# Molecular characterisation of Epstein-Barr virus in endemic Burkitt Lymphoma




Draguma Ismail Legason

Department of Oncology, Medical Sciences Division,

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

The Queen's College

Hillary Term 2026

In loving memory of my mother, who didn’t live to see this finish line, but whose love carried me through the race.

# Declaration

I hereby declare that the contents of this thesis are original and have not been submitted wholly or partly for consideration for any other degree or academic qualification at this or any other university. This thesis is a result of my own work; where specifically indicated within the text, the contributions of other team members have been acknowledged. This thesis contains approximately ~ 49036 words and ~383 pages.

Draguma Ismail Legason

2026

# Acknowledgements

My foremost gratitude goes to Professor Anna Schuh. This D.Phil. journey would not have begun without her support from the very start. The rigour of scientific training and mentorship provided by Professor Schuh and members of her lab laid the foundations for this research and will continue to shape my path in academia. The opportunities to learn from her extended well beyond science. In Professor Schuh, I see a role model whose leadership, integrity, and achievements I deeply admire.

The later chapters of this thesis would not have been possible without the support of my second supervisor, Dr Carol Leung. I had the privilege of witnessing the development of her lab and working alongside her exceptional and dedicated team. From this experience, I learned the importance of discipline, precision, and punctuality—particularly in mouse experiments. I am sincerely grateful to Dr Leung and her team for instilling in me a strong work ethic. Special thanks to Ms Orthi Onupom, Sarah Abdel Malak, and Valerie Leung for their invaluable support.

I have acknowledged many mentors individually throughout this thesis, but I would like to express my sincere appreciation once again for their intellectual generosity. Dr Kate Ridout, thank you for teaching me foundational skills in bioinformatics and for guiding me as I began to connect the broader scientific questions to data analysis. Your continued support and critical review of this thesis were invaluable. I am deeply grateful for your intellectual rigour and mentorship. I also thank Dr Dimitris Vavoulis and Dr Grigore-Aristide Gafencu for the insightful discussions

that shaped my transcriptomic analyses. Dr Adam Burns, you have been more than a mentor. From developing protocols to supervising genomic experiments, your technical guidance and encouragement were instrumental. I still remember our early experiments at the CPHL Uganda lab. Your steady support helped build my confidence and strengthened my belief that I belonged in this academic environment.

To the members of the Schuh group whom I have not named individually, please know that your contributions have not gone unnoticed. From experimental discussions to the practical aspects of maintaining the lab, you were an essential part of this journey. I am grateful to have been part of such a supportive team.

Finally, I could not have completed this journey without the unwavering love and support of my family. My dear wife, Halima, and our children—Bright, Bill, Lucky, and Stryer—have endured four years of my absence while I pursued this degree abroad. Your patience and encouragement sustained me throughout. My father, who never fails to ask, “My son, when are you finishing school?”, has been a constant source of motivation. I know my mother would have asked the same, were she still with us. Though neither of my parents had a formal education, they understood its value and supported me in every way they could. Their love and sacrifices made this achievement possible.

# Molecular characterisation of Epstein-Barr virus in endemic Burkitt Lymphoma

Draguma Ismail Legason, The Queen's College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy, Hillary Term 2026

## Abstract

Epstein–Barr virus (EBV) is a near-universal human pathogen causally linked to multiple malignancies that arise in distinct geographic and clinical contexts. This disparity between widespread infection and restricted cancer incidence represents a central paradox in EBV biology. The current study explored this paradox by integrating comparative viral genomics, mutational signature analysis, tumour transcriptomics, and experimental immunology within a phenotype-aware analytical framework.

Whole-genome analysis, including newly generated data from East African isolates, reaffirmed substantial geographic structuring of EBV genomes, consistent with long-term host–virus coevolution. Comparative analyses identified elevated recombination signals in EBNA1 and LMP1 among Burkitt lymphoma (BL)–associated viruses relative to those from healthy carriers, nominating these loci as potential targets of disease-specific selective pressures. Mutational signature profiling revealed a BL-enriched single-base substitution pattern (SBS_EBV4) characterised by G-rich C>T, T>C, and T>G contexts, with weak similarity to known canonical COSMIC signatures. The biological interpretation remains open and could reflect a composite of host and viral mutational processes, potentially influenced by malaria-associated environmental pressures.

Transcriptomic profiling of BL tumours revealed heterogeneous viral gene expression, including the detection of both latent and lytic transcripts in a subset of cases. However, protein-level validation will be necessary to interpret these findings definitively.

Guided by genomic and transcriptional insights, selected latent and lytic antigens were incorporated into Chimpanzee-derived Adenovirus Oxford-2 (ChAdOx2) and Modified Vaccinia Ankara (MVA) viral vectors for preclinical evaluation. Heterologous prime–boost immunisation elicited robust $CD4^+$ and $CD8^+$ T-cell responses, with multi-antigen constructs inducing broader and more polyfunctional responses than single-compartment formulations.

Collectively, this work provides not only a comprehensive description of the EBV genome in BL, but also insights into the interplay between EBV genome variation and immunogenicity. It reinforces geography-driven EBV diversity, identifies a candidate BL-enriched mutational signature, and highlights heterogeneous viral expression as a potential therapeutic vulnerability. These findings support the development of multi-antigen therapeutic vaccines targeting both latent and lytic compartments and provide a framework for rational intervention in EBV-associated malignancies.

## Table of Contents

**List of Figures** ..... **xii**

**List of Tables** ..... **xvi**

**List of abbreviations and acronyms** ..... **xviii**

**Chapter 1: Introduction** ..... **xxiii**

**1.1 Cancers and origin** ..... **1**

**1.2 Epstein-Barr virus genome** ..... **7**

**1.3 EBV infection** ..... **9**

**1.4 Epidemiology and EBV oncogenesis** ..... **14**

**1.5 EBV-associated malignancies** ..... **20**

1.5.1 Epithelial carcinomas ..... 20

1.5.2 Mesenchymal tumours ..... 24

1.5.3 Lymphoid neoplasms ..... 26

**1.6 Burkitt lymphoma pathogenesis and the rationale for EBV-targeted vaccination** ..... **28**

1.6.1 Hallmark of Burkitt lymphoma ..... 28

1.6.2 Role of infectious agents in Burkitt lymphoma pathogenesis ..... 30

1.6.3 Current treatment of Burkitt lymphoma in Sub-Saharan Africa ..... 33

1.6.4 Clinical outcomes and persistent unmet needs ..... 34

1.6.5 Rationale for immunotherapeutic approaches in Burkitt lymphoma ..... 35

**1.7 Cellular immunity in cancer control** ..... **36**

**1.8 Principles of therapeutic cancer vaccination** ..... **38**

**1.9 Viral vector vaccines and their mechanisms of action** ..... **42**

**1.10 Engineering EBV vaccines for cancer immunotherapy** ..... **47**

**1.11 Aims of the thesis** ..... **48**

**Chapter 2: Materials and Methods** ..... **49**

**2.1 Cell-free DNA workflow** ..... **50**

2.1.1 Materials and equipment ..... 50

2.1.2 Participant recruitment and ethical considerations ..... 51

2.1.3 Sample collection and processing ..... 52
2.1.4 Cell-free DNA extraction and quality assessment ..... 52
2.1.5 Preparation of cell-free DNA libraries for Illumina sequencing ..... 56
2.1.6 Illumina sequencing of cfDNA libraries ..... 62
**2.2 FFPE RNA workflow ..... 64**
2.2.1 Materials and equipment ..... 64
2.2.2 Tissue collection, processing and sectioning ..... 65
2.2.3 RNA Extraction and Quality Assessment ..... 66
2.2.4 FFPE RNA library preparation for Illumina sequencing ..... 69
2.2.4.1 Ribosomal RNA depletion ..... 69
2.2.4.2 RNA-Seq library preparation ..... 71
2.2.4.3 RNA Sequencing ..... 72
**2.3 Burkitt lymphoma diagnosis and EBV antigen expression studies ..... 73**
2.3.1 Materials and equipment ..... 73
2.3.2 Preparation of bond wash solution ..... 74
2.3.3 Histopathologic assessment of tumour sections at the recruitment sites ..... 74
2.3.4 FISH for MYC -translocation ..... 75
2.3.5 Immunohistochemistry for EBV antigens ..... 78
2.3.6 Microscopy and reporting ..... 79
**2.4 EBV vaccine immunogenicity studies ..... 80**
2.4.1 Materials ..... 80
2.4.2 Methods ..... 82
2.4.2.1 Preparation of buffers, peptides and vaccine ..... 82
2.4.2.2 Mice ..... 85
2.4.2.3 Sample size and power calculations ..... 86
2.4.2.4 Vaccinations ..... 86
2.4.2.5 Sample collection and endpoints ..... 87
**Chapter 3: Epstein-Barr Virus Genome Diversity and Evolution ..... 92**
**3.1 Introduction ..... 93**
**3.2 Chapter aims and specific objectives ..... 97**
**3.3 Methods ..... 98**
3.3.1 Study setting, design and ethical approval ..... 98
3.3.2 Patients and samples ..... 98
3.3.3 Samples and nucleic acid extraction ..... 99
3.3.4 EBV probe design and whole genome sequencing ..... 99

3.3.5 EBV whole genome data processing ..... 99
3.3.6 Coverage and sequencing depth across samples..... 103
3.3.7 EBV genome assembly..... 103
3.3.8 Phylogenetic analysis..... 105
3.3.9 Principal component analysis ..... 106
3.3.10 Recombination analysis..... 107
3.3.11 Assessing the impact of recombination on EBV genome evolution .... 110
3.3.12 Assessing EBV mutation frequency..... 112
3.3.13 EBV mutation signatures analysis..... 115
3.3.14 Individual contributions ..... 116
**3.4 Results ..... 117**
3.4.1 Primary study cohort characteristics..... 117
3.4.2 Summary of the analysed EBV genomes ..... 119
3.4.3 Data quality assessment..... 119
3.4.4 EBV population structure and genomic variation..... 121
3.4.5 Higher LD decay is observed in BL-derived EBV genomes compared to other phenotypes ..... 125
3.4.6 Impact of recombination on EBV genome diversity ..... 130
3.4.7 Point mutations and indels concentrate in latency and immune-evasion genes…. ..... 134
3.4.8 EBV mutational signatures and association with Burkitt lymphoma..... 140
**3.5 Discussion..... 151**
3.5.1 Geographic signature strongly influences EBV population structure ..... 151
3.5.2 Recombination in BL challenges the model of strict viral clonality ..... 152
3.5.3 Point mutations and indels concentrate in latency and immune-evasion genes…. ..... 158
3.5.4 Mutational signatures identify BL-enriched viral mutagenic process..... 160
**3.6 Study limitations..... 162**
**3.7 Future direction..... 163**
**3.8 Conclusion..... 164**
**Chapter 4: Epstein-Barr Virus Transcriptional Landscape in Burkitt Lymphoma ..... 165**
**4.1 Introduction..... 166**
**4.2 Chapter aims and objectives ..... 169**

**4.3 Methods ..... 170**
4.3.1 Patients and samples ..... 170
4.3.2 FFPE sectioning ..... 170
4.3.3 Immunohistochemistry ..... 170
4.3.4 RNA extraction ..... 170
4.3.5 Preparing RNA libraries ..... 171
4.3.6 Sequencing libraries ..... 171
4.3.7 RNAseq data processing and DE analysis ..... 171
4.3.8 EBV gene expression ..... 172
4.3.9 Gene set enrichment analysis ..... 173
4.3.10 Evaluating the impact of EBV on host and viral gene expression ..... 174
4.3.11 Individual contributions ..... 175
**4.4 Results ..... 176**
4.4.1 Summary of study samples ..... 176
4.4.2 Data quality and sensitivity analysis ..... 177
4.4.3 EBV gene expression in endemic Burkitt lymphoma ..... 178
4.4.4 Integrated pathway analysis reveals coordinated host cellular programmes associated with EBV gene expression ..... 188
**4.5 Discussion ..... 197**
4.5.1 Heterogeneous EBV gene expression in endemic BL tumours ..... 197
4.5.2 EBV gene activity appears to enhance proliferative and oncogenic pathways in endemic BL ..... 201
4.5.3 Allograft rejection pathway upregulation may suggest antiviral immunity in endemic BL ..... 202
4.5.4 Downregulated pathways reflect dedifferentiation and reduced stress response in endemic BL ..... 203
4.5.5 EBV-dependent subclusters within endemic BL cohort ..... 203
4.6 Study limitations ..... 204
4.7 Future direction ..... 205
**4.8 Conclusions ..... 205**
**Chapter 5: Immunogenicity of ChAdOx2 and MVA Viral Vectors Expressing EBV Latent, Lytic, and Combined Antigens in Preclinical Models ..... 207**
**5.1 Introduction ..... 208**
**5.2 Chapter aims and objectives ..... 213**
**5.3 Methods ..... 214**

5.3.1 Vaccine construct ..... 214
5.3.2 Viral vectors ..... 217
5.3.3 Mice ..... 217
5.3.4 Vaccinations ..... 217
5.3.5 Ex vivo peptide stimulation, surface staining, and intracellular cytokine staining ..... 217
5.3.6 Flow acquisition and data processing ..... 218
5.3.7 Statistical analysis ..... 220
5.3.8 Individual contributions ..... 221
**5.4 Results ..... 222**
5.4.1 Prime-boost vaccination with EBV vaccine induces antigen-specific $CD8^+$ and $CD4^+$ T cell responses in peripheral blood ..... 222
5.4.2 Peptide immunogenicity and antigen hierarchy of vaccine-induced $CD8^+$ T-cell responses ..... 226
5.4.3 Breadth and functional quality of vaccine-induced T-cell responses ..... 232
5.4.4 Heterologous MVA boosting enhances antigen-specific T-cell cytokine responses following ChAd priming ..... 236
5.4.5 End-point splenic antigen-specific T cell responses demonstrate broad antigen recognition ..... 239
**5.5 Discussion ..... 243**
5.5.1 Prime-boost vaccination induces antigen-specific $CD8^+$ and $CD4^+$ T-cell responses ..... 244
5.5.2 Peptide immunogenicity and antigen hierarchy ..... 245
5.5.3 Breadth and functional quality of vaccine-induced T-cell responses ..... 248
5.5.4 Heterologous MVA boosting enhances cytokine responses ..... 249
5.5.5 Establishment of antigen-specific responses in the spleen ..... 250
5.5.6 Study limitations ..... 251
5.5.7 Future direction ..... 253
5.5.8 Conclusion ..... 254
**Chapter 6: General Discussion ..... 255**
**6.1 Research questions and objectives ..... 256**
**6.2 Principal findings of the research ..... 256**
6.2.1 Genomic diversity, recombination, and geographic structure ..... 256
6.2.2 Mutational signatures and disease-specific viral processes ..... 257
6.2.3 Viral transcriptional heterogeneity in BL ..... 258
6.2.4 Implications for vaccine development ..... 259

**6.3 Conclusions and future direction ..... 261**

**Chapter 7: Data Appendix ..... 264**

**7.1 Supplementary figures ..... 265**

**7.2 Supplementary Tables ..... 290**

**Chapter 8: Appendices ..... 320**

**8.1 Period of the Research and Laboratories ..... 321**

**8.2 Sources of funding ..... 322**

**8.3 Abstracts and Scientific Conferences ..... 323**

**8.4 Publications/manuscripts from this thesis ..... 323**

**References ..... 324**

## List of Figures

Figure 1. 1 Model of clonal evolution in neoplasia (adapted from Nowell, 1976) ..... 3

Figure 1. 2 Circular plasmid-like EBV genome ..... 8

Figure 1. 3 Model demonstrating EBV entry into epithelial cells ..... 10

Figure 1. 4 EBV infection cycle and gene expression ..... 13

Figure 1. 5 Activation of multiple cell signalling pathways by the EBV latent membrane protein 1(LMP-1) ..... 17

Figure 1. 6 Pathogenesis model of EBV-driven Burkitt lymphoma ..... 31

Figure 1. 7 Immune escape mechanisms involve disruption of the cancer-immunity cycle at multiple checkpoints ..... 42

Figure 1. 8 Mechanisms of therapeutic cancer vaccines ..... 45

Figure 2. 1 Steps of cfDNA extraction ..... 53

Figure 2. 2 Size distribution of plasma cfDNA analysed by Agilent Bioanalyser 2100 .. 56

Figure 2. 3 Cell-free DNA library preparation workflow for Illumina sequencing, adapted from (Takara Bio Inc. UK) ..... 57

Figure 2. 4 EBV DNA capture workflow using magnetic streptavidin beads, adapted from (IDT Inc. USA) ..... 62

Figure 2. 5 RNA extraction and purification from FFPE sections using the RNeasy FFPE kit, adapted from (Qiagen, Hilden, Germany) ..... 66

Figure 2. 6 Assessment of FFPE RNA quality using DV200 derived from the electropherogram. ..... 69

Figure 2. 7 Ribosomal RNA depletion workflow adapted from (New England Biolabs) 70

Figure 2. 8 Fluorescence in situ hybridisation (FISH) analysis of the MYC (8q24) locus in interphase nuclei ..... 77

Figure 2. 9 Vaccine inserts into viral vector backbone ..... 85

Figure 3. 1 Schematic of the bioinformatic pipeline ..... 102

Figure 3. 2 Phylogeny of the 410 EBV genomes ..... 122
Figure 3. 3 PCs showing region and strain differences ..... 124
Figure 3. 4 Genome-wide linkage disequilibrium decay analysis across EBV clinical phenotypes ..... 128
Figure 3. 5 Pairwise comparison of LD decay rates across phenotypes..... 129
Figure 3. 6 Tanglegram of 219 EBV genomes highlighting recombination-driven shuffling ..... 131
Figure 3. 7 Recombination frequency (ρ/bp) across coding EBV genes..... 133
Figure 3. 8 Comparison of EBV mutation burden across disease phenotypes..... 136
Figure 3. 9 Normalised mean variant density per kb in EBV genomes from BL samples ..... 137
Figure 3. 10 Gene-level mutation enrichment in BL versus Healthy controls ..... 139
Figure 3. 11 COSMIC-style mutational profiles of EBV-derived single-base substitutions ..... 143
Figure 3. 12 Mean contributions of EBV mutational signatures (SBS_EBV1–SBS_EBV4) stratified by Region and Phenotype ..... 145
Figure 3. 13 Beta regression model of EBV mutational signature proportions in disease phenotypes relative to healthy controls ..... 147
Figure 3. 14 Top COSMIC SBS signature matches to de novo EBV signatures ..... 149
Figure 4. 1 Relative normalised EBV transcript expression in cases and controls ..... 180
Figure 4. 2 EBV gene expression in BL compared to non-BL controls (NBL) ..... 181
Figure 4. 3 Expression levels of selected Epstein–Barr virus (EBV) genes in EBV-positive samples ..... 182
Figure 4. 4 Micrographs showing antibody staining for selected EBV antigens ..... 185
Figure 4. 5 Dot plot showing twenty of Hallmark pathways analysed and their enrichment scores ..... 190
Figure 4. 6 Heatmap plot of the twenty-five top leading-edge genes from the downregulated pathways ..... 192

Figure 4. 7 Volcano plots illustrating the effect of EBV infection on gene expression in control samples ..... 194
Figure 4. 8 Integrated model of EBV-driven host pathway modulation ..... 196
Figure 5. 1 Amino acid substitutions mapped across $CD8^+$ and $CD4^+$ T cell epitopes for selected EBV antigens ..... 216
Figure 5. 2 Flow cytometry gating strategy ..... 219
Figure 5. 3 Experimental design and timeline of prime-boost vaccinations ..... 223
Figure 5. 4 Antigen-specific IFN-γ T-cell responses following prime and boost vaccination ..... 225
Figure 5. 5 Peptide immunogenicity profile of vaccine-induced T-cell responses ..... 228
Figure 5. 6 Spearman correlation matrix of peptide-specific response magnitudes ..... 231
Figure 5. 7 Breadth of peptide-specific responses across vaccine groups ..... 233
Figure 5. 8 Polyfunctional $CD8^+$ T-cell responses following peptide stimulation ..... 235
Figure 5. 9 Cytokine responses to EBV peptides in ChAd-primed animals boosted with either ChAd or MVA ..... 238
Figure 5. 10 Splenic $CD8^+$ T-cell cytokine responses following vaccination ..... 241
Figure 5. 11 Splenic $CD4^+$ T-cell cytokine responses following vaccination ..... 242
Supplementary Fig. 1 Participant selection, EBV whole genome sequencing and analysis of the primary study cohort ..... 267
Supplementary Fig. 2 Geographic origins of the EBV genomes analysed in the study. 268
Supplementary Fig. 3 Determining the strain of the EBV genome using the NCBI BLAST algorithm ..... 270
Supplementary Fig. 4 Sequencing depth and coverage across primary and public genomes ..... 272
Supplementary Fig. 5 Per-sample comparison of variant counts identified by three variant calling algorithms ..... 274
Supplementary Fig. 6 Burkitt lymphoma EBV strain-specific LD decay patterns ..... 275

Supplementary Fig. 7 Comparison of LD decay-rate estimates between EBV Type 1 and Type 2 Burkitt lymphoma genomes ..... 276
Supplementary Fig. 8 EBV genome mutation load across six disease phenotypes and healthy carriers ..... 277
Supplementary Fig. 9 NMF rank survey metrics ..... 279
Supplementary Fig. 10 Consensus clustering heatmaps for NMF ranks..... 280
Supplementary Fig. 11 Spearman's rank correlation matrix of EBV mutation signatures ..... 281
Supplementary Fig. 12 Shapiro-Wilk test results of mutational signature distributions in Burkitt lymphoma..... 282
Supplementary Fig. 13 Evaluation of batch effects and correlation between effect size and statistical significance..... 283
Supplementary Fig. 14 Antigen-specific IL-2 T-cell responses following prime and boost vaccination..... 284
Supplementary Fig. 15 Antigen-specific TNF-α T-cell responses following prime and boost vaccination ..... 286
Supplementary Fig. 16 Cytokine responses to EBV peptides in Chad-primed animals boosted with either ChAd or MVA ..... 288

## List of Tables

Table 2. 1 cfDNA reagents, kits and equipment..... 50

Table 2. 2 Thermocycler programme for amplification of cfDNA library ..... 59

Table 2. 3 FFPE RNA reagents, kits and equipment..... 64

Table 2. 4 Monoclonal antibodies used for immunohistochemistry ..... 73

Table 2. 5 Reagents, detection kits and consumables used for immunohistochemistry.. 73

Table 2. 6 Monoclonal antibodies used for IHC and their optimised parameters ..... 79

Table 2. 7 Reagents used for Ex vivo PBMC and splenocyte stimulation assays..... 80

Table 2. 8 List of flow cytometry antibodies and cytokines ..... 81

Table 2. 9 Equipment and consumables used for mouse experiments ..... 81

Table 3. 1 Demographic and clinical features of the primary study cohort ..... 118

Table 3. 2 Reproducibility of Illumina sequencing and variant calling ..... 120

Table 3. 3 Summary of de novo EBV mutation signatures and likely mechanisms ..... 150

Table 4. 1 Clinical characteristics of the cohort samples ..... 177

Table 4. 2 Significantly mutated and actively transcribed EBV genes in BL tumours. 179

Table 4. 3 Expression of selected EBV antigens in primary Burkitt lymphoma samples ..... 184

Table 5. 1 Mouse groupings, vaccine regimens and schedules ..... 223

Table 5. 2 Proportion of mice that generated a positive antigen-specific $CD8^+$ IFNγ response to the EBV peptide pools following prime vaccination ..... 227

Table 5. 3 Rank-based analysis of peptide immunodominance within vaccine groups. 230

Supplementary Table 1. Pairwise comparisons of LD decay rates across EBV-associated clinical phenotypes ..... 291

Supplementary Table 2. Putative recombination events identified in EBV genomes from BL and healthy controls..... 292

Supplementary Table 3. Differential mutation enrichment in Burkitt lymphoma (BL) compared with Healthy controls ..... 293
Supplementary Table 4. Differentially expressed human genes in EBV-positive Burkitt lymphoma ..... 294
Supplementary Table 5. Mixed-effects analysis of vaccine-specific immunodominance hierarchies ..... 298
Supplementary Table 6. Haematopathology assessment and cytogenetic results of the BL tumours ..... 299
Supplementary Table 7. Novel EBV genomes isolated ..... 301
Supplementary Table 8. Public EBV genomes analysed ..... 305

## List of abbreviations and acronyms

| | |
|---|---|
| 26E | 26 epitopes |
| ACK | Ammonium Chloride Potassium |
| ADCC | Antibody Dependent Cellular Cytotoxicity |
| AI-REAL | Aggressive Infection-Related East African Lymphoma |
| AID | Activation-Induced Cytidine Deaminase |
| Alt-EJ | Alternative End Joining |
| ANGPT2 | Angiopoietin 2 |
| APC | Antigen-presenting cell |
| APC | adenomatous polyposis coli |
| APOBEC | apolipoprotein B mRNA editing catalytic polypeptide-enzymes |
| ATM | Ataxia-Telangiectasia mutated |
| AUF-1 | Adenylate Uridylate-rich element RNA-Binding protein 1 |
| BAM | binary alignment map |
| BART | BamHI-A Rightward Transcripts |
| BCL2 | B-Cell Lymphoma 2 |
| BCR | B-cell receptor |
| BED | browser extensible data |
| BER | base excision repair |
| BH | Benjamini-Hochberg |
| BL | Burkitt Lymphoma |
| BMC | Bugando Medical Centre |
| BWA-MEM | Burrows-Wheeler-Aligner-maximum exact matches |
| CAFs | Cancer-associated fibroblasts |
| CAP | College of American Pathologists |
| CDKN2A | Cyclin -Dependent Kinase Inhibitor 2A |
| cDNA | complementary DNA |
| CDS | coding DNA sequence |
| cfDNA | cell-free DNA |
| ChAdOx2 | Chimpanzee-derived Adenovirus Oxford 2 |
| CLL/SLL | Chronic lymphocytic leukaemia/small lymphocytic lymphoma |
| CMV | Cytomegalovirus |
| CNS | central nervous system |
| COSMIC | Catalogue of Somatic Mutations in Cancer |
| CPHL | Central Public Health Laboratories |
| CR2 | complement receptor 2 |

| | |
|---|---|
| CST | Cytometer Setup and Tracking beads |
| CSV | Comma-separated values format |
| CTAR | C-Terminal activating region |
| CTL | cytotoxic T lymphocytes |
| CTLA-4 | cytotoxic T lymphocyte-associated protein 4 |
| CYLD | Cylindromatosis Lysine 63 Deubiquitinase |
| DAB | 3,3'-diaminobenzidine chromogen |
| DCs | Dendritic cells |
| DE | Differential expression |
| DEPC | diethyl pyrocarbonate |
| DLBCL | Diffuse large B -cell lymphoma |
| DMSO | Dimethyl Sulfoxide |
| DMEM | Dulbecco's Modified Eagle Medium |
| DNA | Deoxyribonucleic acid |
| DPBS | Dulbecco’s Phosphate Buffered |
| DSBs | double-strand DNA breaks |
| EBER | Epstein-Barr virus-encoded small RNAs |
| EBNA | Epstein-Barr virus-encoded nuclear antigen |
| EBV | Epstein-Barr Virus |
| EBVaGC | Epstein-Barr virus-associated Gastric Carcinoma |
| EDTA | Ethylenediaminetetraacetic acid |
| EMT | epithelial-mesenchymal transition |
| FDR | False-discovery rate |
| FFPE | Formalin-fixed paraffin-embedded |
| FGF | Functional Genomics Facility |
| FISH | Fluorescence in situ hybridisation |
| FL | Follicular lymphoma |
| FSC-A | Forward scatter- Area |
| FSC-H | Forward scatter- Height |
| GC | Gastric carcinoma |
| GSEA | Gene Set Enrichment Analysis |
| GTR | Generalised -Time Reversible |
| H&E | Haematoxylin and Eosin |
| HIFs | Hypoxia-inducible factors |
| HIV/AIDS | Human immunodeficiency virus/acquired immunodeficiency syndrome |
| HL | Hodgkin Lymphoma |
| HLA | Human leukocyte antigen |
| HNSCC | Head and Neck squamous cell carcinoma |
| HPDI | Highest posterior density intervals |

| | |
|---|---|
| HRP | Horseradish Peroxidase |
| HTA | Human Tissue Act |
| HTS | High Throughput System |
| ICAM-1 | Intercellular Adhesion Molecule 1 |
| ICC | International Consensus Classification |
| IDO | indoleamine 2,3 dioxygenase |
| IDT | Integrated DNA Technologies |
| IFNγ | interferon gamma |
| IgA | immunoglobulin A |
| IgE | immunoglobulin E |
| IgG | immunoglobulin G |
| IgH | immunoglobulin heavy chain |
| IgM | immunoglobulin M |
| IHC | immunohistochemistry |
| IL | interleukin |
| Indels | insertions and deletions |
| IRB | Institutional Review Board |
| ISH | In situ hybridisation |
| IU | infection units |
| JAK3/STAT | Janus Kinase 3/Signal Transducer and Activator of Transcription |
| JNK | c-Jun N-terminal kinase |
| KCMC | Kilimanjaro Christian Medical Centre |
| KSHV | Kaposi sarcoma herpesvirus |
| LD | Linkage disequilibrium |
| LELC | Lymphoepithelioma-like carcinomas |
| LMP | latent membrane protein |
| LOESS | Locally Estimated Scatterplot Smoothing |
| LRT | Likelihood Ratio Test |
| MAF | Minor Allele Frequency |
| MAPK | Mitogen-activated protein kinase |
| MDSCs | myeloid-derived suppressor cells |
| MHC | Major histocompatibility complex |
| ML | Maximum likelihood |
| mL | millilitre |
| MLH1 | MutL homologue 1 |
| MNH | Muhimbili National Hospital |
| mRNA | messenger RNA |
| MSigDB | Molecular Signature Database |
| MTA | Material Transfer Agreement |

| | |
|---|---|
| MUHAS | Muhimbili University of Health and Allied Sciences |
| MVA | Modified Vaccinia Ankara |
| NEB | New England Biolabs |
| NES | normalised enrichment scores |
| NF-kB | Nuclear factor kappa-light-chain enhancer of activated B cells |
| ng | nanograms |
| NHEJ | non-homologous end-joining |
| NIH | National Institute of Health |
| NIHR | National Institute for Health Research |
| NIMR | National Institute for Medical Research |
| NKTCL | Natural Killer T -Cell Lymphoma |
| NMF | Non-negative matrix factorisation |
| NOTCH 1 | Notch Receptor 1 |
| NPC | Nasopharyngeal carcinoma |
| OGG1 | 8-Oxoguanine DNA glycosylase 1 |
| OR | Odds Ratio |
| OriLyt | origin of Lytic replication |
| oriP | origin of replication |
| OxTREC | Oxford Tropical Research Ethics Committee |
| P16 | Tumour suppressor protein 16 |
| PBMCs | Peripheral blood mononuclear cells |
| PBS | phosphate-buffered saline |
| PC | Principal Component |
| PCA | Principal Component Analysis |
| PCR | Polymerase chain reaction |
| PD-1 | programmed cell death protein 1 |
| PD-L1 | programmed death -ligand 1 |
| PDGF | Platelet-Derived Growth Factor |
| PFU | Plague-forming units |
| PI3K/AKT | Phosphoinositide 3-kinase/Protein Kinase B |
| PIK3CA | Phosphatidylinositol-4,5-Bisphosphate 3-kinase catalytic subunit alpha |
| P53 | Tumour suppressor protein 53 |
| PRRs | Pattern recognition receptors |
| QC | Quality Control |
| RAG | Recombination activating genes |
| RB | Retinoblastoma protein |
| RBC | Red blood cell |
| RDP4 | Rapid Recombination Detection Programme |
| RIG-1 | Retinoic acid-inducible gene I proteins |

| | |
|---|---|
| RIP1 | receptor-interacting protein 1 |
| RNA | ribonucleic acid |
| RPMI | Roswell Park Memorial Institute 1640 medium |
| rRNA | ribosomal RNA |
| SBS | Single-base substitution |
| SNP | single nucleotide polymorphism |
| SNV | single nucleotide variant |
| SRA | Sequence Read Archive |
| SSA | Sub-Saharan Africa |
| SSC-A | Side Scatter-Area |
| TAMs | Tumour-associated macrophages |
| TANs | Tumour-associated neutrophils |
| TBE | Transfer Bootstrap Expectation |
| Tfh | follicular helper T cells |
| TGF-β | Transforming growth factor-β |
| THL | Translation Histopathology Laboratory |
| TLRs | Toll-like Receptors |
| TME | Tumour Microenvironment |
| TNFα | Tumour necrosis factor-α |
| TNFR | Tumour necrosis factor receptor |
| TPM | Transcripts per million |
| TRAF | Tumour Necrosis Factor Receptor-Associated Factor |
| TRADD | Tumour Necrosis Factor Receptor-Associated Death Domain |
| UNCST | Uganda National Council for Science and Technology |
| UV | Ultraviolet |
| VAF | Variant allelic fraction |
| VCAM-1 | Vascular Cell Adhesion Molecule-1 |
| VCF | Variant call format |
| VEGF | Vascular endothelial growth factor |
| VVCF | Viral Vector Core Facility |
| WHO | World Health Organisation |
| µg | microgram |
| µL | microlitre |
| µM | micromolar |

# Chapter 1: Introduction

## 1.1 Cancers and origin

Cancer refers to a heterogeneous group of diseases characterised by the uncontrolled proliferation of abnormal cells with the capacity to invade surrounding tissues and, in many cases, metastasise to distant organs. In contrast, a tumour (or neoplasm) is defined as an abnormal mass of tissue arising from dysregulated cell growth. With the exception of most haematological malignancies, cancers form tumours; however, not all tumours are malignant. Benign tumours remain localised and do not invade adjacent tissues, whereas malignant tumours acquire invasive and metastatic potential.

From an evolutionary perspective, cancer is not merely a pathological accident but an inherent vulnerability of multicellular life. Whereas unicellular organisms are evolutionarily driven to maximise survival and replication, multicellular organisms require cellular cooperation and strict regulation of proliferation for organismal survival. Most somatic cells relinquish unlimited replicative potential and instead function within a tightly coordinated network of growth control. Malignancy arises when individual cells escape these regulatory constraints and revert to a more autonomous, self-serving state–effectively representing a breakdown of multicellular cooperation (Aktipis et al., 2015).

Peter Nowell first articulated the evolutionary nature of cancer in his seminal 1976 paper (Nowell, 1976), *"The Clonal Evolution of Tumour Cell Populations."* Nowell proposed that most neoplasms originate from a single transformed cell and that tumour progression results from the stepwise acquisition of genetic alterations

within this founding clone (Figure 1.1). These alterations generate heritable variation among tumour cells, upon which selective pressures act, leading to the expansion of subclones with increased proliferative capacity, survival advantage, invasiveness, or therapeutic resistance (Nowell, 1976).

This framework aligns closely with Darwinian principles of evolution (Darwin, 1964). Just as species evolve through variation, heritability, and natural selection, tumour cell populations evolve through genetic instability–driven variation, clonal inheritance of mutations, and selection within the tissue microenvironment. The tumour ecosystem—including immune surveillance, hypoxia, nutrient limitation, and therapeutic intervention—functions analogously to the selective pressures of the environment in classical evolutionary theory. Consequently, cancer progression can be considered as a process of somatic evolution occurring within the body, governed by the same fundamental principles that shape the evolution of species.

To further conceptualise this, Hanahan and Weinberg proposed the “Hallmarks of Cancer” in their seminal 2000 paper, and updated it in 2011, 2022, and 2026. The framework describes the functional capabilities that cancer cells acquire during tumourigenesis as well as the enabling phenotypic characteristics (Hanahan, 2026; Hanahan & Weinberg, 2000, 2011). The core hallmarks include sustained proliferative signalling, evasion of growth suppressors, resistance to cell death, replicative immortality, induction of angiogenesis, and activation of invasion and metastasis. Subsequent refinements incorporated immune evasion and deregulated cellular energetics as additional hallmarks, alongside enabling characteristics such as genomic instability and tumour-promoting inflammation (Hanahan, 2026).

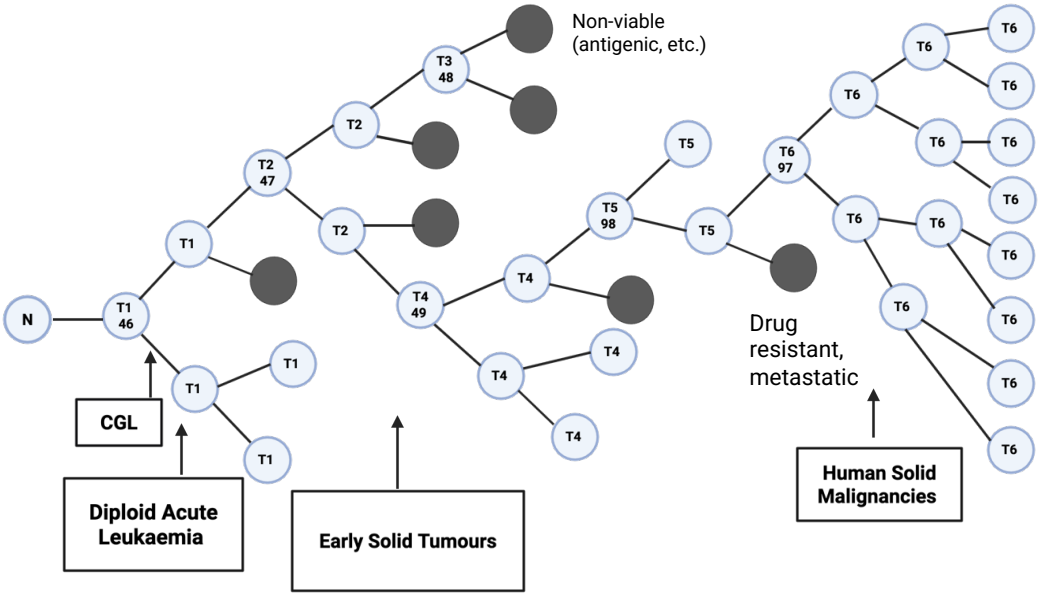



**Figure 1. 1 Model of clonal evolution in neoplasia (adapted from Nowell, 1976)**

In this model, neoplasia originates from a single progenitor cell (N) that undergoes an initial transforming event, giving rise to a founding tumour clone (T1), typically diploid. Successive rounds of cell division generate subclonal populations that accumulate additional genetic alterations. The majority of emerging subclones are eliminated due to metabolic or immunological disadvantage (solid black circles). However, occasional acquisition of mutations conferring a selective growth or survival advantage leads to clonal expansion and dominance of these fitter subpopulations. Tumours with minimal chromosomal alterations (e.g., certain diploid acute leukaemias or chronic granulocytic leukaemia (CGL)) are considered to represent earlier stages of clonal evolution, whereas many solid tumours, frequently characterised by aneuploidy and extensive genomic instability, reflect later evolutionary progression. Figure created using BioRender.

Sustained proliferative signalling is a defining feature of cancer. Many tumours harbour activating mutations or chromosomal rearrangements in oncogenes that chronically stimulate cell cycle progression. Among the most frequently mutated oncogenes are members of the RAS family (KRAS, NRAS, HRAS), which encode small GTPases that regulate proliferative signalling networks. Activating KRAS mutations are particularly prevalent in pancreatic ductal adenocarcinoma (~90%), colorectal cancer (~40–50%), and lung adenocarcinoma (~30–35%) (Lee et al., 2022; Philip et al., 2022; Prior et al., 2020). Other oncogenes that promote

proliferation include PIK3CA, BRAF, BCR-ABL, MYC, CCND1, EGFR, and ERBB2 (HER2) (Bailey et al., 2018; Dhanasekaran et al., 2022; Ostroverkhova et al., 2023; Sinkala, 2023; Wang et al., 2025). These may become oncogenic through point mutations, gene amplification, or chromosomal translocations, which result in constitutive activation or overexpression.

In parallel, cancer cells evade growth-inhibitory signals by inactivating tumour suppressor genes. Proteins such as retinoblastoma protein (RB), adenomatous polyposis coli (APC), and tumour suppressor protein 53 (TP53) function as critical gatekeepers of cell cycle progression and genomic integrity (Joerger et al., 2025; Peng et al., 2025; Weinberg, 1995). APC negatively regulates β-catenin signalling within the WNT pathway, and its loss promotes transcriptional activation of proliferation-associated genes (Hankey et al., 2018; Li et al., 2012). TP53, one of the most frequently mutated genes in human cancer, orchestrates DNA damage responses, apoptosis, and senescence (Kandoth et al., 2013). Disruption of these pathways enables cells to proliferate despite oncogenic stress.

Resistance to apoptosis further facilitates tumour progression. Cancer cells commonly upregulate anti-apoptotic BCL-2 family members (e.g., BCL-2, BCL-XL, MCL-1), thereby neutralising pro-apoptotic effectors such as BAX and BAK (Eskander et al., 2025). This imbalance permits survival despite genomic instability or metabolic stress.

Replicative immortality represents another essential hallmark. Progressive telomere shortening normally limits cellular lifespan by inducing replicative senescence (Jones-Weinert et al., 2025). Loss of TP53 and RB pathways may transiently bypass senescence, but continued telomere attrition ultimately leads to

crisis and genomic instability (Nassour & Karlseder, 2026). Most cancers overcome this barrier by activating telomerase, particularly via upregulation of telomerase reverse transcriptase, thereby restoring telomere length and enabling indefinite replication (Guterres & Villanueva, 2020; Liu et al., 2024; Mishra & Patel, 2025).

As tumours expand, they require increased oxygen and nutrient supply. Under hypoxic conditions, stabilisation of hypoxia-inducible factors (HIFs) induces expression of pro-angiogenic mediators such as vascular endothelial growth factor A (VEGFA) and angiopoietin 2 (ANGPT2), promoting neovascularisation (De Palma et al., 2017). Beyond nutrient delivery, tumour-associated vasculature contributes to immune modulation and facilitates metastatic dissemination (Ileiwat et al., 2022).

Immune evasion has emerged as a central hallmark of cancer (Hanahan, 2022). Tumour cells may upregulate immune checkpoint ligands such as programmed death ligand 1 (PD-L1), leading to T-cell exhaustion and impaired cytotoxic responses (Jiang et al., 2015). Additionally, tumours actively remodel the tumour microenvironment (TME) by recruiting tumour-associated macrophages (TAMs), neutrophils (TANs), cancer-associated fibroblasts (CAFs), and regulatory lymphocytes, thereby generating an immunosuppressive milieu (Hanahan et al., 2025).

Cancer cells also reprogram cellular metabolism, often exhibiting enhanced aerobic glycolysis (the Warburg effect) while retaining mitochondrial oxidative phosphorylation (Bian et al., 2025; Li et al., 2025). Furthermore, phenotypic plasticity enables dedifferentiation and the acquisition of stem-like properties,

which facilitate tumour initiation, therapy resistance, and metastasis (Pérez-González et al., 2023; Subbalakshmi et al., 2024).

Genomic instability and chromosomal instability serve as enabling characteristics that accelerate the acquisition of hallmark capabilities (Hanahan, 2026). Beyond genetic alterations, non-mutational epigenetic reprogramming — including DNA methylation, histone modification, chromatin remodelling, and post-transcriptional regulation — contributes substantially to oncogenic gene expression programmes (Esteller et al., 2024).

Ageing and obesity further influence tumourigenesis. Age-associated accumulation of senescent cells and the senescence-associated secretory phenotype (SASP) promote chronic inflammation and tissue remodelling (Schmitt et al., 2022; Wang et al., 2024). Obesity-related hormonal alterations, including elevated leptin and reduced adiponectin levels, have also been linked to tumour-promoting inflammation and metastatic progression (Donohoe et al., 2017; Harris et al., 2022).

While the hallmarks framework defines the common functional landscape of cancer, the specific molecular events required to achieve these capabilities vary considerably. In virus-associated malignancies, oncogenic viruses may directly provide pre-evolved factors that substitute for host mutational events. Epstein–Barr virus (EBV), in particular, encodes viral proteins and non-coding RNAs capable of driving proliferation (Lo et al., 2021), inhibiting apoptosis (Kelly et al., 2009), modulating immune recognition (Jochum et al., 2012; Strockbine et al., 1998), and reprogramming cellular signalling (Lo et al., 2021) — effectively contributing to multiple cancer hallmarks.

## 1.2 Epstein-Barr virus genome

EBV was first identified in a cell culture derived from Burkitt lymphoma tissue obtained from a Ugandan child in 1964 by Dr Anthony Epstein and his colleagues (Epstein et al., 1964). EBV is a double-stranded DNA virus belonging to the gamma Herpesviridae subfamily and is known to infect humans (Farrell, 2019). It has a 172-kilobase linear genome divided into short and long unique-sequence regions, each comprising terminal and internal repeats (Figure 1.2). This genomic organisation is thought to enhance the coding efficiency and regulatory complexity of EBV by enabling the expression of a diverse repertoire of viral genes and transcripts (Young et al., 2007). EBV encodes slightly more than 80 gene products and dozens of non-coding small RNAs (Kalla & Hammerschmidt, 2012; Münz, 2019).

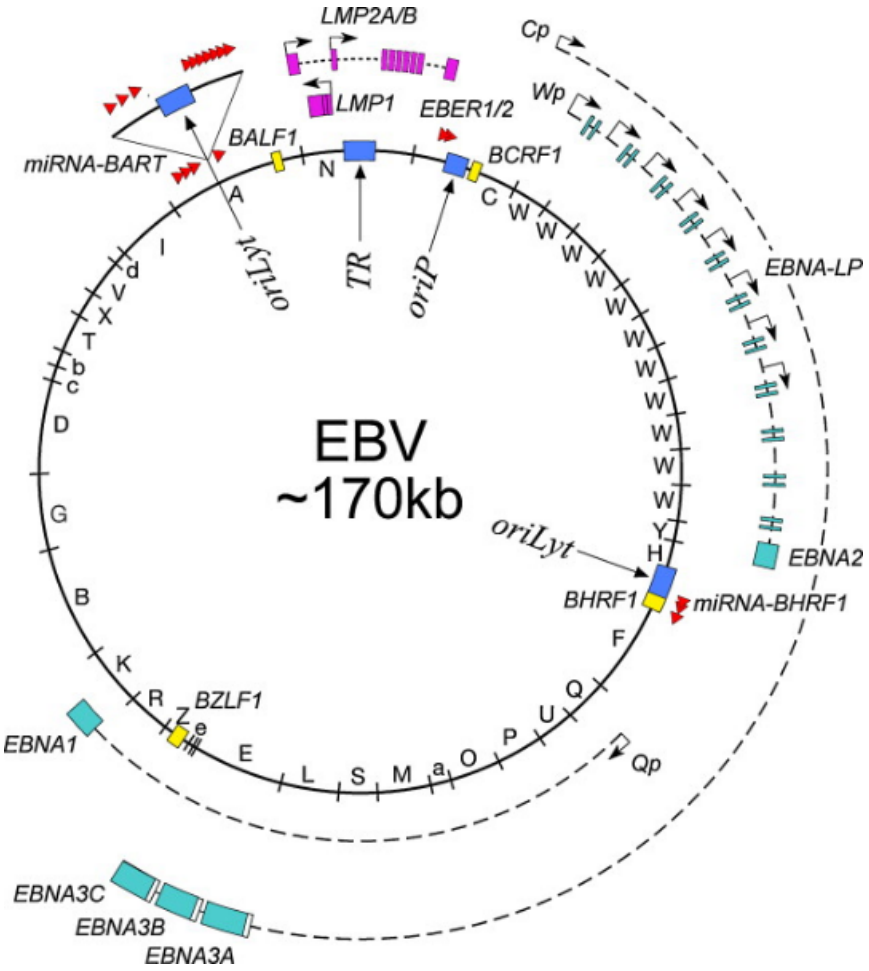



**Figure 1. 2 Circular plasmid-like EBV genome**

The diagram illustrates a representative circular genome of the laboratory EBV strain B95.8, including the 12kb deleted region (inset). The cis-acting elements, e.g., oriP (the plasmid origin of DNA replication), oriLyt (the origin of lytic DNA replication), and the terminal repeat (TR) region, are shown in blue. The inner ring with single letters representing DNA fragments produced by the restriction enzyme BamHI. The latent gene exons (EBNAs and LMPs) are shown in turquoise and magenta boxes, while the lytic gene exons are represented as yellow boxes. Small red triangles mark non-coding RNA gene clusters (EBER-1/EBER-2, miR-BHRF1, and miR-BART). Black arrows indicate various gene promoters. Viral promoters Cp and Wp (W repeats) are active during early infection, transcribing the EBNA-2 and EBNA-LP genes, while the remaining EBNA genes are transcribed from the Cp promoter. Later in the infection, latent membrane proteins are transcribed from the LMP promoters. The Qp promoter is the only active promoter during latency I, transcribing the EBNA-1 gene. Reprinted from European Journal of Cell Biology, Vol. 91, Kalla M., Hammerschmidt W., "Human B cells on their route to latent infection – Early but transient expression of lytic genes of Epstein-Barr virus", pp. 65–69, Copyright (2012), with permission from Elsevier.

## 1.3 EBV infection

EBV transmission primarily occurs through the oral route, although it can also occur via organ transplants and blood if the donor is EBV-infected (Alfieri et al., 1996; Gerber et al., 1969). Kissing and sharing personal items such as toothbrushes, eating utensils, food, and drinks with infected individuals remain the most significant routes of transmission. The primary infection of oral epithelial cells is characterised by lytic viral replication, generating thousands of viral copies that subsequently spread into the bloodstream to infect naïve B cells (Tugizov et al., 2013; Yao et al., 1985).

EBV epithelial infection and entry primarily occur via direct fusion of the viral membrane with the host cell membrane (Chen & Longnecker, 2019), although entry via lipid-raft-dependent endocytosis and micropinocytosis may also occur (Wang et al., 2015). After binding to the host cell membrane receptors, the viral capsid is transcytosed into the cytoplasm (Figure 1.3). The viral nucleocapsid then dissolves, releasing the viral genome into the nucleus through an open nuclear pore.

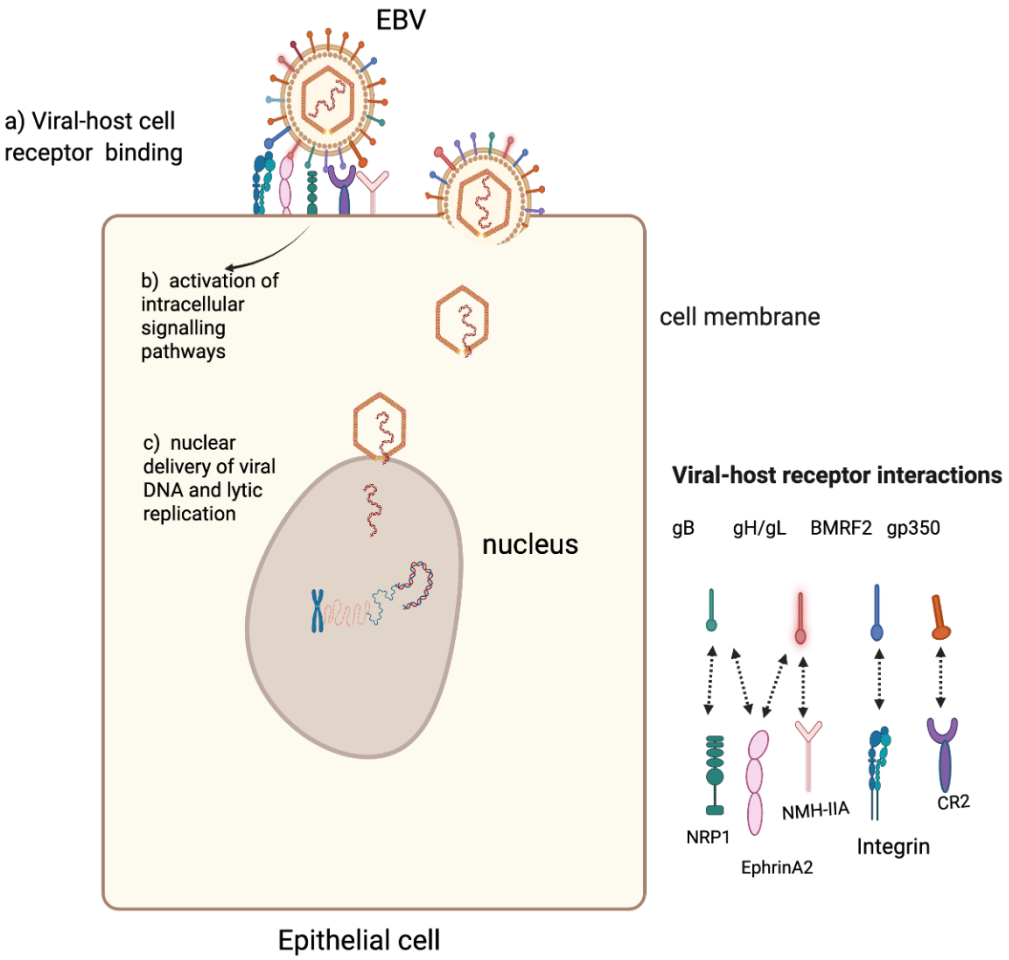



**Figure 1. 3 Model demonstrating EBV entry into epithelial cells**

In step (a), EBV attaches to receptors on the surface of the target host cell. In step (b), binding to neuropilin-1 activates NRP-1-dependent epidermal growth factor receptor (EGFR) signalling, leading to activation of downstream AKT and ERK pathways. The integration of EBV glycoproteins (gH/gL, gp350 or BMRF2) with integrins (e.g., αvβ6) triggers "outside-in" signalling. This leads to activation of Rho, ROCK, Src family kinases, and p38/MAPK, which are necessary for actin remodelling during entry. EphA2 acts as a primary entry receptor, binding directly to EBV gH/gL and activating kinase-dependent endocytic pathways. These signalling events activate actin-binding proteins, such as non-muscle myosin chain IIa (NMHC-IIA), which mediates EBV infection of nasopharyngeal epithelial cells. Once inside (c), the viral nucleocapsid disassembles, and the viral genome is delivered into the nucleus through an open nuclear pore. Figure adapted from Chen & Longnecker (2019) and created using BioRender.

EBV infection of B cells mainly occurs through the interaction of gp350, an abundant viral surface glycoprotein, with complement receptor 2 (CR2; also known as CD21), a B-cell-specific receptor, reviewed in (Connolly et al., 2011). Membrane fusion of EBV is triggered by the binding of the viral glycoprotein (gp42) to human leucocyte antigen class II molecules on target B cells (Connolly et al., 2011). Once in the cytosol, the viral nucleocapsid disassembles, and the viral genome is transported into the host cell nucleus, where initial transcription and replication occur. The immediate-early genes, which encode regulatory proteins, are replicated first, followed by the early genes that encode enzymes required for viral replication (Young et al., 2007). The final set of genes, known as late genes, is responsible for the synthesis of structural proteins (Young et al., 2007). After the initial replication, the viral genome circularises and becomes tethered to the host genome (Young et al., 2007). This period is characterised by minimal expression and replication of the viral genome.

During latency, the EBV genome is usually maintained as a circular episome within the nucleus of infected cells, allowing replication alongside the host genome during cell division (Young et al., 2000). Integration of the viral genome into the host genome has been reported (Baer et al., 1984), albeit rarely; viral integration has been proposed to facilitate synchronous replication with the host genome and may reduce susceptibility to host immune attack (Young et al., 2000). There are five phases of viral latency, categorised as 0, I, IIa, IIb, and III, based on the number of gene products expressed (Figure 1.4). Latency III, also known as complete viral latency, involves six nuclear proteins, three latent membrane proteins, numerous

small RNAs, and dozens of non-coding RNA transcripts (reviewed in (Murata et al., 2021).

After the initial B cell infection, the viral genome circularises or becomes integrated into the host genome, persisting latently during clonal proliferation, germinal centre transit and memory B cell differentiation (Baer et al., 1984; Young et al., 2000). Studies show that minimal viral transcription occurs during this phase of latency (Price & Luftig, 2014; Wang et al., 2019). Although occasional reactivation events occur, according to a study by Inagaki et al. (2021), these do not appear to lead to full lytic viral production (Inagaki et al., 2021). This incomplete lytic cycle, which involves the expression of a select set of key viral genes, is thought to benefit the virus's long-term persistence in the infected cell.

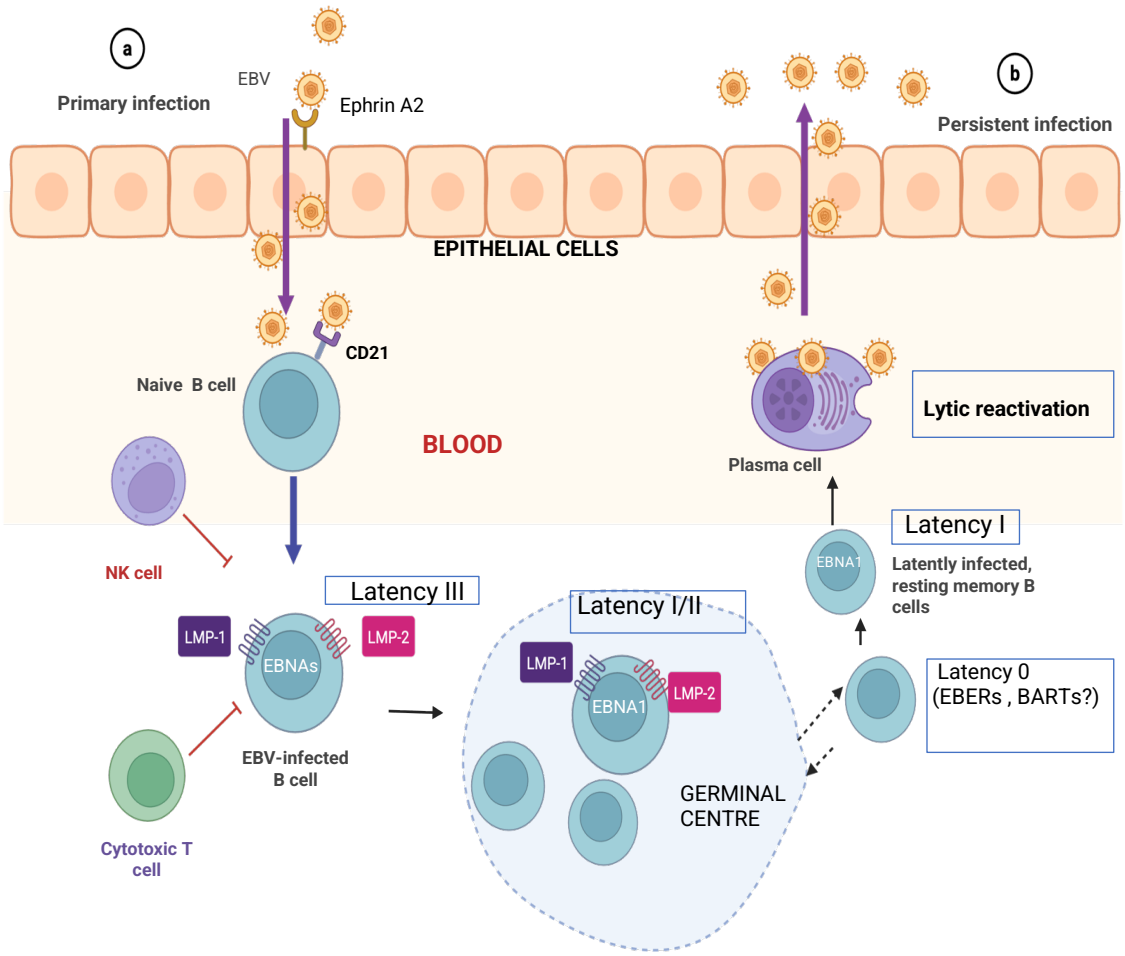



**Figure 1. 4 EBV infection cycle and gene expression**

The schematic illustrates the EBV infection cycle and gene expression patterns. The initial EBV replication in epithelial cells marks the establishment and spread of infection in the systemic blood system. Naïve B cells that encounter EBV initially become activated (atypical lymphocytes), expressing full latency genes of the virus (Latency III). This initial infection triggers a surge in cytokine production and the recruitment of immune cells, including cytotoxic T cells. To escape this immune response, EBV then downregulates its antigen expression and adopts a restricted program, "latency I/II," where only key viral proteins, EBNA-1 (viral episome maintenance) and LMP-1,2A,2B (host cell transformation), are required to sustain the infection. The ultimate goal is a complete shutdown of gene expression (latency 0), which occurs in resting memory B cells. During this phase, some miRNAs, such as EBERs, may be expressed, but their role is not yet entirely understood. Occasionally, reactivation occurs, either due to antigen re-encounter by memory B cells or other cellular triggers that compel memory B cells to differentiate into antibody-secreting plasma cells. EBV replicates with full lytic production of virions that cause further shedding to other cells, including epithelial cells. Adapted from Richardo et al. (2020) and created in BioRender.

## 1.4 Epidemiology and EBV oncogenesis

Approximately 90–95% of adults worldwide are infected with EBV, establishing it as one of the most ubiquitous human herpesviruses (Dunmire et al., 2018; Young & Rickinson, 2004). While primary infection is often asymptomatic, EBV is etiologically linked to a spectrum of lymphoid and epithelial malignancies, including Burkitt lymphoma (BL), Hodgkin lymphoma (HL), nasopharyngeal carcinoma (NPC), and certain gastric carcinomas, as well as autoimmune disorders (Bjornevik et al., 2023; Farrell, 2019; Houen & Trier, 2021; Münz, 2019; Ward et al., 2022; Wong et al., 2022; Young & Rickinson, 2004).

EBV exists as two major viral types, EBV type 1 (EBV-1) and EBV type 2 (EBV-2), originally defined by sequence divergence within the EBNA2 and EBNA3 gene families (Tzellos & Farrell, 2012). The most pronounced divergence occurs in EBNA2, where the proteins share only ~54% amino acid identity between types (Tzellos & Farrell, 2012). These sequence differences are functionally significant: EBV-1 isolates consistently demonstrate greater efficiency in transforming primary B cells into lymphoblastoid cell lines in vitro, largely attributable to differences in EBNA2-mediated transcriptional regulation (Ponnusamy et al., 2019; Tzellos et al., 2014). In contrast, EBV-2 shows reduced B-cell immortalisation capacity but may display altered lytic propensity and T-cell tropism (Romero-Masters et al., 2020; Tzellos & Farrell, 2012).

Globally, EBV-1 predominates across most populations, whereas EBV-2 is more frequently detected in sub-Saharan Africa and among immunocompromised individuals (Smatti et al., 2018; Tzellos & Farrell, 2012). The higher prevalence of EBV-2 in equatorial Africa has led to longstanding speculation regarding its potential association with endemic Burkitt lymphoma. However, molecular epidemiological studies have demonstrated that both EBV types are present in BL tumours, and there is no consistent evidence supporting a direct causal relationship between EBV-2 infection and BL risk (Kaymaz et al., 2020; Liao et al., 2022). Instead, BL pathogenesis appears to be driven primarily by MYC translocation, with EBV contributing through latency-associated survival signals rather than through type-specific oncogenic differences (Münz, 2019; Young & Rickinson, 2004).

EBV infection typically starts in epithelial cells, which primarily facilitate lytic replication (Temple et al., 2014). In the initial stage of the infection, the virus multiplies, leading to cell burst and release of virions that infect naïve B lymphocytes. Following successful infection of B cells, the virus must sustain a lifelong presence by cycling through distinct latency programmes (Figure 1.4). These programmes are characterised by tightly controlled patterns of viral gene expression that promote cell survival, immune evasion, and eventual entry into the long-lived memory B-cell pool.

Interestingly, most known EBV complications occur during latency. The specific disease mechanisms differ by cell type but generally revolve around the expression of highly oncogenic viral proteins, such as latent membrane proteins (LMP-1,

LMP-2A), and the EBNA family of proteins (Lo et al., 2021; Münz, 2019; Murata et al., 2021; Young et al., 2007).

The most aggressive form of latency, known as latency III (the "growth programme"), is characterised by expression of the full complement of latent genes: six EBV nuclear antigens (EBNA-1, 2, 3A, 3B, 3C, and LP), three latent membrane proteins (LMP-1, 2A, and 2B), as well as non-coding RNAs such as EBERs and viral microRNAs (Münz, 2019; Young et al., 2007; Young et al., 2000). Among these, LMP-1 plays a central role in cellular transformation (Kulwichit et al., 1998; Nilsson et al., 1971). Functioning as a constitutively active mimic of CD40, LMP-1 drives B-cell proliferation, survival, and signalling through NF-κB, JAK/STAT, and MAPK pathways (Bangalore-Prakash et al., 2017; Kieser & Sterz, 2015; Lo et al., 2021). This mimicry allows EBV-infected B cells to proliferate independently of normal T-cell help, producing highly immunogenic cells that are rapidly targeted by cytotoxic T lymphocytes. This massive proliferation of antigen-expressing cells, the release of inflammatory cytokines (IFN-γ), and the recruitment of immune cells underlie the symptoms of infectious mononucleosis (fever, lymphadenopathy, and fatigue) (Dunmire et al., 2018; Houen & Trier, 2021). The structure of EBV LMP-1 protein and the various signal transduction pathways affected by LMP-1 are summarised in Figure 1.5.

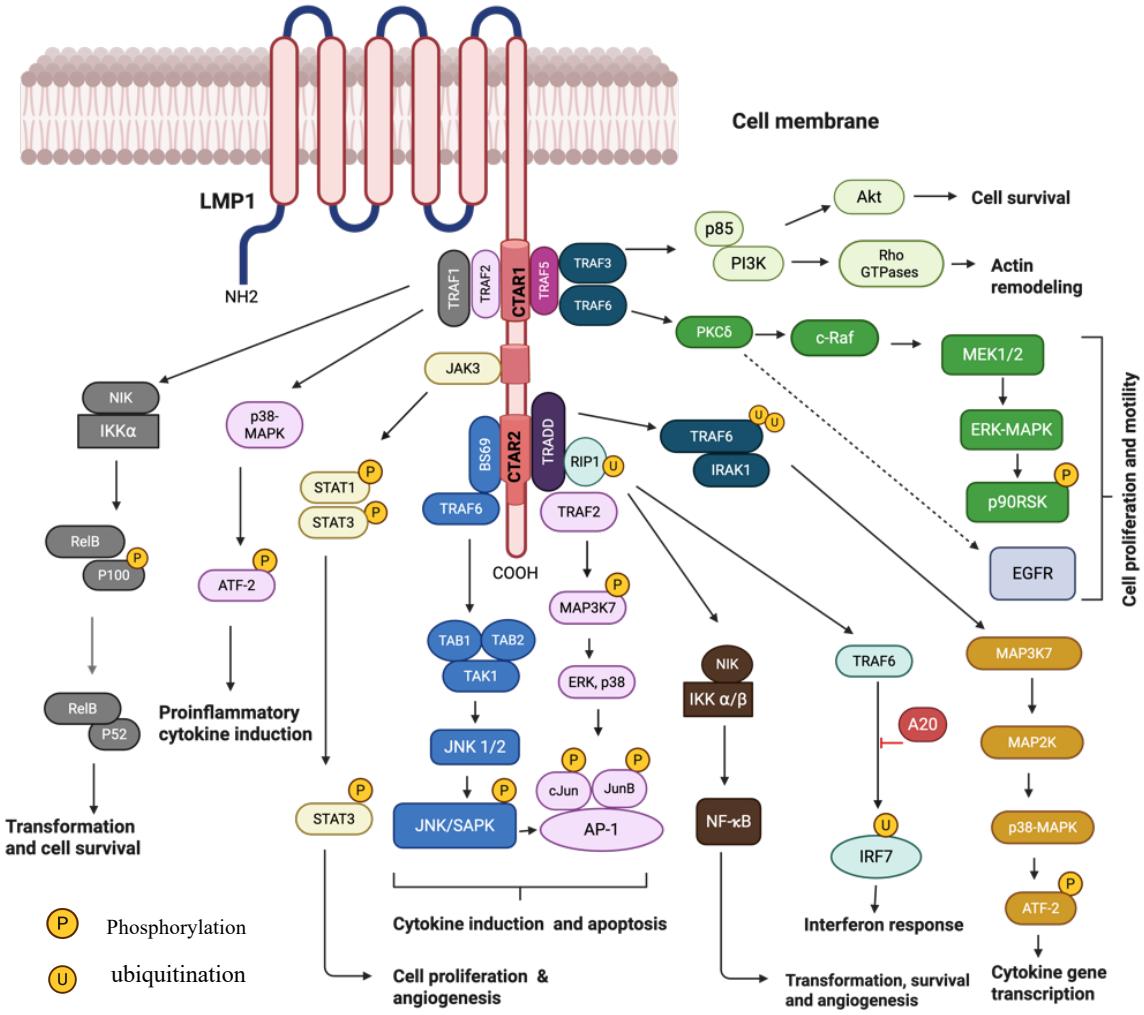



**Figure 1. 5 Activation of multiple cell signalling pathways by the EBV latent membrane protein 1(LMP-1)**

LMP-1 is a 386-amino acid protein, featuring a short N-terminal cytoplasmic domain, six transmembrane domains, and a long C-terminal cytoplasmic domain. The protein acts as a viral mimic of the cellular CD40, a member of the tumour necrosis factor receptor (TNFR) that signals through its cytoplasmic C-terminal activating regions, CTAR1 and CTAR2. CTAR1 mainly recruits TRAF family adaptor proteins to activate PI3K/Akt signalling, the ERK/MAPK cascade (via Raf–MEK–ERK–p90RSK), and non-canonical NF-κB signalling (NIK–IKKα–RelB/p52), thereby promoting cell survival, proliferation, cytoskeletal remodelling, and transformation. CTAR2 signals primarily through TRADD-dependent complexes involving TRAF6, RIP, and related ubiquitination events, leading to activation of TAK1 (MAP3K7). TAK1 then bifurcates into canonical NF-κB signalling (IKKβ–p65/p50), JNK/AP-1 activation, and the p38 MAPK pathway (MKK3/6–p38–ATF-2), resulting in the induction of pro-inflammatory cytokines, as well as promoting cell survival, proliferation, angiogenesis, and invasion. Additionally, CTAR2 activates JAK3/STAT1/STAT3 signalling to facilitate proliferation and angiogenic gene expression. Negative regulation of TRAF6-dependent signalling is mediated by the ubiquitin-editing enzyme A20. Collectively, these pathways contribute to transformation, inflammatory signalling, and tumour progression. Figure adapted from Lo et al. (2021) and created in BioRender.

To escape the immune pressure, EBV must downregulate immunogenic antigens and transition into latency II, or the "immune evasion programme." Here, only a restricted subset of genes (EBNA-1, LMP-1, and LMP-2A) remain expressed, allowing infected cells to persist while remaining less visible to T cells. Further restriction occurs in latency I, where EBNA-1 and non-coding RNAs (EBERs and BARTs) are expressed (Figure 1.4). Studies have shown that these non-coding transcripts are essential elements of the viral immune evasion strategy, as they downregulate the expression of host immune-related genes, such as MHC class I, and also inhibit pro-apoptotic factors (Choi et al., 2013; Kang et al., 2015; Tagawa et al., 2016). Therefore, latency II facilitates persistence within the germinal centre and limits immune recognition during B cell processing. Ultimately, EBV establishes latency 0/I within memory B cells, where viral gene expression is almost completely silenced, creating a stable reservoir.

EBNA-1 plays a unique role across these programmes, as it tethers the episomal viral genome to host chromosomes and ensures its replication and segregation during B-cell division (Young et al., 2007). This is achieved by binding to the origin of plasmid replication (oriP) and to host mitotic chromosomes, enabling EBV episomes to be faithfully partitioned into daughter cells.

Although EBV latency is increasingly recognised as a continuum rather than a set of discrete states, several EBV-associated malignancies exhibit intermediate or dynamic transcriptional programmes (reviewed in (Murata et al., 2021)). For

instance, studies of Burkitt lymphoma tumours have described mixed patterns of EBV gene expression (Abate et al., 2015; Xue et al., 2002), suggesting that viral transcription in these tumours may not always conform to the traditional classical latency pattern associated with BL. The precise molecular characteristics of these intermediate or mixed gene expression states, and their potential link to disease risk, remain poorly understood. Similarly, the overall extent to which mixed latent-lytic cycle programmes are active in vivo remains unclear.

In vitro experiments provide further complexity (Kelly Gemma et al., 2013; Kelly et al., 2006). A transient phase of lytic gene activation, sometimes referred to as the pre-latent phase, has been proposed (Inagaki et al., 2021). This is characterised by weak BZLF1 binding to lytic gene promoters in the context of unmethylated viral DNA, which may facilitate the transition from initial infection to long-term viral silencing (latency)(Murata et al., 2021). Importantly, this activation does not result in productive virion release, but its role in latency establishment and immune modulation remains unclear.

Occasionally, reactivation events occur when memory B cells differentiate into antibody-secreting plasma cells in response to antigen re-exposure, B cell receptor cross-linking, or other cellular stress signals (Murata et al., 2021; Murata & Tsurumi, 2014). In EBV, this process is also driven by BZLF1, which binds preferentially to methylated CpG motifs within lytic gene promoters to trigger reactivation (Murata et al., 2021). This paradoxical behaviour of BZLF1, marked by weak engagement with unmethylated viral DNA during latency establishment and strong activation of methylated promoters during reactivation, raises important

mechanistic questions about how viral gene expression is controlled in different cellular environments.

From a pathogenesis perspective, the key unknowns are whether EBV undergoes a fundamental shift in transcriptional programming in malignant cells, particularly in BL cells, and which cellular or viral factors facilitate this activation, and how these changes influence disease outcomes. Lytic genes, such as BCRF1, BALF1, BHRLF1, BNLF2a, BILF1 and BILF2, as well as non-coding transcripts including EBERs and BARTs, have already been implicated in viral immune evasion, primarily through the downregulation of human leukocyte antigen expression and the inhibition of pro-apoptotic pathways (Choi et al., 2013; Kang et al., 2015; Tagawa et al., 2016). Yet whether these viral products are expressed in a stable, dynamic or stochastic manner within tumours remains an open and essential question for understanding EBV-driven oncogenesis.

## 1.5 EBV-associated malignancies

### 1.5.1 Epithelial carcinomas

Two of the well-characterised epithelial carcinomas associated with EBV infection are nasopharyngeal carcinoma (undifferentiated, NPC) and gastric carcinoma (EBV-associated gastric carcinoma, EBVaGC).

In 2018, the fraction of NPC new cases attributed to EBV infection was approximately 85% (de Martel et al., 2020). More than 90% of NPC tumours have

demonstrable EBV presence, suggesting a high infection rate in endemic settings (Kanda et al., 2019; Wong et al., 2022). NPC affects mainly adults and is believed to be more common in Southeast Asia, Northeast India, North Africa, Northern Canada, Alaska, and the Pacific (Wong et al., 2022). Histologically, NPC originates in the walls of the nasopharynx, characterised by a dense submucosal lymphatic network that leads to early metastasis to nearby lymph nodes. Treatment typically involves radiotherapy, often combined with chemotherapy, but about a third of patients experience relapse due to the extensive metastatic nature of the disease (Le et al., 2019; Lee et al., 2005).

NPC is a typical EBV-driven malignancy, with its molecular basis of pathogenesis involving a complex interaction between viral infection, host genetic factors, and environmental carcinogens (Tao & Chan, 2007; Wong et al., 2021). EBV is consistently present in nearly all NPC tumour cells, indicating its etiological role(Young & Dawson, 2014). NPC tumours generally exhibit latency type II programme of EBV gene expression, with EBNA-1 (episomal maintenance), LMP-1 (CD40 homolog), and LMP-2A (BCR-like signalling) driving proliferation and survival (Brooks et al., 1992; Frappier, 2012). The BART miRNAs are also expressed and are associated with a diminished suppressive immune environment in EBV-positive NPC tumours, thereby allowing the persistence of EBV-infected cells (Cosmopoulos et al., 2009; Wang et al., 2017).

Genetic studies in specific populations, particularly in southern China and Southeast Asia, support this hypothesis by demonstrating that individuals with certain HLA class I/II polymorphisms have reduced immune recognition of EBV

antigens (Ning et al., 2020; Tang et al., 2010; Tse et al., 2009). Epigenetic alterations in tumour suppressor genes, such as RASSF1A and CDKN2A/P16, along with somatic mutations in key immune signalling genes, including PIK3CA, TP53, CYLD, TRAF3, and NF-kB signalling regulators, are common in EBV-positive NPC tumours (Chung et al., 2019; Li et al., 2017; You et al., 2019). Furthermore, the tumours exhibit a predominant infiltration of T and NK cells, although these cells are often exhausted due to the upregulation of PD-1/PD-L1(Fang et al., 2014; Larbcharoensub et al., 2018; Liu et al., 2021).

Nonetheless, EBV-negative NPC occurs, albeit rarely. The clinicopathological features resemble those of conventional head and neck squamous cell carcinoma (HNSCC) (Emfietzoglou et al., 2022; Ngan et al., 2018). However, the dominant molecular features include the presence of driver mutations and chromosomal instability, such as TP53, NOTCH, and PI3K/AKT mutations, loss of CDKN2A, or amplification of Cyclin D1 (Ali et al., 2017; Zheng et al., 2016). Carcinogen exposures, such as tobacco smoking, chemicals like formaldehyde, and alcohol consumption, also appear to play a role (Yu et al., 2022).

Gastric cancer ranks as the fifth highest in incidence and the fourth highest in mortality among cancers worldwide. There are four distinct subtypes: i) microsatellite instability (MSI) gastric cancer, characterised by high mutation rates and hypermethylation at the MutL homologue 1 (MLH1) promoter; ii) genomically stable (GS) gastric cancer, characterised by mutations and/or fusions of proteins associated with the Rho family of GTPases; iii) chromosomal instability (CIN) gastric cancer, which is linked to a higher rate of copy number aberrations; and iv)

the EBV-associated gastric cancer (EBVaGC), characterised by higher DNA hypermethylation compared to other subtypes (Network, 2014). The latter is also described as an EBV-specific epigenotype thought to result from EBV-induced methylation (Kaneda et al., 2012). The causal association with EBV remains under debate, but approximately 10% of all gastric carcinomas worldwide are attributed to EBV infection (Murphy et al., 2009; Network, 2014).

Gastric carcinoma exhibits diverse pathology, with EBV-associated cases forming a distinct molecular subgroup. These tumours are often characterised by chromosomal instability and key mutations, such as PIK3CA (present in up to 80% of cases), JAK2 amplification, and recurrent gains of immune checkpoint ligands PD-L1 (CD274) and PD-L2, which facilitate immune evasion (Dong et al., 2016; Network, 2014; Stanland & Luftig, 2018). The tumour microenvironment features a dense infiltration of cytotoxic T cells and natural killer cells, indicating ongoing immune surveillance, although these immune cells are frequently impaired in function (Salnikov et al., 2023). A notable characteristic of EBV-positive gastric cancer is epigenetic rewiring, characterised by extensive hypermethylation of promoter sites (CpG islands) of tumour suppressors (Ryan et al., 2010), which helps sustain the malignant phenotype.

Consistent across studies (Lo et al., 2021; Stanland & Luftig, 2018; Wang et al., 2017), EBV promotes oncogenesis in NPCs and GC through its oncoproteins (LMP-1 and LMP-2A) and BART-derived microRNAs, which target host regulatory networks. These viral components converge on key immune and survival signalling pathways, including PI3K/AKT, JAK/STAT, and NF-κB,

promoting proliferation, inhibiting apoptosis, and facilitating immune evasion (Fang et al., 2014; Miliotis & Slack, 2020). Together, these effects likely create a supportive environment for EBV persistence, immune escape, and the development of gastric tumours.

Lymphoepithelioma-like carcinomas are poorly differentiated tumours characterised by dense infiltrates of T and B cells, macrophages, and plasma cells (Suster, 2023). Histologically, they resemble NPC but originate outside the nasopharynx, usually affecting the stomach, lungs, and salivary glands, as well as the bladder, liver, and other less frequently involved sites (Dubey et al., 1998; Marioni et al., 2002). The EBV-associated subtypes tend to occur more frequently in some parts of Asia, the Arctic, and North Africa (Hirabayashi et al., 2023); however, they are overall rare compared to NPC and other EBV-associated malignancies.

Like NPCs, the tumours mainly follow a latency II programme linked to EBNA-1, LMP-1, LMP-2A, and BART-derived miRNAs, as well as EBERs. The expression of these viral genes in LELC cells activates key immune signalling pathways, including NF-kB, JAK/STAT, and PI3K/AKT, promoting cell survival, growth, and immune evasion (Choi et al., 2013; Münz, 2019; Tagawa et al., 2016).

### 1.5.2 Mesenchymal tumours

Mesenchymal tumours originate from connective tissue and can vary from benign lesions, such as lipoma, leiomyoma, and fibroma, to malignant sarcomas, including rhabdomyosarcoma, leiomyosarcoma, and angiosarcoma. Sarcomas are rare,

representing less than 1% of adult cancers and about 15% of childhood cancers (Burningham et al.; Jo & Fletcher, 2014). Their molecular pathogenesis involves recurrent genetic alterations, including translocations (e.g., EWSR1-FLI1 fusion in Ewing sarcoma), gain-of-function mutations (e.g., MDM2 and CDK4 in liposarcoma), and loss of tumour suppressor genes (e.g., TP53 and RB1)(Damerell et al., 2021; Delattre et al., 1994; Riggi & Stamenkovic, 2007; Taylor et al., 2011). Dysregulated signalling pathways, such as the PI3K/AKT/mTOR, WNT/β-catenin, Hedgehog, and growth factor signalling pathways (VEGF, PDGF, TGF-β), drive angiogenesis, stromal proliferation, and tumour progression (Osuna & de Álava, 2009).

A specific subset of mesenchymal tumours is associated with EBV, predominantly occurring in immunocompromised individuals such as transplant recipients, patients with HIV/AIDS, and those with congenital immunodeficiencies (Jonigk et al., 2012; Lee et al., 1995; McClain et al., 1995). EBV infects smooth muscles via CD21 and promotes tumour formation through EBNA-1 and small EBV RNAs (EBERs), with variable expression of LMP-1. Like most EBV-driven malignancies, EBNA-1 maintains the viral episome and influences genomic stability, while EBERs act as key non-coding viral effectors. They bind various cellular proteins (e.g., La, RIG-1, AUF-1), are recognised by Toll-like receptors (TLR3), inhibit PKR-mediated apoptosis, and upregulate insulin-like growth factor-1 (IGF-1) (Iwakiri & Takada, 2010; Iwakiri et al., 2009; Samanta & Takada, 2010). Through these interactions, EBERs can induce inflammatory cytokines, enhance cell survival, and support the proliferation of EBV-infected cells.

### 1.5.3 Lymphoid neoplasms

Lymphoid neoplasms are cancers of the blood and lymphatic system, which are primary targets of EBV infection. They are further classified based on cell lineage and developmental stage into precursor lymphoid neoplasms, mature B-cell neoplasms, and mature T-cell and natural killer cell neoplasms (Alaggio et al., 2022). The precursor lymphoid neoplasms include B- and T-lymphoblastic leukaemias (Alaggio et al., 2022). Mature B-cell neoplasms comprise a wide range of B-cell disorders occurring at various stages of differentiation. These include diffuse large B-cell lymphoma (DLBCL), Burkitt lymphoma (BL), follicular lymphoma (FL), Hodgkin lymphoma (HL), mantle cell lymphoma, chronic lymphocytic leukaemia/small lymphocytic lymphoma (CLL/SLL), marginal zone lymphomas, and plasma cell neoplasms (Alaggio et al., 2022; Leoncini et al., 2008; Swerdlow et al., 2008). Mature T-cell and NK-cell lymphomas comprise a rare and heterogeneous group derived from mature post-thymic T and NK cells (Swerdlow et al., 2008).

Histological features of lymphoid neoplasms include effaced lymph node architecture, monomorphic atypical lymphoid cells, abnormal proliferation, high mitotic activity, and necrosis. Specific hallmark features, such as Reed-Sternberg cells in HL (Bárcena, 2023), a starry-sky appearance in BL (Bidot, 2021), proliferation centres in CLL/SLL (Kajtár, 2024), and angiocentric necrosis in NK/T cell lymphoma (Shi, 2020), assist in identifying them under a microscope.

The molecular hallmark of lymphoid neoplasms is the presence of genetic and epigenetic alterations that deregulate antigen receptor pathways and associated survival or proliferation signals, enabling a single transformed lymphoid clone to evade regular checkpoints. In B cells, these genetic changes include chromosomal translocations such as t(8;14) (MYC/IgH) in 80% of Burkitt lymphoma cases, t(14;18) (BCL2/IgH ) in follicular lymphoma, and t(11;14) (Cyclin D1/IgH) in mantle cell lymphoma (Burmeister et al., 2013; Ladanyi et al., 1991; Yin & Luthra, 2013). Somatic hypermutation and immunoglobulin class switch errors during B cell differentiation are common and thought to introduce oncogenic alterations. Driver mutations in genes such as p53, ATM, and RB1, as well as the activation of survival and immune modulation pathways like NF-κB, JAK/STAT, PI3K/AKT, and NOTCH, are frequently observed (Arruga et al., 2018; Han et al., 2010). Conversely, aberrant JAK/STAT, NOTCH1, and epigenetic changes contribute to the development of T-cell/NK lymphomas (de Mel et al., 2019; Montes-Mojarro et al., 2021).

EBV is an established carcinogen in various lymphoid malignancies. For example, nearly all endemic Burkitt lymphoma tumours are EBV-positive, and between 30% and 75% of classical Hodgkin lymphoma tumours are also EBV-positive (de Martel et al., 2020; Wong et al., 2022). Similarly, EBV is found in up to 50% of diffuse large B-cell lymphoma tumours, particularly in endemic regions (Castillo et al., 2018). EBV infection has also been demonstrated in NK/T-cell lymphomas, particularly the nodal and extranodal subtypes. The nodal disease mainly involves the lymph nodes with limited extranodal involvement. In contrast, extranodal disease occurs in the nasal and oral cavities. Studies show EBV positivity in both

tumours ranges from 50% to 95%, and EBV is regarded as a pathogenetic hallmark of this lymphoma (Alaggio et al., 2022), to the extent that EBER ISH is required for diagnosis in the World Health Organisation/International Consensus Classification (WHO/ICC-2022) classification of haematolymphoid tumours.

Unlike B-cell infection, EBV primarily infects NK/T cells via indirect cell-to-cell transfer from infected B cells, possibly utilising low-level CD21 expression (Lee et al., 2018; Tabiasco et al., 2003) or other receptors (van Erp et al., 2019). Once inside, EBV establishes latency II, with LMP1, LMP2A, EBNA1, and EBERs promoting survival, proliferation, and immune evasion, which ultimately contribute to the development of NK/T-cell lymphoma (Murata et al., 2021).

## 1.6 Burkitt lymphoma pathogenesis and the rationale for EBV-targeted vaccination

### 1.6.1 Hallmark of Burkitt lymphoma

Burkitt lymphoma, the focus of this study, is a predominantly childhood cancer endemic to equatorial Africa and parts of Oceania (Mbulaiteye & Devesa, 2022). First described in a Ugandan child more than six decades ago by Denis Burkitt (Burkitt, 1958), BL remains one of the fastest-proliferating human tumours (Grande et al., 2019). The defining molecular hallmark of BL is the translocation of the MYC oncogene from chromosome 8q24 into proximity with one of the three immunoglobulin loci – most frequently the heavy chain locus on chromosome 14q32 (MYC-IGH), or less commonly, the light chain loci on chromosome 2p12 (MYC-IGK) or 22q11 (MYC-IGL) (Alaggio et al., 2022). These rearrangements place MYC under the control of constitutively active immunoglobulin enhancers,

leading to uncontrolled MYC expression and cell-cycle deregulation (Alaggio et al., 2022; Mbulaiteye, 2013).

The molecular events driving MYC translocations are closely linked to normal B-cell development. BL originates from germinal centre B cells, which are specialised for diversifying the antibody repertoire through somatic hypermutation (SHM) and immunoglobulin class switch recombination (CSR) (Basso & Dalla-Favera, 2015). Prior to these processes, each B cell must be able to express a specific B cell receptor (BCR) during early development stages in the bone marrow. The creation of a functional BCR occurs through an immunoglobulin gene rearrangement where genes encoding the variable (V), diversity (D) and joining (J) segments of the immunoglobulin heavy chain locus must rearrange with those of light chain genes (λ or κ) (Nemazee, 2017). This process is mediated by recombination-activating genes RAG1 & RAG2 (Murphy & Weaver, 2016; Nemazee, 2017). The genes encode enzymes that catalyse the assembly of variable (V), diversity (D) and joining (J) segments at the immunoglobulin heavy chain locus, followed by V-J joining at the light chain loci (κ or λ) (Murphy & Weaver, 2016). B cells that successfully express a functional BCR are positively selected to exit the bone marrow and circulate as mature naïve B cells (Basso & Dalla-Favera, 2015). At this point, the B cell receptor specificity is fixed (VDJ recombination is complete), but the immunoglobulin loci are transcriptionally active and remain a recombination hotspot for subsequent events in the germinal centre.

Upon antigenic encounter, naïve B cells are activated and migrate to the germinal centres, where activation-induced cytidine deaminase (AID) initiates SHM and

CSR (Shannon-Lowe & Rickinson, 2019). AID deaminates cytosine to uracil within single-stranded DNA, preferentially targeting immunoglobulin variable and switch regions (Godsmark et al., 2021). These lesions are processed by base excision and mismatch repair pathways. When multiple AID-induced nicks occur on opposite strands, double-strand DNA breaks (DSBs) form (Stavnezer et al., 2008). Under physiological conditions, these DSBs are repaired by the classical non-homologous end-joining (NHEJ) pathway, excising the intervening DNA to achieve isotype switching from IgM to IgG, IgA, or IgE (Stavnezer et al., 2008). However, when the more error-prone alternative end-joining (Alt-EJ) or microhomology-mediated repair pathway is engaged, aberrant recombination can lead to chromosomal translocations, including MYC-IgH, BCL6-IgH and BCL2-IgH, thereby driving lymphomagenesis (Robbiani et al., 2008; Robbiani et al., 2009; Yan et al., 2007).

### 1.6.2 Role of infectious agents in Burkitt lymphoma pathogenesis

Chronic antigenic stimulation by infections such as malaria (e.g., *P. falciparum*– the most frequent cause of malaria in equatorial Africa) remains a significant risk factor in BL pathogenesis. Persistent malaria is thought to contribute to BL pathogenies by upregulating AID activity in B cells (Ariera et al., 2025; Robbiani et al., 2015; Summerauer et al., 2022). Persistent AID activity induces genomic instability, increasing the likelihood of MYC translocations – the defining lesion of endemic BL (Summerauer et al., 2022). The synergistic effect of EBV and malaria likely underpins the striking geographical overlap of BL with malaria-endemic regions of Africa and Papua New Guinea (Kanda et al., 2019; Mbulaiteye, 2013; Peprah et al., 2020), where chronic *P. falciparum* exposure weakens EBV-

specific T-cell surveillance and drives uncontrolled B-cell activation. Figure 1.6 summarises candidate oncogenic processes that contribute to the development of endemic Burkitt lymphoma in Africa.

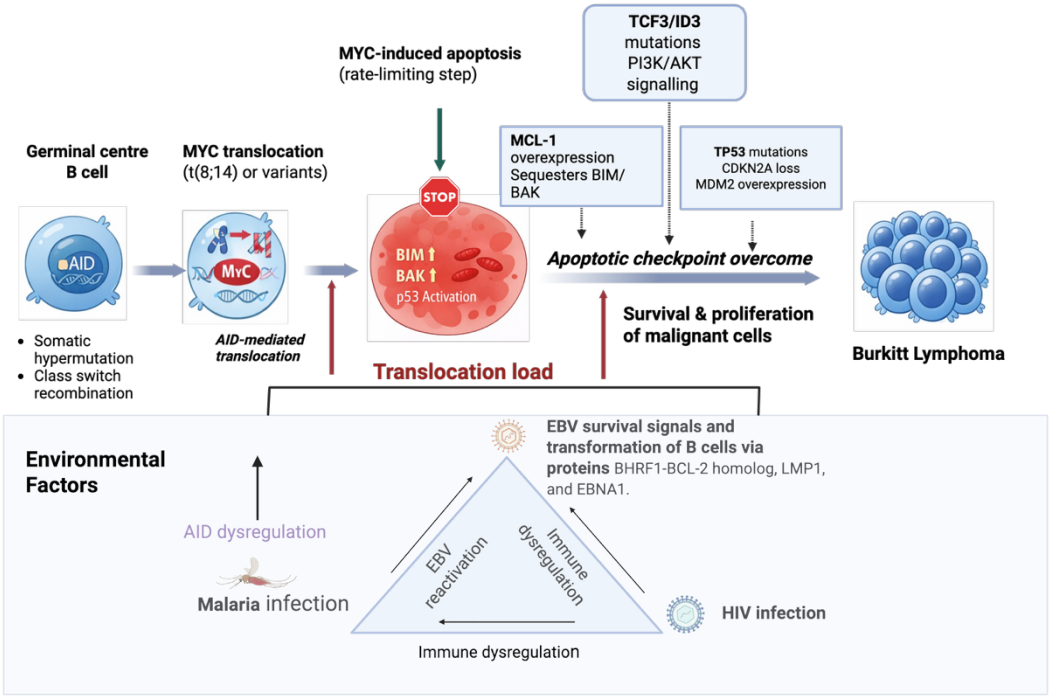



**Figure 1. 6 Pathogenesis model of EBV-driven Burkitt lymphoma**

Burkitt lymphoma is initiated by MYC translocations (t(8;14) or variants), which drive uncontrolled proliferation but simultaneously activate intrinsic apoptotic fail-safe mechanisms. Progression to malignancy requires secondary genetic and viral alterations that suppress MYC-induced apoptosis. These include overexpression of anti-apoptotic MCL-1, TP53 mutations or functional inactivation via CDKN2A loss or MDM2 overexpression, and constitutive PI3K/AKT signalling driven by TCF3/ID3 mutations. In endemic disease, EBV provides additional survival signals through viral proteins, including BHRF1 (a BCL-2 homolog), latent membrane proteins, and nuclear antigens, particularly EBNA1, thereby enabling survival of translocation-positive B cells. Environmental co-factors, including malaria, promote aberrant AID activity, increasing MYC translocation frequency. Collectively, these events overcome apoptotic checkpoints and are believed to enable clonal expansion and the development of Burkitt lymphoma. Figure inspired by Mbulaiteye (2013) and created using BioRender.

Although the oncogenic potential of EBV is well established, the contribution of viral genomic diversity to cancer risk remains incompletely resolved. In vitro experiments demonstrate that viral genotype can affect biological behaviour; for example, EBV Type 1 isolates transform resting B cells more efficiently than Type 2 in tissue culture assays (Romero-Masters et al., 2020), but translating these findings to human disease has proved challenging. Large-scale genomic surveys have documented extensive population-specific EBV sequence variation across the EBV genome, including polymorphisms in latent and lytic loci (e.g., LMP1, BALF2, EBNA genes) that have plausible functional consequences for signalling and immune recognition (da Costa et al., 2015; Liao et al., 2022; Liao et al., 2018; Xu et al., 2019; Xue et al., 2021). Notably, compelling regional associations have been reported for nasopharyngeal carcinoma (NPC): non-synonymous BALF2 variants and several amino-acid substitutions (e.g., BALF2 V317M; BNRF1 G696R; RPMS1 D51E) have been strongly associated with NPC risk in southern China and appear to have undergone local clonal (Xu et al., 2019; Xue et al., 2021). Similarly, EBNA1 and BARF1 variants have been identified in Kenyan BL cases (Simbiri et al., 2015), suggesting a possible contribution to the geographic variation in BL incidence. By contrast, evidence for robust, disease-specific viral determinants in other malignancies such as gastric carcinoma (8-10% of cases) and classical Hodgkin lymphoma (30-75% of cases) remains weaker and more heterogeneous (Begić et al., 2022; Borozan et al., 2018; Kojima et al., 2025); in fact, an early study concluded that there were no significant differences between tumour and matched non-malignant EBV genomes within individuals (Bornkamm et al., 1984).

Importantly, comprehensive phylogenomic analyses of hundreds of complete genomes have repeatedly revealed pronounced geographic population structure—clades and allele frequencies that mirror human population history—rather than clear, globally generalisable disease-specific clustering (Chiara et al., 2016; Correia et al., 2018; Palser et al., 2015; Telford et al., 2020; Zanella et al., 2019).

Taken together, the evidence supports a model in which (i) some viral variants have biological plausibility for altered oncogenicity, (ii) strong regional associations exist for certain cancers (notably NPC and eBL), but (iii) much of the apparent disease signal is confounded by geography, host population structure, and limited sampling, leaving the question of universal EBV oncogenic variants open.

### 1.6.3 Current treatment of Burkitt lymphoma in Sub-Saharan Africa

Burkitt lymphoma remains one of the most common paediatric malignancies in sub-Saharan Africa (SSA), where treatment continues to rely predominantly on cyclophosphamide-based combination chemotherapy regimens developed in the 1960s and 1970s (Olweny et al., 1980; Ziegler et al., 1970). These regimens typically combine low-intensity agents, including cyclophosphamide, vincristine, and methotrexate (COM), with or without cytarabine, and are administered uniformly throughout treatment, without risk-adapted induction or consolidation phases (Ngoma et al., 2012; Ozuah et al., 2020). While generally well tolerated, these protocols do not adequately address high-risk disease, particularly in patients with advanced tumour burden or central nervous system (CNS) involvement, where early relapse and treatment failure remain common.

Attempts to reduce CNS relapse have included intrathecal chemotherapy and, historically, cranial or craniospinal irradiation. However, clinical trials, including early work by Olweny et al.(Olweny et al., 1977), demonstrated limited efficacy and unacceptable long-term toxicity, leading to the abandonment of radiotherapy in paediatric BL. Consequently, treatment strategies in SSA remain constrained by historical protocols, limited supportive care infrastructure, and an inability to intensify therapy safely.

### 1.6.4 Clinical outcomes and persistent unmet needs

Despite incremental improvements in diagnostic accuracy and chemotherapy delivery, long-term survival rates for BL in SSA remain among the lowest globally. Five-year survival remains below 40% with little evidence of sustained improvement over recent decades (Ozuah et al., 2020). Studies conducted under optimised, resource-supported conditions, such as recent longitudinal cohorts from Uganda (McGoldrick et al., 2019) and Malawi (Molyneux et al., 2017), have shown that even when access to diagnosis, chemotherapy, and follow-up is improved, survival outcomes remain comparable to those reported in historical studies.

BL is a highly curable cancer, with nearly 90% of cases in high-income countries achieving cure rates following chemotherapy induction (Dunleavy et al., 2013; Roschewski et al., 2020; Roschewski et al., 2022). This is partly attributable to early precision diagnostics and to generally improved supportive care in these settings. The incorporation of rituximab into BL treatment regimens has

demonstrated clear survival benefits in high-income settings (Hoelzer et al., 2014; Minard-Colin et al., 2020; Ribrag et al., 2016) and, more recently, in SSA. The AI-REAL study demonstrated that rituximab can be administered safely in East African paediatric populations and is associated with improved event-free survival (12-month event-free survival estimated at 67% for Rituximab-COM and 43% for COM-only regimens), particularly in advanced-stage disease (Mawalla et al., 2025). However, rituximab remains costly, requires cold-chain storage, and is associated with increased rates of treatment-related immunosuppression, highlighting the need for alternative strategies that can improve disease control without substantially increasing toxicity.

### 1.6.5 Rationale for immunotherapeutic approaches in Burkitt lymphoma

BL is a biologically compelling candidate for immunotherapeutic intervention. The tumour arises from germinal-centre B cells, is highly proliferative, and in endemic regions is closely associated with EBV infection (Molyneux et al., 2012). These features suggest that augmenting anti-tumour immune responses, by targeting virally infected tumour cells, may complement cytotoxic therapy and potentially improve long-term disease control. Therapeutic cancer vaccines represent one such strategy. Unlike prophylactic vaccines, therapeutic vaccines aim to enhance tumour-specific immunity in individuals with established disease, overcoming deficiencies in immune priming, immune exhaustion, and tumour-mediated immune suppression (Khleif & Gupta, 2025; Pail et al., 2025; Wang & Hao, 2025). Given the limitations of cytotoxic chemotherapy and monoclonal

antibody therapy, therapeutic cancer vaccines may therefore offer a compelling strategy to enhance anti-tumour immunity while reducing systemic toxicity.

## 1.7 Cellular immunity in cancer control

Effective immune-mediated tumour control relies primarily on cellular immunity, particularly $CD8^+$ cytotoxic T lymphocytes, $CD4^+$ helper T cells and natural killer (NK) cells. $CD8^+$ T cells recognise tumour-associated peptides presented on MHC class I molecules (MHC -I) and mediate direct tumour cell killing (Martínez-Lostao et al., 2015). Accordingly, most therapeutic cancer vaccine strategies aim to induce or expand tumour-specific $CD8^+$ T-cell responses.

Unlike $CD8^+$ T cells, which mainly target MHC-I on tumour cells, $CD4^+$ T cells respond to MHC-II. Since most tumour cells do not naturally express MHC-II, the development of anti-tumour $CD4^+$ T cell responses relies heavily on professional antigen-presenting cells (APCs), which form the first line of defence against pathogens that breach the body's epithelial barriers, such as viruses or bacteria. Examples of APCs include macrophages and dendritic cells. In cancer, dendritic cells, in particular, capture tumour-associated antigens (neoantigens), process them, and present them on MHC class II molecules (Banchereau & Steinman, 1998). The $CD4^+$ T cell receptor (TCR) recognises this specific complex; depending on the cytokine environment, naïve CD4+ T cells differentiate into specific anti-tumour lineages. Th1-polarised $CD4^+$ T cells boost macrophage activation through interferon gamma (IFN-γ) secretion, increase MHC expression on tumour cells, and inhibit angiogenesis, thereby cutting off the tumour's blood supply. $CD4^+$ Cytotoxic T lymphocytes ($CD4^+$ CTLs), a

specialised subset that expresses granzyme B and perforin, can directly kill tumour cells when the tumour presents MHC class II, often in an IFN-γ- rich environment. $CD4^+$ T helper cells can also provide additional co-stimulatory signals to DCs via CD40-CD40L interactions, enhancing DC antigen presentation to $CD8^+$ T cells (Schoenberger et al., 1998). The additional proliferation and survival signals from cytokines secreted by $CD4^+$ T cells activate $CD8^+$ T cells, prolonging the cytotoxic response, preventing $CD8^+$ T cell exhaustion, and promoting the development of $CD8^+$ T cell memory. Another specialised subset of $CD4^+$ T cells, follicular helper T cells (Tfh), support humoral immunity by activating B cells.

NK cells provide a complementary innate defence, particularly against tumour cells that evade T-cell recognition through downregulating MHC-I. They mediate cytotoxicity via perforin- and granzyme-dependent mechanisms, antibody-dependent cellular cytotoxicity (ADCC), and bidirectional cross-talk with dendritic cells (Paul & Lal, 2017). Importantly, recent evidence suggests that vaccines and cytokine therapies (such as ANKTIVA) can induce memory-like NK cells (Nangia et al., 2025). These cells are thought to be primed by the vaccine to live longer (months rather than days) and to respond with much greater avidity if they encounter the tumour again. Vaccines often include adjuvants that mimic viral signals. These bind to Toll-like Receptors (TLRs) on NK cells, bypassing the need for an antigen and putting the NK cell on high alert.

## 1.8 Principles of therapeutic cancer vaccination

Therapeutic cancer vaccines aim to prime or expand tumour-specific immune responses in individuals with established disease. Unlike prophylactic vaccines, they must overcome immune tolerance, T-cell exhaustion, and an immunosuppressive tumour microenvironment (TME). By delivering tumour-associated antigens in an immunogenic context, vaccines can promote the de novo generation of tumour-specific $CD8^+$ and $CD4^+$ T cells, increase tumour immune infiltration, and synergise with chemotherapy or immune checkpoint blockade (Dolina et al., 2023). However, the clinical efficacy of therapeutic vaccination is constrained by tumours' capacity to actively evade, suppress, or ignore immune surveillance.

Therapeutic vaccine development against EBV-associated diseases has been explored for several decades, with most clinical trials conducted in NPC (Chia et al., 2012; Elliott et al., 2008; Gu et al., 1995; Hartlage et al., 2015; Hui et al., 2013; Taylor et al., 2004). Early vaccinia-based gp350 vaccines demonstrated safety but were limited by reactogenicity and modest efficacy (Gu et al., 1995). More recent approaches have shown that antigen selection and delivery platforms critically influence immunogenicity and clinical outcome. For example, MVA-based vaccines encoding EBNA1 and LMP2A induced robust T-cell responses in the majority of vaccinated patients (Hui et al., 2013), whereas dendritic-cell-based peptide vaccines showed variable immunogenicity and limited clinical benefit (Chia et al., 2012).

These experiences underscore two key principles: first, that appropriate antigen selection is essential; and second, that vaccine delivery platforms profoundly shape the magnitude, quality and durability of the immune response.

Antigen selection is further complicated by the complex and context-dependent transcriptional programmes of EBV (Abate et al., 2015; Kelly Gemma et al., 2013; Xue et al., 2002). Hence, for an effective vaccine, a careful choice of antigens is critical. EBV-specific immune responses naturally target both latent and lytic antigens. $CD8^+$ T-cell responses are initially dominated by lytic antigens during the acute infection phase, whereas latent antigen responses sustain long-term immune control of viral persistence (Benninger-Döring et al., 1999). $CD4^+$ T-cell responses are readily detectable against lytic proteins (e.g., BZLF1, BMLF1 and BCRF1), envelope proteins (gp350, gp110), and latent antigens (EBNAs, LMP2A) (Cai et al., 2021). Innate lymphocytes also contribute: natural killer (NK) cells expand markedly during acute EBV infection and preferentially target lytically replicating EBV-infected cells, particularly those with reduced MHC class I expression (Williams et al., 2005).

Despite the presence of EBV-specific immune responses, EBV-associated cancers frequently persist and progress, reflecting the broader capacity of tumours to evade immune destruction. Two non-mutually exclusive immune-evasion strategies are widely recognised: active immune suppression and immune exclusion or ignorance (Gajewski et al., 2013).

Immune suppression involves the engagement of inhibitory pathways that dampen effector cell function, including upregulation of immune checkpoint ligands e.g. programmed death-ligand (PD-L1), secretion of immunosuppressive cytokines

(such as IL-10 and TGF-β), metabolic reprogramming of the TME (including hypoxia and tryptophan depletion), and recruitment of regulatory cell populations such as regulatory T cells (Tregs), myeloid-derived suppressor cells (MDSCs), and tolerogenic macrophages (Spranger, 2016; Tang et al., 2020). PD-L1 is a constitutive protein of epithelial cells. However, its upregulation by tumour cells, especially in the TME, delivers potent inhibitory signals to tumours by binding to the PD-1 receptor expressed on T cell surfaces, thereby inducing apoptosis (Dong et al., 2002). Tryptophan is a metabolite for T cells. Its depletion via indoleamine 2,3 dioxygenase (IDO) impairs T cell proliferation (Uyttenhove et al., 2003). Hypoxia results from insufficient blood supply to proliferating tumour cells. In response, tumour cells upregulate the production of hypoxia-inducible factors, which promote the suppressive activities of MDSCs and tumour-associated macrophages (TAMs) (Kumar & Gabrilovich, 2014). On the other hand, MDSCs express a range of inhibitory molecules, including PD-L1, IL-10, nitric oxide and TGF-β, among others. These mechanisms collectively impair T-cell activation, promote functional exhaustion, and limit the durability of vaccine-induced responses.

Meanwhile, tumours can evade immune recognition through immune exclusion or ignorance, in which immune effector cells fail to infiltrate the tumour or to recognise tumour antigens effectively. This may occur due to downregulation of antigen presentation machinery (including MHC class I), loss or heterogeneity of target antigen expression, physical exclusion of immune cells by stromal barriers, or inadequate innate immune activation to support effective antigen presentation (Spranger, 2016). In EBV-associated malignancies, additional viral immune-evasion strategies—such as restricted latency programmes, viral homologues of

immunomodulatory cytokines (e.g., BCRF1 homology of human IL-10, BARF1 acting as macrophage colony-stimulating receptor), and interference with antigen processing (e.g., BNLF2a-mediated TAP inhibition, EBNA1-proteasomal degradation, EBV G-protein coupled receptor-mediated MHC-1 degradation) (Jochum et al., 2012; Ressing et al., 2008; Zuo et al., 2009)—further compound these challenges. Figure 1.7 summarises the immune escape mechanisms utilised by cancer cells to evade recognition and cytotoxic killing by the host's defence systems.

Together, these immune-evasion mechanisms represent a major barrier to the success of therapeutic cancer vaccines and provide a strong rationale for vaccine strategies that (i) incorporate multiple viral antigens to mitigate antigen loss and heterogeneity, (ii) engage both adaptive and innate immune compartments, and (iii) are deployed in rational combination

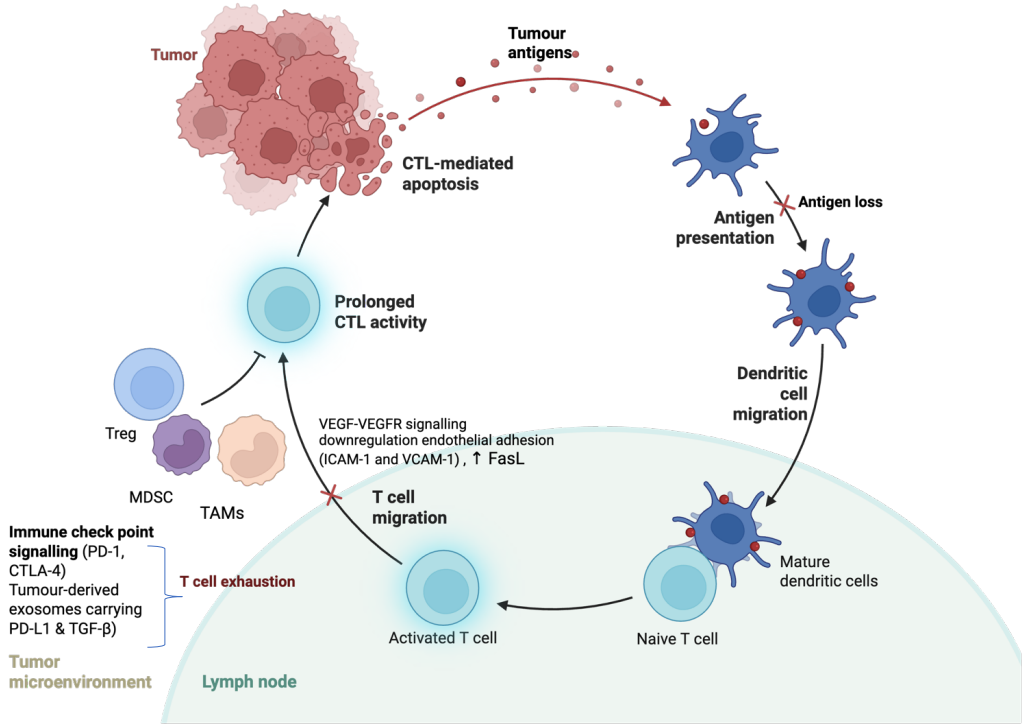



**Figure 1. 7 Immune escape mechanisms involve disruption of the cancer-immunity cycle at multiple checkpoints**

Tumour antigens are captured by antigen-presenting cells (APCs) and presented to naïve T cells in draining lymph nodes; however, defective antigen presentation and antigen loss impair T-cell priming. Activated effector T cells traffic to the tumour, but infiltration is limited by VEGF–VEGFR signalling, which downregulates endothelial adhesion molecules (ICAM-1, VCAM-1) and promotes Fas ligand expression. Within the tumour microenvironment, immune checkpoint signalling (PD-1, CTLA-4), tumour-derived exosomes that carry PD-L1 and TGF-β, and the recruitment of immunosuppressive cells (Tregs, TAMs, MDSCs) suppress effector T-cell function, thereby enabling tumour immune evasion. Figure inspired by Tang et al. (2020) and created using BioRender.

## 1.9 Viral vector vaccines and their mechanisms of action

Multiple EBV vaccine platforms have been explored, including viral vectors (adenoviral, MVA), protein subunits, virus-like particles, mRNA vaccines, and dendritic-cell-based approaches (Cai et al., 2021; Rozman et al., 2022). Viral vectors offer strong immunogenicity and efficient antigen presentation but may be limited by anti-vector immunity. While protein and peptide vaccines are safer, they

often require potent adjuvants to elicit cellular immunity. mRNA platforms provide flexible antigen design and strong innate immune activation but remain logistically challenging in low-resource settings.

This study focuses on viral vector–based vaccines. Recombinant viral vectors represent a relatively recent vaccination strategy and are particularly effective at inducing strong $CD8^+$ T cell–mediated immune responses. These vectors are typically replication-deficient, thereby eliminating the risk of reversion to a pathogenic form associated with live or attenuated vaccines (Folegatti et al., 2022; Tscherne et al., 2024; WHO, 2024). A key advantage of recombinant viral vector technology is its "platform" nature: if a new variant or a completely new disease emerges, scientists don't have to spend years in the lab weakening the virus. They can simply take the existing, proven viral vector and swap out the genetic sequence for the new target protein. This approach significantly accelerates vaccine development compared with traditional attenuation-based methods.

Although the exact mechanisms by which viral vectors activate the immune system are not fully understood, several pathways are well recognised. Following vaccination, therapeutic viral vectors infect cells at the injection site, including antigen-presenting cells (APCs), such as dendritic cells (DCs), thereby inducing endogenous expression of tumour-associated antigens (Hollingsworth & Jansen, 2019). In addition, antigens released from infected cells can be captured by DCs and cross-presented (van Montfoort et al., 2014). Alternatively, DCs may also acquire pre-formed MHC class I–peptide complexes from infected cells via a process known as cross-dressing (Dolan et al., 2006).

Viral infection triggers innate immune sensing pathways, including pattern recognition receptors (PRRs), leading to type I interferon production and DC maturation (Hollingsworth & Jansen, 2019). Activated DCs migrate to the draining lymph nodes, where processed peptides are presented to naïve T cells via MHC class I and II molecules, driving the activation of $CD8^+$ and $CD4^+$ T cells, respectively, in the presence of co-stimulatory signals (Hollingsworth & Jansen, 2019). This leads to the differentiation of naïve $CD8^+$ T cells into cytotoxic T lymphocyte effectors, supported by $CD4^+$ T-cell help. Activated T cells subsequently exit the lymph nodes and traffic to the tumour site in response to chemokine gradients. Upon recognition of a cognate antigen presented by tumour cells, cytotoxic T lymphocytes mediate tumour cell killing and regression by releasing effector cytokines, including IFN-γ and TNF-α, as well as cytotoxic molecules such as perforin and granzyme B (Hollingsworth & Jansen, 2019). Figure 1.8 summarises potential mechanisms by which vaccine-induced cytotoxic T cells kill virally infected tumour cells.

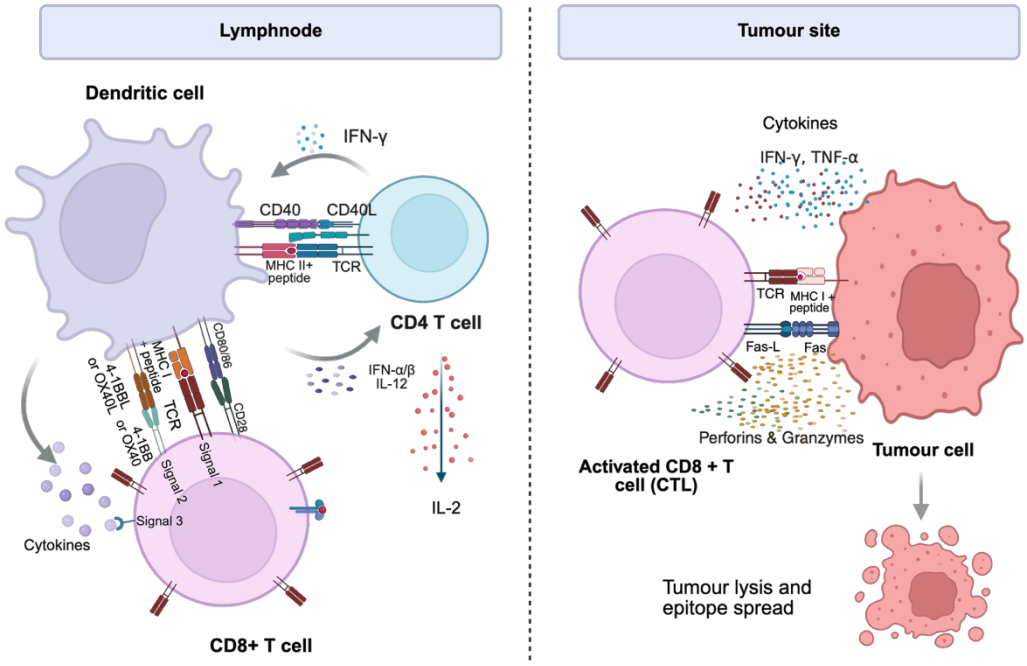



**Figure 1. 8 Mechanisms of therapeutic cancer vaccines**

Vaccination delivers tumour-specific antigens in an inflammatory context, promoting antigen uptake and processing by antigen-presenting cells (APCs), notably dendritic cells (DCs), at the injection site. Activated DCs migrate to tumour-draining lymph nodes, where processed tumour-derived peptides are presented to naïve T cells via MHC molecules. T-cell activation requires three signals: T-cell receptor (TCR) engagement with peptide–MHC complexes (signal 1), co-stimulatory signalling via CD80/86–CD28 and additional co-stimulatory receptors such as OX40 and 4-1BB (signal 2), and inflammatory cytokines including IL-12 and type I interferons (signal 3). These signals promote the proliferation and differentiation of $CD8^+$ T cells into effector cytotoxic T lymphocytes (CTLs) and memory T cells. $CD4^+$ Th1 cells further support CTL expansion and function by secreting IL-2. Activated CTLs traffic to the tumour site, where they recognise tumour cells presenting cognate antigen via MHC class I and mediate tumour cell killing through perforin–granzyme release and Fas–FasL-dependent apoptosis. Figure inspired by Hollingsworth & Jansen, (2019). Created using BioRender.

A range of viral vectors has been evaluated for recombinant vaccine development, including adenoviruses, avian poxviruses (e.g., fowl poxvirus), mammalian poxviruses (e.g., vaccinia virus ), flaviviruses (such as live-attenuated yellow fever 17D virus), and cytomegaloviruses (Bonaldo et al., 2010; Bonaldo et al., 2007; Capone et al., 2013; Hansen et al., 2009). Many of these platforms have

demonstrated acceptable safety and efficacy profiles in humans, both in infectious disease and cancer vaccine settings (Larocca & Schlom, 2011). Among them, the Chimpanzee-derived adenovirus Oxford (ChAdOx) and Modified Vaccinia Ankara (MVA) vectors have shown particular promise due to their capacity to elicit strong $CD8^+$ T cell responses.

Importantly, heterologous prime–boost strategies combining ChAdOx and MVA vectors have been shown to induce more durable T cell responses than homologous regimens (Gilbert et al., 2002; Reyes-Sandoval et al., 2010; Swadling et al., 2014). This enhanced efficacy is attributed to the avoidance of neutralising antibodies against vector capsid proteins that develop after initial vaccination and limit vector infectivity upon repeated administration. While MVA vectors are generally less effective at priming $CD8^+$ T cell responses, they exhibit strong boosting capacity, making them well-suited for use in heterologous vaccine strategies (Reyes-Sandoval et al., 2010). However, a recent finding by McAuliffe et al. indicates that ChAdOx/MVA regimens can extend the duration of protective immunity compared with homologous prime–boost approaches (McAuliffe et al., 2021).

However, a significant limitation of adenoviral vectors is the high prevalence of pre-existing neutralising antibodies in human populations due to pre-exposure to these viruses in natural environments (Nwanegbo et al., 2004). This pre-existing immunity can notably reduce vaccine immunogenicity and effectiveness. Nonetheless, assessment of chimpanzee-derived adenoviral vectors reveals a much lower seroprevalence of neutralising antibodies in humans, making them suitable for vaccine platforms (Dudareva et al., 2009).

## 1.10 Engineering EBV vaccines for cancer immunotherapy

Given EBV's complex life cycle and immune evasion strategies, no single antigen is sufficient for durable tumour control. Therefore, strategies targeting both latent and lytic antigens may enhance vaccine effectiveness. Latent antigen vaccines aim to eradicate viral reservoirs within tumour cells and latently infected memory B cells, while responses to lytic antigens help control viral replication, spread, and immune escape. A vaccine approach combining latent and lytic EBV antigens thus offers the greatest potential to simultaneously engage $CD8^+$ T cells, $CD4^+$ T cells, and NK cells that can recognise and kill tumour cells expressing these viral antigens. This broad epitope approach has both therapeutic and prophylactic relevance and may be particularly advantageous for EBV-associated malignancies with diverse latency patterns, including endemic BL, NPC, gastric carcinoma, Hodgkin lymphoma, and EBV-associated T/NK-cell lymphomas.

Together, these considerations support the development of a multi-antigen EBV therapeutic vaccine delivered via an immunogenic platform that activates complementary cytotoxic lymphocyte populations, thereby addressing both viral persistence and tumour immune escape in endemic BL.

## 1.11 Aims of the thesis

The main aim of this project was to characterise EBV genomic and transcriptional features that contribute to viral diversity and pathogenesis, and to identify viral determinants with potential relevance for immunogenic target discovery and vaccine development, with a particular focus on Burkitt lymphoma.

The specific objectives this thesis aimed to address were:

i) To characterise the genomic diversity and evolutionary dynamics of Epstein–Barr virus, with a focus on BL pathogenesis
ii) To define the Epstein–Barr virus transcriptional landscape and antigen expression in Burkitt lymphoma tumours
iii) To evaluate the immunogenic potential of selected Epstein–Barr virus antigens in murine models

# Chapter 2: Materials and Methods

## 2.1 Cell-free DNA workflow

### 2.1.1 Materials and equipment

Table 2.1 lists the reagent kits and equipment used for cell-free DNA processing, DNA extraction, library preparation, and Illumina sequencing.

**Table 2. 1 cfDNA reagents, kits and equipment**

| Item | Manufacturer | Cat No. |
|---|---|---|
| PAXgene Blood ccfDNA tubes | Qiagen | 768165 |
| QIAamp circulating Nucleic Acid kit | Qiagen | 55114 |
| Qubit dsDNA HS Assay Kit | Thermo Fisher Scientific | Q32854 |
| Bioanalyzer High Sensitivity DNA kit | Agilent Technologies | 5067-5584 |
| ThruPLEX® Tag-Seq HV kit | Takara Bio | R400735 |
| ThruPLEX HV UDI Set A | Takara Bio | R400738 |
| xGen MRD Hyb Panel, 16 rxn | Integrated DNA Technologies | |
| xGen™ Hybridization and Wash Reagents v3, 96 rxn | Integrated DNA Technologies | 10028312 |
| xGen™ Universal Blockers TS, 4x96 rxn | Integrated DNA Technologies | 1075476 |
| xGen™ Library Amplification Primer Mix, 96 rxn | Integrated DNA Technologies | 1077676 |
| xGen™ Human Cot DNA, 650 µL | Integrated DNA Technologies | 1080769 |
| Dynabeads M-270 Streptavidin beads | Thermo Fisher Scientific | 65306 |
| KAPA HiFi HotStart Ready Mix | Roche Diagnostics | 9420398001 |
| Agencourt AMPure XP-PR Purification Beads | Beckman Coulter | A63880 |
| Buffer EB (10 mM Tris-HCl, pH 8.5) | Qiagen | 19086 |
| Molecular Grade Water | Avantor | 436912C-VWR |
| Ethanol absolute 99.8% | Fisher Chemical | 10437341 |
| Illumina MiSeq v2 Reagent Kit 300 Cycles | Illumina | MS-102-2002 |
| QIAvac 24 Plus vacuum system | Qiagen | 19413 |
| Invitrogen Qubit 4 Fluorometer | Thermo Fisher Scientific | Q33238 |
| TapeStation or Bioanalyzer | Agilent Technologies | G2991BA |
| MiSeq System | Illumina | SY-410-1003 |

### 2.1.2 Participant recruitment and ethical considerations

Participants were drawn from the Aggressive Infection-Related East Africa Lymphoma (AI-REAL) study, a prospective case–control study conducted across multiple sites in East Africa. The detailed study design has been published previously (Legason et al., 2022).

Briefly, participant recruitment occurred in two phases. During Phase I (2020–2024), children diagnosed with Burkitt lymphoma (BL) were recruited from four clinical sites: St Mary's Hospital, Lacor (Uganda), Muhimbili National Hospital (Tanzania), Kilimanjaro Christian Medical Centre, Moshi (Tanzania), and Bugando Medical Centre, Mwanza (Tanzania). In Phase II (2024), a limited number of healthy children were recruited only at the Ugandan site.

Appropriate ethical approvals for patient recruitment and study conduct were obtained from the relevant regulatory bodies: the Oxford Tropical Research Ethics Committee (OxTREC; Ref. 15-19), the National Institute for Medical Research in Tanzania (NIMR/HQ/R.8a/Vol.IX/3408), the Uganda National Council for Science and Technology (HS529ES), and the St Mary's Hospital Lacor Institutional Research Ethics Committee, Gulu, Uganda.

Written informed consent was obtained from adult participants or from parents or legal guardians of minors, with assent obtained from children aged 7–17 years. All study procedures complied with the Declaration of Helsinki and applicable international data protection regulations.

### 2.1.3 Sample collection and processing

Peripheral venous blood samples were collected in Paxgene blood ccfDNA tubes (Qiagen) containing a preservative that stabilised blood cells and prevented cellular DNA contamination of the plasma. The samples were transported in secure cold boxes maintained at 2-8 °C to the local research laboratory. At the laboratory, samples were separated initially at a lower speed of 1600 × g for 10 minutes, and the resulting plasma was then centrifuged at a higher speed (4500 × g) for an additional 10 minutes. The double-spun plasma was frozen at -80 degrees until it was required for circulating nucleic acid extraction.

### 2.1.4 Cell-free DNA extraction and quality assessment

Cell-free DNA (cfDNA) consists of apoptotic and necrotic cell products released into the bloodstream (Wan et al., 2017). In neoplastic disease, these products may include tumour fragments, hence the term ‘circulating tumour DNA’ or ctDNA. Frequently, the EBV genome exists as episomal DNA within the host cell, and in some cases, it is integrated into the host genome (Xiao et al., 2016; Xu et al., 2019). Therefore, analysing circulating tumour DNA from apoptotic tumour cells provides a non-invasive way to assess the tumour-associated viral genome.

cfDNA was extracted from 4–5 mL of plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany), following the manufacturer’s instructions (Figure 2.1). This method is based on the selective binding of nucleic

acids to silica-based QIAamp Mini columns under chaotropic conditions, allowing contaminants such as proteins and divalent cations to be removed during subsequent wash steps.

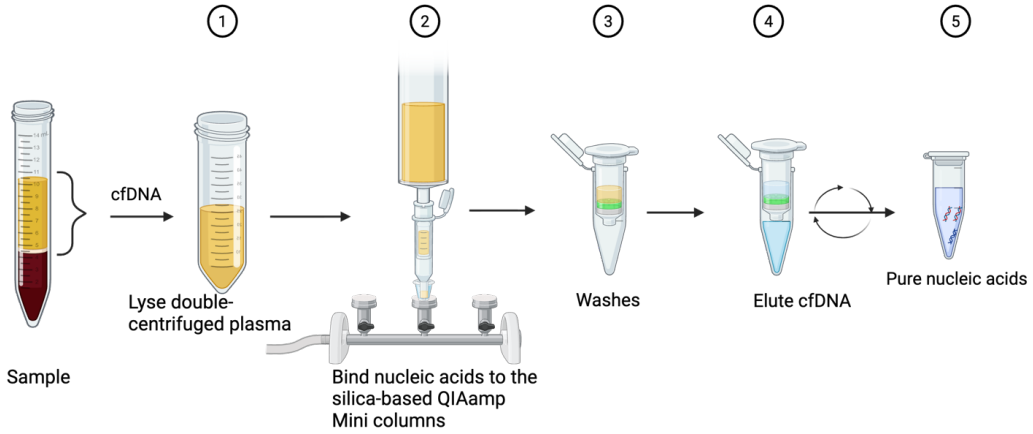



**Figure 2. 1 Steps of cfDNA extraction**

The schema illustrates the cfDNA workflow from sample separation to the extraction of pure nucleic acids. In step (1), lysis buffers and enzymes are added to the 4-5 ml of double-centrifuged plasma sample to break down proteins and release cfDNA. The binding buffer is then introduced to the lysate, which is transferred to a QIAamp mini column connected to a QIAvac 24 Plus vacuum manifold in step (2). After DNA binds to the silica-based column membranes, impurities are eliminated through washing, and the membrane is dried by centrifugation in step (3). The final stages (4 & 5) involve eluting the DNA in an appropriate buffer. Figure drawn using BioRender.

Prior to cfDNA extraction, buffers ACB, ACW1, and ACW2 (supplied as concentrates) were diluted as per the manufacturer's instructions. Buffer ACB was prepared by adding 200 mL of 100% isopropanol to 300 mL of ACB concentrate, then mixing by inversion 10 times. Buffer ACW1 was prepared by adding 25 ml of ethanol (96–100%) to 19 ml of ACW1 concentrate to reach a final volume of 44

ml, and the solution was mixed by inversion 10 times. Buffer ACW2 was prepared by adding 30 ml of ethanol (96–100%) to 13 ml of ACW2 concentrate to reach a final volume of 43 ml, then mixed thoroughly by inverting 10 times.

Carrier RNA used to improve nucleic acid binding to the QIAamp Mini membrane, especially in samples with low levels of target cfDNA, was supplied as 310 µg lyophilised powder. To prepare the lyophilised carrier RNA, the tube was briefly pulse-spun, and 1,550 µl of Buffer AVE was added to 310 µg of carrier RNA powder, yielding a working concentration of 0.2 µg/µl. The solution was gently pipetted and then aliquoted into 0.5 ml Eppendorf tubes at 30 µl each. Aliquots were stored at −20 °C until needed. For each experiment, a fresh aliquot was thawed, and any unused carrier RNA was discarded.

Plasma samples were equilibrated to room temperature and, when required, adjusted to a final volume of 5 mL with phosphate-buffered saline (PBS). ACL buffer supplemented with carrier RNA (approximately 1 mL per 5 mL plasma) was added, and samples were gently inverted approximately 10 times before incubation at 60 °C for 1 hour to facilitate lysis and nucleoprotein dissociation. Following lysis, 9 mL buffer ACB, pre-warmed to 56 °C, was added and mixed thoroughly until homogeneous. Samples were then incubated on ice for 5 minutes to facilitate nucleic acid binding.

The lysate was transferred to a QIAamp Mini column, and vacuum pressure was applied until the lysate passed entirely through the membrane. Columns were washed sequentially with 600 µL of buffer ACW1 and 750 µL of buffer ACW2, followed by an additional wash with 750 µL of absolute ethanol to remove residual

salts and buffer components. Columns were centrifuged at 12,860 × g for 3 minutes, transferred to clean collection tubes, and incubated at 56 °C for 10 minutes to ensure complete drying of the membrane. cfDNA was eluted by applying 40 µL buffer AVE to the column membrane, incubating for 3 minutes, and centrifuging at 12,860 × g for 1 minute. Approximately 35 µL cfDNA was recovered per 5 mL plasma.

The concentration of extracted cfDNA was measured using the Qubit dsDNA High Sensitivity Assay (Thermo Fisher Scientific, USA). This fluorometric assay selectively binds intact double-stranded DNA, with fluorescence intensity proportional to DNA concentration. Working solutions, standards, and controls were prepared in accordance with the manufacturer's instructions (Thermo Fisher Scientific, USA). cfDNA samples were diluted 1:200 in working solution, vortexed, incubated in the dark for 2 minutes, and measured using the Qubit fluorimeter 4.0 (Invitrogen). The assay has a linear detection range of 0.1–120 ng, and control values between 9–11 ng/µL were considered acceptable.

The fragment size distribution of cfDNA was assessed using the Agilent 4200 TapeStation system with the cfDNA ScreenTape Assay (Agilent Technologies, USA). Reagents were equilibrated to room temperature for 30 minutes prior to use. For each sample, 2 µL cfDNA sample buffer was combined with 2 µL cfDNA ladder or sample, vortexed at 2,000 rpm for 1 minute using an IKA MS3 vortexer and centrifuged for 1 minute. Prepared samples were loaded onto the TapeStation instrument for electrophoretic analysis (Figure 2.2).

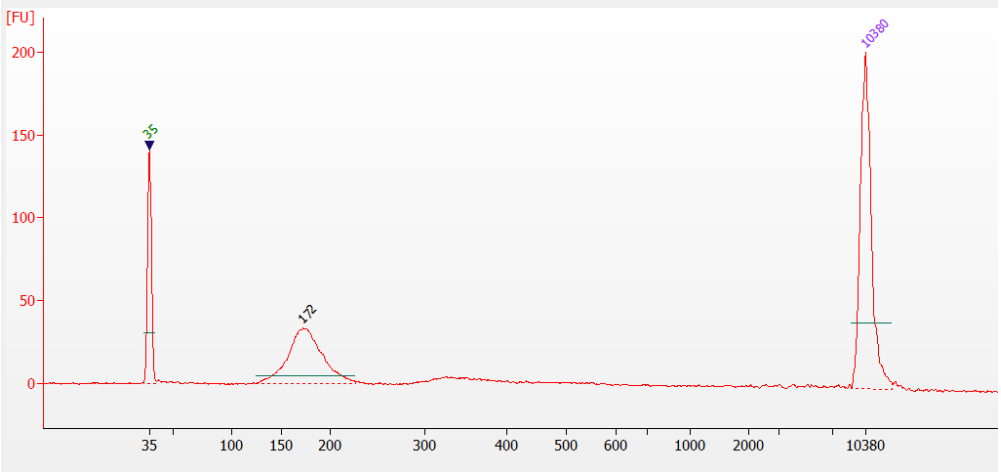



**Figure 2. 2 Size distribution of plasma cfDNA analysed by Agilent Bioanalyser 2100**

Electropherogram derived from a sample analysed by Agilent Bioanalyser 2100. A dominant peak at approximately 172 bp, which corresponds to a mononucleosomal cfDNA fragment, is shown. A smaller secondary peak at around 320 bp, indicative of traces of dinucleosomal fragments, can also be observed. The lower and upper DNA ladders are also shown. The lack of significant high–molecular-weight DNA suggests minimal genomic DNA contamination and confirms good sample integrity. This analysis is essential for the success of downstream cfDNA analysis, ensuring the correct input and concentrations.

### 2.1.5 Preparation of cell-free DNA libraries for Illumina sequencing

cfDNA libraries were prepared using the ThruPLEX Tag-Seq HV kit (Takara Bio, UK), which employs stem-loop adapters rather than conventional Y-adapters to enable rapid, single-tube library construction with minimal sample loss. This design improves ligation efficiency and reduces adapter-derived background noise. Library preparation begins with the enzymatic repair of cfDNA fragments to generate blunt ends. Stem-loop adapters are subsequently ligated to the 5′ ends of the DNA fragments, leaving a single nick at the 3′ end. As these adapters lack overhangs, nonspecific ligation is minimised. The final steps involve extending the ligated fragments, cleaving the stem-loop adapters, incorporating Illumina-compatible sample indices, and performing high-fidelity PCR amplification,

followed by removal of excess adapters and reagents in a single-tube workflow (Figure 2.3).

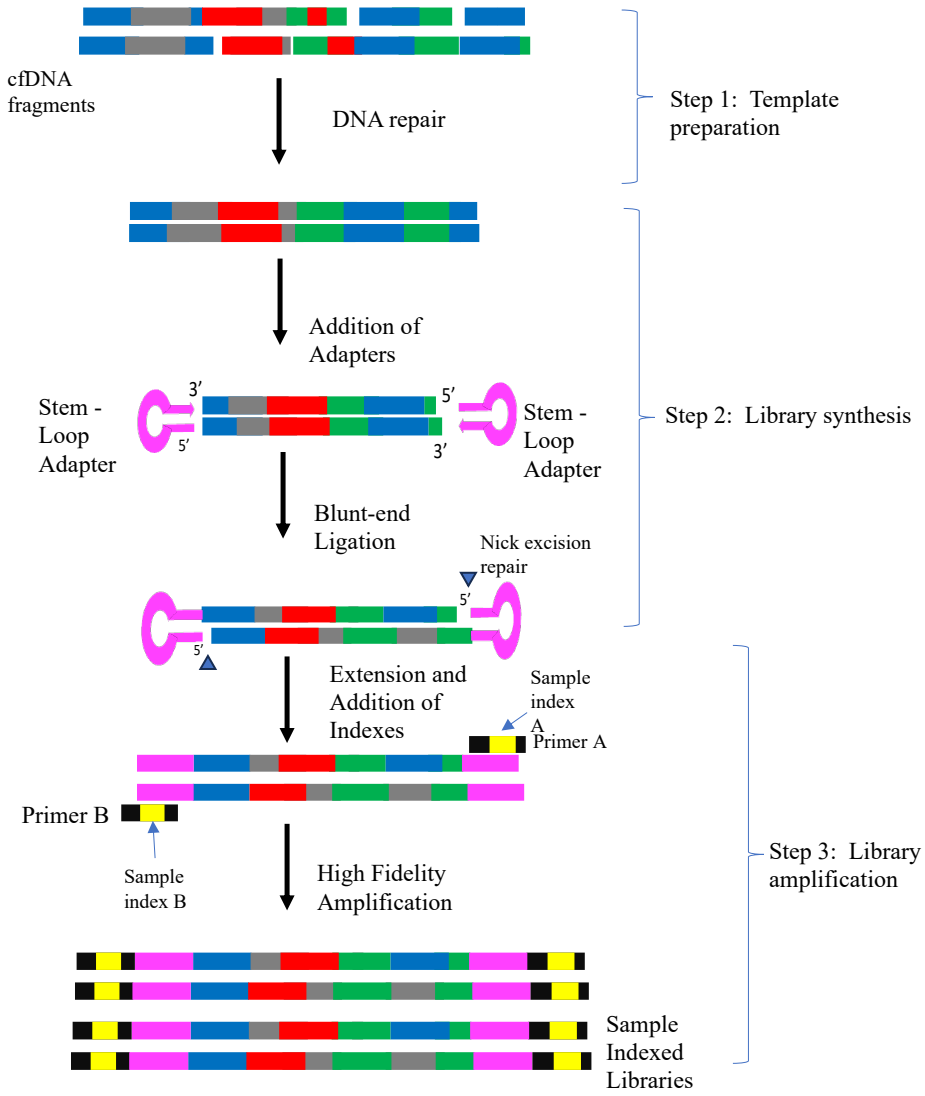



**Figure 2. 3 Cell-free DNA library preparation workflow for Illumina sequencing, adapted from (Takara Bio Inc. UK)**

Diagram drawn using Microsoft PowerPoint.

cfDNA template preparation was performed using a DNA repair enzyme to fill 5′ overhangs and generate blunt-ended fragments suitable for adapter ligation. A total of 10–50 ng cfDNA was used per sample in a final volume of 30 µL, calculated by dividing the required DNA input by the measured sample concentration. Sample volumes were adjusted to 30 µL using elution buffer (EB). Buffer PBT1 and enzyme PET1 were thawed on ice, briefly vortexed, and centrifuged. The template master mix was prepared by combining 5 µL PBT1 with 1 µL PET1 enzyme. 6 µL of this master mix was added to each 30 µL cfDNA sample and gently mixed by pipetting. Samples were incubated in a thermocycler under the following conditions: heated lid at 99 °C; 22 °C for 25 minutes; 55 °C for 20 minutes; followed by a 4 °C hold for up to 2 hours. Stem-loop adapter ligation was performed using ligase SET1. Buffer SBT1 and enzyme SET1 were thawed on ice, vortexed, and briefly centrifuged. The ligation master mix was prepared by combining 2 µL SBT1 buffer with 4 µL SET1 enzyme per reaction.

Following template preparation, 6 µL of the ligation master mix was added to each sample, mixed gently, and incubated in a thermocycler with the following programme: heated lid at 99 °C; 30 °C for 40 minutes; and a 4 °C hold for up to 30 minutes. The final library synthesis step involved extension of the cfDNA templates, cleavage of the stem-loop adapters, and PCR amplification using enzyme AET1, during which Illumina-compatible unique dual indices (UDI) were incorporated. Buffer ABT1 and enzyme AET1 were thawed on ice, vortexed, and centrifuged. The amplification master mix was prepared by combining 46 µL

ABT1 buffer with 2 µL AET1 enzyme per sample. Meanwhile, the index plate was equilibrated to room temperature (25 °C).

Following ligation, 48 µL of amplification master mix was added to each library and mixed thoroughly. The index plate was briefly centrifuged, wiped with ethanol, and pierced to add 10 µL ThruPLEX HV UDI to each sample. After mixing and recording index assignments, samples were amplified in a thermocycler using the conditions described in Table 2.2.

**Table 2. 2 Thermocycler programme for amplification of cfDNA library**

| Step | Temp ℃ | Time | Cycles |
|---|---|---|---|
| Heated lid | 99 | | |
| Step 1 | 72 | 3 min | 1 |
| Step 2 | 85 | 2 min | 1 |
| Step 3 | 98 | 2 min | 1 |
| Step 4 | 98 | 20 sec | 7 (50ng) |
| | | | 8(25ng) |
| | 65 | 75 sec | 9(10ng) |
| Step 5 | 68 | 5 min | 1 |
| Step 6 | 4 | ∞ | 1 |

Amplified libraries were purified using AMPure XP beads. Beads were resuspended, and 100 µL was added to each sample. Samples were mixed by pipetting 10 times and incubated at 25 °C for 5 minutes. Following a brief centrifugation, the tubes were placed on a magnetic rack until the supernatant cleared, after which it was discarded. Beads were washed twice with 200 µL freshly prepared 80% ethanol, incubating for 30 seconds each time. After the second wash, residual ethanol was removed following brief centrifugation. Beads were air-dried

for approximately 2 minutes until no longer shiny but before cracking. DNA was eluted by resuspending beads in 50 µL nuclease-free water, incubating at 25 °C for 2 minutes, and separating on a magnetic rack. Approximately 50 µL of purified library supernatant was transferred to a low-adhesion tube for EBV DNA enrichment.

Target enrichment was performed using a custom xGen hybridisation capture panel designed to cover the entire EBV genome (~172 kb). The panel comprised 120-bp biotinylated probes with a 60-bp overlap, designed by Ms Daisy Jennings at the Oxford Molecular Diagnostic Centre (OMDC) and manufactured by Integrated DNA Technologies (IDT, USA). Up to six indexed libraries were pooled for capture. Hybridisation reagents were equilibrated to room temperature (25 °C), and 7.5 µL human Cot-1 DNA was added to each pooled library to suppress repetitive sequence hybridisation. AMPure XP beads were added at 1.8× the total library volume and incubated at 25 °C for 10 minutes, followed by magnetic separation and two washes with 80% ethanol. The hybridisation mix was prepared by combining 9.5 µL 2× hybridisation buffer, 3 µL hybridisation enhancer, 2 µL blocking oligos, and 4.5 µL EBV probe pool. After a 5-minute incubation at 25 °C, 17 µL of the supernatant was transferred to a PCR tube and hybridised under the following conditions: denaturation at 95 °C for 30 seconds, hybridisation at 65 °C for 20 hours, with a heated lid at 99 °C.

Wash buffers—2× Bead Wash Buffer (BWB), 10× Wash Buffer 1 (WB1), 10× Wash Buffer 2 (WB2), 10× Wash Buffer 3 (WB3), and 10× Stringent Wash Buffer

(SWB)—were diluted to 1× working concentration using Sigma water prior to use. For preparation of a single IDT pool, the following volumes were prepared: 160 µL of 2× BWB was diluted with an equal volume of Sigma water; 28 µL of 10× WB1 was diluted with 252 µL of Sigma water; 16 µL of 10× WB2 was diluted with 144 µL of Sigma water; 16 µL of 10× WB3 was diluted with 144 µL of Sigma water; and 32 µL of 10× SWB was diluted with 288 µL of Sigma water. BWB, WB2 and WB3 were stored at room temperature, while WB1 (110 µL) and Stringent Wash Buffer (SWB; 160 µL) were pre-warmed to 65 °C. Streptavidin M-270 Dynabeads were equilibrated to room temperature for 30 minutes and vortexed. 50 µL beads were washed twice with 1× BWB and resuspended in 17 µL bead resuspension mix. Beads were added to the hybridised library pool and incubated at 65 °C for 60 minutes with periodic mixing.

Following capture, beads underwent sequential washes with pre-heated WB1 and SWB at 65 °C, followed by additional washes with WB1, WB2, and WB3 at room temperature, with alternating vortexing and resting to ensure thorough washing. After the final wash, libraries were resuspended in 20 µL nuclease-free water.

Enriched EBV libraries were amplified using KAPA HiFi HotStart DNA polymerase (Roche Diagnostics) for 18 PCR cycles (Quail et al., 2012). Each reaction contained 25 µL of 2× KAPA ReadyMix, 1.25 µL of xGen amplification primer, and 3.75 µL of nuclease-free water. Amplified libraries were purified using 75 µL AMPure XP beads, washed twice with 125 µL 80% ethanol, air-dried, and eluted in 22 µL EB buffer. Following magnetic separation, 20 µL of purified

libraries were transferred to nuclease-free tubes for sequencing (Summary of steps and IDT workflow in Figure 2.4).

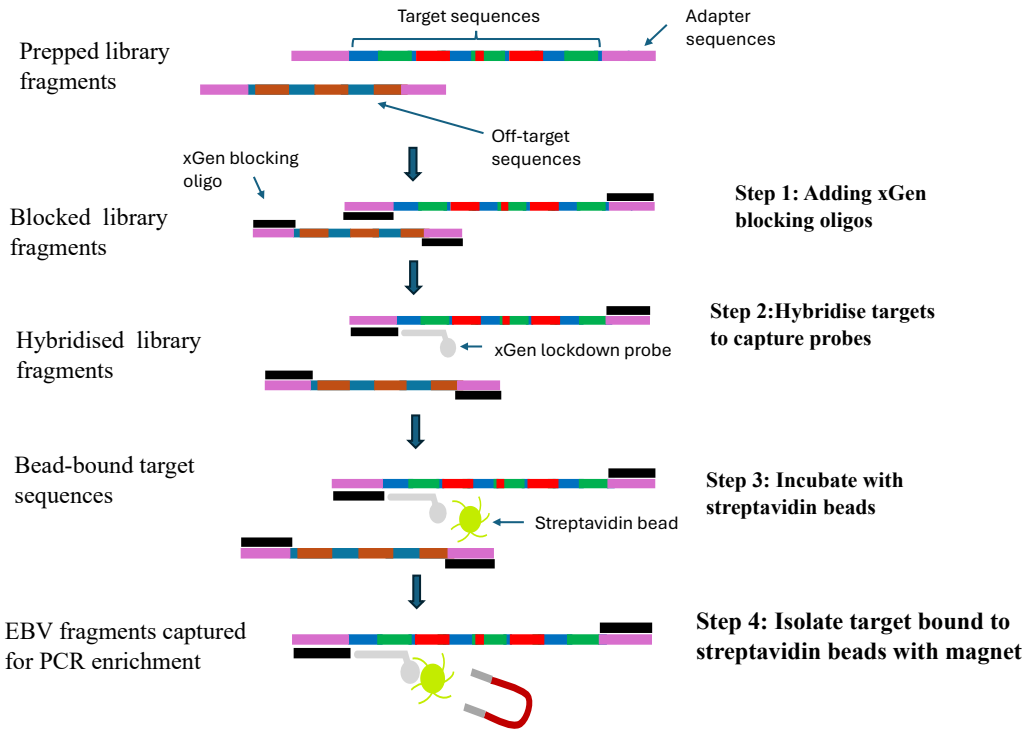



**Figure 2. 4 EBV DNA capture workflow using magnetic streptavidin beads, adapted from (IDT Inc. USA)**

Diagram drawn in Microsoft PowerPoint.

### 2.1.6 Illumina sequencing of cfDNA libraries

Sequencing libraries were diluted in buffer EB to a final concentration of 4 nM, and equal volumes of each library were pooled into a 1.5 mL Eppendorf tube. Library denaturation was performed by combining an equal volume of the pooled

library with 0.2 M NaOH (5 µL) and incubating at room temperature (25 °C) for 5 minutes. The denatured library was then diluted by adding 990 µL pre-chilled HT1 buffer to achieve a final concentration of 20 pM. This solution was further diluted to 10 pM by mixing 300 µL of the denatured library with an equal volume of pre-chilled HT1 buffer.

Prior to sequencing, a 5% Phi X control library was spiked into the final library pool. Libraries were loaded into a MiSeq v2 reagent cartridge, and cluster generation and sequencing-by-synthesis (SBS) were performed on the Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA). Sequencing was carried out using a paired-end configuration, consisting of 150 cycles for Read 1, 8 cycles for index i7, 8 cycles for index i5, and 150 cycles for Read 2.

## 2.2 FFPE RNA workflow

### 2.2.1 Materials and equipment

Table 2.3 shows the reagent kits and equipment used for processing RNA samples from formalin-fixed paraffin-embedded sections and for downstream processes, including Illumina sequencing.

**Table 2. 3 FFPE RNA reagents, kits and equipment**

| Item | Manufacturer | Cat No. | Purpose |
|---|---|---|---|
| RNeasy FFPE kit | Qiagen | 73504 | RNA extraction |
| Deparaffinisation solution | Qiagen | 19093 | Removal of paraffin from sections |
| Qubit RNA High Sensitivity Kit | Thermo Fisher Scientific | Q32852 | RNA quantification |
| TapeStation High Sensitivity RNA Reagent kit & Screen Tape | Agilent | 5067-5576 | RNA fragment analysis |
| NEBNext rRNA Depletion Kit v2 (Human/Mouse/Rat) | New England Biolabs | E7405 | ribosomal RNA depletion and enrichment of non-ribosomal RNA |
| NEBNext Ultra II RNA Library Prep Kit | New England Biolabs | E7765 | RNA library preparation |
| NEBNext® Multiplex Oligos (UDIs) | NEB | E6440S | Required for indexing and library amplification. |
| Agencourt RNA Clean XP Beads | Beckman Coulter | A63880 | RNA purification and size selection |
| Nuclease-free Water | Avantor | 436912C-VWR | RNA elution; free of nucleases |
| 80% Ethanol (freshly prepared) | Fisher Chemical | 10437341 | Washing agent for magnetic beads to remove salts and contaminants during RNA extraction |
| Qubit 4 Fluorometer | Thermo Fisher Scientific | Q33238 | RNA yield measurement |
| TapeStation or Bioanalyzer | Agilent | G2991BA | RNA fragment analysis |
| Nanodrop | Thermo Fisher Scientific | NDLPLUSGL | RNA purity and yield measurements |
| Illumina NovaSeq X Plus | Illumina | 20084804 | paired-end short-read sequencing |
| NovaSeq X Series 1.5B Reagent Kit | Illumina | 20085594 | Consumables (flow cell/buffers) for sequencing run |

### 2.2.2 Tissue collection, processing and sectioning

Diagnostic core needle biopsy specimens from eligible patients identified in Section 2.1.2 were fixed immediately in 10% neutral buffered formalin (pH 6.9–7.1) to preserve tissue morphology and nucleic acid integrity. Samples were processed according to standard operating procedures at the respective accredited pathology laboratories, and formalin-fixed, paraffin-embedded (FFPE) tissue blocks were retrieved for histopathological assessment and downstream molecular analyses.

A total of 26 FFPE tissue blocks containing sufficient tumour material, as determined by histopathological review, were selected for RNA extraction and transcriptome profiling. These samples were analysed to investigate EBV transcriptional activity within tumour tissue and to complement findings obtained from plasma-derived EBV genomic analyses.

For RNA extraction, 10 serial sections per FFPE block, each cut at approximately 5 μm thickness, were cut using a microtome. Sections were floated on diethyl pyrocarbonate (DEPC)-treated, nuclease-free water and mounted onto negatively charged, uncoated glass slides. To minimise cross-contamination, a new microtome blade and fresh DEPC-treated water were used for each tissue block.

Excess paraffin surrounding the tissue sections was carefully trimmed with a sterile scalpel blade, and the sections were immediately transferred into 2 mL nuclease-free microcentrifuge tubes for RNA extraction. All handling steps were performed using RNase-free reagents and consumables.

### 2.2.3 RNA Extraction and Quality Assessment

Total RNA was extracted from FFPE tissue sections using the RNeasy FFPE Kit (Qiagen, Hilden, Germany), following the manufacturer's protocol (Figure 2.5). Prior to first use, buffer RPE was prepared by adding 44 mL of ethanol (96-100%) to the bottle containing 11 mL of buffer RPE concentrate. Meanwhile, the DNase 1 stock solution was prepared by dissolving the lyophilised DNase 1 in 550 µl of RNase-free water, then gently mixed by inverting the vial. The solution was aliquoted into 50 µl volumes and stored at -20 °C; thawed aliquots were kept at 2-8 °C for up to 6 weeks.

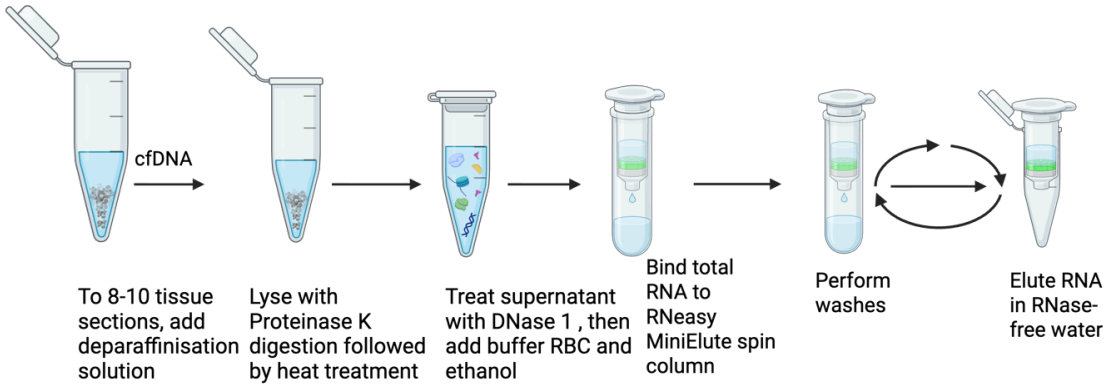



**Figure 2. 5 RNA extraction and purification from FFPE sections using the RNeasy FFPE kit, adapted from (Qiagen, Hilden, Germany)**

Diagram drawn in BioRender.

To the sample, 320 µL of deparaffinisation solution was added. Samples were vortexed briefly, centrifuged to collect the material at the bottom of the tube, incubated at 56 °C for 3 minutes, then cooled to room temperature (25 °C).

Following deparaffinisation, 240 µL buffer PKD was added, and samples were vortexed and centrifuged for 1 minute at 11,000 × g. Tissue digestion and reversal of formalin-induced crosslinks were performed by adding 10 µL proteinase K to the lower aqueous phase and incubating at 56 °C for 15 minutes, followed by 80 °C for 15 minutes on a shaking heating block to facilitate homogenisation.

The lower, uncoloured phase was transferred to a fresh 2 mL microcentrifuge tube, incubated on ice for 3 minutes, and centrifuged at 20,000 × g for 15 minutes to remove residual debris. The clarified supernatant was transferred to a new tube, and genomic DNA was removed by adding 25 µL DNase booster buffer and 10 µL DNase I stock solution, then incubating at room temperature (25°C) for 15 minutes.

To adjust binding conditions, 500 µL buffer RBC was added, followed by 1,200 µL absolute ethanol, and the mixture was thoroughly mixed by pipetting. The lysate was applied in 700 µL aliquots to the RNeasy MinElute spin column, centrifuged for 15 seconds at 8,000 × g, and the flow-through discarded. This step was repeated until the entire sample had passed through the column.

Columns were washed twice with 500 µL of buffer RPE, then centrifuged. To remove residual ethanol, columns were transferred to fresh collection tubes and centrifuged at full speed with the lid open for 5 minutes. RNA was eluted in 30 µL RNase-free water and stored at −30 °C until further processing.

The amount of RNA extracted was measured using the Qubit Fluorimeter (Life Technologies, Carlsbad, CA, USA), and its integrity was checked with an RNA kit on the Agilent 5300 Fragment Analyser or TapeStation (Agilent Technologies, Palo Alto, CA, USA). Reagents were brought to room temperature (25 °C) before use. RNA ladder preparation, sample denaturation, and fragment analysis were conducted following the manufacturer's instructions. RNA integrity metrics, including DV200, were obtained from electropherograms (Figure 2.6).

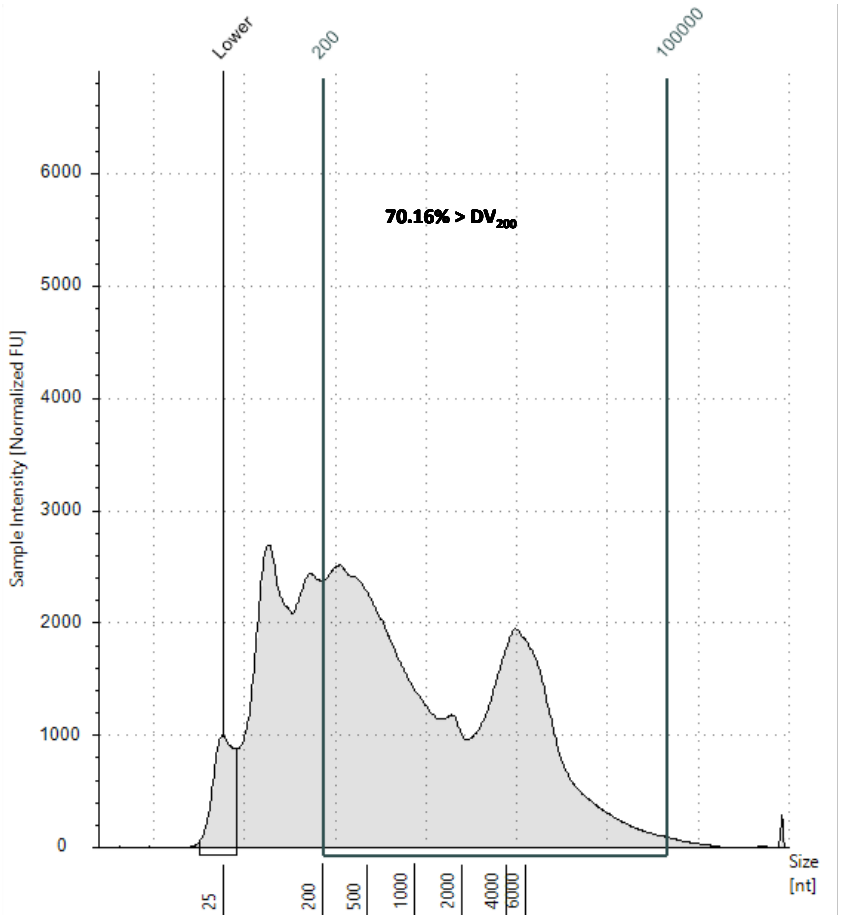



**Figure 2. 6 Assessment of FFPE RNA quality using DV200 derived from the electropherogram.**

The figure shows an electropherogram analysed using High Sensitivity RNA ScreenTape. Since RNA integrity (RIN) scores depend heavily on intact rRNA peaks (18S and 28S), these peaks disappear in FFPE or degraded samples. Therefore, Distribution Value (DV200) provides a better reflection of the usable RNA fragment length for library preparation. DV200 on an Agilent TapeStation electropherogram is the percentage of RNA fragments longer than 200 nucleotides and is calculated directly from the area under the electropherogram curve.

### 2.2.4 FFPE RNA library preparation for Illumina sequencing

#### 2.2.4.1 Ribosomal RNA depletion

To enrich for target RNA species and improve downstream transcriptomic analyses, ribosomal RNA (rRNA) was depleted from total RNA using the

NEBNext rRNA Depletion Kit v2 (Human/Mouse/Rat) (New England Biolabs). This kit employs an RNase H–based depletion strategy, using DNA probes complementary to rRNA sequences, resulting in the selective degradation of cytoplasmic rRNAs (5S, 5.8S, 18S and 28S) and mitochondrial rRNAs (12S and 16S), as well as precursor rRNA transcripts derived from the primary rRNA transcription unit (Figure 2.7).

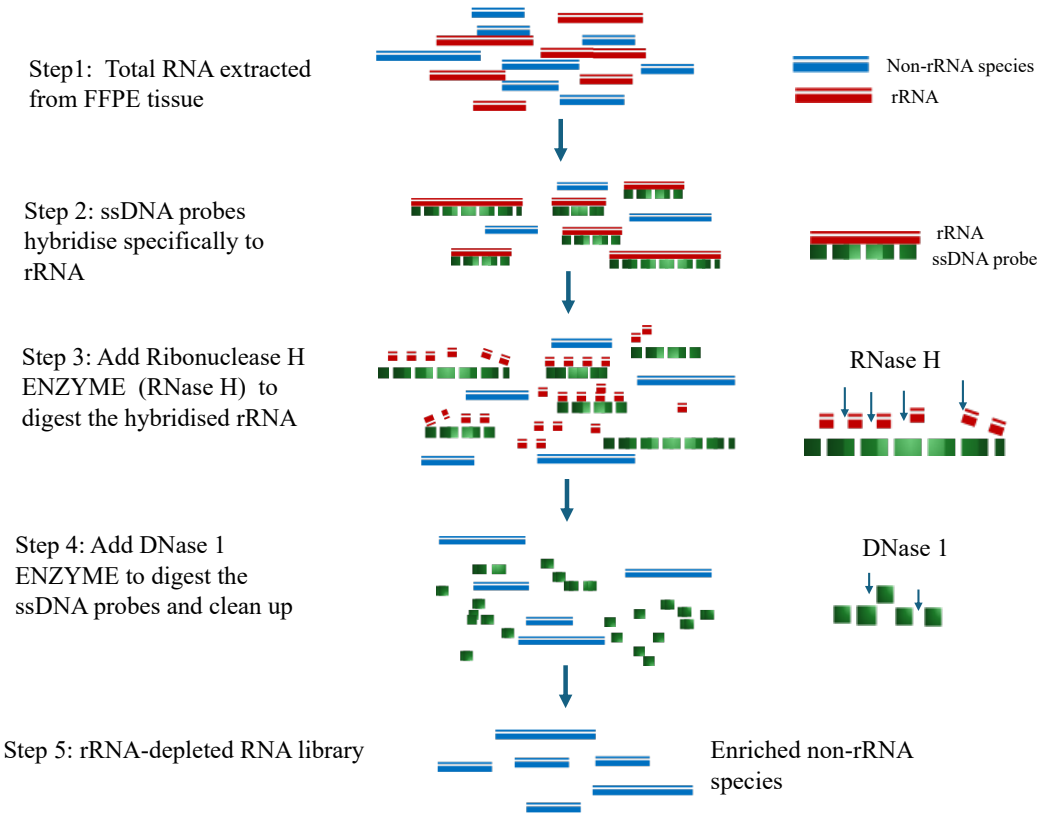



**Figure 2. 7 Ribosomal RNA depletion workflow adapted from (New England Biolabs)**

The diagram illustrates the process of removing ribosomal RNA (rRNA) from total RNA using the NEBNext rRNA Depletion Kit v2 (Human/Mouse/Rat) (New England Biolabs). In step 2, single-stranded DNA probes specifically hybridise to rRNA species. Ribonuclease enzyme is added in step 3 to digest the hybridised rRNA. Subsequently, DNase 1 is used to digest DNA probes, followed by RNA purification for subsequent applications. Diagram created in Microsoft PowerPoint.

Prior to depletion, RNA concentration and integrity were checked as outlined in section 2.2.3. RNA input ranged from 10 ng to 1,000 ng, depending on DV200 values, and samples with DV200 > 30% were generally considered suitable. RNA was diluted in nuclease-free water to a final volume of 11 µL.

For probe hybridisation, 2 µL rRNA depletion solution and 2 µL hybridisation buffer were added to each sample. Samples were heated to 95°C for 2 minutes, ramped to 22°C at 0.1 °C/s, and held at 22°C for 5 minutes. Following hybridisation, RNase H digestion was performed by adding 2 µL RNase H reaction buffer, 2 µL thermostable RNase H, and 1 µL nuclease-free water, then incubating at 50 °C for 30 minutes. Residual DNA and probes were removed by DNase I digestion using 5 µL DNase I reaction buffer and 2.5 µL DNase I, with incubation at 37 °C for 30 minutes. RNA was purified using Agencourt RNA Clean XP beads, washed twice with 80% ethanol, and eluted in 7 µL nuclease-free water, yielding 5 µL rRNA-depleted RNA for library preparation.

#### 2.2.4.2 RNA-Seq library preparation

RNA-Seq libraries were generated using the NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB), following the manufacturer's recommendations for degraded FFPE-derived RNA. rRNA-depleted RNA (5 µL) was combined with 1 µL random primers, denatured at 65 °C for 5 minutes, and cooled to 4 °C. First-strand cDNA synthesis was performed by adding 8 µL nuclease-free water, 4 µL first-strand synthesis reaction buffer, and 2 µL enzyme mix, followed by incubation at 25 °C for 10 minutes, 42 °C for 15 minutes, and 70 °C for 15 minutes. Second-

strand synthesis was performed by adding 8 µL second-strand synthesis buffer and 4 µL enzyme mix, adjusting the volume to 80 µL, and incubating at 16 °C for 1 hour. Double-stranded cDNA was purified using NEBNext sample purification beads, washed twice with 80% ethanol, and eluted in 53 µL 0.1× TE buffer. End repair and adaptor ligation were performed using the NEBNext Ultra II End Prep Module (NEB), followed by ligation with diluted NEBNext adaptors and USER enzyme treatment. Libraries were purified using NEBNext purification beads and eluted in 17 µL 0.1× TE buffer. PCR amplification was performed using NEBNext Ultra II Q5 Master Mix and index primers according to the manufacturer's recommended cycling conditions. Final libraries were purified using NEBNext sample purification beads and eluted in 23 µL 0.1× TE buffer, yielding 20 µL of final library for sequencing.

#### 2.2.4.3 RNA Sequencing

The yield and quality of RNA sequencing libraries were assessed as described in Section 2.2.3 prior to sequencing. Libraries were indexed using unique dual indices, multiplexed, and loaded onto the Illumina NovaSeq X Plus instrument according to the manufacturer's instructions. Sequencing was performed using a paired-end (2 × 150 bp) configuration. Image analysis and base calling were carried out using NovaSeq Control Software v1.3. Raw sequencing data were converted to FASTQ format and demultiplexed using Illumina bcl2fastq, allowing for one mismatch in the index sequence.

## 2.3 Burkitt lymphoma diagnosis and EBV antigen expression studies

### 2.3.1 Materials and equipment

**Table 2. 4 Monoclonal antibodies used for immunohistochemistry**

| EBV antigen | Clone | Host | Supplier | Cat No. |
|---|---|---|---|---|
| EBNA-1 | 1H4 | Rat | Merck Life Sciences | MABF2800 |
| LMP-1 | CS1-4 | Mouse | Abcam | ab78113 |
| LMP2A | 4E11 | Rat | Santa Cruz | sc-101316 |
| EBNA2 | PE2 | Mouse | Abcam | ab90543 |
| BZLF1 | BZ1 | Mouse | Santa Cruz | sc-53904 |

**Table 2. 5 Reagents, detection kits and consumables used for immunohistochemistry**

| Reagent | Supplier | Cat No. |
|---|---|---|
| BOND Dewax Solution | Leica Biosystems | AR9222 |
| BOND Epitope Retrieval Solution 1 (ER1) | Leica Biosystems | AR9961 |
| BOND Epitope Retrieval Solution 2 (ER2) | Leica Biosystems | AR9640 |
| BOND Wash Solution 10× Concentrate | Leica Biosystems | AR9590 |
| Primary Antibody Diluent | Leica Biosystems | AR9352 |
| Normal Goat Serum, 2.5% (Blocking Serum) | Vector Laboratories | S-1000 |
| ImmPRESS Peroxidase Polymer Anti-Rat IgG (Mouse Adsorbed), 15 ml | Vector Laboratories | MP-7444 |
| BOND Polymer Refine Detection Kit | Leica Biosystems | DS9800 |
| Ethanol series (70–100%) | Fisher Chemical | 10437341 |
| Xylene | | |
| Aquatex | Sigma-Aldrich (Merck) | 1.08562 |
| Leica Bond RXm | Leica Biosystems | 49.1501 |

### 2.3.2 Preparation of bond wash solution

Prepared by diluting 1 L of 10x Bond Wash concentrate with 9 L of distilled or deionised water to produce a final working solution. The buffer includes Tris-buffered saline, surfactant, and 3.5% ProClin 950. The final solution has a pH of 7.5–7.7.

### 2.3.3 Histopathologic assessment of tumour sections at the recruitment sites

Lymphoma diagnoses were confirmed through histopathological analysis of open or Tru-Cut biopsy specimens collected at pathology laboratories across AI-REAL recruitment centres in East Africa. To improve histopathology assessment at these locations, the AI-REAL consortium implemented a quarterly quality-assurance programme across three histopathology centres. Initially, competence-based training sessions were conducted for clinical pathologists and histotechnicians at the recruitment sites, providing mentorship and technical support by international expert haematopathologists in techniques such as tissue fixation, embedding, sectioning, and immunohistochemistry.

To maintain high-quality diagnostics, the consortium acquired automated immunohistochemistry (IHC) staining platforms (Ventana Benchmark GX) for all diagnostic centres and ensured a continuous supply chain throughout the study. Tissue samples were fixed in 10% neutral-buffered formalin and embedded in paraffin according to standard protocols. All samples were stained with Haematoxylin and Eosin (H&E) for morphology assessment and complemented by

immunohistochemistry. The diagnostic workflow employed a three-phase scoring system developed by Naresh et al. (2011). The first phase involved typical Burkitt lymphoma (BL) morphology combined with immunomarkers BCL2, CD10, and CD20. Inconclusive cases were referred to a second phase with additional markers (CD38, CD44, Ki-67), while unresolved cases proceeded to a third phase involving FISH analysis for MYC-Ig translocation.

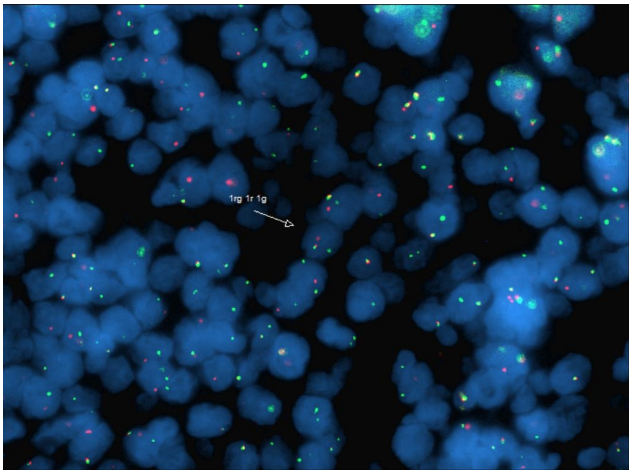
Beyond regular tumour board meetings, the AI-REAL consortium established telepathology networks, allowing study pathologists to seek second opinions and receive expert haematopathology reviews from overseas. This approach ensured accurate and consistent pathology diagnoses across all sites.

### 2.3.4 FISH for MYC -translocation

Fifteen Burkitt lymphoma tumour sections selected for RNA sequencing and EBV protein expression studies were subsequently examined for MYC translocation at the Oxford Regional Genetics Laboratories in the UK (For detailed pathology diagnosis and FISH for MYC, see Supplementary Table 6, Data Appendix). Interphase FISH was performed on formalin-fixed, paraffin-embedded (FFPE) tissue sections using the ZytoLight SPEC MYC Dual Colour Break Apart Probe (ZytoVision GmbH), following the manufacturer's protocol. This dual-colour break-apart probe comprises a ZyGreen-labelled DNA fragment hybridising distal to the MYC breakpoint region and a ZyOrange-labelled fragment hybridising proximal to the breakpoint on chromosome 8q24.21, allowing detection of MYC rearrangements regardless of the fusion partner. Briefly, 4–5 µm FFPE sections

were baked, deparaffinised, and pretreated (with heat and/or enzymatic digestion) to expose target DNA. Slides were co-denatured with the probe and target DNA, then hybridised under controlled temperature and humidity. After hybridisation, the unbound probe was removed through stringent washes, and nuclei were counterstained with DAPI. Hybridisation signals were visualised using a fluorescence microscope with appropriate filters. Normal cells typically showed two fused (green/orange) signals per nucleus, whereas a separated green and orange signal pattern was indicative of a MYC rearrangement (Figure 2.8).

**BL9**




**B14**

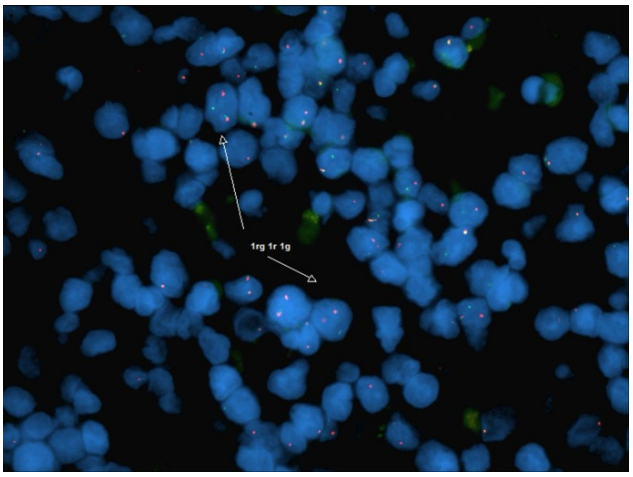



**Figure 2. 8 Fluorescence in situ hybridisation (FISH) analysis of the *MYC* (8q24) locus in interphase nuclei**

Using the ZytoLight SPEC *MYC* Dual Colour Break Apart Probe, a normal cell shows two orange/green fusion (yellow) signals, indicating that both alleles are intact. In the abnormal cells (BL 9 & BL14) shown above, the signal pattern consists of one green and one separate orange, confirming a chromosomal break at the 8q24 locus.

### 2.3.5 Immunohistochemistry for EBV antigens

Reagents and equipment used for immunohistochemistry are listed in Tables 2.4 and 2.5. IHC experiments were carried out at the Oxford Translational Histopathology Laboratories (THL) using the Leica BOND™ automated staining system and the standard protocol F30 (Leica Biosystems Inc., Heidelberg, Germany). FFPE tissue blocks were sectioned at approximately 3 μm with a microtome, and the sections were mounted onto positively charged amino-propyl-tri-ethoxy-silane-coated glass slides. Slides were labelled with histopathology identification numbers and the corresponding antibody markers and then baked overnight at 50 °C before staining.

Commercially available primary antibodies were used, including EBNA-1 rat monoclonal antibody (clone IH4; Merck Life Sciences), LMP-1 mouse monoclonal antibody (clone CS1-4; Abcam), LMP2A rat monoclonal antibody (clone 4E11; Santa Cruz), anti-EBNA2 mouse monoclonal antibody (clone PE2; Abcam), and anti-BZLF1 mouse monoclonal antibody (clone BZ1; Santa Cruz). Optimised antibody dilutions, duration of antigen-epitope retrieval, and pH conditions are shown in Table 2.6.

Tissue sections were pre-treated using heat-mediated antigen retrieval for 20 minutes with either citrate-based buffer (pH 6.0; Epitope Retrieval Solution 1) or EDTA-based buffer (pH 9.0; Epitope Retrieval Solution 2), as appropriate for each antibody. Sections were then incubated with primary antibodies for 30 minutes at room temperature. Antibody binding was visualised using an HRP-conjugated polymer detection system with 3,3′-diaminobenzidine (DAB) as the chromogen.

Sections were counterstained with haematoxylin and mounted using Aquatex mounting medium. Positive and negative controls were included in each staining run to ensure assay specificity and reproducibility. All staining procedures were performed according to standardised protocols using the Leica BOND™ system.

**Table 2. 6 Monoclonal antibodies used for IHC and their optimised parameters**

| EBV antigen | Dilution | Epitope retrieval duration | Primary antibody incubation |
|---|---|---|---|
| EBNA-1 | 1:100 | 20 min, | 20 min |
| LMP-1 | 1:1000 | 20 min | 30 min |
| LMP2A | 1:25 | 20 min | 30 min |
| EBNA2 | 1:100 | 40 min | 30 min |
| BZLF1 | 1:25 | 20 min | 30 min |

***** Antigen retrieval by heat at pH 9 for all antibodies

### 2.3.6 Microscopy and reporting

Each immunohistochemical marker was assessed for staining quality, expected subcellular localisation, staining pattern, intensity, and the proportion of positive tumour cells. Results were classified as positive only when the appropriate staining pattern and a sufficient proportion of tumour cells demonstrated clear, specific reactivity, in accordance with published College of American Pathologists (CAP) guidelines for immunohistochemical assay validation and reporting.

## 2.4 EBV vaccine immunogenicity studies

### 2.4.1 Materials

**Table 2. 7 Reagents used for Ex vivo PBMC and splenocyte stimulation assays**

| Item | Supplier | Cat No. |
|---|---|---|
| Isoflurane | Zoetis, IsoFlo | 50019100 |
| Ammonium-Chloride-Potassium (ACK) lysis buffer | Gibco, Thermo Fisher Scientific | A10492-01 |
| 0.5M EDTA sterile stock solution | VWR Chemicals | E522-100ML |
| Brefeldin A | Abcam | ab193369 |
| Cytofix/Cytoperm solution | BD Biosciences | 554722 |
| DNase I | Roche | 10-104-159-001 |
| Perm Wash buffer | BD Biosciences | 554723 |
| Dimethyl Sulfoxide (DMSO) | Sigma Aldrich | D2650-100ML |
| Dulbecco's Phosphate Buffered Saline (DPBS), 1x | Gibco, Thermo Fisher Scientific | 14190-094 |
| Dulbecco's Modified Eagle Medium (DMEM) | Gibco, Thermo Fisher Scientific | 31966-021 |
| RPMI 1640 medium | Gibco, Thermo Fisher Scientific | 21875-034 |
| LIVE/DEAD™ Fixable Aqua Dead Cell Stain Kit | Thermo Fisher Scientific | L34965 |
| anti-mouse CD28 (α-CD28) | Tonbo Biosciences | 70-0281-U100 |
| anti-mouse CD16/CD32 (Fc block) | BD Biosciences | 553142 |

**Table 2. 8 List of flow cytometry antibodies and cytokines**

| Antibody/Cytokine | Supplier | Clone | Dilution | Fluorochrome | Laser |
|---|---|---|---|---|---|
| CD8 | BioLegend | 53-6.7 | 1:200 | BUV395 | B530 |
| CD4 | BioLegend | GK1.5 | 1:200 | BV421 | V450 |
| IFNγ | BioLegend | XMG1.2 | 1:200 | APC | R670 |
| CD3 | BioLegend | 17A2 | 1:200 | BV650 | V670 |
| TNFα | BioLegend | MP6-XT22 | 1:200 | FITC | B530 |
| IL2 | BioLegend | JES6-5H4 | 1:200 | PE | YG586 |

**Table 2. 9 Equipment and consumables used for mouse experiments**

| Item | Supplier |
|---|---|
| BD Fortessa | BD Biosciences |
| TC20 Automated Cell Counter | Bio-Rad |
| Multi-step pipette | Eppendorf |
| Multi-channel pipettes | Eppendorf |
| Pipettes (P2, P10, P20, P50, P100, P200, P1000) | Eppendorf |
| Sterile filter tips | Thermo Fisher Scientific |
| Centrifuge | Eppendorf |
| BD PlastiPak Luer Slip Syringe | BD Biosciences |
| Cell strainers, 70µm | Falcon, Corning |
| Cell strainers, 100µm | Falcon, Corning |
| Centrifuge tubes, 50ml | Falcon, Corning |
| FACS tubes, Polystyrene, 5ml | Sarstedt |
| Insulin syringes (29G x 12.7mm), 0.5ml | BD Biosciences |
| Scalpel blades, No. 11 | Swann-Morton |
| Syringe needles (23G x 16mm) | Agani, Terumo |
| Microcentrifuge tubes, 1.5ml | Eppendorf |

### 2.4.2 Methods

Commercially available reagents (flow cytometry antibodies for surface and intracellular cell staining), buffers, and consumables used for ex vivo mouse experiments are all listed in Tables 2.7, 2.8, and 2.9, respectively.

#### 2.4.2.1 Preparation of buffers, peptides and vaccine

**a) Preparation of 12.5 mM EDTA for tail bleed**

The evening before bleeding mice, 12.5 mM EDTA solution was prepared by diluting a 0.5 M sterile EDTA stock 1:40 in sterile phosphate-buffered saline (PBS) under a tissue culture hood. The final solution was aliquoted into 200 µL volumes in 1.5 mL Eppendorf tubes, ready for use.

**b) ACK lysis buffer preparation**

To prepare 10x ACK lysis stock solution, 89.9g $NH_4Cl$, 10.0g $KHCO_3$, and 370mg Tetrasodium EDTA were dissolved in 1L of double-distilled water ($ddH_2O$), then the pH was adjusted to 7.3. The 10x stock was stored for up to 6 months at room temperature (15-25 °C) or for longer periods at -20 °C. A 1x working solution of ACK lysis buffer was prepared before use by diluting 1 part of the 10x stock in 9 parts $ddH_2O$.

**c) Peptides**

Peptide libraries corresponding to selected EBV proteins were synthesised by Mimotopes (UK). For each target, peptide libraries were designed as 15-mers with

a 10-amino acid overlap, providing a continuous scan of the selected immunodominant regions or entire protein sequence. Individual peptides were initially dissolved in DMSO at 50 mg/mL and subsequently combined into antigen-specific peptide pools for in vitro T-cell stimulation assays at a final concentration of 4 μg/mL. Five peptide pools were generated: EBNA1, LMP2A, BARF1, BZLF1, and a composite pool comprising 26 epitopes (26E) derived from additional EBV proteins (EBNA2, EBNA3A, EBNA3B, EBNA3C, EBNA-LP, LMP1, BMLF1, and gp350).

**d) Vaccines**

Eight latent and four lytic EBV genes were selected for inclusion in the vaccine constructs to provide broad immunological coverage across the distinct phases of the EBV life cycle. The construct design incorporated a combination of full-length antigens and epitope-string-based antigens, depending on the biological and functional characteristics of each protein and the availability of experimentally validated immunogenic regions.

Full-length coding sequences were used for LMP2A, BARF1 and BZLF1. These antigens were selected in full to preserve native protein architecture and maximise the breadth of potential $CD8^+$ and $CD4^+$ T-cell epitopes presented following intracellular antigen processing. In contrast, for the remaining latent antigens (EBNA1, EBNA2, EBNA3A, EBNA3B, EBNA3C, EBNA-LP, and LMP1) and lytic antigens (BMLF1, BLLF1/gp350), the vaccine design utilised concatenated epitope strings comprising selected human $CD8^+$ and $CD4^+$ T-cell epitopes rather than full-length proteins.

Transgene inserts encoding either selected full-length antigens or concatenated epitope strings were synthesised and cloned into shuttle plasmid vectors prior to recombination into Chimpanzee Adenovirus Oxford 2 (ChAdOx2) and Modified Vaccinia Ankara (MVA) viral vector backbones at the Viral Vector Core Facility (VVCF), Jenner Institute, University of Oxford, using established facility procedures. ChAdOx2 is a replication-deficient chimpanzee adenoviral vector (E1/E3 deleted) that is widely used as a vaccine platform to induce potent cellular immune responses via transgene expression driven by a mammalian promoter. MVA is a highly attenuated poxvirus vector commonly used in heterologous prime–boost vaccination strategies to enhance T-cell immunity following adenoviral priming (McAuliffe et al., 2021; Saunders et al., 2023).

The transgene inserts were designed to encode proteins expressed during both the latent and lytic phases of the EBV life cycle, with the aim of eliciting broad $CD4^+$ and $CD8^+$ T-cell responses. The specific latent and lytic sequences were cloned into appropriate shuttle plasmids by Dr Carol Leung and Dr Emily Steffke prior to recombination into the ChAdOx2 and MVA vector backbones. For the ChAdOx2 constructs, three distinct antigen configurations were generated (Figure 2.9): a latent-only insert, a lytic-only insert, and a combined latent–lytic insert, each of which was recombined into the ChAdOx2 destination vector for adenoviral production. A corresponding EBV latent–lytic insert construct was also recombined into an MVA backbone for use in heterologous prime–boost regimens. Shuttle plasmids and synthetic templates were obtained commercially (GeneArt, Thermo Fisher Scientific, Paisley, UK), and viral vectors were produced according to standard cloning, recombination, amplification, and quality-control procedures.

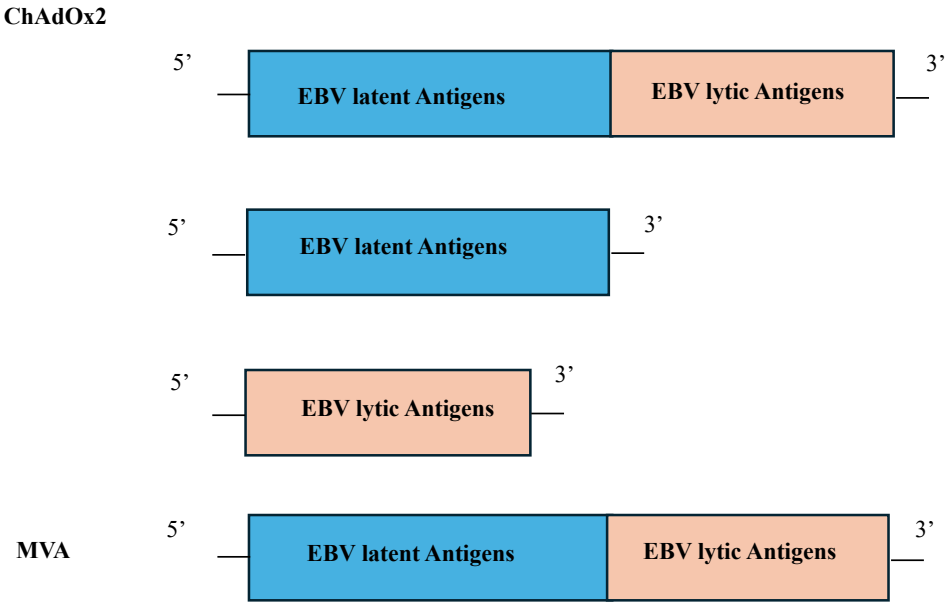



**Figure 2. 9 Vaccine inserts into viral vector backbone**
The design of the ChAdOx2 and MVA viral vectors encoding selected EBV latent and lytic antigens. Three ChAdOx2-EBV antigen configurations were generated: ChAdOx2-EBV-latent only, ChAdOx2-EBV-lytic, and ChAdOx2-EBV-latent-lytic antigens. An EBV-latent-lytic antigen insert construct was recombined into an MVA backbone for use in heterologous prime-boost regimens.

### 2.4.2.2 Mice

Female BALB/cJ mice, aged 8–9 weeks, were used in these experiments. Mice were purchased from Envigo (UK) and housed under standard specific-pathogen-free conditions in the Functional Genomics Facility (FGF) at the University of Oxford. Experimental procedures complied with the UK Animals (Scientific Procedures) Act 1986 and Project License PP7109477.

#### 2.4.2.3 Sample size and power calculations

Experimental group sizes (n = 6 mice per group) follow institutional precedent and prior in-house and published studies of the same assay (McAuliffe, 2020). For transparency, given $\alpha = 0.05$ and two-sided tests, a design with n = 6 per group has ~80% power to detect very large, standardised differences (Cohen's $d \approx 1.8$ for unpaired comparisons and $d \approx 1.43$ for paired comparisons). Because smaller standardised effects have substantially lower power with this sample size, secondary outcomes and small-to-moderate effects are considered exploratory. Nonparametric tests were used (Kruskal–Wallis for >2 groups, and Wilcoxon matched-pairs or Mann–Whitney for pairwise comparisons); the detectable-effect estimates above are parametric approximations provided to aid interpretation and translate to very large nonparametric effect sizes (e.g., Cliff's $\Delta \gtrsim 0.6$–$0.8$).

#### 2.4.2.4 Vaccinations

To assess vaccine immunogenicity, prime-and-boost vaccinations were performed using the ChAdOx2 and MVA viral vectors encoding the selected EBV antigens. A sham vaccination (sterile PBS injection) was assigned to the control groups to compare vaccine-specific T cell responses with those of the treatment groups. Vaccinations were performed under inhalation anaesthesia, using isoflurane. A dose of $10^8$ infectious units (IU) of ChAdOx2 or $10^7$ plaque-forming units (PFU) of MVA virus was given via intramuscular injection (i.m.) into the thigh in a total volume of 50 μL per mouse. All vaccines were prepared using sterile DPBS. Prime and boost vaccinations were administered 4 weeks apart, and peripheral blood

samples were collected between vaccinations to assess antigen-specific T-cell responses.

#### 2.4.2.5 Sample collection and endpoints

**a) Blood sampling**

Mice were bled 14 days after vaccination to assess vaccine-induced T cell responses. Mice were first warmed in a heat box for approximately 15 minutes at 38 °c to promote vasodilation and enhance blood flow. A small incision was made in the lateral tail vein using a sterile scalpel blade. Ten drops (~100 µL) of blood were collected into 1.5 ml Eppendorf tubes containing 200 µL of freshly prepared 12.5 mM EDTA in PBS to prevent coagulation (refer to Section 2.4.2.1for EDTA preparation). No more than 10% of the mouse's total blood volume was collected at any one time, and within a month, no more than 15% of the total blood volume was collected, as specified in the project licence (PPL).

Blood samples were briefly centrifuged at 500 × g for 10 seconds to pellet the cells, then transferred into pre-labelled FACS tubes. Red blood cell (RBC) lysis was carried out by adding 500 µL of 1× ACK lysis buffer (prepared as described in Section 2.4.2.1) and incubating at room temperature (25 °C) for 5 minutes. A further 500 µL of 1× ACK lysis buffer was added, followed by another 5-minute incubation. During lysis, tubes were intermittently vortexed to facilitate RBC disruption. Lysis was halted by adding 3 mL of ice-cold PBS, and the samples were centrifuged at 500 × g for 5 minutes. The supernatant was discarded, and the pellet

was briefly vortexed. This lysis step was repeated with 1 mL of ACK buffer to ensure complete removal of RBCs. Cells were then washed with 3 mL of ice-cold PBS and resuspended in complete RPMI medium for subsequent assays.

**b) Spleens**

At the termination of experiments, mice were euthanised by cervical dislocation, and spleens were removed via surgical incision and transferred into 1.5 mL Eppendorf tubes containing 1ml of 1X PBS kept on ice. To prepare single-cell suspensions, spleens were mashed between two glass slides, then passed through a 70µm cell strainer into a 50ml Falcon tube and washed with 15 ml of ice-cold 1x PBS. Cells were then pelleted by spinning the tubes at 500 g for 5 minutes and then resuspended in 1ml of ACK buffer to lyse red cells. The lysis reaction was stopped by adding 29 ml of ice-cold PBS, after which each cell sample was passed through a 70 µm cell strainer again. The samples were then centrifuged at 500g for 5 minutes to pellet the cells, after which they were resuspended in complete RPMI medium for downstream analysis.

**c) Ex vivo peptide stimulations**

Peptides were thawed on ice and working stimulation mixtures were prepared by diluting peptide stocks 1:2 to achieve a final concentration of 4 µg/mL. Peptide pools included EBNA1, LMP2A, BARF1, BZLF1, and a pool of 26 epitopes derived from additional EBV antigens (EBNA2, EBNA3A, EBNA3B/3C, EBNA-LP, LMP1, BMLF1 and gp350).

PBCs or splenocytes resuspended in complete RMPI 1640 medium were transferred into 96-well V-bottom plates. Cells were stimulated by adding 50 µL of an αCD28 (2 µg/ml) + DNase1 (20 µg/ml) mix, followed by 100 µL of peptide stimulation mix or DMSO vehicle control. Plates were incubated at 37 °C in a tissue culture incubator (5% $CO_2$) for 5 hours. After 1 hour of incubation, brefeldin A (2 µL of a 100× stock diluted in R10 medium) was added to all wells to inhibit Golgi transport and promote intracellular cytokine accumulation. Following the stimulation period, plates were centrifuged at 500 × g for 5 min, the supernatants were discarded, and cells were washed once with 200 µL of ice-cold PBS.

**d) Intracellular cytokine staining**

Following ex vivo stimulation, cells were stained with fluorescently conjugated antibodies against surface and intracellular markers. Samples were kept on ice throughout this stage, and centrifugations were performed at 4 ℃.

Cells were washed with 200 µL of ice-cold PBS, and then Fc receptors were blocked via incubation in 50 µL anti-mouse CD16/CD32 antibody (5µg/mL) diluted in 1 x PBS for 10 min at 4 °C, then washed with 150 µL ice-cold PBS. Surface staining was performed using fluorochrome-conjugated antibodies (anti-CD4, anti-CD8) and Live/Dead Aqua viability dye diluted in PBS. Cells were incubated with 50 µL of surface antibody cocktail for 20 min at 4 °C in the dark, then washed twice with 200 µl of ice-cold PBS.

For fixation and permeabilisation, cells were resuspended in 100 µL of Cytofix/Cytoperm buffer (BD Biosciences) and incubated for 20 min at 4 °C in the dark. Cells were washed twice with 1× BD Perm/Wash buffer (prepared by diluting 10× stock 1:10 in distilled water), centrifuging at 500 × g for 5 min between washes. Intracellular staining was performed using fluorochrome-conjugated antibodies (TNF-α, IFN-γ, IL-2) diluted in 1× BD Perm/Wash buffer. Cells were incubated with 50 µL of intracellular antibody cocktail for 20 min at 4 °C in the dark, followed by two washes with BD Perm/Wash buffer. Finally, cells were resuspended in 120 µL of BD Perm/Wash buffer and acquired by flow cytometry.

**e) Flow acquisition, T cell phenotyping and cytokine measurements**

Flow cytometry data were acquired on a BD Fortessa flow cytometer after standard startup and quality-control procedures. The instrument was powered on at least 30 minutes prior to acquisition to allow laser stabilisation, and cytometer performance was verified using Cytometer Setup and Tracking (CST) beads according to the manufacturer's instructions. Compensation was calculated using antibody-capture beads for all fluorochromes in the panel, including the Live/Dead Aqua viability dye, with single-stained and unstained controls. Photomultiplier tube voltages were optimised to place positive signals within the linear detection range, and compensation matrices were calculated and applied in FACSDiva software prior to sample acquisition.

For T cell phenotyping and intracellular cytokine analysis, samples were acquired in tube or plate mode using the high-throughput sampler (HTS). Events were

recorded at a high flow rate following voltage optimisation, and the compensation values were confirmed to be within acceptable limits. Data were collected using standard laser and filter configurations, and, where applicable, the maximum allowable number of events per well was obtained. Flow cytometry data were subsequently analysed using FlowJo software (Tree Star, Ashland, Oregon) to identify $CD4^+$ and $CD8^+$ T cell populations and to quantify intracellular cytokine expression (including IFN-γ, TNF-α, and IL-2) following peptide stimulation.

Lymphocytes were gated on a forward scatter area (FSC-A) versus side scatter area (SSC-A) pseudo colour dot plot. Single cells were gated using a forward scatter area (FSC-A) versus forward scatter height (FSC-H) plot. Dead cells were excluded through Live/Dead Aqua staining. T cells were gated on the CD4 versus CD8 plot. Antigen-specific T cells responding to antigen stimulation were identified by gating on IFN-γ+, TNF-α+, and IL-2+ events within the $CD4^+$ populations. Polyfunctional T-cell responses were analysed using Boolean gating in FlowJo (Tree Star, Ashland, Oregon), and background-subtracted T-cell responses to different EBV peptides were analysed using SPICE (v6, NIH) to assess their functional composition.

# Chapter 3: Epstein-Barr Virus Genome Diversity and Evolution

## 3.1 Introduction

In this chapter, I explored the role of specific EBV genomic features, their impact on viral genome evolution, and potential contribution to the risk of Burkitt lymphoma. EBV is a highly prevalent human herpesvirus that establishes lifelong infection, with adult infection rates approaching 95% (de-Thé et al., 1975; Kuri et al., 2020; Smatti et al., 2018; Winter et al., 2019; Wong et al., 2022). Despite this widespread distribution, the burden of EBV-associated malignancies exhibits notable geographical variation (Kanda et al., 2019). Endemic BL, for instance, primarily affects children in sub-Saharan Africa, where its incidence is significantly higher than in other regions of the world (Mbulaiteye & Devesa, 2022).

Malaria has long been proposed as a major cofactor in the development of BL (Burkitt, 1969; Torgbor et al., 2014); however, the strength of this association remains weak. Much of the evidence linking malaria to BL derives from ecological observations (Burkitt, 1962), raising the possibility of an ecological fallacy, as individual-level data supporting a direct causal relationship remain limited (Peprah et al., 2020; Redmond et al., 2020). Outside endemic areas, BL is relatively rare and generally associated with immunodeficiency, particularly in the context of HIV infection (Ziegler et al., 1982). In contrast, nasopharyngeal carcinoma, extranodal NK/T-cell lymphomas, and a subset of gastric carcinomas, predominantly adult-onset cancers, are most prevalent in East and Southeast Asia and parts of North

Africa (Wong et al., 2022). These malignancies highlight the complex interplay of viral, host, and environmental determinants of EBV-driven oncogenesis.

The oncogenic potential of EBV has been well established in both experimental and clinical contexts (Sadreddini et al., 2016; Tsao et al., 2012). The virus can transform normal B lymphocytes and epithelial cells into immortalised clones (Danve et al., 2001; Kulwichit et al., 1998; Nilsson et al., 1971), underscoring its inherent capacity to dysregulate cellular growth pathways. However, despite this oncogenic potential, only a small proportion of individuals infected with EBV develop EBV-associated malignancies. This striking discrepancy between the near-universal prevalence of infection and the restricted geographical and clinical occurrence of EBV-driven cancers remains a central paradox in EBV biology.

A growing body of evidence suggests that genetic diversity within the EBV genome influences viral behaviour, immune evasion, and oncogenic potential across different populations. For example, sequence variants in immunogenic and regulatory regions of the virus have been linked to increased oncogenicity and poorer clinical outcomes among NPC patients in Southeast Asia (Xu et al., 2019; Xue et al., 2021). Sequence divergence in the viral *EBNA-2* gene not only defines viral subtypes but also modulates viral behaviour by enhancing transformation efficiency (Rickinson et al., 1987; Sample et al., 1990). Additionally, early genetic studies linked EBNA2 deletions to enhanced c-MYC expression, the principal oncoprotein driving BL (Kelly et al., 2002; Kelly et al., 2009). Collectively, these observations suggest that specific high-risk viral strains may contribute to disease

susceptibility, yet the molecular basis of such associations remains poorly understood.

To elucidate how EBV diversity arises and is maintained across populations, it is important to consider the fundamental mechanisms that generate genomic variation. Two primary processes drive EBV evolution: mutations and genetic recombination (or genetic reshuffling). In most tumours, EBV genomes exhibit clonal patterns as the virus replicates synchronously with the host genome during latency (Bahri et al., 2022; Pathmanathan et al., 1995). Clonal evolution involves the vertical accumulation of mutations, resulting in phylogenetic clustering of tumour-associated EBV genomes with minimal intra-tumoral diversity. In contrast, recombination entails the exchange of homologous or non-homologous genetic material between co-infecting viral strains (Pérez-Losada et al., 2015). Hypothetically, EBV's largely latent replication strategy should constrain recombination, leading to predominantly clonal phylogenetic structures. However, recent phylogenetic studies have revealed substantial mosaicism among circulating EBV strains, with some clustering more strongly by region than by disease phenotype (Briercheck et al., 2024; Telford et al., 2020). This suggests that both mutation and recombination may act in concert to shape EBV evolution and adaptation in disease-specific contexts.

The enzymatic drivers of EBV mutations and recombination are not fully understood. B cells frequently express activation-induced cytidine deaminase (AID) and apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC)

enzymes, both host cytidine deaminases that can induce C→U or G→A transitions (Godsmark et al., 2021). AID is essential for antibody diversification in germinal centre B cells (Stavnezer et al., 2008), while APOBEC enzymes act as innate antiviral factors that restrict viral replication through hypermutation (Shilova et al., 2022). In cancer, APOBEC activity can also produce characteristic mutational signatures (COSMIC SBS2 and SBS13) in host DNA(Alexandrov et al., 2020). EBV infection has been shown to modulate AID and APOBEC expression (Chen et al., 2016; Kikuchi et al., 2015; Summerauer et al., 2022; Wakae et al., 2020), potentially promoting mutagenesis in both viral and host genomes during chronic infection. Thus, analysis of EBV mutational signatures in viral and human tumour DNA provides a valuable framework for inferring the mutagenic processes underlying viral evolution and tumour adaptation, with significant implications for therapy and diagnosis.

Using comparative genomics, phylogenetics, and mutational signature analysis, I sought to identify genomic features distinguishing EBV lineages associated with BL from those circulating in other populations. The analyses incorporated 64 newly assembled genomes derived from circulating tumour DNA of BL cases and from plasma of EBV-positive healthy children in East Africa. These were combined with 346 publicly available EBV genomes isolated from primary tumours, or plasma, collected across Africa, Asia, South America, Central America, and North America.

## 3.2 Chapter aims and specific objectives

The specific aims and objectives of this chapter were:

i) To describe the differences in the viral genomes across geographic regions and disease phenotypes

ii) To assess the frequency of EBV genome recombination in Burkitt lymphoma compared to other EBV-associated malignancies

iii) To identify viral variants enriched in endemic Burkitt lymphoma

iv) Using COSMIC single-base substitution models, characterise the viral mutational patterns and

v) Map point mutations, insertions and deletions across EBV genomes from primary clinical samples (BL cases) for subsequent protein translation and preclinical immunogenicity studies.

## 3.3 Methods

### 3.3.1 Study setting, design and ethical approval

This chapter analysed data collected from a prospective case-control study in East Africa (AI-REAL). The study design and detailed protocol have been previously published (Legason et al., 2022). The analysis employed a nested case-control molecular epidemiological design, combining whole-genome sequence data from the primary study with publicly available genomes for comparative phylogenomic analyses.

The AI-REAL study setting and participant recruitment are described in Section 2.1.2 of the General Methods. Appropriate ethical approvals were obtained in accordance with the research guidelines and the local IRB regulations of the participating recruitment centres in East Africa. All study protocols adhered to the Declaration of Helsinki and international data protection regulations.

### 3.3.2 Patients and samples

Fifty-three cfDNA samples from histologically confirmed Burkitt lymphoma and other lymphoma cases were initially analysed for EBV presence using a customised EBV and lymphoma panel (see Supplementary Fig.1 for study design, participant selection, and data analysis). The liquid biopsy diagnostic panel design targeted EBV genes EBER-1, EBER-2, and/or EBNA-2 for genotyping the two known EBV strains (Legason et al., 2022). In healthy controls, EBV in plasma samples was detected using a quantitative PCR kit from 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. The primary sample set consisted of 64 cfDNA samples from cases plus 11 healthy carriers, which were processed as described in section 2.1.3.

### 3.3.3 Samples and nucleic acid extraction

cfDNA extraction and quality assessment are described in section 2.1.4.

### 3.3.4 EBV probe design and whole genome sequencing

The EBV capture probes were designed with a 60-bp overlap and a 120-bp length, comprising 2,862 probes that covered the entire genome. The custom panel was synthesised by Integrated DNA Technologies (IDT). Sequencing libraries were prepared using the Thruplex Tag-Seq kit (Takara Bio) and normalised prior to EBV capture and enrichment, which employed the IDT xGen Hybridisation protocol (IDT); details are in Sections 2.1.5. Sequencing libraries were multiplexed into up to 6 pools to achieve sufficient coverage and sequencing depth, as detailed in Section 2.1.6.

### 3.3.5 EBV whole genome data processing

The primary EBV whole-genome dataset was combined with publicly available data to create a total of 410 complete genomes for comparative genomic analyses.

All EBV genomes, including publicly accessible ones, were derived from primary tumour material or plasma cfDNA.

FASTQ files for public genomes were downloaded from the Sequence Read Archive (SRA) using BioProject IDs PRJNA552587, PRJNA522388, and PRJNA1063319, representing five geographic regions: Africa, Asia, South America, Central America and North America (see Supplementary Fig. 2 for the geographic origin and composition of the public EBV genomes). All public and primary sequence files were processed using an in-house bioinformatics pipeline (Figure 3.1).

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 and EBNA3s contigs from NC_009334.1) using the BWA-MEM alignment method (Vasimuddin et al., 2019). Coverage statistics were computed from aligned binary alignment map (BAM) files using samtools coverage (Danecek et al., 2021).

Variant calling was performed using Genome Analysis Toolkit (GATK) v4.6.0.0 with the Mutect2 algorithm, followed by variant filtering using the FilterMutectCalls (Benjamin et al., 2019). Variants were called against the EBV reference genomes for Type 1 and Type 2, as well as a custom hybrid reference genome, as previously mentioned. Reads with a mapping quality below 20 and

base quality below 10 were discarded. Variants were only considered if they had a minimum callable depth of 10 reads (TLOD score ≥ 3). Base quality recalibration used a base quality score threshold of 18.

Post-calling filtering was carried out using an optimal F-score threshold with a false discovery rate of 5%. Variants were discarded if they had low allele fraction, strand bias, mapping quality bias, clustered events, contamination, read orientation artefacts, or polymerase slippage within short tandem repeats. Additional filters removed variants with a low median base quality (<20), abnormal fragment-length distributions, or inadequate read-position support.

To further improve variant reliability, somatic variants were cross-validated using two additional variant callers, VarDict (Lai et al., 2016), and VarScan (Koboldt et al., 2012). Variant call sets were intersected using bcftools (v1.12) (Danecek et al., 2021), retaining only variants detected by at least two callers. Functional annotation was performed using SnpEff (v5.2c), and repeat annotations were added using RepeatMasker tracks (Cingolani et al., 2012).

The final variant caller files (vcfs) included annotated, filtered, and combined variants. Summary QC statistics, including mean sequencing depth and coverage, were visualised in R v4.4.1, using ggplot2 (Team, 2016).

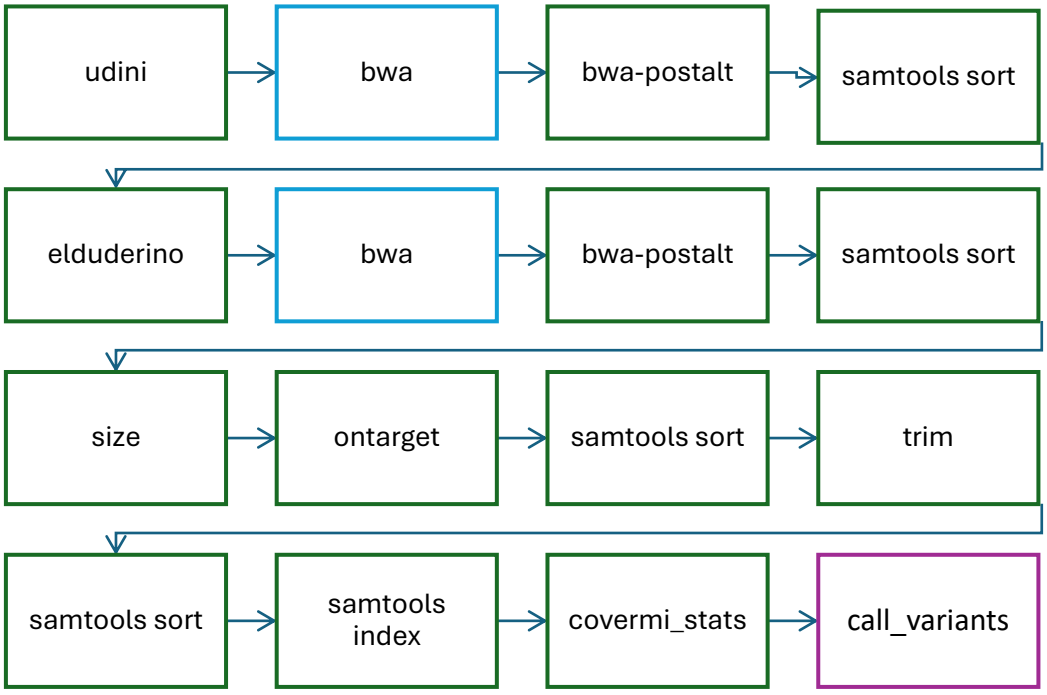



**Figure 3. 1 Schematic of the bioinformatic pipeline**

Schematic of the in-house bioinformatics analysis pipeline. The steps in green denote QC or pre-processing steps. Blue represents alignment steps. Purple indicates variant calling. The first tool, udini, extracts unique molecular indices (UMIs) from the sample FASTQ file and combines all reads for a sample into a single interleaved FASTQ file. BWA aligns the FASTQ from udini to the EBV reference genome. The next tool, elduderino, deduplicates the reads and collapses the read families into consensus reads. The deduplicated FASTQ file is re-aligned with BWA before being passed to "size," a locally developed Python script. Size outputs the fragment size distribution of cfDNA. The next tool, "ontarget," calculates off-target rates from the deduplicated Sequence Alignment/Map (SAM) file and BED files. The last set of tools, "trim" and "covermi_stats," are used to trim the last base (often due to poor quality) and generate coverage statistics, including mean depth and percentage of targets covered at 30x, 100x, 500x, 1000x, and 2000x. The last step involved variant calling recalibration and annotation following the GATK workflow v4.6.0.

### 3.3.6 Coverage and sequencing depth across samples

Genome coverage and sequencing depth were determined from aligned BAM files using samtools version 1.17 (Li et al., 2009). For each sample, the samtools coverage function was applied to the sorted BAM file to calculate coverage statistics across the EBV reference genome. Metrics were extracted for EBV chromosome (chrEBV): reference name, start and end coordinates, number of mapped reads, number of covered bases, proportion of covered bases, mean sequencing depth, mean base quality, and mean mapping quality. Results were compiled into a comma-separated values (CSV) file for visualisation in R v4.4.1 using ggplot (Team, 2016).

### 3.3.7 EBV genome assembly

The EBV genome assembly was performed using SPAdes v4.2.0, an open-source de novo genome assembly toolkit (Prjibelski et al., 2020). Before assembly, the software corrects read errors by identifying and fixing misreads. The reads are then divided into short, overlapping sequences called k-mers. These k-mers were used to construct the de Bruijn graph; for short reads, k-mers of length 21, 33, and 55 were employed, with an average read length of 143 bp. The k-mers are subsequently assembled into contigs.

The quality of the contig assemblies was evaluated using QUAST v5.3.0, a genome-quality assessment tool (Gurevich et al., 2013). Key assembly metrics, including genome fraction, number of misassemblies, mismatches per 100 kb, GC

content, and duplication ratio, were extracted from QUAST reports. Genome fraction represents the proportion of the reference genome covered by the assembly, while misassemblies and mismatches per 100 kb reflect structural and base-level errors, respectively (Mikheenko et al., 2016).

To identify high-quality assemblies, empirical filtering criteria were applied to the QUAST metrics because there are no universal published thresholds for similar datasets. Assemblies had to meet the following criteria: genome fraction >70%, fewer than 1 misassembly, fewer than 1000 mismatches per 100 kb, GC content within ±2% of the reference genome, and a duplication ratio ≤1. A composite quality score was calculated from these metrics, and assemblies with an overall score ≥50% were retained for downstream analyses.

To order contigs against the relevant reference genomes, BLAST databases were created from EBNA-2 sequences of EBV Types 1 (NC_007605.1) and 2 (NC_009334.1). The contigs were aligned to each reference with blastn (v2.17.0), part of the NCBI BLAST+ suite (Camacho et al., 2009).

Samples exhibiting the highest similarity to a specific reference genome were assigned a genotype call. This categorisation was vital for the public genomes, which lacked sample metadata and EBV strain information. Subsequently, the assembled contigs were ordered and oriented according to Type 1 or Type 2 reference genomes, depending on the sample genotype, using ABACAS v1.3.1 (an Algorithm-Based Automatic Contiguation of Assembled Sequences) (Zhou et al.,

2025), resulting in a single pseudomolecule FASTA sequence. Samples displaying mixed genotypes were excluded from further analysis because there were no tools available to appropriately assemble their genomes. Contigs were trimmed to remove leading and trailing ambiguous bases and other sequencing artefacts using trimAl v1.4 rev 15 (Capella-Gutiérrez et al., 2009).

### 3.3.8 Phylogenetic analysis

The assembled genomes were aligned using the multiple sequence alignment tool MAFFT v7.520 (Katoh et al., 2002). A total of 410 genomes were aligned, including the EBV Type 1 and 2 reference genomes and an outgroup, Macacine herpesvirus 4 (NC_006146.1). The multiple sequence alignment file was trimmed using trimAl v1.4 rev 15 (Capella-Gutiérrez et al., 2009), and known EBV repeat regions were masked using RepeatMasker version 4.1.9 in default mode (Smit, 2015). The maximum likelihood of the phylogenetic tree was then inferred using the Randomised Accelerated Maximum Likelihood (RAxML-NG v.1.2.2) tool with the Generalised Time Reversible (GTR-G) model (Togkousidis et al., 2023). The Transfer Bootstrap Expectation (TBE) approach was used to assess branch support, and nodes were considered to have strong support if their TBE exceeded 70% (Lemoine et al., 2018). The phylogenetic tree was visualised and annotated using R packages: ape (v5.8.1) for tree import and manipulation (Paradis & Schliep, 2019), including re-rooting using the Macacine herpesvirus-4 as an outgroup, ggtree (v3.12.0) for tree visualisation (Yu et al., 2017), ggtreeExtra (v1.14.0) for adding metadata annotations (Xu et al., 2021), and tidyverse (v2.0) for data wrangling and factor reordering (Wickham et al., 2019).

### 3.3.9 Principal component analysis

Pre-processed and filtered variant caller files (VCFs) were merged using bcftools v1.12 (Danecek et al., 2021). Multiallelic sites introduced by VCF merging were split into biallelic sites, and indels were left-aligned using bcftools norm (Danecek et al., 2021). The normalised merged file was then converted into PLINK binary format with genotype calls. Variants were filtered by minor allele frequency (>5%) and linkage disequilibrium (LD). The purpose of LD pruning prior to principal component analysis (PCA) was to reduce redundancy among highly correlated variants while still retaining sufficient independent variants to capture the EBV population structure. In large genomes such as human datasets, stringent thresholds (e.g., $r^2 > 0.2$) are commonly used. However, for a compact viral genome such as EBV (~172kb), stringent pruning would remove a large proportion of variants and significantly reduce the number of informative variants available for PCA.

To address this, a less stringent LD pruning ($r^2 > 0.6$ within a 1000-bp sliding window) was applied to balance removing strongly correlated variants with retaining sufficient genetic variation for population structure analysis. This approach followed a previous EBV genomic study (Xu et al., 2019), in which SNVs were similarly filtered for minor allele frequency (MAF >5%) and LD pruning ($r^2 > 0.6$) prior to PCA, yielding several hundred informative variants for downstream analysis. Importantly, LD pruning does not eliminate LD entirely but reduces excessive redundancy among highly correlated variants. By using a relaxed threshold ($r^2 > 0.6$), 776 variants were retained for PCA.

All analyses were conducted in PLINK v1.9 (Chang et al., 2015), a whole-genome analysis toolset. The PCA was visualised in R using tidyverse, patchwork, scale and RColorBrewer packages (Team, 2016).

### 3.3.10 Recombination analysis

Recombination frequency was estimated from patterns of linkage disequilibrium (LD) decay across EBV genomes from different disease phenotypes and EBV-positive healthy participants. LD quantifies the non-random association of alleles at two loci; loci that are physically close are more likely to co-occur on the same haplotype and therefore exhibit higher LD, and LD decays with increasing physical distance in proportion to recombination activity. This general rationale is well established in population genetics (Slatkin, 2008).

For viral genomes, LD measures how often two mutations co-occur within the same sequenced genome (haplotype), relative to what would be expected under independence. Recombination, often caused by template switching by the viral polymerase during replication, breaks these associations and results in an observable decay of LD with increasing genomic distance. Using LD decay in viruses has proven to be a valuable method for estimating relative recombination rates across different viruses and genomic regions (Patiño-Galindo et al., 2021; Romero & Feder, 2024).

Merged pre-processed VCFs were filtered for MAF > 1%, and pairwise LD was calculated for each phenotype ($r^2 > 0.1$) within 10-kb sliding windows. The choice

of a stricter LD threshold ($r^2 > 0.1$) aimed to retain only weakly correlated variants and remove strongly linked sites. This method has been shown to reduce redundant sites and improve the resolution of LD decay patterns, making subtle recombination signals detectable that would otherwise be masked by strong LD (Vos et al., 2017).

Pairwise LD between single-nucleotide polymorphisms (SNPs) was assessed using the squared correlation coefficient ($r^2$) in PLINK v1.9 (Chang et al., 2015). For each phenotype group, pairwise SNP comparisons were extracted, and the genomic distance between SNPs was calculated as the absolute difference in base-pair coordinates. To characterise LD decay across the viral genome, SNP pairs separated by more than 10 kb were excluded, as LD at greater distances is typically weak in recombining viral populations and contributes little to LD decay estimation.

To summarise LD as a function of genomic distance, SNP pairs were grouped into 1-kb distance bins, and the mean $r^2$ within each bin was calculated. Binning pairwise LD values is a common method in population genomics to reduce noise in pairwise LD estimates and facilitate visualisation of LD decay curves across genomic distance. This approach is particularly useful for viral genomes, where dense SNP comparisons can produce highly variable pairwise LD estimates. Previous comparative analyses of EBV genomes have shown substantial genome-wide LD patterns despite recombination, supporting the use of distance-based LD decay analyses to characterise viral population structure (Wegner et al., 2019; Xu et al., 2019).

To generate smooth LD decay curves and estimate uncertainty, LOESS smoothing (a nonparametric locally weighted regression method) was applied to the binned $r^2$ values (Cleveland & Devlin, 1988). Confidence intervals were derived through a bootstrap process (n = 200), with SNP pairs resampled with replacement within each phenotype group. For each bootstrap sample, the LOESS model was refitted, and predicted $r^2$ values were calculated across a range of genomic distances. The average smoothed curve and bootstrap percentile confidence intervals were then obtained. Since short-distance SNP pairs are often underrepresented in large LD datasets after subsampling, stratified sampling by distance bins (500 bp) was used prior to smoothing to maintain the density of short-distance comparisons.

To compare LD decay among phenotypes, an exponential decay model was fitted:

$r^2 = a \times e^{(-b \times d)}$ where,

$r^2$: is the linkage disequilibrium between a SNP pair,

d: is the genomic distance,

a: represents the intercept (LD at zero distance), and

b: is the LD decay rate parameter.

Models were fitted using nonlinear least squares (NLS) implemented in R v4.4.1 (Team, 2016). For robust estimates of b, the fitting was repeated across bootstrap replicates (n = 200), generated by resampling SNP pairs within each phenotype group. The bootstrap distribution of b values provided an estimate of uncertainty in LD decay rates. Differences among phenotype groups in LD decay rates were

tested using a Kruskal–Wallis rank-sum test, a non-parametric method suitable for comparing multiple groups (Kruskal & Wallis, 1952). When the Kruskal–Wallis test was significant, pairwise Dunn's post hoc tests identified specific differences, with p-values adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini & Hochberg, 1995). Bootstrap distributions of the decay parameter b were visualised using boxplots in R.

Importantly, LD values were calculated from pooled SNP pairs, and $r^2$ values are population-level summaries across samples that cannot be easily linked to individual sample covariates; population structure could not be explicitly incorporated into the models. Additionally, phenotype and geographic origin were partially confounded in the dataset (e.g., NPC genomes mainly originated from Asian populations, whereas BL genomes were mostly from African populations). Consequently, the observed differences in LD decay should be interpreted as associations rather than definitive phenotype-specific effects.

### 3.3.11 Assessing the impact of recombination on EBV genome evolution

The impact of recombination on EBV genome evolution was evaluated using the Rapid Recombination Detection Programme (RDP4 v4.101) (Martin et al., 2015). RDP4 employs an ensemble of methods, including phylogeny-based (RDP and Bootscan) and statistical models (GENECONV, MaxChi, Chimaera, 3Seq, and SiSscan), to identify recombination events and breakpoints in DNA sequences. EBV genomes from 141 BL cases and 78 EBV-positive healthy controls were analysed. Multiple sequence alignments were generated using MAFFT v7.520

(Katoh et al., 2002) and trimmed using Gblocks v0.91 (Castresana, 2000) before loading into RDP4 with default window and step sizes of 200 and 20, respectively. The output from RDP4, which included recombination events, breakpoint positions, recombinant sequences, and major and minor parental sequences, was exported in CSV format for further analysis. Putative recombination events were only considered if at least five methods, including at least one phylogenetic-based method, detected the event with a threshold p-value $< 0.05$ after Bonferroni correction.

Phylogenetic trees, with and without recombinant regions, were inferred using the maximum likelihood (ML) method implemented in RAxML, employing the GTR model and 1000 bootstrap replicates (Togkousidis et al., 2023). The ML tree topologies were compared using the normalised Robinson-Foulds (RF) (Robinson & Foulds, 1981) implemented in the phangorn package in R (Schliep, 2011). The RF metric quantifies the proportion of bipartitions that differ between two trees, with values ranging from 0 (identical trees) to 1 (completely distinct topologies). The trees were visualised using R packages ape (v5.8.1), phangorn (v2.12.1) and phytools (v2.4.4) (Revell, 2024; Schliep, 2011).

Next, the impact of recombination on EBV protein-coding sequences was evaluated. Coding DNA sequences (CDS) were extracted from the multiple sequence alignment files using a custom Python pipeline. The alignment files were parsed with Biopython v1.85 (Cock et al., 2009), and the reference genome (NC_007605.1) was used to map alignment coordinates to genomic positions. Gene

coordinates from a BED annotation file were grouped by gene, and exons were merged to produce contiguous CDS sequences, including wraparound genes such as LMP-2A. For each gene, concatenated CDS alignments were exported as FASTA files using pandas v2.3.0 (Team, 2020) for BED parsing and defaultdict structures for sequence assembly. A phylogenetic tree for each CDS fasta alignment was inferred with IQ-TREE v3.0.1 (Wong et al., 2025), employing the default GTR+G substitution model and 1000 bootstraps.

The trees and fasta alignments were then used to estimate recombination parameters in ClonalFrameML v1.13 (Didelot & Wilson, 2015). The per-site recombination rate (ρ/bp) was calculated from the ratio of recombination to mutation rates ($\rho/\theta$) and the average length of imported DNA ($\partial$). Gene-specific 95% highest posterior density intervals (HPDI) were derived by propagating uncertainty from the posterior distributions of $\rho/\theta$ and $\partial$. Genome-wide 95% HPDI estimates for ρ/bp were also determined for each group. Visualisation was carried out in R using ggplot, with genes ordered by their per-site recombination rate (ρ/bp), and recombination hotspots defined as genes with a per-site recombination rate exceeding the upper limit of the genome-wide HPDI, as established previously (Berenstein et al., 2018).

### 3.3.12 Assessing EBV mutation frequency

Mutation burden was compared across BL, NPC, NKTCL, HL, GC, Healthy, and Other EBV-associated malignancies. For each group, the total number of unique variants was calculated from aggregated variant lists. Because cohort sizes differed

across groups, raw variant counts were not considered directly comparable. To account for cross-group variant sharing, a fractional-weighting approach was employed. For each variant, the number of groups in which it was observed (k) was determined, and a weight of 1/k was assigned to that variant within each group. Weighted variant counts per group were then calculated as the sum of fractional weights across all variants observed in that group. This procedure reduces inflation of variant counts caused by broadly shared variants.

To further adjust for differences in cohort size, both raw and weighted variant counts were normalised by the number of samples in each group, yielding per-sample mutation-burden estimates. Four metrics were therefore evaluated: (i) raw unique variant counts, (ii) fractional weighted variant counts, (iii) raw variants per sample, and (iv) weighted variants per sample. These complementary measures enabled comparison of overall variant diversity, sharing-adjusted diversity, and sample-size–normalized mutation burden across disease groups.

To examine the overlap of mutations and identify those unique to each phenotype, variant lists were compiled for each sample using bcftools (v1.12) (Danecek et al., 2021). The lists were then converted into a binary presence/absence matrix, where 1 indicated the presence of a given variant in a group, and 0 indicated its absence. The resulting matrix was used as input to the UpSet (v1.4.0) visualisation in R (Conway et al., 2017).

To assess genome-wide variant distribution, the average number of variants per window across samples was calculated. The resulting mean variant density per kb was then visualised as a line plot to emphasise regions with higher or lower

mutation frequencies. R packages dplyr (v1.1.4), tidyr (v1.3.1), and ggplot (v3.5.2) were used for data manipulation and plotting (Ginestet, 2011; Siddiqui, 2025).

Next, gene-level mutation enrichment in Burkitt lymphoma compared to EBV-positive healthy controls was assessed using Fisher's exact test. For each gene, a 2 × 2 contingency table was created to compare the number of observed mutations in the gene of interest with those in all other genes between the BL and healthy samples. Specifically, the table included i) the number of mutations in the given gene in BL, ii) the number of mutations in all other genes in BL, iii) the number of mutations in the given gene in healthy controls, and iv) the number of mutations in all other genes in healthy controls. The odds ratios (ORs) and 95% confidence intervals were estimated from Fisher's exact test, with effect sizes expressed as $\log_2$-transformed odds ratios (log2OR). P-values were adjusted for multiple comparisons across genes using the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995). Genes with FDR < 0.05 were regarded as statistically significant. Volcano plots were generated to visualise enrichment (log2OR > 0) and depletion (log2OR < 0) in BL relative to healthy controls. All analyses were conducted in R v4.4.1 using the tidyverse (v2.0), ggrepel (v0.9.6), and patchwork (v1.3.2) libraries (Pedersen, 2019; Slowikowski et al., 2018).

### 3.3.13 EBV mutation signatures analysis

To start deconvoluting EBV mutation signatures, a mutation matrix was generated from individual sample VCFs. Variants were parsed using cyVCF2 v0.30.28 (Pedersen & Quinlan, 2017), and Biopython v1.85 (Cock et al., 2009) was used to extract trinucleotide sequence context from the custom hybrid EBV reference genome (NC_007605.1 plus contigs from EBNA2 and EBNA3s from NC_009334.1). Mutations were normalised to the pyrimidine context and categorised into the six base substitution classes (C>A, C>G, C>T, T>A, T>C, T>G). These were then expanded into 96 trinucleotide contexts following the COSMIC convention (Alexandrov et al., 2020). For each sample, mutation counts across all contexts were aggregated into a sample-by-context matrix, which was exported for downstream mutational signature analysis.

The R package MutSignature (v2.1.5) was employed to extract de novo EBV mutation signatures using the non-negative matrix factorisation (NMF) algorithm (Fantini et al., 2020). To identify the optimal number of signatures, a rank survey was conducted for k = 2 to 5 over 50 iterations, using both the original and randomised mutation matrices (Gaujoux & Seoighe, 2010). The best number of signatures (k = 4) was selected based on cophenetic correlation, silhouette width, consensus clustering stability, and residual sum of squares (Fantini et al., 2020). Signature exposure proportions were normalised within each sample and combined with the metadata for statistical analysis. The normality of signature exposures was tested using the Shapiro-Wilk test (Shapiro & Wilk, 1965). Because the data were not normally distributed, a non-parametric approach was employed. Beta

regression models were fitted for each signature, adjusting for strain and region as fixed effects, to assess each signature's contribution to the total number of mutations across phenotypes. Significant phenotype-SBS associations (FDR < 0.05) were reported with their effect estimates.

To contextualise the de novo EBV signatures, a cosine similarity analysis was performed, comparing each signature with the well-curated COSMIC v3.2 reference SBS signatures database, using the lsa (v0.73.3) R package (Wild, 2007). Cosine similarity values of 0.90 or higher were deemed strong matches (Alexandrov et al., 2020). Bar plots and heatmaps were used to visualise the best matches.

### 3.3.14 Individual contributions

In this chapter, sample collection and processing, including EBV whole genome sequencing, were carried out at St. Mary's Hospital, Lacor, Central Public Health Laboratory (CPHL; Uganda), and Muhimbili University of Health and Allied Sciences (MUHAS; Tanzania), two of the collaborating sites for the AI-REAL study in East Africa. As a senior research assistant on the AI-REAL programme, I participated in these activities in Uganda. The EBV whole-genome sequencing panel (EBV probes) was developed by the OMDC laboratory at Oxford, led by Ms Daisy Jennings and Ms Helene Dreau. Fastq data processing, adapter trimming, BAM file alignment to EBV reference genomes, variant calling and annotation, and translation of consensus sequences were performed by Dr Kate Ridout.

## 3.4 Results

### 3.4.1 Primary study cohort characteristics

The primary study cohort consisted of 64 participants: 41 with Burkitt lymphoma, 2 with Hodgkin lymphoma, 10 with other lymphomas, and 11 healthy controls positive for EBV. The median age was 9 years (IQR: 6.0–11.0). There were more males than females (64% versus 36%), and 77% of the cohort was from Uganda. Clinical features such as weight loss, cachexia, and night sweats were common among BL cases ($P < 0.05$), and over half of the cases had advanced-stage disease (Table 3.1).

**Table 3. 1 Demographic and clinical features of the primary study cohort**

| Characteristics | **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

### 3.4.2 Summary of the analysed EBV genomes

The data from the primary study cohort were combined with published data to create a set of 410 EBV whole genomes, consisting of 332 samples from individuals with malignancies and 78 from EBV-positive healthy controls, all derived from EBV DNA isolated from primary samples (for information on analysis samples, see Supplementary Tables 7 & 8, Data Appendix). Illumina paired-end FASTQ data from the novel isolates have been submitted to the NCBI Sequence Read Archive under BioProject ID PRJNA1333409, to be released in September 2026. Most public EBV genomes lacked accompanying metadata, such as patient age, sex or gender, and EBV strain. Using the NCBI BLAST+ suite as described in Section 3.3.7, 266 (76.9%) of these genomes were identified as Type 1 and 80 (23.1%) as Type 2 (Supplementary Fig. 3). Seven samples with mixed or undetermined strains were excluded from further analysis.

### 3.4.3 Data quality assessment

To evaluate data quality, coverage statistics were computed from aligned BAM files as outlined in Section 3.3.6. The average sequencing depth of the primary study samples was 291x, with EBV genome coverage at 92.5%. In comparison, the corresponding metrics were 386.2x and 91.4% for the public genome dataset (Supplementary Fig. 4, Data Appendix).

Concordance between variant callers was evaluated by comparing variant positions across VarScan, VarDict, and Mutect2. Variants were classified as unique to a single caller, shared by two callers, or shared by all three callers. Overall, 65.07 %

of variants were detected by at least two callers, while 26.23 % were identified by all three callers, indicating substantial agreement among variant calling algorithms. Varscan and VarDict generally detected more variants than Mutect2, suggesting higher sensitivity (Supplementary Fig. 5, Data Appendix). However, Mutect2 showed greater agreement with the combined high-confidence set across several samples, suggesting a more conservative calling approach. To increase variant call confidence and reduce caller-specific artefacts, only variants detected by at least two independent algorithms were retained for downstream analysis.

To assess the accuracy of sequencing and variant calling, four primary study samples were processed in duplicate, and variant calling was performed to evaluate reproducibility. The average coefficient of variation between the first and second runs for the four sample duplicates was 2.3%, indicating higher sequencing reproducibility (Table 3.2).

**Table 3. 2 Reproducibility of Illumina sequencing and variant calling**

| Sample | Run1 (nvariants) | Run2 (nvariants) | 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 |

* Calculations included the number of variants (nvariants), absolute difference (AbsDiff), per cent difference (PercDiff), mean variant count, standard deviation (SD), and coefficient of variation (CV) of variant counts from the two runs.

### 3.4.4 EBV population structure and genomic variation

A total of 410 EBV genomes, including reference genomes (NC_007605.1 for EBV Type 1 and NC_009334.1 for Type 2) and two well-characterised EBV-positive B cell lines, Daudi and Raji (both belonging to Type 1), were used to construct the phylogenetic tree. These genomes represented five regions: Africa (Kenya, Malawi, Uganda), Asia (China), Central America (Guatemala), North America (USA), and South America (Peru). Because phenotypes were not uniformly distributed across regions in the available data, all phenotype comparisons are presented here as hypothesis-generating unless supported by stratified analyses by geography and EBV type.

The phylogenetic results revealed a clear geographic population structure of EBV genomes, clustering by region regardless of the disease phenotype (Figure 3.2). Within African samples, no distinct patterns were observed between cases and EBV-positive healthy controls. Similar patterns were also seen in Asian samples, with no evident phenotypic clustering. Among the 141 BL samples, 2 were from Asia and formed a subcluster with Asian NPCs; the African BL cases clustered with healthy African samples. In terms of strains, the larger proportion of African type 2 EBV genomes formed a separate subcluster from type 1 genomes, regardless of disease phenotype. Interestingly, four of the NPC-associated EBV genomes from Asia clustered with type 2 African EBV genomes. Local alignment with the NCBI BLAST suite confirmed that these genomes had the closest matches to type 2 EBV. Consequently, they were classified as type 2 genomes.

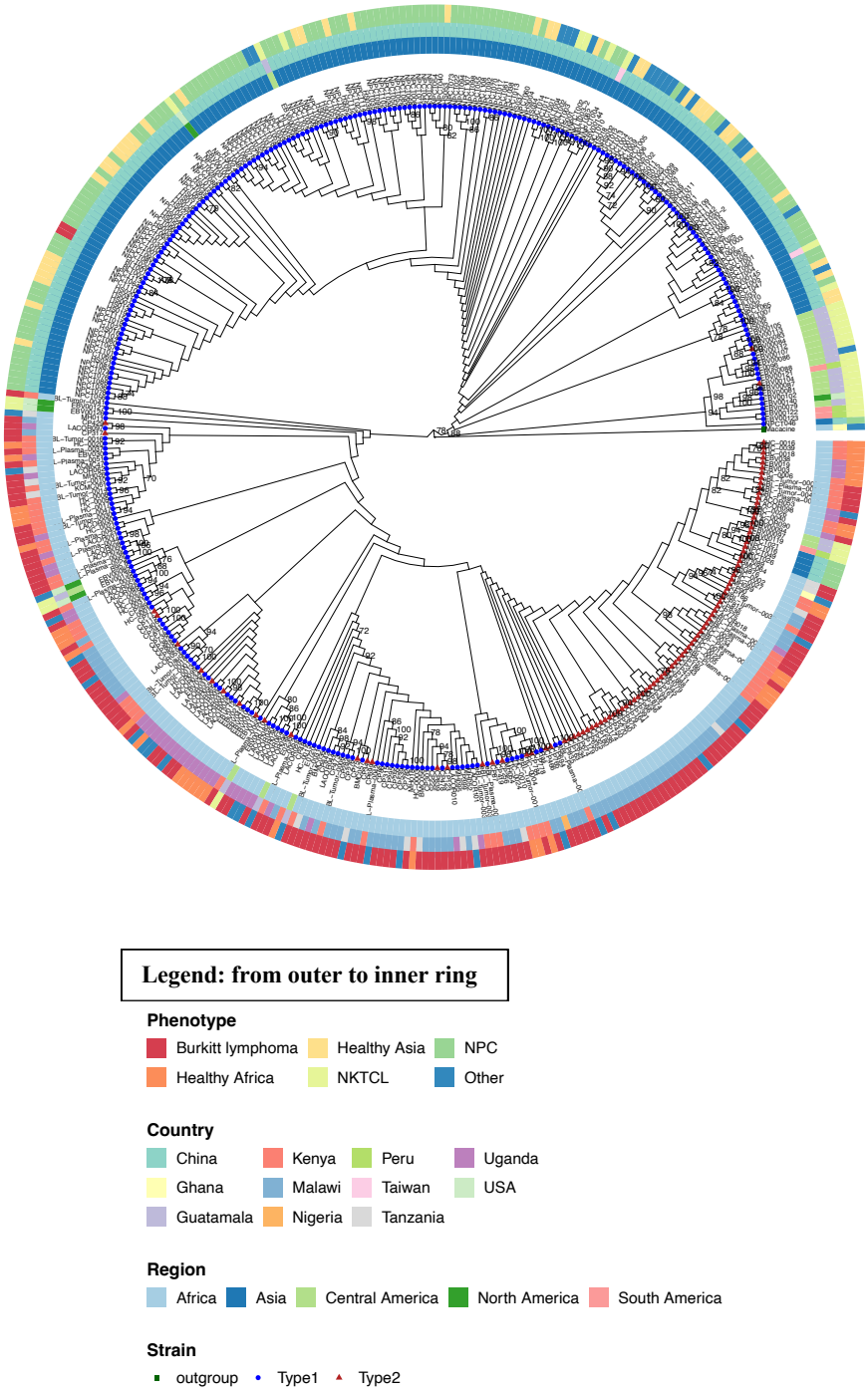



**Figure 3. 2 Phylogeny of the 410 EBV genomes**

The heatmaps: the outer ring represents sample phenotypes, the middle ring denotes the country of origin, and the inner ring indicates the region. The phenotypes analysed in this study include eBL, EBV-positive Healthy controls, HL, NKTCL, NPC, and other EBV-associated malignancies. The tips are annotated with EBV type (type 1 blue solid circle, type 2 firebrick triangle), and the most recent ancestor (outgroup green box) is indicated. The support values shown at the nodes represent TBE values computed from RAxML-ng; nodes with support values greater than 70% are displayed. Figure created in R using ggtree (v3.12.0).

To corroborate the phylogenetic findings, I performed principal component analysis on the 410 genomes. A total of 36,100 variants were loaded from the binary PLINK file. Thirty-four thousand three hundred twenty-nine (34,329) variants were excluded due to minor allele frequency thresholds (MAF < 5%), and a further 995 were pruned due to linkage disequilibrium ($r^2 > 0.6$), resulting in a final set of 776 variants used to generate the relationship matrix. In the PCA, the samples or genome regions of origin contributed most to the genetic variation (Figure 3.3). The first two principal components (PC1 and PC2) accounted for roughly 34% of the total variance. PC1 differentiated African strains from Asian strains, explaining 18.8% of the variation. PC2 accounted for 15.4% of the genetic variation and separated American EBV strains from those of Asian and African origins. Consistent with the phylogeny, no clear phenotypic clustering was observed. EBV strains (Type 1 versus Type 2) were spread across regions and did not define principal axes of variation. Even in higher-order PCs (PC3 versus PC4 and PC5 versus PC6), strain clustering was weak, indicating that intra-strain variation is subtle compared to regional differences.

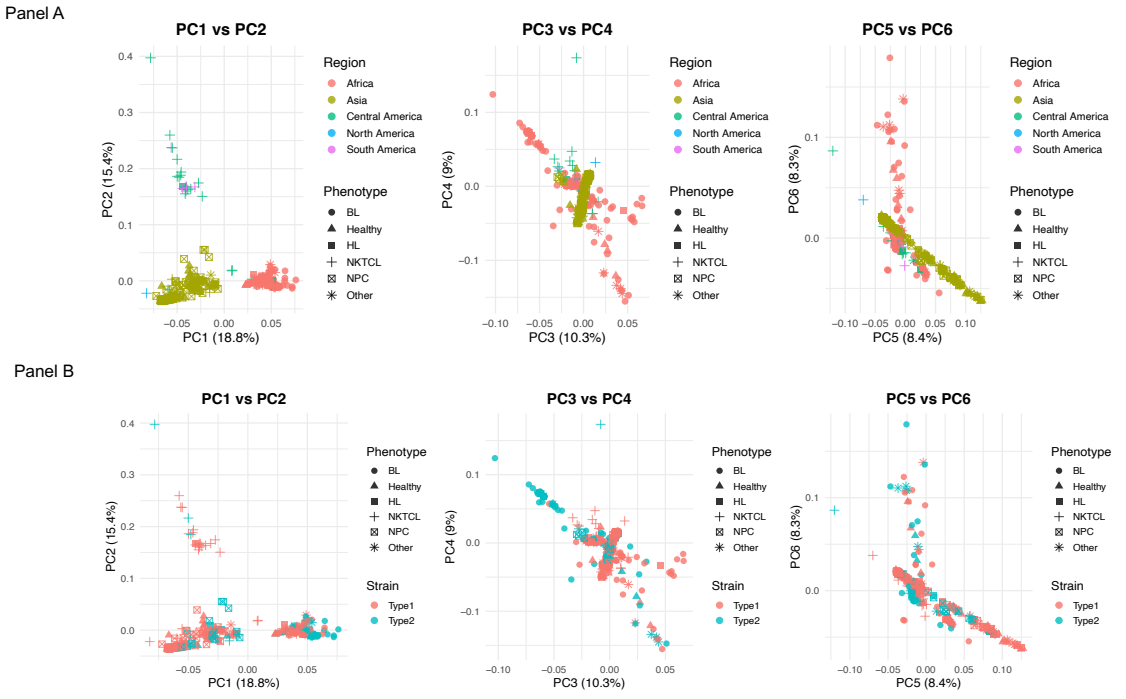



**Figure 3. 3 PCs showing region and strain differences**

**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.

### 3.4.5 Higher LD decay is observed in BL-derived EBV genomes compared to other phenotypes

A total of 410 EBV genomes were analysed; 141 from BL, 113 from NPC, 12 from HL, 34 from NKTCL, 32 from other EBV-associated malignancies, and 78 from healthy EBV-positive carriers. Genome-wide LD decay was examined across EBV genomes from different clinical phenotypes by plotting pairwise LD ($r^2$) against genomic distance. LD decay curves showed clear differences among phenotypes, with LD generally decreasing with increasing genomic distance (Figure 3.4).

Among the groups analysed, NKTCL genomes exhibited the highest overall LD ($r^2$ ~ 0.78) and the slowest decay with genomic distance, maintaining elevated LD even at larger genomic distances ($r^2 > 0.68$). In contrast, healthy EBV genomes showed the lowest LD levels ($r^2$ ~0.37, dropping to ~0.2 at 10kb) and the fastest decay, consistent with weaker linkage between distant loci and greater evidence of recombination. Similar to EBV healthy populations, BLs showed relatively low LD, $r^2$ ~0.4, dropping to ~0.5 at 10kb at the largest genomic distances, indicating frequent recombination. Intermediate LD decay patterns were observed for NPC ($r^2$ ~ 0.45), HL ($r^2$ ~ 0.47), and other phenotypes ($r^2$ ~ 0.46), suggesting varying degrees of recombination and population structure across viral populations.

To quantify these differences, an exponential decay model was fitted to the LD decay curves for each phenotype, and the decay-rate b was estimated using

nonlinear least-squares regression. Bootstrapping was performed to obtain distributions of b values for each group.

Bootstrap distributions of the LD decay parameter b differed across phenotype groups (Figure 3.5). A Kruskal–Wallis test indicated a significant difference in decay rates among groups (Kruskal–Wallis chi-squared = 792.44, $p < 2.2 \times 10^{-16}$). Pairwise comparisons using Dunn's post hoc test with Benjamini–Hochberg correction identified significant differences between several phenotype pairs (Supplementary Table 1, Data Appendix). In particular, NKTCL genomes showed significantly lower decay rates (Dunn's FDR < $1.9 \times 10^{-66}$) than healthy genomes, consistent with the elevated LD observed across the genome. While BL exhibited faster decay than other phenotypes, it was similar to that observed in NPC genomes (Dunn's FDR = 0.42), suggesting comparable recombination patterns in these two common EBV-associated malignancies. Additional pairwise differences were observed among other phenotype groups, reflecting heterogeneity in LD structure across EBV populations. Summary statistics for the decay parameter b (median and interquartile range) are provided in Figure 3.5 and the annexed table.

However, because BL samples contained both Type 1 and Type 2 EBV strains, an additional strain-stratified analysis was performed to assess whether EBV strain contributed to the observed decay patterns. When BL genomes were stratified by EBV strain, significant differences in LD decay were observed between Type 1 and Type 2 viruses (Supplementary Fig.6). Type 2 EBV genomes exhibited consistently higher LD decay rates than Type 1 genomes, indicating a more rapid breakdown of linkage disequilibrium with genomic distance. Further analysis of bootstrap

distributions of decay parameter b demonstrated significant differences in decay-rate estimates between Type 1 and Type 2 BL-derived EBV genomes (Mann–Whitney U test, $p < 2.2 \times 10^{-16}$; Supplementary Fig.7). These findings indicate that EBV strain contributes substantially to the LD structure observed within BL-associated viral populations.

Overall, the observed variation in LD decay among EBV phenotypes and strains suggests differences in the underlying evolutionary dynamics of EBV populations. Slower LD decay may reflect reduced recombination, clonal expansion, or population bottlenecks associated with particular disease contexts, whereas faster LD decay is consistent with greater recombination and haplotype turnover. However, because phenotype groups were partially confounded with geographic origin in the dataset, these findings should be interpreted cautiously, as both disease-associated evolution and underlying population structure are likely to contribute to the observed LD patterns.

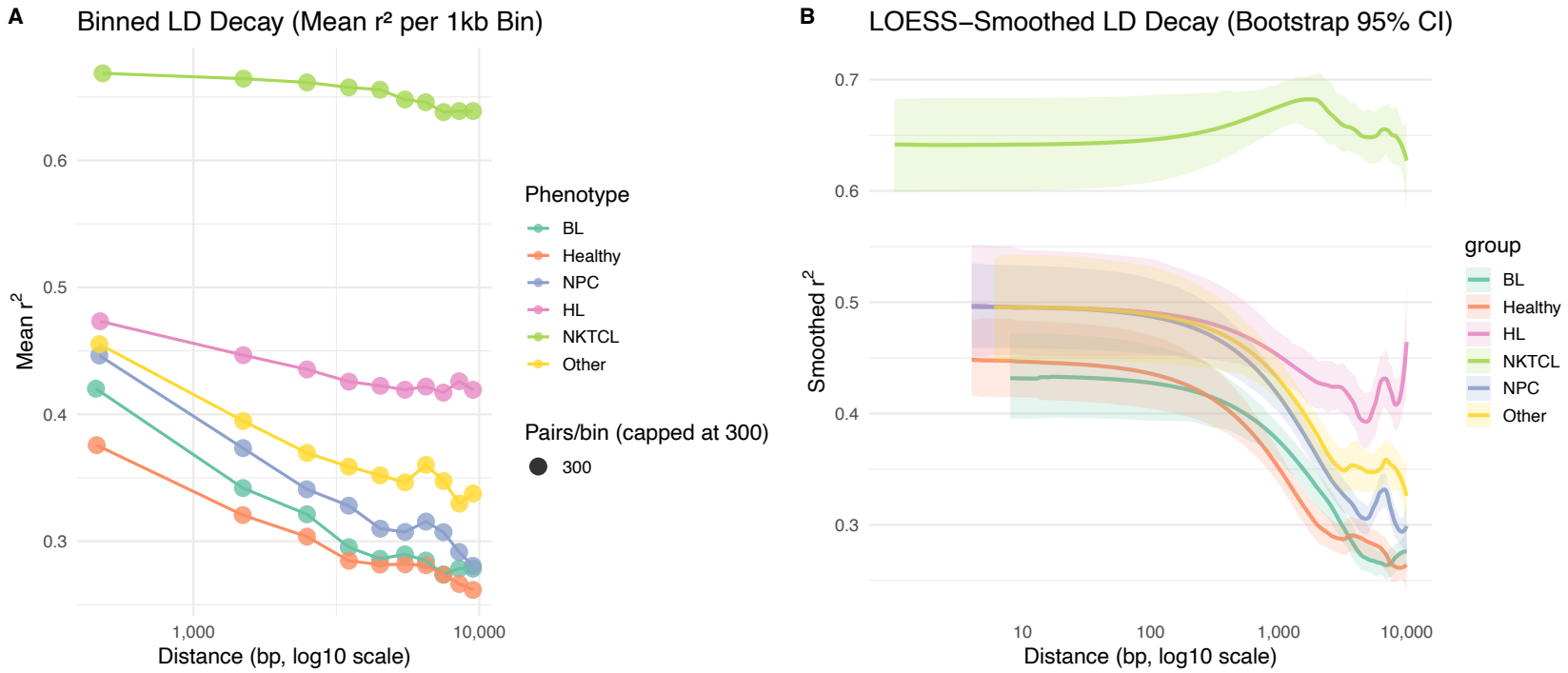
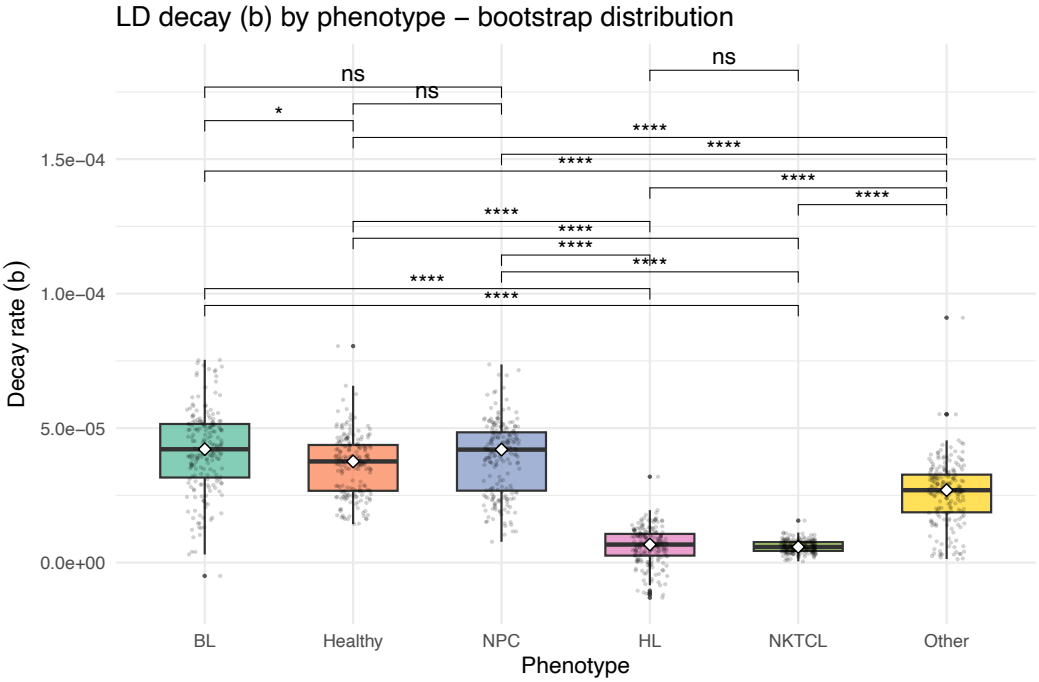



**Figure 3. 4 Genome-wide linkage disequilibrium decay analysis across EBV clinical phenotypes**

(A) Pairwise LD ($r^2$) plotted against genomic distance for each phenotype group. SNP pairs were grouped into 1-kb distance bins, and the mean $r^2$ within each bin was calculated to reduce noise in pairwise LD estimates. (B) LOESS-smoothed LD decay curves showing the relationship between $r^2$ and genomic distance. Shaded regions indicate bootstrap confidence intervals derived from resampling SNP pairs within each phenotype group. LD decreases with increasing genomic distance across all phenotypes, although the rate of decay varies among groups. Phenotypes included: BL (n=141), NPC (n=113), HL (n= 12), NKTCL (n= 34), Other EBV-associated malignancies (n = 32) and Healthy EBV-positive carriers (n = 78). Points represent mean $r^2$ values per bin, and smooth curves show LOESS fits for each group. Point size reflects the number of SNP pairs contributing to each distance bin (pairs/bin). Both panels are plotted on a $\log_{10}$ distance scale.




| Phenotype | N_boot | Median_b | IQR |
|---|---|---|---|
| BL | 199 | 4.21e−05 | 1.99e−05 |
| Healthy | 200 | 3.76e−05 | 1.70e−05 |
| NPC | 200 | 4.20e−05 | 2.17e−05 |
| HL | 200 | 6.71e−06 | 8.07e−06 |
| NKTCL | 200 | 5.84e−06 | 3.26e−06 |
| Other | 200 | 2.69e−05 | 1.40e−05 |

**Figure 3. 5 Pairwise comparison of LD decay rates across phenotypes**

The LD decay parameter b (exponential r^2 = a e^(-b x distance)) was estimated by fitting the model to 1-kb binned mean r^2 for each phenotype, with bootstrapping of the bins (n = 200). Boxplots display the bootstrap distribution of b for each phenotype; median (IQR) and bootstrap sample size (n) are shown. Statistical comparisons employed a Kruskal–Wallis test with Dunn's post-hoc tests and Benjamini–Hochberg correction; stars denote adjusted significance levels (**** p ≤ 1e-4; *** p ≤ 1e-3; ** p ≤ 1e-2; * p ≤ 0.05; ns = not significant). Since phenotype and geographic origin are partly confounded in the dataset (see Methods), these associations should be interpreted with caution.

### 3.4.6 Impact of recombination on EBV genome diversity

To evaluate the impact of recombination on EBV genome diversification, analyses were restricted to the BL and healthy control datasets. This restriction was necessary because other phenotype groups in the dataset were strongly confounded with geographic origin (e.g., NPC genomes were predominantly derived from Asian populations), whereas BL and healthy control genomes were largely sampled from African populations, providing a more balanced comparison with reduced geographic bias.

A total of 219 EBV genomes from BL and healthy controls were analysed, as described in Section 3.3.11. Putative recombination events were detected across all 219 genomes, of which 43 events were statistically significant (Supplementary Table 2, Data Appendix).

To assess the influence of recombination on EBV genome diversity, phylogenetic trees were reconstructed using the complete genome alignment and a second alignment in which recombinant regions were masked as described in Section 3.3.11. The resulting phylogenies were then compared to evaluate topological differences (Figure 3.6). The Robinson–Foulds distance between the two trees was 0.884, indicating substantial divergence in tree topology and suggesting that recombination contributes strongly to phylogenetic incongruence and genome diversification in EBV.

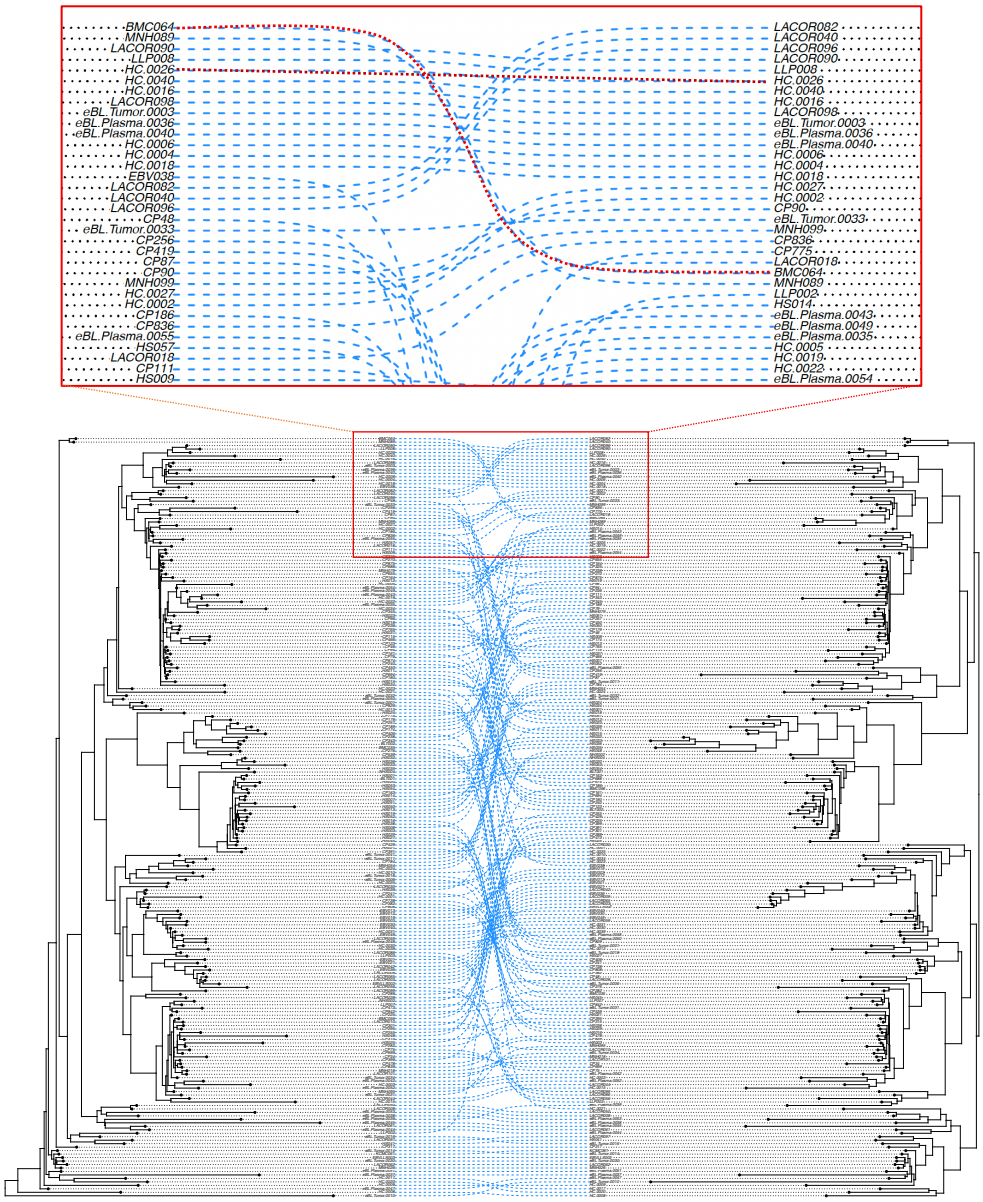



**Figure 3. 6 Tanglegram of 219 EBV genomes highlighting recombination-driven shuffling**

Tanglegram comparing maximum-likelihood phylogenies reconstructed from the complete genome alignment (left) and an alignment in which recombinant regions were masked (right). Lines connect identical genomes across the two trees. The magnified inset focuses on a region of high topological contrast: parallel lines indicate taxa with stable phylogenetic positions unaffected by recombination, while crossing lines (entanglements) identify taxa where genetic shuffling has significantly altered their ancestral placement. This visual divergence illustrates the heterogeneous impact of recombination across the EBV dataset. Differences in tree topology assessed using the Robinson–Foulds distance were 0.884, indicating substantial recombination-driven reshaping of EBV phylogenetic structure.

Given the robust recombination signals observed in BL-derived EBV genomes, the subsequent inquiry was whether recombination events occurred randomly across the viral genome. To evaluate this hypothesis, coding DNA sequences were extracted from the multiple FASTA alignments. Maximum likelihood phylogenetic trees were constructed for each coding DNA sequence alignment, as outlined in Section 3.3.11. The genome-wide recombination rate was estimated using ClonalFrameML v1.13 (Didelot & Wilson, 2015). ClonalFrameML was selected for its principal approach, which differentiates recombinant events from mutations within a phylogenetic framework, thereby accounting for the latter when assessing the influence of recombination on EBV evolution, as previously documented (Zanella et al., 2019).

The coding DNA sequence estimates revealed highly diverse recombination patterns across the EBV genome, as indicated by wide highest posterior density intervals (HPDIs; Figure 3.7). Of these, EBV latency and immune evasion genes (EBNA-1 and LMP-1) showed the highest level of recombinant events, well above the genome-wide 95% HPDIs. The remaining genes exhibited a wide variation in recombination rates (broad 95% HPDIs) in both cases and controls.

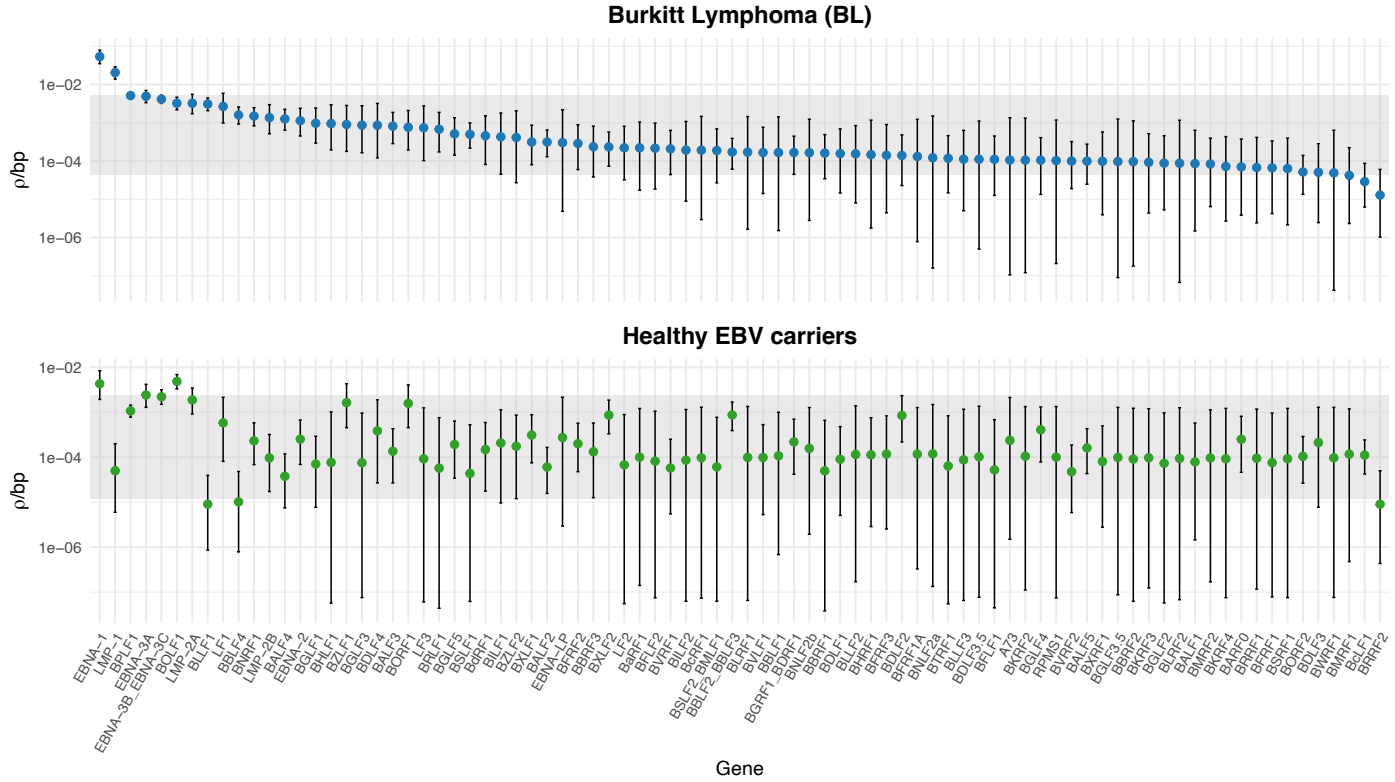



**Figure 3. 7 Recombination frequency (ρ/bp) across coding EBV genes**
The plots show per site (ρ/bp) recombination rates for each coding EBV gene. Each point represents the posterior mean of ρ per base pair (ρ/bp) for a gene, as estimated by ClonalFrameML. Top panel: EBV genes from BL. Bottom panel: EBV genes from EBV-positive healthy controls. Genes are ordered by their recombination rate. Vertical lines indicate the 95% highest posterior density intervals (HPDIs). The shaded grey area in each panel represents the genome-wide 95% HPDI for ρ/bp. Statistical significance is considered when the gene's HPDI interval does not overlap with the genome-wide 95% HPDI (grey area). The gene-specific error bars show intragenic heterogeneity in the recombination rate. The large 95% HPDI reflects more variability in the ρ/bp signal.

### 3.4.7 Point mutations and indels concentrate in latency and immune-evasion genes

To investigate differences in EBV genome diversity across disease contexts, mutation burden was compared across BL, NPC, NKTCL, HL, GC, healthy EBV-positive controls, and other EBV-associated malignancies. Because cohort sizes differed substantially among groups and many variants were shared between phenotypes, both raw variant counts and normalised mutation metrics were examined.

As expected, the total number of unique variants per phenotype group varied considerably, largely reflecting differences in cohort size. EBV genomes from BL and NKTCLs displayed the highest number of unique variants (10,068 versus 10, 486). When normalised by the cohort size, per-sample mutation loads were: NKTCL (349.1), HL (117.0), EBV-positive Healthy (87.0), BL (84.0), GC (49.2) and NPC (21.1), respectively. A descriptive summary of mutation frequencies across phenotypes is presented in Supplementary Fig.8 (Data Appendix).

To further account for variant sharing across phenotypes, a fractional weighting scheme was applied, in which variants detected in multiple groups were down-weighted proportionally by the number of groups in which they occurred. This approach reduced inflation of mutation counts caused by widely shared variants while preserving group-specific signals (Figure 3.8). Across all seven phenotypes, 813 variants were shared among groups (Figure 3.8), highlighting the presence of a core set of EBV polymorphisms circulating across diverse clinical contexts.

To investigate how mutations were distributed across the EBV genome in BL samples, mutation density was calculated in 1-kb genomic windows and normalised by the number of genomes analysed. This analysis revealed that mutation density was not uniformly distributed across the viral genome in BL (Figure 3.9). Regions corresponding to latency-associated and immune-modulatory genes (e.g., LMP1, EBNA1, BNLF2a, and BCRF1) exhibited relatively higher mutation densities, whereas several genomic regions involved in viral replication and genome maintenance (e.g., the EBNA1 G-rich repeat, LMP2A/B, and LF3s) appeared comparatively conserved.

Importantly, these observations represent descriptive patterns of mutation distribution rather than evidence of positive or negative selection. Mutation enrichment alone cannot be used to infer selective pressure, as such conclusions require formal evolutionary analyses (e.g., dN/dS-based selection inference) and functional validation to determine the biological impact of the observed variants. Therefore, while the distribution of mutations across functional regions of the EBV genome is noteworthy, no inference regarding selective pressure is made in this

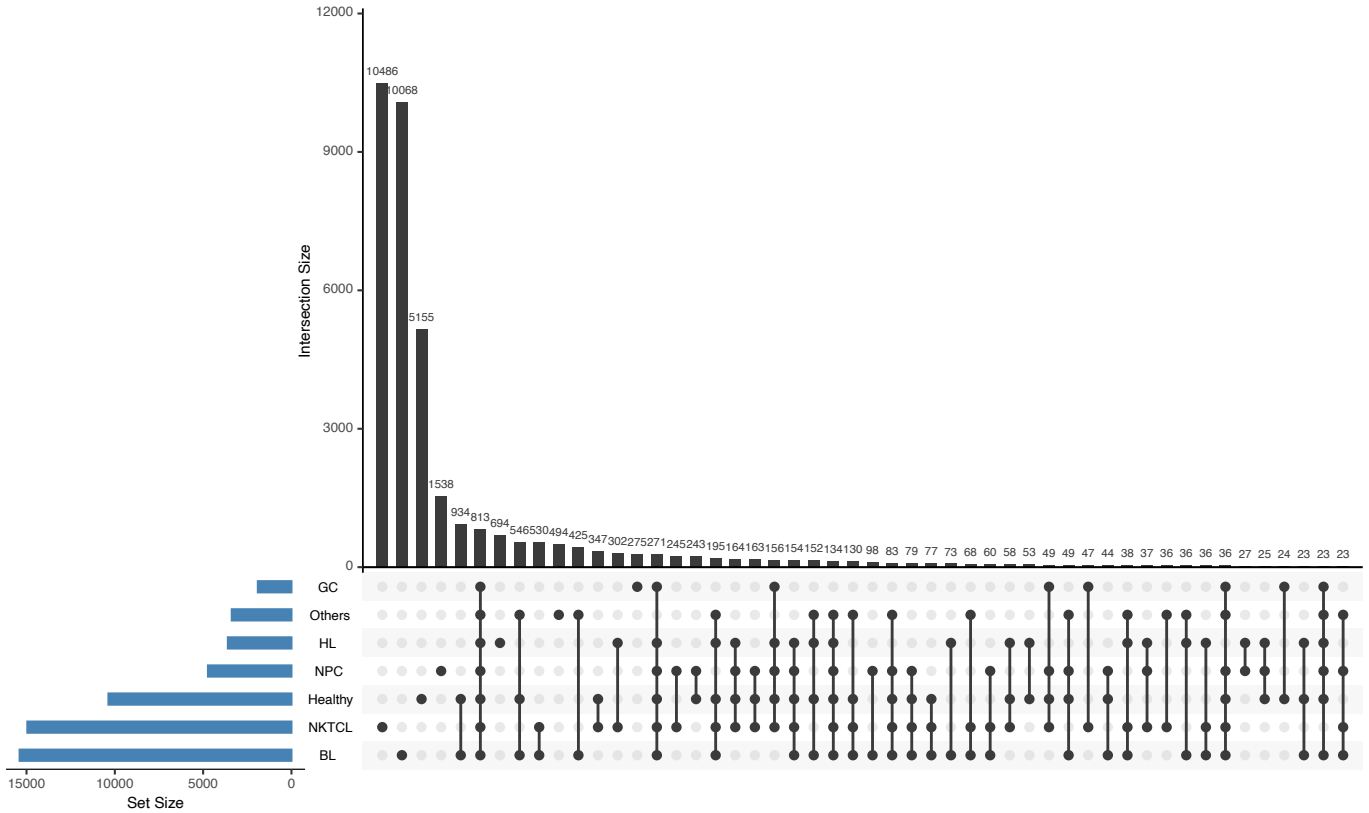



**Figure 3. 8 Comparison of EBV mutation burden across disease phenotypes**

The set size bar chart (horizontal, left side) displays the size of an individual set (mutation counts per group). The intersection matrix (middle, bottom) shows sets (phenotypes, rows) and intersections (connecting dots, columns) ordered by intersection size (largest to smallest). The intersection size bar chart (vertical, top) displays each vertical bar as the size of the intersection (count), as indicated by the corresponding column below.

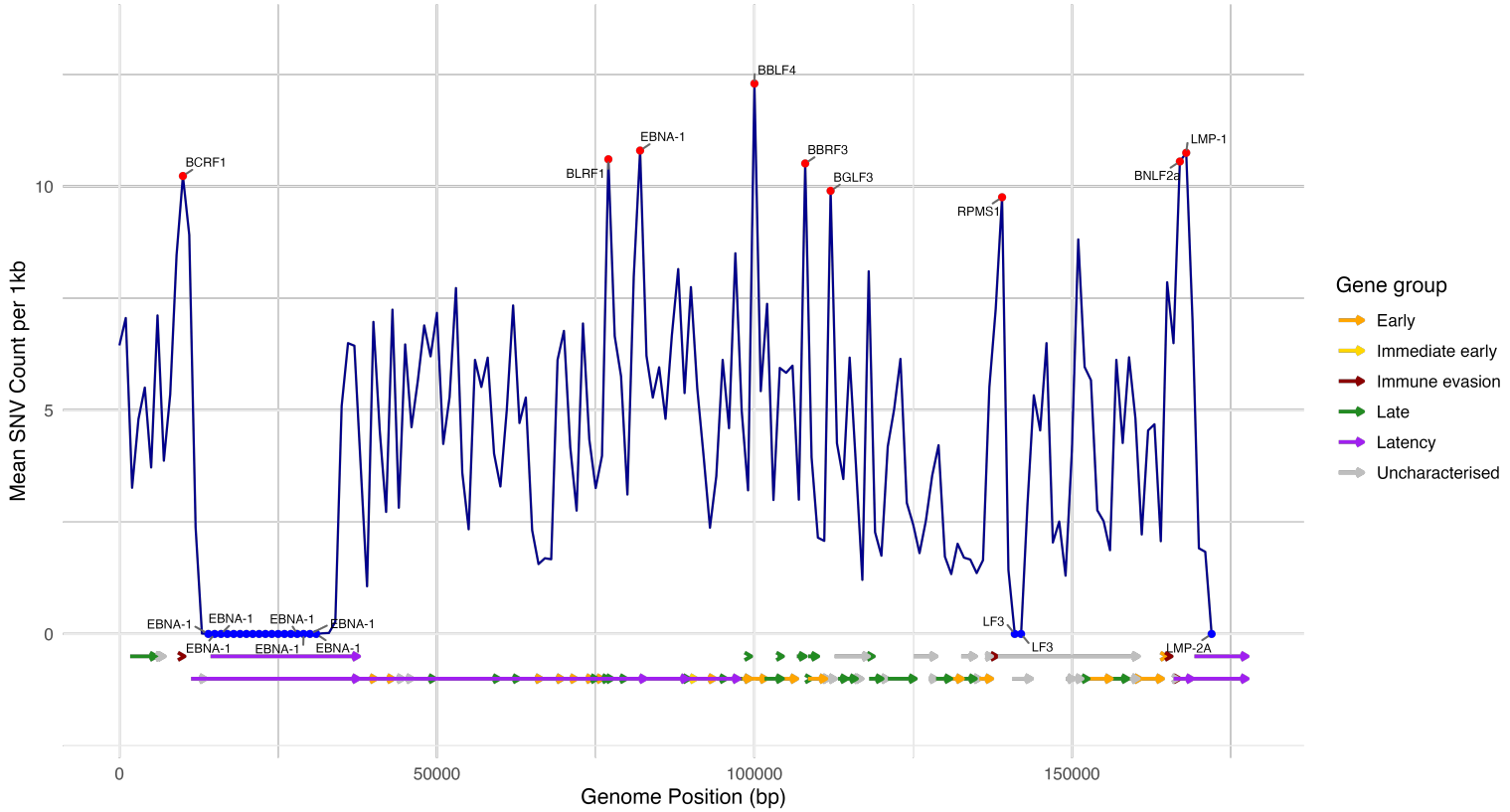



**Figure 3. 9 Normalised mean variant density per kb in EBV genomes from BL samples**

The line plot displays the mean SNV count per 1 kb across the EBV genome in eBL (normalised across the genome length). Red dots highlight hypervariable regions with high SNV density, while blue dots indicate SNV deserts. Labels identify genes within these regions. The gene track was annotated with bed-defined coordinates of the B95.8 (NC_007605.1) reference. Arrows point to EBV genes, coloured by function (latency, early, immediate early, late, immune evasion, uncharacterised).

To formally evaluate differences in mutational burden between BL and EBV-positive healthy controls, a gene-level mutation enrichment analysis was performed as described in Section 3.3.12. This comparison was restricted to BL and healthy control genomes because these cohorts were the most comparable within the dataset, with both groups predominantly derived from plasma cfDNA samples and largely originating from African populations, thereby minimising potential confounding effects observed in other phenotype groups.

Differential mutation enrichment analysis identified 19 genes significantly enriched in BL relative to Healthy controls (FDR < 0.05) (Figure 3.10). The strongest enrichment signal was observed for BHLF1 (OR = 7.68, $\log_2$OR = 2.94, 95% CI: 2.61–3.28, FDR = $5.35 \times 10^{-120}$), indicating a markedly higher mutation frequency in BL genomes. Several additional genes demonstrated moderate but highly significant enrichment, including LMP1 (OR = 1.56, $\log_2$OR = 0.64, FDR = $2.59 \times 10^{-24}$), EBNA-3A (OR = 1.51, $\log_2$OR = 0.60, FDR = $3.90 \times 10^{-23}$), and BLLF1 (OR = 1.35, $\log_2$OR = 0.44, FDR = $1.67 \times 10^{-11}$). Other significantly enriched loci included BBLF4, BRLF1, BNLF2a, BDLF3, BALF5, and BBLF2/BBLF3, all showing positive $\log_2$ odds ratios and FDR-adjusted p-values below 0.001.

Several genes also displayed relatively strong effect sizes ($\log_2$OR ≥ 1), including BNLF2a ($\log_2$OR = 1.16), BKRF3 ($\log_2$OR = 1.01), BSRF1 ($\log_2$OR = 1.23), and BLRF2 ($\log_2$OR = 0.98), corresponding to approximately two-fold or greater enrichment of mutations in BL genomes. Conversely, a subset of genes was significantly depleted in BL relative to Healthy controls. In particular, LMP2B,

BHRF1, and BOLF1 exhibited negative $\log_2$ odds ratios with significant FDR-adjusted p-values, indicating higher mutation frequencies in Healthy samples (Supplementary Table 3, Data Appendix).

Overall, this analysis revealed a distinct pattern of gene-level mutation enrichment and depletion distinguishing BL from Healthy EBV genomes, highlighting several loci that may contribute to BL-associated mutational processes.

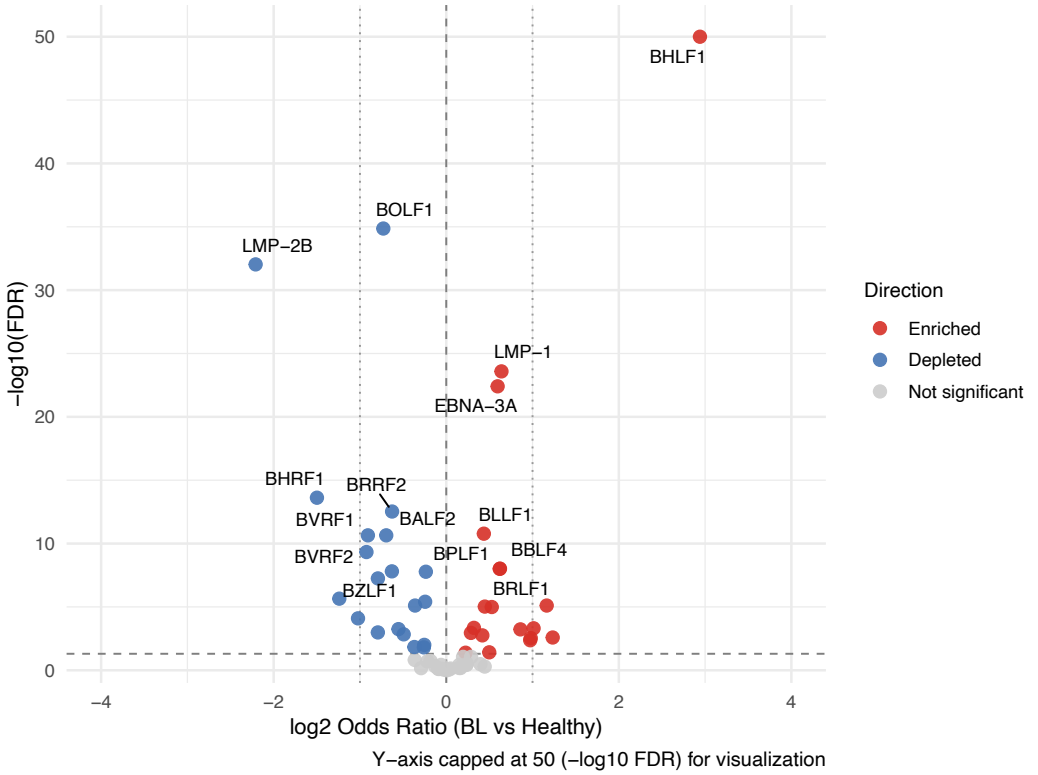



**Figure 3. 10 Gene-level mutation enrichment in BL versus Healthy controls**

Volcano plot showing gene-level mutation enrichment in BL relative to Healthy samples. The x-axis represents the $\log_2$ odds ratio (log2OR) from Fisher's exact test, where positive values indicate enrichment in BL and negative values indicate depletion relative to Healthy controls. The y-axis shows $-\log_{10}$ false discovery rate (FDR)–adjusted p-values. Each point represents one gene. Genes significantly enriched in BL (FDR < 0.05 and log2OR > 0) are shown in red, significantly depleted genes (FDR < 0.05 and log2OR < 0) are shown in blue, and non-significant genes are shown in grey. Horizontal dashed lines indicate the FDR significance threshold (0.05), and vertical dashed lines indicate no enrichment (log2OR = 0). Dotted vertical lines denote a two-fold change threshold (|log2OR| = 1).

### 3.4.8 EBV mutational signatures and association with Burkitt lymphoma

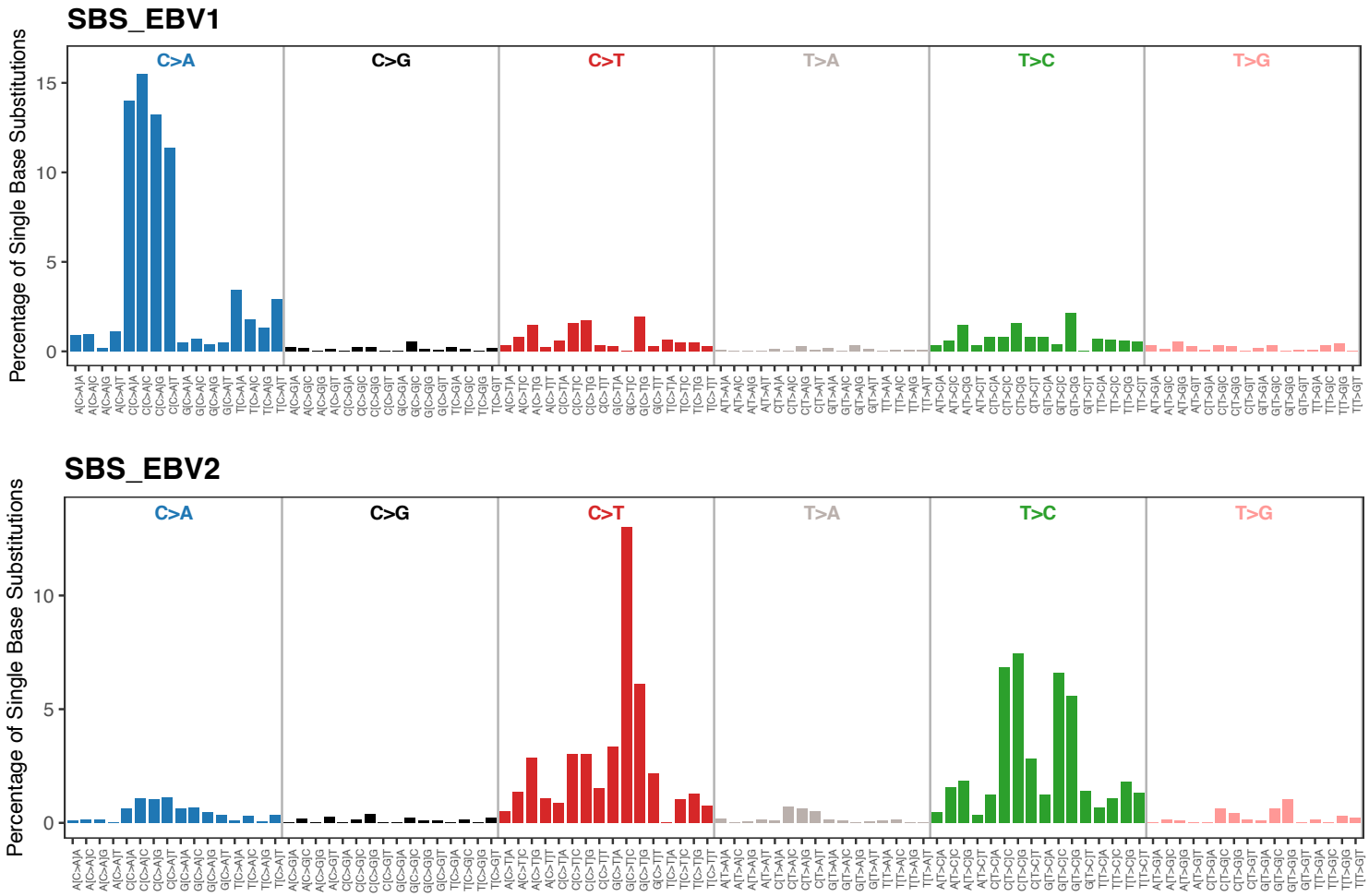
Having identified a distinct pattern of gene-level mutation enrichment in Burkitt lymphoma (BL) compared to EBV-positive Healthy controls, I then aimed to determine whether the underlying mutational processes varied across disease phenotypes. To examine this, non-negative matrix factorisation (NMF) was applied to the EBV mutational spectra derived from 410 viral genomes as described in Section 3.3.13.

First, an NMF rank survey (k = 2–5) was performed to determine the optimal number of mutational signatures using both the original dataset and a randomised control dataset. Model performance was evaluated using clustering stability (cophenetic correlation coefficient and dispersion), reconstruction accuracy (residual sum of squares and explained variance), and sample coherence (silhouette width). Across metrics, ranks k = 3 and k = 4 consistently demonstrated the highest stability and optimal reconstruction, whereas higher ranks yielded reduced clustering robustness and minimal gains in explained variance (Supplementary Fig.9, Data Appendix).

Reproducibility across independent NMF runs further supported the stability of these ranks, with both k = 3 and k = 4 yielding highly consistent signatures (silhouette ≥ 0.99; Supplementary Fig.10, Data Appendix). Correlation analysis confirmed that the extracted signatures were non-redundant and captured distinct mutational patterns (Supplementary Fig.11, Data Appendix).

Based on overall stability, interpretability, and minimal redundancy, k = 4 was selected for downstream analyses. The final model identified four distinct EBV mutational signatures (SBS-EBV1 to SBS-EBV4) across BL, NPC, HL, NKTCL and Healthy EBV-positive controls (Figure 3.11).

SBS-EBV1 was characterised by prominent C>A and C>T substitutions, particularly in CpC and CpG contexts, consistent with oxidative stress–associated mutagenesis. SBS-EBV2 exhibited a strong enrichment of C>T transitions at CpG sites, indicative of spontaneous deamination of methylated cytosines. SBS-EBV3 was dominated by C>T transitions with moderate C>A and T>C contributions, suggesting mixed endogenous mutational activity. SBS-EBV4 demonstrated a bias toward C>T, T>C, and T>G substitutions within G-rich or CpG contexts, a pattern consistent with combined oxidative and deamination-driven damage.


SBS_EBV1
Percentage of Single Base Substitutions
C>A
C>G
C>T
T>A
T>C
T>G
SBS_EBV2
Percentage of Single Base Substitutions
C>A
C>G
C>T
T>A
T>C
T>G

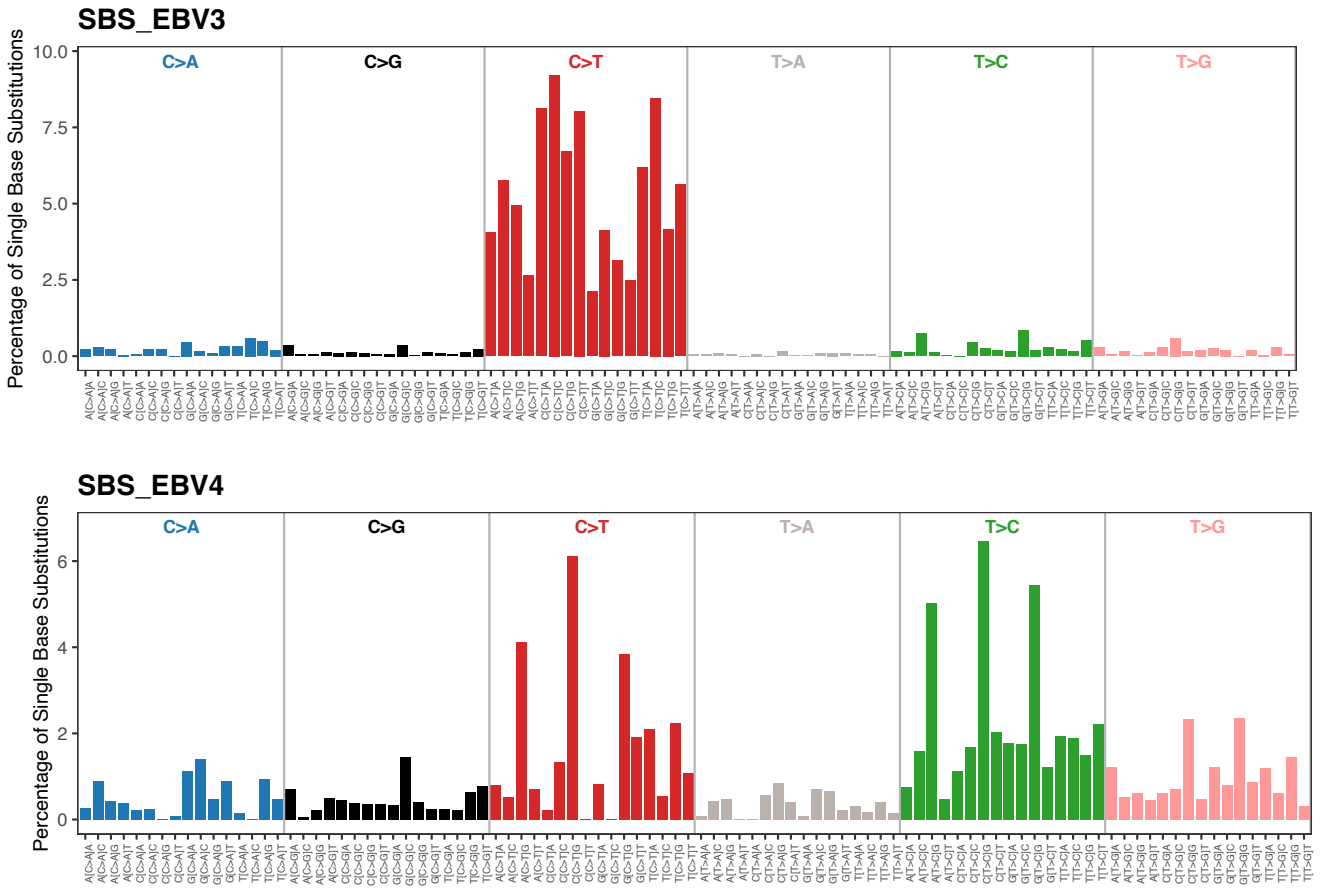



**Figure 3. 11 COSMIC-style mutational profiles of EBV-derived single-base substitutions**
Each plot shows the contribution of the 96 trinucleotide mutation contexts for each signature (SBS_EBV1–SBS_EBV4) as percentages of total substitutions within that signature. Substitutions are grouped by type (C>A, C>G, C>T, T>A, T>C, T>G) and coloured consistently across signatures. Vertical lines separate substitution groups, and coloured labels indicate substitution type.

To assess whether mutational signature contributions varied across EBV-associated clinical phenotypes, the proportional contributions of each signature were calculated and normalised. Samples were then grouped by region and phenotype, and the mean contributions were computed. This analysis revealed striking differences in mean signature profiles across the five geographic regions—Africa, Asia, Central America, North America, and South America—and among the phenotypes: BL, NPC, NKTCL, HL, and EBV-positive Healthy controls (Figure 3.12).

Overall, SBS_EBV4 accounted for the highest proportion of total mutational activity (40–60%) in BLs, HL, NKTCL, and NPCs, as well as across Africa, Asia, and Central America. SBS_EBV3 was enriched in NKTCL, particularly in the Americas, accounting for 21–72% of total mutational activity, suggesting a potential NKTCL-specific mutational process. SBS_EBV2 displayed moderate contributions across geographies, with peak activity (45.7%) observed in EBV-positive healthy controls from Africa. In contrast, SBS_EBV1 activity remained near background levels across several regions and phenotypes, with elevations primarily observed in African BL (21.7%) and HL (17.6%) cases, suggesting a shared mutational pattern.

These geographic and phenotype-specific differences in mutational contexts and signature proportions suggest distinct, context-dependent EBV mutational processes and mechanisms of oncogenesis, warranting further investigation.

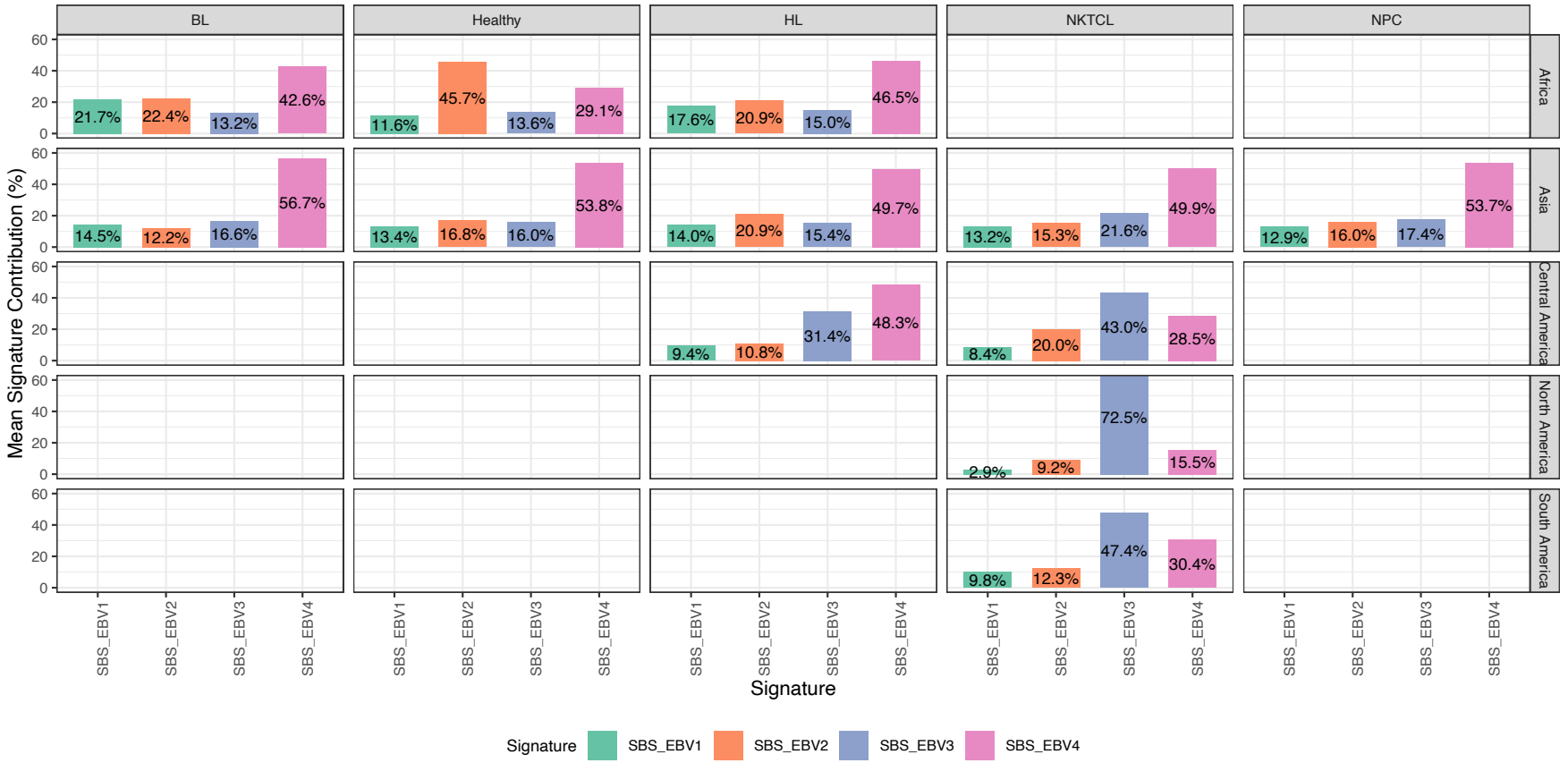



**Figure 3. 12 Mean contributions of EBV mutational signatures (SBS_EBV1–SBS_EBV4) stratified by Region and Phenotype**
Each stacked bar represents the average proportion of each signature across all samples within the corresponding Region × Phenotype group. Percentages are indicated within each bar segment. Signatures are coloured consistently across groups to facilitate comparison of relative contributions.

To determine whether the observed differences in EBV mutational signature proportions were statistically significant, formal hypothesis testing was undertaken. Normality of signature contribution distributions was first assessed using the Shapiro–Wilk test, which demonstrated significant deviation from normality (Shapiro–Wilk $p < 0.05$; Supplementary Fig.12, Data Appendix). Given the bounded nature of proportional data and violations of normality assumptions, beta regression models were used to assess associations between signature contributions and clinical phenotype, adjusting for region and viral strain as fixed effects.

Group comparisons revealed clear phenotype-specific differences in mutational signature activity (Figure 3.13). SBS_EBV1, SBS_EBV3, and SBS_EBV4 were significantly enriched in BL compared with EBV-positive healthy controls (adjusted p-values $<0.001$), consistent with earlier gene-level mutation enrichment findings. In contrast, SBS_EBV2 was significantly enriched in EBV-positive healthy individuals and comparatively reduced in BL, suggesting that this signature may be associated with non-malignant or latency-associated viral processes.

Other EBV-associated malignancies (NPC, HL, and NKTCL) did not demonstrate statistically significant enrichment of any signature relative to EBV-positive healthy controls after adjustment for strain and region. Although modest shifts in signature contributions were observed in some groups, these differences did not reach statistical significance.

Taken together, these findings indicate that BL is characterised by a distinct mutational signature profile, marked by enrichment of SBS_EBV1, SBS_EBV3, and SBS_EBV4, whereas EBV-positive healthy carriage is distinguished by relative enrichment of SBS_EBV2.

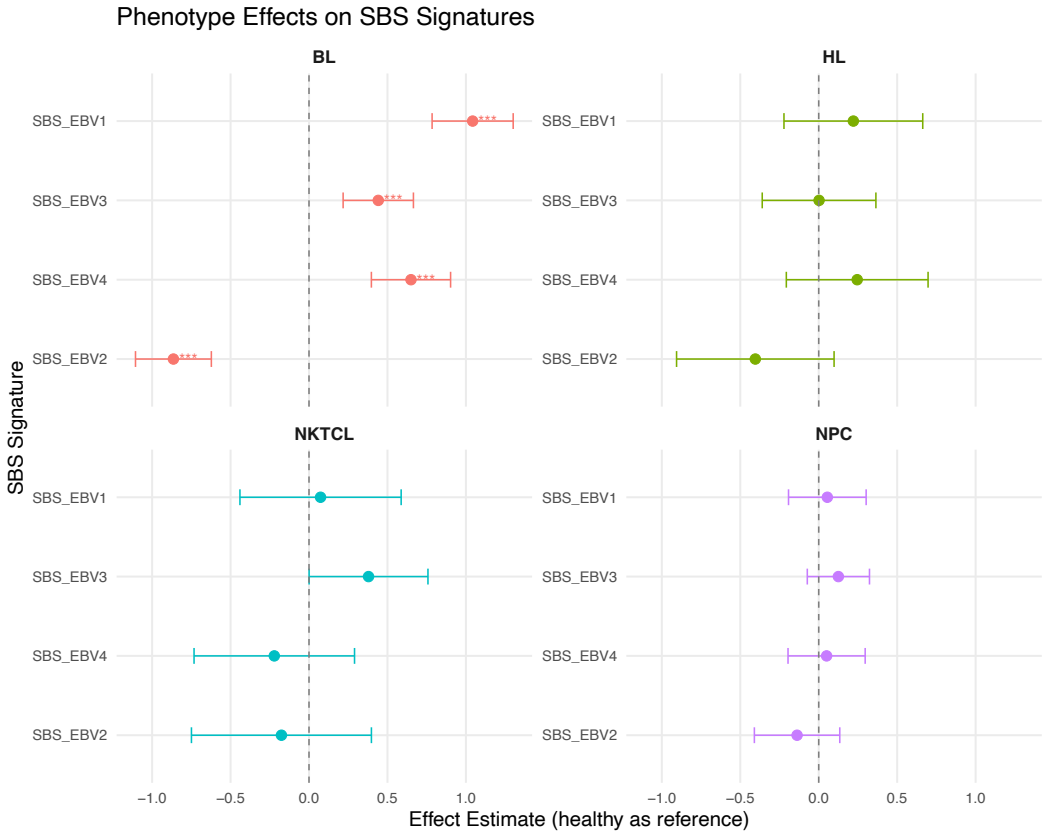



**Figure 3. 13 Beta regression model of EBV mutational signature proportions in disease phenotypes relative to healthy controls**

The plot displays effect size estimates (β coefficients) from beta regression models assessing the association between EBV mutational signatures and phenotypes. Bars show the direction and magnitude of effects relative to the reference group (healthy controls), with positive values indicating increased signature activity and negative values indicating decreased activity relative to healthy controls. Effect sizes (β coefficients) with 95% confidence intervals are shown. Disease phenotypes analysed included Burkitt lymphoma (BL, n=141), Hodgkin lymphoma (HL, n=12), natural killer T cell lymphomas (NKTCL, n=34), and nasopharyngeal carcinoma (NPC, n=113).

To relate the EBV-derived mutational signatures (SBS_EBV1–4) to established mutational processes in human cancers, a cosine similarity analysis was performed by fitting the EBV mutation signatures to the well-curated COSMIC SBS v3.2 reference database (Alexandrov et al., 2020).

The analysis revealed that SBS_EBV1 and SBS_EBV3 closely resembled known cancer-associated signatures (Figure 3.14). Specifically, SBS_EBV1 showed strong similarity to SBS45 (cosine similarity = 0.91), a signature of unknown aetiology characterised by G → T (C > A) transversions that are common in reactive oxygen species (ROS)-induced oxidation (Alexandrov et al., 2020). Meanwhile, SBS_EBV3 closely matched SBS30 (cosine similarity = 0.89), which is attributed to defective base excision repair (BER) caused by NTHL1 (endonuclease III-like 1) deficiency (Alexandrov et al., 2020; Drost et al., 2017).

In contrast, SBS_EBV2 and SBS_EBV4 showed only weak similarity to known COSMIC signatures, with their closest matches, albeit the weakest, being SBS42 (cosine similarity = 0.65) and SBS54 (cosine similarity = 0.71), respectively. Notably, none of the EBV mutational signatures matched the canonical APOBEC-associated COSMIC signatures (SBS2 and SBS13), which are characterised by a predominance of C > T and C > G substitutions within TpCpW trinucleotide contexts (Alexandrov et al., 2020).

A summary of de novo EBV mutational signatures identified and their likely mechanisms is provided in Table 3.3.

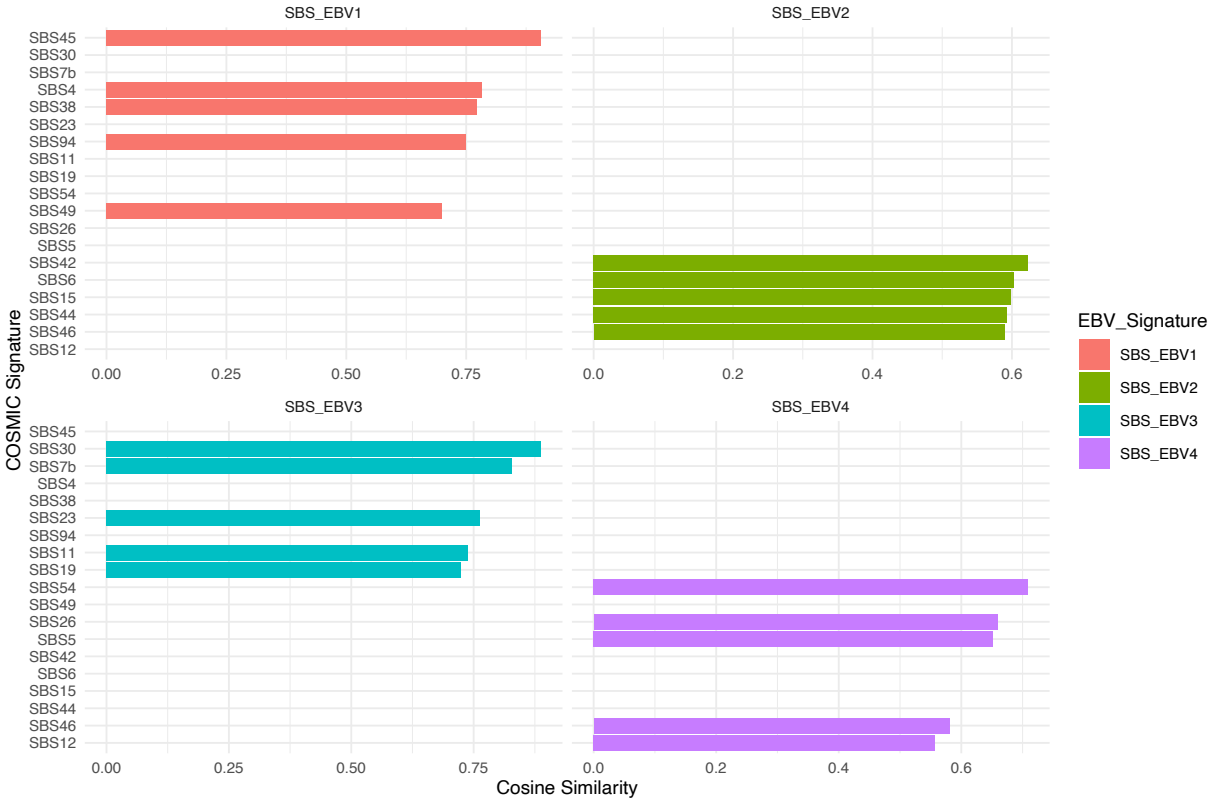



**Figure 3. 14 Top COSMIC SBS signature matches to de novo EBV signatures**

The bar plots show cosine similarity between each EBV-derived mutational signature (SBS_EBV1 to SBS_EBV4) and the top five matches to COSMIC single-base substitution (SBS) signatures. The x-axis represents the cosine similarities, and the y-axis represents the Cosmic SBS signatures. De novo SBS_EBV1 matched COSMIC SBS 45 (cosine similarity = 0.91), and SBS_EBV3 matched COSMIC SBS30 (cosine similarity = 0.89). SBS_EBV2 and SBS_EBV4 showed weak matches to COSMIC SBS 42 (cosine similarity = 0.65) and 54 (cosine similarity = 0.71), suggesting distinct mutational signatures in EBV-associated malignancies.

**Table 3. 3 Summary of de novo EBV mutation signatures and likely mechanisms**

| Signature | Closest COSMIC match | Mutation pattern | Key Contexts | Underlying mechanism-COSMIC | Likely EBV mechanism |
|---|---|---|---|---|---|
| SBS_EBV1 | SBS 45 (cosine 0.91 | C>A transversions | - High in C[C>A]C, C[C>A]A, C[C>A]T, C[C>T]C, C[C>T]G- Enrichment at CpC and CpG contexts | Sequencing artefacts related to oxidative damage to guanines (Alexandrov et al., 2020), and linked to survival urological carcinomas (Karihtala et al., 2023) | ? Likely a sequencing artifact |
| SBS_EBV2 | No match | C>T and C>T/C>G at CpG contexts | - High in G[C>T]C, C[C>T]C, C[C>T]G, C[T>C]G, C[T>C]C | | ? Likely basal replication associated substitutions in episomal maintenance |
| SBS_EBV3 | SBS30 (cosine 0.89) | C>T transitions, moderate C>A, T>C | - Enrichment in A[C>T]C, C[C>T]C, T[C>T]C, T[T>C]T | Defective or suppressed BER (NTHL1 or OGG1-like loss) (Drost et al., 2017; Lirussi & Nilsen, 2023) | ? Viral genome replication under BER impairment, possibly due to viral interference with NTHL1-like activity. |
| SBS_EBV4 | No match | C>T, C>G, and T>G mutations highly dominant | - Strong peaks at G[C>G]C, C[C>T]G, A[T>C]G, G[T>C]G, C[T>G]G | | ROS-induced and transcription-linked lesions, may reflect EBV lytic reactivation, with oxidative burst and viral polymerase activity. |

*SBS_EBV2 and SBS_EBV4 showed weak matches to COSMIC SBS42 and SBS54. Hence, they represent a distinct mutational spectrum in EBV-positive BL.

## 3.5 Discussion

This chapter integrates EBV whole-genome data from multiple clinical contexts using a unified analytical pipeline and adds 64 newly sequenced EBV genomes from children in Uganda and Tanzania. The primary objective was to determine whether specific EBV genomic features correlate with the epidemiological distribution of BL, thereby providing molecular insight into population-level associations and EBV-driven tumourigenesis.

Notably, several observations align with the established EBV genomics literature: EBV population structure is strongly influenced by geography; EBNA1 and LMP1 are among the most variable viral genes; and EBV recombination can increase genomic diversity, particularly in regions with high rates of mixed infections.

### 3.5.1 Geographic signature strongly influences EBV population structure

The analysis of 410 EBV genomes from Africa, Asia, South America, Central America, and North America revealed a population structure primarily driven by geography rather than disease association, consistent with prior reports (Briercheck et al., 2024; Chiara et al., 2016; Palser et al., 2015). Phylogenetic clustering demonstrated clear regional segregation of EBV lineages, with African sequences comprising both Type 1 and Type 2 strains but lacking disease-specific clustering. Principal component analysis of mutational profiles mirrored this geographic stratification. Together, these findings underscore the dominant role of long-term host–virus co-evolution in shaping EBV diversity. Founder effects, population

bottlenecks, and region-specific immune selection pressures likely underpin this structure.

### 3.5.2 Recombination in BL challenges the model of strict viral clonality

Although recombination is recognised as a source of genetic diversity in various herpesviruses (Koelle et al., 2017; Lassalle et al., 2016; Norberg et al., 2015; Smith et al., 2013), its role in EBV evolution, particularly in tumour-associated viral populations, remains comparatively underexplored. At the molecular level, recombination has been well-characterised in alpha- and beta-herpesviruses, particularly herpes simplex virus 1 (HSV-1), where homologous recombination contributes to genome maintenance, repair, and diversification during latency and lytic replication (Brown, 2014; Wilkinson & Weller, 2004). In contrast, EBV genomes in tumour contexts have often been implicitly considered relatively stable following malignant transformation, with viral persistence assumed to occur largely through clonal expansion of infected cells (Bahri et al., 2022).

However, analysis of 410 EBV genomes across BL, NPC, HL, NKTCL, healthy carriers, and other EBV-associated malignancies revealed a considerably more dynamic evolutionary landscape. Genome-wide linkage disequilibrium (LD) decay analyses showed marked phenotype-specific differences in recombination rates. LD generally decreased with increasing genomic distance across all phenotype groups , although the magnitude and rate of decay varied substantially between disease contexts. BL-derived EBV genomes generally exhibited relatively rapid

LD decay compared with several other phenotypes, suggesting increased historical recombination and reduced long-range haplotype conservation. Subsequent strain-stratified analyses further identified significant differences between Type 1 and Type 2 BL-associated EBV genomes, indicating that EBV strain background contributes substantially to LD structure within BL-associated viral populations. In contrast, HL- and NKTCL-derived EBV genomes showed comparatively slower LD decay, consistent with greater haplotype conservation and reduced recombination. Together, these findings challenge the paradigm of strictly clonal EBV persistence in tumour-associated EBV genomes and instead support the presence of ongoing or historically frequent genomic exchange within at least a subset of EBV-associated malignancies.

Several biological mechanisms could contribute to the recombination patterns observed. Recombination in herpesviruses is generally facilitated during active viral replication and may arise through lytic reactivation, superinfection, or co-infection of host cells with genetically distinct viral strains (Pérez-Losada et al., 2015; Yao et al., 1996). Sporadic lytic activation has previously been documented even in predominantly latent EBV-associated tumours (Abate et al., 2015; Labrecque et al., 1999; Xue et al., 2002), supporting the plausibility of such mechanisms in vivo. Importantly, however, the present analyses do not directly demonstrate the presence of mixed infection or co-infection within individual tumours. Rather, LD decay and phylogenetic incongruence analyses were used to infer recombination frequency and recombinant population structure at the population level. Although intermediate-frequency polymorphisms and strain-

associated heterogeneity may be compatible with the coexistence of multiple viral populations, these features alone cannot definitively establish evidence of mixed infection.

An important limitation in interpreting these findings is that EBV genomes in this study were derived from cfDNA/ctDNA and sequenced using Illumina short-read technology. While short-read sequencing permits robust detection of population-level polymorphism and LD structure, it does not reliably enable reconstruction of complete viral haplotypes or long-range phasing of distant variants across the EBV genome. Consequently, it is not possible to determine definitively whether recombinant genomes arose through co-infection within tumour cells, recombination elsewhere in the same host, or transmission of already-recombinant viral genomes from previous hosts prior to infection. Likewise, although variant allele frequency patterns may suggest the coexistence of multiple viral populations, short-read data alone cannot reliably distinguish mixed infection from intra-host diversification, sequencing artefacts, or subclonal viral evolution without haplotype-resolved analyses. Furthermore, because cfDNA may contain viral DNA fragments originating from multiple anatomical compartments, the observed genomic diversity may not exclusively reflect tumour-associated viral populations. Inclusion of paired tumour and non-tumour samples, such as PBMC-derived EBV genomes, together with long-read sequencing or single-molecule haplotyping approaches, would substantially improve the resolution of within-host viral compartmentalisation, mixed infection dynamics, and the cellular contexts in which recombination occurred.

Phylogenetic incongruence analyses further reinforced the central role of recombination in shaping EBV evolution, consistent with earlier reports (Agwati et al., 2022; Zanella et al., 2019). Masking recombinant regions substantially altered tree topologies among 219 genomes, including BL and healthy control EBV genomes, indicating that recombination acts alongside point mutation in structuring EBV population diversity. Notably, recombinant segments were enriched within genes associated with immune evasion, latency maintenance, and oncogenic signalling, including the EBNA1 and LMP1 regions, both of which are subject to strong immune surveillance and are critical for viral persistence and host transformation (Ressing et al., 2008; Sausen et al., 2023). This enrichment suggests that recombination is unlikely to be entirely random and may instead be selectively retained when novel haplotypic combinations confer adaptive advantages under host immune pressure.

Importantly, these findings can be reconciled with those of Wegner et al. (2019), who reported the apparently paradoxical coexistence of widespread recombination and extensive linkage disequilibrium across EBV genomes. In their study, the persistence of LD was proposed to reflect coevolution among functionally or immunologically linked loci, potentially maintained through host immune selection (Wegner et al., 2019). Within this framework, recombination may occur frequently, yet only particular allele combinations are preferentially retained because they optimise viral fitness under selective pressure. This may explain why loci such as EBNA1 and LMP1,despite functioning as recombination hotspots in the present

analysis, can simultaneously exhibit strong LD patterns. In this model, recombination continuously generates genomic diversity, whereas selection filters and preserves advantageous haplotypic combinations. Methodological differences may additionally contribute to the contrasting observations between studies. Wegner et al. focused primarily on genome-wide and site-level coevolutionary signals inferred from correlated mutations, whereas the present study employed LD decay, phylogenetic incongruence, and gene-level recombination inference analyses to examine recombination dynamics across multiple genomic scales and within a phenotype-aware framework. Moreover, LD structure in EBV is likely influenced not only by recombination and selection but also by population structure, including geographic stratification and host ancestry, and strain-specific evolutionary histories, all of which may preserve haplotypes despite ongoing genetic exchange.

Another important consideration is the unequal sample size across phenotype groups. While BL (n = 141), NPC (n = 113), and healthy carrier (n = 78) cohorts provided relatively robust estimates of LD decay, smaller groups such as HL (n = 12) and, to a lesser extent, NKTCL (n = 34) may yield less stable estimates of recombination dynamics and LD structure. Because LD decay analyses rely on the density and distribution of informative variant pairs, smaller cohorts are more susceptible to sampling variance, reduced haplotype diversity, and inflated uncertainty in decay-rate estimates. To mitigate this, bootstrap resampling and nonparametric statistical approaches were used throughout the analysis. Nevertheless, phenotype-specific comparisons with smaller cohorts should be

interpreted with caution and, where possible, validated on larger independent datasets. In addition, because EBV strain composition differed across some phenotype groups, strain-specific population structure may partially contribute to the observed LD decay differences, as demonstrated by the significant divergence in LD decay patterns between Type 1 and Type 2 BL-associated EBV genomes.

The strain-specific LD decay patterns observed in BL may also have broader implications for understanding EBV-associated oncogenesis. Faster LD decay in Type 2 BL-associated EBV genomes may reflect increased recombination activity, distinct transmission histories, altered selective pressures, or differences in viral population structure compared with Type 1 genomes. Because recombination can reshape immune-associated loci and latency genes, these strain-specific genomic architectures may influence viral persistence, immune evasion, and potentially disease pathogenesis. More broadly, these findings highlight the importance of incorporating EBV strain background into future genomic and epidemiological studies of BL. Improved understanding of strain-specific recombination landscapes may ultimately inform the development of EBV-targeted vaccine strategies and therapeutic approaches, although additional functional studies will be required to establish direct clinical relevance.

Taken together, these findings support a model in which EBV evolution reflects a balance between recombination-driven diversification and immune-mediated constraint. In this framework, recombination appears pervasive and biologically

active, even within tumour-associated viral populations, yet its long-term evolutionary consequences are modulated by selective pressures acting on key functional loci. Consequently, EBV genomes may undergo frequent genomic exchange while maintaining structured linkage disequilibrium, shaped by host immunity, viral fitness, strain background, and population structure. These results provide further evidence that recombination is a central force in EBV evolution and underscore the importance of considering both viral strain heterogeneity and disease context when interpreting EBV genomic diversity.

### 3.5.3 Point mutations and indels concentrate in latency and immune-evasion genes

Complementary mutation enrichment analyses further supported a model of adaptive viral evolution in BL. Nineteen genes were significantly enriched for mutations in BL compared to EBV-positive healthy controls. These included LMP1, EBNA3A, BHLF1, and BLLF1, genes associated with viral transformation capability and immunomodulation. LMP1, in particular, has been consistently identified as one of the most polymorphic regions of the EBV genome across populations (da Costa et al., 2015; Khine et al., 2025; Lei et al., 2015; Liao et al., 2022; Liao et al., 2018).

Importantly, recent large-scale genomic analyses suggest that this diversity is not random. Liao et al. (2022) demonstrated that EBV genomes in African BL cases cluster into a restricted set of LMP1 sequence patterns, indicating selective constraints and possible functional convergence. Consistent with this, a recent large

genomic study identified recurrent point mutations and deletions affecting LMP1 and EBNA genes across multiple EBV-associated diseases (Khine et al., 2025), supporting the idea that shared selective pressures, particularly immune evasion and host adaptation, shape variation in these loci.

Functionally, LMP1's oncogenic roles in host transformation are well established (Dawson et al., 2003; Dirmeier et al., 2003; Khine et al., 2025; Zhang et al., 2023). It activates several oncogenic signalling pathways, including NF-κB, JAK/STAT, and JNK/AP-1, promoting survival, proliferation, metastasis, and immune evasion (Lo et al., 2021). The concentration of polymorphisms within this gene likely reflects adaptive tuning of signalling strength and immune visibility, rather than neutral variation.

The other genes mutated in our study also play vital roles in maintaining the viral genome and driving oncogenesis. EBNA3A is crucial for B-cell immortalisation and represses tumour suppressor pathways (Maruo et al., 2011; Skalska et al., 2013; Wang et al., 2016). Enrichment of mutations in BHLF1 (required for lytic replication complex assembly) and BLLF1 (gp350, mediating host cell entry) suggests selective modulation of both replication and host-interaction pathways. The involvement of BNLF2a, which suppresses MHC class I antigen presentation (Bell et al., 2009), further supports the role of immune-driven selection in shaping EBV genomic variation.

Together, these findings support a co-dominant evolutionary model in which recombination and point mutations synergistically shape EBV diversity in tumour contexts. Rather than remaining evolutionarily static, BL-associated EBV genomes appear to undergo ongoing diversification under host immune pressure and tumour microenvironmental constraints.

### 3.5.4 Mutational signatures identify BL-enriched viral mutagenic process

Using a COSMIC-aligned single-base substitution framework, four EBV-associated mutational signatures (SBS_EBV1–4) were identified across five phenotypes. SBS_EBV1 and SBS_EBV3 showed similarity to COSMIC SBS45 and SBS30, respectively. SBS30 has been linked to defective base excision repair (BER), particularly in the context of NTHL1 deficiency. SBS30 has been linked to defective base excision repair (BER), particularly in the context of NTHL1 deficiency (Drost et al., 2017; Lirussi & Nilsen, 2023), whereas SBS45 is widely considered an oxidative artefact (Yaacov et al., 2024).

The enrichment of SBS_EBV3 in BL remains to be fully explained but may reflect impaired BER and the accumulation of oxidative lesions during viral persistence. EBV latency-associated proteins, including LMP1 and EBNA1, have been reported to suppress BER enzymes, such as OGG1 and NTHL1 (Wang et al., 2020*)*, whereas lytic proteins (e.g., BZLF1 and BGLF5) can downregulate broader DNA repair pathways (Rowe et al., 2007; Wu et al., 2010). These observations together provide

a mechanistic basis by which EBV infection may contribute to BER dysfunction and oxidative DNA damage.

SBS_EBV2 was enriched in healthy EBV carriers and comparatively reduced in BL, suggesting it may represent a latency-associated or non-malignant mutational process. Its weak resemblance to SBS42 (cosine similarity = 0.65) remains hypothesis-generating and warrants further investigation.

In contrast, SBS_EBV4 was significantly enriched in BL and is characterised by substitutions occurring in G-rich or CpG contexts. This pattern may represent a composite mutational process involving oxidative stress, replication-associated errors, and dysregulated host DNA repair. EBV infection has been shown to induce reactive oxygen species (ROS), contributing to DNA damage and genomic instability, partly through modulation of host signalling pathways such as NF-κB (Cerimele et al., 2005; Chen et al., 2016; Gruhne et al., 2009). In parallel, EBV can promote aberrant expression of activation-induced cytidine deaminase (AID) in infected B cells (Kikuchi et al., 2015), leading to off-target cytosine deamination.

Chronic immune stimulation associated with malaria, a known cofactor in endemic BL, has also been linked to sustained germinal centre activity and prolonged AID expression (Ariera et al., 2025; Robbiani et al., 2015; Summerauer et al., 2022). These converging influences raise the possibility that AID-mediated deamination, together with error-prone DNA repair and ROS-associated damage, may contribute to the composite mutational spectrum observed in SBS_EBV4. While AID activity

is primarily directed towards immunoglobulin loci, its deregulation can result in widespread off-target mutagenesis across the host genome, and potentially within replicating viral episomes. However, the extent to which AID directly contributes to EBV genome mutagenesis in vivo remains unclear and a subject for further research.

Together, these findings support a model in which EBV infection and malaria-associated immune activation may synergistically enhance AID activity and oxidative DNA damage, thereby shaping the mutational landscape of BL, while not excluding contributions from additional mutational processes.

## 3.6 Study limitations

Further to the study limitations highlighted under specific sections, several general limitations should be considered when interpreting the study findings. First, public datasets were processed using a unified pipeline, but sequencing artefacts and platform-specific biases may persist despite stringent variant filtering. Second, geographic representation was uneven across phenotypes, and paired host ancestry data were unavailable, limiting causal inference. Stratified analyses within regions and types (and/or meta-analyses across strata) are necessary to strengthen the findings. Third, reference genome bias may affect variant calling. This study used the laboratory strains (NC_007605.1 and NC_009934.1) as reference genomes, both originating from samples outside the study regions. Future analyses could incorporate a geographically specific reference panel to minimise reference bias.

Finally, in the primary study cohort, BL diagnosis was established using a combination of haematopathological evaluation and molecular assays, including next-generation sequencing and fluorescence in situ hybridisation to detect *MYC* translocations. However, comparable diagnostic validation may not be consistently applied across the publicly available datasets analysed, which may include samples classified based on differing criteria or originating from heterogeneous lymphoma cohorts. This could introduce potential misclassification and represents a limitation of the study.

## 3.7 Future direction

The findings of this chapter highlight several key areas for future research. To mention just a few, future studies could explore the following:

- Validate the BL-enriched mutational signatures and recombination signals using independent cohorts, particularly from malaria-endemic regions
- Integrate host genomic and immunogenetic data to disentangle host–virus co-evolution and determine whether specific HLA backgrounds drive selection at EBNA1, LMP1, and other key loci.
- Use single-cell or long-read sequencing approaches to resolve mixed infections and directly characterise intratumour viral heterogeneity and recombination events.
- Evaluate whether BL-enriched variants influence immune recognition, replication kinetics, or oncogenic signalling in functional assays.

- Prospective longitudinal studies in high-risk populations could investigate whether the specific EBV genomic features identified here precede BL development, clarifying their potential role in disease initiation.

## 3.8 Conclusion

This study provides a comprehensive genomic characterisation of EBV in endemic BL, demonstrating that recombination and point mutations jointly shape viral evolution. Elevated recombination rates in BL-associated EBV challenge the traditional view of strict viral clonality and instead suggest dynamic viral adaptation within the tumour microenvironment. The concurrent enrichment of mutations in latency- and immune-evasion genes further supports the presence of ongoing selective pressures acting on the viral genome.

Collectively, these findings advance our understanding of EBV evolution in oncogenic contexts, highlighting the interplay between recombination, mutation, and immune selection. By defining BL-associated viral evolutionary dynamics, this work provides a framework for improved diagnostic biomarkers, refined epidemiological models, and the potential development of vaccine or therapeutic strategies targeting both latent and lytic EBV antigens.

# Chapter 4: Epstein-Barr Virus Transcriptional Landscape in Burkitt Lymphoma

## 4.1 Introduction

Following the results of Chapter 3, which revealed an increased frequency of recombination events, higher mutation rates, and distinct mutational signatures in EBV genomes from BL tumours—indicating a transcriptionally active viral programme rather than a classical "silent" latency programme associated with the disease—I aimed to examine the viral transcriptional activity underlying these genomic features using archival formalin-fixed, paraffin-embedded (FFPE) tumour sections from East African children diagnosed with BL and other EBV-associated lymphomas in the AI-REAL study.

EBV is known to establish three canonical latency programmes: Latency III, II, and I, characterised by distinct patterns of expression of the following viral genes: EBNA-1, EBNA-2, EBNA-3s, EBNA-LP, LMP-1, LMP2A, and LMP-2B (Münz, 2019; Young et al., 2007). Classically, BL is characterised by latency I, where EBNA-1 is exclusively expressed (Münz, 2019). This restricted latency pattern is believed to be critical for maintaining the viral episome during host-cell division while minimising immune recognition, thereby enabling long-term viral persistence within tumour cells (Münz, 2019).

However, emerging evidence challenges this simplistic model. Experimental studies have reported that a subset of BL tumours display transcriptional activity from the Wp promoter, typically associated with the broader Latency III programme (Kelly et al., 2002; Kelly et al., 2013). Interestingly, these tumours frequently harbour EBNA-2 deletions, yet retain expression of most other latent genes, a mechanism thought to mitigate EBNA-2's antagonistic effect on c-MYC-

driven oncogenesis (Kelly et al., 2006). The presence of such “mixed” latency states suggests that some BL cells originate from a Latency III progenitor and subsequently evolve toward a restricted Latency I phenotype under selective pressures for tumour growth and immune evasion (Kelly et al., 2013).

This concept has been further supported by single-cell transcriptomic analyses and studies of other BL variants and nasopharyngeal carcinoma, which reveal striking heterogeneity in EBV gene expression both within and between tumours (Bell et al., 2006; Kelly et al., 2006; Niedobitek et al., 1995). From a pathogenetic perspective, one can question whether EBV is undergoing a fundamental shift in its transcriptional programme within malignant cells, what viral and host factors could drive such a shift, and how these transcriptional changes influence cellular pathways and disease outcomes.

A seminal study from Malawi further underscored the complexity of EBV transcriptional programmes in Burkitt lymphoma (Xue et al., 2002). In addition to identifying non-canonical patterns of viral gene expression, some relapsed tumours displayed markedly reduced or undetectable EBV genomes (Xue et al., 2002), suggesting possible viral loss or immune-mediated clearance during tumour evolution. This observation supports the proposed ‘hit-and-run’ model of EBV-driven oncogenesis, in which transient viral activity initiates transformation but is not required for tumour maintenance (Ferreira et al., 2021; Mundo et al., 2017).

Moreover, the concept of polymicrobial aetiology of BL has gained increasing attention. In Ugandan BL cases, Abate et al. reported that up to 40% of tumours harboured co-infection with other herpesviruses, including cytomegalovirus (CMV) and Kaposi sarcoma herpesvirus (KSHV), alongside EBV (Abate et al., 2015).These tumours also displayed lytic EBV gene expression signatures, suggesting activation of viral replication. Because herpesviruses encode functionally analogous transcription factors such as Rta (immediate-early lytic activator), encoded by BRLF1 in EBV and ORF50 in KSHV (Lan et al., 2004), these findings also raise the possibility of cross-reactivation of viral transcriptional programmes in co-infected cells.

Lytic-phase viral products such as BARFI, BCRF1, BALF1, BHRLF1 and several non-coding RNAs (EBERs, BARTs*)* are known to modulate host immunity, apoptosis and oncogenic signalling (Choi et al., 2013; Kang et al., 2015; Tagawa et al., 2016). Yet whether these viral genes are expressed in a stable, dynamic or stochastic manner within tumour tissue remains an open and critical question for understanding EBV-driven oncogenesis.

To further explore these hypotheses, this study characterised the EBV transcriptional landscape in BL and other EBV-associated lymphomas using archival FFPE tumour samples from East African patients.

## 4.2 Chapter aims and objectives

The specific aims and objectives of this chapter were to:

i) Identify the repertoire of viral genes and transcripts expressed in BL tumours
ii) Compare EBV transcriptional profiles between BL and other EBV-associated B cell malignancies with similar pathogenic mechanisms
iii) Relate these transcriptional findings to the genomic features identified in Chapter 3.4.

Together, these analyses aim to refine our understanding of EBV's role in BL pathogenesis and identify candidate viral transcripts that influence host oncogenic pathways or serve as targets for therapeutic and vaccine development.

## 4.3 Methods

### 4.3.1 Patients and samples

The patients whose tumour biopsies were analysed in this study were identified from the AI-REAL study. FFPE sections were collected and processed according to standard hospital procedures described in Chapter 2, Section 2.2.2.

### 4.3.2 FFPE sectioning

Preparation and processing of FFPE sections for immunohistochemical staining and FISH studies are discussed in Chapter 2, Section 2.3.5.

### 4.3.3 Immunohistochemistry

Immunohistochemistry was performed on the selected markers using Leica Biosystems (Leica Biosystems Inc., Heidelberg, Germany). Details of the commercial monoclonal antibodies used are provided in Chapter 2, Section 2.3.5.

### 4.3.4 RNA extraction

Extraction of FFPE RNA and quality assessment are described in Chapter 2, Section 2.2.3.

### 4.3.5 Preparing RNA libraries

Depletion of rRNA and library preparation procedures are discussed in Chapter 2, Section 2.2.4.

### 4.3.6 Sequencing libraries

The sequencing libraries were multiplexed and loaded onto the Illumina NovaSeq XPlus instrument as described in Chapter 2, Section 2.2.4.3.

### 4.3.7 RNAseq data processing and DE analysis

Paired-end reads were trimmed to remove adapter sequences and low-quality bases using Trimmomatic v0.39 (Bolger et al., 2014). The processed reads were then mapped to a hybrid human–EBV transcriptome index using Kallisto v0.51.1(Bray et al., 2016) to generate transcript-level abundance estimates. These estimates were imported into R using the *tximport* package v1.36.1(Soneson et al., 2016) with length-scaled TPM normalisation. Transcript-level abundances were subsequently summarised to gene-level counts using a tx2gene mapping file.

Differential expression analysis was performed in DESeq2 (v1.42.0), in R (Love et al., 2014). The design formula included both batch and disease group effects:

design ~ batch + covariates + group,

controls set as reference.

The covariates evaluated included age, sex and EBV infection status. Genes with minimal expression (counts ≤1 in most samples) were removed to improve

dispersion estimation and statistical power. Only genes with counts >1 in at least three samples were included in differential expression analysis.

To account for potential batch effects (due to differential treatments of samples, as they were processed in three different batches) and covariate effects, expression values were adjusted using the limma package v3.64.3 (Ritchie et al., 2015). Specifically, batch was modelled as a categorical variable, while age, sex, and EBV status were included as binary covariates. The outcome (group) was retained in the design matrix to preserve the biological variation of interest. The resulting corrected expression matrix was used for downstream visualisation, including heatmaps and gene-level plots.

PCA was conducted on variance-stabilised counts both before and after batch correction to visualise batch and group separation. Batch-associated contrasts were analysed to evaluate systematic differences across batches. Results comparing "batch2 vs batch1" and "batch3 vs batch1" were sourced from DESeq2 and reviewed to confirm minimal batch-driven expression shifts. Effect size correlation and overlap between models with and without batch adjustment were also computed. Batch effects were visualised using ggplot implemented in R.

### 4.3.8 EBV gene expression

To assess viral gene expression, variance-stabilised normalised counts were imported into R (v4.4.1) and merged with sample metadata containing group assignments (BL cases vs. non-BL controls) and EBV positivity status. Only EBV-positive samples were included in this analysis. Differences in normalised

expression between BL cases and EBV-positive controls were assessed for selected EBV genes using the Wilcoxon rank-sum test. Boxplots with individual data points were generated using ggplot2 and annotated with significance levels. For global visualisation, expression values for all EBV genes were transformed into z-scores (row-wise normalisation) and displayed using the ComplexHeatmap package (v2.18.0)(Gu et al., 2016). Genes were grouped by functional class (latent, immediate early, early, or late), and hierarchical clustering was applied to both genes and samples within each group to reveal functional expression patterns.

### 4.3.9 Gene set enrichment analysis

Ranked gene lists based on Wald statistic from DESeq2 were analysed using the fgsea R package (v1.34.2) (Korotkevich et al., 2016) to identify significantly enriched Hallmark gene sets from the Molecular Signatures Database (MSigDB v25.1.1). The Hallmark collection was selected because it provides a concise, non-redundant representation of key biological processes, reducing redundancy and improving the robustness and interpretability of enrichment analyses (Liberzon et al., 2015). A fast gene-set enrichment analysis fgsea algorithm (fgsea v1.30.0 ) was applied to compute the enrichment scores, with p-values adjusted using permutation-based FDR correction (Sergushichev, 2016). Pathways with an FDR–adjusted p-value $< 0.05$ were considered statistically significant. The top five upregulated and downregulated Hallmark pathways were visualised using a dotplot generated with the ggplot function in R.

To explore the expression patterns of genes driving these enrichments, leading-edge genes were extracted from the top five upregulated and downregulated pathways (based on normalised enrichment score, NES). For each pathway, leading-edge genes were combined with pre-ranked gene statistics and then sorted by absolute enrichment score. The top 25 genes per pathway were selected for visualisation. Normalised expression values from DESeq2 were $\log_2$-transformed before heatmap generation. These values were scaled by row (z-score normalisation) to emphasise relative differences across samples. Genes were annotated by pathway membership, and columns were hierarchically clustered by correlation distance, while rows remained in the predefined pathway order to maintain biological interpretability. Heatmaps were created using the R package pheatmap (v 1.0.12) (Kolde, 2019).

### 4.3.10 Evaluating the impact of EBV on host and viral gene expression

Because most BL cases (15 of 16) were EBV-positive, it was not practical to directly assess the impact of EBV infection on host gene expression in the full cohort. Therefore, the analysis was conducted independently of case-control status, using only the control subset, which included 10 samples (3 EBV-negative and 7 EBV-positive). Gene-level counts were imported using tximport and summarised to gene-level abundance. Lowly expressed genes were filtered out (requiring counts >1 in at least 3 samples), and only protein-coding human genes and EBV genes were retained. The effect of EBV on gene expression in controls was assessed using DESeq2 with a likelihood ratio test (LRT), comparing a full model that included

EBV status with a reduced model excluding it. The LRT chi-square statistic and FDR-adjusted p-values (padj) were used to identify genes significantly associated with EBV infection. The proportion of genes significantly affected by EBV, including counts of up- and downregulated genes, was calculated. Results were visualised using a volcano plot with the top significant genes highlighted.

### 4.3.11 Individual contributions

FFPE tumour sections analysed in this chapter were obtained from Ugandan cases (both BL and non-BL) enrolled in the AI-REAL study at St Mary's Hospital, Lacor, Gulu. Our collaborators shipped FFPE samples under a Material Transfer Agreement (MTA Ref: R61914/CN023) and an HTA licence: 12217. Sections were processed, and H&E staining and immunohistochemistry were performed by the THL under framework contract No. 388. FISH analysis was carried out at the Genetics Laboratory, Churchill Hospital. I performed total RNA extraction at the Oxford Molecular Diagnostic Centre. RNA library preparation and sequencing were outsourced to GENEWIZ UK Ltd, a subsidiary of Azenta Life Sciences. Raw data, including paired-end FASTQ files, were processed, and transcript counts were generated by Dr Dimitris Vavoulis. I conducted statistical analyses, including differential gene expression and gene set enrichment analysis.

## 4.4 Results

### 4.4.1 Summary of study samples

FFPE sections were available for twenty-six primary study samples analysed in Chapter 3. These included sixteen BL tumour sections and ten other lymphoma samples, comprising two HL, two DLBCL, one high-grade B-cell lymphoma, one lymphoma of unspecified type, three sections of non-specific lymphadenitis, and two splenic sections with unspecified pathologies. In terms of age, sex distribution, and EBV infection status, no statistically significant differences were observed (Table 4.1). MYC translocation was confirmed by FISH in 15 out of 16 of the BL samples analysed. EBV infection status of the tumour samples was assessed via an indirect approach utilising a targeted sequencing panel of the cell-free DNA as outlined in Section 2.1 of Chapter 2. Accordingly, 22 out of 26 tumour samples (15 BL and 7 non-BL) were EBV positive by sequencing (Table 4.1). All tumours sequenced were EBV type 1-positive, as determined by EBNA-2 genotyping. RNA sequence data (Illumina paired-end FASTQs) were submitted to the NCBI Gene Expression Omnibus (GEO) under the accession number GSE309323.

**Table 4. 1 Clinical characteristics of the cohort samples**

| Characteristics | Overall N = 26[1] | Burkitt Lymphoma N = 16[1] | Non Burkitt Lymphoma N = 10[1] | p-value[2] |
|---|---|---|---|---|
| **Age, years** | 12 (9, 13) | 12 (10, 15) | 11 (6, 13) | 0.4 |
| **Sex** | | | | >0.9 |
| Female | 8 (31%) | 5 (31%) | 3 (30%) | |
| Male | 18 (69%) | 11 (69%) | 7 (70%) | |
| **MYC_IGH** | 15 (58%) | 15 (94%) | 0 (0%) | <0.001 |
| **EBVstatus** | 22 (85%) | 15 (94%) | 7 (70%) | 0.3 |

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

[2]Wilcoxon rank sum test; Fisher's exact test

### 4.4.2 Data quality and sensitivity analysis

To evaluate batch effects, principal component analysis was applied on normalised expression counts before and after batch correction as described in Section 4.3.7. The results revealed that before batch-effect correction, PC1 accounted for 49% of the total variance, largely driven by batch-associated differences (Supplementary Fig.13, Data Appendix). After batch-effect correction, the proportion of variance explained by PC1 decreased to 31%, indicating successful mitigation of batch-driven variation. PC2, which explains approximately 12% of the variance in both analyses, primarily reflects biological differences between experimental groups, suggesting that group-associated expression patterns are preserved following correction. Meanwhile, batch-associated differential expression was minimal, with only a small number of genes (< 0.63% of 18353) showing significant changes after multiple testing correction (adjusted $p < 0.05$). These results indicate that batch-

related variation did not introduce systematic shifts in gene expression, supporting the effectiveness of the normalisation and batch-correction procedures.

### 4.4.3 EBV gene expression in endemic Burkitt lymphoma

Analysis of the viral transcriptome revealed a heterogeneous pattern of EBV gene expression in eBL, contrasting with the strict latency I programme classically associated with this disease (Münz, 2019; Young et al., 2007). In addition to the abundant non-coding RNAs EBER1 and EBER2, EBV-positive BL tumours exhibited detectable expression of both latent and lytic genes across most samples analysed (Figure 4.1), suggesting a broader transcriptional profile than expected under a canonical latency I model.

Among the 22 EBV-positive tumours identified by whole-genome sequencing, 17 (77.3%) expressed at least one latency-associated gene. In addition to EBNA1, 9 of 15 BL tumours (60%) showed detectable expression of other latent genes, including LMP1, LMP2A, EBNA2, the EBNA3 family, and EBNA-LP, suggesting a deviation from a strict latency I programme characterised mainly by EBNA1 expression. This pattern indicates the presence of a non-canonical or expanded latency state in a significant subset of BL cases.

Lytic activity was also observed. BZLF1 transcripts were detected in 9 of 15 BL tumours (60%), consistent with intermittent or partial lytic reactivation. Other early and late lytic genes, such as BCRF1 and BILF2, were also detectable, although generally at lower transcript levels than latent genes. Overall, most tumour samples

showed transcript abundance in the range of >1 to ≤100 normalised counts, indicating low-to-moderate viral gene expression.

In contrast, viral gene expression was generally lower in non-BL samples, except for two aggressive B-cell malignancies (NBL1 and NBL2; diffuse large B-cell lymphoma and lymphoma-unspecified), which showed comparatively elevated EBV transcription (Figure 4.2). Importantly, several viral genes identified as being significantly affected by mutation or recombination (Table 4.2) were also actively transcribed in BL tumours, suggesting potential functional relevance of these genomic alterations.

**Table 4. 2 Significantly mutated and actively transcribed EBV genes in BL tumours**

| Gene | EBV cycle | Log2 enrichment (OR) | FDR-adjusted P value |
|---|---|---|---|
| BHLF1 | early-lytic | 2.94 | 5.35E-120 |
| LMP-1 | latency | 0.64 | 2.59E-24 |
| EBNA-3A | latency | 0.6 | 3.90E-23 |
| BLLF1 | late-lytic | 0.44 | 1.67E-11 |
| BBLF4 | early-lytic | 0.62 | 9.52E-09 |
| BRLF1 | immediate-early | 0.62 | 9.94E-09 |
| BNLF2a | early-lytic | 1.16 | 7.80E-06 |

Collectively, these transcriptomic data indicate that EBV latency in endemic BL does not conform strictly to a discrete latency I state but rather exists along a continuum characterised by mixed latent and lytic gene expression. This dynamic transcriptional landscape may reflect episodic reactivation, reinfection, co-infection, or immunosuppressive host factors that shape viral gene expression during tumour development and progression.

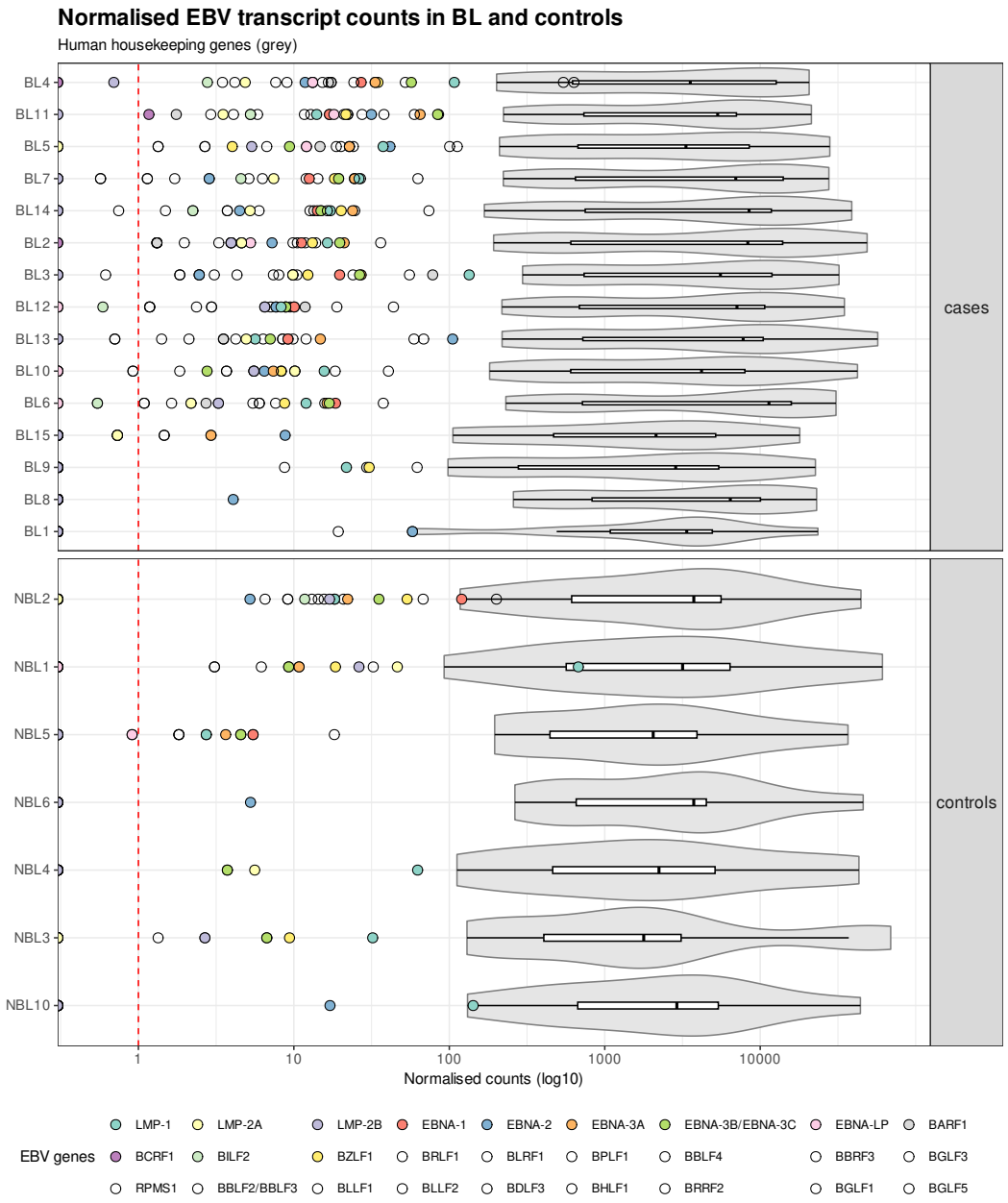



**Figure 4. 1 Relative normalised EBV transcript expression in cases and controls**

The plot shows the distribution of normalised read counts (log10 scale) for EBV genes relative to human housekeeping genes (grey). Normalised counts of the EBV genes are shown for 15 BL tumours (cases) and 7 controls, analysed using FFPE ribosome-depleted total RNA sequencing. The grey violin and boxplot display the expression levels of 10 selected human housekeeping genes (ACTB, GAPDH, HPRT1, B2M, PP1A, RPLP0, RPL13A, TBP, GUSB, and SDHA). The red dashed line marks a normalised count of 1. Latent and immune-evasion genes are coloured, whereas lytic genes remain uncoloured to facilitate comparison.

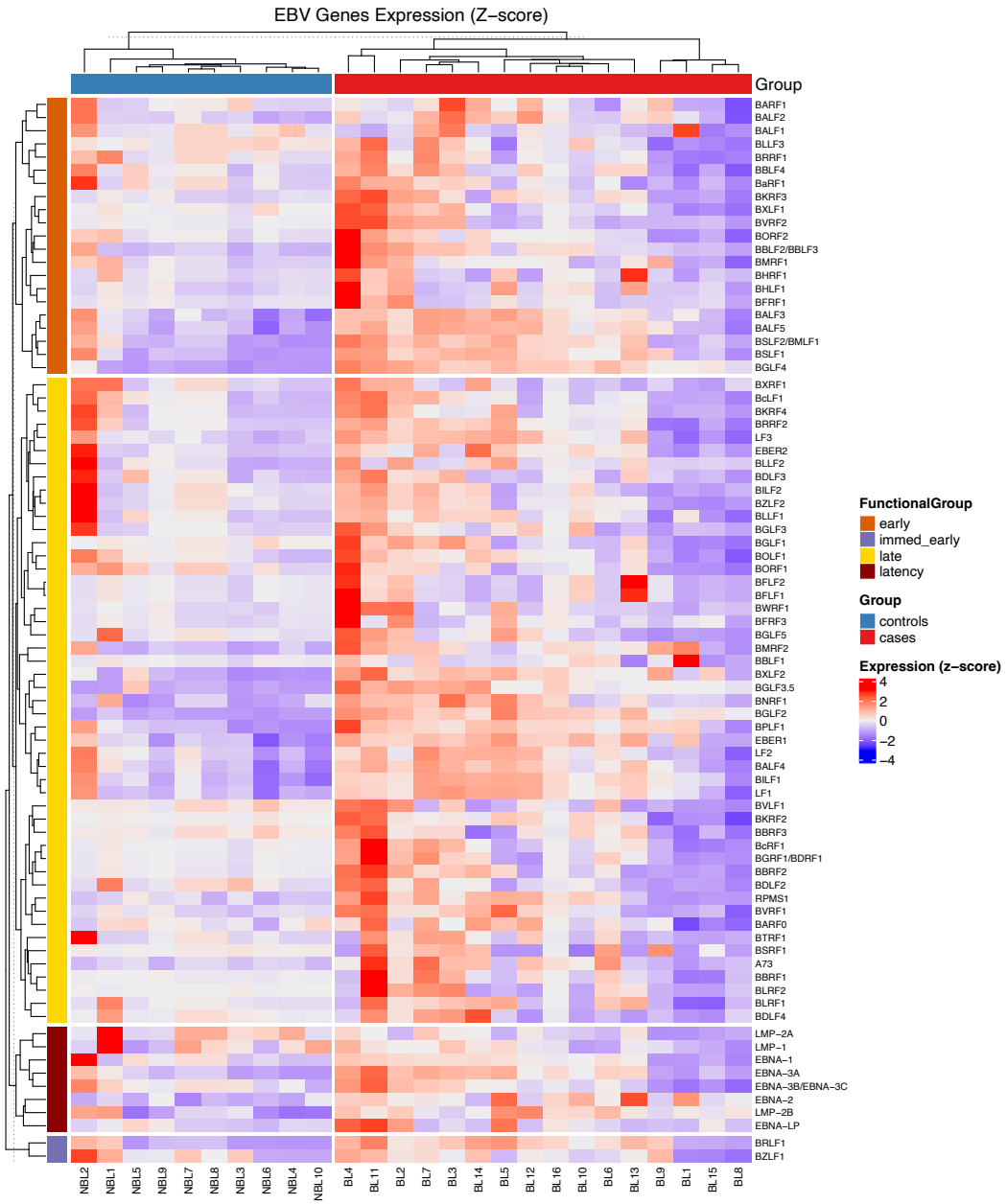



**Figure 4. 2 EBV gene expression in BL compared to non-BL controls (NBL)**

The heatmap shows the scaled, normalised expression levels of viral genes. Rows display genes grouped by functional categories: latency (dark red), immediate early (purple), early (orange), late (yellow), and others (grey). Columns represent individual samples, grouped by category: cases coloured red and controls blue. Cases are BL, while controls include other lymphomas such as HL, DLBCL, and non-specific types. Both rows and columns are hierarchically clustered to show patterns of co-expression.

Next, the expression levels of selected EBV genes were compared between Burkitt lymphoma cases and controls (non-BL malignancies). This analysis aimed to identify viral genes that may influence BL pathogenesis and could serve as potential therapeutic targets. The Wilcoxon rank-sum test was used to compare transcript levels between the groups, with $p < 0.05$ considered statistically significant. Interestingly, the results showed that latency-associated viral genes EBNA-2 ($p < 0.01$) and LMP-2A ($p < 0.05$), and lytic activator genes BRLF1 ($p < 0.05$), were significantly more highly expressed in BL tumours, supporting their role in BL pathogenesis (Figure 4.3).

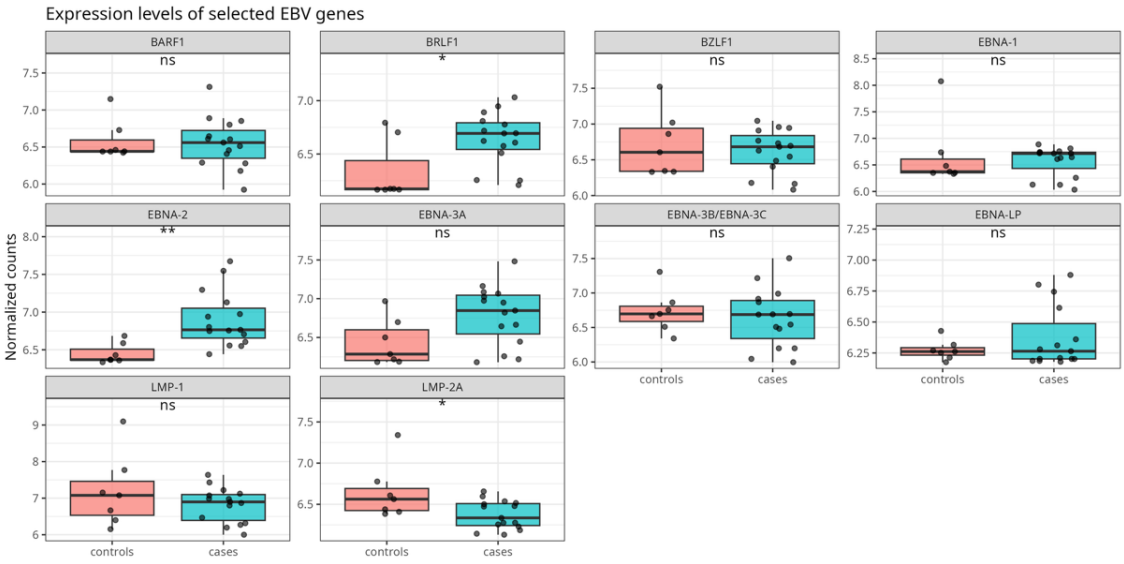



**Figure 4. 3 Expression levels of selected Epstein–Barr virus (EBV) genes in EBV-positive samples**

Boxplots show normalised expression counts for each EBV gene, comparing cases (BL) and controls (NBL). Each point represents an individual sample. Statistical differences were assessed using the Wilcoxon rank-sum test. Significance levels are indicated as $p < 0.05$ (*), $p < 0.01$(**), $p < 0.001$ (***), and non-significant (ns).

To corroborate the transcriptomic results, EBV protein expression was evaluated using commercially available monoclonal antibodies. FFPE sections were processed as described in Section 2.3.5 of the methods. The results were variable (Table 4.3 and micrographs in Figure 4.4): 4/15 (26.7%) sections showed detectable EBNA1 and LMP1. While EBNA2, a viral transcription factor for latent genes, was detectable in 2/15 (13.3%) of the tumour sections examined. Interestingly, BZLF1 was reactive in 2/15 samples, including BL5, which also showed positivity for EBNA1, EBNA2, and LMP1, suggesting a lytic reactivation in these tumours. The staining patterns in 6 sections stained for LMP1 and 3 sections stained for EBNA1 appeared non-specific and were therefore considered indeterminate. LMP2A staining failed completely in all sections examined, despite some positivity in the control section, suggesting poor-quality tissue. Similarly, staining for BARF1, an EBV oncoprotein, was also unsuccessful. This may be explained by the fact that BARF1 is secreted; consequently, only trace amounts of the protein are present in FFPE sections that are poorly preserved for antigen detection (Strockbine et al., 1998).

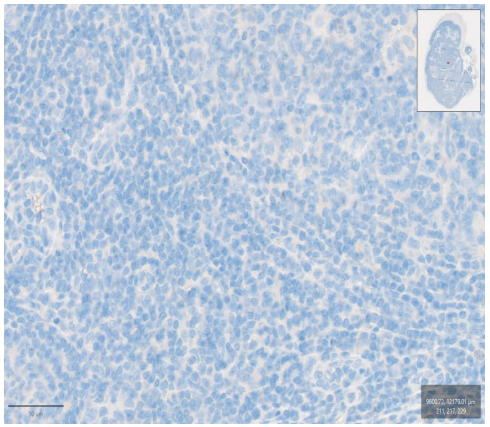
**Table 4. 3 Expression of selected EBV antigens in primary Burkitt lymphoma samples**

| | EBNA1 | | LMP1 | | LMP2A | | BZLF1 | | EBNA2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample | RNA | IHC | RNA | IHC | RNA | IHC | RNA | IHC | RNA | IHC |
| BL1 | - | - | - | - | - | - | - | - | + | - |
| BL2 | + | - | + | - | + | - | + | - | + | - |
| BL3 | + | - | + | - | + | - | + | - | + | - |
| BL4 | + | - | + | - | + | - | + | - | + | - |
| BL5 | + | + | + | + | + | - | + | + | + | + |
| BL6 | + | - | + | + | + | - | + | - | + | - |
| BL7 | + | + | + | - | + | - | + | - | + | - |
| BL8 | - | - | - | - | - | - | - | - | + | - |
| BL9 | - | ǂ | + | - | - | - | + | - | - | - |
| BL10 | + | - | + | - | + | - | + | - | + | - |
| BL11 | + | - | + | + | + | - | + | + | + | - |
| B12 | + | ǂ | + | ǂ | - | - | - | - | + | - |
| BL13 | + | - | + | + | + | - | - | - | + | - |
| BL14 | + | + | + | - | + | - | + | - | + | + |
| BL15 | + | + | - | - | - | - | - | - | + | - |
| Pos n (%) | 12/15 (80) | 4/15 (26.7) | 12/15 (80) | 4/15 (26.7) | 10/15 (66.7) | 0/15 (0) | 10/15 (66.7) | 2/15 (13.3) | 14/15 (93.3) | 2/15 (13.3) |

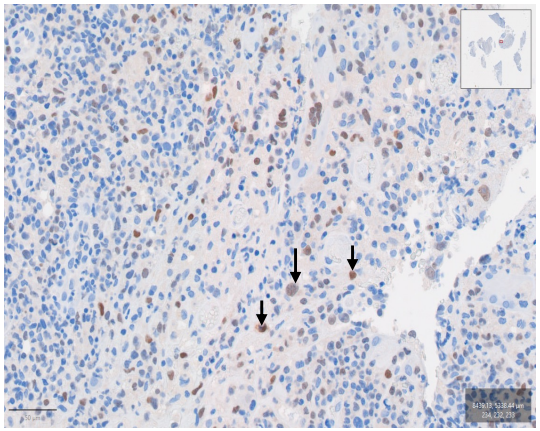
*RNA-tested by FFPE RNA sequencing, IHC- tested by immunohistochemistry, Pos + , Neg - , result not conclusive ǂ.

**Figure 4. 4 Micrographs showing antibody staining for selected EBV antigens**

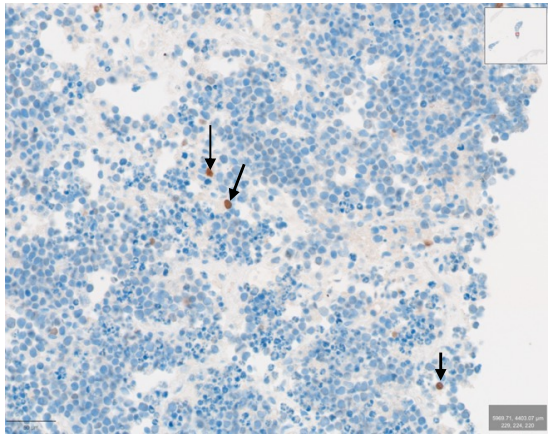
Fifteen BL tumour sections (FFPE) were processed at THL using the Leica BOND™ automated staining system and the standard protocol F30 (Leica Biosystems). Sections were cut approximately 3 μm thick with a microtome and mounted onto positively charged amino-propyl-tri-ethoxy-silane-coated glass slides. Slides were heated at 50 °C prior to staining. Commercially available primary antibodies used included EBNA1 mouse monoclonal antibody (clone IH4; Merck Life Sciences), LMP-1 mouse monoclonal antibody (clone CS1-4; Abcam), LMP2A rat monoclonal antibody (clone 4E11; Santa Cruz), anti-EBNA2 mouse monoclonal antibody (clone PE2; Abcam), and anti-BZLF1 mouse monoclonal antibody (clone BZ1; Santa Cruz). Slides were scanned using a Leica scanner and analysed with QuPath software v0.5.1, an open-source software for digital pathology image analysis. The micrographs display the staining patterns and positivity in selected tumour sections. Arrows indicate positive-staining cells. The scale bar represents 50 μm.



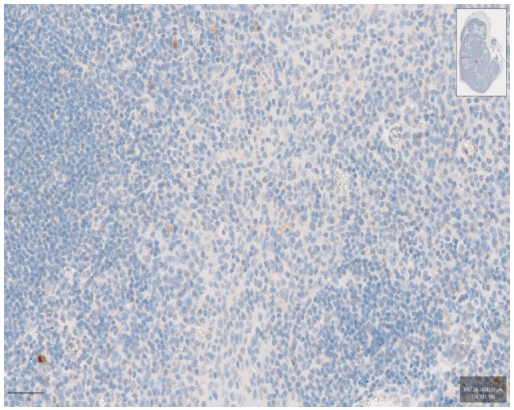
EBNA1-secondary antibody-only (negative control)



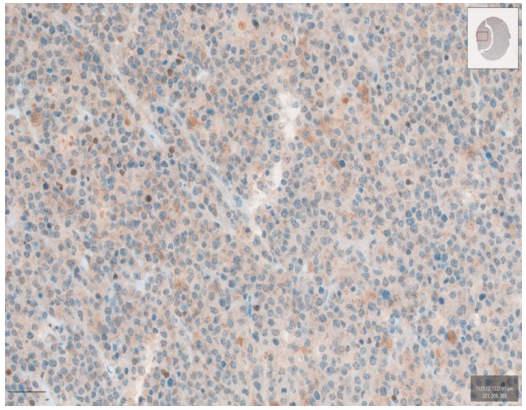
EBNA1 –positive control (nuclear localisation)



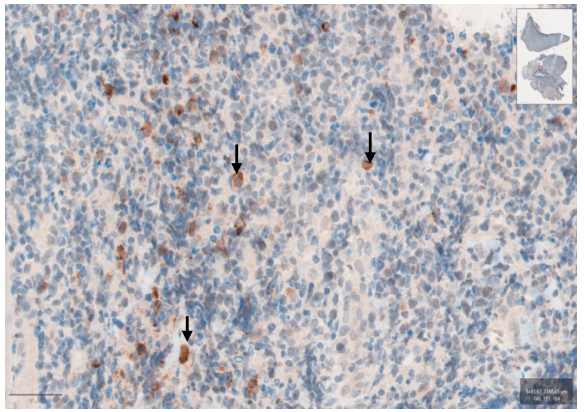
EBNA1-positive tumour (BL7)



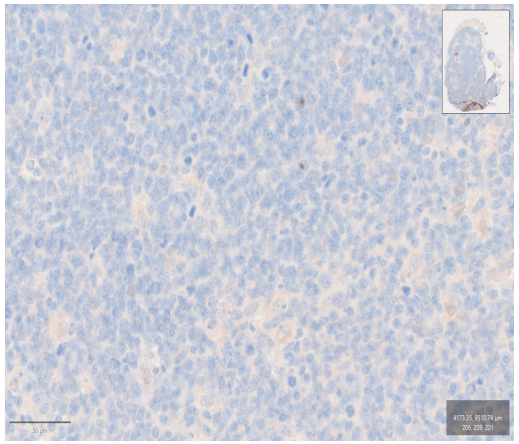
LMP1-secondary antibody-only (negative control)



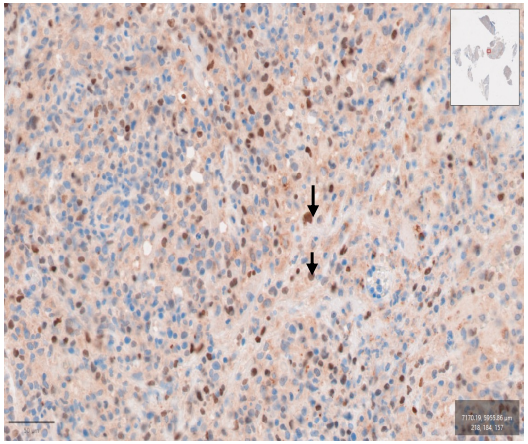
LMP1 –positive control (membranous/cytoplasmic staining)



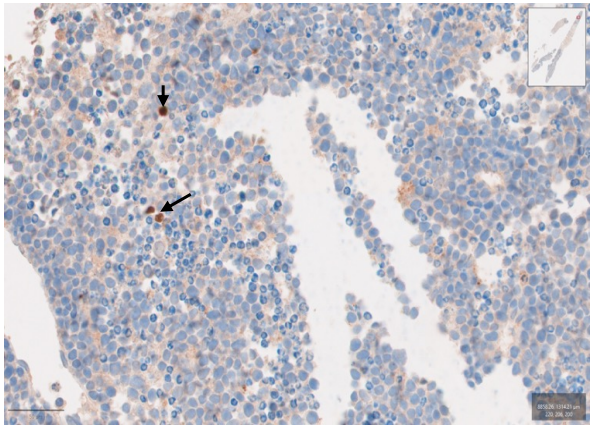
LMP1-positive tumour (BL15)



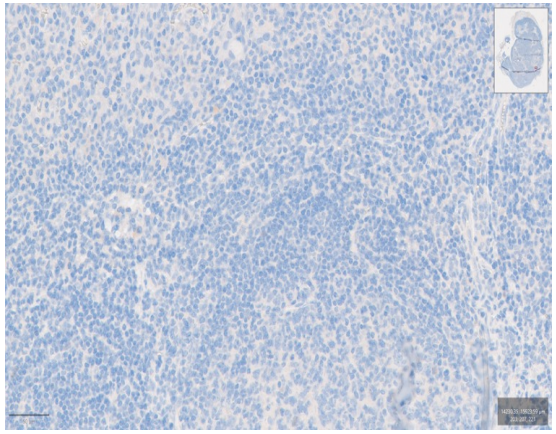
EBNA2-secondary antibody-only (negative control)



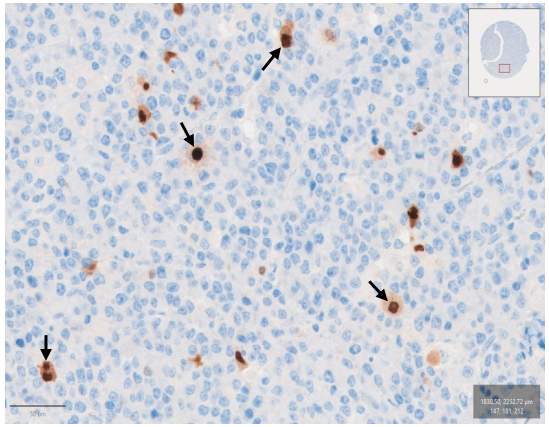
EBNA2 –positive control (nuclear staining )



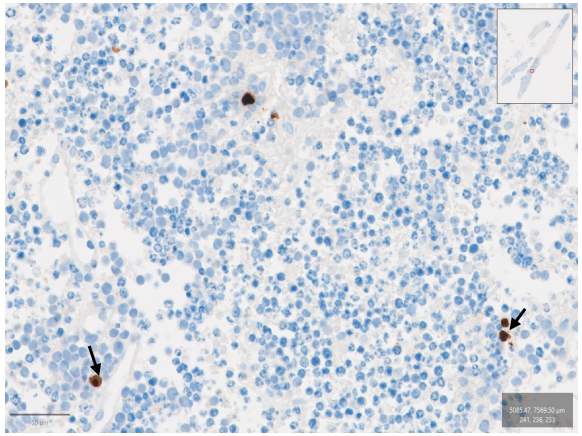
EBNA2 –positive tumour (BL5)



BZLF1-secondary antibody-only (negative control)



BZLF1-positive control (nuclear localisation)



BZLF1-positive tumour (BL5)

### 4.4.4 Integrated pathway analysis reveals coordinated host cellular programmes associated with EBV gene expression

To investigate the host transcriptional programmes associated with EBV gene activity, differential gene expression analysis was performed, followed by gene set enrichment analysis (GSEA) as described in sections 4.3.7 and 4.3.9 of the methods. Sixteen BL and 10 non-BL tumours were analysed in total. The analysis identified 131 genes that were differentially expressed between BL and non-BL samples (Supplementary Table 4, Data Appendix). Leading-edge analysis following GSEA of differentially expressed genes revealed significant enrichment of genes within the Hallmark MYC targets v1 (NES: 3.14, padj = $2.62 \times 10^{-16}$) and v2 ( NES: 2.42, padj = $3.64 \times 10^{-7}$) gene sets, consistent with the well-established MYC-driven biology of Burkitt lymphoma (Basso & Dalla-Favera, 2015). Additionally, GSEA revealed significant enrichment of other proliferative and oncogenic pathways, including E2F targets (NES: 3.14, padj = $5.85 \times 10^{-29}$) and G2M checkpoint (NES: 2.56, padj = $1.34 \times 10^{-15}$), reflecting enhanced cell-cycle progression potentially driven by EBV activity, in agreement with previous findings (Abate et al., 2015). Interestingly, the Allograft Rejection pathway (NES = 2.01, padj = $4.8 \times 10^{-8}$) was among the most significantly enriched pathways in EBV-positive BL samples (Figures 4.5). This pathway includes genes involved in antigen presentation (HLA-DRA, TAP2), T-cell activation (CD3D, CD28, CD40LG), cytokine signalling (IFNG, IL2RA, CXCL9), and cytotoxic effector function (PRF1, GZMA, FASLG) (Carnel et al., 2023; Wang et al., 2008).

Although classically defined in the context of organ transplant rejection, in our B-cell lymphoma dataset, the finding likely reflects heightened cytotoxic T-cell/interferon-driven immune responses associated with EBV infection. The overlap supports a model in which EBV drives host immune recognition and effector pathways analogous to “rejection” signatures observed in transplantation.

Conversely, pathways related to epithelial-mesenchymal transition (EMT; NES: -2.6, padj = $3.87 \times 10^{-19}$), apical junction integrity (NES: -1.99, padj = $9.47 \times 10^{-8}$), myogenesis (NES: -2.1, padj = $2.65 \times 10^{-9}$), coagulation (NES: -2.03, padj = $1.06 \times 10^{-6}$), and UV/DNA damage response (NES: -1.92, padj = $8.83 \times 10^{-6}$) were negatively enriched in BL compared with controls (Figure 4.6). These results indicate a relative suppression of cell differentiation, adhesion, and stress-response programmes in BL, consistent with its highly proliferative and metabolically focused phenotype (López et al., 2022; Ott et al., 2013).

Interestingly, hierarchical clustering of tumour samples based on EBV infection status revealed two distinct subclusters within the BL group (Figure 4.6). One subcluster exhibited strong enrichment of MYC target genes and oncogenic signalling pathways, while the second subcluster (BL1, BL9, and BL15) showed moderate to normal MYC activity. These findings suggest that active EBV gene expression is associated with enhanced proliferative and oncogenic transcriptional programmes in BL.

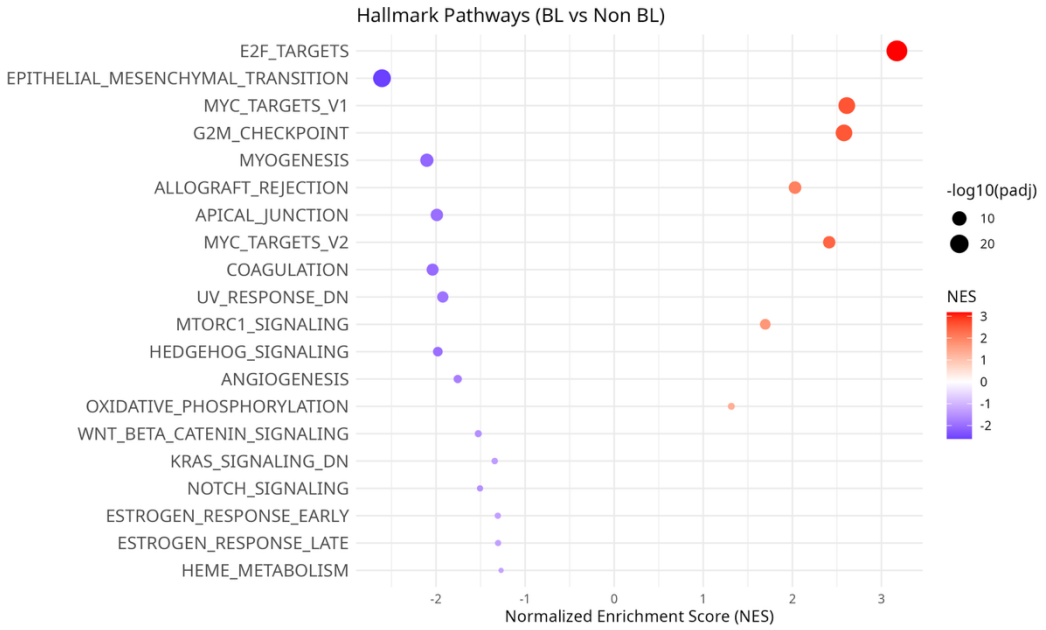
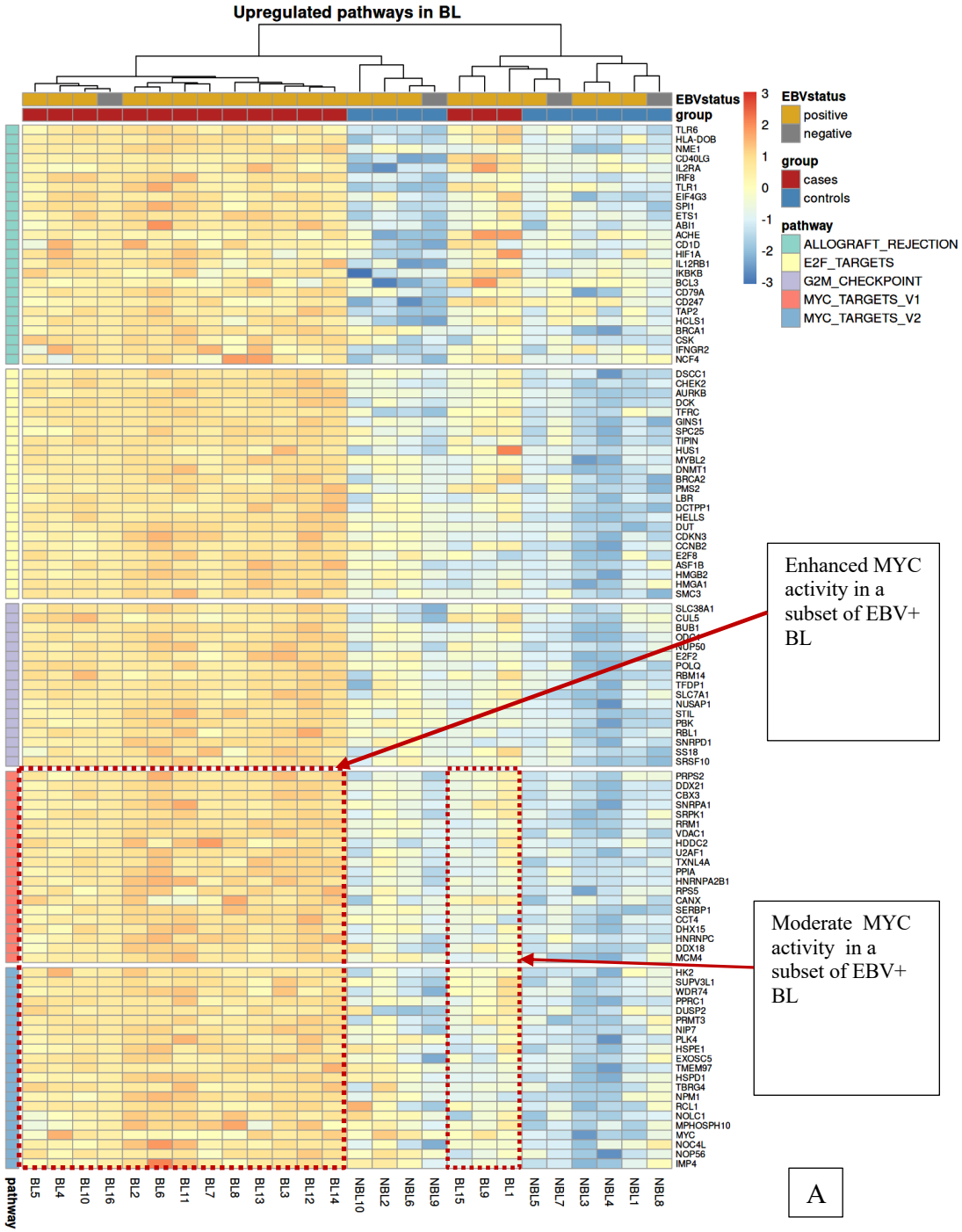



**Figure 4. 5 Dot plot showing twenty of Hallmark pathways analysed and their enrichment scores**

The top 20 Hallmark pathways significantly enriched in BL cases compared to non-BL lymphoma controls are shown on the vertical axis. The horizontal axis represents the normalised enrichment scores, which are obtained by normalising the enrichment scores (ES) against scores from gene set permutations to account for gene set size and correlation. Twenty thousand permutations were performed for each gene set, and FDR-corrected p-values are reported alongside the normalised enrichment scores. Positively enriched pathways are depicted as red dots, while negatively enriched pathways are shown as blue dots. The size of the dots indicates the -log(adjusted p-values); larger dots correspond to smaller p-values, while smaller dots indicate larger p-values.


Upregulated pathways in BL
EBVstatus
group
TLR6
HLA-DOB
NME1
CD40LG
IL2RA
IRF8
TLR1
EIF4G3
SPI1
ETS1
ABI1
ACHE
CD1D
HIF1A
IL12RB1
IKBKB
BCL3
CD79A
CD247
TAP2
HCLS1
BRCA1
CSK
IFNGR2
NCF4
DSCC1
CHEK2
AURKB
DCK
TFRC
GINS1
SPC25
TIPIN
HUS1
MYBL2
DNMT1
BRCA2
PMS2
LBR
DCTPP1
HELLS
DUT
CDKN3
CCNB2
E2F8
ASF1B
HMGB2
HMGA1
SMC3
SLC38A1
CUL5
BUB1
ORC6
NUP50
E2F2
POLQ
RBM14
TFDP1
SLC7A1
NUSAP1
STIL
PBK
RBL1
SNRPD1
SS18
SRSF10
PRPS2
DDX21
CBX3
SNRPA1
SRPK1
RRM1
VDAC1
HDDC2
U2AF1
TXNL4A
PPIA
HNRNPA2B1
RPS5
CANX
SERBP1
CCT4
DHX15
HNRNPC
DDX18
MCM4
HK2
SUPV3L1
WDR74
PPRC1
DUSP2
PRMT3
NIP7
PLK4
HSPE1
EXOSC5
TMEM97
HSPD1
TBRG4
NPM1
RCL1
NOLC1
MPHOSPH10
MYC
NOC4L
NOP56
IMP4
pathway
BL5
BL4
BL10
BL16
BL2
BL6
BL11
BL7
BL8
BL13
BL3
BL12
BL14
NBL10
NBL2
NBL6
NBL9
BL15
BL9
BL1
NBL5
NBL7
NBL3
NBL4
NBL1
NBL8
3
2
1
0
-1
-2
-3
EBVstatus
positive
negative
group
cases
controls
pathway
ALLOGRAFT_REJECTION
E2F_TARGETS
G2M_CHECKPOINT
MYC_TARGETS_V1
MYC_TARGETS_V2
Enhanced MYC activity in a subset of EBV+ BL
Moderate MYC activity in a subset of EBV+ BL
A

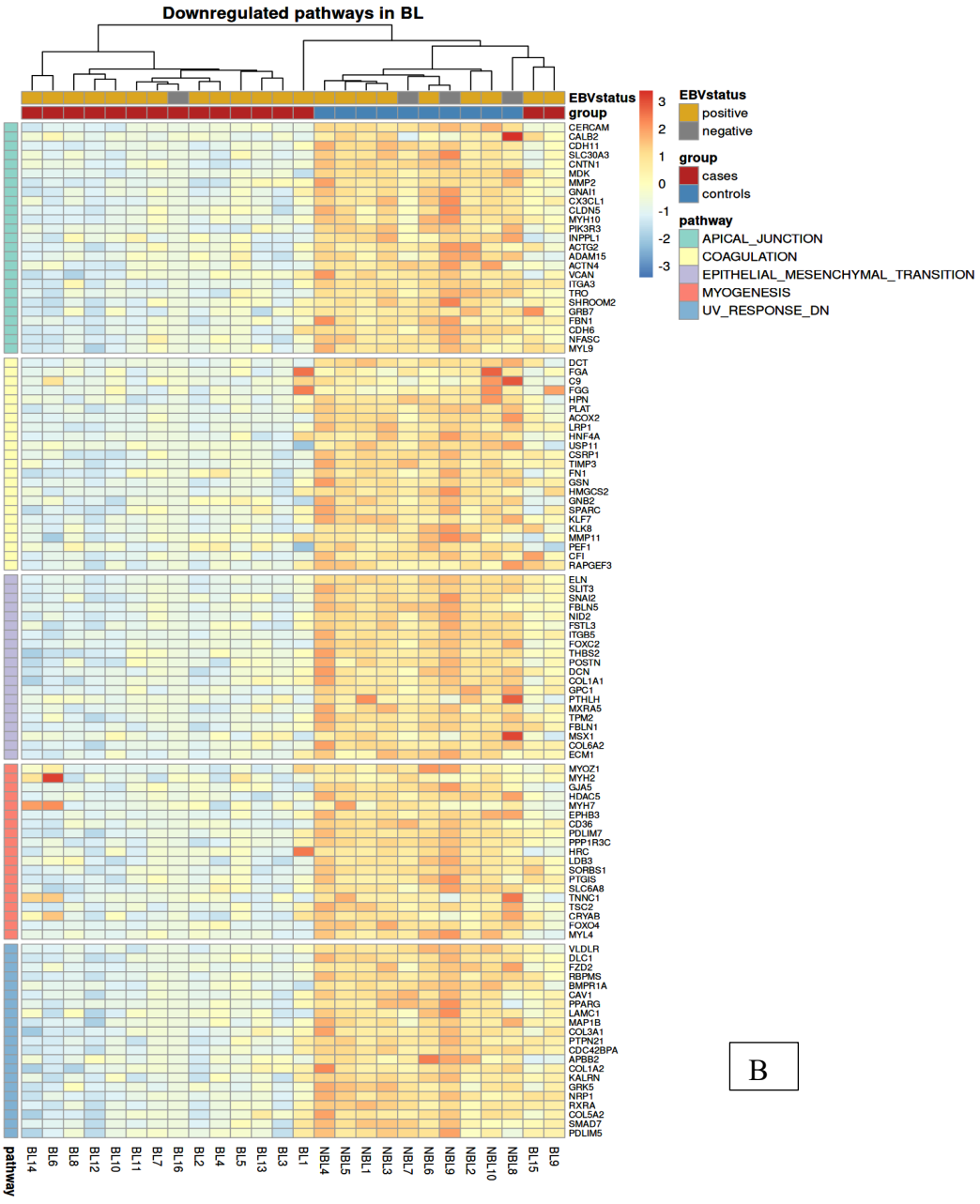



**Figure 4. 6 Heatmap plot of the twenty-five top leading-edge genes from the downregulated pathways**

Heatmaps show the top five up- **(A)** and downregulated **(B)** Hallmark pathways in Burkitt lymphoma (BL), clustered by sample group and EBV status. The 25 leading-edge genes for each pathway were included, and the genes were clustered using Spearman's rank correlation distance. Expression values were scaled across genes to highlight relative patterns of up- and downregulation. The upregulated pathways included: E2F, MYC v1, MYC v2, G2M, and Allograft rejection. The downregulated pathways included epithelial-mesenchymal transition (EMT), coagulation, Apical junction integrity, myogenesis, and UV response. BL cases are shown in red, while controls are displayed in blue. EBV status is indicated by golden-orange for positive tumours and grey for negative tumours.

To corroborate gene set enrichment analysis results and isolate EBV signals that might influence host transcription programmes, differential expression analysis was performed, adjusting for EBV status, on the control dataset comprising 7 EBV-positive and 3 EBV-negative tumour samples, as described in section 4.3.10 of the methods. The LRT chi-square statistic and FDR-adjusted p-values (padj) were used to identify genes significantly associated with EBV infection. Overall, 1217 of 17334 evaluated genes showed differential expression between EBV-positive and EBV-negative controls (Figure 4.7). Of these, 251 were upregulated and 966 downregulated, representing 7.02% of EBV-driven genes after adjusting for sex and age (padj $< 0.05$). While BL and other B-cell lymphomas are biologically distinct, enrichment of pathways related to proliferation, immune modulation, and survival in EBV-positive controls suggests that these transcriptional programmes may reflect conserved EBV-driven processes relevant to B-cell transformation.

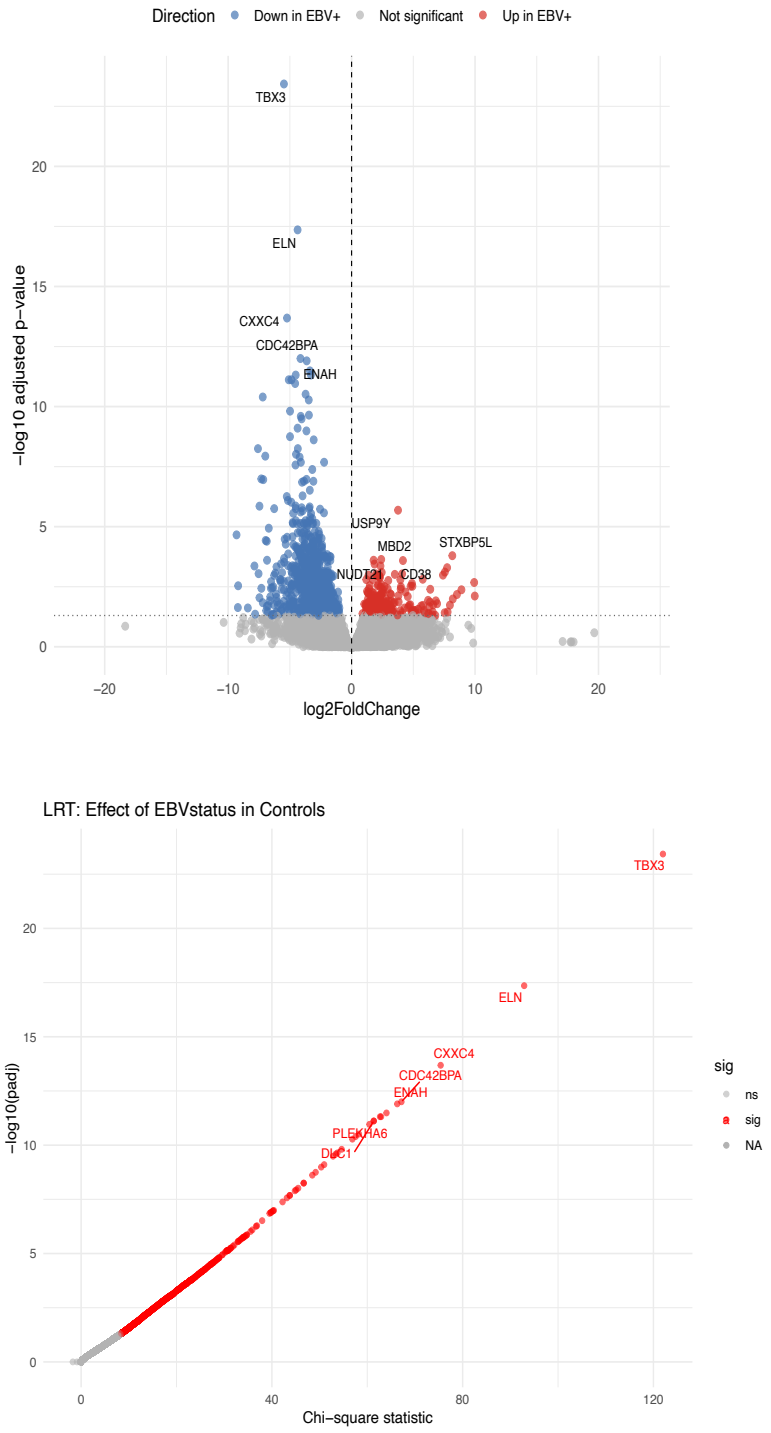



**Figure 4. 7 Volcano plots illustrating the effect of EBV infection on gene expression in control samples**

Volcano plots (**top**) illustrating the effect of EBV infection on gene expression in control samples. Log2 fold-change versus –log10(FDR-adjusted p-value) shows upregulated (red) and downregulated (blue) genes in EBV-positive controls, with the top significant genes labelled. LRT chi-square statistic versus –log10(FDR-adjusted p-value) (**bottom**) highlights genes significantly associated with EBV status. Non-significant genes are shown in grey.

Taken together, pathway enrichment and network-level analyses show a coordinated regulation of host cellular processes related to the cell cycle, tissue structure, and stress responses in the context of EBV gene expression. The key focus is the convergence of MYC- and E2F-related transcriptional programmes, coupled with activation of the G2/M checkpoint and pathways involved in cellular growth and division. Meanwhile, pathways associated with epithelial–mesenchymal transition, apical junction organisation, myogenesis, and tissue remodelling were consistently decreased, along with suppression of DNA-damage and stress-response signatures. Viral latent and lytic factors, including EBNA and LMP family proteins and EBV-encoded miRNAs, are linked to these host pathways, indicating distinct yet overlapping roles in shaping the observed transcriptional landscape. A schematic summary illustrating these pathway interactions and functional effects is provided in Figure 4.8.

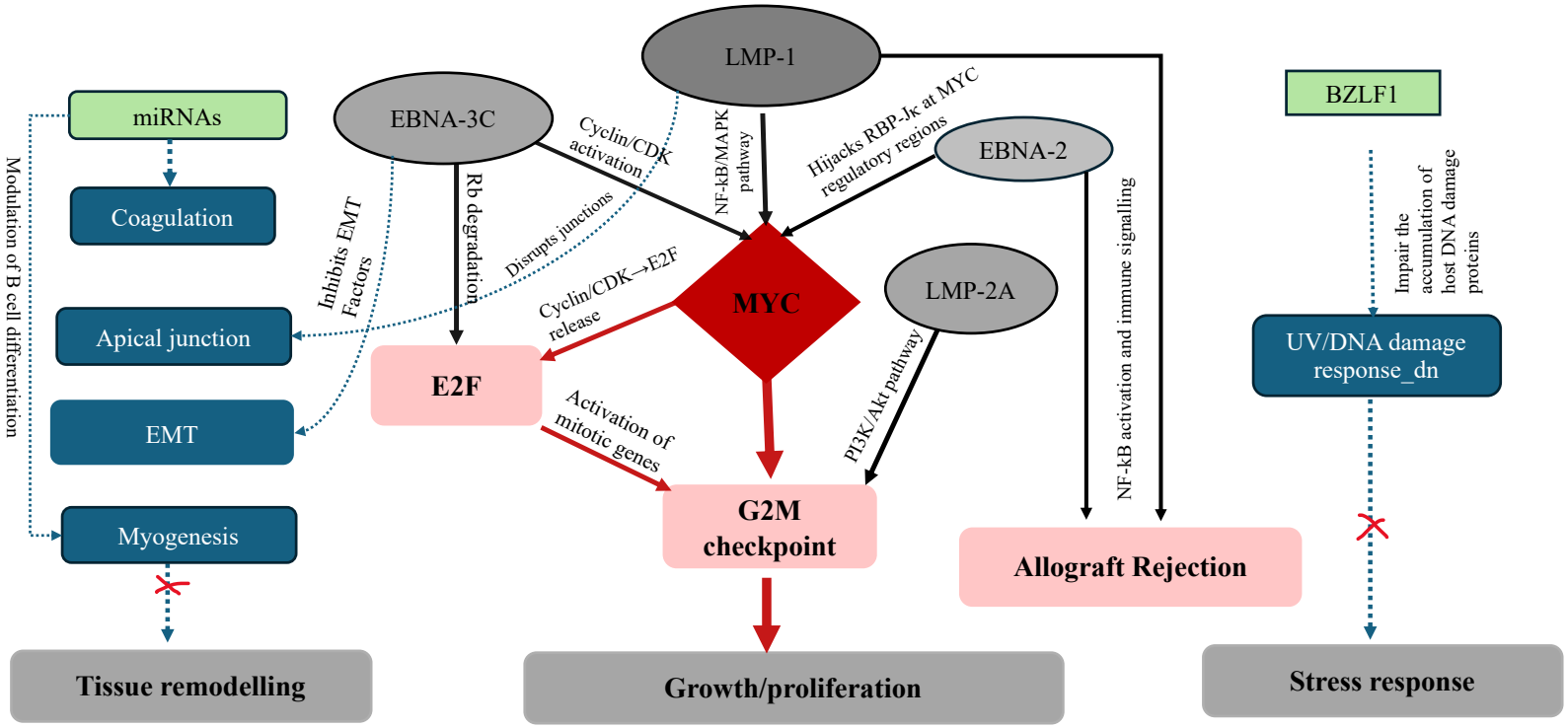



**Figure 4. 8 Integrated model of EBV-driven host pathway modulation**

A schematic summary of the major pathway interactions identified through Gene Set Enrichment and network-level analysis. Pink boxes indicate upregulated pathways, while the diamond-red box represents MYC, the key regulator of the cell cycle. Downregulated pathways are shown with dark blue boxes. EBV latent genes are depicted by grey ellipses, and lytic genes or microRNAs are represented by green boxes. Solid arrows denote activation, whereas dashed arrows indicate pathway inhibition. E2F targets include genes that regulate cell-cycle progression, including S-phase entry and DNA synthesis. MYC targets encompass genes involved in cell cycle progression and proliferation. Genes associated with EMT, which are involved in cell migration and invasion, are downregulated, reflecting BL's focus on proliferation rather than invasion. Figure drawn in Microsoft PowerPoint.

## 4.5 Discussion

This study examined the viral and host transcriptional programmes in EBV-positive tumours collected from Ugandan children diagnosed with Burkitt lymphoma and B-cell malignancies. By relating viral and human transcriptomic profiles to the genomic features identified in Chapter 3, elevated mutation rates, frequent recombination, and distinct mutational signatures, this analysis provides insights into how EBV activity may shape lymphomagenesis in this population.

Traditionally, EBV's latent replication strategy in BL has been viewed as genetically stable, with minimal recombination or genomic diversification (Bahri et al., 2022; Pathmanathan et al., 1995). However, our whole-genome sequencing data challenge this assumption, showing evidence of recombination and mutation patterns inconsistent with strict latency. Instead, the findings support an updated model in which EBV latency may exist on a functional continuum, dynamically influenced by reinfection, co-infections, and tumour-specific host-virus interactions.

### 4.5.1 Heterogeneous EBV gene expression in endemic BL tumours

Analysis of the viral transcriptome revealed marked heterogeneity in EBV activity across the BL series. Over 60% (10/15) of our tumours showed evidence of both latent and lytic transcripts, and immunostaining for EBNA-1, EBNA-2, LMP1, LMP2A and BZLF1 (where material permitted) confirmed protein expression in a

subset of cases. Many monoclonal antibody stains failed or were faint, consistent with antigen degradation and the known limitations of formalin-fixed or long-stored material for some EBV antibodies (Xue et al., 2002). Nonetheless, BZLF1 (Zta), the immediate-early lytic switch protein, was detected by IHC in a minority of cells in two sections, supporting the presence of focal lytic reactivation within at least some tumours.

These observations align with previous reports describing non-canonical EBV transcriptional programmes in endemic BL (Abate et al., 2015; Labrecque et al., 1999; Xue et al., 2002). Xue et al. documented broadly "promiscuous" EBV transcriptional profiles in Malawian eBLs, including frequent detection of both latent and lytic transcripts, while Abate et al, using RNA-seq with extended PCR and IHC validation, similarly demonstrated expression of lytic proteins in a substantial proportion of tumours and highlighted the heterogeneous and focal nature of viral staining within tumour tissue (Abate et al., 2015). Earlier immunocytochemical studies of tumour aspirates likewise reported high frequencies of Zta/BZLF1 positivity and suggested associations between lytic gene activation and favourable chemotherapy responses in some African BL cohorts (Labrecque et al., 1999). Collectively, these findings support the concept that mixed latent lytic states occur in a biologically meaningful subset of eBLs and may contribute to tumour heterogeneity and possibly clinical response.

The biological triggers underlying focal lytic reactivation in BL remain uncertain. Several mechanisms described in the literature may contribute, including malaria-

driven polyclonal B-cell activation, which can promote EBV reactivation (Chêne et al., 2007), concurrent infections within the tumour microenvironment (Abate et al., 2015) and cellular stresses such as hypoxia or reactive oxygen species (McKenzie & El-Guindy, 2015; Murata & Tsurumi, 2014), all of which have been linked experimentally to induction of Zta /Rta expression and activation of the lytic cycle. Abate et al. additionally reported frequent detection of other herpesviruses within the tumour microenvironment, raising the possibility that coinfection-associated inflammatory signalling may facilitate EBV reactivation in vivo.

An important limitation of the present study, however, is that transcriptional profiling was performed using bulk RNA-seq, which provides an averaged transcriptomic signal across all cells within the tumour specimen and therefore cannot assign EBV gene expression to specific cellular populations. Consequently, the observed latency II and lytic transcripts cannot be definitively attributed to the malignant Burkitt cells themselves. Some viral transcripts may instead arise from infiltrating non-malignant lymphocytes, plasma cells, or other EBV-infected bystander cells present within the tumour microenvironment. Likewise, even within the malignant compartment, the detected lytic or non-latency-I transcriptional programmes may reflect expression within only a small subset of tumour cells rather than the dominant tumour population. This distinction is biologically important because the detection of latency II or lytic transcripts does not necessarily imply that these programmes are required for tumour maintenance. Such expression could represent a transcriptionally permissive or “dead-end” state within a minority of cells, rather than a fundamental driver of oncogenesis.

Alternatively, EBV may have contributed primarily during earlier stages of transformation, with persistent viral transcription playing only a limited role in established tumours, consistent with “hit-and-run” or “hit-and-stay” models of viral oncogenesis (Ambinder, 2000; Ferreira et al., 2021). The present data therefore support ongoing EBV transcriptional activity in endemic BL but do not establish whether this activity is essential for tumour survival or progression. These uncertainties also have implications for EBV-directed immunotherapeutic or vaccination strategies targeting lytic or latency II antigens. If such transcripts are predominantly expressed in non-malignant cells or in a small, non-essential tumour subpopulation, then therapies directed against these antigens may have limited impact on the major malignant clone. Conversely, if these transcriptionally active populations contribute to tumour persistence, immune evasion, or reseeding, they may still represent clinically relevant therapeutic targets. Clarifying the cellular source and functional significance of these viral transcriptional states will therefore be critical for the rational design of EBV-targeted therapies.

Future studies using single-cell RNA sequencing or spatial transcriptomics approaches could help overcome these limitations by resolving viral transcription at cellular resolution and distinguishing malignant from non-malignant EBV-infected populations. Integration of such approaches with EBER in situ hybridisation, immunophenotyping, and long-read sequencing would further improve understanding of intra-tumoural EBV heterogeneity and help determine whether focal lytic activation reflects biologically important tumour states or epiphenomena of the tumour microenvironment.

Taken together, the present data confirm and extend prior work showing that EBV transcriptional programmes in endemic BL are not uniformly restricted to latency I and that focal lytic programme activation is sufficiently frequent to be considered in models of BL pathogenesis, tumour heterogeneity, and therapeutic targeting.

### 4.5.2 EBV gene activity appears to enhance proliferative and oncogenic pathways in endemic BL

Differential expression and pathway enrichment analyses revealed that EBV gene activity is tightly linked to host transcriptional programmes regulating proliferation, immunomodulation, and oncogenic features conserved across B cell phenotypes (Shannon-Lowe et al., 2017; Young et al., 2016).

Genes associated with MYC targets (v1 and v2), E2F targets, and the G2M checkpoint were strongly enriched in BL cases, consistent with the well-established MYC-driven biology of BL (Basso & Dalla-Favera, 2015). The E2F pathway, which governs the G1-to-S cell cycle transition, is co-opted by MYC and further amplified through EBV's oncoproteins such as LMP-1, EBNA-2 and EBNA-3C (Liang et al., 2016; Maruo et al., 2011). LMP-1 stabilises c-MYC protein and activates several proliferation and survival pathways, including NF-κB, JAK/STAT, and PI3K/Akt pathways (Cahir-McFarland et al., 2000; Lo et al., 2021; Xiao et al., 2014). EBNA-2 and EBNA-3C are capable of modulating MYC, activity of enhancer RNAs (eRNA) and E2F-mediated transcription to sustain

proliferation while repressing cell cycle checkpoints (Bajaj et al., 2008; Liang et al., 2016; Maruo et al., 2011; Skalska et al., 2010). EBV miRNAs further potentiate these effects by targeting PTEN and activating AKT signalling (Cai et al., 2015).

### 4.5.3 Allograft rejection pathway upregulation may suggest antiviral immunity in endemic BL

Interestingly, alongside proliferative and oncogenic signalling signatures, BLs exhibited an enrichment in the allograft rejection pathway. Although this pathway was initially characterised in the context of transplant rejection, its associated genes, such as HLA-DRA, CD40LG, IFN-γ, and PRF1, form core components of the host antiviral immune response. The significant enrichment of the allograft rejection pathway in EBV-positive BL emphasises a transcriptional programme akin to T cell-mediated antiviral responses. Viral oncoproteins such as LMP-1 and EBNA-2 activate NF-κB and interferon signalling cascades, eliciting a host response that mimics allograft rejection (Šimičić et al., 2024). However, chronic immune pressure in established EBV-associated tumours can also lead to immune editing, including downregulation of HLA and co-stimulatory molecules (Staege et al., 2002; Zuo et al., 2009) (Kelly et al., 2013). This duality of activation and evasion illustrates how EBV simultaneously stimulates and suppresses immune surveillance, enabling persistent infection and tumour progression.

### 4.5.4 Downregulated pathways reflect dedifferentiation and reduced stress response in endemic BL

Conversely, pathways involved in differentiation, tissue remodelling, and stress responses, such as epithelial-mesenchymal transition (EMT), apical junction integrity, myogenesis, UV/DNA damage response, and coagulation, were significantly underrepresented in BL tumours. This repression is consistent with a germinal centre origin of B cells, which naturally lack mesenchymal or epithelial differentiation programmes. MYC-driven transcriptional reprogramming likely reinforces this dedifferentiated state by silencing genes involved in DNA repair, cytoskeletal remodelling, and cellular adhesion. Additionally, suppression of p53 and stress-response signalling, whether via p53 mutations or EBV-mediated inhibition of DNA repair, further supports a model of unchecked proliferation coupled with attenuated damage sensing (Price & Luftig, 2014; Kalla et al., 2010).

### 4.5.5 EBV-dependent subclusters within endemic BL cohort

Hierarchical clustering of tumour transcriptomes based on EBV infection status revealed two potential subclusters within the BL group. One subset displayed strong MYC and cell-cycle activation, while the other showed nearly background levels of EBV transcription and reduced MYC activity. These findings suggest that active viral gene expression may enhance oncogenic signalling, whereas reduced viral transcription could indicate immune editing or clonal tumour evolution. To further assess EBV's role in host transcription and expression, a subset analysis was performed on the non-BL lymphoma subgroup. The analysis indicated that

1217 genes, approximately 7% of the total genes examined, showed significant differences in expression between EBV-positive and EBV-negative tumour samples, further supporting the idea that active EBV gene expression promotes proliferation and oncogenic transcription in B cells.

## 4.6 Study limitations

In addition to the limitations discussed in specific sections , there are further general study limitations to consider when interpreting the results. First, the sample size was modest, and EBV transcript levels were generally lower than those of housekeeping transcripts, reflecting typical FFPE degradation and fragmentation of short viral RNA fragments. Second, RNA-based detection does not equate to translation; therefore, orthogonal protein-level validation, such as IHC/ISH or Western blot, would be necessary to strengthen the findings. Third, differential expression and gene enrichment analyses may be influenced by differences in cellular origins and baseline transcriptional programmes across lymphoma phenotypes; therefore, pathway-level rather than gene-level interpretations were prioritised. When possible, analyses were performed within homogeneous subsets to specifically examine EBV-driven transcriptional changes, thereby reducing confounding from lymphoma subtype. Finally, RNA derived from FFPE tissue is prone to degradation and modification, which may affect transcript quantification. To mitigate these challenges, stringent RNA quality criteria, FFPE-optimised library preparation, and batch correction were employed.

## 4.7 Future direction

The present findings of heterogeneous transcriptional states in a subset of BL tumours are both fascinating and intriguing. Given the often focal and subclonal nature of lytic expression, future studies should combine sensitive RNA approaches (single-cell RNA-Seq / RT-qPCR) with careful protein-level mapping (IHC with defined cutoffs and assessment of fraction of positive cells) to support claims of productive viral replication in BL tumours from endemic regions.

Future studies should also functionally dissect the immunological pathways, particularly the Allograft Rejection signature, to determine whether they represent actionable vulnerabilities in EBV-associated lymphomas. Targeting the crosstalk between viral signalling and host immune pathways may open new therapeutic strategies for BL.

## 4.8 Conclusions

Together, the genomic and transcriptomic data support a dynamic model of EBV latency that exists on a continuum between latent and lytic states, with transcriptional plasticity driving both viral evolution and tumour adaptation.

The enrichment of proliferative and immune activation pathways, particularly the Allograft Rejection signature, suggests ongoing host–virus interactions that may be driving genomic alterations and viral diversification.

Accordingly, the data support the hypothesis that EBV-mediated oncogenesis in BL results from a convergence of two processes: viral gene expression that sustains cell-cycle and survival pathways through MYC, E2F, and PI3K/Akt activation, and the concomitant suppression of differentiation and stress-response genes that enables unchecked proliferation and immune evasion.

# Chapter 5: Immunogenicity of ChAdOx2 and MVA Viral Vectors Expressing EBV Latent, Lytic, and Combined Antigens in Preclinical Models

## 5.1 Introduction

The results presented in Chapter 4 revealed dynamic EBV transcriptional states and protein expression profiles across endemic Burkitt lymphoma and other EBV-associated lymphomas. These findings highlight the co-expression of viral antigens associated with both latent and lytic phases of the EBV life cycle, reinforcing the importance of antigen selection as a central consideration in EBV therapeutic vaccine design. Notably, the identification of viral proteins expressed within tumour contexts suggests an opportunity to target infected tumour cells while simultaneously addressing persistent viral reservoirs, such as latently infected memory B cells and sites of ongoing viral replication.

Building on these transcriptomic and immunohistochemical observations, this chapter evaluated the immunogenicity of selected EBV antigens as candidate components of a therapeutic vaccine strategy intended for use alongside standard Burkitt lymphoma treatment. The main goal was to assess whether immune responses can be elicited against viral proteins identified earlier, thereby informing vaccine designs that may help reduce disease relapse in EBV-associated malignancies.

The therapeutic potential of EBV-targeted vaccination has been most extensively demonstrated in NPC. A phase I clinical trial conducted in the UK evaluated an MVA vector encoding EBNA1 and LMP2 (MVA-EL) and demonstrated that this

approach was safe and capable of boosting EBV-specific T-cell responses, even in patients with advanced disease (Taylor et al., 2014). Vaccine-induced increases in antigen-specific T-cell frequencies were observed in more than half of participants, accompanied by phenotypic and functional diversification of both $CD4^+$ and $CD8^+$ T-cell populations (Taylor et al., 2014). Similar immunogenicity was reported in a cohort of Chinese NPC patients, most of whom showed enhanced vaccine-specific T-cell responses following MVA-EL immunisation (Hui et al., 2013).

Conversely, a phase II trial employing an adenovirus-based vaccine encoding EBNA1 and LMP1/2 delivered via transduced dendritic cells showed limited immunogenicity, with no significant expansion of peripheral antigen-specific T-cell responses, despite modest clinical activity in some patients (Chia et al., 2012). Earlier studies using dendritic cells pulsed with selected human leukocyte antigen (HLA)-restricted LMP2 peptide epitopes reported stronger epitope-specific T-cell responses and tumour regression in NPC patients (Lin et al., 2002). More recently, vaccination with CD137-ligand–expressing dendritic cells pulsed with EBV antigens resulted in enhanced memory T-cell responses and reduced circulating EBV DNA levels in patients with recurrent or metastatic NPC (Nickles et al., 2022).

Alongside viral vector– and dendritic cell–based approaches, mRNA vaccine platforms are increasingly being explored for EBV-associated malignancies. Candidates such as WGc-043 are currently under investigation in EBV-positive relapsed or refractory lymphomas (Clinical Trial Reg. No. NCT06788600).

mRNA-based vaccines offer several advantages, including efficient expression of large or complex antigens, such as EBNA1 and LMP2, which can be challenging to deliver using traditional viral vectors.

Together, these studies underline two key principles in EBV therapeutic vaccine development: the critical importance of antigen selection and the substantial influence of vaccine delivery platforms on the magnitude, quality, and durability of elicited immune responses. Despite these advances, developing an EBV therapeutic vaccine capable of inducing robust and durable $CD8^+$ T-cell responses remains an important unmet goal.

During primary EBV infection, $CD8^+$ T-cell responses primarily target lytic antigens, including immediate-early proteins such as BZLF1 and BRLF1, early antigens such as BCRF1, and late proteins such as the viral envelope glycoprotein gp350 (Cai et al., 2021). Conversely, immune responses to latent antigens, including EBNA1 and the latent membrane proteins LMP1 and LMP2, typically develop later and are vital for long-term immune control of viral persistence (Benninger-Döring et al., 1999), reinforcing their importance as vaccine targets.

Unlike earlier EBV vaccine strategies focusing on individual antigens, this chapter's approach incorporates both latent and lytic EBV antigens. By targeting proteins expressed during viral latency and active replication, this strategy aims to address both persistent viral reservoirs and ongoing viral replication.

To inform antigen selection, consensus EBV sequences derived from primary clinical isolates of BL cases recruited in Uganda and Tanzania within the AI-REAL study were examined (Chapter 3). Common clinical variants and BL expression patterns were used to guide antigen prioritisation. For experimental assessment of immunogenicity, eight latent viral antigens and four lytic antigens were selected as described in Section d) of the Methods. The corresponding $CD8^+$ and $CD4^+$ T cell epitopes, including peptide sequences, protein positions, and HLA restriction elements for the selected antigens, were derived from previously published primary studies of EBV-seropositive human donors (Blake et al., 2000; Brooks et al., 2016; Burrows et al., 1994; Hislop et al., 2007; Meij et al., 2002). These epitopes were originally identified through systematic epitope mapping and functional T cell assays in healthy carriers, as summarised in Hislop et al. (2007) and related primary studies, including Brooks et al. (2016). As described in Figure 4 of Hislop et al. (2007), several of these epitopes were confirmed through systematic screening of Caucasian or Chinese donors.

With the exception of LMP2A, BARF1, and BZLF1, conserved epitope strings rather than full-length protein sequences were incorporated into the vaccine constructs for all selected EBV antigens. LMP2A, BARF1, and BZLF1 were included as full-length sequences because they exhibited greater sequence conservation across the EBV genomes analysed in this study, and because preserving the intact protein structure was considered biologically advantageous for vaccine design. LMP2A and BARF1 are functionally important EBV proteins

implicated in viral persistence, immune modulation, and oncogenic signalling (Fish et al., 2020; Wei et al., 1994), whereas BZLF1 functions as the immediate-early lytic switch protein that initiates the transition from latency to lytic replication (Murata & Tsurumi, 2014). The inclusion of the full-length BZLF1 sequence was therefore intended not only to provide broad epitope presentation but also, potentially, to promote lytic viral gene expression following vector transduction, thereby increasing expression of downstream immunogenic lytic antigens within infected cells. Conceptually, this strategy could broaden immune recognition of EBV-positive tumour cells beyond latency-restricted antigen programmes. However, this assumption depends on efficient transduction of tumour cells by the viral vectors and sufficient induction of lytic gene expression within the malignant compartment in vivo, neither of which was directly assessed in the present study.

Transgene inserts encoding either full-length or concatenated conserved epitope strings were synthesised and cloned into shuttle plasmids, then recombined into ChAdOx2 and MVA viral vectors to generate latent-only, lytic-only, and combined latent–lytic adenoviral constructs, alongside a corresponding MVA vector for heterologous prime–boost regimens. Peptide libraries corresponding to selected EBV antigens or epitopes were used to stimulate mouse peripheral blood mononuclear cells and splenocytes ex vivo to assess antigen-specific T-cell responses elicited by the different vaccine constructs.

## 5.2 Chapter aims and objectives

The overall aim of this chapter was to evaluate antigen-specific T-cell immune responses induced by vaccination with ChAdOx2 vectors encoding latent, lytic, or combined latent–lytic EBV antigens, using homologous and heterologous prime–boost vaccination strategies in a preclinical model.

The specific objectives of this chapter were:

i) To assess the magnitude, breadth, and cytokine profile of EBV-specific $CD4^+$ and $CD8^+$ T-cell responses induced by vaccination with ChAdOx2 EBV constructs.

ii) To determine whether heterologous ChAdOx2/MVA prime-boost enhances antigen-specific T-cell responses compared with homologous prime-boost with ChAdOx2.

## 5.3 Methods

### 5.3.1 Vaccine construct

As previously stated, eight latent and four lytic EBV genes were selected for inclusion in the vaccine construct. Antigen selection was informed by (i) viral gene expression patterns observed in primary endemic BL tumour samples (Chapter 4), and (ii) analysis of common clinical sequence variants identified in EBV genomes derived from these cases (Chapter 3).

Accordingly, sequence data generated in the present study were integrated with previously established EBV immunogenicity datasets to refine antigen design rather than to identify entirely novel epitopes de novo. Previously validated $CD8^+$ and $CD4^+$ T-cell epitopes described by Hislop et al. (2007) and Brooks et al. (2016) were evaluated for sequence conservation across geographically and clinically diverse EBV isolates, including BL-associated genomes analysed in this study. Particular emphasis was placed on selecting epitopes with broad predicted HLA coverage and representation of both $CD8^+$ and $CD4^+$ T-cell responses in order to maximise population-level immunogenic breadth.

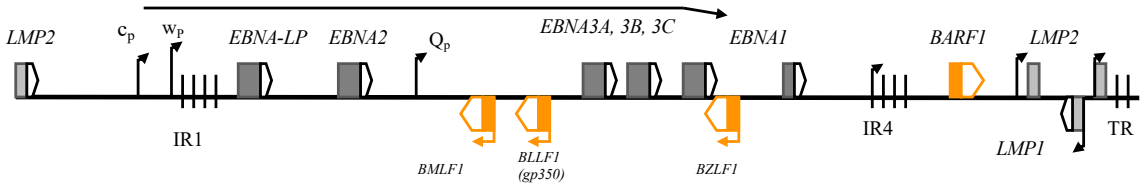
To assess conservation of candidate epitopes, EBV whole-genome sequences were aligned to the B95.8 (NC_007605.1; Type 1) and AG876 (NC_09334.1; Type 2) reference genomes using MAFFT v7.520 (Katoh et al., 2002). Gene-specific exon coordinates were extracted from the aligned fasta files using BED-defined genomic coordinates implemented in Python. Coding DNA sequences (CDS) were translated into protein sequences using Biopython (NCBI translation table 1), and

protein multifasta files were subsequently aligned using MAFFT v7.520 to generate consensus protein sequences for each antigen.

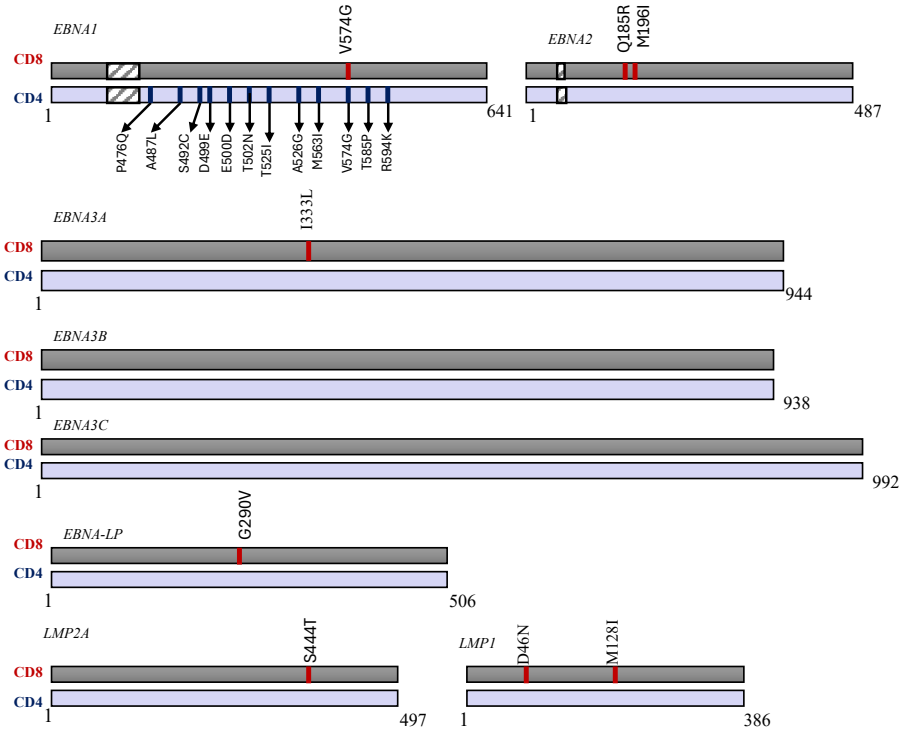
Amino acid substitutions relative to reference sequences were identified from the aligned protein datasets. To assess overall protein conservation, reference and query sequences were analysed using the NCBI Constrained-based Multiple Alignment Tool (COBALT) (Papadopoulos & Agarwala, 2007), and percentage identity, coverage, and alignment length were recorded. Identified amino acid substitutions were mapped onto candidate epitope sequences, and only epitopes fully conserved across BL-derived isolates (i.e., containing no amino acid substitutions) were retained for inclusion in the vaccine constructs (Figure 5.1).

Transgene inserts encoding the selected full-length antigens or conserved epitope strings were synthesised, cloned into shuttle vectors, and prepared for recombination into viral backbones as described in Section 2.4.2.1(d) of the General Methods. Four viral vector constructs were generated: ChAdOx2-Latent, ChAdOx2-Lytic, combined ChAdOx2-Latent-Lytic, and MVA-Latent-Lytic. Peptide panels corresponding to the selected antigens were subsequently used for ex vivo T cell stimulation assays.

A) Selected EBV genes




B) Latent antigen epitopes




C) Lytic antigen epitopes

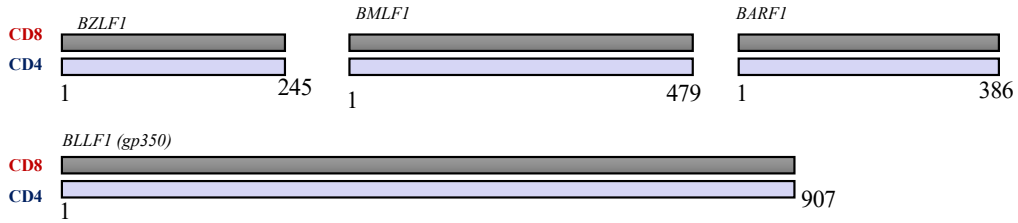



**Figure 5. 1 Amino acid substitutions mapped across CD8$^+$ and CD4$^+$ T cell epitopes for selected EBV antigens**

The BamHI restriction map of the EBV genome (A) illustrates the genomic locations of selected genes. EBV latent genes (grey boxes), lytic genes (orange boxes), and their corresponding promoters are depicted. Arrows indicate the direction of transcription. B & C display the latent and lytic antigens and epitopes identified with single amino acid substitutions. The hatched boxes in EBNA1 and EBNA2 represent the glycine-alanine and proline-rich regions, respectively. The position of amino acid changes within mapped epitopes—red for CD8 T cell epitopes and blue lines for CD4 T cell epitopes—is shown. Figure drawn in Microsoft PowerPoint.

### 5.3.2 Viral vectors

Production and purification of recombinant ChAdOx2 and MVA vectors were performed by the VVCF according to standard operating protocols as stated in Section 2.4.2.1 (d) of Chapter 2.

### 5.3.3 Mice

The choice of mice, housing conditions, and experimental procedures are described in Section 2.4.2.2 of Chapter 2.

### 5.3.4 Vaccinations

Mice received 50 µl intramuscular injections of $10^8$ IU of ChAdOx2 or $10^7$ PFU of MVA into the thigh. Control groups were sham vaccinated with sterile PBS. The prime and boost doses were spaced 4 weeks apart, with blood samples collected between doses to evaluate T-cell responses, as detailed in Section 2.4.2.4 of Chapter 2.

### 5.3.5 Ex vivo peptide stimulation, surface staining, and intracellular cytokine staining

Peripheral blood mononuclear cells (PBMCs) and bulk splenocytes were harvested and processed as described in Methods Section 2.4.2.4. Cells were stimulated using 4 µg/mL of each peptide pool in the presence of DNA1 (20 µg/mL, Roche) and co-stimulatory anti-CD28 (2 µg/mL, Tonbo Biosciences) at 37°C for 5 hours, with brefeldin A (10 µg/mL, BioLegend) added during the last 4 hours to enable

intracellular cytokine accumulation. Surface and intracellular cytokine staining were performed according to the protocol outlined in Section 2.4.2.4 (c & d). Intracellular cytokine antibodies used included anti-TNF-α (clone MP6-XT22), anti-IFN-γ (clone XMG1.2), and anti-IL-2 (clone JES6-5H4), all from BioLegend, UK.

### 5.3.6 Flow acquisition and data processing

Samples were acquired on a BD Fortessa flow cytometer and compensated using single-stain controls as described in Section 2.4.2.5(e). Lymphocytes were gated on forward/side scatter, singlets were selected using forward scatter height vs area, live cells were gated using a viability dye, and $CD8^+$ T cells were gated using $CD3^+$and $CD8^+$ markers (Figure 5.2). For cytokine production, Boolean gating was applied to the three cytokines (IFN-γ, IL-2, TNF-α) to create seven possible positive combinations (three single positives, three double positives, and one triple positive). FlowJo v10.10 (Tree Star, Ashland, Oregon). Frequency values for each Boolean subset were exported from FlowJo as percentages of the parent $CD8^+$ T-cell gate. Data were further processed in Microsoft Excel using Power Query to generate a long-format dataset in which each row represented a single Boolean cytokine combination for an individual sample.

Background correction was performed using matched DMSO-treated samples. For each sample and Boolean cytokine combination, the frequency observed in the DMSO control was subtracted from the corresponding peptide-stimulated condition on a per-sample basis. Negative values resulting from background

subtraction were set to zero. This ensured that background correction accounted for sample-specific baseline cytokine production. Background-subtracted data were analysed using SPICE (v6, NIH). Boolean cytokine combinations were visualised as pie charts to assess the functional composition of antigen-specific $CD8^+$ T-cell responses.

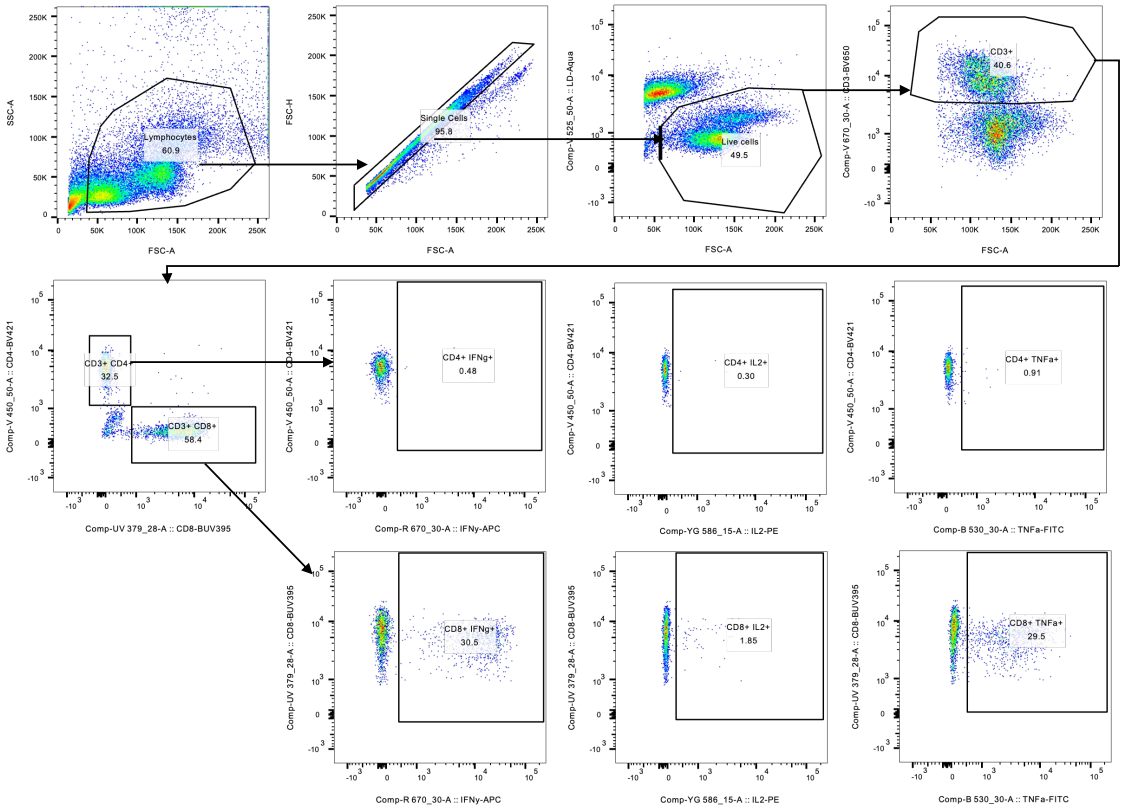



**Figure 5. 2 Flow cytometry gating strategy**

Cells were initially gated based on FSC-A and SSC-A properties to exclude debris. Subsequently, single cells were gated to exclude doublets and aggregates using a FSC-H vs. FSC-A plot. Live single cells were further identified using a LIVE/DEAD Aqua marker. T cells were identified as $CD3^+$ and subsequently differentiated into $CD8^+$ or $CD4^+$ subsets. Antigen-specific T cells responding to stimulation were identified by gating on IFN-γ+, TNF-α+, and IL-2+ events within the $CD8^+$ and $CD4^+$ populations. The gates for cytokine-positive cells were set based on the staining signals observed in unstimulated cells from unvaccinated control animals.

### 5.3.7 Statistical analysis

All statistical analyses were performed using Prism (v10.6.1, GraphPad), R (v4.4.1) (Team, 2016), and associated packages unless otherwise stated. Flow cytometry data were analysed using FlowJo (v10.10 Tree Star, Ashland, Oregon). T-cell polyfunctionality analyses were performed using SPICE (v6, NIH). Microsoft Excel was used for data organisation and preliminary formatting.

Comparisons between vaccinated and unvaccinated groups were conducted using nonparametric tests because the immunological response data were not normally distributed. Group comparisons involving more than two independent groups were performed using the Kruskal–Wallis test followed by Dunn's multiple comparisons test. For repeated-measures comparisons within subjects (e.g., peptide responses within animals), Friedman's test was used with Dunn's post hoc correction for multiple comparisons. Paired comparisons were analysed using the Wilcoxon matched-pairs signed-rank test, and unpaired two-group comparisons (e.g., ChAdOx versus MVA boost responses) were analysed using the Mann–Whitney U test. All tests were two-tailed, and p-values were adjusted for multiple comparisons where applicable.

Peptide immunodominance was assessed using a rank-based analytical framework implemented in R using the rstatix package v0.7.3 (Kassambara, 2019). For statistical comparisons, immunodominance was defined as a consistent within-subject hierarchy of peptide-specific responses across animals, quantified by rank-based concordance statistics. Peptide-specific responses were rank-transformed within each animal such that the highest response received rank 1. The concordance

of peptide rankings across animals was then evaluated using Friedman's test for repeated measures. The strength of concordance was quantified using Kendall's coefficient W, calculated from the Friedman $\chi^2$ statistic.

To assess whether immunodominance hierarchies were influenced by vaccine group (because vaccine compositions varied), a rank-based linear mixed-effects model was fitted using the lmerTest package v3.1.3 (Kuznetsova et al., 2017). Fixed effects were evaluated using Type III analysis of variance with Satterthwaite approximation for degrees of freedom. Significant peptide × vaccine interactions were followed by post hoc pairwise comparisons using estimated marginal means (emmeans) with BH adjustment (Benjamini & Hochberg, 1995). Associations between antigen-specific $CD8^+$ INF-γ response magnitudes were evaluated using Spearman's rank correlation across vaccinated mice (n= 36), with correlation matrices visualised as heatmaps. Statistical significance was defined as adjusted $p < 0.05$ unless otherwise specified.

### 5.3.8 Individual contributions

Mouse experiments were conducted under the supervision of Dr Carol Leung. Vaccine preparations, including vector cloning, were carried out by Dr Leung and Dr Emily Steffke at Ludwig Cancer Institute. Vaccinations were administered with the assistance of Ms Sarah Abdel Malak and Ms Orthi Onupom. I collected peripheral blood samples (tail bleed) and performed intracellular cytokine assays, including flow cytometry acquisitions. Follow-up experiments using different vaccine construct designs were undertaken by Ms Onupom. I performed statistical analyses and interpreted the final dataset.

## 5.4 Results

### 5.4.1 Prime-boost vaccination with EBV vaccine induces antigen-specific $CD8^+$ and $CD4^+$ T cell responses in peripheral blood

To evaluate immunogenicity and determine the most effective vaccine design, mice were primed with a panel of ChAd-based vaccines and boosted with either homologous ChAd or heterologous modified vaccinia Ankara (MVA) (see Figure 5.3 and Table 5.1 for experimental design and vaccination schedules). The ChAd constructs encoded either latent antigens, lytic antigens, or both. Meanwhile, the MVA vector encoded both latent and lytic antigens, and was used in a heterologous prime–boost regimen. Prime and boost vaccinations were administered four weeks apart.

To assess the immunogenicity of the vaccine constructs, antigen-specific T-cell responses were measured in peripheral blood using intracellular cytokine staining two weeks after priming (Day 14) and two weeks after boosting (Day 42). The primary immunogenicity endpoint was the frequency of antigen-specific IFN-γ–producing $CD8^+$ T cells measured by intracellular cytokine staining. Comparisons between vaccine groups were conducted using the Kruskal–Wallis test with Dunn's multiple comparisons against the PBS control. IL-2 and TNF-α responses were assessed as secondary, exploratory endpoints to provide additional insight into the functional breadth of vaccine-induced T-cell responses. To assess response specificity, prime and boost responses in unvaccinated control mice (PBS) were analysed in parallel using paired non-parametric tests (Wilcoxon matched-pairs signed-rank test).

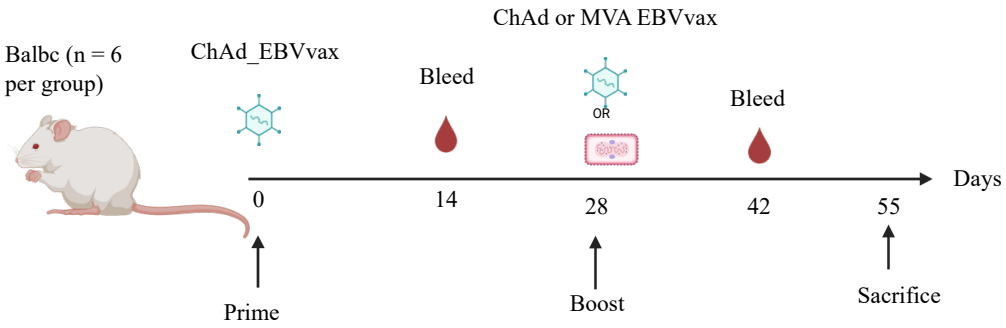



**Figure 5. 3 Experimental design and timeline of prime-boost vaccinations**

Female BALB/c mice, 8-9 weeks old, inbred, received a primary vaccination via intramuscular injection (i.m.) with 10^8 IU of ChAdOx2-EBVvax or a PBS sham. Four weeks after the initial vaccination, mice received a booster dose with either ChAdOx2-EBVvax, 10^7 PFU of MVA-EBVvax, or PBS sham. Vaccinations were administered according to the experimental groupings in Table 5.1. To evaluate the vaccine response, mice were bled 14 days after vaccination. The figure was assembled with BioRender.

**Table 5. 1 Mouse groupings, vaccine regimens and schedules**

| Group | Prime | Boost | Mouse number |
|---|---|---|---|
| 1 | PBS | PBS | 1-6 |
| 2 | ChAdOx2-Latent | ChAdOx2-Latent | 7-12 |
| 3 | ChAdOx2-Lytic | ChAdOx2-Lytic | 13-18 |
| 4 | ChAdOx2-Latent-Lytic | ChAdOx2-Latent-Lytic | 19-24 |
| 5 | | MVA-Latent-Lytic | 25-30 |
| 6 | | No Boost | 31-36 |
| 7 | ChAdOx2-Latent (R) + ChAdOx2-Lytic (L) | ChAdOx2-Latent (R) + ChAdOx2-Lytic (L) | 37-42 |

Following priming, antigen-specific $CD8^+$ IFN-γ responses were detectable against both latent and lytic antigens, with broader responses observed in mice receiving combined latent-lytic antigen constructs (Figure 5.4). Boosting increased the magnitude of circulating antigen-specific responses in vaccinated groups, confirming effective priming and recall of T-cell immunity. While responses in

unvaccinated mice (PBS-treated) remained minimal at both time points, confirming the overall antigen specificity of the vaccine constructs, rare, low-magnitude responses to BARF1 and BZLF1 were detected in isolated mice that received the ChAd-latent construct, particularly within the CD4+ T-cell compartment. Because these antigens were not encoded by the latent-only vaccine, these sporadic responses were interpreted as likely background or cross-reactive rather than evidence of genuine off-target priming. Notably, boosting with the MVA vector significantly enhanced $CD4^+$ IFN-γ responses to the EBNA1 antigen (adjusted $p < 0.001$).

Assessment of secondary cytokine endpoints revealed complementary patterns of vaccine-induced T-cell immunity (Supplementary Fig. 14 & 15, Data Appendix). IL-2 and TNF-α responses were observed across multiple antigens and vaccine regimens, often encompassing a broader distribution of responding cells than IFN-γ alone. In particular, IL-2 and TNF-α responses highlighted increased $CD4^+$ T-cell involvement and more consistent responses to latent antigens, including EBNA1 and LMP2A, following both priming and boosting. While the overall hierarchy of vaccine regimens was broadly concordant with that defined by IFN-γ responses, analysis of IL-2 and TNF-α suggested that IFN-γ measurements preferentially captured effector responses, whereas secondary cytokine profiling revealed additional functional breadth within the vaccine-elicited T-cell response. The 26E constructs did not elicit consistent antigen-specific responses across cytokine readouts and were therefore excluded from subsequent post-boost analyses.

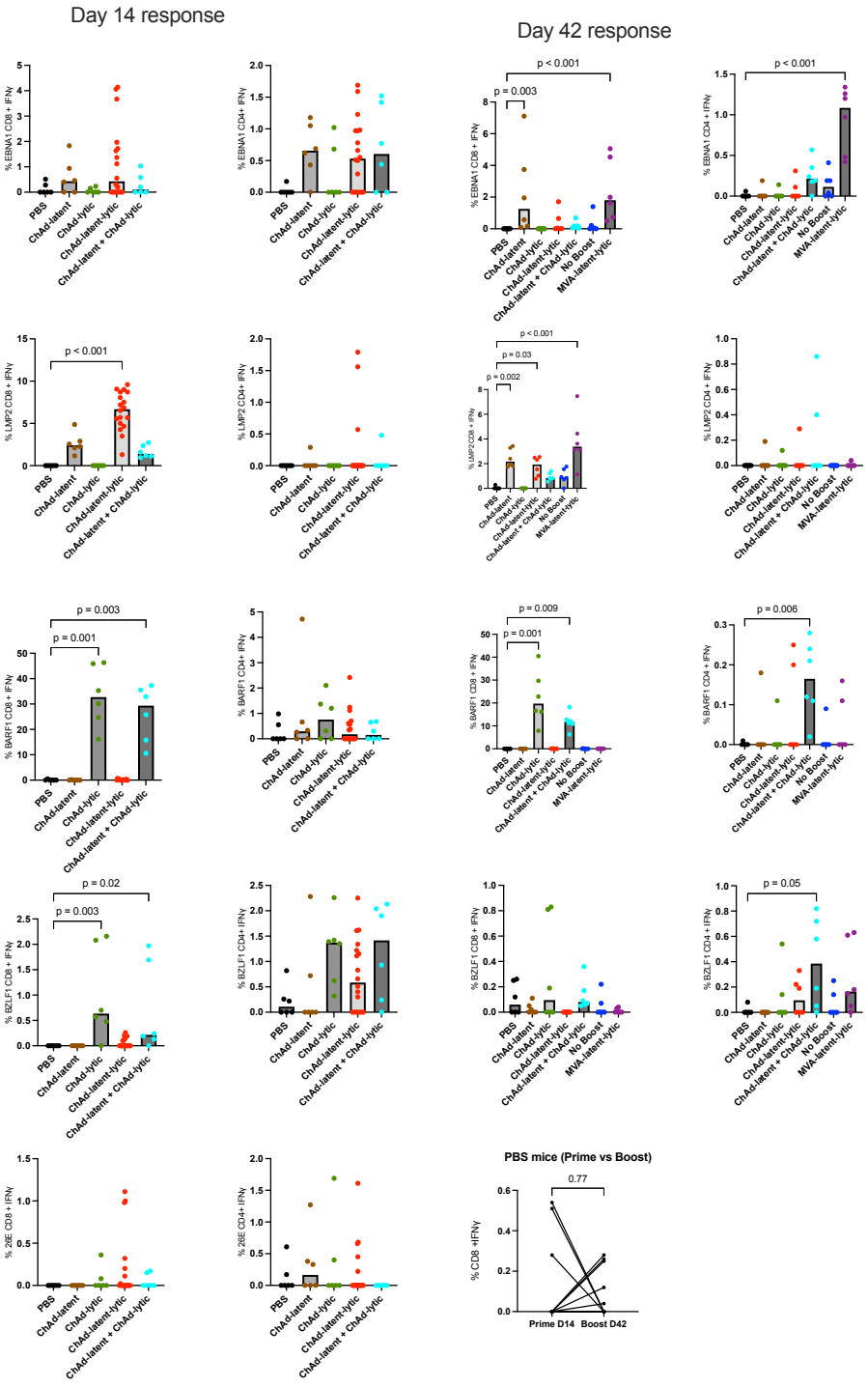



**Figure 5. 4 Antigen-specific IFN-γ T-cell responses following prime and boost vaccination**

Antigen-specific IFN-γ responses were measured by intracellular cytokine staining two weeks after priming (Day 14) and two weeks after booster immunisation (Day 42). Peripheral blood mononuclear cells were stimulated ex vivo with peptide pools corresponding to latent (EBNA1, LMP2A) and lytic (BARF1, BZLF1, 26E) antigens. Each dot represents an individual mouse; horizontal lines indicate medians. Statistical comparisons between vaccine groups were performed using Kruskal–Wallis testing with Dunn's multiple-comparison correction; only significant comparisons versus PBS controls are shown.

### 5.4.2 Peptide immunogenicity and antigen hierarchy of vaccine-induced CD8$^+$ T-cell responses

Responder frequency analysis was performed to assess the proportion of mice mounting antigen-specific CD8$^+$ IFN-γ responses to each peptide pool across vaccination regimens following prime vaccination. Mice were classified as responders if background-subtracted IFN-γ frequencies were ≥0.05% and at least two-fold above DMSO controls. Differences in responder proportions between vaccine regimens were assessed using two-tailed Fisher's exact tests, with effect sizes reported as differences in proportions.

LMP2A-specific responses were detected in all LMP2A-encoding vaccine constructs, with 100% of mice responding regardless of the inclusion of lytic antigens, indicating robust and preserved LMP2A immunogenicity in multi-antigen constructs (Table 5.2). EBNA1-specific CD8$^+$ IFN-γ responses were detected in 3/6 (50%) mice receiving the ChAd Latent vaccine and in 11/18 (61%) mice receiving the latent–lytic construct. Although responder frequency was numerically higher in the latent-lytic group, this difference was not statistically significant (Fisher's exact test, $p = 0.67$; difference in proportions = 11.1 %), indicating that inclusion of lytic antigens did not compromise EBNA1 immunogenicity.

For lytic antigens, responder frequencies were higher in the ChAd Lytic group than in the Latent–lytic group for both BARF1 (6/6 [100%] vs 9/18 [50%]) and BZLF1 (5/6 [83%] vs 7/18 [39%]). Although these differences did not reach statistical significance (Fisher's exact test, $p = 0.052$; difference in proportions = 50%, for

BARF1 and P = 0.155, difference in proportions = 44.4%, for BZLF1), the magnitude and direction of the effects were consistent with enhanced lytic antigen targeting in the lytic-only constructs. PBS-treated mice showed minimal responses across all peptide pools. Together, these data indicate that vaccine-encoded antigens drive antigen-specific $CD8^+$ T-cell responses in a manner consistent with construct composition, with multi-antigen vaccination preserving latent antigen immunogenicity while promoting lytic antigen recognition.

**Table 5. 2 Proportion of mice that generated a positive antigen-specific $CD8^+$ IFNγ response to the EBV peptide pools following prime vaccination**

| Vaccine regimen | 26E pep responders | BARF1 responders | BZLF1 responders | EBNA1 responders | LMP2A responders |
|---|---|---|---|---|---|
| | n (%) | n (%) | n (%) | n (%) | n (%) |
| ChAd Latent | 0/6 (0%) | NA | NA | 3/6 (50%) | 6/6 (100%) |
| ChAd Latent–Lytic | 6/18 (33%) | 9/18 (50%) | 7/18 (39%) | 11/18 (61%) | 18/18 (100%) |
| ChAd Latent (R)+ChAd Lytic (L) | 2/6 (33%) | 6/6 (100%) | 5/6 (83%) | 3/6 (50%) | 6/6 (100%) |
| ChAd Lytic | 2/6 (33%) | 6/6 (100%) | 5/6 (83%) | NA | NA |

* Mice were classified as responders if background-subtracted IFN-γ frequencies were ≥ 0.05% and at least two-fold above DMSO controls. PBS mice typically showed ≤ 1 response across peptide stimulations.

To further evaluate whether some peptides were more likely to be recognised than others, an analysis of the within-mouse antigen-specific immunogenicity across peptide pools was conducted among vaccinated animals. Cytokine responses were background-subtracted using matched DMSO controls, and negative values were set to zero as previously stated.

The analysis showed that, within individual mice, $CD8^+$ and $CD4^+$ T-cell cytokine magnitudes differed significantly across peptide stimulations, as assessed by Friedman's test (global P values shown), indicating the presence of a structured antigen hierarchy (Figure 5.5). Inspection of median responses revealed that BARF1 consistently elicited the strongest $CD8^+$ IFN-γ, IL-2, and TNFα responses, whereas $CD4^+$ responses were more broadly distributed, with lytic antigens, particularly BZLF1, eliciting the highest IL-2 and TNFα responses.

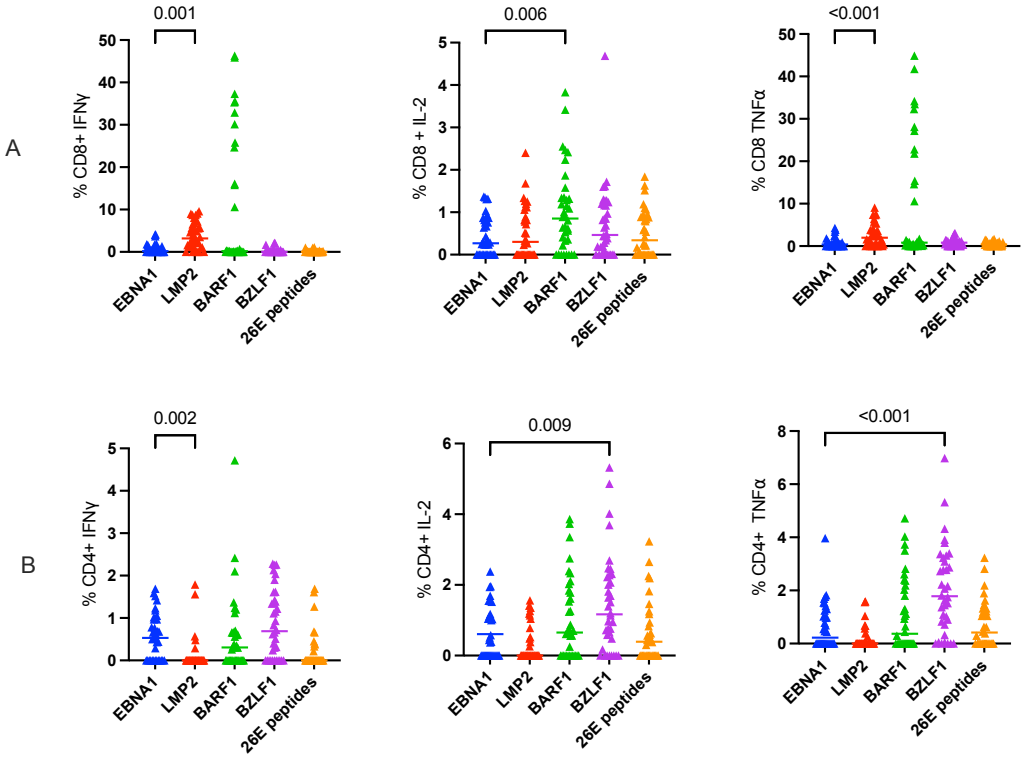



**Figure 5. 5 Peptide immunogenicity profile of vaccine-induced T-cell responses**

The plots show antigen-specific $CD8^+$ (Panel A) and $CD4^+$ (Panel B) T-cell cytokine responses following ex vivo stimulation with peptide pools corresponding to latent (EBNA1, LMP2A) and lytic (BARF1, BZLF1, 26E) antigens. Each dot represents an individual mouse; bars indicate group medians. Differences across peptide pools within the same mice were assessed using Friedman's test (global P values shown). Statistical comparisons across peptides were performed on the combined vaccinated cohort. PBS-treated mice served as negative controls.

Given this peptide hierarchy, the question was whether these observations were random or if vaccine composition influenced immunodominance. To address these, a rank-based analytical framework was employed to analyse $CD8^+$ IFNγ responses to the five peptide pools. Peptide-specific IFNγ responses were rank-transformed within each animal and compared across vaccines using Friedman's test and Kendall's W to assess concordance of rankings, as previously described in Section 5.3.7. The analysis showed that $CD8^+$ IFNγ responses to peptides were highly structured within a vaccine group (Friedman's $p < 0.001$ for all groups), and concordance of peptide rankings across mice was strong to near consensus as assessed by Kendall's W (0. 73 to 0. 88), indicating a non- random dominance pattern within each vaccine group (Table 5.3).

To assess whether immunodominance hierarchies differed between vaccines, a rank-based linear mixed model with fixed effects for peptide and vaccine and a random intercept for mice was employed. Fixed effects were estimated using Satterthwaite's method, and significant peptide × vaccine contrasts were evaluated using estimated marginal means, as detailed in Section 5.3.7.

The results showed that peptide immunogenicity and dominance hierarchy varied by vaccine type ($F_{(12, 180)} = 30.53$, $p < 2 \times 10^{-16}$). Post hoc contrasts revealed that LMP2A was dominant in the Chad Latent- only and Latent- Lytic constructs, whereas BARF1 emerged dominant in Chad Latent (R) + Chad lytic (L) and Chad lytic- only vaccine constructs, confirming vaccine-specific reprogramming of peptide hierarchy (Supplementary Table 5, Data Appendix).

To further characterise the structure of peptide responses beyond rank ordering, the association between peptide response ($CD8^+$ IFNγ) magnitudes was assessed using Spearman's rank correlation (Figure 5.6). Correlation analyses included mice immunised with latent-only (n = 6), lytic-only (n = 6), and combined latent-lytic vaccine constructs (n = 24), with responses grouped according to both vaccine category and peptide stimulation condition. This analysis revealed a strong positive correlation between BARF1 and BZLF1 responses ($\rho = 0.73$, adjusted $p < 0.001$), suggesting coordinated lytic antigen recognition in lytic vaccine constructs. In contrast, LMP 2 responses were inversely associated with BARF 1 ($\rho = -0.53$, adjusted $p = 0.002$) and BZLF1 ($\rho = -0.42$, adjusted $p = 0.022$), indicating reciprocal relationships between latent- and lytic- associated responses. Together, these findings support a model in which vaccines may not only reprogram peptide dominance hierarchies but also differentially structure co-variation among peptide-specific responses.

**Table 5. 3 Rank-based analysis of peptide immunodominance within vaccine groups**

| Vaccine | n | k | Friedman $X^2$ | df | P value | Kendall's W |
|---|---|---|---|---|---|---|
| Chad latent-only | 6 | 5 | 21.10 | 4 | 0.00030 | 0.88 |
| Chad lytic-only | 6 | 5 | 20.92 | 4 | 0.00033 | 0.87 |
| Chad latent(R)+ Chad lytic (L) | 6 | 5 | 18.45 | 4 | 0.00101 | 0.77 |
| Chad latent-lytic | 18 | 5 | 52.24 | 4 | 0.00000 | 0.73 |

* n = number of mice per vaccine group, k = number of peptides ranked per mouse, and df = degrees of freedom.

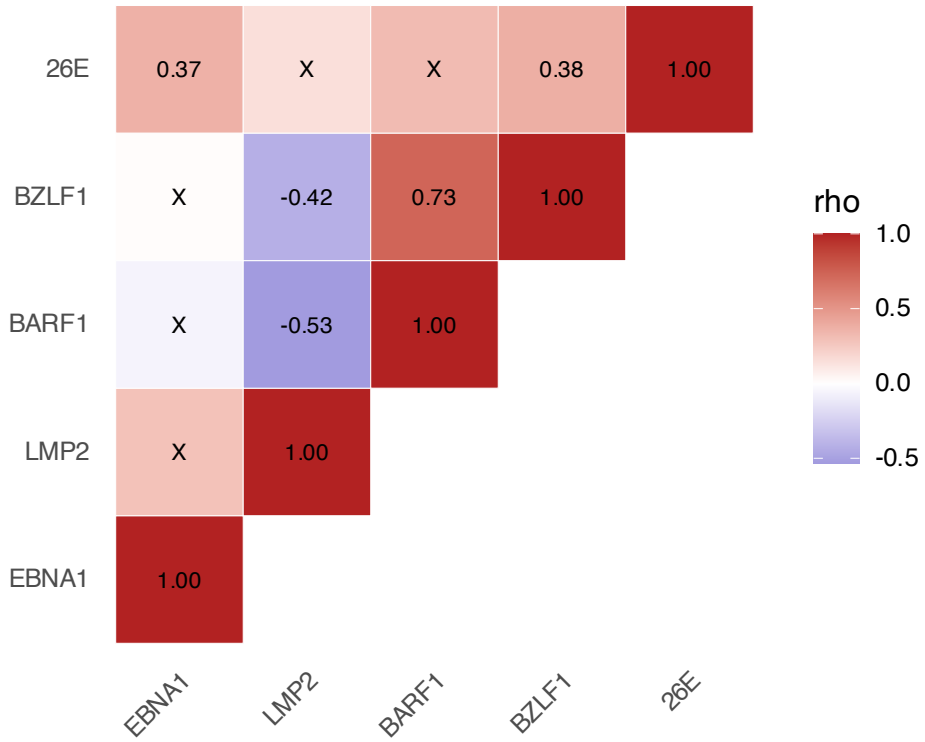



**Figure 5. 6 Spearman correlation matrix of peptide-specific response magnitudes**

Spearman's rank correlation analysis was performed to assess associations between antigen-specific CD8+ INF-γ response magnitudes across vaccinated animals (n = 36). Correlation analyses included mice immunised with latent-only (n = 6), lytic-only (n = 6), and combined latent–lytic vaccine constructs (n = 24), with responses grouped by both vaccine category and peptide stimulation condition. The heatmap displays correlation coefficients (ρ), with colour intensity corresponding to the strength and direction of association. Positive correlations are shown in red and negative correlations in blue. Non-significant correlations after Benjamini–Hochberg correction (adjusted $p > 0.05$) are indicated by "x".

### 5.4.3 Breadth and functional quality of vaccine-induced T-cell responses

While peptide-specific responder frequencies indicate which antigens are recognised across a cohort, they do not capture the diversity of antigen recognition or the functional complexity of responses within individual animals. Since both response diversity and functional complexity are associated with vaccine quality (Wang et al., 2026), response breadth was assessed following prime vaccination using post-prime PBMC samples collected prior to booster immunisation.

In this study, response breadth was defined as the number of EBV peptide pools eliciting a positive $CD8^+$ IFN-γ response within an individual mouse. This operational definition aligns with T-cell immunogenicity studies in other viral systems, where breadth is quantified as the number of distinct peptide regions recognised by antigen-specific T cells following infection or vaccination (Tarke et al., 2024).

Analysis of post-prime response breadth demonstrated significant differences across vaccine groups (Kruskal–Wallis test, $\chi^2 = 16.6$, $p = 0.002$). Post hoc Dunn's multiple comparisons testing demonstrated that vaccinated groups exhibited significantly greater breadth than PBS controls. In particular, the ChAd Latent-Lytic and ChAd Latent + ChAd Lytic groups showed the highest breadth values, with several animals responding to four or five peptide pools (Figure 5.7). In contrast, PBS-treated animals demonstrated minimal breadth (0–1 peptides), while ChAd Latent and ChAd Lytic groups showed intermediate breadth distributions.

Notably, combined vaccine regimens consistently generated broader post-prime peptide recognition profiles than single-program vaccine constructs, suggesting that multi-antigen exposure enhances epitope coverage during the primary vaccine response. Together, these findings indicate that vaccine composition strongly influences the breadth of peptide-specific $CD8^+$ T-cell responses following priming immunisation.

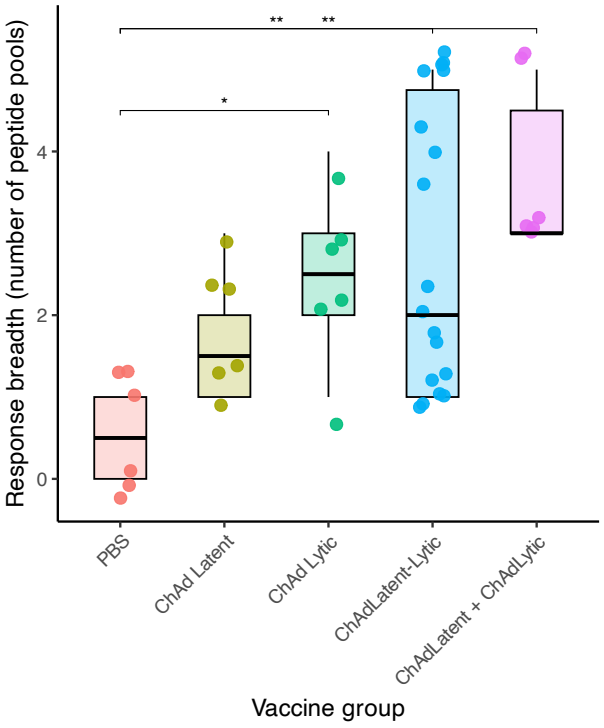



**Figure 5. 7 Breadth of peptide-specific responses across vaccine groups**
Breadth was defined as the number of peptides that elicited detectable responses per mouse (range, 0–5). Each point represents an individual mouse; boxes indicate the interquartile range, with the median shown by the central line. Overall group differences were assessed using the Kruskal–Wallis test followed by Dunn's multiple comparison test with Benjamini–Hochberg adjustment. Significant pairwise comparisons are indicated.

To further characterise response quality, antigen-specific T cells were analysed for co-expression of IFN-γ, IL-2, and TNFα as described in Section 5.3.5. Polyfunctional analysis was performed following background subtraction using matched DMSO controls, and responses were evaluated based on IFN-γ, IL-2, and TNF-α production within the $CD8^+$ T-cell compartment.

Across multiple EBV-derived peptides, $CD8^+$ T-cell responses varied not only in magnitude but also in functional composition between vaccine groups (Figure 5.8). Vaccines encoding latent antigens were linked to a higher proportion of bi- and tri-functional $CD8^+$ T cells, whereas responses induced by lytic antigen constructs were more often dominated by monofunctional subsets or were undetectable after background subtraction. Control conditions showed minimal or no responses across all peptides analysed.

Notably, vaccines encoding multiple antigens (e.g., ChAd-latent-lytic or ChAd-latent + ChAd-lytic) consistently showed a higher proportion of tri-functional T cells, indicating induction of functionally diverse effector responses. Collectively, these findings demonstrate that multi-antigen vaccination not only broadens antigen recognition but also promotes the development of polyfunctional T-cell responses, aligned with improved vaccine-induced T-cell quality.

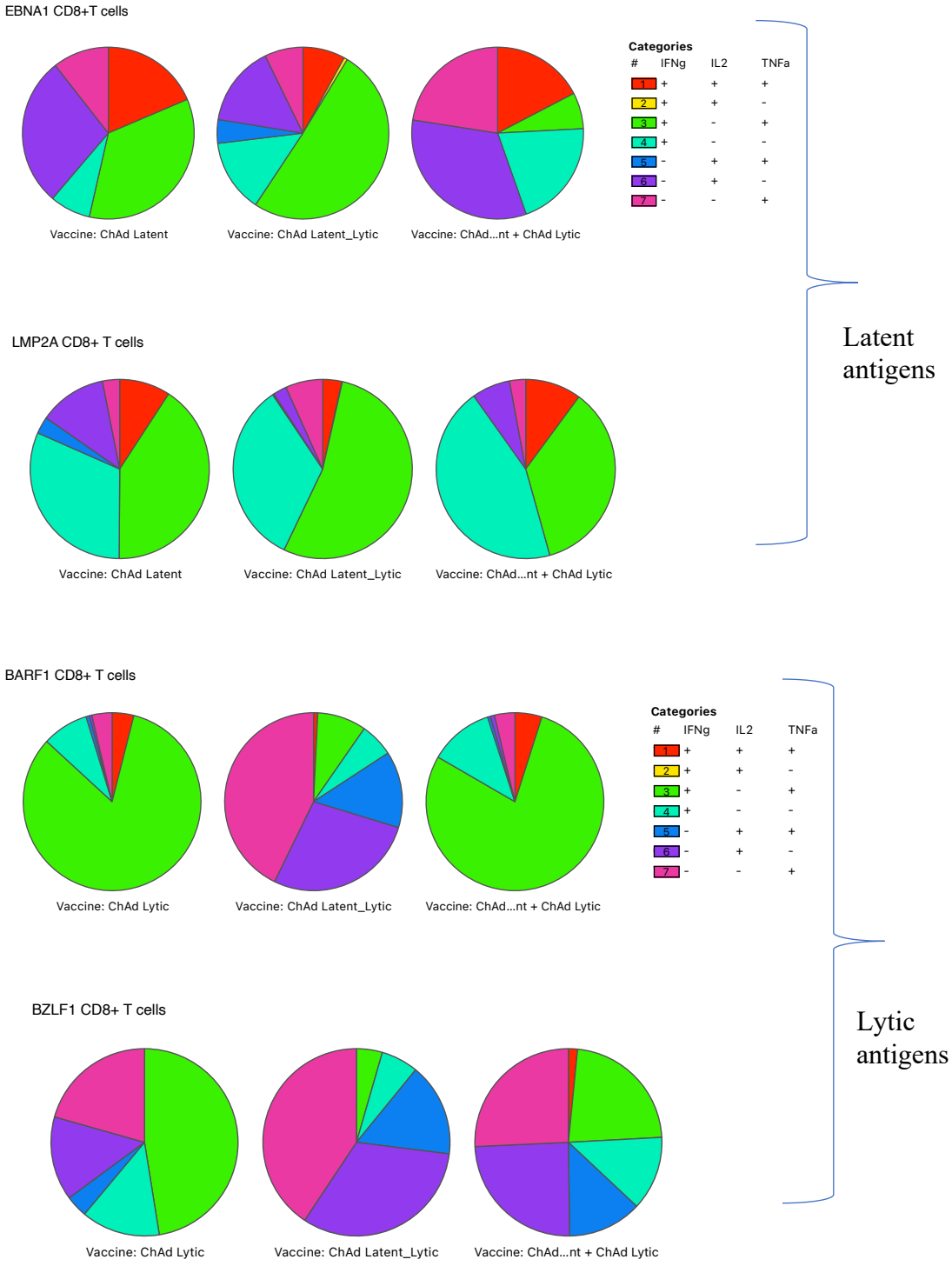



**Figure 5. 8 Polyfunctional CD8⁺ T-cell responses following peptide stimulation**
Pie charts illustrate the mean background-subtracted frequency of IFN-γ–, IL-2–, and TNF-α–producing CD8⁺ T cells for each vaccine group in response to the specified EBV peptides. Colours indicate the number of cytokines produced per cell (monofunctional, bifunctional, or trifunctional), and slice sizes show the absolute contribution of each functional subset. Each pie reflects the mean response across all samples within a vaccine group. Responses to 26 epitope peptides were below the limits of detection and hence not shown here.

### 5.4.4 Heterologous MVA boosting enhances antigen-specific T-cell cytokine responses following ChAd priming

Recent studies have highlighted the benefits of heterologous vaccine regimens and schedules (Jeon et al., 2025; Ramirez-Valdez et al., 2023). In particular, adenoviral priming followed by MVA boosting has been shown to elicit robust peripheral anti-tumour T-cell effector responses in preclinical models (McAuliffe et al., 2021). Enhanced $CD8^+$ T-cell responses in heterologous prime–boost protocols are thought to arise from synergistic activation of innate and adaptive immune pathways, thereby overcoming the immunological limitations associated with homologous boosting (Hofer et al., 2021; Jeon et al., 2025; Rühl et al., 2019).

To assess the impact of boosting on antigen-specific T-cell responses, mice primed with a ChAd latent-lytic vaccine were boosted with either homologous ChAd or heterologous MVA, and cytokine production by $CD8^+$ and $CD4^+$ T cells was measured after peptide stimulation.

MVA-boosted animals exhibited higher frequencies of IFN-γ-, IL-2-, and TNF-α-producing $CD8^+$ T cells following EBNA1 stimulation (Mann–Whitney U test; $p < 0.05$ for all comparisons) (Figure 5.9). A similar, albeit less pronounced, trend was observed for LMP2A-specific $CD8^+$ T-cell responses, with increased cytokine production following MVA boosting. In contrast, $CD8^+$ T-cell responses to BARF1 and BZLF1 peptides were consistently low and clustered near the assay limit of detection across all cytokines, with no discernible differences between ChAd- and MVA-boosted groups (Supplementary Fig.16, Data Appendix).

Within the $CD4^+$ T-cell compartment, boosting effects were more selective. EBNA1-specific $CD4^+$ T-cell responses were significantly enhanced after MVA boosting, with increased production of IFN-γ, IL-2, and TNF-α ($p < 0.01$ for all cytokines). In contrast, responses to the other EBV antigens remained near background levels regardless of the boost vector, indicating limited immunogenicity for subsequent booster doses.

Together, these data demonstrate that heterologous MVA boosting following ChAd priming selectively enhances the magnitude of antigen-specific T-cell cytokine responses, with the most robust effects observed for EBNA1 across both $CD8^+$ and $CD4^+$ T-cell lineages. Importantly, the boost-dependent enhancement was antigen-specific rather than global, as responses to BARF1 and BZLF1 remained minimal in both boost groups.

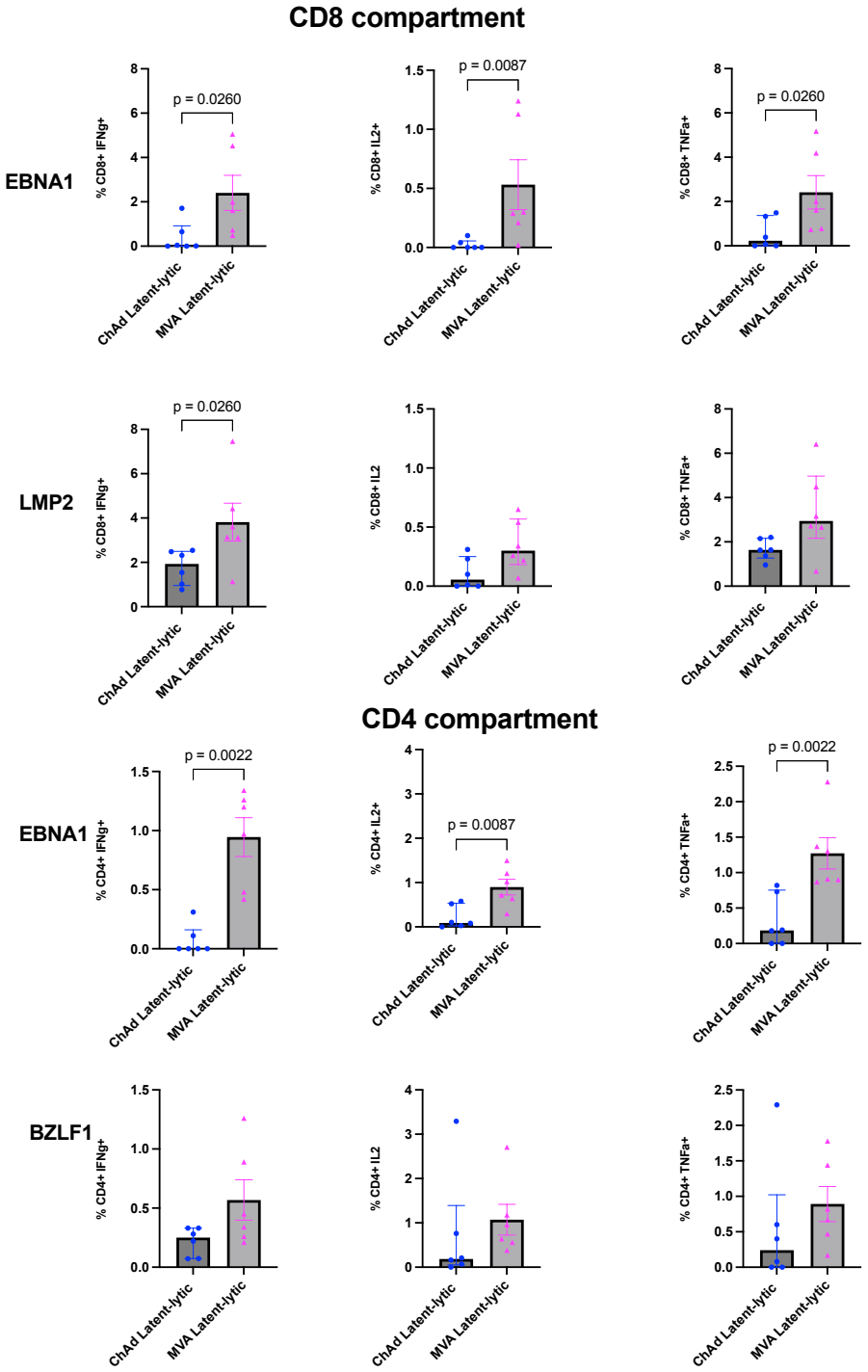



**Figure 5. 9 Cytokine responses to EBV peptides in ChAd-primed animals boosted with either ChAd or MVA**

Mice were primed with a ChAd latent-lytic vaccine and, four weeks later, boosted with either a ChAd latent-lytic or an MVA vector encoding the same antigen (latent-lytic). PBMCs were stimulated ex vivo with peptide pools corresponding to latent (EBNA1, LMP2A) and lytic (BARF1, BZLF1) antigens 2 weeks post-boost vaccination to assess T-cell cytokine responses. Rows = peptides (EBNA1, LMP2A, BZLF1); columns = cytokines. Points = individual animals; bars = median ± IQR. Comparisons (ChAd vs MVA) were made using two-tailed Mann–Whitney U tests; p-values are shown.

### 5.4.5 End-point splenic antigen-specific T cell responses demonstrate broad antigen recognition

To assess the impact of vaccination on tissue-resident antigen-specific immunity, spleens were harvested at the termination of experiments, and splenocytes were stimulated ex vivo with peptide pools for EBNA1, LMP2A, BARF1, and BZLF1. Cytokine production was measured by intracellular staining as described in Section 5.3.4. Comparisons between vaccine groups were performed using the Kruskal–Wallis test followed by Dunn's post hoc correction for multiple comparisons. PBS-treated mice served as controls (unvaccinated).

The analysis showed that vaccination elicited strong, broad $CD8^+$ T-cell responses to multiple EBV antigens in the spleen (Figure 5.10). Compared with PBS controls, vaccinated mice had significantly higher numbers of $CD8^+$ T cells producing IFN-γ, IL-2, and TNF-α in response to peptide epitopes from EBNA1, LMP2A, BARF1, and BZLF1. EBNA1-specific $CD8^+$ T cells exhibited consistent responses across all cytokines, with notable increases after vaccination, especially in mice receiving heterologous MVA boosts. Unlike in blood, LMP2A responses by $CD8^+$ T cells producing IL-2 and TNF-α were observed in the boosted groups.

While heterologous MVA boost further enhanced cytokine responses for some antigens, many lytic antigen-specific $CD8^+$ T-cell responses were already strong after ChAd priming alone, indicating effective induction of splenic $CD8^+$ immunity without additional boosters. The presence of IL-2-producing $CD8^+$ T cells across multiple antigens suggests the development of diverse, potentially memory-like T-cell populations in secondary lymphoid tissues.

Analysis of splenic $CD4^+$ T cells revealed broad antigen-specific cytokine responses to both latent and lytic EBV antigens. Vaccinated animals showed significantly higher frequencies of

$CD4^+$ T cells producing IFN-γ, IL-2, and TNF-α than controls across all four EBV antigens tested (Figure 5.11). EBNA1 elicited the strongest $CD4^+$ responses, with heterologous MVA boosting markedly increasing cytokine production. LMP2A-specific $CD4^+$ responses were detectable but lower, while BARF1 and BZLF1-specific $CD4^+$ T-cell responses varied in the spleen. Overall, vaccination induced broad, multifunctional EBV-specific $CD8^+$ and $CD4^+$ T-cell responses in the spleen compared to blood.

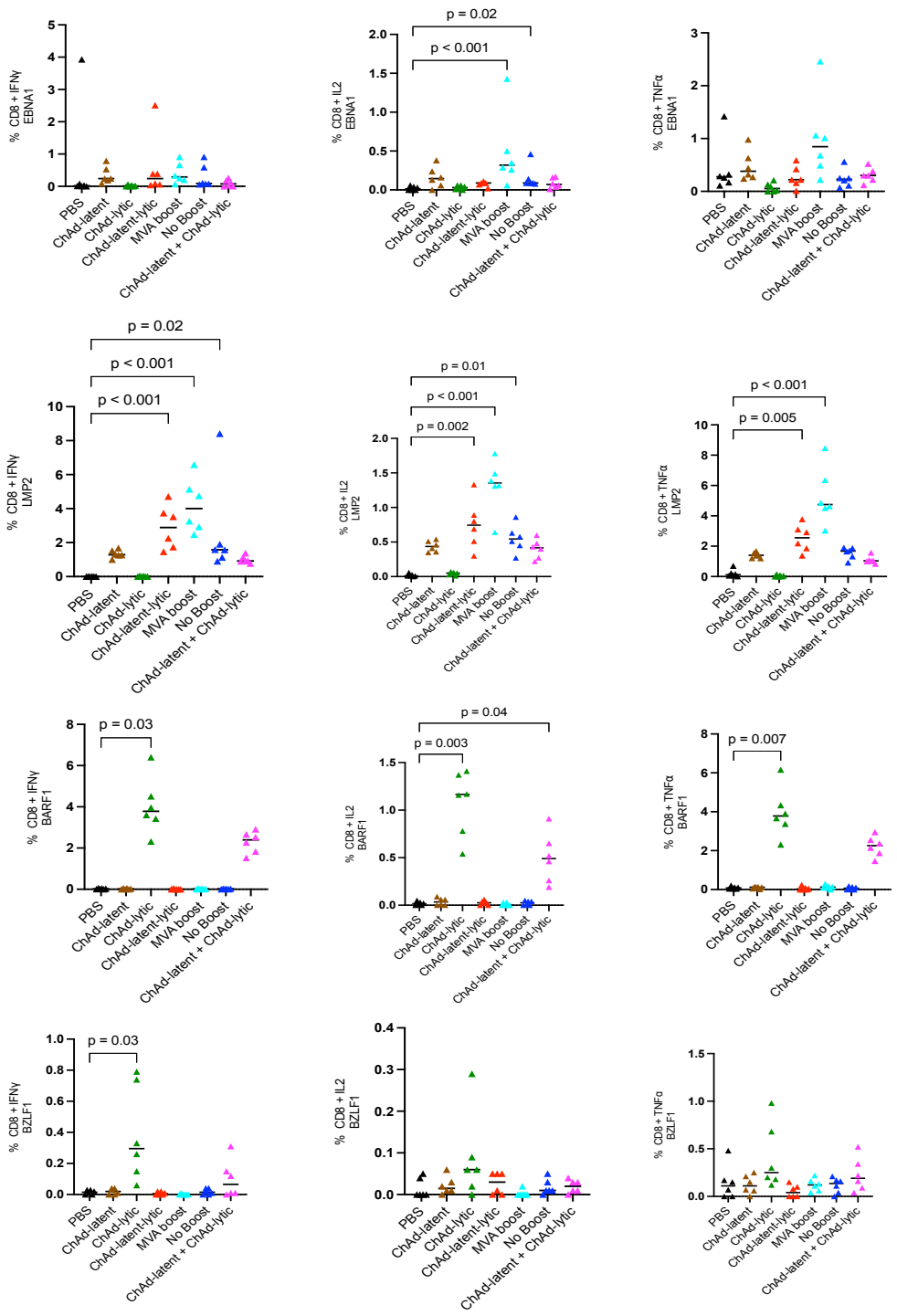



**Figure 5. 10 Splenic CD8⁺ T-cell cytokine responses following vaccination**

Frequencies of IFN-γ-, IL-2-, and TNF-α-producing CD8⁺ T cells in the spleen were measured following stimulation with EBNA1, LMP2A, BARF1, or BZLF1 peptide pools. Animals were vaccinated with ChAd vectors encoding latent, lytic, or combined latent–lytic EBV antigens, with or without MVA boosting. Each symbol represents an individual animal; horizontal lines indicate the median. PBS-treated animals served as controls. Statistical comparisons were performed using the Kruskal-Wallis test, with Dunn's multiple comparisons test. Significant p-values are shown.

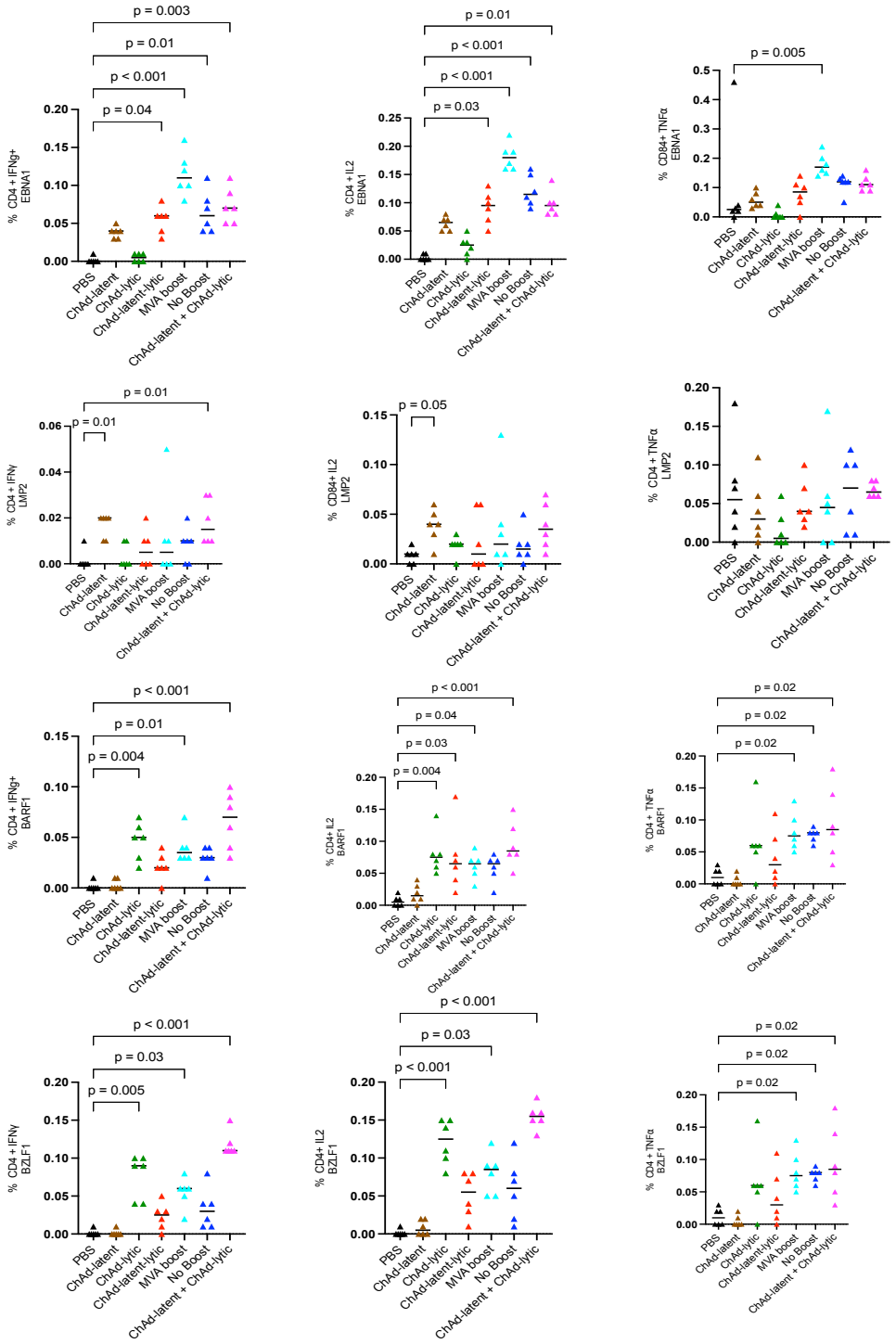



**Figure 5. 11 Splenic CD4⁺ T-cell cytokine responses following vaccination**

Frequencies of IFN-γ-, IL-2-, and TNF-α-producing CD4⁺ T cells in the spleen were quantified following peptide stimulation with EBNA1, LMP2A, BARF1, or BZLF1. Data are shown for animals vaccinated with ChAd EBV constructs with or without heterologous MVA boosting. Each symbol represents an individual animal; horizontal lines indicate the median. Statistical comparisons were performed using the Kruskal-Wallis test, with Dunn's multiple comparisons test. Significant p-values are shown.

## 5.5 Discussion

Viral vector vaccines targeting EBV-associated malignancies have been widely evaluated in preclinical models and have demonstrated the capacity to induce antigen-specific T-cell responses (Pan et al., 2009; Ranieri et al., 1999; Rühl et al., 2019). However, the majority of these studies have focused on single-antigen approaches, limiting their applicability to the complex biology of EBV infection, which involves both latent and lytic phases. Given the heterogeneous antigen expression patterns observed across EBV-associated diseases, multi-antigen vaccine strategies are increasingly being explored to broaden immune targeting and potentially enhance protective efficacy.

This chapter evaluated the immunogenicity of ChAdOx2-based viral vectors encoding distinct combinations of EBV latent and lytic antigens, together with an MVA vector in a heterologous prime–boost regimen. Specifically, three ChAdOx2 constructs encoding latent-only, lytic-only, or combined latent–lytic antigens were assessed, alongside an MVA vector encoding both antigen classes. The magnitude, specificity, breadth, and functional quality of vaccine-induced T-cell responses were analysed following prime-boost vaccination. The findings presented herein provide insight into how antigen composition and boosting strategy influence the qualitative and quantitative features of EBV-specific cellular immunity.

### 5.5.1 Prime-boost vaccination induces antigen-specific $CD8^+$ and $CD4^+$ T-cell responses

Prime–boost vaccination with EBV constructs elicited measurable antigen-specific $CD8^+$ and $CD4^+$ T-cell responses in peripheral blood. Adenoviral vectors are well-established inducers of cellular immunity due to efficient antigen expression and strong innate immune activation, thereby priming robust T-cell responses (Coughlan, 2020). The observed induction of IFN-γ-, IL-2-, and TNFα-producing T cells following vaccination suggests effective antigen presentation and functional activation of T cells. The presence of both $CD8^+$ and $CD4^+$ responses is particularly significant in the context of EBV, as both recognise a wide range of EBV antigens during lytic and latent infection phases. Specifically, $CD8^+$ T cells mediate cytotoxic effector function and control EBV infection by recognising endogenously derived short peptides presented via MHC class I molecules on the surface of infected cells (Liu et al., 2023). On the one hand, $CD4^+$ T cells provide essential support for $CD8^+$ expansion, differentiation into effector cells, memory formation, and long-term immune control (Laidlaw et al., 2016; Liu et al., 2023; Long et al., 2019). The simultaneous induction of both compartments suggests that the vaccine constructs successfully engaged multiple arms of the adaptive immune response.

These findings are consistent with prior viral vector vaccine studies demonstrating strong T-cell priming following adenoviral immunisation (Isogawa et al., 2025; Tatsis et al., 2007; van der Gracht et al., 2020). Importantly, boosting enhanced response magnitude to latent antigens (particularly EBNA1), indicating effective recall of primed T-cell populations.

### 5.5.2 Peptide immunogenicity and antigen hierarchy

Although overall immunogenicity was confirmed, responses varied across different EBV antigens. Responder frequency analysis showed that vaccine-encoded antigens were preferentially recognised, consistent with structured peptide immunogenicity. Responses to LMP2A were detected in 100% of regimens encoding this antigen, indicating that latent antigen immunogenicity was preserved even within multi-antigen constructs. While EBNA1 responses were less frequent than those to LMP2A, they were maintained in both latent-only and combined latent–lytic constructs, suggesting that inclusion of additional antigens did not compromise EBNA1 recognition. Among lytic antigens, BARF1 and BZLF1 showed higher responder frequencies in lytic-only constructs than in combined latent–lytic constructs; although these differences did not reach statistical significance, the directionality of the effect is consistent with enhanced lytic antigen targeting in the absence of competing antigen classes.

Analysis of response magnitude further revealed a structured antigen hierarchy. In $CD8^+$ T cells, lytic antigens, particularly BARF1, elicited the strongest cytokine responses, whereas EBNA1 and LMP2A responses were comparatively lower. To determine whether these patterns reflected random variation or ordered antigen structure, peptide-specific $CD8^+$ IFN-γ responses were rank-transformed within individual mice. The analysis showed a highly ordered peptide recognition within and between vaccine groups. LMP2A was dominant in latent-only and latent–lytic constructs, whereas BARF1 dominated in lytic-only constructs.

Correlation analyses further demonstrated that peptide-specific responses were not independent. Positive correlations between BARF1 and BZLF1 responses suggested coordinated recognition of lytic antigens. In contrast, inverse associations between LMP2A and lytic antigens suggested apparent reciprocal structuring of latent and lytic responses at the population level. However, these inverse relationships were substantially attenuated when analysed within individual vaccine groups, indicating that the apparent negative correlations were driven primarily by differences in vaccine construct composition (e.g., latent-only versus lytic-only immunisation groups) rather than by intrinsic antigen competition within individual mice.

Overall, these results indicate that lytic antigens are more immunogenic than latent antigens in this model. This observation is consistent with the well-established concept of antigen hierarchy in EBV infection. Human studies of primary EBV infection have shown that $CD8^+$ T-cell responses are preferentially directed toward lytic-cycle antigens, particularly immediate-early and early proteins, with comparatively weaker responses to late lytic-cycle or latent proteins (Hislop et al., 2007; Pudney et al., 2005; Steven et al., 1997). In the study by Pudney et al., $CD8^+$ responses were mapped across a broad panel of EBV lytic proteins and revealed a clear hierarchy in which immediate-early antigens elicited the strongest responses, followed by early antigens, while late antigens were rarely dominant. This hierarchy closely mirrored the efficiency with which these antigens were processed and presented on MHC class I molecules in lytically infected cells, suggesting that immunodominance is strongly influenced by antigen-presentation efficiency rather than antigen abundance alone (Pudney et al., 2005).

The preferential targeting of immediate-early and early lytic proteins may also mirror their biological importance for immune regulation. Since these proteins are expressed at the earliest

phases of viral replication, recognition of infected cells at this point enables $CD8^+$ T cells to destroy infected cells before viral replication and virion production are complete (Long et al., 2019). In comparison, many late lytic proteins are expressed after the viral immune evasion mechanism starts to hinder antigen processing and MHC class I presentation, thereby decreasing their detectability by $CD8^+$ T cells (Quinn et al., 2014).

Our data are broadly consistent with this paradigm, as $CD8^+$ T-cell responses to the lytic antigens BARF1 and BZLF1 were several orders of magnitude higher than those observed for the latent antigens LMP2A and EBNA1. Interestingly, within the latent antigen vaccine constructs, LMP2A elicited stronger responses than EBNA1, suggesting a secondary hierarchy among latent antigens themselves. This is also consistent with previous work showing that EBNA1 is a relatively poor $CD8^+$ T-cell target, in part due to mechanisms that limit its proteasomal processing and presentation (Blake et al., 2000).

Nevertheless, these findings should be interpreted with caution, given the experimental system used. EBV is a strictly human pathogen and does not naturally infect mice, meaning that antigen processing, epitope repertoire, and immunodominance hierarchies in murine models may not fully recapitulate those observed in human infection. Despite this limitation, the strong skewing toward lytic antigen recognition observed here parallels the immunodominance patterns described in human EBV infection and therefore supports the relevance of these findings for understanding EBV-specific T-cell immunity.

A further limitation relates to the assessment of immunodominance using a single inbred mouse strain. Because immune responses in inbred murine models are restricted by a narrow,

genetically uniform MHC repertoire (Mestas & Hughes, 2004), the hierarchy of dominant and subdominant EBV epitopes observed in this study may not accurately reflect the diversity of antigen processing and presentation in human populations. Consequently, the relative immunogenicity of individual epitopes or antigen constructs identified here may be influenced by strain-specific MHC bias and therefore may not fully predict immunodominance patterns across genetically diverse human HLA backgrounds. Although the vaccine design incorporated epitopes with broad predicted HLA coverage and representation of both $CD8^+$ and $CD4^+$ T-cell responses, additional studies in multiple mouse strains, HLA-transgenic models, or human ex vivo systems would be necessary to validate the breadth and reproducibility of the observed immune responses. Ultimately, evaluation in more immunologically diverse models will be important for determining the translational relevance of the identified immunodominant EBV antigens and epitopes.

### 5.5.3 Breadth and functional quality of vaccine-induced T-cell responses

Responder frequency across individual peptide pools does not reflect the number of distinct antigens recognised within a given animal. To evaluate the breadth of antigen-specific $CD8^+$ T-cell responses, the number of distinct antigens or peptide pools recognised by T cells within an individual mouse was analysed. Response breadth was defined as the number of EBV peptide pools eliciting a positive $CD8^+$ IFN-γ response per mouse, as previously stated.

The analysis showed that the breadth of the $CD8^+$ T-cell responses varied with vaccine composition. Single-antigen class constructs (lytic or latent only) induced responses to limited peptide pools, whereas multi-antigen constructs (latent-lytic) elicited broader recognition patterns, with some mice responding to all tested peptide pools. This observation aligns with

the rationale underlying multivalent vaccine design: broader antigen targeting reduces the likelihood that viral mutation or epitope variation will abrogate immune control (Kaur et al., 2024; Rella et al., 2024).

Beyond antigen diversity, functional quality was assessed through polyfunctionality analysis. Vaccine-induced $CD8^+$ and $CD4^+$ T cells comprised mono-, bi-, and tri-functional subsets capable of co-producing IFN-γ, IL-2, and TNFα. Multi-cytokine-producing T cells were particularly evident in regimens that encoded multiple antigens, particularly combined latent and lytic constructs. Polyfunctional T cells are frequently associated with superior anti-tumour immunity, enhanced proliferative capacity, and improved memory potential (Duvall et al., 2008; Holtkamp et al., 2024). Thus, the presence of polyfunctional subsets suggests qualitative robustness of the induced response.

Together, breadth and polyfunctionality analyses indicate that multi-antigen vaccination strategies enhance both the diversity and functional complexity of EBV-specific T-cell immunity.

### 5.5.4 Heterologous MVA boosting enhances cytokine responses

One of the main barriers to using the same viral vector twice (for prime and boost) is that the first dose induces neutralising antibodies and T cells against the delivery vehicle itself (Zhao et al., 2021). However, employing two different vectors to deliver the same antigen, one for priming and the other for boosting (e.g., Adenovirus/MVA), has proven highly effective at enhancing antigen-specific T-cell responses by ensuring that the second dose is not neutralised before it can deliver its payload (Jeon et al., 2025; Zhao et al., 2021). This synergy was

evaluated here by priming mice with the ChAdOx2-EBV latent-lytic vaccine and then boosting with the MVA-EBV latent-lytic vaccine.

The analysis showed that MVA boost significantly enhanced antigen-specific T cell responses to EBNA1 and LMP2A, without fundamentally altering the antigen hierarchy. This suggests that boosting amplifies preexisting specificities rather than reshaping antigen-targeting patterns. Such an approach may therefore be effective in enhancing vaccine potency while maintaining immunological balance across antigen targets. These findings are consistent with previous studies demonstrating the immunological advantages of adenovirus/MVA prime–boost regimens in viral vaccine development (McAuliffe et al., 2021; Rühl et al., 2019).

### 5.5.5 Establishment of antigen-specific responses in the spleen

While peripheral blood provided insight into circulating responses, assessment of splenic responses enabled evaluation of established immunity within secondary lymphoid tissues. At the endpoint, antigen-specific T-cell responses were detectable in the spleen, albeit at lower magnitudes than in peripheral blood. The antigen structure closely mirrored the patterns observed in peripheral blood, with LMP2A dominating latent-only and latent-lytic responses, whereas BARF1 dominated lytic responses in the $CD8^+$ T cell compartment.

In the $CD4^+$ T cell compartment, EBNA1 elicited the strongest and often broader cytokine responses, with enhanced responses observed in MVA-boosted mice. While responses to other antigens varied, they reflected antigen-dependent priming and recall. The presence of IL-2- and TNF-α-producing $CD4^+$ T cells across antigens indicates the induction of functionally capable helper T-cell responses in lymphoid tissue (Laidlaw et al., 2016).

### 5.5.6 Study limitations

An important consideration in interpreting EBV immunogenicity findings is the biological relevance of non-EBNA1 EBV antigens as therapeutic vaccine targets in endemic BL. EBNA1 is consistently expressed in latency I tumours and is therefore widely regarded as a stable tumour-associated viral antigen (Young et al., 2007). In contrast, expression of latency II and lytic antigens appears considerably more heterogeneous, as reported by prior studies (Abate et al., 2015; Xue et al., 2002) and, based on the bulk RNA-seq analyses presented in Chapter 4, may occur only within a subset of tumour cells or potentially within non-malignant EBV-infected cells in the tumour microenvironment. Consequently, inclusion of these antigens within the vaccine constructs was based on the underlying assumption that at least some non-latency-I transcriptional programmes may remain biologically relevant to tumour persistence, immune evasion, intra-tumoural heterogeneity, or viral reactivation dynamics. However, the present data cannot determine whether expression of latency II or lytic antigens represents a necessary tumour-maintenance programme, a transient "dead-end" transcriptional state, or residual viral activity that could be silenced without affecting tumour survival. It therefore remains possible that some targeted antigens are dispensable for the dominant malignant clone. Nevertheless, even transient or minority antigen expression may still be immunologically important. Lytic or non-canonical latent gene expression could contribute disproportionately to antigen presentation, viral dissemination, or maintenance of tumour-supportive microenvironments, and targeting multiple viral programmes may reduce the likelihood of immune escape through antigen restriction or transcriptional silencing. This broader multi-antigen strategy was therefore intended to maximise immunological coverage across

heterogeneous EBV-associated tumour states, while acknowledging that the functional contribution of individual antigens to BL pathogenesis remains incompletely resolved.

An additional consideration relevant to clinical translation is the potential risk of molecular mimicry and autoimmune responses associated with certain EBV antigens, particularly EBNA1 and EBNA2, which have been implicated in autoimmune diseases such as rheumatoid disease, systemic lupus erythematosus and multiple sclerosis through cross-reactive immune responses against host proteins (Lum et al., 2026; Munir et al., 2025; Soldan & Lieberman, 2023; Wang et al., 2021). In the present study, selected conserved epitope strings derived from EBNA1, EBNA2 and other polymorphic antigens were included based on prior experimental immunogenicity validation, predicted HLA coverage, and relative sequence conservation across clinical EBV isolates. This epitope-focused strategy was intended to maximise immunogenic breadth while potentially reducing unnecessary exposure to non-essential or poorly conserved protein regions that could contribute to off-target immune responses. Nevertheless, the possibility of cross-reactive immune responses cannot be entirely excluded, and the present study did not directly evaluate the autoimmune potential or host-protein homology of the selected epitopes. Therefore, further preclinical safety assessment, including evaluation of potential molecular mimicry and autoreactive T-cell responses, would be necessary before clinical application of EBV-targeted vaccine strategies.

Additional limitations should also be acknowledged. Mouse group sizes were relatively modest, which may have limited statistical power for subgroup comparisons, including Fisher's exact tests of responder frequency. Furthermore, murine vaccination models do not fully recapitulate the complexity of human EBV immunobiology, particularly with respect to viral latency programmes, HLA diversity, and tumour–immune microenvironment interactions.

Functional assessment of vaccine immunogenicity in this thesis was primarily based on cytokine production assays and did not include detailed evaluation of cytotoxic effector function, memory differentiation, or in vivo killing capacity of antigen-specific T cells. In addition, although tumour challenge experiments and therapeutic efficacy studies have been performed in prophylactic and therapeutic mouse models, these analyses fall outside the scope of the present thesis. Consequently, a direct correlation between the observed immunogenicity and protective or therapeutic efficacy against EBV-associated tumours was not established.

### 5.5.7 Future direction

Tumour challenge experiments using a vaccine design informed by this study have been performed in both prophylactic and therapeutic mouse models to assess vaccine-mediated protection; however, these data are beyond the scope of the present thesis and will be described in future work.

Future work will explore additional functional markers, such as cytotoxic mediators or exhaustion markers, to gain deeper insights into T-cell quality. Further optimisation studies of antigen balance and boosting intervals are necessary to improve the breadth and polyfunctionality of the EBV vaccine construct.

### 5.5.8 Conclusion

This study demonstrates that multi-antigen EBV vaccination strategies induce robust, hierarchically structured, and functionally diverse T-cell responses. Vaccine composition governs antigen-targeting patterns, whereas heterologous boosting enhances magnitude without compromising specificity. Importantly, antigen-specific responses are maintained in lymphoid tissue at endpoint, supporting the potential of multi-antigen prime–boost approaches for EBV vaccine development.

# Chapter 6: General Discussion

## 6.1 Research questions and objectives

The primary aim of this thesis was to characterise Epstein–Barr virus (EBV) genomic and transcriptional features that influence viral diversity and disease development, and to identify viral determinants with potential relevance for immunogenic target discovery and vaccine development in EBV-positive Burkitt lymphoma (BL). This work addressed five interconnected objectives: (i) characterising EBV genomic variation, including point mutations and recombination; (ii) mapping hypervariable and conserved regions; (iii) identifying BL-enriched mutational signatures; (iv) defining viral transcriptional profiles in BL tumours; and (v) evaluating the immunogenicity of selected viral antigens in preclinical models.

EBV infects approximately 95% of adults worldwide (de-Thé et al., 1975; Smatti et al., 2018; Winter et al., 2019), yet EBV-associated malignancies occur in specific geographic and clinical contexts. This discrepancy between near-universal infection and restricted cancer incidence represents a central paradox in EBV biology. The present study approached this paradox by integrating comparative genomics, mutational signature analysis, tumour transcriptomics, and experimental immunology within a phenotype-aware framework.

## 6.2 Principal findings of the research

### 6.2.1 Genomic diversity, recombination, and geographic structure

Analysis of a comprehensive EBV whole-genome dataset, including newly sequenced East African isolates, reaffirmed that EBV population structure is strongly shaped by geography.

This finding underscores the importance of long-term host–virus coevolution in shaping viral diversity and has practical implications for vaccine design, suggesting that region-specific sequence variation should be considered when selecting antigen targets.

Although recombination has previously been described in EBV (Agwati et al., 2022; Yao et al., 1996), this study identified relatively elevated recombination signals in EBNA1 and LMP1 among BL-associated viruses compared with healthy carriers. Both genes encode proteins central to viral persistence, immune modulation, and oncogenic signalling (Ressing et al., 2008; Sausen et al., 2023). Their enhanced diversity and recombination in BL nominate them as candidate loci under selective pressure and reinforce their relevance as vaccine targets. However, further region-stratified analyses will be required to confirm disease-specific recombination patterns.

### 6.2.2 Mutational signatures and disease-specific viral processes

To characterise mutational processes operating in EBV genomes, a COSMIC-aligned mutational signature framework was applied. Four EBV-associated single-base substitution signatures were identified. Notably, SBS_EBV1, SBS_EBV3, and SBS_EBV4 were significantly enriched in BL compared with EBV-positive healthy controls, whereas SBS_EBV2 was enriched in healthy individuals. Other EBV disease phenotypes showed modest enrichment for these mutational signatures, possibly due to limited statistical power.

SBS_EBV4, characterised by substitutions in G-rich or CpG contexts, may reflect a combination of mutational processes, including oxidative stress, replication-associated damage, and immune-mediated editing. In particular, EBV-associated induction of reactive oxygen species and abnormal activation-induced cytidine deaminase (AID) activity (Cerimele

et al., 2005; Chen et al., 2016; Gruhne et al., 2009; Kikuchi et al., 2015), potentially intensified by chronic immune stimulation in malaria-endemic regions, could contribute to the observed mutation spectrum. While the relatively low mutation burden per genome warrants cautious interpretation, these findings suggest that BL is associated with a distinct mutational environment that may influence viral evolution. However, replication of these findings in independent BL cohorts, particularly from endemic regions, will be necessary to validate our claims.

### 6.2.3 Viral transcriptional heterogeneity in BL

Transcriptomic profiling of BL tumours revealed heterogeneous patterns of EBV gene expression, including detection of both latent and lytic transcripts in subsets of tumours. These observations align with earlier reports from BL-endemic Equatorial Africa, in which a subset of tumours exhibits non-canonical antigen expression that has not previously been linked to the disease. This was first observed in Malawian children diagnosed with endemic BL (Labrecque et al., 1999; Xue et al., 2002), although later in a subset of Kenyan and Ugandan children (Abate et al., 2015).

Perhaps the observations received little scientific attention, given the limited technological capabilities for molecular analysis at the time. Here, our data revealed this pattern through both transcriptomics and whole-genome sequencing, although protein-level expression was limited, partly due to the nature of FFPE-degraded material and to the fact that few EBV monoclonal antibodies are commercially validated for immunohistochemistry.

Together, these findings support the concept that EBV transcriptional states in BL are dynamic and may involve occasional reactivation.

Further analyses of transcriptomic data demonstrated strong enrichment of MYC target genes, E2F targets, and G2M checkpoint pathways in BL, consistent with canonical MYC-driven oncogenesis (Basso & Dalla-Favera, 2015). Additionally, enrichment of immune-related pathways, such as those associated with allograft rejection, likely reflects host immune activation downstream of viral oncoprotein signalling (Šimičić et al., 2024). Stratified analysis by EBV transcriptional activity identified potential molecular subclusters within BL, suggesting that active viral gene expression may modulate the intensity of oncogenic signalling. These findings imply that EBV activity contributes variably to tumour biology and may influence therapeutic responsiveness.

### 6.2.4 Implications for vaccine development

The high global prevalence of EBV complicates the development of prophylactic vaccines, particularly in regions where primary infection occurs in early infancy and maternally derived antibodies may interfere with vaccine efficacy. These epidemiological considerations support increasing interest in therapeutic vaccine strategies aimed at enhancing anti-tumour immunity rather than preventing primary infection.

Guided by genomic and transcriptomic findings presented in this thesis, selected latent and lytic EBV antigens were incorporated into viral vector vaccine constructs for preclinical evaluation. Prime–boost immunisation using ChAdOx2 and MVA vectors elicited robust antigen-specific $CD4^+$ and $CD8^+$ T-cell responses. Multi-antigen (latent–lytic) constructs elicited broader, more polyfunctional responses than single-class antigen formulations, supporting the rationale for multivalent vaccine design. Polyfunctional T cells have been associated with enhanced anti-tumour control and durable immunity (Duvall et al., 2008;

Holtkamp et al., 2024), and were consistently observed, particularly following multi-antigen vaccine regimens.

Within the latent vaccine formulations, LMP2A- and EBNA1-directed responses were among the most immunogenic, while BARF1 and BZLF1 elicited strong responses within the lytic regimens, supporting their prioritisation as candidate vaccine targets. Notably, full-length BZLF1 was incorporated into the construct design and may therefore offer mechanistic advantages beyond direct antigen-specific T-cell priming. In principle, expression of intact BZLF1 following vector transduction could promote induction of the EBV lytic cycle in infected cells, thereby increasing expression of downstream immunogenic lytic antigens and potentially broadening immune-mediated tumour recognition. Conceptually, this resembles a "kick-and-kill" strategy in which lytic induction enhances immune visibility of EBV-positive tumour cells. However, this proposed mechanism remains hypothetical in the context of the present study and would depend on efficient transduction of EBV-positive tumour cells and successful induction of lytic gene expression in vivo, neither of which was directly assessed here.

Potential safety considerations associated with the expression of intact EBV proteins should also be acknowledged. Although certain EBV antigens, including LMP2A (Fish et al., 2020; Yun et al., 2019) and BARF1(Wei et al., 1994), possess known immunomodulatory or oncogenic functions in their native viral conformation, the vaccine constructs were delivered using replication-deficient ChAdOx2 and MVA vectors designed for transient antigen expression rather than persistent transformation. In addition, most polymorphic EBV antigens were incorporated as epitope strings rather than full-length proteins, thereby reducing expression of intact functional viral proteins. Nevertheless, formal safety evaluation of vector-

mediated EBV antigen expression, particularly regarding unintended cellular effects or the induction of lytic replication, would be required prior to clinical translation.

Collectively, these data support the feasibility of a multi-antigen therapeutic vaccine strategy targeting both latent and lytically active EBV-associated tumour states, while also highlighting the need for further mechanistic and translational studies to define the biological relevance and therapeutic vulnerability of individual viral antigen programmes in endemic BL .

## 6.3 Conclusions and future direction

By integrating viral genomics, mutational processes, tumour transcriptional states, and immunogenicity testing, this work provides a cohesive framework linking EBV diversity to disease biology and vaccine design. BL emerges as a phenotype characterised by distinct viral mutational signatures, enhanced diversity in key oncogenic loci, heterogeneous transcriptional activity, and identifiable immunogenic targets. These findings demonstrate that although infection is nearly universal, viral genomic evolution and transcriptional behaviour differ across phenotypes, shaping disease risk and therapeutic opportunities.

Several key conclusions emerge from this work. First, EBV population structure remains strongly shaped by geography, reinforcing the importance of host–virus co-evolution and demonstrating that regional viral diversity should be considered during the development and optimisation of EBV-targeted immunotherapies and vaccine strategies. However, the degree of conservation observed across many immunogenic regions suggests that fully region-specific vaccine formulations may not necessarily be required, particularly when conserved epitopes

with broad HLA coverage are prioritised. In this context, host HLA diversity may ultimately be a more important determinant of vaccine responsiveness than regional EBV sequence variation alone.

Second, recombination and mutational signature analyses indicate that EBV genomes in BL are subject to distinct selective and mutagenic pressures compared with healthy carriage. Third, transcriptional profiling demonstrates that EBV activity in BL is heterogeneous and may influence the intensity of oncogenic signalling. Finally, preclinical vaccine experiments demonstrate that multi-antigen viral vector constructs can elicit robust, broad, and polyfunctional T-cell responses targeting both latent and lytic antigens.

Together, these findings support a model in which EBV-driven oncogenesis reflects a dynamic interplay between viral evolution, host immune pressure, and tumour microenvironmental constraints. Rather than functioning as a static passenger genome, EBV appears to undergo phenotype-specific adaptation that may shape disease risk and therapeutic vulnerability.

However, several limitations warrant consideration.

i) The relatively low mutation burden per EBV genome limits the resolution of mutational signatures and necessitates validation in larger, independent cohorts–particularly from malaria endemic settings.
ii) Transcriptomic evidence of lytic activity requires orthogonal protein-level validation to distinguish true productive replication from abortive or partial reactivation.
iii) Additionally, murine immunogenicity models, while informative, do not fully recapitulate human HLA diversity or tumour immune suppression. Translation of vaccine strategies into human settings will therefore require careful evaluation of antigen selection, dosing, and combination with existing therapies.

Future research should prioritise:

i) Expansion of geographically diverse EBV whole-genome datasets to refine recombination and mutational analyses.
ii) Functional interrogation of candidate BL-enriched signatures to clarify their mechanistic drivers.
iii) Single-cell and spatial transcriptomic approaches to resolve intra-tumour heterogeneity of EBV transcription.
iv) Evaluation of multi-antigen vaccine constructs in humanised or tumour-bearing models to better approximate therapeutic settings.
v) Integration of viral genomics with host germline and immune repertoire data to further resolve the EBV paradox.

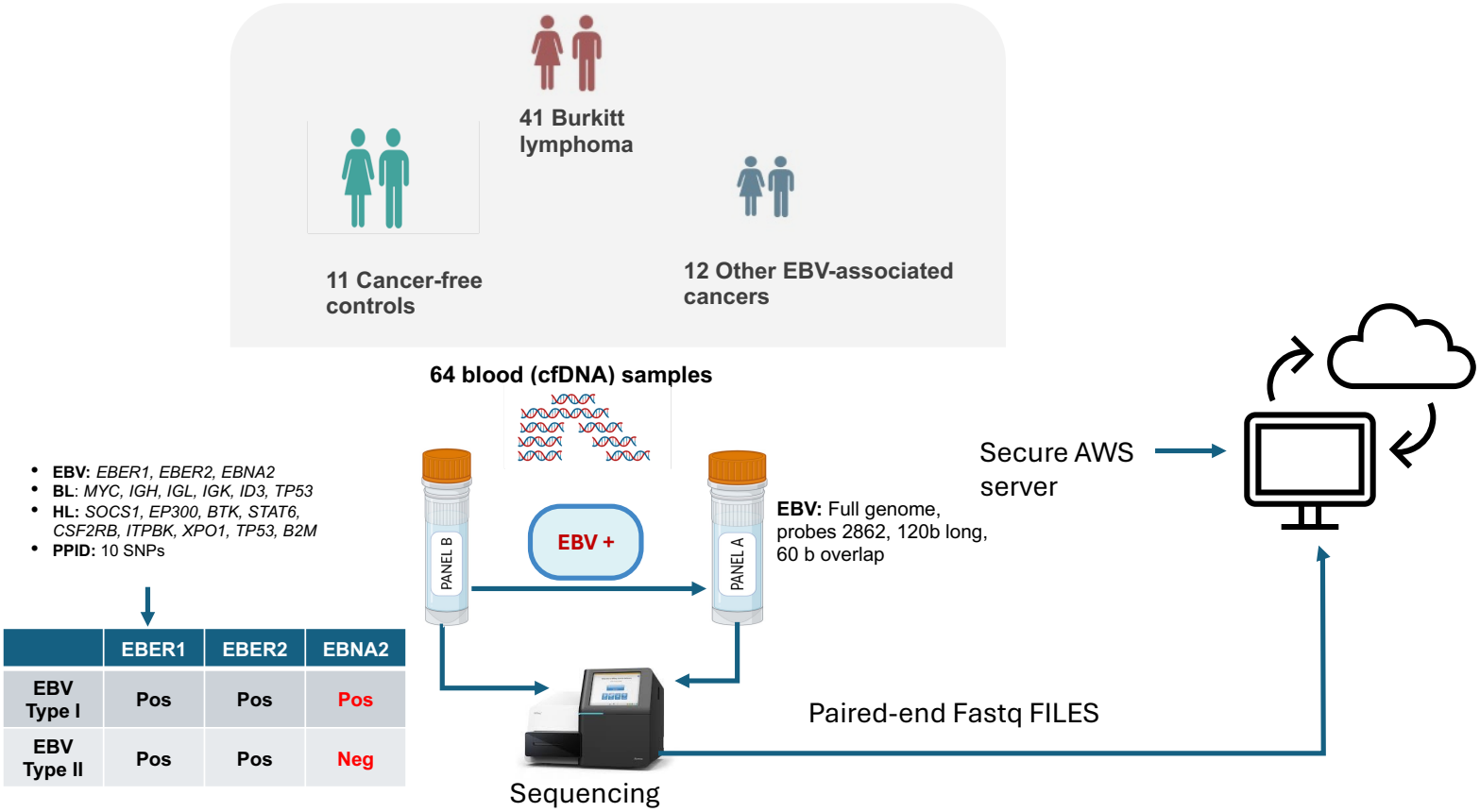
Ultimately, this work supports a shift from viewing EBV-associated malignancies solely through a host-centric lens toward a more integrated host–virus systems perspective. By identifying disease-associated viral features and demonstrating the feasibility of multi-antigen immunotherapeutic strategies, this thesis lays a foundation for rational vaccine development and precision targeting of EBV-driven cancers.

# Chapter 7: Data Appendix

## 7.1 Supplementary figures

**Supplementary Fig. 1 Participant selection, EBV whole genome sequencing and analysis of the primary study cohort**

Diagrammatic scheme of the study recruitment, sample processing, and analysis. Cases included EBV-positive Burkitt lymphoma and other EBV-positive lymphomas, as diagnosed by histopathology, in the AI-REAL study. Controls were cancer-free children identified from the community. Tumour biopsies and peripheral blood were collected from cases, while controls provided only peripheral blood. Peripheral blood was collected in PAXgene cfDNA blood tubes and processed within 3 hours of collection to minimise genomic DNA contamination. cfDNA was extracted using Qiagen circulating nucleic acid kits and analysed for the presence of EBV using a diagnostic NGS panel (for cases) and a quantitative PCR kit from GeneProof (Thermo Fisher Scientific). The diagnostic panel included MYC targets, immunoglobulin loci, and frequently mutated genes in lymphoma, as well as abundant EBV transcripts (EBER1, EBER2, EBNA2) to assist in diagnosing EBV-positive Burkitt lymphoma (Panel B). Results showing EBER1, EBER2, 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.

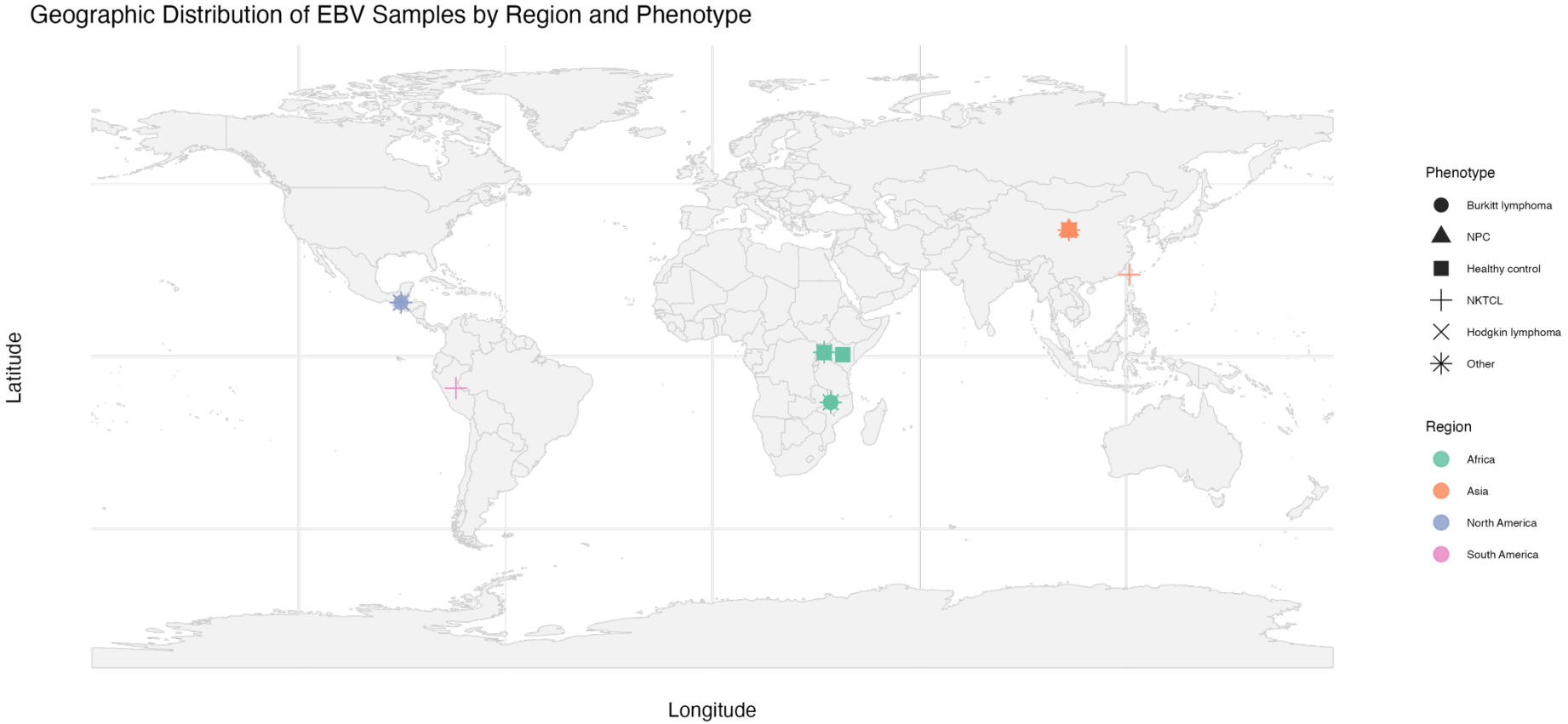



| | EBER1 | EBER2 | EBNA2 |
|---|---|---|---|
| EBV Type I | Pos | Pos | Pos |
| EBV Type II | Pos | Pos | Neg |

**Supplementary Fig. 1 Participant selection, EBV whole genome sequencing and analysis of the primary study cohort**

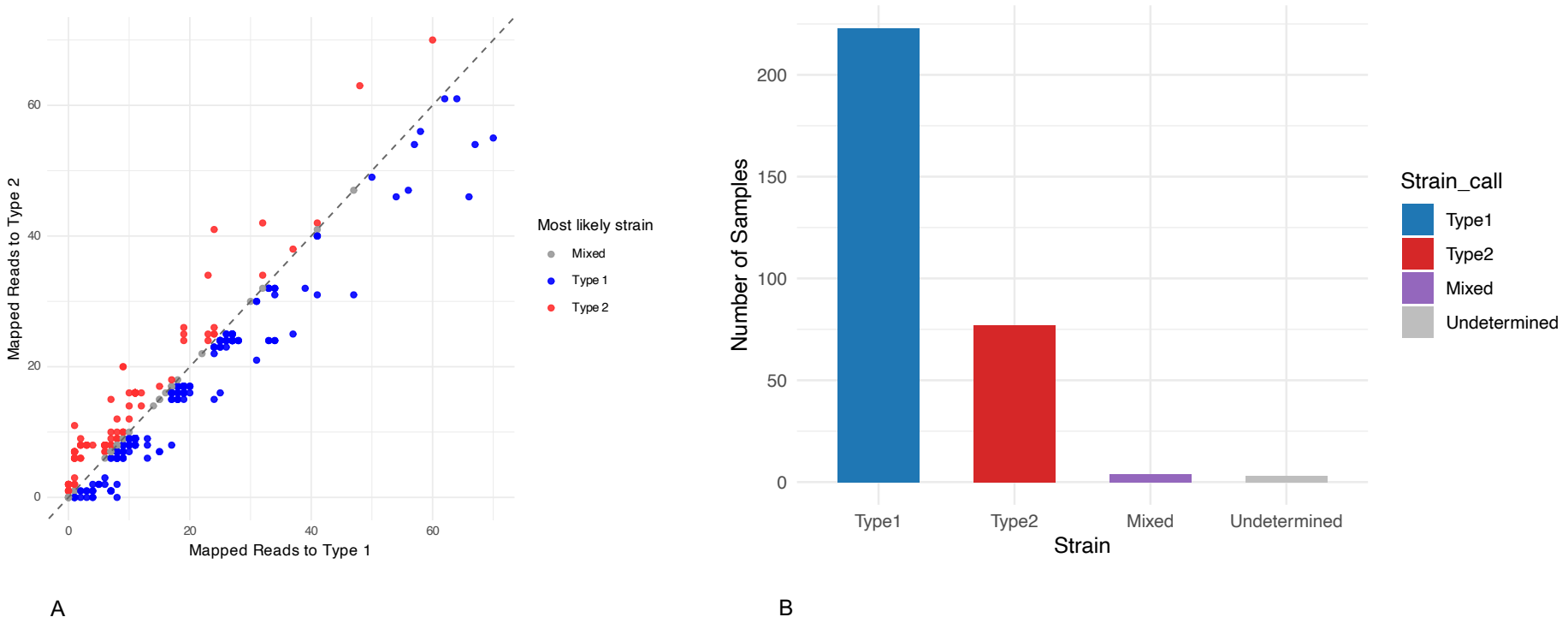
**Supplementary Fig. 2 Geographic origins of the EBV genomes analysed in the study**

Map showing the geographic origins and proportion of samples analysed in the study. Three hundred forty-six publicly available EBV whole genome sequences (raw FASTQs) from primary clinical isolates (tumour biopsy or plasma) were downloaded from the NCBI Sequence Read Archive under BioProject IDs PRJNA552587, PRJNA522388, and PRJNA1063319, and processed using the OMDC bioinformatics pipeline. 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.


Geographic Distribution of EBV Samples by Region and Phenotype
Latitude
Longitude
Phenotype
Burkitt lymphoma
NPC
Healthy control
NKTCL
Hodgkin lymphoma
Other
Region
Africa
Asia
North America
South America

**Supplementary Fig. 3 Determining the strain of the EBV genome using the NCBI BLAST algorithm**

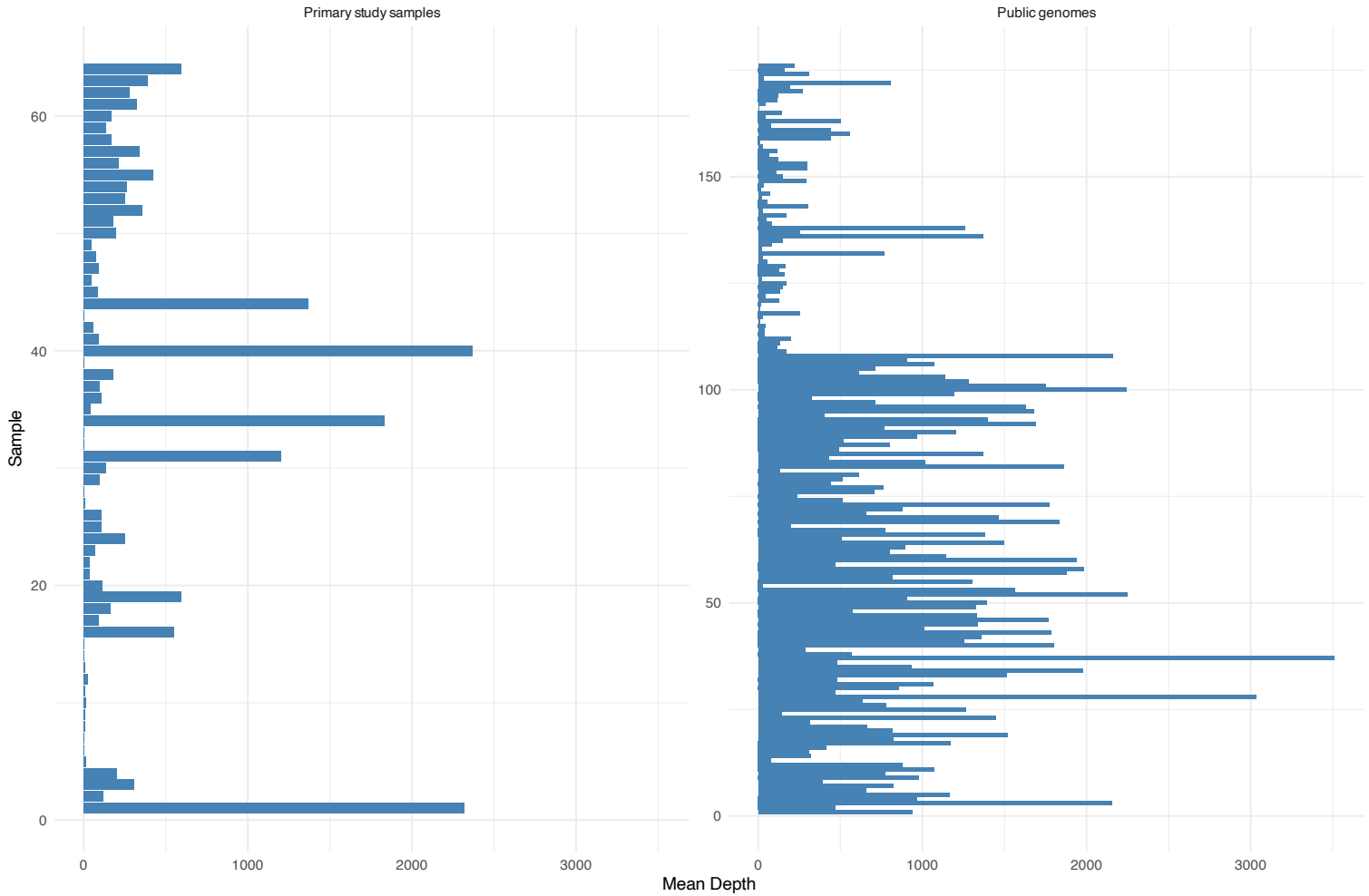
Because the public genomes lacked strain information, assembled EBNA2 contigs were used to determine EBV strain identity by sequence similarity to type-specific EBNA2 reference sequences. Two reference FASTA files, corresponding to EBNA2 from NC_007605.1 (Type 1) and NC_009334.1 (Type 2), were formatted into nucleotide BLAST databases using makeblastdb (BLAST+). For each sample, contigs in FASTA format were queried against both the Type 1 and Type 2 EBNA2 databases using blastn with an E-value cutoff of 1e-5, returning up to five targets per query in tabular output format. For each query set the highest-scoring hit was selected by sorting on bitscore and taking the top result. The per cent identity (pident) of the top hit to the Type 1 and Type 2 databases was recorded. Samples were classified as “Type1” or “Type2” when the top per cent identity to one reference exceeded that to the other; if the top per cent identities were equal, the strain was reported as “Undetermined”. All steps were executed in a bash pipeline with set -euo pipefail for reproducibility.

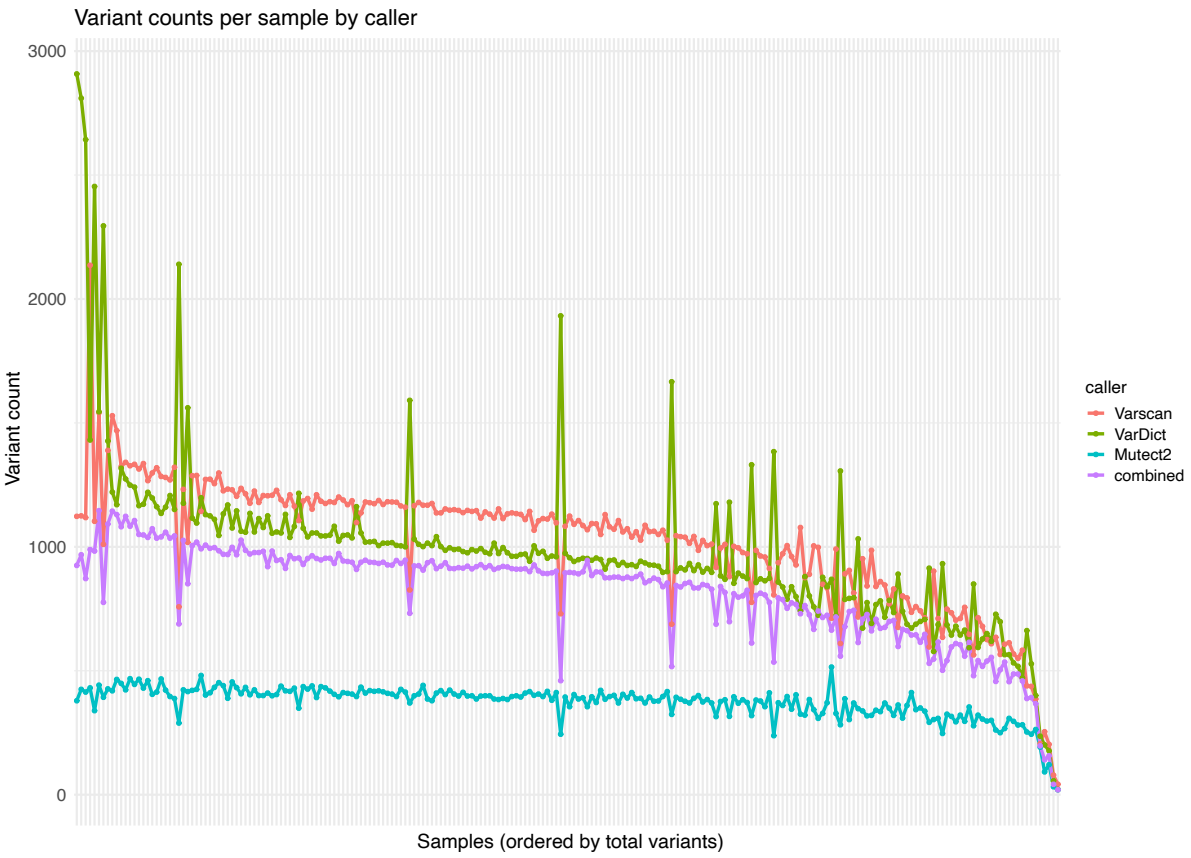



Scatter plot [A] of EBV contigs aligned to the Type 1 and 2 EBNA-2 reference. Bar plot [B] shows the number of genotypes called as type 1, type 2, mixed or undetermined by the BLAST. EBV contigs were aligned to the EBNA2 reference, and hits with the highest matches were assigned a strain call. This way,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

**Supplementary Fig. 4 Sequencing depth and coverage across primary and public genomes**

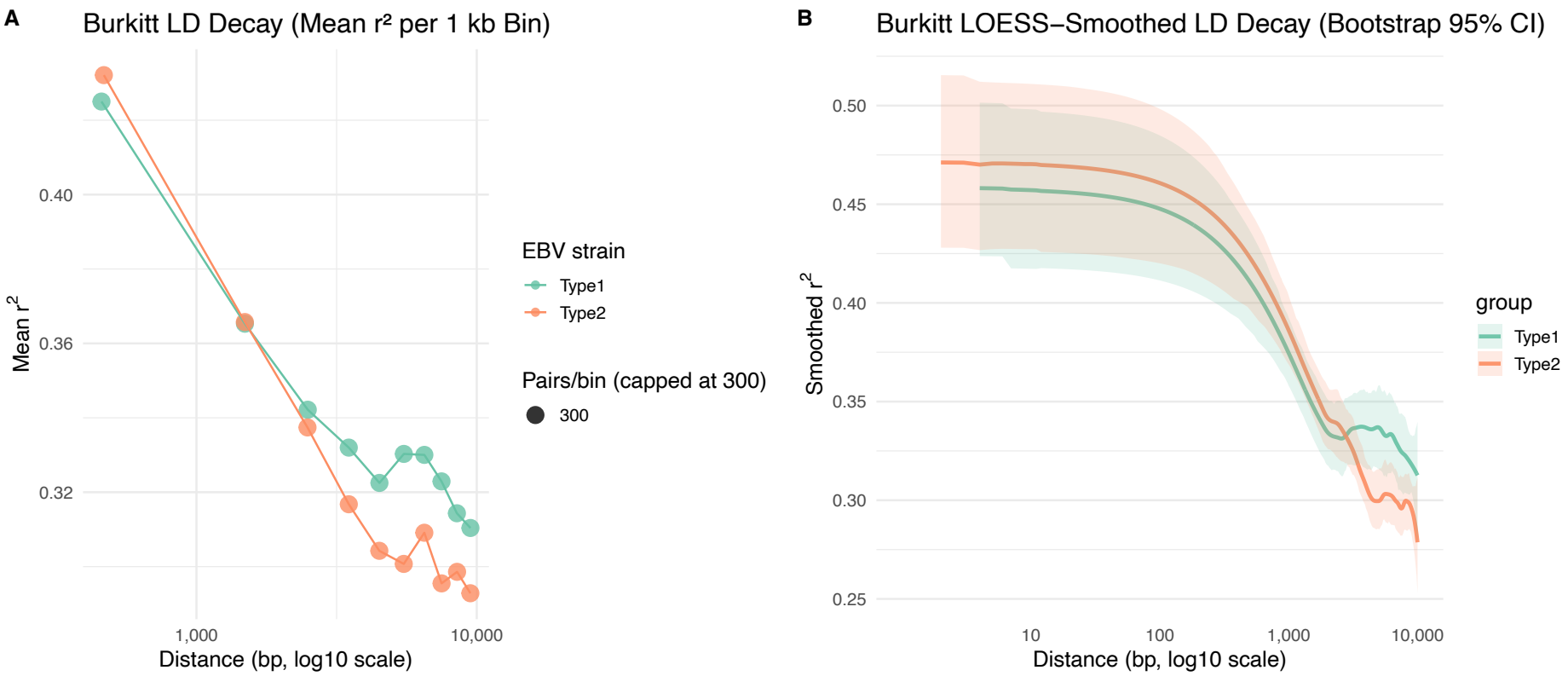
Genome coverage and sequencing depth were calculated from aligned BAM files using samtools. For each sample, the samtools coverage function was applied to the sorted BAM file to compute coverage statistics across the EBV reference genome. Metrics were extracted specifically for the reference sequence labelled *chrEBV*. The following parameters were recorded for each sample: reference name, start and end coordinates, number of mapped reads, number of covered bases, proportion of covered bases (coverage), mean sequencing depth, mean base quality, and mean mapping quality. Results were compiled into a comma-separated values (CSV) file for downstream analysis.


Primary study samples
Public genomes
Sample
Mean Depth
0
20
40
60
0
50
100
150
0
1000
2000
3000



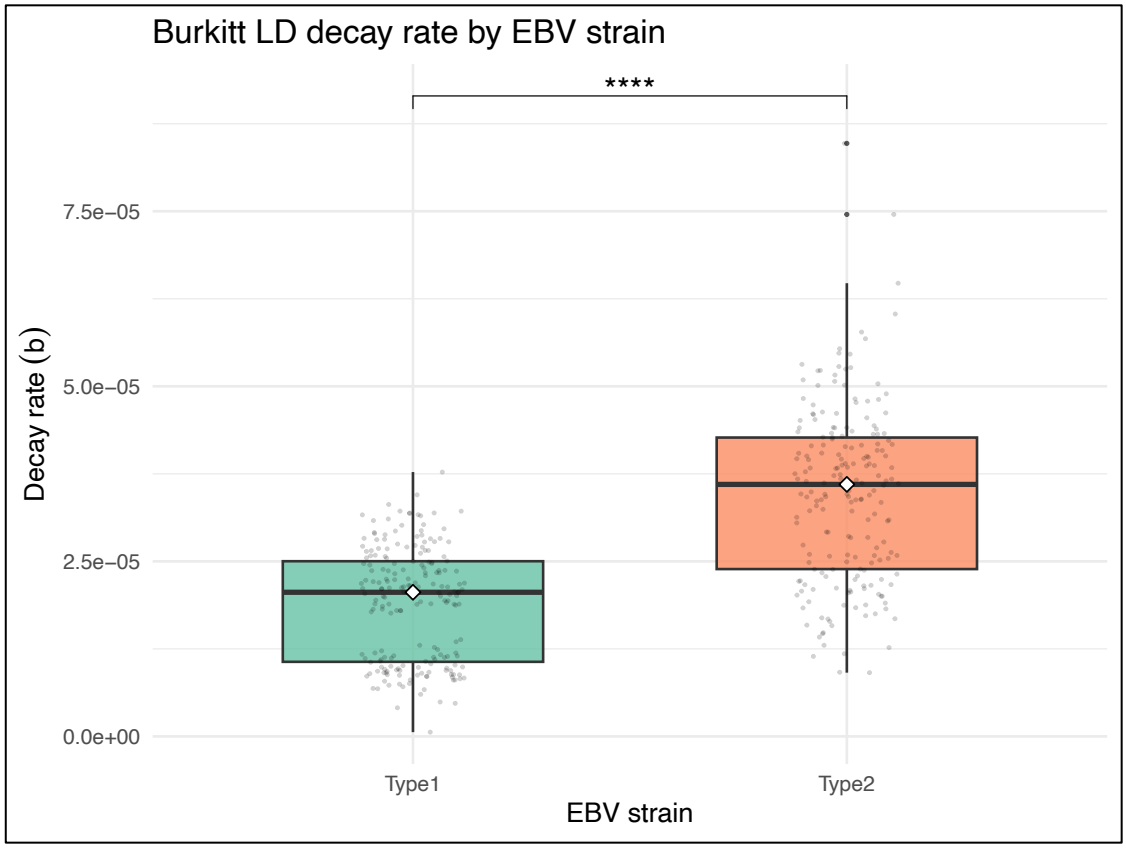

**Supplementary Fig. 5 Per-sample comparison of variant counts identified by three variant calling algorithms**

The number of variants detected per sample by Varscan, VarDict, and Mutect2 was compared. While overall variant counts varied between callers, a significant proportion of variants were shared by at least two callers. Variants identified by at least two callers were deemed high-confidence and used for further analyses. Varscan and VarDict generally detected more variants than Mutect2, suggesting higher sensitivity. However, Mutect2 showed greater agreement with the combined high-confidence set across several samples, suggesting a more conservative calling approach. To increase variant call confidence and reduce caller-specific artefacts, only variants detected by at least two independent algorithms were retained for downstream analysis.




**Supplementary Fig. 6 Burkitt lymphoma EBV strain-specific LD decay patterns**

(A) Binned linkage disequilibrium (LD) decay plots showing mean pairwise LD (r^2) across genomic distance for EBV Type 1 and Type 2 genomes derived from Burkitt lymphoma (BL) samples. Points represent mean r^2 values within 1 kb distance bins, with point size proportional to the number of SNP pairs per bin. (B) LOESS-smoothed LD decay curves with bootstrap-derived 95% confidence intervals showing the decline in LD with increasing genomic distance. Type 2 EBV genomes exhibited more rapid LD decay than Type 1 genomes, consistent with increased recombination or reduced haplotype conservation.



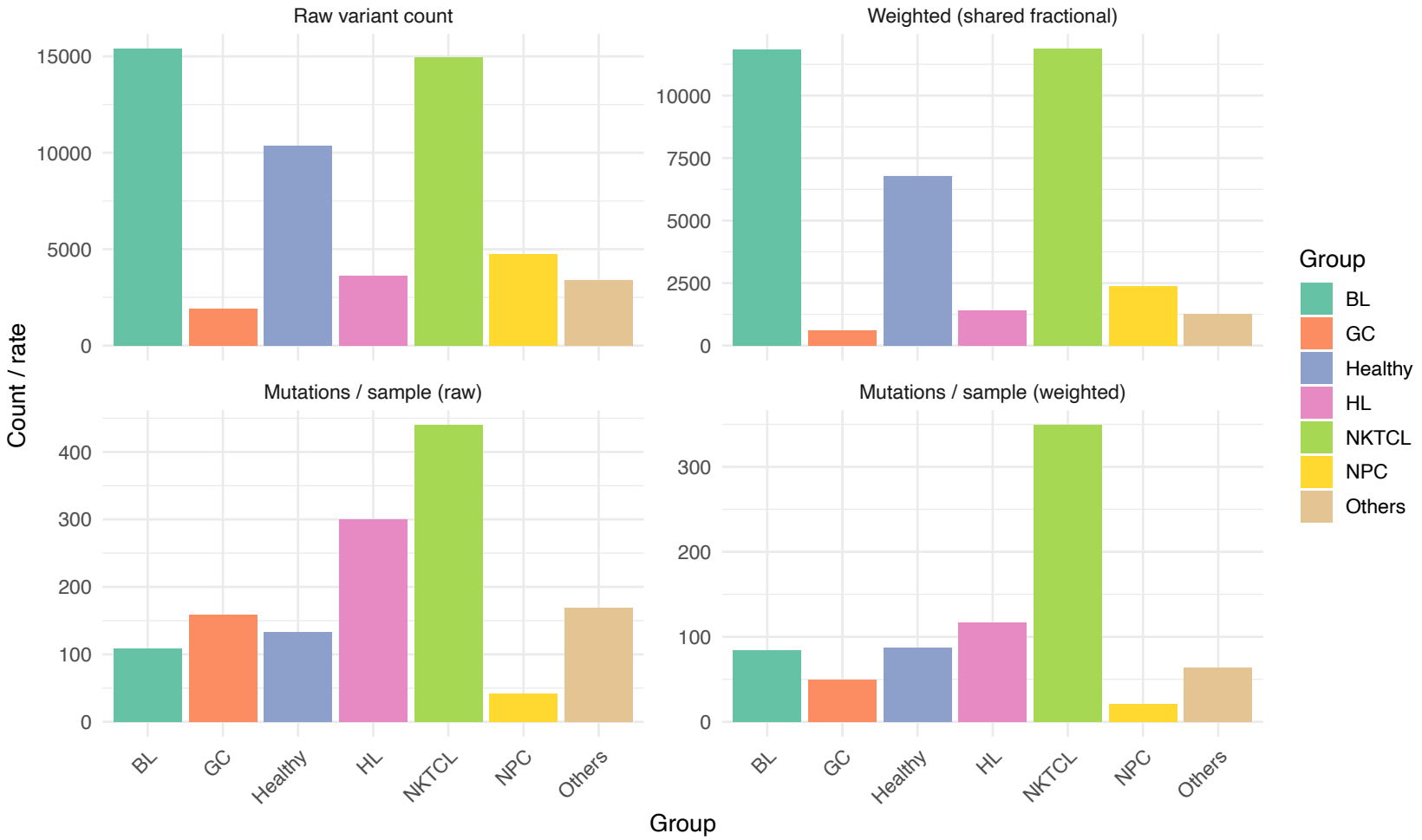

**Supplementary Fig. 7 Comparison of LD decay-rate estimates between EBV Type 1 and Type 2 Burkitt lymphoma genomes**

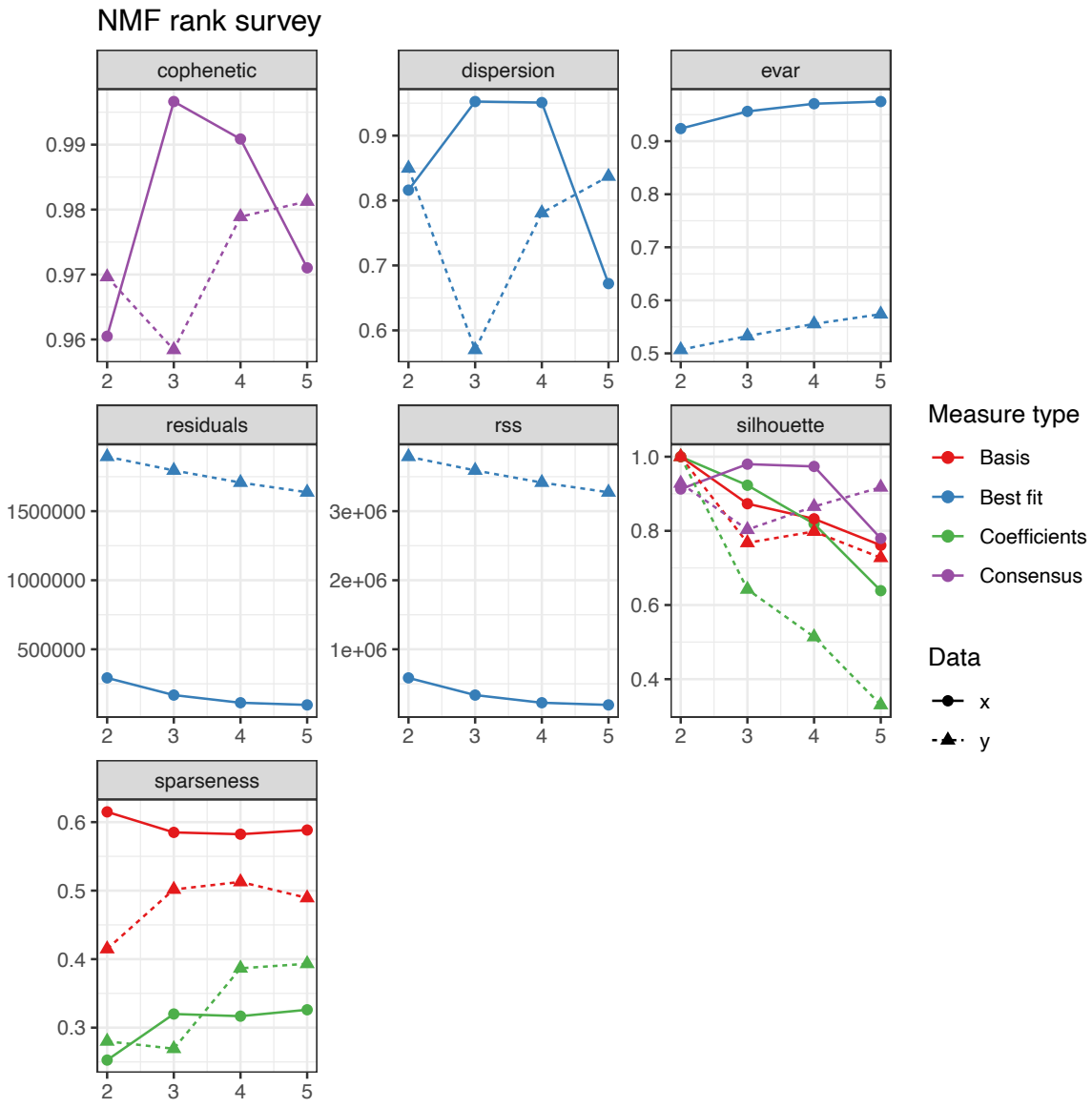
Distribution of bootstrap-estimated LD decay-rate parameters (b) derived from non-linear exponential modelling of LD decay curves in Type 1 and Type 2 BL-associated EBV genomes. Higher b values indicate faster LD decay. Type 2 genomes showed significantly higher decay-rate estimates than Type 1 genomes (Mann–Whitney U test, $p < 2.2 \times 10^{-16}$). The box plots show the median and interquartile range, with overlaid points representing bootstrap replicates.

**Supplementary Fig. 8 EBV genome mutation load across six disease phenotypes and healthy carriers**

Phenotypes assessed included BL, NPC, NKTCL, HL, GC, Healthy, and other EBV-associated cancers. For each group, raw variant counts were computed. To account for cross-group variant sharing, a fractional-weighting approach was employed, and weighted variant counts per group were then calculated as the sum of fractional weights across all variants observed in that group. This method reduced inflation of variant counts caused by broadly shared variants. To further adjust for differences in cohort size, both raw and weighted variant counts were normalised by the number of samples in each group, yielding per-sample mutation-burden estimates.

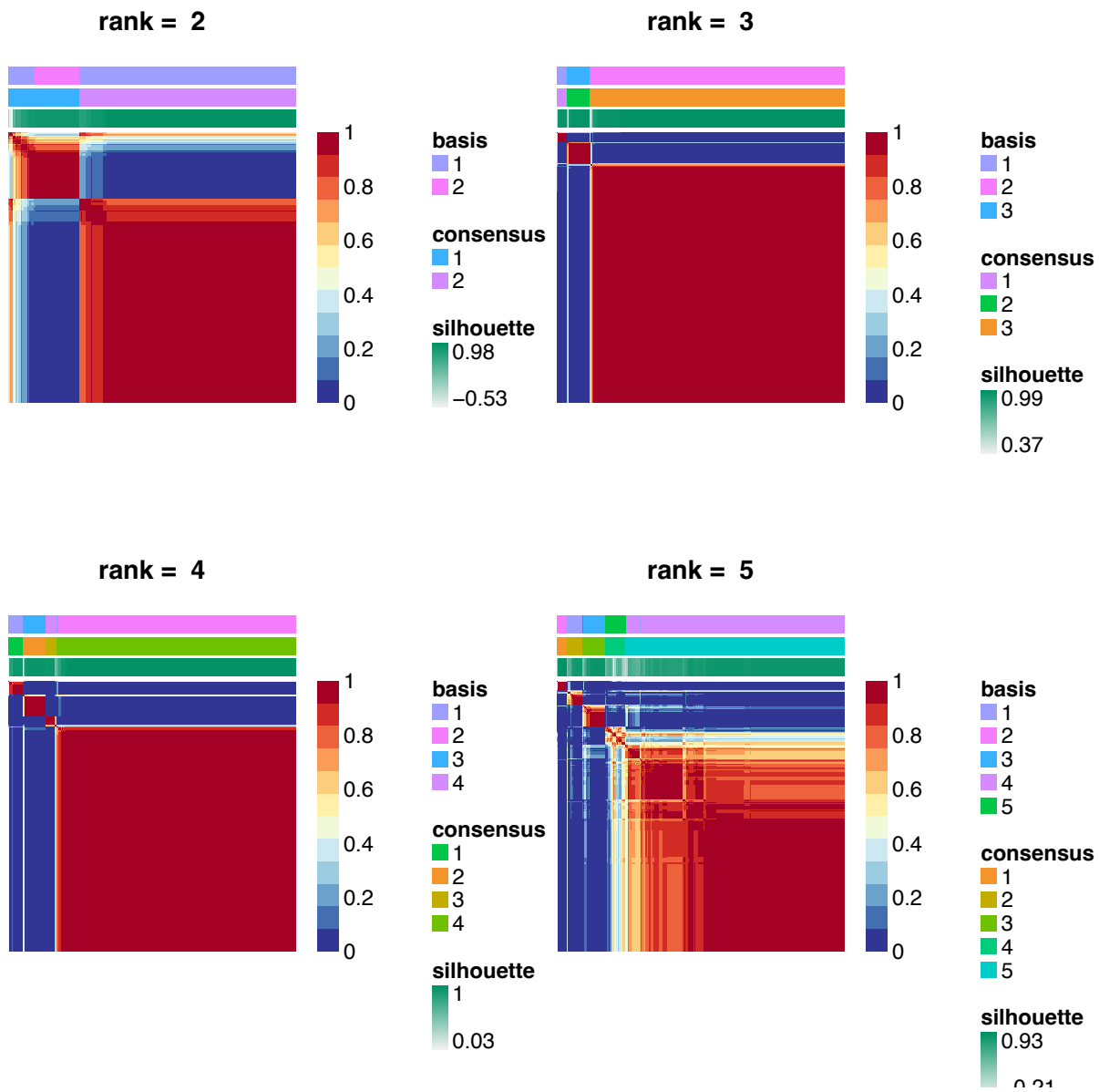
Bar plots illustrate mutation patterns across six disease phenotypes and healthy carriers. The top plots present total raw counts and weighted counts (shared fractions). The bottom plots show per-sample (raw) variant counts and weighted variant counts after adjusting for sharing.


Raw variant count
Weighted (shared fractional)
Mutations / sample (raw)
Mutations / sample (weighted)
Count / rate
Group
BL
GC
Healthy
HL
NKTCL
NPC
Others
0
5000
10000
15000
2500
7500
100
200
300
400



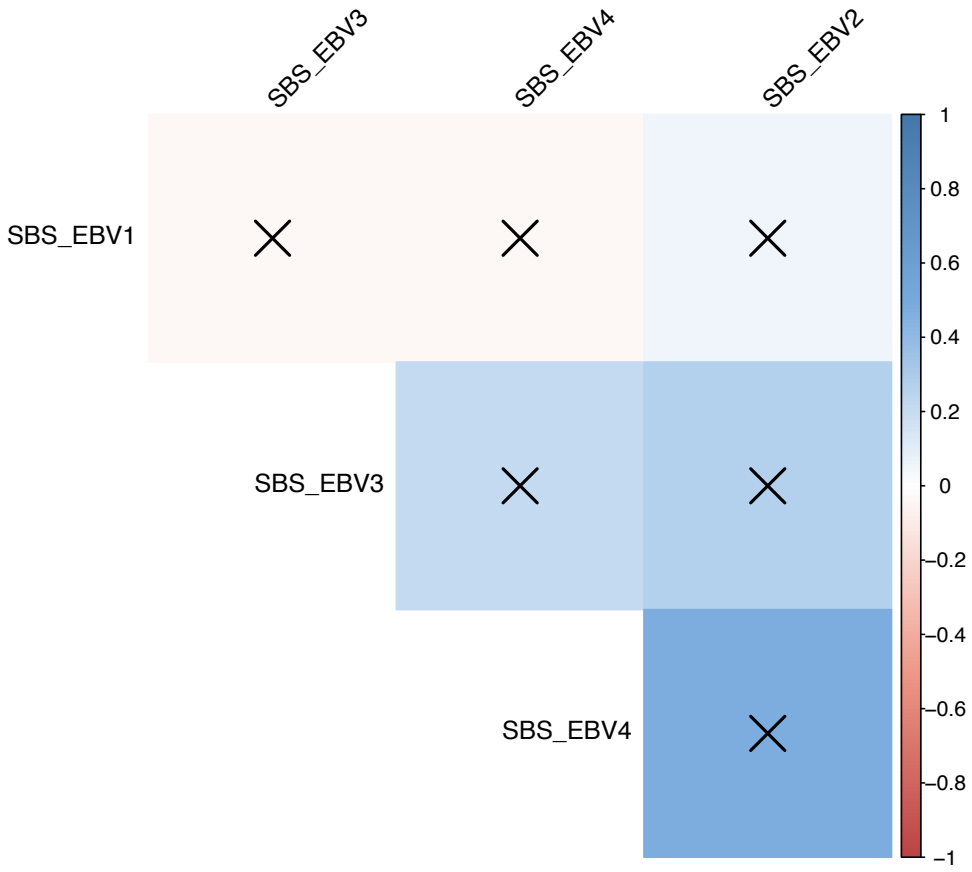

**Supplementary Fig. 9 NMF rank survey metrics**

Metric performance across 50 NMF runs for k = 2 to 5. The cophenetic correlation coefficient (CCC), dispersion, silhouette width, residual sum of squares (RSS), explained variance (evar), and sparseness of the factorised matrices are presented. Rank k = 3 indicated the most parsimonious solution. However, rank k = 4 exhibited a more comprehensive data structure while maintaining comparable stability, interpretability, and clustering quality.



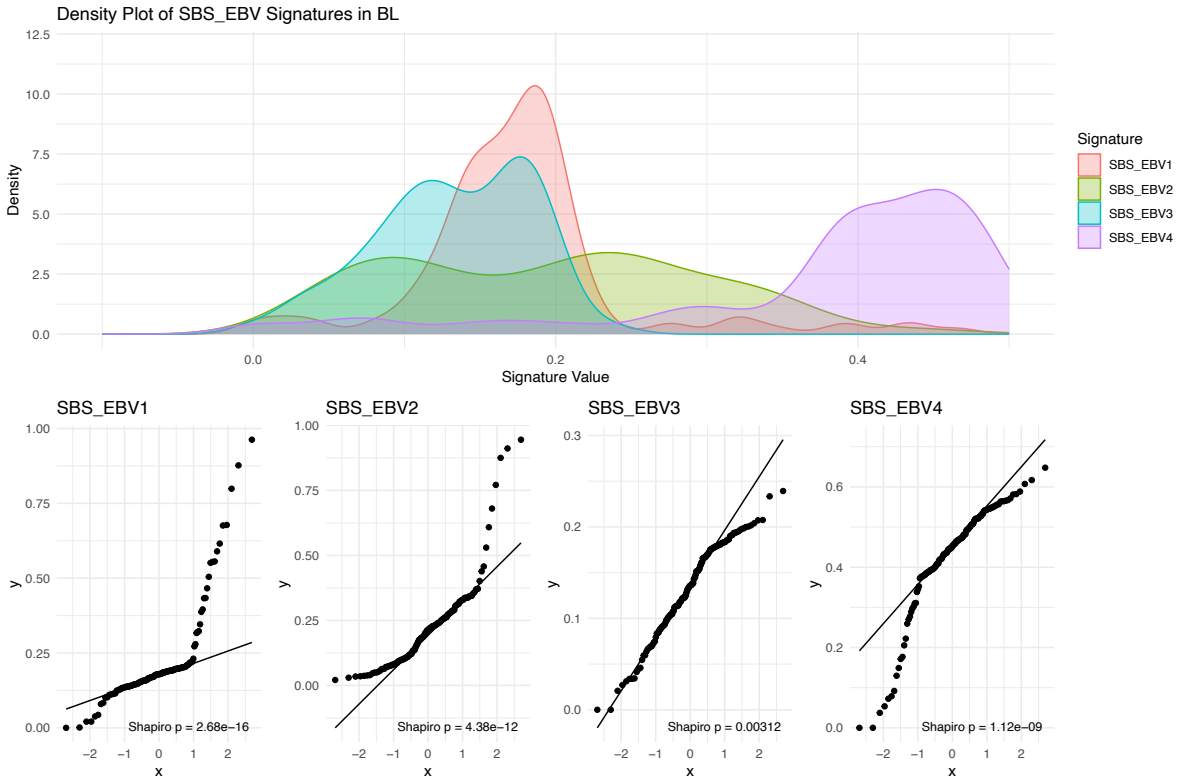

**Supplementary Fig. 10 Consensus clustering heatmaps for NMF ranks**

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 the frequency of co-clustering across NMF runs. Silhouette widths are shown for each rank, reflecting cluster stability. Rank k = 4 demonstrated the highest silhouette score (1.0) with distinct consensus blocks, supporting its selection as the optimal number of mutational signatures.



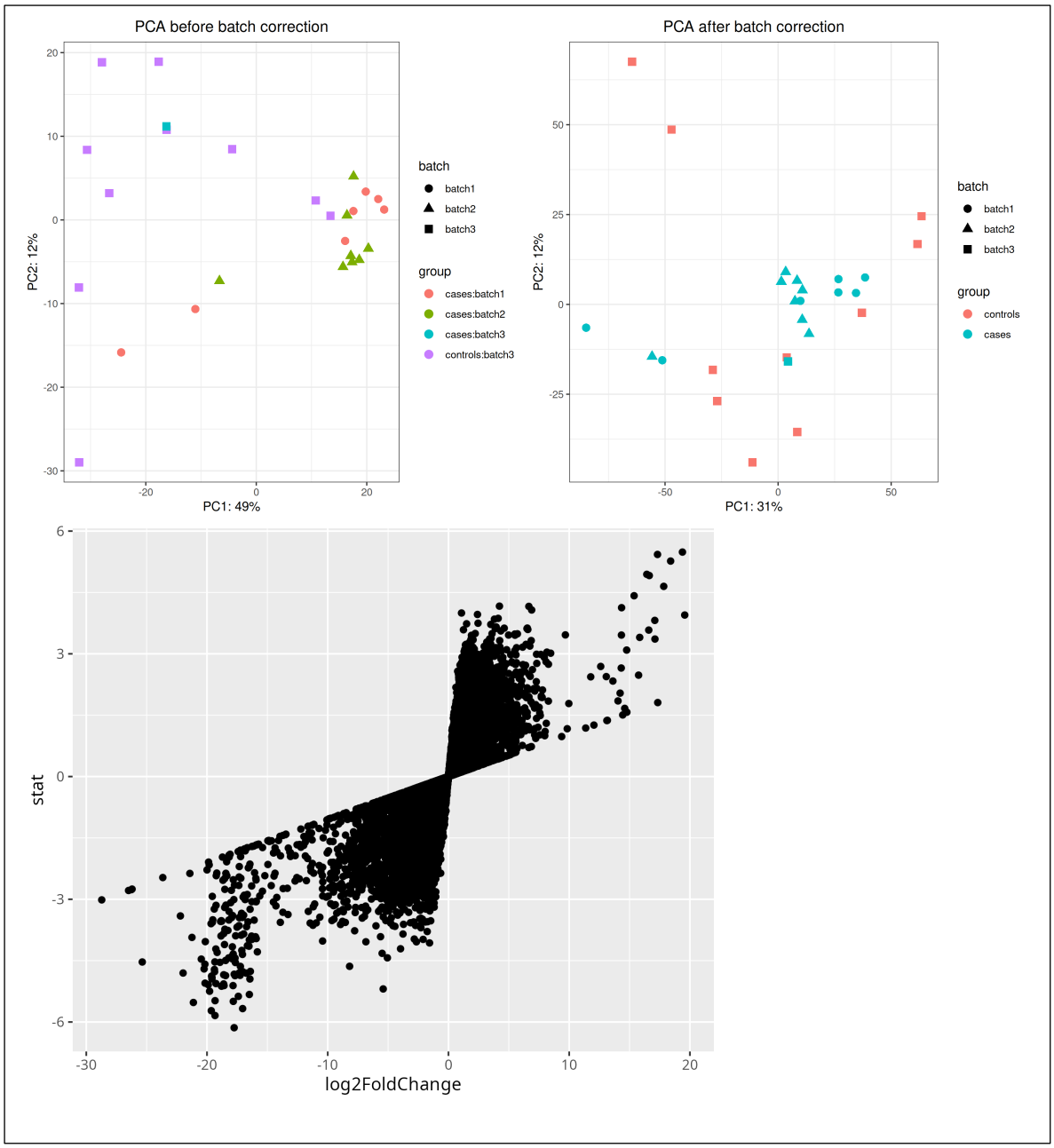

**Supplementary Fig. 11 Spearman's rank correlation matrix of EBV mutation signatures**

The heatmap displays pairwise Spearman's correlation coefficients between signatures (SBS-EBV1-4). The colour scale indicates the strength and direction of correlation. Blue = high positive correlation, white = no correlation, and red = negative correlation. A strong positive correlation between signatures suggests a shared mutational process, while negative correlations may indicate mutually exclusive activity across samples.




**Supplementary Fig. 12 Shapiro-Wilk test results of mutational signature distributions in Burkitt lymphoma**

Density plots (top) and quantile–quantile (QQ) plots (bottom) for EBV-specific signatures SBS_EBV1 to SBS_EBV4 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 deviate significantly from normality ($p < 0.05$), indicating that non-parametric methods are appropriate for downstream comparative analyses.




**Supplementary Fig. 13 Evaluation of batch effects and correlation between effect size and statistical significance**

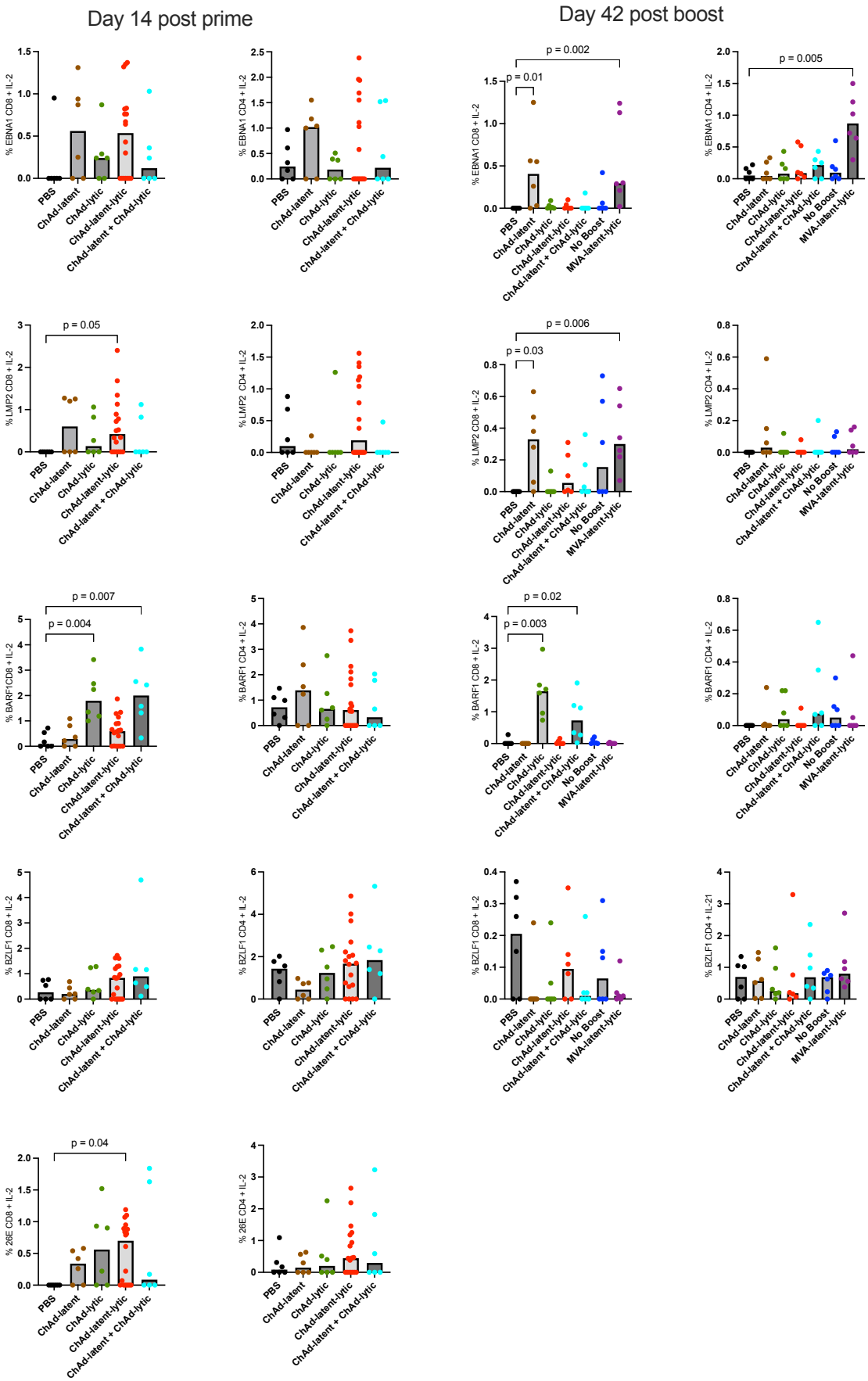
**Top:** The scatter plots display the principal components (PC1 vs PC2), with axes labelled by the proportion of variance explained by each PC. Points are coloured by group and shaped by batch. Before batch correction, the first principal component (PC1) accounted for 49% of the variance, indicating strong batch-related differences. (B) After correction, the variance explained by PC1 decreased to 31%, confirming effective mitigation of batch-related variation, while PC2 (12%) consistently represented biological group differences between samples.

**Bottom**: Scatter plot showing the relationship between $\log_2$ fold change and Wald statistic from DESeq2 differential expression analysis. Each point represents a gene, with the x-axis showing the estimated $\log_2$ fold change (cases vs. controls) and the y-axis showing the corresponding Wald test statistic. The plot illustrates the correlation between effect size and statistical significance across all expressed genes.

**Supplementary Fig. 14 Antigen-specific IL-2 T-cell responses following prime and boost vaccination**

The figure shows antigen-specific IL-2 T cell responses following prime and boost vaccinations, measured by intracellular cytokine staining at 2 weeks after days 14 and 42, respectively. Cytokine measurements were performed on PBMCs following intracellular staining. The plots have been organised by peptide and T cell lineage. The peptides stimulated include EBNA1, LMP2A, BARF1, BZLF1, 26E antigen epitopes. Each dot represents an individual mouse; horizontal lines indicate medians. Statistical comparisons between vaccine groups were performed using Kruskal–Wallis testing with Dunn's multiple-comparison correction; only significant comparisons versus PBS controls are shown.

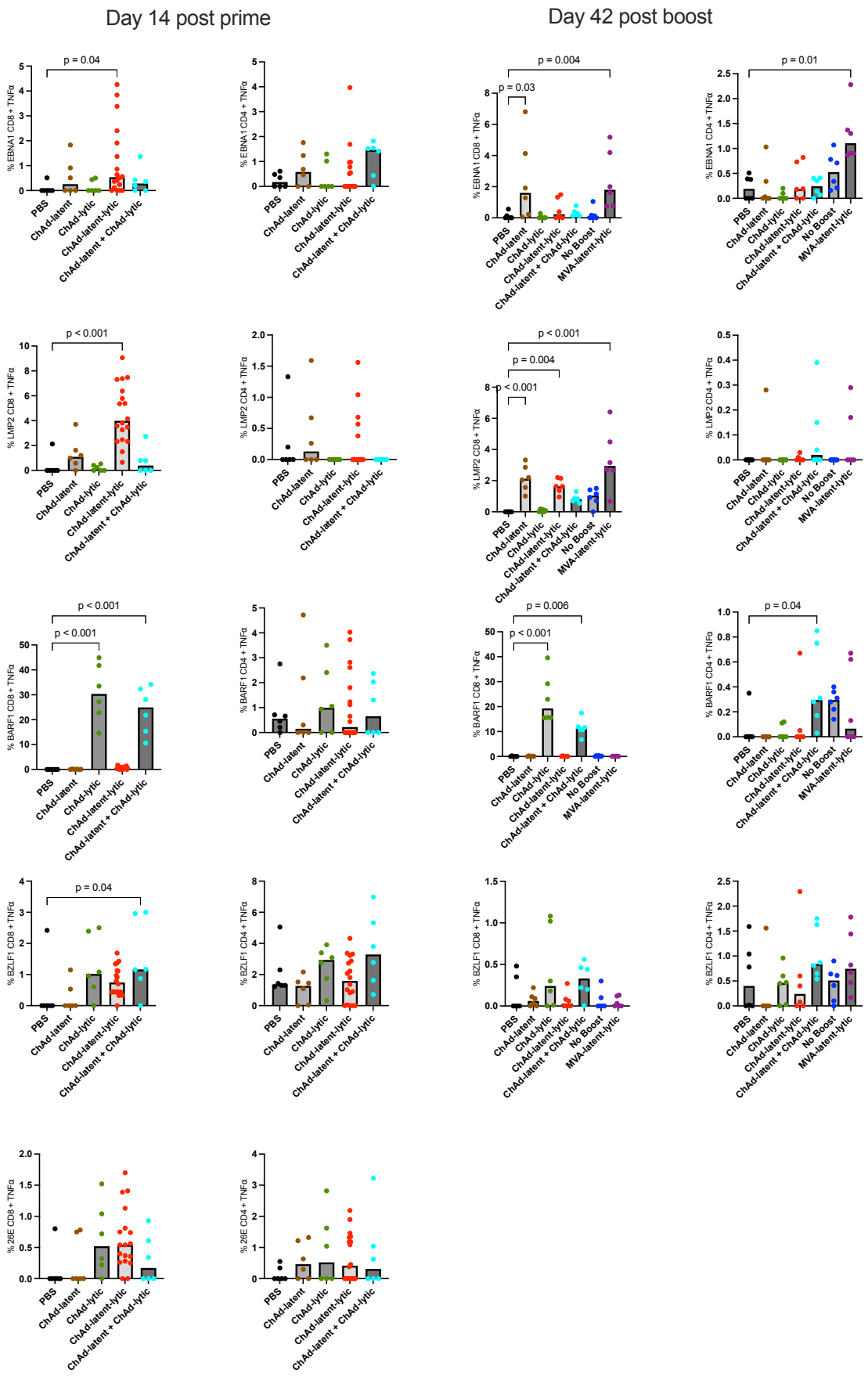
**Antigen-specific IL-2 T-cell responses following prime and boost vaccination**



**Supplementary Fig. 15 Antigen-specific TNF-α T-cell responses following prime and boost vaccination**

The figure shows antigen-specific TNF-α T-cell responses following prime and boost vaccinations, measured by intracellular cytokine staining at 2 weeks after days 14 and 42, respectively. Cytokine measurements were performed on PBMCs following intracellular staining. The plots have organised by peptide and T cell lineage. The peptides stimulated include EBNA1, LMP2A, BARF1, BZLF1, 26E antigen epitopes. Each dot represents an individual mouse; horizontal lines indicate medians. Statistical comparisons between vaccine groups were performed using Kruskal–Wallis testing with Dunn's multiple-comparison correction; only significant comparisons versus PBS controls are shown.

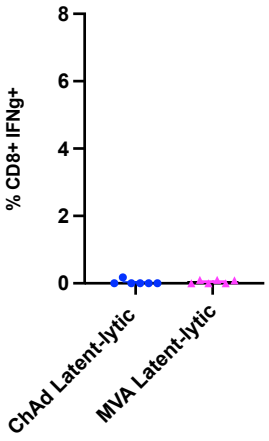
**Antigen-specific TNFα T-cell responses following prime and boost vaccination**

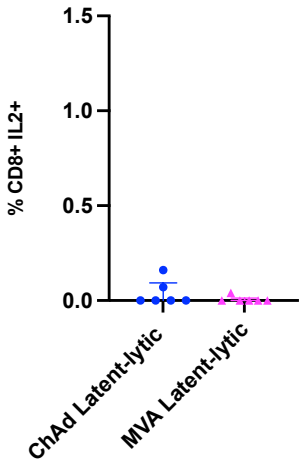


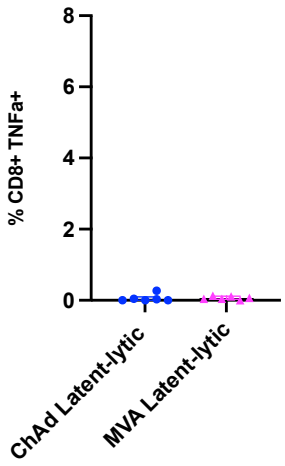
**Supplementary Fig. 16 Cytokine responses to EBV peptides in Chad-primed animals boosted with either ChAd or MVA**

Mice were primed with a ChAd latent-lytic vaccine and, four weeks later, boosted with either a ChAd latent-lytic or an MVA vector encoding the same antigen (latent-lytic). Rows = peptides (BARF1, BZLF1, LMP2A); columns = cytokines. Points = individual animals; bars = median ± IQR. Comparisons (ChAd vs MVA) were made using two-tailed Mann–Whitney U tests; p-values are shown.

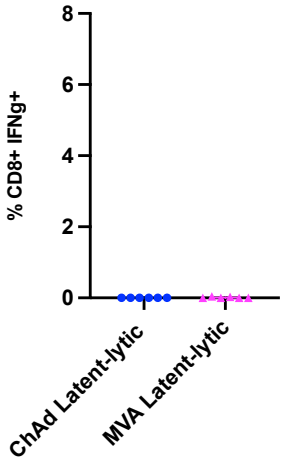
BARF1

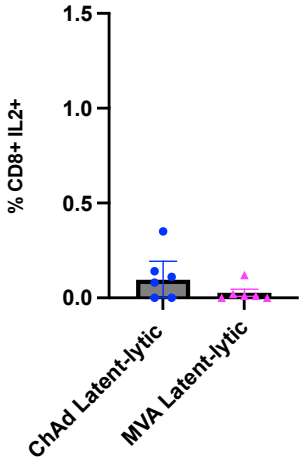


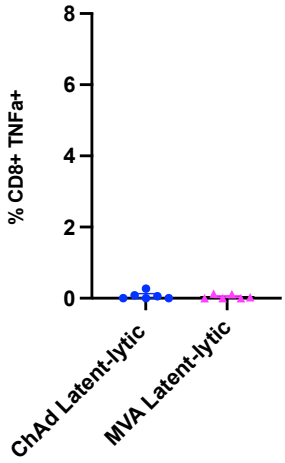



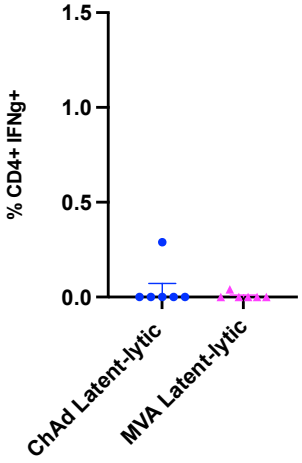

BZLF1

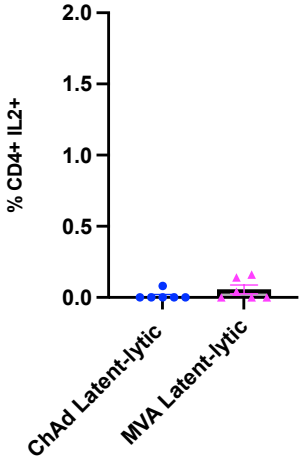


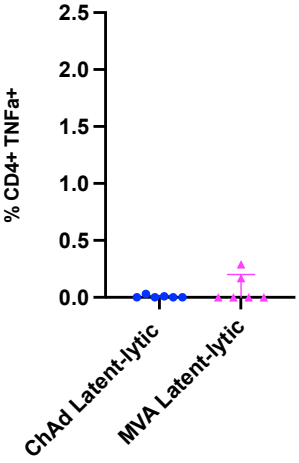




LMP2



## 7.2 Supplementary Tables

**Supplementary Table 1. Pairwise comparisons of LD decay rates across EBV-associated clinical phenotypes**

| Comparison | Group1 | Group2 | FDR adjusted P-value | signif |
|---|---|---|---|---|
| BL - NKTCL | BL | NKTCL | 2.81E-83 | **** |
| BL - HL | BL | HL | 1.44E-77 | **** |
| NKTCL - NPC | NKTCL | NPC | 6.51E-77 | **** |
| HL - NPC | HL | NPC | 2.81E-71 | **** |
| Healthy - NKTCL | Healthy | NKTCL | 1.90E-66 | **** |
| Healthy - HL | Healthy | HL | 3.40E-61 | **** |
| NKTCL - Other | NKTCL | Other | 5.94E-30 | **** |
| HL - Other | HL | Other | 1.67E-26 | **** |
| BL - Other | BL | Other | 1.37E-15 | **** |
| NPC - Other | NPC | Other | 8.23E-13 | **** |
| Healthy - Other | Healthy | Other | 6.82E-09 | **** |
| BL - Healthy | BL | Healthy | 0.03376729 | * |
| Healthy - NPC | Healthy | NPC | 0.198599404 | ns |
| BL - NPC | BL | NPC | 0.424990745 | ns |
| HL - NKTCL | HL | NKTCL | 0.469591473 | ns |

**** $p \leq 1\times10^{-4}$, *** $p \leq 1\times10^{-3}$, ** $p \leq 1\times10^{-2}$, * $p \leq 0.05$, ns = not significant

**Supplementary Table 2. Putative recombination events identified in EBV genomes from BL and healthy controls**

| Breakpoint | 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 |

**Supplementary Table 3. Differential mutation enrichment in Burkitt lymphoma (BL) compared with Healthy controls**

| Gene | OR | log2OR | CI.lower | CI.upper | p.value | FDR | Sig.direction |
|---|---|---|---|---|---|---|---|
| BHLF1 | 7.68 | 2.94 | 2.61 | 3.28 | 8.50E-122 | 5.35E-120 | Enriched |
| BOLF1 | 0.60 | -0.73 | -0.84 | -0.62 | 4.36E-37 | 1.37E-35 | Depleted |
| LMP-2B | 0.22 | -2.21 | -2.61 | -1.82 | 4.43E-34 | 9.30E-33 | Depleted |
| LMP-1 | 1.56 | 0.64 | 0.52 | 0.77 | 1.64E-25 | 2.59E-24 | Enriched |
| EBNA-3A | 1.51 | 0.60 | 0.48 | 0.72 | 3.10E-24 | 3.90E-23 | Enriched |
| BHRF1 | 0.35 | -1.50 | -1.89 | -1.11 | 2.30E-15 | 2.41E-14 | Depleted |
| BRRF2 | 0.65 | -0.63 | -0.79 | -0.46 | 3.30E-14 | 2.97E-13 | Depleted |
| BLLF1 | 1.35 | 0.44 | 0.31 | 0.56 | 2.12E-12 | 1.67E-11 | Enriched |
| BALF2 | 0.62 | -0.70 | -0.89 | -0.50 | 3.40E-12 | 2.26E-11 | Depleted |
| BVRF1 | 0.53 | -0.91 | -1.17 | -0.65 | 3.58E-12 | 2.26E-11 | Depleted |
| BVRF2 | 0.53 | -0.92 | -1.21 | -0.64 | 8.17E-11 | 4.68E-10 | Depleted |
| BBLF4 | 1.54 | 0.62 | 0.41 | 0.84 | 1.81E-09 | 9.52E-09 | Enriched |
| BRLF1 | 1.54 | 0.62 | 0.41 | 0.83 | 2.05E-09 | 9.94E-09 | Enriched |
| BZLF1 | 0.65 | -0.63 | -0.84 | -0.42 | 3.47E-09 | 1.56E-08 | Depleted |
| BPLF1 | 0.85 | -0.24 | -0.31 | -0.16 | 3.99E-09 | 1.68E-08 | Depleted |
| BSLF2/BMLF1 | 0.58 | -0.79 | -1.07 | -0.52 | 1.43E-08 | 5.65E-08 | Depleted |
| BBLF1 | 0.42 | -1.24 | -1.75 | -0.73 | 6.03E-07 | 2.24E-06 | Depleted |
| LMP-2A | 0.84 | -0.24 | -0.34 | -0.15 | 1.12E-06 | 3.92E-06 | Depleted |
| BNLF2a | 2.24 | 1.16 | 0.64 | 1.72 | 2.45E-06 | 7.80E-06 | Enriched |
| EBNA-2 | 0.78 | -0.36 | -0.51 | -0.21 | 2.48E-06 | 7.80E-06 | Depleted |
| BDLF3 | 1.36 | 0.45 | 0.25 | 0.64 | 3.15E-06 | 9.45E-06 | Enriched |
| BALF5 | 1.44 | 0.53 | 0.30 | 0.76 | 3.57E-06 | 1.02E-05 | Enriched |
| BGLF4 | 0.49 | -1.02 | -1.52 | -0.53 | 2.92E-05 | 8.00E-05 | Depleted |
| BBLF2/BBLF3 | 1.25 | 0.32 | 0.15 | 0.49 | 1.71E-04 | 4.49E-04 | Enriched |
| BKRF3 | 2.02 | 1.01 | 0.45 | 1.61 | 1.94E-04 | 4.88E-04 | Enriched |
| BXLF1 | 0.68 | -0.55 | -0.85 | -0.26 | 2.34E-04 | 5.66E-04 | Depleted |
| BARF1 | 1.81 | 0.86 | 0.38 | 1.37 | 2.54E-04 | 5.93E-04 | Enriched |
| BBRF2 | 0.58 | -0.79 | -1.25 | -0.34 | 4.59E-04 | 0.001033149 | Depleted |
| BNRF1 | 1.22 | 0.29 | 0.12 | 0.45 | 5.19E-04 | 0.001127151 | Enriched |
| BXLF2 | 0.71 | -0.49 | -0.78 | -0.21 | 6.98E-04 | 0.001466613 | Depleted |
| BBRF1 | 1.34 | 0.42 | 0.17 | 0.67 | 8.69E-04 | 0.001766023 | Enriched |
| BSRF1 | 2.35 | 1.23 | 0.43 | 2.12 | 0.001320738 | 0.002600203 | Enriched |
| BLRF2 | 1.97 | 0.98 | 0.34 | 1.67 | 0.001516195 | 0.002894553 | Enriched |
| BaRF1 | 1.96 | 0.97 | 0.32 | 1.68 | 0.002306059 | 0.004272992 | Enriched |
| BGLF1 | 0.84 | -0.25 | -0.44 | -0.07 | 0.005493877 | 0.009888979 | Depleted |
| BKRF4 | 0.77 | -0.37 | -0.65 | -0.09 | 0.008364389 | 0.01463768 | Depleted |
| BALF3 | 0.84 | -0.26 | -0.45 | -0.06 | 0.009204661 | 0.0156728 | Depleted |
| BMRF2 | 1.41 | 0.50 | 0.06 | 0.95 | 0.023165102 | 0.038405301 | Enriched |
| BGLF5 | 1.17 | 0.22 | 0.03 | 0.42 | 0.025520165 | 0.041224883 | Enriched |

**Supplementary Table 4. Differentially expressed human genes in EBV-positive Burkitt lymphoma**

| gene | baseMean | log2FoldChange | lfcSE | stat | p.value | padj |
|---|---|---|---|---|---|---|
| SPATS1 | 16.39 | -17.75 | 2.89 | -6.14 | 0.00000 | 0.0000 |
| HOXA3 | 98.01 | -19.34 | 3.31 | -5.84 | 0.00000 | 0.0000 |
| PTPRH | 21.04 | -19.65 | 3.43 | -5.72 | 0.00000 | 0.0001 |
| NPHP3-ACAD11 | 40.25 | -17.05 | 3.01 | -5.67 | 0.00000 | 0.0001 |
| C17orf114 | 2.30 | 19.38 | 3.53 | 5.49 | 0.00000 | 0.0001 |
| EVA1A | 25.72 | -19.33 | 3.53 | -5.48 | 0.00000 | 0.0001 |
| HOXA10 | 33.63 | -21.13 | 3.83 | -5.52 | 0.00000 | 0.0001 |
| VSTM5 | 10.02 | -17.82 | 3.24 | -5.50 | 0.00000 | 0.0001 |
| PKIB | 6.40 | -17.40 | 3.24 | -5.37 | 0.00000 | 0.0001 |
| PNLDC1 | 11.03 | -16.49 | 3.10 | -5.33 | 0.00000 | 0.0002 |
| DLL3 | 15.13 | 18.40 | 3.49 | 5.27 | 0.00000 | 0.0002 |
| CHRM4 | 11.00 | -19.79 | 3.77 | -5.25 | 0.00000 | 0.0002 |
| CERCAM | 159.07 | -5.41 | 1.04 | -5.19 | 0.00000 | 0.0003 |
| NME1-NME2 | 16.78 | -18.80 | 3.67 | -5.13 | 0.00000 | 0.0003 |
| PIERCE2 | 4.74 | -17.84 | 3.49 | -5.11 | 0.00000 | 0.0003 |
| TRPC4 | 13.25 | -18.59 | 3.64 | -5.11 | 0.00000 | 0.0003 |
| KCNA1 | 12.11 | -19.89 | 3.91 | -5.09 | 0.00000 | 0.0004 |
| FLRT3 | 30.89 | -20.16 | 3.99 | -5.05 | 0.00000 | 0.0004 |
| NPIPA7 | 24.41 | -18.62 | 3.67 | -5.07 | 0.00000 | 0.0004 |
| NYAP1 | 10.99 | -19.26 | 3.81 | -5.06 | 0.00000 | 0.0004 |
| CHRFAM7A | 30.61 | -19.59 | 3.96 | -4.95 | 0.00000 | 0.0006 |
| GABRA6 | 3.01 | 16.43 | 3.32 | 4.94 | 0.00000 | 0.0006 |
| JPH1 | 12.39 | -16.45 | 3.32 | -4.95 | 0.00000 | 0.0006 |
| FAM180A | 5.60 | -19.59 | 3.99 | -4.91 | 0.00000 | 0.0006 |
| SLC39A2 | 2.49 | 16.64 | 3.39 | 4.91 | 0.00000 | 0.0006 |
| CYP17A1 | 12.94 | -17.68 | 3.63 | -4.87 | 0.00000 | 0.0007 |
| EYA2 | 37.51 | -19.62 | 4.02 | -4.88 | 0.00000 | 0.0007 |
| FOXD1 | 8.31 | -18.46 | 3.79 | -4.87 | 0.00000 | 0.0007 |
| KCNJ12 | 9.46 | -17.26 | 3.55 | -4.86 | 0.00000 | 0.0007 |
| KLK1 | 16.84 | -18.61 | 3.84 | -4.85 | 0.00000 | 0.0007 |
| ZBBX | 63.51 | -17.75 | 3.67 | -4.83 | 0.00000 | 0.0008 |
| DRC7 | 14.46 | -16.75 | 3.49 | -4.81 | 0.00000 | 0.0008 |
| LRRC24 | 18.56 | -17.34 | 3.61 | -4.80 | 0.00000 | 0.0008 |
| VSTM2L | 8.02 | -21.98 | 4.58 | -4.80 | 0.00000 | 0.0008 |
| APOL5 | 3.68 | -17.03 | 3.57 | -4.77 | 0.00000 | 0.0009 |
| UGT2B15 | 9.15 | -19.37 | 4.06 | -4.77 | 0.00000 | 0.0009 |
| KLK7 | 13.26 | -16.41 | 3.45 | -4.76 | 0.00000 | 0.0009 |
| IGFN1 | 25.43 | -19.39 | 4.11 | -4.72 | 0.00000 | 0.0011 |

Differentially expressed human genes in EBV-positive Burkitt lymphoma cont….

| gene | baseMean | log2FoldChange | lfcSE | stat | p.value | padj |
|---|---|---|---|---|---|---|
| KLHDC8A | 28.13 | -20.24 | 4.30 | -4.70 | 0.00000 | 0.0012 |
| ATP6V1G3 | 2.99 | 17.83 | 3.84 | 4.65 | 0.00000 | 0.0015 |
| HOXC4 | 129.68 | -8.19 | 1.77 | -4.64 | 0.00000 | 0.0015 |
| MAPK15 | 59.92 | -17.24 | 3.71 | -4.64 | 0.00000 | 0.0015 |
| DCT | 8.25 | -20.17 | 4.40 | -4.59 | 0.00000 | 0.0019 |
| MYOZ1 | 10.26 | -17.65 | 3.88 | -4.55 | 0.00001 | 0.0022 |
| SLC22A14 | 5.36 | -18.93 | 4.17 | -4.54 | 0.00001 | 0.0022 |
| ENSG00000293606 | 72.67 | -19.35 | 4.27 | -4.53 | 0.00001 | 0.0022 |
| LCN8 | 7.81 | -25.36 | 5.59 | -4.53 | 0.00001 | 0.0022 |
| MFSD6L | 5.83 | -18.46 | 4.11 | -4.50 | 0.00001 | 0.0026 |
| KCNK9 | 7.40 | -20.47 | 4.59 | -4.46 | 0.00001 | 0.0029 |
| RSPH1 | 36.53 | -17.87 | 4.00 | -4.46 | 0.00001 | 0.0029 |
| CDH7 | 45.60 | -5.06 | 1.14 | -4.43 | 0.00001 | 0.0032 |
| HOXA4 | 15.99 | -17.68 | 3.98 | -4.44 | 0.00001 | 0.0032 |
| TPPP3 | 87.49 | -18.20 | 4.11 | -4.43 | 0.00001 | 0.0032 |
| TMSB4Y | 8.88 | 15.38 | 3.48 | 4.42 | 0.00001 | 0.0033 |
| ZNF488 | 6.17 | -17.92 | 4.08 | -4.39 | 0.00001 | 0.0037 |
| FAM43B | 8.16 | -17.06 | 3.92 | -4.35 | 0.00001 | 0.0044 |
| SULT1A1 | 8.56 | -17.83 | 4.11 | -4.34 | 0.00001 | 0.0046 |
| LAMC3 | 222.94 | -5.49 | 1.27 | -4.32 | 0.00002 | 0.0048 |
| LGR5 | 21.27 | -19.16 | 4.46 | -4.30 | 0.00002 | 0.0052 |
| GCM2 | 3.91 | -15.83 | 3.69 | -4.29 | 0.00002 | 0.0055 |
| IDI2 | 5.16 | -17.12 | 4.03 | -4.25 | 0.00002 | 0.0063 |
| HOXB5 | 16.58 | -19.27 | 4.57 | -4.22 | 0.00002 | 0.0073 |
| DENND2A | 193.32 | -3.97 | 0.94 | -4.21 | 0.00003 | 0.0073 |
| FAM72C | 204.65 | 6.66 | 1.60 | 4.16 | 0.00003 | 0.0087 |
| TAFA5 | 5.80 | -17.90 | 4.30 | -4.16 | 0.00003 | 0.0087 |
| TMEM145 | 62.83 | 4.22 | 1.01 | 4.17 | 0.00003 | 0.0087 |
| MGARP | 16.69 | -16.52 | 3.98 | -4.15 | 0.00003 | 0.0091 |
| PHYHIP | 17.47 | -16.65 | 4.03 | -4.13 | 0.00004 | 0.0095 |
| MT1G | 30.51 | 14.34 | 3.47 | 4.13 | 0.00004 | 0.0096 |
| LIPJ | 9.60 | -18.55 | 4.52 | -4.11 | 0.00004 | 0.0104 |
| FREY1 | 16.68 | 6.89 | 1.69 | 4.07 | 0.00005 | 0.0119 |
| GDI1 | 847.54 | -1.57 | 0.39 | -4.07 | 0.00005 | 0.0120 |
| PAK4 | 320.95 | -2.65 | 0.66 | -4.04 | 0.00005 | 0.0131 |
| TCF21 | 154.59 | -6.84 | 1.69 | -4.04 | 0.00005 | 0.0131 |
| PCDHB6 | 19.88 | -20.14 | 4.99 | -4.04 | 0.00005 | 0.0132 |

Differentially expressed human genes in EBV-positive Burkitt lymphoma cont….

| gene | baseMean | log2FoldChange | lfcSE | stat | p.value | padj |
|---|---|---|---|---|---|---|
| PTPRD | 179.35 | -10.43 | 2.59 | -4.02 | 0.00006 | 0.0139 |
| CIMAP1C | 3.53 | -16.45 | 4.12 | -4.00 | 0.00006 | 0.0150 |
| INTS10 | 936.07 | 1.08 | 0.27 | 4.00 | 0.00006 | 0.0150 |
| C1orf167 | 19.71 | -16.50 | 4.13 | -3.99 | 0.00007 | 0.0151 |
| ENSG00000256514 | 41.02 | -16.28 | 4.09 | -3.98 | 0.00007 | 0.0154 |
| FCSK | 270.99 | -2.10 | 0.53 | -3.98 | 0.00007 | 0.0154 |
| NAGLU | 256.93 | -2.83 | 0.71 | -3.97 | 0.00007 | 0.0158 |
| NEU3 | 977.93 | 2.40 | 0.61 | 3.96 | 0.00007 | 0.0162 |
| NPTX2 | 5.80 | -15.92 | 4.02 | -3.96 | 0.00007 | 0.0162 |
| SLC7A3 | 20.87 | 19.57 | 4.96 | 3.95 | 0.00008 | 0.0168 |
| AMPD1 | 12.27 | -21.25 | 5.40 | -3.93 | 0.00008 | 0.0178 |
| CCDC183 | 88.99 | -5.63 | 1.44 | -3.92 | 0.00009 | 0.0186 |
| LYPD6B | 30.65 | -15.99 | 4.08 | -3.92 | 0.00009 | 0.0186 |
| PVALEF | 4.59 | -16.31 | 4.18 | -3.90 | 0.00010 | 0.0196 |
| GPR20 | 6.45 | -18.84 | 4.83 | -3.90 | 0.00010 | 0.0196 |
| FAM81B | 9.15 | -17.63 | 4.53 | -3.89 | 0.00010 | 0.0198 |
| TMPRSS15 | 42.26 | -17.16 | 4.43 | -3.88 | 0.00011 | 0.0207 |
| SARS2 | 199.33 | 4.12 | 1.07 | 3.87 | 0.00011 | 0.0215 |
| PRKG1 | 142.25 | -3.77 | 0.98 | -3.85 | 0.00012 | 0.0226 |
| SMIM39 | 11.69 | -16.96 | 4.41 | -3.85 | 0.00012 | 0.0226 |
| ZCCHC10 | 279.79 | 3.81 | 0.99 | 3.85 | 0.00012 | 0.0226 |
| MYO18B | 28.18 | -18.61 | 4.85 | -3.84 | 0.00012 | 0.0233 |
| THSD8 | 1.73 | 17.09 | 4.48 | 3.82 | 0.00014 | 0.0251 |
| MMUT | 196.26 | -1.76 | 0.46 | -3.79 | 0.00015 | 0.0278 |
| TRABD2B | 43.27 | -7.77 | 2.06 | -3.77 | 0.00016 | 0.0296 |
| RBM24 | 17.06 | -18.27 | 4.86 | -3.76 | 0.00017 | 0.0304 |
| DNAJC27 | 136.61 | 2.45 | 0.65 | 3.75 | 0.00018 | 0.0318 |
| NKAIN1 | 18.58 | -18.52 | 4.96 | -3.73 | 0.00019 | 0.0332 |
| PSME4 | 1934.61 | 1.50 | 0.40 | 3.73 | 0.00019 | 0.0332 |
| MAIP1 | 75.51 | 3.51 | 0.94 | 3.72 | 0.00020 | 0.0348 |
| BRD3OS | 309.60 | -2.28 | 0.62 | -3.70 | 0.00022 | 0.0376 |
| MYH2 | 55.99 | -17.51 | 4.76 | -3.68 | 0.00023 | 0.0393 |
| H3C14 | 287.97 | -16.72 | 4.56 | -3.67 | 0.00024 | 0.0409 |
| VLDLR | 236.55 | -4.43 | 1.21 | -3.66 | 0.00025 | 0.0418 |
| ITPRIPL1 | 364.75 | 3.95 | 1.08 | 3.66 | 0.00026 | 0.0422 |
| FGA | 66.41 | -17.35 | 4.76 | -3.64 | 0.00027 | 0.0429 |
| MAP2 | 96.86 | -6.00 | 1.65 | -3.65 | 0.00026 | 0.0429 |

Differentially expressed human genes in EBV-positive Burkitt lymphoma cont….

| gene | baseMean | log2FoldChange | lfcSE | stat | p.value | padj |
|---|---|---|---|---|---|---|
| MOB2 | 236.29 | -2.33 | 0.64 | -3.64 | 0.00027 | 0.0429 |
| PCDHB14 | 53.86 | -4.59 | 1.26 | -3.65 | 0.00027 | 0.0429 |
| NHSL2 | 337.67 | -2.73 | 0.75 | -3.64 | 0.00027 | 0.0431 |
| GCOM1 | 14.65 | -11.23 | 3.09 | -3.64 | 0.00028 | 0.0435 |
| AARD | 20.38 | 6.53 | 1.80 | 3.63 | 0.00028 | 0.0443 |
| ZNF823 | 41.68 | 4.62 | 1.27 | 3.63 | 0.00029 | 0.0443 |
| CCDC188 | 31.18 | -8.94 | 2.47 | -3.62 | 0.00030 | 0.0451 |
| ENSG00000293570 | 86.64 | 3.88 | 1.07 | 3.62 | 0.00030 | 0.0451 |
| TERT | 1804.61 | 6.50 | 1.80 | 3.62 | 0.00030 | 0.0451 |
| NES | 254.76 | -3.72 | 1.03 | -3.61 | 0.00030 | 0.0452 |
| ENSG00000285982 | 50.98 | -19.67 | 5.47 | -3.60 | 0.00032 | 0.0477 |
| MYT1 | 53.06 | 6.57 | 1.83 | 3.59 | 0.00032 | 0.0477 |
| DIMT1 | 548.22 | 1.24 | 0.35 | 3.59 | 0.00033 | 0.0486 |
| ELN | 369.12 | -4.79 | 1.34 | -3.58 | 0.00034 | 0.0489 |
| NRK | 49.93 | -11.46 | 3.20 | -3.58 | 0.00034 | 0.0489 |
| RPRML | 5.47 | 16.59 | 4.63 | 3.58 | 0.00034 | 0.0489 |
| PMF1-BGLAP | 19.72 | -10.95 | 3.06 | -3.58 | 0.00035 | 0.0490 |
| CPAMD8 | 124.82 | -3.15 | 0.88 | -3.57 | 0.00035 | 0.0493 |

**Supplementary Table 5. Mixed-effects analysis of vaccine-specific immunodominance hierarchies**

| contrast | estimate | SE | df | t.ratio | p.value |
|---|---|---|---|---|---|
| **Chad latent-only** | | | | | |
| BARF1 - BZLF1 | -0.25 | 0.366 | 162 | -0.682 | 0.5512 |
| BARF1 - EBNA1 | 1.167 | 0.366 | 162 | 3.183 | 0.0025 |
| BARF1 - LMP2 | 2.667 | 0.366 | 162 | 7.276 | <.0001 |
| BARF1 - X26E | -0.25 | 0.366 | 162 | -0.682 | 0.5512 |
| BZLF1 - EBNA1 | 1.417 | 0.366 | 162 | 3.865 | 0.0003 |
| BZLF1 - LMP2 | 2.917 | 0.366 | 162 | 7.958 | <.0001 |
| BZLF1 - X26E | 0 | 0.366 | 162 | 0 | 1 |
| EBNA1 - LMP2 | 1.5 | 0.366 | 162 | 4.093 | 0.0002 |
| EBNA1 - X26E | -1.417 | 0.366 | 162 | -3.865 | 0.0003 |
| LMP2 - X26E | -2.917 | 0.366 | 162 | -7.958 | <.0001 |
| **Chad latent-lytic** | | | | | |
| BARF1 - BZLF1 | -0.611 | 0.212 | 162 | -2.888 | 0.0055 |
| BARF1 - EBNA1 | 0.611 | 0.212 | 162 | 2.888 | 0.0055 |
| BARF1 - LMP2 | 2.444 | 0.212 | 162 | 11.552 | <.0001 |
| BARF1 - X26E | -0.222 | 0.212 | 162 | -1.05 | 0.2952 |
| BZLF1 - EBNA1 | 1.222 | 0.212 | 162 | 5.776 | <.0001 |
| BZLF1 - LMP2 | 3.056 | 0.212 | 162 | 14.441 | <.0001 |
| BZLF1 - X26E | 0.389 | 0.212 | 162 | 1.838 | 0.0755 |
| EBNA1 - LMP2 | 1.833 | 0.212 | 162 | 8.664 | <.0001 |
| EBNA1 - X26E | -0.833 | 0.212 | 162 | -3.938 | 0.0002 |
| LMP2 - X26E | -2.667 | 0.212 | 162 | -12.603 | <.0001 |
| **Chad latent (R)+Chad lytic (L)** | | | | | |
| BARF1 - BZLF1 | -2.417 | 0.366 | 162 | -6.594 | <.0001 |
| BARF1 - EBNA1 | -2.75 | 0.366 | 162 | -7.504 | <.0001 |
| BARF1 - LMP2 | -1.333 | 0.366 | 162 | -3.638 | 0.0006 |
| BARF1 - X26E | -3.5 | 0.366 | 162 | -9.55 | <.0001 |
| BZLF1 - EBNA1 | -0.333 | 0.366 | 162 | -0.91 | 0.3644 |
| BZLF1 - LMP2 | 1.083 | 0.366 | 162 | 2.956 | 0.0045 |
| BZLF1 - X26E | -1.083 | 0.366 | 162 | -2.956 | 0.0045 |
| EBNA1 - LMP2 | 1.417 | 0.366 | 162 | 3.865 | 0.0003 |
| EBNA1 - X26E | -0.75 | 0.366 | 162 | -2.046 | 0.047 |
| LMP2 - X26E | -2.167 | 0.366 | 162 | -5.912 | <.0001 |
| **Chad lytic-only** | | | | | |
| BARF1 - BZLF1 | -1.25 | 0.366 | 162 | -3.411 | 0.0012 |
| BARF1 - EBNA1 | -2.833 | 0.366 | 162 | -7.731 | <.0001 |
| BARF1 - LMP2 | -3.25 | 0.366 | 162 | -8.868 | <.0001 |
| BARF1 - X26E | -2.667 | 0.366 | 162 | -7.276 | <.0001 |
| BZLF1 - EBNA1 | -1.583 | 0.366 | 162 | -4.32 | 0.0001 |
| BZLF1 - LMP2 | -2 | 0.366 | 162 | -5.457 | <.0001 |
| BZLF1 - X26E | -1.417 | 0.366 | 162 | -3.865 | 0.0003 |
| EBNA1 - LMP2 | -0.417 | 0.366 | 162 | -1.137 | 0.2858 |
| EBNA1 - X26E | 0.167 | 0.366 | 162 | 0.455 | 0.6499 |
| LMP2 - X26E | 0.583 | 0.366 | 162 | 1.592 | 0.1418 |

Degrees-of-freedom method: Kenward-roger . P value adjustment: BH method for 10 tests

**Supplementary Table 6. Haematopathology assessment and cytogenetic results of the BL tumours**

The table shows the haematopathology report and FISH analysis of fifteen BL tumour sections selected for RNA sequencing and EBV protein expression studies.

Lymphoma diagnoses were confirmed through histopathological analysis of biopsy specimens collected at AI-REAL recruitment centres in East Africa. Tissue samples were fixed in 10% neutral buffered formalin and embedded in paraffin. All samples were stained with Haematoxylin and Eosin (H&E), and lymphoma diagnosis was complemented by immunohistochemistry. The diagnostic workflow used a three-phase scoring system by Naresh et al. (2011). The first phase involved typical BL morphology and immunomarkers BCL2, CD10, and CD20. Inconclusive cases underwent a second phase with additional markers (CD38, CD44, Ki-67), while unresolved cases proceeded to a third phase involving FISH analysis for MYC-Ig translocation. Interphase FISH was performed on FFPE tissue sections using the ZytoLight SPEC MYC Dual Colour Break Apart Probe, following the manufacturer's protocol. This dual-colour probe detects MYC rearrangements on chromosome 8q24.21, regardless of fusion partner.

| Sample | H&E review | Tumour content | Cytogenetic report |
|---|---|---|---|
| BL1 | Small to medium monomorphic lymphoid cells with starry sky appearance, immunostains CD10+, BCL2 -ve (Naresh diagnostic algorithm score 6) | 50% | MYC rearrangement in 25 out of 40 cells examined. Of these 25 abnormal cells, gain of the derivative MYC signal was seen in 17 cells. |
| BL2 | Small to medium lymphoid cell proliferation with condensed chromatin, starry sky appearance. IHC : Ki67 greater than 95%, CD10+, BCL2-ve (diagnostic algorithm study score 6). | 80-90% | MYC rearrangement in 13 out of 20 cells examined |
| BL3 | Small to medium sized cells, having a basophilic cytoplasm and foci of tingible body macrophages which confer a starry sky appearance to the tumour. Naresh morphology score 2. | 50-90% | MYC rearrangement in 28 out of 40 cells examined |
| BL4 | Starry sky tumour composed of primitive lymphoid cells | 90% | MYC rearrangement in 34 out of 40 cells examined |
| BL5 | Small to medium sized cells, having a basophilic cytoplasm and foci of tingible body macrophages which confer a starry sky appearance to the tumour. Naresh morphology score 2 | 80-90% | MYC rearrangement in 31 out of 40 cells examined |
| BL6 | Small to medium sized cells, starry sky appearance. The tumour cells have basophilic cytoplasm | 80-90% | MYC rearrangement in 31 out of 40 cells examined |
| BL7 | Round to ovoid cells with rather scanty cytoplasm and indistinct nucleoli | 40-90% | MYC rearrangement in 30 out of 40 cells examined |
| BL8 | Medium sized monotonous cells with round nuclei and finely clumped chromatin. There are numerous mitotic bodies and apoptotic bodies. | 80-90% | MYC rearrangement in 28 out of 40 cells examined |
| BL9 | Atypical lymphoid cells of small to medium sized cells, starry sky appearance. The tumour cells have basophilic cytoplasm | 70-90% | MYC rearrangement in 32 out of 40 cells examined |
| BL10 | Diffuse infiltrate of round cells with rather scanty cytoplasm with foci of necrosis | 70-90% | MYC rearrangement in all 20 cells examined |
| BL11 | kidney replaced by a tumour composed of round to ovoid cells with rather scanty cytoplasm and indistinct nucleoli | 60-70% | MYC rearrangement in 34 out of 40 cells examined |
| BL12 | Starry sky tumour composed of round cells with scanty cytoplasm | >90% | MYC rearrangement in 33 out of 40 cells examined |
| BL13 | Tumour composed of round to ovoid cells with rather scanty cytoplasm and indistinct nucleoli | 60-70% | MYC rearrangement in 34 out of 40 cells examined |
| BL14 | A starry sky tumour composed of primitive lymphoid cells | 80-90% | MYC rearrangement in 36 out of 40 cells examined |
| BL15 | A starry sky tumour composed of primitive lymphoid cells | >90% | MYC rearrangement in 31 out of 40 cells examined |

**Supplementary Table 7. Novel EBV genomes isolated**

The table presents sample metadata for 64 newly sequenced EBV genomes from children with and without Burkitt lymphoma in Uganda and Kenya. EBV whole-genome sequencing was conducted using a custom capture panel developed at the Oxford Molecular Diagnostics Laboratory under the AI-REAL project. Samples included cfDNA samples, free of genomic DNA contamination (collected in PAXgene DNA tubes) and extracted at our collaborating molecular diagnostic laboratories in Uganda and Tanzania. The raw paired-end Illumina FASTQ files for the 64 genomes are published under the BioProject ID PRJNA1333409. Meanwhile, assembled genomes are published in GenBank. HHV4 = Human gamma Herpesvirus 4.

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Host Age | Host Sex | Region | Country | Accession | BioProject |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BMC035 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 5 | Male | Africa | Tanzania | SRR35585212 | PRJNA1333409 |
| BMC056 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 3 | Male | Africa | Tanzania | SRR35585211 | PRJNA1333409 |
| BMC057 | Paired_FastQ | Illumina | Plasma cfDNA | Lymphoma NOS | HHV4 | Type1 | 11 | Male | Africa | Tanzania | SRR35585200 | PRJNA1333409 |
| BMC064 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 9 | Male | Africa | Tanzania | SRR35585189 | PRJNA1333409 |
| EBV001 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type1 | 20 | Male | Africa | Uganda | SRR35585178 | PRJNA1333409 |
| EBV003 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type2 | 7 | Male | Africa | Uganda | SRR35585167 | PRJNA1333409 |
| EBV012 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type1 | 9 | Male | Africa | Uganda | SRR35585156 | PRJNA1333409 |
| EBV013 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type2 | 9 | Female | Africa | Uganda | SRR35585151 | PRJNA1333409 |
| EBV019 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type1 | 8 | Female | Africa | Uganda | SRR35585150 | PRJNA1333409 |
| EBV021 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type2 | 7 | Female | Africa | Uganda | SRR35585149 | PRJNA1333409 |
| EBV030 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type1 | 7 | Female | Africa | Uganda | SRR35585210 | PRJNA1333409 |
| EBV036 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type1 | 0 | Female | Africa | Uganda | SRR35585209 | PRJNA1333409 |
| EBV038 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type2 | 7 | Male | Africa | Uganda | SRR35585208 | PRJNA1333409 |
| EBV040 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type1 | 11 | Female | Africa | Uganda | SRR35585207 | PRJNA1333409 |
| EBV049 | Paired_FastQ | Illumina | Plasma cfDNA | Healthy | HHV4 | Type1 | 10 | Male | Africa | Uganda | SRR35585206 | PRJNA1333409 |
| EBVLLS002 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 10 | Male | Africa | Uganda | SRR35585205 | PRJNA1333409 |
| EBVLLS003 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 13 | Male | Africa | Uganda | SRR35585204 | PRJNA1333409 |
| KCMC061 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 7 | Male | Africa | Tanzania | SRR35585203 | PRJNA1333409 |
| KCMC47 | Paired_FastQ | Illumina | Plasma cfDNA | Lymphadenitis | HHV4 | Type1 | 2 | Female | Africa | Tanzania | SRR35585202 | PRJNA1333409 |
| LACOR003 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 8 | Male | Africa | Uganda | SRR35585201 | PRJNA1333409 |
| LACOR008 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 24 | Female | Africa | Uganda | SRR35585199 | PRJNA1333409 |
| LACOR009 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 13 | Male | Africa | Uganda | SRR35585198 | PRJNA1333409 |
| LACOR010 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 12 | Female | Africa | Uganda | SRR35585197 | PRJNA1333409 |
| LACOR018 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 9 | Female | Africa | Uganda | SRR35585196 | PRJNA1333409 |

Novel EBV genomes isolated cont…

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Host Age | Host Sex | Region | Country | Accession | BioProject |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LACOR025 | Paired_FastQ | Illumina | Plasma cfDNA | DLBCL | HHV4 | Type1 | 11 | Female | Africa | Uganda | SRR35585195 | PRJNA1333409 |
| LACOR028 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 4 | Male | Africa | Uganda | SRR35585194 | PRJNA1333409 |
| LACOR030 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 11 | Male | Africa | Uganda | SRR35585193 | PRJNA1333409 |
| LACOR031 | Paired_FastQ | Illumina | Plasma cfDNA | DLBCL | HHV4 | Type1 | 4 | Male | Africa | Uganda | SRR35585192 | PRJNA1333409 |
| LACOR034 | Paired_FastQ | Illumina | Plasma cfDNA | Lymphadenitis | HHV4 | Type2 | 12 | Male | Africa | Uganda | SRR35585191 | PRJNA1333409 |
| LACOR037 | Paired_FastQ | Illumina | Plasma cfDNA | Lymphoblastic lymphoma | HHV4 | Type2 | 13 | Female | Africa | Uganda | SRR35585190 | PRJNA1333409 |
| LACOR040 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 13 | Female | Africa | Uganda | SRR35585188 | PRJNA1333409 |
| LACOR042 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 9 | Male | Africa | Uganda | SRR35585187 | PRJNA1333409 |
| LACOR044 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 4 | Female | Africa | Uganda | SRR35585186 | PRJNA1333409 |
| LACOR050 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 9 | Male | Africa | Uganda | SRR35585185 | PRJNA1333409 |
| LACOR052 | Paired_FastQ | Illumina | Plasma cfDNA | Kidney neoplasm | HHV4 | Type1 | 9 | Male | Africa | Uganda | SRR35585184 | PRJNA1333409 |
| LACOR053 | Paired_FastQ | Illumina | Plasma cfDNA | Lymphoblastic lymphoma | HHV4 | Type1 | 9 | Male | Africa | Uganda | SRR35585183 | PRJNA1333409 |
| LACOR055 | Paired_FastQ | Illumina | Plasma cfDNA | Rhabdomyosarcoma | HHV4 | Type2 | 10 | Male | Africa | Uganda | SRR35585182 | PRJNA1333409 |
| LACOR056 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 12 | Male | Africa | Uganda | SRR35585181 | PRJNA1333409 |
| LACOR059 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 11 | Female | Africa | Uganda | SRR35585180 | PRJNA1333409 |
| LACOR061 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 3 | Male | Africa | Uganda | SRR35585179 | PRJNA1333409 |
| LACOR062 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 7 | Male | Africa | Uganda | SRR35585177 | PRJNA1333409 |
| LACOR064 | Paired_FastQ | Illumina | Plasma cfDNA | Lymphoma NOS | HHV4 | Type1 | 3 | Male | Africa | Uganda | SRR35585176 | PRJNA1333409 |
| LACOR065 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 8 | Male | Africa | Uganda | SRR35585175 | PRJNA1333409 |
| LACOR082 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 4 | Male | Africa | Uganda | SRR35585174 | PRJNA1333409 |
| LACOR086 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 5 | Male | Africa | Uganda | SRR35585173 | PRJNA1333409 |
| LACOR090 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 6 | Female | Africa | Uganda | SRR35585172 | PRJNA1333409 |
| LACOR096 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 9 | Female | Africa | Uganda | SRR35585171 | PRJNA1333409 |
| LACOR097 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 12 | Female | Africa | Uganda | SRR35585170 | PRJNA1333409 |

Novel EBV genomes isolated cont…

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Host Age | Host Sex | Region | Country | Accession | BioProject |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LACOR098 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 12 | Male | Africa | Uganda | SRR35585169 | PRJNA1333409 |
| LACOR099 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 13 | Male | Africa | Uganda | SRR35585168 | PRJNA1333409 |
| LACOR101 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 6 | Male | Africa | Uganda | SRR35585166 | PRJNA1333409 |
| LLP002 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 8 | Female | Africa | Uganda | SRR35585165 | PRJNA1333409 |
| LLP003 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 10 | Female | Africa | Uganda | SRR35585164 | PRJNA1333409 |
| LLP007 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 4 | Male | Africa | Uganda | SRR35585163 | PRJNA1333409 |
| LLP008 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 4 | Male | Africa | Uganda | SRR35585162 | PRJNA1333409 |
| MNH011 | Paired_FastQ | Illumina | Plasma cfDNA | Hodgkin lymphoma | HHV4 | Type1 | 13 | Male | Africa | Tanzania | SRR35585161 | PRJNA1333409 |
| MNH018 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 6 | Male | Africa | Tanzania | SRR35585160 | PRJNA1333409 |
| MNH026 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 7 | Female | Africa | Tanzania | SRR35585159 | PRJNA1333409 |
| MNH052 | Paired_FastQ | Illumina | Plasma cfDNA | Hodgkin lymphoma | HHV4 | Type1 | 10 | Male | Africa | Tanzania | SRR35585158 | PRJNA1333409 |
| MNH054 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 13 | Male | Africa | Tanzania | SRR35585157 | PRJNA1333409 |
| MNH079 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 8 | Female | Africa | Tanzania | SRR35585155 | PRJNA1333409 |
| MNH089 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 10 | Female | Africa | Tanzania | SRR35585154 | PRJNA1333409 |
| MNH095 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type1 | 4 | Male | Africa | Tanzania | SRR35585153 | PRJNA1333409 |
| MNH099 | Paired_FastQ | Illumina | Plasma cfDNA | Burkitt lymphoma | HHV4 | Type2 | 4 | Male | Africa | Tanzania | SRR35585152 | PRJNA1333409 |

**Supplementary Table 8. Public EBV genomes analysed**

The table presents metadata for the 346 primary -isolate EBV genomes downloaded from the NCBI Sequence Read Archive (SRA) under BioProject Ids PRJNA552587, PRJNA522388, and PRJNA1063319. The Illumina-paired-end raw FASTQ files were downloaded using FASTQ dump and processed using our custom (OMDC) bioinformatics pipeline together with the newly sequenced isolates. The tissues phenotypes included Burkitt lymphoma 9BL) , Nasopharyngeal carcinoma (NPC), Gastric Carcinoma (GC), Natural Killer T-cell lymphoma (NKTCL), Hodgkin lymphoma (HL), Diffuse large B cell lymphoma (DLBCL), other EBV-associated malignancies and plasma or saliva from Healthy controls. Two BL cell lines (Raji and Daudi) were processed alongside the primary isolates for comparison. LCL = lymphoblastoid cell line , HHV4= Human gamma herpesvirus 4. na = not applicable.

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLT001 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Asia | China | 36 | Male | SRR8574951 | SRP185895 |
| BLT002 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Asia | China | 34 | Male | SRR8574952 | SRP185895 |
| CP111 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562496 | SRP483990 |
| CP115 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562495 | SRP483990 |
| CP123 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562468 | SRP483990 |
| CP161 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562464 | SRP483990 |
| CP163 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562463 | SRP483990 |
| CP164 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562461 | SRP483990 |
| CP174 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562459 | SRP483990 |
| CP179 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562458 | SRP483990 |
| CP180 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562457 | SRP483990 |
| CP186 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562455 | SRP483990 |
| CP189 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562454 | SRP483990 |
| CP225 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562433 | SRP483990 |
| CP236 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562453 | SRP483990 |
| CP238 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562452 | SRP483990 |
| CP240 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562450 | SRP483990 |
| CP241 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562449 | SRP483990 |
| CP255 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562432 | SRP483990 |
| CP256 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562448 | SRP483990 |
| CP257 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562431 | SRP483990 |
| CP270 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562446 | SRP483990 |
| CP276 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562445 | SRP483990 |
| CP282 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562444 | SRP483990 |
| CP289 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562443 | SRP483990 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CP295 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562442 | SRP483990 |
| CP301 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562441 | SRP483990 |
| CP315 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562439 | SRP483990 |
| CP317 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562438 | SRP483990 |
| CP325 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562401 | SRP483990 |
| CP34 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562398 | SRP483990 |
| CP343 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562396 | SRP483990 |
| CP345 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562395 | SRP483990 |
| CP382 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562393 | SRP483990 |
| CP391 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562392 | SRP483990 |
| CP399 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562391 | SRP483990 |
| CP406 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562390 | SRP483990 |
| CP419 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562389 | SRP483990 |
| CP42 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562430 | SRP483990 |
| CP429 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562388 | SRP483990 |
| CP450 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562386 | SRP483990 |
| CP46 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562384 | SRP483990 |
| CP466 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562382 | SRP483990 |
| CP473 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562381 | SRP483990 |
| CP478 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562380 | SRP483990 |
| CP48 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562379 | SRP483990 |
| CP539 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562375 | SRP483990 |
| CP655 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562371 | SRP483990 |
| CP674 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562370 | SRP483990 |
| CP688 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562369 | SRP483990 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CP694 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562582 | SRP483990 |
| CP699 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562580 | SRP483990 |
| CP71 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562579 | SRP483990 |
| CP738 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562578 | SRP483990 |
| CP75 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562577 | SRP483990 |
| CP765 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562575 | SRP483990 |
| CP77 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562573 | SRP483990 |
| CP775 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562572 | SRP483990 |
| CP778 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562571 | SRP483990 |
| CP782 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562570 | SRP483990 |
| CP784 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562569 | SRP483990 |
| CP79 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562568 | SRP483990 |
| CP808 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562567 | SRP483990 |
| CP812 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562566 | SRP483990 |
| CP829 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562565 | SRP483990 |
| CP834 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562564 | SRP483990 |
| CP835 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562561 | SRP483990 |
| CP836 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562560 | SRP483990 |
| CP839 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562559 | SRP483990 |
| CP842 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562558 | SRP483990 |
| CP86 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Africa | Malawi | na | na | SRR27562557 | SRP483990 |
| CP87 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562556 | SRP483990 |
| CP874 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562555 | SRP483990 |
| CP875 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562554 | SRP483990 |
| CP90 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562553 | SRP483990 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CP94 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562552 | SRP483990 |
| CP96 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Malawi | na | na | SRR27562435 | SRP483990 |
| E0096 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562551 | SRP483990 |
| E0235 | Paired_FastQ | Illumina | tissue biopsy | DLBCL | HHV4 | Type2 | public | Central America | Guatemala | na | na | SRR27562537 | SRP483990 |
| E0720 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562532 | SRP483990 |
| eBL-Plasma-0035 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Kenya | 8 | Male | SRR9641908 | SRP212943 |
| eBL-Plasma-0036 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Kenya | 10 | Male | SRR9641907 | SRP212943 |
| eBL-Plasma-0037 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 11 | Male | SRR9641906 | SRP212943 |
| eBL-Plasma-0038 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 3 | Female | SRR9641905 | SRP212943 |
| eBL-Plasma-0039 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 12 | Male | SRR9641904 | SRP212943 |
| eBL-Plasma-0040 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Kenya | 6 | Male | SRR9641903 | SRP212943 |
| eBL-Plasma-0042 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 6 | Male | SRR9641856 | SRP212943 |
| eBL-Plasma-0043 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Kenya | 7 | Male | SRR9641855 | SRP212943 |
| eBL-Plasma-0044 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 4 | Male | SRR9641854 | SRP212943 |
| eBL-Plasma-0045 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 5 | Male | SRR9641853 | SRP212943 |
| eBL-Plasma-0047 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 9 | Male | SRR9641861 | SRP212943 |
| eBL-Plasma-0048 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Kenya | 5 | Male | SRR9641863 | SRP212943 |
| eBL-Plasma-0049 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 10 | Male | SRR9641864 | SRP212943 |
| eBL-Plasma-0050 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 10 | Female | SRR9641865 | SRP212943 |
| eBL-Plasma-0051 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 10 | Male | SRR9641866 | SRP212943 |
| eBL-Plasma-0052 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 8 | Female | SRR9641867 | SRP212943 |
| eBL-Plasma-0053 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 6 | Female | SRR9641868 | SRP212943 |
| eBL-Plasma-0054 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Kenya | 5 | Male | SRR9641869 | SRP212943 |
| eBL-Plasma-0055 | Paired_FastQ | Illumina | Plasma | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 4 | Male | SRR9641870 | SRP212943 |
| eBL-Tumor-0001 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 7 | Male | SRR9641813 | SRP212943 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| eBL-Tumor-0003 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type2 | public | Africa | Kenya | 6 | Male | SRR9641819 | SRP212943 |
| eBL-Tumor-0006 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 5 | Male | SRR9641822 | SRP212943 |
| eBL-Tumor-0010 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 0.8 | Male | SRR9641831 | SRP212943 |
| eBL-Tumor-0011 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 8 | Female | SRR9641832 | SRP212943 |
| eBL-Tumor-0014 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 11 | Male | SRR9641827 | SRP212943 |
| eBL-Tumor-0016 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 8 | Male | SRR9641823 | SRP212943 |
| eBL-Tumor-0018 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 6 | Male | SRR9641836 | SRP212943 |
| eBL-Tumor-0030 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 10 | Male | SRR9641845 | SRP212943 |
| eBL-Tumor-0031 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 8 | Female | SRR9641846 | SRP212943 |
| eBL-Tumor-0032 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 8 | Female | SRR9641851 | SRP212943 |
| eBL-Tumor-0033 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 11 | Male | SRR9641852 | SRP212943 |
| eBL-Tumor-0034 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 6 | Male | SRR9641849 | SRP212943 |
| eBL-Tumor-0041 | Paired_FastQ | Illumina | tissue biopsy | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 6 | Male | SRR9641857 | SRP212943 |
| EBV00079 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562531 | SRP483990 |
| EBV00081 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562529 | SRP483990 |
| EBV00083 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562527 | SRP483990 |
| EBV00084 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type2 | public | Central America | Guatemala | na | na | SRR27562526 | SRP483990 |
| EBV00085 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562525 | SRP483990 |
| EBV00086 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562524 | SRP483990 |
| EBV00087 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562522 | SRP483990 |
| EBV00088 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562521 | SRP483990 |
| EBV00089 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562520 | SRP483990 |
| EBV00094 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562491 | SRP483990 |
| EBV00097 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562489 | SRP483990 |
| EBV00100 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | North America | USA | na | na | SRR27562487 | SRP483990 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EBV00102 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562484 | SRP483990 |
| EBV00104 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type2 | public | Central America | Guatemala | na | na | SRR27562482 | SRP483990 |
| EBV00105 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562481 | SRP483990 |
| EBV00107 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type2 | public | Central America | Guatemala | na | na | SRR27562479 | SRP483990 |
| EBV00110 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562476 | SRP483990 |
| EBV00112 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562473 | SRP483990 |
| EBV00113 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562472 | SRP483990 |
| EBV00119 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | South America | Peru | na | na | SRR27562467 | SRP483990 |
| EBV00121 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | South America | Peru | na | na | SRR27562402 | SRP483990 |
| EBV00122 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | South America | Peru | na | na | SRR27562400 | SRP483990 |
| EBV00123 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | South America | Peru | na | na | SRR27562581 | SRP483990 |
| EBV00126 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | USA | na | na | SRR27562518 | SRP483990 |
| EBV00127 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | USA | na | na | SRR27562517 | SRP483990 |
| EBV00135 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | USA | na | na | SRR27562511 | SRP483990 |
| EBV00140 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Central America | Guatemala | na | na | SRR27562505 | SRP483990 |
| EBV00144 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Asia | Taiwan | na | na | SRR27562502 | SRP483990 |
| EBV00145 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Asia | Taiwan | na | na | SRR27562501 | SRP483990 |
| GCT001 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 43 | Male | SRR8574954 | SRP483990 |
| GCT002 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 51 | Female | SRR8574947 | SRP483990 |
| GCT003 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 26 | Male | SRR8574948 | SRP483990 |
| GCT004 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 44 | Male | SRR8574949 | SRP483990 |
| GCT005 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 63 | Male | SRR8574950 | SRP483990 |
| GCT006 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574955 | SRP483990 |
| GCT009 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574699 | SRP483990 |
| GCT010 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 42 | Female | SRR8574700 | SRP483990 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GCT011 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 60 | Male | SRR8574701 | SRP483990 |
| GCT012 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 55 | Male | SRR8574702 | SRP483990 |
| GCT013 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 67 | Male | SRR8574703 | SRP483990 |
| GCT014 | Paired_FastQ | Illumina | tissue biopsy | Other Cancer | HHV4 | Type1 | public | Asia | China | 69 | Male | SRR8574704 | SRP483990 |
| HC-0001 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 2.6 | Female | SRR9641902 | SRP212943 |
| HC-0002 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 4 | Female | SRR9641895 | SRP212943 |
| HC-0003 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 3 | Male | SRR9641896 | SRP212943 |
| HC-0004 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 3 | Female | SRR9641897 | SRP212943 |
| HC-0005 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 2 | Female | SRR9641898 | SRP212943 |
| HC-0006 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 2 | Female | SRR9641893 | SRP212943 |
| HC-0008 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 4 | Female | SRR9641875 | SRP212943 |
| HC-0009 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 5 | Male | SRR9641876 | SRP212943 |
| HC-0010 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 3 | Male | SRR9641877 | SRP212943 |
| HC-0013 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 2.4 | Female | SRR9641880 | SRP212943 |
| HC-0015 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 4 | Male | SRR9641882 | SRP212943 |
| HC-0016 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 4 | Female | SRR9641873 | SRP212943 |
| HC-0017 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 4 | Male | SRR9641874 | SRP212943 |
| HC-0018 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 5 | Male | SRR9641888 | SRP212943 |
| HC-0019 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 3 | Male | SRR9641887 | SRP212943 |
| HC-0020 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 3.6 | Male | SRR9641890 | SRP212943 |
| HC-0021 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 4 | Female | SRR9641889 | SRP212943 |
| HC-0022 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 5 | Male | SRR9641884 | SRP212943 |
| HC-0023 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 3 | Female | SRR9641883 | SRP212943 |
| HC-0024 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 3 | Male | SRR9641886 | SRP212943 |
| HC-0025 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 4 | Male | SRR9641885 | SRP212943 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HC-0026 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 4 | Male | SRR9641892 | SRP212943 |
| HC-0027 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 3 | Female | SRR9641891 | SRP212943 |
| HC-0030 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type1 | public | Africa | Kenya | 3 | Male | SRR9641915 | SRP212943 |
| HC-0031 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 5 | Female | SRR9641916 | SRP212943 |
| HC-0034 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 2 | Male | SRR9641911 | SRP212943 |
| HC-0039 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 5 | Male | SRR9641815 | SRP212943 |
| HC-0040 | Paired_FastQ | Illumina | Plasma | Healthy control | HHV4 | Type2 | public | Africa | Kenya | 4 | Female | SRR9641814 | SRP212943 |
| HLT001 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Asia | China | 35 | Male | SRR8574707 | SRP185895 |
| HLT002 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Asia | China | 69 | Male | SRR8574779 | SRP185895 |
| HLT005 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Asia | China | 54 | Male | SRR8574780 | SRP185895 |
| HLT006 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Asia | China | 52 | Male | SRR8574775 | SRP185895 |
| HLT007 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Asia | China | 55 | Male | SRR8574774 | SRP185895 |
| HLT010 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Asia | China | 59 | Male | SRR8574776 | SRP185895 |
| HLT011 | Paired_FastQ | Illumina | tissue biopsy | Hodgkin lymphoma | HHV4 | Type1 | public | Asia | China | 60 | Male | SRR8574772 | SRP185895 |
| HS001 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 29 | Female | SRR8574816 | SRP185895 |
| HS003 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 60 | Female | SRR8574967 | SRP185895 |
| HS007 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 56 | Male | SRR8574812 | SRP185895 |
| HS008 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 40 | Male | SRR8574813 | SRP185895 |
| HS009 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 46 | Male | SRR8574810 | SRP185895 |
| HS011 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 36 | Male | SRR8574929 | SRP185895 |
| HS012 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 56 | Male | SRR8574809 | SRP185895 |
| HS013 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 37 | Male | SRR8574849 | SRP185895 |
| HS014 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 72 | Male | SRR8574848 | SRP185895 |
| HS015 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 58 | Male | SRR8574847 | SRP185895 |
| HS016 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 70 | Male | SRR8574846 | SRP185895 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HS018 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 43 | Male | SRR8574852 | SRP185895 |
| HS019 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 35 | Male | SRR8574851 | SRP185895 |
| HS020 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574850 | SRP185895 |
| HS021 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 56 | Male | SRR8574844 | SRP185895 |
| HS023 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 47 | Male | SRR8574727 | SRP185895 |
| HS024 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 56 | Male | SRR8574782 | SRP185895 |
| HS025 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 60 | Male | SRR8574783 | SRP185895 |
| HS027 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 45 | Male | SRR8574750 | SRP185895 |
| HS029 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 46 | Male | SRR8574769 | SRP185895 |
| HS032 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 40 | Male | SRR8574887 | SRP185895 |
| HS033 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574724 | SRP185895 |
| HS034 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 40 | Male | SRR8574918 | SRP185895 |
| HS035 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574921 | SRP185895 |
| HS036 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 73 | Male | SRR8574920 | SRP185895 |
| HS037 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574923 | SRP185895 |
| HS038 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 40 | Male | SRR8574726 | SRP185895 |
| HS039 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574925 | SRP185895 |
| HS041 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 43 | Male | SRR8574927 | SRP185895 |
| HS045 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574728 | SRP185895 |
| HS048 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 50 | Male | SRR8574960 | SRP185895 |
| HS050 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 47 | Female | SRR8574907 | SRP185895 |
| HS051 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 49 | Female | SRR8574958 | SRP185895 |
| HS052 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 67 | Female | SRR8574963 | SRP185895 |
| HS053 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 63 | Female | SRR8574930 | SRP185895 |
| HS054 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 47 | Female | SRR8574961 | SRP185895 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HS057 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 58 | Female | SRR8574966 | SRP185895 |
| NHS002 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 58 | Male | SRR8574840 | SRP185895 |
| NHS004 | Paired_FastQ | Illumina | Saliva | Healthy control | HHV4 | Type1 | public | Asia | China | 32 | Female | SRR8574729 | SRP185895 |
| NKLT002 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Asia | China | 48 | Male | SRR8574770 | SRP185895 |
| NKLT003-2 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Asia | China | 46 | Male | SRR8574725 | SRP185895 |
| NKLT004 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574747 | SRP185895 |
| NKLT006 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Asia | China | 24 | Male | SRR8574819 | SRP185895 |
| NKLT007 | Paired_FastQ | Illumina | tissue biopsy | NKTCL | HHV4 | Type1 | public | Asia | China | 48 | Male | SRR8574854 | SRP185895 |
| NNPCT001 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 46 | Male | SRR8574752 | SRP185895 |
| NNPCT002 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574718 | SRP185895 |
| NNPCT003 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 45 | Male | SRR8574719 | SRP185895 |
| NNPCT004 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 46 | Female | SRR8574720 | SRP185895 |
| NNPCT005 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 17 | Male | SRR8574721 | SRP185895 |
| NPCT001 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 41 | Male | SRR8574736 | SRP185895 |
| NPCT002 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 51 | Male | SRR8574737 | SRP185895 |
| NPCT003 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 44 | Male | SRR8574738 | SRP185895 |
| NPCT004 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 35 | Male | SRR8574731 | SRP185895 |
| NPCT005 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574732 | SRP185895 |
| NPCT006 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 54 | Male | SRR8574733 | SRP185895 |
| NPCT007 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574734 | SRP185895 |
| NPCT008 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 73 | Male | SRR8574743 | SRP185895 |
| NPCT009 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 58 | Male | SRR8574744 | SRP185895 |
| NPCT010 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 51 | Male | SRR8574829 | SRP185895 |
| NPCT011 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 71 | Male | SRR8574828 | SRP185895 |
| NPCT012 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 43 | Female | SRR8574831 | SRP185895 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NPCT013 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 34 | Male | SRR8574830 | SRP185895 |
| NPCT014 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 27 | Male | SRR8574825 | SRP185895 |
| NPCT015 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 69 | Male | SRR8574824 | SRP185895 |
| NPCT016 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type2 | public | Asia | China | 62 | Male | SRR8574827 | SRP185895 |
| NPCT017 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 43 | Male | SRR8574826 | SRP185895 |
| NPCT018 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574823 | SRP185895 |
| NPCT019 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 43 | Male | SRR8574822 | SRP185895 |
| NPCT020-2 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 69 | Male | SRR8574794 | SRP185895 |
| NPCT021 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type2 | public | Asia | China | 66 | Male | SRR8574795 | SRP185895 |
| NPCT022 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 72 | Male | SRR8574792 | SRP185895 |
| NPCT023 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 47 | Male | SRR8574793 | SRP185895 |
| NPCT024 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 50 | Male | SRR8574790 | SRP185895 |
| NPCT025 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 66 | Male | SRR8574791 | SRP185895 |
| NPCT026 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type2 | public | Asia | China | 31 | Male | SRR8574788 | SRP185895 |
| NPCT027 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 56 | Male | SRR8574789 | SRP185895 |
| NPCT028-2 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 53 | Male | SRR8574786 | SRP185895 |
| NPCT029 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 46 | Male | SRR8574787 | SRP185895 |
| NPCT031 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 47 | Female | SRR8574917 | SRP185895 |
| NPCT032 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 44 | Female | SRR8574916 | SRP185895 |
| NPCT033 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 51 | Male | SRR8574915 | SRP185895 |
| NPCT035 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574914 | SRP185895 |
| NPCT036 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 48 | Female | SRR8574913 | SRP185895 |
| NPCT037 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 58 | Male | SRR8574912 | SRP185895 |
| NPCT038 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574911 | SRP185895 |
| NPCT039 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 37 | Male | SRR8574910 | SRP185895 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NPCT040 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 28 | Male | SRR8574909 | SRP185895 |
| NPCT041 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574908 | SRP185895 |
| NPCT042 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 65 | Male | SRR8574876 | SRP185895 |
| NPCT043 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 42 | Male | SRR8574877 | SRP185895 |
| NPCT045 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 38 | Female | SRR8574878 | SRP185895 |
| NPCT046 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 52 | Female | SRR8574879 | SRP185895 |
| NPCT047 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 62 | Female | SRR8574880 | SRP185895 |
| NPCT048 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 39 | Female | SRR8574881 | SRP185895 |
| NPCT049 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type2 | public | Asia | China | 57 | Male | SRR8574882 | SRP185895 |
| NPCT050 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 38 | Male | SRR8574883 | SRP185895 |
| NPCT052 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 36 | Male | SRR8574885 | SRP185895 |
| NPCT053 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 63 | Male | SRR8574748 | SRP185895 |
| NPCT054 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 57 | Male | SRR8574749 | SRP185895 |
| NPCT055 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 34 | Female | SRR8574746 | SRP185895 |
| NPCT056 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 31 | Male | SRR8574740 | SRP185895 |
| NPCT057 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 60 | Male | SRR8574742 | SRP185895 |
| NPCT058 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 44 | Male | SRR8574919 | SRP185895 |
| NPCT059 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574940 | SRP185895 |
| NPCT060 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 61 | Male | SRR8574762 | SRP185895 |
| NPCT061 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 28 | Male | SRR8574761 | SRP185895 |
| NPCT062 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 35 | Male | SRR8574768 | SRP185895 |
| NPCT063 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 47 | Female | SRR8574767 | SRP185895 |
| NPCT064 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 31 | Male | SRR8574766 | SRP185895 |
| NPCT065 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 44 | Male | SRR8574765 | SRP185895 |
| NPCT066 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 38 | Male | SRR8574818 | SRP185895 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NPCT067 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Female | SRR8574804 | SRP185895 |
| NPCT068 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 42 | Male | SRR8574802 | SRP185895 |
| NPCT069 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 46 | Male | SRR8574803 | SRP185895 |
| NPCT070 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 31 | Male | SRR8574800 | SRP185895 |
| NPCT071 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 60 | Female | SRR8574801 | SRP185895 |
| NPCT072 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 51 | Male | SRR8574798 | SRP185895 |
| NPCT073 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 44 | Female | SRR8574799 | SRP185895 |
| NPCT074 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 66 | Female | SRR8574796 | SRP185895 |
| NPCT075 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 32 | Male | SRR8574805 | SRP185895 |
| NPCT076 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 43 | Male | SRR8574806 | SRP185895 |
| NPCT077 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 47 | Male | SRR8574837 | SRP185895 |
| NPCT078 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 39 | Male | SRR8574836 | SRP185895 |
| NPCT080 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 29 | Male | SRR8574838 | SRP185895 |
| NPCT081 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 45 | Female | SRR8574833 | SRP185895 |
| NPCT082 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 34 | Male | SRR8574832 | SRP185895 |
| NPCT083 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 42 | Male | SRR8574835 | SRP185895 |
| NPCT084 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 45 | Male | SRR8574834 | SRP185895 |
| NPCT085 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 59 | Female | SRR8574842 | SRP185895 |
| NPCT086 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Female | SRR8574841 | SRP185895 |
| NPCT087 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 63 | Male | SRR8574866 | SRP185895 |
| NPCT088 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 51 | Male | SRR8574867 | SRP185895 |
| NPCT089 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 25 | Female | SRR8574868 | SRP185895 |
| NPCT090 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 34 | Male | SRR8574869 | SRP185895 |
| NPCT091 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 25 | Female | SRR8574870 | SRP185895 |
| NPCT092 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 30 | Male | SRR8574871 | SRP185895 |

Public EBV genomes analysed cont….

| Sample | Data TYPE | Platform | Sample TYPE | Phenotype | Isolate | EBV Type | Source | Region | Country | Host Age | Host Sex | Accession | SRA Study |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NPCT094 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 71 | Male | SRR8574873 | SRP185895 |
| NPCT096 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Male | SRR8574875 | SRP185895 |
| NPCT098 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 61 | Female | SRR8574926 | SRP185895 |
| NPCT099 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 41 | Male | SRR8574904 | SRP185895 |
| NPCT100 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 34 | Female | SRR8574903 | SRP185895 |
| NPCT101 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 39 | Male | SRR8574902 | SRP185895 |
| NPCT102 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 59 | Male | SRR8574901 | SRP185895 |
| NPCT103 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 66 | Female | SRR8574900 | SRP185895 |
| NPCT104 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 33 | Male | SRR8574899 | SRP185895 |
| NPCT105 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 56 | Male | SRR8574890 | SRP185895 |
| NPCT106 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 62 | Male | SRR8574889 | SRP185895 |
| NPCT107 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 49 | Female | SRR8574817 | SRP185895 |
| NPCT108 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 47 | Male | SRR8574957 | SRP185895 |
| NPCT109 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 43 | Female | SRR8574939 | SRP185895 |
| NPCT110 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 42 | Female | SRR8574922 | SRP185895 |
| NPCT111 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 48 | Male | SRR8574945 | SRP185895 |
| NPCT112 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 42 | Male | SRR8574946 | SRP185895 |
| NPCT113 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 29 | Male | SRR8574943 | SRP185895 |
| NPCT114 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 51 | Female | SRR8574944 | SRP185895 |
| NPCT115 | Paired_FastQ | Illumina | tissue biopsy | NPC | HHV4 | Type1 | public | Asia | China | 47 | Male | SRR8574807 | SRP185895 |
| Raji | Paired_FastQ | Illumina | LCL | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Nigeria | 11 | Male | SRR9641909 | SRP212943 |
| Daudi | Paired_FastQ | Illumina | LCL | Burkitt lymphoma | HHV4 | Type1 | public | Africa | Kenya | 16 | Male | SRR9641871 | SRP212943 |

# Chapter 8: Appendices

## 8.1 Period of the Research and Laboratories

The participants whose data were analysed in Chapters 3 and 4 were recruited in Uganda and Tanzania between 2020 and 2024 as part of the AI-REAL study, a collaboration involving four East African regional oncology centres: St Mary's Hospital Lacor, Gulu (Uganda); Muhimbili National Hospital (MNH), Dar es Salaam (Tanzania); Kilimanjaro Christian Medical Centre (KCMC), Tanzania; Bugando Medical Centre (BMC), Mwanza (Tanzania); and the University of Oxford. In addition to strengthening research capacity, the AI-REAL consortium supported the establishment of the first cancer genomics laboratories in Uganda and Tanzania. Whole-genome analysis of cfDNA samples was subsequently performed within these newly established laboratories.

During this period, I worked as a research assistant and benefited directly from the consortium's capacity-building programme. I coordinated participant recruitment for the AI-REAL study between 2020 and 2022, prior to commencing my doctoral studies in the UK. I later contributed to laboratory experiments at the Uganda site (Central Public Health Laboratories, Kampala), where I led the isolation and sequencing of 64 EBV genomes, predominantly from Ugandan samples.

The twenty-six tumour sections analysed in Chapter 4 were obtained in 2024 from St Mary's Hospital, Lacor, Gulu, Uganda. The tissues were imported into the UK under a Material Transfer Agreement (MTA Ref: R61914/CN023) and under HTA licence number 12217. Tissue processing and immunohistochemistry were performed at Oxford Translational Histopathology Laboratory under a framework contract, while FISH for MYC translocation was conducted at the Genetics Laboratory, Churchill Hospital. RNA library preparation and sequencing were outsourced to GENEWIZ UK Ltd (Azenta Life Sciences).

Additional whole-genome analysis using long-read Oxford Nanopore Technology was carried out in Professor Schuh's laboratory (OMDC). The vaccine preclinical work was undertaken in Dr Leung's laboratory at the Centre for Immuno-Oncology between 2024 and 2025.

## 8.2 Sources of funding

The AI-REAL consortium was funded by the UK National Institute for Health Research (NIHR) Global Health Programme (NIHR-RIGHT award 200133), using UK international development funding from the UK Government to support global health research.

This work was also supported by Commonwealth funding and a University of Oxford studentship covering three years of my doctoral studies.

I additionally received a £3,000 training grant from Cancer Research UK (CRUK) to undertake a bioinformatics course as part of the Oxford Biomedical Data Science Training Programme.

## 8.3 Abstracts and Scientific Conferences

- **Legason ID,** Burns A, Ridout K, Jennings D, Ogwang MD, Otim I, Nabalende H, Achola C, Ayitewala A, Sseruyange J, Chamba CC. Genomic landscape of Epstein-Barr virus in endemic Burkitt lymphoma. Blood. 2023 Nov 2;142:2985.

## 8.4 Publications/manuscripts from this thesis

- 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. Evidence of latent-lytic replication in EBV-positive Burkitt lymphoma from whole genome and transcriptome sequencing. (*submitted*)

# References

## References

Abate, F., Ambrosio, M. R., Mundo, L., Laginestra, M. A., Fuligni, F., Rossi, M., Zairis, S., Gazaneo, S., De Falco, G., Lazzi, S., Bellan, C., Rocca, B. J., Amato, T., Marasco, E., Etebari, M., Ogwang, M., Calbi, V., Ndede, I., Patel, K., . . . Rabadan, R. (2015). Distinct Viral and Mutational Spectrum of Endemic Burkitt Lymphoma. *PLOS Pathogens*, *11*(10), e1005158. https://doi.org/10.1371/journal.ppat.1005158

Agwati, E. O., Oduor, C. I., Ayieko, C., Ong'echa, J. M., Moormann, A. M., & Bailey, J. A. (2022). Profiling genome-wide recombination in Epstein Barr virus reveals type-specific patterns and associations with endemic-Burkitt lymphoma. *Virology Journal*, *19*(1), 208.

Aktipis, C., Boddy, A. M., Jansen, G., Hibner, U., Hochberg, M. E., Maley, C. C., & Wilkinson, G. S. (2015). Cancer across the tree of life: cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1673).

Alaggio, R., Amador, C., Anagnostopoulos, I., Attygalle, A. D., Araujo, I. B. d. O., Berti, E., Bhagat, G., Borges, A. M., Boyer, D., Calaminici, M., Chadburn, A., Chan, J. K. C., Cheuk, W., Chng, W.-J., Choi, J. K., Chuang, S.-S., Coupland, S. E., Czader, M., Dave, S. S., . . . Xiao, W. (2022). The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia*, *36*(7), 1720-1748. https://doi.org/10.1038/s41375-022-01620-2

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Alexandrov, L. B., . . .

Consortium, P. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94-101. https://doi.org/10.1038/s41586-020-1943-3

Alfieri, C., Tanner, J., Carpentier, L., Perpete, C., Savoie, A., Paradis, K., Delage, G., & Joncas, J. (1996). Epstein-Barr virus transmission from a blood donor to an organ transplant recipient with recovery of the same virus strain from the recipient's blood and oropharynx. *Blood*, *87*(2), 812-817.

Ali, S. M., Yao, M., Yao, J., Wang, J., Cheng, Y., Schrock, A. B., Chirn, G. W., Chen, H., Mu, S., & Gay, L. (2017). Comprehensive genomic profiling of different subtypes of nasopharyngeal carcinoma reveals similarities and differences to guide targeted therapy. *Cancer*, *123*(18), 3628-3637.

Ambinder, R. F. (2000). Gammaherpesviruses and "Hit-and-Run" oncogenesis. *The American journal of pathology*, *156*(1), 1-3.

Ariera, B., Guyah, B., Rahkola, J., Arao, I., Waomba, K., Koech, E., Samayoa-Reyes, G., Sabourin, K. R., Ogolla, S., & Rochford, R. (2025). Sustained activation induced cytidine deaminase (AID) expression in B cells following Plasmodium falciparum malaria infection in Kenyan children. *The Journal of Immunology*, *214*(5), 926-935.

Arruga, F., Vaisitti, T., & Deaglio, S. (2018). The NOTCH Pathway and Its Mutations in Mature B Cell Malignancies. *Front Oncol*, *8*, 550. https://doi.org/10.3389/fonc.2018.00550

Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. J., Hatfull, G., Hudson, G. S., Satchwell, S. C., Séguin, C., Tuffnell, P. S., & Barrell, B. G. (1984). DNA sequence and expression of the B95-8 Epstein—Barr virus genome. *Nature*, *310*(5974), 207-211. https://doi.org/10.1038/310207a0

Bahri, R., Boyer, F., Halabi, M. A., Chaunavel, A., Feuillard, J., Jaccard, A., & Ranger-Rogez, S. (2022). Epstein-Barr Virus (EBV) is mostly latent and clonal in angioimmunoblastic T cell lymphoma (AITL). *Cancers*, *14*(12), 2899.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., . . . Ding, L. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371-385.e318. https://doi.org/10.1016/j.cell.2018.02.060

Bajaj, B. G., Murakami, M., Cai, Q., Verma, S. C., Lan, K., & Robertson, E. S. (2008). Epstein-Barr virus nuclear antigen 3C interacts with and enhances the stability of the c-Myc oncoprotein. *J Virol*, *82*(8), 4082-4090. https://doi.org/10.1128/jvi.02500-07

Banchereau, J., & Steinman, R. M. (1998). Dendritic cells and the control of immunity. *Nature*, *392*(6673), 245-252.

Bangalore-Prakash, P., Stunz, L. L., Mambetsariev, N., Whillock, A. L., Hostager, B. S., & Bishop, G. A. (2017). The oncogenic membrane protein LMP1 sequesters TRAF3 in B-cell lymphoma cells to produce functional TRAF3 deficiency. *Blood Advances*, *1*(27), 2712-2723.

Bárcena, C., de Leval, L. (2023, May 28th, 2025). *CHL lymphocyte rich*. PathologyOutlines.com Retrieved October 4th, 2025. from https://www.pathologyoutlines.com/topic/lymphomanonblrhl.html

Basso, K., & Dalla-Favera, R. (2015). Germinal centres and B cell lymphomagenesis. *Nature Reviews Immunology*, *15*(3), 172-184. https://doi.org/10.1038/nri3814

Begić, V., Korać, P., Gašparov, S., Rozman, M., Simicic, P., & Zidovec-Lepej, S. (2022). Molecular Characterisation of Epstein–Barr Virus in Classical Hodgkin Lymphoma. *International Journal of Molecular Sciences*, *23*(24), 15635. https://www.mdpi.com/1422-0067/23/24/15635

Bell, A. I., Groves, K., Kelly, G. L., Croom-Carter, D., Hui, E., Chan, A. T., & Rickinson, A. B. (2006). Analysis of Epstein–Barr virus latent gene expression in endemic Burkitt's lymphoma and nasopharyngeal carcinoma tumour cells by

using quantitative real-time PCR assays. *Journal of General Virology*, *87*(10), 2885-2890.

Bell, M. J., Abbott, R. J., Croft, N. P., Hislop, A. D., & Burrows, S. R. (2009). An HLA-A2-restricted T-cell epitope mapped to the BNLF2a immune evasion protein of Epstein-Barr virus that inhibits TAP. *Journal of Virology*, *83*(6), 2783-2788.

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). Calling somatic SNVs and indels with Mutect2. *Biorxiv*, 861054.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Benninger-Döring, G., Pepperl, S., Deml, L., Modrow, S., Wolf, H., & Jilg, W. (1999). Frequency of CD8+ T lymphocytes specific for lytic and latent antigens of Epstein–Barr virus in healthy virus carriers. *Virology*, *264*(2), 289-297.

Berenstein, A. J., Lorenzetti, M. A., & Preciado, M. V. (2018). Recombination rates along the entire Epstein Barr virus genome display a highly heterogeneous landscape. *Infection, Genetics and Evolution*, *65*, 96-103.

Bian, X., Zhuo, Y., Zhou, L., Zhu, Y. Z., Li, N., & Wu, J. L. (2025). Carboxylic acid metabolism in cancer: Mechanisms, microenvironment interactions, and therapeutic opportunities. *Metabolism*, *171*, 156334. https://doi.org/10.1016/j.metabol.2025.156334

Bidot, S. A., S. . (2021). *Burkitt lymphoma*. PathologyOutlines.com. Retrieved October 4th, 2025 from https://www.pathologyoutlines.com/topic/lymphomaburkitt.html

Bjornevik, K., Münz, C., Cohen, J. I., & Ascherio, A. (2023). Epstein-Barr virus as a leading cause of multiple sclerosis: mechanisms and implications. *Nat Rev Neurol*, *19*(3), 160-171. https://doi.org/10.1038/s41582-023-00775-5

Blake, N., Haigh, T., Shaka'a, G., Croom-Carter, D., & Rickinson, A. (2000). The importance of exogenous antigen in priming the human CD8+ T cell response:

lessons from the EBV nuclear antigen EBNA1. *The Journal of Immunology*, *165*(12), 7078-7087.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bonaldo, M. C., Martins, M. A., Rudersdorf, R., Mudd, P. A., Sacha, J. B., Piaskowski, S. M., Costa Neves, P. C., Veloso de Santana, M. G., Vojnov, L., Capuano, S., 3rd, Rakasz, E. G., Wilson, N. A., Fulkerson, J., Sadoff, J. C., Watkins, D. I., & Galler, R. (2010). Recombinant yellow fever vaccine virus 17D expressing simian immunodeficiency virus SIVmac239 gag induces SIV-specific CD8+ T-cell responses in rhesus macaques. *J Virol*, *84*(7), 3699-3706. https://doi.org/10.1128/jvi.02255-09

Bonaldo, M. C., Mello, S. M., Trindade, G. F., Rangel, A. A., Duarte, A. S., Oliveira, P. J., Freire, M. S., Kubelka, C. F., & Galler, R. (2007). Construction and characterization of recombinant flaviviruses bearing insertions between E and NS1 genes. *Virol J*, *4*, 115. https://doi.org/10.1186/1743-422x-4-115

Bornkamm, G. W., von Knebel-Doeberitz, M., & Lenoir, G. M. (1984). No evidence for differences in the Epstein-Barr virus genome carried in Burkitt lymphoma cells and nonmalignant lymphoblastoid cells from the same patients. *Proc Natl Acad Sci U S A*, *81*(15), 4930-4934. https://doi.org/10.1073/pnas.81.15.4930

Borozan, I., Zapatka, M., Frappier, L., & Ferretti, V. (2018). Analysis of Epstein-Barr virus genomes and expression profiles in gastric adenocarcinoma. *Journal of Virology*, *92*(2), 10.1128/jvi. 01239-01217.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525-527.

Briercheck, E. L., Ravishankar, S., Ahmed, E. H., Carías Alvarado, C. C., Barrios Menéndez, J. C., Silva, O., Solórzano-Ortiz, E., Siliézar Tala, M. M., Stevenson, P., & Xu, Y. (2024). Geographic EBV variants confound disease-specific variant

interpretation and predict variable immune therapy responses. *Blood Advances*, *8*(14), 3731-3744.

Brooks, J. M., Long, H. M., Tierney, R. J., Shannon-Lowe, C., Leese, A. M., Fitzpatrick, M., Taylor, G. S., & Rickinson, A. B. (2016). Early T cell recognition of B cells following Epstein-Barr virus infection: identifying potential targets for prophylactic vaccination. *PLOS Pathogens*, *12*(4), e1005549.

Brooks, L., Yao, Q., Rickinson, A., & Young, L. (1992). Epstein-Barr virus latent gene transcription in nasopharyngeal carcinoma cells: coexpression of EBNA1, LMP1, and LMP2 transcripts. *Journal of Virology*, *66*(5), 2689-2697.

Brown, J. C. (2014). The role of DNA repair in herpesvirus pathogenesis. *Genomics*, *104*(4), 287-294.

Burkitt, D. (1958). A sarcoma involving the jaws in African children. *British journal of surgery*, *46*(197), 218-223.

Burkitt, D. (1962). A “tumour safari” in East and Central Africa. *British journal of cancer*, *16*(3), 379.

Burkitt, D. P. (1969). Etiology of Burkitt's lymphoma—an alternative hypothesis to a vectored virus. *Journal of the National Cancer Institute*, *42*(1), 19-28.

Burmeister, T., Molkentin, M., Schwartz, S., Gökbuget, N., Hoelzer, D., Thiel, E., & Reinhardt, R. (2013). Erroneous class switching and false VDJ recombination: molecular dissection of t(8;14)/MYC-IGH translocations in Burkitt-type lymphoblastic leukemia/B-cell lymphoma. *Mol Oncol*, *7*(4), 850-858. https://doi.org/10.1016/j.molonc.2013.04.006

Burningham, Z., Hashibe, M., Spector, L., & Schiffman, J. The epidemiology of sarcoma. Clin Sarcoma Res. 2012; 2 (1): 14. In: Epub 2012/10/06. https://doi. org/10.1186/2045-3329-2-14 PMID: 23036164.

Burrows, S., Gardner, J., Khanna, R., Steward, T., Moss, D., Rodda, S., & Suhrbier, A. (1994). Five new cytotoxic T cell epitopes identified within Epstein-Barr virus nuclear antigen 3. *Journal of General Virology*, *75*(9), 2489-2493.

Cahir-McFarland, E. D., Davidson, D. M., Schauer, S. L., Duong, J., & Kieff, E. (2000). NF-κB inhibition causes spontaneous apoptosis in Epstein–Barr virus-transformed lymphoblastoid cells. *Proceedings of the National Academy of Sciences*, *97*(11), 6055-6060.

Cai, J., Zhang, B., Li, Y., Zhu, W., Akihisa, T., Li, W., Kikuchi, T., Liu, W., Feng, F., & Zhang, J. (2021). Prophylactic and Therapeutic EBV Vaccines: Major Scientific Obstacles, Historical Progress, and Future Direction. *Vaccines*, *9*(11), 1290. https://www.mdpi.com/2076-393X/9/11/1290

Cai, L., Ye, Y., Jiang, Q., Chen, Y., Lyu, X., Li, J., Wang, S., Liu, T., Cai, H., Yao, K., Li, J. L., & Li, X. (2015). Epstein-Barr virus-encoded microRNA BART1 induces tumour metastasis by regulating PTEN-dependent pathways in nasopharyngeal carcinoma. *Nat Commun*, *6*, 7353. https://doi.org/10.1038/ncomms8353

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*(1), 421.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972-1973.

Capone, S., D'Alise, A. M., Ammendola, V., Colloca, S., Cortese, R., Nicosia, A., & Folgori, A. (2013). Development of chimpanzee adenoviruses as vaccine vectors: challenges and successes emerging from clinical trials. *Expert Rev Vaccines*, *12*(4), 379-393. https://doi.org/10.1586/erv.13.15

Carnel, N., Lancia, H. H., Guinier, C., & Benichou, G. (2023). Pathways of Antigen Recognition by T Cells in Allograft Rejection. *Transplantation*, *107*(4), 827-837. https://doi.org/10.1097/tp.0000000000004420

Castillo, J. J., Beltran, B. E., Miranda, R. N., Young, K. H., Chavez, J. C., & Sotomayor, E. M. (2018). EBV-positive diffuse large B-cell lymphoma, not otherwise specified: 2018 update on diagnosis, risk-stratification and management.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular biology and evolution*, *17*(4), 540-552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Cerimele, F., Battle, T., Lynch, R., Frank, D. A., Murad, E., Cohen, C., Macaron, N., Sixbey, J., Smith, K., Watnick, R. S., Eliopoulos, A., Shehata, B., & Arbiser, J. L. (2005). Reactive oxygen signaling and MAPK activation distinguish Epstein-Barr Virus (EBV)-positive versus EBV-negative Burkitt's lymphoma. *Proc Natl Acad Sci U S A*, *102*(1), 175-179. https://doi.org/10.1073/pnas.0408381102

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-13015-10047-13748.

Chen, J., & Longnecker, R. (2019). Epithelial cell infection by Epstein–Barr virus. *FEMS Microbiology Reviews*, *43*(6), 674-683. https://doi.org/10.1093/femsre/fuz023

Chen, X., Kamranvar, S. A., & Masucci, M. (2016). Oxidative stress enables Epstein–Barr virus-induced B-cell transformation by posttranscriptional regulation of viral and cellular growth-promoting factors. *Oncogene*, *35*(29), 3807-3816.

Chêne, A., Donati, D., Guerreiro-Cacais, A. O., Levitsky, V., Chen, Q., Falk, K. I., Orem, J., Kironde, F., Wahlgren, M., & Bejarano, M. T. (2007). A molecular link between malaria and Epstein–Barr virus reactivation. *PLOS Pathogens*, *3*(6), e80.

Chia, W. K., Wang, W. W., Teo, M., Tai, W. M., Lim, W. T., Tan, E. H., Leong, S. S., Sun, L., Chen, J. J., Gottschalk, S., & Toh, H. C. (2012). A phase II study evaluating the safety and efficacy of an adenovirus-ΔLMP1-LMP2 transduced dendritic cell vaccine in patients with advanced metastatic nasopharyngeal carcinoma. *Ann Oncol*, *23*(4), 997-1005. https://doi.org/10.1093/annonc/mdr341

Chiara, M., Manzari, C., Lionetti, C., Mechelli, R., Anastasiadou, E., Chiara Buscarinu, M., Ristori, G., Salvetti, M., Picardi, E., D'Erchia, A. M., Pesole, G., & Horner, D. S. (2016). Geographic Population Structure in Epstein-Barr Virus Revealed by Comparative Genomics. *Genome Biology and Evolution*, *8*(11), 3284-3291. https://doi.org/10.1093/gbe/evw226

Choi, H., Lee, H., Kim, S. R., Gho, Y. S., & Lee, S. K. (2013). Epstein-Barr virus-encoded microRNA BART15-3p promotes cell apoptosis partially by targeting BRUCE. *Journal of Virology*, *87*(14), 8135-8144.

Chung, A. K., OuYang, C. N., Liu, H., Chao, M., Luo, J. D., Lee, C. Y., Lu, Y. J., Chung, I. C., Chen, L. C., & Wu, S. M. (2019). Targeted sequencing of cancer-related genes in nasopharyngeal carcinoma identifies mutations in the TGF-β pathway. *Cancer medicine*, *8*(11), 5116-5127.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). 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*(2), 80-92.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, *83*(403), 596-610.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., & Wilczynski, B. (2009). Biopython: freely available

Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422.

Connolly, S. A., Jackson, J. O., Jardetzky, T. S., & Longnecker, R. (2011). Fusing structure and function: a structural view of the herpesvirus entry machinery. *Nature Reviews Microbiology*, *9*(5), 369-381. https://doi.org/10.1038/nrmicro2548

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938-2940.

Correia, S., Bridges, R., Wegner, F., Venturini, C., Palser, A., Middeldorp, J. M., Cohen, J. I., Lorenzetti, M. A., Bassano, I., & White, R. E. (2018). Sequence variation of Epstein-Barr virus: viral types, geography, codon usage, and diseases. *Journal of Virology*, *92*(22), 10.1128/jvi. 01132-01118.

Cosmopoulos, K., Pegtel, M., Hawkins, J., Moffett, H., Novina, C., Middeldorp, J., & Thorley-Lawson, D. A. (2009). Comprehensive profiling of Epstein-Barr virus microRNAs in nasopharyngeal carcinoma. *Journal of Virology*, *83*(5), 2357-2367.

Coughlan, L. (2020). Factors which contribute to the immunogenicity of non-replicating adenoviral vectored vaccines. *Frontiers in Immunology*, *11*, 909.

da Costa, V. G., Marques-Silva, A. C., & Moreli, M. L. (2015). The Epstein-Barr virus latent membrane protein-1 (LMP1) 30-bp deletion and XhoI-polymorphism in nasopharyngeal carcinoma: a meta-analysis of observational studies. *Systematic Reviews*, *4*(1), 46. https://doi.org/10.1186/s13643-015-0037-z

Damerell, V., Pepper, M. S., & Prince, S. (2021). Molecular mechanisms underpinning sarcomas and implications for current and future therapy. *Signal Transduction and Targeted Therapy*, *6*(1), 246. https://doi.org/10.1038/s41392-021-00647-8

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Danve, C., Decaussin, G., Busson, P., & Ooka, T. (2001). Growth transformation of primary epithelial cells with a NPC-derived Epstein-Barr virus strain. *Virology*, *288*(2), 223-235. https://doi.org/10.1006/viro.2001.1072

Darwin, C. (1964). *On the origin of species: A facsimile of the first edition*. Harvard University Press.

Dawson, C. W., Tramountanis, G., Eliopoulos, A. G., & Young, L. S. (2003). Epstein-Barr Virus Latent Membrane Protein 1 (LMP1) Activates the Phosphatidylinositol 3-Kinase/Akt Pathway to Promote Cell Survival and Induce Actin Filament Remodeling *. *Journal of Biological Chemistry*, *278*(6), 3694-3704. https://doi.org/10.1074/jbc.M209840200

de Martel, C., Georges, D., Bray, F., Ferlay, J., & Clifford, G. M. (2020). Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *The Lancet Global Health*, *8*(2), e180-e190.

de Mel, S., Hue, S. S., Jeyasekharan, A. D., Chng, W. J., & Ng, S. B. (2019). Molecular pathogenic pathways in extranodal NK/T cell lymphoma. *J Hematol Oncol*, *12*(1), 33. https://doi.org/10.1186/s13045-019-0716-7

De Palma, M., Biziato, D., & Petrova, T. V. (2017). Microenvironmental regulation of tumour angiogenesis. *Nature Reviews Cancer*, *17*(8), 457-474. https://doi.org/10.1038/nrc.2017.51

de-Thé, G., Day, N. E., Geser, A., Lavoué, M. F., Ho, J. H., Simons, M. J., Sohier, R., Tukei, P., Vonka, V., & Zavadova, H. (1975). Sero-epidemiology of the Epstein-Barr virus: preliminary analysis of an international study - a review. *IARC Sci Publ (1971)*(11 Pt 2), 3-16.

Delattre, O., Zucman, J., Melot, T., Garau, X. S., Zucker, J.-M., Lenoir, G. M., Ambros, P. F., Sheer, D., Turc-Carel, C., & Triche, T. J. (1994). The Ewing family of tumors--a subgroup of small-round-cell tumors defined by specific chimeric transcripts. *New England Journal of Medicine*, *331*(5), 294-299.

Dhanasekaran, R., Deutzmann, A., Mahauad-Fernandez, W. D., Hansen, A. S., Gouw, A. M., & Felsher, D. W. (2022). The MYC oncogene—the grand orchestrator of cancer growth and immune evasion. *Nature reviews Clinical oncology*, *19*(1), 23-36.

Didelot, X., & Wilson, D. J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology*, *11*(2), e1004041.

Dirmeier, U., Neuhierl, B., Kilger, E., Reisbach, G., Sandberg, M. L., & Hammerschmidt, W. (2003). Latent Membrane Protein 1 Is Critical for Efficient Growth Transformation of Human B Cells by Epstein-Barr Virus1. *Cancer research*, *63*(11), 2982-2989.

Dolan, B. P., Gibbs, K. D., Jr., & Ostrand-Rosenberg, S. (2006). Dendritic cells cross-dressed with peptide MHC class I complexes prime CD8+ T cells. *J Immunol*, *177*(9), 6018-6024. https://doi.org/10.4049/jimmunol.177.9.6018

Dolina, J. S., Lee, J., Brightman, S. E., McArdle, S., Hall, S. M., Thota, R. R., Zavala, K. S., Lanka, M., Premlal, A. L. R., & Greenbaum, J. A. (2023). Linked CD4+/CD8+ T cell neoantigen vaccination overcomes immune checkpoint blockade resistance and enables tumor regression. *The Journal of clinical investigation*, *133*(17).

Dong, H., Strome, S. E., Salomao, D. R., Tamura, H., Hirano, F., Flies, D. B., Roche, P. C., Lu, J., Zhu, G., & Tamada, K. (2002). Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion. *Nature medicine*, *8*(8), 793-800.

Dong, M., Wang, H. Y., Zhao, X. X., Chen, J. N., Zhang, Y. W., Huang, Y., Xue, L., Li, H. G., Du, H., Wu, X. Y., & Shao, C. K. (2016). Expression and prognostic roles of PIK3CA, JAK2, PD-L1, and PD-L2 in Epstein-Barr virus-associated gastric carcinoma. *Hum Pathol*, *53*, 25-34. https://doi.org/10.1016/j.humpath.2016.02.007

Donohoe, C. L., Lysaght, J., O'Sullivan, J., & Reynolds, J. V. (2017). Emerging Concepts Linking Obesity with the Hallmarks of Cancer. *Trends in Endocrinology & Metabolism*, *28*(1), 46-62. https://doi.org/10.1016/j.tem.2016.08.004

Drost, J., Van Boxtel, R., Blokzijl, F., Mizutani, T., Sasaki, N., Sasselli, V., de Ligt, J., Behjati, S., Grolleman, J. E., & van Wezel, T. (2017). Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science*, *358*(6360), 234-238.

Dubey, P., Ha, C. S., Ang, K. K., El-Naggar, A. K., Knapp, C., Byers, R. M., & Morrison, W. H. (1998). Nonnasopharyngeal lymphoepithelioma of the head and neck. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *82*(8), 1556-1562.

Dudareva, M., Andrews, L., Gilbert, S. C., Bejon, P., Marsh, K., Mwacharo, J., Kai, O., Nicosia, A., & Hill, A. V. (2009). Prevalence of serum neutralizing antibodies against chimpanzee adenovirus 63 and human adenovirus 5 in Kenyan children, in the context of vaccine vector efficacy. *Vaccine*, *27*(27), 3501-3504. https://doi.org/10.1016/j.vaccine.2009.03.080

Dunleavy, K., Pittaluga, S., Shovlin, M., Steinberg, S. M., Cole, D., Grant, C., Widemann, B., Staudt, L. M., Jaffe, E. S., & Little, R. F. (2013). Low-intensity therapy in adults with Burkitt's lymphoma. *New England Journal of Medicine*, *369*(20), 1915-1925.

Dunmire, S. K., Verghese, P. S., & Balfour, H. H. (2018). Primary Epstein-Barr virus infection. *Journal of clinical virology*, *102*, 84-92. https://doi.org/https://doi.org/10.1016/j.jcv.2018.03.001

Duvall, M. G., Precopio, M. L., Ambrozak, D. A., Jaye, A., McMichael, A. J., Whittle, H. C., Roederer, M., Rowland-Jones, S. L., & Koup, R. A. (2008). Polyfunctional T cell responses are a hallmark of HIV-2 infection. *European Journal of Immunology*, *38*(2), 350-363. https://doi.org/https://doi.org/10.1002/eji.200737768

Elliott, S. L., Suhrbier, A., Miles, J. J., Lawrence, G., Pye, S. J., Le, T. T., Rosenstengel, A., Nguyen, T., Allworth, A., & Burrows, S. R. (2008). Phase I trial of a CD8+ T-cell peptide epitope-based vaccine for infectious mononucleosis. *Journal of Virology*, *82*(3), 1448-1457.

Emfietzoglou, R., Pettas, E., Georgaki, M., Papadopoulou, E., Theofilou, V. I., Papadogeorgakis, N., Piperi, E., Lopes, M. A., & Nikitakis, N. G. (2022). Lymphoepithelial subtype of oral squamous cell carcinoma: report of an EBV-negative case and literature review. *Dentistry journal*, *10*(9), 165.

Eskander, G., Abdelhamid, S. G., Wahdan, S. A., & Radwan, S. M. (2025). Insights on the crosstalk among different cell death mechanisms. *Cell Death Discovery*, *11*(1), 56. https://doi.org/10.1038/s41420-025-02328-9

Esteller, M., Dawson, M. A., Kadoch, C., Rassool, F. V., Jones, P. A., & Baylin, S. B. (2024). The Epigenetic Hallmarks of Cancer. *Cancer Discov*, *14*(10), 1783-1809. https://doi.org/10.1158/2159-8290.Cd-24-0296

Fang, W., Zhang, J., Hong, S., Zhan, J., Chen, N., Qin, T., Tang, Y., Zhang, Y., Kang, S., Zhou, T., Wu, X., Liang, W., Hu, Z., Ma, Y., Zhao, Y., Tian, Y., Yang, Y., Xue, C., Yan, Y., . . . Zhang, L. (2014). EBV-driven LMP1 and IFN-γ up-regulate PD-L1 in nasopharyngeal carcinoma: Implications for oncotargeted therapy. *Oncotarget*, *5*(23), 12189-12202. https://doi.org/10.18632/oncotarget.2608

Fantini, D., Vidimar, V., Yu, Y., Condello, S., & Meeks, J. J. (2020). MutSignatures: an R package for extraction and analysis of cancer mutational signatures. *Scientific Reports*, *10*(1), 18217. https://doi.org/10.1038/s41598-020-75062-0

Farrell, P. J. (2019). Epstein–Barr virus and cancer. *Annual Review of Pathology: Mechanisms of Disease*, *14*(1), 29-53.

Ferreira, D. A., Tayyar, Y., Idris, A., & McMillan, N. A. (2021). A "hit-and-run" affair–a possible link for cancer progression in virally driven cancers. *Biochimica et biophysica acta (BBA)-Reviews on Cancer*, *1875*(1), 188476.

Fish, K., Comoglio, F., Shaffer III, A. L., Ji, Y., Pan, K.-T., Scheich, S., Oellerich, A., Doebele, C., Ikeda, M., & Schaller, S. J. (2020). Rewiring of B cell receptor signaling by Epstein–Barr virus LMP2A. *Proceedings of the National Academy of Sciences*, *117*(42), 26318-26327.

Folegatti, P. M., Jenkin, D., Morris, S., Gilbert, S., Kim, D., Robertson, J. S., Smith, E. R., Martin, E., Gurwith, M., & Chen, R. T. (2022). Vaccines based on the replication-deficient simian adenoviral vector ChAdOx1: Standardized template with key considerations for a risk/benefit assessment. *Vaccine*, *40*(35), 5248-5262. https://doi.org/https://doi.org/10.1016/j.vaccine.2022.06.008

Frappier, L. (2012). Role of EBNA1 in NPC tumourigenesis. Seminars in Cancer Biology,

Gajewski, T. F., Schreiber, H., & Fu, Y. X. (2013). Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol*, *14*(10), 1014-1022. https://doi.org/10.1038/ni.2703

Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC bioinformatics*, *11*(1), 367.

Gerber, P., Walsh, J., Rosenblum, E., & Purcell, R. (1969). Association of EB-virus infection with the post-perfusion syndrome. *The Lancet*, *293*(7595), 593-596.

Gilbert, S. C., Schneider, J., Hannan, C. M., Hu, J. T., Plebanski, M., Sinden, R., & Hill, A. V. (2002). Enhanced CD8 T cell immunogenicity and protective efficacy in a mouse malaria model using a recombinant adenoviral vaccine in heterologous prime-boost immunisation regimes. *Vaccine*, *20*(7-8), 1039-1045. https://doi.org/10.1016/s0264-410x(01)00450-9

Ginestet, C. (2011). ggplot2: elegant graphics for data analysis. In: Oxford University Press.

Godsmark, G., LA, D. E. S. R., & Mowla, S. (2021). Activation-Induced Cytidine Deaminase Promotes Proliferation and Enhances Chemoresistance and Migration in B-cell Lymphoma. *Anticancer Res*, *41*(1), 237-247. https://doi.org/10.21873/anticanres.14770

Grande, B. M., Gerhard, D. S., Jiang, A., Griner, N. B., Abramson, J. S., Alexander, T. B., Allen, H., Ayers, L. W., Bethony, J. M., Bhatia, K., Bowen, J., Casper, C., Choi, J. K., Culibrk, L., Davidsen, T. M., Dyer, M. A., Gastier-Foster, J. M., Gesuwan, P., Greiner, T. C., . . . Staudt, L. M. (2019). Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood*, *133*(12), 1313-1324. https://doi.org/10.1182/blood-2018-09-871418

Gruhne, B., Sompallae, R., Marescotti, D., Kamranvar, S. A., Gastaldello, S., & Masucci, M. G. (2009). The Epstein-Barr virus nuclear antigen-1 promotes genomic instability via induction of reactive oxygen species. *Proc Natl Acad Sci U S A*, *106*(7), 2313-2318. https://doi.org/10.1073/pnas.0810619106

Gu, S. Y., Huang, T. M., Ruan, L., Miao, Y. H., Lu, H., Chu, C. M., Motz, M., & Wolf, H. (1995). First EBV vaccine trial in humans using recombinant vaccinia virus expressing the major membrane antigen. *Dev Biol Stand*, *84*, 171-177.

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847-2849. https://doi.org/10.1093/bioinformatics/btw313

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Guterres, A. N., & Villanueva, J. (2020). Targeting telomerase for cancer therapy. *Oncogene*, *39*(36), 5811-5824. https://doi.org/10.1038/s41388-020-01405-w

Han, S. S., Yun, H., Son, D. J., Tompkins, V. S., Peng, L., Chung, S. T., Kim, J. S., Park, E. S., & Janz, S. (2010). NF-kappaB/STAT3/PI3K signaling crosstalk in iMyc E mu B lymphoma. *Mol Cancer*, *9*, 97. https://doi.org/10.1186/1476-4598-9-97

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discov*, *12*(1), 31-46. https://doi.org/10.1158/2159-8290.Cd-21-1059

Hanahan, D. (2026). Hallmarks of cancer—Then and now, and beyond. *Cell*.

Hanahan, D., Michielin, O., & Pittet, M. J. (2025). Convergent inducers and effectors of T cell paralysis in the tumour microenvironment. *Nature Reviews Cancer*, *25*(1), 41-58. https://doi.org/10.1038/s41568-024-00761-z

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57-70.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646-674.

Hankey, W., Frankel, W. L., & Groden, J. (2018). Functions of the APC tumor suppressor protein dependent and independent of canonical WNT signaling: implications for therapeutic targeting. *Cancer and Metastasis Reviews*, *37*(1), 159-172.

Hansen, S. G., Vieville, C., Whizin, N., Coyne-Johnson, L., Siess, D. C., Drummond, D. D., Legasse, A. W., Axthelm, M. K., Oswald, K., Trubey, C. M., Piatak, M., Jr., Lifson, J. D., Nelson, J. A., Jarvis, M. A., & Picker, L. J. (2009). Effector memory T cell responses are associated with protection of rhesus monkeys from

mucosal simian immunodeficiency virus challenge. *Nat Med*, *15*(3), 293-299. https://doi.org/10.1038/nm.1935

Harris, B. H. L., Macaulay, V. M., Harris, D. A., Klenerman, P., Karpe, F., Lord, S. R., Harris, A. L., & Buffa, F. M. (2022). Obesity: a perfect storm for carcinogenesis. *Cancer Metastasis Rev*, *41*(3), 491-515. https://doi.org/10.1007/s10555-022-10046-2

Hartlage, A. S., Liu, T., Patton, J. T., Garman, S. L., Zhang, X., Kurt, H., Lozanski, G., Lustberg, M. E., Caligiuri, M. A., & Baiocchi, R. A. (2015). The Epstein–Barr Virus lytic protein BZLF1 as a candidate target antigen for vaccine development. *Cancer immunology research*, *3*(7), 787-794.

Hirabayashi, M., Georges, D., Clifford, G. M., & de Martel, C. (2023). Estimating the Global Burden of Epstein-Barr Virus–Associated Gastric Cancer: A Systematic Review and Meta-Analysis. *Clinical Gastroenterology and Hepatology*, *21*(4), 922-930.e921. https://doi.org/https://doi.org/10.1016/j.cgh.2022.07.042

Hislop, A. D., Taylor, G. S., Sauce, D., & Rickinson, A. B. (2007). Cellular Responses to Viral Infection in Humans: Lessons from Epstein-Barr Virus. *Annual Review of Immunology*, *25*(Volume 25, 2007), 587-617. https://doi.org/https://doi.org/10.1146/annurev.immunol.25.022106.141553

Hoelzer, D., Walewski, J., Döhner, H., Viardot, A., Hiddemann, W., Spiekermann, K., Serve, H., Dührsen, U., Hüttmann, A., Thiel, E., Dengler, J., Kneba, M., Schaich, M., Schmidt-Wolf, I. G. H., Beck, J., Hertenstein, B., Reichle, A., Domanska-Czyz, K., Fietkau, R., . . . Leukemia, f. t. G. M. S. G. f. A. A. L. (2014). Improved outcome of adult Burkitt lymphoma/leukemia with rituximab and chemotherapy: report of a large prospective multicenter trial. *Blood*, *124*(26), 3870-3879. https://doi.org/10.1182/blood-2014-03-563627

Hofer, T., Rossi, M., Carboni, S., Di Berardino Besson, W., von Laer, D., Wollmann, G., Derouazi, M., & Santiago-Raber, M.-L. (2021). Heterologous prime-boost

vaccination with a peptide-based vaccine and viral vector reshapes dendritic cell, CD4+ and CD8+ T cell phenotypes to improve the antitumor therapeutic effect. *Cancers*, *13*(23), 6107.

Hollingsworth, R. E., & Jansen, K. (2019). Turning the corner on therapeutic cancer vaccines. *npj Vaccines*, *4*(1), 7. https://doi.org/10.1038/s41541-019-0103-y

Holtkamp, S. J., Lagoda, F. R., Lister, A., Harish, P., Kleymann, U., Pesch, T., Soon, C. F., Pirmohamed, M., Naisbitt, D., & Trautwein, M. (2024). Polyfunctional T cells and unique cytokine clusters imprint the anti rAAV2/rAAV9 vector immune response. *Frontiers in Immunology*, *15*, 1450524.

Houen, G., & Trier, N. H. (2021). Epstein-Barr Virus and Systemic Autoimmune Diseases [Review]. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.587380

Hui, E. P., Taylor, G. S., Jia, H., Ma, B. B., Chan, S. L., Ho, R., Wong, W.-L., Wilson, S., Johnson, B. F., & Edwards, C. (2013). Phase I trial of recombinant modified vaccinia ankara encoding Epstein–Barr viral tumor antigens in nasopharyngeal carcinoma patients. *Cancer research*, *73*(6), 1676-1688.

Ileiwat, Z. E., Tabish, T. A., Zinovkin, D. A., Yuzugulen, J., Arghiani, N., & Pranjol, M. Z. I. (2022). The mechanistic immunosuppressive role of the tumour vasculature and potential nanoparticle-mediated therapeutic strategies. *Front Immunol*, *13*, 976677. https://doi.org/10.3389/fimmu.2022.976677

Inagaki, T., Sato, Y., Ito, J., Takaki, M., Okuno, Y., Yaguchi, M., Masud, H. A. A., Watanabe, T., Sato, K., & Iwami, S. (2021). Direct evidence of abortive lytic infection-mediated establishment of Epstein-Barr virus latency during B-cell infection. *Frontiers in Microbiology*, *11*, 575255.

Isogawa, M., Onodera, T., Ainai, A., Kotaki, R., Kanno, T., Saito, S., Tobiume, M., Tokunaga, K., Hara, M., Hirota, Y., Suzuki, T., Takahashi, Y., & Tsuru, T. (2025). Prolonged effects of adenoviral vector priming on T-cell cytokine

production in heterologous adenoviral vector/mRNA COVID-19 vaccination regimens. *Scientific reports*, *15*(1), 18684. https://doi.org/10.1038/s41598-025-00054-x

Iwakiri, D., & Takada, K. (2010). Role of EBERs in the pathogenesis of EBV infection. *Advances in cancer research*, *107*, 119-136.

Iwakiri, D., Zhou, L., Samanta, M., Matsumoto, M., Ebihara, T., Seya, T., Imai, S., Fujieda, M., Kawa, K., & Takada, K. (2009). Epstein-Barr virus (EBV)–encoded small RNA is released from EBV-infected cells and activates signaling from toll-like receptor 3. *Journal of Experimental Medicine*, *206*(10), 2091-2099.

Jeon, J. H., Kim, S., Kim, S.-Y., Shin, K.-S., Park, B., Chang, S., Kang, C.-Y., Kim, Y.-J., Park, J.-E., & Youk, S. (2025). Heterologous prime-boost vaccination drives stromal activation and adaptive immunity against SARS-CoV-2 variants. *Frontiers in Immunology*, *16*, 1597417.

Jiang, Y., Li, Y., & Zhu, B. (2015). T-cell exhaustion in the tumor microenvironment. *Cell Death & Disease*, *6*(6), e1792-e1792. https://doi.org/10.1038/cddis.2015.162

Jo, V. Y., & Fletcher, C. D. M. (2014). WHO classification of soft tissue tumours: an update based on the 2013 (4th) edition. *Pathology - Journal of the RCPA*, *46*(2), 95-104. https://doi.org/10.1097/pat.0000000000000050

Jochum, S., Moosmann, A., Lang, S., Hammerschmidt, W., & Zeidler, R. (2012). The EBV immunoevasins vIL-10 and BNLF2a protect newly infected B cells from immune recognition and elimination. *PLOS Pathogens*, *8*(5), e1002704.

Joerger, A. C., Stiewe, T., & Soussi, T. (2025). TP53: the unluckiest of genes? *Cell Death & Differentiation*, *32*(2), 219-224.

Jones-Weinert, C., Mainz, L., & Karlseder, J. (2025). Telomere function and regulation from mouse models to human ageing and disease. *Nature Reviews Molecular Cell Biology*, *26*(4), 297-313. https://doi.org/10.1038/s41580-024-00800-5

Jonigk, D., Laenger, F., Maegel, L., Izykowski, N., Rische, J., Tiede, C., Klein, C., Maecker-Kolhoff, B., Kreipe, H., & Hussein, K. (2012). Molecular and clinicopathological analysis of Epstein-Barr virus–associated posttransplant smooth muscle tumors. *American Journal of Transplantation*, *12*(7), 1908-1917.

Kajtár, B. (2024). *CLL / SLL.* PathologyOutlines.com. Retrieved October 4th, 2025. from https://www.pathologyoutlines.com/topic/lymphomacll.html

Kalla, M., & Hammerschmidt, W. (2012). Human B cells on their route to latent infection–early but transient expression of lytic genes of Epstein-Barr virus. *European journal of cell biology*, *91*(1), 65-69.

Kanda, T., Yajima, M., & Ikuta, K. (2019). Epstein-Barr virus strain variation and cancer. *Cancer science*, *110*(4), 1132-1139.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., & Wyczalkowski, M. A. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*(7471), 333-339.

Kaneda, A., Matsusaka, K., Aburatani, H., & Fukayama, M. (2012). Epstein–Barr virus infection as an epigenetic driver of tumorigenesis. *Cancer research*, *72*(14), 3445-3450.

Kang, D., Skalsky, R. L., & Cullen, B. R. (2015). EBV BART MicroRNAs Target Multiple Pro-apoptotic Cellular Genes to Promote Epithelial Cell Survival. *PLOS Pathogens*, *11*(6), e1004979. https://doi.org/10.1371/journal.ppat.1004979

Karihtala, P., Kilpivaara, O., & Porvari, K. (2023). Mutational signatures and their association with survival and gene expression in urological carcinomas. *Neoplasia*, *44*, 100933. https://doi.org/10.1016/j.neo.2023.100933

Kassambara, A. (2019). rstatix: Pipe-friendly framework for basic statistical tests. *CRAN: Contributed packages*.

Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, *30*(14), 3059-3066.

Kaur, A., Kumar, A., Kumari, G., Muduli, R., Das, M., Kundu, R., Mukherjee, S., & Majumdar, T. (2024). Rational design and computational evaluation of a multi-epitope vaccine for monkeypox virus: Insights into binding stability and immunological memory. *Heliyon*, *10*(16). https://doi.org/10.1016/j.heliyon.2024.e36154

Kaymaz, Y., Oduor, C. I., Aydemir, O., Luftig, M. A., Otieno, J. A., Ong'echa, J. M., Bailey, J. A., & Moormann, A. M. (2020). Epstein-Barr Virus Genomes Reveal Population Structure and Type 1 Association with Endemic Burkitt Lymphoma. *Journal of Virology*, *94*(17), e02007-02019. https://doi.org/10.1128/jvi.02007-19

Kelly, G., Bell, A., & Rickinson, A. (2002). Epstein–Barr virus–associated Burkitt lymphomagenesis selects for downregulation of the nuclear antigen EBNA2. *Nature medicine*, *8*(10), 1098-1104.

Kelly Gemma, L., Stylianou, J., Rasaiyaah, J., Wei, W., Thomas, W., Croom-Carter, D., Kohler, C., Spang, R., Woodman, C., Kellam, P., Rickinson Alan, B., & Bell Andrew, I. (2013). Different Patterns of Epstein-Barr Virus Latency in Endemic Burkitt Lymphoma (BL) Lead to Distinct Variants within the BL-Associated Gene Expression Signature. *Journal of Virology*, *87*(5), 2882-2894. https://doi.org/10.1128/jvi.03003-12

Kelly, G. L., Long, H. M., Stylianou, J., Thomas, W. A., Leese, A., Bell, A. I., Bornkamm, G. W., Mautner, J., Rickinson, A. B., & Rowe, M. (2009). An Epstein-Barr virus anti-apoptotic protein constitutively expressed in transformed cells and implicated in burkitt lymphomagenesis: the Wp/BHRF1 link. *PLOS Pathogens*, *5*(3), e1000341.

Kelly, G. L., Milner, A. E., Baldwin, G. S., Bell, A. I., & Rickinson, A. B. (2006). Three restricted forms of Epstein–Barr virus latency counteracting apoptosis in c-myc-expressing Burkitt lymphoma cells. *Proceedings of the National Academy of Sciences*, *103*(40), 14935-14940.

Kelly, G. L., Stylianou, J., Rasaiyaah, J., Wei, W., Thomas, W., Croom-Carter, D., Kohler, C., Spang, R., Woodman, C., Kellam, P., Rickinson, A. B., & Bell, A. I. (2013). Different patterns of Epstein-Barr virus latency in endemic Burkitt lymphoma (BL) lead to distinct variants within the BL-associated gene expression signature. *J Virol*, *87*(5), 2882-2894. https://doi.org/10.1128/jvi.03003-12

Khine, H. T., Sato, Y., Hamada, M., Umeda, M., Iizuka, A., Son, S., Arai, H., Kojima, Y., Watanabe, T., & Naruse, A. (2025). Association of Epstein-Barr virus genomic alterations with human pathologies. *Blood*, *146*(13), 1533-1545.

Khleif, S. N., & Gupta, S. (2025). Cancer vaccines as enablers of immunotherapy. *Nature Immunology*, 1-13.

Kieser, A., & Sterz, K. R. (2015). The Latent Membrane Protein 1 (LMP1). In C. Münz (Ed.), *Epstein Barr virus volume 2: one herpes virus: many diseases* (pp. 119-149). Springer International Publishing. https://doi.org/10.1007/978-3-319-22834-1_4

Kikuchi, K., Ishige, T., Ide, F., Ito, Y., Saito, I., Hoshino, M., Inoue, H., Miyazaki, Y., Nozaki, T., & Kojima, M. (2015). Overexpression of Activation-Induced Cytidine Deaminase in MTX-and Age-Related Epstein-Barr Virus-Associated B-Cell Lymphoproliferative Disorders of the Head and Neck. *Journal of oncology*, *2015*(1), 605750.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: somatic

mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, *22*(3), 568-576.

Koelle, D. M., Norberg, P., Fitzgibbon, M. P., Russell, R. M., Greninger, A. L., Huang, M.-L., Stensland, L., Jing, L., Magaret, A. S., & Diem, K. (2017). Worldwide circulation of HSV-2× HSV-1 recombinant strains. *Scientific reports*, *7*(1), 44084.

Kojima, Y., Hamada, M., Naruse, A., Goto, K., Khine, H. T., Arai, H., Akutsu, Y., Satou, A., Nakaguro, M., Kato, S., Kodera, Y., Yatabe, Y., Torii, Y., Kawada, J.-i., Murata, T., Kimura, H., Takiguchi, S., Inagaki, H., Kataoka, H., & Okuno, Y. (2025). The landscape of 142 Epstein–Barr viral whole genomes in gastric cancer. *Journal of Gastroenterology*, *60*(1), 55-65. https://doi.org/10.1007/s00535-024-02170-3

Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0. 12. In.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2016). Fast gene set enrichment analysis. *Biorxiv*, 060012.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical association*, *47*(260), 583-621.

Kulwichit, W., Edwards, R. H., Davenport, E. M., Baskar, J. F., Godfrey, V., & Raab-Traub, N. (1998). Expression of the Epstein–Barr virus latent membrane protein 1 induces B cell lymphoma in transgenic mice. *Proceedings of the National Academy of Sciences*, *95*(20), 11963-11968.

Kumar, V., & Gabrilovich, D. I. (2014). Hypoxia-inducible factors in regulation of immune responses in tumour microenvironment. *Immunology*, *143*(4), 512-519.

Kuri, A., Jacobs, B. M., Vickaryous, N., Pakpoor, J., Middeldorp, J., Giovannoni, G., & Dobson, R. (2020). Epidemiology of Epstein-Barr virus infection and infectious mononucleosis in the United Kingdom. *BMC Public Health*, *20*(1), 912.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of statistical software*, *82*(13), 1 - 26. https://doi.org/10.18637/jss.v082.i13

Labrecque, L. G., Xue, S. a., Kazembe, P., Phillips, J., Lampert, I., Wedderburn, N., & Griffin, B. E. (1999). Expression of Epstein-Barr virus lytically related genes in African Burkitt's lymphoma: Correlation with patient response to therapy. *International journal of cancer*, *81*(1), 6-11.

Ladanyi, M., Offit, K., Jhanwar, S. C., Filippa, D. A., & Chaganti, R. S. K. (1991). MYC Rearrangement and Translocations Involving Band 8q24 in Diffuse Large Cell Lymphomas. *Blood*, *77*(5), 1057-1063. https://doi.org/https://doi.org/10.1182/blood.V77.5.1057.1057

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J. C., & Dry, J. R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research*, *44*(11), e108-e108.

Laidlaw, B. J., Craft, J. E., & Kaech, S. M. (2016). The multifaceted role of CD4(+) T cells in CD8(+) T cell memory. *Nat Rev Immunol*, *16*(2), 102-111. https://doi.org/10.1038/nri.2015.10

Lan, K., Kuppers, D. A., Verma, S. C., & Robertson, E. S. (2004). Kaposi's sarcoma-associated herpesvirus-encoded latency-associated nuclear antigen inhibits lytic replication by targeting Rta: a potential mechanism for virus-mediated control of latency. *J Virol*, *78*(12), 6585-6594. https://doi.org/10.1128/jvi.78.12.6585-6594.2004

Larbcharoensub, N., Mahaprom, K., Jiarpinitnun, C., Trachu, N., Tubthong, N., Pattaranutaporn, P., Sirachainan, E., & Ngamphaiboon, N. (2018). Characterization of PD-L1 and PD-1 Expression and CD8+ Tumor-infiltrating

Lymphocyte in Epstein-Barr Virus-associated Nasopharyngeal Carcinoma. *Am J Clin Oncol*, *41*(12), 1204-1210. https://doi.org/10.1097/coc.0000000000000449

Larocca, C., & Schlom, J. (2011). Viral vector-based therapeutic cancer vaccines. *Cancer J*, *17*(5), 359-371. https://doi.org/10.1097/PPO.0b013e3182325e63

Lassalle, F., Depledge, D. P., Reeves, M. B., Brown, A. C., Christiansen, M. T., Tutill, H. J., Williams, R. J., Einer-Jensen, K., Holdstock, J., & Atkinson, C. (2016). Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. *Virus evolution*, *2*(1), vew017.

Le, Q. T., Colevas, A. D., O'Sullivan, B., Lee, A. W., Lee, N., Ma, B., Siu, L. L., Waldron, J., Lim, C.-M., & Riaz, N. (2019). Current treatment landscape of nasopharyngeal carcinoma and potential trials evaluating the value of immunotherapy. *JNCI: Journal of the National Cancer Institute*, *111*(7), 655-663.

Lee, A. W., Sze, W., Au, J. S., Leung, S., Leung, T., Chua, D. T., Zee, B. C., Law, S. C., Teo, P. M., & Tung, S. Y. (2005). Treatment results for nasopharyngeal carcinoma in the modern era: the Hong Kong experience. *International Journal of Radiation Oncology* Biology* Physics*, *61*(4), 1107-1116.

Lee, E. S., Locker, J., Nalesnik, M., Reyes, J., Jaffe, R., Alashari, M., Nour, B., Tzakis, A., & Dickman, P. S. (1995). The association of Epstein–Barr virus with smooth-muscle tumors occurring after organ transplantation. *New England Journal of Medicine*, *332*(1), 19-25.

Lee, J. H., Choi, J., Ahn, Y. O., Kim, T. M., & Heo, D. S. (2018). CD21-independent Epstein-Barr virus entry into NK cells. *Cell Immunol*, *327*, 21-25. https://doi.org/10.1016/j.cellimm.2018.01.011

Lee, J. K., Sivakumar, S., Schrock, A. B., Madison, R., Fabrizio, D., Gjoerup, O., Ross, J. S., Frampton, G. M., Napalkov, P., & Montesion, M. (2022). Comprehensive

pan-cancer genomic landscape of KRAS altered cancers and real-world outcomes in solid tumors. *NPJ precision oncology*, *6*(1), 91.

Legason, I. D., Ogwang, M. D., Chamba, C., Mkwizu, E., El Mouden, C., Mwinula, H., Chirande, L., Schuh, A., & Chiwanga, F. (2022). 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), 1-9.

Lei, H., Li, T., Li, B., Tsai, S., Biggar, R. J., Nkrumah, F., Neequaye, J., Gutierrez, M., Epelman, S., Mbulaiteye, S. M., Bhatia, K., & Lo, S. C. (2015). Epstein-Barr virus from Burkitt Lymphoma biopsies from Africa and South America share novel LMP-1 promoter and gene variations. *Sci Rep*, *5*, 16706. https://doi.org/10.1038/srep16706

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, *556*(7702), 452-456.

Leoncini, L., Raphael, M., Stein, H., Harris, N., Jaffe, E., & Kluin, P. (2008). WHO classification of tumours of haematopoietic and lymphoid tissues. *World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissue. IARC, Lyon*, 262-264.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y. Y., Chung, G. T., Lui, V. W., To, K. F., Ma, B. B., Chow, C., Woo, J. K., Yip, K. Y., Seo, J., Hui, E. P., Mak, M. K., Rusan, M., Chau, N. G., Or, Y. Y., Law, M. H., Law, P. P., Liu, Z. W., Ngan, H. L., Hau, P. M., . . . Lo, K. W. (2017). Exome and genome sequencing of nasopharynx cancer identifies NF-κB pathway

activating mutations. *Nat Commun*, *8*, 14121. https://doi.org/10.1038/ncomms14121

Li, Z., Li, W., Song, L., & Zhu, W. (2012). Cilia, adenomatous polyposis coli and associated diseases. *Oncogene*, *31*(12), 1475-1483.

Li, Z., Munim, M. B., Sharygin, D. A., Bevis, B. J., & Vander Heiden, M. G. (2025). Understanding the Warburg Effect in Cancer. *Cold Spring Harb Perspect Med*, *15*(12). https://doi.org/10.1101/cshperspect.a041532

Liang, J., Zhou, H., Gerdt, C., Tan, M., Colson, T., Kaye, K. M., Kieff, E., & Zhao, B. (2016). Epstein–Barr virus super-enhancer eRNAs are essential for MYC oncogene expression and lymphoblast proliferation. *Proceedings of the National Academy of Sciences*, *113*(49), 14121-14126. https://doi.org/doi:10.1073/pnas.1616697113

Liao, H.-M., Liu, H., Chin, P.-J., Li, B., Hung, G.-C., Tsai, S., Otim, I., Legason, I. D., Ogwang, M. D., & Reynolds, S. J. (2022). Epstein-Barr Virus in Burkitt Lymphoma in Africa Reveals a Limited Set of Whole Genome and LMP-1 Sequence Patterns: Analysis of Archival Datasets and Field Samples From Uganda, Tanzania, and Kenya. *Frontiers in Oncology*, *12*.

Liao, H.-M., Liu, H., Lei, H., Li, B., Chin, P.-J., Tsai, S., Bhatia, K., Gutierrez, M., Epelman, S., & Biggar, R. J. (2018). Frequency of EBV LMP-1 promoter and coding variations in Burkitt lymphoma samples in Africa and South America and peripheral blood in Uganda. *Cancers*, *10*(6), 177.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, Jill P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6), 417-425. https://doi.org/https://doi.org/10.1016/j.cels.2015.12.004

Lin, C.-L., Lo, W.-F., Lee, T.-H., Ren, Y., Hwang, S.-L., Cheng, Y.-F., Chen, C.-L., Chang, Y.-S., Lee, S. P., & Rickinson, A. B. (2002). Immunization with Epstein-Barr Virus (EBV) peptide-pulsed dendritic cells induces functional CD8+ T-cell

immunity and may lead to tumor regression in patients with EBV-positive nasopharyngeal carcinoma. *Cancer research*, *62*(23), 6952-6958.

Lirussi, L., & Nilsen, H. L. (2023). DNA glycosylases define the outcome of endogenous base modifications. *International Journal of Molecular Sciences*, *24*(12), 10307.

Liu, M., Wang, R., & Xie, Z. (2023). T cell-mediated immunity during Epstein–Barr virus infections in children. *Infection, Genetics and Evolution*, *112*, 105443. https://doi.org/https://doi.org/10.1016/j.meegid.2023.105443

Liu, M., Zhang, Y., Jian, Y., Gu, L., Zhang, D., Zhou, H., Wang, Y., & Xu, Z. X. (2024). The regulations of telomerase reverse transcriptase (TERT) in cancer. *Cell Death Dis*, *15*(1), 90. https://doi.org/10.1038/s41419-024-06454-7

Liu, Y., He, S., Wang, X. L., Peng, W., Chen, Q. Y., Chi, D. M., Chen, J. R., Han, B. W., Lin, G. W., Li, Y. Q., Wang, Q. Y., Peng, R. J., Wei, P. P., Guo, X., Li, B., Xia, X., Mai, H. Q., Hu, X. D., Zhang, Z., . . . Bei, J. X. (2021). Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. *Nat Commun*, *12*(1), 741. https://doi.org/10.1038/s41467-021-21043-4

Lo, A. K.-F., Dawson, C. W., Lung, H. L., Wong, K.-L., & Young, L. S. (2021). The role of EBV-encoded LMP1 in the NPC tumor microenvironment: from function to therapy. *Frontiers in Oncology*, *11*, 640207.

Long, H. M., Meckiff, B. J., & Taylor, G. S. (2019). The T-cell Response to Epstein-Barr Virus-New Tricks From an Old Dog. *Front Immunol*, *10*, 2193. https://doi.org/10.3389/fimmu.2019.02193

López, C., Burkhardt, B., Chan, J. K. C., Leoncini, L., Mbulaiteye, S. M., Ogwang, M. D., Orem, J., Rochford, R., Roschewski, M., & Siebert, R. (2022). Burkitt lymphoma. *Nature Reviews Disease Primers*, *8*(1), 78. https://doi.org/10.1038/s41572-022-00404-3

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.

Lum, F. M., Sattarnezhad, N., Ho, P. P., Orr, N., Robinson, W. H., Lanz, T. V., & Steinman, L. (2026). Multiple molecular mimics in Epstein Barr Nuclear Antigen-1, and the pathogenesis of multiple sclerosis. *Proceedings of the National Academy of Sciences*, *123*(11), e2519445123.

Marioni, G., Mariuzzi, L., Gaio, E., Portaleone, S., Pertoldi, B., & Staffieri, A. (2002). Lymphoepithelial carcinoma of the larynx. *Acta oto-laryngologica*, *122*(4), 429-434.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution*, *1*(1). https://doi.org/10.1093/ve/vev003

Martínez-Lostao, L., Anel, A., & Pardo, J. (2015). How Do Cytotoxic Lymphocytes Kill Cancer Cells? *Clinical Cancer Research*, *21*(22), 5047-5056. https://doi.org/10.1158/1078-0432.Ccr-15-0685

Maruo, S., Zhao, B., Johannsen, E., Kieff, E., Zou, J., & Takada, K. (2011). Epstein-Barr virus nuclear antigens 3C and 3A maintain lymphoblastoid cell growth by repressing p16INK4A and p14ARF expression. *Proceedings of the National Academy of Sciences*, *108*(5), 1919-1924.

Mawalla, W. F., Achola, C., Nabalende, H., Otim, I., Legason, I. D., Ogwang, M. D., Aol, P. M., Sandi, G., Mwamtemi, H., & Mahawi, S. (2025). Rituximab for children with EBV-positive Burkitt lymphoma in East Africa. *Blood Advances*, *9*(10), 2393-2401.

Mbulaiteye, S. M. (2013). Burkitt Lymphoma: beyond discoveries. *Infect Agent Cancer*, *8*(1), 35.

Mbulaiteye, S. M., & Devesa, S. S. (2022). Burkitt lymphoma incidence in five continents. *Hemato*, *3*(3), 434-453.

McAuliffe, J. (2020). *Developing viral vectored vaccines for MAGE-expressing tumours* University of Oxford ]. Oxford.

McAuliffe, J., Chan, H. F., Noblecourt, L., Ramirez-Valdez, R. A., Pereira-Almeida, V., Zhou, Y., Pollock, E., Cappuccini, F., Redchenko, I., Hill, A. V., Leung, C. S. K., & Van den Eynde, B. J. (2021). Heterologous prime-boost vaccination targeting MAGE-type antigens promotes tumor T-cell infiltration and improves checkpoint blockade therapy. *J Immunother Cancer*, *9*(9). https://doi.org/10.1136/jitc-2021-003218

McClain, K. L., Leach, C. T., Jenson, H. B., Joshi, V. V., Pollock, B. H., Parmley, R. T., DiCarlo, F. J., Chadwick, E. G., & Murphy, S. B. (1995). Association of Epstein–Barr virus with leiomyosarcomas in young people with AIDS. *New England Journal of Medicine*, *332*(1), 12-18.

McGoldrick, S. M., Mutyaba, I., Adams, S. V., Larsen, A., Krantz, E. M., Namirembe, C., Mooka, P., Nabakooza, S., Ndagire, M., Mubiru, K., Nabwana, M., Nankinga, R., Gerdts, S., Gordon-Maclean, C., Geriga, F., Omoding, A., Sessle, E., Kambugu, J., Uldrick, T. S., . . . Casper, C. (2019). Survival of children with endemic Burkitt lymphoma in a prospective clinical care project in Uganda. *Pediatr Blood Cancer*, *66*(9), e27813. https://doi.org/10.1002/pbc.27813

McKenzie, J., & El-Guindy, A. (2015). Epstein-Barr virus lytic cycle reactivation. *Epstein Barr virus volume 2: one herpes virus: many diseases*, 237-261.

Meij, P., Leen, A., Rickinson, A. B., Verkoeijen, S., Vervoort, M. B. H. J., Bloemena, E., & Middeldorp, J. M. (2002). Identification and prevalence of CD8+ T-cell responses directed against Epstein-Barr virus-encoded latent membrane protein 1 and latent membrane protein 2. *International journal of cancer*, *99*(1), 93-99. https://doi.org/https://doi.org/10.1002/ijc.10309

Mestas, J., & Hughes, C. C. (2004). Of mice and not men: differences between mouse and human immunology. *The Journal of Immunology*, *172*(5), 2731-2738.

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088-1090.

Miliotis, C. N., & Slack, F. J. (2020). Multi-layered control of PD-L1 expression in Epstein-Barr virus-associated gastric cancer. *J Cancer Metastasis Treat*, *6*(13). https://doi.org/10.20517/2394-4722.2020.12

Minard-Colin, V., Auperin, A., Pillon, M., Burke, G. A., Barkauskas, D. A., Wheatley, K., Delgado, R. F., Alexander, S., Uyttebroeck, A., & Bollard, C. M. (2020). Rituximab for high-risk, mature B-cell non-Hodgkin's lymphoma in children. *New England Journal of Medicine*, *382*(23), 2207-2219.

Mishra, A., & Patel, T. N. (2025). Telomerase in cancer- ongoing quest and future discoveries. *Mol Biol Rep*, *52*(1), 161. https://doi.org/10.1007/s11033-025-10251-6

Molyneux, E., Schwalbe, E., Chagaluka, G., Banda, K., Israels, T., Depani, S., Mittermayer-Vassallo, K., Windebank, K., Mvula, J., & Njiram'madzi, J. (2017). The use of anthracyclines in the treatment of endemic Burkitt lymphoma. *British journal of haematology*, *177*(6), 984-990.

Molyneux, E. M., Rochford, R., Griffin, B., Newton, R., Jackson, G., Menon, G., Harrison, C. J., Israels, T., & Bailey, S. (2012). Burkitt's lymphoma. *Lancet*, *379*(9822), 1234-1244. https://doi.org/10.1016/s0140-6736(11)61177-x

Montes-Mojarro, I. A., Fend, F., & Quintanilla-Martinez, L. (2021). EBV and the Pathogenesis of NK/T Cell Lymphoma. *Cancers (Basel)*, *13*(6). https://doi.org/10.3390/cancers13061414

Mundo, L., Ambrosio, M. R., Picciolini, M., Lo Bello, G., Gazaneo, S., Del Porro, L., Lazzi, S., Navari, M., Onyango, N., & Granai, M. (2017). Unveiling another missing piece in EBV-driven lymphomagenesis: EBV-encoded microRNAs expression in EBER-negative Burkitt lymphoma cases. *Frontiers in Microbiology*, *8*, 229.

Munir, A., Khan, S., Saleem, A., Nusrat, H., Khan, S. A., Sayyed, H., Khalid, A., Javed, B., & Hidayat, F. (2025). The Role of Epstein–Barr Virus Molecular Mimicry in Various Autoimmune Diseases. *Scandinavian journal of immunology*, *101*(4), e70016.

Münz, C. (2019). Latency and lytic replication in Epstein–Barr virus-associated oncogenesis. *Nature Reviews Microbiology*, *17*(11), 691-700.

Murata, T., Sugimoto, A., Inagaki, T., Yanagi, Y., Watanabe, T., Sato, Y., & Kimura, H. (2021). Molecular basis of epstein–barr virus latency establishment and lytic reactivation. *Viruses*, *13*(12), 2344.

Murata, T., & Tsurumi, T. (2014). Switching of EBV cycles between latent and lytic states. *Reviews in medical virology*, *24*(3), 142-153.

Murphy, G., Pfeiffer, R., Camargo, M. C., & Rabkin, C. S. (2009). Meta-analysis shows that prevalence of Epstein–Barr virus-positive gastric cancer differs based on sex and anatomic location. *Gastroenterology*, *137*(3), 824-833.

Murphy, K., & Weaver, C. (2016). *Janeway's immunobiology*. Garland science.

Nangia, C. S., Seery, T. E., Moini, K., Blanton, R., Adlard, K., Nourani, A., Spilman, P., Bhar, P., Zhang, H., Sender, L. S., Reddy, S. B., & Soon-Shiong, P. (2025). QUILT 3.076 phase 1 study of memory-like cytokine-enriched natural killer (M-CENK) cells plus N-803 in locally advanced or metastatic solid tumors. *Journal of Clinical Oncology*, *43*(16_suppl), TPS2680-TPS2680. https://doi.org/10.1200/JCO.2025.43.16_suppl.TPS2680

Nassour, J., & Karlseder, J. (2026). Telomere Crisis Shapes Cancer Evolution. *Cold Spring Harb Perspect Biol*, *18*(3). https://doi.org/10.1101/cshperspect.a041688

Nemazee, D. (2017). Mechanisms of central tolerance for B cells. *Nature Reviews Immunology*, *17*(5), 281-294.

Network, C. G. A. R. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202.

Ngan, H.-L., Wang, L., Lo, K.-W., & Lui, V. W. Y. (2018). Genomic landscapes of EBV-associated nasopharyngeal carcinoma vs. HPV-associated head and neck cancer. *Cancers*, *10*(7), 210.

Ngoma, T., Adde, M., Durosinmi, M., Githang'a, J., Aken'Ova, Y., Kaijage, J., Adeodou, O., Rajab, J., Brown, B. J., Leoncini, L., Naresh, K., Raphael, M., Hurwitz, N., Scanlan, P., Rohatiner, A., Venzon, D., & Magrath, I. (2012). Treatment of Burkitt lymphoma in equatorial Africa using a simple three-drug combination followed by a salvage regimen for patients with persistent or recurrent disease. *British journal of haematology*, *158*(6), 749-762. https://doi.org/https://doi.org/10.1111/j.1365-2141.2012.09236.x

Nickles, E., Dharmadhikari, B., Yating, L., Walsh, R. J., Koh, L. P., Poon, M., Tan, L. K., Wang, L. Z., Ang, Y., Asokumaran, Y., Chong, W. Q., Huang, Y., Loh, K. S., Tay, J., Soo, R., Koh, M., Ho, L. P., Chan, M., Niam, M., . . . Schwarz, H. (2022). Dendritic cell therapy with CD137L-DC-EBV-VAX in locally recurrent or metastatic nasopharyngeal carcinoma is safe and confers clinical benefit. *Cancer Immunol Immunother*, *71*(6), 1531-1543. https://doi.org/10.1007/s00262-021-03075-3

Niedobitek, G., Agathanggelou, A., Rowe, M., Jones, E., Jones, D., Turyaguma, P., Oryema, J., Wright, D., & Young, L. (1995). Heterogeneous expression of Epstein-Barr virus latent proteins in endemic Burkitt's lymphoma.

Nilsson, K., Klein, G., Henle, W., & Henle, G. (1971). The establishment of lymphoblastoid lines from adult and fetal human lymphoid tissue and its dependence on EBV. *International journal of cancer*, *8*(3), 443-450.

Ning, L., Ko, J. M., Yu, V. Z., Ng, H. Y., Chan, C. K., Tao, L., Lam, S. Y., Leong, M. M., Ngan, R. K., Kwong, D. L., Lee, A. W., Ng, W. T., Cheng, A., Tung, S., Lee, V. H., Lam, K. O., Kwan, C. K., Li, W. S., Yau, S., . . . Lung, M. L. (2020). Nasopharyngeal carcinoma MHC region deep sequencing identifies HLA and

novel non-HLA TRIM31 and TRIM39 loci. *Commun Biol*, *3*(1), 759. https://doi.org/10.1038/s42003-020-01487-y

Norberg, P., Depledge, D. P., Kundu, S., Atkinson, C., Brown, J., Haque, T., Hussaini, Y., MacMahon, E., Molyneaux, P., & Papaevangelou, V. (2015). Recombination of globally circulating varicella-zoster virus. *Journal of Virology*, *89*(14), 7133-7146.

Nowell, P. C. (1976). The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science*, *194*(4260), 23-28.

Nwanegbo, E., Vardas, E., Gao, W., Whittle, H., Sun, H., Rowe, D., Robbins, P. D., & Gambotto, A. (2004). Prevalence of neutralizing antibodies to adenoviral serotypes 5 and 35 in the adult populations of The Gambia, South Africa, and the United States. *Clinical and Vaccine Immunology*, *11*(2), 351-357.

Olweny, C., Atine, I., Kaddu-Mukasa, A., Katongole-Mbidde, E., Lwanga, S., Johansson, B., Onyango, J., Høst, H., Norin, T., & Willey, B. (1977). Cerebrospinal Irradiation of Burkitt's Lymphoma: Failure in preventing central nervous system relapse. *Acta radiologica: therapy, physics, biology*, *16*(3), 225-231.

Olweny, C. L., Katongole-Mbidde, E., Otim, D., Lwanga, S. K., Magrath, I. T., & Ziegler, J. L. (1980). Long-term experience with Burkitt's lymphoma in Uganda. *International journal of cancer*, *26*(3), 261-266.

Ostroverkhova, D., Przytycka, T. M., & Panchenko, A. R. (2023). Cancer driver mutations: predictions and reality. *Trends in molecular medicine*, *29*(7), 554-566.

Osuna, D., & de Álava, E. (2009). Molecular pathology of sarcomas. *Reviews on recent clinical trials*, *4*(1), 12-26.

Ott, G., Rosenwald, A., & Campo, E. (2013). Understanding MYC-driven aggressive B-cell lymphomas: pathogenesis and classification. *Blood*, *122*(24), 3884-3891. https://doi.org/10.1182/blood-2013-05-498329

Ozuah, N. W., Lubega, J., Allen, C. E., & El-Mallawany, N. K. (2020). Five decades of low intensity and low survival: adapting intensified regimens to cure pediatric Burkitt lymphoma in Africa. *Blood Advances*, *4*(16), 4007-4019.

Pail, O., Lin, M. J., Anagnostou, T., Brown, B. D., & Brody, J. D. (2025). Cancer vaccines and the future of immunotherapy. *The Lancet*, *406*(10499), 189-202. https://doi.org/https://doi.org/10.1016/S0140-6736(25)00553-7

Palser, A. L., Grayson, N. E., White, R. E., Corton, C., Correia, S., Ba Abdullah, M. M., Watson, S. J., Cotten, M., Arrand, J. R., Murray, P. G., Allday, M. J., Rickinson, A. B., Young, L. S., Farrell, P. J., & Kellam, P. (2015). Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. *Journal of Virology*, *89*(10), 5222-5237. https://doi.org/10.1128/jvi.03614-14

Pan, J., Zhang, Q., Zhou, J., Ma, D., Xiao, X., & Wang, D. W. (2009). Recombinant adeno-associated virus encoding Epstein-Barr virus latent membrane proteins fused with heat shock protein as a potential vaccine for nasopharyngeal carcinoma. *Molecular Cancer Therapeutics*, *8*(9), 2754-2761. https://doi.org/10.1158/1535-7163.Mct-08-1176

Papadopoulos, J. S., & Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, *23*(9), 1073-1079.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526-528.

Pathmanathan, R., Prasad, U., Sadler, R., Flynn, K., & Raab-Traub, N. (1995). Clonal proliferations of cells infected with Epstein–Barr virus in preinvasive lesions related to nasopharyngeal carcinoma. *New England Journal of Medicine*, *333*(11), 693-698.

Patiño-Galindo, J. Á., Filip, I., & Rabadan, R. (2021). Global patterns of recombination across human viruses. *Molecular biology and evolution*, *38*(6), 2520-2531.

Paul, S., & Lal, G. (2017). The molecular mechanism of natural killer cells function and its importance in cancer immunotherapy. *Frontiers in Immunology*, *8*, 1124.

Pedersen, B. S., & Quinlan, A. R. (2017). Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *The American Journal of Human Genetics*, *100*(3), 406-413.

Pedersen, T. L. (2019). Package 'patchwork'. *R package http://CRAN. R-project. org/package= patchwork. Cran*, *418*.

Peng, L., Gao, R., Han, L., Fu, T., & Bian, C. (2025). Roles of β-catenin protein in non-small cell lung cancer. *Oncology Letters*, *30*(6), 565.

Peprah, S., Ogwang, M. D., Kerchan, P., Reynolds, S. J., Tenge, C. N., Were, P. A., Kuremu, R. T., Wekesa, W. N., Sumba, P. O., Masalu, N., Kawira, E., Magatti, J., Kinyera, T., Otim, I., Legason, I. D., Nabalende, H., Dhudha, H., Ally, H., Genga, I. O., . . . Mbulaiteye, S. M. (2020). Risk factors for Burkitt lymphoma in East African children and minors: A case-control study in malaria-endemic regions in Uganda, Tanzania and Kenya. *International journal of cancer*, *146*(4), 953-969. https://doi.org/10.1002/ijc.32390

Pérez-González, A., Bévant, K., & Blanpain, C. (2023). Cancer cell plasticity during tumor progression, metastasis and response to therapy. *Nat Cancer*, *4*(8), 1063-1082. https://doi.org/10.1038/s43018-023-00595-y

Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F., & González-Candelas, F. (2015). Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, *30*, 296-307.

Philip, P. A., Azar, I., Xiu, J., Hall, M. J., Hendifar, A. E., Lou, E., Hwang, J. J., Gong, J., Feldman, R., & Ellis, M. (2022). Molecular characterization of KRAS wild-type tumors in patients with pancreatic adenocarcinoma. *Clinical Cancer Research*, *28*(12), 2704-2714.

Ponnusamy, R., Khatri, R., Correia, P. B., Wood, C. D., Mancini, E. J., Farrell, P. J., & West, M. J. (2019). Increased association between Epstein-Barr virus EBNA2 from type 2 strains and the transcriptional repressor BS69 restricts EBNA2 activity. *PLOS Pathogens*, *15*(7), e1007458.

Price, A. M., & Luftig, M. A. (2014). Dynamic Epstein–Barr virus gene expression on the path to B-cell transformation. *Advances in virus research*, *88*, 279-313.

Prior, I. A., Hood, F. E., & Hartley, J. L. (2020). The frequency of Ras mutations in cancer. *Cancer research*, *80*(14), 2969-2974.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes de novo assembler. *Current protocols in bioinformatics*, *70*(1), e102.

Pudney , V. A., Leese , A. M., Rickinson , A. B., & Hislop , A. D. (2005). CD8+ immunodominance among Epstein-Barr virus lytic cycle antigens directly reflects the efficiency of antigen presentation in lytically infected cells. *Journal of Experimental Medicine*, *201*(3), 349-360. https://doi.org/10.1084/jem.20041542

Quinn, L. L., Zuo, J., Abbott, R. J., Shannon-Lowe, C., Tierney, R. J., Hislop, A. D., & Rowe, M. (2014). Cooperation between Epstein-Barr virus immune evasion proteins spreads protection from CD8+ T cell recognition across all three phases of the lytic cycle. *PLOS Pathogens*, *10*(8), e1004322. https://doi.org/10.1371/journal.ppat.1004322

Ramirez-Valdez, R. A., Baharom, F., Khalilnezhad, A., Fussell, S. C., Hermans, D. J., Schrager, A. M., Tobin, K. K. S., Lynn, G. M., Khalilnezhad, S., Ginhoux, F., Van den Eynde, B. J., Leung, C. S. K., Ishizuka, A. S., & Seder, R. A. (2023). Intravenous heterologous prime-boost vaccination activates innate and adaptive immunity to promote tumor regression. *Cell Rep*, *42*(6), 112599. https://doi.org/10.1016/j.celrep.2023.112599

Ranieri, E., Herr, W., Gambotto, A., Olson, W., Rowe, D., Robbins, P. D., Kierstead, L. S., Watkins, S. C., Gesualdo, L., & Storkus, W. J. (1999). Dendritic cells transduced with an adenovirus vector encoding Epstein-Barr virus latent membrane protein 2B: a new modality for vaccination. *J Virol*, *73*(12), 10416-10425. https://doi.org/10.1128/jvi.73.12.10416-10425.1999

Redmond, L. S., Ogwang, M. D., Kerchan, P., Reynolds, S. J., Tenge, C. N., Were, P. A., Kuremu, R. T., Masalu, N., Kawira, E., & Otim, I. (2020). Endemic Burkitt lymphoma: a complication of asymptomatic malaria in sub-Saharan Africa based on published literature and primary data from Uganda, Tanzania, and Kenya. *Malaria journal*, *19*(1), 1-14.

Rella, S. A., Kulikova, Y. A., Minnegalieva, A. R., & Kondrashov, F. A. (2024). Complex vaccination strategies prevent the emergence of vaccine resistance. *Evolution*, *78*(10), 1722-1738. https://doi.org/10.1093/evolut/qpae106

Ressing, M. E., Horst, D., Griffin, B. D., Tellam, J., Zuo, J., Khanna, R., Rowe, M., & Wiertz, E. J. (2008). Epstein-Barr virus evasion of CD8+ and CD4+ T cell immunity via concerted actions of multiple gene products. Seminars in cancer biology,

Revell, L. J. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, *12*, e16505.

Reyes-Sandoval, A., Berthoud, T., Alder, N., Siani, L., Gilbert, S. C., Nicosia, A., Colloca, S., Cortese, R., & Hill, A. V. (2010). Prime-boost immunization with adenoviral and modified vaccinia virus Ankara vectors enhances the durability and polyfunctionality of protective malaria CD8+ T-cell responses. *Infect Immun*, *78*(1), 145-153. https://doi.org/10.1128/iai.00740-09

Ribrag, V., Koscielny, S., Bosq, J., Leguay, T., Casasnovas, O., Fornecker, L.-M., Recher, C., Ghesquieres, H., Morschhauser, F., Girault, S., Gouill, S. L., Ojeda-Uribe, M., Mariette, C., Cornillon, J., Cartron, G., Verge, V., Chassagne-

Clément, C., Dombret, H., Coiffier, B., . . . Salles, G. (2016). Rituximab and dose-dense chemotherapy for adults with Burkitt's lymphoma: a randomised, controlled, open-label, phase 3 trial. *The Lancet*, *387*(10036), 2402-2411. https://doi.org/10.1016/S0140-6736(15)01317-3

Rickinson, A. B., Young, L. S., & Rowe, M. (1987). Influence of the Epstein-Barr virus nuclear antigen EBNA 2 on the growth phenotype of virus-transformed B cells. *J Virol*, *61*(5), 1310-1317. https://doi.org/10.1128/jvi.61.5.1310-1317.1987

Riggi, N., & Stamenkovic, I. (2007). The biology of Ewing sarcoma. *Cancer letters*, *254*(1), 1-10.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, *43*(7), e47-e47.

Robbiani, D. F., Bothmer, A., Callen, E., Reina-San-Martin, B., Dorsett, Y., Difilippantonio, S., Bolland, D. J., Chen, H. T., Corcoran, A. E., Nussenzweig, A., & Nussenzweig, M. C. (2008). AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. *Cell*, *135*(6), 1028-1038. https://doi.org/10.1016/j.cell.2008.09.062

Robbiani, D. F., Bunting, S., Feldhahn, N., Bothmer, A., Camps, J., Deroubaix, S., McBride, K. M., Klein, I. A., Stone, G., & Eisenreich, T. R. (2009). AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Molecular cell*, *36*(4), 631-641.

Robbiani, D. F., Deroubaix, S., Feldhahn, N., Oliveira, T. Y., Callen, E., Wang, Q., Jankovic, M., Silva, I. T., Rommel, P. C., & Bosque, D. (2015). Plasmodium infection promotes genomic instability and AID-dependent B cell lymphoma. *Cell*, *162*(4), 727-737.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, *53*(1-2), 131-147.

Romero, E. V., & Feder, A. F. (2024). Elevated HIV viral load is associated with higher recombination rate in vivo. *Molecular biology and evolution*, *41*(1), msad260.

Romero-Masters, J. C., Huebner, S. M., Ohashi, M., Bristol, J. A., Benner, B. E., Barlow, E. A., Turk, G. L., Nelson, S. E., Baiu, D. C., Van Sciver, N., Ranheim, E. A., Gumperz, J., Sherer, N. M., Farrell, P. J., Johannsen, E. C., & Kenney, S. C. (2020). B cells infected with Type 2 Epstein-Barr virus (EBV) have increased NFATc1/NFATc2 activity and enhanced lytic gene expression in comparison to Type 1 EBV infection. *PLOS Pathogens*, *16*(2), e1008365. https://doi.org/10.1371/journal.ppat.1008365

Roschewski, M., Dunleavy, K., Abramson, J. S., Powell, B. L., Link, B. K., Patel, P., Bierman, P. J., Jagadeesh, D., Mitsuyasu, R. T., & Peace, D. (2020). Multicenter study of risk-adapted therapy with dose-adjusted EPOCH-R in adults with untreated Burkitt lymphoma. *Journal of Clinical Oncology*, *38*(22), 2519-2529.

Roschewski, M., Staudt, L. M., & Wilson, W. H. (2022). Burkitt's lymphoma. *New England Journal of Medicine*, *387*(12), 1111-1122.

Rowe, M., Glaunsinger, B., van Leeuwen, D., Zuo, J., Sweetman, D., Ganem, D., Middeldorp, J., Wiertz, E. J. H. J., & Ressing, M. E. (2007). Host shutoff during productive Epstein–Barr virus infection is mediated by BGLF5 and may contribute to immune evasion. *Proceedings of the National Academy of Sciences*, *104*(9), 3366-3371. https://doi.org/doi:10.1073/pnas.0611128104

Rozman, M., Korać, P., Jambrosic, K., & Židovec Lepej, S. (2022). Progress in Prophylactic and Therapeutic EBV Vaccine Development Based on Molecular Characteristics of EBV Target Antigens. *Pathogens*, *11*(8). https://doi.org/10.3390/pathogens11080864

Rühl, J., Citterio, C., Engelmann, C., Haigh, T., Dzionek, A., Dreyer, J., Khanna, R., Taylor, G. S., Wilson, J. B., & Leung, C. S. (2019). Heterologous prime-boost vaccination protects against EBV antigen–expressing lymphomas. *The Journal of clinical investigation*, *129*(5), 2071-2087.

Ryan, J. L., Jones, R. J., Kenney, S. C., Rivenbark, A. G., Tang, W., Knight, E. R. W., Coleman, W. B., & Gulley, M. L. (2010). Epstein-Barr virus-specific methylation of human genes in gastric cancer cells. *Infect Agent Cancer*, *5*(1), 27. https://doi.org/10.1186/1750-9378-5-27

Sadreddini, S., Jadidi-Niaragh, F., Younesi, V., Pourlak, T., Afkham, A., Shokri, F., & Yousefi, M. (2016). Evaluation of EBV transformation of human memory B-cells isolated by FACS and MACS techniques. *J Immunotoxicol*, *13*(4), 490-497. https://doi.org/10.3109/1547691x.2015.1132288

Salnikov, M., Prusinkiewicz, M. A., Lin, S., Ghasemi, F., Cecchini, M. J., & Mymryk, J. S. (2023). Tumor-infiltrating T cells in EBV-associated gastric carcinomas exhibit high levels of multiple markers of activation, effector gene expression, and exhaustion. *Viruses*, *15*(1), 176.

Samanta, M., & Takada, K. (2010). Modulation of innate immunity system by Epstein–Barr virus-encoded non-coding RNA and oncogenesis. *Cancer science*, *101*(1), 29-35.

Sample, J., Young, L., Martin, B., Chatman, T., Kieff, E., Rickinson, A., & Kieff, E. (1990). Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. *Journal of Virology*, *64*(9), 4084-4092. https://jvi.asm.org/content/64/9/4084

https://jvi.asm.org/content/64/9/4084.short

http://www.ncbi.nlm.nih.gov/pubmed/2166806

Saunders, J. E., Gilbride, C., Dowall, S., Morris, S., Ulaszewska, M., Spencer, A. J., Rayner, E., Graham, V. A., Kennedy, E., & Thomas, K. (2023). Adenoviral vectored vaccination protects against Crimean-Congo Haemorrhagic Fever disease in a lethal challenge model. *EBioMedicine*, *90*.

Sausen, D., Poirier, M., Spiers, L., & Smith, E. (2023). Mechanisms of T cell evasion by Epstein-Barr virus and implications for tumor survival. *Frontiers in immunology*, *14*, 1289313.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592-593.

Schmitt, C. A., Wang, B., & Demaria, M. (2022). Senescence and cancer - role and therapeutic opportunities. *Nat Rev Clin Oncol*, *19*(10), 619-636. https://doi.org/10.1038/s41571-022-00668-4

Schoenberger, S. P., Toes, R. E., Van der Voort, E. I., Offringa, R., & Melief, C. J. (1998). T-cell help for cytotoxic T lymphocytes is mediated by CD40–CD40L interactions. *Nature*, *393*(6684), 480-483.

Sergushichev, A. A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *Biorxiv*, *60012*, 1-9.

Shannon-Lowe, C., & Rickinson, A. (2019). The global landscape of EBV-associated tumors. *Frontiers in Oncology*, *9*, 713.

Shannon-Lowe, C., Rickinson, A. B., & Bell, A. I. (2017). Epstein-Barr virus-associated lymphomas. *Philos Trans R Soc Lond B Biol Sci*, *372*(1732). https://doi.org/10.1098/rstb.2016.0271

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591-611.

Shi, M., Jevremovic, D. (2020). *T cell large granular lymphocytic leukemia*. PathologyOutlines.com. Retrieved October 3rd, 2025. from https://www.pathologyoutlines.com/topic/lymphomanonbtlgl.html.

Shilova, O., Tsyba, D., & Shilov, E. (2022). Mutagenic activity of AID/APOBEC deaminases in antiviral defense and carcinogenesis. *Molecular biology*, *56*(1), 46-58.

Siddiqui, N. (2025). Dplyr and TidyR. In *Cultural Analytics in R: A Tidy Approach* (pp. 57-95). Springer.

Simbiri, K. O., Smith, N. A., Otieno, R., Wohlford, E. E., Daud, I. I., Odada, S. P., Middleton, F., & Rochford, R. (2015). Epstein–Barr Virus Genetic Variation in Lymphoblastoid Cell Lines Derived from Kenyan Pediatric Population. *PLoS One*, *10*(5), e0125420.

Šimičić, P., Batović, M., Stojanović Marković, A., & Židovec-Lepej, S. (2024). Deciphering the Role of Epstein-Barr Virus Latent Membrane Protein 1 in Immune Modulation: A Multifaced Signalling Perspective. *Viruses*, *16*(4). https://doi.org/10.3390/v16040564

Sinkala, M. (2023). Mutational landscape of cancer-driver genes across human cancers. *Scientific reports*, *13*(1), 12742.

Skalska, L., White, R. E., Franz, M., Ruhmann, M., & Allday, M. J. (2010). Epigenetic repression of p16INK4A by latent Epstein-Barr virus requires the interaction of EBNA3A and EBNA3C with CtBP. *PLOS Pathogens*, *6*(6), e1000951.

Skalska, L., White, R. E., Parker, G. A., Sinclair, A. J., Paschos, K., & Allday, M. J. (2013). Induction of p16INK4a is the major barrier to proliferation when Epstein-Barr virus (EBV) transforms primary B cells into lymphoblastoid cell lines. *PLOS Pathogens*, *9*(2), e1003187.

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477-485. https://doi.org/10.1038/nrg2361

Slowikowski, K., Schep, A., Hughes, S., Lukauskas, S., Irisson, J.-O., Kamvar, Z. N., Ryan, T., Christophe, D., Hiroaki, Y., & Gramme, P. (2018). Package ggrepel. *Automatically position non-overlapping text labels with 'ggplot2*.

Smatti, M. K., Al-Sadeq, D. W., Ali, N. H., Pintus, G., Abou-Saleh, H., & Nasrallah, G. K. (2018). Epstein-Barr Virus Epidemiology, Serology, and Genetic Variability of LMP-1 Oncogene Among Healthy Population: An Update. *Front Oncol*, *8*, 211. https://doi.org/10.3389/fonc.2018.00211

Smatti, M. K., Al-Sadeq, D. W., Ali, N. H., Pintus, G., Abou-Saleh, H., & Nasrallah, G. K. (2018). Epstein–Barr virus epidemiology, serology, and genetic variability of LMP-1 oncogene among healthy population: an update. *Frontiers in Oncology*, *8*, 211.

Smit, A. (2015). Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. In.

Smith, L., McWhorter, A., Shellam, G., & Redwood, A. (2013). The genome of murine cytomegalovirus is shaped by purifying selection and extensive recombination. *Virology*, *435*(2), 258-268.

Soldan, S. S., & Lieberman, P. M. (2023). Epstein–Barr virus and multiple sclerosis. *Nature Reviews Microbiology*, *21*(1), 51-64.

Soneson, C., Love, M. I., & Robinson, M. D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521.

Spranger, S. (2016). Mechanisms of tumor escape in the context of the T-cell-inflamed and the non-T-cell-inflamed tumor microenvironment. *International immunology*, *28*(8), 383.

Staege, M. S., Lee, S. P., Frisan, T., Mautner, J., Scholz, S., Pajic, A., Rickinson, A. B., Masucci, M. G., Polack, A., & Bornkamm, G. W. (2002). MYC overexpression imposes a nonimmunogenic phenotype on Epstein–Barr virus-infected B cells. *Proceedings of the National Academy of Sciences*, *99*(7), 4550-4555.

Stanland, L. J., & Luftig, M. A. (2018). Molecular features and translational outlook for Epstein-Barr virus-associated gastric cancer. *Future Virol*, *13*(11), 803-818. https://doi.org/10.2217/fvl-2018-0071

Stavnezer, J., Guikema, J. E., & Schrader, C. E. (2008). Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.*, *26*, 261-292.

Steven, N. M., Annels, N. E., Kumar, A., Leese, A. M., Kurilla, M. G., & Rickinson, A. B. (1997). Immediate Early and Early Lytic Cycle Proteins Are Frequent Targets of the Epstein-Barr Virus–induced Cytotoxic T Cell Response. *Journal of Experimental Medicine*, *185*(9), 1605-1618. https://doi.org/10.1084/jem.185.9.1605

Strockbine, L. D., Cohen, J. I., Farrah, T., Lyman, S. D., Wagener, F., DuBose, R. F., Armitage, R. J., & Spriggs, M. K. (1998). The Epstein-Barr virus BARF1 gene encodes a novel, soluble colony-stimulating factor-1 receptor. *J Virol*, *72*(5), 4015-4021. https://doi.org/10.1128/jvi.72.5.4015-4021.1998

Subbalakshmi, A. R., Ramisetty, S., Mohanty, A., Pareek, S., Do, D., Shrestha, S., Khan, A., Talwar, N., Tan, T., Vishnubhotla, P., Singhal, S. S., Salgia, R., & Kulkarni, P. (2024). Phenotypic Plasticity and Cancer: A System Biology Perspective. *J Clin Med*, *13*(15). https://doi.org/10.3390/jcm13154302

Summerauer, A. M., Jäggi, V., Ogwang, R., Traxel, S., Colombo, L., Amundsen, E., Eyer, T., Subramanian, B., Fehr, J., Mantel, P. Y., Idro, R., & Bürgler, S. (2022). Epstein-Barr virus and malaria upregulate AID and APOBEC3 enzymes, but only AID seems to play a major mutagenic role in Burkitt lymphoma. *Eur J Immunol*, *52*(8), 1273-1284. https://doi.org/10.1002/eji.202249820

Suster, D., Mackinnon, A. C. (2023). *Lymphoepithelial carcinoma*. PathologyOutlines.com. Retrieved October 4th, 2025 from **https://www.pathologyoutlines.com/topic/mediastinumlel.html**

Swadling, L., Capone, S., Antrobus, R. D., Brown, A., Richardson, R., Newell, E. W., Halliday, J., Kelly, C., Bowen, D., Fergusson, J., Kurioka, A., Ammendola, V., Del Sorbo, M., Grazioli, F., Esposito, M. L., Siani, L., Traboni, C., Hill, A., Colloca, S., . . . Barnes, E. (2014). A human vaccine strategy based on chimpanzee adenoviral and MVA vectors that primes, boosts, and sustains functional HCV-specific T cell memory. *Sci Transl Med*, *6*(261), 261ra153. https://doi.org/10.1126/scitranslmed.3009185

Swerdlow, S. H., Campo, E., Harris, N. L., Jaffe, E. S., Pileri, S. A., Stein, H., Thiele, J., & Vardiman, J. W. (2008). *WHO classification of tumours of haematopoietic and lymphoid tissues* (Vol. 2). International agency for research on cancer Lyon, France.

Tabiasco, J., Vercellone, A., Meggetto, F., Hudrisier, D., Brousset, P., & Fournié, J. J. (2003). Acquisition of viral receptor by NK cells through immunological synapse. *J Immunol*, *170*(12), 5993-5998. https://doi.org/10.4049/jimmunol.170.12.5993

Tagawa, T., Albanese, M., Bouvet, M., Moosmann, A., Mautner, J., Heissmeyer, V., Zielinski, C., Lutter, D., Hoser, J., Hastreiter, M., Hayes, M., Sugden, B., & Hammerschmidt, W. (2016). Epstein-Barr viral miRNAs inhibit antiviral CD4+ T cell responses targeting IL-12 and peptide processing. *J Exp Med*, *213*(10), 2065-2080. https://doi.org/10.1084/jem.20160248

Tang, M., Zeng, Y., Poisson, A., Marti, D., Guan, L., Zheng, Y., Deng, H., Liao, J., Guo, X., Sun, S., Nelson, G., de Thé, G., Winkler, C. A., O'Brien, S. J., Carrington, M., & Gao, X. (2010). Haplotype-dependent HLA susceptibility to nasopharyngeal carcinoma in a Southern Chinese population. *Genes Immun*, *11*(4), 334-342. https://doi.org/10.1038/gene.2009.109

Tang, S., Ning, Q., Yang, L., Mo, Z., & Tang, S. (2020). Mechanisms of immune escape in the cancer immune cycle. *International Immunopharmacology*, *86*, 106700. https://doi.org/https://doi.org/10.1016/j.intimp.2020.106700

Tao, Q., & Chan, A. T. (2007). Nasopharyngeal carcinoma: molecular pathogenesis and therapeutic developments. *Expert Rev Mol Med*, *9*(12), 1-24. https://doi.org/10.1017/s1462399407000312

Tarke, A., Ramezani-Rad, P., Alves Pereira Neto, T., Lee, Y., Silva-Moraes, V., Goodwin, B., Bloom, N., Siddiqui, L., Avalos, L., Frazier, A., Zhang, Z., da Silva Antunes, R., Dan, J., Crotty, S., Grifoni, A., & Sette, A. (2024). SARS-CoV-2 breakthrough infections enhance T cell response magnitude, breadth, and epitope repertoire. *Cell Reports Medicine*, *5*(6). https://doi.org/10.1016/j.xcrm.2024.101583

Tatsis, N., Fitzgerald, J. C., Reyes-Sandoval, A., Harris-McCoy, K. C., Hensley, S. E., Zhou, D., Lin, S. W., Bian, A., Xiang, Z. Q., Iparraguirre, A., Lopez-Camacho, C., Wherry, E. J., & Ertl, H. C. (2007). Adenoviral vectors persist in vivo and maintain activated CD8+ T cells: implications for their use as vaccines. *Blood*, *110*(6), 1916-1923. https://doi.org/10.1182/blood-2007-02-062117

Taylor, B. S., Barretina, J., Maki, R. G., Antonescu, C. R., Singer, S., & Ladanyi, M. (2011). Advances in sarcoma genomics and new therapeutic targets. *Nature Reviews Cancer*, *11*(8), 541-557.

Taylor, G., Haigh, T., Gudgeon, N., Phelps, R., Lee, S., Steven, N., & Rickinson, A. (2004). Dual stimulation of Epstein-Barr Virus (EBV)-specific CD4+-and CD8+-T-cell responses by a chimeric antigen construct: potential therapeutic vaccine for EBV-positive nasopharyngeal carcinoma. *Journal of Virology*, *78*(2), 768-778.

Taylor, G. S., Jia, H., Harrington, K., Lee, L. W., Turner, J., Ladell, K., Price, D. A., Tanday, M., Matthews, J., Roberts, C., Edwards, C., McGuigan, L., Hartley, A.,

Wilson, S., Hui, E. P., Chan, A. T. C., Rickinson, A. B., & Steven, N. M. (2014). A recombinant modified vaccinia ankara vaccine encoding Epstein-Barr Virus (EBV) target antigens: a phase I trial in UK patients with EBV-positive cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *20*(19), 5009-5022. https://doi.org/10.1158/1078-0432.ccr-14-1122-t

Team, R. C. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *http://www. R-project. org/*.

Team, T. P. D. (2020). pandas-dev/pandas: Pandas. *Zenodo, February*.

Telford, M., Hughes, D. A., Juan, D., Stoneking, M., Navarro, A., & Santpere, G. (2020). Expanding the geographic characterisation of Epstein–Barr virus variation through gene-based approaches. *Microorganisms*, *8*(11), 1686.

Temple, R. M., Zhu, J., Budgeon, L., Christensen, N. D., Meyers, C., & Sample, C. E. (2014). Efficient replication of Epstein–Barr virus in stratified epithelium in vitro. *Proceedings of the National Academy of Sciences*, *111*(46), 16544-16549.

Togkousidis, A., Kozlov, O. M., Haag, J., Höhler, D., & Stamatakis, A. (2023). Adaptive RAxML-NG: accelerating phylogenetic inference under maximum likelihood using dataset difficulty. *Molecular biology and evolution*, *40*(10), msad227.

Torgbor, C., Awuah, P., Deitsch, K., Kalantari, P., Duca, K. A., & Thorley-Lawson, D. A. (2014). A Multifactorial Role for P. falciparum Malaria in Endemic Burkitt's Lymphoma Pathogenesis. *PLOS Pathogens*, *10*(5), e1004170. https://doi.org/10.1371/journal.ppat.1004170

Tsao, S. W., Tsang, C. M., Pang, P. S., Zhang, G., Chen, H., & Lo, K. W. (2012). The biology of EBV infection in human epithelial cells. *Semin Cancer Biol*, *22*(2), 137-143. https://doi.org/10.1016/j.semcancer.2012.02.004

Tscherne, A., Meyer zu Natrup, C., Kalodimou, G., & Volz, A. (2024). Generating MVA-vector vaccine candidates and testing them in animal models. In *Vaccinia, Mpox, and Other Poxviruses: Methods and Protocols* (pp. 297-340). Springer.

Tse, K.-P., Su, W.-H., Chang, K.-P., Tsang, N.-M., Yu, C.-J., Tang, P., See, L.-C., Hsueh, C., Yang, M.-L., & Hao, S.-P. (2009). Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21. 3. *The American Journal of Human Genetics*, *85*(2), 194-203.

Tugizov, S. M., Herrera, R., & Palefsky, J. M. (2013). Epstein-Barr virus transcytosis through polarized oral epithelial cells. *Journal of Virology*, *87*(14), 8179-8194.

Tzellos, S., Correia, P. B., Karstegl, C. E., Cancian, L., Cano-Flanagan, J., McClellan, M. J., West, M. J., & Farrell, P. J. (2014). A single amino acid in EBNA-2 determines superior B lymphoblastoid cell line growth maintenance by Epstein-Barr virus type 1 EBNA-2. *Journal of Virology*, *88*(16), 8743-8753.

Tzellos, S., & Farrell, P. J. (2012). Epstein-Barr virus sequence variation—biology and disease. *Pathogens*, *1*(2), 156-175.

Uyttenhove, C., Pilotte, L., Théate, I., Stroobant, V., Colau, D., Parmentier, N., Boon, T., & Van den Eynde, B. J. (2003). Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2, 3-dioxygenase. *Nature medicine*, *9*(10), 1269-1274.

van der Gracht, E. T. I., Schoonderwoerd, M. J. A., van Duikeren, S., Yilmaz, A. N., Behr, F. M., Colston, J. M., Lee, L. N., Yagita, H., van Gisbergen, K. P. J. M., Hawinkels, L. J. A. C., Koning, F., Klenerman, P., & Arens, R. (2020). Adenoviral vaccines promote protective tissue-resident memory T cell populations against cancer. *Journal for ImmunoTherapy of Cancer*, *8*(2), e001133. https://doi.org/10.1136/jitc-2020-001133

van Erp, E. A., van Kampen, M. R., van Kasteren, P. B., & de Wit, J. (2019). Viral Infection of Human Natural Killer Cells. *Viruses*, *11*(3), 243. https://www.mdpi.com/1999-4915/11/3/243

van Montfoort, N., van der Aa, E., & Woltman, A. M. (2014). Understanding MHC class I presentation of viral antigens by human dendritic cells as a basis for rational design of therapeutic vaccines. *Front Immunol*, *5*, 182. https://doi.org/10.3389/fimmu.2014.00182

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS),

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G., van Eck, H. J., & van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics*, *130*(1), 123-135.

Wakae, K., Kondo, S., Pham, H. T., Wakisaka, N., Que, L., Li, Y., Zheng, X., Fukano, K., Kitamura, K., Watashi, K., Aizaki, H., Ueno, T., Moriyama-Kita, M., Ishikawa, K., Nakanishi, Y., Endo, K., Muramatsu, M., & Yoshizaki, T. (2020). EBV-LMP1 induces APOBEC3s and mitochondrial DNA hypermutation in nasopharyngeal cancer. *Cancer Med*, *9*(20), 7663-7671. https://doi.org/10.1002/cam4.3357

Wan, J. C., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R., & Rosenfeld, N. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer*, *17*(4), 223-238.

Wang, A., Welch, R., Zhao, B., Ta, T., Keleş, S., & Johannsen, E. (2016). Epstein-Barr virus nuclear antigen 3 (EBNA3) proteins regulate EBNA2 binding to distinct RBPJ genomic sites. *Journal of Virology*, *90*(6), 2906-2919.

Wang, B., Han, J., Elisseeff, J. H., & Demaria, M. (2024). The senescence-associated secretory phenotype and its physiological and pathological implications. *Nat Rev Mol Cell Biol*, *25*(12), 958-978. https://doi.org/10.1038/s41580-024-00727-x

Wang, C., & Hao, P. (2025). Therapeutic vaccine strategies in cancer: advances, challenges, and future directions. *Clinical Cancer Bulletin*, *4*(1), 10.

Wang, C., Li, D., Zhang, L., Jiang, S., Liang, J., Narita, Y., Hou, I., Zhong, Q., Zheng, Z., & Xiao, H. (2019). RNA sequencing analyses of gene expression during Epstein-Barr virus infection of primary B lymphocytes. *Journal of Virology*, *93*(13), 10.1128/jvi. 00226-00219.

Wang, E., Worschech, A., & Marincola, F. M. (2008). The immunologic constant of rejection. *Trends in Immunology*, *29*(6), 256-262. https://doi.org/10.1016/j.it.2008.03.002

Wang, H.-B., Zhang, H., Zhang, J.-P., Li, Y., Zhao, B., Feng, G.-K., Du, Y., Xiong, D., Zhong, Q., & Liu, W.-L. (2015). Neuropilin 1 is an entry factor that promotes EBV infection of nasopharyngeal epithelial cells. *Nature communications*, *6*(1), 6240.

Wang, J., Nagy, N., & Masucci, M. G. (2020). The Epstein-Barr virus nuclear antigen-1 upregulates the cellular antioxidant defense to enable B-cell growth transformation and immortalization. *Oncogene*, *39*(3), 603-616. https://doi.org/10.1038/s41388-019-1003-3

Wang, X., Niu, Y., Hu, Y., Zhu, C., Yang, X., Shi, H., Yan, Y., Zhou, P., Ding, L., Zhang, M., Wu, M., Tang, S., Xing, M., & Zhou, D. (2026). Enhancing the breadth of protection in mice with a multivalent influenza vaccine. *Communications Biology*, *9*(1), 173. https://doi.org/10.1038/s42003-025-09451-4

Wang, Y., Guo, Z., Shu, Y., Zhou, H., Wang, H., & Zhang, W. (2017). BART miRNAs: an unimaginable force in the development of nasopharyngeal carcinoma. *European Journal of Cancer Prevention*, *26*(2), 144-150.

Wang, Y., Rozen, V., Zhao, Y., & Wang, Z. (2025). Oncogenic activation of PIK3CA in cancers: Emerging targeted therapies in precision oncology. *Genes & Diseases*, *12*(2), 101430.

Wang, Z., Kennedy, P. G., Dupree, C., Wang, M., Lee, C., Pointon, T., Langford, T. D., Graner, M. W., & Yu, X. (2021). Antibodies from multiple sclerosis brain identified Epstein-Barr virus nuclear antigen 1 & 2 epitopes which are recognized by oligoclonal bands. *Journal of Neuroimmune Pharmacology*, *16*(3), 567-580.

Ward, B. J. H., Schaal, D. L., Nkadi, E. H., & Scott, R. S. (2022). EBV Association with Lymphomas and Carcinomas in the Oral Compartment. *Viruses*, *14*(12). https://doi.org/10.3390/v14122700

Wegner, F., Lassalle, F., Depledge, D. P., Balloux, F., & Breuer, J. (2019). Coevolution of Sites under Immune Selection Shapes Epstein–Barr Virus Population Structure. *Molecular Biology and Evolution*, *36*(11), 2512-2521. https://doi.org/10.1093/molbev/msz152

Wei, M. X., Moulin, J.-C., Decaussin, G., Berger, F., & Ooka, T. (1994). Expression and tumorigenicity of the Epstein-Barr virus BARF1 gene in human Louckes B-lymphocyte cell line. *Cancer Research*, *54*(7), 1843-1848.

Weinberg, R. A. (1995). The retinoblastoma protein and cell cycle control. *Cell*, *81*(3), 323-330.

WHO. (2024, 27 November 2024). *MVA-BN (Modified Vaccinia Ankara – Bavarian Nordic) smallpox and mpox vaccine: interim guidance*. Retrieved 13 March 2026 from **https://www.who.int/publications/i/item/B09202**

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of open source software*, *4*(43), 1686.

Wild, F. (2007). An LSA package for R. Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07),

Wilkinson, D. E., & Weller, S. K. (2004). Recruitment of cellular recombination and repair proteins to sites of herpes simplex virus type 1 DNA replication is dependent on the composition of viral proteins within prereplicative sites and correlates with the induction of the DNA damage response. *Journal of Virology*, *78*(9), 4783-4796.

Williams, H., McAulay, K., Macsween, K. F., Gallacher, N. J., Higgins, C. D., Harrison, N., Swerdlow, A. J., & Crawford, D. H. (2005). The immune response to primary EBV infection: a role for natural killer cells. *British journal of haematology*, *129*(2), 266-274.

Winter, J. R., Taylor, G. S., Thomas, O. G., Jackson, C., Lewis, J. E. A., & Stagg, H. R. (2019). Predictors of Epstein-Barr virus serostatus in young people in England. *BMC Infect Dis*, *19*(1), 1007. https://doi.org/10.1186/s12879-019-4578-y

Wong, K. C. W., Hui, E. P., Lo, K. W., Lam, W. K. J., Johnson, D., Li, L., Tao, Q., Chan, K. C. A., To, K. F., King, A. D., Ma, B. B. Y., & Chan, A. T. C. (2021). Nasopharyngeal carcinoma: an evolving paradigm. *Nat Rev Clin Oncol*, *18*(11), 679-695. https://doi.org/10.1038/s41571-021-00524-x

Wong, T. K., Ly-Trong, N., Ren, H., Baños, H., Roger, A. J., Susko, E., Bielow, C., De Maio, N., Goldman, N., & Hahn, M. W. (2025). IQ-TREE 3: phylogenomic inference software using complex evolutionary models.

Wong, Y., Meehan, M. T., Burrows, S. R., Doolan, D. L., & Miles, J. J. (2022). Estimating the global burden of Epstein-Barr virus-related cancers. *J Cancer Res Clin Oncol*, *148*(1), 31-46. https://doi.org/10.1007/s00432-021-03824-y

Wu, C. C., Liu, M. T., Chang, Y. T., Fang, C. Y., Chou, S. P., Liao, H. W., Kuo, K. L., Hsu, S. L., Chen, Y. R., Wang, P. W., Chen, Y. L., Chuang, H. Y., Lee, C. H., Chen, M., Wayne Chang, W. S., & Chen, J. Y. (2010). Epstein-Barr virus DNase

(BGLF5) induces genomic instability in human epithelial cells. *Nucleic Acids Res*, *38*(6), 1932-1949. https://doi.org/10.1093/nar/gkp1169

Xiao, K., Yu, Z., Li, X., Li, X., Tang, K., Tu, C., Qi, P., Liao, Q., Chen, P., & Zeng, Z. (2016). Genome-wide analysis of Epstein-Barr virus (EBV) integration and strain in C666-1 and Raji cells. *Journal of Cancer*, *7*(2), 214.

Xiao, L., Hu, Z. y., Dong, X., Tan, Z., Li, W., Tang, M., Chen, L., Yang, L., Tao, Y., Jiang, Y., Li, J., Yi, B., Li, B., Fan, S., You, S., Deng, X., Hu, F., Feng, L., Bode, A. M., . . . Cao, Y. (2014). Targeting Epstein–Barr virus oncoprotein LMP1-mediated glycolysis sensitizes nasopharyngeal carcinoma to radiation therapy. *Oncogene*, *33*(37), 4568-4578. https://doi.org/10.1038/onc.2014.32

Xu, M., Yao, Y., Chen, H., Zhang, S., Cao, S.-M., Zhang, Z., Luo, B., Liu, Z., Li, Z., & Xiang, T. (2019). Genome sequencing analysis identifies Epstein–Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. *Nature Genetics*, *51*(7), 1131-1136.

Xu, M., Zhang, W.-L., Zhu, Q., Zhang, S., Yao, Y.-y., Xiang, T., Feng, Q.-S., Zhang, Z., Peng, R.-J., & Jia, W.-H. (2019). Genome-wide profiling of Epstein-Barr virus integration by targeted sequencing in Epstein-Barr virus associated malignancies. *Theranostics*, *9*(4), 1115.

Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L., Tang, W., Feng, T., Chen, M., & Zhan, L. (2021). ggtreeExtra: compact visualization of richly annotated phylogenetic data. *Molecular biology and evolution*, *38*(9), 4039-4042.

Xue, S. a., Labrecque, L. G., Lu, Q. L., Ong, S. K., Lampert, I. A., Kazembe, P., Molyneux, E., Broadhead, R. L., Borgstein, E., & Griffin, B. E. (2002). Promiscuous expression of Epstein-Barr virus genes in Burkitt's lymphoma from the central African country Malawi. *International journal of cancer*, *99*(5), 635-643.

Xue, W.-Q., Wang, T.-M., Huang, J.-W., Zhang, J.-B., He, Y.-Q., Wu, Z.-Y., Liao, Y., Yuan, L.-L., Mu, J., & Jia, W.-H. (2021). A comprehensive analysis of genetic diversity of EBV reveals potential high-risk subtypes associated with nasopharyngeal carcinoma in China. *Virus evolution*, *7*(1), veab010-veab010. https://doi.org/10.1093/ve/veab010

Yaacov, A., Ben Cohen, G., Landau, J., Hope, T., Simon, I., & Rosenberg, S. (2024). Cancer mutational signatures identification in clinical assays using neural embedding-based representations. *Cell Rep Med*, *5*(6), 101608. https://doi.org/10.1016/j.xcrm.2024.101608

Yan, C. T., Boboila, C., Souza, E. K., Franco, S., Hickernell, T. R., Murphy, M., Gumaste, S., Geyer, M., Zarrin, A. A., & Manis, J. P. (2007). IgH class switching and translocations use a robust non-classical end-joining pathway. *Nature*, *449*(7161), 478-482.

Yao, Q., Rickinson, A., & Epstein, M. (1985). A re-examination of the Epstein-Barr virus carrier state in healthy seropositive individuals. *International journal of cancer*, *35*(1), 35-42.

Yao, Q., Tierney, R., Croom-Carter, D., Cooper, G., Ellis, C., Rowe, M., & Rickinson, A. (1996). Isolation of intertypic recombinants of Epstein-Barr virus from T-cell-immunocompromised individuals. *Journal of virology*, *70*(8), 4895-4903.

Yin, C. C., & Luthra, R. (2013). Molecular detection of t(14;18)(q32;q21) in follicular lymphoma. *Methods Mol Biol*, *999*, 203-209. https://doi.org/10.1007/978-1-62703-357-2_13

You, R., Liu, Y. P., Lin, D. C., Li, Q., Yu, T., Zou, X., Lin, M., Zhang, X. L., He, G. P., Yang, Q., Zhang, Y. N., Xie, Y. L., Jiang, R., Wu, C. Y., Zhang, C., Cui, C., Wang, J. Q., Wang, Y., Zhuang, A. H., . . . Chen, M. Y. (2019). Clonal Mutations Activate the NF-κB Pathway to Promote Recurrence of

Nasopharyngeal Carcinoma. *Cancer Res*, *79*(23), 5930-5943. https://doi.org/10.1158/0008-5472.Can-18-3845

Young, L. S., Arrand, J. R., & Murray, P. G. (2007). EBV gene expression and regulation. In A. Arvin, G. Campadelli-Fiume, E. Mocarski, P. S. Moore, B. Roizman, R. Whitley, & K. Yamanishi (Eds.), *Human herpesviruses: biology, therapy, and immunoprophylaxis*. Cambridge University Press. http://www.ncbi.nlm.nih.gov/books/NBK47431/

Young, L. S., & Dawson, C. W. (2014). Epstein-Barr virus and nasopharyngeal carcinoma. *Chinese journal of cancer*, *33*(12), 581.

Young, L. S., Dawson, C. W., & Eliopoulos, A. G. (2000). The expression and function of Epstein-Barr virus encoded latent genes. *Molecular Pathology*, *53*(5), 238-247. https://doi.org/10.1136/mp.53.5.238

Young, L. S., & Rickinson, A. B. (2004). Epstein–Barr virus: 40 years on. *Nature Reviews Cancer*, *4*(10), 757-768.

Young, L. S., Yap, L. F., & Murray, P. G. (2016). Epstein–Barr virus: more than 50 years old and still providing surprises. *Nature Reviews Cancer*, *16*(12), 789-802. https://doi.org/10.1038/nrc.2016.92

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in ecology and evolution*, *8*(1), 28-36.

Yu, H., Yin, X., Mao, Y., Chen, M., Tang, Q., & Yan, S. (2022). The global burden of nasopharyngeal carcinoma from 2009 to 2019: an observational study based on the Global Burden of Disease Study 2019. *European Archives of Oto-Rhino-Laryngology*, *279*(3), 1519-1533. https://doi.org/10.1007/s00405-021-06922-2

Yun, S.-m., Kim, Y. S., & Hur, D. Y. (2019). LMP1 and 2A induce the expression of Nrf2 through Akt signaling pathway in epstein-barr virus–Transformed B Cells. *Translational oncology*, *12*(5), 775-783.

Zanella, L., Riquelme, I., Buchegger, K., Abanto, M., Ili, C., & Brebi, P. (2019). A reliable Epstein-Barr Virus classification based on phylogenomic and population analyses. *Scientific reports*, *9*(1), 9829.

Zhang, J., Sommermann, T., Li, X., Gieselmann, L., de la Rosa, K., Stecklum, M., Klein, F., Kocks, C., & Rajewsky, K. (2023). LMP1 and EBNA2 constitute a minimal set of EBV genes for transformation of human B cells [Original Research]. *Frontiers in Immunology*, *Volume 14 - 2023*. https://doi.org/10.3389/fimmu.2023.1331730

Zhao, W., Zhang, P., Bai, S., Lv, M., Wang, J., Chen, W., Yu, Q., & Wu, J. (2021). Heterologous prime-boost regimens with HAdV-5 and NDV vectors elicit stronger immune responses to Ebola virus than homologous regimens in mice. *Arch Virol*, *166*(12), 3333-3341. https://doi.org/10.1007/s00705-021-05234-4

Zheng, H., Dai, W., Cheung, A. K. L., Ko, J. M. Y., Kan, R., Wong, B. W. Y., Leong, M. M. L., Deng, M., Kwok, T. C. T., & Chan, J. Y.-W. (2016). Whole-exome sequencing identifies multiple loss-of-function mutations of NF-κB pathway regulators in nasopharyngeal carcinoma. *Proceedings of the National Academy of Sciences*, *113*(40), 11283-11288.

Zhou, W., Zheng, D., Liu, Q., Lu, D., Liu, Y., Lin, P., Huang, Y., Peng, X., Bao, J. J., & Cai, C. (2025). ABACUS: An electronic structure analysis package for the AI era. *The Journal of Chemical Physics*, *163*(19).

Ziegler, J., Miner, R., Rosenbaum, E., Lennette, E., Shillitoe, E., Casavant, C., Drew, W. L., Mintz, L., Gershow, J., & Greenspan, J. (1982). Outbreak of Burkitt's-like lymphoma in homosexual men. *The Lancet*, *320*(8299), 631-633.

Ziegler, J. L., Morrow Jr, R. H., Fass, L., Kyalwazi, S. K., & Carbone, P. P. (1970). Treatment of Burkitt's tumor with cyclophosphamide. *Cancer*, *26*(2), 474-484.

Zuo, J., Currin, A., Griffin, B. D., Shannon-Lowe, C., Thomas, W. A., Ressing, M. E., Wiertz, E. J., & Rowe, M. (2009). The Epstein-Barr virus G-protein-coupled receptor contributes to immune evasion by targeting MHC class I molecules for degradation. *PLOS Pathogens*, *5*(1), e1000255.