The DNA virome varies with human genes and environments

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[1,2]. Variable host control of their ongoing abundance ("load") can lead to clearance or disease[3–5]. We analyzed the viral DNA load of 31 common viruses in human blood and saliva using whole-genome sequencing data from UK Biobank (n=490,401), All of Us (n=414,817), and SPARK (n=12,519). Viral DNA load varied markedly with age, time of day, and season; most viruses were also present at greater abundance in men than in women. Human genetic variation at dozens of loci associated with DNA load of seven viruses: Epstein-Barr virus (EBV, 45 loci), human herpesvirus 7 (HHV-7, 37 loci), HHV-6B, Merkel cell polyomavirus, and three anelloviruses. Variation at the major histocompatibility complex (MHC) locus generated the strongest associations ($p = 5.8\times10^{-9}$ to 2.5×10^{-1459}), which were specific to each virus. The *HLA-B*08:01* allele also exhibited a host-virus genetic interaction with EBV subtype ($p = 7.4\times10^{-70}$). Other human genetic effects implicated 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 supported a strong causal effect of EBV DNA load on increased risk of Hodgkin lymphoma ($p = 1.8\times10^{-3}$) but not multiple sclerosis ($p = 0.52$). This suggests that higher chronic EBV load increases lymphoma risk, whereas associations of EBV infection with autoimmune conditions reflect host immune responses to particular viral epitopes.

Introduction

After infection by a virus, the long-term balance between viral replication and antiviral immune recognition within a person can result in spontaneous clearance[3], stable persistence around a set point of viral load, as with lifelong Epstein-Barr virus (EBV) infection[4], or progression to disease, such as the progression from human immunodeficiency virus (HIV) infection to acquired immunodeficiency syndrome (AIDS)[5]. EBV infection is a large risk factor for Hodgkin

lymphoma[6] and multiple sclerosis[7] years after primary infection, and is associated with other autoimmune disorders for reasons that are not well understood[8].

Large-scale DNA sequencing studies of blood samples from thousands of individuals have begun to characterize the "blood DNA virome"—the population of DNA viruses and retroviruses that is resident within each person's blood system—which spans a broad spectrum of viral species[9,10]. Two commonly observed taxa are herpesviruses, which are latently present in most humans, and anelloviruses, which are thought to be commensal or even symbiotic[11]. The amount of a virus present within an individual—the viral "load"—varies over time and is a potential biomarker of immune function[12–14]. Increased viral load may indicate reduced immunocompetence and can lead to disease (such as AIDS or liver cancer from viral hepatitis[15]), 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—and in other tissues such as saliva—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[16,17] and hepatitis C[18], and chronic persistence (vs. spontaneous clearance) of hepatitis B[19] and C[20]. 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[21,22]. The drivers of viral load variation in tissues other than blood are also largely unexplored.

Results

The DNA virome in blood and saliva

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 (mean 32.5x to 42x) of 490,401 blood samples in UK Biobank[23] (UKB), 365,918 blood samples and 48,899 saliva samples in the All of Us (AoU) data set[24], and 12,519 saliva samples in SPARK[25] (**Fig. 1a**). We aligned sequencing reads of potential non-human origin to a reference panel we assembled of 31 common viruses (27 DNA viruses and 4 retroviruses; **Extended Data Fig. 1a** and **Supplementary Table 1**), selected based on earlier WGS-based blood virome studies[9,10]. 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 such read pairs from a given biological sample. 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[26] and torque teno viruses (TTVs, a common form of anelloviruses)[14].

As in other cohorts[9,10], 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**, **Extended Data Fig. 1b**, and **Supplementary Tables 2** and **3**). This was expected despite the near-ubiquity of herpesviruses and anelloviruses in humans across populations[11] (e.g., 94.7% EBV seroprevalence in UKB) given the low fractions of latently infected cells that carry viral DNA[4] (**Methods**): viral reads are stochastically observed at rates proportional to the number of copies of a viral genome present in a person's blood. Despite this sparsity, several properties of the data set, such as enrichment of viral reads in seropositive individuals and consistent depth-of-coverage within viral genomes, supported genuine viral origin of these reads (**Extended Data Fig. 1c,d** and **2** and **Supplementary Note 1**). Additionally, the large majority of the sequencing reads ascertained to be of viral origin had been aligned to a viral reference genome with high mapping quality (**Supplementary Note 2**). Different viruses tended not to be co-observed much more often than by chance, with the exception of anelloviruses (**Extended Data Fig. 1e**). A small fraction of blood samples (675 out of 490,401 in UKB) exhibited unusually large numbers of parvovirus B19-derived reads[9,10]; these donors had greatly depleted reticulocyte counts ($p = 1.4\times10^{-67}$, **Extended Data Fig. 1f**), consistent with active viral infection of red blood cell precursors by B19V[27]. High, ~15x coverage of the HHV-6A or HHV-6B genome by WGS reads was also observed in ~1% of donors (**Fig. 1c** and **Extended Data Fig. 1b,g**), indicating inheritance of haplotypes containing copies of these viral genomes that have integrated into human chromosomes (endogenous HHV-6; eHHV-6)[28]. These carriers (0.32% of donors with eHHV-6A and 1.03% with eHHV-6B, comparable to previous work[28,29]) were excluded from analyses of HHV-6 viral DNA load.

In contrast to blood, saliva WGS data contained viral DNA sequences that were highly prevalent and abundant in SPARK (**Fig. 1d,e** and **Supplementary Table 4**) and AoU (**Extended Data Fig. 1h** and **Supplementary Table 5**). 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 originates 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 appear to occur continuously, such that HHV-7 DNA was detectable in nearly all saliva samples from adults (**Extended Data Fig. 1h**), whereas for EBV, viral shedding in saliva appears 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[30]. 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[31]. Endogenous HHV-6A and HHV-6B were observed in SPARK at rates similar to UKB (0.37% and 0.78%, respectively; **Fig. 1e** and **Extended Data Fig. 1i**). Viral DNA load levels were broadly independent of autism spectrum disorder status in SPARK (**Extended Data Fig. 1j,k**) and were consistent between SPARK and AoU saliva samples (**Fig. 1e** and **Extended Data Fig. 1h**).

Viral DNA load varies with age, sex, and time

The measurement of the prevalence and abundance of diverse viruses in the blood and saliva of so many people made it possible to recognize correlations with age, sex, and circadian and seasonal dynamics through cross-sectional analyses of the large cohort. Age and sex exhibited strong associations with viral DNA load (**Fig. 2a-d**). In blood samples from UKB, which recruited participants 40–70 years old, viral DNA prevalence of EBV and TTVs was greater in older individuals ($p = 6.7\times10^{-356}$ and 1.4×10^{-451}; regression), as recently observed for TTVs[14], whereas viral DNA prevalence was greater in younger individuals for HHV-7 ($p = 6.7\times10^{-200}$) and HHV-6B ($p = 1.7\times10^{-28}$), suggesting that viral load of HHV-7 and HHV-6B decline with age (**Fig. 2b**). In saliva samples from SPARK, which represented infancy to old age, viral DNA loads within age tranches (measured by viral read prevalence or mean abundance) exhibited trajectories that varied dramatically with age and across viruses (**Fig. 2a** and **Extended Data Fig. 1d** and **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 DNA prevalence continued to increase with age ($p = 7.6\times10^{-140}$), whereas HHV-6B DNA prevalence dropped sharply and steadily from childhood onward ($p = 6.4\times10^{-315}$, **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 ($p = 6.1\times10^{-585}$, **Fig. 2a**), consistent with the pattern in UKB blood samples (**Fig. 2b**), whereas HHV-7 viral abundance in saliva began dropping beginning in childhood ($p = 6.6\times10^{-194}$, **Extended Data Fig. 3c**), suggesting diminishing intensity or frequency of lytic events.

Notably, viral DNA 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[32]. These sex differences in viral DNA load appeared to emerge in adolescence (**Fig. 2a**). Viral DNA prevalence also displayed substantial circadian and seasonal variation (**Fig. 2e,f** and **Extended Data Fig. 3e-l**) that was robust to potential confounders (**Supplementary Note 3**), although confounding by unmeasured visit-related factors cannot be fully excluded. Viral read prevalence increased by ~1.2–1.3-fold from 9AM to 8PM (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 (Kronos[33] harmonic regression $p = 1.2\times10^{-9}$ and 3.3×10^{-23}; **Fig. 2e**). EBV viral read prevalence was also ~1.3-fold higher in winter compared to summer months, whereas HHV-7 showed different, more modest seasonal variation (Kronos $p = 2.9\times10^{-101}$ and 9.8×10^{-12}; **Fig. 2f**). These patterns replicated in AoU saliva samples (for EBV; **Extended Data Fig. 3k,l**) and were not driven by cyclical variation in lymphocyte percentages in UKB blood samples, which

exhibited weaker and different patterns (**Extended Data Fig. 3m,n**). 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 DNA load also exhibited striking, several-fold variation across genetic ancestries, with different viruses enriched in different ancestry groups (**Extended Data Fig. 4**, **Supplementary Table 6**, and **Supplementary Note 4**). These patterns were robust to controlling for available socioeconomic variables (**Supplementary Note 4)**.

Human genetics shapes viral DNA load

Almost all human traits are shaped by genetic variation, in ways that can elucidate molecular mechanisms. To identify such human genetic effects on viral DNA load, we performed genome-wide association analysis (GWAS) on viral DNA load phenotypes in UKB and AoU (**Methods**). For viral DNA load in blood, we then meta-analyzed results across UKB and AoU. Common variation (MAF>0.1%) at 82 loci in the human genome associated (at $p<5\times10^{-8}$) with viral DNA load of one or more viruses in blood (of six tested): 45 loci for EBV, 3 loci for HHV-6B, 37 loci for HHV-7, 6 for TTV (TUS01), 6 for TTV (HD14a), and 2 for TTV (VT-416) (**Fig. 3a-c**, **Extended Data Fig. 5**, and **Supplementary Table 7**). Common variants also associated with viral DNA load in saliva samples from AoU for EBV (2 loci), HHV-6B (3 loci), HHV-7 (12 loci), and MCPyV (2 loci) **Extended Data Fig. 6**). We verified that these associations were not driven by misalignment of human reads to viral genomes (**Extended Data Fig. 7a,b** and **Supplementary Note 5**), were concordant between UKB and AoU (**Extended Data Fig. 7c-e**), reflected distinct effects in blood versus saliva (**Extended Data Fig. 7f-i**), and did not overlap any of the 10 telomeric sites of eHHV-6 chromosomal integration we identified in UKB (**Extended Data Fig. 7j,k** and **Supplementary Note 6**).

For six of the above seven viruses (all except HHV-6B), the MHC region harbored the strongest associations between common human genetic variants and viral DNA load (**Fig. 3a-c**, **Extended Data Figs. 5** and **6**, and **Supplementary Table 7**). The MHC region as a whole contributed large proportions of viral DNA load heritability (56% for EBV, 32% for HHV-7, and 14–30% for the anelloviruses in UKB; **Fig. 3d**), as previously observed for HIV[17]. 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 8**). For EBV viral DNA 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[34]; *DRB1*04:04* might encode an isoform with higher affinity for EBV protein gp42, facilitating infection. Several other independent associations with EBV DNA 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 associations 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[22], but only a subset of these serological measurements (VCA-p18 and ZEBRA) exhibited association patterns similar to that of EBV DNA load (**Extended Data Fig. 8** and **Supplementary Note 7**), suggesting that levels of some but not all antibodies are influenced by viral DNA load.

The strong associations of HLA alleles with viral DNA load allowed us to characterize how these genetic effects vary across individuals as a function of age and other genotypes (**Supplementary Note 8** and **Extended Data Fig. 9a-h**). Some associations of variants in the MHC region with EBV DNA load also exhibited specificity to blood or saliva: in particular, the *HLA-A*02:01* allele appeared to decrease EBV DNA load in blood ($p = 2.9\times10^{-431}$) but have no effect on EBV DNA load in saliva ($p = 0.099$; **Extended Data Fig. 9i**), consistently across cohorts (**Extended Data Fig. 9j**). 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 DNA load in saliva.

Haplotypes at the *ERAP1–ERAP2–LNPEP* locus (on chromosome 5), which encodes the ERAP1, ERAP2, and LNPEP peptidases that process peptides for MHC display[35], associated strongly with DNA 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)[36]. In contrast, EBV DNA 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[37].

Beyond the very large effects from the *HLA* and *ERAP1–ERAP2–LNPEP* loci, each virus associated with additional human genomic loci (**Fig. 3a-c**, **Extended Data Fig. 5** and **6**, and **Supplementary Table 7**). 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 DNA load), likely encode alleles that affect survival and proliferation of infected host cells[38]. In contrast, *PML*, *SP110*, and *MX2* (associated with HHV-7 and/or HHV-6B DNA load) encode nuclear proteins with roles in restricting herpesvirus replication[39,40]. Associations at genes that influence apoptosis (*FAS*, *TNFAIP3*, and *TBKBP1*) suggest that virally-induced apoptosis signaling might be particularly important for anellovirus survival[41]. At the *ABO* locus, which determines the presence of blood group antigens bound by some viruses[42], alleles that produce type A and type B antigens both associated with higher HHV-7 DNA load (**Extended Data Fig. 9k,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[43,44].

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

load: *TERT* and *ASAH2B* for EBV, and *MX2* and *ZNF584* for HHV-7 ($p < 2.8\times10^{-6}$; **Supplementary Table 9**). These associations appeared to be virus-specific, except for *ZNF584*, for which rare coding variants also associated with increased EBV DNA 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 DNA load, the burden test demonstrated concordant associations of loss-of-function variants in these genes (**Extended Data Fig. 9m,n**).

The strong ancestry differences in viral DNA load (**Extended Data Fig. 4** and **Supplementary Note 4**) led us to wonder whether GWAS of EBV and HHV-7 DNA load in African-ancestry AoU participants (n=77,573) might identify specific loci contributing to these differences. These analyses identified a novel association at *ACKR1*: the rs2814778-CC genotype that generates the malaria-protective Duffy-null phenotype[45] appeared to explain 28% [21%–35%] and 45% [38%–52%] of the African-versus-European differences in EBV and HHV-7 DNA prevalence, respectively, and these associations are only partially explained by neutropenia in Duffy-null individuals[46] (**Extended Data Fig. 10** and **Supplementary Note 9**).

HLA-B*08:01 **association is EBV type-specific**

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[47] 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 DNA load of one EBV type but not the other, as has been observed for hepatitis C[48]. 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*)[49] (**Fig. 4a**, **Extended Data Fig. 11a,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. 4b,c** and **Extended Data Fig. 11c**), as expected[50]. EBV type 2 was relatively more prevalent in individuals born in Africa[50] compared with individuals born in Europe ($p = 5.4\times10^{-7}$) or North America ($p = 0.042$; **Fig. 4b**), and its relative prevalence also varied regionally among individuals born in the United Kingdom (**Fig. 4c**) independently of overall EBV prevalence (**Extended Data Fig. 11d**). Interestingly, quantitative antibody responses to EBV antigens were slightly higher in individuals with type 2 EBV than type 1 (**Extended Data Fig. 11e-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 suggested that HLA variation confers type-specific protection against EBV infection (**Fig. 4d-g**). In analyses of EBV type 1 and type 2 DNA load separately, association patterns in the MHC region (**Fig. 4d,e**) were broadly similar to those we had observed for total EBV DNA 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 DNA load were compared against genetic data from those with higher type 1 DNA

load—generated a readily interpretable association pattern led by the MHC class I *B*08:01* allele ($p = 7.4\times10^{-70}$; **Fig. 4f**). *B*08:01* appeared to be protective against type 1 EBV (OR = 0.62 [0.59–0.66] for observing a type 1 read; $p = 7.1\times10^{-72}$) but associated with increased risk for type 2 EBV (OR = 1.63 [1.48–1.80]; $p = 4.1\text{x}10^{-24}$), 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[51,52]. 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 associations of variation in both MHC class I and II regions (**Fig. 4g**).

EBV DNA load is a risk factor for Hodgkin lymphoma

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. 5a** and **Supplementary Tables 10–13**). Most of the disease associations involved higher viral DNA loads in affected individuals (**Extended Data Fig. 12a**), and many associated diseases involved inflammation or weakened immune function, such as AIDS, anemia, diabetes, and renal failure (**Fig. 5a**). Organ transplantation associated strongly with elevated anellovirus DNA load, presumably reflecting use of post-transplant immunosuppression medication[14]. Concordantly, individuals taking common immunosuppressive drugs had elevated EBV and anellovirus DNA loads (**Supplementary Table 14**); this could explain some of the observed disease associations. Carrier status for endogenous HHV-6B or HHV-6A did not significantly associate with any biological or clinical phenotypes (**Supplementary Table 10**).

Many of the strongest associations of EBV and HHV-7 DNA load with disease phenotypes were directly or indirectly related to smoking (e.g., nicotine dependence, respiratory disorders, and lung cancer; **Fig. 5a** and **Supplementary Tables 10 and 11**). 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. 5b**). In contrast, smoking exposure associated with decreased prevalence of HHV-7 DNA ($p = 9.6\times10^{-22}$; **Fig. 5b**). Cigarettes smoked per day associated with similar trends ($p = 3.3\times10^{-563}$ and 2.7×10^{-9} for increased EBV prevalence and decreased HHV-7 prevalence; **Fig. 5c**), suggesting that smoking has strong, opposite effects on viral DNA load for EBV and HHV-7. These associations replicated in AoU blood samples (**Extended Data Fig. 12b**), 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 ($p = 4.2\times10^{-8}$, **Extended Data Fig. 12c,d**).

The large number of genetic effects on EBV DNA load identified by our GWAS provided an opportunity to investigate the causal direction of phenotypic associations: i.e., whether viral DNA load affects (or is affected by) the clinical phenotypes with which it associates. We did this by using these DNA load-influencing SNPs as instrument variables for Mendelian randomization,

asking whether or not the genetic variants that affect viral DNA load also affect the traits with which viral DNA load associates in populations[53]. While most viral DNA 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 DNA load also consistently confer risk for disease offers evidence that viral DNA load is causally upstream.

To use this approach to assess the potential effect of EBV DNA load on autoimmune diseases and blood cancers epidemiologically associated with EBV infection[2], we analyzed GWAS summary statistics from a meta-analysis of FinnGen[54], UK Biobank[55], and the Million Veteran Program[56] using four MR methods, excluding the MHC region to guard against pleiotropy from linkage disequilibrium[53] (**Supplementary Table 15**). For multiple sclerosis (MS), for which prior EBV infection is a large risk factor[7], these analyses indicated that latent viral DNA load after primary infection is unlikely to further influence the risk of developing MS (MR-Egger $p = 0.52$, **Fig. 5d** and **Extended Data Fig. 12e**) despite UKB participants with MS having elevated EBV DNA load ($p = 0.005$). For systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA), which also associated with elevated EBV DNA load in UKB ($p = 4.6\times10^{-7}$ and 2.2×10^{-23}, respectively), MR analyses were inconclusive about causality (**Extended Data Fig. 12f,g**).

In contrast, all analyses (across the four Mendelian randomization methods) suggested that viral DNA load after EBV infection significantly influences the development of Hodgkin lymphoma (HL; $p = 1.2\times10^{-3}$ to 0.014; **Fig. 5e,f** and **Supplementary Table 15**). 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 (**Extended Data Fig. 12h-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 (OR = 2.01 [1.40, 2.84] for incident HL among EBV DNA-positive vs. DNA-negative individuals; **Fig. 5g**). Elevated EBV viral DNA 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, we have characterized the major factors that are likely to shape these constituents of the human virome. Human genetics, age, sex, time of day, month of the year, and smoking behavior consistently associate strongly with viral DNA 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. These results also corroborate parallel studies of EBV in UKB and AoU[57–59]; here, we extended analyses to 30 other viruses, and for EBV, we also observed interactions of genetic effects with biosample type and with EBV genetic variation.

The largely distinct, polygenic human genetic effects that we observed for DNA load of several common viruses (EBV, HHV-6B, HHV-7, MCPyV, and anelloviruses) expand upon earlier work identifying key loci influencing control of HIV[16,17] and hepatitis C[18,48] load. The MHC region contributed the strongest genetic effects for each virus we studied except HHV-6B, similar to other GWAS of infectious disease phenotypes[60], but the specific HLA alleles that influenced viral DNA load varied across viruses (**Supplementary Table 8**). These virus-specific effects at the MHC probably contribute to the striking variation of viral DNA 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[61]. Beyond the MHC, polygenic influences on DNA 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[39,40], associated with DNA load of HHV-7 but not EBV. We also highlight the utility of using pairs of related phenotypes, such as type 1 and type 2 EBV DNA load (**Fig. 4**) or EBV DNA load in blood and saliva (**Extended Data Fig. 9i,j**), as a mechanism to accomplish fine-mapping of a complex genetic association.

The large number of polygenic effects we observed for EBV DNA load—45 human genomic loci—enabled Mendelian randomization to evaluate causality of associations with clinical phenotypes (**Fig. 5**). These analyses suggested that multiple sclerosis risk from EBV infection may be mediated largely by adaptive immune response to an EBV infection[62] without an additional influence of subsequent lifetime viral DNA load. In contrast, EBV DNA load appears to be a causal risk factor for Hodgkin lymphoma (HL). This result contrasts with an MR analysis from a parallel study that did not observe an effect of EBV DNA load on HL[58]; this discrepancy might be explained by increased statistical power here (from HL GWAS summary statistics from a 2.6-fold larger meta-analysis, as well as more EBV DNA load GWAS hits available as genetic instruments) and inclusion versus exclusion of the MHC region. This result could potentially be validated *in vitro* by exposing germinal center B-cells to increasing titers of EBV and measuring the number of Hodgkin and Reed-Sternberg (HRS) cells produced[63]; alternatively, it could be studied in humanized mice[64] if HRS generation requires a tissue niche. An effect on HL is not unexpected given the risk of EBV-positive HL conferred by infectious mononucleosis[6], but it suggests that interindividual differences in lifetime management of DNA load affect health risks beyond infection itself. If so, interventions that reduce EBV DNA load could potentially decrease the risk of HL. Notably, smoking has previously been found to be a risk factor for EBV-positive HL (OR=1.81 [1.27 - 2.56] for current smokers), but not for EBV-negative HL (OR=1.02 [0.72–1.44])[65] and more weakly for NHL (OR=1.10 [1.00-1.20])[66], suggesting that the effect of smoking on EBV DNA 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[67] may be worth evaluation for potential prophylactic benefit against HL.

Our analyses of human viruses commonly observable from population DNA sequencing data had several limitations that will need future efforts to resolve using analytical approaches that more completely capture viral genetic variation, additional cohorts of greater scale, and samples

obtained from a wider range of body sites (**Supplementary Note 10**). 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 hepatitis B and retroviruses such as HIV and HTLV.

Figure Legends

Figure 1: Detection and quantification of viral DNA 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 (n=365,918) cohorts and saliva samples from the Simons Foundation Powering Autism Research (SPARK, n=12,519) and All of Us (n=48,899) cohorts. 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 associations with viral DNA load.

a) Prevalence of viral DNA sequences by age and sex in saliva-derived WGS from SPARK (n=12,519 total, 58-1,582 participants per age-sex bin). 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 (n=482,882 total, 22,896-63,810 participants per age-sex bin). Error bars, 95% CIs.
c) Prevalence of viral DNA sequences by sex in blood-derived WGS from UKB (EBV, HHV-7, HHV-6B, TTV, n=222,094 men and 263,132 women) and AoU (CMV, n=143,918 men and 220,649 women) and saliva-derived WGS from AoU (EBV, CMV, MCPyV, HSV-1, n=17,335 men and 31,425 women) and SPARK (HHV-7, HHV-6B, n=3,378 fathers and 3,380 mothers). Error bars, 95% CIs.
d) Abundance of viral DNA sequences by sex in saliva-derived WGS from AoU (EBV, n=9,973 men and 15,856 women; CMV, n=409 men and 622 women; MCPyV, n=1,840 men and 1,185 women) and SPARK (HHV-7, n=3,288 fathers and 3,217 mothers; HHV-6B, n=1,945 fathers and 1,762 mothers). Centres indicate mean and 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, n=490,136 total, 6,618-52,949 participants per hour). Error bars, 95% CIs.
f) Prevalence of viral DNA sequences in UKB blood WGS data by month of sample collection (n=490,136 total, 27364-51195 participants per month). Error bars, 95% CIs. P

values, two-sided linear regression (**c**,**d**) or one-sided ANOVA test between models with and without collection hour (**e**) or month (**f**) (see **Methods**).

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

a) Genome-wide associations of human genetic variants with EBV DNA load in blood (inverse-normal transformed abundance), meta-analyzed across UKB (n=453,770) and AoU (n=201,168) European-ancestry sub-cohorts.
b) Genome-wide associations with HHV-7 DNA load in blood (inverse-normal transformed abundance), meta-analyzed across UKB (n=453,770) and AoU (n=200,091) European-ancestry sub-cohorts.
c) Genome-wide associations with anellovirus TUS01 DNA load in blood (inverse-normal transformed abundance), meta-analyzed across UKB (n=453,770) and AoU (n=200,091) European-ancestry sub-cohorts.
d) Estimates of heritability from common genetic variants for viral DNA load phenotypes from variants within versus outside of the MHC region of the human genome, defined as 24–35Mb on chromosome 6 (n=419,421 unrelated European-ancestry UKB participants). Error bars, 95% CIs.
e) Associations of variants in the MHC region of the human genome with EBV DNA load in blood in European-ancestry UKB participants.
f) Analogous to e, for HHV-7 DNA load.
g) Associations of common genetic variants at the *ERAP1–ERAP2–LNPEP* locus with EBV DNA load (green) and HHV-7 DNA 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: EBV type 1 versus 2 analysis identifies a type-specific association with *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 (n=12,519). 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-DNA-positive individuals) by location of birth. In each group, the proportion was estimated as the number of UKB blood WGS samples containing two or more EBV type 2-specific reads divided by the number containing two or more type 1-specific or two or more type-2-specific reads (total n=11,511 participants). Estimates are shown for countries of birth with more than 35 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 (n=9,769).

d) Associations of variants in the MHC region of the human genome with EBV type 1 positivity (presence of two or more WGS reads that mapped to type-informative regions of the EBV type 1 genome) in UKB blood WGS (n=418,403 unrelated European ancestry participants). 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 (n=9,233 unrelated European ancestry participants).
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. P values, two-sided linear regression (**d**-**g**).

Figure 5: Associations of viral DNA 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). Several of the strongest independent associations are shown in the plot; the full set of associations is provided in **Supplementary Tables 10** and **11**.
b) Prevalence of detectable EBV positivity (in green) and HHV-7 positivity (in blue) by smoking pack-years in UKB (n=379,233 European ancestry participants). Never smokers are indicated by the square point on the left of the plot.
c) Prevalence of detectable EBV positivity (in green) and HHV-7 positivity (in blue) by cigarettes smoked per day in UKB (n=438,101 European ancestry participants). Never or former smokers are indicated by the square points on the left.
d) Estimates of the causal effect of EBV DNA load (exposure; n=638,825) on risk for multiple sclerosis (outcome; meta-analysis of 7,907 cases and 1,474,810 controls) 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). ConMix, contamination mixture; IVW, inverse-variance weighted.
e) Estimates of the causal effect of EBV DNA load on risk for Hodgkin lymphoma (meta-analysis of 2,529 cases and 1,159,394 controls). MR odds ratios are expected to be overestimated due to measurement noise (**Methods**).
f) Effect sizes for 44 common genetic variants (at distinct non-MHC loci)—used as instrument variables in Mendelian randomization—for EBV DNA 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; n=174 incident cases), 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 in all panels. P values, two-sided linear regression (**b**,**c**).

Extended Data Figure Legends

Extended Data Figure 1: Viral DNA 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.[68].
b) Prevalence and abundance of viral DNA sequences in blood samples from AoU (n=365,918).
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) Prevalence of viral DNA sequences in saliva-derived WGS by age (rounded to the nearest month) in SPARK for children under 3 years old (total n=576; n=1, 2, 5, 6, 4, 8, 4, 4 for the 0-month to 7-month bins; 7≤n≤45 for remaining bins).
e) Pairwise enrichments of co-observing reads from two viruses in the same UKB sample. * for $p < 0.01$.
f) Reticulocyte counts of European ancestry UKB participants (n=434,573, y-axis) stratified by number of reads mapping to the parvovirus B19 genome (x-axis). P-value is from regression of log-transformed reticulocyte counts on log-transformed B19 reads with sex, age, age-squared, assessment center, and 20 genetic principal components as covariates.
g) 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.
h) Prevalence and abundance of viral DNA sequences in saliva samples from AoU (n=48,899).
i) Analogous to g, for saliva-derived WGS from SPARK (n=12,519).
j) Prevalence and abundance of viral DNA sequences in saliva samples from children without autism spectrum disorder in SPARK (n=2,220).
k) Prevalence and abundance of viral DNA sequences in saliva samples from children with autism spectrum disorder in SPARK (n=3,540).

Extended Data 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. Additional metrics on coverage in each cohort can be found in **Supplementary Tables 2-5**.

Extended Data Figure 3: Additional data supporting age, sex, circadian, and seasonal associations with viral DNA load in saliva.

a) Abundance of viral DNA sequences in saliva-derived WGS by age (rounded to the nearest year) in SPARK. Sample sizes for each virus are in the following panel legends. 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 (n=4,753). Error bars, 95% CIs.
c) Analogous to b, for HHV-7 DNA sequences (n=9,978).
d) Analogous to b, for HHV-6B DNA sequences (n=8,528).
e) Prevalence of EBV DNA sequences in UKB blood WGS by time of day of sample collection (rounded to the nearest hour) split by assessment center (n=490,136).
f) Prevalence of EBV DNA sequences in UKB blood WGS data by month of sample collection split by assessment center.
g) Prevalence of HHV-7 DNA sequences in UKB blood WGS by time of day of sample collection (rounded to the nearest hour) split by assessment center.
h) Prevalence of HHV-7 DNA sequences in UKB blood WGS data by month of sample collection split by assessment center.
i) Prevalence of EBV DNA sequences in UKB blood WGS by time of day of sample collection (rounded to the nearest hour) split by season (winter, Dec-Feb; spring, Mar-May; summer, Jun-Aug; autumn, Sep-Nov).
j) Prevalence of HHV-7 DNA sequences in UKB blood WGS by time of day of sample collection (rounded to the nearest hour) split by season.
k) 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.
l) 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.
m) Lymphocyte percentage in UKB blood samples by sample collection time of day (n=472,115). Error bars, 95% CIs.
n) Lymphocyte percentage in UKB blood samples by sample collection month. Error bars, 95% CIs.

Extended Data Figure 4: Variation in viral DNA 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, n=474,353). Error bars, 95% CIs. Per-ancestry prevalences and abundances can be found in **Supplementary Table 6**.
b) Prevalence of EBV DNA positivity in blood-derived WGS data by genetic ancestry in AoU (n=365,918). Error bars, 95% CIs.
c) Prevalence of EBV DNA positivity in saliva-derived WGS data by genetic ancestry in AoU (n=48,899). Error bars, 95% CIs.

d) Abundance of EBV DNA sequences in saliva-derived WGS data by genetic ancestry in AoU (n=25,915). Error bars, 95% CIs.
e) Analogous to c, for saliva-derived WGS in adults (≥20 years old) from SPARK (n=5,691).
f) Analogous to d, for saliva-derived WGS in adults from SPARK (n=2,730).
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 plus American ancestry, n=184,514). Error bars, 95% Cis.
h) Analogous to g, for proportion of estimated African ancestry admixture for AoU participants with estimated >90% European plus African ancestry (n=168,651). 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 (n=474,353). Error bars, 95% CIs.
j) Prevalence of HHV-7 DNA positivity in blood-derived WGS data by genetic ancestry in AoU (n=365,918). Error bars, 95% CIs.
k) Prevalence of anellovirus DNA positivity in blood-derived WGS data by genetic ancestry in UKB (n=474,353). Error bars, 95% CIs.
l) Prevalence of anellovirus DNA positivity in blood-derived WGS data by genetic ancestry in AoU (n=365,918). Error bars, 95% CIs.

Extended Data Figure 5: Genome-wide association analyses of HHV-6B and anellovirus (HD14a and VT416 TTV strains) DNA load in blood.

a) Genome-wide associations of human genetic variants with HHV-6B DNA load in blood (inverse-normal transformed abundance), meta-analyzed across UKB (n=447,190) and AoU (n=198,059) European-ancestry sub-cohorts.
b) Analogous to a, for the anellovirus HD14a strain (UKB n=453,770; AoU n=200,091).
c) Analogous to a, for the anellovirus VT416 strain (UKB n=453,770; AoU n=200,091).

Extended Data Figure 6: Genome-wide association analyses of viral DNA 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=33,164). 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 DNA load (presence/absence of any reads from the virus) and quantitative viral DNA 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 (n=16,282).
c) Analogous to a, for HHV-7 positivity (n=33,050).
d) Analogous to a, for HHV-7 abundance (n=29,989).
e) Analogous to a, for HHV-6B positivity (n=32,711). Carriers of endogenous HHV6-A or HHV6-B germline integrations were excluded.
f) Analogous to a, for Merkel cell polyomavirus positivity (n=33,050).

Extended Data Figure 7: Robustness of genetic associations with EBV and HHV-7 DNA load and GWAS for HHV-6 integration sites.

a) Effect sizes for index variants at 44 independent non-MHC EBV DNA load-associated loci for association with presence of reads aligned to the left (x-axis) or right (y-axis) half of the EBV genome (n=419,249 unrelated European ancestry participants).
b) Analogous to a, for index variants at 34 independent non-MHC HHV-7 DNA load-associated loci. Two lower frequency variants of large effect size were excluded to facilitate visualization of the remaining variants.
c) Effect sizes for index variants at 45 independent EBV blood DNA load-associated loci for association with EBV DNA load in blood samples from UKB (x-axis; n=453,770) and AoU (y-axis; n=201,181). Line shown is the median ratio of effect sizes between UKB and AoU.
d) Analogous to c, for index variants at 35 independent HHV-7 blood DNA load-associated loci. Two lower frequency variants of large effect size were excluded to facilitate visualization of the remaining variants.
e) Analogous to c, for index variants at 3 independent HHV-6B blood DNA load-associated loci.
f) Effect sizes for index variants at 45 independent EBV blood or saliva DNA load-associated loci for association with EBV DNA load in blood samples from UKB+AoU (x-axis; n=453,770+201,168) and EBV presence in saliva samples from AoU (y-axis; n=33,164). Line shown is the median ratio of effect sizes between UKB+AoU blood and AoU saliva; technical variation (e.g., in sample processing or DNA sequencing for saliva vs. blood samples) would generally be expected to increase or decrease signal-to-noise to the same extent across different genetic effects, which would not explain the observed heterogeneity in saliva vs. blood effect size ratios.
g) Analogous to f, for index variants at 40 independent HHV-7 blood or saliva DNA load-associated loci. Two lower frequency variants of large effect size were excluded to facilitate visualization of the remaining variants.
h) Analogous to f, for index variants at 5 independent HHV-6B blood or saliva DNA load-associated loci.
i) Genetic correlations between viral phenotypes in UKB, AoU, and SPARK. UKB and AoU were included separately when statistical power was sufficient within each cohort; for HHV6B and TTV (VT416), only the meta-analysis had sufficient power. Borders around cells indicate the same phenotype in different cohorts (e.g., UKB and AoU blood samples). * indicates $p < 0.1$ and ** for $p < 0.01$.
j) Genome-wide associations of human genetic variants with carrier status for endogenous HHV-6A germline integration (eHHV-6A) in unrelated European-ancestry UKB participants (n=419,249). Peritelomeric peaks indicate haplotypes in linkage disequilibrium with telomeric HHV-6A insertions.
k) Analogous to c, for endogenous HHV-6B germline integration (eHHV-6B).

Extended Data Figure 8: Analyses of serology data and comparison with viral DNA 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, n=8,100) and EBV DNA load in UKB blood WGS (x-axis; n=453,770). Serology measurements were from unrelated European ancestry participants and associations were performed with BOLT-LMM using standard covariates.
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 DNA load (x-axis; n=453,770).
f) Quantitative antibody response to EBV-encoded EBNA-1 (y-axis) stratified by EBV DNA fragment count in UKB blood WGS (x-axis) in EBV seropositive individuals (n=8,325). 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) in HHV-7 seropositive individuals (n=8,341). P-value is from a linear regression model including standard covariates and rank-based inverse normal transformed HHV-7 read counts.

Extended Data Figure 9: Non-additive genetic effects and associations of blood type and coding variants with viral DNA load in blood.

a) Prevalence of detectable EBV positivity by age in UKB blood WGS, stratified by *HLA-DRB1*04:04* genotype (n=452,318 European ancestry participants). 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) Effect sizes of index variants at 44 independent non-MHC loci associated with EBV DNA load in UKB blood WGS in *DRB1*04:04* carriers (x-axis, n=34,089) and *DRB1*04:04* non-carriers (y-axis, n=371,339).
d) Effect sizes of index variants at 23 independent non-MHC loci associated with HHV-7 DNA load in UKB blood WGS in chr6:31345435 T/T homozygotes (x-axis, n=167,251) and C/C homozygotes (y-axis, n=57,034). Three lower frequency variants of large effect size were excluded to facilitate visualization of the remaining variants.
e) Effect sizes of variants in the MHC region in *DRB1*04:04* carriers (x-axis, n=34,089) and *DRB1*04:04* non-carriers (y-axis, n=371,339) on EBV DNA 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.
f) Prevalence of detectable EBV positivity by *HLA-DRB1*04:04* and *HLA-DRB1* T210 genotype combinations in UKB blood WGS (n=454,545 European ancestry participants). Error bars, 95% CIs.

g) Prevalence of detectable EBV positivity by *HLA-DRB1*04:04* and *HLA-DQA1* E198 genotype combinations in UKB blood WGS (n=454,545 European ancestry participants). Error bars, 95% CIs.
h) Prevalence of EBV DNA positivity by *HLA-DRB1*04:04* and *HLA-DPB1* V105 genotype combinations in UKB (n=454,545 European ancestry participants). Error bars, 95% CIs.
i) Associations of common genetic variants in the MHC region with EBV DNA load in blood (UKB, n=453,770 European ancestry participants; x-axis) and saliva (AoU, n=33,168 European ancestry participants; 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).
j) Effect sizes of *HLA-A*02:01* on EBV DNA load in blood (green points; UKB, n=453,770 European ancestry participants; AoU, n=201,181 European ancestry participants) and saliva (gray points; AoU, n=33,168 European ancestry participants; SPARK, n=9,209 European ancestry participants). 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). Error bars, 95% CIs.
k) Prevalence of HHV-7 DNA positivity by genetically determined blood type in UKB (n= 456,988 European ancestry participants). Blood type was determined as in ref[69].
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 DNA load for the *CPVL* Y168H missense variant and *CPVL* loss-of-function variants (aggregated across such variants in a burden test; n=453,770).
n) Effect size on EBV DNA 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[70].

Extended Data Figure 10: Genome-wide association analyses of viral DNA load in blood samples from African ancestry participants.

a) Genome-wide associations of human genetic variants with EBV DNA positivity in blood-derived WGS from African-ancestry AoU participants (n=77,573).
b) Genome-wide associations of human genetic variants with HHV-7 DNA positivity in blood-derived WGS from African-ancestry AoU participants (n=77,573).
c) Prevalence of detectable EBV and HHV-7 positivity (reads mapping to each viral genome) by genetic ancestry and Duffy-null genotype in AoU blood WGS (n=251,209 participants). The AFR sub-cohort analyzed here consisted of individuals with 70–90% African ancestry (n=51,052 participants) to minimize correlation of Duffy-null genotype with admixture proportion. Error bars, 95% CIs.
d) Number of reads mapping to the EBV (n=18,723 participants) and HHV-7 (n=33,823 participants) genomes by genetic ancestry and Duffy-null genotype in AoU saliva WGS. The AFR sub-cohort analyzed here consisted of individuals with 70–90% African ancestry (n=51,052 participants) to minimize correlation of Duffy-null genotype with admixture proportion. Error bars, 95% CIs.

Extended Data Figure 11: 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 (n=490,401). 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=4,757). 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 reads (as a fraction of those positive for 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 (n=456,923). 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.

Extended Data Figure 12: Additional data regarding associations of viral DNA 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 (n=365,918). 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 (n=48,899) 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 DNA load (x-axis; n=638,825) and risk for multiple sclerosis (y-axis; meta-analysis of 7,907 cases and 1,474,810 controls). 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 DNA load (exposure) on risk for systemic lupus erythematosus (outcome; meta-analysis of 3,491 cases and 1,361,247 controls) 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[71], we do not conclude causality, but rather that viral DNA load is more likely to affect the risk of systemic lupus erythematosus compared to MS.
g) Estimates of the causal effect of EBV DNA load on risk for rheumatoid arthritis (meta-analysis of 35,875 cases and 1,296,933 controls). Error bars, 95% CIs.
h) Estimates of the causal effect of EBV DNA load on risk for lymphoid leukemia (meta-analysis of 5,966 cases and 1,248,154 controls). Error bars, 95% CIs.
i) Estimates of the causal effect of EBV DNA load on risk for non-Hodgkin lymphoma (meta-analysis of 13,198 cases and 1,273,062 controls). Error bars, 95% CIs.
j) Estimates of the causal effect of EBV DNA load on risk for chronic lymphocytic leukemia (meta-analysis of 4,467 cases and 1,248,724 controls). 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. Damania, B., Kenney, S. C. & Raab-Traub, N. Epstein-Barr virus: Biology and clinical disease. *Cell* **185**, 3652–3670 (2022).
3. 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).
4. 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).
5. 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).
6. Hjalgrim, H. *et al.* Characteristics of Hodgkin's Lymphoma after Infectious Mononucleosis. *N. Engl. J. Med.* **349**, 1324–1332 (2003).
7. Bjornevik, K. *et al.* Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* **375**, 296–301 (2022).
8. 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).
9. Moustafa, A. *et al.* The blood DNA virome in 8,000 humans. *PLoS Pathog.* **13**, e1006292 (2017).
10. 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).
11. 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).

12. Gulley, M. L. & Tang, W. Laboratory Assays for Epstein-Barr Virus-Related Disease. *J. Mol. Diagn. JMD* **10**, 279–292 (2008).
13. Saag, M. S. *et al.* HIV viral load markers in clinical practice. *Nat. Med.* **2**, 625–629 (1996).
14. 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).
15. 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).
16. Fellay, J. *et al.* A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science* **317**, 944 (2007).
17. 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).
18. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399–401 (2009).
19. 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).
20. Thomas, D. L. *et al.* Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature* **461**, 798–801 (2009).
21. 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).
22. Kachuri, L. *et al.* The landscape of host genetic factors involved in immune response to common viral infections. *Genome Med.* **12**, 93 (2020).
23. Carss, K. *et al.* Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* 1–10 (2025) doi:10.1038/s41586-025-09272-9.

24. Bick, A. G. *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
25. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research: Neuron. https://www.cell.com/neuron/fulltext/S0896-6273(18)30018-7.
26. 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).
27. Chisaka, H., Morita, E., Yaegashi, N. & Sugamura, K. Parvovirus B19 and the pathogenesis of anaemia. *Rev. Med. Virol.* **13**, 347–359 (2003).
28. Flamand, L. *et al.* Endogenous human herpesviruses 6A/B. *J. Virol.* **0**, e01054-25 (2025).
29. Liu, X. *et al.* Endogenization and excision of human herpesvirus 6 in human genomes. *PLOS Genet.* **16**, e1008915 (2020).
30. 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).
31. 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).
32. Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).
33. Bastiaanssen, T. F. S. *et al.* Kronos: A computational tool to facilitate biological rhythmicity analysis. 2023.04.21.537503 Preprint at https://doi.org/10.1101/2023.04.21.537503 (2023).
34. 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).
35. Saveanu, L. *et al.* IRAP identifies an endosomal compartment required for MHC class I cross-presentation. *Science* **325**, 213–217 (2009).

36. 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).
37. 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).
38. Willis, T. G. & Dyer, M. J. S. The role of immunoglobulin translocations in the pathogenesis of B-cell malignancies. *Blood* **96**, 808–822 (2000).
39. Scherer, M. & Stamminger, T. Emerging Role of PML Nuclear Bodies in Innate Immune Signaling. *J. Virol.* **90**, 5850–5854 (2016).
40. Schilling, M. *et al.* Human MxB Protein Is a Pan-herpesvirus Restriction Factor. *J. Virol.* **92**, e01056-18 (2018).
41. Verstrepen, L. *et al.* Expression, biological activities and mechanisms of action of A20 (TNFAIP3). *Biochem. Pharmacol.* **80**, 2009–2020 (2010).
42. Tan, M. & Jiang, X. Norovirus and its histo-blood group antigen receptors: an answer to a historical puzzle. *Trends Microbiol.* **13**, 285–293 (2005).
43. Kubiniok, P. *et al.* Understanding the constitutive presentation of MHC class I immunopeptidomes in primary tissues. *iScience* **25**, 103768 (2022).
44. 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).
45. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295**, 302–304 (1976).
46. Reich, D. *et al.* Reduced Neutrophil Count in People of African Descent Is Due To a Regulatory Variant in the Duffy Antigen Receptor for Chemokines Gene. *PLoS Genet.* **5**, e1000360 (2009).

47. 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).
48. 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).
49. Zanella, L. *et al.* A reliable Epstein-Barr Virus classification based on phylogenomic and population analyses. *Sci. Rep.* **9**, 9829 (2019).
50. Zimber, U. *et al.* Geographical prevalence of two types of Epstein-Barr virus. *Virology* **154**, 56–66 (1986).
51. Moss, D. J. *et al.* Cytotoxic T-cell clones discriminate between A- and B-type Epstein-Barr virus transformants. *Nature* **331**, 719–721 (1988).
52. An Epstein-Barr virus-specific cytotoxic T cell epitope in EBV nuclear antigen 3 (EBNA 3). *J. Exp. Med.* **171**, 345–349 (1990).
53. Sanderson, E. *et al.* Mendelian randomization. *Nat. Rev. Methods Primer* **2**, 6 (2022).
54. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
55. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
56. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024).
57. Nyeo, S. S. *et al.* Population-scale sequencing resolves determinants of persistent EBV DNA. *Nature* 1–9 (2026) doi:10.1038/s41586-025-10020-2.
58. Schmidt, A. *et al.* Host control of latent Epstein-Barr virus infection. 2025.07.19.25331823 Preprint at https://doi.org/10.1101/2025.07.19.25331823 (2025).
59. Yasumizu, Y. *et al.* Host Genetic Architecture between Epstein–Barr Virus Activity and Multiple Sclerosis Reveals Shared Pathways. 2025.12.11.25342083 Preprint at https://doi.org/10.64898/2025.12.11.25342083 (2025).

60. 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).
61. 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).
62. Lanz, T. V. *et al.* Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. *Nature* **603**, 321–327 (2022).
63. Küppers, R. *et al.* Identification of Hodgkin and Reed-Sternberg cell-specific genes by gene expression profiling. *J. Clin. Invest.* **111**, 529–537 (2003).
64. Fujiwara, S., Imadome, K.-I. & Takei, M. Modeling EBV infection and pathogenesis in new-generation humanized mice. *Exp. Mol. Med.* **47**, e135–e135 (2015).
65. 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).
66. 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).
67. 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).
68. 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).
69. Kamitaki, N. *et al.* Human and bacterial genetic variation shape oral microbiomes and health. *Nature* 1–11 (2026) doi:10.1038/s41586-025-10037-7.
70. Salzer, U. & Grimbacher, B. TACI deficiency — a complex system out of balance. *Curr. Opin. Immunol.* **71**, 81–88 (2021).

71. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **44**, 313–329 (2020).

72. Manghi, P. *et al.* Large-scale metagenomic analysis of oral microbiomes reveals markers for autism spectrum disorders. *Nat. Commun.* **15**, 9743 (2024).

73. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

74. Gray, G. C. *et al.* Adult Adenovirus Infections: Loss of Orphaned Vaccines Precipitates Military Respiratory Disease Epidemics. *Clin. Infect. Dis.* **31**, 663–670 (2000).

75. Kamminga, S., van der Meijden, E., Feltkamp, M. C. W. & Zaaijer, H. L. Seroprevalence of fourteen human polyomaviruses determined in blood donors. *PLoS ONE* **13**, e0206273 (2018).

76. Gossai, A. *et al.* Seroepidemiology of Human Polyomaviruses in a US Population. *Am. J. Epidemiol.* **183**, 61–69 (2016).

77. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

78. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

79. 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).

80. Hujoel, M. L. A. *et al.* Insights into DNA repeat expansions among 900,000 biobank participants. *Nature* 1–10 (2026) doi:10.1038/s41586-025-09886-z.

81. Conley, A. B. *et al.* Rye: genetic ancestry inference at biobank scale. *Nucleic Acids Res.* **51**, e44 (2023).

82. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8**, e64683 (2013).

83. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).

84. 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).

85. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

86. Hujoel, M. L. A. *et al.* Influences of rare copy-number variation on human complex traits. *Cell* **185**, 4233-4248.e27 (2022).

87. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).

88. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

89. 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).

90. 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).

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

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

93. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).

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

95. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

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

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

98. 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

Ethics

This research complies with all relevant ethical regulations. The study protocol (NHSR-8429) was determined to be not human subjects research by the Broad Institute Office of Research Subject Protection as all data analyzed were previously collected and de-identified. Use of SPARK data for this research was approved by SFARI (project 3350.2).

UK Biobank, All of Us, and SPARK 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[23] was performed on libraries prepared with NEBNext Ultra™ II PCR-free kit (New England Biolabs) using 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 NIH All of Us v8 cohort[24] was performed on blood and saliva-derived DNA from 414,817 individuals (365,918 blood-derived and 48,899 saliva-derived samples). Sequencing reads from libraries prepared with PCR Free Kapa HyperPrep library construction kit 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.

WGS from the Simons Foundation Powering Autism Research (SPARK) cohort of the Simons Foundation Autism Research Initiative (SFARI)[25] was performed on saliva-derived DNA from 12,519 individuals. Sequencing reads from libraries prepared with Illumina DNA PCR-Free Library Prep kit 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[72] (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).

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[55] using the TOPMed reference panel[73] (as the final UKB WGS genotype call set was not available at the time of analysis). For AoU and SPARK, we analyzed genotypes previously called from WGS using DRAGEN (AoU[24]) and DeepVariant (SPARK[25]).

Selection of viral reference genomes

We selected a panel of 31 viruses for which to profile viral DNA load (**Supplementary Table 1**) based on previous work that identified the most prevalent viruses observed in blood-derived DNA sequencing data from 8,240 individuals[9]. Additional viral reference genomes were included to more comprehensively represent viral families previously observed. For example, adenovirus was represented by Human adenovirus type 7 and Human adenovirus E (type 4) reference genomes, as these are two of the most commonly observed in adults[74], and polyomavirus was represented by a set of common types (1/BK, 2/JC, 3/KI, 4/WU, 5/MC, 6, 7)[75,76]. Undetected herpesviruses (HSV-2 and VZV) were added given the high general prevalence of EBV and HHV-7. 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[68] (**Extended Data Fig. 1a**).

The rationale for prioritizing a smaller set of viruses to profile (rather than attempting to comprehensively characterize viral diversity) was that our main goal was to identify effects of human genetics, age, sex, and environmental exposures (such as smoking) on viral DNA load, and we only had statistical power to detect such effects for commonly observed viruses. Working with a smaller set of common viruses allowed us to perform careful QC on WGS-based quantifications of viral DNA load, which was important given the potential for read alignment artifacts.

Measurement of viral DNA presence and abundance in WGS samples

In the UK Biobank (blood), All of Us (blood and saliva), and SPARK (saliva) WGS data sets, 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 (merged into a single reference for alignment) using BWA-MEM[77] (v0.7.18) with 4 threads (-t 4). We took this read-mapping-based approach following Moustafa et al.[9], who observed minimal additional detection of viral sequences from blood-derived WGS upon using de novo assembly followed by protein-based search.

After realignment, 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 aggregated across all samples in each data set (UKB, AoU blood, AoU saliva, and SPARK, each analyzed separately). To do so, for each sample, we used mosdepth[78] (v0.3.9) to compute depth-of-coverage in each 500bp window of each viral reference genome (--by 500); we skipped per-base depth output (-n) and mate overlap/CIGAR corrections (--fast-mode), restricted to reads passing default filters on SAM flags (-F 1796, which excludes duplicate reads), and filtered to reads with mapping quality ≥5 (-Q 5) for which their mates mapped to the same reference genome with an insert size in the range 100–1000bp (-l 100 -u 1000). The results of these filters were not sensitive to the choice of the mapping quality threshold; using a more stringent threshold (-Q 20) affected only a few percent of reads attributed to common viruses and had a negligible effect on downstream genetic association analyses.

Upon computing the 500bp-resolution coverage profile of each viral genome of each sample, we computed the coverage profile of each virus (i.e., for each 500bp bin, what fraction of samples had nonzero coverage) within each cohort (UKB, AoU blood, AoU saliva, and SPARK). For each virus, for each cohort, we flagged a subset of 500bp regions for exclusion based on having coverage exceeding the following threshold:

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

where $Q1$ and $Q3$ are the first and third quartiles of the distribution of alignment coverage across all 500bp regions of that virus' genome in that cohort. This expression corresponds to a lenient "Tukey fence," which is an outlier removal boundary defined based on adding a multiple of the interquartile range (Q3-Q1) to the third quartile (Q3). We used a lenient Tukey fence to retain regions with modest elevation of coverage (as such regions may still have a majority of alignments derived from the viral genome), and we added a constant offset of 5 to handle situations in which the first and third quartiles have the same value. Applying this bin-level filtering strategy per virus per cohort helped handle cohort-specific error modes of false positive viral alignments that might arise from heterogeneity in WGS data generation and processing (e.g., details of how reads had previously been aligned to the human reference genome, which impacted which reads did and did not map to human chromosomes).

For association analyses of viral DNA load with biological and clinical phenotypes, these measures of viral DNA 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 DNA load in blood, the number of viral read pairs mapping to each genome (obtained by summing mosdepth 500bp-bin depth values across non-excluded bins and multiplying by 500/300 = (bin size)/(bases per read pair)) 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 pair 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 (UKB or AoU for blood; AoU or SPARK for saliva) 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[4] with roughly 100 episomes per infected cell[67]. Assuming B-cells comprise roughly 5% of all white blood cells[79] 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; for such samples, 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 hour-of-day and month-of-year associations (**Fig. 2e,f** and **Extended Data Fig. 3k-n**) 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 top genetic principal components as covariates (20 PCs for UKB; 16 PCs for AoU). Associations were confirmed with Kronos (v1.0.0), which was run with default settings; to handle covariates, viral phenotypes were first adjusted for covariate effects estimated using harmonic regression.

Association of viral DNA prevalence with genetic ancestry

For UK Biobank, genetically-inferred ancestry was determined as described previously in ref.[80]. 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[24] and ancestry admixture estimates from Rye[81] 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 (n=5,225)[82] was first converted to VCF format with variants lifted over to hg19 and merged into multiallelic sites where appropriate. A small number of individuals (n=136) with >2 alleles for at least one multiallelic site were excluded from the reference panel. Imputation of HLA alleles onto phased SNP-array haplotypes[83] was done with BEAGLE[84] (v5.4) using default parameters, 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 DNA load phenotypes in UK Biobank

Abundances of reads aligning to reference genomes for EBV, HHV-6B, 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 (447,190 for HHV-6B after removal of individuals with endogenous HHV-6 integration). 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[85] (v2.5) to account for relatedness, using the following covariates: age, age squared, sex, genotype array, assessment center, and 20 genetic PCs. SNP-array genotypes were used for model-fitting, and LD scores derived from European-ancestry 1KGP samples were used for test statistic calibration. 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[69].

To identify index variants outside the MHC region (**Supplementary Table 7**), we first iteratively selected 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[86]. The less-strongly associated variant *i* was dropped if its approximate conditional association was no longer genome-wide significant ($p < 5 \times 10^{-8}$). Identified index variants were annotated with nearby genes using GENCODE 39 (ref.[87]) 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[88]. 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 8**).

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[89] (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, with the --remlNoRefine option set. 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[90] 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[91]. The specific burden masks analyzed here for association with viral DNA load used a minor allele frequency threshold of MAF<0.001 and included missense variants with PrimateAI-3D[92] scores >0.7 merged with loss-of-function SNPs, indels, and CNVs, using only the MANE Select transcript[93]. 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 5.7×10^{-7} (adjusting for five viruses included in these analyses: EBV, HHV-7, and three TTVs).

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

GWAS was performed on European-ancestry All of Us participants for the following phenotypes: inverse normal transformed EBV, HHV-6B, HHV-7, and anellovirus strains TUS01, VT416, and HD14a abundances in blood WGS (n=201,168 for EBV, 198,059 for HHV-6B, and 200,091 for

other viruses), EBV, HHV-6B, HHV-7, and MCPyV DNA positivity (i.e., presence of any EBV reads) in saliva WGS (n=33,164 for EBV, 32,711 for HHV-6B, and 33,050 for other viruses), and inverse normal transformed EBV (n=16,282) and HHV-7 (n=29,989) abundances in DNA-positive saliva WGS samples. Variants from the Allele Count/Allele Frequency (ACAF) threshold callset present in the TOPMed-r2 imputation panel (to exclude those in regions of poor mappability) 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 SNP-array genotypes (minor allele frequency >1%, missingness <10% in European-ancestry samples) as model SNPs, with the following covariates: age, age squared, sex, sequencing site, and 16 genetic PCs. For EBV GWAS, samples without a sex call of XX or XY were assigned indicator variables, whereas they were excluded from GWAS for other viruses due to a minor change in the analytical pipeline during the course of the project.

GWAS was also performed on African-ancestry All of Us participants for inverse normal transformed EBV and HHV-7 abundances in blood WGS (n=77,573). Variants from the Allele Count/Allele Frequency (ACAF) threshold callset with minor allele frequency >0.5% in gnomAD v4.1 African-ancestry samples were filtered to those with minor allele frequency >0.1% and allele count ≥40 in the subset of African ancestry samples used. BOLT-LMM (using the --lmmInfOnly flag, as the non-infinitesimal mixed model provided a negligible increase in statistical power) was run with SNP-array genotypes (minor allele frequency >1%, missingness <10% in African-ancestry samples) as model SNPs, with the following covariates: age, age squared, sex, sequencing site, and 16 genetic PCs.

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

As an auxiliary analysis, we also performed GWAS of viral DNA load phenotypes from SPARK saliva WGS. 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[69]. 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.

We observed that GWAS power in SPARK saliva WGS was much lower than in AoU saliva WGS, as expected given the much smaller sample size. We therefore chose not to meta-analyze saliva viral DNA load GWAS results across AoU and SPARK because the potential power gain was modest (given the much smaller size of the SPARK cohort) and might be negated by the age heterogeneity of SPARK (mostly children) versus AoU (only adults).

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

Associations with EBV, HHV-6B, HHV-7, and anellovirus strains TUS01, VT416, and HD14a DNA load in blood in All of Us were meta-analyzed with those from UK Biobank using METAL[94] (v2020-05-05) in standard error mode (SCHEME STDERR), restricting to variants with minor allele frequency greater than 0.1% (for EBV and HHV-7) or 1% (for HHV-6B and anelloviruses) in both cohorts (ADDFILTER MAF > 0.001 or 0.01), and applying genomic control correction within input studies (GENOMICCONTROL ON). 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 5**) was filtered.

To compute genetic correlation between viral phenotypes in UKB, AoU, and SPARK, LDSC[95] (v2.0.0) was run with standard settings and pairs of viral summary statistics as input.

Generation of lymphocyte percentage phenotype in All of Us

A lymphocyte percentage phenotype was generated from the “Lymphocytes/100 leukocytes in Blood by Automated count” phenotype (OMOP Concept Id: 3037511). Only entries with "percent", "percent of white blood cells", "Percent", and "Percentage unit" as units were kept, discarding entries with other or missing units as well as values outside the range [0,100]. For individuals with multiple valid measurements, we took the median value. This left 170,196 people with lymphocyte percentage values, among whom 24,789 African-ancestry individuals had blood-derived WGS available and were used to evaluate the extent to which the Duffy-null effects on EBV and HHV-7 DNA load were mediated by effects on lymphocyte percentage.

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 and using 4 computational threads). Read alignments were filtered to those for which both the read and its mate were mapped (samtools view -F 12) and were then collated within 500bp bins of each of the two reference genomes using mosdepth with the same parameters as above, with the exception that the filter on insert size (-l 100 -u 1000) was dropped, as insert size was undefined in situations in which a read mapped to EBV type 1 and its mate to EBV type 2 or vice versa (e.g., if only one read in the pair fell within a type-informative region, and its non-type-informative mate was mapped arbitrarily to type 1 or type 2). 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 the numbers of type-1-specific and type-2-specific reads for each sample were computed by summing 500bp depths

across these regions (separately for the two EBV genomes), multiplying by 500/150 = (bin size)/(bases per read), and rounding to the nearest integer. In UKB, samples with at least two type-1-specific reads were designated as positive for EBV type 1, and analogously for type 2.

To evaluate the risk of misclassification between type 1 and type 2, we examined the distribution of type 2 vs. type 1 reads in SPARK saliva samples, making use of the fact that saliva samples frequently have hundreds of EBV reads that should typically come from only one EBV type (depending on whether the individual is infected with a type 1 or type 2 EBV strain). This analysis showed that as expected, nearly all samples had read counts heavily skewed to either type 1 or type 2 (typically >99% type 1 or >99% type 2), indicating a low rate of misclassification of reads.

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 determined to be EBV type 2-positive to the total number of samples determined to be either EBV type 1-positive or type 2-positive. For country-level analyses, Fields 1647 and 20115 were used.

Genetic association analyses of MHC variants with EBV type DNA load phenotypes

Variants in the MHC region of the human genome (including imputed HLA alleles) were tested for association with three binary EBV DNA load phenotypes derived from EBV type-specific read alignments. The first two phenotypes coded EBV type 1-positive individuals (respectively, type 2-positive individuals) as cases and all other individuals as controls. The third phenotype, used in case-case association analyses of EBV type 1- versus type 2-positivity, coded individuals who were type 2-positive and lacked type 1-specific alignments as cases (n=1,366), and those who were type 1-positive and lacked type 2-specific alignments as controls (n=9,817); a small number of individuals determined to be positive for both type 1 and type 2 were excluded (n=126). 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 DNA 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). Bonferroni correction was applied to the full set of tests performed across all eight viruses. Association tests were performed using logistic regression with Firth correction using the Wald approximation (pl=FALSE) 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, and Bonferroni correction was applied considering all pairwise tests of quantitative blood phenotypes with viral DNA positivity phenotypes.

Binary immunosuppressive drug phenotypes were generated by aggregating synonymous terms from self-reported medication data (Category 100075). Specifically, methotrexate was generated from the union of individuals reporting “methotrexate” or “mtx – methotrexate”; cyclosporin from “cya - cyclosporin”, “ciclosporin product”, “csa - cyclosporin a”, “cya - cyclosporin a”, “cyclosporin”, “cyclosporin product”, or “ciclosporin”; and corticosteroids from “prednisone”, “prednisolone”, “methylprednisolone”, “prednisolone product”, “dexamethasone”, “fludrocortisone”, “hydrocortisone”, “cortisone product”, “hydrocortisone product”, or “cortisone”.

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 DNA 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 DNA load in UKB and AoU. The MHC region was excluded from MR analyses given the likelihood of linkage disequilibrium generating pleiotropic effects of MHC haplotypes on many immune-related phenotypes. To minimize the impact of sample overlap between cohorts used in the exposure GWAS (EBV DNA load in UKB+AoU blood WGS) and outcome GWAS (disease phenotypes in FinnGen+UKB+MVP), we regenerated GWAS summary statistics for EBV DNA load in UKB after restricting to a control-only sub-cohort[96]. 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.

GWAS summary statistics for outcome phenotypes of interest were downloaded from https://mvp-ukbb.finngen.fi/ after applying for access. In these GWAS, FinnGen phenotypes had been harmonized over ICD-8, -9 and -10, cancer-specific ICD-O-3, (NOMESCO) procedure codes, Finnish-specific Social Insurance Institute (KELA) drug reimbursement codes and ATC-codes collected from various registries. MVP phenotypes had been defined by ICD-9 and -10 codes from electronic health records grouped into corresponding phecodes, with case status defined as having two or more phecode-mapped ICD-9 or -10 codes. Meta-analyses had been performed by identifying phenotypes with concordant endpoints, as described at https://finngen.gitbook.io/documentation/methods/meta-analysis. We used the MendelianRandomization R package[97] (v0.10.0) with default settings to generate estimates of causal effect sizes of EBV DNA 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[98].

We caution that effect size estimates from Mendelian randomization (here, the increase in log-odds of disease per s.d. increase in EBV reads; **Fig. 5e**) are expected to be overestimated when the exposure variable (here, EBV DNA load) is measured with high noise. For example, were the exposure phenotype to be randomly permuted in a subset of samples, this would leave the s.d. of the exposure unchanged, but it would shrink down the effect sizes (betas) of the instrumental variables. The betas of IVs are used as independent variables in the regression analysis used to estimate MR effect sizes (x-axis of **Fig. 5f**), such that shrinking the betas of the IVs would then increase the slope of the MR regression, causing the estimated effect size from MR to increase. This behavior contrasts with how measurement noise in the exposure variable impacts direct analyses of association with the outcome variable (**Fig. 5g**): in such analyses, noise reduces the observed association.

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/), METAL (v2020-05-05, https://genome.sph.umich.edu/wiki/METAL), 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/), MVP-Finngen-UKBB meta-analysis summary statistics (https://mvp-ukbb.finngen.fi/), and T1DGC HLA imputation panel (https://repository.niddk.nih.gov/study/173). The following data resources are publicly available: human reference genome build GRCh38 (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/), TOPMed-r2 imputation panel variant list (https://imputation.biodatacatalyst.nhlbi.nih.gov/), gnomAD v4.1 variant call set (https://gnomad.broadinstitute.org/), LD score resources https://alkesgroup.broadinstitute.org/LDSCORE/), NCBI Virus for reference sequences (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/), PrimateAI-3D scores (https://primateai3d.basespace.illumina.com/), GENCODE 39 definitions (https://www.gencodegenes.org/), and GTEx expression and splice quantitative trait associations (https://gtexportal.org/home/). Full viral DNA load GWAS summary statistics are available from the GWAS Catalog under accessions GCST90809801 to GCST90809829.

Acknowledgments

We thank C. Usher for edits to text and figures. This research was conducted using the UK Biobank Resource under application number 40709. We are grateful to all of the families in SPARK, the SPARK clinical sites and SPARK staff. We appreciate obtaining access to SPARK genetic data on SFARI Base. 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.

Competing interests

The authors declare no competing interests.

Figure 1

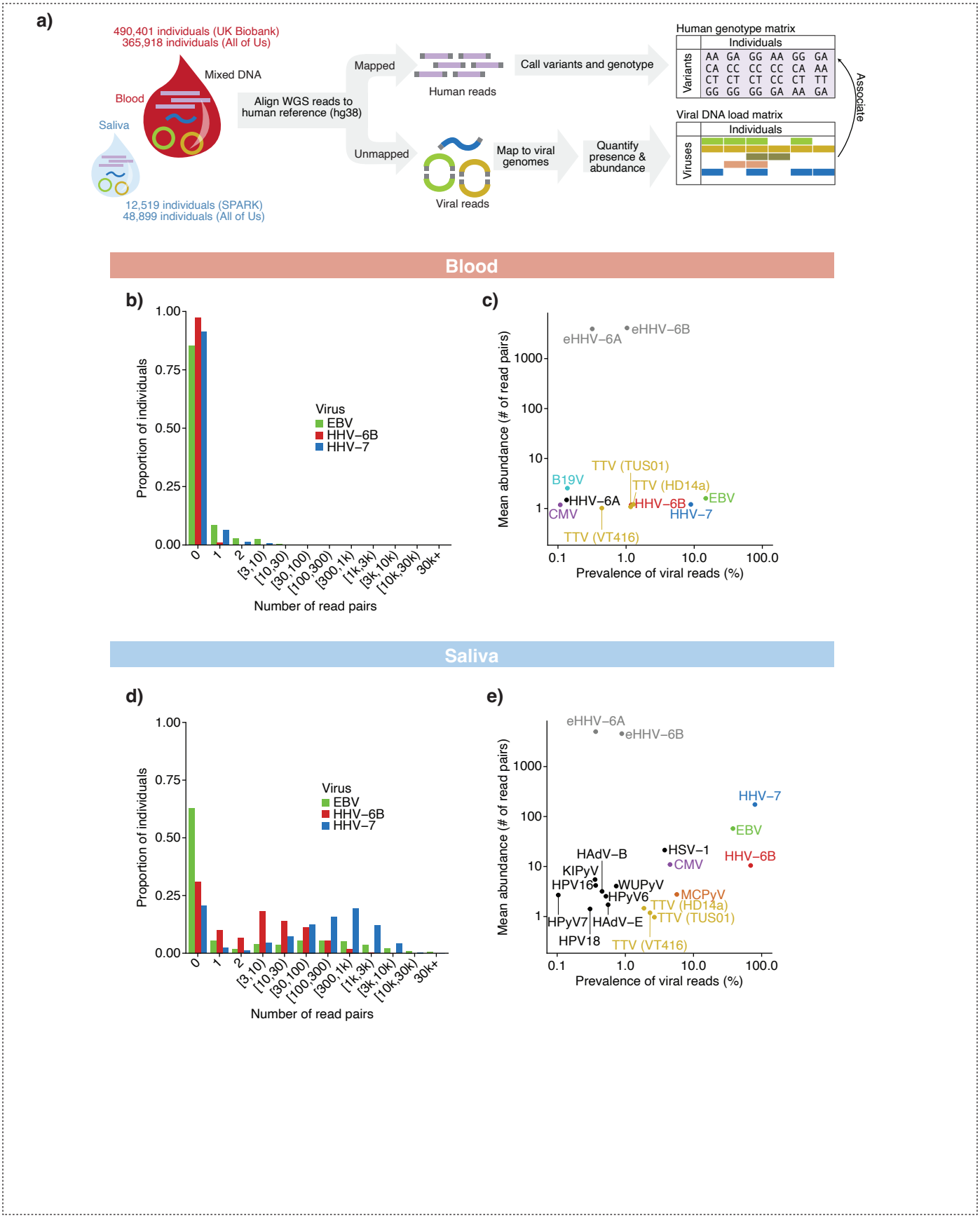

a)
490,401 individuals (UK Biobank)
365,918 individuals (All of Us)
Mixed DNA
Blood
Saliva
12,519 individuals (SPARK)
48,899 individuals (All of Us)
Align WGS reads to human reference (hg38)
Mapped
Unmapped
Human reads
Viral reads
Call variants and genotype
Map to viral genomes
Quantify presence & abundance
Human genotype matrix
Individuals
Variants
Viral DNA load matrix
Viruses
Associate
Blood
b)
Proportion of individuals
Virus
EBV
HHV−6B
HHV−7
Number of read pairs
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
Virus
EBV
HHV−6B
HHV−7
Number of read pairs
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

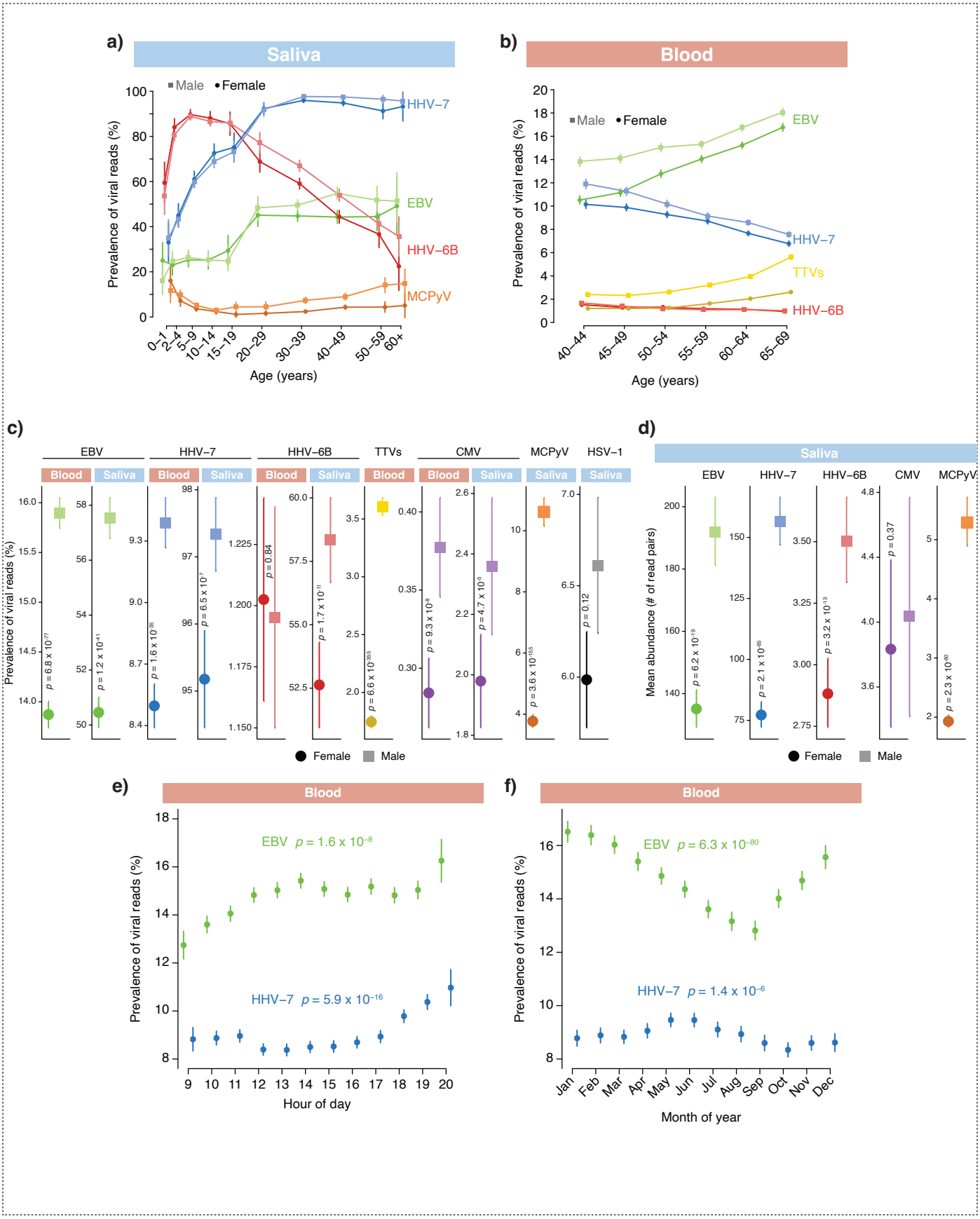

a)
Saliva
Male
Female
Prevalence of viral reads (%)
HHV−7
EBV
HHV−6B
MCPyV
0–1
2–4
5–9
10–14
15–19
20–29
30–39
40–49
50–59
60+
Age (years)
b)
Blood
Male
Female
Prevalence of viral reads (%)
EBV
HHV−7
TTVs
HHV−6B
40–44
45–49
50–54
55–59
60–64
65–69
Age (years)
c)
EBV
HHV−7
HHV−6B
TTVs
CMV
MCPyV
HSV−1
Blood
Saliva
Prevalence of viral reads (%)
p = 6.8 x 10−77
p = 1.2 x 10−41
p = 1.6 x 10−35
p = 6.5 x 10−7
p = 0.84
p = 1.7 x 10−11
p = 6.6 x 10−355
p = 9.3 x 10−8
p = 4.7 x 10−5
p = 3.6 x 10−155
p = 0.12
Female
Male
d)
Saliva
EBV
HHV−7
HHV−6B
CMV
MCPyV
Mean abundance (# of read pairs)
p = 6.2 x 10−19
p = 2.1 x 10−65
p = 3.2 x 10−13
p = 0.37
p = 2.3 x 10−60
Female
Male
e)
Blood
Prevalence of viral reads (%)
EBV p = 1.6 x 10−8
HHV−7 p = 5.9 x 10−16
Hour of day
f)
Blood
Prevalence of viral reads (%)
EBV p = 6.3 x 10−80
HHV−7 p = 1.4 x 10−6
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month of year

Figure 3

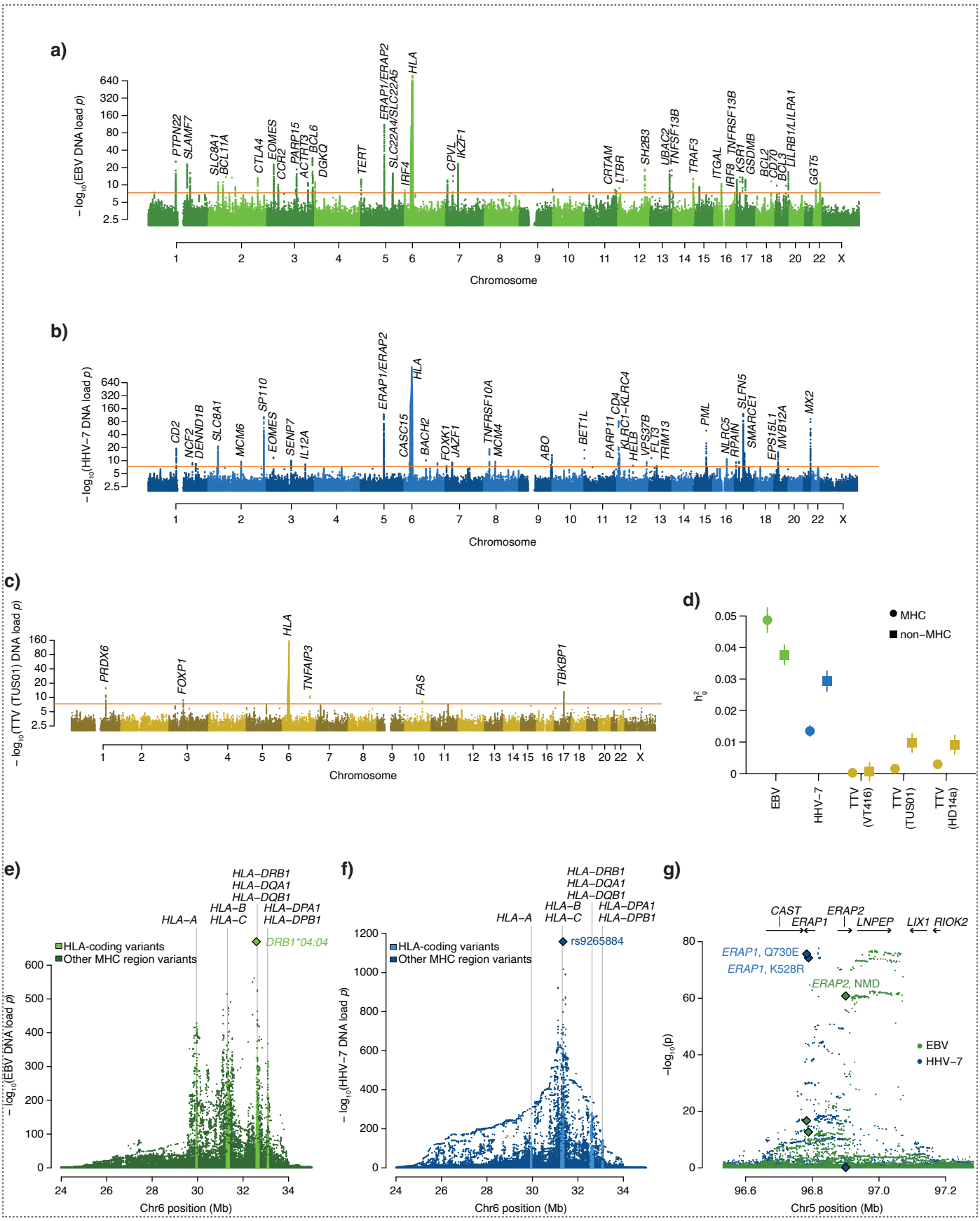

a)
– log10(EBV DNA load p)
PTPN22
SLAMF7
SLC8A1
BCL11A
CTLA4
EOMES
CCR2
PARP15
ACTRT3
BCL6
DGKQ
TERT
ERAP1/ERAP2
SLC22A4/SLC22A5
IRF4
HLA
CPVL
IKZF1
CRTAM
LTBR
SH2B3
UBAC2
TNFSF13B
TRAF3
ITGAL
IRF8
TNFRSF13B
KSR1
GSDMB
BCL2
CD70
BCL3
LILRB1/LILRA1
GGT5
Chromosome
b)
– log10(HHV–7 DNA load p)
CD2
NCF2
DENND1B
SLC8A1
MCM6
SP110
EOMES
SENP7
IL12A
ERAP1/ERAP2
CASC15
HLA
BACH2
FOXK1
JAZF1
TNFRSF10A
MCM4
ABO
BET1L
PARP11
CD4
KLRC1–KLRC4
HELB
VPS37B
FLT3
TRIM13
PML
NLRC5
RPAIN
SLFN5
SMARCE1
EPS15L1
MVB12A
MX2
Chromosome
c)
– log10(TTV (TUS01) DNA load p)
PRDX6
FOXP1
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 DNA 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 DNA 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

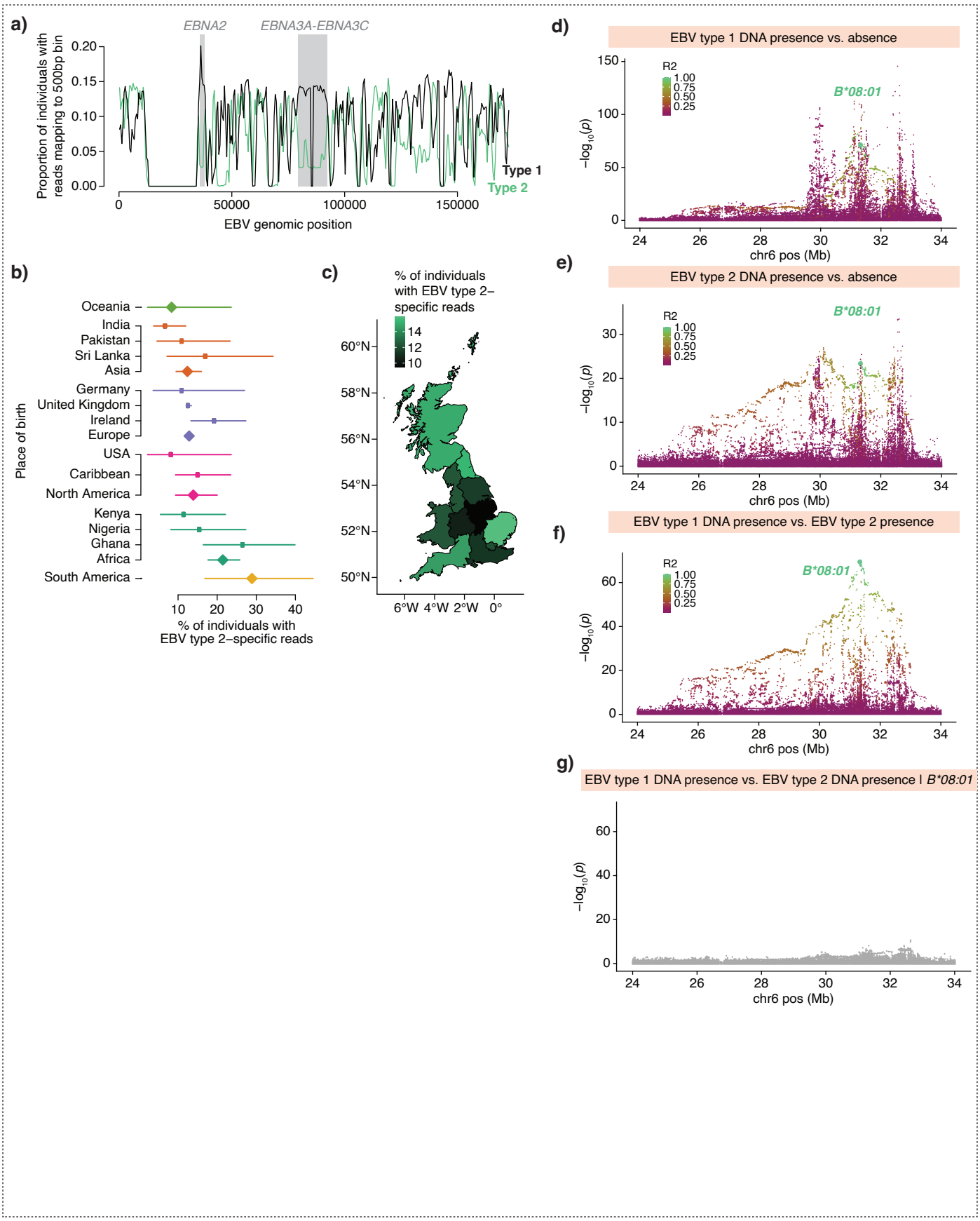
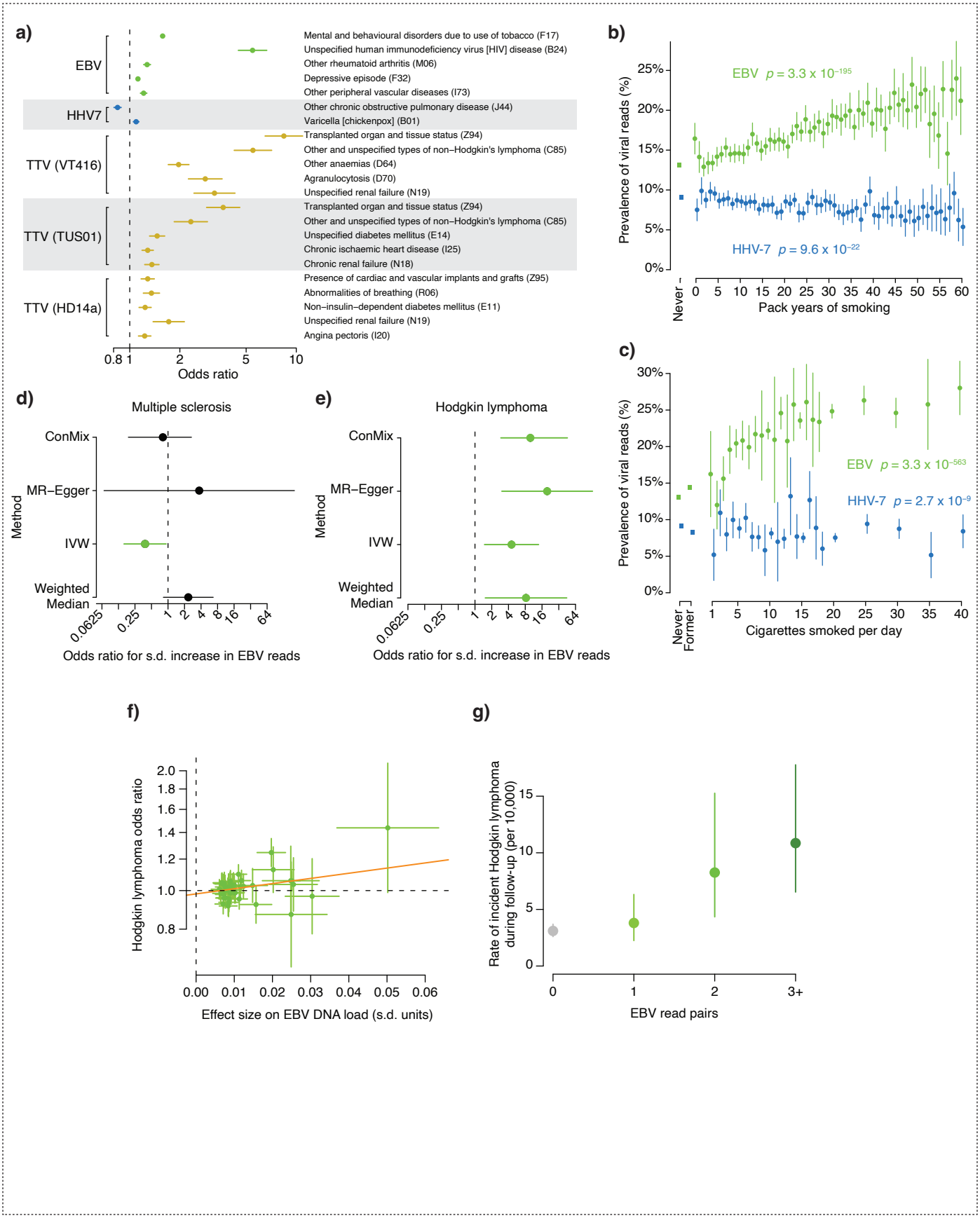
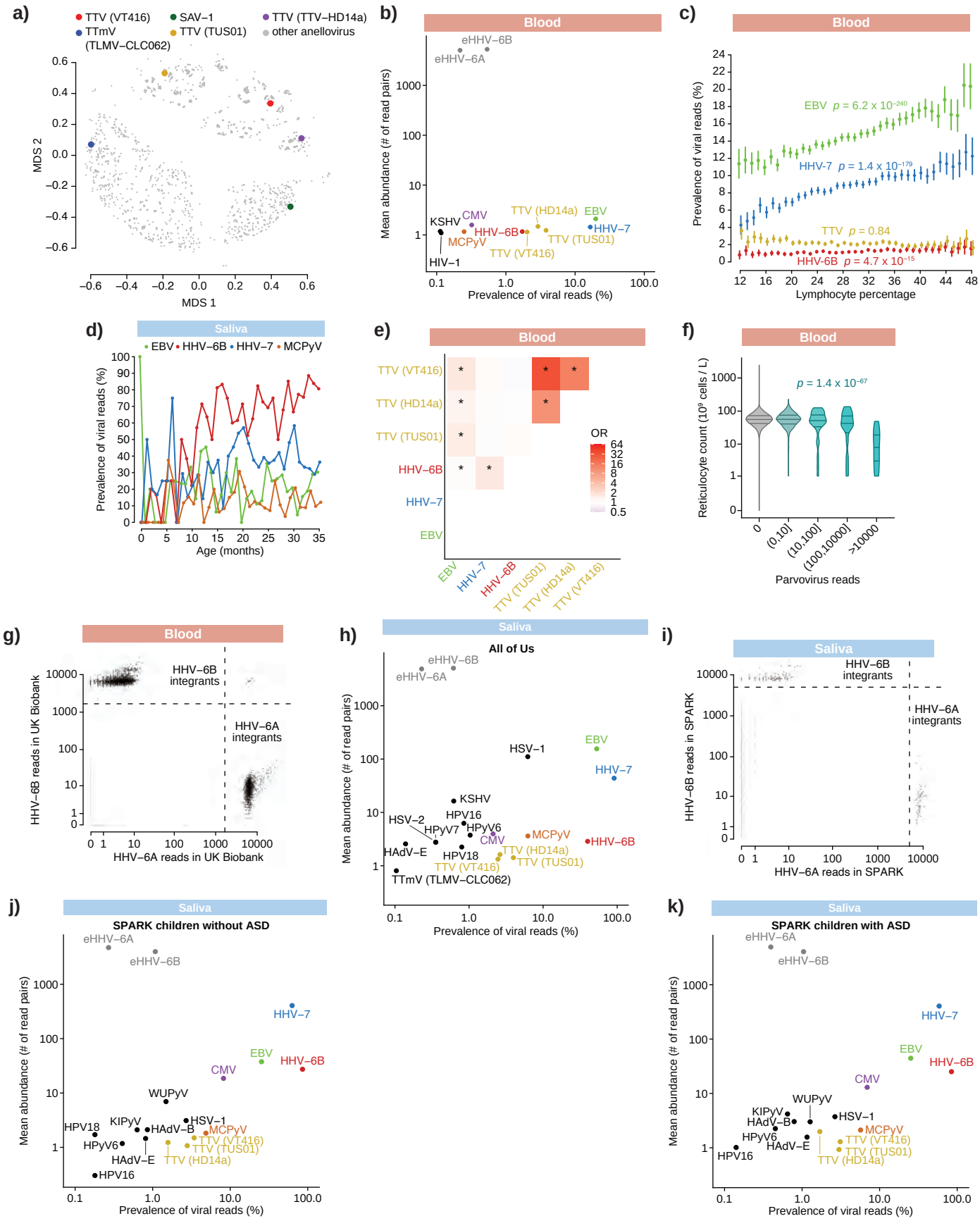
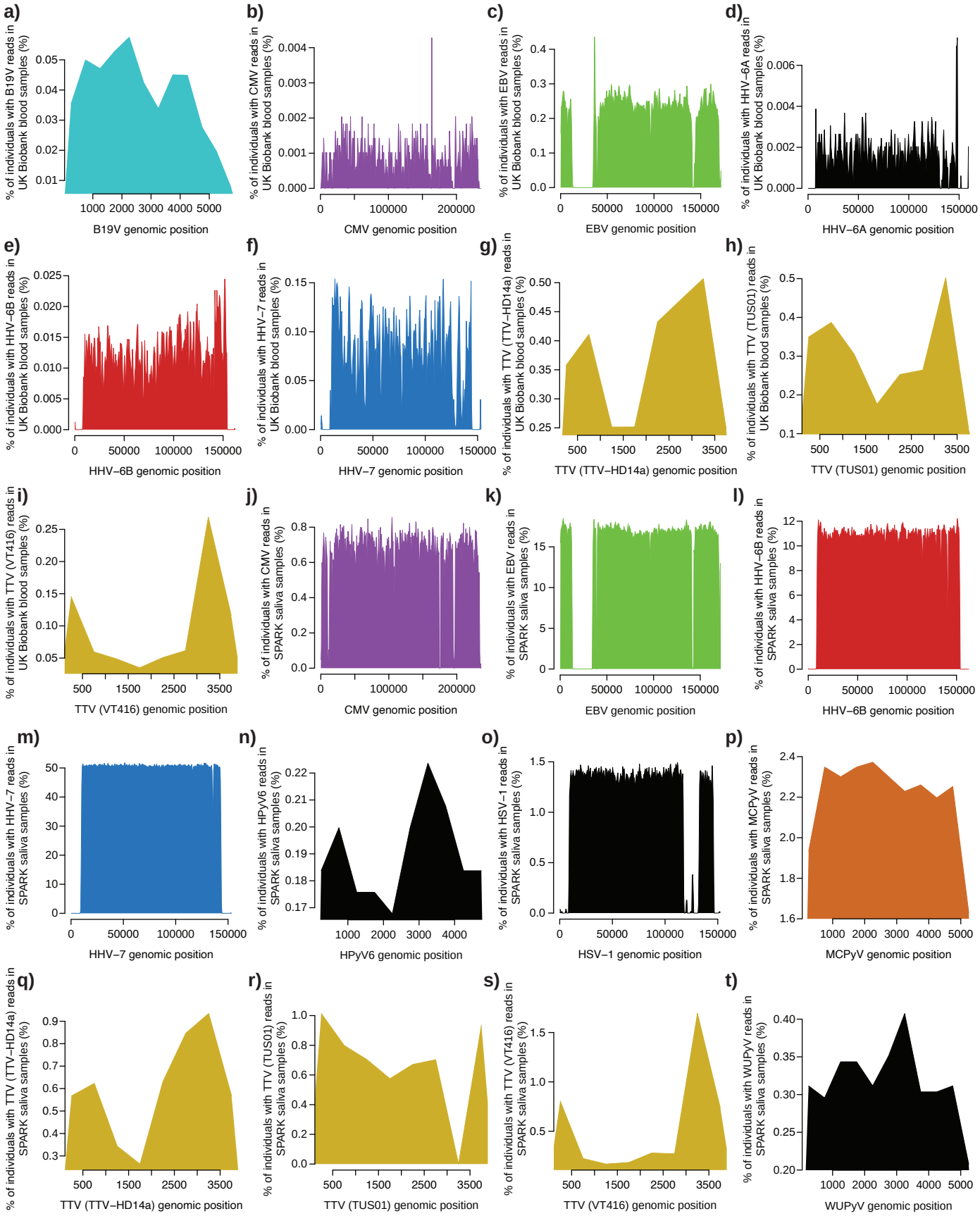
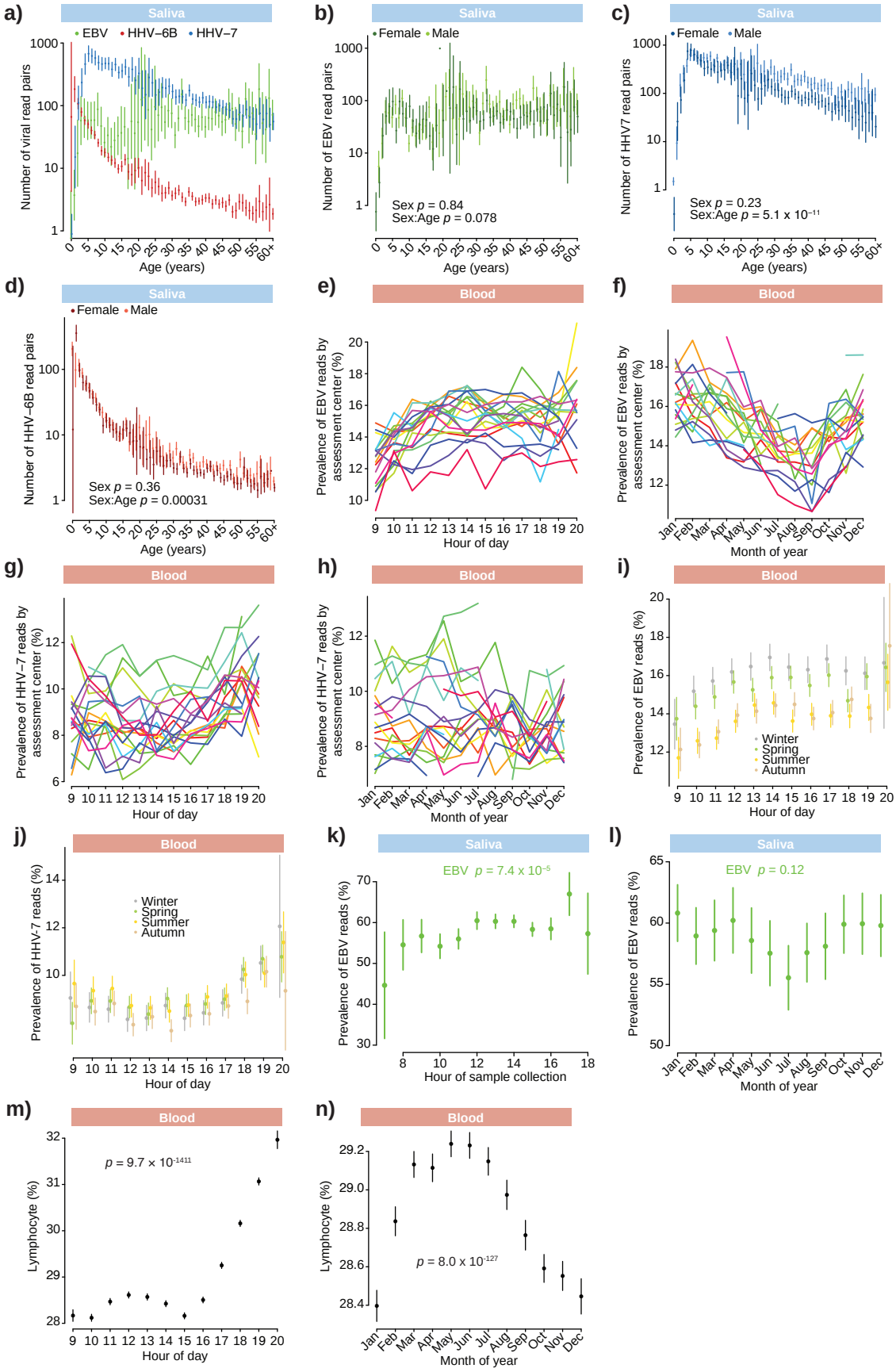

a)
EBNA2
EBNA3A-EBNA3C
Proportion of individuals with reads mapping to 500bp bin
0.20
0.15
0.10
0.05
0.00
0
50000
100000
150000
EBV genomic position
Type 1
Type 2
b)
Place of birth
Oceania
India
Pakistan
Sri Lanka
Asia
Germany
United Kingdom
Ireland
Europe
USA
Caribbean
North America
Kenya
Nigeria
Ghana
Africa
South America
10
20
30
40
% of individuals with EBV type 2–specific reads
c)
% of individuals with EBV type 2–specific reads
14
12
10
60°N
58°N
56°N
54°N
52°N
50°N
6°W
4°W
2°W
0°
d)
EBV type 1 DNA presence vs. absence
R2
1.00
0.75
0.50
0.25
B*08:01
−log10(p)
150
100
50
0
24
26
28
30
32
34
chr6 pos (Mb)
e)
EBV type 2 DNA presence vs. absence
R2
1.00
0.75
0.50
0.25
B*08:01
−log10(p)
30
20
10
0
24
26
28
30
32
34
chr6 pos (Mb)
f)
EBV type 1 DNA presence vs. EBV type 2 presence
R2
1.00
0.75
0.50
0.25
B*08:01
−log10(p)
60
40
20
0
24
26
28
30
32
34
chr6 pos (Mb)
g)
EBV type 1 DNA presence vs. EBV type 2 DNA presence | B*08:01
−log10(p)
60
40
20
0
24
26
28
30
32
34
chr6 pos (Mb)

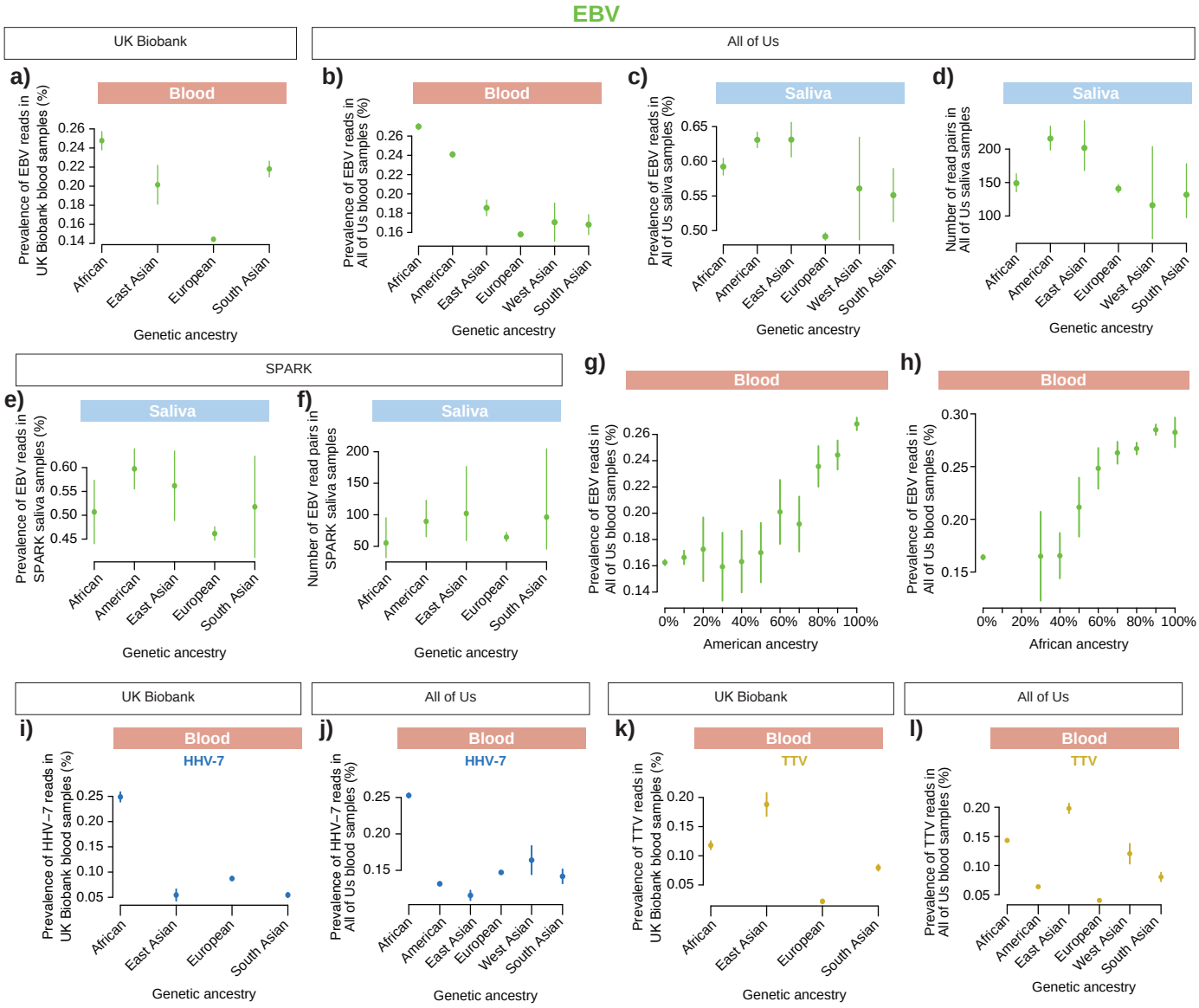
Figure 5


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
ConMix
MR−Egger
IVW
Weighted Median
Method
0.0625 0.25 1 2 4 8 16 64
Odds ratio for s.d. increase in EBV reads
e)
Hodgkin lymphoma
ConMix
MR−Egger
IVW
Weighted Median
Method
0.0625 0.25 1 2 4 8 16 64
Odds ratio for s.d. increase in EBV reads
f)
Hodgkin lymphoma odds ratio
2.0 1.8 1.6 1.4 1.2 1.0 0.8
0.00 0.01 0.02 0.03 0.04 0.05 0.06
Effect size on EBV DNA 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

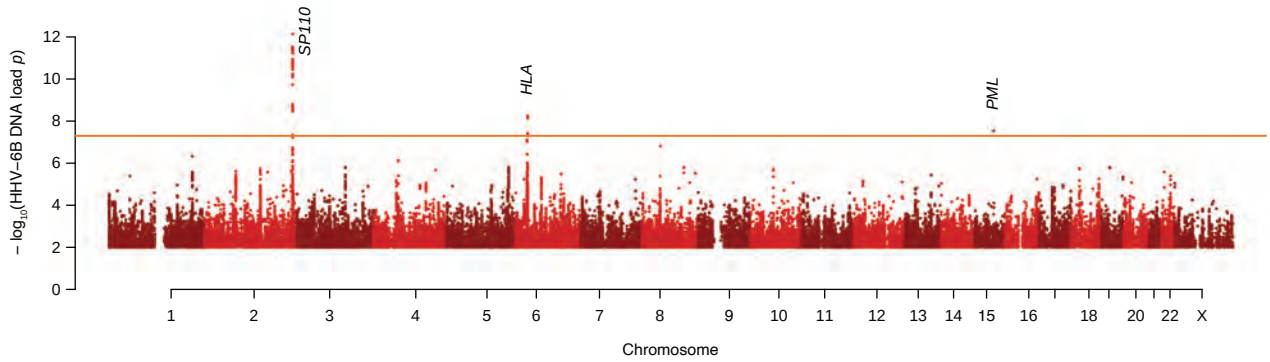

a)
TTV (VT416)
SAV−1
TTV (TTV−HD14a)
TTmV (TLMV−CLC062)
TTV (TUS01)
other anellovirus
MDS 1
MDS 2
b)
Blood
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
eHHV−6B
eHHV−6A
KSHV
CMV
HHV−6B
MCPyV
HIV−1
TTV (HD14a)
EBV
HHV−7
TTV (TUS01)
TTV (VT416)
c)
Blood
Prevalence of viral reads (%)
Lymphocyte percentage
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
d)
Saliva
EBV
HHV−6B
HHV−7
MCPyV
Prevalence of viral reads (%)
Age (months)
e)
Blood
TTV (VT416)
TTV (HD14a)
TTV (TUS01)
HHV−6B
HHV−7
EBV
OR
f)
Blood
Reticulocyte count (10⁹ cells / L)
p = 1.4 x 10−67
0
(0,10]
(10,100]
(100,10000]
>10000
Parvovirus reads
g)
Blood
HHV−6B reads in UK Biobank
HHV−6A reads in UK Biobank
HHV−6B integrants
HHV−6A integrants
h)
Saliva
All of Us
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
eHHV−6B
eHHV−6A
HSV−1
EBV
HHV−7
KSHV
HPV16
HSV−2
HPyV7
HPyV6
CMV
MCPyV
HHV−6B
HAdV−E
HPV18
TTV (HD14a)
TTV (TUS01)
TTV (VT416)
TTmV (TLMV−CLC062)
i)
Saliva
HHV−6B reads in SPARK
HHV−6A reads in SPARK
HHV−6B integrants
HHV−6A integrants
j)
Saliva
SPARK children without ASD
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
eHHV−6A
eHHV−6B
HHV−7
EBV
HHV−6B
CMV
WUPyV
HPV18
KIPyV
HSV−1
HAdV−B
MCPyV
TTV (VT416)
HPyV6
TTV (TUS01)
HAdV−E
TTV (HD14a)
HPV16
k)
Saliva
SPARK children with ASD
Mean abundance (# of read pairs)
Prevalence of viral reads (%)
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)


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

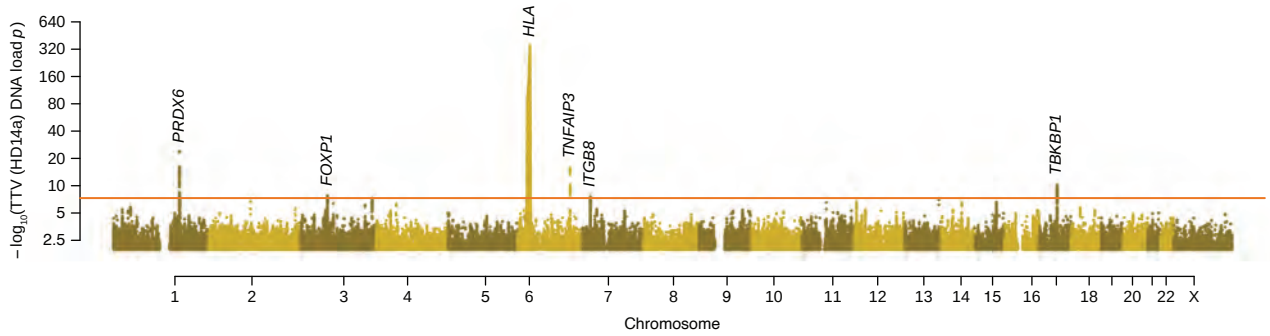

a)
Saliva
EBV
HHV−6B
HHV−7
Number of viral read pairs
Age (years)
b)
Saliva
Female
Male
Number of EBV read pairs
Sex p = 0.84
Sex:Age p = 0.078
Age (years)
c)
Saliva
Female
Male
Number of HHV7 read pairs
Sex p = 0.23
Sex:Age p = 5.1 x 10−11
Age (years)
d)
Saliva
Female
Male
Number of HHV−6B read pairs
Sex p = 0.36
Sex:Age p = 0.00031
Age (years)
e)
Blood
Prevalence of EBV reads by assessment center (%)
Hour of day
f)
Blood
Prevalence of EBV reads by assessment center (%)
Month of year
g)
Blood
Prevalence of HHV−7 reads by assessment center (%)
Hour of day
h)
Blood
Prevalence of HHV−7 reads by assessment center (%)
Month of year
i)
Blood
Prevalence of EBV reads (%)
Winter
Spring
Summer
Autumn
Hour of day
j)
Blood
Prevalence of HHV-7 reads (%)
Winter
Spring
Summer
Autumn
Hour of day
k)
Saliva
EBV p = 7.4 x 10−5
Prevalence of EBV reads (%)
Hour of sample collection
l)
Saliva
EBV p = 0.12
Prevalence of EBV reads (%)
Month of year
m)
Blood
p = 9.7 x 10-1411
Lymphocyte (%)
Hour of day
n)
Blood
p = 8.0 x 10-127
Lymphocyte (%)
Month of year


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 (%)
0.26
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
150
100
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 (%)
0.26
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

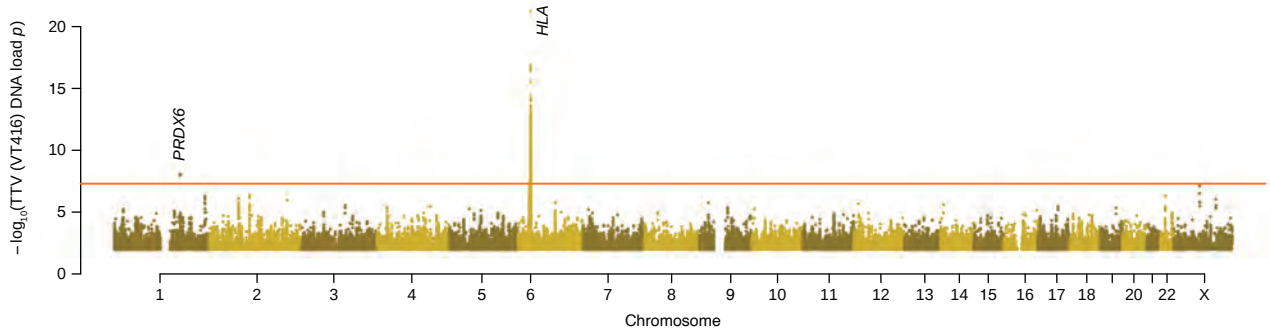
a)

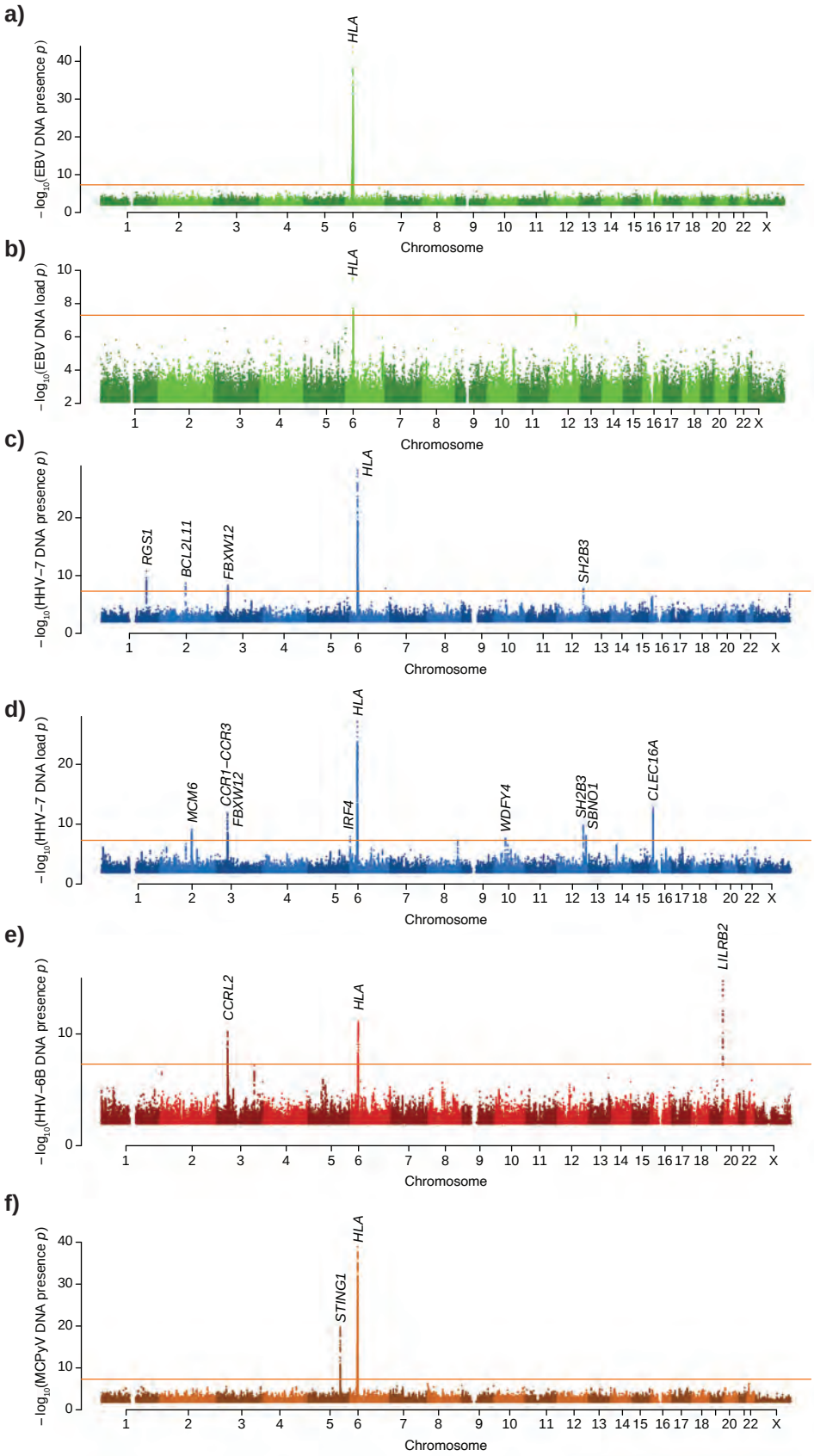

SP110
HLA
PML
– log10(HHV–6B DNA load p)
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 X
Chromosome

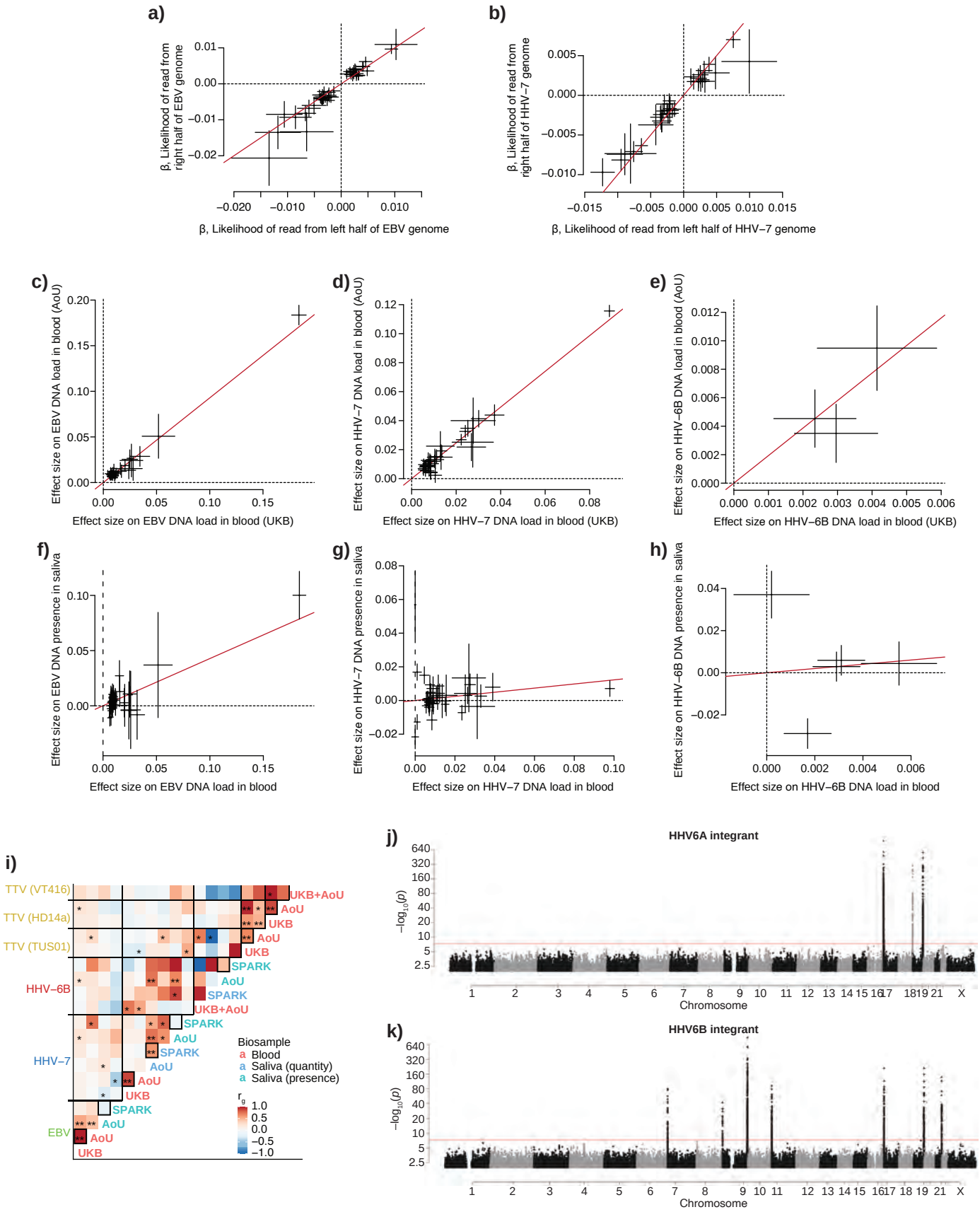
b)

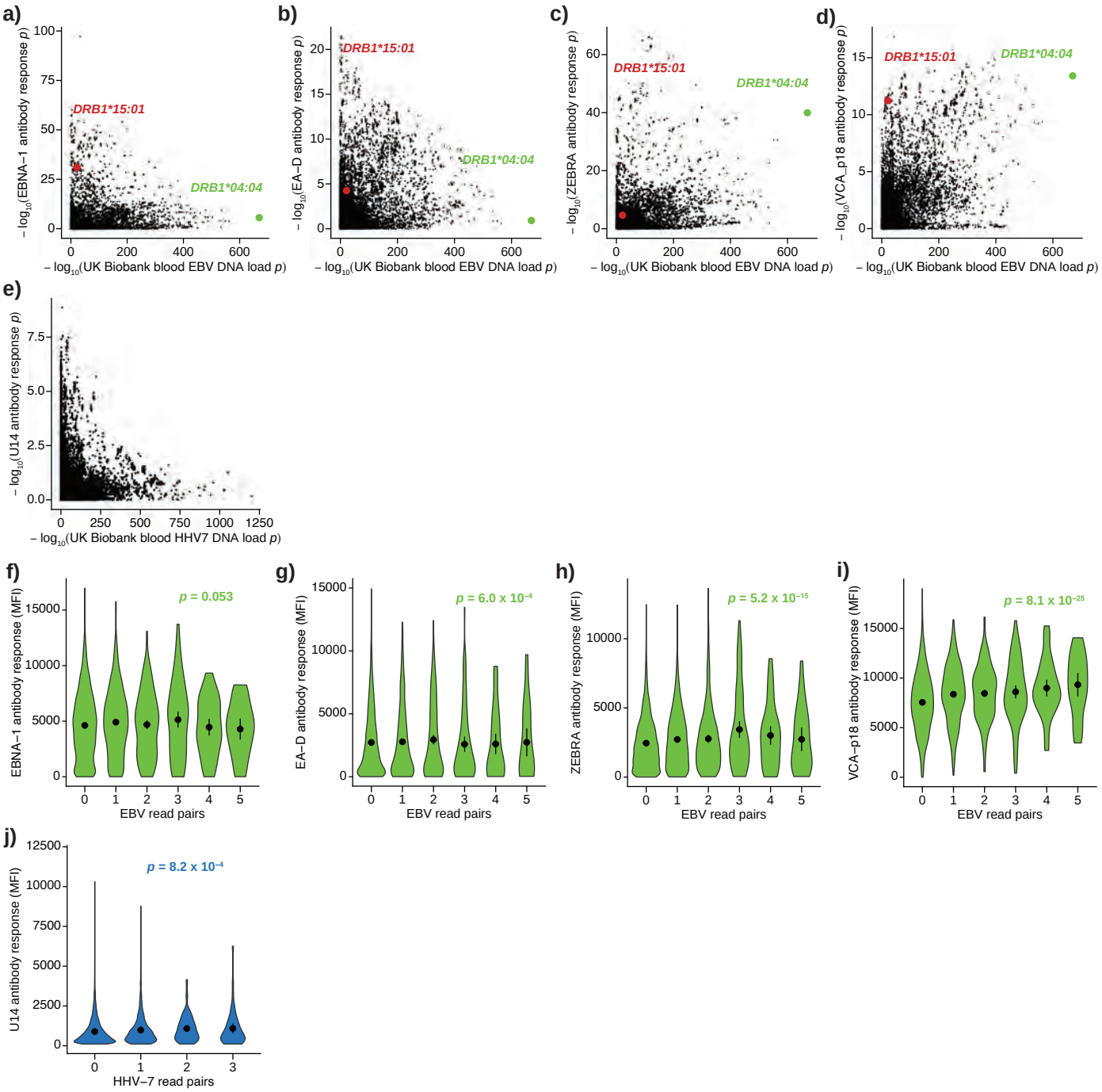

HLA
PRDX6
FOXP1
TNFAIP3
ITGB8
TBKBP1
– log10(TTV (HD14a) DNA load p)
640
320
160
80
40
20
10
5
2.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 X
Chromosome

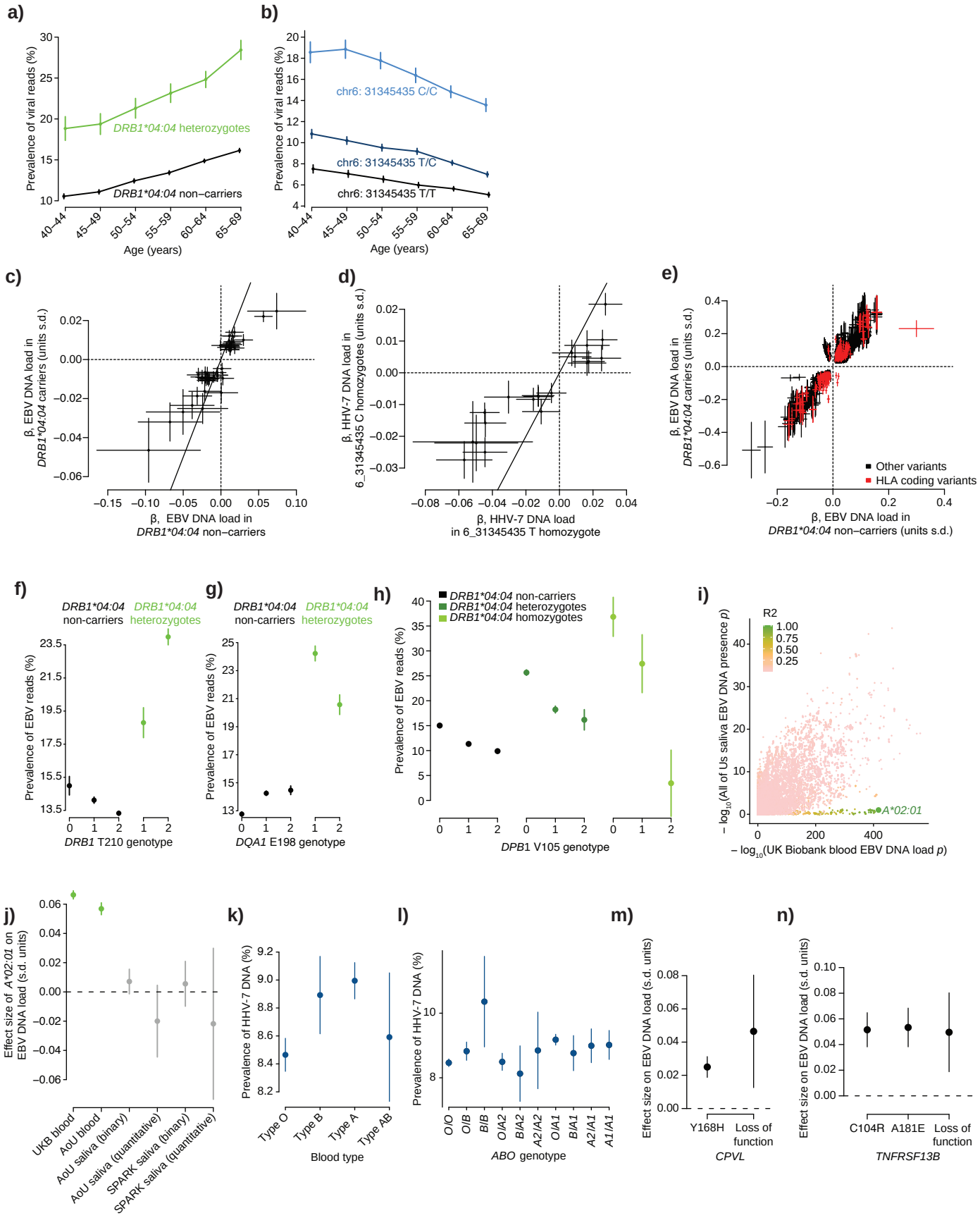
c)

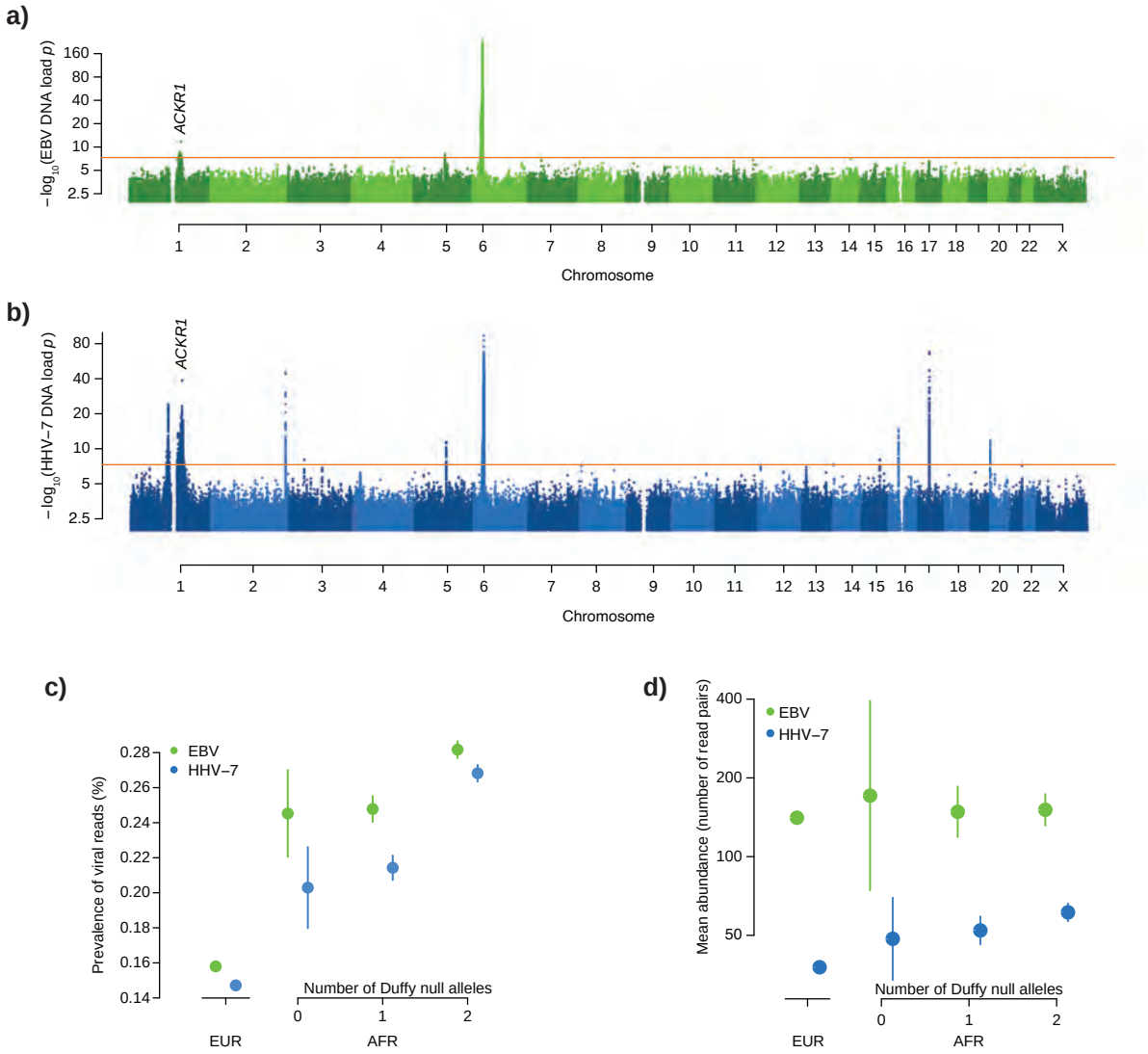

HLA
PRDX6
– log10(TTV (VT416) DNA load p)
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 22 X
Chromosome

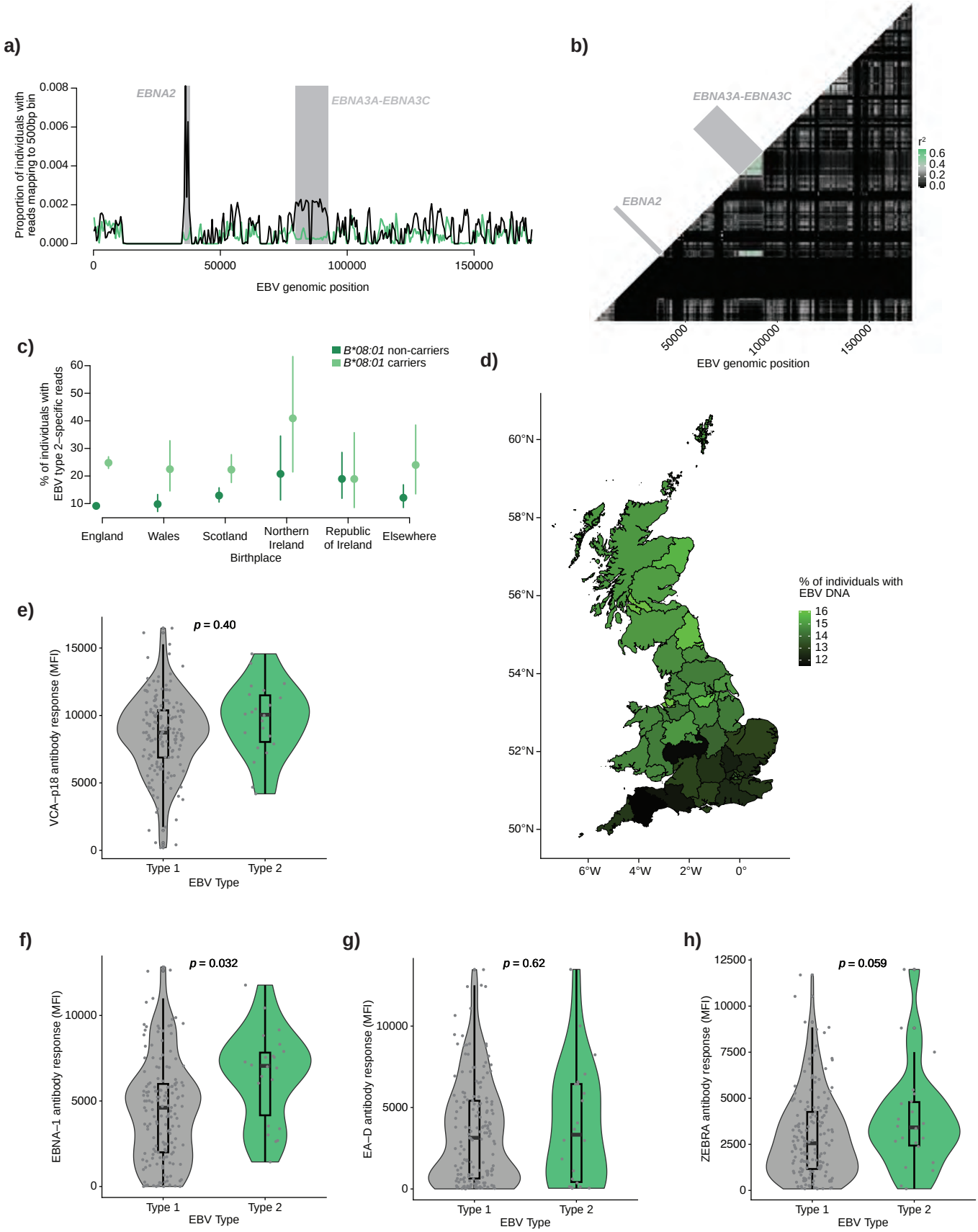

a)
HLA
– log10(EBV DNA presence p)
Chromosome
b)
HLA
– log10(EBV DNA load p)
Chromosome
c)
HLA
RGS1
BCL2L11
FBXW12
SH2B3
– log10(HHV–7 DNA presence p)
Chromosome
d)
HLA
MCM6
CCR1–CCR3
FBXW12
IRF4
WDFY4
SH2B3
SBNO1
CLEC16A
– log10(HHV–7 DNA load p)
Chromosome
e)
CCRL2
HLA
LILRB2
– log10(HHV–6B DNA presence p)
Chromosome
f)
HLA
STING1
– log10(MCPyV DNA presence p)
Chromosome

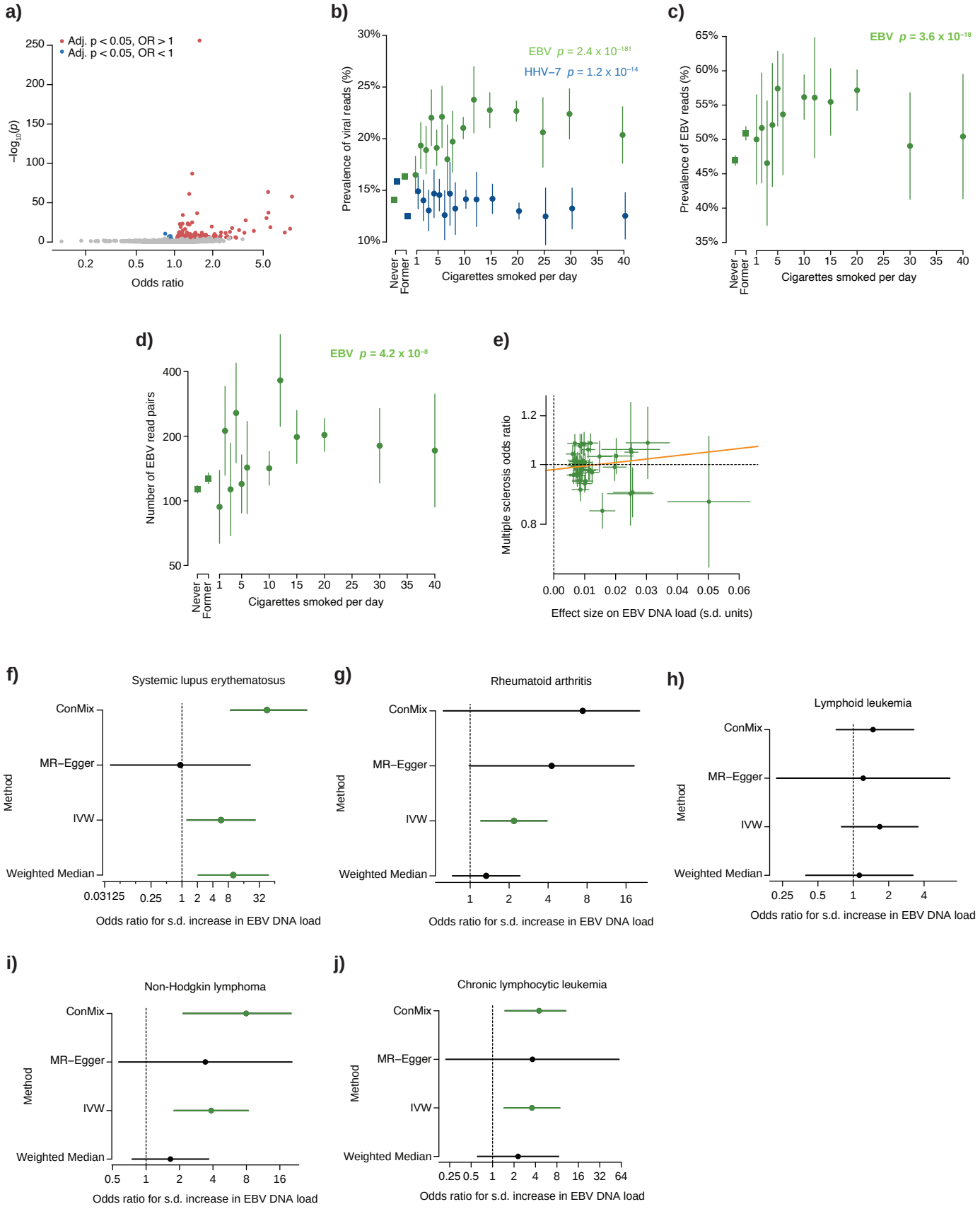

a)
β, Likelihood of read from right half of EBV genome
β, Likelihood of read from left half of EBV genome
b)
β, Likelihood of read from right half of HHV−7 genome
β, Likelihood of read from left half of HHV−7 genome
c)
Effect size on EBV DNA load in blood (AoU)
Effect size on EBV DNA load in blood (UKB)
d)
Effect size on HHV−7 DNA load in blood (AoU)
Effect size on HHV−7 DNA load in blood (UKB)
e)
Effect size on HHV−6B DNA load in blood (AoU)
Effect size on HHV−6B DNA load in blood (UKB)
f)
Effect size on EBV DNA presence in saliva
Effect size on EBV DNA load in blood
g)
Effect size on HHV−7 DNA presence in saliva
Effect size on HHV−7 DNA load in blood
h)
Effect size on HHV−6B DNA presence in saliva
Effect size on HHV−6B DNA load in blood
i)
TTV (VT416)
TTV (HD14a)
TTV (TUS01)
HHV−6B
HHV−7
EBV
UKB+AoU
AoU
UKB
SPARK
Biosample
a Blood
a Saliva (quantity)
a Saliva (presence)
r_g
1.0
0.5
0.0
−0.5
−1.0
j)
HHV6A integrant
$-\log_{10}(p)$
Chromosome
k)
HHV6B integrant
$-\log_{10}(p)$
Chromosome


a)
b)
c)
d)
e)
DRB1*15:01
DRB1*04:04
– log10(EBNA–1 antibody response p)
– log10(EA–D antibody response p)
– log10(ZEBRA antibody response p)
– log10(VCA_p18 antibody response p)
– log10(U14 antibody response p)
– log10(UK Biobank blood EBV DNA load p)
– log10(UK Biobank blood HHV7 DNA load p)
f)
g)
h)
i)
j)
p = 0.053
p = 6.0 x 10–4
p = 5.2 x 10–15
p = 8.1 x 10–25
p = 8.2 x 10–4
EBNA–1 antibody response (MFI)
EA–D antibody response (MFI)
ZEBRA antibody response (MFI)
VCA–p18 antibody response (MFI)
U14 antibody response (MFI)
EBV read pairs
HHV–7 read pairs


a)
Prevalence of viral reads (%)
DRB1*04:04 heterozygotes
DRB1*04:04 non–carriers
Age (years)
b)
chr6: 31345435 C/C
chr6: 31345435 T/C
chr6: 31345435 T/T
c)
β, EBV DNA load in DRB1*04:04 carriers (units s.d.)
β, EBV DNA load in DRB1*04:04 non–carriers
d)
β, HHV-7 DNA load in 6_31345435 C homozygotes (units s.d.)
β, HHV-7 DNA load in 6_31345435 T homozygote
e)
β, EBV DNA load in DRB1*04:04 carriers (units s.d.)
Other variants
HLA coding variants
β, EBV DNA load in DRB1*04:04 non–carriers (units s.d.)
f)
DRB1*04:04 non-carriers
DRB1*04:04 heterozygotes
Prevalence of EBV reads (%)
DRB1 T210 genotype
g)
DQA1 E198 genotype
h)
DRB1*04:04 non-carriers
DRB1*04:04 heterozygotes
DRB1*04:04 homozygotes
DPB1 V105 genotype
i)
R2
1.00
0.75
0.50
0.25
A*02:01
– log10(All of Us saliva EBV DNA presence p)
– log10(UK Biobank blood EBV DNA load p)
j)
Effect size of A*02:01 on EBV DNA load (s.d. units)
UKB blood
AoU blood
AoU saliva (binary)
AoU saliva (quantitative)
SPARK saliva (binary)
SPARK saliva (quantitative)
k)
Prevalence of HHV-7 DNA (%)
Type O
Type B
Type A
Type AB
Blood type
l)
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 DNA load (s.d. units)
Y168H
Loss of function
CPVL
n)
C104R
A181E
Loss of function
TNFRSF13B


a)
−log10 (EBV DNA load p)
160
80
40
20
10
5
2.5
ACKR1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 22 X
Chromosome
b)
−log10 (HHV−7 DNA load p)
80
40
20
10
5
2.5
ACKR1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 X
Chromosome
c)
Prevalence of viral reads (%)
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
EBV
HHV−7
Number of Duffy null alleles
0 1 2
EUR
AFR
d)
Mean abundance (number of read pairs)
400
200
100
50
EBV
HHV−7
Number of Duffy null alleles
0 1 2
EUR
AFR


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)
B*08:01 non-carriers
B*08:01 carriers
% of individuals with EBV type 2–specific reads
60
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 = 0.40
VCA–p18 antibody response (MFI)
15000
10000
5000
0
Type 1
Type 2
EBV Type
f)
p = 0.032
EBNA–1 antibody response (MFI)
10000
5000
0
Type 1
Type 2
EBV Type
g)
p = 0.62
EA–D antibody response (MFI)
10000
5000
0
Type 1
Type 2
EBV Type
h)
p = 0.059
ZEBRA antibody response (MFI)
12500
10000
7500
5000
2500
0
Type 1
Type 2
EBV Type


a)
Adj. p < 0.05, OR > 1
Adj. p < 0.05, 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 DNA load (s.d. units)
f)
Systemic lupus erythematosus
ConMix
MR−Egger
IVW
Weighted Median
Method
Odds ratio for s.d. increase in EBV DNA load
g)
Rheumatoid arthritis
Odds ratio for s.d. increase in EBV DNA load
h)
Lymphoid leukemia
Odds ratio for s.d. increase in EBV DNA load
i)
Non-Hodgkin lymphoma
Odds ratio for s.d. increase in EBV DNA load
j)
Chronic lymphocytic leukemia
Odds ratio for s.d. increase in EBV DNA load