

1 Heterogeneity of the Epstein-Barr virus major internal repeat reveals evolutionary mechanisms
2 of EBV and a functional defect in the prototype EBV strain B95-8.

3
4 Mohammed Ba abdullah¹, Richard Palermo¹, Anne Palser^{2,a}, Nicholas E Grayson², Paul
5 Kellam^{1,2,3}, Samantha Correia¹, Agnieszka Szymula^{1b}, Rob White^{1#}

6
7 ¹ Section of Virology, Imperial College Faculty of Medicine, St Mary's Hospital, London, UK.

8 ² Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.

9 ³ Kymab, Babraham Research Campus, Cambridge UK

10

11 # Address correspondence to Rob White, robert.e.white@imperial.ac.uk

12 ^a Present address: Kymab, Babraham Research Campus, Cambridge UK.

13 ^b Present address: Division of Infectious Diseases, Department of Medicine, Brigham and
14 Women's Hospital, Harvard Medical School, Boston, MA, USA.

15

16 Running Title: Heterogeneity of EBV internal repeat 1

17

18 Word Count (Abstract): 254

19 Word Count (Importance): 147

20 Word Count (Manuscript): 9484

21

22

Abstract:

Epstein-Barr virus (EBV) is a ubiquitous pathogen of humans that can cause several types of lymphoma and carcinoma. Like other herpesviruses, EBV has diversified both through co-evolution with its host, and genetic exchange between virus strains. Sequence analysis of the EBV genome is unusually challenging, because of the large number and length of repeat regions within the virus. Here we describe the sequence assembly and analysis of the large internal repeat of EBV (IR1 or BamW repeats) from over 70 strains.

Diversity of the latency protein EBNA-LP resides predominantly within the exons downstream of IR1. The integrity of the putative BWRF1 ORF is retained in over 80% of strains, and deletions truncating IR1 always spare BWRF1. Conserved regions include the IR1 latency promoter (Wp), and one zone upstream of and two within BWRF1.

IR1 is heterogeneous in 70% of strains, and this heterogeneity arises from sequence exchange between strains as well as spontaneous mutation, with inter-strain recombination more common in tumour-derived viruses. This genetic exchange often incorporates regions of <1kb, and allelic gene conversion changes the frequency of small regions within the repeat, but not close to the flanks. These observations suggest that IR1 – and by extension EBV – diversifies through both recombination and breakpoint repair, while concerted evolution of IR1 is driven by gene conversion of small regions. Finally, the prototype EBV strain B95-8 contains four non-consensus variants within a single IR1 repeat unit, including a STOP codon in EBNA-LP.

Repairing IR1 improves EBNA-LP levels and the quality of transformation by the B95-8 BAC.

Importance

Epstein-Barr virus (EBV) infects the majority of the world population, but only causes illness in a small minority. Nevertheless, over 1% of cancers worldwide are attributable to EBV. Recent sequencing projects investigating virus diversity, to see if different strains have different disease impacts, have excluded regions of repeating sequence, as they are more technically challenging. Here we analyse the sequence of the largest repeat in EBV (IR1). We first characterised the variations in protein sequences encoded across IR1. In studying variations within the repeat of each strain, we identified a mutation in the main laboratory strain of EBV that impairs virus function, and suggest that tumour-associated viruses may be more likely to contain DNA mixed from two strains. Patterns of this mixing suggest that sequences can spread between strains (and also within the repeat) by copying sequence from another strain (or repeat unit) to repair DNA damage.

Introduction

Epstein-Barr virus is a human herpesvirus that infects the vast majority of the human population. Usually, such infection is asymptomatic, but EBV can cause a number of malignancies, including both lymphomas – Burkitt lymphoma (BL), Hodgkin lymphoma (HL) and immunoblastic lymphomas such as post-transplant lymphoproliferative disease (PTLD) or diffuse large B cell lymphoma (DLBCL) – and carcinomas such as nasopharyngeal carcinoma (NPC) and gastric cancer (GC). While EBV-associated disease is relatively uncommon in the immunocompetent, the near-ubiquitous nature of EBV infection makes it a considerable worldwide health burden (1).

Many of these cancers are geographically restricted: EBV-associated BL is prevalent predominantly in malarial regions of Africa and Asia/Australasia; NPC incidence is dramatically higher in south China and nearby regions of southeast Asia, and unevenly distributed elsewhere around the world (2); the prevalence of EBV-GC (compared to total GC) is higher in the Americas (3). While the global asymmetry in BL incidence is attributed to malarial co-infection, the reasons for other geographic variations in the incidence of EBV-associated diseases remain unclear. They may relate to environmental differences, behavioural or cultural factors, genetics of the population, differences between virus strains, or to some interaction between these possible causes. Sequences of EBV genes/genomes vary geographically around the world (4), likely reflecting the close host-pathogen relationship exemplified by the co-speciation typical of herpesviruses (5). Two major subgroups of EBV – type 1 and type 2 – have been defined according to their EBNA2 and EBNA3 gene sequences. Biologically, type 2 viruses transform B cells less efficiently (6), and have recently been shown to also transform T cells (7). However, the diversity of the main viral oncogenes (EBNA2, EBNA3s and LMP1) have not yet revealed

any clear associations with disease beyond supporting this geographic diversity (8). In recent years, viral genome sequencing has become a sufficiently cost effective proposition to allow these ideas to be revisited.

When published in 1984, the sequence of the B95-8 strain of EBV was the longest contiguous DNA sequence known (9). This was updated to provide a more representative type 1 EBV genome (a B95-8/Raji hybrid) and type 2 (AG876) virus strains (10, 11). However, the advent of short read sequencing technology has resulted in the more widespread sequencing of various herpesvirus genomes, notably of CMV (12, 13), herpes simplex viruses (HSV) 1 and 2 (14-17), varicella zoster virus (VZV) (18) as well as EBV (4, 19-22). The consensus approach for assembling herpesvirus genomes comprises de novo contig assembly, followed by gap-filling approaches and genome assembly driven by known consensus genome structures. This assembly can then be annotated as a framework to address biological questions that arise from the genome sequence. This approach is exemplified by the VirGA protocol (applied to HSV1) (15), but similar approaches have been followed for CMV (13) and EBV (4).

One of the major challenges for genome sequencing - particularly using short read libraries - is the accurate assembly of repetitive regions. Many viruses contain repetitive regions, particularly at their termini. Of the human herpesviruses, EBV contains perhaps the most repeat regions, yet sequencing the repeats is both troublesome and important, as many of these regions are replication origins, or encode proteins (or parts of proteins) that play a major role in virus biology, particularly viral latency and persistence (Figure 1A). Accurately assembling these regions remains the largest barrier to producing complete EBV genomes: the recent publication of 71 virus genomes blanked out over 20 repeat regions to facilitate comparisons between the strains (4). Current viral genomes have often used Sanger sequencing to bridge these gaps, and

105 more recently long read technology (PacBio) was used to sequence across the EBV repeats in
106 two BAC-cloned viruses (23). However, even these methods struggle to resolve many of the
107 EBV repeats, due to their large size and complexity.

108 In order to develop a methodology for more systematically extracting the sequence of
109 repetitive regions from short read data, we have chosen to resolve and analyse internal repeat 1
110 (IR1, also known as the BamW repeats), the largest of EBV's repeat regions. IR1 is composed of
111 typically 5.6 to 8.6 copies of a 3kb repeat unit (24). Upstream of the repeat lies the major latency
112 promoter, Cp and the first two exons of the extensively spliced EBNA transcript (exons C1 and
113 C2). Each repeat unit hosts a pair of exons (W1 and W2) after the immediate early latency
114 promoter Wp, which initiates transcription at exon W0. Exons W1 and W2 encode the repeat
115 domain of EBNA-LP, whose non-repetitive C terminus is encoded by the Y exons immediately
116 downstream of IR1. A complex and incompletely characterised pattern of alternative splicing
117 and alternative promoter usage across IR1 and beyond dictates the transcription of the six EBV
118 nuclear antigen (EBNA) genes during latency (25), as well as the anti-apoptotic protein BHRF1
119 during the lytic cycle (26). In addition to these transcribed exons, the B95-8 IR1 contains an
120 extended open reading frame (BWRF1) that lacks a start codon and promoter (9, 27, 28). At the
121 5' end of BWRF1 is an extended hairpin approximately 500 bases long (27). As yet there is no
122 direct evidence that either BWRF1 or the structured region are biologically significant. It is
123 known that some component(s) of IR1 are essential for the transformation of cells by EBV, as
124 viruses with 1.6 or fewer repeats are completely unable to transform B cells, while more than 4.6
125 repeats are required for maximal transformation efficiency (24). This deficiency is attributed to
126 the reduced transcription of EBNA-LP and EBNA2 caused by having fewer Wp promoters.

127 Here we aimed to develop a strategy for both assembling and analysing EBV repeat
128 regions. As well as identifying some idiosyncrasies of individual viruses, including a defect in
129 the main lab strain B95-8, this analysis of the sequence characteristics and diversity of IR1
130 defined conserved regions, and characterised the diversity of EBNA-LP and BWRFL sequences.
131 In addition, IR1 is often heterogeneous within each strain, and this heterogeneity probably arises
132 both through spontaneous mutation and sequence exchange between viruses. By treating the
133 repeats as an internally controlled sequence, we can infer that inter-strain genetic exchange may
134 be more prevalent in tumour-derived strains, and speculate that genetic exchange in IR1, both
135 within and between viruses, can arise through the repair of DNA damage.
136

137

138 **Materials and Methods**

139

140 **DNA sequence and sample classification.**

141 The sequence data analysed herein is described in detail in a previous report (4) and sequence
142 data are available (accession number ERP001026) at the European Nucleotide archive
143 (<http://www.ebi.ac.uk/ena>). Briefly the EBV sequences come from a collection of cell lines and
144 primary tumours from around the world, although predominantly from Europe, Australia, China
145 and East Africa. For the purposes of the analyses we describe, all spontaneous LCLs (sLCLs) are
146 defined as being ‘non-tumour derived’, while all long term cell lines are defined as tumour
147 derived. Since some of the sLCLs were from patients with EBV malignancies, this is necessarily
148 an approximation, but there is no evidence that the virus strain in the patient’s circulation (the
149 one used to establish the sLCL) is the same as the one causing the malignancy. This information
150 is published previously, but is included in Supplementary table ST1 for convenience. For the
151 purposes of analyses requiring unrelated or distinct strains, however, we have combined certain
152 strains because they have common origins: So, for diversity analyses, Jijoye and P3HR1-HH514-
153 C16, are combined into a single origin (Jijoye); B95-8, LCL B95-8 delEBER, X50-7 and BL36
154 (see text) are combined into a single B95-8 entry; TRL1-pre and TRL1-post have identical
155 sequences so are described as a single strain, TRL1 (see table ST1). For transparency all
156 sequences (including the very similar strains) are included in the multiple sequence alignments
157 and minor variant analyses. Additionally, we received Illumina HiSeq reads for the parental B95-
158 8 cell line from Alan Chiang (processed as described in (21)).

159

160 **Assembly of an IR1 template sequence.**

161 Reads were mapped to template sequences using Burrows-Wheeler Aligner (v0.7.9a), using the
162 BWA-MEM algorithm (29). The sequence of the IR1 region was determined by creating a
163 theoretical template, comprising 716 nucleotides upstream of IR1, 1.6 repeat units of IR1 and
164 456 nucleotides downstream. This was achieved using a python script to undertake the following
165 assembly: For each strain, reads were mapped to the first repeat unit of the B95-8 IR1 (position
166 12001-15072 of NC_007605.1), and the consensus of the repeat unit extracted using SAMtools
167 (v0.1.19) (30). This was then concatamerized to give one intact repeat, and a second partial
168 repeat whose end is defined by sequence at the end of the B95-8 IR1 (AGGCCCAGCCCCCTC).
169 Flanking sequences were then added from the published genomes (4). Finally, reads were
170 remapped to this template with BWA and the consensus extracted by SAM tools.

171

172 **Read depth analysis for identifying deleted regions.**

173 Since deletions can be spanned by low levels of contaminating reads after BWA
174 alignment, we devised a read depth analysis (RDA) to identify sudden changes in read depth that
175 might indicate a missed deletion. RDA calculates the ratio of read depths at adjacent nucleotides
176 for every position in the sequence, and this can be graphed across the sequence to visualize the
177 data. Positions where the $\log_2$ of this ratio exceeded ± 0.5 were flagged for analysis. RDA was
178 able to confirm known deletions in non-repetitive regions (such as the EBNA3B deletion in
179 TRL1 – not shown), but gave a weak signal in the repeats, and a strong signal at the junction
180 between IR1 and the non-repetitive part of the genome. Pile-up files were visualized at the
181 position of the RDA signal using Geneious (version 7.1.8) which allows the identification of
182 reads that span the sequence, and hence the identification of the real sequence at the deletion.

183

184 **Analyses of IR1 templated multiple sequence alignment.**

185 Single nucleotide polymorphism (SNP) densities (excluding indels) were counted from a
186 multiple sequence alignment of 85 strains (Supplementary data file SD2), using an in house
187 script. Phylogenetic analysis was conducted using Geneious, with a Jukes-Cantor genetic
188 distance model and neighbour-joining method. Identity plots, comparing a virus with a panel of
189 other viruses were generated using T-RECs (31). This approach is scalable to an unlimited
190 number of comparator strains, but random mutations can cause deviation from 100% similarity
191 that may complicate interpretation. Bootscan analysis (32) was performed using the manual
192 bootscan method in RDP4 (v4.82) (33). The four strains used as comparators in the bootscan
193 analysis were carefully selected from different branches of the phylogenetic tree to maximise the
194 discriminatory power of the bootscan. Use of different or additional strains substantially affected
195 the output, so this approach was usable for visualisation but not for discovery in our dataset.
196 Identity plots and bootscan analyses were generated using a scanning window size of 200 and
197 400 bp and step size of 20 and 30 respectively.

198 The BWRF1 ORF was calculated at its maximum possible size, allowing any non-STOP
199 codon to initiate the ORF. All ORFs exceeding 600 nucleotides (200 aa) were identified, and are
200 presented in table ST1). To define the likelihood of obtaining an ORF as long as that of BWRF1,
201 we used the following calculation: The distribution of bases in IR1 is 12.7% A, 44% C, 28.6% G
202 and 14.8% T. Based on this, the cumulative chance of the next codon after any position being a
203 stop codon is 1.31%. Therefore, the odds of an ORF starting at a specified position having a
204 minimum length n codons is 0.9869^n .

205 To estimate the size of IR1 (ie number of repeat units) the read depth across the W exons was
206 compared to unique regions of the template (Cp) and the genome (BALF5 gene) and corrected
207 according to southern blot-based estimates of IR1 size for a subset of samples (4). However, we
208 are not confident that this method is consistently reliable, as samples with low read depth often
209 gave higher IR1 repeat numbers (not shown).

210

211 **Identification of minor variants (MVs).**

212 Minor variants are bases calls in a sample that occur less frequently than the consensus
213 nucleotide, but are more prevalent than the background of technical error. This might typically
214 arise from either samples containing multiple virus strains, quasispecies groups or – in the case
215 of IR1 – from differences in sequence between repeat units within IR1. Reads from the
216 sequencing described in (4) were aligned to the consensus using BWA and the resultant pileup
217 files were generated from SAMtools version 0.1.19 (30), output with a minimum Phred quality
218 score of 20 to consider minor variants. Then, the mpileup2SNP algorithm in VarScan v2.3.7 (34)
219 was used to output all variants mapped on both the forward (+) and reverse (-) strands of the
220 reference templates using default parameters. We then filtered MVs to remove likely false
221 negatives. Positions consistently exhibiting MVs in all strains were assumed to be technical
222 errors. Such positions were identified and excluded, based on having a coefficient of variation of
223 less than one. Technical errors are also often more prevalent on one strand than the other, being
224 exacerbated by local template structure that differs between the strands. This read bias at each
225 MV was scored by calculating $\frac{(DR+/VR+)}{(DR-/VR-)}$ where DR(+) and DR(-) are numbers of consensus-
226 supporting reads on the plus and minus strands respectively, while VR(+) and VR(-) are numbers
227 of variant-supporting reads on the plus and minus strands. All MV data for all strains were

228 visualised by plotting read bias against the MV frequency using RStudio (v0.98.1103). This
229 initially highlighted a substantial number of MVs with very high frequency but strong read bias
230 that were identified as being close to the edges of IR1 – the read bias derived from the
231 predominant mapping of these variants to reads spanning the junction between IR1 and the
232 adjacent unique part of the genome (not shown). These were used to correct the consensus, and
233 are regarded as SNPs. Overall, a relatively high background level – up to around 7% – of
234 underlying MVs were seen that corresponded to inter-strain SNPs. This implied some degree of
235 sample cross contamination or barcode exchange during multiplexing or library generation.
236 Since genuine MVs have a minimum frequency of between 8% and 20% (since strains have
237 between 5 and 11 repeat units in IR1), we instituted a cut-off that escalated with increasing read
238 bias, according to the equation $\sqrt{Read_{bias}} \times Variant_{frequency} \geq 7$.

239

240 **Infection of PBLs with HB9 and EBV wild-type IR1.**

241 Peripheral blood leukocytes were isolated from blood on a ficoll gradient, washed in RPMI
242 supplemented with 1% FCS and stored in RPMI/20%FCS/10% DMSO. Frozen PBLs from four
243 donors were recovered and seeded at 2×10^6 cells per well in a 24 well plate and infected with
244 2×10^5 Raji infectious units of either HB9 or WT^w EBV (35). Medium was replaced on the
245 following day and twice weekly thereafter. Cells were kept growing in RPMI medium
246 supplemented with Penicillin/Streptomycin and L-Glutamine 10% foetal calf serum.
247 Cyclosporine A (100 ng/ml) was added in the medium for the first 2 weeks. Proliferation was
248 assessed by staining cells – either CD19-purified B cells, or peripheral blood leukocytes with
249 CellTrace Violet (Life Technologies), and analysed 8 days post infection by flow cytometry as
250 described elsewhere (35).

251

252 **Western blotting and cloning of B95-8 W repeat and LP.**

253 Lymphoblastoid cell lines were seeded at a defined density (3×10^5 cells per ml) and left for 24
254 hours before harvesting. To generate the EBNA-LP size ladder, three plasmids – pSG5-LP2W
255 (36) was gift from Andy Bell, while the C-terminally FLAG-tagged constructs pSG5LP with
256 four W repeats (37) and pRSP693 with 7 W exon pairs (38) were both provided by Paul Ling –
257 were transfected into 293 cells, and cells harvested after 48 hours. Cells were washed once in
258 PBS and the pellet resuspended in RIPA lysis buffer. Protein was quantitated using the DC
259 protein Assay (Bio-Rad) using bovine gamma-globulin as a protein standard. 30 µg of Protein
260 was loaded into a 0.75 mm thickness 12.5% polyacrylamide mini-gel, electrophoresed on a
261 Protean2 mini gel (Bio-Rad), and then electroblotted onto Amersham Protran nitrocellulose (GE
262 healthcare). Membrane was blocked with powdered milk in PBS/0.5% Tween20, and hybridised
263 with supernatant from the JF186 hybridoma cell line (1:100). Blot was washed and then re-
264 probed with anti-mouse antibody conjugated to Horseradish peroxidase. Blot was washed and
265 chemiluminescent signal produced using ECL reagents (GE healthcare), and developed on
266 Amersham Hyperfilm MP. The blot was stripped and re-probed with anti-gamma tubulin as a
267 loading control.

268 IR1 repeat units were cloned by digestion with BamHI into a pBR322-based plasmid.

269 Sequence was determined using fluorescent dideoxynucleotide DNA sequencing (CSC

270 Genomics, Hammersmith Hospital, London).

271

272

Results

Templated assembly of IR1 shows that deletion breakpoints cluster in IR1.

Our previous assembly of 71 EBV genomes produced a set of aligned genomes in which the repeat regions remain unresolved, replaced by blank sequence (4). This approach was taken because the number and complexity resulting from variations in repeat structure and length, combined with relatively short reads (76 nucleotide paired end reads) could not be reliably built by the available de novo assemblers, and the repeats disrupted genome alignment and subsequent ORF analysis. IR1 typically comprises 4-7 complete repeat units (Figure 1B), containing BWRF1 and Wp regions, and a downstream partial repeat – the BWRF1 part of the repeat unit (4, 24). To simplify analysis of this region for analysis, a template was generated (based on previous sequence and mapping of reads to a standardised IR1 repeat – see methods) comprising one complete and one partial repeat unit, flanked by the unique regions of the virus genome that exist either side of IR1 (Figure 1). The reads were mapped to this template to confirm the consensus sequence of IR1 for each strain, and these consensus sequences were compared in a multiple sequence alignment.

Overall we found the junctions between IR1 and the unique region of the genome to be structurally conserved, except in a few strains which exhibit large deletions of the flanks of IR1. For instance, the P3HR1 and Daudi strains have previously been reported to contain deletions of the downstream junction between IR1 and the unique sequence (39-41), while X50-7 has a deletion at the C promoter end of IR1 (42). We have found that these deletions are sometimes missed where there is a large read depth, and sufficient cross-contaminating reads to span the deletion. These erroneous reads may arise from sample cross contamination, or from barcode exchange and ‘jumping PCR’ during library preparation and sequencing, which can be avoided

296 by double barcoding (43). These hidden deletions are still characterized by a sudden change in
297 read depth at each end, so we devised a read depth analysis (which compares the number of reads
298 mapped to adjacent nucleotides across the template to identify sudden changes in read depth) to
299 identify the location of large deletions. The read depth analysis for X50-7 (figure 2A) illustrates
300 how this effectively identifies the junction of a mismatched deletion in unique regions of the
301 genome, and that this signal is much weaker in repetitive regions, and erroneous signals occur at
302 the junctions between unique and repeat sequences. This approach corrected a falsely identified
303 deletion in the original HL01 sequence, and identified novel deletions in L591 and at one end of
304 the Cheptages IR1, in addition to confirming the known deletions in P3HR1, Daudi and X50-7.
305 The deletion in the Cp end of IR1 in Cheptages occurred in two repeats (IR1 and the family of
306 repeats of oriP) so could only be resolved by visually scanning the sorted mapped reads rather
307 than by RDA, which loses signal in repeats. The deletion junctions are summarised in Table 1,
308 and shown schematically in figure 2B.

309 Curiously, all 6 of the breakpoints in the 5 strains with internal deletions lay within a
310 <1kb region upstream of the hairpin, and downstream of the W0 exon. There were no strains
311 with breakpoints in the hairpin, in BWRF1 or in Wp. The odds of all six independent breakpoints
312 clustering within this 1kb part of the repeat are around 1 in 700, and the odds of 6 breakpoints
313 clustering in any 1kb stretch can be approximated to one in 3^5 , or $p < 0.005$. Whether a biological
314 process influences this bias (such as different frequencies of breakpoint generation, or selection
315 for some function of the BWRF1 region that does not tolerate partial copies) is not clear.

316

317 **Conserved regions of IR1 include Wp, W2 and regions of BWRF1 of unknown function.**

318 After aligning the consensus sequences for IR1 of newly assembled and previously
319 published EBV strains (85 strains in total - Supplementary data file SD2), the degree of diversity
320 between strains was assessed. Variations in the length of the repeat unit are largely restricted to
321 two regions; NC_007605.1 position 13210-40 within BWRF1, (which has three distinct variant
322 length genotypes and two strains (Makau and sLCL1.19) with independent >10bp deletions), and
323 the region between exons W0 and W1 (14450-14500) which has two distinct single base indels.

324 Analysing the distribution of indels and single nucleotide polymorphisms (SNPs), shows
325 considerable variation in the degree of sequence conservation across the repeats (Figure 3A). The
326 polymorphism density ranges from 0 to 12 SNPs/100bp. The Wp promoter (upstream of and
327 including exon W0) is highly conserved, as is a region encompassing the W2 exon and the
328 adjacent intronic sequence that is processed into a stable RNA - EBV-sisRNA1 (44). Other
329 highly conserved regions were identified immediately upstream of the IR1 hairpin (position
330 12225-12350) and within the final third of the BWRF1 ORF (13290-13450), immediately
331 downstream of the indels in Makau and sLCL1.19 described above.

332

333 **The main annotated features of IR1 are fairly well conserved, but more diverse than**
334 **previously reported.**

335 The 512 nucleotide hairpin region (whose ends are defined by a repeated CGGGCACCCC
336 motif (27)) appears to be retained in all of the virus strains analysed. The SNP frequency across
337 this region is moderate, and its conservation was assessed by aligning the hairpin to its own
338 reverse complement (Supplementary data file SD3). There are no examples of polymorphisms on
339 one strand of the hairpin being matched by a complementing polymorphism on the other strand.
340 However, there are several SNPs that are predicted to change G-T into G-C base pairs on the

transcribed strand, and changes in mismatched bases that retain the mismatch. Overall, the hairpin is retained within IR1 of all strains with only slight variations, suggesting that it offers some sort of selective advantage for EBV. However, there is no indication of important structures based on co-conservation of bases.

A second major feature of IR1 is BWRF1, the large uncharacterised open reading frame with no canonical start codon that partly overlaps the hairpin in the second reportedly stable intron of IR1, EBV-sisRNA-2 (44). Since BWRF1 has no ATG to define a start codon, we defined the ORF length as the distance between in-frame stop codons within IR1. These ORFs fall into two main groups (models 1 and 2 - Figure 3B and supplementary table ST2), dependent on a TAA to TAC polymorphism at NC_007605.1 position 12540 defining two different lengths of ORF. The smaller group (model 2), with a shorter (384 codon) ORF includes B95-8 (model 2A) and the type 2 viruses (model 2B), while about 60% of strains contain a much longer (514 codon) ORF that also spans the hairpin (model 1A). However, 9 of the 78 distinct strains contain an indel causing a frameshift mutation that truncates BWRF1 (models 1X, 2X and 1D where X indicates the disrupted ORF). This suggests that, at least for these strains, the presence of an intact full-length BWRF1 ORF is not essential for the virus lifecycle. There are also other individual strains carrying indels that change the potential start and end positions of BWRF1, but all of these retain an ORF of at least 1000 nucleotides. The high GC content (72.6%) of IR1 means reduces the frequency of stop codons, meaning that extended ORFs are more likely to occur randomly than in a lower GC environment. Nevertheless, the chance of an ORF of over 380 codons (seen in 86% of strains) appearing in this position by chance is still less than 1%, so it seems likely that there is some selection pressure to retain the BWRF1 ORF in most strains, or to retain certain regions of the BWRF1 ORF in all strains.

364 The EBV nuclear antigen leader protein (EBNA-LP) is encoded across the W repeat
365 region, combining repeating exon pairs from IR1 (typically 2-7 copies of exons W1 and W2)
366 with a non-repetitive C terminal domain (encoded by exons Y1 and Y2). It is among the first
367 viral proteins translated after infection, during the establishment of latency. The diversity of the
368 EBNA-LP sequence is summarised in Figure 3C with a new nomenclature for EBNA-LP
369 subtypes proposed in table 2.

370 Previous analyses of EBNA-LP sequences identified two distinct protein isoforms of
371 EBNA-LP based on the presence of G8/T12 or V8/A12 respectively in exon W1 (45). These
372 have been called type 1 and type 2 variants respectively, even though the so-called type 2 variant
373 is found in many type 1 strains. Our analysis confirms that these are the two W1 exon
374 polymorphisms. Only one strain has a SNP in exon W2 - an African type 2 sLCL with a Q54R
375 change. This same change is found in the EBNA-LP of chimpanzee lymphocryptoviruses (36,
376 38), so may represent a circulating variant.

377 The Y exons of EBNA-LP exhibit more diversity than has previously been described
378 (Figure 3C; table 2). This analysis identified four major sequence subgroups of the Y exons
379 (defined mainly by exon Y2) that are herein designated A, B, C and Z. Subgroup A was
380 previously reported in Akata (36) and also contains a synonymous polymorphism in exon Y1.
381 Group B is found in the prototype B95-8 strain, and group Z in type 2 EBVs (11). The C subtype
382 is novel, characterised by V95E and V102I amino acid differences from B95-8, yet surprisingly
383 is the most common one in the samples, found in both African and European-origin strains,
384 including the Mutu cell line (19). Several of these four main subtypes have rare variants. Most
385 intriguingly, three NPC-derived samples (M81, C666.1 and GD2) share a common IRDH to
386 MRAS change at amino acids 86-89. The new C666 sequence conflicts with a previously

published partial C666 sequence, which reported MRDH at this position (46), while another C666 sequence (Accession AB828190) confirms the MRAS motif we observed at this position, although the genome region containing the Y exons is inverted in this sequence (47). Other less common Y exon subgroups are noted in figure 3C, and the designations for all strains are given in Supplementary table ST1. There also seem to be certain restrictions as to which exon combinations are compatible: most notably, the type 1 W1 exon seems to require the B subtype of Y exons, while Y exon subtypes A, C and Z require type 2 W exons. Whether these relationships reflect historic opportunities for recombination, or some sort of functional incompatibility remains to be assessed.

All of the variations described above were visualized against a phylogenetic tree of the IR1 template sequence (Figure 4). Some anomalies were apparent, such as the mainly type 2 strain Jijoye, which has B95-8-like Y exons and segregates between the type 1 and type 2 strains. If this combination were found in circulating (i.e. non disease-associated) virus types, this would suggest that EBNA-LP and EBNA2 genotypes do not exhibit sequence co-dependency. However, it remains possible that this mixture represents a disease-associated or lab-acquired mixture of two strains, rather than a genuine example of a circulating virus strain.

Based on the observed variations, we propose a nomenclature for EBNA-LP that indicates both the W1 polymorphism (1 or 2) and the Y exon subgroup (A, B, C or Z – see figure 3C, Table 2 and Supplementary table ST1). Subtypes of these main variants are then indicated by superscripted letters indicating a characteristic amino acid change. For instance, the EBNA-LP variant found in the NPC-derived C666 and M81 virus strains is type 2B^M, with M representing the MRAS motif in exon Y2.

409 From mapping both the BWRF1 and EBNA-LP variants onto the phylogenetic tree
410 (Figure 4), it is apparent that certain EBNA-LP and BWRF1 subtypes usually co-segregate in a
411 manner that defines the genetic variation of this region. While there is no linkage between the
412 EBNA-LP and BWRF1 isoforms that is always true, EBNA-LP subtypes 2A and 2C are usually
413 associated with the longer BWRF1 ORF, while subtype 2Z (type 2 EBV) is generally associated
414 with the shorter BWRF1 (model 2). In contrast, EBNA-LP subtype 1B can be found in
415 combination with all the commonest BWRF1 models (1A, 2A and 2X). Since no function has
416 been ascribed to the BWRF1 ORF – certainly not one that is linked to EBNA-LP function –
417 these correlations between BWRF1 and EBNA-LP types can currently only be considered as
418 reflecting wider sequence conservation/linkage, rather than implying a functional
419 interdependence.

420
421 **IR1 repeats are frequently not homogeneous, but variations at the flanks of IR1 are**
422 **protected from gene conversion.**

423 Attempting to sequence large repeats with short read data means that variations within
424 repeats are missed when using simple consensus alignment or contig assembly methods. Repeats
425 are often subject to concerted evolution, where they evolve together as a result of ectopic gene
426 conversion, a process by which the sequences are homogenised through DNA repair processes
427 templated by other repeat units (48, 49). When reads are mapped to a consensus sequence, reads
428 from repeat units with an alternative sequence will be mapped to the same position as the
429 consensus reads (illustrated in Figure 5A). Therefore, heterogeneity within repeats will be
430 observable as a non-consensus nucleotides at a frequency that is proportional to the percentage of
431 repeats containing that alternative nucleotide. These ‘minor variants’ (MVs) should be equally

432 prevalent on both forward and reverse reads. False positive MVs were eliminated by identifying
433 and excluding error-prone sites, ensuring MVs were represented on both strands, and setting an
434 identification threshold of over 7% (see Methods and Figure 5B), since any MV occurring in at
435 least one strain would have a frequency of 8-20%, based on the observation that IR1 repeat
436 number is typically 5.6 to 6.6, and not more than the ~11.6 copies found in B95-8 (4, 9, 50,
437 Supplementary table ST1).

438 In defining the MVs, several positions near the flanks of IR1 were identified that
439 exhibited high MV frequency (not shown), but also a high degree of read bias caused by read
440 pairs that spanned the junctions between IR1 and the adjacent unique regions of the genome.
441 After correcting these nucleotides in the template, it was apparent that these variants were found
442 only at the edge of IR1 close to the unique sequence, and usually not as MVs elsewhere in IR1.
443 There were 22 positions within 200 bp of either edge of IR1 that were different from the
444 corresponding internal position (Supplementary Table ST3). Two of these (one of which has
445 been reported previously (28)) are found only in the B95-8-related strains and BL36 (previously
446 described as an intertypic recombinant). By mapping the flank SNPs that are shared by more
447 than two independent strains onto the phylogenetic tree (Figure 4) it is clear that these generally
448 co-segregate with the phylogeny. Notably, the further these edge-associated variants are from the
449 ends of the repeat, the less tightly they were associated with the phylogeny. For instance, the
450 SNP in group g is 160 bp from the edge of IR1, and is found only in a subset of related strains. In
451 some cases the edge-associated variants were propagated to purity throughout IR1 (e.g. sLCL-
452 BL1.03). In contrast, there were no examples of groups some of whose members had reverted
453 SNPs near flanks to the normal base call. Thus these edge variants are evolutionarily stable and
454 not subject to gene conversion, but were able to seed gene conversion events elsewhere in the

455 repeats. These flank variants were therefore classified as SNPs rather than MVs, to prevent them
456 from skewing our analyses of repeat heterogeneity elsewhere within IR1.

457

458 **Heterogeneity in IR1 is more common in tumours and tumour-derived cell lines, and arises**
459 **though inter strain recombination as well as spontaneous mutation.**

460 Even after excluding the flank-associated SNPs, only 21 of the 70 distinct strains tested
461 were homogeneous (i.e. had no MVs in their repeats), and most were sLCLs. MV numbers were
462 then compared according to whether the virus was isolated directly from tumours (or long
463 established cell lines) or were isolated more recently from blood (i.e. sLCLs) or saliva. While 70-
464 80% of the samples had two or fewer heterogeneous positions per repeat unit in IR1, nine viruses
465 (7 of which were tumour-derived) had at least 5 (and up to 21) MVs per repeat unit (figure 6A).
466 To assess whether number of MVs was simply a function of the number of repeats in IR1, the
467 length of IR1 was estimated, based on the read depth in the W exons relative to Cp
468 (Supplementary table ST1). Strains with long IR1 repeats were no more likely to contain (or
469 lack) MVs. Conversely the strains with the most MVs often had a lower proportion of reads
470 mapping to IR1 (ie fewer IR1 repeat units).

471 To further characterise the origin of these variants, the MVs were stratified into those
472 arising from either spontaneous mutation, or from some form of inter-strain genetic exchange
473 that we will describe as ‘recombination-like’. This stratification is based on the assumption that
474 sporadic mutations will be randomly distributed through IR1, whereas MVs arising through
475 inter-strain recombination will also be observed in our data set as SNPs between strains. Even
476 though this approximation is limited – there are undoubtedly SNPs that have not yet been
477 identified, and there may be mutagenic processes within the cell that target certain sequences

478 more often – this analysis suggested that viruses with few MVs mostly exhibit sporadic
479 mutations, whereas those with many MVs appear to have arisen through genetic exchange
480 between strains (figures 6B and 6C). Cheptages exhibits the largest number of mutation-
481 associated MVs that are not accompanied by any recombination-like changes. Since this strain
482 also contains deletions at both ends of IR1, this is further evidence that Cheptages has
483 experienced a highly mutagenic environment. In contrast, strains that show substantial evidence
484 of inter-strain recombination tend to also have large numbers of apparent mutations (figure 6C
485 and Supplementary table ST4). This could mean that the circumstances that predispose to genetic
486 exchange also predispose to mutational events. However, it also remains possible that these
487 apparent mutations represent previously unseen SNPs.

488 Since a recombination-like signature was typically associated with mutations, the strains
489 were split into three classes: homogeneous, recombination-like, or mutation only. The
490 distribution of these classes within the ‘tumour’ and ‘circulating’ groups suggests that inter-strain
491 recombination is more prevalent in cancer-associated viruses than in circulating strains, whereas
492 the latter were more likely to have a homogeneous IR1.

493

494 **Changes in IR1 arise regionally and can efficiently purify to homogeneity in IR1.**

495 Looking in more depth at the nature of inter-strain recombination within IR1, we noted
496 that HL05 (a Hodgkin lymphoma biopsy) contains a cluster of 5 recombination-like MVs within
497 a 200bp region. We similarly noted a cluster of three MVs in the prototype EBV strain B95-8
498 (two of which are recombination-like) separated by only 126 bp (discussed in more detail later).
499 Other strains contained more widespread distribution of MVs, suggesting a larger region of
500 genetic exchange. Intriguingly two of these (TRL1 and sLCL-IS2.01) appeared to combine SNPs

501 from type 1 and type 2 viruses. TRL1 is a type 1 EBV containing a deletion within EBNA3B,
502 sequenced from two independent cell lines grown from an aggressive post-transplant DLBCL
503 (51, 52) while sLCL-IS2.01 is a spontaneous LCL established from the blood of a PTLD patient
504 containing a type 2 EBV (4). These strains also show evidence of clustering of the minor
505 variants, with different groups of MVs exhibiting different MV frequencies (Supplementary
506 Table ST5).

507 In order to analyse these variations, we used T-RECs to visualise the virus strains that
508 most closely resembled both the consensus and minor variant versions of the two strains (Figure
509 7). Both TRL1 and sLCL-IS2.01 showed evidence of having substantial regions of intertypic
510 recombination, TRL1 acquiring predominantly type 2 IR1 sequences in a type 1 background, and
511 sLCL-IS2.01 acquiring type 1 sequences in the type 2 background. To more clearly visualise the
512 transitions between type 1 and type 2 sequence without the signals caused by spontaneous
513 mutations, we used a bootscan analysis to view the transitions between the most closely related
514 type 1 and type 2 strains (Figure 8). This analysis clearly showed that both TRL1 and sLCL-
515 IS2.01 had acquired a region of the other type from Cp (the junction upstream of C1 is apparent
516 in the bootscan plots) to a position in IR1. However, the cross-type acquisition of sequence is no
517 longer represented by a single stretch of sequence. While the region from Wp to exon W2 is too
518 highly conserved to discriminate between strains, it is clear that the region from exon W2 to the
519 start of BWRF1, and the end of BWRF1 into Wp in sLCL-IS2.01 is entirely type 2, whereas the
520 centre of BWRF1 is almost entirely Akata-like, with an MV frequency of 17% for the type 2
521 sequence (Supplementary table ST5) suggesting that only a single repeat unit retains type 2
522 sequence here. For TRL1, the situation is more complex, showing an almost entirely type 2 IR1

sequence, with the ends of BWRF1 having type 1 sequence in one repeat unit, but the centre of BWRF1 being type 1 in two repeat units.

While it is not possible to conclusively establish how this combination of type 1 and type 2 sequences was acquired, in both cases, the patchy nature of the type 1 and type 2 sequence transitions is most easily explained by a two-step process, with the virus first acquiring a large region of sequence (for TRL1, a region from Cp to a middle repeat unit in IR1; for IS2.01, the region could be shorter), presumably by recombination, and the repeat subsequently undergoing ectopic gene conversion between repeat units, where some regions are templated to the original sequence, while others propagate the introduced region. Notably, in all five of the analysed strains, the blocks with a common MV frequency are relatively short – less than 500 bp. Indeed, for HL05, the only part of IR1 with evidence of inter-strain exchange is a small region, which could have been acquired in a single allelic gene conversion event between different virus genomes, rather than recombination followed by gene conversion.

The prototype strain B95-8 contains low frequency polymorphisms that include a stop codon in EBNA-LP.

The prototype EBV strain (B95-8) contains five heterogeneous positions two of which match SNPs in other virus genomes. The same set of MVs is found in one B95-8-infected cell line that we sequenced, one sequenced by the Chiang laboratory (21), and in cells containing B95-8-deLEBER, a recombinant subclone of the B95-8 BAC originally produced in the Hammerschmidt laboratory (53, 54). In addition, the identical MVs were found in BL36 – an African BL previously reported to be intertypic recombinant (4, 55) – and two of these MVs were also found in X50-7, an umbilical cord LCL established by co-culture with X-ray-irradiated B95-8 cells

546 (56). In all of these cell lines, four of the five SNPs share the same frequency in each genome
547 (around 8% in the parental B95-8 and around 20% in the other lines), which most likely
548 corresponds to an occurrence of once per IR1 (i.e. each sequence variant occurs in only one
549 repeat unit, such as is illustrated in Figure 5A). The MV in the hairpin region of B95-8 (position
550 12647) occurs at a higher frequency (Figure 9A).

551 These MVs within IR1 of B95-8 are distributed across the repeat unit (Figure 9B), but the
552 four MVs of equal frequency were twice (from independent digests) cloned from the EBV BAC
553 in a single BamHI fragment, demonstrating that they were all in the same repeat unit. Of the
554 three MVs that cluster towards the end of BWRFL1, two are SNPs in at least one other strain,
555 lying in a relatively polymorphic region of IR1. Most strikingly, however, the more distant MV
556 (G14617T – highlighted red in table ST4) changes the last codon of the W1 exon of EBNA-LP
557 from GAG (glutamate) to a TAG stop codon. As described elsewhere (35), we have generated a
558 B95-8 BAC (WT^w) in which the original IR1 has been replaced with 6.6 copies of the consensus
559 IR1 repeat unit, removing all MVs. LCLs established with WT^w exhibit both a higher and more
560 consistent level of EBNA-LP protein, and a more consistent range of larger EBNA-LP isoforms
561 (Figure 9C), suggesting that the STOP codon alters the character of EBNA-LP expression in
562 LCLs. Additionally, LCLs establish more rapidly with WT^w than with the parental HB9 BAC
563 (not shown). Consistent with this subjective observation, a dye dilution-based proliferation assay
564 shows that after 8 days, infected cells (those that have divided at least once) have typically
565 progressed through more cell divisions when infected with WT^w than with HB9 (Figure 9D).

566

567 **Discussion**

568 **Assembling repeat regions from short read data.**

569 Repeats are challenging to assemble from short read data, but we have shown here that
570 they can offer considerable insights into the biology of the virus and perhaps of virus-associated
571 disease. We have identified some strategies to improve the quality of repeat sequence assembly,
572 by considering flank and internal variations in sequence. Repeats also represent an internally
573 controlled sequence space, where detection of minor variants offers insights into events
574 occurring elsewhere in the virus genome. For instance, our data set – taken from clonal cell lines
575 – exhibited widespread low level variants, likely due to cross contamination or index exchange
576 facilitated by the high degree of multiplexing used in these sequencing reactions (43), and
577 identified error-prone sequence positions, suggesting that considerable care is needed when
578 searching for rare variants or quasispecies in virus populations.

579 Minor variant detection has only recently begun to be applied to herpesviruses at the
580 whole genome level: For instance, analysis of farmed carp showed infection by multiple strains
581 of Cyprinid herpesvirus 3 to be normal in moribund fish (57). Similarly, CMV infection of
582 humans is typified by infection with a combination of strains that is often stable over time
583 (reviewed in (58)), but can vary between different organs within individuals both in strain
584 combination and different degrees of virus diversity (12, 59). However, these studies do not
585 clearly distinguish between acquisition of new mutations and the presence of multiple diverse
586 strains, and do not make use of haplotype variation or shared MV frequency to explore the
587 causes of these observations, or provide analyses to estimate sample or technical cross-
588 contamination that could confound them.

589 One of the challenges in assessing the accuracy of our study is the lack of read data for
590 IR1 in other virus strains. We were able to confirm the B95-8-associated MVs by subcloning and
591 Sanger sequencing repeat units from the B95-8 BAC, and by analysing B95-8 sequence data
592 generated in another centre (21). Additionally, YCCEL1 was recently cloned as a BAC and this
593 BAC sequenced using PacBio long read technology. In agreement with our observations for
594 YCCEL1, the PacBio sequence did not find any MV SNPs within IR1. However, the assembled
595 YCCEL1 BAC sequence did contain a 1607 bp deletion in the 8th repeat unit of IR1 (Accession
596 number AP015016; (23)). We found no evidence for this truncated repeat in our reads, and
597 subsequent restriction and PCR analysis of the BAC was unable to detect this deleted repeat unit
598 (T Kanda; personal communication) suggesting that our assembly was accurate, and by
599 extension that the long read sequence of YCCEL1 required more careful manual curation, at
600 least for IR1.

601 **Biological implications of IR1 sequence diversity for EBV.**

602 The internal repeat regions of EBV are functionally poorly understood. Variable numbers
603 of W1-W2 exon pairs are known to comprise the upstream region of the EBNA transcripts (60).
604 It is alternatively spliced either to allow translation of EBNA-LP or to be retained as a 5' UTR
605 upstream of EBNA2 or one of the other EBNAs (25, 61). The W0 exon (and its associated
606 promoter, Wp) is at its most active immediately after infection, but its activity declines as
607 transcription switches to the upstream Cp (62). Very recently, transcripts spliced between W1
608 exons were described that produce the anti-apoptotic protein BHRF1 during the EBV lytic cycle
609 (26). The W1-W2 intron is reported to be a persistent RNA (sisRNA1) in latency III cells, and
610 there is some evidence that the larger W2-W1 exons may also produce a stable RNA, sisRNA2
611 (44).

612 The high degree of conservation observed in the Wp promoter and the W1-W2 intron
613 support the possibility that they have a specific conserved function. In contrast, the conservation
614 of the BWRF1 ORF in only 80% of strains casts some doubt on its importance, at least in its
615 current (hypothetical) context. Even so, it is unlikely that such a large ORF would arise by
616 chance (and be conserved in so many strains). It is also striking that all six examples of genome
617 deletions that delete the flanks of IR1 are in the small zone that leave the hairpin-BWRF1 region
618 of the genome intact (figure 2B), suggesting that this part of IR1 represents a single functional
619 entity, either as a DNA structure or acting through sisRNA2. Because there are multiple copies
620 of this region in IR1, it also implies that partial copies may act in some sort of dominant negative
621 manner, or that the first and last units are more important. A case for the first repeat unit being
622 more important has previously been made, in that deleting the regulatory region upstream of Wp
623 was found to be tolerated in internal IR1 repeat units, but mutation of the first repeat unit had a
624 detrimental effect on B cell transformation and Cp activity (63), implying that the integrity of the
625 BWRF1 region may be important for proper activity of the Wp and Cp promoters.

626 This BWRF1 region is the source of the reportedly stable intronic RNA, sisRNA2 (44).
627 Conceptually, sisRNA2 has similarities to the latency-associated transcript (LAT) of HSV,
628 which is critical for HSV latency in neurons. HSV-LAT is also an intron-derived persistent RNA
629 that is processed by splicing events in addition to the simple excision of the intron (64) (65). Two
630 regions of BWRF1 are highly conserved, while all of the sequence changes that disrupt BWRF1,
631 including the larger indels found in a few strains are all in the central polymorphic region. It
632 therefore remains possible that some parts of the BWRF1 ORF may be used to generate
633 translation products if sisRNA-2 is processed to excise the less conserved regions. Additionally,
634 some mechanism – perhaps mediated by an RNA structure associated with the hairpin – would

635 be required to facilitate BWRF1 translation. The polypeptides encoded by HSV-LAT are
636 reported to be untranslated (65), but detailed analysis of the sisRNA-2 would be required to test
637 whether this is also true for EBV.

638 In contrast to BWRF1, the repeat domain of EBNA-LP is highly conserved across IR1,
639 with most of its diversity residing in the C-terminal Y exons. This is distinct from the Rhesus
640 lymphocryptovirus, whose EBNA-LP transcript contained a final W1/2 exon pair that differed
641 substantially from the first two (66). Our MV analysis found no evidence of similar W exon
642 diversity in EBV, and indeed only two strains (X50-7 and sLCL1.12) contained an MV in exon
643 W2, while the stop codon in B95-8-related strains was the only MV in exon W1 in any virus.
644 Our analysis reinforces previous observations that the 'type 2' W exon sequence is not confined
645 to type 2 viruses. These are easily distinguished, as the monoclonal antibody JF186 only binds to
646 the type 1 variant (45), whereas antibody 4D3 detects an epitope in exon W2 (67), so can be used
647 to confirm this difference between strains ((67) and our unpublished observations). Using our
648 new nomenclature to describe EBNA-LP variants highlights that there appear to be some
649 restrictions on which W and Y exon variants are found together, which suggests possible
650 functional co-dependence. Particularly intriguing is the B^M subgroup, which is found only in
651 three NPC-derived viruses, and appears to be compatible with both type 1 and type 2 W exons.
652 However, functional assays and/or structural information will be required to establish whether
653 different EBNA-LP W and Y exon pairs are cross compatible.

654 **Functional consequences of IR1 heterogeneity.**

655 Only limited attempts have previously been made to investigate the heterogeneity of IR1
656 or the other EBV repeats, although a recent study reported a relatively low proportion of
657 positions containing MVs (68). While the genome sequence of B95-8 reports heterogeneity in

the terminal repeats (9), contemporaneous investigations searching for heterogeneity relied on restriction analysis, and found no evidence of IR1 heterogeneity in B95-8 or JHU-1 EBV strains (28, 69). However, the early sequences of B95-8 IR1 did disagree over whether B95-8 position 12647 was a G or a T (9, 27, 28). We found that this position is approximately a 50:50 split in the parental B95-8 viruses (Figure 9). It may appear more surprising that the STOP codon in EBNA-LP has not previously been reported. However, because EBNA-LP transcription is characterised by a poorly characterised pattern of alternative splicing and exon skipping (25), variations in the predominant EBNA-LP sizes in LCLs has become regarded as normal, concealing this mutation. We have shown that an intact IR1 of 6.6 repeat units produces EBNA-LP with predominantly of 4, 5, and 6 repeats, whereas only the HB9 seemed to select single smaller sized EBNA-LPs. Curiously, we even see evidence of EBNA-LP with more repeat domains than IR1 repeat units, which are most likely produced by splicing between (rather than within) transcripts. It is not formally possible to say whether the EBNA-LP produced in HB9-LCLs is truncated or not. However, we have found that deletion of the Y exons from the B95-8 BAC (in the context of heterogeneous IR1) results in very low protein levels (35), suggesting that any truncated EBNA-LP made may not be stable. Nevertheless, we have observed that the MVs in B95-8 do appear to have a detrimental impact on its transforming ability. Since these minor variants are retained in the B95-8 cell line, it is possible that the reduced levels of EBNA-LP are only significant during initial transformation, or that they are favoured in the marmoset background of the B95-8 cell line. Nevertheless, a great deal of research has been performed using the B95-8 BAC as a basis for genetic manipulation of the virus (70). It is therefore important to consider that these experiments have been done in a background of a modestly

680 defective EBNA-LP, and may be advisable that future genetic studies use an EBV-BAC with an
681 intact EBNA-LP.

682 The pattern of IR1 heterogeneity (including the B95-8-specific flank SNP) is retained in a
683 number of B95-8-related strains. The commonly used X50-7 strain, whose genome is integrated,
684 and has lost Cp (42), is missing three of the B95-8 MVs, but since the genome was UV irradiated
685 prior to transformation (56), it seems likely that these MVs were lost during homology driven
686 DNA repair. The B95-8-type MV heterogeneity is completely reproduced in the intertypic
687 recombinant BL36, which was not known to be B95-8 associated. Further examination of the
688 BL36 sequence supports the hypothesis that its genome arose by super-infection of a type 2 BL
689 with the B95-8 virus resulting in intertypic recombination between the two: Between position
690 10024 (where BL36 contains 14 bp palindromic insertion disrupting the vIL10 gene – possibly a
691 scar from the recombination) and position 43611, the BL36 and B95-8 sequences are identical.
692 This includes five SNPs – three in IR1 and two in the Y exons – that are only found in B95-8.
693 We therefore suggest that BL36 is a lab artefact, and is not representative of a natural circulating
694 strain. Nevertheless, additional examples of intertypic recombinants remain. They have been
695 reproducibly isolated from both the blood and throat of T cell immunodeficient Caucasian HIV
696 patients (71), as well as from two NPC patients (one with a mutant EBNA3A) and from one
697 healthy Chinese individual with type 1 EBNA 2 and EBNA3A with type 2 EBNA3B and 3C
698 (72), and African sLCL-1.18 from this dataset, which has a typical type 1 IR1 and EBNA2 with
699 type 2 EBNA3s. Since type 2 EBNA2 gives a reduced efficiency of LCL outgrowth (6), it
700 remains possible that at least some of these examples may be artefactual strains produced as a
701 result of the selection pressures associated with LCL outgrowth. However, despite our

702 identification of BL36 an artefactual intertypic recombinant, the evidence for their existence in
703 the wild remains firm, but would benefit from further corroboration.

704 Further corroboration is also required for our observation that tumour-derived viruses
705 appear more likely to have evidence of inter-strain genetic exchange in IR1. The non-random
706 sampling of these strains make any statistical comparison of the two groups uninformative, but
707 there are other examples of changes in virus genomes are associated with disease. For instance,
708 deletion of EBNA2 and the end of EBNA-LP is seen in approximately 10% of BL (73-75), while
709 EBNA3B mutation is associated with DLBCL, and also found in HL and BL (52). Indeed HL08
710 contains a small deletion in EBNA3B that has distinct haplotypes on either side (52), implicating
711 an aberrant genetic exchange between viruses in its origin. The association of inter-strain
712 recombinants with disease could indicate that aberrations in the process of genetic exchange
713 could produce pathogenic mutations. Equally, it is possible that the processes that promote
714 oncogenesis also promote the exchange of DNA between distinct EBV strains. Nevertheless, this
715 would still require the oncogenic cell to be co-infected by two distinct viruses, which – if it is
716 random – seems likely to be a relatively rare event. Further studies, such as by comparing the
717 sequences of EBV in tumours with that in the patient's blood and/or saliva, will be required to
718 establish whether co-infection and genetic exchange between viruses may contribute to
719 pathogenesis.

720 **Mechanisms and implications of inter-strain genetic exchange in IR1.**

721 Whole genome analyses supported earlier suggestions that inter-strain recombination is a
722 major component of diversity in EBV (4), HSV1 (15) and CMV(13), and by extension this
723 potentially makes evolutionary analysis of herpesvirus genomes enormously complicated. By
724 using the repeat units as internal controls, we have been able to deduce patterns of the

725 accumulation of changes in virus genomes, classifying them as either spontaneous mutations or
726 inter-strain genetic exchange. Our analysis is necessarily an oversimplification, as we do not yet
727 have a comprehensive list of EBV SNPs, so some SNP-like changes will have been missed.
728 However, the presence in a few strains of large clusters of exchanged SNPs has offered
729 considerable insight into the processes by which EBV strains can exchange genetic information,
730 and thereby gain diversity.

731 Inter-strain recombination would allow large regions of virus genomes to be exchanged,
732 as seen in BL36, and probably TRL1 and sLCL-IS2.01 (Figures 7 and 8). However, most of the
733 regions of SNP-like MVs exchanged in IR1 are much smaller, between 150 and 1000 bp in
734 length, best exemplified by the distinct region introduced in HL05. Small inter-strain exchanges
735 of sequence have been reported elsewhere, such as a type 1 virus reported to contain a type 2
736 sequence at the end of EBNA3A that spanned 882 nucleotides (76). However, if these small
737 exchanges of sequence are commonplace throughout the EBV genome (and those of other
738 herpesviruses), it makes the development of a nomenclature describing herpesvirus strain
739 variation and relationships between strains extremely complicated.

740 Classically, the diversification of repeats and their subsequent purification to a new
741 sequence has been explained by the process of crossover fixation - the expansion and contraction
742 of repeats by crossover recombination. This process at the gene level - a 'gene accordion' model
743 - has been reported to be a key mechanism in the diversification of poxvirus genes, and requires
744 a linear DNA intermediate (77). However, this model would predict the MV diversity to span
745 whole repeat units, which we do not generally observe, with the possible exception of some parts
746 of TRL1. Genetic repeats are also driven to homogeneity through allelic gene conversion, which
747 is mechanistically based around the homologous repair of double stranded breaks (78). The small

748 patches of linked MVs in IR1 are in the size range typical of synthesis-dependent strand
749 annealing-mediated break repair (78). Alternative methods of templated DNA break repair such
750 as single strand annealing or DNA-break induced replication (BIR) (reviewed in (49, 79)) could
751 also cause these patterns. In particular, BIR generates an extended ssDNA intermediate that may
752 be more prone to mutation, which could explain the cluster of apparent spontaneous mutations in
753 HL04, and more generally the observation that strains with evidence of genetic exchange also
754 contain spontaneous mutations.

755 Most analyses of herpesvirus recombination have been performed in HSV1. These
756 analyses have all used models of coinfection during the lytic virus lifecycle (reviewed in (80)). In
757 contrast, EBV also undergoes a substantial proportion of its genome replication in latency in B
758 cells, where viral genome replication is similar to that of the chromosome, rather than during the
759 lytic cycle. This difference is supported by sequence analysis of diverse herpesvirus genomes,
760 which found that gammaherpesviruses – but not alphaherpesviruses – are enriched for distinct
761 sequences reported to promote genetic exchange (81), including one also involved in the B cell
762 specific process of antibody class switch recombination (82). Thus the apparently higher
763 frequency of small repair-driven genetic exchange may be a reflection of the biological
764 differences in latency between alpha- and gammaherpesviruses.

765 The sequence diversity of the repeat leads us to propose a model where there is an initial
766 introduction of genetic sequence into the repeat region, followed by the purification through gene
767 conversion of sequences across IR1. In the case of TRL1 and sLCL-IS2.01, the mixed genotype
768 likely arose through a large recombination event introducing Cp and one or more repeat units. In
769 contrast, the small region in HL04 seems most likely to have been introduced through allelic
770 gene conversion (i.e. repair in one virus genome templated by another genome), although a

771 larger region could have been introduced by recombination, and then much of it lost as the repeat
772 mostly purified to its original sequence. It appears likely that ectopic gene conversion drives the
773 repeat towards homogeneity, with repeat units in the same virus genome templating repairs.

774 Our analysis of IR1 also hints at the limits of gene conversion in this setting, as SNPs
775 within 120 bp of the flanks of IR1 appear to mainly be fixed in their respective populations,
776 while those a little further out are found in some – but not all – related strains. It suggests that the
777 concerted evolution of IR1 typically relies on a considerable region of homology (perhaps 100-
778 200 bp) on either side of any variant in the target region, fitting a gene conversion model.
779 However, three strains (HL11, L591 and sLCL-BL1.03) have propagated a flank SNP to purity
780 across IR1, suggesting that the slow conversion of SNPs in the flank by internal sequences
781 contrasts with potentially rapid concerted evolution in the body of the repeat, indicated by the
782 overall low degree of IR1 heterogeneity.

783 In summary, our in depth analysis of IR1 sequence variation has uncovered a number of
784 surprising features of EBV biology. We have identified a new variant of EBNA-LP, and suggest
785 that it is important to the virus to retain the BWRF1 region intact, but not for the BWRF1 ORF
786 to be retained in full. We observe that IR1 is often heterogeneous, and that this heterogeneity
787 offers considerable insights into viral evolution, suggesting that the exchange of genetic
788 information between strains resembles gene conversion, and demonstrate the subsequent
789 concerted evolution of the repeats. We suggest that inter-strain recombination may contribute to
790 pathogenesis, and we have identified a mutation in EBNA-LP of the prototype strain of EBV. We
791 hope that the sequence analysis of further virus strains will help to clarify the validity (or
792 otherwise) of these observations.

793

794 Acknowledgements

795 We would like to thank: Prof Paul Farrell for suggestions and critical reading of the
796 manuscript; James Abbot of the Imperial College bioinformatics department for technical advice
797 on repeat sequencing; Alan Chiang for provision of his B95-8 short read data; Teru Kanda for his
798 prompt reanalysis of the YCCEL1 BAC; Darren Martin for the rapid response to our bug report
799 on RDP4.

800

801 Funding Information.

802 This work was primarily funded by the Medical Research Council (grant numbers
803 MR/L008432/1 and MR/N010388/1) to REW, which supports RP and AS. MBa was supported
804 by a PhD studentship from King Abdulaziz City for Science and Technology, Saudi Arabia. REW
805 and AS were supported by Imperial College Trust; SC by a Medical Research Council PhD
806 Studentship; AP, NG and PK supported by the Wellcome Trust (grant number 098051).

807

808 References

- 809 1. **de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, Plummer M.** 2012.
810 Global burden of cancers attributable to infections in 2008: a review and synthetic
811 analysis. *Lancet Oncol* **13**:607–615.
- 812 2. **Chang ET, Adami H-O.** 2006. The enigmatic epidemiology of nasopharyngeal
813 carcinoma. *Cancer Epidemiol Biomarkers Prev* **15**:1765–1777.
- 814 3. **Camargo MC, Murphy G, Koriyama C, Pfeiffer RM, Kim WH, Herrera-Goepfert R,**
815 **Corvalan AH, Carrascal E, Abdirad A, Anwar M, Hao Z, Kattoor J, Yoshiwara-**
816 **Wakabayashi E, Eizuru Y, Rabkin CS, Akiba S.** 2011. Determinants of Epstein-Barr
817 virus-positive gastric cancer: an international pooled analysis. *Br J Cancer* **105**:38–43.
- 818 4. **Palser AL, Grayson NE, White RE, Corton C, Correia S, Ba Abdullah MM, Watson**
819 **SJ, Cotten M, Arrand JR, Murray PG, Allday MJ, Rickinson AB, Young LS, Farrell**
820 **PJ, Kellam P.** 2015. Genome diversity of Epstein-Barr virus from multiple tumor types
821 and normal infection. *J Virol* **89**:5222–5237.

- 822 5. **McGeoch DJ, Dolan A, Ralph AC.** 2000. Toward a Comprehensive Phylogeny for
823 Mammalian and Avian Herpesviruses. *J Virol* **74**:10401–10406.
- 824 6. **Rickinson AB, Young LS, Rowe M.** 1987. Influence of the Epstein-Barr virus nuclear
825 antigen EBNA 2 on the growth phenotype of virus-transformed B cells. *J Virol* **61**:1310–
826 1317.
- 827 7. **Coleman CB, Smith NA, Rochford R.** 2014. Epstein-Barr Virus Type 2 Latently Infects
828 T Cells, Inducing an Atypical Activation Characterized by Expression of Lymphotactic
829 Cytokines. *J Virol* **89**:2301–2312.
- 830 8. **Tzellos S, Farrell PJ.** 2012. Epstein-barr virus sequence variation-biology and disease.
831 *Pathogens* **1**:156–174.
- 832 9. **Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G,
833 Hudson GS, Satchwell SC, Séguin C, Tuffnell PS, Barrell BG.** 1984. DNA sequence
834 and expression of the B95-8 Epstein-Barr virus genome. *Nature* **310**:207–211.
- 835 10. **Elgueta Karstegl C, Niller HH, Huang D.** 2003. Updated Epstein-Barr virus (EBV)
836 DNA sequence and analysis of a promoter for the BART (CST, BARF0) RNAs of EBV. *J*
837 *Gen Virol* **84**:1443–1450.
- 838 11. **Dolan A, Addison C, Gatherer D, Davison AJ, McGeoch DJ.** 2006. The genome of
839 Epstein-Barr virus type 2 strain AG876. *Virology* **350**:164–170.
- 840 12. **Renzette N, Pokalyuk C, Yamamoto AY, Britt WJ, Jensen JD.** 2015. Limits and
841 patterns of cytomegalovirus genomic diversity in humans. *Proc Natl Acad Sci U S A*
842 **112**:E4120–E4128.
- 843 13. **Sijmons S, Thys K, Van Loock M.** 2015. High-throughput analysis of human
844 cytomegalovirus genome diversity highlights the widespread occurrence of gene-
845 disrupting mutations and pervasive recombination. *J Virol* **89**:7673–7695.
- 846 14. **Tafari YR, Shreve JT, Enquist LW.** 2015. Rapid genome assembly and comparison
847 decode intrastrain variation in human alphaherpesviruses. *MBio* **6**:e02213–14.
- 848 15. **Szpara ML, Gatherer D, Ochoa A, Greenbaum B, Dolan A, Enquist LW.** 2014.
849 Evolution and diversity in human herpes simplex virus genomes. *J Virol* **88**:1209–1227.
- 850 16. **Lamers SL, Saif S, Young S, Henn M, Laeyendecker O, Knipe DM.** 2015. Genome
851 Sequencing and Analysis of Geographically Diverse Clinical Isolates of Herpes Simplex
852 Virus 2. *J Virol* **89**:8219–8232.
- 853 17. **Kolb AW, Cuellar JA, Brandt CR.** 2015. Genomic, phylogenetic, and recombinational
854 characterization of herpes simplex virus 2 strains. *J Virol* **89**:6427–6434.
- 855 18. **Zell R, Taudien S, Pfaff F, Wutzler P, Platzer M, Sauerbrei A.** 2011. Sequencing of 21
856 Varicella-Zoster Virus Genomes Reveals Two Novel Genotypes and Evidence of

- 857 Recombination. *J Virol* **86**:1608–1622.
- 858 19. **Wang X, Strong MJ, Baddoo M, Baribault C, Taylor CM.** 2013. Whole-genome
859 sequencing of the Akata and Mutu Epstein-Barr virus strains. *J Virol* **87**:1172–1182.
- 860 20. **Hung G-C, Lo S-C.** 2013. Identification and characterization of EBV genomes in
861 spontaneously immortalized human peripheral blood B lymphocytes by NGS technology.
862 *BMC Genomics* **14**:804.
- 863 21. **Kwok H, Wu CW, Kellam P, Chiang AKS.** 2014. Genomic diversity of Epstein-Barr
864 virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. *J Virol*
865 **88**:10662–10672.
- 866 22. **Simbiri KO, Smith NA, Otieno R, Wohlford EEM, Daud II, Odada SP, Middleton F,**
867 **Rochford R.** 2015. Epstein-barr virus genetic variation in lymphoblastoid cell lines
868 derived from kenyan pediatric population. *PLoS ONE* **10**:e0125420.
- 869 23. **Kanda T, Furuse Y, Oshitani H.** 2016. Highly Efficient CRISPR/Cas9-Mediated
870 Cloning and Functional Characterization of Gastric Cancer-Derived Epstein-Barr Virus
871 Strains. *J Virol* **90**:4383–4393.
- 872 24. **Kao K-Y, Nagra JK.** 2011. Epstein-Barr virus BamHI W repeat number limits
873 EBNA2/EBNA-LP coexpression in newly infected B cells and the efficiency of B-cell
874 transformation: a rationale for the multiple W repeats in wild-type virus strains. *J Virol*
875 **85**:12362–12375.
- 876 25. **Rogers RP, Woisetschlaeger M, Speck SH.** 1990. Alternative splicing dictates
877 translational start in Epstein-Barr virus transcripts. *EMBO J* **9**:2273–2277.
- 878 26. **O'Grady T, Wang X, Höner Zu Bentrup K, Baddoo M, Concha M, Flemington EK.**
879 2016. Global transcript structure resolution of high gene density genomes through multi-
880 platform data integration. *Nucleic Acids Res* **44**:e145–e145.
- 881 27. 1982. Long internal direct repeat in Epstein-Barr virus DNA. *J Virol* **44**:286–294.
- 882 28. **Jones MD, Griffin BE.** 1983. Clustered repeat sequences in the genome of Epstein Barr
883 virus. *Nucleic Acids Res* **11**:3919–3937.
- 884 29. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-
885 MEM. *arXiv*.
- 886 30. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,**
887 **Durbin R, 1000 Genome Project Data Processing Subgroup.** 2009. The Sequence
888 Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.
- 889 31. **Tsimpidis M, Bachoumis G, Mimouli K, Kyriakopoulou Z, Robertson DL,**
890 **Markoulatos P, Amoutzias GD.** 2017. T-RECs: rapid and large-scale detection of
891 recombination events among different evolutionary lineages of viral genomes. *BMC*

- 892 Bioinformatics **18**:13.
- 893 32. **Martin DP, Williamson C.** 2005. A modified bootscan algorithm for automated
894 identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum*
895 *Retroviruses* **21**:98–102.
- 896 33. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B.** 2015. RDP4: Detection and
897 analysis of recombination patterns in virus genomes. *Virus Evolution* **1**:vev003.
- 898 34. **Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis**
899 **ER, Ding L, Wilson RK.** 2012. VarScan 2: somatic mutation and copy number alteration
900 discovery in cancer by exome sequencing. *Genome Res* **22**:568–576.
- 901 35. **Szymula A, Palermo RD, Groves IJ, abdullah MB, Holder BS, White RE.** 2017.
902 Epstein-Barr virus nuclear antigen EBNA-LP is essential for transforming naive B cells,
903 and facilitates recruitment of transcription factors to the viral genome. *bioRxiv* DOI:
904 10.1101–176099.
- 905 36. **McCann EM, Kelly GL, Rickinson AB, Bell AI.** 2001. Genetic analysis of the Epstein-
906 Barr virus-coded leader protein EBNA-LP as a co-activator of EBNA2 function. *J Gen*
907 *Virol* **82**:3067–3079.
- 908 37. **Harada S.** 1997. Epstein-Barr virus nuclear protein LP stimulates EBNA-2 acidic
909 domain-mediated transcriptional activation. *J Virol* **71**:6611–6618.
- 910 38. **Ling PD, Tan J, Peng R.** 2009. Nuclear-cytoplasmic shuttling is not required for the
911 Epstein-Barr virus EBNA-LP transcriptional coactivation function. *J Virol* **83**:7109–7116.
- 912 39. **Rabson M, Gradoville L, Heston L, Miller G.** 1982. Non-immortalizing P3J-HR-1
913 Epstein-Barr virus: a deletion mutant of its transforming parent, Jijoye. *J Virol* **44**:834–
914 844.
- 915 40. **Jeang KT, Hayward SD.** 1983. Organization of the Epstein-Barr virus DNA molecule.
916 III. Location of the P3HR-1 deletion junction and characterization of the NotI repeat units
917 that form part of the template for an abundant 12-O-tetradecanoylphorbol-13-acetate-
918 induced mRNA transcript. *J Virol* **48**:135–148.
- 919 41. **Jones MD, Foster L, Sheedy T, Griffin BE.** 1984. The EB virus genome in Daudi
920 Burkitt's lymphoma cells has a deletion similar to that observed in a non-transforming
921 strain (P3HR-1) of the virus. *EMBO J* **3**:813–821.
- 922 42. **Yandava CN, Speck SH.** 1992. Characterization of the deletion and rearrangement in the
923 BamHI C region of the X50-7 Epstein-Barr virus genome, a mutant viral strain which
924 exhibits constitutive BamHI W promoter activity. *J Virol* **66**:5646–5650.
- 925 43. **Kircher M, Sawyer S, Meyer M.** 2012. Double indexing overcomes inaccuracies in
926 multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**:e3.

- 927 44. **Moss WN, Steitz JA.** 2013. Genome-wide analyses of Epstein-Barr virus reveal
928 conserved RNA structures and a novel stable intronic sequence RNA. *BMC Genomics*
929 **14**:543.
- 930 45. **Finke J, Rowe M, Kallin B, Ernberg I, Rosén A, Dillner J, Klein G.** 1987. Monoclonal
931 and polyclonal antibodies against Epstein-Barr virus nuclear antigen 5 (EBNA-5) detect
932 multiple protein species in Burkitt's lymphoma and lymphoblastoid cell lines. *J Virol*
933 **61**:3870–3878.
- 934 46. **Tso KK-Y, Yip KY-L, Mak CK-Y, Chung GT-Y, Lee S-D, Cheung S-T, To KF, Lo**
935 **KW.** 2013. Complete genomic sequence of Epstein-Barr virus in nasopharyngeal
936 carcinoma cell line C666-1. *Infect Agents Cancer* **8**:29.
- 937 47. **Xiao K, Yu Z, Li X, Li X, Tang K, Tu C, Qi P, Liao Q, Chen P, Zeng Z, Li G, Xiong**
938 **W.** 2016. Genome-wide Analysis of Epstein-Barr Virus (EBV) Integration and Strain in
939 C666-1 and Raji Cells. *J Cancer* **7**:214–224.
- 940 48. **Hastings PJ.** 2010. Mechanisms of Ectopic Gene Conversion. *Genes* **1**:427–439.
- 941 49. **Verma P, Greenberg RA.** 2016. Noncanonical views of homology-directed DNA repair.
942 *Genes Dev* **30**:1138–1154.
- 943 50. **Given D, Kieff E.** 1979. DNA of Epstein-Barr virus. VI. Mapping of the internal tandem
944 reiteration. *J Virol* **31**:315–324.
- 945 51. **Gottschalk S, Ng CY, Perez M, Smith CA, Sample C, Brenner MK, Heslop HE,**
946 **Rooney CM.** 2001. An Epstein-Barr virus deletion mutant associated with fatal
947 lymphoproliferative disease unresponsive to therapy with virus-specific CTLs. *Blood*
948 **97**:835–843.
- 949 52. **White RE, Rämer PC, Naresh KN, Meixlsperger S, Pinaud L, Rooney C, Savoldo B,**
950 **Coutinho R, Bödör C, Gribben J, Ibrahim HA, Bower M, Nourse JP, Gandhi MK,**
951 **Middeldorp J, Cader FZ, Murray P, Münz C, Allday MJ.** 2012. EBNA3B-deficient
952 EBV promotes B cell lymphomagenesis in humanized mice and is found in human
953 tumors. *J Clin Invest* **122**:1487–1502.
- 954 53. **Delecluse HJ, Hilsendegen T, Pich D, Zeidler R, Hammerschmidt W.** 1998.
955 Propagation and recovery of intact, infectious Epstein-Barr virus from prokaryotic to
956 human cells. *Proc Natl Acad Sci U S A* **95**:8245–8250.
- 957 54. **Gregorovic G, Pattle S, Chiang AKS, Kracht M, Russ R, Farrell PJ.** 2011. Cellular
958 gene expression that correlates with EBER expression in Epstein-Barr Virus-infected
959 lymphoblastoid cell lines. *J Virol* **85**:3535–3545.
- 960 55. **Lenoir GM, Vuillaume M, Bonnardel C.** 1985. The use of lymphomatous and
961 lymphoblastoid cell lines in the study of Burkitt's lymphoma. *IARC Sci Publ* 309–318.
- 962 56. **Wilson G, Miller G.** 1979. Recovery of Epstein-Barr virus from nonproducer neonatal

- human lymphoid cell transformants. *Virology* **95**:351–358.
57. **Hammoumi S, Vallaey S, Santika A, Leleux P, Borzym E, Klopp C, Avarre J-C.** 2016. Targeted genomic enrichment and sequencing of CyHV-3 from carp tissues confirms low nucleotide diversity and mixed genotype infections. *PeerJ* **4**:e2516.
58. **Renzette N, Gibson L, Jensen JD, Kowalik TF.** 2014. Human cytomegalovirus intrahost evolution-a new avenue for understanding and controlling herpesvirus infections. *Curr Opin Virol* **8**:109–115.
59. **Ross SA, Novak Z, Pati S, Patro RK, Blumenthal J, Danthuluri VR, Ahmed A, Michaels MG, Sánchez PJ, Bernstein DI, Tolan RW, Palmer AL, Britt WJ, Fowler KB, Boppana SB.** 2011. Mixed infection and strain diversity in congenital cytomegalovirus infection. *J Infect Dis* **204**:1003–1007.
60. **Bodescot M, Perricaudet M.** 1986. Epstein-Barr virus mRNAs produced by alternative splicing. *Nucleic Acids Res* **14**:7103–7114.
61. **Sample J, Hummel M, Braun D, Birkenbach M, Kieff E.** 1986. Nucleotide sequences of mRNAs encoding Epstein-Barr virus nuclear proteins: a probable transcriptional initiation site. *Proc Natl Acad Sci U S A* **83**:5096–5100.
62. **Woisetschlaeger M, Yandava CN, Furmanski LA, Strominger JL, Speck SH.** 1990. Promoter switching in Epstein-Barr virus during the initial stages of infection of B lymphocytes. *Proc Natl Acad Sci U S A* **87**:1725–1729.
63. **Yoo LI, Woloszynek J, Templeton S, Speck SH.** 2002. Deletion of Epstein-Barr Virus Regulatory Sequences Upstream of the EBNA Gene Promoter Wp1 Is Unfavorable for B-Cell Immortalization. *J Virol* **76**:11763–11769.
64. **Farrell MJ, Dobson AT, Feldman LT.** 1991. Herpes simplex virus latency-associated transcript is a stable intron. *Proc Natl Acad Sci U S A* **88**:790–794.
65. **Spivack JG, Woods GM, Fraser NW.** 1991. Identification of a novel latency-specific splice donor signal within the herpes simplex virus type 1 2.0-kilobase latency-associated transcript (LAT): translation inhibition of LAT open reading frames by the intron within the 2.0-kilobase LAT. *J Virol* **65**:6800–6810.
66. **Peng R, Gordadze AV, Fuentes Pananá EM, Wang F, Zong J, Hayward GS, Tan J, Ling PD.** 2000. Sequence and functional analysis of EBNA-LP and EBNA2 proteins from nonhuman primate lymphocryptoviruses. *J Virol* **74**:379–389.
67. **Shaku F, Matsuda G, Furuya R, Kamagata C, Igarashi M, Tanaka M, Kanamori M, Nishiyama Y, Yamamoto N, Kawaguchi Y.** 2005. Development of a monoclonal antibody against Epstein-Barr virus nuclear antigen leader protein (EBNA-LP) that can detect EBNA-LP expressed in P3HR1 cells. *Microbiol Immunol* **49**:477–483.
68. **Chiara M, Manzari C, Lionetti C, Mechelli R, Anastasiadou E, Chiara Buscarinu M,**

- 999 **Ristori G, Salvetti M, Picardi E, D'Erchia AM, Pesole G, Horner DS.** 2016.
1000 Geographic Population Structure in Epstein-Barr Virus Revealed by Comparative
1001 Genomics. *Genome Biol Evol* **8**:3284–3291.
- 1002 69. **Hayward SD, Noguee L, Hayward GS.** 1980. Organization of repeated regions within the
1003 Epstein-Barr virus DNA molecule. *J Virol* **33**:507–521.
- 1004 70. **Feederle R, Bartlett EJ, Delecluse H-J.** 2010. Epstein-Barr virus genetics: talking about
1005 the BAC generation. *Herpesviridae* **1**:6.
- 1006 71. **Yao QY, Tierney RJ, Croom-Carter D, Cooper GM, Ellis CJ, Rowe M, Rickinson**
1007 **AB.** 1996. Isolation of intertypic recombinants of Epstein-Barr virus from T-cell-
1008 immunocompromised individuals. *J Virol* **70**:4895–4903.
- 1009 72. **Midgley RS, Blake NW, Yao QY, Croom-Carter D, Cheung ST, Leung SF, Chan AT,**
1010 **Johnson PJ, Huang D, Rickinson AB, Lee SP.** 2000. Novel intertypic recombinants of
1011 Epstein-Barr virus in the Chinese population. *J Virol* **74**:1544–1548.
- 1012 73. **Kelly GL, Croom-Carter DSG, Hammerschmidt W, Bell AI.** 2005. Epstein-Barr virus
1013 nuclear antigen 2 (EBNA2) gene deletion is consistently linked with EBNA3A, -3B, and -
1014 3C expression in Burkitt's lymphoma cells and with increased resistance to apoptosis. *J*
1015 *Virol* **79**:10709–10717.
- 1016 74. **Kelly G, Bell A, Rickinson A.** 2002. Epstein-Barr virus-associated Burkitt
1017 lymphomagenesis selects for downregulation of the nuclear antigen EBNA2. *Nat Med*
1018 **8**:1098–1104.
- 1019 75. **Kelly GL, Long HM, Stylianou J, Thomas WA, Leese A, Bell AI, Bornkamm GW,**
1020 **Mautner J, Rickinson AB, Rowe M.** 2009. An Epstein-Barr virus anti-apoptotic protein
1021 constitutively expressed in transformed cells and implicated in Burkitt lymphomagenesis:
1022 the Wp/BHRF1 link. *PLoS Pathog* **5**:e1000341.
- 1023 76. **Görzer I, Niesters HGM, Cornelissen JJ, Puchhammer-Stöckl E.** 2006.
1024 Characterization of Epstein-Barr virus Type I variants based on linked polymorphism
1025 among EBNA3A, -3B, and -3C genes. *Virus Res* **118**:105–114.
- 1026 77. **Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, Geballe AP,**
1027 **Malik HS.** 2012. Poxviruses deploy genomic accordions to adapt rapidly against host
1028 antiviral defenses. *Cell* **150**:831–841.
- 1029 78. **Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP.** 2009. Gene Conversion
1030 in Evolution and Disease. *In* Encyclopedia of Life Sciences (ELS). John Wiley & Sons
1031 Ltd. DOI: 10.1002/9780470015902.a0005100.pub2
- 1032 79. **Chan K, Gordenin DA.** 2015. Clusters of Multiple Mutations: Incidence and Molecular
1033 Mechanisms. *Annu Rev Genet* **49**:243–267.
- 1034 80. **Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F.** 2015.

1035 Recombination in viruses: mechanisms, methods of study, and evolutionary consequences.
1036 Infect Genet Evol **30**:296–307.

1037 81. **Brown JC.** 2014. The role of DNA repair in herpesvirus pathogenesis. Genomics
1038 **104**:287–294.

1039 82. **Chuzhanova N, Chen J-M, Bacolla A, Patrinos GP, Férec C, Wells RD, Cooper DN.**
1040 2009. Gene conversion causing human inherited disease: evidence for involvement of
1041 non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage
1042 and repair. Hum Mutat **30**:1189–1198.

1043

1044 **Figure legends**

1045 **Figure 1. Schematic representations of the IR1 region of EBV.** (A) Schematic representation
1046 of the EBV genome showing IR1 composed of the typical 5.6 repeat units (white boxes) with the
1047 other major repeats of EBV (Internal repeats [IR]2-4, FR of *oriP* and TR; yellow boxes), with
1048 latency genes shown as ORFs (red arrows) or spliced transcripts (black arrows), and the major
1049 latency promoters (Qp, Cp) shown as bent arrows. (B) Schematic view of the IR1 region with its
1050 flanks. The BWRF1 (large red arrow) is stepped to represent the two major forms of the BWRF1
1051 ORF: the left end of the arrow indicates the longer form (previously reported in AG876) and the
1052 widened arrow indicates the first position at which the shorter (B95-8) isoform might start. (C)
1053 The structure of the simplified IR1 template (1.6 repeat units) against which reads are mapped.
1054 Purple box indicates the large hairpin sequence in IR1. Latent transcript exons (W1, W2, C1, C2,
1055 Y1, Y2) are indicated, with exon W0 being the exon initiated at the Wp promoter, which lies
1056 within IR1.

1057 **Figure 2. Characterisation of strains with deletions affecting IR1.** (A) Read depth analysis.
1058 Graph showing the read depth ratio at adjacent nucleotides from X50-7 reads mapped against the
1059 prototype EBV genome NC_007605 on the Y axis against EBV genome position on the X-axis.
1060 Scheme above graph shows positions of genomic features. Numbers in the graph show positions

1061 for each RDA signal peak, those in bold are edges of the X50-7 deletion. **(B)** Schematic
1062 representation of deletions that include parts of IR1. Black box indicates the position of the
1063 deletion; the dotted box indicates uncertainty over the position of the deletion, as the junction lies
1064 within a repetitive region. Positional details of these deletions are found in table 1. For scale,
1065 note that there are 3072 bp between identical features in the repeat.

1066 **Figure 3. Polymorphisms in IR1, BWRF1 and EBNA-LP. (A)** The distribution of
1067 polymorphisms (SNPs found in more than 1 strain) in IR1 is shown: SNP counts in a 100bp
1068 sliding window are against the position in the IR1 template. In the schematic, arrows indicate
1069 positions of Exons/ORFs; dark vertical lines represent edges of IR1 repeat units. **(B)**
1070 Visualisation of all ORFs of over 600nt in length across the BWRF1 region of IR1. Since there
1071 are no AUG start codons in most of the BWRF1, ORFs are defined at their maximum possible
1072 length (i.e. maximum distance between STOP codons). The frequency of each model
1073 (percentage) is shown to the right. **(C)** The amino acid sequence variants of the EBNA-LP
1074 protein encoded by the viruses are shown. All sequences are shown relative to the B95.8
1075 consensus (dots indicating no change). The protein subtype according to the nomenclature
1076 proposed here (commonest subtypes in bold - see table 2) and name of an example strain are
1077 given to the left. The number (out of 74) of unrelated strains encoding each variant is shown to
1078 the right (see supplementary table ST1 for strain-specific information). Numbers at the top
1079 indicate amino-acid position (according to a W1W2Y1Y2 exon structure), with amino acids 26-
1080 49 omitted [gap in sequence] as they are identical in all strains. Exon structure is shown as boxes
1081 below.

1082 **Figure 4. Phylogenetic tree showing the IR1 diversity between strains.** Phylogenetic tree of
1083 the templated sequence with branch lengths representative of degrees of difference. Non-

1084 independent strains, and strains without a complete IR1 are excluded. The scale bar represents
1085 the tree distance corresponding to 1 nucleotide substitution per kb. Geographic origin of samples
1086 is shown by the colour of the strain name (see key). To the right of the tree, for each strain, its
1087 EBNA-LP variant is shown by an alphanumeric (table 2; figure 2C). BWRf1 subtypes are
1088 shown by coloured boxes. The left side of the BWRf1 boxes are colour-coded for the 3 major
1089 groups (1, 2 and 3 – see supplementary table ST2), with the presence of the indels characteristic
1090 of type 2 indicated by a dark border. The right part of the BWRf1 box is a different colour where
1091 subgroups are distinct from the major ones (red to indicate the disrupted BWRf1 ORF in 1X and
1092 2X). Strains with a type 2 EBNA2 are labelled with empty black boxes. Orange boxes indicate
1093 strains with common SNPs in the flanks, with the letter showing which SNP they contain (see
1094 table ST3). A solid orange box represents a strain in which the flank SNP has propagated
1095 throughout IR1.

1096 **Figure 5. Defining minor variants. (A)** Visual representation of the nature of minor variants
1097 (MVs), explaining their identification. Top section shows a repeat region in which one of 5
1098 repeat units contains the minor variant T rather than the consensus G. When reads (horizontal
1099 grey bars) from sequencing this repeat are mapped against the consensus template (the
1100 representation of IR1 containing 1.6 repeats), 80% of reads will be G (black), but 20% will be
1101 the minor variant T (red). **(B)** Defining a threshold for calling minor variants. The graph plots the
1102 read bias (x-axis) against the MV frequency (y-axis) for every potential MV with a frequency of
1103 over 1% across 76 strains. Red points are potential MVs at positions that show an elevated MV
1104 frequency in all strains (error prone positions). Green points are MVs that are seen as SNPs in at
1105 least one strain. All other MVs are represented by blue points. The dashed line represents the cut-
1106 off above which we accepted the MVs as likely to be genuine, based on the equation

1107 $\sqrt{\text{Read}_{\text{Bias}} \times \text{Variant}_{\text{Frequency}}} \geq 7$. Potential MVs that were designated as SNPs in the flanks of IR1 are
1108 excluded.

1109 **Figure 6. Minor variant diversity in spontaneous LCLs compared with tumour-derived**
1110 **samples.** Graphical representations showing the number of minor variants in IR1 repeat region in
1111 each independent strain (defined in methods). Samples are grouped as being either spontaneous
1112 LCLs (including one saliva-derived sequence - blue bars) or tumour-derived (red bars). The
1113 number of MVs is shown on the x-axis. Y axis shows the percentage of strains with that number
1114 of MVs. Graphs display (A) total MVs; (B) MVs that resemble SNPs (Recombination-like
1115 MVs); and (C) MVs that resemble spontaneous mutations (Mutation-like MVs). (D) Pie charts
1116 indicating the types of genetic variations found in EBV IR1 from sLCL and tumour groups. The
1117 type of variation is classified as either homogeneous repeats (yellow), at least one
1118 recombination-like MV (blue) and only mutation-like MVs.

1119 **Figure 7. T-RECs analysis of genetic exchange in IR1 of sLCL-IS2.01 and TRL1.** The
1120 consensus sequence for sLCL-IS2.01 (A) and TRL1 (C) was modified to instead contain all of
1121 the minor variant bases (B, D). These sequences were compared with a set of four viruses (see
1122 key, right of the plots) from different clades on the phylogenetic tree, to identify regions of
1123 similarity using T-RECs. The similarity plots show the percentage identity of the test strain to
1124 each comparator strain within a 200 bp scanning window.

1125 **Figure 8. Bootstrap analysis of genetic exchange in IR1 of sLCL-IS2.01 and TRL1.** The
1126 consensus sequence for sLCL-IS2.01 (A) and TRL1 (C) were modified to contain all of the
1127 minor variant bases (B and D respectively). These two sequences were compared with a set of
1128 four viruses from different clades on the phylogenetic tree, to identify regions of similarity, using
1129 Bootscan. This analysis simulates phylogenetic tree assembly for a 200 bp sliding window, with

the Y axis indicating the percentage of occasions each strain is the nearest neighbour on the tree, with x-axis marking position, illustrated by the IR1 schematic below each plot.

Figure 9. Minor variants in IR1 of the prototype EBV strain B95-8. (A) Graphs showing the MV frequencies across the template sequence of IR1 of B95-8. The sequence variations are shown with the base of the published B95-8 sequence left, and the newly identified variant right, with the Y axis showing the frequency of the newly identified nucleotide. The x-axis shows the position of that variant in the IR1 template multiple sequence alignment (Supplementary data SD1). Note that this means the alternative nucleotide at position 649/3726 is the dominant allele in the HB9 BAC-based genome (right). At other positions, the different percentage frequencies in MVs are due to different numbers of IR1 repeat units between the B95-8 cell line (11.6 repeat units) and the HB9 clone of the B95-8 BAC (6.6 repeats). **(B)** Graphical representation of the B95-8 MVs relative to major features of IR1. For each MV, the left base indicates the published nucleotide, the right indicates the alternative variant, above the numerical position in the first repeat unit of the prototype EBV sequence (NC007605) below. Bases in green indicate those changes that have been observed as SNPs in other strains. The G/T in red indicates the STOP codon in W1 introduced by the MV 'T' at that position. The blue line and Y axis indicates the SNP density (from Figure 3A) across the repeat for comparison. **(C)** Western blot comparing EBNA-LP levels and sizes in LCLs 30 days after infection in four independent donors using either the original B95-8 EBV-BAC (HB9) or a recombinant BAC from which all MVs have been removed (WT^w). The EBNA-LP ladder is from 293 cells transfected with three different sized EBNA-LP variants: the number of W1W2 domains in each band plus any tag is indicated to the right. The blot was stripped and re-probed with gamma tubulin as a loading control (lower panel). **(D)** Flow cytometry traces showing dilution of the cellTrace violet stain on B cells 8 days

1153 after infection, comparing equal cell numbers of WT^w (solid line) with parental HB9 (dotted
1154 line). Y axis shows cell count, X axis shows log₁₀ CellTrace violet fluorescence. The left hand
1155 plot is from CD19^{+ve} (gated) live cells infected as mixed lymphocytes. Right plot is from CD19-
1156 purified B cells infected and grown in the absence of other cells, based on data also published in
1157 (35).

1158

1159 **Table 1: Deletions overlapping the flanks of IR1**

1160

Strain	Upstream (BamC) deletions ^a		Downstream (BamY) deletions ^a	
	Upstream	IR1	IR1	Downstream
X50-7	9479 in oriP rep* region	14428 after exon W0		
L591	11632 in exon C2	12193 50 bp before hairpin		
Cheptages	7777 in oriP FR	14669 in W1-W2 intron	33604 between W2 and hairpin	38301 [†] within IR2
Daudi			33127 W1-W2 intron	40536 oriLyt upstream element
P3HR1			33351 after exon W2	40157 nr BHLF1 start

1161

1162

1163

1164

1165

1166

1167

1168

^a The positions in prototype sequence NC007605 that are equivalent to the first and last bases of the deletion, above the nearest EBV genome feature.

[†] First position within repeat: sequence recurs every 125bp, so precise junction is uncertain

Table 2: Subtypes of EBNA-LP.

Example strain	W1	W2	Y1	Y2	Overall subtype	Number (/74)
B95-8*	1.1	1	1.1	B	1B	13
sLCL-IM1.05*	1.2	1	1.1	B	1B	3
HL05	1.2	1	1.1	B-P	1B ^P	1
GD2	1.1	1	1.1	B-M	1B ^M	1
Jijoye	2	1	1.1	B	2B	2
C666.1 [#]	2	1	1.1	B-M	2B ^M	2
Mutu	2	1	1.1	C	2C	29
sLCL-IS1.18	2	1	1.1	C-I	2C ^I	2
sLCL-IM1.17	2	1	1.1	C-H	2C ^H	1
Akata	2	1	1.2	A	2A	13
sLCL-2.14**	2	1	2.1	Z	2Z	4
sLCL-IS2.01*	2	1	2.2	Z	2Z	3
sLCL-2.16	2	1 ^R	2.2	Z	2 ^R Z	1

1169

1170

1171

1172

1173

1174

Notes:

* Adjacent shaded rows differ in nucleotide sequence but are identical at the amino-acid level.

[#] Includes M81 strain.

[†] Includes prototype type 2 EBV strain AG876.

















