Host Variation in Interferon, MHC Class I, Glycosylation, and Viral Transcription Genes Predict HIV Persistence

David A. Siegel[1], Cassandra Thanh[4], Eunice Wan[2], Rebecca Hoh[1], Kristen Hobbs[4], Tony Pan[4], Erica A. Gibson[4], Deanna L. Kroetz[3], Peter W. Hunt[4], Jeffrey Marin[5], Frederick Hecht[1], Christopher Pilcher[1], Jeffrey Milush[4], Maureen Martin[6], Mary Carrington[6,7], Satish Pillai[8], Michael P. Busch[8], Mars Stone[8], Meei-Li Huang[9,10], Pavitra Roychoudhury[9,10], Kevin Haworth[9,10], Keith R. Jerome[9,10], Hans-Peter Kiem[9,10], Timothy J. Henrich[4], Steven G. Deeks[1], Sulggi Lee[1]

[1] Department of Medicine, Division of HIV, Infectious Diseases & Global Medicine, University of California San Francisco, 995 Potrero Avenue, San Francisco, CA 94110, USA. [2] Institute of Human Genetics, University of California San Francisco, CA 94143, USA. [3] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, CA 94158. [4] Department of Medicine, Division of Experimental Medicine, University of California San Francisco, CA 94110, USA. [5] Department of Biostatistics & Epidemiology, University of California San Francisco, CA 94158, USA. [6] Basic Science Program, Frederick National Laboratory for Cancer Research, National Cancer Institute, Frederick, MD and Laboratory of Integrative Cancer Immunology, Center for Cancer Research, National Cancer Institute, Bethesda, MD USA. [7] Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [8] Vitalant Blood Bank, San Francisco, CA 94118 USA. [9] Department of Laboratory Medicine and Pathology, University of Washington, Seattle WA 98195, USA. [10] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle WA 98109, USA.

Abstract

Background: Prior host genomewide association studies have failed to observe an association with the HIV reservoir.

Methods: Custom whole exome sequencing and direct HLA typing were performed from 202 HIV+ ART-suppressed individuals and associated with 4 measures of the circulating CD4+ T cell reservoir: total DNA, unspliced (us)RNA, RNA/DNA, and intact DNA. Common variant, gene-based, and HLA analyses were performed using linear mixed models adjusted for sex, timing of ART initiation, nadir CD4 count, input cell count, and ancestry.

Results: HIV total DNA was associated with variants in genes involved in type I interferon (*MX1, PPP1CB, DDX3X*) and MHC class I peptide recognition (*LRMP*), while HIV usRNA was associated with a variant related to lymphocyte lymph node migration (*PLVAP*; this SNP was also <30kb from *BST2*, which encodes tetherin, an HIV host restriction factor). Gene-based analyses demonstrated significant associations with type I interferon (intact DNA), glycosylation (total DNA), and retroviral transcription (usRNA). HLA "protective" B*57:01 and "risk" C*07 alleles, as well as *CCR5Δ32,* were associated with HIV usRNA and total DNA, but not intact DNA.

Conclusions: Host genetic variation in type I interferon, MHC class I, glycosylation, residual viral transcription, and lymphocyte lymph node migration may influence the circulating HIV CD4+ T cell reservoir.

Keywords: HIV reservoir, host genetics, whole exome sequencing, type I interferon, MHC class I, glycosylation

CONFLICTS: The authors do not have a commercial or other association that might pose a

conflict of interest.

FUNDING: This work was supported in part by the National Institutes of Health: K23GM112526 (SAL), the DARE Collaboratory (U19 AI096109; SGD), the Division of Intramural Research of the National Institutes (MC), UM1 AI126623 (HPK, KRJ), and NIH/NIAID R01A141003 (TJH). This project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under Contract No. HHSN261200800001E (MC). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research (MC). This work was also supported by the amfAR Research Consortium on HIV Eradication a.k.a. ARCHE (108072-50-RGRL; SGD) and a Collaboration for AIDS Vaccine Discovery (CAVD) grant from the Bill & Melinda Gates Foundation (INV-008500), the Reservoir Assay Validation and Evaluation Network (RAVEN) Study Group. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reprints: Reprint requests can be directed to Dr. Sulggi Lee, the corresponding author (contact information above).

Acknowledgements

The authors wish to acknowledge the participation of all the study participants who contributed to this work as well as the clinical research staff of the SCOPE and Options who made this research possible. All funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. All authors provided critical feedback in finalizing

the report. SAL conceived and designed the study with critical feedback from SGD, PWH, TH, DK, KRJ, and SH. SGD, JM, FH, CP, RH, and SAL coordinated the collection, management, and quality control processes for the cohort clinical data and SGD, JM, FH, CP, MPB, MS provided biospecimens. SAL and EW performed the whole exome sequencing assays. SAL, CT, JM, KB, TP, EAG performed participant sample processing, SAL and EW performed the whole exome sequencing, and DAS and SAL performed quality control analyses and the genomic association analyses for the study. SAL, SAL, CT, TH, KRJ, MLH, PR performed the HIV reservoir measurements, and PR, DAS, TJH, and SAL analyzed these data. MM and MC performed the HLA typing to determine HLA alleles for the analyses. DAS and SAL wrote the report. All authors provided critical feedback in finalizing the manuscript.

1. Background

Although antiretroviral therapy (ART) prolongs life, it does not fully restore health, as evidenced by persistently high levels of immune activation [1] and increased rates of non-AIDS-related mortality [2] observed in HIV-infected compared to uninfected individuals [3-6]. Persistent HIV may contribute to ongoing inflammation, immune activation, and subsequent clinical outcomes, even during effective ART [5-8]. However, factors associated with the HIV reservoir size, including the timing of ART initiation (smaller reservoir sizes with starting therapy during acute versus chronic infection [9, 10]), and host factors such as the level of T cell activation and T cell function only explain a fraction of the person-to-person variability in HIV reservoir size [11]. Identifying host genetic predictors of the HIV reservoir in ART-suppressed individuals may shed light on additional (potentially modifiable factors) that may contribute to HIV persistence.

Most genetic studies in HIV have focused on viral control or HIV disease progression in untreated HIV-infected individuals [12-21]. These studies identified several key single nucleotide polymorphisms (SNPs) in the human Major Histocompatibility Complex (MHC), or human leukocyte antigen (HLA)-B and -C regions as well as deletions in the C-C chemokine receptor type 5 gene (*CCR5Δ32*) [22-25] and a SNP in the HLA complex 5 (*HCP5*) gene [13]. However, since the study was performed in untreated individuals, the identified SNPs were likely measuring host genetic predictors of (active) HIV replication rather than of the persistent HIV reservoir. There has been one genomewide study of the HIV reservoir during treated disease to date [26]. The study evaluated 797 HIV-infected individuals, of which 194 also had whole exome sequencing performed, but observed no association between host genetic variants (including HLA or *CCR5Δ32*) and total HIV DNA; this included longitudinal measures of total HIV DNA over time. This study measured the HIV reservoir from total peripheral blood mononuclear cells (PBMCs) and imputed HLA alleles from genotypes rather than performed direct HLA typing.

Here, were performed custom whole exome sequencing (including intronic, 50 kilobase

upstream, and 50 kilobase downstream regulatory regions of select candidate genes) and direct HLA typing of 25 polymorphic exons of HLA-A & B, C, E, DPA1, DPB1, DQA1, DQB1, DRB1, and DRB3, 4, 5. We quantified the HIV reservoir from CD4-enriched cells (from cryopreserved PBMCs) and included four measures of the HIV reservoir: cell-associated total DNA ("defective" and "intact"), unspliced (full length transcript prior to alternative splicing) RNA, "intact" DNA (an estimate of the frequency of potentially "replication-competent" virus with intact HIV genomes) [27], and RNA/DNA (an estimate of the "transcriptional activity" of the HIV reservoir) from 202 highly characterized HIV+ ART-suppressed individuals in the University of California, San Francisco (UCSF) SCOPE and Options HIV Cohorts (**Figure 1**). We performed both "common" variant (polymorphisms observed at minor allele frequency, MAF, ≥5%) and rare-variant, gene-based analyses. Overall, our findings support prior data suggesting an important role for type I interferon signaling [28], the "transcriptionally active" HIV reservoir [29-32], and glycosylation [33-35]) in the maintenance of the HIV reservoir. In addition, our findings link previously reported genetic associations with viral control during untreated HIV disease (e.g., with MHC class I alleles [13, 20] and *CCR5Δ32* [23, 24, 36]), now with the HIV reservoir during treated disease, and highlight additional novel pathways that warrant further study in downstream functional studies.

3. Results

Study population

A total of 202 HIV-infected ART-suppressed individuals from the UCSF SCOPE and Options HIV+ cohort were included in the study. Consistent with our San Francisco-based HIV patient population, participants were mostly male (94%) with median age of 46 (**Table S1**). We included individuals who initiated ART during early (<6 months) and chronic (≥6 months) HIV infection. Samples included in this study were collected at a median of 5.1 years of ART

suppression from individuals who had initiated ART a median of 2.0 years from the estimated date of detected HIV infection. Participants had a median nadir CD4+ count of 341 cells/mm^3 and pre-ART HIV RNA of 4.3 log_{10}copies/mL. Frequencies of self-reported ethnicity were White/European American (63%), Black/African American (12%), Hispanic/Latino (11%), Mixed Ethnicity/Multiracial (6%), Asian (4%), Pacific Islander (1.5%), Native American (<1%), and Middle Eastern (<1%).

Earlier initiation of ART was associated with lower circulating CD4+ T cell HIV reservoir size

There is currently no "gold standard" for measuring the HIV reservoir; the reservoir largely consists of "defective" virus that harbors mutations prohibiting the production of infectious virus [37, 38]. We estimated the frequency of HIV total DNA ("defective" plus "intact" DNA) and unspliced RNA (full length transcript prior to alternative splicing) using an in-house quantitative polymerase chain reaction (qPCR) TaqMan assay using HIV-1 long terminal repeat (LTR)-specific primers [39]. Cell-associated HIV total DNA and unspliced RNA, simultaneously extracted from the same sample aliquot, were statistically significantly correlated (Spearman r=0.59, p=$8.3x10^{-20}$, **Figure S1a**). The frequency of "intact" HIV DNA was estimated using a droplet digital PCR (ddPCR)-based assay. HIV intact DNA was also associated with HIV unspliced RNA (Spearman r=0.25, p=$7.0x10^{-4}$, **Figure S1b**). Of the 199 individuals in the full cohort, 182 individuals also had HIV intact DNA quantification by ddPCR performed. The association between HIV total DNA and intact DNA was weaker (Spearman r=0.10, p=0.19, **Figure S1c**). Earlier timing of ART initiation was statistically significantly associated with lower HIV reservoir measures (HIV total DNA: R=0.28; p=$7.6x10^{-5}$, unspliced RNA: R=0.28,p=$5.9x10^{-5}$, and intact DNA: R=0.21; p=$3.8x10^{-3}$; **Figure 2**), and higher nadir CD4+ T cell counts (HIV total DNA: r=-0.28, p=$7.1x10^{-5}$, unspliced RNA: R=-0.23, p=$9.9x10^{-4}$, RNA/DNA: -0.15, p=$3.3x10^{-2}$,

and intact DNA: R=-0.27, $p=2.5x10^{-4}$; **Figure S2**). Of note, duration of ART suppression was weakly associated with these HIV reservoir measures as well, but the correlations between HIV reservoir measures and pre-ART viral load were more substantial (**Figure S3** and **Figure S4**, respectively).

Custom whole exome sequencing

We designed custom primers and probes to sequence an additional 54 megabases of intronic, intergenic, and untranslated regions of putative genes associated with the HIV reservoir, including genes previously associated with HIV integration into the host genome, host HIV antiviral restriction factors, and genes involved in the regulation of cell cycle progression, in addition to performing whole exome sequencing of all exons in the human genome (**Table S2**) [40-43]. These additional custom regions included intronic and regulatory regions (50 kilobases upstream and 50 kilobases downstream of each gene) for the analysis of potential variants influencing key gene expression quantitative trait loci [44].

A total of 1,279,156 variants from 23,733 genes from 199 participants passed quality control, with a total of 46% intronic, 21% intergenic, 19% exonic, 6% untranslated region (UTR), 6% non-coding RNA variants that were included in the final analysis (**Figure 3a**). Among the exonic variants, 53% were nonsynonymous and 42% were synonymous variants (**Figure 3b**).

Principal component analysis demonstrated that the largest cluster was for individuals with European ancestry (**Figure 4)**. Individuals with African ancestry had higher values of PC1, while individuals with Asian, Native American, or admixed ancestry had higher values of PC2. We performed a secondary principal component analysis to identify potential subpopulations among those with European ancestry, but these additional analyses did not identify separate clusters, suggesting a relatively homogenous population for the study participants with European ancestry (**Figure 4b**).

HLA "protective" B*57:01 and "risk" C*07 alleles were associated with cell-associated HIV total DNA and unspliced RNA

HLA typing was performed from extracted genomic DNA from cryopreserved PBMCs, following the PCR-SSOP (sequence-specific oligonucleotide probing) typing and PCR-SBT (sequence based typing) protocols recommended by the 13th International Histocompatibility Workshop [45, 46]. We used HLA typing to determine whether HLA alleles known to be associated with viral control [20] were also associated with the HIV reservoir measured studied here (**Tables S3-S6**). We examined alleles that were previously reported to be protective (B*57:01, B*27:05, B*14, C*08:02, B*52, and A*25) or risk (B*35 and C*07) alleles in their association with viral control in the absence of therapy [20]. Of the alleles that we examined, the most statistically significant associations were found with B*57:01 in total HIV DNA and HIV unspliced RNA ($p=0.032$, $\beta=-1.6$ and $p=4.2\times10^{-4}$, $\beta=-1.5$, respectively, in the full cohort, and $p=0.041$, $\beta=-1.4$ and $p=5.3\times10^{-4}$, $\beta=-1.5$, respectively, in the European ancestry male subcohort; these correspond to a false discovery rate (FDR) $q=0.13$, 0.0033, 0.12, and 0.0032, respectively, after adjusting for multiple testing), with a protective effect, consistent with prior findings related to viral control in untreated HIV disease. C*07 was also associated with HIV reservoir size ($p=0.012$, $\beta=0.76$ and $p=0.034$, $\beta=0.41$ for total HIV DNA and HIV unspliced RNA, respectively, in the European ancestry male dataset, or $q=0.072$ and 0.10, respectively), but with a risk effect, consistent with previous associations with viral control in untreated HIV disease. We also observed an association between HIV reservoir size (total HIV DNA and HIV unspliced RNA) and B*35 ($p=0.021$, $\beta=-0.81$ and $p=0.032$, $\beta=-0.44$, respectively; $q=0.13$ for both), but the allele was protective (negative association), rather than a risk allele. None of the alleles were statistically significantly associated with HIV RNA/DNA (a biomarker of transcriptional activity in relation to total HIV reservoir size) or HIV intact DNA measures. Moreover, additional analyses employing a composite HLA variable (the aggregate of the total

number of protective alleles minus the total number of risk alleles was analyzed in relation to each of these HIV reservoir measures), did not identify statistically significant associations.

CCR5Δ32 was significantly associated with measures of HIV total DNA and unspliced RNA

Deletions in the C-C chemokine receptor type 5 gene (*CCR5Δ32*) have previously been shown to be associated with HIV viral control in the absence of therapy [23, 24, 36]. Although not genome-wide significant, we observed a statistically significant association between the *CCR5Δ32* deletion and cell-associated HIV DNA and RNA (HIV DNA: p=0.0043, β=-1.3 and p=0.045, β=-0.86, and HIV RNA: p=0.0087, β=-0.78 and p=0.12, β=-0.41, for the European ancestry and full cohorts, respectively). In each of these cases the p-value was two-sided but the effect estimate was negative, which is consistent with the direction of the association for a protective allele. We also examined the lncRNA SNP rs1015164 which has previously been shown to be associated with HIV-1 viral load during untreated infection [25]. Again, we observed a statistically significant association with cell-associated HIV RNA (p=0.027, β=-0.39 and p=0.051, β=-0.30 in the European ancestry and full cohorts, respectively).

HIV reservoir measures were associated with common genetic variants related to type I interferon responses, MHC class I recognition, and lymphocyte lymph node migration

For the common variant (MAF ≥0.05) genetic analyses, standard quality control metrics demonstrated that the genomic inflation factor, or lambda GC, was close to 1 for all measures of the HIV reservoir, suggesting adequate control of systematic bias in the analysis (**Figure 5**). In the overall study population, the strongest genetic associations with HIV reservoir size were with total HIV DNA. We observed 49 SNPs that met a Benjamini Hochberg adjusted false discovery rate [47], q_{BH}, cutoff of 0.25 (**Table 1**), and 44 of these SNPs were in linkage

disequilibrium(LD) in the MX Dynamin Like GTPase 1 (*MX1*) gene (**Table S7**). Most of the SNPs demonstrated inverse associations with the HIV reservoir measures (**Table 1**), but as these are DNA variant associations and the relationships between these SNPs and genes were previously unknown, the question of whether the corresponding gene or protein expression directly correlates with a smaller or larger HIV reservoir size remains to be determined. The exception is with *MX1*, which demonstrated SNPs with both positive and negative associations with the HIV reservoir (**Table 1, Table S7**). Many of these *MX1* SNPs were in known expression quantitative trait loci (eQTL), regions associated with *MX1* gene expression variability [48]. Comparing associations between SNPs and HIV reservoir to associations between SNPs and *MX1* gene expression (in whole blood from previously reported data [48]), we found that when *MX1* SNPs were found to be positively associated with HIV total DNA, the same SNPs had been found to be positively associated with *MX1* gene expression; and when *MX1* SNPs were negatively associated with HIV total DNA, they were similarly negatively associated with *MX1* gene expression from these eQTL data (**Table S7**). This consistently points to a positive (i.e., same direction) association between *MX1* gene expression and the size of the circulating HIV CD4+ T cell reservoir.

The *MX1* gene encodes an interferon-induced GTP-metabolizing protein [49] known to inhibit viral replication in various diseases (e.g., influenza and hepatitis B) and is controlled by type I or type III interferons [50]. We also observed variants in two other genes that are linked to type I interferon responses: both *DDX3X* (DEAD-box helicase 3 X-linked), which plays a key role in innate antimicrobial immunity by regulating the production of type I IFNs [51], and *PPP1CB* (Protein Phosphatase 1 Catalytic Subunit Beta), for which the encoded protein reduces the antiviral potency of MX2 against HIV-1 [52], were significantly associated with HIV total DNA. In addition, variation in *LRMP* (Lymphoid Restricted Membrane Protein, otherwise known as *IRAG2* or Inositol 1,4,5-Triphosphate Receptor Associated 2), which encodes for a protein regulating MHC class I peptide recognition [53] and *AKAP6 (*A-Kinase Anchor Protein 6), which

encodes a scaffolding protein that regulates protein kinase signaling [54, 55], were also associated with HIV total DNA. Finally, a polymorphism in *PLVAP* (Plasmalemma Vesicle Associated Protein), which encodes for a protein that regulates the transmigration of lymphocytes into lymph nodes [56] was associated with HIV usRNA; this SNP also lies <30kb of *BST2* (Bone Marrow Stromal Cell Antigen 2), which encodes for the host HIV restriction factor, tetherin [57].

Of note, the SNPs in *MX1* were also associated with HIV total DNA in the European ancestry male subpopulation. Although the top SNP (rs469390) association was in this case positive (**Table 1**), eQTLs continue to point to a positive association between *MX1* gene expression and HIV reservoir size (**Table S7** shows that the eQTL effect is positive whenever the SNP effect is positive, and negative when the SNP effect is negative). In addition, a SNP in *OSBP* (oxysterol binding protein) was associated with HIV total DNA in the European male ancestry subpopulation. *OSBP* is a lipid transport gene shown to be associated with *in vitro* HIV-1 infection of monocyte-derived macrophages (MDMs) [58]. Finally, *PKHD1,* was associated with HIV total DNA in the overall study population, but we did not find a direct relationship with HIV in the published literature to date. PKHD1 is predicted to have a transmembrane-spanning domain and an immunoglobulin-like plexin-transcription-factor domain.

HIV reservoir measures were associated with gene sets linked to type I IFN signaling (STAT signaling, glycosylation) and the "active reservoir" (virus entry into host cells, HIV retroviral transcription)

We then performed multi-SNP rare variant analyses using the Variant Set Mixed Model Association Test (SMMAT) [59], a similar rare variant analysis method to SKAT-O [60]. Although no single gene was statistically significantly associated with HIV reservoir measures, gene set enrichment analyses (GSEA) [43, 61] demonstrated statistically significant associations with HIV total DNA, unspliced RNA, and intact DNA (FDR q-value <0.25) (**Figure 6, Table S8)**.

Interestingly, two gene sets that are again linked to type I IFN signaling (glycosylation [33] and HIV total DNA; STAT signaling [62] and HIV intact DNA) were statistically significant. HIV total DNA was associated with "N-Glycan Processing" (normalized enrichment score, NES=1.76, q=0.13), consistent with recent emerging data of the critical role of glycosylation in HIV persistence. Indeed, glycomic biomarkers were shown to predict time-to-viral rebound during treatment interruption, a clinical measure of the size of the HIV reservoir in previously ART-suppressed HIV-infected individuals, in at least 2 studies [35, 63]. A recent study suggests that changes in IFN-α signaling may induce glycomic alterations that then modulate the host immune response (e.g., CD8+ T and NK cell responses) [33]. HIV unspliced RNA was associated with retroviral genome replication ("Single-Stranded Viral RNA Replication via Double-Stranded DNA Intermediate," NES=1.73, q=0.16), which is in agreement with growing evidence of the transcriptionally "active," rather than just "latent," or quiescent reservoir [29-32]. Intact DNA, a measure of the potentially replication-competent reservoir, was statistically significantly associated with several gene sets (**Figure 6, Table S8)**. These included two gene sets related to STAT signaling ("Regulation of Peptidyl Serine Phosphorylation of STAT Protein," NES=1.67, q=0.19; "Serine Phosphorylation of STAT Protein," NES=1.63, q=0.25), a key pathway by which HIV blocks the host's innate antiviral type I IFN signaling response [64], as discussed above. HIV intact DNA was also associated with genes indicative of viral infectivity ("Negative Regulation of Viral Entry into Host Cells," NES=1.81, q=0.027), which was the most statistically significant gene set from all of the analyses (**Figure 6, Table S8)**. Finally, HIV intact DNA was associated with gene sets reflective of abnormal inflammation and metabolic processes that might link chronic treated HIV disease to the degree of HIV persistence [65]. These included genes associated with fatty acid metabolism ("Linoleic Acid Metabolic Process," NES=1.71, q=0.22), lower levels of dietary linoleic acid have been linked to cardiovascular risk [66]; cellular response to lead ("Response to Lead Ion," NES=1.65, q=0.21), genes reflecting the impact of this metal on immune cells (e.g., Th2 cell development, Th1 cell proliferation [67, 68]) supported by several studies

suggesting that lead exposure increases the incidence of infectious disease, allergy, autoimmune, and cancer [69]; and genes associated with new bone formation ("Replacement Ossification," NES=1.69, q=0.18), a metabolic process known to be abnormally in patients with chronic HIV [70] and a key potential cellular HIV reservoir (bone marrow) [71-73].

4. Discussion

This is the largest whole exome sequencing host genetic analysis of the HIV reservoir measured from CD4+ T cells in the blood to date. We identified several variants and gene sets associated with the defective and intact HIV reservoir size (total DNA), HIV transcription (unspliced RNA and RNA/DNA ratio), and the potentially inducible reservoir (intact DNA). We performed common genetic variant (MAF ≥5%) and multi-SNP rare variant gene-based analyses and identified several statistically significant associations with variants and genes involving innate immunity (type I IFN, STAT signaling), MHC class I recognition (including previously described HLA "protective" B*57:01 and "risk" C*07 alleles), *CCR5Δ32*, glycosylation, ongoing viral transcription, and lymphocyte lymph node migration, suggesting an important role for these pathways in the maintenance of the circulating HIV CD4+ T cell reservoir.

We identified several variants in genes associated with type I interferon signaling to be associated with total DNA. These include several SNPs in the *MX1* gene, which encodes a potent IFN-induced host restriction factor that is upregulated in HIV-infected compared to uninfected individuals [74] and in HIV-infected participants receiving a toll-like receptor 9 agonist (a potent stimulator of type I IFN) [75, 76]. MX1 has also been associated with higher viremia among HIV-infected individuals [77], HIV-1 persistence in latently infected cell lines [78], immune cell exhaustion [79] and T cell persistence [80]. Genetic variants in two other type I IFN-related genes were statistically significantly associated with HIV total DNA in our study: *DDX3X*, an X-chromosome-encoded RNA helicase which regulates the production of type I IFNs [51] and

PPP1CB, for which the encoded protein regulates the antiviral potency of MX2 [52] (closely related to MX1 which directly inhibits HIV infection prior to integration [52]). In addition, we observed HIV total DNA levels to be associated with genetic variants associated with MHC class I peptide recognition (*LRMP*) [53] and viral assembly/disassembly (*AKAP6*) [54, 55]. HIV unspliced RNA was associated with variants in *PLVAP*, which encodes a protein regulating the transmigration of lymphocytes into lymph nodes [56] and lies <30kb of *BST2*, which encodes for the host HIV restriction factor, tetherin [57]. We also identified additional genes linking HIV DNA measures to type I IFN responses in the rare variant gene-based analyses: glycosylation (IFN-α has been shown to alter host glycosylation during treated HIV disease, potentially contributing to modulation of the host immune response [33]) and STAT signaling (type I IFNs activate JAK-STAT signaling pathways [62]). HIV usRNA was also associated with gene sets related to retroviral transcription and viral entry into host cells. We also found that previous genetic associations with viral control during untreated HIV disease – HLA "protective" B*57:01 and "risk" C*07 alleles, as well as with CCR5, were associated with HIV usRNA and total DNA, but not intact DNA in our study.

The common variant analyses found mostly inversely associations with the HIV reservoir size (**Table 1**). These included variants in genes associated with innate immunity (*MX1*, *DDX3X*, *PPP1CB*), MHC class I recognition (*LRMP*, or *IRAG2*), protein kinase signaling (*AKAP6*), and transmigration of lymphocytes into lymph nodes (*PLVAP*, also associated with the HIV-1 host restriction factor, tetherin). However, the clinical and biologic relevance of the direction of these associations is unclear, with the exception of *MX1*, for which there are previously published expression quantitative trait loci, or eQTL, "functional interpretation," and for which there were several SNPs in LD. Therefore, whether genes or proteins associated with these DNA variants directly correlate with smaller HIV reservoir size needs to be evaluated in further functional studies. Similarly, in the rare variant gene-based analyses, HIV reservoir size

was associated with gene sets reflecting glycosylation (total DNA), viral replication (usRNA), viral infection of host cells (intact DNA), innate immunity (STAT signaling; intact DNA), and metabolic processes (fatty acid metabolism, bone formation, and cellular response to lead; all with intact DNA). Whether the direction of these associations directly translates into a change in function in these biologic pathways remains to be seen.

The most striking finding from our common variant analysis was the identification of several SNPs in the human interferon-inducible myxovirus resistance 1 gene, *MX1*, also known as *MXA*, as well as additional genes linked to type I IFN signaling (**Table 1, Table S7**). The encoded protein, MX1, is closely related to MX2 (MXB), a potent host restriction factor that inhibits HIV infection prior to integration [52, 81, 82]. MX1 itself is long recognized as a potent antiviral factor, inhibiting replication of several RNA viruses, including influenza A and measles viruses, and DNA viruses, such as hepatitis B virus [50]. Indeed, *MX1* gene expression has been shown to be upregulated in HIV-infected individuals compared to HIV-uninfected individuals [74], in HIV-infected individuals with high vs low viremia [77], and with HIV-1 latency in latently infected cell lines [78]. eQTL function at these loci [48] suggest that elevated levels of MX1 expression may be associated with higher total HIV DNA levels, perhaps by causing greater rates of immune cell exhaustion in chronically infected individuals [79, 80]. We also observed novel associations for HIV total DNA with two other genes linked to type I IFN signaling, *DDX3X* and *PPP1CB*. DDX3X (DEAD Box Helicase 3 X-Linked, an RNA helicase), an X-chromosome-encoded RNA helicase regulates type I IFN production [51] while PPP1CB (Protein Phosphatase 1 Catalytic Subunit Beta), regulates the antiviral potency of MX2 [52]. Prior work has also demonstrated that DDX3X may be involved in HIV-1 translation, delivering HIV-1 RNA from the nucleus to the cytoplasm [83]. In Hela cells, DDX3X expression was upregulated after transfection with HIV-1 Tat, enhanced HIV Gag production, and in another study [84], enhanced HIV replication in MT4 and Jurkat T cells. *PPP1CB* encodes a subunit of the protein PP1, which was similarly shown to regulate HIV-1 in Jurkat cells, HeLa cells, and other cell lines [84]. PP1CB interacts with HIV

Tat and is required for Tat-induced HIV-1 transcription in cultured cells; suppression of PP1 has been shown to inhibit HIV-1 transcription.

The two other genes associated with HIV total DNA were *LRMP* (Lymphoid-Restricted Membrane Protein), which encodes for a protein important for the delivery of peptides to MHC class I molecules [53], and *AKAP6* (A-Kinase Anchoring Protein 6) encodes for a protein that binds to regulatory subunits of protein kinase A [54, 55]. *LRMP* is differentially expressed in HIV-1 infected lymph nodes from patients before and after ART initiation, compared with untreated controls [85]. Interestingly, HIV-infected patients who underwent ART interruption were found to have differential *LRMP* gene expression between HIV-infected individuals with slow versus rapid CD4 decline during treatment interruption [86]. Given its role in peptide delivery to MHC class I molecules, *LRMP* may play an important role in CTL recognition of HIV-infected cells [53]. *AKAP6* (A-Kinase Anchoring Protein 6) encodes for a protein that binds to regulatory subunits of protein kinase A and anchors them to the nuclear membrane or sarcoplasmic reticulum. PKA activation is related to many cell cycle functions and has been associated with HIV-1 infection [87, 88], T cell proliferation [89], and HIV persistence [90, 91]. Among males with European ancestry, we also observed an association between total HIV DNA and a novel SNP in *OSBP*, a gene that encodes oxysterol-binding protein (OSBP), which is implicated in cholesterol metabolism and anti-HIV viral defenses HIV [58]. *OSBP* encodes a lipid transport protein required for the replication of several RNA viruses in humans, such as hepatitis C virus (HCV), encephalomyocarditis virus (EMCV), and Zika virus [92]. *OSBP* blocked HIV entry into host cells when its interaction with another key protein regulating intracellular cholesterol homeostasis, vesicle-associated membrane protein-associated protein A (VAPA) was disrupted [93]. Indeed, sterol metabolism has been shown to modulate susceptibility to HIV infection. In a study comparing highly exposed seronegative (HESN) to “lower risk” HIV-uninfected controls, HESNs demonstrated significantly higher expression of *OSBP*, which correlated with a reduced susceptibility to *in vitro* HIV-1 infection of PBMCs and monocyte-derived macrophages (MDMs)

[58].

The gene-based analyses also identified HIV reservoir associations with type I IFN responses. HIV total DNA was associated with a gene set reflecting glycosylation ("N-glycan processing"), which is modulated by IFN-α during treated HIV disease [33] and may be linked to host inflammation and immune activation during treated HIV disease. HIV intact DNA was also associated with two gene sets reflecting STAT signaling (Regulation of Peptidyl Serine Phosphorylation of STAT Protein" and "Serine Phosphorylation of STAT Protein"), key immune pathways activated by type I IFNs [62]. In addition, HIV usRNA was associated with gene sets reflecting viral entry into host cells (the most statistically significant gene set overall), as well as with retroviral transcription (Single-Stranded Viral RNA Replication via Double-Stranded DNA Intermediate).

There have been recent studies suggesting an important role for glycoprotein metabolism in the maintenance of the HIV reservoir [34, 94]. Glycoproteins are proteins with covalently attached glycans (oligosaccharide chains) to amino acid side-chains. Glycosylation is the process by which the attached carbohydrate is (post-translationally) modified. Indeed, the composition of glycans in the IgG's Fc region has been shown to modify functional and FcγR binding properties in HIV-infected patients [95, 96]. N-glycan processing is associated with HIV proliferation, as the envelope (Env) glycoprotein of HIV is densely glycosylated with N-glycans and has been shown to modulate host antibody response to the HIV virus [97, 98]. Finally, in a recent study, HIV-infected individuals with early viral rebound exhibited higher levels of total HIV-specific IgGs carrying inflammatory Fc glycans, while delayed rebounders showed an enrichment of highly functional antibodies [34].

STAT signaling is another pathway that has been implicated in contributing to HIV persistence [99]. IFNs are key innate immune cytokines that activate the JAK/STAT pathway inducing the expression of IFN-stimulated genes, key players in the host's antiviral response to limit viral infection in IFN-α-stimulated microenvironments. STAT signaling has been shown to

be disrupted by HIV Vpu, Nef [100] and Vif [101] to block these antiviral type I IFN responses [102-105]. Thus, modulation of these pathways may play a role in future HIV cure strategies. Indeed, JAK/STAT inhibitors have been hypothesized to be potentially useful in HIV eradication, given the availability of these inhibitors and experience with its use in other diseases such as cancer [106].

HIV infectivity ("Negative Regulation of Viral Entry into Host Cells") and viral replication ("Single-Stranded Viral RNA Replication via Double-Stranded DNA Intermediate") gene sets highlight the potential role of the transcriptionally and translationally "active reservoir," an emerging key player in HIV cure strategies. Previously, the HIV reservoir was thought to consist mostly of quiescent, latently-infected cells that could reactivate and lead to systemic infection with interruption of ART [107]. New methods, such as combined HIV integration and viral outgrowth assays, now suggest that transcriptionally and translationally active cells might be a major source of residual viremia during suppressive ART [108], and these "active" cells can persist for years (previously thought to be short-lived and cleared by immune recognition/cytotoxic response) [29-32].

For the other gene sets from our multi-SNP analysis, immunologically and/or clinically relevant associations with HIV persistence were less clear. These remaining statistically significant gene sets (all associated with HIV intact DNA, in particular) might reflect measures of abnormal metabolic processes that might be caused by, or be the result of, persistent HIV. These include gene sets associated with bone formation, fatty acid metabolism, cellular response to lead (or at least reflecting known changes to Th1 and Th2 responses due to lead toxicity [67, 68]). Changes in bone mineral density in patients living with HIV have been well documented, appear accelerated compared to the general population, and may be caused by HIV itself [109, 110]. Lower levels of gamma-linolenic acid have been associated with lower CD4 counts and higher rates of hospitalization and mortality in HIV-infected patients [111]. Iron overload has been shown to downregulate HIV-1 gene expression by decreasing the levels of

viral RNAs and also to modulate the expression of viral cofactors such as the REV cofactor eIF5A (which correlated with iron-induced inhibition of HIV-1 gene expression) [112]. Nonetheless, these findings suggest several intriguing novel hypotheses regarding how viral or host factors interact to influence the peripheral HIV reservoir. As with the common variant analysis results, however, the direction of these associations does not necessarily imply function; these data need to be further validated in functional studies.

We also observed associations consistent with prior reports linking viral control during untreated HIV disease (with HLA class I alleles [13, 20] and *CCR5Δ32* [23, 24, 36]), now with the HIV reservoir. The observed association with *CCR5Δ32* here are consistent with, and in the same direction (positively associated), as a prior study in untreated individuals [113], but are in contrast to another study which found no association between *CCR5Δ32* and viral control during untreated HIV disease [114]. Our results are also in contrast to a recent genomewide study which found no association between *CCR5Δ32* and the HIV reservoir (measured as total DNA from PBMCs) [26]. A major difference between these studies is that only the first former study included HIV-infected individuals treated during acute infection (i.e., timing of ART initiation in relation to HIV acquisition), while the latter two studies included individuals who initiated ART during chronic infection (or presumed chronic infection, based on treatment guidelines for the historical cohorts). Neither of these latter studies were therefore able to perform multivariate analyses accounting for timing of ART initiation in relation to HIV infection, a key predictor of the HIV reservoir size [9, 10], which may have influenced the divergent results.

We also observed a statistically significant protective association of B*57:01 and a risk association of C*07 (as previously described in untreated disease on viral setpoint [13, 20]) in relation to the HIV reservoir size (total DNA and unspliced RNA). We also observed a protective association between HLA-B*35 and HIV reservoir size, which is in contrast to its previously reported risk association in untreated HIV disease. Interestingly, we observed this association only with measures of HIV total DNA and unspliced RNA, but not with intact DNA. This

discrepancy might simply be due to the lower frequency of intact DNA compared to these other measures, but might also support prior work that suggests that even defective provirus can produce proteins (e.g., HIV Nef) that are “visible” to the host immune system (e.g., cells expressing MHC class I molecules) [115-117]. The aforementioned genomewide association study which did not observe a statistically significant association with *CCR5Δ32* also reported no association with HLA alleles in relation to the HIV reservoir size. However, the study only included quantification of total HIV-1 DNA levels from PBMCs, did not adjust for timing of ART initiation in relation to HIV infection), and imputed, rather directly performed HLA typing in their study [26].

Our study has some limitations that deserve mention. First, the study design is cross-sectional, and although trends were observed in relation to HIV reservoir size as a function of duration of ART suppression (e.g., a negative association between total HIV DNA levels and nadir CD4+ T cell count; a positive association with pre-ART viral load), these associations might simply reflect artifact due to historical ART initiation guidelines which were based on initiating ART only at lower CD4+ T cell counts [118]. Nonetheless, our study participants initiated ART from 1995-2013, when guidelines based on CD4 thresholds for ART initiation were relatively stable in the U.S. [119], and for all analyses, we carefully adjusted for timing of ART initiation, nadir CD4, and pre-ART viral load, which were all associated with reservoir size (our data suggested that duration of ART was not associated with the measures of HIV reservoir included in our study). A second limitation is that the direction of the observed associations between host DNA variants and the HIV reservoir are difficult to interpret since many of the relationships between these SNPs and genes are unknown. The exception to this was *MX1*, for which many of the SNPs were in known eQTL regions [48]; indeed, based on the eQTL data, all of the *MX1* SNPs showed a consistent direction between *MX1* gene expression and the size of the HIV reservoir (both positive or both negative). Further studies are needed to functionally validate whether these variants/genes correspond to gene and/or protein expression

that directly correlates with a smaller or larger HIV reservoir size. A third limitation is the heterogenous nature of our U.S.-based study population (**Figure 4**); population stratification can lead to false positive associations due to ancestry rather than from a true association [120]. In order to address this, we performed a comprehensive analysis adjusted for ancestry using established methods [121], but also presented our results stratified by ancestry focused on the largest genetically homogeneous group, males with European ancestry. Another important limitation was our ability to detect rare variants associated with the HIV reservoir size. Unlike cancer or cardiovascular disease registries, national and international HIV/AIDS surveillance registries have been challenged by changing methods and data sources and lack of coordination of biospecimen repositories and research cohorts [122], We included over 200 individuals in our study who met eligibility criteria, balancing assay cost against biospecimen availability and optimization of study design to detect an association with HIV reservoir size. We also performed gene set enrichment analyses, which grouped rare variants into immunologically and biologically relevant pathways for analyses, a method that can enhance the ability to identify potentially hypothesis-generating associations for further study in follow-up experiments. Finally, host genetic variants associated with the tissue HIV reservoir, where the majority of the HIV reservoir persists during ART suppression [123], may be different than the host genetic variants we found to be associated with measures of the peripheral HIV reservoir. Tissue-based HIV studies, let alone host genetic tissue-based studies, can be challenging to enroll and costly due to the invasive nature of the biospecimen sampling and the complexity of performing tissue-based assays. Recent data suggest that the tissue HIV reservoir is often reflected in the peripheral blood HIV reservoir [124-126], and in an autopsy-based study, large, clonal, intact virus emerged in blood after treatment interruption, suggesting that rebounding virus derived from circulating reservoirs quickly populated deep tissue reservoirs throughout the body via the blood [127]. This study found that differences across compartments varied according to participant, with no consistent clonality trends across tissues with or without sustained viral

suppression. By employing measures of HIV intact DNA by ddPCR, which estimates the frequency of intact integrated HIV DNA that can go on to produce a replication-competent virion, our study might at least provide some comparative data for future studies able to perform larger tissue-based analyses.

Our study has several strengths. Given the extremely low frequency of reservoir cells in virally suppressed subjects, precise measurement of the size of the reservoir is challenging but critical to determine. Since the majority of the viral genomes detected in resting CD4+ T cells are defective, PCR-based assays for HIV DNA overestimate the size of the reservoir [128] while the “gold standard” viral outgrowth assay (which provide a minimal estimate of the frequency of resting CD4+ T cells with replication competent HIV) underestimates the size of the reservoir and, given the large number of cells required, cannot be used to measure viral persistence in tissues or in rare populations of cells [129, 130]. Our study is the first to perform host genetic analyses in relation to several detailed measures of the HIV reservoir- cell-associated HIV unspliced RNA, total DNA, RNA/DNA ratio, as well as measures estimating the frequency of HIV intact (i.e., replication-competent) DNA. We also performed simultaneous RNA and DNA extraction from circulating CD4+ T cells from a single PBMC aliquot. This allowed a relatively unbiased comparison of the association between HIV cell-associated RNA and DNA measures, in contrast to earlier studies performing these extractions separately (which may explain the much higher dynamic range of HIV unspliced RNA compared to total DNA levels in our study compared to prior studies (**Figure S1**) [131]. GWAS genotyping attempts to capture the majority of common SNPs (polymorphisms with MAF ≥5% in the population) in the genome. However, many SNPs associated with disease have been found in intronic or intergenic regions, or “expression quantitative trait loci” (eQTL) regions [132], which are not adequately captured by GWAS genotyping alone. Performing whole genome sequencing (WGS) remains cost-prohibitive for performing large studies and the millions of variants collected from these data must be filtered to determine true associations [133]. Here, we performed custom whole exome

sequencing to include common as well as rare variants, which included over 50 MB of custom sequencing or intronic and flanking regulatory regions of putative candidate genes involved in HIV persistence (**Table S2**). Our study also performed direct HLA typing via targeted next generation sequencing of 25 polymorphic exons of HLA-A, B, C, E, DPA1, DPB1, DQA1, DQB1, DRB1, and DRB3, 4, 5 genes on the Illumina MiSeq platform. Additionally, our analyses included key HIV clinical covariates that were significantly associated with the reservoir size, including data on timing of ART initiation (as early as within weeks to months of the estimated date of detected HIV infection), nadir CD4+ T cell count, and pre-ART viral load (**Figures 2, S2, S4**).

A key challenge of HIV eradication strategies is the persistence of residual HIV despite ART-mediated viral suppression. In contrast to another study evaluating host genetic predictors of the peripheral HIV reservoir in ART-suppressed individuals using total HIV DNA from PBMCs, our analyses using cell-associated HIV DNA and RNA measures from circulating CD4+ T cells demonstrated statistically significant associations with previously reported immune pathways associated with the HIV reservoir, including a role for type I IFN signaling [28], glycosylation [33, 35, 63], and the transcriptionally and translationally active HIV reservoir [29-32, 108], in the establishment and maintenance of the HIV reservoir. Further studies are needed to confirm whether these host DNA variants predict actual functional (i.e., transcriptional and/or translational) changes to the circulating HIV CD4+ T cell reservoir, and whether they also predict the tissue HIV reservoir.

2. Methods

Study participants

Participants were sampled from the well-established, highly characterized San Francisco-based SCOPE and OPTIONS cohorts, which include HIV-infected individuals who

initiated ART during chronic (>2 years) or early (<6 months) HIV infection. Inclusion criteria were laboratory-confirmed HIV-1 infection, availability of at least one aliquot of cryopreserved 10^6 PBMCs, and plasma HIV RNA levels below the limit of assay quantification for at least 24 months at the time of biospecimen collection. Exclusion criteria were recent hospitalization, infection requiring antibiotics, vaccination, or exposure to immunomodulatory drugs in the six months prior to biospecimen timepoint. The research was approved by the University of California Committee on Human Research (CHR), and all participants provided written informed consent.

Custom whole exome host DNA sequencing

Briefly, cryopreserved PBMCs were enriched for CD4+ T cells (StemCell, Vancouver, Canada), and genomic DNA was extracted from CD4+ T cells (AllPrep Universal Kit, Qiagen, Hilden, Germany). Whole exome sequencing was using the extracted genomic host DNA. Using targeted exome capture with the SeqCap EZ Exome + UTR Library (Roche NimbleGen, Wilmington, MA) with custom addition of 50 Mb regulatory regions, exon-enriched sequencing libraries were generated and then sequenced on the Illumina HiSeq 2000 system (Illumina, Inc., San Diego, CA). The custom region included 442 candidate genes related to cell cycle regulation, HIV host restriction factors, and HIV-host integration, which were selected based on Gene Ontology (GO) Consortium experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP) indicating experimental evidence directly supporting the annotations (http://geneontology.org/docs/guide-go-evidence-codes/) (**Table S2**), as well as previously associated HLA variants associated with viral setpoint and/or HIV disease progression [134, 135].

HLA typing

Targeted sequencing using locus-specific primers were used to amplify a total of 25 polymorphic exons of HLA-A & B (exons 1 to 4), C (exons 1 to 5), E (exon 3), DPA1 (exon 2), DPB1 (exons 2 to 4), DQA1 (exon 1 to 3), DQB1 (exons 2 & 3), DRB1 (exons 2 & 3), and DRB3, 4, 5 (exon 2) genes with Fluidigm Access Array (Fluidigm Singapore PTE Ltd, Singapore). The 25 Fluidigm PCR amplicons are pooled and subjected to sequencing on an Illumina MiSeq sequencer (Illumina, San Diego, CA 92122 USA). HLA alleles and genotypes are called using the Omixon HLA Explore (version 2.0.0) software (Omixon, Budapest, Hungary).

Total HIV DNA and unspliced RNA from circulating CD4+ T cells

Cryopreserved PBMCs were enriched for CD4+ T cells using magnetic bead separation (StemCell Technologies) and host genomic DNA extracted from purified CD4+ T cells as described above (AllPrep Universal Kit, Qiagen, Hilden, Germany). Participant specimens were assayed with up to 800 ng total cellular RNA or DNA in replicate reaction wells and copy number determined by extrapolation against a 7-point standard curve (1–10,000 copies/second) performed in triplicate.

HIV intact DNA from circulating CD4+ T cells

Cell-associated HIV intact DNA was quantified from purified CD4+ T cells as above. Five regions on the HIV genome, including highly conserved regions, and positions that are frequently deleted or hypermutated, were interrogated to quantify potentially intact provirus [27]. Optimized restriction enzyme digestion was used to prepare the genomic DNA for droplet formation while minimizing the amount of shearing within the viral genome. The protocol targeted 5 regions in the HIV genome across two droplet digital (dd)PCR assays. Droplet generation and thermocycling were performed according to manufacturer instructions. This

multiplex ddPCR assay allowed the analysis of potentially replication-competent ("intact") proviral genomes by quantifying the number of droplets positive for 3 targets per assay. Two targets in a housekeeping gene (*RPP30)* were used to quantify all cells, and a target in the T cell receptor D gene (*TRD)* was used to identify all non-T cells, in order to normalize the HIV copy numbers per $1x10^6$ CD4+ T cells. A DNA shearing index (DSI) was then calculated and mathematically corrected for residual DNA shearing as measured by *RPP30* targets to calculate the estimated number of intact proviral genomes per million CD4+ T cells after correcting for shearing.

Data processing and quality control

The bcbio bioinformatics pipeline [136] was used to perform DNA alignment, which included the Burroughs-Welcome Aligner (BWA) tool [137] and the GenomeAnalysisToolkit (GATK) HaplotypeCaller joint variant calling method [138]. Reads were initially mapped to reference genome b37, then transposed to human genome assembly GRCh38 using Picard tools [139]. SNPs and insertions or deletions (indels) were then filtered by variant quality score recalibration (VQSR) using GATK [140]. The whole genome data analysis toolset, PLINK [141], was then used to validate the chromosomal sex of each individual, filter out individuals with excessive heterozygosity, and SNPS violating Hardy-Weinberg equilibrium (HWE) at a p-value threshold of 1×10^{-8}. The VCFtools suite of functions were then used to summarize data, run calculations, convert data, and filter out data, and convert data, and filter out relevant SNPs [142].

The GENESIS analysis pipeline [121] was used to analyze the relatedness and ancestries of the individuals in the study. All individuals were determined to be unrelated (kinship estimates < 0.05) aside from one pair of siblings, so one sibling was randomly removed from the study. The remaining 199 unrelated individuals had diverse and mixed ancestries, as demonstrated in

Figure 4a. To attempt to increase the power of our study, we identified the largest homogeneous subgroup from the first two principal components (PCs), a group of 126 men with European ancestry (**Figure 4b**). We performed the rest of our analysis in parallel on the full 199 individuals from diverse ancestries and genders and the 126 males with European ancestry. We used principal component analysis (PCA) to account for genetic heterogeneity in two more ways: First, our model includes a genetic effects term with a genetic relatedness matrix (GRM) to account for relatedness. Second, we included the first five PCs as covariates in our models to account for population-level effects.

Single SNP common variant analyses

Single SNP associations were calculated with GENESIS "assocTestSingle". For HIV total DNA, unspliced RNA, RNA/DNA, and intact DNA, respectively, the outcome variables were: $\log_{10}$((DNA copies/10^6 CD4+ T cells + offset)); $\log_{10}$((RNA copies/10^6 CD4+ T cells + offset)); $\log_{10}$((RNA copies/10^6 CD4+ T cells + offset) / (DNA copies/10^6 CD4+ T cells + offset)); $\log_{10}$((Intact DNA copies/10^6 CD4+ T cells + offset)). The offsets for RNA and DNA counts were given by the smallest nonzero measured values of RNA and DNA, respectively, to avoid divergences in the logarithm. Covariates include sex for the full population, time to antiretroviral therapy (ART) initiation, nadir CD4 count, and 5 genetic principal components (PCs). A Gaussian link function was used, and a genetic relatedness matrix (GRM) was included as calculated by PC-AiR and PC-Relate. Results were filtered to allele frequencies $\geq$5%. SNP annotations were then obtained using Annovar, a bioinformatics software tool to interpret SNPS, insertions, deletions, and copy number variants in a genome [143].

Gene-based rare variant analyses

Gene level rare variant multi-SNP associations were calculated with the GENESIS

software package "assocTestAggregate" function implementing the variant Set Mixed Model Association Test (SMMAT), an adjusted gene-based test combined with a burden test [59], for alleles with frequencies < 5% with weights following the beta distribution parameters of a_1=1 and a_2=25 as suggested by Ref [144]. Genes with a total of fewer than 5 alternate alleles were omitted. The same covariates, GRM, and regression family were used as for the single SNP associations. Outcomes were quantile-normalized to follow a normal distribution. Gene regions were defined according to the intersection of Bioconductor's UCSC hg38 genes[145]. Exons for genes with custom sequencing were defined according to (Roche NimbleGen, Wilmington, MA). Exons for genes without custom sequencing were defined according to the intersection of exonic regions defined by Bioconductor and (Roche NimbleGen, Wilmington, MA).

Gene set enrichment analyses (GSEA) were performed using the Molecular Signatures Database (MSigDB) [43, 146]. We used unsigned SMMAT p-values and focused on Gene Ontology (GO) Biological Processes (BP) gene sets that were up-regulated for low p-values. For all gene set analyses, introns and flanking regions of ±50kb were included in the SMMAT p-value calculations for each gene to account for potential regulatory regions and SNPs with smaller effects. GSEAPreranked was run with default parameters on the SMMAT gene-level $-log_{10}(P)$.

HLA analysis

Multivariate regression models were fit using the python statsmodels OLS function [147] with covariates for sex, timing of ART initiation, nadir CD4+ T cell count, and 3 genetic PCs.

FIGURE LEGENDS

Figure 1. Study participant sample selection flowchart. Specific inclusion and exclusion criteria are listed for each selection step and for HIV reservoir measure analysis.

Figure 2. Cell-associated HIV total DNA, unspliced RNA, RNA/DNA, and intact DNA were associated with timing of ART initiation. Spearman correlation and p-values are shown in the inset boxes.

Figure 3. Distribution of all of the total number of variants (a), as well as all exonic variants (b) sequenced in the study population.

Figure 4. Genetic principal component analysis (PCA) plots of the full study population (with legend with self-identified race/ethnicity) (a) and of the largest homogenous ancestral population, males of mostly European ancestry (b). Recalculated European ancestry male PCA plot is shown in panel b, from lower left dashed box in panel a.

Figure 5. Quantile-quantile (QQ) plots (a, c) and Manhattan plots (b, d) of the total study population (a-b) and of European ancestry males (c-d). QQ plots: the blue line represents the expected $-\log_{10}$ p-values while the black lines denote the expected error bars. Manhattan plots: the horizontal black line delineates the genome-wide significance of p-value of 5×10^{-8}, while the gray line represents a false discovery rate (FDR) statistical significance p-value of 0.25.

Figure 6. Gene set enrichment analysis (GSEA) was used to identify associations between genes, ordered by p-value under a host multi-SNP rare variant (minor allele frequency, MAF, <5%) analysis using HIV total DNA, unspliced RNA, RNA/DNA, and intact DNA as outcomes;

and biological processes gene sets, in the total study population (a) and among males of European ancestry (b). Horizontal red dashed line represents the GSEA false discovery rate (FDR) statistical significance level of q=0.025. Gene sets with q <0.25 are labeled in the figures below.

REFERENCES

1. Yukl, S.A., et al., *Differences in HIV burden and immune activation within the gut of HIV-positive patients receiving suppressive antiretroviral therapy.* J Infect Dis, 2010. **202**(10): p. 1553-61.
2. Hunt, P.W., et al., *T cell activation is associated with lower CD4+ T cell gains in human immunodeficiency virus-infected patients with sustained viral suppression during antiretroviral therapy.* J Infect Dis, 2003. **187**(10): p. 1534-43.
3. Lohse, N., et al., *Survival of persons with and without HIV infection in Denmark, 1995-2005.* Ann Intern Med, 2007. **146**(2): p. 87-95.
4. Antiretroviral Therapy Cohort, C., *Causes of death in HIV-1-infected patients treated with antiretroviral therapy, 1996-2006: collaborative analysis of 13 HIV cohort studies.* Clin Infect Dis, 2010. **50**(10): p. 1387-96.
5. Eisele, E. and R.F. Siliciano, *Redefining the viral reservoirs that prevent HIV-1 eradication.* Immunity, 2012. **37**(3): p. 377-88.
6. Volberding, P.A. and S.G. Deeks, *Antiretroviral therapy and management of HIV infection.* Lancet, 2010. **376**(9734): p. 49-62.
7. *Causes of death in HIV-1-infected patients treated with antiretroviral therapy, 1996-2006: collaborative analysis of 13 HIV cohort studies.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2010. **50**(10): p. 1387-96.
8. Lohse, N., et al., *Survival of persons with and without HIV infection in Denmark, 1995-2005.* Annals of internal medicine, 2007. **146**(2): p. 87-95.
9. Chun, T.W., et al., *Decay of the HIV reservoir in patients receiving antiretroviral therapy for extended periods: implications for eradication of virus.* J Infect Dis, 2007. **195**(12): p. 1762-4.
10. Hocqueloux, L., et al., *Long-term antiretroviral therapy initiated during primary HIV-1 infection is key to achieving both low HIV reservoirs and normal T cell counts.* J Antimicrob Chemother, 2013. **68**(5): p. 1169-78.
11. Cohn, L.B., N. Chomont, and S.G. Deeks, *The Biology of the HIV-1 Latent Reservoir and Implications for Cure Strategies.* Cell Host Microbe, 2020. **27**(4): p. 519-530.
12. Dalmasso, C., et al., *Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study.* PLoS One, 2008. **3**(12): p. e3907.
13. Fellay, J., et al., *A whole-genome association study of major determinants for host control of HIV-1.* Science, 2007. **317**(5840): p. 944-7.
14. Fellay, J., et al., *Host genetic determinants of T cell responses to the MRKAd5 HIV-1 gag/pol/nef vaccine in the step trial.* J Infect Dis, 2011. **203**(6): p. 773-9.
15. Fellay, J., et al., *Common genetic variation and the control of HIV-1 in humans.* PLoS Genet, 2009. **5**(12): p. e1000791.
16. Herbeck, J.T., et al., *Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS.* J Infect Dis, 2010. **201**(4): p. 618-26.
17. Le Clerc, S., et al., *Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03).* J Infect Dis, 2009. **200**(8): p. 1194-201.
18. Limou, S., et al., *Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02).* J Infect Dis, 2009. **199**(3): p. 419-26.
19. Pelak, K., et al., *Host determinants of HIV-1 control in African Americans.* J Infect Dis, 2010. **201**(8): p. 1141-9.
20. International, H.I.V.C.S., et al., *The major genetic determinants of HIV-1 control affect HLA class I peptide presentation.* Science, 2010. **330**(6010): p. 1551-7.

21. Petrovski, S., et al., *Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population.* AIDS, 2011. **25**(4): p. 513-8.
22. de Roda Husman, A.M., et al., *Association between CCR5 genotype and the clinical course of HIV-1 infection.* Ann Intern Med, 1997. **127**(10): p. 882-90.
23. Dean, M., et al., *Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study.* Science, 1996. **273**(5283): p. 1856-62.
24. Rappaport, J., et al., *32 bp CCR-5 gene deletion and resistance to fast progression in HIV-1 infected heterozygotes.* Lancet, 1997. **349**(9056): p. 922-3.
25. Kulkarni, S., et al., *CCR5AS lncRNA variation differentially regulates CCR5, influencing HIV disease outcome.* Nat Immunol, 2019. **20**(7): p. 824-834.
26. Thorball, C.W., et al., *Host Genomics of the HIV-1 Reservoir Size and Its Decay Rate During Suppressive Antiretroviral Treatment.* J Acquir Immune Defic Syndr, 2020. **85**(4): p. 517-524.
27. Levy, C.H., S.; Roychoudhury, P.; Reeves, D. B.; Amstuz, C.; Zhu, H.; et al., *A highly multiplexed droplet digital PCR assay to measure the intact HIV-1 proviral reservoir.* Cell Reports Medicine, 2021. **2**.
28. Gondim, M.V.P., et al., *Heightened resistance to host type 1 interferons characterizes HIV-1 at transmission and after antiretroviral therapy interruption.* Sci Transl Med, 2021. **13**(576).
29. Pace, M.J., et al., *HIV reservoirs and latency models.* Virology, 2011. **411**(2): p. 344-54.
30. McManus, W.R., et al., *HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy.* J Clin Invest, 2019. **129**(11): p. 4629-4642.
31. Kearney, M.F., et al., *Origin of Rebound Plasma HIV Includes Cells with Identical Proviruses That Are Transcriptionally Active before Stopping of Antiretroviral Therapy.* J Virol, 2016. **90**(3): p. 1369-76.
32. Wiegand, A., et al., *Single-cell analysis of HIV-1 transcriptional activity reveals expression of proviruses in expanded clones during ART.* Proc Natl Acad Sci U S A, 2017. **114**(18): p. E3659-E3668.
33. Giron, L.B., et al., *Interferon-alpha alters host glycosylation machinery during treated HIV infection.* EBioMedicine, 2020. **59**: p. 102945.
34. Bartsch, Y.C., et al., *Viral Rebound Kinetics Correlate with Distinct HIV Antibody Features.* mBio, 2021. **12**(2).
35. Offersen, R., et al., *HIV Antibody Fc N-Linked Glycosylation Is Associated with Viral Rebound.* Cell Rep, 2020. **33**(11): p. 108502.
36. de Roda Husman, A.M., et al., *Association between CCR5 genotype and the clinical course of HIV-1 infection.* Annals of internal medicine, 1997. **127**(10): p. 882-90.
37. Ho, Y.C., et al., *Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure.* Cell, 2013. **155**(3): p. 540-51.
38. Bruner, K.M., et al., *Defective proviruses rapidly accumulate during acute HIV-1 infection.* Nat Med, 2016. **22**(9): p. 1043-9.
39. Malnati, M.S., et al., *A universal real-time PCR assay for the quantification of group-M HIV-1 proviral load.* Nat Protoc, 2008. **3**(7): p. 1240-8.
40. Kulpa, D.A., et al., *PD-1 coinhibitory signals: the link between pathogenesis and protection.* Semin Immunol, 2013. **25**(3): p. 219-27.
41. Abdel-Mohsen, M., et al., *Select host restriction factors are associated with HIV persistence during antiretroviral therapy.* AIDS, 2015. **29**(4): p. 411-20.
42. Maldarelli, F., et al., *HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells.* Science, 2014. **345**(6193): p. 179-83.
43. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

44. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
45. Martin, M.P., et al., *Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS.* Nat Genet, 2002. **31**(4): p. 429-34.
46. Martin, M.P., et al., *Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1.* Nat Genet, 2007. **39**(6): p. 733-40.
47. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
48. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
49. Staeheli, P. and O. Haller, *Human MX2/MxB: a Potent Interferon-Induced Postentry Inhibitor of Herpesviruses and HIV-1.* J Virol, 2018. **92**(24).
50. Verhelst, J., P. Hulpiau, and X. Saelens, *Mx proteins: antiviral gatekeepers that restrain the uninvited.* Microbiol Mol Biol Rev, 2013. **77**(4): p. 551-66.
51. Szappanos, D., et al., *The RNA helicase DDX3X is an essential mediator of innate antimicrobial immunity.* PLoS Pathog, 2018. **14**(11): p. e1007397.
52. Betancor, G., et al., *MX2-mediated innate immunity against HIV-1 is regulated by serine phosphorylation.* Nat Microbiol, 2021. **6**(8): p. 1031-1042.
53. Collins, K.L., *How HIV evades CTL recognition.* Curr HIV Res, 2003. **1**(1): p. 31-40.
54. Perfettini, J.L., et al., *NF-kappaB and p53 are the dominant apoptosis-inducing transcription factors elicited by the HIV-1 envelope.* J Exp Med, 2004. **199**(5): p. 629-40.
55. Troger, J., et al., *A-kinase anchoring proteins as potential drug targets.* Br J Pharmacol, 2012. **166**(2): p. 420-33.
56. Rantakari, P., et al., *The endothelial protein PLVAP in lymphatics controls the entry of lymphocytes and antigens into lymph nodes.* Nat Immunol, 2015. **16**(4): p. 386-96.
57. Perez-Caballero, D., et al., *Tetherin inhibits HIV-1 release by directly tethering virions to cells.* Cell, 2009. **139**(3): p. 499-511.
58. Saulle, I., et al., *Sterol metabolism modulates susceptibility to HIV-1 Infection.* AIDS, 2020. **34**(11): p. 1593-1602.
59. Chen, H., et al., *Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies.* Am J Hum Genet, 2019. **104**(2): p. 260-274.
60. Lee, S., et al., *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.* Am J Hum Genet, 2012. **91**(2): p. 224-37.
61. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-40.
62. Platanias, L.C., *Mechanisms of type-I- and type-II-interferon-mediated signalling.* Nat Rev Immunol, 2005. **5**(5): p. 375-86.
63. Giron, L.B., et al., *Plasma and antibody glycomic biomarkers of time to HIV rebound and viral setpoint.* AIDS, 2020. **34**(5): p. 681-686.
64. Schneider, W.M., M.D. Chevillotte, and C.M. Rice, *Interferon-stimulated genes: a complex web of host defenses.* Annu Rev Immunol, 2014. **32**: p. 513-45.
65. Hunt, P.W., S.A. Lee, and M.J. Siedner, *Immunologic Biomarkers, Morbidity, and Mortality in Treated HIV Infection.* J Infect Dis, 2016. **214 Suppl 2**: p. S44-50.
66. Farvid, M.S., et al., *Dietary linoleic acid and risk of coronary heart disease: a systematic review and meta-analysis of prospective cohort studies.* Circulation, 2014. **130**(18): p. 1568-78.
67. Heo, Y., W.T. Lee, and D.A. Lawrence, *Differential effects of lead and cAMP on development and activities of Th1- and Th2-lymphocytes.* Toxicol Sci, 1998. **43**(2): p. 172-85.
68. Hsiao, C.L., K.H. Wu, and K.S. Wan, *Effects of environmental lead exposure on T-helper cell-*

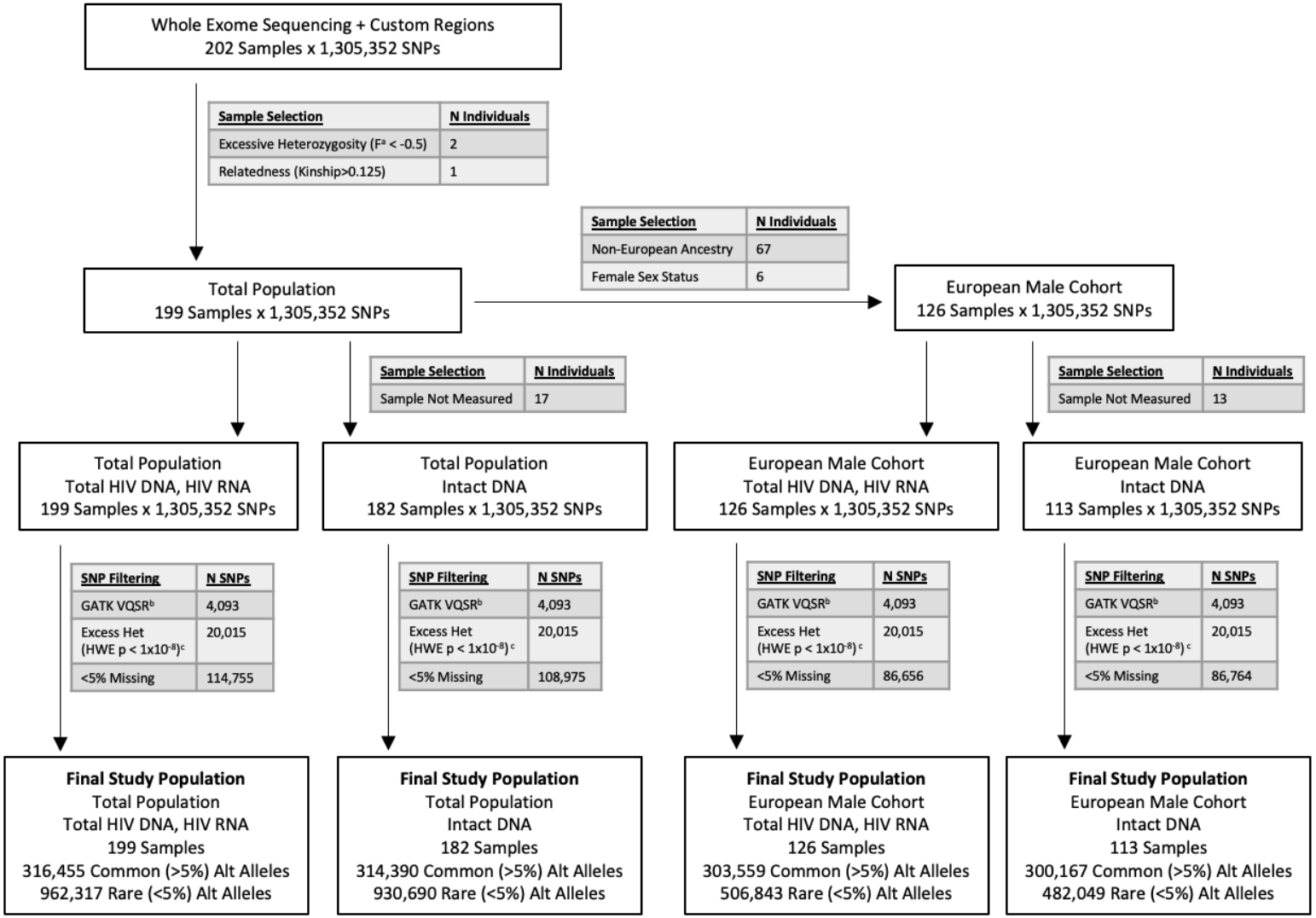
specific cytokines in children. J Immunotoxicol, 2011. **8**(4): p. 284-7.
69. Fenga, C., et al., *Immunological effects of occupational exposure to lead (Review).* Mol Med Rep, 2017. **15**(5): p. 3355-3360.
70. McComsey, G.A., et al., *Bone disease in HIV infection: a practical review and recommendations for HIV care providers.* Clin Infect Dis, 2010. **51**(8): p. 937-46.
71. Hutter, G., et al., *Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation.* N Engl J Med, 2009. **360**(7): p. 692-8.
72. Gupta, R.K., et al., *HIV-1 remission following CCR5Delta32/Delta32 haematopoietic stem-cell transplantation.* Nature, 2019. **568**(7751): p. 244-248.
73. Salgado, M., et al., *Mechanisms That Contribute to a Profound Reduction of the HIV-1 Reservoir After Allogeneic Stem Cell Transplant.* Ann Intern Med, 2018. **169**(10): p. 674-683.
74. Zhao, G., et al., *The dynamic changes of interferon lambdas related genes and proteins in JAK/STAT pathway in both acute and chronic HIV-1 infected patients.* AIDS Res Ther, 2017. **14**(1): p. 31.
75. Vibholm, L., et al., *Short-Course Toll-Like Receptor 9 Agonist Treatment Impacts Innate Immunity and Plasma Viremia in Individuals With Human Immunodeficiency Virus Infection.* Clin Infect Dis, 2017. **64**(12): p. 1686-1695.
76. Krarup, A.R., et al., *The TLR9 agonist MGN1703 triggers a potent type I interferon response in the sigmoid colon.* Mucosal Immunol, 2018. **11**(2): p. 449-461.
77. Van Hecke, C., et al., *Early treated HIV-1 positive individuals demonstrate similar restriction factor expression profile as long-term non-progressors.* EBioMedicine, 2019. **41**: p. 443-454.
78. Lee, S.Y., et al., *Selection of biomarkers for HIV-1 latency by integrated analysis.* Genomics, 2019. **111**(3): p. 327-333.
79. Wang, S., et al., *An atlas of immune cell exhaustion in HIV-infected individuals revealed by single-cell transcriptomics.* Emerg Microbes Infect, 2020. **9**(1): p. 2333-2347.
80. Szabo, P.A., et al., *Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease.* Nat Commun, 2019. **10**(1): p. 4706.
81. Goujon, C., et al., *Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection.* Nature, 2013. **502**(7472): p. 559-62.
82. Liu, Z., et al., *The interferon-inducible MxB protein inhibits HIV-1 infection.* Cell Host Microbe, 2013. **14**(4): p. 398-410.
83. Yedavalli, V.S., et al., *Requirement of DDX3 DEAD box RNA helicase for HIV-1 Rev-RRE export function.* Cell, 2004. **119**(3): p. 381-92.
84. Ammosova, T., et al., *Nuclear protein phosphatase-1 regulates HIV-1 transcription.* J Biol Chem, 2003. **278**(34): p. 32189-94.
85. Li, Q., et al., *Functional genomic analysis of the response of HIV-1-infected lymphatic tissue to antiretroviral therapy.* J Infect Dis, 2004. **189**(4): p. 572-82.
86. Vahey, M.T., et al., *CD4+ T-cell decline after the interruption of antiretroviral therapy in ACTG A5170 is predicted by differential expression of genes in the ras signaling pathway.* AIDS Res Hum Retroviruses, 2008. **24**(8): p. 1047-66.
87. Cartier, C., et al., *Active cAMP-dependent protein kinase incorporated within highly purified HIV-1 particles is required for viral infectivity and interacts with viral capsid protein.* J Biol Chem, 2003. **278**(37): p. 35211-9.
88. Barnitz, R.A., et al., *Protein kinase A phosphorylation activates Vpr-induced cell cycle arrest during human immunodeficiency virus type 1 infection.* J Virol, 2010. **84**(13): p. 6410-24.
89. Skalhegg, B.S., et al., *Isozymes of cyclic AMP-dependent protein kinases (PKA) in human lymphoid cell lines: levels of endogenous cAMP influence levels of PKA subunits and growth in lymphoid cell lines.* J Cell Physiol, 1998. **177**(1): p. 85-93.
90. Lim, H., et al., *Synergistic reactivation of latent HIV-1 provirus by PKA activator dibutyryl-cAMP in combination with an HDAC inhibitor.* Virus Res, 2017. **227**: p. 1-5.
91. Giroud, C., et al., *HIV-1-associated PKA acts as a cofactor for genome reverse transcription.*

Retrovirology, 2013. **10**: p. 157.
92. Roberts, B.L., et al., *Differing activities of oxysterol-binding protein (OSBP) targeting anti-viral compounds.* Antiviral Res, 2019. **170**: p. 104548.
93. Amini-Bavil-Olyaee, S., et al., *The antiviral effector IFITM3 disrupts intracellular cholesterol homeostasis to block viral entry.* Cell Host Microbe, 2013. **13**(4): p. 452-64.
94. Vadrevu, S.K., et al., *Frontline Science: Plasma and immunoglobulin G galactosylation associate with HIV persistence during antiretroviral therapy.* J Leukoc Biol, 2018. **104**(3): p. 461-471.
95. Ackerman, M.E., et al., *Polyfunctional HIV-Specific Antibody Responses Are Associated with Spontaneous HIV Control.* PLoS Pathog, 2016. **12**(1): p. e1005315.
96. Chung, A.W., et al., *Dissecting Polyclonal Vaccine-Induced Humoral Immunity against HIV Using Systems Serology.* Cell, 2015. **163**(4): p. 988-98.
97. Quinones-Kochs, M.I., L. Buonocore, and J.K. Rose, *Role of N-linked glycans in a human immunodeficiency virus envelope glycoprotein: effects on protein function and the neutralizing antibody response.* J Virol, 2002. **76**(9): p. 4199-211.
98. Shen, R., et al., *HIV-1 envelope glycan moieties modulate HIV-1 transmission.* J Virol, 2014. **88**(24): p. 14258-67.
99. Gavegnano, C., et al., *Novel mechanisms to inhibit HIV reservoir seeding using Jak inhibitors.* PLoS Pathog, 2017. **13**(12): p. e1006740.
100. Nguyen, N.V., J.T. Tran, and D.J. Sanchez, *HIV blocks Type I IFN signaling through disruption of STAT1 phosphorylation.* Innate Immun, 2018. **24**(8): p. 490-500.
101. Gargan, S., et al., *HIV-1 Promotes the Degradation of Components of the Type 1 IFN JAK/STAT Pathway and Blocks Anti-viral ISG Induction.* EBioMedicine, 2018. **30**: p. 203-216.
102. Chaudhuri, A., et al., *STAT1 signaling modulates HIV-1-induced inflammatory responses and leukocyte transmigration across the blood-brain barrier.* Blood, 2008. **111**(4): p. 2062-72.
103. Pericle, F., et al., *HIV-1 infection induces a selective reduction in STAT5 protein expression.* J Immunol, 1998. **160**(1): p. 28-31.
104. Percario, Z., et al., *Human immunodeficiency virus type 1 (HIV-1) Nef activates STAT3 in primary human monocyte/macrophages through the release of soluble factors: involvement of Nef domains interacting with the cell endocytotic machinery.* J Leukoc Biol, 2003. **74**(5): p. 821-32.
105. Kohler, J.J., et al., *Human immunodeficiency virus type 1 (HIV-1) induces activation of multiple STATs in CD4+ cells of lymphocyte or monocyte/macrophage lineages.* J Leukoc Biol, 2003. **73**(3): p. 407-16.
106. Schwartz, D.M., et al., *JAK inhibition as a therapeutic strategy for immune and inflammatory diseases.* Nat Rev Drug Discov, 2017. **17**(1): p. 78.
107. Eriksson, S., et al., *Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies.* PLoS Pathog, 2013. **9**(2): p. e1003174.
108. Groebner, J., et al. *Resting and activated CD4+ T cells both have silent and active HIV proviruses in vivo. Abstract # 304. Spotlight E4.* in *Conference on Retroviruses and Opportunistic Infections.* . 2021. Boston, MA.
109. Triant, V.A., et al., *Fracture prevalence among human immunodeficiency virus (HIV)-infected versus non-HIV-infected patients in a large U.S. healthcare system.* J Clin Endocrinol Metab, 2008. **93**(9): p. 3499-504.
110. Grant, P.M., et al., *Long-term Bone Mineral Density Changes in Antiretroviral-Treated HIV-Infected Individuals.* J Infect Dis, 2016. **214**(4): p. 607-11.
111. Kabagambe, E.K., et al., *Plasma n-6 Fatty Acid Levels Are Associated With CD4 Cell Counts, Hospitalization, and Mortality in HIV-Infected Patients.* J Acquir Immune Defic Syndr, 2016. **73**(5): p. 598-605.
112. Mancone, C., et al., *Iron overload down-regulates the expression of the HIV-1 Rev cofactor eIF5A in infected T lymphocytes.* Proteome Sci, 2017. **15**: p. 18.
113. Wang, C., et al., *Decreased HIV type 1 transcription in CCR5-Delta32 heterozygotes during suppressive antiretroviral therapy.* J Infect Dis, 2014. **210**(11): p. 1838-43.

114. Henrich, T.J., et al., *CCR5-Delta32 Heterozygosity, HIV-1 Reservoir Size, and Lymphocyte Activation in Individuals Receiving Long-term Suppressive Antiretroviral Therapy.* J Infect Dis, 2016. **213**(5): p. 766-70.
115. Pollack, R.A., et al., *Defective HIV-1 Proviruses Are Expressed and Can Be Recognized by Cytotoxic T Lymphocytes, which Shape the Proviral Landscape.* Cell Host Microbe, 2017. **21**(4): p. 494-506 e4.
116. Thomas, A.S., et al., *T-cell responses targeting HIV Nef uniquely correlate with infected cell frequencies after long-term antiretroviral therapy.* PLoS Pathog, 2017. **13**(9): p. e1006629.
117. Hiener, B., et al., *Identification of Genetically Intact HIV-1 Proviruses in Specific CD4(+) T Cells from Effectively Treated Participants.* Cell Rep, 2017. **21**(3): p. 813-822.
118. Panel on Antiretroviral Guidelines for Adults and Adolescents, D.o.H.a.H.S., *Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV.*
119. Richardson, E.T., P.M. Grant, and A.R. Zolopa, *Evolution of HIV treatment guidelines in high- and low-income countries: converging recommendations.* Antiviral Res, 2014. **103**: p. 88-93.
120. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-9.
121. Gogarten, S.M., et al., *Genetic association testing using the GENESIS R/Bioconductor package.* Bioinformatics, 2019. **35**(24): p. 5346-5348.
122. Brookmeyer, R., *Measuring the HIV/AIDS epidemic: approaches and challenges.* Epidemiol Rev, 2010. **32**: p. 26-37.
123. Pantaleo, G., et al., *Lymphoid organs function as major reservoirs for human immunodeficiency virus.* Proc Natl Acad Sci U S A, 1991. **88**(21): p. 9838-42.
124. Evering, T.H., et al., *Absence of HIV-1 evolution in the gut-associated lymphoid tissue from patients on combination antiviral therapy initiated during primary infection.* PLoS Pathog, 2012. **8**(2): p. e1002506.
125. Josefsson, L., et al., *The HIV-1 reservoir in eight patients on long-term suppressive antiretroviral therapy is stable with few genetic changes over time.* Proc Natl Acad Sci U S A, 2013. **110**(51): p. E4987-96.
126. Imamichi, H., et al., *Lack of compartmentalization of HIV-1 quasispecies between the gut and peripheral blood compartments.* J Infect Dis, 2011. **204**(2): p. 309-14.
127. Chaillon, A., et al., *HIV persists throughout deep tissues with repopulation from multiple anatomical sources.* J Clin Invest, 2020. **130**(4): p. 1699-1712.
128. Eriksson, S., et al., *Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies.* PLoS pathogens, 2013. **9**(2): p. e1003174.
129. Ho, Y.-C., et al., *Replication-Competent Noninduced Proviruses in the Latent Reservoir Increase Barrier to HIV-1 Cure.* Cell, 2013. **155**(3): p. 540-551.
130. Laird, G.M., et al., *Rapid quantification of the latent reservoir for HIV-1 using a viral outgrowth assay.* PLoS pathogens, 2013. **9**(5): p. e1003398.
131. Lee, S.A., et al., *Anti-HIV Antibody Responses and the HIV Reservoir Size during Antiretroviral Therapy.* PLoS One, 2016. **11**(8): p. e0160192.
132. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS genetics, 2010. **6**(4): p. e1000888.
133. Gonzaga-Jauregui, C., J.R. Lupski, and R.A. Gibbs, *Human genome sequencing in health and disease.* Annual review of medicine, 2012. **63**: p. 35-61.
134. Martin, M.P. and M. Carrington, *Immunogenetics of HIV disease.* Immunol Rev, 2013. **254**(1): p. 245-64.
135. Martin, M.P., et al., *Killer cell immunoglobulin-like receptor 3DL1 variation modifies HLA-B*57 protection against HIV-1.* J Clin Invest, 2018. **128**(5): p. 1903-1912.
136. Chapman, B.e.a., *bcbio/bcbio-nextgen*. 2021, Zenodo.
137. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv:1303.3997, 2013.

138. Poplin, R., et al., *Scaling accurate genetic variant discovery to tens of thousands of samples.* bioRxiv, 2017. **201178**.
139. *Picard toolkit*. 2019; Available from: http://broadinstitute.github.io/picard/.
140. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
141. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
142. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
143. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
144. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet, 2011. **89**(1): p. 82-93.
145. Bioconductor Core Team, B.P.M.M.B.O., *Txdb.hsapiens.ucsc.hg38.knowngene*. 2017.
146. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. **34**(3): p. 267-73.
147. Seabold; Skipper; Perktold, J. *statsmodels: Econometric and statistical modeling with python*. in *Proceedings of the 9th Python in Science Conference*. 2010.

Figure 1. Study participant sample selection flowchart. Specific inclusion and exclusion criteria are listed for each selection step and for HIV reservoir measure analysis.



[a] F = inbreeding coefficient
[b] VQSR = Variant Quality Score Recalibration
[c] HWE = Hardy-Weinberg equilibrium

Figure 2. Cell-associated HIV total DNA, unspliced RNA, RNA/DNA, and intact DNA were associated with timing of ART initiation. Spearman correlation and p-values are shown in the inset boxes.

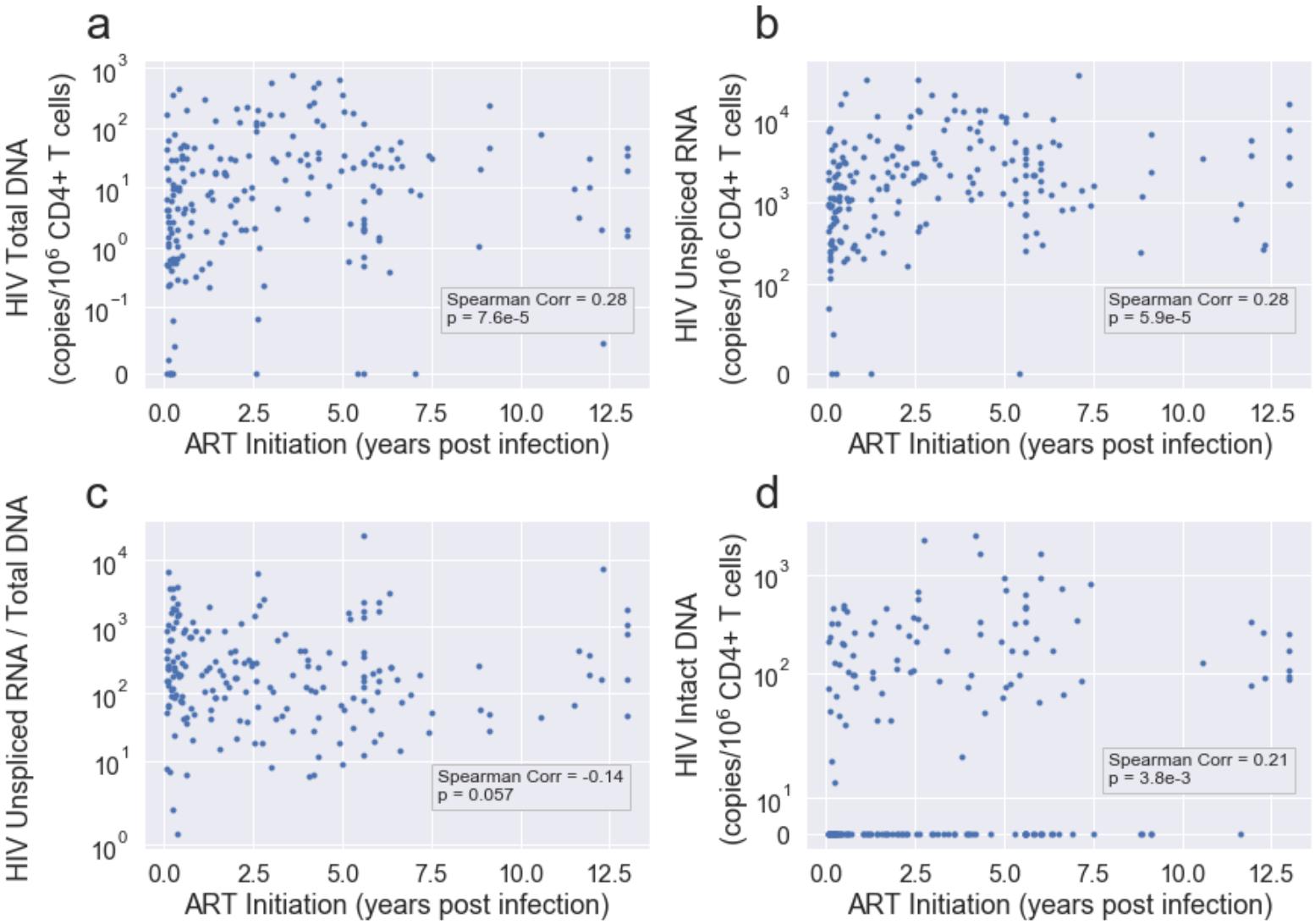


Figure 3. Distribution of the total number of host variants (a), as well as all exonic variants (b) sequenced in the study population.

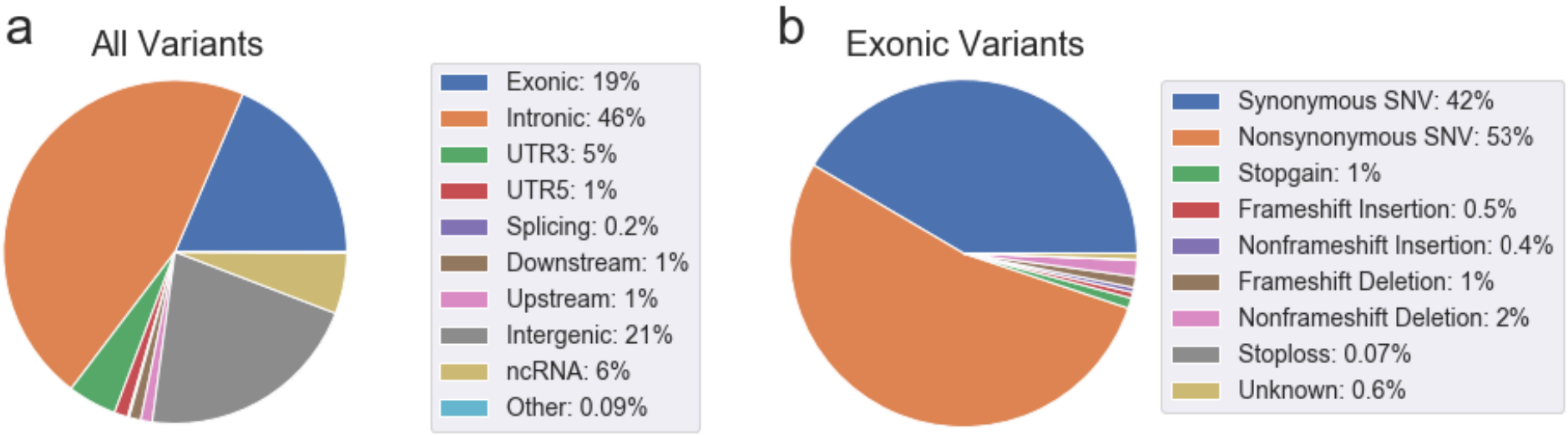


Figure 4. Genetic principal component analysis (PCA) plots of the full study population (with legend with self-identified race/ethnicity) (a) and of the largest homogenous ancestral population, males of mostly European ancestry (b). Recalculated European ancestry male PCA plot is shown in panel b, from lower left dashed box in panel a.

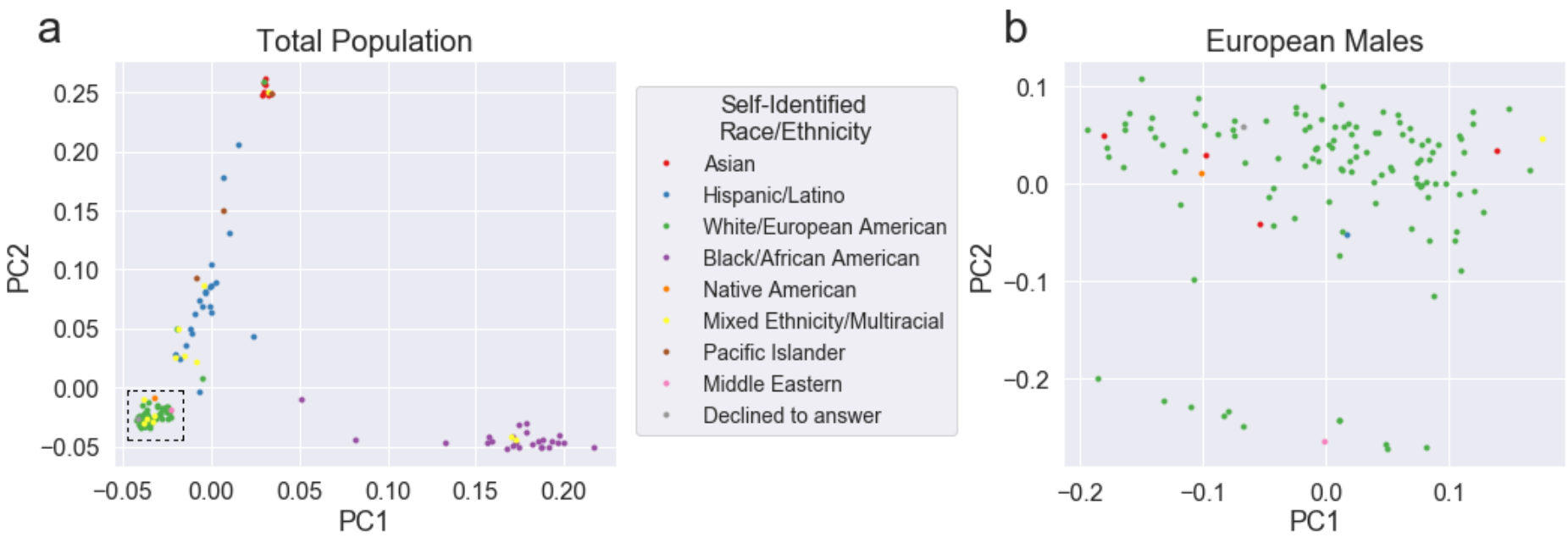


Figure 5. Quantile-quantile (QQ) plots (a, c) and Manhattan plots (b, d) of the total study population (a-b) and of European ancestry males (c-d). QQ plots: the blue line represents the expected -$\log_{10}$ p-values while the black lines denote the expected error bars. Manhattan plots: the horizontal black line delineates the genome-wide significance of p-value of 5×10^{-8}, while the gray line represents a false discovery rate (FDR) statistical significance p-value of 0.25.

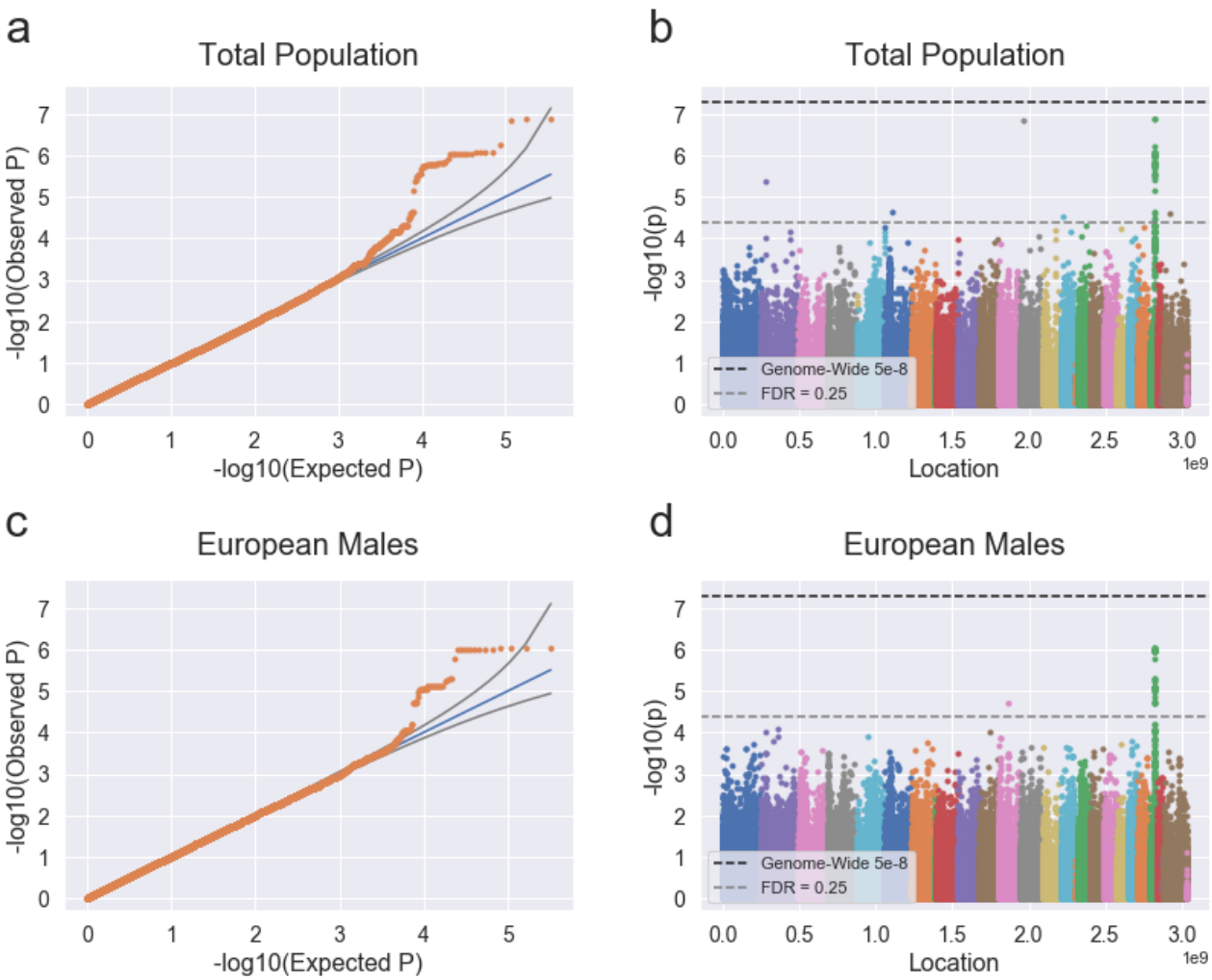


Table 1. Statistically significant (false discovery rate, FDR,<0.25) common single nucleotide polymorphisms (SNPs) were associated with cell-associated HIV total DNA (upper table) and unspliced RNA (lower table) in the full cohort. Two additional SNPs were significantly associated with total HIV DNA in the subpopulation of European ancestry males (middle table). The top SNP for each gene is shown. UTR= untranslated region; PVE = proportion of phenotype variance explained; p = p-value; q = FDR q-value.

SNP	Chrom	Position	Closest Gene	Gene Location	Freq	Effect	PVE	p	q	Description
HIV TOTAL DNA										
Full Cohort										
rs10670165	chr21	41421873	*MX1*	Intron/Exon/ 5'UTR	0.45	-1.2	0.15	1.3e-7	0.02	Antiviral. Upregulated in HIV+ compared to uninfected individuals [74] and associated with higher viremia among HIV+ individuals [77]. Associated with HIV-1 latency [78]. MX2, but not MX1, was shown to be capable of directly acting against HIV-1 virus [81-82].
rs74867009	chr12	25063777	*LRMP*	5'UTR	0.06	-2.5	0.15	1.5e-7	0.02	Differentially expressed in lymphatic tissue [85] and peripheral blood mononuclear cells (PBMCs) [86] in HIV+ individuals in response to ART initiation and cessation, respectively.
rs751660317	chr2	28786774	*PPP1CB*	Intronic	0.07	-1.8	0.11	4.1e-6	0.03	Encodes a subunit of PP1. Depletion of PPP1CB was shown to reduce the antiviral potency of MX2 against HIV [52]. PP1 is involved in transcription of HIV-1; inhibition of PP1 inhibits HIV-1 transcription [84].
rs776025235	chr6	51638799	*PKHD1*	Intronic	0.09	-1.7	0.09	2.5e-5	0.17	Predicted to have a transmembrane-spanning domain and an immunoglobulin-like plexin-transcription-factor domain. We could not find a direct relationship with HIV in the literature.
N/A	chrX	41382082	*DDX3X*; *NYX*	Intergenic	0.18	-1.4	0.09	2.5e-5	0.17	*DDX3X* regulates the production of type I IFNs [51] and encodes a protein that shuttles HIV-1 RNA from the nucleus to the cytoplasm. Is upregulated in HIV-infected cells; knockdown of *DDX3X* suppresses HIV-1 replication [83]. Also plays a key role in innate antimicrobial immunity [51].
rs17506750	chr14	32599402	*AKAP6*	Intronic	0.07	-1.8	0.09	3.0e-5	0.20	Binds to regulatory subunits of protein kinase A (PKA) and anchors them to the nuclear membrane. PKA activation has been associated with HIV-1 infection, T cell proliferation, and dysfunction [54-55].
European Ancestry Male Cohort										
rs469390	chr21	41446003	*MX1*	Intron/Exon/ 5'UTR	0.54	1.3	0.20	1.0e-6	0.03	Antiviral. Upregulated in HIV+ compared to uninfected individuals [74] and associated with higher viremia among HIV+ individuals [77]. Associated with HIV-1 latency [78]. MX2, but not MX1, was shown to be capable of directly acting against HIV-1 virus [81-82].
N/A	chr11	59574427	*OSBP*	3'UTR	0.15	-1.8	0.15	1.9e-5	0.14	Oxysterol binding protein involved in intracellular lipid transport. Associated with *in vitro* HIV-1 infection of monocyte-derived macrophages (MDMs) from highly (HIV)-exposed seronegatives (HESNs) [58]. Is required for the replication of several human viruses such as hepatitis C (HCV), encephalomyocarditis (EMCV), Zika, etc. [92].
HIV UNSPLICED RNA										
Full Cohort										
N/A	chr19	17374631	*PLVAP*	Intronic	0.10	-1.2	0.14	6.9e-7	0.21	Endothelial membrane protein, also controls the entry of lymphocytes and antigens into lymph nodes [56]. Located within 30kb of *BST2* (aka tetherin), which is known to inhibit HIV-1 release by directly tethering virions to cells [57].

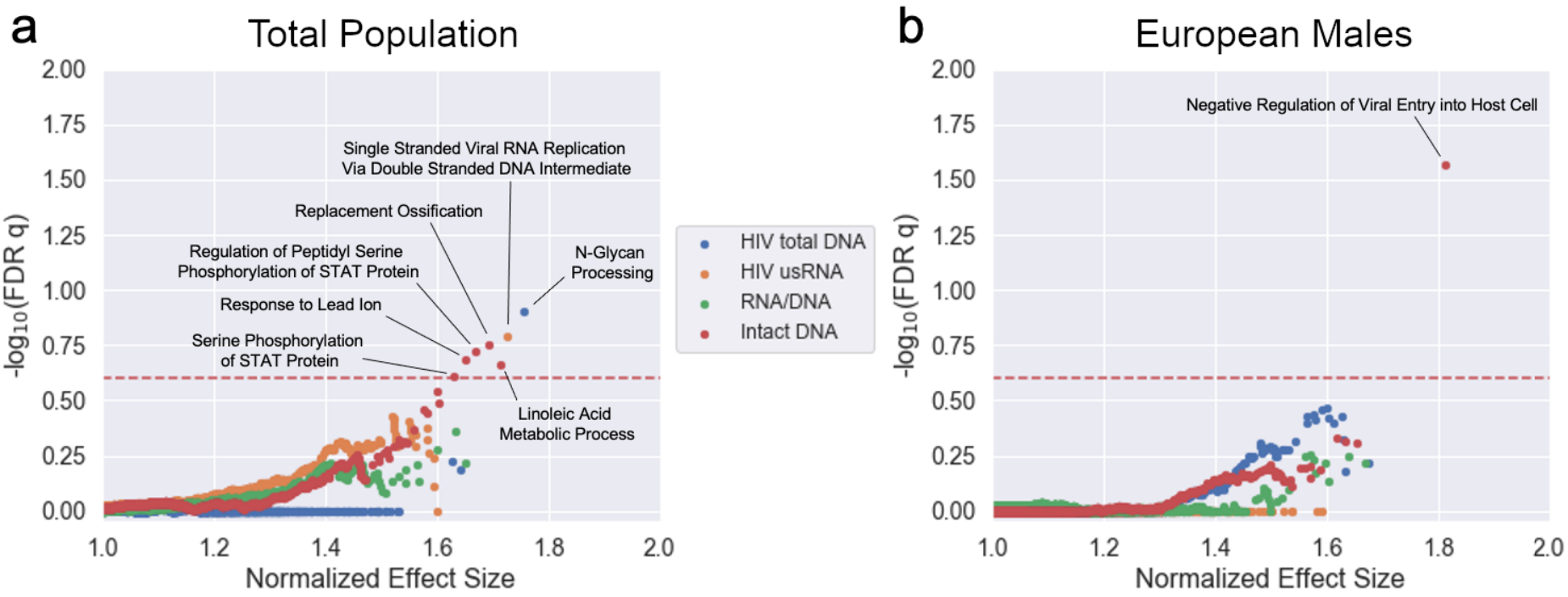



Figure 6. Gene set enrichment analysis (GSEA) was used to identify associations between genes, ordered by p-value under a host multi-SNP rare variant (minor allele frequency, MAF, <5%) analysis using HIV total DNA, unspliced RNA, RNA/DNA, and intact DNA as outcomes; and biological processes gene sets, in the total study population (a) and among males of European ancestry (b). Horizontal red dashed line represents the GSEA false discovery rate (FDR) statistical significance level of q=0.025. Gene sets with q < 0.25 are labeled in the figures below.