

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

Sina Rüeger<sup>1,2#</sup>, Christian Hammer<sup>3#</sup>, Alexis Loetscher<sup>2,4#</sup>, Paul J McLaren<sup>5,6</sup>, Dylan Lawless<sup>1,2</sup>, Olivier Naret<sup>1,2</sup>, Nina Khanna<sup>7</sup>, Enos Bernasconi<sup>8</sup>, Matthias Cavassini<sup>9</sup>, Huldrych F. Günthard<sup>10,11</sup>, Christian R. Kahlert<sup>12</sup>, Andri Rauch<sup>13</sup>, Daniel P. Depledge<sup>14</sup>, Sofia Morfopoulou<sup>14</sup>, Judith Breuer<sup>14</sup>, Evgeny Zdobnov<sup>2,4</sup>, Jacques Fellay<sup>1,2,15\*</sup> and the Swiss HIV Cohort Study

<sup>1</sup> School of Life Sciences, EPFL, Lausanne, Switzerland

<sup>2</sup> Swiss Institute of Bioinformatics, Switzerland

<sup>3</sup> Genentech Inc, 1 DNA Way, South San Francisco, CA, USA

<sup>4</sup> Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

<sup>5</sup> JC Wilt Infectious Diseases Research Centre, Public Health Agency of Canada, Winnipeg, MB, Canada

<sup>6</sup> Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, MB, Canada

<sup>7</sup> Department of Infectious Diseases and Hospital Epidemiology and Department of Biomedicine, University and University Hospital of Basel, Switzerland

<sup>8</sup> Division of Infectious Diseases, Regional Hospital of Lugano, Lugano, Switzerland

<sup>9</sup> Service of Infectious Diseases, Lausanne University Hospital and University of Lausanne, Switzerland

<sup>10</sup> Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland

<sup>11</sup> Institute of Medical Virology, University of Zurich, Zurich, Switzerland;

<sup>12</sup> Division of Infectious Diseases, Cantonal Hospital of St. Gallen, St. Gallen, Switzerland

<sup>13</sup> Department of Infectious Diseases, Bern University Hospital, University of Bern, Switzerland

<sup>14</sup> Division of Infection and Immunity, University College London, London, UK

<sup>15</sup> Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Switzerland

# These authors contributed equally to this work.

\* Corresponding author.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

# 1 Abstract

2 Epstein-Barr virus (EBV) is one of the most common viruses latently infecting  
3 humans. Little is known about the impact of human genetic variation on the  
4 large inter-individual differences observed in response to EBV infection. To  
5 search for a potential imprint of host genomic variation on the EBV sequence,  
6 we jointly analyzed paired viral and human genomic data from 268 HIV-  
7 coinfecting individuals with CD4+ T cell count  $<200/\text{mm}^3$  and elevated EBV  
8 viremia. We hypothesized that the reactivated virus circulating in these patients  
9 could carry sequence variants acquired during primary EBV infection, thereby  
10 providing a snapshot of early adaptation to the pressure exerted on EBV by the  
11 individual immune response. We searched for associations between host and  
12 pathogen genetic variants, taking into account human and EBV population  
13 structure. Our analyses revealed significant associations between human and  
14 EBV sequence variation. Three polymorphic regions in the human genome were  
15 found to be associated with EBV variation: one at the amino acid level  
16 (BRLF1:p.Lys316Glu); and two at the gene level (burden testing of rare variants  
17 in BALF5 and BBRF1). Our findings confirm that jointly analyzing host and  
18 pathogen genomes can identify sites of genomic interactions, which could help  
19 dissect pathogenic mechanisms and suggest new therapeutic avenues.

# 1 Introduction

2 Human genetic variation plays a key role in determining individual responses after exposure  
3 to infectious agents. Even though susceptibility or resistance to a microbial challenge is the  
4 final result of dynamic interactions between host, pathogen, and environment, human genetic  
5 polymorphisms have been shown to have an important, directly quantifiable impact on the  
6 outcome of various infections [1,2].

7 Genome-wide association studies (GWAS) have proven powerful to identify genetic regions  
8 implicated in a wide range of complex traits in both health and disease [3]. In the field of  
9 infectious diseases, several clinical and laboratory phenotypes have been investigated,  
10 including, for example disease susceptibility [4,5], clinical outcomes [6], adaptive immunity  
11 [7,8,9] or drug response [10]. In chronically infected patients, however, the pathogen genome  
12 itself provides a promising complementary target to investigate the impact of host genomic  
13 diversity on infection. While one part of the variation observed in pathogen DNA or RNA  
14 sequence is present at the transmission event, another fraction is acquired during the course  
15 of an infection, resulting at least partially from selective pressure exerted by the host response  
16 on the infectious agent. The phenomenon of within-host evolution has been extensively  
17 investigated for both viruses [11,12,13] and bacteria [14,11]. Pathogen genomic variation can  
18 thus be considered an intermediate phenotype that is detectable as a footprint of within-host  
19 evolution. This can serve as a basis for a joint association analyses of host and pathogen  
20 genome variation, which we called genome-to-genome (G2G) analysis [15], a more powerful  
21 approach than using a clinical outcome alone. A global description of the adaptive forces  
22 acting on a pathogen genome during natural infection holds the potential to identify novel  
23 therapeutic and diagnostic targets and could inform vaccine design efforts [16].

24 A G2G analysis for the quickly evolving human immunodeficiency virus (HIV) identified strong  
25 associations of single nucleotide polymorphisms (SNPs) in the HLA class I region with multiple  
26 amino acid variants across the viral genome [15]. More recent work showed an impact of  
27 variation in the HLA class II and interferon lambda 4 (IFNL4) loci on hepatitis C virus (HCV)  
28 sequence diversity [17,18,19]. While the rate of evolutionary change in RNA viruses is higher  
29 than in DNA viruses [20], the latter also present considerable amounts of inter- and intra-host  
30 variation. Among herpesviruses, it has been shown that human cytomegalovirus (HCMV) has  
31 higher genomic variability than other DNA viruses [21]. Recent genome sequencing efforts  
32 demonstrated that the same holds true for Epstein-Barr virus (EBV) [22,23].

33 EBV is a widespread human pathogen that causes infectious mononucleosis in about 10% of  
34 individuals during primary infection. EBV infection occurs most often early in life, with about  
35 30% of children being seropositive by age 5, 50% by age 10 and up to 80% by age 18 [24].  
36 This human infecting herpesvirus has also been associated with post-transplant  
37 lymphoproliferative disease [25] and could play a role in some autoimmune diseases [26,27].  
38 In addition, EBV has oncogenic properties and is implicated in the pathogenesis of multiple  
39 cancer types, predominantly Burkitt's lymphoma, Hodgkin's and non-Hodgkin's lymphoma,  
40 nasopharyngeal carcinoma and gastric carcinoma [28,29]. More than 5% of the 2 million  
41 infection-associated new cancer cases in 2008 could be attributed to EBV [30]; it was also  
42 estimated to have caused 1.8% of cancer deaths in 2010, i.e. more than 140,000 cases [31].

43 The EBV genome is approximately 170 Kbp long and encodes at least 80 proteins, not all of  
44 which have been definitively identified or characterized. After primary infection, the EBV  
45 genome persists in B cells as multicopy episomes that replicate once per cell cycle. In this

1 latent mode, only a small subset of viral genes is expressed. Latent EBV can then reactivate  
2 to a lytic cycle, which involves higher gene expression and genome amplification for packaging  
3 into new infectious viral particles [32].

4 A small number of host genomic analyses of EBV infection have been recently published,  
5 demonstrating that human genetic diversity plays a role in disease outcome. A study in 270  
6 EBV isolates from southern China identified two non-synonymous EBV variants within the  
7 *BALF2* gene that were strongly associated with the risk of nasopharyngeal carcinoma [33].  
8 Another group investigated the co-evolution of worldwide EBV strains [34] and found extensive  
9 linkage disequilibrium (LD) throughout EBV genomes. Furthermore, they observed that genes  
10 in strong LD were enriched in immunogenic genes, suggesting adaptive immune selection and  
11 epistasis. In a pediatric study of 58 Endemic Burkitt lymphoma cases and 40 healthy controls,  
12 an EBV genome GWAS identified 6 associated variants in the genes *EBNA1*, *EBNA2*, *BcLF1*,  
13 and *BARF1* [35]. Finally, the narrow-sense heritability of the humoral immune response  
14 against EBV was estimated to be 0.28 [9,36].

15 Here, we present the first global analysis of paired human and EBV genomes. We studied full  
16 EBV genomes together with their respective host genomic variation in a cohort of 268  
17 immunocompromised, HIV-coinfected patients. We chose untreated HIV-coinfected patients  
18 because EBV reactivation leading to viremia is more prevalent in immunosuppressed  
19 individuals than in an average population. Our analysis reveals three novel host genomic loci  
20 that are associated with variation in EBV amino acids or genes.

## 21 **Materials and Methods**

### 22 **Study participants, sample preparation**

23 The Swiss HIV Cohort Study (SHCS) is a nationwide, prospective cohort study of HIV-infected  
24 patients that enrolled >20,000 individuals since its establishment in 1988 and prospectively  
25 followed them at 6-month intervals [37]. For this project, SHCS participants were identified  
26 based on written consent for human genetic testing and availability of a peripheral blood  
27 mononuclear cell (PBMC) sample at time of advanced immunosuppression (i.e., with CD4+ T  
28 cell count below 200/mm<sup>3</sup>) in the absence of antiretroviral treatment.

29 We obtained demographic and clinical information from the SHCS database. These included  
30 sex, age, longitudinal HIV viral load results (number of RNA copies per ml of plasma),  
31 longitudinal CD4+ T cell counts (number of cells per mm<sup>3</sup> of blood), and history of opportunistic  
32 infections.

33 The SHCS has been approved by the Ethics Committees of all participating institutions. Each  
34 study participant provided written informed consent for genetic testing.

### 35 **EBV genome quantification, enrichment and sequencing**

36 DNA was extracted from PBMCs using the MagNA Pure 96 DNA and the Viral NA Small  
37 Volume Kit (Roche, Basel, Switzerland). Cellular EBV load was then determined using  
38 quantitative real-time PCR. Samples that yielded > 100 viral copies / ul were selected for EBV  
39 genome sequencing.

1 We used the previously described enrichment procedure to increase the relative abundance  
2 of EBV compared to host DNA [38]. Shortly, baits covering the EBV type 1 and 2 reference  
3 genomes were used to selectively capture viral DNA according to the SureSelect Illumina  
4 paired-end sequencing library protocol. Samples were then multiplexed and sequenced on an  
5 Illumina NextSeq sequencer [38].

## 6 EBV sequence analyses

7 We chose a reference-based approach to call variants in the pathogen data. Since EBNA-2  
8 and EBNA-3s are highly variable between EBV-1 and EBV-2 strains, we suspected that reads  
9 sequenced from these genes would map only to their corresponding type. In an attempt to  
10 attenuate the reference-bias this could cause, we constructed two references, one with the  
11 whole genome of the EBV-1 strain B95-8 (accession NC\_007605) and EBNA-2 and EBNA-3s  
12 sequences from EBV-2 strain AG876 (accession NC\_009334) and another one with the whole  
13 genome of AG876 with the EBNA-2 and EBNA-3s sequences of B95-8.

14 The read libraries were processed through Trimmomatic [39] to remove remnant PCR tags,  
15 TagDust [40] to eliminate low complexity reads and CD-HIT [41] to filter out duplicate reads.  
16 The remaining sequence reads were aligned to the constructs described in the previous  
17 paragraph. Following GATK best practices [42,43], we mapped the read libraries using BWA  
18 mem [44]. We cleaned the regions around InDels using GATK v3.8's IndelRealigner [45]. As  
19 a last pre-processing step, we applied bwa-postalt.js, a BWA script that adjusts mapping  
20 quality score in function of alignments on ALT haplotypes.

21 Because patients can be infected by multiple EBV strains [46], we used BWA's ALT-aware  
22 ability. In short, reads mapping to an ALT contig were always marked as supplementary  
23 alignment, regardless of mapping quality, unless they did not map to the primary assembly.  
24 This makes it easy to find unambiguously mapped reads, which we used as markers to  
25 quantify type 1 and type 2 EBV reads in all samples.

$$r1 = \frac{\# T1}{(\# T1 + \# T2) \cdot L1}$$

$$r2 = \frac{\# T2}{(\# T1 + \# T2) \cdot L2}$$

26  $r = r1 - r2$

27 where #T1 and #T2 are the unambiguous read counts against type 1 and 2 haplotypes,  
28 respectively, L1 and L2 are the length of type 1 and 2 haplotypes, respectively, and r1 and r2  
29 are the type 1 and 2 ratios, respectively. The score r is the relative abundance between type  
30 1 and type 2.

## 31 Definition of EBV amino acid variants

32 Since no gold standard variant set exists for EBV nor any closely related viral species, variant  
33 calling was performed using three different variant callers (GATK haplotypcaller, SNVer [47]  
34 and VarScan2 [48]) and by selecting as *bona fide* variant set the intersection of the three. The  
35 identified EBV variants were annotated using snpEff [49]. Nucleotide variants were

1 transformed into binary amino acid matrices using in-house python scripts. The whole pipeline  
2 is written in Snakemake [50] and Python [51].

3 This approach was benchmarked using synthetic libraries generated from B95-8 and AG876  
4 using ART Illumina and RNFTools, at a range of coverage between 10X and 250X and 5  
5 different admixture conditions, 100% B95-8 or AG876, 75% - 25 % and 50%-50%. Assessing  
6 the true number of variants between EBV-1 and EBV-2 strains is not trivial because of the  
7 high variability in EBNA-2 and EBNA-3s regions. Therefore, we rated the variant callers and  
8 the consensus of the three mentioned callers on self-consistency. The performances of the  
9 runs were measured using the ratio of the variant counts to the size of the union of all variants  
10 called by a specific tested tool.

11 By using EBV type 2 as a reference, we focused on two types of variation in EBV strains: 1)  
12 single amino acid variants; and 2) burden of very rare amino acid variants (present in only 1  
13 sample) in each viral gene (Figure 1). We call these datasets *EBV amino acids* and *EBV*  
14 *genes*, respectively. Both datasets contain binary values, with a value of 1 standing for "variant  
15 present" and 0 for "no variant present". Positions with a coverage of less than 6x were set to  
16 "missing" and samples with more than 80% missing positions were excluded entirely. The  
17 positions covered by less than 6 reads were considered missing and imputed using the  
18 imputePCA function implemented in the missMDA R-package [52]. In total, we obtained 4392  
19 amino acid variants and 83 gene variants. However, to limit the risk of model overfitting and  
20 because of low statistical power due to sample size we only included in the downstream  
21 association analyses the 575 *EBV amino acids* with an amino acid frequency of more than  
22 10% and 52 *EBV genes*.

23

24

## A. EBV genes and amino acids Example



## B. Data Example BNF1 gene

Nucleotides	ID	SNVs					
		SNV 1	SNV 2	SNV 3	SNV 4	SNV 5	SNV 6
	1	G	A	G	T	T	G
	2	G	C	G	T	T	G
	3	G	A	G	T	C	G
	4	G	A	G	T	T	G

Amino acids	ID	Amino acid changes	
		Amino acid 1 Glu > Ala	Amino acid 2 Ser > Leu
	1	0	0
	2	1	0
	3	0	1
	4	0	0

## C. Datasets

### 575 EBV amino acids

1 = variant present  
0 = no variant present

### 52 EBV genes

Only aggregates amino acids that appear in one individual.  
1 = at least one variant present  
0 = no variant present

*non-synonymous amino acid variation*

**Figure 1:** Illustration of EBV sequence variation.

A) The EBV genome is about 170 Kbp long and contains 83 genes, for a total of 4392 amino acid residues. As an example, we focus on the *BNRF1* gene and on two amino acid changes: Glu→Ala and Ser→Leu. We know for each sample the genomic variants across the whole genome, as illustrated with the colored nucleotides. Using the nucleotide information and a reference genome we can compute the amino acid changes.

B) We compare each individual (ID) to reference data and encode an amino acid as 1 if that individual has a non-synonymous change, and a 0 if not. This process returns us a matrix containing binary values, with individuals as row, and amino acids as columns. In our example, individual 2 has an amino acid change Glu→Ala and individual 3 an amino acid change Ser→Leu.

C) To transform the data into outcomes for the G2G analysis we can use the amino acid matrix as it is (*EBV amino acids* dataset) or remove all amino acid columns that appear in more than 1 individual and then pool amino acids per gene (1 = variant present, *EBV genes* dataset).

## Human genotyping and imputation

A subset of 84 participants had been genotyped in the context of previous studies on several platforms. For the remaining 196 samples, human genomic DNA was isolated from PBMCs with the QIA Symphony DSP DNA Kit (Qiagen, Hilden, Germany), and genotyped using Illumina OmniExpress (v1.1) BeadChip arrays.

1 Genotype imputation was performed on the Sanger imputation server independently for all  
2 genotyping platforms, using EAGLE2 [53] for pre-phasing and PBWT [54] with the 1000  
3 Genomes Phase 3 reference panel [55]. Low-quality imputed variants were excluded based  
4 on imputation INFO score ( $< 0.8$ ). All datasets were merged, only keeping markers that were  
5 genotyped or imputed for all genotyping platforms. SNPs were excluded on the basis of per-  
6 individual missingness ( $> 3\%$ ), genotype missingness ( $> 1\%$ ), marked deviation from Hardy-  
7 Weinberg equilibrium ( $p < 1 \times 10^{-6}$ ) and minor allele frequency  $< 5\%$  (Table 1). All quality control  
8 procedures were performed using PLINK 2.0 [56].

## 9 Association analyses

10 We used the mixed model association implementation for binary and continuous outcomes in  
11 GCTA (v1.92) [57,58] to search for potential associations between human SNPs and EBV  
12 variants. The model can be expressed with the following equation:

$$13 \quad y_k = \alpha X + \beta^{(kl)} g_l + \eta + \varepsilon$$

14 where the outcome  $y$  is a binary vector indicating whether an EBV variant is present (1) or not  
15 (0);  $X$  is a matrix that contains all covariates,  $\alpha$  represents all fixed effects of all covariates  
16 (including an intercept term),  $g$  is the SNP genotype vector with coded additive allele dosages  
17 0, 1 or 2,  $\beta$  is the (fixed) effect of the SNP to be tested for association,  $\eta$  is the polygenic  
18 (random) effect and  $\varepsilon$  the error term. This mixed model was estimated for each EBV variant  
19 ( $k$ ) and SNP ( $l$ ), and integrated over all  $L$  SNPs and  $K$  EBV variants. To estimate  $\eta$ , the host  
20 genetic relationship matrix (GRM) was calculated from QC preprocessed genotype data using  
21 GCTA [57].

22 The use of a mixed effects association model allows to account for population stratification of  
23 the host genome. To control for population stratification among EBV genomes, we included  
24 the first six principal components (PCs) of EBV genetic variation to the covariate matrix  $X$  [59].  
25 Other covariates were sex, age and EBV type. PCs were calculated from EBV amino acid  
26 variants using the convexLogisticPCA function from the R package logisticPCA [60] in R [61].  
27 As data preparation for PC computation, we removed variants with less than 5% or more than  
28 95% frequency. Missing amino acid values were imputed with the imputePCA function from  
29 the R package missMDA [52].

30 Significance was assessed using the usual genome-wide significance threshold in European  
31 populations of  $5 \times 10^{-8}$  and dividing it by the effective number of GWASs performed [62]. We  
32 used FINEMAP [63] to determine the most likely causal SNP(s) in a 2-Mb-wide window around  
33 each significant SNP. FINEMAP requires GWAS summary statistics and LD estimations as  
34 input. To estimate LD between SNPs, we used LDstore [64]. We performed eQTL lookups for  
35 host SNPs in eQTLGen [65], EUGENE [66] and GTEx [67].

36 Unless otherwise specified, all data preparation and analyses were performed using R [61].

## 37 Code availability

38 EBV data preparation: [https://gitlab.com/ezlab/vir\\_var\\_calling](https://gitlab.com/ezlab/vir_var_calling)

39 G2G Analysis: <https://github.com/sinarueeger/G2G-EBV-manuscript>



# 1 Results

## 2 Study participants and human genetic data

3 PBMC samples from 778 SHCS participants were screened for the presence of cellular EBV  
4 DNA using RT-PCR. A total of 290 of them were identified as viremic for EBV (>2000 copies).  
5 We obtained good quality human genotyping and EBV sequencing data for 268 of them, which  
6 were included in the association analyses. The study cohort comprised 206 male and 62  
7 female individuals, between the ages of 20 and 78 (median 40) (Table 1 and Supplementary  
8 Figure S1).

9 We applied standard GWAS quality control (QC) procedures that yielded information for  
10 4'291'179 SNPs (Table 1 and Supplementary Figure S2, which shows the distribution of the  
11 minor allele frequency spectrum after QC).

## 12 EBV genomic diversity and variant calling

13 Genome coverage was very uneven between the samples. Mean depth varied from less than  
14 6x for 14 samples, up to more than 500x in 5 others. We also observed fluctuation in coverage  
15 above 6x, which we used to exclude 12 samples in which less than 20% of the EBV genome  
16 was sufficiently covered (Supplementary Figure S7a). In addition, the coverage in the first  
17 sequencing batch was not uniform.

18 We estimated the clonality of EBV strain in each sample by taking advantage of the high  
19 divergence between EBNA<sub>s</sub> T1 and T2 haplotypes. Among the 282 sequenced samples,  
20 57.1% were predominantly (9:1) infected by T1 EBV, while 5.7% were mostly infected by T2  
21 EBV (Supplementary Figure S6). The remaining 37.2% were infected by multiple strains or by  
22 recombinant viruses. This approach does not allow to stratify further than the EBNA types.

23 The variant calling pipeline was adapted to output variants by minimizing the impact of the  
24 admixture ratio and of the low coverage observed in the SHCS samples. Variants were called  
25 against EBV-2, as EBV-2 was able to call more variants than EBV-1 (Supplementary Figure  
26 S7d). The benchmark experiments against AG876 (EBV-2) yielded a total of 961 different  
27 variants. The most conservative was SNVer (783 variants), while the most sensitive was  
28 BCFtools (930 variants). The variant callers can be prone to artifacts [68], which was  
29 specifically observed in SNVer (Supplementary Figure S7c) in these datasets. To reduce the  
30 probability of calling artifacts, we chose to use the *bona fide* intersection of GATK HC, SNVer  
31 and VarScan2. This approach is likely to be impacted by low coverage. The recall is stable at  
32 around 95% at 25X coverage upwards and reasonable (10%) at 20X (Supplementary Figure  
33 S7c). Hence, low coverage has an impact, specifically, half potential variants called, on only  
34 15% of the SHCS sample. However, this approach is very conservative, since it outputs only  
35 88%, 85% and 79% of the variants called by SNVer, GATK HC and VarScan2, respectively.

36 On average, around 800 amino acid variants were called for each sample, with slight  
37 differences correlating with the clonality of the samples and the coverage above 6X  
38 (Supplementary Figure S7d). The variant counts against the AG876 construct (EBV-2) were  
39 generally higher in mixed infections and EBV-1 strains (Supplementary Figure S7d A). The  
40 variant counts were generally lower in the samples included in the first sequencing batch,  
41 which is likely due to the fluctuating coverage. However, overall, the number of variants was

1 found to be comparable across the samples, ranging from 400 to 1500 for the 77% samples  
2 with a 6X coverage above 80% (Supplementary Figure S7d). Under 80% coverage, the variant  
3 counts hardly exceed 500 but rarely drops under 200 either. It is therefore likely that we missed  
4 variants using our approach. The positions covered by less than 6 reads were considered  
5 missing and imputed afterwards using the imputePCA function implemented in the missMDA  
6 R-package.

7 We analyzed EBV variation using two approaches: single marker analysis of *EBV amino acids*,  
8 *to investigate common viral variation*, and burden testing of very rare amino acid variants in  
9 *EBV genes* (Table 1, Supplementary Figure S3). Applying logistic principal component  
10 analysis of viral genomic structure showed a single main cluster (Supplementary Figure S4).

## 11 **Genome-to-Genome association analysis**

12 We tested for associations between each EBV variant and human SNPs. We studied 575 EBV  
13 amino acids and 52 EBV genes, for a total of 627 GWASs. The effective number of GWASs  
14 performed was 458. As covariates, we included the first six EBV principal components (51.4%  
15 deviance explained), sex, age, type 1 vs 2 of EBV (Supplementary Figure S1). The sample  
16 size ranged between 120 and 268, with a median sample size of 264. Sample size variation  
17 was due to variable missingness in the EBV data. Genomic inflation factors for each of the  
18 627 GWASs ranged between 0.92 and 1.12.

19 Significant associations ( $p < 1.09 \times 10^{-10}$ ) were identified between a total of 25 human SNPs  
20 and viral variants mapping to three EBV regions (Table 2): the *EBV genes* **BALF5** (Figure 2a)  
21 and **BBRF1** (Figure 2b) and the *EBV amino acid* **BRLF1:p.Lys316Glu** (Figure 2c). The minor  
22 allele frequency of all significant host SNPs was between 0.05 and 0.10. The genomic inflation  
23 factors of the three GWASs ranged between 0.95 and 0.96 (Q-Q plots shown in  
24 Supplementary Figure S5).

25

26

27

28

29

30

31

32

33

34

35

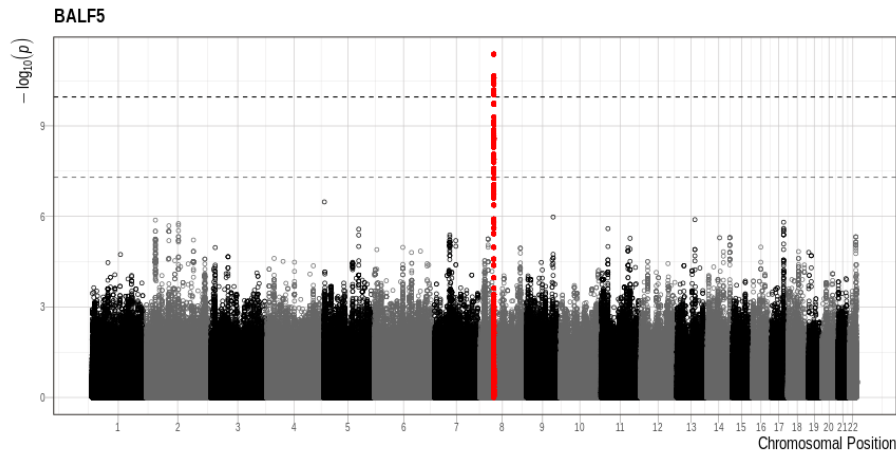
36

37

38

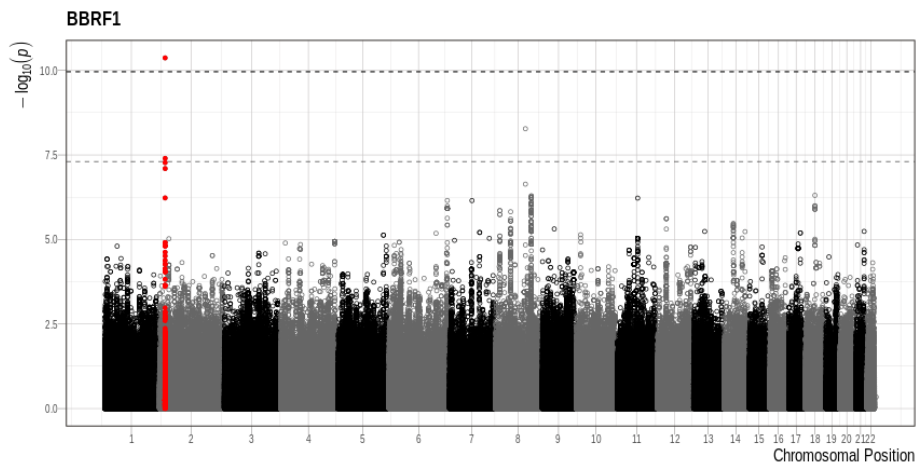
1  
2

**A**



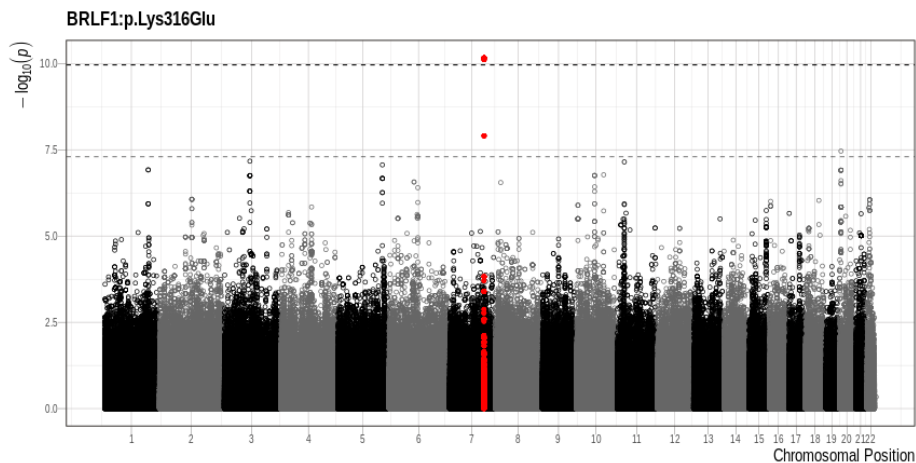
3  
4

**B**



5  
6

**C**



7

8 **Figure 2:** Significant associations - A: *BALF5*, B: *BBRF1*, C: *BRLF1*:p.Lys316Glu.

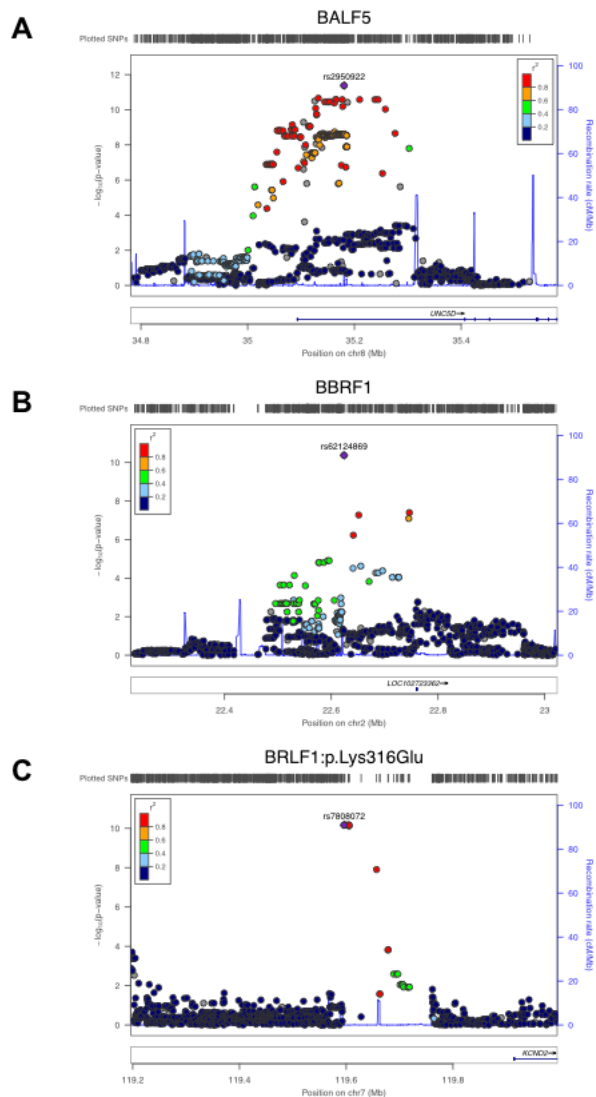
9 The x-axis represents the chromosomal position and the y-axis displays the  $-\log_{10}(p)$ .  
10 Colour alternates between chromosomes. Regions that contain statistically significant SNP  
11 are presented in red (top SNP +/- 400 Kbp). The light grey dashed line represents the GWAS  
12 significance threshold of  $5 \times 10^{-8}$ , the dark grey dashed line the G2G threshold of  $1.09 \times 10^{-10}$ .

1 Strong associations were observed between 17 SNPs in the *UNC5D* region on chromosome  
2 8 and the occurrence of very rare functional variants in the EBV **BALF5** gene (Figures 3a, 4a),  
3 which is involved in viral DNA replication during the late phase of lytic infection. *UNC5D* is a  
4 poorly characterized gene expressed mainly in neuronal tissues, which encodes a protein that  
5 has been shown to regulate p53-dependent apoptosis in neuroblastoma cells [69]. The top  
6 associated SNP, rs2950922 (OR = 1.31, 95%-CI = 1.21-1.41, P =  $4.2 \times 10^{-12}$ , effect allele G),  
7 is an eQTL for *UNC5D* in esophageal tissue (GTEx, [67]).

8 Rare amino acid variation in **BBRF1** was found to be associated with a single SNP,  
9 rs62124869 (OR = 1.29, 95%-CI = 1.19-1.39, P =  $4.2 \times 10^{-11}$ , effect allele C), which maps to the  
10 non-coding RNA gene *LINC01830* (Long Intergenic Non-Protein Coding RNA 1830) on  
11 chromosome 2 (Figures 3b, 4b).

12 Finally, 7 SNPs mapping to a non-coding region of chromosome 7 were found to be associated  
13 with the EBV amino acid variant **BRLF1:p.Lys316Glu** (Figure 2c, 3c). The top SNP,  
14 rs6466720, had a p-value of  $6.85 \times 10^{-11}$  and an OR of 1.41 (95%-CI = 1.28-1.58, effect allele  
15 G). **BRLF1** controls lytic reactivation of EBV from latency and regulates viral transcription.  
16 **BRLF1:p.Lys316Glu** has not been described previously, but variation at the nearby residue  
17 377 (**BRLF1:p.Glu377Ala**) has been shown to be prevalent in cases of nasopharyngeal and  
18 gastric carcinomas in Chinese samples [70]. **BRLF1:p.Lys316Glu** and **BRLF1:p.Glu377Ala**  
19 are in moderate LD ( $r^2=0.55$ ) in our dataset.

20



1  
2

3 **Figure 3:** Locuszoom plots.

4 Locuszoom plots for the three EBV association signals highlighted in red in Figure 2 (A:  
5 *BALF5*, B: *BBRF1*, C: *BRLF1*:p.Lys316Glu).

## 6 Discussion

7 Because immunosuppression - and in particular T cell deficiency - favors EBV reactivation  
8 from its latent B cell reservoir, EBV viremia is frequently detected in (untreated) HIV-infected  
9 individuals with advanced disease and low CD4+ T cell counts. We hypothesized that the  
10 reactivated virus circulating in these patients could carry sequence variants acquired during  
11 primary EBV infection, thereby providing a snapshot of early adaptation to the pressure  
12 exerted on EBV by the individual immune response.

13 To search for a potential imprint of host genomic variation on the viral sequence, we jointly  
14 analyzed genomic information obtained from paired EBV and human samples. Viral sequence  
15 variation can be seen as an intermediate phenotype, closer to potentially causal host  
16 polymorphisms than clinically observable outcomes like viral load or disease phenotypes. As

1 such, it allows the detection of more subtle associations, less likely to be obscured by  
2 environmental influences. In our G2G analysis, we used variation at EBV amino acid residues  
3 as outcome in multiple parallel GWAS, which allowed us to obtain effect estimations between  
4 each human genetic variant and EBV variation.

5 We identified two EBV genes and one EBV amino acid as associated with three regions of the  
6 human genome, spanning altogether 25 SNPs. For the GWAS with **BALF5** as the outcome,  
7 the associated human genomic region contains eQTLs for the nearby gene *UNC5D*. This gene  
8 is poorly characterized but has been shown to play a role in the regulation of apoptosis [69].  
9 The other two EBV genes (**BRLF1**, **BBRF1**) provided little indication of what underlying  
10 mechanism might be at play.

11 Our study is limited by its small sample size and by the complexity of correcting for human  
12 and EBV population stratification. Indeed, if not carefully controlled for, the existence of  
13 population structure in the host and pathogen genome might create spurious associations or  
14 decrease real signals in G2G analyses, resulting in both type I and type II errors. With a mixed  
15 model approach and the inclusion of pathogen principal components as covariates, the  
16 genomic inflation factors of our GWAS ranged between 0.92 and 1.12. This wide range of  
17 genomic inflation factors is likely due to a combination of small sample size and complex  
18 statistical model. To prevent false positives, we adjusted for genomic inflation when extracting  
19 significant SNPs and used a conservative G2G significance threshold of  $5 \times 10^{-8}$  divided by the  
20 effective number of GWAS performed. Although viral genetic variation is a more precise  
21 phenotype to study than traditional outcomes, it comes at the price of decreased power due  
22 to the high-dimensional outcome. The significance threshold is thus much lower than in a  
23 single GWAS. To limit the number of statistical tests performed, we restricted our analysis to  
24 common gene and amino acid variation.

25 Our analyses have been performed using historical samples collected from untreated HIV-  
26 infected individuals. Considering the natural history of EBV infection in humans and its high  
27 likelihood to be acquired during the first 2 decades of life, we postulate that intra-host  
28 adaptation of EBV happened before HIV infection, i.e. with a normally functioning immune  
29 system. At the time of sample collection, all study participants had advanced  
30 immunosuppression with low CD4+ T cell counts ( $< 200$  cells/mm<sup>3</sup> of blood). We therefore  
31 assume an absence of selective pressure on EBV at that time. These assumptions limit  
32 obviously the generalizability of our findings to non-HIV-infected population. Similar studies  
33 performed during primary EBV infection or in other specific population (e.g. bone-marrow  
34 transplant recipients) would help better delineate the global impact of intra-host selection on  
35 EBV sequence variation.

36 Our study provides a preliminary list of statistical associations between the EBV and the  
37 human genomes. The cataloguing of the sites of host-pathogen genomic conflict is potentially  
38 useful for further functional exploration, as has been demonstrated for HIV and HCV infections.  
39 Our results require replication and validation in different cohorts and settings. Importantly,  
40 larger sample sizes will be needed to increase power and provide more robust estimations.

## 41 **Data availability**

42 The datasets generated during and/or analysed during the current study are available in the  
43 following Zenodo repositories: G2G results are in <https://doi.org/10.5281/zenodo.4289138>,  
44 pathogen data in <https://doi.org/10.5281/zenodo.4011995>.

## 1 References

- 2 1. Chapman, S. J. & Hill, A. V. S. Human genetic susceptibility to infectious disease. *Nat.*  
3 *Rev. Genet.* **13**, 175–188 (2012).
- 4 2. Casanova, J.-L. & Abel, L. The human genetic determinism of life-threatening infectious  
5 diseases: genetic heterogeneity and physiological homogeneity? *Hum Genet* **139**, 681–694  
6 (2020).
- 7 3. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation.  
8 *The American Journal of Human Genetics* **101**, 5–22 (2017).
- 9 4. Timmann, C. *et al.* Genome-wide association study indicates two novel resistance loci for  
10 severe malaria. *Nature* **489**, 443–446 (2012).
- 11 5. McLaren, P. J. *et al.* Association Study of Common Genetic Variants and HIV-1  
12 Acquisition in 6,300 Infected Cases and 7,200 Controls. *PLoS Pathogens* **9**, e1003515  
13 (2013).
- 14 6. McLaren, P. J. *et al.* Polymorphisms of large effect explain the majority of the host genetic  
15 contribution to variation of HIV-1 virus load. *PNAS* **112**, 14658–14663 (2015).
- 16 7. Rubicz, R. *et al.* A Genome-Wide Integrative Genomic Study Localizes Genetic Factors  
17 Influencing Antibodies against Epstein-Barr Virus Nuclear Antigen 1 (EBNA-1). *PLoS*  
18 *Genetics* **9**, e1003147 (2013).
- 19 8. Zhou, Y. *et al.* Genetic loci for Epstein-Barr virus nuclear antigen-1 are associated with  
20 risk of multiple sclerosis. *Mult. Scler.* **22**, 1655–1664 (2016).
- 21 9. Hammer, C. *et al.* Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of  
22 Humoral Response to Common Viruses. *The American Journal of Human Genetics* **97**, 738–  
23 743 (2015).
- 24 10. Ge, D. *et al.* Genetic variation in *IL28B* predicts hepatitis C treatment-induced viral  
25 clearance. *Nature* **461**, 399–401 (2009).
- 26 11. Alizon, S., Luciani, F. & Regoes, R. R. Epidemiological and clinical consequences of  
27 within-host evolution. *Trends in Microbiology* **19**, 24–32 (2011).
- 28 12. Fraser, C. *et al.* Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary  
29 Perspective. *Science* **343**, 1243727 (2014).
- 30 13. Farci, P. *et al.* The Outcome of Acute Hepatitis C Predicted by the Evolution of the Viral  
31 Quasispecies. *Science* **288**, 339–344 (2000).
- 32 14. Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution  
33 of bacterial pathogens. *Nature Reviews Microbiology* **14**, 150–162 (2016).
- 34 15. Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic  
35 variation, HIV-1 sequence diversity, and viral control. *Elife* **2**, e01123 (2013).
- 36 16. Cohen, J. I. Epstein–barr virus vaccines. *Clin Transl Immunology* **4**, e32 (2015).

- 1 17. Ansari, M. A. *et al.* Genome-to-genome analysis highlights the effect of the human innate  
2 and adaptive immune systems on the hepatitis C virus. *Nature Genetics* **49**, 666–673 (2017).
- 3 18. Ansari, M. A. *et al.* Interferon lambda 4 impacts the genetic diversity of hepatitis C virus.  
4 *Elife* **8**, (2019).
- 5 19. Chaturvedi, N. *et al.* Adaptation of hepatitis C virus to interferon lambda polymorphism  
6 across multiple viral genotypes. *eLife* **8**, e42542 (2019).
- 7 20. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses:  
8 patterns and determinants. *Nature Reviews Genetics* **9**, 267–276 (2008).
- 9 21. Cudini, J. *et al.* Human cytomegalovirus haplotype reconstruction reveals high diversity  
10 due to superinfection and evidence of within-host recombination. *Proc Natl Acad Sci USA*  
11 **116**, 5693 (2019).
- 12 22. Kwok, H. *et al.* Genomic Diversity of Epstein-Barr Virus Genomes Isolated from Primary  
13 Nasopharyngeal Carcinoma Biopsy Samples. *Journal of Virology* **88**, 10662–10672 (2014).
- 14 23. Palser, A. L. *et al.* Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types  
15 and Normal Infection. *Journal of Virology* **89**, 5222–5237 (2015).
- 16 24. Balfour, H. H. *et al.* Age-specific prevalence of Epstein-Barr virus infection among  
17 individuals aged 6-19 years in the United States and factors affecting its acquisition. *J.*  
18 *Infect. Dis.* **208**, 1286–1293 (2013).
- 19 25. Green, M. & Michaels, M. G. Epstein-Barr Virus Infection and Posttransplant  
20 Lymphoproliferative Disorder: EBV and PTL. *American Journal of Transplantation* **13**, 41–  
21 54 (2013).
- 22 26. Pender, M. P. The essential role of Epstein-Barr virus in the pathogenesis of multiple  
23 sclerosis. *Neuroscientist* **17**, 351–367 (2011).
- 24 27. Pender, M. P. & Burrows, S. R. Epstein-Barr virus and multiple sclerosis: potential  
25 opportunities for immunotherapy. *Clin Transl Immunology* **3**, e27 (2014).
- 26 28. Young, L. S. & Rickinson, A. B. Epstein-Barr virus: 40 years on. *Nat. Rev. Cancer* **4**,  
27 757–768 (2004).
- 28 29. Ko, Y.-H. EBV and human cancer. *Experimental & Molecular Medicine* **47**, e130 (2015).
- 29 30. de Martel, C. *et al.* Global burden of cancers attributable to infections in 2008: a review  
30 and synthetic analysis. *The Lancet Oncology* **13**, 607–615 (2012).
- 31 31. Khan, G. & Hashim, M. J. Global burden of deaths from Epstein-Barr virus attributable  
32 malignancies 1990-2010. *Infectious Agents and Cancer* **9**, 38 (2014).
- 33 32. Hammerschmidt, W. & Sugden, B. Replication of Epstein-Barr viral DNA. *Cold Spring*  
34 *Harb Perspect Biol* **5**, a013029 (2013).
- 35 33. Xu, M. *et al.* Genome sequencing analysis identifies Epstein-Barr virus subtypes  
36 associated with high risk of nasopharyngeal carcinoma. *Nature Genetics* **1** (2019)  
37 doi:[10.1038/s41588-019-0436-5](https://doi.org/10.1038/s41588-019-0436-5).



- 1 34. Wegner, F., Lassalle, F., Depledge, D. P., Balloux, F. & Breuer, J. Co-evolution of sites  
2 under immune selection shapes Epstein-Barr Virus population structure. *Mol Biol Evol*  
3 (2019) doi:[10.1093/molbev/msz152](https://doi.org/10.1093/molbev/msz152).
- 4 35. Kaymaz, Y. *et al.* Epstein Barr virus genomes reveal population structure and type 1  
5 association with endemic Burkitt lymphoma. *bioRxiv* 689216 (2019) doi:[10.1101/689216](https://doi.org/10.1101/689216).
- 6 36. Hayward, T. A. *et al.* Antibody response to common human viruses is shaped by genetic  
7 factors. *Journal of Allergy and Clinical Immunology* **143**, 1640–1643 (2019).
- 8 37. The Swiss HIV Cohort Study *et al.* Cohort Profile: The Swiss HIV Cohort Study.  
9 *International Journal of Epidemiology* **39**, 1179–1189 (2010).
- 10 38. Depledge, D. P. *et al.* Specific Capture and Whole-Genome Sequencing of Viruses from  
11 Clinical Samples. *PLOS ONE* **6**, e27805 (2011).
- 12 39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina  
13 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 14 40. Lassmann, T. TagDust2: a generic method to extract reads from sequencing data. *BMC*  
15 *Bioinformatics* **16**, 24 (2015).
- 16 41. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-  
17 generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
- 18 42. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-  
19 generation DNA sequencing data. *Nat Genet* **43**, 491–498 (2011).
- 20 43. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the  
21 Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.1-  
22 11.10.33 (2013).
- 23 44. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.  
24 *arXiv:1303.3997 [q-bio]* (2013).
- 25 45. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing  
26 next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
- 27 46. Correia, S. *et al.* Natural Variation of Epstein-Barr Virus Genes, Proteins, and Primary  
28 MicroRNA. *J Virol* **91**, (2017).
- 29 47. Wei, Z., Wang, W., Hu, P., Lyon, G. J. & Hakonarson, H. SNVer: a statistical tool for  
30 variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic*  
31 *Acids Res.* **39**, e132 (2011).
- 32 48. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of  
33 individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
- 34 49. Cingolani, P. *et al.* A program for annotating and predicting the effects of single  
35 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain  
36 w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
- 37 50. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine.  
38 *Bioinformatics* **28**, 2520–2522 (2012).
- 39 51. *Python Language Reference*. (Python Software Foundation).

- 1 52. Josse, J. & Husson, F. missMDA: A Package for Handling Missing Values in Multivariate  
2 Data Analysis. *Journal of Statistical Software* **70**, 1–31 (2016).
- 3 53. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium  
4 panel. *Nature Genetics* **48**, 1443–1448 (2016).
- 5 54. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-  
6 Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
- 7 55. The 1000 Genomes Project Consortium. A global reference for human genetic variation.  
8 *Nature* **526**, 68–74 (2015).
- 9 56. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer  
10 datasets. *GigaScience* **4**, 7 (2015).
- 11 57. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide  
12 Complex Trait Analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).
- 13 58. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and  
14 pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100–106  
15 (2014).
- 16 59. Naret, O. *et al.* Correcting for Population Stratification Reduces False Positive and False  
17 Negative Results in Joint Analyses of Host and Pathogen Genomes. *Front. Genet.* **9**, (2018).
- 18 60. Landgraf, A. J. & Lee, Y. Dimensionality Reduction for Binary Data through the  
19 Projection of Natural Parameters. *arXiv:1510.06112 [stat]* (2015).
- 20 61. R Development Core Team. *R: A Language and Environment for Statistical Computing*.  
21 (R Foundation for Statistical Computing, 2008).
- 22 62. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic  
23 association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32**,  
24 361–369 (2008).
- 25 63. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from  
26 genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
- 27 64. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using  
28 Summary Statistics from Genome-wide Association Studies. *The American Journal of*  
29 *Human Genetics* **101**, 539–551 (2017).
- 30 65. Vösa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL  
31 meta-analysis. <http://biorxiv.org/lookup/doi/10.1101/447367> (2018) doi:[10.1101/447367](https://doi.org/10.1101/447367).
- 32 66. Ferreira, M. A. R. *et al.* Gene-based analysis of regulatory variants identifies 4 putative  
33 novel asthma risk genes related to nucleotide synthesis and signaling. *J. Allergy Clin.*  
34 *Immunol.* **139**, 1148–1157 (2017).
- 35 67. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement:  
36 The GTEx Project. *Biopreserv Biobank* **13**, 311–319 (2015).
- 37 68. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation  
38 Sequencing Data. *Scientific Reports* **7**, 43169 (2017).

1 69. Wang, H. *et al.* Unc5D regulates p53-dependent apoptosis in neuroblastoma cells. *Mol*  
2 *Med Rep* **9**, 2411–2416 (2014).

3 70. Jia, Y. *et al.* Sequence analysis of the Epstein-Barr virus (EBV) BRLF1 gene in  
4 nasopharyngeal and gastric carcinomas. *Virology* **7**, 341 (2010).

## 5 **Acknowledgments**

6 This study was supported by the Leenaards Foundation (Leenaards Prize 2015 to JF and  
7 EZ). This study has also been partly financed within the framework of the Swiss HIV Cohort  
8 Study, supported by the Swiss National Science Foundation (grant #177499), by SHCS  
9 project #743 and by the SHCS research foundation. The data are gathered by the Five  
10 Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private  
11 physicians (listed in <http://www.shcs.ch/180-health-care-providers>).

12  
13 **Members of the Swiss HIV Cohort Study:** Aebi-Popp K, Anagnostopoulos A, Battegay M,  
14 Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G,  
15 Egger M, Elzi L, Fehr J, Fellay J, Furrer H, Fux CA, Günthard HF (President of the SHCS),  
16 Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M,  
17 Kahlert CR (Chairman of the Mother & Child Substudy), Kaiser L, Keiser O, Klimkait T,  
18 Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini  
19 C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of  
20 the Scientific Board), Rudin C, Scherrer AU (Head of Data Centre), Schmid P, Speck R,  
21 Stöckle M (Chairman of the Clinical and Laboratory Committee), Tarr P, Trkola A, Vernazza  
22 P, Wandeler G, Weber R, Yerly S.

## 23 **Author contributions**

24 EZ and JF conceived and supervised the work.  
25 SR and CH performed the association analyses.  
26 DPD, SM, JB performed EBV sequencing and curated the data.  
27 AL prepared and analysed the viral sequencing data.  
28 PJM, DL, ON contributed to the design of the work.  
29 NK, EB, AC, MC, HFG, CRK, AR collected the samples and associated data.  
30 SR, CH, AL and JF wrote the paper.  
31 All authors reviewed the manuscript and approved the submission.

## 32 **Additional information**

### 33 **Competing interests**

34 CH is an employee of Genentech.  
35  
36

# 1 Figure legends

2 **Figure 1:** Illustration of EBV sequence variation.

3 A) The EBV genome is about 170 Kbp long and contains 83 genes, for a total of 4392 amino  
4 acid residues. As an example, we focus on the *BNRF1* gene and on two amino acid changes:  
5 Glu→Ala and Ser→Leu. We know for each sample the genomic variants across the whole  
6 genome, as illustrated with the colored nucleotides. Using the nucleotide information and a  
7 reference genome we can compute the amino acid changes.

8 B) We compare each individual (ID) to reference data and encode an amino acid as 1 if that  
9 individual has a non-synonymous change, and a 0 if not. This process returns us a matrix  
10 containing binary values, with individuals as row, and amino acids as columns. In our example,  
11 individual 2 has an amino acid change Glu→Ala and individual 3 an amino acid change  
12 Ser→Leu.

13 C) To transform the data into outcomes for the G2G analysis we can use the amino acid matrix  
14 as it is (*EBV amino acids* dataset) or remove all amino acid columns that appear in more than  
15 1 individual and then pool amino acids per gene (1 = variant present, *EBV genes* dataset).

16

17 **Figure 2:** Significant associations - A: *BALF5*, B: *BBRF1*, C: *BRLF1*:p.Lys316Glu.

18 The x-axis represents the chromosomal position and the y-axis displays the  $-\log_{10}(\text{p-value})$ .  
19 Colour alternates between chromosomes. Regions that contain statistically significant SNP  
20 are presented in red (top SNP +/- 400 Kbp). The light grey dashed line represents the GWAS  
21 significance threshold of  $5 \times 10^{-8}$ , the dark grey dashed line the G2G threshold of  $1.09 \times 10^{-10}$ .

22

23 **Figure 3:** Locuszoom plots.

24 Locuszoom plots for the three EBV association signals highlighted in red in Figure 2 (A:  
25 *BALF5*, B: *BBRF1*, C: *BRLF1*:p.Lys316Glu).

26

27

28

29

30

31

32

33

34

35

36

37

38

# 1 Tables

dataset	variable	counts	mean	median	sd	min	max
Pathogen genome (83 Rare EBV gene variation)	Variant frequency		0.067	0.049	0.057	0	0.37
Pathogen genome (575 EBV amino acids)	Variant frequency		0.26	0.24	0.12	0.093	0.5
Covariates	Sex	male: 206, female: 62					
Covariates	AGE		42.02	40.79	10.98	20.25	77.52
Covariates	PC1		-8.95	-7.12	32.53	-	76.11
Covariates	PC2		7.39	8.4	30.41	-	61.8
Covariates	PC3		4.45	3.19	21.41	-	51.11
Covariates	PC4		-3.91	-6.69	20.72	-	52.11
Covariates	PC5		-5.09	-8.39	18.76	-51.7	56.02
Covariates	PC6		-3.16	-3.21	18.65	-	35.68
Covariates	EBV type		0.54	0.95	0.67	-1	1

2

3 **Table 1:** Summary of pathogen variants, host SNPs and covariates for 268 individuals. For  
4 each covariate, we indicate the number of individuals measured, and distribution (mean,  
5 median, standard deviation, minimum, maximum for quantitative, frequency for sex). For  
6 aggregated EBV genes the frequency is shown, for host SNPs the MAF distribution is  
7 presented.

Locus	EBV dataset	EBV outcome	SNP	Chr	OR**	p	Finemapped SNP	Finemapped probability	Effect allele	EAF	n	gene	Consequence type
2	gene (binary, variants < 1 sample)	BALF5	rs2950922*	8	1.30739	4.2E-12	TRUE	0.13317	G	0.10821	268	UNC5D***	intron_variant
3	gene (binary, variants < 1 sample)	BBRF1	rs62124869*	2	1.29103	4.3E-11	TRUE	0.71847	C	0.0541	268	LINC01830	non_coding_transcript_variant
1	amino acid (binary)	BRLF1:p.Lys316Glu	rs7808072*	7	1.41346	6.8E-11	FALSE		T	0.06762	244		
1	amino acid (binary)	BRLF1:p.Lys316Glu	rs6466720	7	1.42464	7.7E-11	TRUE	0.04783	G	0.06967	244		

1

2 \* SNP with locus-wide lowest P-value

3 \*\* Odds ratio (exp(b) for logistic mixed effects model) in SHCS

4 \*\*\* Gene and tissue of eqtl association (P-value of association), from GTEx:

5 UNC5D in Esophagus\_Muscularis (P=1.15243e-13), UNC5D in Esophagus\_Gastroesophageal\_Junction (P=2.61983e-05)

6

7 **Table 2:** Summary of G2G analysis results. Top SNP and/or fine-mapped SNP per locus represented with: EBV dataset, EBV outcome,  
8 chromosome, SNP identifier, odds ratio, p-value, whether this SNP is a top SNP or a fine-mapped SNP, the causal probability from FINEMAP,  
9 effect allele, effect allele frequency, sample size, corresponding gene, variant consequence, associated eQTL gene, associated eQTL associated  
10 eQTL gene and p-value in GTEx (from <https://gtexportal.org/home/>) [67]. See Supplementary Table S1 for detailed information about all 25  
11 variants and Supplementary Table S2 for fine-mapping results.



# A. EBV genes and amino acids Concept

EBV genome

Genes

**BNRF1**

**BCRF1**

52 genes in EBV

**LMP1**

Amino acids

Glu > Ala

Ser > Leu

Wikipedia: EBV genome - EBV genome - EBV genome

Reference

G A G T C G

SNVs

G C G T T G

# B. Data Example: BCRF1.gene

ID	SNV 1	SNV 2	SNV 3	SNV 4	SNV 5	SNV 6
1	G	A	G	T	T	G
2	G	C	G	T	T	G
3	G	A	G	T	C	G
4	G	A	G	T	T	G

Individuals

ID	Amino acid 1 Glu > Ala	Amino acid 2 Ser > Leu
1	0	0
2	1	0
3	0	1
4	0	0

Amino acids

# C. Datasets

**575 EBV amino acids**

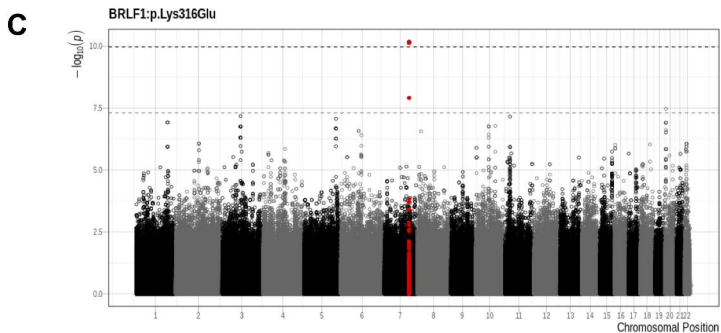
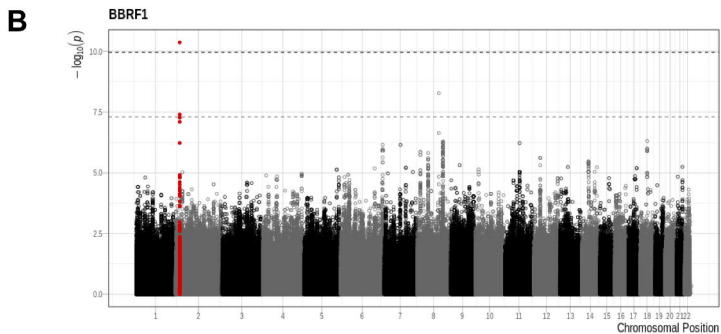
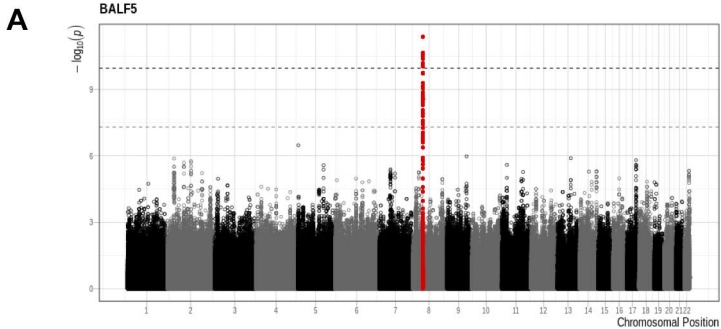
- 1 = variant present
- 0 = no variant present

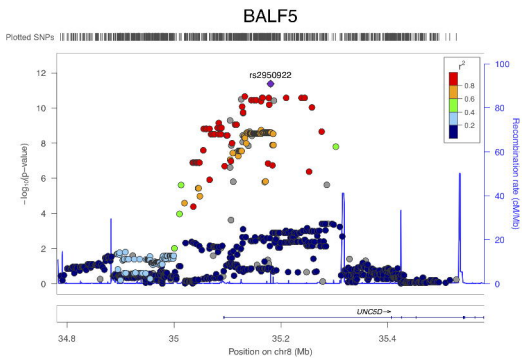
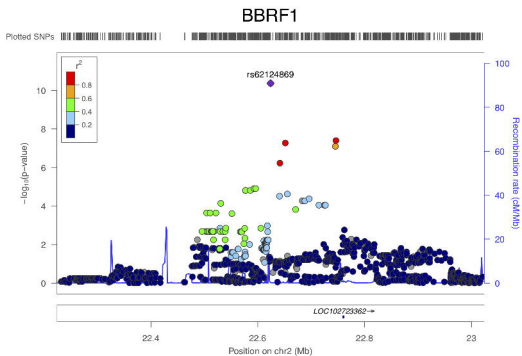
**52 EBV genes**

- Only aggregates amino acids that appear in one individual.
- 1 = at least one variant present
- 0 = no variant present

aggregation across individuals





**A****B****C**