## Supplementary Information

# Evidence of latent-lytic replication in EBV-positive Burkitt lymphoma from whole genome and transcriptome sequencing

Ismail D. Legason[1,3,5*], Adam Burns[1‡], Dimitris Vavoulis[1,], Silvia Halim[1], Helene Dreau[1], Daisy Jennings[1], Hadijah Nabalende[3], Isaac Otim[3], Martin D. Ogwang[3], Julius Sseruyange[4], Alisen Ayitewala[4], Caroline Achola[4], Susan Nabadda[4], Emmanuel Josephat[6], Heavenlight Christopher[6], William F. Mawalla[6], Clara Chamba[6], Eric Magorosa[7], Leah Mnango[7], Lulu Chirande[7], Hadija M. Mwamtemi[7], Daniel Mbwambo[8], Priscus Mapendo[8], Alex Mremi[8,9], Elifuraha Mkwizu[9], Paul Shadrack Ntemi[10], Heronima Joas[10], Edrick Mtalemwa Elias[11], Carol SK Leung[2], Kate Ridout[1‡,] and Anna Schuh[1‡] on behalf of the AIREAL Consortium.

Affiliations
[1] Department of Oncology, University of Oxford, UK
[2] Centre for Immuno-Oncology, Nuffield Department of Medicine, University of Oxford, UK
[3] Department of Paediatrics, St. Mary's Hospital, Lacor, Uganda
[4] Central Public Health Laboratory(CPHL), Ministry of Health, Kampala, Uganda
[5] Faculty of Health Sciences, Muni University, Arua, Uganda
[6] Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania
[7] Muhimbili National Hospital, Dar es Salaam, Tanzania
[8] Department of Pathology, Kilimanjaro Christian Medical Centre, Moshi, Tanzania
[9] School of Medicine, KCMC University, Moshi, Tanzania
[10] Department of Oncology, Bugando Medical Centre, Mwanza, Tanzania
[11] Department of Pathology, Catholic University of Health and Allied Sciences, Mwanza, Tanzania

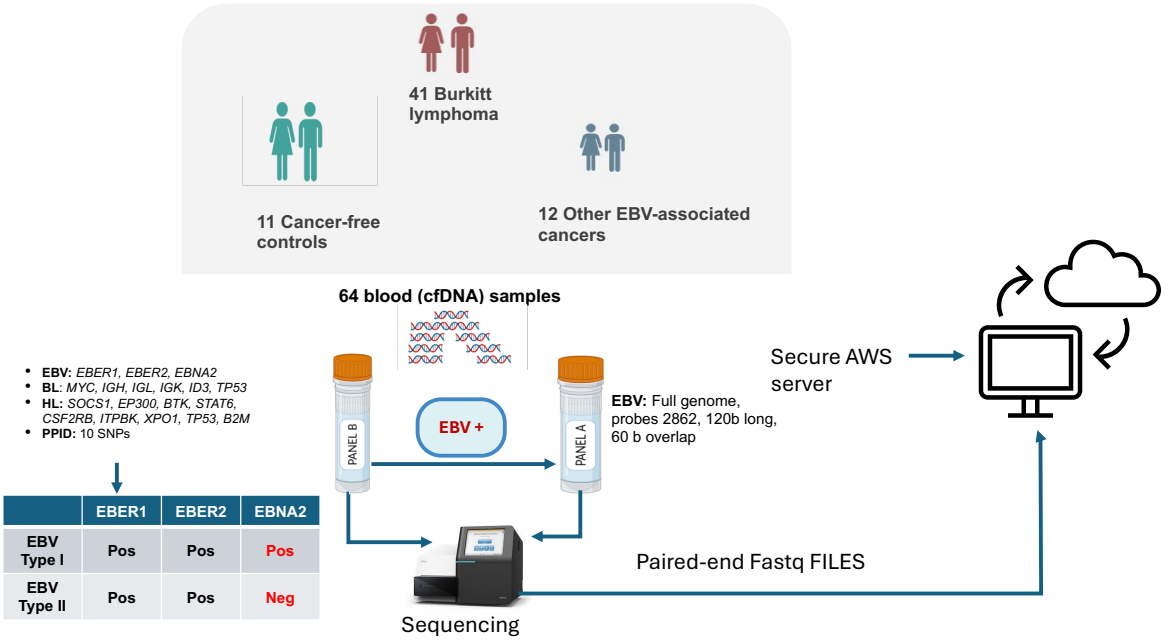
*Correspondence to ismail.draguma@queens.ox.ac.uk
‡These authors jointly supervised this work.

# Supplementary methods

## Section 1: Study Recruitment and Selection of EBV-Positive Samples

The detailed study protocol and design have been described in our previous publication[1]. Briefly, participants in the current study were enrolled in two phases. Phase I (2020–2024) was a hospital-based case-control study that assessed liquid biopsies for diagnosing EBV-driven lymphomas, known as the AI-REAL study. Cases (Burkitt lymphoma) were recruited from four East African sites: St Mary's Hospital Lacor in Uganda, Muhimbili National Hospital in Tanzania, Kilimanjaro Christian Medical Centre in Moshi, and Bugando Medical Centre in Mwanza. In Phase II (2024), a limited number of healthy controls, mainly children, were recruited from community health facilities and matched by age and sex to BL cases. This phase was only at the Ugandan site. Both studies received ethical approval from the relevant bodies: OxTREC in the UK, NIMR in Tanzania, UNCST, and the St Mary's Hospital Lacor Ethics Committee in Uganda. Participants gave informed consent or assent for minors aged 7–17. Protocols conformed to the Declaration of Helsinki and data protection standards regulations.

We selected 53 circulating tumour DNA samples initially analysed for EBV presence using a customised EBV and lymphoma panel. Samples that tested positive for EBV genes *EBER-1*, *EBER-2* and/or *EBNA-2* were selected for whole genome sequencing. For healthy controls, EBV in plasma samples was detected using a quantitative PCR kit supplied by GeneProof (Thermo Fisher Scientific). Plasma EBV DNA copies per mL were calculated from the standard curve and multiplied by $10^3$. Samples with at least 1000 viral copies per mL were considered for whole-genome sequencing. Although 79 samples were initially analysed, only 11 had viral copies exceeding 1000 per ml, and these were prioritised for whole-genome sequencing. So, a total of 64 EBV genomes were sequenced in this study. Cell-free DNA was extracted using QIAamp Circulating Nucleic Acid Kit (Qiagen), and libraries were prepared using ThruPLEX Tag-Seq Kit (Takara Bio). We performed EBV capture and enrichment using an Integrated DNA Technologies (IDT) -manufactured custom EBV whole-genome panel, which comprised 2,862 probes, each 120 bp long, designed with a 60 bp overlap. The pooled libraries, a maximum of 6, were then normalised prior to sequencing on the Illumina MiSeq platform. Paired-end FASTQ files were uploaded to Illumina Basespace and later analysed on a secure web-based server, AWS.




**Suppl. Fig. 1 Experimental design and analysis**

Diagrammatic scheme of the study recruitment, sample processing, and analysis. First, circulating cell-free DNA was analysed using a customised panel targeting the MYC, Immunoglobulin loci, and frequently mutated genes in lymphoma (Panel B). The panel, primarily diagnostic, included abundantly expressed EBV genes, such as EBER-1, EBER-2, and the latent gene EBNA-2 probes, to aid in the diagnosis of EBV. Results showing EBER 1, EBER 2, and EBNA-2 positivity were interpreted as indicative of a Type 1 EBV infection. Conversely, the absence of EBNA-2 in the presence of EBERs was interpreted as a Type 2 strain infection. Samples (BL) with sufficient quantities (> 2 EBV copies per cell) were considered positive and prioritised for EBV whole genome sequencing. Forty-one BL cases, along with twelve other lymphomas, including HL and DLBCL, were selected and analysed in the present study.

## Section 2. EBV Whole Genome Data Processing

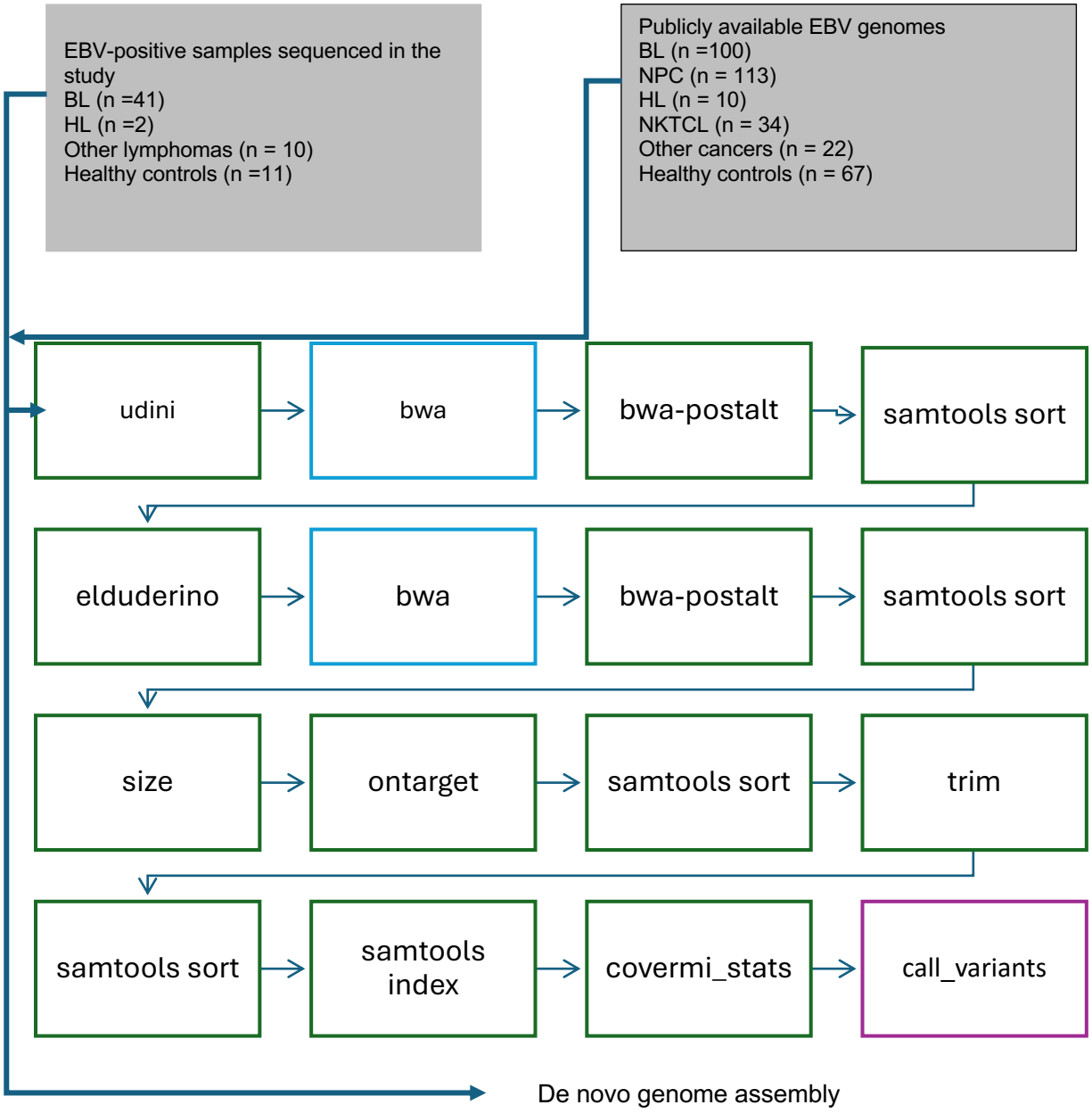
We combined sequencing data from our study with 346 publicly available sequencing datasets downloaded from the Sequence Read Archive (SRA), resulting in a total of 410 high-quality genomes. The public genome set was primarily generated from three studies: a case-control study in Kenya [2], an NPC population study in China[3] , and a hospital-based cohort of lymphoma samples from Malawi, Guatemala, Peru, Taiwan, and the USA [4].

FASTQ files were trimmed for adapter sequences and low-quality bases (Phred score > 20), followed by aligning the reads to the EBV reference genomes (NC_007605.1 for type 1, NC_009334.1 for type 2), as well as a custom hybrid reference (NC_007605.1 combined with EBNA2, EBNA3s contigs from NC_009334.1) using the BWA-MEM alignment method[5]. The

aligned reads were sorted and indexed for further analysis. Variant calling was conducted on the aligned BAM files using three different tools: VarScan[6], VarDict[7], and Mutect2[8]. Functional annotation was performed using SNPEff version 5.1[9], with annotation databases for both type 1 (NC_007605.1) and type 2 (NC_009334.1) reference genomes. The final VCF files contained variants identified by at least two callers and filtered for a mapping quality score of 60 or higher.

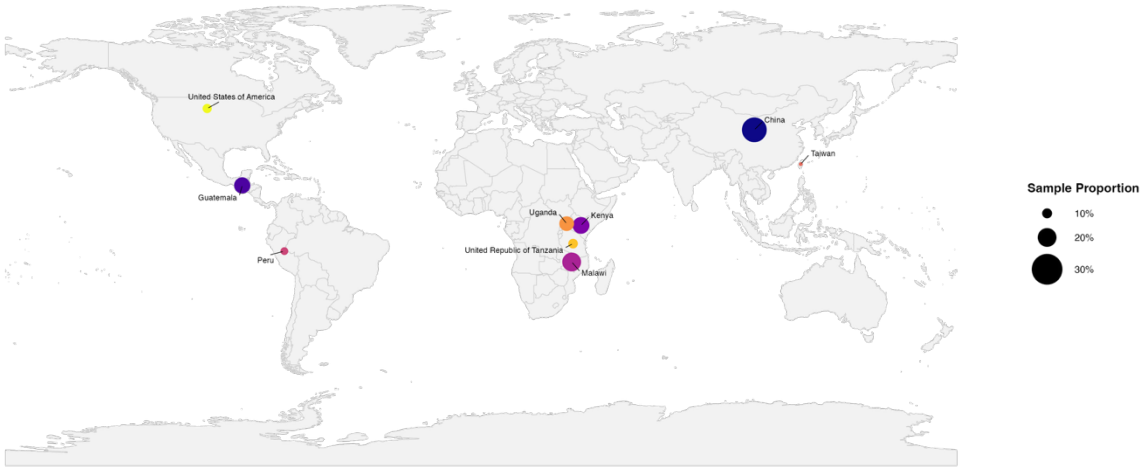
Genome assembly was performed using SPAdes v4.2.0, which constructs a de Bruijn graph by decomposing sequencing reads into k-mers of multiple sizes to assemble contigs[10]. In our dataset, the average read length was 143 bp, and the default k-mer sizes used by SPAdes were 21, 33, and 55 bp. The assembly quality was assessed with QUAST v5.3.0, a genome quality evaluation tool[11]. A composite score was calculated based on four QUAST parameters: genome fraction (GF%) $\geq$ 70%, misassemblies $\leq$ 1, mismatches per 100 kb < 1000, and GC content within 2% of the reference, as well as a duplication ratio ≤1. Contigs achieving an overall score of 50% or higher were retained for further analysis.

To order contigs against the relevant reference genome, a BLAST database was created using *EBNA-2* EBV types 1 and 2 contigs, and the contigs were aligned to each reference using blastn, part of the NCBI BLAST+suite[12]. The resulting *EBNA-2* matches were used to group the contigs into types 1 and 2. This categorisation was crucial for the public genomes, which lacked sample metadata information. Subsequently, the assembled contigs were ordered and oriented according to the reference genome using ABACAS[13], resulting in a single pseudomolecule FASTA sequence. The assembly quality, including genome coverage, alignment rate, mean depth, and average base quality of aligned reads, was evaluated with minimap2 v2.29-r1283[14]. Contigs were trimmed to eliminate leading and trailing ambiguous bases, along with other sequencing artefacts, using trimAl v1.4 rev 15[15]. Only assemblies with at least 60% coverage of the EBV reference genome and a mean mapping quality of at least 30 were selected for phylogenetic analysis.




**Suppl. Fig. 2 Overview of the bioinformatics pipeline**

Diagrammatic representation of the analysis pipeline. Paired-end FASTQ files were initially adapter-trimmed and filtered for quality (Phred score >20), then aligned to an EBV reference genome (NC_007605.1, for Type 1 and NC_009334.1, for Type 2). The steps in green are QC or pre-processing steps. Blue represents alignment steps. Purple indicates variant calling. Only variants present in at least two callers were considered for downstream analysis. FASTQ files that passed quality filters were used to assemble viral genomes using de novo tools.




**Suppl. Fig. 3 Geographic origins of the EBV genomes analysed in the study**

Map showing the geographic origins and proportion of samples analysed in the study. EBV genomes represented five regions: Africa, Asia, South America, Central America, and North America. Burkitt lymphoma samples were almost exclusively from Africa, with 21.3% from Uganda, 7.8% from Tanzania, 23.4% from Kenya, and 53.2% from Malawi. NPCs were solely from Asia, while 79.4% of the NKTCL were from the Americas. Healthy controls came from Uganda (14.1%), Kenya (35.9%), and Asia (50%). The distribution of other malignancies varied across the five regions.

## Section 3: FFPE RNA Extraction, Library Preparation, and Sequencing

We extracted total RNA using the RNeasy FFPE kit (Qiagen) according to the manufacturer's instructions. RNA samples were quantified using the Qubit 4.0 Fluorometer (Life Technologies, Carlsbad, CA, USA), and RNA integrity was checked with the RNA Kit on the Agilent 5300 Fragment Analyser (Agilent Technologies, Palo Alto, CA, USA). rRNA depletion was performed using the NEBNext rRNA Depletion Kit (H/M/R). Sequencing libraries were prepared using NEBNext Ultra II RNA Library Prep Kit for Illumina, following the manufacturer's recommendations for degraded FFPE-derived total RNA (NEB, Ipswich, MA, USA). Briefly, first and second strand cDNA were synthesised. cDNA fragments were end-repaired, adenylated at the 3' ends, and a universal adapter was ligated to the cDNA fragments, followed by index addition and library enrichment with limited-cycle PCR. Sequencing libraries were validated using the NGS Kit on the Agilent 5300 Fragment Analyser (Agilent Technologies, Palo Alto, CA, USA) and quantified using the Qubit 4.0 Fluorometer (Invitrogen, Carlsbad, CA). The sequencing libraries were multiplexed and loaded onto the Illumina NovaSeq XPlus instrument following the manufacturer's instructions. The samples were sequenced using a 2x150 Pair-End (PE) configuration. The NovaSeq Control Software

v1.3 conducted image analysis and base calling. Raw sequence data generated from Illumina NovaSeq were converted into fastq files and de-multiplexed using the Illumina bcl2fastq programme. One mismatch was permitted for index sequence identification.

## Section 4. EBV Transcriptome Data Processing and Analysis

Paired-end RNA-Seq reads were first trimmed for adapters and low-quality bases using standard Illumina trimming tools. Transcript-level quantification was performed using Kallisto (v0.51.1)[16]. Reads were pseudo-aligned to the hybrid human-EBV transcriptome index to generate abundance estimates. The transcript-level abundance of all samples from Kallisto was aggregated to gene-level counts and imported using the tximport function in R, along with the gene annotation. Gene expression analysis was performed in DESeq2 (v1.48.1)[17]. Normalised counts were used to generate heatmaps with ComplexHeatmap (v2.24.1)[18]

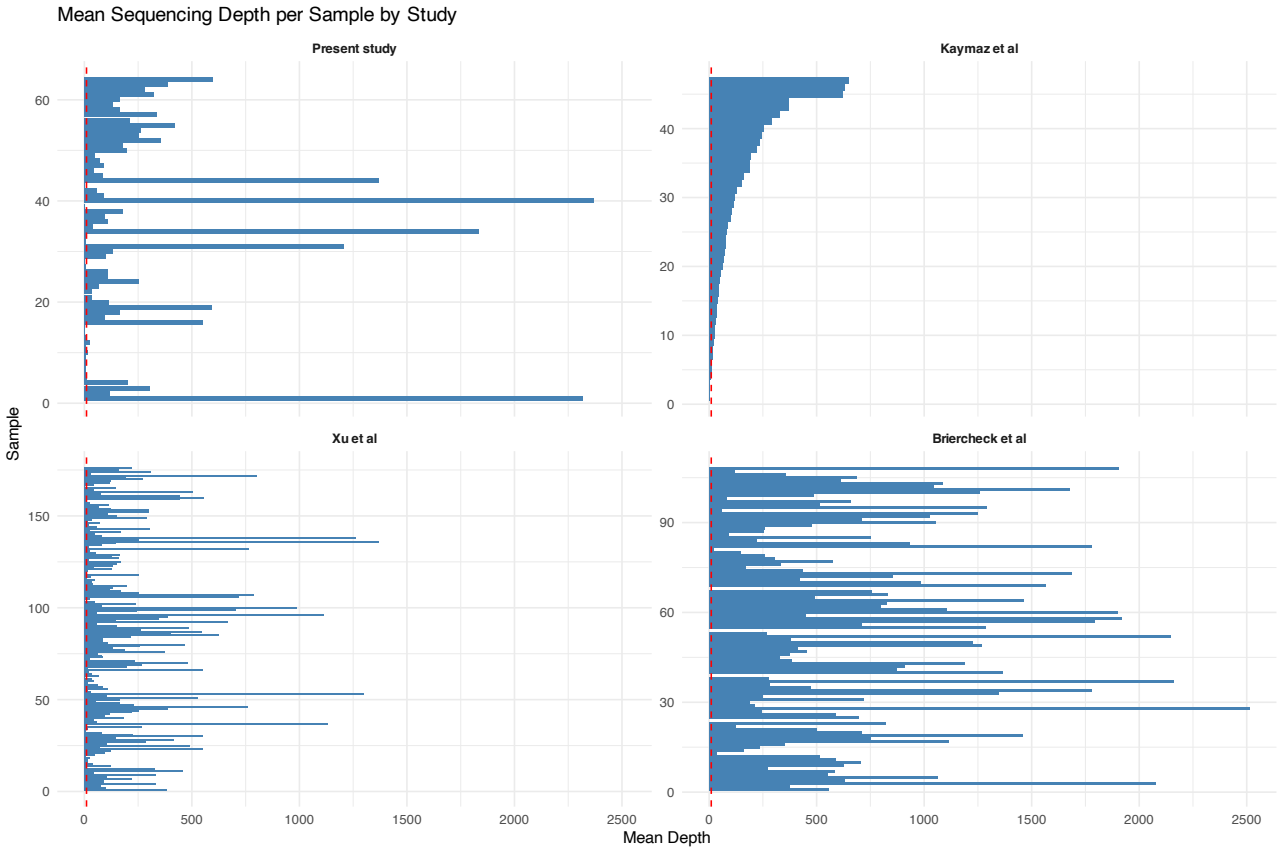
## Supplementary data, Tables and Figures

**Suppl. Table 1. Demographic and clinical features of cases compared with EBV-positive healthy controls**

| Characteristic | Overall N = 64[1] | BL N = 41[1] | Healthy control N = 11[1] | HL N = 2[1] | Other lymphoma N = 10[1] | p-value[2] |
|---|---|---|---|---|---|---|
| **Age, years** | 9.0 (6.0, 11.0) | 8.0 (5.0, 11.0) | 8.0 (7.0, 10.0) | 11.5 (10.0, 13.0) | 9.5 (4.0, 11.0) | 0.5 |
| **Sex** | | | | | | 0.5 |
| Female | 23 (36%) | 14 (34%) | 6 (55%) | 0 (0%) | 3 (30%) | |
| Male | 41 (64%) | 27 (66%) | 5 (45%) | 2 (100%) | 7 (70%) | |
| **Country** | | | | | | 0.018 |
| Tanzania | 15 (23%) | 11 (27%) | 0 (0%) | 2 (100%) | 2 (20%) | |
| Uganda | 49 (77%) | 30 (73%) | 11 (100%) | 0 (0%) | 8 (80%) | |
| **Weight loss** | 46 (72%) | 33 (80%) | 1 (9.1%) | 2 (100%) | 10 (100%) | <0.001 |
| **Cachexia** | 18 (28%) | 11 (27%) | 0 (0%) | 2 (100%) | 5 (50%) | 0.004 |
| **Night sweats** | 32 (50%) | 21 (51%) | 0 (0%) | 2 (100%) | 9 (90%) | <0.001 |
| **Tumour site** | | | | | | <0.001 |
| Abdominal | 29 (55%) | 22 (54%) | 0 (NA%) | 1 (50%) | 6 (60%) | |
| Axilla | 2 (3.8%) | 1 (2.4%) | 0 (NA%) | 1 (50%) | 0 (0%) | |
| Inguinal | 1 (1.9%) | 0 (0%) | 0 (NA%) | 0 (0%) | 1 (10%) | |
| Jaw | 16 (30%) | 16 (39%) | 0 (NA%) | 0 (0%) | 0 (0%) | |
| Neck | 3 (5.7%) | 0 (0%) | 0 (NA%) | 0 (0%) | 3 (30%) | |
| Other | 2 (3.8%) | 2 (4.9%) | 0 (NA%) | 0 (0%) | 0 (0%) | |
| **HIV** | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| **Sickle cell** | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| **Tumour stage** | | | | | | 0.7 |
| stage a | 3 (7.5%) | 3 (8.3%) | 0 (NA%) | 0 (NA%) | 0 (0%) | |
| stage b | 9 (23%) | 9 (25%) | 0 (NA%) | 0 (NA%) | 0 (0%) | |
| stage c | 20 (50%) | 17 (47%) | 0 (NA%) | 0 (NA%) | 3 (75%) | |
| stage d | 8 (20%) | 7 (19%) | 0 (NA%) | 0 (NA%) | 1 (25%) | |

[1]Median (Q1, Q3); n (%)

[2]Kruskal-Wallis rank sum test; Fisher's exact test



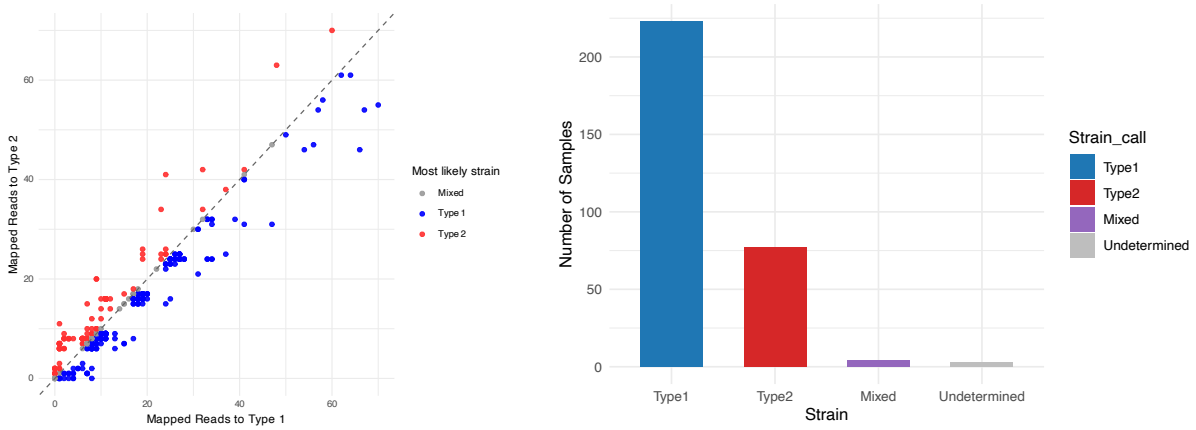

**Suppl. Fig. 4 Sequencing metrics**

The plots show sequencing metrics of the present study and three other studies that contributed the public EBV genome data sets analysed in the study. The mean number of reads and base coverage was 368126.3 and 159001.5, achieving a mean sequencing depth of 291x and 92.5% genome coverage in the present study. In comparison, the corresponding metrics were 141.8 x and 75.04% in the study by Kaymaz et al., 220.6 x and 91.6% in the study by Xu et al., and 776.92 x and 98.6% in the study by Briercheck et al, respectively. Individual samples with a mean sequencing depth of at least 10x and 75% genome coverage across the studies were selected for evaluation.

**Suppl. Table 2. Reproducibility of variant calling between two technical replicates**

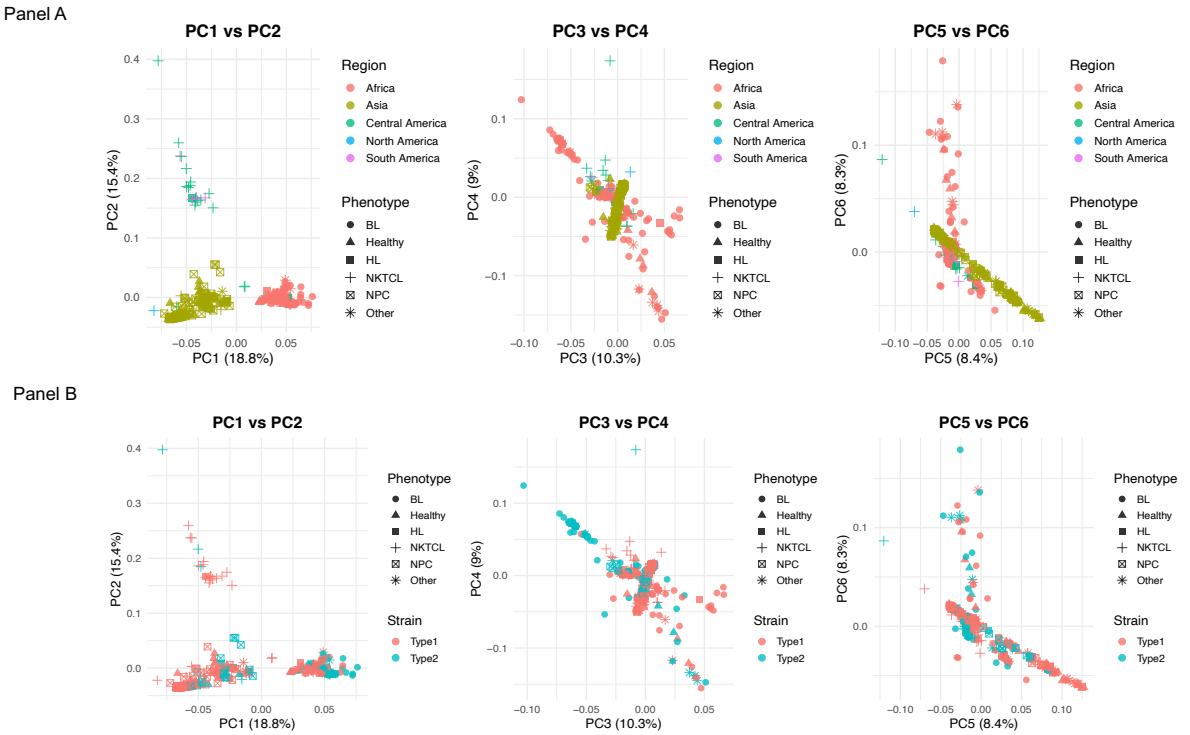
| Sample | Run1 | Run2 | AbsDiff | PercDiff | Mean | SD | CV (%) |
|---|---|---|---|---|---|---|---|
| MNH011 | 749 | 766 | 17 | 2.24 | 757.5 | 12.02 | 1.59 |
| MNH052 | 600 | 607 | 7 | 1.16 | 603.5 | 4.95 | 0.82 |
| KCMC047 | 683 | 721 | 38 | 5.41 | 702 | 26.87 | 3.83 |
| BMC035 | 709 | 679 | 30 | 4.32 | 694 | 21.21 | 3.06 |

Samples were processed in duplicate, and variant calling performed to check accuracy. The absolute difference (AbsDiff), percentage difference (PercDiff), mean variant count, standard deviation (SD), and coefficient of variation (CV) of variant counts from the two runs were calculated.




**Suppl. Fig. 5 Plots show the metric performance for EBV strain calls in public genomes**

Scatter plot [A] of EBV contigs aligned to the Type 1 and 2 EBNA2 reference. Bar plot [B] shows the number of genotypes called as type 1, type 2, mixed or undetermined by the metric. Viral contigs were aligned to a custom EBNA2 reference contig, and hits with the highest matches were assigned a strain call. This way, 266 (76.9%) of the public EBV genomes were classified as type 1 and 80 (23.1%) as type 2. The mixed or undetermined genomes were removed from further analysis.




**Suppl. Fig. 6 Principal component analysis (PCA) of 410 Epstein–Barr virus (EBV) genomes**

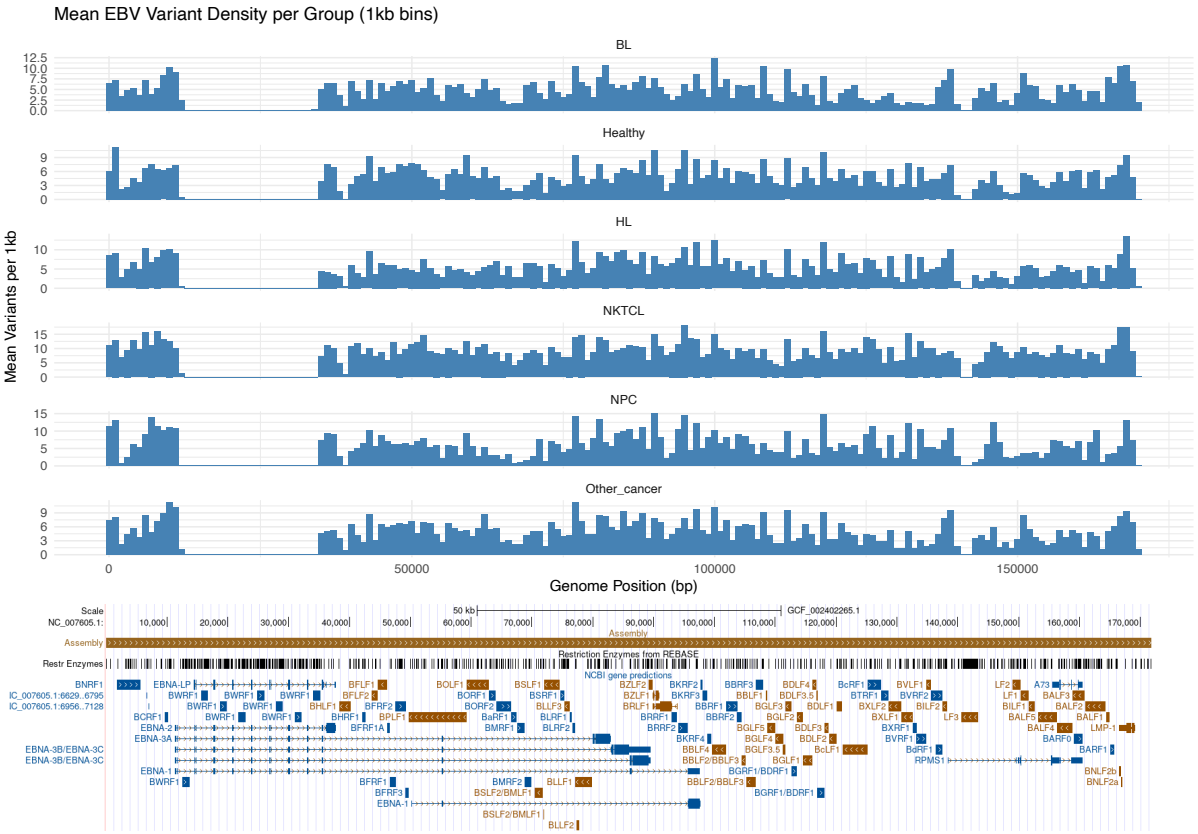
**Panel A:** Scatter plots show PC1 vs PC2, PC3 vs PC4, and PC5 vs PC6, with axes labelled by the percentage of variance explained. Points are coloured by geographic region (Africa, Asia, Central America, North America, South America) and shaped by clinical phenotype (BL, Healthy, HL, NKTCL, NPC, Other). PC1 and PC2 primarily separate genomes by regional origin, while higher-order PCs capture finer substructure and potential within-region variation

**Panel B:** Scatter plots show PC1 vs PC2, PC3 vs PC4, and PC5 vs PC6 with points coloured by strain (Type 1 vs Type 2) and shaped by host phenotype. The first six PCs explain 18.8%, 15.4%, 10.3%, 9.0%, 8.4%, and 8.3% of the variance, respectively. Minimal separation is observed by strain, indicating that geographic variation rather than strain type is the dominant driver of viral diversity.

**Suppl. Table 3. Putative recombination events detected in EBV genomes derived from BL and EBV-positive healthy controls**

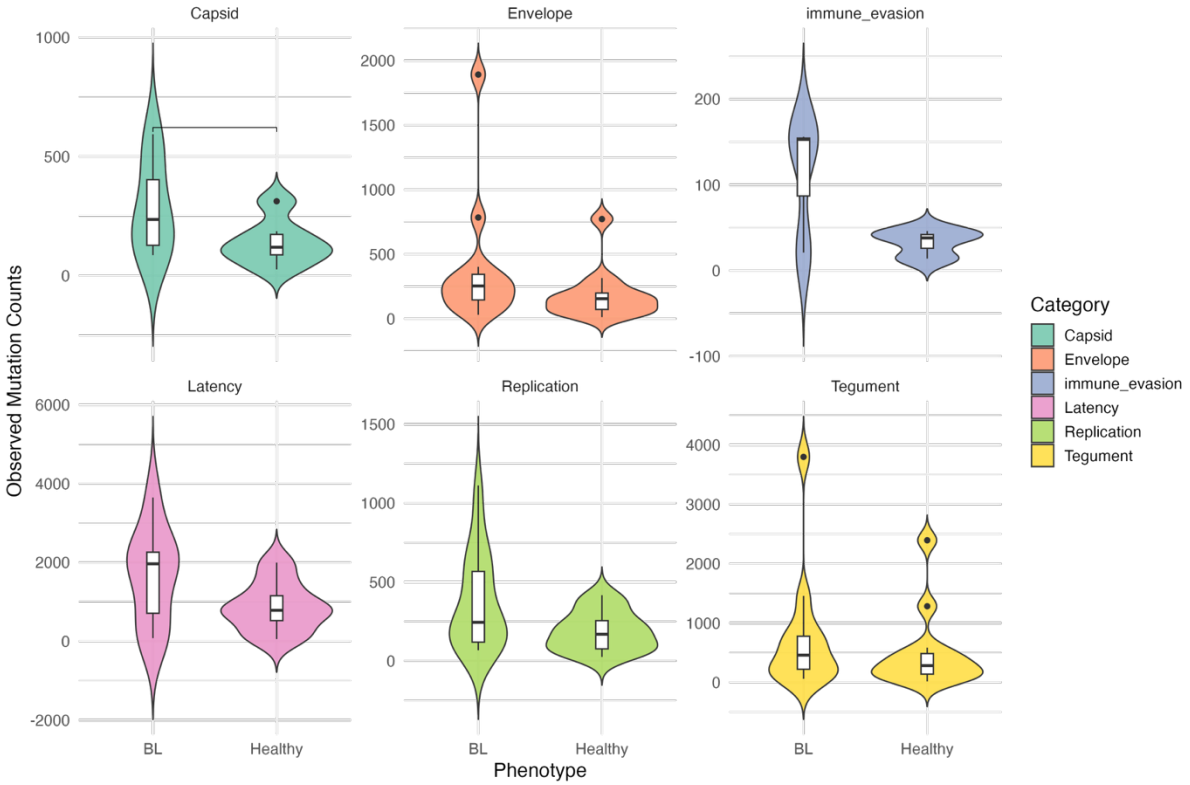
| Event(n) | RDP | GENECONV | MaxChi | Chimaera | 3Seq | Bootscan | SiSscan |
|---|---|---|---|---|---|---|---|
| 1 | 3.93E-31 | 4.07E-08 | 2.26E-11 | 4.22E-10 | 7.44E-11 | NS | 2.59E-20 |
| 2 | 2.64E-29 | 2.90E-29 | 8.28E-11 | 1.22E-10 | 7.44E-11 | 1.22E-21 | 6.61E-20 |
| 3 | 5.65E-28 | 1.53E-27 | 1.56E-07 | 5.43E-08 | 7.44E-11 | 1.35E-27 | 2.45E-08 |
| 5 | 7.57E-26 | 6.80E-22 | 2.46E-12 | 6.00E-13 | 7.44E-11 | 1.42E-06 | 6.01E-03 |
| 6 | 1.60E-24 | 4.49E-25 | 4.26E-15 | 6.97E-10 | 1.48E-23 | NS | 3.07E-32 |
| 7 | 7.49E-23 | 1.30E-12 | 2.58E-11 | 1.37E-12 | 7.44E-11 | NS | 7.58E-09 |
| 8 | 5.67E-28 | 1.04E-21 | 2.00E-14 | 9.44E-12 | 2.85E-20 | NS | 1.12E-09 |
| 9 | 3.42E-24 | 8.98E-24 | 1.38E-11 | 9.92E-12 | 1.28E-22 | 5.48E-07 | 1.36E-12 |
| 10 | 2.25E-23 | 8.65E-18 | 2.56E-09 | 8.51E-10 | 7.44E-11 | 1.58E-04 | 4.53E-10 |
| 11 | 5.32E-22 | 6.32E-16 | 2.04E-10 | 1.37E-10 | 1.49E-10 | NS | 5.09E-05 |
| 12 | 1.96E-20 | 1.02E-13 | 4.64E-06 | 1.61E-04 | 9.99E-03 | NS | 4.01E-03 |
| 13 | 5.57E-20 | 5.48E-09 | 1.52E-14 | 9.78E-15 | 7.44E-11 | NS | 2.35E-09 |
| 14 | 3.64E-19 | 6.09E-17 | 3.34E-09 | 2.26E-09 | 7.44E-11 | 2.44E-04 | 7.52E-08 |
| 15 | 1.21E-18 | 1.42E-16 | 7.75E-07 | 6.08E-07 | 2.24E-13 | NS | 2.53E-08 |
| 16 | 1.47E-19 | 8.03E-15 | 2.73E-07 | 3.24E-07 | 7.44E-11 | 1.75E-11 | 1.84E-07 |
| 17 | 1.35E-16 | 2.57E-09 | 1.23E-06 | 1.35E-10 | 7.44E-11 | NS | 2.08E-03 |
| 18 | 2.90E-15 | 3.59E-12 | 1.35E-09 | 6.62E-05 | 2.23E-10 | NS | 1.29E-07 |
| 19 | 2.95E-15 | 2.91E-09 | 8.93E-03 | 1.53E-03 | 5.95E-10 | NS | 4.06E-02 |
| 21 | 4.17E-15 | 1.75E-07 | 1.95E-04 | NS | 8.18E-07 | 1.41E-04 | NS |
| 22 | 5.61E-14 | 5.14E-14 | 1.45E-08 | 8.44E-09 | 8.51E-06 | NS | 1.07E-09 |
| 23 | 9.55E-14 | 1.58E-11 | 1.28E-13 | 3.00E-11 | 7.44E-11 | NS | 2.22E-06 |
| 24 | 1.72E-13 | 8.89E-11 | 1.25E-03 | 1.10E-03 | 1.06E-08 | 1.28E-03 | NS |
| 25 | 2.70E-13 | 9.88E-14 | 5.58E-10 | 5.09E-09 | 3.58E-14 | NS | 1.82E-11 |
| 27 | 3.56E-13 | 4.63E-11 | 6.47E-10 | 9.02E-07 | 1.97E-08 | 1.91E-04 | 2.96E-10 |
| 28 | 1.45E-12 | 8.54E-11 | 1.11E-03 | 1.06E-03 | 1.88E-07 | NS | 3.12E-03 |
| 29 | 4.03E-02 | NS | 6.22E-11 | 2.22E-04 | 7.66E-09 | NS | 3.16E-02 |
| 31 | 6.48E-11 | 8.56E-08 | 2.43E-06 | 3.66E-06 | 8.53E-03 | 6.50E-11 | 2.35E-08 |
| 32 | 2.41E-10 | 7.15E-05 | 6.37E-06 | 1.14E-06 | 5.25E-07 | NS | 3.79E-04 |
| 33 | 3.25E-10 | 1.54E-07 | 1.11E-05 | 7.84E-05 | 1.01E-04 | NS | 8.46E-09 |
| 34 | 4.16E-11 | 6.90E-10 | 1.24E-08 | 3.53E-07 | 3.28E-05 | NS | 2.11E-11 |
| 36 | 1.22E-09 | 3.08E-05 | 3.36E-07 | 5.66E-06 | 1.91E-08 | NS | 7.61E-04 |
| 37 | 7.51E-03 | NS | 4.19E-09 | 2.26E-07 | 1.13E-06 | NS | 2.78E-35 |
| 38 | 1.58E-08 | 1.41E-06 | 1.68E-05 | 1.14E-05 | 6.56E-06 | NS | 1.37E-09 |
| 39 | 1.28E-08 | 1.81E-07 | 1.71E-04 | 3.21E-04 | 3.80E-02 | NS | 7.66E-13 |
| 40 | 3.43E-04 | 3.07E-03 | 5.48E-07 | 9.46E-05 | 4.93E-03 | NS | 7.03E-06 |
| 41 | 6.54E-07 | 7.23E-03 | 6.81E-03 | 7.81E-03 | 1.02E-02 | 1.59E-03 | NS |
| 48 | 5.02E-06 | 1.66E-04 | 1.78E-03 | 1.64E-03 | NS | NS | 2.19E-04 |
| 49 | 1.72E-05 | 4.84E-05 | 7.17E-03 | 3.95E-03 | 2.88E-05 | NS | 1.01E-03 |
| 54 | 3.58E-05 | 1.20E-03 | 7.67E-03 | 5.98E-03 | NS | NS | 1.90E-02 |
| 55 | 4.09E-05 | 6.91E-05 | 2.67E-03 | 1.44E-03 | 6.61E-04 | NS | 1.30E-02 |
| 56 | 5.35E-05 | 7.92E-04 | 3.45E-02 | 4.59E-02 | 4.48E-03 | NS | 3.80E-04 |
| 57 | 1.56E-03 | 5.45E-05 | 3.60E-03 | 1.93E-04 | 8.67E-06 | 1.50E-03 | 3.78E-07 |
| 62 | 8.73E-05 | 3.11E-02 | 1.30E-02 | 4.99E-02 | 3.05E-04 | NS | 5.19E-03 |

Events detected by at least five methods. RDP and Bootscan are phylogeny-based. The rest are statistical models. *P* values with Bonferroni correction are shown. NS- No significant P-value was recorded for this recombination event using this method.



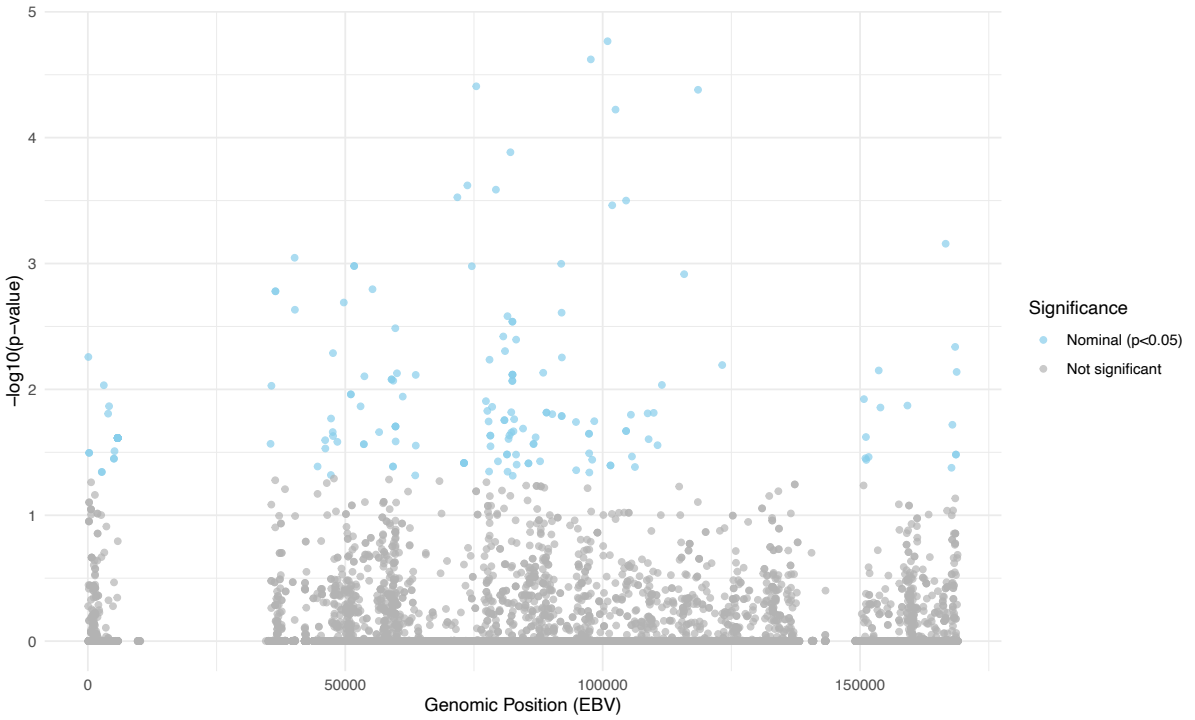

**Suppl. Fig. 7 Mutation frequency across EBV genomes**

The bar plots depict the average EBV variant density per 1 kb window across the viral genome, organised by sample group. Each bar represents the mean number of variants per 1 kb genomic window (NC_007605.1 reference) for all samples within that group. Facets illustrate group-specific variant density profiles, highlighting regions of increased mutation (hypervariable regions) and conserved segments (mutation deserts). The areas with zero counts correspond to regions of poor quality that could not be aligned to the reference genomes. Below the bar plot, the NCBI gene annotation track for EBV NC_007605.1 displays the relative positions of the EBV genes.




**Suppl. Fig. 8 Distribution of EBV gene mutations by functional group in BL and Healthy samples**

Violin plots with overlaid boxplots show the observed mutation counts per sample for six EBV gene categories: Capsid, Envelope, Immune Evasion, Latency, Replication, and Tegument. Pairwise comparisons between BL and healthy samples were performed using the Wilcoxon rank-sum test, and p-values were adjusted using the Benjamini-Hochberg FDR method. The lowest unadjusted p-value was observed in the Replication category ($p = 0.029$, FDR = 0.174, Cliff's $\Delta = 0.386$), but no category reached statistical significance after FDR correction. Effect sizes (Cliff's Delta) indicated small-to-moderate differences, with the largest effect observed in the Immune Evasion category ($\Delta = 0.556$).



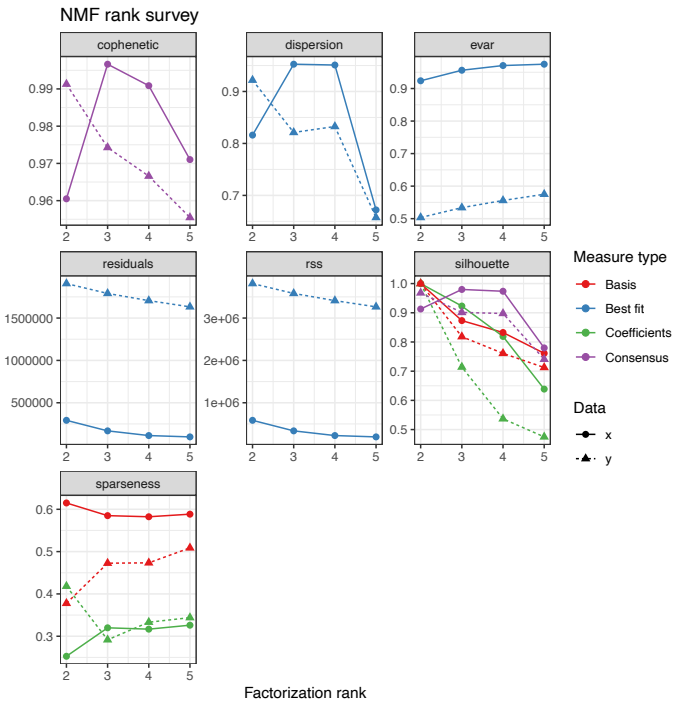

**Suppl. Fig. 9 EBV genome-wide association study of BL and Healthy controls**

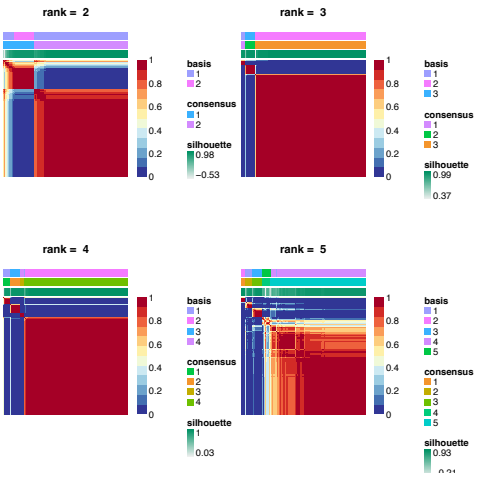
Manhattan plot of logistic regression results comparing EBV genomes from Burkitt Lymphoma (BL) and Healthy Controls, adjusted for Region, Country, EBV strain and population structure (PC1 &2). Each point represents a tested genomic locus on the EBV genome, with the x-axis showing genomic position and the y-axis the $-\log_{10}$(p-value). Loci reaching nominal significance are annotated with blue dots. Grey dots represent non-significant loci.

## Suppl. Table 4. EBV-genome-wide association study of BL and EBV-positive healthy controls

| Position | Ref/Alt genotype | Alt freq in BL | Alt freq in healthy | estimate | St. Error | statistic | p.value | OR | FDR | Gene annotation | Amino acid change |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 100957 | G>A | 95.0% | 66.7% | 2.448 | 0.569 | 4.299 | 1.71E-05 | 11.57 | 1.16E-01 | BBLF4 | p.Ala339Thr |
| 97698 | T>C | 82.3% | 51.3% | 4.532 | 1.073 | 4.225 | 2.39E-05 | 92.91 | 1.16E-01 | BKRF2 | p.Ile10Thr |
| 75448 | A>G | 95.0% | 71.8% | 2.534 | 0.616 | 4.113 | 3.91E-05 | 12.61 | 1.16E-01 | BLLF3 | p.Thr247Ala |
| 118531 | T>C | 88.7% | 61.5% | 2.284 | 0.557 | 4.098 | 4.17E-05 | 9.81 | 1.16E-01 | BDLF3 | p.Phe83Ser |
| 102502 | A>G | 97.9% | 75.6% | 3.007 | 0.749 | 4.013 | 5.99E-05 | 20.23 | 1.34E-01 | BBRF1 | p.Lys196Arg |
| 82084 | T>C | 55.3% | 10.3% | 2.927 | 0.765 | 3.825 | 1.31E-04 | 18.67 | 2.43E-01 | EBNA-3A | p.Val681Ala |
| 73726 | A>G | 75.2% | 59.0% | 1.890 | 0.515 | 3.673 | 2.40E-04 | 6.62 | 3.50E-01 | BSLF1 | p.Met290Val |
| 79265 | G>C | 46.1% | 56.4% | 2.346 | 0.642 | 3.653 | 2.59E-04 | 10.44 | 3.50E-01 | BLLF1 | p.Glu201Gln |
| 71780 | C>A | 99.3% | 80.8% | 4.305 | 1.190 | 3.617 | 2.98E-04 | 74.10 | 3.50E-01 | BSLF2/BMLF1 | p.Leu39Met |
| 104550 | T>G | 92.2% | 65.4% | 1.856 | 0.515 | 3.602 | 3.16E-04 | 6.40 | 3.50E-01 | BBLF2/BBLF3 | p.Ile691Met |
| 101862 | G>A | 89.4% | 64.1% | 1.817 | 0.508 | 3.579 | 3.45E-04 | 6.15 | 3.50E-01 | BBLF4 | p.Arg37Lys |
| 166625 | T>G | 98.6% | 85.9% | 3.320 | 0.979 | 3.391 | 6.96E-04 | 27.67 | 6.48E-01 | BNLF2b | p.Phe71Cys |
| 40166 | T>G | 78.7% | 18.0% | 2.156 | 0.649 | 3.320 | 9.00E-04 | 8.63 | 7.33E-01 | BHLF1 | p.Leu35Arg |
| 91950 | A>G | 49.7% | 16.7% | 2.184 | 0.664 | 3.289 | 1.01E-03 | 8.88 | 7.33E-01 | BRLF1 | p.Lys316Glu |
| 51733 | A>G | 0.0% | 1.3% | -2.234 | 0.682 | -3.278 | 1.05E-03 | 0.11 | 7.33E-01 | BPLF1 | p.Thr2503Ala |
| 51733 | dupC | 0.0% | 1.3% | -2.234 | 0.682 | -3.278 | 1.05E-03 | 0.11 | 7.33E-01 | BPLF1 | p.Thr2503fs |
| 74587 | G>A | 36.9% | 51.3% | 2.683 | 0.819 | 3.277 | 1.05E-03 | 14.63 | 7.33E-01 | BSLF1 | p.Ala3Thr |
| 115832 | T>G | 99.3% | 84.6% | 3.773 | 1.166 | 3.235 | 1.22E-03 | 43.53 | 7.98E-01 | BGLF1 | p.Asp85Glu |
| 55297 | A>G | 97.2% | 80.8% | 2.323 | 0.736 | 3.156 | 1.60E-03 | 10.21 | 9.76E-01 | BPLF1 | p.Ile1315Val |
| 36440 | delCCC | 0.7% | 0.0% | -3.498 | 1.112 | -3.145 | 1.66E-03 | 0.03 | 9.76E-01 | EBNA-2 | p.Pro77del |
| 36440 | delCCC | 0.7% | 0.0% | -3.498 | 1.112 | -3.145 | 1.66E-03 | 0.03 | 9.76E-01 | EBNA-2 | p.Pro77del |
| 49713 | C>A | 96.5% | 84.6% | 2.336 | 0.757 | 3.084 | 2.04E-03 | 10.34 | 1.00E+00 | BFRF3 | p.His165Gln |
| 40218 | C>T | 73.1% | 2.6% | 2.661 | 0.874 | 3.045 | 2.33E-03 | 14.31 | 1.00E+00 | BHLF1 | p.His18Tyr |
| 92028 | G>T | 45.4% | 2.6% | 2.456 | 0.811 | 3.028 | 2.46E-03 | 11.66 | 1.00E+00 | BRLF1 | p.Ala290Ser |
| 81517 | T>C | 63.1% | 59.0% | 1.789 | 0.594 | 3.009 | 2.62E-03 | 5.98 | 1.00E+00 | EBNA-3A | p.Phe492Ser |
| 82473 | A>G | 1.4% | 34.6% | 1.265 | 0.425 | 2.978 | 2.90E-03 | 3.54 | 1.00E+00 | EBNA-3A | p.Thr811Ala |
| 82473 | A>G | 1.4% | 34.6% | 1.265 | 0.425 | 2.978 | 2.90E-03 | 3.54 | 1.00E+00 | EBNA-3A | p.Thr811Ala |
| 59726 | A>C | 1.4% | 14.1% | -2.783 | 0.946 | -2.941 | 3.27E-03 | 0.06 | 1.00E+00 | BOLF1 | p.Ile1076Leu |
| 80698 | T>C | 30.5% | 56.4% | 1.914 | 0.661 | 2.895 | 3.80E-03 | 6.78 | 1.00E+00 | EBNA-3A | p.Leu219Pro |
| 83192 | A>C | 35.5% | 2.6% | 2.803 | 0.975 | 2.876 | 4.02E-03 | 16.50 | 1.00E+00 | EBNA-3B/EBNA-3C | p.Glu43Ala |
| 168476 | G>T | 76.6% | 10.3% | 1.441 | 0.509 | 2.834 | 4.60E-03 | 4.23 | 1.00E+00 | LMP-1 | p.Met129Ile |
| 81039 | A>C | 65.3% | 62.8% | 1.770 | 0.630 | 2.809 | 4.96E-03 | 5.87 | 1.00E+00 | EBNA-3A | p.Ile333Leu |

Variants with normal statistical significance ($p < 0.05$) are shown.



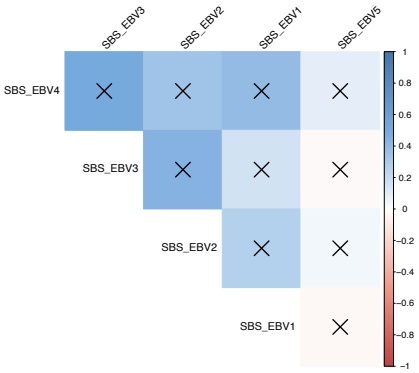




**Suppl. Fig. 11 Consensus clustering heatmaps for NMF ranks 2 to 5 in EBV mutational signature analysis.**

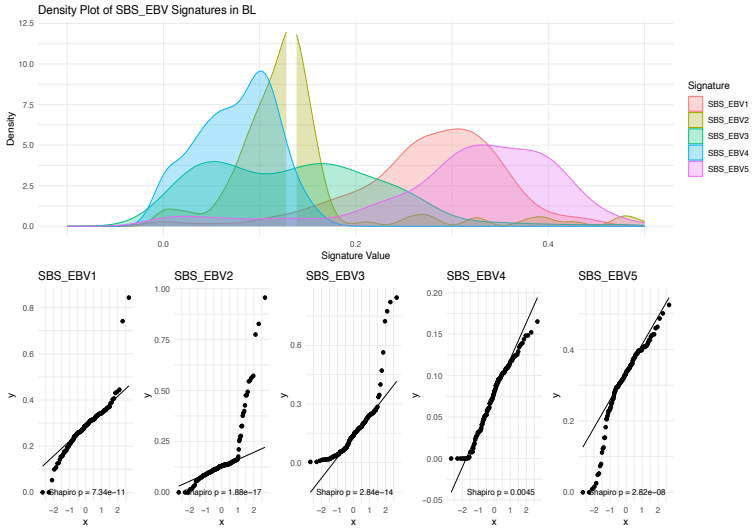
Each panel displays the consensus matrix and basis matrix clustering results for ranks k = 2 to k = 5. Columns and rows represent samples, with colours indicating co-clustering frequency across NMF runs. Silhouette widths are shown for each rank, reflecting cluster stability. Rank k = 5 yielded a high silhouette score (0.93) and distinct consensus blocks, supporting its selection as the optimal number of mutational signatures.

**Suppl. Fig. 10 Non-negative matrix factorisation (NMF) rank survey for EBV mutational signature analysis**

Quality metrics were evaluated across NMF ranks k = 2 to 5, including cophenetic correlation, dispersion, silhouette width, residual sum of squares (RSS), explained variance (evar), and sparseness of the basis and coefficient matrices. Rank k = 5 was selected as the optimal factorisation rank based on the highest cophenetic correlation before a sharp drop, high silhouette width, and a favourable balance between RSS and explained variance, indicating stable and interpretable mutational signatures.

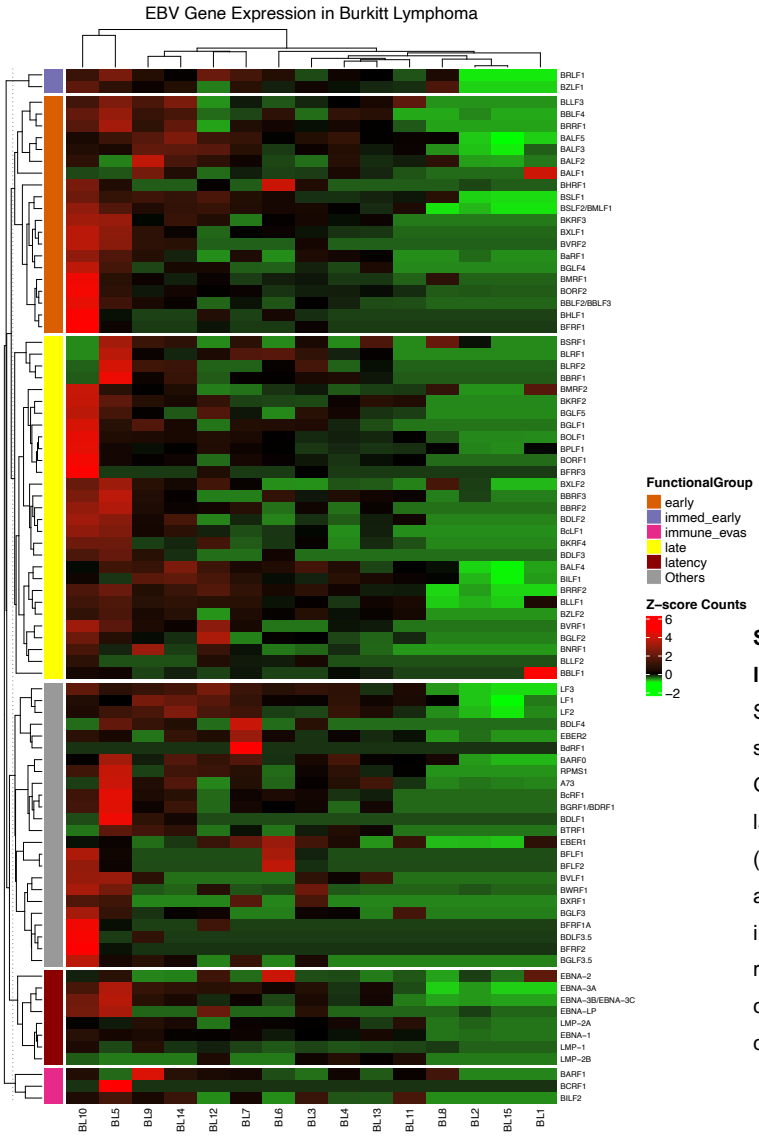



**Suppl. Fig. 12 Spearman correlation matrix for the de novo EBV mutation signatures.** The heatmap displays the pairwise Spearman correlation coefficients (ρ) between five de novo extracted mutational signatures (SBS_EBV1 to SBS_EBV5) across all samples. The colour scale shows the strength and direction of the correlation, with blue indicating positive correlation, red indicating negative correlation, and white indicating no correlation. Strong positive correlations imply co-occurrence or shared mutational processes, while negative correlations may suggest mutually exclusive activity across samples.




**Suppl Fig. 13 Distribution and normality assessment of EBV mutation signature contributions in BL samples**.

Density plots (top) and quantile–quantile (QQ) plots (bottom) for five EBV-specific SBS signatures (SBS_EBV1 to SBS_EBV5) show the distribution of signature contributions in BL samples. Shapiro–Wilk p-values are reported for each signature to assess deviation from normality. All signatures significantly deviate from a normal distribution ($p < 0.05$), indicating that non-parametric methods are appropriate for downstream comparative analyses.




**Suppl. Fig. 14 EBV gene expression in Burkitt lymphoma samples.**

Scaled normalised counts of EBV genes are shown for 15 Burkitt lymphoma samples analysed. Genes are clustered by functional categories: latency (dark red), immediate early (purple), early (orange), late (yellow), immune evasion (pink), and uncharacterized (grey). Columns represent individual Burkitt lymphoma (BL) samples, and rows represent EBV genes. Both rows and columns are hierarchically clustered to highlight co-expression patterns.

References


1. Legason, I.D. *et al.* A protocol to clinically evaluate liquid biopsies as a tool to speed up diagnosis of children and young adults with aggressive infection-related lymphoma in East Africa "(AI-REAL)". *BMC cancer* **22**, 1-9 (2022).
2. Kaymaz, Y. *et al.* Epstein-Barr Virus Genomes Reveal Population Structure and Type 1 Association with Endemic Burkitt Lymphoma. *Journal of virology* **94**, e02007-19 (2020).
3. Xu, M. *et al.* Genome sequencing analysis identifies Epstein–Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. *Nature Genetics* **51**, 1131-1136 (2019).
4. Briercheck, E.L. *et al.* Geographic EBV variants confound disease-specific variant interpretation and predict variable immune therapy responses. *Blood Advances* **8**, 3731-3744 (2024).
5. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314-324 (IEEE, 2019).
6. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576 (2012).
7. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research* **44**, e108-e108 (2016).
8. Benjamin, D. *et al.* Calling somatic SNVs and indels with Mutect2. *Biorxiv*, 861054 (2019).
9. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *fly* **6**, 80-92 (2012).
10. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* **70**, e102 (2020).
11. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).
12. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).
13. Zhou, W. *et al.* ABACUS: An Electronic Structure Analysis Package for the AI Era. *arXiv preprint arXiv:2501.08697* (2025).
14. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
15. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
16. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525-527 (2016).
17. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
18. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).