# Genes and environment profoundly affect the human virome

Nolan Kamitaki[1,2,3,4,5,6], David Tang[1,2,3,4], Steven A McCarroll[3,5,6,7], Po-Ru Loh[1,2,3]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.
[2]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA, USA.
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
[5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[6]Department of Genetics, Harvard Medical School, Boston, MA, USA.
[7]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA.

Correspondence should be addressed to N.K. (nolan_kamitaki@hms.harvard.edu), S.A.M. (smccarro@broadinstitute.org), or P.-R.L. (poruloh@broadinstitute.org).

## Abstract

Many viruses have adapted to persist in infected humans for life. Variable host control of their abundance (or load) can lead to clearance or disease. Here, we analyzed the viral load of 31 DNA viruses in human blood and saliva using whole-genome sequencing data from UK Biobank (n=490,401), SPARK (n=12,519), and All of Us (n=414,817). Viral load varied markedly with age, time of day, and season, and was higher in men than women for most viruses. Human genetic variation at dozens of genomic loci associated with load of seven viruses: Epstein-Barr virus (EBV, 45 loci), human herpesvirus 7 (HHV-7, 24 loci), HHV-6B, Merkel cell polyomavirus, and three anelloviruses. Variation at the human leukocyte antigen (HLA) complex generated the strongest associations ($p = 1.1\times10^{-8}$ to $9.1\times10^{-851}$). Effects of *HLA* alleles exhibited specificity to different viruses and varied by age and body site, and *HLA-B*08:01* also exhibited a host-virus genetic interaction with EBV subtype ($p = 2.3\times10^{-44}$). Other human genetic effects involved genes encoding proteins that process peptides for antigen presentation, such as *ERAP1* (HHV-7, $p = 2.7\times10^{-78}$) and *ERAP2* (EBV, $p = 4.6\times10^{-111}$). Mendelian randomization analyses indicated that although EBV infection strongly increases risk of multiple sclerosis (MS), EBV DNA load is unlikely to further modulate risk of MS ($p = 0.52$). In contrast, EBV viral load exhibited a strong causal effect on increased risk of Hodgkin lymphoma (OR = 19.81 [3.04–129] per s.d. increase in EBV load, $p = 1.8\times10^{-3}$). This suggests that higher chronic EBV viral load increases lymphoma risk, whereas associations of EBV infection with autoimmune conditions reflect host immune responses to particular viral epitopes.

## Introduction

Diverse viruses commonly populate numerous body sites in healthy humans[1–3]. After primary infection, some viruses, including Epstein-Barr virus (EBV) and other herpesviruses, persist for life within a reservoir of latently infected host cells, occasionally reactivating to spread within and

among people[4]. The balance between viral replication and immune recognition within an individual can result in spontaneous clearance of the virus[5], stable persistence at a set point of viral load[6], or progression to disease, such as the progression from human immunodeficiency virus (HIV) infection to acquired immunodeficiency syndrome (AIDS)[7–9]. EBV infection is a large risk factor for Hodgkin lymphoma[10] and multiple sclerosis[11] years after primary infection, and is associated with other autoimmune disorders for reasons that are not fully understood[12].

Large-scale DNA sequencing studies of blood samples from thousands of individuals have begun to characterize the “blood DNA virome”—the population of viruses that is resident within each person’s blood system—which spans a broad spectrum of viral species[13–17]. Two commonly observed taxa are herpesviruses, which are latently present in most humans, and anelloviruses, which are thought to be commensal or even symbiotic[2,3]. The amount of a virus present within an individual—the viral “load”—varies over time and is a potential biomarker of immune function[18–21]. Increased viral load may indicate reduced immunocompetence and can lead to disease (such as AIDS or liver cancer from viral hepatitis[22]), either as a direct result of inadequate control of an existing virus or through susceptibility to new infection.

The extent to which viral load in blood varies among individuals and across the lifespan is less well-characterized, and the factors that drive such variation—and its potential effects on human health—are largely unknown. Human genetic variants in the major histocompatibility complex (MHC) region on chromosome 6 and a few other genomic loci have been shown to influence load of HIV[23,24] and hepatitis C[25,26], chronic persistence (vs. spontaneous clearance) of hepatitis B[27,28] and C[29,30], and severity of COVID-19[31]. Less is known about variation in load of those viruses that commonly reside latently in healthy individuals, though immune (antibody) response to EBV and other common viruses is known to be influenced by host genetics[32–36]. The drivers of viral load variation in tissues other than blood are largely unexplored.

In this work, we performed the largest analysis of human viromes to date, by quantifying the load of 31 DNA viruses in blood or saliva samples from over 900,000 individuals. These analyses uncovered effects of age, sex, genetic ancestry, and smoking on viral load—effects that differed across viruses and between blood and saliva. Analyses also identified polygenic effects of host genetic variation across the human genome on load of several viruses and revealed a causal association between EBV viral load and increased risk of Hodgkin lymphoma.

## Results

### The DNA virome in 856,319 blood and 61,418 saliva samples

Blood and saliva are compartments of the body in which many DNA viruses transit between latent infection and lytic activation. Blood and saliva are also the two primary sources of genomic DNA in human population genetic studies. To quantify viral DNA load from widely available whole genome sequencing (WGS) data, we analyzed sequencing reads previously generated from high-coverage WGS of 490,401 blood samples in UK Biobank[37] (UKB), 365,918

blood samples and 48,899 saliva samples in the All of Us (AoU) data set[38], and 12,519 saliva samples in SPARK[39] (**Fig. 1a**). We aligned sequencing reads of potential non-human origin to a reference panel we assembled of 31 common DNA viruses (**Supplementary Fig. 1a** and **Supplementary Table 1**), many of which were previously observed in earlier WGS-based blood virome studies[13–17]. In the following analyses, we use the term "prevalence" to refer to the proportion of research participants with at least one WGS read pair aligning to a given viral genome and "abundance" to be the number of WGS read pairs from a given sample aligned to a given virus. The viral DNA fragments counted in these measurements of viral DNA load could be derived from viral genomes present within host cells or from cell-free infectious particles (virions) present in bodily fluids. In blood, viral DNA is known to be more frequently host-cell-derived for EBV[40] and torque teno viruses (TTVs, a common form of anelloviruses)[20].

As in other cohorts[13–17], viral DNA was infrequently detected in blood WGS data from UKB and AoU, with most individuals having no sequencing reads confidently attributable to any virus (**Fig. 1b,c** and **Supplementary Fig. 1b**). This was expected despite the near-ubiquity of herpesviruses and anelloviruses in humans across populations[2] (e.g., 94.7% EBV seroprevalence in UK Biobank) given the low fractions of latently infected cells that carry viral DNA[6] (**Methods**): stochastically observed viral reads reflect the number of copies of a viral genome that are present in a person's blood. Despite this sparsity, several properties of the data set supported genuine viral origin of these reads: (i) reads mapped evenly across each viral genome (indicating successful filtering of alignment artifacts; **Supplementary Fig. 2**); (ii) prevalence of viral reads associated strongly with seropositivity for viral antigens measured in UKB (EBV: OR = 21.5, $p = 5.8\times10^{-28}$; HHV-7: OR = 2.5, $p = 9.7\times10^{-6}$; CMV: OR = 6.5, $p = 0.053$); and (iii) prevalence of viral reads tracked with lymphocyte percentage (measured by hematological assays in UKB) for herpesviruses known to latently reside in lymphocytes (EBV in B cells and HHV-6B and HHV-7 in T cells[41]) but not for anelloviruses, which infect a broader range of cell types[42] (**Supplementary Fig. 1c**). A small fraction of blood samples (675 out of 490,401 in UKB) exhibited unusually large numbers of parvovirus B19-derived reads[13,14,17]; these donors had greatly depleted reticulocyte counts ($p = 1.4\times10^{-67}$, **Supplementary Fig. 1d**), consistent with active viral infection of red blood cell precursors by B19V[43]. High, ~15x coverage of the HHV-6A or HHV-6B genome by WGS reads was also observed in ~1% of donors (**Fig. 1c** and **Supplementary Fig. 1b,e**), indicating inheritance of haplotypes containing copies of these viral genomes that have integrated into human chromosomes (endogenous HHV-6; eHHV-6)[44]. These carriers (0.32% of donors with eHHV-6A and 1.03% with eHHV-6B, comparable to previous work[44,45]) were excluded from analyses of HHV-6 viral load.

In contrast to blood, saliva WGS data contained viral DNA sequences that were highly prevalent and abundant in SPARK (**Fig. 1d,e**) and AoU (**Supplementary Fig. 1i**). Several herpesviruses exhibited strikingly bimodal distributions of observed viral read counts, with saliva WGS samples tending to either contain no reads from the virus or to contain many such reads (tens, hundreds, or thousands; **Fig. 1d**). This suggests that for these viruses that infect >90% of people, viral DNA in saliva is coming not from constantly present latent infections, but rather from frequent bursts of active (lytic) viral replication and release that vary widely in intensity. For HHV-7, these lytic episodes appeared to occur continuously, such that HHV-7 DNA was detectable in nearly

all saliva samples from adults (**Supplementary Fig. 1i**), whereas for EBV, viral shedding in saliva appeared to be common but intermittent, with roughly half of saliva samples exhibiting no EBV DNA (**Fig. 1d**). In contrast to herpesviruses, DNA from anelloviruses was detected in saliva only modestly more frequently than in blood, typically at low abundances (one or a few reads; **Fig. 1c,e**), consistent with anelloviruses having a distinct replication cycle[42]. Several viruses that were rarely detected in blood (<0.1% prevalence) were detected in saliva at low to moderate prevalence (**Fig. 1c,e**). Merkel cell polyomavirus (MCPyV)—which commonly infects skin cells and can cause Merkel cell carcinoma—was seen in 5.7% of saliva samples (**Fig. 1e**), supporting a previous study[46]. Endogenous HHV-6A and HHV-6B were observed at rates similar to UKB (0.37% and 0.78%, respectively; **Fig. 1e** and **Supplementary Fig. 1f**). Viral load levels were broadly independent of autism spectrum disorder status in SPARK (**Supplementary Fig. 1g,h**) and replicated in AoU saliva samples (**Supplementary Fig. 1i**).

**Viral load is shaped by age, sex, circadian and seasonal effects**

The measurement of the prevalence and abundance of diverse viruses in the blood and saliva of so many people made it possible to recognize effects of age, sex, and circadian and seasonal dynamics through cross-sectional analyses of the large cohort. Age and sex exhibited strong effects on viral load (**Fig. 2a-d**). In blood samples from UKB, which recruited participants 40–70 years old, viral load of EBV and TTVs were greater in older individuals and thus appear to increase with age (as recently observed for TTVs[20]), whereas HHV-7 and HHV-6B viral load were greater in younger individuals and thus appear to decline with age (**Fig. 2b**). In saliva samples from SPARK, which represented infancy to old age, viral load (measured by viral read prevalence or mean abundance) exhibited trajectories that varied dramatically with age and across viruses (**Fig. 2a** and **Supplementary Fig. 3a-d**). Most viruses appear to increase rapidly in prevalence in the first several years of life (**Fig. 2a**), presumably reflecting primary infection, after which EBV load continued to increase with age, whereas HHV-6B load dropped sharply and steadily from childhood onward (**Fig. 2a**), perhaps indicating increasing control by the host's adaptive immune system over life. HHV-7 viral read prevalence in saliva increased steadily to >95% before modestly dropping in middle age (**Fig. 2a**), consistent with the pattern in UKB blood samples (**Fig. 2b**), whereas HHV-7 viral abundance in saliva began dropping beginning in childhood (**Supplementary Fig. 3c**), suggesting diminishing intensity or frequency of lytic events.

Notably, viral load was consistently higher in men than in women, across seven viruses and in both blood and saliva (**Fig. 2a-d**). This aligns with observations of stronger immune responses and lower viral load of HIV and hepatitis B in women relative to men[47]. These sex differences in viral load appeared to emerge in adolescence (**Fig. 2a**). Viral load also displayed substantial circadian and seasonal variation (**Fig. 2e,f** and **Supplementary Fig. 3e,f**). Viral read prevalence increased by ~20–30% from morning to evening (based on recorded times of blood sample acquisition) for both EBV and HHV-7, but with somewhat different trajectories over the course of the day (**Fig. 2e**). EBV viral read prevalence was also ~30% higher in winter compared to summer months, whereas HHV-7 showed different, more modest seasonal variation (**Fig. 2f**). These patterns replicated in AoU saliva samples (for EBV; **Supplementary Fig. 3e,f**) and were

not driven by cyclical variation in lymphocyte percentages of white blood cells, which exhibited weaker and different patterns (**Supplementary Fig. 3g,h**). The high prevalence of EBV reads in saliva with similar circadian patterns as blood, together with the bimodal abundance of EBV DNA in saliva (**Fig. 1d**), suggests lytic events may be quite frequent and play a larger role in determining systemic load.

Viral load also exhibited striking variation across genetic ancestries, with different viruses enriched in different ancestry groups (**Supplementary Fig. 4**). EBV DNA was least commonly observed in European-ancestry donors and most commonly observed in blood samples from African-ancestry individuals, varying in prevalence by nearly 2-fold across ancestries in both UKB (**Supplementary Fig. 4a**) and AoU (**Supplementary Fig. 4b**). Surprisingly, these ancestry effects were somewhat different in saliva samples, with individuals of East Asian and American ancestries exhibiting the highest EBV prevalence and abundance in saliva, consistently across AoU and SPARK (**Supplementary Fig. 4c-f**). In admixed AoU participants, EBV viral read prevalence in blood WGS tracked with proportions of African and American admixture (**Supplementary Fig. 4g,h**). Other viruses exhibited even more variation in blood viral load across ancestries: HHV-7 viral read prevalence was 2–4 fold higher in African-ancestry individuals compared to other ancestries (**Supplementary Fig. 4i,j**), and TTV viral read prevalence exhibited a 5–9-fold range, from 2.19% (2.15–2.23%) in European-ancestry UKB participants to 18.8% (16.8–20.8%) in participants of East Asian ancestry (**Supplementary Fig. 4k,l**).

**Many distinct genetic effects shape load of different viruses**

Almost all human traits are shaped by genetic variation, in ways that can elucidate molecular mechanisms. To identify such human genetic effects on viral load, we performed genome-wide association analysis (GWAS) on viral load phenotypes in UKB, SPARK, and AoU (**Methods**). Common variation (MAF>0.1%) at 73 loci in the human genome associated (at $p<5\times10^{-8}$) with viral DNA load of one or more viruses in blood (of five tested): 45 loci for EBV (meta-analyzed across UKB+AoU), 24 loci for HHV-7, five for TTV (TUS01), four for TTV (HD14a), and one for TTV (VT-416) (**Fig. 3a-c**, **Supplementary Fig. 5**, and **Supplementary Table 2**). Common variants also associated with viral load in saliva samples from SPARK for EBV, HHV-6B, HHV-7, and MCPyV (1–2 loci each; **Supplementary Fig. 6**). We verified that these associations were not driven by misalignment of human reads to viral genomes (**Supplementary Fig. 7a,b** and **Supplementary Note 1**) and did not overlap any of the 10 telomeric sites of eHHV-6 chromosomal integration we identified in UKB (**Supplementary Fig. 7c,d** and **Supplementary Note 2**).

For all seven viruses, the MHC region harbored the strongest common human genetic effects on viral load (**Fig. 3a-c**, **Supplementary Fig. 5** and **6**, and **Supplementary Table 2**). The MHC region as a whole contributed large proportions of viral load heritability (56% for EBV, 32% for HHV-7, and 14–30% for the anelloviruses in UKB; **Fig. 3d**), as previously observed for HIV[24]. Association patterns in the MHC region—the specific associating SNPs and their relative strengths of association—were unique to each virus (**Fig. 3e,f** and **Supplementary Table 3**).

For EBV viral load in blood, the top genetic association was generated by the human leukocyte antigen (HLA) *DRB1*04:04* allele (**Fig. 3e**), which associated with almost two-fold increased prevalence of EBV reads (23.0% (s.e. 0.2%) in heterozygous carriers versus 13.6% (s.e. 0.05%) in noncarriers). DRB1 is a co-receptor that EBV binds to enter human cells[48]; *DRB1*04:04* might encode an isoform with higher affinity for EBV protein gp42, facilitating infection. Several other independent associations with EBV load were centered in or near other HLA genes (**Fig. 3e**). This pattern was also observed (but less pronounced) for HHV-7 (**Fig. 3f**). Some of these effects likely reflect the role of the HLA system in viral antigen presentation and immune response: specific HLA alleles that strongly enable presentation of peptides from a given virus may drive distinct effects of HLA alleles on load of different viruses. Genetic variation in the MHC region has also been observed to associate with measurements of several anti-EBV antibodies[35,36], but only a subset of these serological measurements (VCA-p18 and ZEBRA) exhibited association patterns similar to that of EBV viral load (**Supplementary Fig. 8a-j** and **Supplementary Note 3**), suggesting that levels of some but not all antibodies are influenced by viral load.

Haplotypes at the *ERAP1–ERAP2–LNPEP* locus (on chromosome 5), which encodes the ERAP1, ERAP2, and LNPEP peptidases that process peptides for MHC display[49], associated strongly with viral load of EBV ($p = 1.8\times10^{-77}$ in UKB) and HHV-7 ($p = 2.7\times10^{-78}$; **Fig. 3g**). For HHV-7, the association signal appeared to be driven by two missense variants in *ERAP1* (in high linkage disequilibrium with each other, $r^2$=0.7) that affect substrate preference (Q730E) and enzymatic activity (K528R)[50]. In contrast, EBV viral load associated more strongly with variation near *ERAP2* and *LNPEP*, including a common cryptic splice variant (rs2248374) that generates a truncated *ERAP2* transcript that undergoes nonsense-mediated decay[51].

Beyond the very large effects from the *HLA* and *ERAP1–ERAP2–LNPEP* loci, each virus associated with additional human genomic loci, which were generally virus-specific except for associations at the *EOMES*, *PRDX6*, and *SLC8A1* loci (**Fig. 3a-c**, **Supplementary Fig. 5** and **6**, and **Supplementary Table 2**). These loci contained many genes with known or likely roles in viral infection, immune evasion, and proliferation. Some of these loci, such as the B-cell lymphoma (*BCL*) genes (associated with EBV viral load), likely encode alleles that affect survival and proliferation of infected host cells[52]. In contrast, *PML*, *SP110*, and *MX2* (associated with HHV-7 viral load) encode nuclear proteins with roles in restricting herpesvirus replication[53,54]. Associations at genes that influence apoptosis (*FAS*, *TNFAIP3*, and *TBKBP1*) suggest that virally-induced apoptosis signaling might be particularly important for anellovirus survival[55]. At the *ABO* locus, which determines the presence of blood group antigens bound by some viruses[56,57], alleles that produce type A and type B antigens both associated with higher HHV-7 viral load (**Supplementary Fig. 8k,l**). Associations at the *LILR* (leukocyte immunoglobulin-like receptor) and *CPVL* (carboxypeptidase vitellogenic like) genes, like the strong associations at *HLA* and *ERAP1*/*ERAP2*, indicated the importance of genetic variation that modulates viral peptide presentation and recognition[58,59].

Gene-level burden tests of rare loss-of-function and missense variants additionally identified four genes in which rare protein-coding variants associated with increased viral load: *TERT* and

*ASAH2B* for EBV, and *MX2* and *ZNF584* for HHV-7 ($p < 2.8\times10^{-6}$; **Supplementary Table 4**). These associations appeared to be virus-specific, except for *ZNF584*, for which rare coding variants also associated with increased EBV viral load ($p = 3.1\times10^{-3}$). At two loci (*CPVL* and *TNFRSF13B*) at which a missense variant generated the top common-variant association with EBV viral load, the burden test demonstrated concordant effects of loss-of-function variants in these genes (**Supplementary Fig. 8m,n**).

**Complex effects of MHC variation on viral load**

The strong effects of genetics and age on viral load led us to wonder how genetic effects on viral load might themselves vary across donors. For both EBV and HHV-7, the effect size of the lead variant in the MHC region (*HLA-DRB1*04:04* for EBV) was greater in age ranges with higher mean viral load (**Fig. 4a,b**). Variants at MHC appeared to generally act multiplicatively (rather than additively) with the effects of genetic variation elsewhere in the human genome (**Supplementary Fig. 9a,b**), deviating from the standard model of additive polygenic effects. The strong effects from the MHC region also exhibited complex interactions with one another. Variants in the MHC region generally associated with effects on EBV viral load that were consistent among *DRB1*04:04* carriers and non-carriers, but we observed several exceptions potentially reflecting interactions between the set of presented peptides presented by *DRB1*04:04* and other alleles in generating an effective immune response. (**Supplementary Fig. 9c**). The *DRB1* T210 allele associated with decreased EBV load in noncarriers of *DRB1*04:04* but with increased EBV load in carriers (**Fig. 4c**), and the *DQA1* E198 allele exhibited a similar reversal (**Fig. 4d**). The *DPB1* V105 allele also appeared to display strongly non-additive interactions with *DRB1*04:04* (**Supplementary Fig. 9d**).

Some associations of variants in the MHC region with EBV DNA load also exhibited specificity to blood or saliva. Although most variants had broadly concordant associations with EBV viral load in blood and saliva, the *HLA-A*02:01* allele appeared to decrease EBV viral load in blood ($p = 2.9\times10^{-431}$) but have no effect on EBV load in saliva ($p = 0.099$; **Fig. 4e**). The effect size of *HLA-A*02:01* in blood was consistent across UKB and AoU, and the absence of an effect in saliva was consistent in SPARK and AoU (**Fig. 4f**). One possible explanation is that *A*02:01* could present a peptide from a latent-phase EBV antigen, which could help control infection in blood but have little impact on recognition of the lytically infected plasma or epithelial cells that more directly influence viral load in saliva.

***HLA-B*08:01* reduces the load of EBV type 1 but not EBV type 2**

EBV is known to exist in two main strains, type 1 and type 2, which differ primarily in the sequences of key latency-associated EBV genes[60,61] and may interact differently with the host immune system. This raises the possibility that some human genetic variants involved in host-virus genetic interactions—e.g., HLA alleles with affinities for specific viral peptides—might associate specifically with viral load of one EBV type but not the other, as has been observed for hepatitis C[26]. To search for such effects, we separately quantified type 1 and type 2 EBV by realigning WGS reads to EBV type 1 and type 2 reference genomes and identifying reads that

aligned within two type-informative genomic regions (*EBNA2* and *EBNA3A–EBNA3C*)[62] (**Fig. 5a**, **Supplementary Fig. 10a,b**, and **Methods**). The prevalence of reads from EBV type 2 relative to type 1 varied significantly by an individual's place of birth (**Fig. 5b,c** and **Supplementary Fig. 10c**), as expected[63]. EBV type 2 was relatively more prevalent in individuals born in Africa[63] compared with individuals born in Europe ($p = 5.0\times10^{-8}$) or North America ($p = 0.010$; **Fig. 5b**), and its relative prevalence also varied regionally among individuals born in the United Kingdom (**Fig. 5c**) independently of overall EBV prevalence (**Supplementary Fig. 10d**). Interestingly, quantitative antibody responses to EBV antigens were slightly higher in individuals with type 2 EBV than type 1 (**Supplementary Fig. 10e-h**), suggesting that type 2 EBV might generate a stronger autoimmune response for autoantigen epitope mimics.

Genetic association analyses leveraging this additional information about EBV type revealed HLA variation conferring type-specific protection against EBV infection (**Fig. 5d-g**). In analyses of EBV type 1 and type 2 viral load separately, association patterns in the MHC region (**Fig. 5d,e**) were broadly similar to those we had observed for total EBV viral load (**Fig. 3e**), presumably reflecting the contributions of many human genetic effects that were difficult to resolve. However, a case-case association analysis—in which genetic data from individuals with higher type 2 load were compared against genetic data from those with higher type 1 load—generated a readily interpretable association pattern driven by the MHC class I *B*08:01* allele ($p = 9.4\times10^{-48}$; **Fig. 5f**). *B*08:01* appeared to be protective against type 1 EBV (OR = 0.74 [0.71–0.76] for observing a type 1 read; $p = 6.4\times10^{-57}$) but associated with increased risk for type 2 EBV (OR = 1.25 [1.16–1.36]; $p = 4.0\times10^{-8}$), suggesting that it may have high affinity for one or more peptides derived from the type 1 EBNA2 or EBNA3 proteins, but not homologous type 2 proteins. A possible candidate is an antigen from EBNA-3A that enables type 1-specific cytotoxic T-lymphocyte activity and is B*08:01-restricted[64,65]. Conditional association analyses that included *B*08:01* as a covariate suggested that this haplotype accounts for most of the type-specific association signal in the MHC region, with some remaining effects of variation in both MHC class I and II regions (**Fig. 5g**).

**EBV DNA load is a causal risk factor for Hodgkin lymphoma and not multiple sclerosis**

We next sought to identify associations between viral DNA load and the abundant biological and clinical phenotypes available for the 490,401 UKB research participants. We identified 181 significant associations between viral DNA load and disease phenotypes, and 366 with blood count, biomarker, and metabolite measurements (Bonferroni-corrected $p<0.05$; **Fig. 6a** and **Supplementary Tables 5–8**). Most of the disease associations involved higher viral loads in affected individuals (**Supplementary Fig. 11a**), and many associated diseases involved inflammation or weakened immune function, such as AIDS, anemia, diabetes, and renal failure (**Fig. 6a**). Organ transplantation associated strongly with elevated anellovirus load, presumably reflecting use of post-transplant immunosuppression medication[21]. Carrier status for endogenous HHV-6B or HHV-6A did not significantly associate with any biological or clinical phenotypes (**Supplementary Table 5**).

Many of the strongest associations of EBV and HHV-7 load with disease phenotypes were directly or indirectly related to smoking (e.g., nicotine dependence, respiratory disorders, and lung cancer; **Fig. 6a** and **Supplementary Tables 5 and 6**). Cumulative smoking exposure (pack-years) strongly associated with increased prevalence of EBV DNA, which was nearly twice as high for the heaviest smokers compared to nonsmokers ($p = 3.3\times10^{-195}$; **Fig. 6b**). In contrast, smoking exposure associated with decreased prevalence of HHV-7 DNA ($p = 9.6\times10^{-22}$; **Fig. 6b**). Cigarettes smoked per day associated with similar trends ($p = 3.3\times10^{-563}$ and $2.7 \times 10^{-9}$ for increased EBV prevalence and decreased HHV-7 prevalence; **Fig. 6c**), suggesting that smoking has strong, opposite effects on viral load for EBV and HHV-7. These effects replicated in AoU blood samples (**Supplementary Fig. 11b**), and in AoU saliva samples, smoking associated with an increase not only in EBV prevalence ($p = 3.6\times10^{-18}$) but also in EBV abundance among EBV-DNA-positive individuals ($p = 4.2\times10^{-8}$, **Supplementary Fig. 11c,d**).

The large number of genetic effects on EBV load identified by our GWAS provided an opportunity to investigate the causal direction of phenotypic associations: i.e., whether viral load affects (or is affected by) the clinical phenotypes with which it associates. We did this by using these load-influencing SNPs as instrument variables for Mendelian randomization, asking whether or not the genetic variants that affect vial load also affect the traits with which viral load associates in populations[66]. While most viral load-phenotype associations are likely a consequence of pathology, treatment, or environmental exposures (such as smoking) leading to elevated viral DNA presence in blood, finding that the alleles that increase risk for viral load also consistently confer risk for disease offers evidence that viral load is causally upstream.

To use this approach to assess the potential effect of EBV load on autoimmune diseases and blood cancers epidemiologically associated with EBV infection[4], we analyzed GWAS summary statistics from a meta-analysis of FinnGen[67], UK Biobank[68], and the Million Veteran Program[69] using four MR methods, excluding the MHC region to guard against pleiotropy from linkage disequilibrium[66] (**Supplementary Table 9**). For multiple sclerosis (MS), for which prior EBV infection is a large risk factor[11], these analyses indicated that latent viral load after primary infection is unlikely to further influence the risk of developing MS (MR-Egger $p = 0.52$, **Fig. 6d** and **Supplementary Fig. 11e**) despite UKB participants with MS having elevated EBV viral load ($p = 0.005$). For systemic lupus erythematosus and rheumatoid arthritis, which also associated with elevated EBV load in UKB ($p = 4.6\times10^{-7}$ and $2.2\times10^{-23}$, respectively), MR analyses were inconclusive about causality (**Supplementary Fig. 11f,g**).

In contrast, all analyses (across the four Mendelian randomization methods) suggested that viral load after EBV infection significantly influences the development of Hodgkin lymphoma (HL; MR-Egger OR = 19.8 [3.04–129] per s.d. increase of EBV load, $p = 1.8\times10^{-3}$; **Fig. 6e,f**). Effect estimates for non-Hodgkin lymphoma, chronic lymphocytic leukemia, and lymphoid leukemia were smaller and not consistently significant across MR approaches, suggesting specificity for HL (**Supplementary Fig. 11h-j**). Moreover, EBV DNA abundance in people without HL at the time of blood draw for WGS was quantitatively predictive of their likelihood to go on to develop HL (**Fig. 6g**). Elevated EBV viral load in blood—which could reflect a higher proportion of

latently infected B-cells—may increase the opportunities for oncogenic transformation leading to HL.

## Discussion

Here, by profiling common DNA viruses in blood and saliva in population sequencing data (**Fig. 1**), we have characterized the major factors that shape these constituents of the human virome. Human genetics (**Fig. 3-5**), age (**Fig. 2a,b**), sex (**Fig. 2c,d**), time of day (**Fig. 2e**), month of the year (**Fig. 2f**), and smoking behavior (**Fig. 6b,c**) consistently exert strong influences on viral load across many viruses, but the relative magnitudes and even directions of these effects are often virus-specific. These results on common, latent viruses, made possible by cross-sectional analyses across over 900,000 biobank participants, complement previous insights from targeted studies of highly pathogenic viruses.

The largely distinct, polygenic human genetic effects that we observed for load of several common viruses (EBV, HHV-7, and anelloviruses) expand upon earlier work identifying key loci influencing control of HIV[23,24] and hepatitis C[25,26] load. The MHC region contributed the strongest genetic effects for each virus we studied, similar to other GWAS of infectious disease phenotypes[70], but the specific HLA alleles that influenced viral load varied across viruses (**Supplementary Table 3**). These virus-specific effects at the MHC probably contribute to the striking variation of viral load that we observed across genetic ancestries, which likely reflects the balancing selection that has shaped human genetic variation affecting the immune system: alleles that confer an advantage against a particular virus may increase susceptibility to other pathogens[71]. Beyond the MHC, polygenic influences on load of different viruses were likewise largely distinct, sometimes surprisingly so: for example, variation at *MX2* and *PML*, which encode host factors thought to be broadly restrictive against herpesviruses[53,54], associated with load of HHV-7 but not EBV. We also want to highlight the utility of using pairs of related phenotypes, such as type 1 and type 2 EBV viral load (**Fig. 5**) or EBV viral load in blood and saliva (**Fig. 4e,f**), as a mechanism to accomplish fine-mapping of a complex genetic association.

The large number of polygenic effects we observed for EBV viral load—45 human genomic loci—enabled Mendelian randomization to evaluate causality of associations with clinical phenotypes (**Fig. 6**). These analyses suggested that multiple sclerosis risk from EBV infection may be mediated largely by adaptive immune response to an EBV infection[72] without an additional influence of subsequent lifetime viral load. In contrast, EBV viral load appears to be a causal risk factor for Hodgkin lymphoma (HL). This is not unexpected given the risk of EBV-positive HL conferred by infectious mononucleosis[10], but it suggests that interindividual differences in lifetime management of viral load affect health risks beyond infection itself. If so, interventions that reduce EBV viral load could potentially decrease the risk of HL. Notably, smoking has previously been found to be a risk factor for EBV-positive HL[73] (OR=1.81), but not for EBV-negative HL and more weakly for NHL[74] (OR=1.07), suggesting that the effect of smoking on EBV viral load might be one mechanism underlying HL subtype-specific risk. Likewise, antivirals such as acyclic nucleoside analogs that inhibit herpesvirus DNA

polymerases and reduce EBV viral load in blood[75] may be worth evaluation for potential prophylactic benefit against HL.

Our analyses of the human virome from population WGS data had several limitations. While biobank WGS data sets offer the opportunity to profile viromes at an unprecedented scale (requiring only computational analyses of existing data), the individual-level viral load phenotypes that we generated were noisy, stochastic observations, often of only a single viral DNA fragment present in a sequenced blood sample. These analyses will need to be revisited in future studies of even larger population biobank cohorts that increase the number of genetic instruments available for Mendelian randomization, which will improve the resolution and robustness of MR analyses. Saliva samples, in which we found viral DNA to be more abundant, allowed assessing DNA load for more viruses with greater precision, but most WGS data have been generated from blood. Studying the virome at higher resolution and in other body sites will require other data sources[76,77]. Additionally, although we could infer average trajectories of viral load with age, time of day, and time of year using cross-sectional analyses, studying individual-level viral load trajectories will require longitudinal sampling[6]. The analytical pipeline that we utilized also imposed limitations: future work could expand the set of viral genomes considered and undertake further strain-level genetic association analyses, which we used here to fine-map complex genetic associations at HLA with EBV. Finally, the cohorts we studied here, from the UK and United States, provided limited representation of viruses that are endemic in other parts of the world. We envision that biobank WGS data will continue to be a powerful resource for virology and epidemiology, particularly in other geographic regions with a greater burden of certain DNA viruses such as HIV, hepatitis B, hepatitis C, and HTLV-1.

## Figure Legends

**Figure 1: Detection and quantification of viral load in blood and saliva WGS.**

a) Detection (presence/absence) and quantification (abundance) of viral DNA sequences in whole-genome sequencing data derived from human blood and saliva samples, including blood samples from the UK Biobank (n=490,401) and All of Us cohorts (n=365,918) and saliva samples from the Simons Foundation Powering Autism Research (SPARK, n=12,519) and All of Us cohorts (n=48,899). Sequence reads that did not map to the human genome were aligned against a reference panel of 31 viral genomes to identify and tabulate viral DNA fragments.
b) Distribution of viral DNA fragment counts (read pairs) from EBV, HHV-7, and HHV-6B genomes in blood-derived WGS from UKB participants (n=490,401).
c) Prevalence (population frequency) and mean abundance of viral DNA sequences in UKB blood WGS. Mean abundance indicates the geometric mean of the number of observed read pairs among samples in which the virus was detected. For HHV-6, data from carriers of chromosomally integrated endogenous HHV-6A and HHV-6B are indicated separately in gray. Viruses that appear in subsequent figures are indicated with colors.
d) Distribution of viral DNA fragment counts (read pairs) from EBV, HHV-7 and HHV-6B genomes in saliva-derived WGS from SPARK participants (n=12,519).
e) Prevalence and abundance of viral DNA sequences in SPARK saliva WGS.

**Figure 2: Age, sex, time of day, and time of year effects on viral load.**

a) Prevalence of viral DNA sequences by age and sex in saliva-derived WGS from SPARK. Error bars, 95% CIs.
b) Prevalence of viral DNA sequences by age and sex in blood-derived WGS from UKB. Anellovirus (TTV) measurements group together reads aligned to the TUS01, HD14a, and VT416 reference genomes. Error bars, 95% CIs.
c) Prevalence of viral DNA sequences by sex in blood-derived WGS from UKB (EBV, HHV-7, HHV-6B, TTV) and AoU (CMV) and saliva-derived WGS from AoU (EBV, CMV, MCPyV, HSV-1) and SPARK (HHV-7, HHV-6B). Error bars, 95% CIs.
d) Abundance of viral DNA sequences by sex in saliva-derived WGS from AoU (EBV, CMV, MCPyV) and SPARK (HHV-7, HHV-6B). Error bars, 95% CIs.
e) Prevalence of viral DNA sequences in UKB blood WGS by time of day of sample collection (rounded to the nearest hour). P-values are from an ANOVA test between models with and without collection hour (see **Methods**). Error bars, 95% CIs.
f) Prevalence of viral DNA sequences in UKB blood WGS data by month of sample collection. P-values are from an ANOVA test between models with and without collection month. Error bars, 95% CIs.

**Figure 3: Genome-wide association analyses of viral load in blood.**

a) Genome-wide associations of human genetic variants with EBV load in blood (inverse-normal transformed abundance), meta-analyzed across UKB (n=453,770) and AoU (n=201,170) European-ancestry sub-cohorts.

b) Genome-wide associations with HHV-7 load in blood (inverse-normal transformed abundance) in European-ancestry UKB participants.
c) Genome-wide associations with anellovirus TUS01 load in blood (inverse-normal transformed abundance) in European-ancestry UKB participants.
d) Estimates of heritability from common genetic variants for viral-load phenotypes from variants within versus outside of the MHC region of the human genome (defined as 24–35Mb on chromosome 6). Error bars, 95% CIs.
e) Associations of variants in the MHC region of the human genome with EBV load in blood in European-ancestry UKB participants.
f) Analogous to e, for HHV-7 load.
g) Associations of common genetic variants at the *ERAP1–ERAP2–LNPEP* locus with EBV load (green) and HHV-7 load (blue) in blood. Arrows indicate direction of transcription for nearby genes. Associations of two missense variants in *ERAP1* and a cryptic splice variant in *ERAP2* are indicated with diamonds.

**Figure 4: Complex effects of MHC variation on viral load in blood and saliva.**

a) Prevalence of detectable EBV positivity (reads mapping to the EBV genome) by age in UKB blood WGS, stratified by *HLA-DRB1*04:04* genotype. Error bars, 95% CIs.
b) Prevalence of detectable HHV-7 positivity by age in UKB blood WGS, stratified by genotype for a SNP in the MHC region at chr6:31345435. Error bars, 95% CIs.
c) Prevalence of detectable EBV positivity by *HLA-DRB1*04:04* and *HLA-DRB1* T210 genotype combinations in UKB blood WGS. Error bars, 95% CIs.
d) Prevalence of detectable EBV positivity by *HLA-DRB1*04:04* and *HLA-DQA1* E198 genotype combinations in UKB blood WGS. Error bars, 95% CIs.
e) Associations of common genetic variants in the MHC region with EBV load in blood (UKB, n=201,185; x-axis) and saliva (AoU, n=33,168; y-axis). Variants (dots in the scatter plot) are colored by linkage disequilibrium with rs12153924 (green-to-pink shading), which tags the *HLA-A*02:01* allele ($r^2$=0.99).
f) Effect sizes of *HLA-A*02:01* on EBV load in blood (green points; UKB, n=453,770; AoU, n=201,170) and saliva (gray points; AoU, n=33,164; SPARK, n=9,209). For saliva, effect sizes are shown for both a binary presence/absence phenotype and a quantitative phenotype (inverse-normal transformed abundance, restricted to individuals whose saliva WGS had detectable EBV DNA).

**Figure 5: EBV type 1 versus 2 analysis identifies a type-specific effect of *HLA-B*08:01*.**

a) Frequencies of observing WGS reads mapping to each 500bp segment of the EBV type 1 (black line) and type 2 (green line) reference genomes in saliva WGS samples from SPARK. Highlighted in gray are two regions used to distinguish EBV types, both containing genes with known type-specific sequences, *EBNA2* and *EBNA3A–EBNA3C*. Read alignments in these two regions show consistent ratios of type 1-aligned to type 2-aligned reads, whereas other EBV genomic regions (which are not type-informative) do not.
b) Proportion of UKB participants positive for EBV type 2 (among EBV-positive individuals) by location of birth. In each group, the proportion was estimated as the number of UKB

blood WGS samples containing EBV type 2-specific reads divided by the number containing any type-informative reads (total n=18,802). Estimates are shown for countries of birth with more than 50 samples containing a type-specific read, and for continental groupings (aggregated across all countries; diamonds). Error bars, 95% CIs.

c) Proportion of UKB participants positive for EBV type 2 by region of birth within the United Kingdom.
d) Associations of variants in the MHC region of the human genome with EBV type 1 positivity (presence of DNA sequences that map to type-informative regions of the EBV type 1 genome) in UKB blood WGS. Variants are colored by linkage disequilibrium relative to *HLA-B*08:01* (green-to-purple shading), indicated by the large green dot.
e) Associations of variants with EBV type 2 positivity.
f) Associations of variants with EBV type 2 positivity, using individuals with EBV type 1 positivity as controls.
g) Conditional associations of variants with EBV type 2 positivity, using individuals with EBV type 1 positivity as controls, after including *HLA-B*08:01* genotype as a covariate.

**Figure 6: Associations of viral load with clinical phenotypes and smoking.**

a) Associations of clinical conditions with detectable viral positivity (presence of viral DNA sequences in blood WGS) in European-ancestry UKB participants (n=453,770). For each virus commonly observed in UKB, several of the strongest independent associations are shown in the plot; the full set of associations is provided in **Supplementary Tables 5** and **6**. Error bars, 95% CIs.
b) Prevalence of detectable EBV positivity (in green) and HHV-7 positivity (in blue) by smoking pack-years in UKB. Prevalence among individuals who have never smoked is indicated by the square point on the left of the plot. Error bars, 95% CIs.
c) Prevalence of detectable EBV positivity (in green) and HHV-7 positivity (in blue) by cigarettes smoked per day in UKB. Prevalences among individuals who have never smoked or formerly smoked are indicated by the square points on the left. Error bars, 95% CIs.
d) Estimates of the causal effect of EBV load (exposure) on risk for multiple sclerosis (outcome) using different Mendelian randomization approaches (y-axis) with 44 non-MHC loci as instrument variables. An estimate is plotted in green if its 95% CI does not overlap an odds ratio of 1 (no effect). Error bars, 95% CIs.
e) Estimates of the causal effect of EBV load on risk for Hodgkin lymphoma. Error bars, 95% CIs.
f) Effect sizes for 44 common genetic variants (at distinct non-MHC loci)—used as instrument variables in Mendelian randomization—for EBV load (x-axis) and risk for Hodgkin lymphoma (y-axis). The line shown is from the MR-Egger test; the y-intercept is not significantly different from OR=1, indicating low pleiotropy of the genetic instruments.
g) Incidence rates of Hodgkin lymphoma (during ~15 years of follow-up), stratified by the number of EBV DNA fragments in UKB blood WGS, for individuals without a Hodgkin lymphoma diagnosis at the time of sample collection. Error bars, 95% CIs.

## Supplementary Figure Legends

**Supplementary Figure 1: Viral load quantification, associations with blood counts, and identification of endogenous HHV-6 chromosomal integrants.**

a) Locations of the five representative anellovirus reference genomes included in the viral reference panel relative to other anellovirus genomes in the multidimensional scaling projection from ref.[78].
b) Prevalence and abundance of viral DNA sequences in blood samples from AoU.
c) Prevalence of viral DNA sequences in blood by lymphocyte percentage (rounded to the nearest integer) in European-ancestry UKB participants (n=441,026). P-values are from a model including sex, age, age-squared, assessment center, and 20 genetic principal components as covariates. Error bars, 95% CIs.
d) Reticulocyte counts of UKB participants (y-axis) stratified by number of reads mapping to the parvovirus B19 genome (x-axis).
e) Numbers of reads aligned to the HHV-6A (x-axis) and HHV-6B (y-axis) genomes in blood-derived WGS from UKB. Threshold lines shown were used to determine carriers of endogenous HHV-6A or HHV-6B. A few individuals carrying both integrants are in the top right quadrant.
f) Analogous to e, for saliva-derived WGS from SPARK.
g) Prevalence and abundance of viral DNA sequences in saliva samples from children without autism spectrum disorder in SPARK.
h) Prevalence and abundance of viral DNA sequences in saliva samples from children with autism spectrum disorder in SPARK.
i) Prevalence and abundance of viral DNA sequences in saliva samples from AoU.

**Supplementary Figure 2: Coverage of viral genomes by aligned WGS reads.**

For each virus commonly detected in UKB blood WGS or SPARK saliva WGS, the proportion of samples containing reads in 500bp regions spanning the indicated virus' genome is plotted. Evenness of coverage indicates that reads are arising from sparse sampling of viral genomes across people; misalignments from homologous sequences in the human genome would appear as spikes in viral genomic coverage.

**Supplementary Figure 3: Additional data supporting age, sex, circadian, and seasonal effects on viral load in saliva.**

a) Abundance of viral DNA sequences in saliva-derived WGS by age (rounded to the nearest year) in SPARK. Error bars, 95% CIs.
b) Abundance of EBV DNA sequences by age in SPARK saliva WGS, stratified by sex. P-values are from a joint model containing both sex and an interaction term between sex and age. Error bars, 95% CIs.
c) Analogous to b, for HHV-7 DNA sequences.
d) Analogous to b, for HHV-6B DNA sequences.

e) Prevalence of EBV DNA sequences by sample collection time of day (rounded to the nearest hour) in AoU saliva WGS (n=18,751). P-values are from an ANOVA test between models with and without hour of the day. Error bars, 95% CIs.
f) Prevalence of EBV DNA sequences by sample collection month of year in AoU saliva WGS. P-values are from an ANOVA test between models with and without collection month. Error bars, 95% CIs.
g) Lymphocyte percentage in UKB blood samples by sample collection time of day. Error bars, 95% CIs.
h) Lymphocyte percentage in UKB blood samples by sample collection month. Error bars, 95% CIs.

**Supplementary Figure 4: Variation in viral load in blood and saliva by genetic ancestry.**

a) Prevalence (population frequency) of detectable EBV DNA positivity (reads that align to the EBV genome) in blood-derived WGS data by ancestry in UKB (genetically inferred). Error bars, 95% CIs.
b) Prevalence of EBV DNA positivity in blood-derived WGS data by genetic ancestry in AoU. Error bars, 95% CIs.
c) Prevalence of EBV DNA positivity in saliva-derived WGS data by genetic ancestry in AoU. Error bars, 95% CIs.
d) Abundance of EBV DNA sequences in saliva-derived WGS data by genetic ancestry in AoU. Error bars, 95% CIs.
e) Analogous to c, for saliva-derived WGS in SPARK.
f) Analogous to d, for saliva-derived WGS in SPARK.
g) Prevalence of EBV DNA positivity blood-derived WGS data from AoU by proportion of estimated American ancestry admixture (for individuals with estimated >90% European and/or American ancestry). Error bars, 95% Cis.
h) Analogous to g, for proportion of estimated African ancestry admixture for AoU participants with estimated >90% European and/or African ancestry. Admixture proportions of 10-30% African ancestry were excluded due to low sample numbers (n<500).
i) Prevalence of HHV-7 DNA positivity in blood-derived WGS data by genetic ancestry in UKB. Error bars, 95% CIs.
j) Prevalence of HHV-7 DNA positivity in blood-derived WGS data by genetic ancestry in AoU. Error bars, 95% CIs.
k) Prevalence of anellovirus DNA positivity in blood-derived WGS data by genetic ancestry in UKB. Error bars, 95% CIs.
l) Prevalence of anellovirus DNA positivity in blood-derived WGS data by genetic ancestry in AoU. Error bars, 95% CIs.

**Supplementary Figure 5: Genome-wide association analyses of anellovirus load in blood for the HD14a and VT416 TTV strains.**

a) Genome-wide associations of human genetic variants with anellovirus HD14a load in blood (inverse-normal transformed abundance) in European-ancestry UKB participants.

b) Analogous to a, for the anellovirus VT416 strain.

**Supplementary Figure 6: Genome-wide association analyses of viral load in saliva.**

a) Genome-wide associations of human genetic variants with EBV DNA positivity in saliva-derived WGS from European-ancestry AoU participants (n=33,168). Given the bimodal distribution of viral DNA load in saliva, in which samples containing virally-aligned reads often had hundreds or thousands of reads, we performed GWAS both on binarized viral load (presence/absence of any reads from the virus) and quantitative viral load phenotypes (inverse-normal transformed abundance, restricted to individuals whose saliva WGS had detectable viral DNA) to capture effects on the frequency and amount of secreted virus.
b) Analogous to a, for EBV abundance.
c) Genome-wide associations of human genetic variants with HHV-6B DNA positivity in saliva-derived WGS from European-ancestry SPARK participants (n=9,081). Carriers of endogenous HHV6-A or HHV6-B germline integrations were excluded.
d) Analogous to c, for HHV-6B abundance (n=6,258).
e) Analogous to c, for HHV-7 abundance (n=7,360).
f) Analogous to c, for Merkel cell polyomavirus positivity (n=9,209).

**Supplementary Figure 7: Robustness of genetic associations with EBV and HHV-7 load and GWAS for HHV-6 integration sites.**

a) Effect sizes for index variants at 44 independent non-MHC EBV load-associated loci for association with presence of reads aligned to the left (x-axis) or right (y-axis) half of the EBV genome.
b) Analogous to a, for index variants at 20 independent non-MHC HHV-7 load-associated loci. Three lower frequency variants of large effect size were excluded to facilitate visualization of the remaining variants.
c) Genome-wide associations of human genetic variants with carrier status for endogenous HHV-6A germline integration (eHHV-6A) in European-ancestry UKB participants. Peritelomeric peaks indicate haplotypes in linkage disequilibrium with telomeric HHV-6A insertions.
d) Analogous to c, for endogenous HHV-6B germline integration (eHHV-6B).

**Supplementary Figure 8: Analyses of serology data and effects of blood type and coding variants on viral load in blood.**

a) Associations of variants in the MHC region of the human genome with quantitative antibody response to EBV-encoded EBNA-1 in UKB serology measurements (y-axis) and EBV load in UKB blood WGS (x-axis).
b) Analogous to a, for quantitative antibody response to EA-D.
c) Analogous to a, for quantitative antibody response to ZEBRA.
d) Analogous to a, for quantitative antibody response to VCA-p18.
e) Associations to quantitative antibody response to HHV-7 encoded U14 (y-axis) and HHV-7 load (x-axis).

f) Quantitative antibody response to EBV-encoded EBNA-1 (y-axis) stratified by EBV DNA fragment count in UKB blood WGS (x-axis). P-value is from a linear regression model with standard covariates and rank-based inverse normal transformed EBV read counts. MFI, median fluorescence intensity.
g) Analogous to f, for quantitative antibody response to EBV-encoded EA-D.
h) Analogous to f, for quantitative antibody response to EBV-encoded ZEBRA.
i) Analogous to f, for quantitative antibody response to EBV-encoded VCA-p18.
j) Quantitative antibody response to HHV-7-encoded U14 (y-axis) stratified by HHV-7 DNA fragment count in UKB blood WGS (x-axis). P-value is from a linear regression model including standard covariates and rank-based inverse normal transformed HHV-7 read counts.
k) Prevalence of HHV-7 DNA positivity by genetically determined blood type in UKB.
l) Prevalence of HHV-7 DNA positivity by genotype combinations of common blood type alleles (*O*, *B*, *A1*, and *A2*) in UKB.
m) Effect size on EBV load for the *CPVL* Y168H missense variant and *CPVL* loss-of-function variants (aggregated across such variants in a burden test).
n) Effect size on EBV load for the *TNFRSF13B* C104R and A181E missense variants and *TNFRSF13B* loss-of-function variants. *TNFRSF13B* coding variants (C104R, $p = 2.3\times10^{-14}$; A181E, $p = 3.5\times10^{-12}$) lead to impaired ligand binding and downstream signaling[79].

**Supplementary Figure 9: Non-additive genetic effects on viral load in blood.**

a) Effect sizes of index variants at 44 independent non-MHC loci associated with EBV load in UKB blood WGS in *DRB1*04:04* carriers (x-axis) and *DRB1*04:04* non-carriers (y-axis).
b) Effect sizes of index variants at 23 independent non-MHC loci associated with HHV-7 load in UKB blood WGS in chr6:31345435 T/T homozygotes (x-axis) and C/C homozygotes (y-axis). Three lower frequency variants of large effect size were excluded to facilitate visualization of the remaining variants.
c) Effect sizes of variants in the MHC region in *DRB1*04:04* carriers (x-axis) and *DRB1*04:04* non-carriers (y-axis) on EBV load in UKB blood WGS. Only variants with associations that reached genome-wide significance in both sample subsets are shown. HLA coding variants and classical alleles are highlighted in red.
d) Prevalence of EBV DNA positivity by *HLA-DRB1*04:04* and *HLA-DPB1* V105 genotype combinations in UKB. Error bars, 95% CIs.

**Supplementary Figure 10: Validation of type-informative EBV genome regions and assessment of quantitative antibody responses for EBV type 1 and type 2.**

a) Frequencies of observing WGS reads mapping to each 500bp segment of the EBV type 1 (black line) and type 2 (green line) reference genomes in blood WGS samples from UKB. Highlighted in gray are two regions used to distinguish EBV types, both containing genes with known type-specific sequences, *EBNA2* and *EBNA3A–EBNA3C*. Read alignments in these two regions show consistent ratios of type 1-aligned to type 2-aligned reads, whereas other EBV genomic regions (which are not type-informative) do not.

b) Pairwise correlations in the abundances of DNA sequences mapping to 500bp segments of the EBV type 2 genome in saliva WGS samples from SPARK with >1 type-specific alignment (n=4757). For each sample, read counts for each 500bp bin were first normalized by the maximum number of reads in any 500bp region across the EBV type 1 and type 2 genomes. Regions used for downstream EBV type analyses (containing *EBNA2* and *EBNA3A*–*EBNA3C*) are highlighted. The correlation heatmap shows that reads aligning to the EBV type 2 genome in these two regions tended to occur in the same individuals.
c) Proportion of UKB participants positive for EBV type 2 (as a fraction of those with EBV reads specific to either type 1 or type 2) by country of birth within the United Kingdom and Republic of Ireland, stratified by *HLA-B*08:01* allele carrier status. The effect of *B*08:01* on EBV type 1 versus type 2 (decreasing the proportion of individuals with EBV type 1 relative to type 2) appeared to be present across all geographic regions of the United Kingdom, which vary in rates of EBV type 2 prevalence (**Fig. 5c**). Error bars, 95% CIs.
d) Proportion of UKB participants positive for EBV DNA by region of birth within the United Kingdom.
e) Distribution of quantitative antibody response to VCA-p18 for EBV seropositive UKB participants with type-informative EBV DNA sequences exclusively from the EBV type 1 (black) or type 2 (green) genomes (n=335 with available serology measurements). P-values are from linear regression including age, age-squared, sex, assessment center, and 20 ancestry PCs as covariates. MFI, median fluorescence intensity.
f) Analogous to c, for quantitative antibody response to EBNA-1.
g) Analogous to c, for quantitative antibody response to EA-D.
h) Analogous to c, for quantitative antibody response to ZEBRA.

**Supplementary Figure 11: Additional data regarding associations of viral load with clinical phenotypes and smoking.**

a) Associations of clinical conditions with detectable viral positivity (presence of viral DNA sequences in blood WGS) in European-ancestry UKB participants (n=453,770). Significant associations (Bonferroni-corrected $p<0.05$) are colored by direction of effect.
b) Prevalence of detectable EBV positivity (in green) and HHV-7 positivity (in blue) by cigarettes smoked per day in AoU blood samples. Prevalences among individuals who have never smoked or formerly smoked are indicated by the square points on the left. Error bars, 95% CIs.
c) Prevalence of detectable EBV positivity in AoU saliva WGS by cigarettes smoked per day. Prevalences among individuals who have never smoked or formerly smoked are indicated by the square points to the left. Error bars, 95% CIs.
d) Analogous to b, for mean EBV abundance (geometric mean among samples with at least one read pair).
e) Effect sizes for 44 common genetic variants (at distinct non-MHC loci)—used as instrument variables in Mendelian randomization—for EBV load (x-axis) and risk for multiple sclerosis (y-axis). The line shown is from the MR-Egger test; the y-intercept is not significantly different from OR=1, indicating low pleiotropy of the genetic instruments.

f) Estimates of the causal effect of EBV load (exposure) on risk for systemic lupus erythematosus (outcome) using different Mendelian randomization approaches (y-axis) with 44 non-MHC loci as instrument variables. An estimate is plotted in green if its 95% CI does not overlap an odds ratio of 1 (no effect). Error bars, 95% CIs. However, as these results replicated inconsistently across MR approaches with different assumptions[80], we do not conclude causality, but rather that viral load is more likely to affect the risk of systemic lupus erythematosus compared to MS.
g) Estimates of the causal effect of EBV load on risk for rheumatoid arthritis. Error bars, 95% CIs.
h) Estimates of the causal effect of EBV load on risk for lymphoid leukemia. Error bars, 95% CIs.
i) Estimates of the causal effect of EBV load on risk for non-Hodgkin lymphoma. Error bars, 95% CIs.
j) Estimates of the causal effect of EBV load on risk for chronic lymphocytic leukemia. Error bars, 95% CIs.

## References


1. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nat. Rev. Microbiol.* **19**, 514–527 (2021).
2. Koonin, E. V., Dolja, V. V. & Krupovic, M. The healthy human virome: from virus–host symbiosis to disease. *Curr. Opin. Virol.* **47**, 86–94 (2021).
3. Kandathil, A. J. & Thomas, D. L. The Blood Virome: A new frontier in biomedical science. *Biomed. Pharmacother.* **175**, 116608 (2024).
4. Damania, B., Kenney, S. C. & Raab-Traub, N. Epstein-Barr virus: Biology and clinical disease. *Cell* **185**, 3652–3670 (2022).
5. Grebely, J. *et al.* Hepatitis C virus clearance, reinfection, and persistence, with insights from studies of injecting drug users: towards a vaccine. *Lancet Infect. Dis.* **12**, 408–414 (2012).
6. Khan, G., Miyashita, E. M., Yang, B., Babcock, G. J. & Thorley-Lawson, D. A. Is EBV persistence in vivo a model for B cell homeostasis? *Immunity* **5**, 173–179 (1996).
7. Coombs, R. W. *et al.* Plasma viremia in human immunodeficiency virus infection. *N. Engl. J. Med.* **321**, 1626–1631 (1989).
8. Ho, D. D., Moudgil, T. & Alam, M. Quantitation of human immunodeficiency virus type 1 in the blood of infected persons. *N. Engl. J. Med.* **321**, 1621–1625 (1989).
9. Clark, S. J. *et al.* High titers of cytopathic virus in plasma of patients with symptomatic primary HIV-1 infection. *N. Engl. J. Med.* **324**, 954–960 (1991).
10. Hjalgrim, H. *et al.* Characteristics of Hodgkin's Lymphoma after Infectious Mononucleosis. *N. Engl. J. Med.* **349**, 1324–1332 (2003).
11. Bjornevik, K. *et al.* Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* **375**, 296–301 (2022).
12. Robinson, W. H., Younis, S., Love, Z. Z., Steinman, L. & Lanz, T. V. Epstein–Barr virus as a potentiator of autoimmune diseases. *Nat. Rev. Rheumatol.* **20**, 729–740 (2024).

13. Moustafa, A. *et al.* The blood DNA virome in 8,000 humans. *PLoS Pathog.* **13**, e1006292 (2017).
14. Liu, S. *et al.* Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* **175**, 347-359.e14 (2018).
15. Tong, X. *et al.* Peripheral Blood Microbiome Analysis via Noninvasive Prenatal Testing Reveals the Complexity of Circulating Microbial Cell-Free DNA. *Microbiol. Spectr.* **10**, e00414-22 (2022).
16. Guo, J. *et al.* The blood virome of 10,585 individuals from the ChinaMAP. *Cell Discov.* **8**, 113 (2022).
17. Linthorst, J., Baksi, M. M. M., Welkers, M. R. A. & Sistermans, E. A. The cell-free DNA virome of 108,349 Dutch pregnant women. *Prenat. Diagn.* **43**, 448–456 (2023).
18. Gulley, M. L. & Tang, W. Laboratory Assays for Epstein-Barr Virus-Related Disease. *J. Mol. Diagn. JMD* **10**, 279–292 (2008).
19. Saag, M. S. *et al.* HIV viral load markers in clinical practice. *Nat. Med.* **2**, 625–629 (1996).
20. Focosi, D. *et al.* Assessment of prevalence and load of torquetenovirus viraemia in a large cohort of healthy blood donors. *Clin. Microbiol. Infect.* **26**, 1406–1410 (2020).
21. Strassl, R. *et al.* Quantification of Torque Teno Virus Viremia as a Prospective Biomarker for Infectious Disease in Kidney Allograft Recipients. *J. Infect. Dis.* **218**, 1191–1199 (2018).
22. Chen, C.-J. *et al.* Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level. *JAMA* **295**, 65–73 (2006).
23. Fellay, J. *et al.* A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science* **317**, 944 (2007).
24. McLaren, P. J. *et al.* Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc. Natl. Acad. Sci.* **112**, 14658–14663 (2015).

25. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399–401 (2009).
26. Ansari, M. A. *et al.* Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat. Genet.* **49**, 666–673 (2017).
27. Kamatani, Y. *et al.* A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat. Genet.* **41**, 591–595 (2009).
28. Li, Y. *et al.* Genome-wide association study identifies 8p21.3 associated with persistent hepatitis B virus infection among Chinese. *Nat. Commun.* **7**, 11664 (2016).
29. Duggal, P. *et al.* Genome-Wide Association Study of Spontaneous Resolution of Hepatitis C Virus Infection: Data From Multiple Cohorts. *Ann. Intern. Med.* **158**, 235–245 (2013).
30. Thomas, D. L. *et al.* Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature* **461**, 798–801 (2009).
31. Niemi, M. E. K. *et al.* Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).
32. Rubicz, R. *et al.* A Genome-Wide Integrative Genomic Study Localizes Genetic Factors Influencing Antibodies against Epstein-Barr Virus Nuclear Antigen 1 (EBNA-1). *PLOS Genet.* **9**, e1003147 (2013).
33. Scepanovic, P. *et al.* Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. *Genome Med.* **10**, 59 (2018).
34. Sallah, N. *et al.* Distinct genetic architectures and environmental factors associate with host response to the γ2-herpesvirus infections. *Nat. Commun.* **11**, 3849 (2020).
35. Kachuri, L. *et al.* The landscape of host genetic factors involved in immune response to common viral infections. *Genome Med.* **12**, 93 (2020).
36. Mentzer, A. J. *et al.* Identification of host–pathogen-disease relationships using a scalable multiplex serology platform in UK Biobank. *Nat. Commun.* **13**, 1818 (2022).

37. Carss, K. *et al.* Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* 1–10 (2025) doi:10.1038/s41586-025-09272-9.
38. Bick, A. G. *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
39. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research: Neuron. https://www.cell.com/neuron/fulltext/S0896-6273(18)30018-7.
40. Kanakry, J. A. *et al.* The clinical significance of EBV DNA in the plasma and peripheral blood mononuclear cells of patients with or without EBV diseases. *Blood* **127**, 2007–2017 (2016).
41. Mori, Y. & Yamanishi, K. HHV-6A, 6B, and 7: pathogenesis, host response, and clinical disease. in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Arvin, A. et al.) (Cambridge University Press, Cambridge, 2007).
42. Spandole, S., Cimponeriu, D., Berca, L. M. & Mihăescu, G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Arch. Virol.* **160**, 893–908 (2015).
43. Chisaka, H., Morita, E., Yaegashi, N. & Sugamura, K. Parvovirus B19 and the pathogenesis of anaemia. *Rev. Med. Virol.* **13**, 347–359 (2003).
44. Flamand, L. *et al.* Endogenous human herpesviruses 6A/B. *J. Virol.* **0**, e01054-25 (2025).
45. Liu, X. *et al.* Endogenization and excision of human herpesvirus 6 in human genomes. *PLOS Genet.* **16**, e1008915 (2020).
46. Loyo, M. *et al.* Quantitative detection of Merkel cell virus in human tissues and possible mode of transmission. *Int. J. Cancer* **126**, 2991–2996 (2010).
47. Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).
48. Li, Q. *et al.* Epstein-Barr virus uses HLA class II as a cofactor for infection of B lymphocytes. *J. Virol.* **71**, 4657–4662 (1997).

49. Saveanu, L. *et al.* IRAP identifies an endosomal compartment required for MHC class I cross-presentation. *Science* **325**, 213–217 (2009).
50. Stamogiannos, A., Koumantou, D., Papakyriakou, A. & Stratikos, E. Effects of polymorphic variation on the mechanism of Endoplasmic Reticulum Aminopeptidase 1. *Mol. Immunol.* **67**, 426–435 (2015).
51. Andrés, A. M. *et al.* Balancing Selection Maintains a Form of ERAP2 that Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation. *PLoS Genet.* **6**, e1001157 (2010).
52. Willis, T. G. & Dyer, M. J. S. The role of immunoglobulin translocations in the pathogenesis of B-cell malignancies. *Blood* **96**, 808–822 (2000).
53. Scherer, M. & Stamminger, T. Emerging Role of PML Nuclear Bodies in Innate Immune Signaling. *J. Virol.* **90**, 5850–5854 (2016).
54. Schilling, M. *et al.* Human MxB Protein Is a Pan-herpesvirus Restriction Factor. *J. Virol.* **92**, e01056-18 (2018).
55. Verstrepen, L. *et al.* Expression, biological activities and mechanisms of action of A20 (TNFAIP3). *Biochem. Pharmacol.* **80**, 2009–2020 (2010).
56. Tan, M. & Jiang, X. Norovirus and its histo-blood group antigen receptors: an answer to a historical puzzle. *Trends Microbiol.* **13**, 285–293 (2005).
57. Wu, S.-C. *et al.* Blood group A enhances SARS-CoV-2 infection. *Blood* **142**, 742–747 (2023).
58. Kubiniok, P. *et al.* Understanding the constitutive presentation of MHC class I immunopeptidomes in primary tissues. *iScience* **25**, 103768 (2022).
59. Hirayasu, K. & Arase, H. Functional and genetic diversity of leukocyte immunoglobulin-like receptor and implication for disease associations. *J. Hum. Genet.* **60**, 703–708 (2015).
60. Sample, J. *et al.* Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. *J. Virol.* **64**, 4084–4092 (1990).

61. Adldinger, H. K., Delius, H., Freese, U. K., Clarke, J. & Bornkamm, G. W. A putative transforming gene of Jijoye virus differs from that of Epstein-Barr virus prototypes. *Virology* **141**, 221–234 (1985).
62. Zanella, L. *et al.* A reliable Epstein-Barr Virus classification based on phylogenomic and population analyses. *Sci. Rep.* **9**, 9829 (2019).
63. Zimber, U. *et al.* Geographical prevalence of two types of Epstein-Barr virus. *Virology* **154**, 56–66 (1986).
64. Moss, D. J. *et al.* Cytotoxic T-cell clones discriminate between A- and B-type Epstein-Barr virus transformants. *Nature* **331**, 719–721 (1988).
65. An Epstein-Barr virus-specific cytotoxic T cell epitope in EBV nuclear antigen 3 (EBNA 3). *J. Exp. Med.* **171**, 345–349 (1990).
66. Sanderson, E. *et al.* Mendelian randomization. *Nat. Rev. Methods Primer* **2**, 6 (2022).
67. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
68. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
69. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024).
70. Kwok, A. J., Mentzer, A. & Knight, J. C. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nat. Rev. Genet.* **22**, 137–153 (2021).
71. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2010).
72. Lanz, T. V. *et al.* Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. *Nature* **603**, 321–327 (2022).

73. Kamper-Jørgensen, M. *et al.* Cigarette smoking and risk of Hodgkin lymphoma and its subtypes: a pooled analysis from the International Lymphoma Epidemiology Consortium (InterLymph). *Ann. Oncol.* **24**, 2245–2255 (2013).
74. Morton, L. M. *et al.* Cigarette Smoking and Risk of Non-Hodgkin Lymphoma: A Pooled Analysis from the International Lymphoma Epidemiology Consortium (InterLymph). *Cancer Epidemiol. Biomarkers Prev.* **14**, 925–933 (2005).
75. Hoshino, Y. *et al.* Long-term administration of valacyclovir reduces the number of Epstein-Barr virus (EBV)-infected B cells but not the number of EBV DNA copies per B cell in healthy volunteers. *J. Virol.* **83**, 11857–11861 (2009).
76. Kumata, R., Ito, J., Takahashi, K., Suzuki, T. & Sato, K. A tissue level atlas of the healthy human virome. *BMC Biol.* **18**, 55 (2020).
77. Pyöriä, L. *et al.* Unmasking the tissue-resident eukaryotic DNA virome in humans. *Nucleic Acids Res.* **51**, 3223–3239 (2023).
78. Arze, C. A. *et al.* Global genome analysis reveals a vast and dynamic anellovirus landscape within the human virome. *Cell Host Microbe* **29**, 1305-1315.e6 (2021).
79. Salzer, U. & Grimbacher, B. TACI deficiency — a complex system out of balance. *Curr. Opin. Immunol.* **71**, 81–88 (2021).
80. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **44**, 313–329 (2020).
81. Manghi, P. *et al.* Large-scale metagenomic analysis of oral microbiomes reveals markers for autism spectrum disorders. *Nat. Commun.* **15**, 9743 (2024).
82. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
83. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

84. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
85. Kleiveland, C. R. Peripheral Blood Mononuclear Cells. in *The Impact of Food Bioactives on Health: in vitro and ex vivo models* (eds. Verhoeckx, K. et al.) (Springer, Cham (CH), 2015).
86. Insights into the causes and consequences of DNA repeat expansions from 700,000 biobank participants | bioRxiv. https://www.biorxiv.org/content/10.1101/2024.11.25.625248v1.full.
87. Conley, A. B. *et al.* Rye: genetic ancestry inference at biobank scale. *Nucleic Acids Res.* **51**, e44 (2023).
88. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8**, e64683 (2013).
89. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
90. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
91. Kamitaki, N. *et al.* Human and bacterial genetic variation shape oral microbiomes and health. 2025.03.31.25324952 Preprint at https://doi.org/10.1101/2025.03.31.25324952 (2025).
92. Hujoel, M. L. A. *et al.* Influences of rare copy-number variation on human complex traits. *Cell* **185**, 4233-4248.e27 (2022).
93. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
94. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
95. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).

96. Tang, D., Kamitaki, N., Mukamel, R. E., Rubinacci, S. & Loh, P.-R. Patterns and drivers of 43,617 mosaic chromosomal alterations in blood. 2025.07.30.25332451 Preprint at https://doi.org/10.1101/2025.07.30.25332451 (2025).

97. Hujoel, M. L. A. *et al.* Protein-altering variants at copy number-variable regions influence diverse human phenotypes. *Nat. Genet.* **56**, 569–578 (2024).

98. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).

99. Kamitaki, N. *et al.* Human and bacterial genetic variation shape oral microbiomes and health. 2025.03.31.25324952 Preprint at https://doi.org/10.1101/2025.03.31.25324952 (2025).

100. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

101. Burgess, S., Davies, N. M. & Thompson, S. G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol.* **40**, 597–608 (2016).

102. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).

103. Rees, J. M. B., Wood, A. M., Dudbridge, F. & Burgess, S. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PloS One* **14**, e0222362 (2019).

# Methods

## UK Biobank, SPARK, and All of Us WGS data

All whole genome sequencing data analyzed in this work was generated in previous studies. For all cohorts, PCR-free methods were used in library preparation and sequencing was done on Illumina NovaSeq 6000 machines. Whole genome sequencing (WGS) from UK Biobank[37] was performed on blood-derived DNA from 490,401 individuals at the deCODE facility in Reykjavik, Iceland and the Wellcome Sanger Institute (Sanger) in Cambridge, UK. Sequencing reads were aligned to human reference build GRCh38 graph genome with Illumina DRAGEN Bio-IT Platform Germline Pipeline v3.7.8. Samples were sequenced to an average coverage of 32.5x.

WGS from the Simons Foundation Powering Autism Research (SPARK) cohort of the Simons Foundation Autism Research Initiative (SFARI)[39] was performed on saliva-derived DNA from 12,519 individuals. Sequencing reads were aligned to human reference build GRCh38 with BWA-MEM by the New York Genome Center (NYGC) using Centers for Common Disease Genomics project standards. Samples were sequenced to an average coverage of 42x. Details of saliva sample collection, DNA extraction, and sequencing were described in ref[81] (which analyzed data from sequencing waves WGS1-3 of the SPARK integrated WGS (iWGS) v1.1 data set; here we analyzed WGS1-5, which included additional samples included in subsequent sequencing waves).

WGS from the NIH All of Us v8 cohort[38] was performed on blood and saliva-derived DNA from 414,817 individuals (365,918 blood-derived and 48,899 saliva-derived samples). Sequencing reads were aligned to human reference build GRCh38 with Illumina DRAGEN Bio-IT Platform Germline Pipeline v3.4.12. Samples were sequenced to an average coverage of 37.9x.

Genotypes of genome-wide human genetic variants (SNPs and indels) were previously generated for all three data sets. For UKB, we analyzed genotypes previously imputed into UKB SNP-array data[68] using the TOPMed reference panel[82] (as the final UKB WGS genotype call set was not available at the time of analysis). For SPARK and AoU, we analyzed genotypes previously called from WGS using DeepVariant (SPARK[39]) and DRAGEN (AoU[38]).

## Selection of viral reference genomes

The panel of 31 DNA viruses included in analyses of viral load built on previous work that identified the most prevalent viruses observed in blood-derived WGS data[13]. Additional viral reference genomes were included to more comprehensively represent viral families previously observed (herpesviruses and polyomaviruses). To maximize representation of the diversity of anellovirus genomes while limiting the size of the viral reference panel, we selected a representative from each of the five recently described anellovirus clades with available NCBI reference genomes[78] (**Supplementary Fig. 1a**). All viral reference genomes (**Supplementary Table 1**) were merged into a single reference for alignment.

**Measurement of viral DNA presence and abundance in WGS samples**

In the UK Biobank (blood-derived samples), All of Us (blood-derived samples), and SPARK (saliva-derived samples) cohorts, we first extracted unmapped reads (i.e., reads that did not align to the GRCh38 human reference genome) and reads that aligned to the chrEBV decoy contig. We realigned these reads to the reference panel of 31 viral genomes using BWA-MEM[83] (v0.7.18). These alignments were then filtered to those having mapping quality >5, both mates of a read pair aligned to the same reference genome, and an insert size in the range 100–1000bp. Each virus' genome was then scanned to identify regions with excessive numbers of alignments suggestive of accumulated misalignments originating from some other source of DNA (e.g., mismapped human DNA) by first computing alignment coverage across an entire cohort (UK Biobank, All of Us, and SPARK each analyzed separately) in each 500bp window of the viral genome using mosdepth[84] (v0.3.9). 500bp regions with alignment coverage that exceeded the following threshold were then excluded from analysis:

$$4 \times (Q3 - Q1) + Q3 + 5$$

where $Q1$ and $Q3$ are the first and third quartiles of the distribution of alignment coverage across all 500bp regions of that virus' genome. This approach is a lenient application of Tukey's fences to allow for inclusion of some regions which are still likely to have a majority of alignments derived from that viral genome. The constant offset avoids excluding many regions that likely contain true alignments in the situation where the first and third quartiles have the same value.

For association analyses of viral load with biological and clinical phenotypes, these measures of viral load were then converted into binary "viral DNA positivity" indicators of presence or absence of reads from a given virus in each individual. For genetic association analyses of viral load in blood, the number of viral read pairs mapping to each genome was inverse normal transformed to capture quantitative information about viral abundance while limiting the influence of outlier samples. In saliva samples, in which viral reads were often much more abundant, two phenotypes were generated for genetic association analyses: viral DNA positivity (binary presence or absence of reads), and a quantitative abundance metric in which we applied inverse normal transform to non-zero values (i.e., masking individuals with no reads from a given virus) after normalizing read counts for library size. This allowed for the possibility of observing effects on viral prevalence but not abundance and vice versa. For associations with sex (**Fig. 2c,d**), the cohort used for analysis was chosen to maximize prevalence and/or representation of an age range with a larger sex difference; for SPARK, which used a family design, analyses of sex effects were restricted to parents.

**Estimation of the number of EBV-derived reads expected to be present in blood WGS**

To assess the reasonableness of the distribution of viral read counts observed in blood WGS (typically 0 or 1 read pair per sample, even for near-ubiquitous herpesviruses; **Fig. 1b**), we roughly estimated the number of EBV-derived reads expected per WGS sample as follows. On

average, EBV is present in 1 out of 100,000 B-cells[6] with roughly 100 episomes per infected cell[75]. Assuming B-cells comprise roughly 5% of all white blood cells[85] yields an expected $5x10^{-5}$ EBV genomes per white blood cell, or $2.5x10^{-5}$ EBV genomes per haploid human genome in blood-derived WGS. The EBV reference genome is 171,823 bp, so WGS at 30x coverage of the human genome by 2x150bp read pairs should produce an expected 0.4 read pairs from the EBV genome per sample.

**Association of viral DNA prevalence with sample collection time**

Collection time for blood samples in UK Biobank was obtained from Field 3166. For analyses of collection time for saliva-derived WGS in the All of Us cohort, we excluded a large fraction of samples (62%) that we determined were likely to have been mailed, such that recorded collection times corresponded to receipt of samples rather than time of saliva sampling. Specifically, samples were excluded if they lacked an in-person physical measurement for heart rate or had a heart rate measurement time separated by more than a day from WGS sample collection time. The recorded collection times for the 18,751 remaining samples were converted from UTC to local time by taking the modal time zone of the state containing the 3-digit zip code for the corresponding individual.

P-values for association between sample collection time and viral DNA prevalence (**Fig. 2e,f** and **Supplementary Fig. 3e,f**) were calculated with ANOVA by comparing models with and without hour of the day or month of the year. Both hour of the day and month of the year were encoded as a series of indicator variables to allow for non-linear relationships with viral prevalence. All models included age, age-squared, sex, assessment center, and the top 20 genetic principal components as covariates.

**Association of viral DNA prevalence with genetic ancestry**

For UK Biobank, genetically-inferred ancestry was determined as described previously in ref.[86]. In brief, 20 genome-wide ancestry PCs were used to identify groups of individuals within a Euclidean distance radius from the center of individuals within each self-reported ethnicity category, with the distance threshold chosen to include a large fraction of individuals in that self-reported ethnicity category.

For All of Us participants, previously generated genetically-inferred ancestry[38] and ancestry admixture estimates from Rye[87] were used.

For SPARK participants, genetic ancestry was inferred using Euclidean distance on 10 genome-wide ancestry PCs to cluster centers determined from the subset of individuals who self-reported race. For each ancestry group, a cluster center was chosen to have the median PC coordinate for each principal component among individuals who self-reported a corresponding race. As in UKB, all individuals that fell within a Euclidean distance radius that enclosed a large majority of individuals self-reporting that race were then assigned to that genetically-inferred ancestry. For European ancestry, the radius was set to include 90% of individuals who self-

reported as “White” (n=8,157). For African ancestry, the radius was set to include 90% of individuals who self-reported as “African American” (n=345). For American ancestry, the radius was set to include 75% of individuals who self-reported as “Hispanic” or “Native American” (n=906). For East Asian ancestry, the radius was set to include 75% of individuals who self-reported as “Asian” and were separated from the majority of samples on PC2 (n=342). For South Asian ancestry, the radius was set to include 75% of individuals who self-reported as “Asian” and were separated from the majority of samples on PC4 (n=144).

**HLA allele imputation in UK Biobank**

The T1DGC reference panel for HLA allele imputation[88] was first converted to VCF format with variants lifted over to hg19 and merged into multiallelic sites where appropriate. A small number of haplotypes (n=136) with >2 alleles for at least one multiallelic site were excluded from the reference panel. Imputation of HLA alleles onto phased array genotypes was done with BEAGLE[89] (v5.4), after which imputed alleles were converted back to biallelic variants for genetic association analysis and lifted over to hg38.

**Genome-wide association analyses of viral load phenotypes in UK Biobank**

Abundances of reads aligning to reference genomes for EBV, HHV-7, and anellovirus strains TUS01, VT416, and HD14a were inverse normal transformed into quantitative phenotypes for genome-wide association analyses. Individuals were excluded based on the following criteria: not having European genetic ancestry, not having available TOPMed-imputed genotypes (including for chromosome X), and/or having withdrawn, leaving 453,770 individuals for genetic association analyses. TOPMed-imputed variants for these individuals were filtered to require minor allele frequency > 0.001 and INFO > 0.3. Linear mixed model association tests were performed with BOLT-LMM[90] (v2.5) to account for relatedness, using the following covariates: age, age squared, sex, genotype array, assessment center, and 20 genetic PCs. GWAS of germline-inherited endogenous HHV-6A and HHV-6B carrier status were performed using linear regression with the same covariates in individuals of European genetic ancestry, excluding one from each pair of relatives with second-degree or closer relatedness[91].

To identify index variants outside the MHC region (**Supplementary Table 2**), we first iteratively selecting the strongest association and removed any variants within 1 Mb. Index variant pairs within 3Mb were then evaluated to determine whether they represented independent associations using the following approximation of the association strength of the index variant *i* conditional on the more strongly associated index variant *j*:

$$\chi^2_{i|j} \approx \chi^2_i (1 - r_{ij} sign(\beta_i \beta_j) \sqrt{\frac{\chi^2_j}{\chi^2_i}})^2$$

as previously described[92]. The less-strongly associated variant *i* was dropped if its approximate conditional association was no longer genome-wide significant. Identified index variants were

annotated with nearby genes using GENCODE 39 (ref.[93]) definitions for protein-coding genes, long non-coding RNAs, and microRNAs. Index variants were annotated as eQTLs and sQTLs using the v10 release of GTEx[94]. Follow-up genetic association analyses of variants in the MHC region including both TOPMed-imputed variants and imputed HLA alleles were performed using linear regression with BOLT-LMM on unrelated European-ancestry individuals with the same covariates (**Supplementary Table 3**).

**Partitioning of heritability between MHC and non-MHC variation**

To partition heritability between common variants within and outside the MHC region of the human genome, BOLT-REML[95] (v2.5) was run on SNP-array genotypes for unrelated European-ancestry UK Biobank participants with variants within the range chr6:24,000,000-chr6:35,000,000 assigned to one component (MHC region) and all other variants assigned to a second component. Age, age-squared, sex, assessment center, genotyping array, and the top 20 genetic ancestry principal components were included as covariates.

**Rare variant association analyses in UK Biobank**

Gene-level burden masks were generated as previously described[96] using genotypes of rare protein-coding variants in the UKB DRAGEN WGS data set and genotypes of CNVs previously ascertained from UKB whole-exome sequencing data[97]. The specific burden masks analyzed here for association with viral load used a minor allele frequency threshold of MAF<0.001 and included missense variants with PrimateAI-3D[98] scores >0.7 merged with loss-of-function SNPs, indels, and CNVs. Association tests were performed using linear regression (implemented in BOLT-LMM) on unrelated European-ancestry individuals with age, age-squared, sex, assessment center, genotyping array, and the 20 genetic principal components included as covariates. Genes in the MHC region (chr6:24,000,000-chr6:35,000,000) were excluded given the potential for linkage disequilibrium with strong common variant associations, leaving 17,589 gene burden masks and a Bonferroni threshold of $5.7 \times 10^{-7}$.

**Genome-wide association analyses of viral load phenotypes in SPARK**

Abundances of reads aligning to reference genomes for HHV-7, HHV-6B, and Merkel cell polyomavirus were transformed into up to two GWAS phenotypes for each virus: a binary phenotype encoding the presence/absence of any viral reads (HHV-6B, n=9,081 after sample exclusions; MCPyV, n=9,209) and a quantitative phenotype comprising inverse normal transformed non-zero values (HHV-6B, n=6,258; HHV-7, n=7,360). Individuals with non-European genetic ancestry were excluded using a more permissive ancestry definition based on genomic PC1 and PC2 to maximize power, leaving 9,209 individuals for genetic association analyses. For HHV-6B, carriers of eHHV-6B were also excluded. Variants called by DeepVariant were filtered and used to generate ancestry principal components as previously described[99]. Linear mixed model association tests were performed using BOLT-LMM to account for relatedness, using the following covariates: age, square root of age, age squared, sex, sequencing batch, percentage of mapped reads, and 10 genetic PCs.

**Genome-wide association analyses of viral load phenotypes in All of Us**

GWAS was performed on European-ancestry All of Us participants for the following phenotypes: inverse normal transformed EBV abundances in blood WGS (n=201,170), EBV DNA positivity (i.e., presence of any EBV reads) in saliva WGS (n=33,164), and inverse normal transformed EBV abundances in EBV DNA-positive saliva WGS samples (n=16,282). Variants from the Allele Count/Allele Frequency (ACAF) threshold callset were filtered to those with minor allele frequency >0.1% and allele count >40 in the subset of European ancestry samples used. BOLT-LMM was run with the following covariates: age, age squared, sex, sequencing site, and 16 genetic PCs.

**Meta-analysis of GWAS results from UK Biobank and All of Us**

Associations with EBV viral load in blood in All of Us were meta-analyzed with those from UK Biobank using METAL[100] (v2020-05-05) in standard error mode, restricting to variants with minor allele frequency greater than 0.1% in both cohorts. One significantly associated locus in UKB that disagreed in direction of effect between cohorts (and between EBV presence phenotypes generated from left and right halves of the EBV genome, see **Supplementary Note 1**) was filtered.

**Measurement of EBV type 1 and type 2-specific alignments**

In both UK Biobank and SPARK, unmapped reads and reads that aligned to the chrEBV decoy contig were realigned to the EBV type 1 (NC_007605.1) and EBV type 2 (NC_009334.1) reference genomes using BWA-MEM (providing both reference genomes simultaneously). Reads with mapping quality >5 were kept, with no requirement that both reads in a pair align to the same genome (to account for the high homology between the EBV type 1 and type 2 genomes). Alignments were counted in 500bp bins (based on coverage per 500bp bin computed by mosdepth) and a 1kb offset was added to alignment bin coordinates in the type 1 genome starting at position 85000 to account for differences between the two reference genomes in the *EBNA3A–EBNA3C* region. Alignments to 500bp bins within the viral genomic regions 35501-38000 and 79001-92500 (corresponding to *EBNA2* and *EBNA3A–EBNA3C*) were considered to be type-specific, and samples containing any alignments in these regions were designated as positive for the EBV type of the reference with the alignment.

**Variation in EBV type 1 versus type 2 frequency by birthplace of UK Biobank participants**

For analyses of individuals born in the United Kingdom, geographic boundaries for nine regions in England, Wales, Scotland, and Northern Ireland were obtained as a GeoJSON file corresponding to “NUTS, level 1 (January 2018) Boundaries UK BFC” (for EBV type analyses) or “NUTS, level 2 (January 2018) Boundaries UK BFC” (for total EBV analyses) from the Open Geography portal from the Office for National Statistics (ONS) (https://geoportal.statistics.gov.uk). Birth coordinates for individuals born in the United Kingdom

(Fields 129 and 130) were assigned to regions with these boundaries using the R package sf (v1.0-20), and the proportion of participants positive for EBV type 2 (among EBV-positive individuals) was estimated as the ratio of the number of samples containing an EBV type 2-specific alignment to the number of samples containing either an EBV type 1-specific or type 2-specific alignment. For country-level analyses, Fields 1647 and 20115 were used.

**Genetic association analyses of MHC variants with EBV type viral load phenotypes**

Variants in the MHC region of the human genome (including imputed HLA alleles) were tested for association with three binary EBV viral load phenotypes derived from EBV type-specific read alignments. The first two phenotypes coded individuals with EBV type 1-specific alignments (respectively, type 2-specific alignments) as cases and all other individuals as controls. The third phenotype, used in case-case association analyses of EBV type 1- versus type 2-specific alignments, coded individuals with type 2-specific alignments but not type 1-specific alignments were as cases, and those with type 1-specific alignments but not type 2-specific alignments as controls (n=15,612). Association tests were performed using linear regression (implemented in BOLT-LMM) on unrelated European-ancestry individuals with age, age squared, sex, genotype array, assessment center, and 20 genetic PCs as covariates.

**Association analyses of viral load with biological and clinical phenotypes in UK Biobank**

Binarized viral DNA positivity phenotypes were tested for association with binary disease phenotypes in ICD-10 categories derived from UK Biobank “first occurrence” data fields (fields under Category 1712, which merged data from electronic health records and self-report) and cancer registry data (field 40006). For each virus, we tested only binary disease phenotypes for which at least 5 cases were expected among viral DNA-positive individuals for that virus (comprising 493-1413 tests for each of eight common viruses tested: EBV, HHV-6B, HHV-7, three TTVs, and eHHV-6A and eHHV-6B). Association tests were performed using logistic regression with Firth correction using the Wald approximation as implemented in the logistf R package (v1.26.0) on European-ancestry participants with age, age-squared, sex, assessment center, and the 20 genetic PCs as covariates.

Viral DNA positivity phenotypes were also tested for association with quantitative blood phenotypes in UK Biobank: blood counts (Category 100081), blood biochemistry (Category 17518), and NMR metabolomics (Category 220). Association tests were performed using linear regression on European-ancestry individuals with the above covariates.

Quantitative smoking phenotypes in UK Biobank were generated from the pack-years and cigarettes per day phenotypes by encoding never-smokers (and, for cigarettes per day, former smokers) as 0 values. These smoking phenotypes were tested for association with viral DNA positivity phenotypes using linear regression on European-ancestry individuals with the above covariates.

**Mendelian randomization to identify causal effects of EBV viral load on disease**

Instrument variables (IVs) for Mendelian randomization analyses were identified as the 44 lead variants at non-MHC loci from the GWAS meta-analysis of EBV viral load in UKB and AoU. The MHC region was excluded from MR analyses given the likelihood of linkage disequilibrium generating pleiotropic effects of MHC variants on many immune-related phenotypes. To minimize the impact of sample overlap between cohorts used in the exposure GWAS (EBV viral load in UKB and AoU blood WGS) and outcome GWAS (disease phenotypes in FinnGen+UKB+MVP), we regenerated GWAS summary statistics for EBV viral load in UKB after restricting to a control-only sub-cohort[101]. Specifically, we reran BOLT-LMM after removing UKB participants with the following ICD-10 phenotypes: G35, C81, C82, C83, C84, C85, C91, M05, M06, M32. GWAS results from this control-only UKB sub-cohort were then meta-analyzed with All of Us associations to generate effect sizes and standard errors for use in Mendelian randomization.

Outcome summary statistics for phenotypes of interest were downloaded from https://mvp-ukbb.finngen.fi/ after applying for access. The MendelianRandomization R package[102] (v0.10.0) was used to generate estimates of causal effect sizes of EBV viral load on disease phenotypes using the weighted median, inverse variance weighed (IVW), Mendelian Randomization-Egger (MR-Egger), and contamination mixture (ConMix) approaches. For weighted median, IVW, and MR-Egger, robust regression with penalized weights was used to account for invalid IVs[103].

## Code availability

The following publicly available software resources were used: bwa (v0.7.18, https://bio-bwa.sourceforge.net/), mosdepth (v0.3.9, https://github.com/brentp/mosdepth), bcftools (v1.14, http://www.htslib.org/), samtools (v1.15.1, http://www.htslib.org/), plink (v1.90b6.26 and v2.00a3.7, https://www.cog-genomics.org/plink/), BEAGLE (v5.4, https://faculty.washington.edu/browning/beagle/beagle.html), BOLT-LMM (v2.5, https://alkesgroup.broadinstitute.org/BOLT-LMM/), qqman R package (v0.1.8, https://cran.r-project.org/web/packages/qqman/index.html), sf R package (v1.0-20, https://cran.r-project.org/web/packages/sf/index.html), logistf R package (v1.26.1, https://cran.r-project.org/web/packages/logistf/index.html), and MendelianRandomization R package (v0.10.0, https://cran.r-project.org/web/packages/MendelianRandomization/index.html).

## Data availability

The following data resources are available by application: UK Biobank (http://www.ukbiobank.ac.uk/), All of Us Research Program (https://allofus.nih.gov/), SFARI SPARK (https://www.sfari.org/resource/spark/), and MVP-Finngen-UKBB meta-analysis summary statistics (https://mvp-ukbb.finngen.fi/). Full virome summary statistics will be uploaded to the GWAS Catalog upon publication and are currently available here: https://data.broadinstitute.org/lohlab/virome_summary_statistics/

## Acknowledgments

We thank C. Usher for edits to text and figures. This research was conducted using the UKB Resource under application number 40709. N.K. was supported by US NIH Fellowship F31 DE034283. S.A.M. was supported by US NIH grant R01 HG006855 and the Howard Hughes Medical Institute. P.-R.L. was supported by US NIH grants R56 HG012698 and R01 HG013110 and a Burroughs Wellcome Fund Career Award at the Scientific Interfaces. The funders had no role in study design, data collection and analysis, the decision to publish or the preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Computational analyses were performed on the O2 High Performance Compute Cluster supported by the Research Computing Group at Harvard Medical School (http://rc.hms.harvard.edu), the UKB Research Analysis Platform, and the *All of Us* Researcher Workbench. The *All of Us* Research Program is supported by the NIH, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA no. AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center:5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; and 1 OT2 OD025276. In addition, the *All of Us* Research Program would not be possible without the partnership of its participants.

## Author contributions

N.K., S.A.M. and P.-R.L. conceived the study design. N.K. did the computational and statistical analyses. D.T. annotated functional coding variants and generated the gene burden masks for rare variant associations in UK Biobank. N.K., S.A.M. and P.-R.L. wrote the manuscript with contributions from all authors.

**Figure 1**

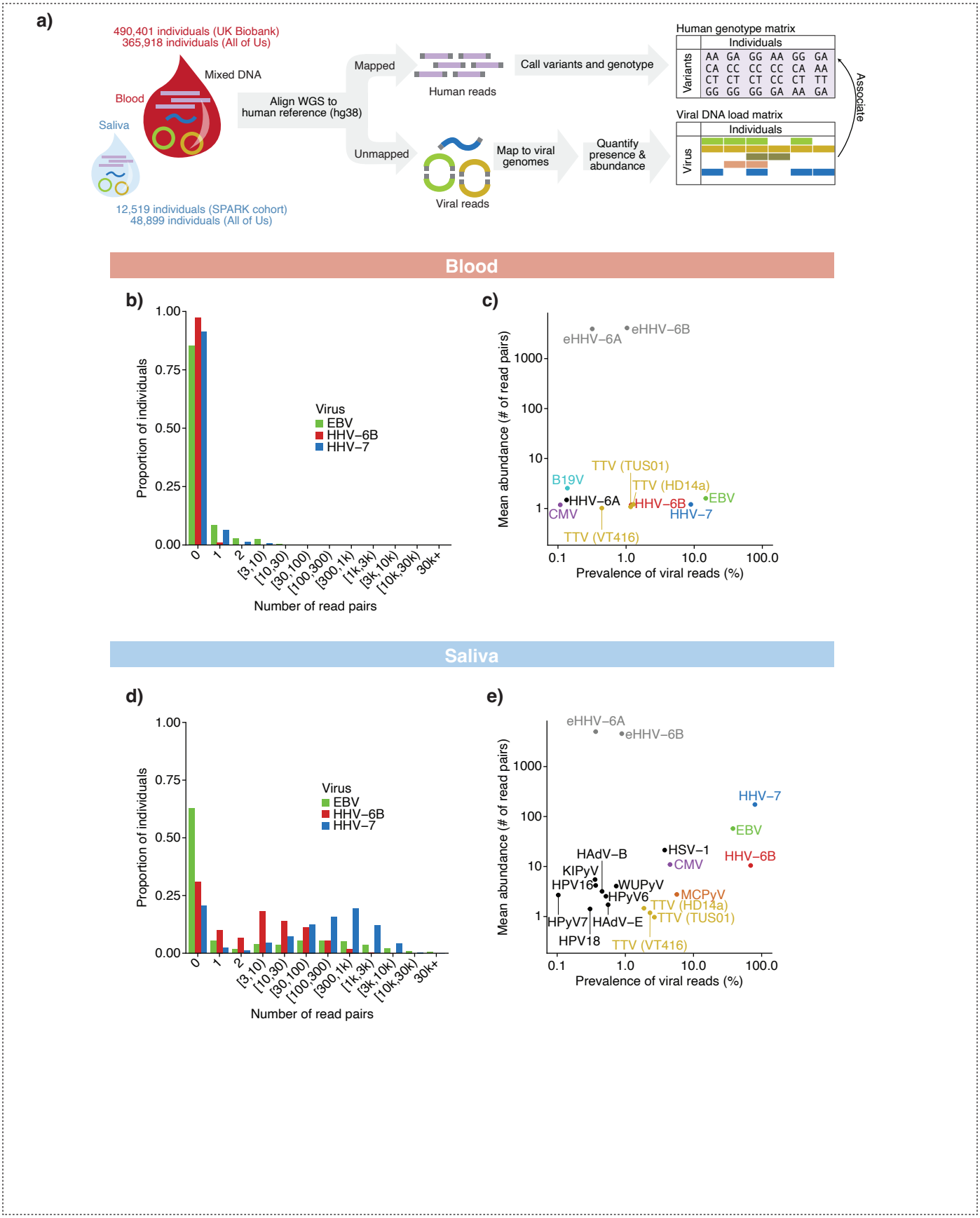

a)
490,401 individuals (UK Biobank)
365,918 individuals (All of Us)
Mixed DNA
Blood
Saliva
12,519 individuals (SPARK cohort)
48,899 individuals (All of Us)
Align WGS to human reference (hg38)
Mapped
Unmapped
Human reads
Viral reads
Call variants and genotype
Map to viral genomes
Quantify presence & abundance
Human genotype matrix
Individuals
Variants
Viral DNA load matrix
Virus
Associate
Blood
b)
Proportion of individuals
Number of read pairs
Virus
EBV
HHV−6B
HHV−7
c)
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
eHHV−6A
eHHV−6B
TTV (TUS01)
B19V
TTV (HD14a)
HHV−6A
HHV−6B
EBV
CMV
HHV−7
TTV (VT416)
Saliva
d)
Proportion of individuals
Number of read pairs
Virus
EBV
HHV−6B
HHV−7
e)
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
eHHV−6A
eHHV−6B
HHV−7
EBV
HSV−1
HAdV−B
CMV
HHV−6B
KIPyV
HPV16
WUPyV
HPyV6
MCPyV
TTV (HD14a)
TTV (TUS01)
HPyV7
HAdV−E
HPV18
TTV (VT416)

**Figure 2**

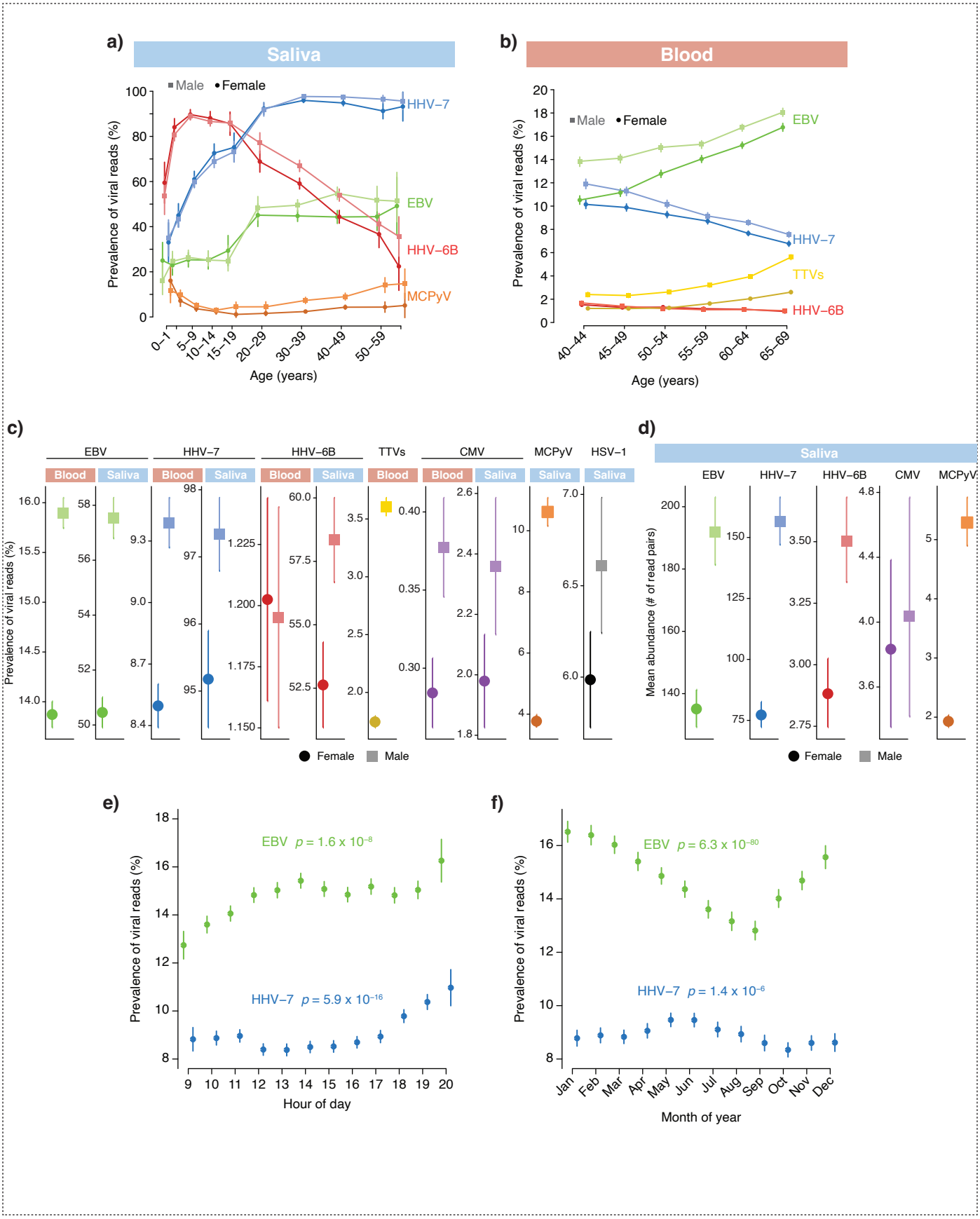

a)
Saliva
Male
Female
Prevalence of viral reads (%)
HHV−7
EBV
HHV−6B
MCPyV
0–1
5–9
10–14
15–19
20–29
30–39
40–49
50–59
Age (years)
b)
Blood
Male
Female
Prevalence of viral reads (%)
EBV
HHV−7
TTVs
HHV−6B
40–44
45–49
50–54
55–59
60–64
65–69
Age (years)
c)
EBV
HHV−7
HHV−6B
TTVs
CMV
MCPyV
HSV−1
Blood
Saliva
Prevalence of viral reads (%)
Female
Male
d)
Saliva
EBV
HHV−7
HHV−6B
CMV
MCPyV
Mean abundance (# of read pairs)
Female
Male
e)
EBV p = 1.6 x 10−8
HHV−7 p = 5.9 x 10−16
Prevalence of viral reads (%)
Hour of day
f)
EBV p = 6.3 x 10−80
HHV−7 p = 1.4 x 10−6
Prevalence of viral reads (%)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month of year

**Figure 3**

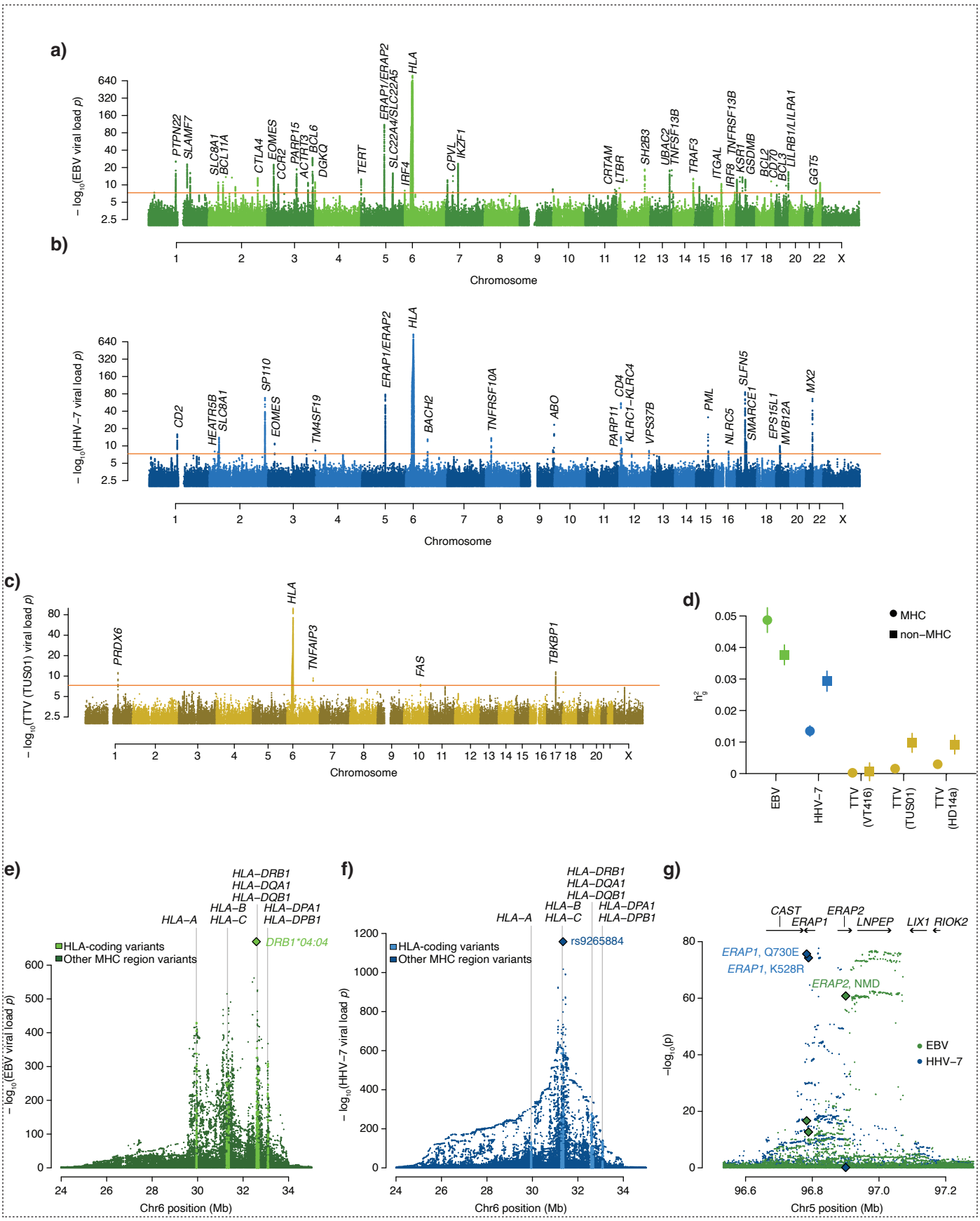

a)
– log10(EBV viral load p)
PTPN22
SLAMF7
SLC8A1
BCL11A
CTLA4
EOMES
CCR2
PARP15
ACTR1A
BCL6
DGKQ
TERT
ERAP1/ERAP2
SLC22A4/SLC22A5
IRF4
HLA
CPVL
IKZF1
CRTAM
LTBR
SH2B3
UBAC2
TNFSF13B
TRAF3
ITGAL
IRF8
TNFRSF13B
KSR1
GSDMB
BCL2
CD70
BCL3
LILRB1/LILRA1
GGT5
Chromosome
b)
– log10(HHV–7 viral load p)
CD2
HEATR5B
SLC8A1
SP110
EOMES
TM4SF19
ERAP1/ERAP2
HLA
BACH2
TNFRSF10A
ABO
PARP11
CD4
KLRC1–KLRC4
VPS37B
PML
NLRC5
SLFN5
SMARCE1
EPS15L1
MVB12A
MX2
Chromosome
c)
– log10(TTV (TUS01) viral load p)
PRDX6
HLA
TNFAIP3
FAS
TBKBP1
Chromosome
d)
MHC
non–MHC
h2g
EBV
HHV–7
TTV (VT416)
TTV (TUS01)
TTV (HD14a)
e)
HLA–A
HLA–B
HLA–C
HLA–DRB1
HLA–DQA1
HLA–DQB1
HLA–DPA1
HLA–DPB1
DRB1*04:04
HLA-coding variants
Other MHC region variants
– log10(EBV viral load p)
Chr6 position (Mb)
f)
HLA–A
HLA–B
HLA–C
HLA–DRB1
HLA–DQA1
HLA–DQB1
HLA–DPA1
HLA–DPB1
rs9265884
HLA-coding variants
Other MHC region variants
– log10(HHV–7 viral load p)
Chr6 position (Mb)
g)
CAST
ERAP1
ERAP2
LNPEP
LIX1
RIOK2
ERAP1, Q730E
ERAP1, K528R
ERAP2, NMD
EBV
HHV–7
–log10(p)
Chr5 position (Mb)

**Figure 4**

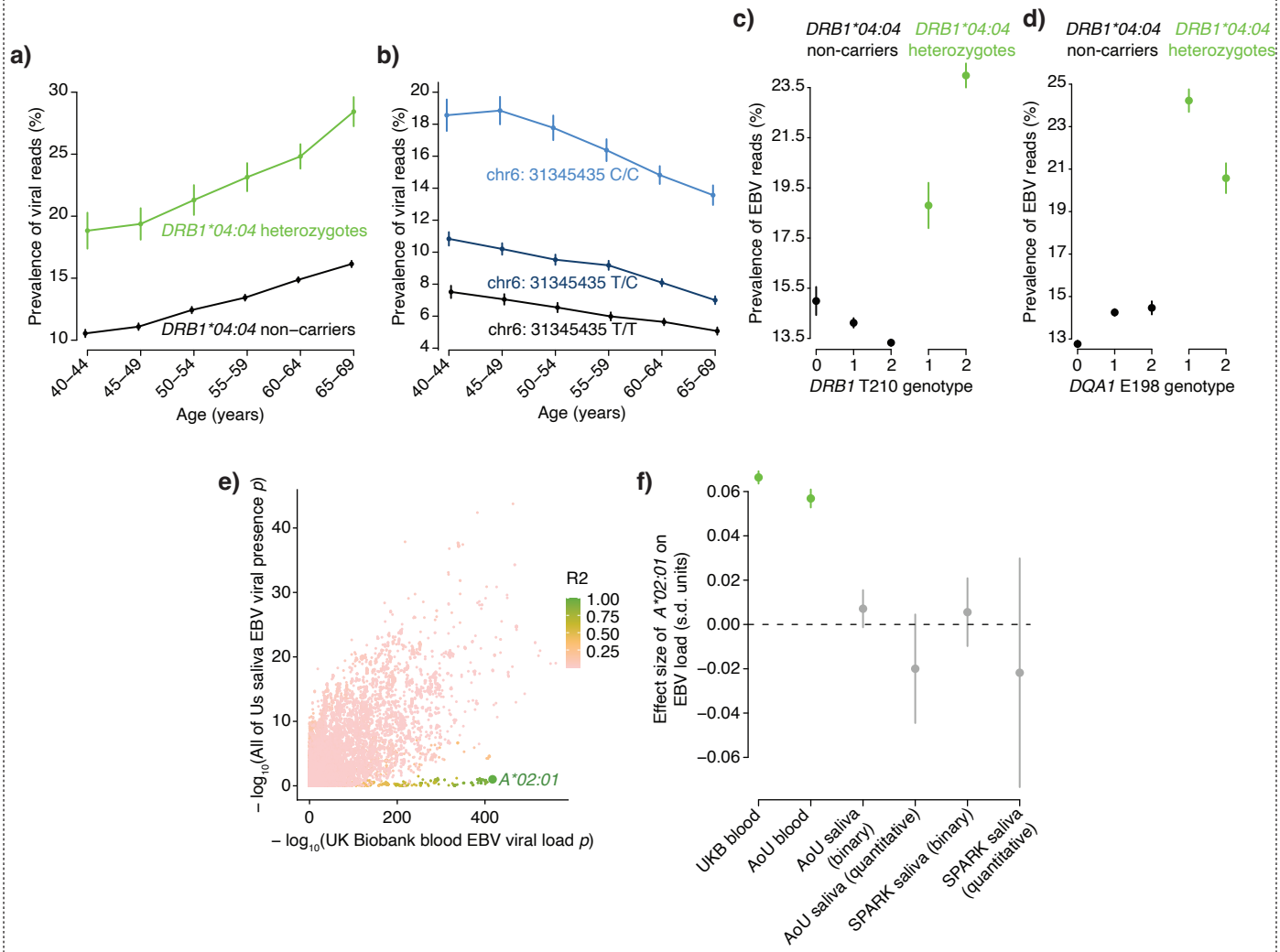
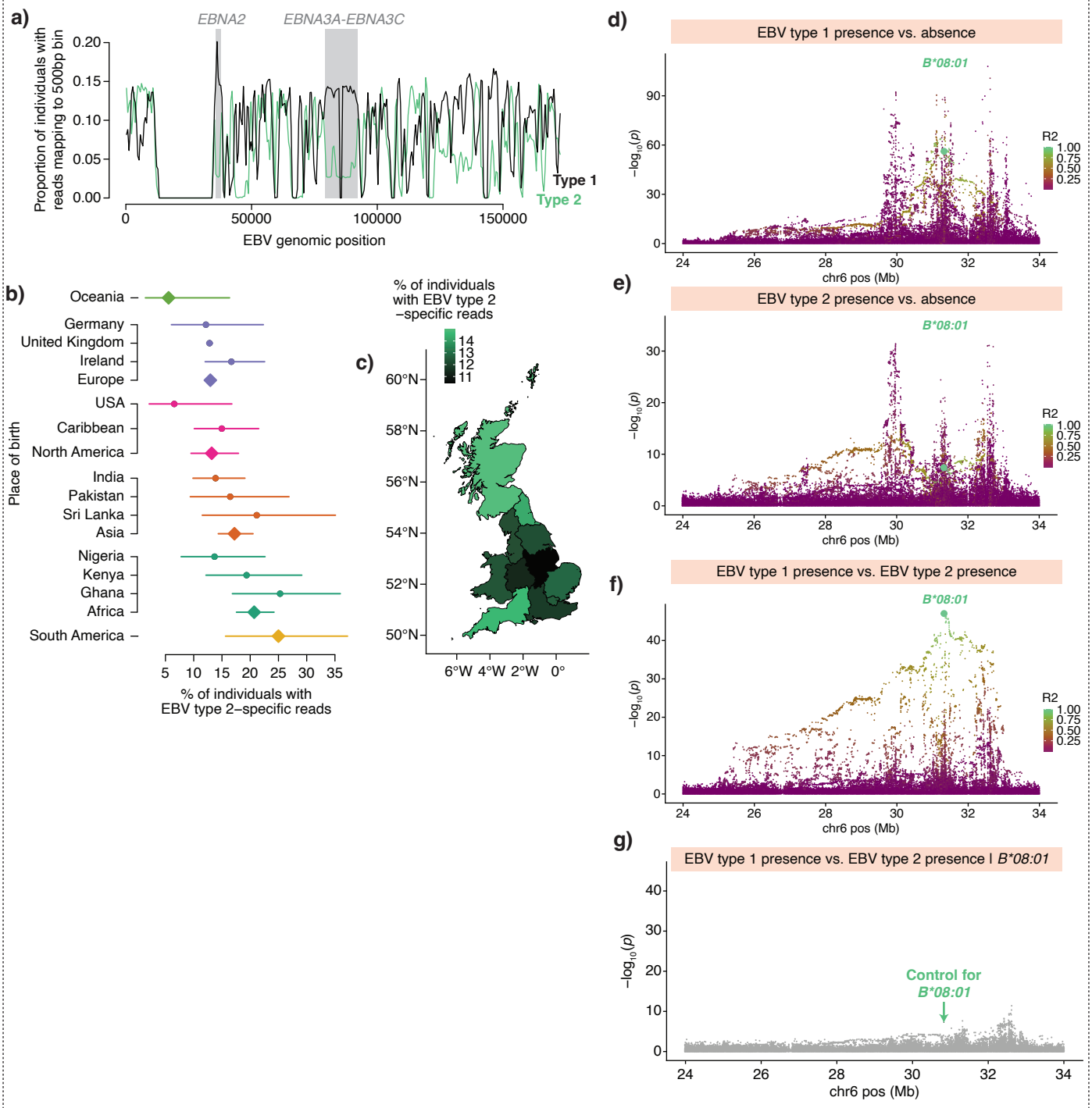
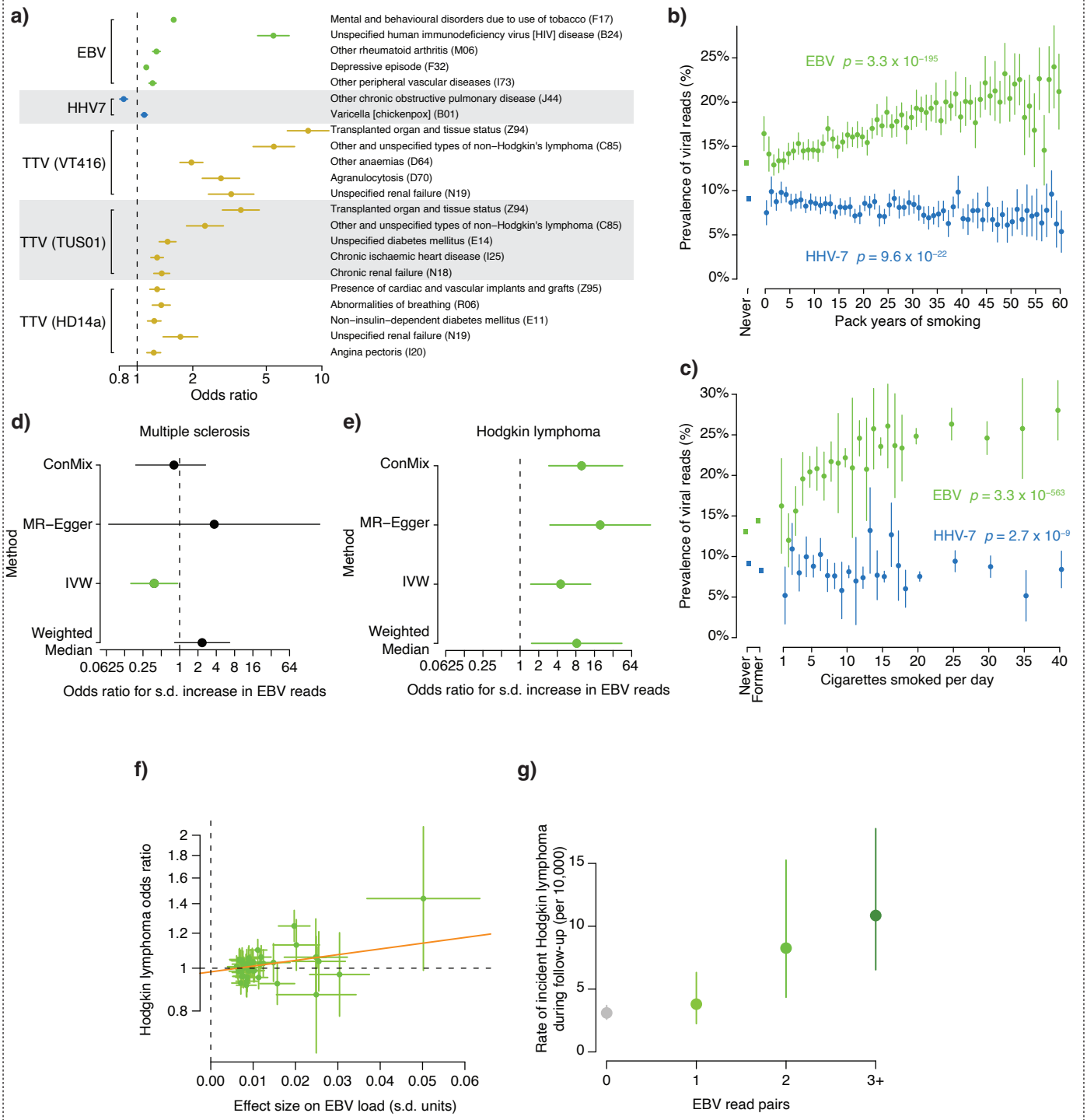

a)
Prevalence of viral reads (%)
DRB1*04:04 heterozygotes
DRB1*04:04 non–carriers
Age (years)
40–44
45–49
50–54
55–59
60–64
65–69
b)
Prevalence of viral reads (%)
chr6: 31345435 C/C
chr6: 31345435 T/C
chr6: 31345435 T/T
Age (years)
c)
DRB1*04:04 non-carriers
DRB1*04:04 heterozygotes
Prevalence of EBV reads (%)
DRB1 T210 genotype
d)
DRB1*04:04 non-carriers
DRB1*04:04 heterozygotes
Prevalence of EBV reads (%)
DQA1 E198 genotype
e)
– log10(All of Us saliva EBV viral presence p)
– log10(UK Biobank blood EBV viral load p)
R2
1.00
0.75
0.50
0.25
A*02:01
f)
Effect size of A*02:01 on EBV load (s.d. units)
UKB blood
AoU blood
AoU saliva (binary)
AoU saliva (quantitative)
SPARK saliva (binary)
SPARK saliva (quantitative)

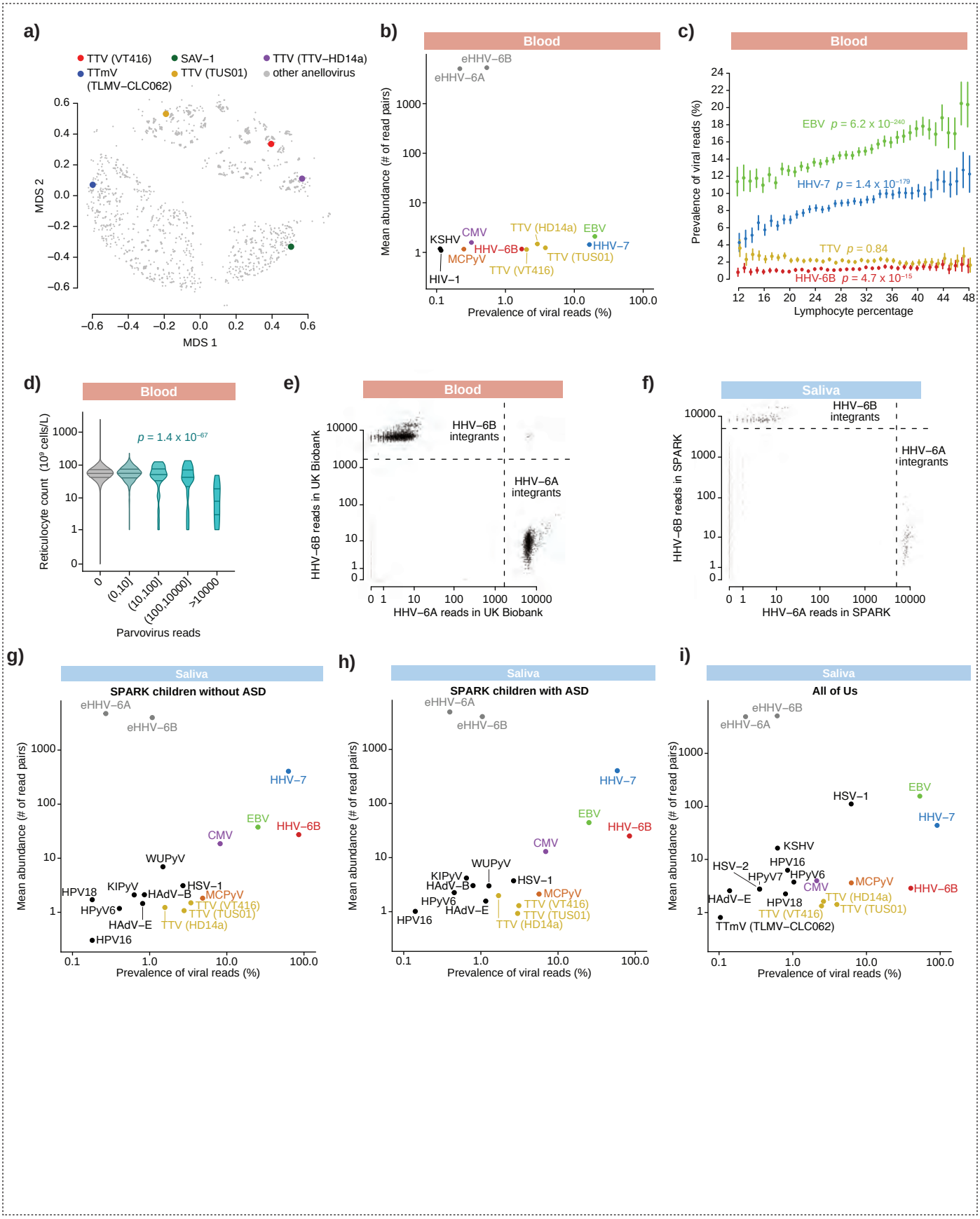
**Figure 5**


a)
EBNA2
EBNA3A-EBNA3C
Proportion of individuals with reads mapping to 500bp bin
0.20
0.15
0.10
0.05
0.00
0
50000
100000
150000
EBV genomic position
Type 1
Type 2
b)
Oceania
Germany
United Kingdom
Ireland
Europe
USA
Caribbean
North America
India
Pakistan
Sri Lanka
Asia
Nigeria
Kenya
Ghana
Africa
South America
Place of birth
5 10 15 20 25 30 35
% of individuals with EBV type 2–specific reads
c)
% of individuals with EBV type 2 –specific reads
14
13
12
11
60°N
58°N
56°N
54°N
52°N
50°N
6°W 4°W 2°W 0°
d)
EBV type 1 presence vs. absence
B*08:01
R2
1.00
0.75
0.50
0.25
−log10(p)
chr6 pos (Mb)
e)
EBV type 2 presence vs. absence
B*08:01
f)
EBV type 1 presence vs. EBV type 2 presence
B*08:01
g)
EBV type 1 presence vs. EBV type 2 presence I B*08:01
Control for B*08:01

## Figure 6


a)
EBV
HHV7
TTV (VT416)
TTV (TUS01)
TTV (HD14a)
Mental and behavioural disorders due to use of tobacco (F17)
Unspecified human immunodeficiency virus [HIV] disease (B24)
Other rheumatoid arthritis (M06)
Depressive episode (F32)
Other peripheral vascular diseases (I73)
Other chronic obstructive pulmonary disease (J44)
Varicella [chickenpox] (B01)
Transplanted organ and tissue status (Z94)
Other and unspecified types of non−Hodgkin's lymphoma (C85)
Other anaemias (D64)
Agranulocytosis (D70)
Unspecified renal failure (N19)
Transplanted organ and tissue status (Z94)
Other and unspecified types of non−Hodgkin's lymphoma (C85)
Unspecified diabetes mellitus (E14)
Chronic ischaemic heart disease (I25)
Chronic renal failure (N18)
Presence of cardiac and vascular implants and grafts (Z95)
Abnormalities of breathing (R06)
Non−insulin−dependent diabetes mellitus (E11)
Unspecified renal failure (N19)
Angina pectoris (I20)
0.8 1 2 5 10
Odds ratio
b)
EBV p = 3.3 x 10−195
HHV-7 p = 9.6 x 10−22
Prevalence of viral reads (%)
25% 20% 15% 10% 5% 0%
Never 0 5 10 15 20 25 30 35 40 45 50 55 60
Pack years of smoking
c)
EBV p = 3.3 x 10−563
HHV-7 p = 2.7 x 10−9
Prevalence of viral reads (%)
30% 25% 20% 15% 10% 5% 0%
Never Former 1 5 10 15 20 25 30 35 40
Cigarettes smoked per day
d)
Multiple sclerosis
Method
ConMix
MR−Egger
IVW
Weighted Median
0.0625 0.25 1 2 4 8 16 64
Odds ratio for s.d. increase in EBV reads
e)
Hodgkin lymphoma
Method
ConMix
MR−Egger
IVW
Weighted Median
0.0625 0.25 1 2 4 8 16 64
Odds ratio for s.d. increase in EBV reads
f)
Hodgkin lymphoma odds ratio
2 1.8 1.6 1.4 1.2 1 0.8
0.00 0.01 0.02 0.03 0.04 0.05 0.06
Effect size on EBV load (s.d. units)
g)
Rate of incident Hodgkin lymphoma during follow-up (per 10,000)
15 10 5 0
0 1 2 3+
EBV read pairs

## Figure S1


a)
TTV (VT416)
SAV−1
TTV (TTV−HD14a)
TTmV (TLMV−CLC062)
TTV (TUS01)
other anellovirus
MDS 1
MDS 2
b)
Blood
eHHV−6B
eHHV−6A
KSHV
CMV
HHV−6B
MCPyV
TTV (HD14a)
EBV
HHV−7
TTV (TUS01)
TTV (VT416)
HIV−1
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
c)
Blood
EBV p = 6.2 x 10−240
HHV-7 p = 1.4 x 10−179
TTV p = 0.84
HHV-6B p = 4.7 x 10−15
Prevalence of viral reads (%)
Lymphocyte percentage
d)
Blood
p = 1.4 x 10−67
Reticulocyte count (10⁹ cells/L)
Parvovirus reads
e)
Blood
HHV−6B integrants
HHV−6A integrants
HHV−6B reads in UK Biobank
HHV−6A reads in UK Biobank
f)
Saliva
HHV−6B integrants
HHV−6A integrants
HHV−6B reads in SPARK
HHV−6A reads in SPARK
g)
Saliva
SPARK children without ASD
eHHV−6A
eHHV−6B
HHV−7
EBV
HHV−6B
CMV
WUPyV
HPV18
KIPyV
HAdV−B
HSV−1
MCPyV
HPyV6
TTV (VT416)
TTV (TUS01)
HAdV−E
TTV (HD14a)
HPV16
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
h)
Saliva
SPARK children with ASD
eHHV−6A
eHHV−6B
HHV−7
EBV
HHV−6B
CMV
WUPyV
KIPyV
HSV−1
HAdV−B
MCPyV
HPyV6
TTV (VT416)
HAdV−E
TTV (TUS01)
HPV16
TTV (HD14a)
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
i)
Saliva
All of Us
eHHV−6B
eHHV−6A
HSV−1
EBV
HHV−7
KSHV
HSV−2
HPV16
HPyV7
HPyV6
CMV
MCPyV
HHV−6B
HAdV−E
HPV18
TTV (HD14a)
TTV (TUS01)
TTV (VT416)
TTmV (TLMV−CLC062)
Mean abundance (# of read pairs)
Prevalence of viral reads (%)

## Figure S2

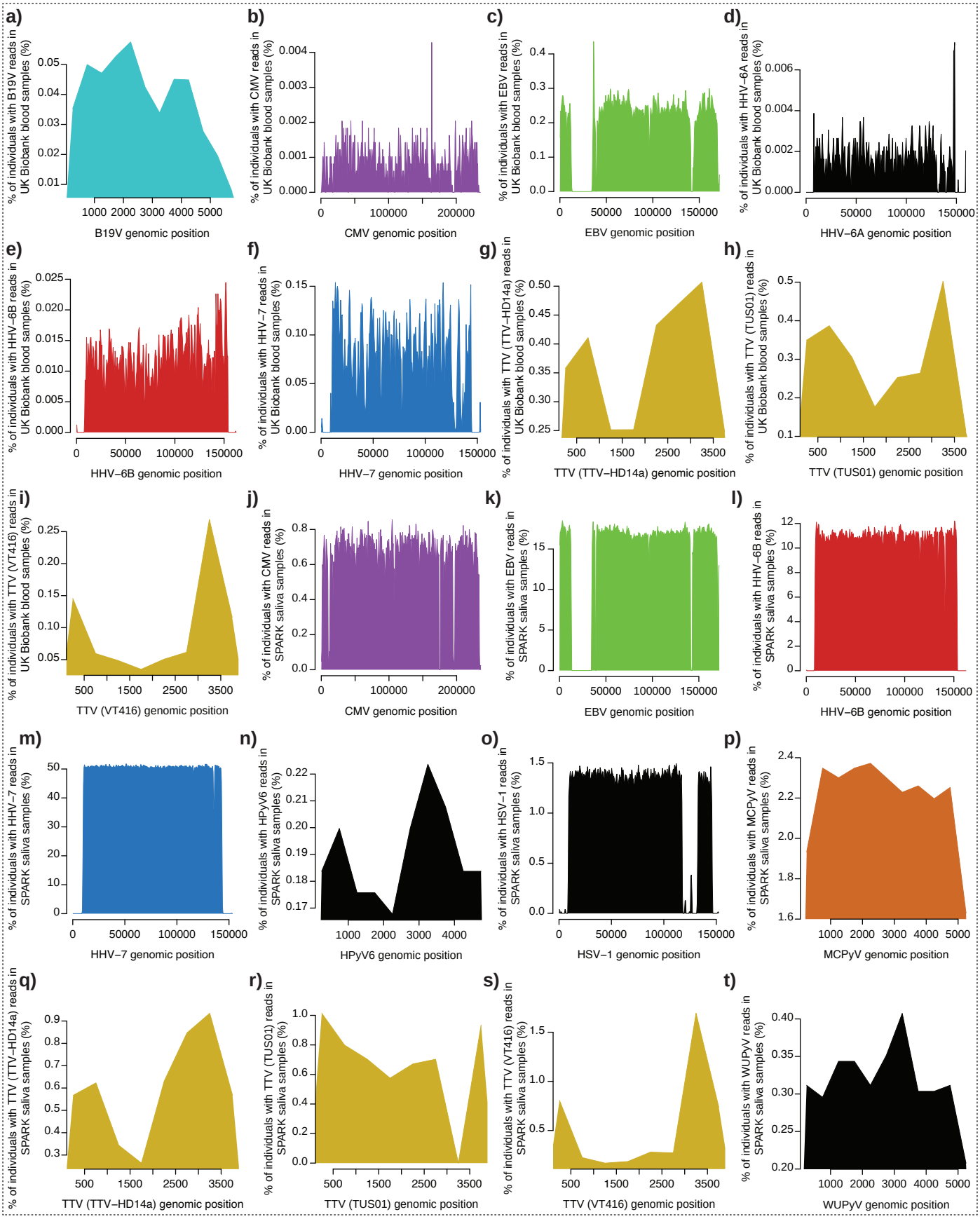
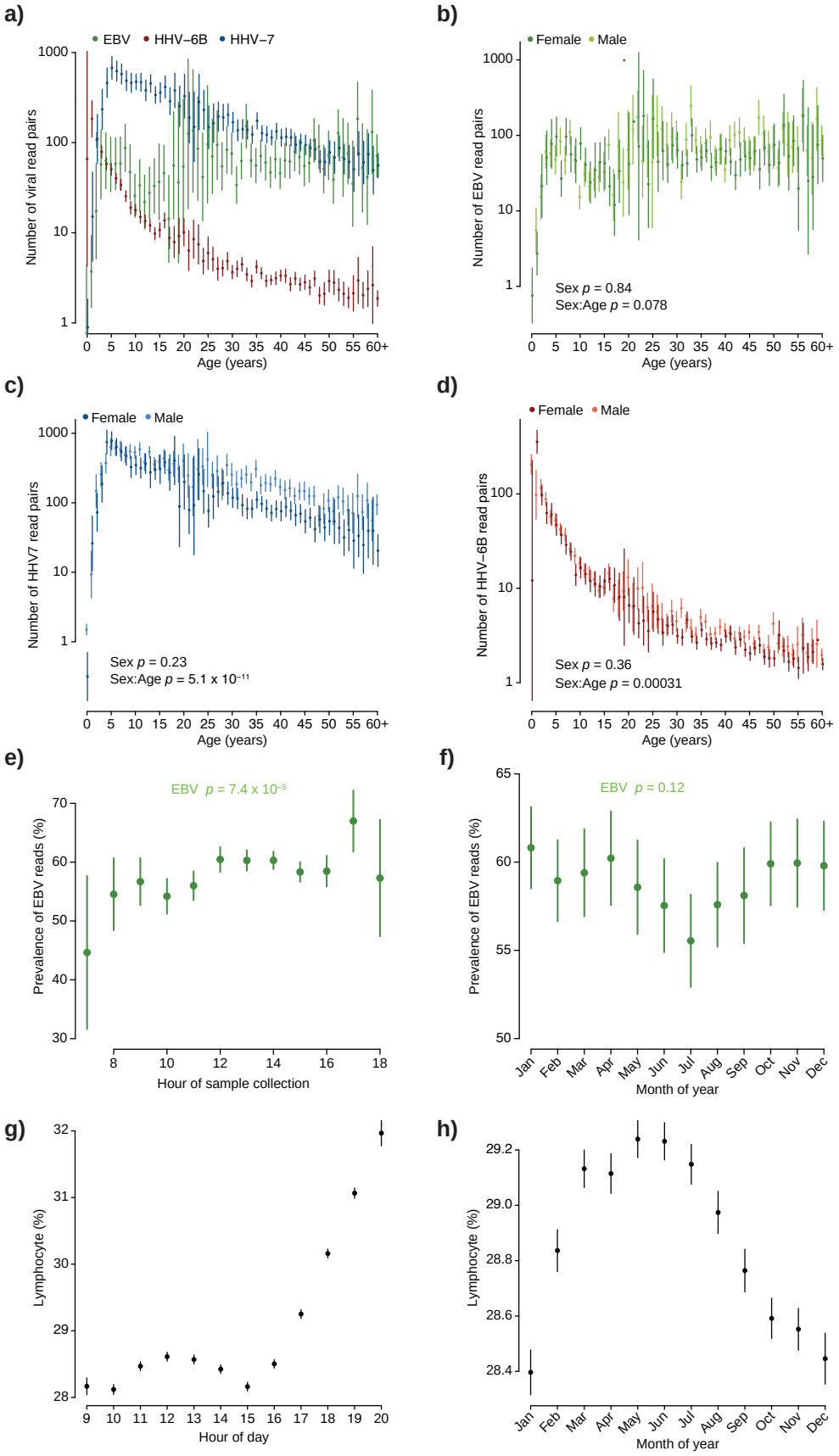

a)
% of individuals with B19V reads in UK Biobank blood samples (%)
B19V genomic position
b)
% of individuals with CMV reads in UK Biobank blood samples (%)
CMV genomic position
c)
% of individuals with EBV reads in UK Biobank blood samples (%)
EBV genomic position
d)
% of individuals with HHV−6A reads in UK Biobank blood samples (%)
HHV−6A genomic position
e)
% of individuals with HHV−6B reads in UK Biobank blood samples (%)
HHV−6B genomic position
f)
% of individuals with HHV−7 reads in UK Biobank blood samples (%)
HHV−7 genomic position
g)
% of individuals with TTV (TTV−HD14a) reads in UK Biobank blood samples (%)
TTV (TTV−HD14a) genomic position
h)
% of individuals with TTV (TUS01) reads in UK Biobank blood samples (%)
TTV (TUS01) genomic position
i)
% of individuals with TTV (VT416) reads in UK Biobank blood samples (%)
TTV (VT416) genomic position
j)
% of individuals with CMV reads in SPARK saliva samples (%)
CMV genomic position
k)
% of individuals with EBV reads in SPARK saliva samples (%)
EBV genomic position
l)
% of individuals with HHV−6B reads in SPARK saliva samples (%)
HHV−6B genomic position
m)
% of individuals with HHV−7 reads in SPARK saliva samples (%)
HHV−7 genomic position
n)
% of individuals with HPyV6 reads in SPARK saliva samples (%)
HPyV6 genomic position
o)
% of individuals with HSV−1 reads in SPARK saliva samples (%)
HSV−1 genomic position
p)
% of individuals with MCPyV reads in SPARK saliva samples (%)
MCPyV genomic position
q)
% of individuals with TTV (TTV−HD14a) reads in SPARK saliva samples (%)
TTV (TTV−HD14a) genomic position
r)
% of individuals with TTV (TUS01) reads in SPARK saliva samples (%)
TTV (TUS01) genomic position
s)
% of individuals with TTV (VT416) reads in SPARK saliva samples (%)
TTV (VT416) genomic position
t)
% of individuals with WUPyV reads in SPARK saliva samples (%)
WUPyV genomic position

**Figure S3**


a)
EBV
HHV−6B
HHV−7
Number of viral read pairs
Age (years)
b)
Female
Male
Number of EBV read pairs
Sex p = 0.84
Sex:Age p = 0.078
Age (years)
c)
Female
Male
Number of HHV7 read pairs
Sex p = 0.23
Sex:Age p = 5.1 x 10⁻¹¹
Age (years)
d)
Female
Male
Number of HHV−6B read pairs
Sex p = 0.36
Sex:Age p = 0.00031
Age (years)
e)
EBV p = 7.4 x 10⁻⁵
Prevalence of EBV reads (%)
Hour of sample collection
f)
EBV p = 0.12
Prevalence of EBV reads (%)
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month of year
g)
Lymphocyte (%)
Hour of day
h)
Lymphocyte (%)
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month of year

## Figure S4

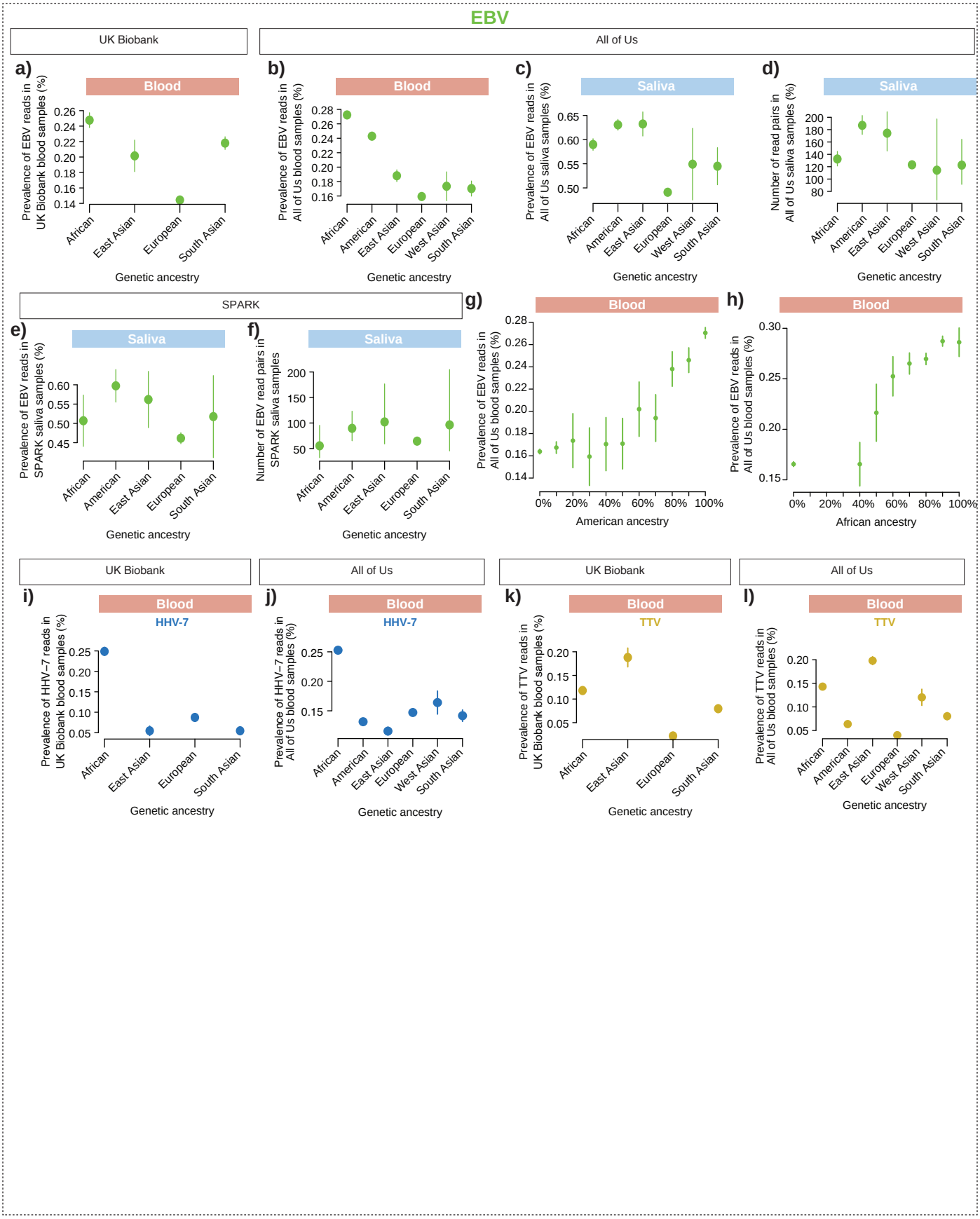

EBV
UK Biobank
All of Us
a)
Blood
Prevalence of EBV reads in UK Biobank blood samples (%)
0.26
0.24
0.22
0.20
0.18
0.16
0.14
African
East Asian
European
South Asian
Genetic ancestry
b)
Blood
Prevalence of EBV reads in All of Us blood samples (%)
0.28
0.26
0.24
0.22
0.20
0.18
0.16
African
American
East Asian
European
West Asian
South Asian
Genetic ancestry
c)
Saliva
Prevalence of EBV reads in All of Us saliva samples (%)
0.65
0.60
0.55
0.50
African
American
East Asian
European
West Asian
South Asian
Genetic ancestry
d)
Saliva
Number of read pairs in All of Us saliva samples (%)
200
180
160
140
120
100
80
African
American
East Asian
European
West Asian
South Asian
Genetic ancestry
SPARK
e)
Saliva
Prevalence of EBV reads in SPARK saliva samples (%)
0.60
0.55
0.50
0.45
African
American
East Asian
European
South Asian
Genetic ancestry
f)
Saliva
Number of EBV read pairs in SPARK saliva samples
200
150
100
50
African
American
East Asian
European
South Asian
Genetic ancestry
g)
Blood
Prevalence of EBV reads in All of Us blood samples (%)
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0%
20%
40%
60%
80%
100%
American ancestry
h)
Blood
Prevalence of EBV reads in All of Us blood samples (%)
0.30
0.25
0.20
0.15
0%
20%
40%
60%
80%
100%
African ancestry
UK Biobank
All of Us
UK Biobank
All of Us
i)
Blood
HHV-7
Prevalence of HHV−7 reads in UK Biobank blood samples (%)
0.25
0.20
0.15
0.10
0.05
African
East Asian
European
South Asian
Genetic ancestry
j)
Blood
HHV-7
Prevalence of HHV−7 reads in All of Us blood samples (%)
0.25
0.20
0.15
African
American
East Asian
European
West Asian
South Asian
Genetic ancestry
k)
Blood
TTV
Prevalence of TTV reads in UK Biobank blood samples (%)
0.20
0.15
0.10
0.05
African
East Asian
European
South Asian
Genetic ancestry
l)
Blood
TTV
Prevalence of TTV reads in All of Us blood samples (%)
0.20
0.15
0.10
0.05
African
American
East Asian
European
West Asian
South Asian
Genetic ancestry

**Figure S5**

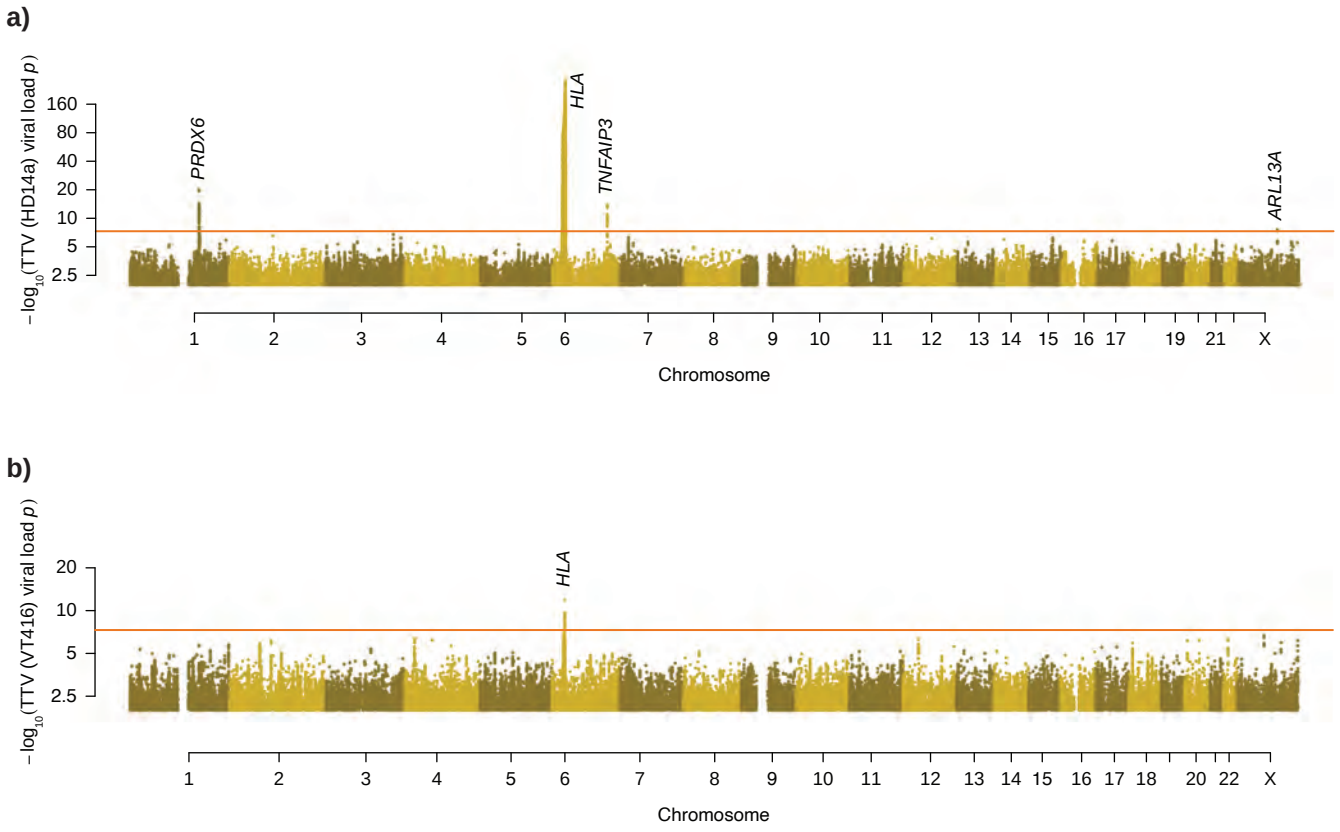

a)
−log10(TTV (HD14a) viral load p)
160
80
40
20
10
5
2.5
PRDX6
HLA
TNFAIP3
ARL13A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 19 21 X
Chromosome
b)
−log10(TTV (VT416) viral load p)
20
10
5
2.5
HLA
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 22 X
Chromosome

**Figure S6**

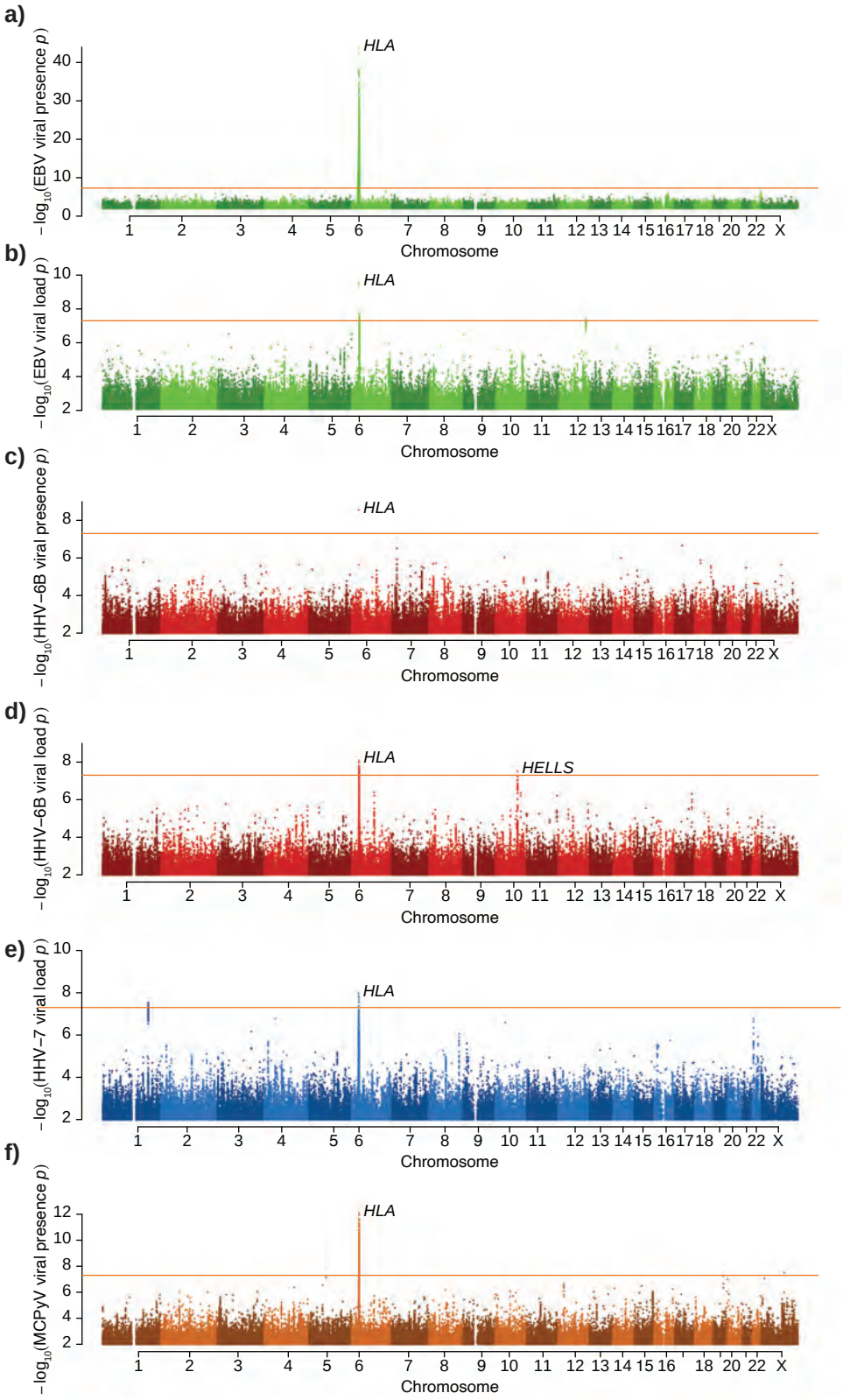
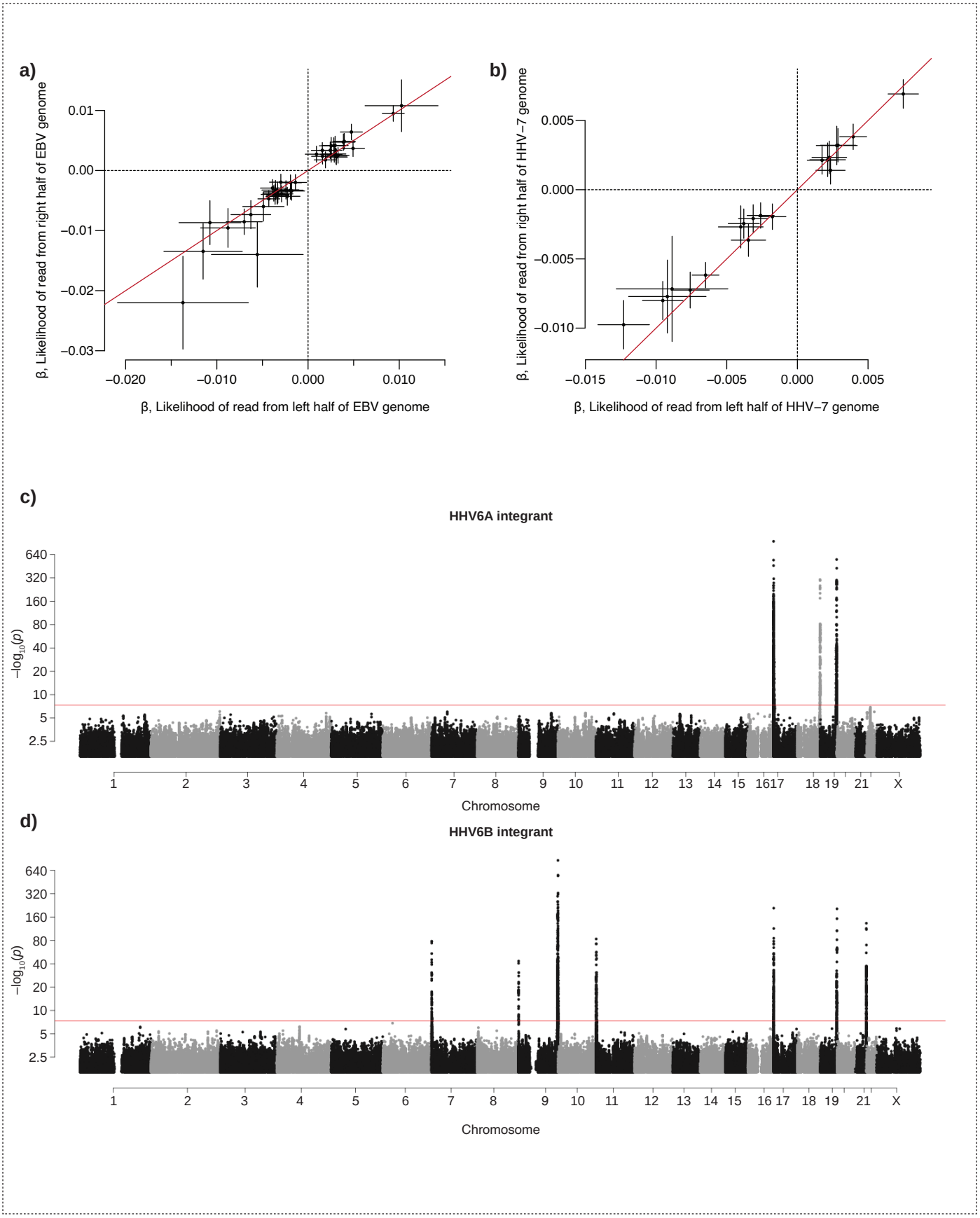
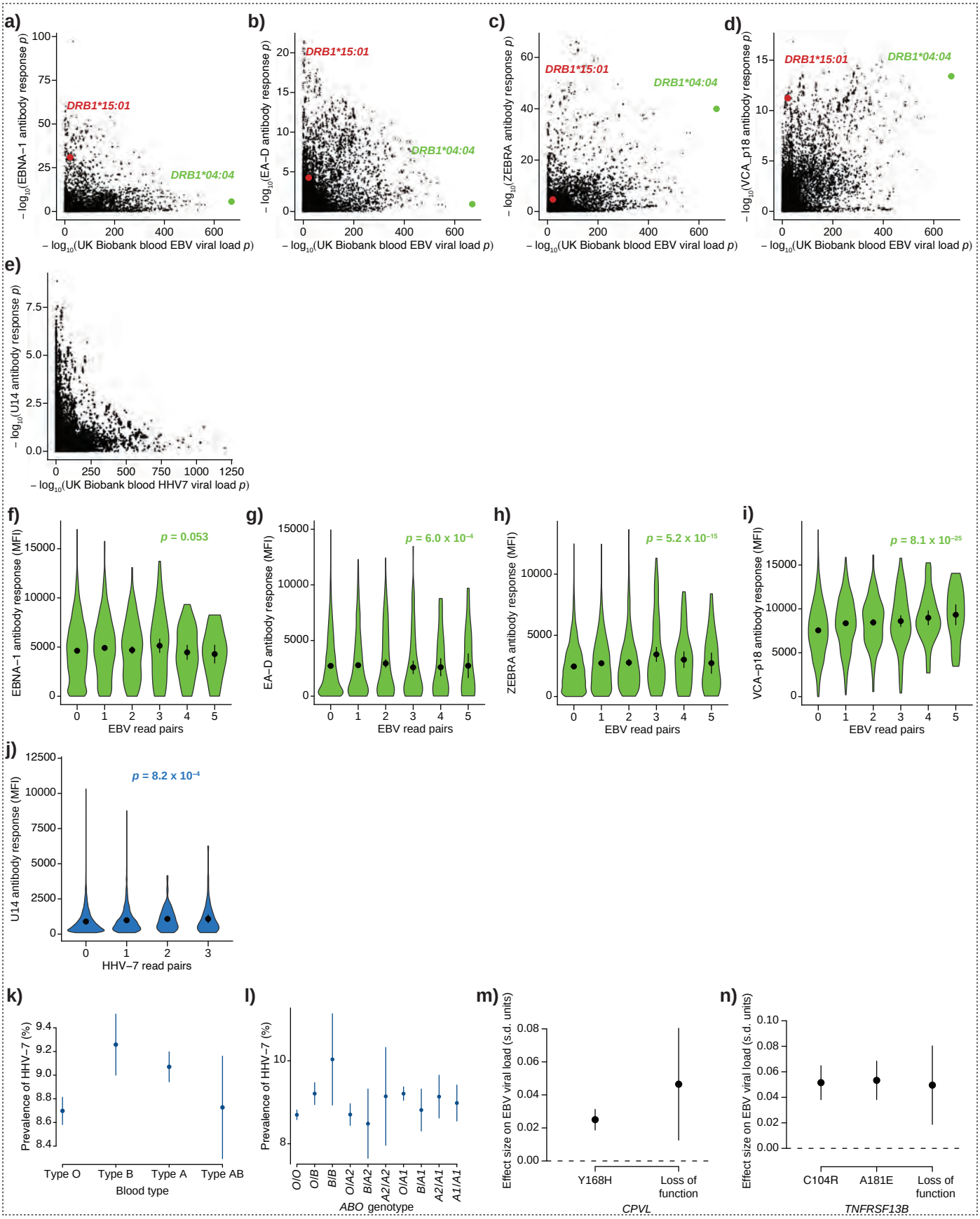

a)
HLA
−log10(EBV viral presence p)
Chromosome
b)
HLA
−log10(EBV viral load p)
Chromosome
c)
HLA
−log10(HHV−6B viral presence p)
Chromosome
d)
HLA
HELLS
−log10(HHV−6B viral load p)
Chromosome
e)
HLA
−log10(HHV−7 viral load p)
Chromosome
f)
HLA
−log10(MCPyV viral presence p)
Chromosome

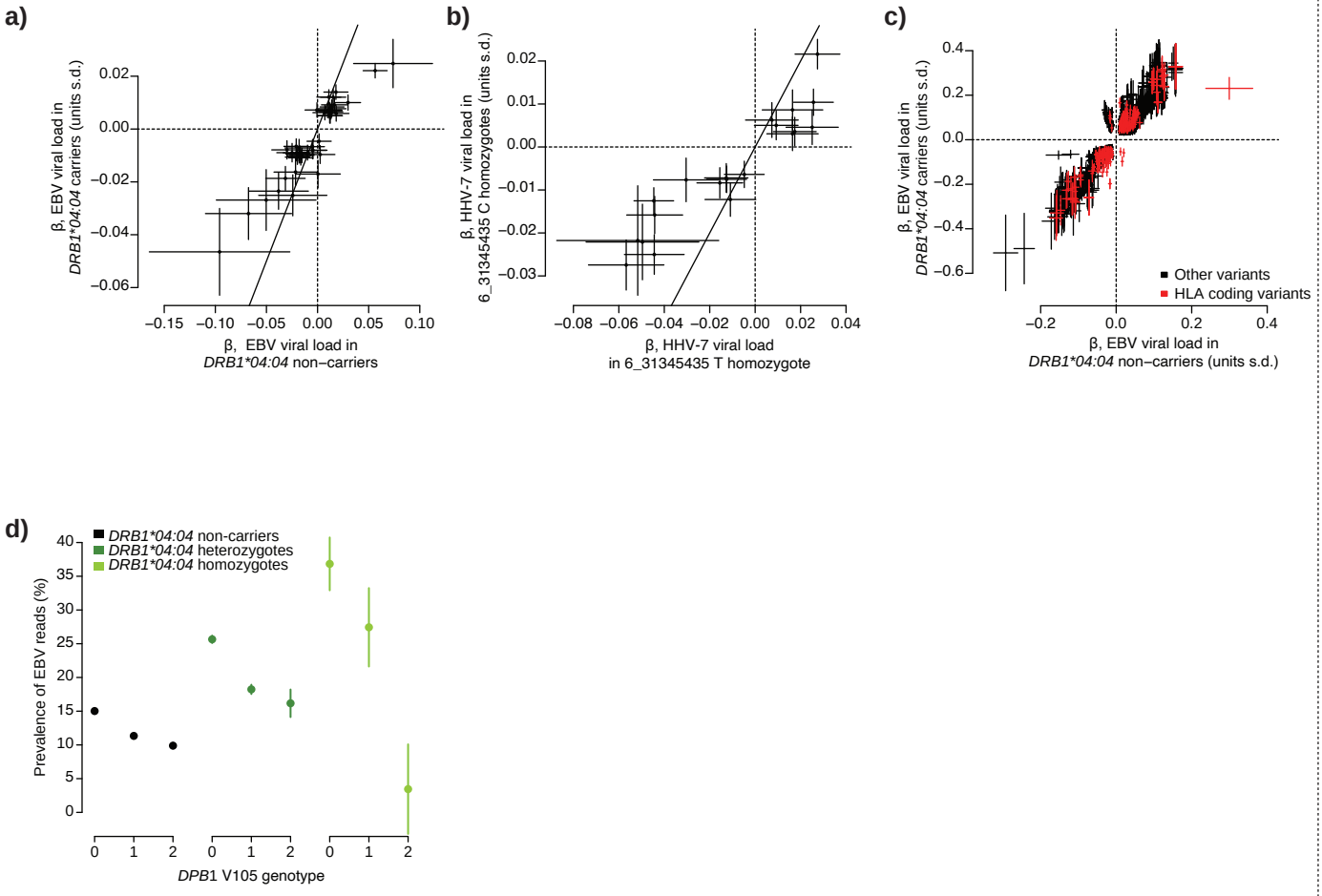
**Figure S7**


a)
β, Likelihood of read from right half of EBV genome
0.01
0.00
−0.01
−0.02
−0.03
−0.020
−0.010
0.000
0.010
β, Likelihood of read from left half of EBV genome
b)
β, Likelihood of read from right half of HHV−7 genome
0.005
0.000
−0.005
−0.010
−0.015
−0.010
−0.005
0.000
0.005
β, Likelihood of read from left half of HHV−7 genome
c)
HHV6A integrant
640
320
160
80
40
20
10
5
2.5
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 X
Chromosome
d)
HHV6B integrant
640
320
160
80
40
20
10
5
2.5
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 X
Chromosome

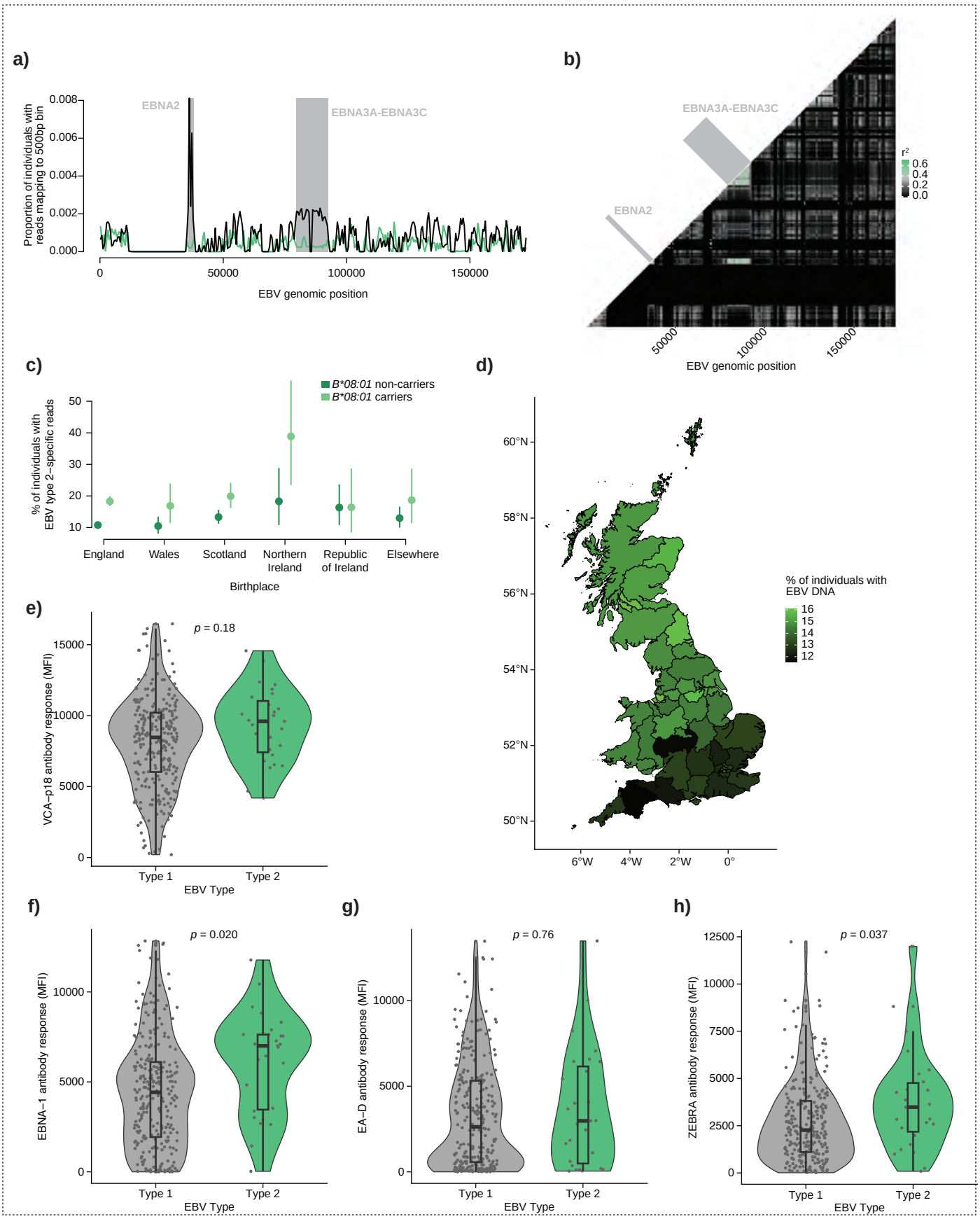
## Figure S8


a)
– log10(EBNA–1 antibody response p)
DRB1*15:01
DRB1*04:04
– log10(UK Biobank blood EBV viral load p)
b)
– log10(EA–D antibody response p)
DRB1*15:01
DRB1*04:04
– log10(UK Biobank blood EBV viral load p)
c)
– log10(ZEBRA antibody response p)
DRB1*15:01
DRB1*04:04
– log10(UK Biobank blood EBV viral load p)
d)
– log10(VCA_p18 antibody response p)
DRB1*15:01
DRB1*04:04
– log10(UK Biobank blood EBV viral load p)
e)
– log10(U14 antibody response p)
– log10(UK Biobank blood HHV7 viral load p)
f)
EBNA–1 antibody response (MFI)
p = 0.053
EBV read pairs
g)
EA–D antibody response (MFI)
p = 6.0 x 10–4
EBV read pairs
h)
ZEBRA antibody response (MFI)
p = 5.2 x 10–15
EBV read pairs
i)
VCA–p18 antibody response (MFI)
p = 8.1 x 10–25
EBV read pairs
j)
U14 antibody response (MFI)
p = 8.2 x 10–4
HHV–7 read pairs
k)
Prevalence of HHV–7 (%)
Type O
Type B
Type A
Type AB
Blood type
l)
Prevalence of HHV–7 (%)
O/O
O/B
B/B
O/A2
B/A2
A2/A2
O/A1
B/A1
A2/A1
A1/A1
ABO genotype
m)
Effect size on EBV viral load (s.d. units)
Y168H
Loss of function
CPVL
n)
Effect size on EBV viral load (s.d. units)
C104R
A181E
Loss of function
TNFRSF13B

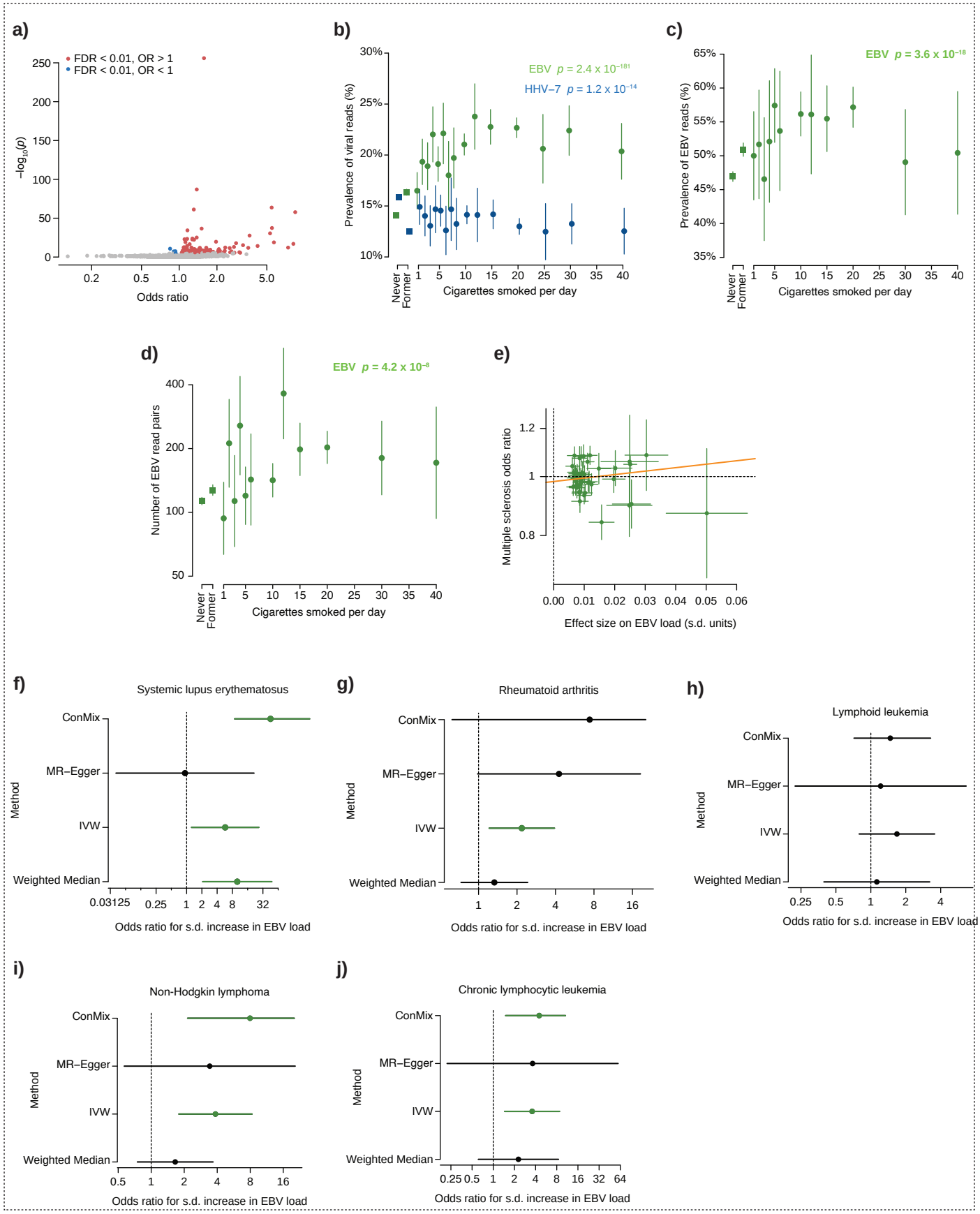
## Figure S9


a)
β, EBV viral load in DRB1*04:04 carriers (units s.d.)
β, EBV viral load in DRB1*04:04 non−carriers
b)
β, HHV-7 viral load in 6_31345435 C homozygotes (units s.d.)
β, HHV-7 viral load in 6_31345435 T homozygote
c)
β, EBV viral load in DRB1*04:04 carriers (units s.d.)
β, EBV viral load in DRB1*04:04 non−carriers (units s.d.)
Other variants
HLA coding variants
d)
DRB1*04:04 non-carriers
DRB1*04:04 heterozygotes
DRB1*04:04 homozygotes
Prevalence of EBV reads (%)
DPB1 V105 genotype

## Figure S10


a)
Proportion of individuals with reads mapping to 500bp bin
0.008
0.006
0.004
0.002
0.000
EBNA2
EBNA3A-EBNA3C
0
50000
100000
150000
EBV genomic position
b)
EBNA3A-EBNA3C
EBNA2
r²
0.6
0.4
0.2
0.0
50000
100000
150000
EBV genomic position
c)
% of individuals with EBV type 2–specific reads
B*08:01 non-carriers
B*08:01 carriers
50
40
30
20
10
England
Wales
Scotland
Northern Ireland
Republic of Ireland
Elsewhere
Birthplace
d)
60°N
58°N
56°N
54°N
52°N
50°N
6°W
4°W
2°W
0°
% of individuals with EBV DNA
16
15
14
13
12
e)
p = 0.18
VCA–p18 antibody response (MFI)
15000
10000
5000
0
Type 1
Type 2
EBV Type
f)
p = 0.020
EBNA–1 antibody response (MFI)
10000
5000
0
Type 1
Type 2
EBV Type
g)
p = 0.76
EA–D antibody response (MFI)
10000
5000
0
Type 1
Type 2
EBV Type
h)
p = 0.037
ZEBRA antibody response (MFI)
12500
10000
7500
5000
2500
0
Type 1
Type 2
EBV Type

## Figure S11


a)
FDR < 0.01, OR > 1
FDR < 0.01, OR < 1
−log10(p)
Odds ratio
b)
EBV p = 2.4 x 10−181
HHV−7 p = 1.2 x 10−14
Prevalence of viral reads (%)
Never
Former
Cigarettes smoked per day
c)
EBV p = 3.6 x 10−18
Prevalence of EBV reads (%)
Cigarettes smoked per day
d)
EBV p = 4.2 x 10−8
Number of EBV read pairs
Cigarettes smoked per day
e)
Multiple sclerosis odds ratio
Effect size on EBV load (s.d. units)
f)
Systemic lupus erythematosus
ConMix
MR−Egger
IVW
Weighted Median
Method
Odds ratio for s.d. increase in EBV load
g)
Rheumatoid arthritis
Odds ratio for s.d. increase in EBV load
h)
Lymphoid leukemia
Odds ratio for s.d. increase in EBV load
i)
Non-Hodgkin lymphoma
Odds ratio for s.d. increase in EBV load
j)
Chronic lymphocytic leukemia
Odds ratio for s.d. increase in EBV load