PNAS

Supplementary Information for

Human herpesvirus diversity is altered in HLA class I binding peptides.

William H. Palmer*, Marco Telford, Arcadi Navarro, Gabriel Santpere, Paul J. Norman

William H. Palmer

Email: William.H.Palmer@cuanschutz.edu

This PDF file includes:

Supplementary Methods

Figures S1 to S6

Legends for Datasets S1 to S5

Other supplementary materials for this manuscript include the following:

Datasets S1 to S5

Supplementary Methods

Herpesvirus genomes

From each genome, we extracted the coding sequences (CDS) (as annotated in the NCBI Virus database), corrected the coding frame with seqtk (v1.3), and translated the CDS to proteins with faTrans, for input to Orthofinder (v2.4.0) to identify homologues [135]. Homologous CDS were translation-aligned using MUSCLE (v3.8.31) [136], then checked by eye for misalignments. We removed regions of the CDS that were consistent with sequencing and assembly errors, such as strings of misaligned bases preceding a gap or early stop codon. We also treated highly repetitive regions with structural variants as missing data – for EBV this included the repeat regions in BLLF1, EBNA1, LMP1, EBNA3B, and EBNA3C and for VZV, ORF11, ORF14, and ORF22. Finally, for some EBV and HCMV genes that have highly divergent types (e.g. the EBNA2 and EBNA3 genes that identify Type 1 versus Type 2 EBV, or HCMV genes with divergent haplotypes [41]) we performed analyses two ways: 1) with "types" split into separate alignments (e.g. Type 1 and Type 2 would have separate alignments), and 2) including all "types" in a single multiple sequence alignment. The analysis of polymorphism rates presented in Figure 2 use the former, to avoid silent site saturation from highly divergent haplotypes of the same gene. Silent site saturation does not impact the calculation of F_{ST}, and therefore Figure 3 presents analyses using the latter.

Analysis of polymorphism rates

We used a Poisson mixed model to model mutation counts, similar to the SnIPRE method [62], except without corresponding divergence data. This linear modelling approach estimates parameters relative to a baseline intercept, with interactions between predictors estimated relative to the effect associated with each individual predictor. For example, consider the relationship between β_0, β^N, β_{Latent}, and β^N_{Latent} (Fig. 1). Here, the fixed predictor β^N, associated with the state of a mutation being nonsynonymous, estimates the nonsynonymous effect on mutation frequency relative to the intercept β_0, which captures the synonymous mutation rate of lytic genes. Similarly, β_{Latent} is the estimated difference in synonymous diversity between lytic and latent genes. Finally, β^N_{Latent} measures differences in the rate of nonsynonymous mutations in latency genes that would not be expected based on the estimates of β_0, β^N, and β_{Latent} alone. In the most basic form, we fitted the model as:

$$log(\mu_{ij}) = \beta_0 + \beta^N i + \beta_{length} x_{ij} + u^S_{Gene:j} + u^N_{Gene:j} i + \varepsilon_{ij} \quad [1]$$

where μ_{ij} is the expected number of synonymous (i = 0) or nonsynonymous (i = 1) mutations in gene *j*. This model estimates the intercept β_0 (density of synonymous polymorphisms) and β^N (N = nonsynonymous; a fixed effect of a mutation being nonsynonymous versus synonymous). For example, given widespread constraint, β^N is expected to be estimated as negative, reflecting pN/pS < 1. Parameter x_{ik} is the logarithm of the number of synonymous (i = 0) or nonsynonymous (i = 1) sites in gene *k* and the fixed effect β_{length} models the relationship between the number of observed mutations and the number of sites. We also fit random effects associated with each observation of mutation counts synonymous: $u^S_{Gene:j}$ and nonsynonymous: $u^N_{Gene:j}$, to control for correlated mutation counts within each gene. The two gene-specific random effects were assumed to come from a multivariate normal distribution with estimated (co)variance matrix.

In the first model (Fig. 1), we compared general patterns of polymorphism across the three viruses. Synonymous and nonsynonymous polymorphisms were counted and used as input for a model in MCMCglmm (v2.32) [137]. Counts were modelled as in equation 1 with the following additional fixed effects: β_{Virus}, β_{Latent}, $\beta_{Latent:Virus}$, β^N_{Latent}, β^N_{Virus}, and $\beta^N_{Latent:Virus}$ which estimate the differences in synonymous and nonsynonymous mutation counts across viruses and latent versus lytic gene classes. In a separate model, the estimated F_{ST} value for each gene was included as a fixed predictor, alongside the interaction between F_{ST} and the other fixed predictors. These analyses were repeated with divergent high-LD HCMV genes removed, with qualitatively similar results.

Each CDS alignment was then split into the gene regions predicted to bind HLA or regions predicted not to bind HLA. The number of segregating synonymous and nonsynonymous polymorphisms were individually summed for each subregion of the alignment. These counts were modelled, for each virus separately, as described in equation 1, with the following additional fixed predictors: β_{HLA}, β_{Latent}, $\beta_{Latent:HLA}$, β^N_{HLA}, β^N_{Latent}, and $\beta^N_{Latent:HLA}$, which are plotted in Fig. 2. This also resulted in four observations per gene (nonsynonymous and synonymous counts for regions predicted to bind HLA, or to not bind HLA), each associated with a random effect. As above, we estimated the 4x4 covariance matrix, assuming random effects are derived from a multivariate normal distribution.

Finally, regions targeted by HLA were further split by whether they were predicted to bind HLA-A, -B, or -C. These were modelled as above, except with fixed predictors associated with each HLA protein (A, B, or C) and an expanded covariance matrix to control for 16 observations per gene. We also fit this model using only sites uniquely recognized by HLA-A, HLA-B, or HLA-C. We determined significance of each fixed effect by assessing the proportion of MCMC iterations that overlap zero (pMCMC) and report 95% highest

posterior density intervals as 95% confidence intervals (CI). Throughout the results, pMCMC values are presented as "p" values for brevity.

Analysis of population differentiation

We used MCMCglmm to combine estimates of F_{ST} across genes and herpesviruses to determine if peptides recognized by HLA are associated with altered patterns of population differentiation. By combining estimates across viruses and fitting virus-specific parameters, we have greater power to detect differential population structure in HLA-binding regions while accounting for different baseline levels of population differentiation in each virus. We calculated F_{ST} in HLA-binding and non-binding regions of each EBV and HCMV gene that had at least three non-singleton amino acid polymorphisms in each region (Fig. S5). For VZV, we used a filter of one non-singleton amino acid polymorphism, because VZV had a higher F_{ST} combined with a greater number of populations represented by a single isolate. These filters were utilized to exclude genes with poor F_{ST} estimates, and those with F_{ST} estimates driven solely by synonymous polymorphism. The resulting F_{ST} distribution for each virus was bimodal, with some genes exhibiting low F_{ST} values caused by a preponderance of rare polymorphism. These genes were also removed prior to analyses (Fig. S5).

We analyzed F_{ST} with the following mixed model:

$$\hat{F}_{ST:klmn} = \beta_0 + \beta_{Virus:k} + \beta_{HLA:l} + \beta_{Latent:m} + \beta_{HLA:Virus:kl} + \beta_{Latent:Virus:km} + \beta_{HLA:Latent:lm} + u_{Gene:n} + u_{Gene:HLA:ln} + \varepsilon_{Virus:klmn} \quad [2]$$

with fixed effects associated with each virus, k ($\beta_{Virus:k}$), HLA binding (β_{HLA}), whether the gene is expressed during latency (β_{Latent}), the interaction between virus and HLA binding ($\beta_{HLA:Virus}$), the interaction between each virus and latent expression ($\beta_{Latent:Virus}$), and the interaction between latent expression and HLA binding ($\beta_{HLA:Latent}$). We fit random effects associated with HLA-targeted ($u_{Gene:HLA:ln}$) and non-targeted ($u_{Gene:n}$) regions of each gene, n and allowed these estimates to covary. Because average F_{ST} varied in EBV, HCMV, and VZV, we fit independent error variances for each virus. To plot normalized F_{ST} (Fig. 3B) we subtracted the posterior mean of $\beta_{Latent:Virus}$ and β_{Virus} from $\hat{F}_{ST:klmn}$. To identify individual genes with evidence of increased F_{ST} in HLA-binding peptides (Fig. 3D), we subtracted u_{Gene} from $u_{Gene:HLA}$ and determined the proportion of MCMC chain iterations that were greater than 0.

Analysis of the relationship between HLA allotype frequency and HLA recognition of latent peptides

To determine if predicted binding of latent peptides by HLA was dependent on HLA allotype frequency, we counted the number of latent peptides from each isolate recognized by each HLA allotype. We removed isolates whose assemblies excluded > 1 latency gene, resulting in 147 EBV, 155 HCMV, and 147 VZV isolates from 15, 13, and 17 populations respectively (Fig. S1). To reduce false positive HLA-binding peptides, we also employed a more stringent filter for inclusion of HLA peptides (percentile rank of 0.05), which is expected to include 50% of major T cell epitopes [125,126]. For each virus, we modelled the number of latent peptides (μ_{st}) from each isolate:allotype pair (s: t) using a Poisson mixed model:

$$log(\mu_{st}) = \beta_0 + \beta_{Freq}x_t + \beta_X y_s + \beta_{loci}z_s + u_{HLA:t} + u_{PC:HLA:tu} + u_{Isolate:s} + \varepsilon_{stu} \quad [3]$$

Fixed effects β_X, β_{loci}, and β_{Freq} were associated with the number of residues that were "missing data" in the assembly (y_s), number of latent genes present in the assembly (z_s), and the frequency of the HLA allotype in the population from which that isolate was derived (x_t). We included a random effect associated with each isolate and each HLA allotype, to control for multiple observations of each. We also fit random slopes associated with the principal components (PCs) of population relatedness (Fig. S6) for each HLA allotype. The number of PCs included was determined by Deviance Information Criterion (DIC) scores. PCs were calculated from a matrix of all HLA allotype frequencies across populations. The number of latent peptides per EBV isolate was zero-inflated. Therefore, in this analysis, we used a hurdle poisson model, which models two latent response variables: the mean of a zero-truncated Poisson distribution and the probability that an observation is zero (Pr(0)). The hurdle poisson model was fit with the same fixed and random effects as in equation 3.

Analysis of IEDB epitopes

We modelled these epitope frequencies using a hurdle binomial model, with a structure analogous to the hurdle Poisson model (eq. 3), except without the β_X and β_{loci} fixed effects and with β_{Latent} and $\beta_{Latent:Freq}$, to compare epitope frequencies between latent and lytic proteins across HLA allotype frequencies. We also used the following random effect structure: $u_{epitopeID:t} + u_{PC:epitopeID:tu}$, to account for multiple observations of each unique epitope across populations.

Supplementary Figures

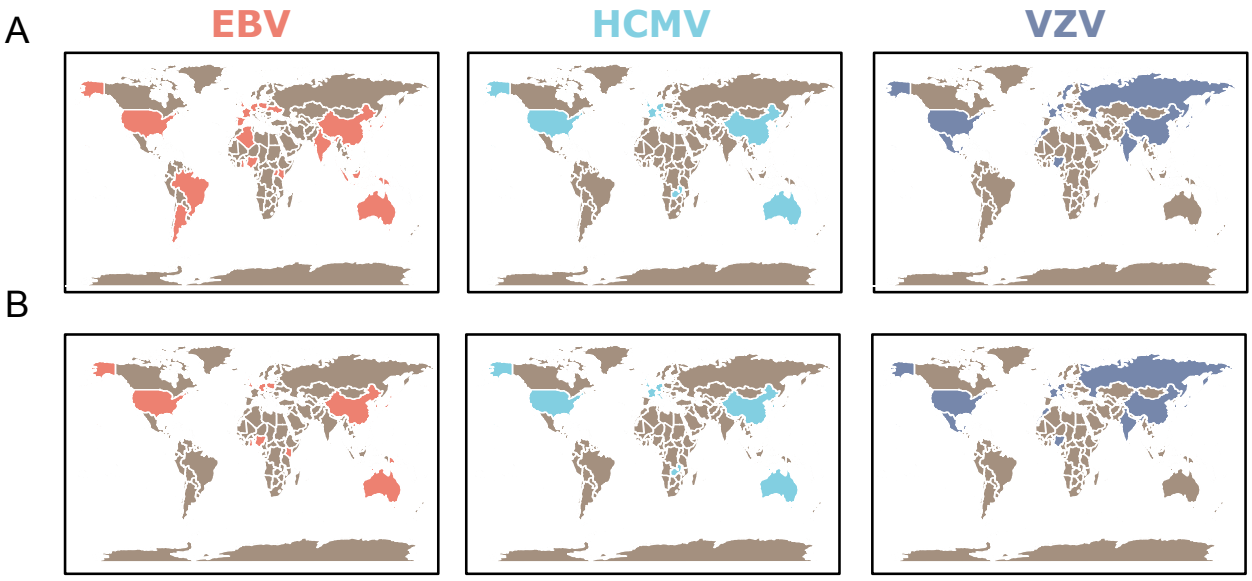

Figure S1: Populations included in the calculations of pN/pS, F_{ST} and the number of latent peptides per isolate. (A) Shaded are the populations from which EBV (red), HCMV (cyan), and VZV (blue) isolates were derived. These isolates were used in the calculation of pN/pS and F_{ST} for Fig. 2 and Fig. 3, respectively. (B) Filtering out of isolates with incomplete assemblies led to a reduced set of populations, which was used to calculate the number of distinct latent peptides per isolate (Fig. 4).

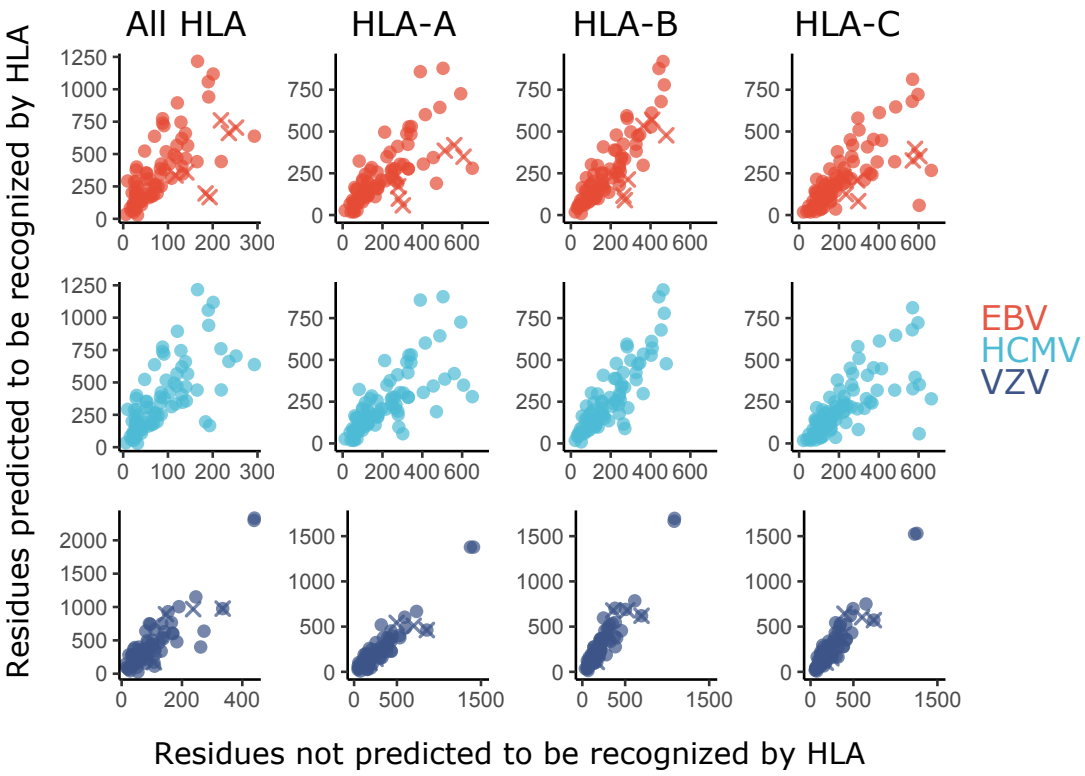

Figure S2: Number of herpesvirus residues predicted to bind any HLA. Each herpesvirus gene was split into two alignments based on recognition, or not, by any HLA class I protein, or individually by HLA-A, HLA-B, or HLA-C. Plotted are the number of residues, per protein, predicted to bind or not to bind HLA.

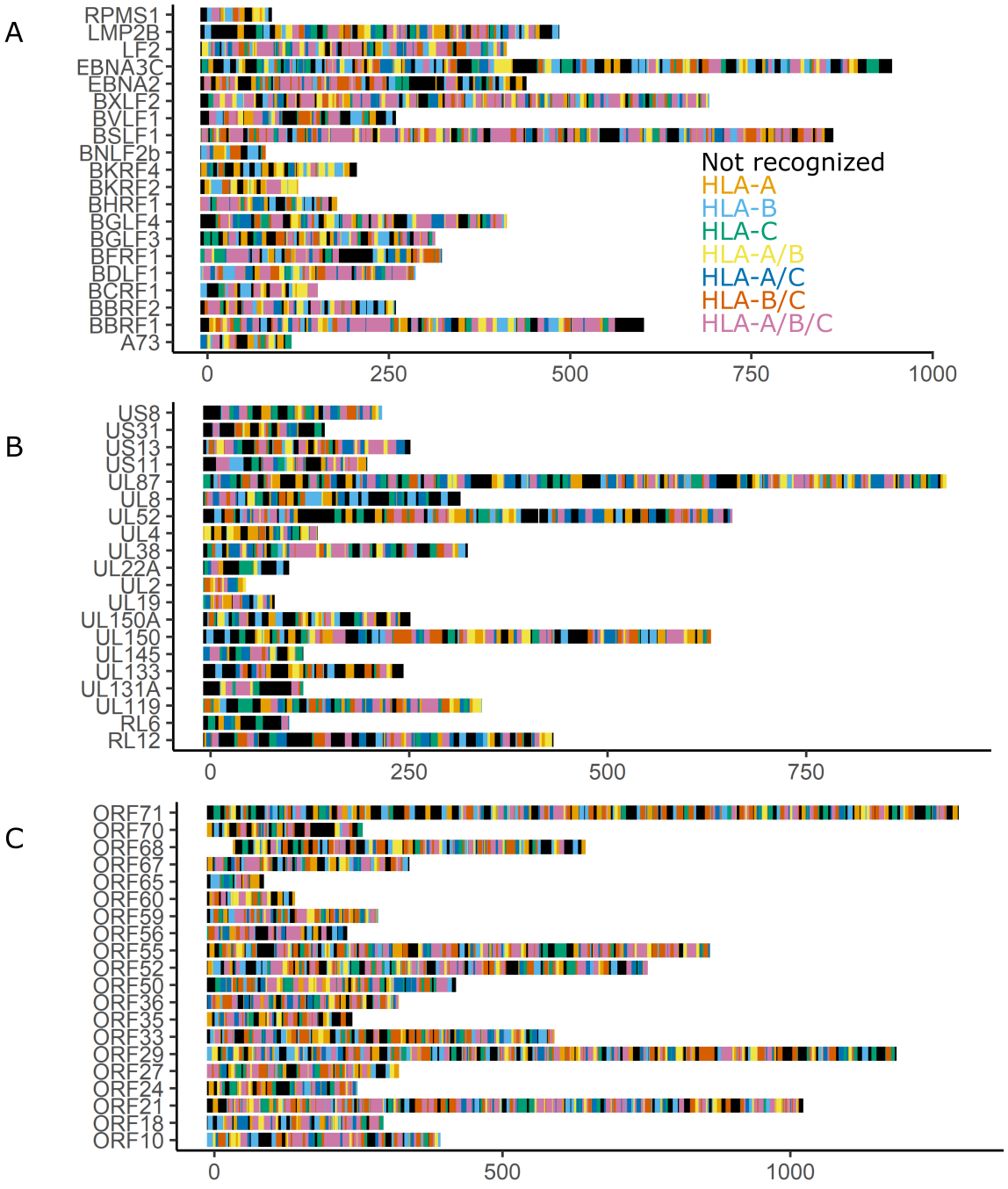

Figure S3: Examples of herpesvirus genes predicted to bind HLA. Twenty genes from EBV (A), HCMV (B), or VZV (C) were chosen at random to illustrate the patterns of predicted HLA binding across the gene. Residues are color-coded by which HLA proteins they are predicted to bind. For a full table of coordinates of HLA-binding regions on each herpesvirus protein, see Table S3.

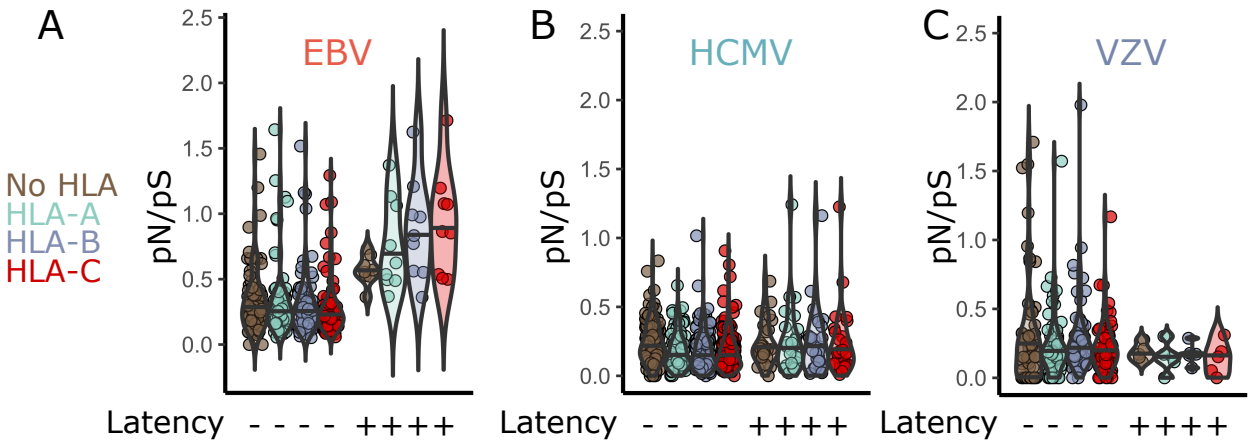

Figure S4: pN/pS for HLA-A, HLA-B, and HLA-C binding regions of viral proteins. The ratio of the rates of nonsynonymous to synonymous polymorphism (pN/pS) was calculated in HLA binding and non-binding regions of EBV (A), HCMV (B), and VZV (C) proteins. Herpesvirus genes were classified by their expression during latency or lytic replication. SnIPRE-like analysis of this data is presented in Fig. 2E.

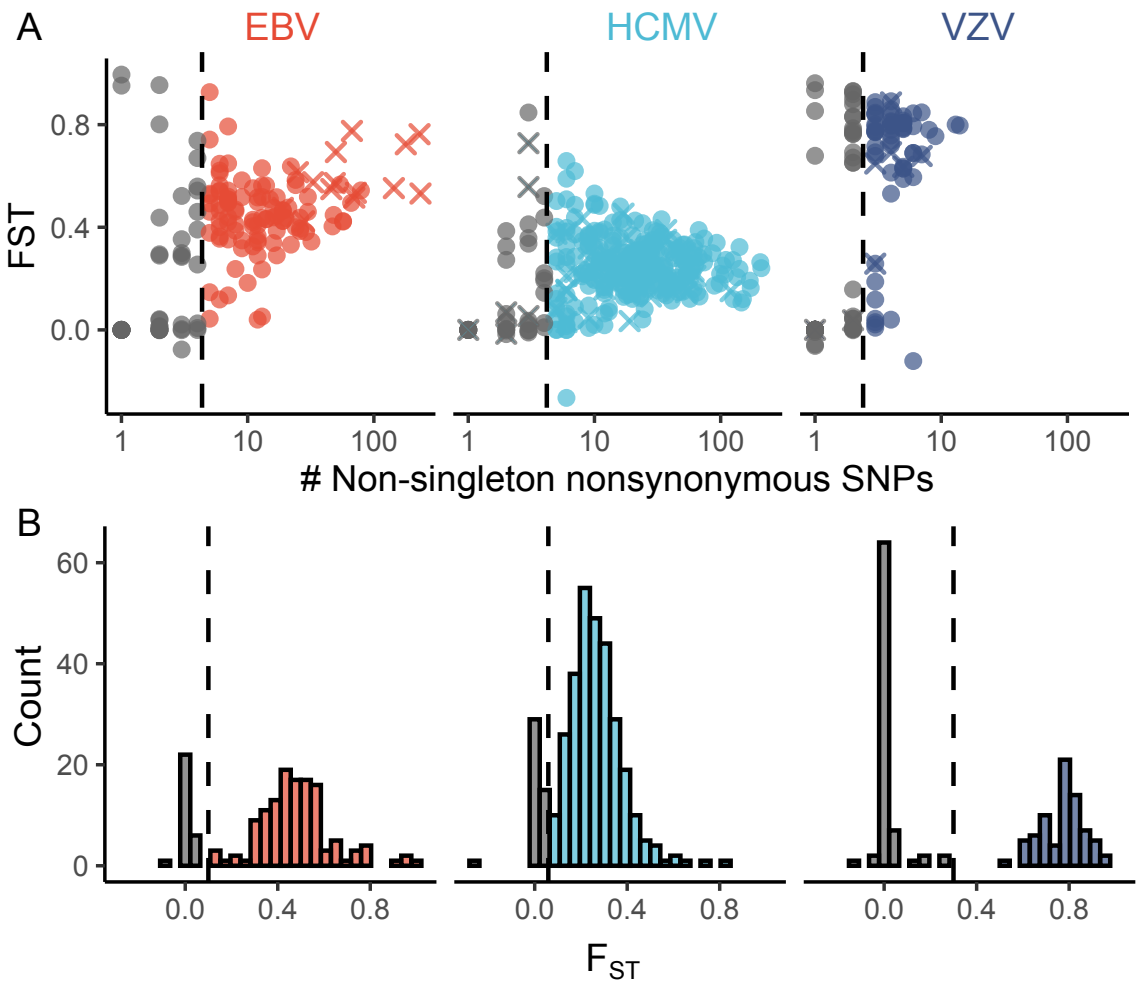

Figure S5: Gene filtering for F_{ST} analyses. (A) Gene alignments were filtered out (grey) if the number of nonsynonymous non-singleton (i.e. mutations present in more than one virus genome) were lower than the thresholds defined by the dashed line. (B) This resulted in a bimodal distribution of F_{ST} values, one centered at zero and the other approximately at the average genome-wide F_{ST} for each virus. We filtered out the former (grey).

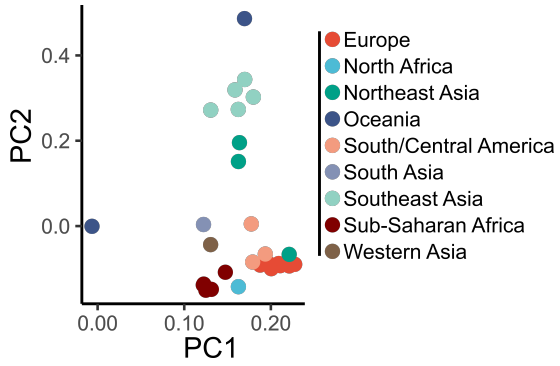

Figure S6: Principal components analysis of HLA allotype frequency. Principal components analysis was performed on HLA allotype frequencies across the populations considered (Table S2). Populations are color-coded by major geographic regions.

Supplemental Datasets

Dataset S1: Genbank ID and geographical data for each virus isolate

Dataset S2: HLA allotype frequencies from allelefrequencies.net

Dataset S3: Nonsynonymous and synonymous polymorphism counts for each virus gene

Dataset S4: F_{ST} estimates for each virus gene

Dataset S5: Number of latent peptides recognized across HLA allotype:Isolate pairs