxford University Press.
- Peterson RE, Kuchenbaecker K, Walters RK et al. Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell* 2019;**179**:589–603. <https://doi.org/10.1016/j.cell.2019.08.051>
- Poggensee G, Kücherer C, Werning J et al. Impact of transmission of drug-resistant HIV on the course of infection and the treatment success. Data from the German HIV-1 Seroconverter study. *HIV Med* 2007;**8**:511–9. <https://doi.org/10.1111/j.1468-1293.2007.00504.x>
- Ponstingl H, Ning Z. SMALT-a new mapper for DNA sequencing reads. *F1000 Posters* 2010;**1**.
- Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: Lessons from human GWAS. *Nat Rev Genet* 2017;**18**: 41–50. Nature Publishing Group.
- Redd AD, Quinn TC, Tobian AA. Frequency and implications of HIV superinfection. *Lancet Infect Dis* 2013;**13**:622–8 Elsevier. [https://doi.org/10.1016/S1473-3099\(13\)70066-5](https://doi.org/10.1016/S1473-3099(13)70066-5)
- Rindler AE, Kusejko K, Kuster H et al. The interplay between replication capacity of HIV-1 and surrogate markers of disease. *J Infect Dis* 2022;**226**:1057–68. <https://doi.org/10.1093/infdis/jiac100>
- Scherrer AU, Traytel A, Braun DL et al. Cohort profile update: The Swiss HIV cohort study (SHCS). *Int J Epidemiol* 2022;**51**:33–34j. <https://doi.org/10.1093/ije/dyab141>
- Stapleton JT. Human Pegivirus type 1: A common human virus that is beneficial in immune-mediated disease? *Front Immunol* 2022;**13**:887760. Frontiers. <https://doi.org/10.3389/fimmu.2022.887760>
- Struck D, Lawyer G, Ternes A-M et al. COMET: Adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res* 2014;**42**:e144. Oxford University Press.
- Golubchik T. *Typewriter*. 2018. <https://github.com/tgolubch/type-writer>
- Taylor BS, Sobieszczyk ME, McCutchan FE et al. The challenge of HIV-1 subtype diversity. *N Engl J Med* 2008;**358**:1590–602 Massachusetts Medical Society. <https://doi.org/10.1056/NEJMra0706737>
- Touloumi G, Pantazis N, Pillay D et al. Impact of HIV-1 subtype on CD4 count at HIV seroconversion, rate of decline, and viral load set point in European seroconverter cohorts. *Clin Infect Dis* 2013;**56**: 888–97. <https://doi.org/10.1093/cid/cis1000>
- Ueno T, Idegami Y, Motozono C et al. Altering effects of antigenic variations in HIV-1 on antiviral effectiveness of HIV-specific CTLs. *J Immunol* 2007;**178**:5513–23. <https://doi.org/10.4049/jimmunol.178.9.5513>
- Ueno T, Motozono C, Dohki S et al. CTL-mediated selective pressure influences dynamic evolution and pathogenic functions of HIV-1 Nef. *J Immunol* 2008;**180**:1107–16. Am Assoc Immunol.
- UK Register of HIV Seroconverters (UKRHS) Steering Committee. The UK register of HIV Seroconverters: Methods and analytical issues. *Epidemiol Infect* 1996;**117**:305–12. Cambridge University Press.
- Wei X, Ghosh SK, Taylor ME et al. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 1995;**373**:117–22. Nature Publishing Group. <https://doi.org/10.1038/373117a0>
- Wensing AM, Calvez V, Günthard HF et al. 2015 update of the drug resistance mutations in HIV-1. *Top Antivir Med* 2015;**23**:132–41.
- Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**:1–12. BioMed Central.
- Wymant C, Bezemer D, Blanquart F et al. A highly virulent variant of HIV-1 circulating in the Netherlands. *Science* 2022;**375**:540–5. <https://doi.org/10.1126/science.abk1688>
- Wymant C, Blanquart F, Golubchik T et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evol* 2018;**4**:vey007. Oxford University Press.
- Wymant C, Hall M, Ratmann O et al. PHYLOSCANNER: Inferring transmission from within-and between-host pathogen genetic diversity. *Mol Biol Evol* 2018;**35**:719–33. Oxford University Press.
- Yang W-L, Kouyos RD, Böni J et al. Persistence of transmitted HIV-1 drug resistance mutations associated with fitness costs and viral genetic backgrounds. *PLoS Pathog* 2015;**11**:e1004722. Public Library of Science.
- Zhao L, Wymant C, Blanquart F et al. Phylogenetic estimation of the viral fitness landscape of HIV-1 set-point viral load. *Virus Evol* 2022;**8**:veac022.