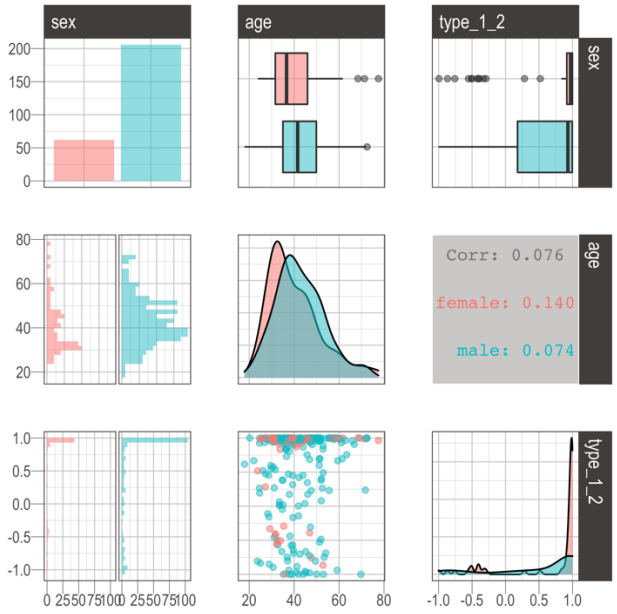
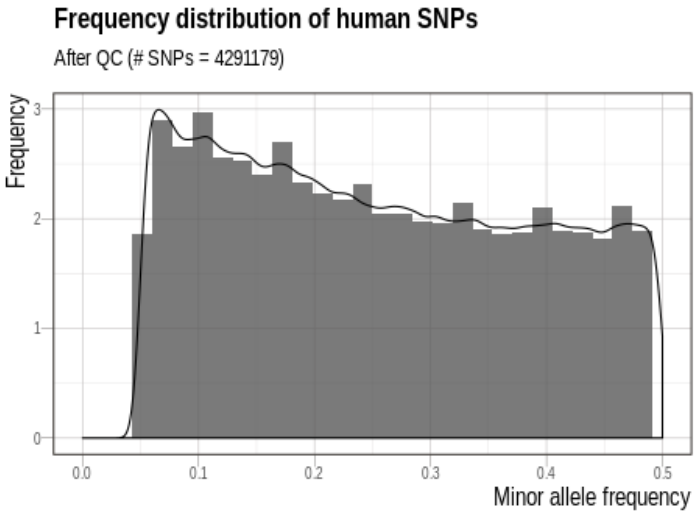
The influence of human genetic variation on Epstein-Barr virus sequence diversity

Sina Rüeger, Christian Hammer, Alexis Loetscher, Paul J McLaren, Dylan Lawless, Olivier Naret, Nina Khanna, Enos Bernasconi, Matthias Cavassini, Huldrych F. Günthard, Christian R. Kahlert, Andri Rauch, Daniel P. Depledge, Sofia Morfopoulou, Judith Breuer, Evgeny Zdobnov, Jacques Fellay and the Swiss HIV Cohort Study

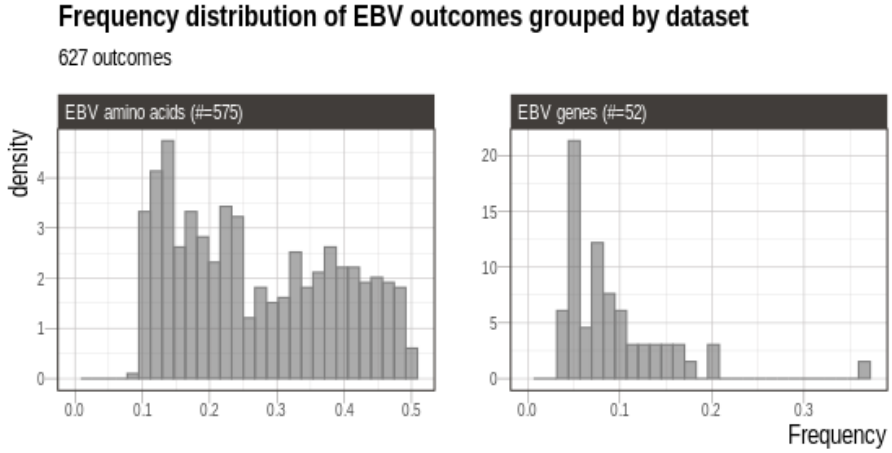
Supplementary Figures



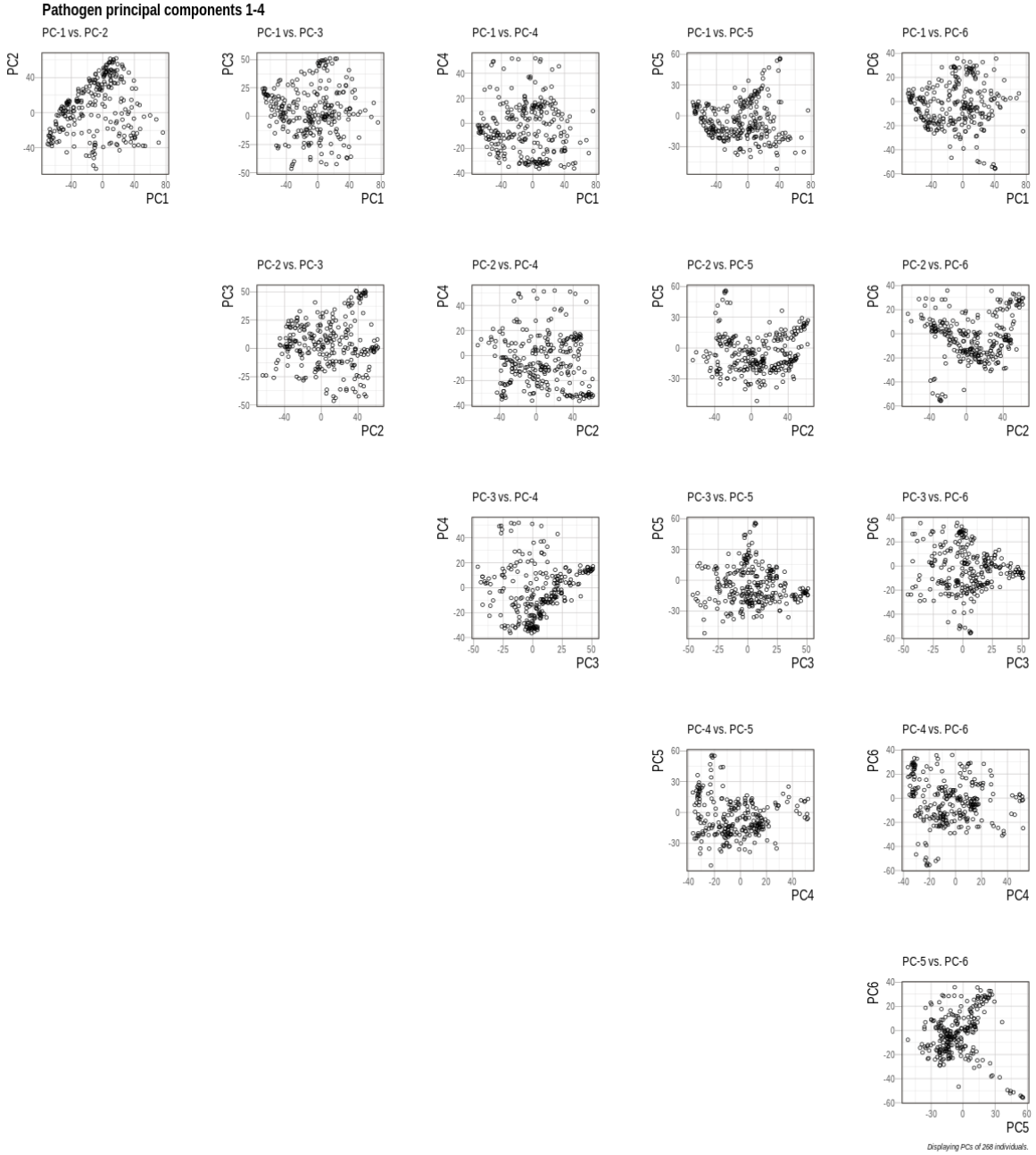
Supplementary Figure S1: Description of all covariates (sex, age, EBV type 1 vs 2). Each covariate is presented with all other covariates. Colour represents sex (red female, turquoise male). This Figure was produced using R [65].



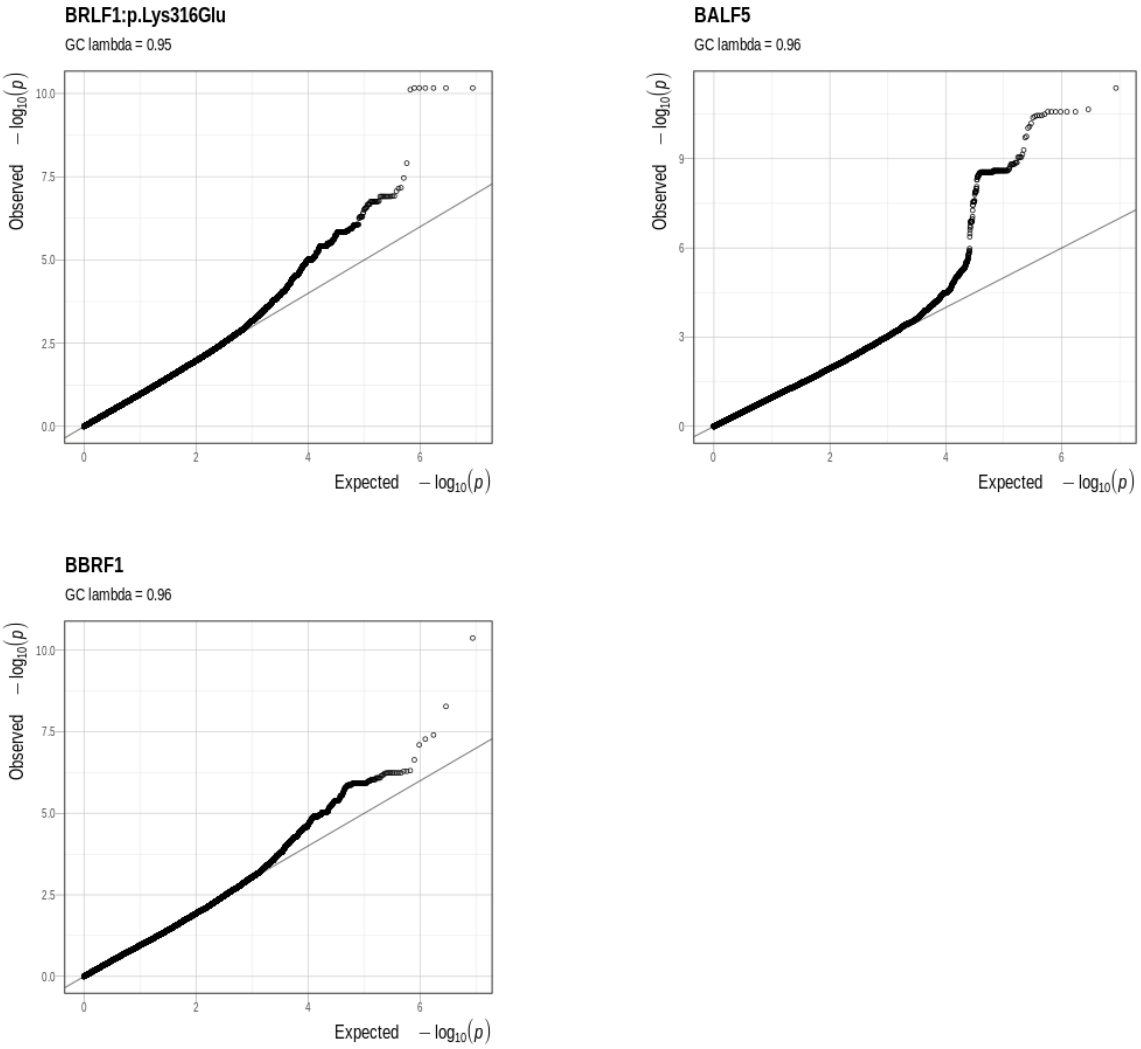
Supplementary Figure S2: Summary of host SNP minor frequency after QC. The x-axis represents the minor allele frequency, the y-axis the frequency. This Figure was produced using R [65].



Supplementary Figure S3: Summary of 627 EBV gene and amino acid variants grouped into the two EBV datasets: *EBV amino acids* dataset (LHS); *EBV genes* (RHS). Each value represents the frequency for one outcome. This Figure was produced using R [65].

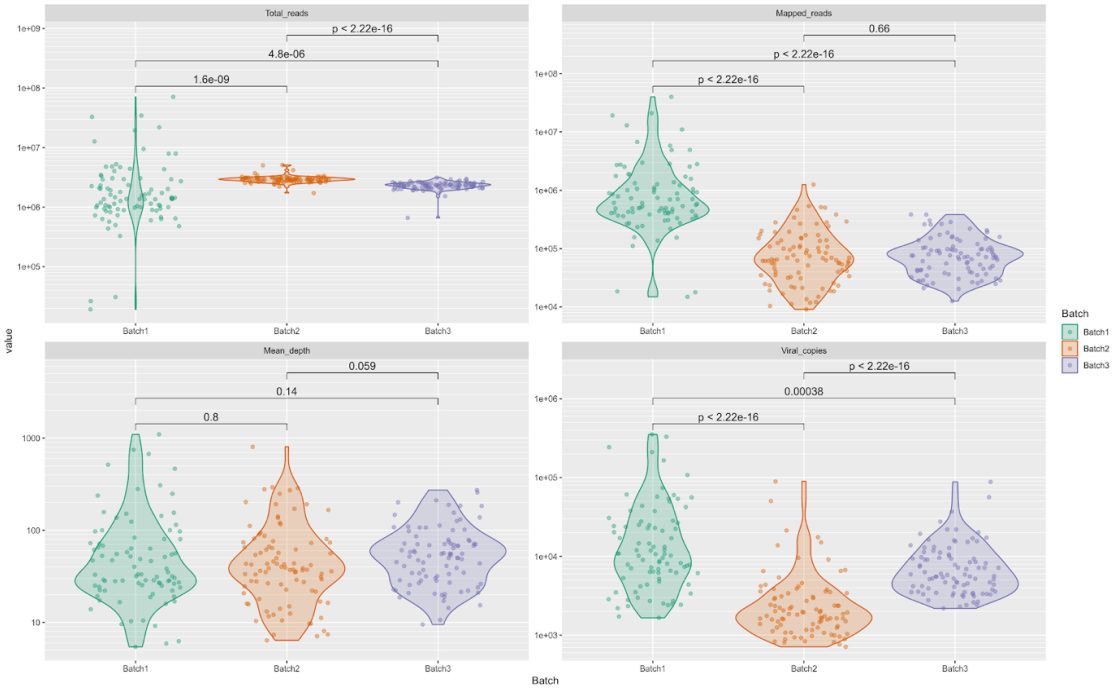


Supplementary Figure S4: First six pathogen principal components (PCs) plotted against each other. Each dot represents one individual. Principal components were computed based on EBV amino acid variants. This Figure was produced using R [65].

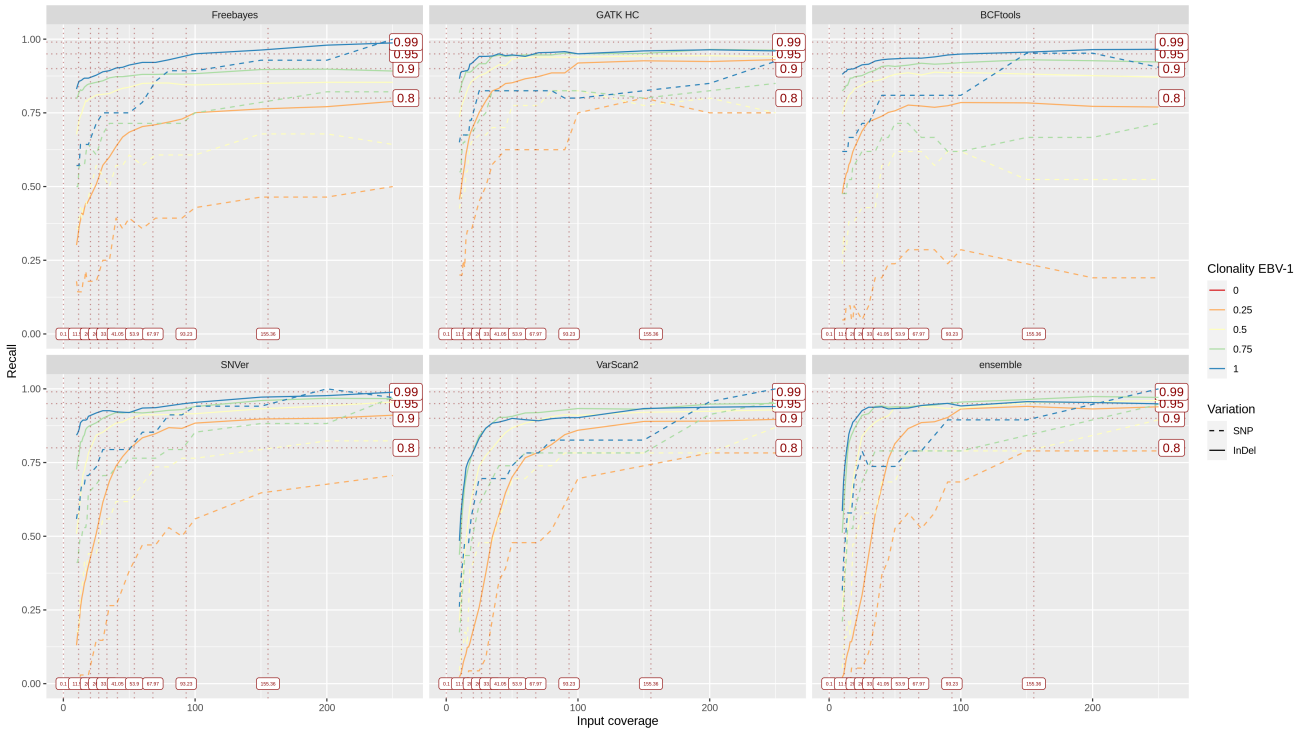


Supplementary Figure S5: Q-Q plots for all three EBV GWASs containing at least one G2G significant signals. Each point represents one SNP association. The x-axis displays the expected, the y-axis the observed -log10(p-value). The grey line is the identity line. This Figure was produced using R [65].

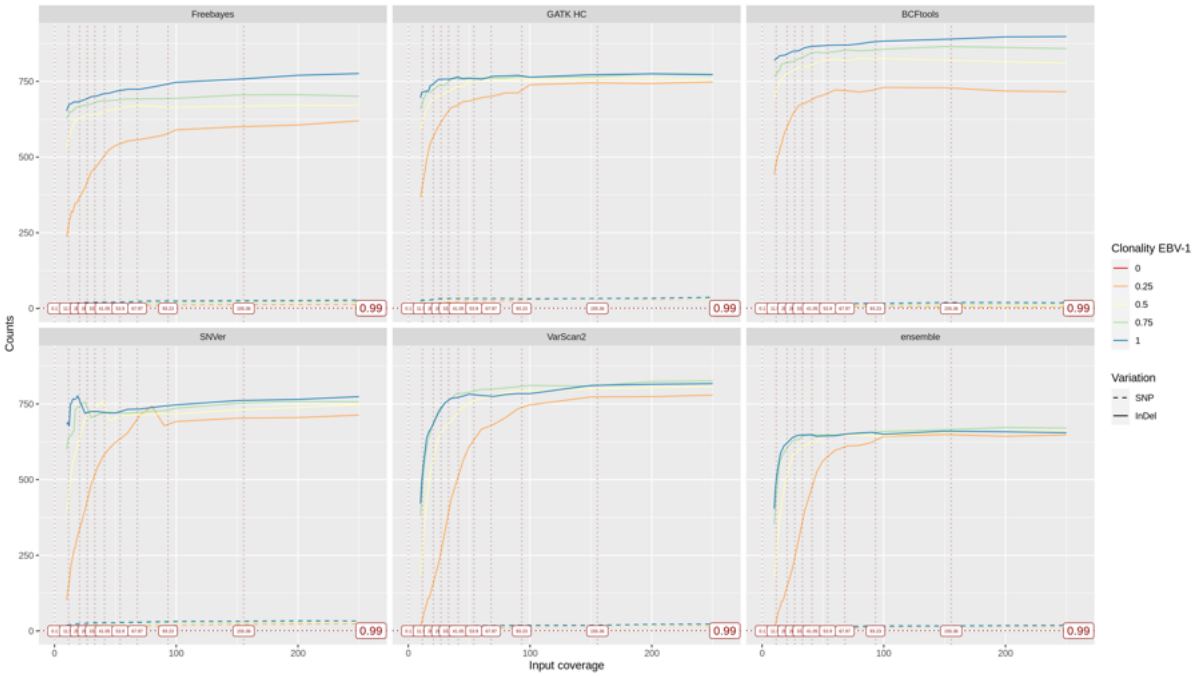
Supplementary Figure S6: Clonality of type 1 (red) and type 2 (turquoise) EBV strain in the blood of HIV patients. The relative abundance is calculated from reads mapping unambiguously to alternates haplotypes, as in formula 1. The heatmap shows that a substantial number of samples have mixed clonality and that 3 of them might be recombinant strains. This Figure was produced using R [65].



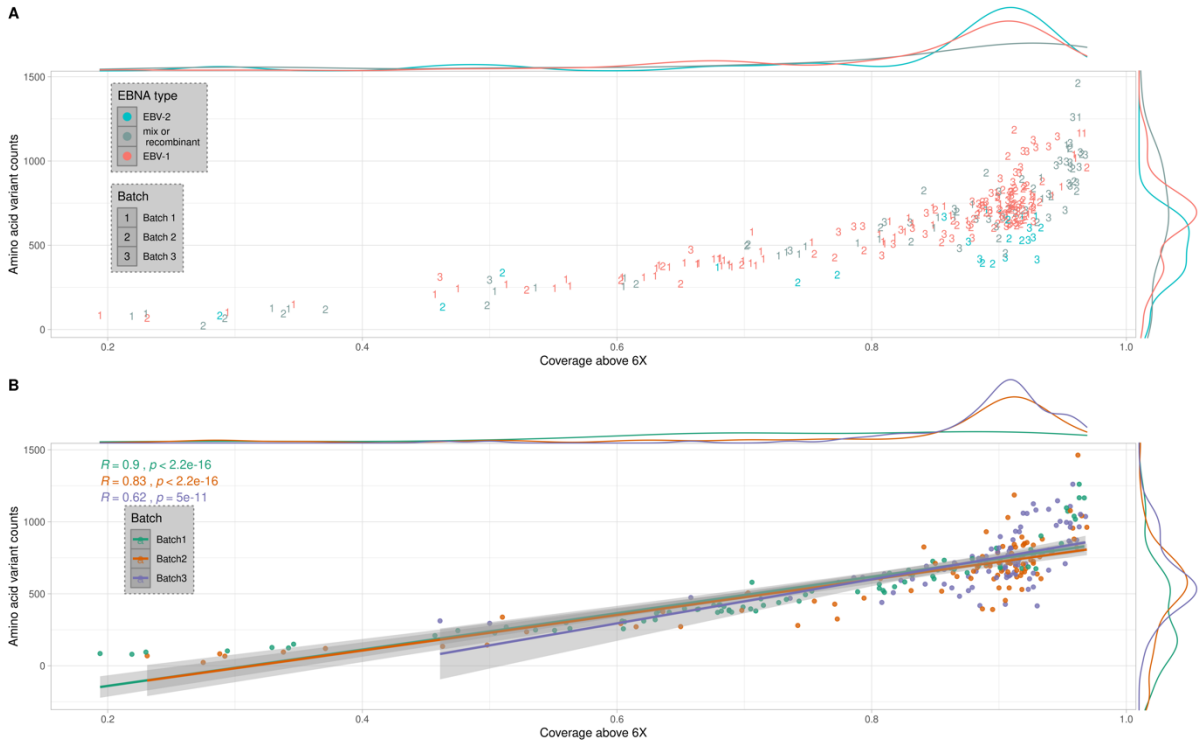
Supplementary Figure S7a: Statistics about coverage against AG876 and number of viral copies separated by sequencing batches. The p-values have been obtained by comparing the distributions, by batch, using the Mann-Whitney's U test. This Figure was produced using R [65].



Supplementary Figure S7b: Performance of Freebayes, GATK HC, BCFtools, SNVer, VarScan2 against the synthetic read libraries. The ensemble plot shows the intersection of GATK HC, SNVer and VarScan2. The reads were mapped using BWA mem against AG876. The clonality (blue to red gradient) is referred as the percentage of EBV-1 reads in the libraries. The vertical axis is the recall, namely, the variant counts of a caller in a given dataset over the size of the union of all variants found in all datasets by this caller. The vertical dashed lines are the decile of the mean read depths of EBV reads in the SHCS samples. This Figure was produced using R [65].



Supplementary Figure S7c: Raw variant counts by GATK HC, BCFtools, SNVer, VarScan2 against the synthetic read libraries. The reads were mapped using BWA mem againts AG876. The noticeable artifact observed in SNVer did not affect the result of the ensemble, let alone the result of the other callers. This Figure was produced using R [65].



Supplementary Figure S7d: Correlation between coverage and amino acid variant counts. In A, the data points are colored in function of their clonality estimated by read counts against alternate haplotypes (see Supplementary Figure S6). The samples were considered mixed (gray) when the relative abundance of neither of EBV-1 (red) nor EBV-2 (turquoise) did not exceed 90%. In B, the same data is colored by batches. The results of the Pearson correlations tests between the coverage and the amino acid variant counts were colored accordingly. This Figure was produced using R [65].