Genome-Wide Analysis of Wild-Type Epstein–Barr Virus Genomes Derived from Healthy Individuals of the 1000 Genomes Project

Gabriel Santpere[1], Fleur Darre[1], Soledad Blanco[2], Antonio Alcami[2], Pablo Villoslada[3], M. Mar Albà[4,5] and Arcadi Navarro[1,5,6,*]

[1]Institut de Biologia Evolutiva (Universitat Pompeu Fabra – CSIC), Barcelona, Spain

[2]Centro de Biologia Molecular Severo Ochoa, CSIC-UAM, Madrid, Spain

[3]Center for Neuroimmunology, Institut d'investigacions Biomèdiques August Pi I Sunyer (IDIBAPS), Barcelona, Spain

[4]Research Programme on Biomedical Informatics, Hospital del Mar Research Institute (IMIM), Universitat Pompeu Fabra (UPF), Barcelona, Spain

[5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

[6]Centre de Regulació Genòmica (CRG) Barcelona, Catalonia, Spain

*Corresponding author: E-mail: arcadi.navarro@upf.edu.

Accepted: March 17, 2014

Data deposition: The DNA sequences have been deposited at GenBank under the accession KF602052–KF602060 and EBV complete genome sequences in supplementary material S9, Supplementary Material online.

Abstract

Most people in the world (~90%) are infected by the Epstein–Barr virus (EBV), which establishes itself permanently in B cells. Infection by EBV is related to a number of diseases including infectious mononucleosis, multiple sclerosis, and different types of cancer. So far, only seven complete EBV strains have been described, all of them coming from donors presenting EBV-related diseases. To perform a detailed comparative genomic analysis of EBV including, for the first time, EBV strains derived from healthy individuals, we reconstructed EBV sequences infecting lymphoblastoid cell lines (LCLs) from the 1000 Genomes Project. As strain B95-8 was used to transform B cells to obtain LCLs, it is always present, but a specific deletion in its genome sets it apart from natural EBV strains. After studying hundreds of individuals, we determined the presence of natural EBV in at least 10 of them and obtained a set of variants specific to wild-type EBV. By mapping the natural EBV reads into the EBV reference genome (NC007605), we constructed nearly complete wild-type viral genomes from three individuals. Adding them to the five disease-derived EBV genomic sequences available in the literature, we performed an in-depth comparative genomic analysis. We found that latency genes harbor more nucleotide diversity than lytic genes and that six out of nine latency-related genes, as well as other genes involved in viral attachment and entry into host cells, packaging, and the capsid, present the molecular signature of accelerated protein evolution rates, suggesting rapid host–parasite coevolution.

Key words: recombination, selection, human herpesvirus 4, EBV, Illumina reads, whole-genome analysis.

Introduction

Human herpesvirus type 4, or Epstein–Barr virus (EBV), is a dsDNA virus with a genome size of around 170 kb. The EBV infects more than 90% of adults worldwide (Chang et al. 2009). Usually, the primary infection occurs early in life and is asymptomatic in children, while in young adults can produce infectious mononucleosis (IM). EBV preferentially infects B cells but can also attach to and enter into epithelial cells (Borza and Hutt-Fletcher 2002). Upon infection, EBV establishes life-long latency in the form of episomes residing in the host cell's nucleus. Only a small subset of the ~90 coding regions of the virus are expressed in latency: six nuclear proteins (*EBNA-1*, *-2*, *-3A*, *-3B*, *-3C*, and *-LP*) and three membrane proteins (*LMP-1*, *-2A*, and *-2B*) (Young and Rickinson 2004).

Infection by EBV has been related to a number of complex diseases including multiple sclerosis (MS) and different types of cancer such as Burkitt's lymphoma (BL), Hodgkins' lymphoma (HL), or nasopharyngeal carcinoma (NPC) and gastric

carcinoma (GC). The prevalence of these diseases varies with geography. For instance, NPC is much more prevalent in Asia, whereas BL is found most commonly in equatorial Africa (Chang et al. 2009); another notorious example is given by MS, which has greater prevalence in higher latitudes in Caucasian populations, while it is extremely rare in the Tropics (Ascherio and Munger 2010). A number of studies have tried to relate genetic variability of EBV strains to the prevalence of these diseases or to broadly defined geographical regions. Such efforts have focused on sequencing genes that play important roles in essential viral processes, such as *BZLF1* (Gutiérrez et al. 2002; Martini et al. 2007), *EBNA-1* (Habeshaw et al. 1999; Brennan et al. 2010; Lorenzetti et al. 2010; Wang, Liu, Xing et al. 2010), *EBNA-2* (Aitken et al. 1994), *EBNA-3A*, *-3B*, and *-3C* (Görzer et al. 2006), *LMP-1* (Edwards et al. 1999, 2004), and *LMP-2a* (Wang, Liu, Jia et al. 2010). However, the few studies that have classified EBV strains using combinations of variants from more than one gene suggest that recombination does take place and that many variant combinations are possible, thus increasing the difficulty of typing EBV strains (Gutiérrez et al. 2000; Chang et al. 2009; Sawada et al. 2011).

The study of the worldwide distribution of the genetic variability of EBV and its potential contributions to the risk of disease would benefit from the analysis of whole EBV genomes. So far, only seven EBV strains have been completely sequenced: the B95-8/Raji strain from an American patient of IM; AG876 and MUTU from African individuals with BL; and the GD1, GD2, HKNPC1, and AKATA strains from Asian individuals with NPC or BL (Baer et al. 1984; Zeng et al. 2005; Dolan et al. 2006; Liu et al. 2011; Kwok et al. 2012; Lin et al. 2013) (supplementary table S1.1, Supplementary Material online). All these strains were derived from diseased individuals.

We set out to perform a comprehensive comparative genomic study of EBV. Rather than limiting ourselves to the published strains, we add new information by reconstructing EBV sequences extracted from lymphoblastoid cell lines (LCLs) belonging to healthy individuals who were fully sequenced within the 1000 Genomes Project (1kG Project) (Durbin et al. 2010). To obtain a steady source of DNA for genetic studies, cell cultures of LCLs were established by transforming each donor's B cells with the B95-8 EBV strain. Thus, that strain is always present in all individual LCL cultures and may have eliminated all traces of endogenous EBV strains. However, B95-8 presents a specific 12-kb deletion in its genome that was spontaneously produced in the laboratory (Skare et al. 1982), allowing us to distinguish between natural and artificial EBV strains. We were able to determine the presence of natural EBV in at least ten individuals. A complete set of natural nucleotide variants was obtained from three of them. Adding these three new sequences to the already known complete EBV genomes, we were able to carry out analyses exploring genome diversity, phylogenetic relationships, and possible recombination events among EBV strains. Finally, we conducted an analysis of rates of protein evolution of viral genes.

Materials and Methods

Origin of Individual LCLs

The samples sequenced within the 1000 Genomes Project came from LCLs established from blood at the nonprofit Coriell Institute for Medical Research. We used sequence reads from the 1kG release 20101123, with 1,103 available individual alignments to the reference human assembly. The analysis shown here is based on the 929 individuals who were sequenced using Illumina technology with paired-end sequenced DNA libraries. We retrieved reads that do not map to the human genome and are labeled as "unmapped" for each individual in the 1kG Project database.

Mapping

Unmapped reads were mapped against the EBV reference genome (NC_007605), a composite of the B95-8 strain and 12 kb from the Raji strain to correct the nonnatural B95-8-specific deletion from position 139,724 to 151,554. Previous to mapping, we masked the reference EBV genome using RepeatMasker (Smit et al. 1996–2010) to remove low-complexity regions. We also masked a large region of repeats between positions 12,001 and 35,355. Mapping was performed using BWA (Li and Durbin 2010) with default parameters. After that, we removed duplicated read-pairs using SAMtools (Li et al. 2009). We kept only paired-reads with both reads mapping uniquely to the EBV genome. We determined INDELs and realigned reads falling around them using GATK (DePristo et al. 2011), and finally, we recalibrated base quality scores of the aligned reads by considering empirical Phred scores at the level of cycle effect (position in read) and dinucleotide mismatch distribution (the effect of the preceding nucleotide).

Variant Calling

We called variants from pileup files produced by SAMtools based on cut-offs at the level of coverage ($\geq$10), phred-score ($\geq$20), and a minimum number of reads supporting a variant ($\geq$2). We filtered out those variants with extreme strand bias and an allele balance lower than 5%. INDELs were called with the same cut-offs using VarScan (Koboldt et al. 2009). Variants calling in the B95-8-specific deletion were refined using VarScan relaxing coverage cut-offs. Only variants present in the nonrepetitive region of EBV were used for subsequent analysis. In addition to the already masked region identified by RepeatMasker, we also ignored variants in specific repeats in the EBV genome reference genome (NC_007605) according to GenBank annotation. The total

length of the evaluated EBV genome was 135,486 bp out of ~172 kb (~79% of the total).

PCR Amplification, Cloning and Sanger Sequencing

hDNA templates NA19114, NA19315, and NA19384 were purchased from Coriell Cell Repositories. Oligonucleotide primer pairs were EBNA-2Fw: 5′-ggatgcctggacacaagag-3′ and EBNA-2Rev: 5′-tgtgctggtgctgctgg-3′, BBLF4Fw: 5′-agacgatgcagggaatgc-3′ and BBLF4Rev: 5′-agaggcgctctcgtccac-3′, LF2Fw: 5′-agccactgaggaagatctgg-3′ and LF2Rev: 5′-gaagcttaccccggaggag-3′. The High Fidelity Taq DNA Polymerase KAPA HiFi (Cultek) was used and PCR amplification conditions comprised an initial denaturation of 3 min at 95 °C, followed by 35 cycles of 98 °C for 20 s, 60 °C for 15 s, and at 72 °C for 15 s with a final extension at 72 °C for 5 min. The PCR products were purified according to the QIAquick Gel Extraction kit instructions (Qiagen). The DNA fragments of interest were ligated into the pGEM-T easy vector at 4 °C overnight according to manufacturer's instructions (pGEM-T Easy cloning kit, Promega) (in the case of the amplified product using *EBNA-2* primers with the NA19114 template, DNA was digested with BspHI and the undigested fragment was selected for ligation). The products were transformed into competent DH5α bacteria. Positive recombinants were selected on LB-ampicillin/IPTG/X-gal plates and plasmid DNA was purified from overnight culture using QIAGEN Miniprep kit (Qiagen). The presence of the fragment of interest was verified by RsaI and EcoRI restriction digestion. Positive DNA plasmids were sequenced by Macrogen, Inc.

Analysis and Representation

Individual LCLs EBV genome sequences were obtained by incorporating called SNVs into the reference EBV sequence. Most statistical analyses were performed using R packages. Manipulation of sequence intervals was carried out with IRanges (Pages et al.). Principal component analysis (PCA) was performed using the R package prcomp (R Development Core Team 2012), on SNVs without scaling variables to avoid overcontribution of rare variants (McVean 2009). Significance level of all principal components (PCs) was computed calculating the Tracy-Widom statistic to each PC eigenvalue (Johnstone 2001; Patterson et al. 2006). We estimated putative recombination breakpoints among EBV strains using Recco (Maydt and Lengauer 2006). Each recombination block was compared pairwise measuring genetic identity with the seqinr R package (Charif and Lobry 2007). Plots were performed using the ggplot2 package (Wickham 2009) and Circos software (Connors et al. 2009).

Phylogenetic trees were constructed using Phyml (Guindon et al. 2010) with the command line: *phyml –i OUR_SEQs.phy -d nt -b 1000 -m K80 -v e -c 1 -a e -o n -s NNI* with 1,000 bootstrap. We performed d*N*/d*S* analysis using model 0 in codeml from PAML (Yang 2007), which gives one ω ratio for all lineages and considering the global tree resulting from the alignment of the whole EBV sequences. The NSites parameter, which sets variation between sites, was used to test one ω ratio (NSites = 0) or different ωs values between codons (NSites = 7, which allows ω between 0 and 1; and NSites = 8, which allows ω to be greater than 1). Different models were compared as usual in these cases. First NSites 0 versus NSites 0 fixing ω at 0.22 or 1 were compared with a log likelihood test with one degree of freedom. NSites 7 versus 8 was compared with a log likelihood test with two degrees of freedom. Robustness of ω estimates were assessed by producing all possible topologies with eight final branches using PAUP* 4.0b10 (Swofford 2003) and reestimating ω in each of them.

Finally, alignment visualization tools included IGV viewer (Thorvaldsdóttir et al. 2013) and Geneious Pro 5.6.3.

Results

Identification of Natural EBV Strains

To evaluate the presence of natural EBV strains, we measured coverage in the region corresponding to the 12-kb B95-8 deletion and compared it with the average coverage in the rest of the masked viral genome (127,219 bp). Overall EBV coverage varied greatly among the analyzed individuals, ranging from 0× to 1,622.45×. The mean and median coverage were 111.9× and 65.9×, respectively. We required that each strain presented a minimum median coverage of 20× to be included in the analysis, which left a set of 784 individuals. The average median coverage for them was 123.3×, with 90% of the individuals presenting a coverage >20× in at least 99.9% of the genome (supplementary material S2, Supplementary Material online).

Uniquely mapping reads (mapped only once) in the B95-8 deletion indicate the presence of wild-type EBV. However, the B95-8 deletion contains two repeat regions (IR4 and DRright) that could generate false positives. Thus, coverage was calculated only for the single-copy region immediately after these repeats (7 kb between positions 144,445 and 151,554). The results suggested the presence of natural strains in some LCLs (fig. 1). The median coverage in the deletion region in the 784 individuals was 0.002×. A total of 61 LCLs showed no uniquely mapping reads at the deletion coupled with an overall median EBV coverage of at least 20×, suggesting that they are exclusively infected by the transforming strain. Traces of wild-type EBV were detected in 711 LCLs, but in these cases mean coverage in the deletion region was <1×, not high enough for variant calling. Finally, 12 individuals (1.53%) presented a mean coverage greater than 1× in the B95-8 deletion, with five of them reaching a coverage >10×. We set a threshold for mean deletion coverage plus one standard deviation (2.1×) to bring any individual to further analysis. This left a set of ten individuals with detectable natural EBV levels.

All ten naturally infected LCLs are shown in table 1 ordered according to the proportion of natural EBV coverage, as estimated by the ratio between average coverage present in the B95-8-specific deletion and overall EBV average coverage.

Five out of the ten cell lines contained a proportion of the natural virus that is around or below the 5% allele balance cut-off that was used to call variants (see Materials and Methods). However, in the five LCLs with higher natural EBV load, the proportion of natural EBV ranged from ~16% to ~80%. These were four Kenyans (LWK, individuals NA19384, NA19380, NA19315, and NA19474) and one Yoruba (YRI, individual NA19114). For these individuals, we estimated the expected proportion of their EBV genomes covered by natural EBV, assuming that coverage follows a Poisson distribution. The probability of having all EBV genome covered by at least two reads of the natural strain is given by

$$P(X \geq 2) = \left(1 - e^{-\lambda} - \lambda e^{-\lambda}\right)^{\kappa}$$

where λ is the average coverage per nucleotide at the B95-8 deletion and κ the total length of the viral genome in base pair. This simple calculation suggests that, with a 95–100% probability, all endogenous viral genome is covered by at least two natural EBV reads in individuals NA19114, NA19315, and NA19384, with the figure dropping to 84.5% and 11.5%, respectively, for individuals NA19380 and NA19474. In conclusion, natural EBV can be detected in at least three cell lines with reliability high enough for variant calling. Of course, this is a probabilistic approach, and even if probabilities are quite high it is not possible to have absolute guarantee that there are no regions with no coverage from the natural genome.

Measuring Nucleotide Diversity

All evidence indicates that variability levels in the transforming strain B95-8 are extremely low. Only 279 positions were putatively polymorphic among the 61 LCLs, which were infected solely with the transforming strain. Out of these 279 variants, only 7 showed an allele balance for the alternative allele greater than 0.8, suggesting that they are sequencing artifacts. Moreover, all 279 variants, but one, were present in a single individual. Five individuals share this common variant, which is found at position 70,269 in a region with no annotated functional elements. In summary, each of these 61 individuals is infected by a B95-8 strain which is nearly identical to the reference. Interestingly, this allows for an estimation of

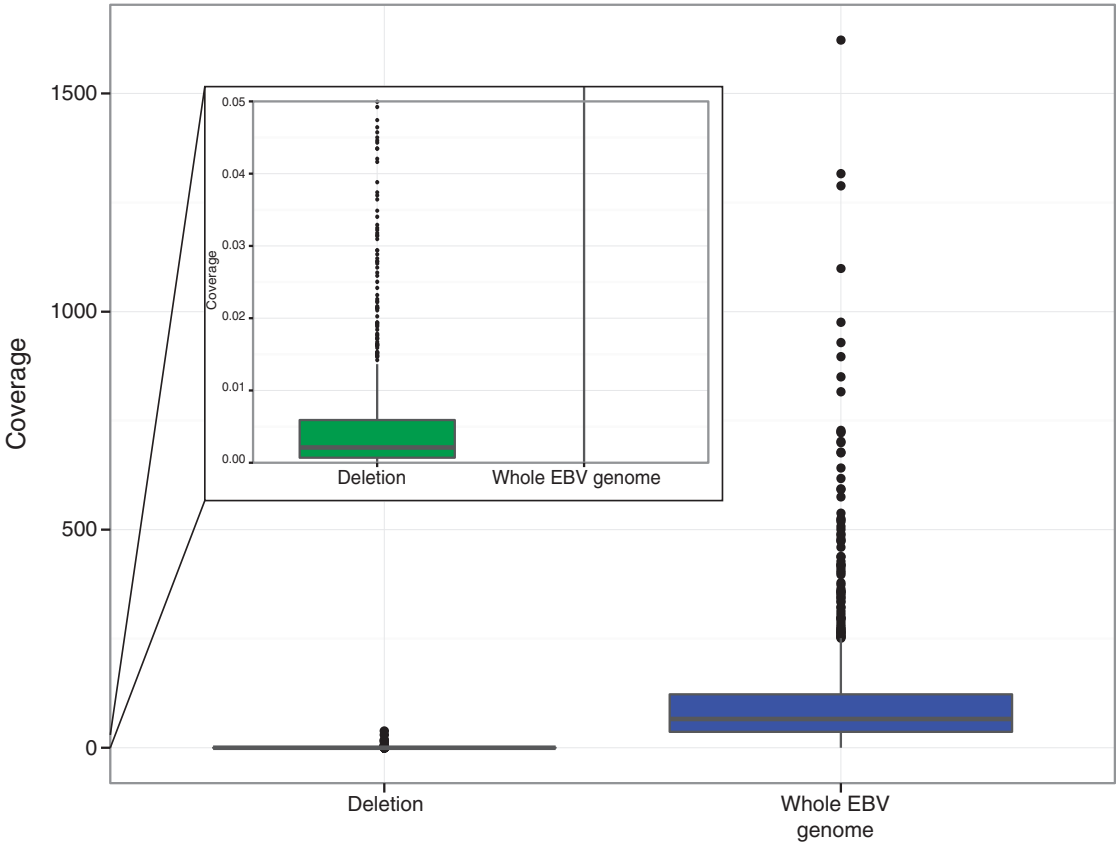

Fig. 1.—Boxplot of the coverage for all LCLs in the whole EBV genome (right) and in the B95-8-specific deletion (left). The inner panel displays a zoom in the coverage scale that shows that coverage at B95-8 is nonzero, suggesting the presence of natural EBV in some LCLs.

Table 1

Basic Statistics for the Ten Naturally Infected Individuals

IND	Origin	Mean EBV	Mean Coverage in Deletion	% of Natural EBV	No. of Variants	Mean Variants AB (%)	Ts/Tv
NA20783	TSI	277.67	5.91	2.13	11	0.34	1.2
NA18507	YRI	115.22	3.08	2.67	27	0.06	1.455
NA20348	ASW	75.53	3.30	4.36	121	0.07	1.814
NA18923	YRI	68.24	3.18	4.66	77	0.07	2.208
NA20524	TSI	119.84	6.21	5.18	170	0.06	1.698
NA19114	YRI	246.15	38.38	15.59	445	0.17	1.967
NA19474	LWK	56.91	13.67	24.03	602	0.20	1.894
NA19315	LWK	62.84	18.04	28.70	628	0.25	1.881
NA19380	LWK	38.44	16.39	42.65	556	0.35	1.78
NA19384	LWK	36.16	29.56	81.75	576	0.70	1.73

NOTE.—Shadowed rows indicated individuals with a higher endogenous EBV proportion. AB, Allele balance; TSI, Toscani in Italia; ASW, Americans of African ancestry in SW USA; YRI, Yoruba in Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya.

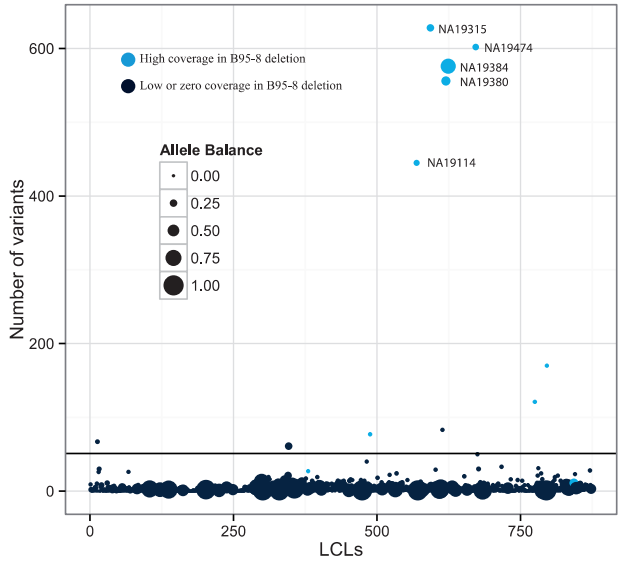

FIG. 2.—Scatter plot where each dot indicates the number of variants called in each LCL. Light blue dots represent LCLs with strong evidence of the presence of natural EBV strains (from ten different individuals). The size of the dot indicates the mean allele balance found in each LCL. Five LCLs clearly show a large amount of genetic variation, as they harbor the greatest proportion of natural EBV.

the amount of artifactual variants expected in individuals that carry the natural strain: given that each one of these 61 LCLs presents five low-quality variants on average, a similar number of false positives should be expected in individuals that also carry variants from the endogenous, wild-type EBV.

In sharp contrast to the lines infected solely with the transforming strain, diversity levels were much higher in all the LCLs showing high coverage at the B95-8 deletion. Allele balances were also higher, indicating that these variants had been confidently called (table 1 and fig. 2). In total, we observed 1,534 single nucleotide variants (SNVs) in the 10 LCLs with the highest coverage at B95-8 deletion. As expected if the endogenous virus contributed them, these are high-quality variants and their allele balance is highly correlated with the proportion of natural EBV reads over total EBV coverage ($R^2 = 0.79$, $P = 0.00056$, see table 1).

Further evidence that the nucleotide variants detected in the 10 LCLs with coverage in the B95-8 deletion are real comes from their transition to transversion ratios (Ti/Tv), see table 1. The Ti/Tv ratio for variants in the 61 LCLs infected solely with the B95-8 strain is quite low (~1), indicating that these variants are mostly sequencing errors. In the 10 individuals with evidence of natural coinfection, the Ti/Tv ratio is around 1.8 (1.77), closer to the ratio of 2, which is expected in nature. For NA19114, NA19315, and NA19384 individuals, with the highest proportion of endogenous EBV, the Ti/Tv ratio in their 1,099 variants is even higher (1.86). Finally, we observed that not a single variant is shared between the two groups of LCLs: the 61 LCLs infected only with B95-8 and the 10 LCLs coinfected with endogenous EBV. The number of variants called and the Ti/Tv ratios for each subset of individuals can be found in supplementary table S1.2, Supplementary Material online. In summary, all these observations, together with the practical absence of variability in the B95-8 strain, indicate that rather than being mutations acquired in vitro, the vast majority of the 1,535 variants are due to the endogenous EBV virus which originally infected the individuals sampled within the 1000 Genomes Project. Therefore, with the exception of the RepeatMasked regions, we can reconstruct the full EBV genome from these three strains. Supplementary material S3, Supplementary Material online, contains a detailed explanation about the effects of applying different cut-offs on the accuracy of our variant calling in our three LCLs.

Experimental Validation of the Presence of Natural EBV Sequences

To prove experimentally the coexistence of different strains in these LCLs and to validate the variants that we had

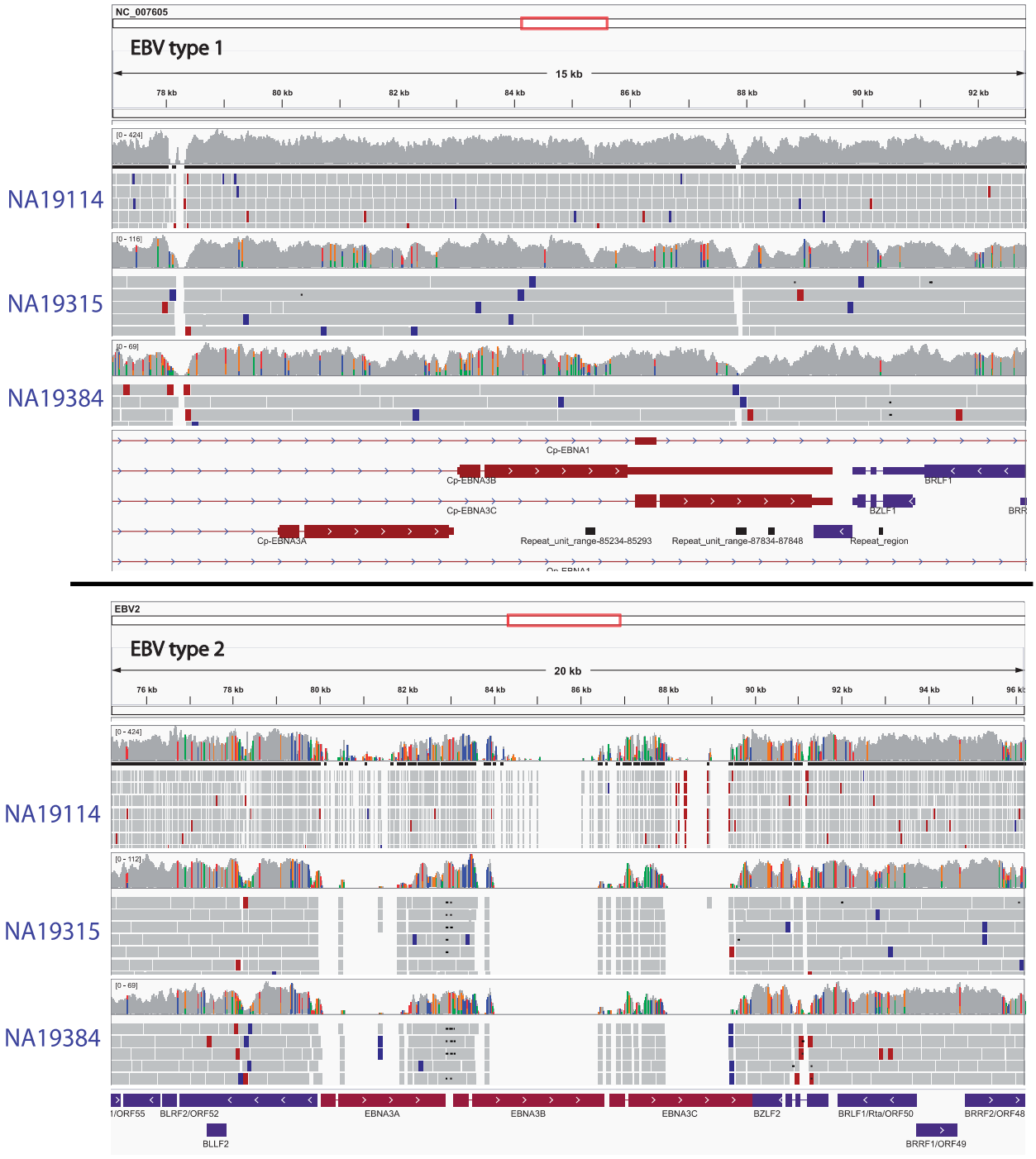

Fig. 3.—Genome Browser snapshot of the reads from the three individuals from which we obtained high-quality EBV genomes aligned against EBV type 1 (top panel) and type 2 (bottom panel) reference genome. The snapshot shows the *EBNA-3*region, which was used to discriminate EBV type 1 from type 2. Large regions with no coverage are observed in the type 2 mappings, demonstrating that all three cell lines are infected only by type 1 EBV strains.

determined, the DNA of the three LCLs for which we had performed high-quality variant calling was obtained from the Coriell Institute. Primers were designed to PCR amplify fragments of the *EBNA-2*, *LF2*, and *BBLF4* genes, spanning regions containing called variants. By cloning and sequencing we obtained and analyzed sequences from these individuals (supplementary fig. S4.1, Supplementary Material online). All the sequences we predict for the natural strain in these individuals are always found and, thus, they are confirmed. Moreover, all of the variants that we had previously identified

in silico for NA19315, NA19384, and NA19114 were present and, crucially, we did not find any new unpredicted variants at a level of confidence that would allow for their calling, which vouches for the completeness of our in silico analysis. Additionally, the transforming strain sequence was detected in fragments of the *EBNA-2* and *BBLF4* genes from all individuals, and it was invariably identical to the B95-8 reference sequence. As expected, *LF2*, a gene located in the B95-8 deletion, only presented natural sequences in all three LCLs. These results confirm both the presence of natural viruses and the high accuracy of our variant calling for individuals NA19315, NA19114, and NA19834.

Classification of All the Available EBV Strains According to Geography

We started by classifying the three genomes we had obtained according to current criteria. To do so, we used a common method of distinguishing EBV types that is based on the fact that nucleotide differences in the *EBNA-2* and *EBNA-3* genes are much higher between than within types, with large regions being exclusive of one or other EBV type. Mapping EBV reads from the diagnostic region *EBNA-3A-3B-3C* of the three called individuals against the EBV type 2 reference genome (fig. 3) shows no uniquely mapping reads in the large regions exclusive of the EBV type 2 version of *EBNA-3* genes, indicating that all natural EBV genomes present in these individuals are type 1. Many substrains of EBV type 1 have been defined by polymorphisms at genes such as *EBNA-1*, *EBNA-2*, *LMP-1*, *LMP-2a*, or *BZLF1*. From such studies, some particular substitutions are found predominantly in different regions of the world. All the polymorphism from the three strains reported here are consistent with the known African origin of the donor individuals. (For a detailed analysis of polymorphisms in these four genes, see supplementary material S5, Supplementary Material online.)

To take advantage from full-genome information in classifying EBV genomes, we performed a PCA including the reference sequence plus all fully sequenced EBV strains that are available in literature and presented low counts of ambiguous nucleotides (Lin et al. 2013): AKATA, GD1, MUTU, and AG876 (fig. 4 and supplementary table S1.1, Supplementary Material online). The first PC explains around 37% of variance ($P = 0.004$) and separates type 1 from type 2 strains. The second component explains around 27% of the variance ($P = 0.014$) and separates Asian strains from African and reference strains.

Phylogenetic Relationships between EBV Strains

Constructing a phylogenetic tree with the eight high-quality EBV sequences (three from the present study and five previously available best-quality sequences from diseased individuals) confirms and expands the PCA results presented earlier. It is immediately obvious from figure 5*a* that the AG876 branch is longer than any other, indicating greater divergence, that NA19315 shares a common ancestor with MUTU, and that all African plus B95-8/Raji strains are clearly separated from the two Asian strains (AKATA and GD1). We repeated the phylogenetic tree using only the 12 kb in the B95-8 deletion (fig. 5*b*). Asian strains still appear separated from the rest, NA19315 and MUTU remain in the same clade, but the

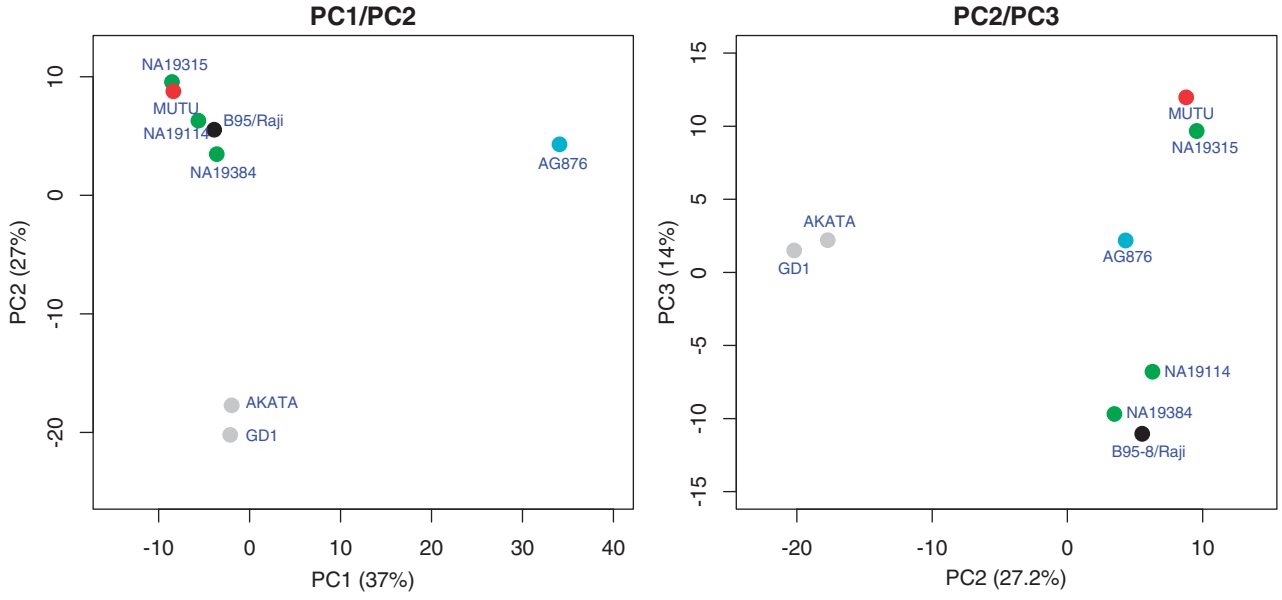

Fig. 4.—PCA using all the SNVs in eight complete EBV strains. PC1 discriminates between type-1 and type-2 strains, while PC2 separates African from Asian strains. A PC3 (marginally significant) does not place strains across a clear geographical axis, as NA18384 and MUTU have Kenyan origin. NA19384, NA19315, and NA19114 in green; MUTU in red, GD1 and AKATA in gray; type 2 EBV in blue; and type 1 reference genome in black.

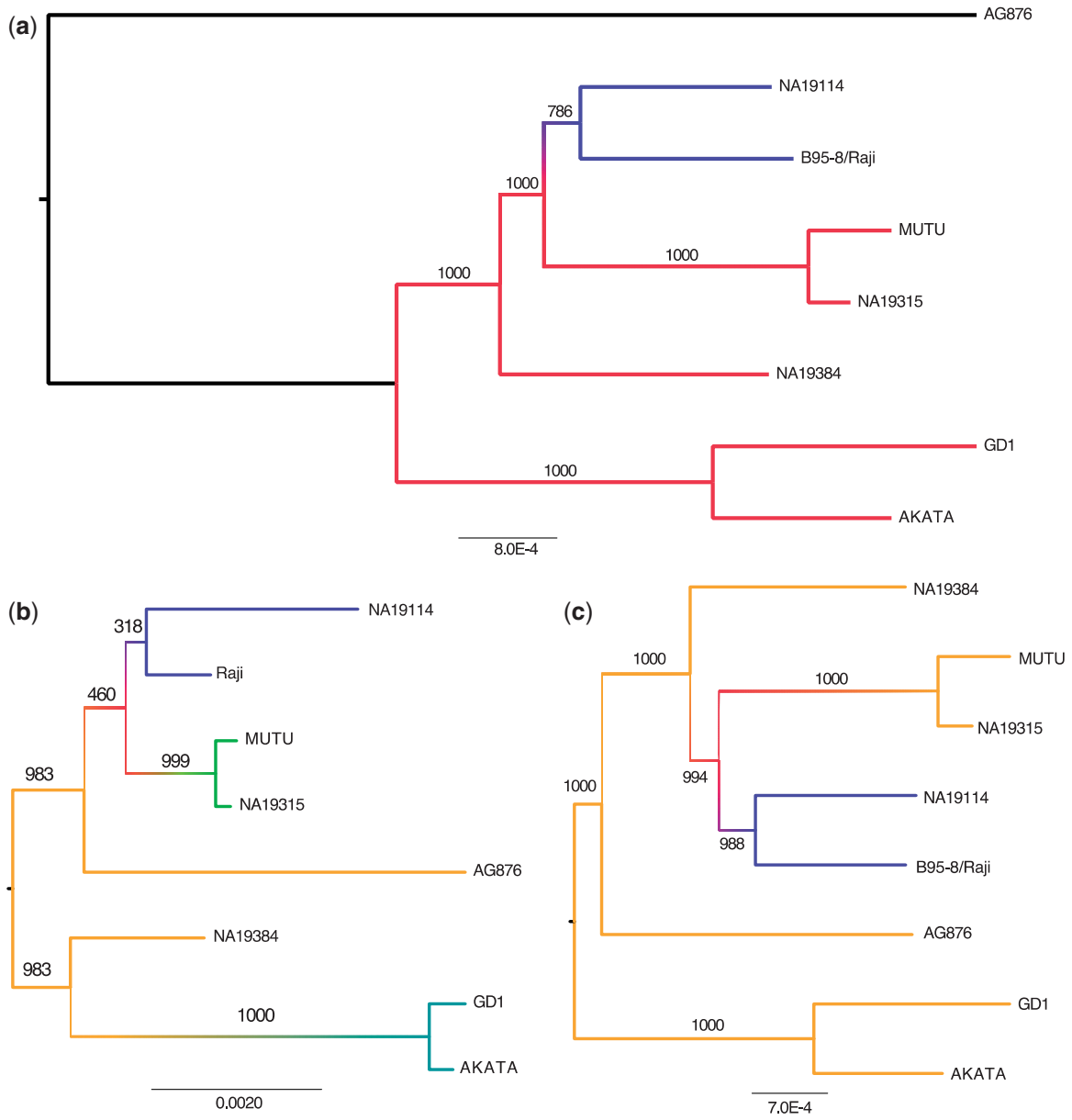

Fig. 5.—Phylogenetic tree representing genetic distance between our three EBV strains and published sequences. Numbers in internal branches indicate the consistency of the branch out of 1,000 bootstraps. Color gradient represents bootstrap support. Trees are rooted using the midpoint. (*a*) Phylogenetic relationships using the whole EBV genome. (*b*) Using only the B95-8-specific deletion, which in the reference EBV genome corresponds to a Raji EBV strain fragment from Nigeria. (*c*) Excluding latency genes, which contain most intertypic EBV variation.

classification of NA19384, NA19114, AG876, and Raji fragment is unclear, probably due to lack of statistical power. However, NA19114 and the Raji fragment still group together. It is noteworthy that the AG876 branch is now much shorter and becomes integrated with most African strains.

Finally, we reconstructed the tree excluding latency genes, where most of intertypic variation is found (Dolan et al. 2006), and observed (fig. 5*c*) that the AG876 branch is again shorter and all African strains are clearly separated from Asians. However, the tree from figure 5*b* shows that NA19384, from Kenya, shares a most recent common ancestor with Asian strains. These findings constitute clear proof that different genomic regions present different genealogies. The most likely cause is, of course, intertypic and intratypic recombination events.

Recombination between EBV Genomes

To try reconstructing the recombinational history of all the available EBV genome sequences, we studied levels of identity

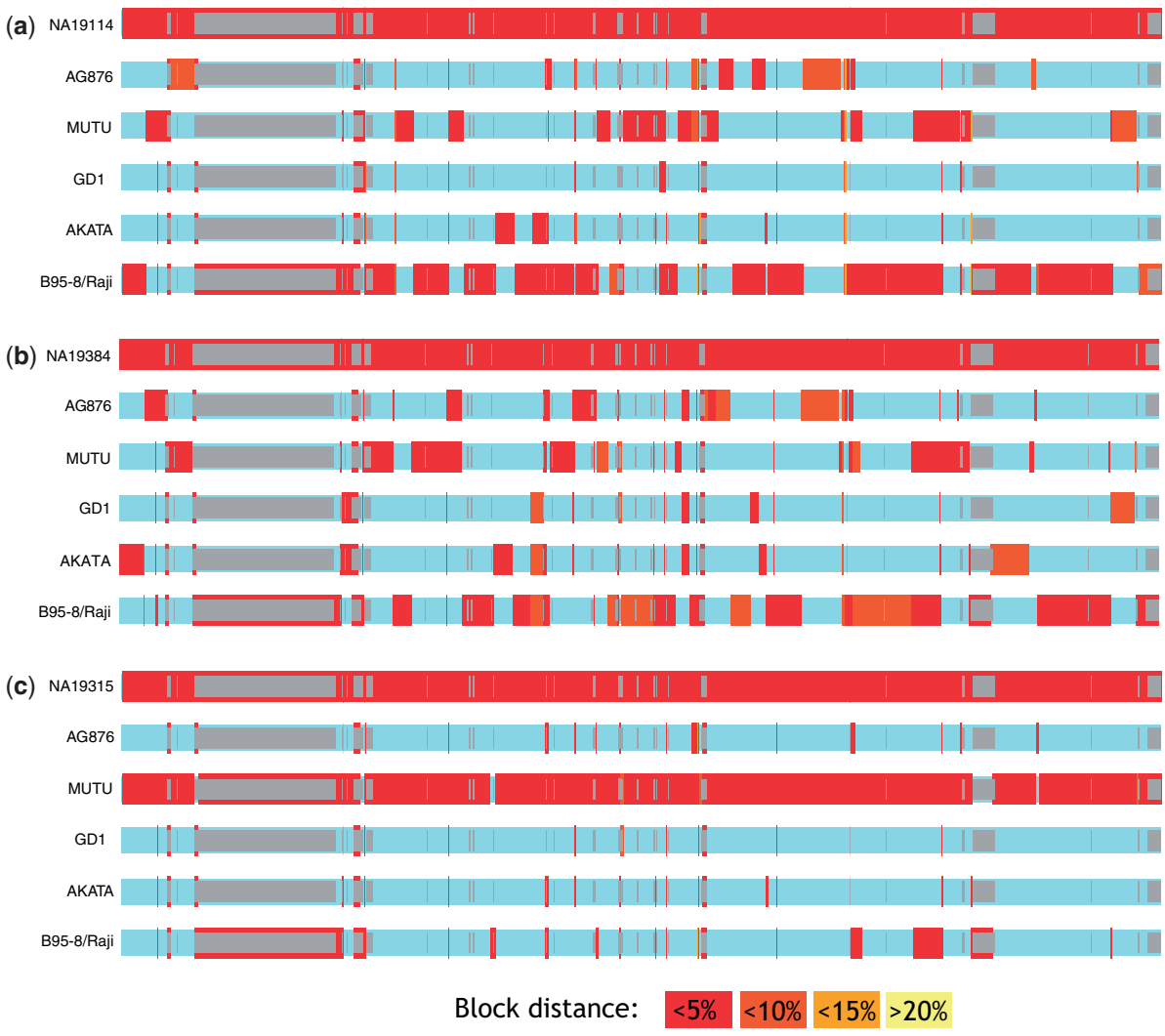

Fig. 6.—Recco analysis on the three African EBV strains. We consider one sequence at a time to be the recombinant product of the rest of sequences. Gray regions are masked due to repeats. The scale color is proportional to genetic distance based on identity, red indicates short distance, and yellow indicates large distance. The first row is the one considered the recombinant product in all panels.

amongst sequences. We used Recco (Maydt and Lengauer 2006) to infer putative recombination breakpoints considering each of our three genomes (one at a time) as the recombinant product of the five good quality strains published so far. Then, we determined "haplotype" blocks as pieces of sequence between any two breakpoints identified in any EBV genome. Finally, we calculated an identity distance matrix for each one of these blocks. Figure 6*a* and *b* shows that the EBV genomes from NA19114 and NA19384 share haplotype blocks mainly with B95-8/Raji and MUTU strains, with fewer blocks closer to the Asian strains (GD1 and AKATA); while NA19315 shows almost 100% block identity with MUTU (fig. 6*c*). The same analysis performed on AG876 indicates that its *EBNA-2* and *EBNA-3* blocks are divided among all other genomes and show very low sequence identity (supplementary fig. S6.1*f*, Supplementary Material online). The rest of the EBV type-2 genome is divided in haplotype blocks distributed evenly among the Asian and B95-8/Raji strains. The genome of the B95-8/Raji strain appears as a composite of MUTU and AG876 haplotypes plus a few AKATA and GD1 blocks (supplementary fig. S6.1*a*, Supplementary Material online). AKATA and GD1 strains can be reciprocally explained (supplementary fig. S6.1*d* and e, Supplementary Material online).

The Genomic Distribution of EBV Nucleotide Variants

As usual, different classes of genes present varying levels of genetic diversity. We calculated Watterson's estimate of θ, the parameter of neutral molecular evolution (Watterson 1975)

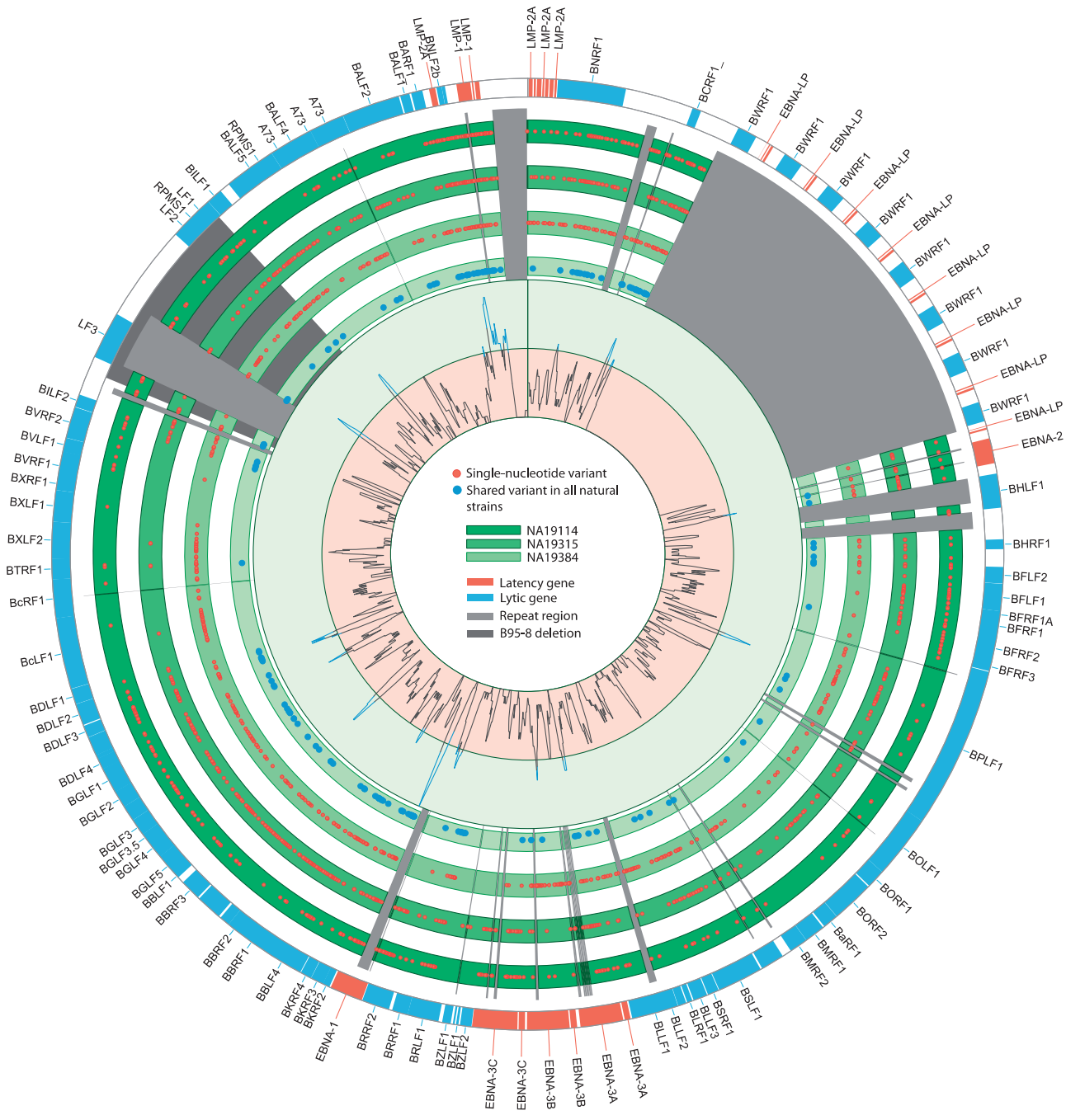

Fig. 7.—Circos plot showing genome-wide single-nucleotide variant diversity in the three LCLs with the highest natural EBV load. Each green circle corresponds to one individual. Red dots indicate variable positions. Blue dots indicate variants present in all three EBV sequences. The innermost circle shows diversity on 500-bp windows considering all variants together. In the outer circle, latency genes (red) and lytic genes (blue) are represented. Light gray shadows indicate repetitive regions, and dark gray sector indicates the location of the B95-8-specific deletion.

using SNVs called in all coding regions and separately for latency and lytic genes, as well as for UTRs and introns (supplementary table S1.3, Supplementary Material online, and fig. 7). Latency genes harbor more diversity than any other element in the genome and they double the amount of diversity found in the rest of coding elements. This trend may be indicative of very different evolutionary pressures in the two classes of genes.

As for genomic elements, they also present different levels of variation (supplementary tables S1.3 and S1.4, Supplementary Material online, and a more detailed description in supplementary material S5, Supplementary Material

online). Focusing on gene-disrupting mutations, we found a mutation introducing a premature stop codon in the *RPMS1* gene of the NA19315 sequence (nucleotide C160T, residue Q54X). Although the *RPSM1* ORF can be translated in vitro, the RPMS1 protein has never been detected in natural infections (Al-Mozaini et al. 2009). Finally, two different frameshift deletions were found in the *LF3* coding region for NA19315 and NA19384. Both mutations create a new frame that produces a slightly longer version of LF3 version (supplementary material S5, Supplementary Material online). It has been shown that the *LF3* coding frame is not conserved in AKATA, and many start and stop codons in the 5′ of the gene in both AKATA and MUTU suggest a primary noncoding function for *LF3* (Lin et al. 2013).

Rates of Protein Evolution

Given that different genes present different diversity levels, it makes sense to evaluate whether they are evolving at different rates. The most widely used method to study rates of protein evolution is to estimate the ratio between dN, the rate of nonsynonymous substitutions along a branch of the phylogenetic tree, and the corresponding dS, the rate of synonymous substitutions. Using the multiple sequence alignments of our three EBV sequences coming from healthy individuals plus the five known high-quality EBV strains, we calculated ω (the dN/dS ratio), for all coding regions by means of PAML's codeml (Yang 2007) and using the global tree. Supplementary table S1.4, Supplementary Material online, shows the average ω value for all genes and for two different groups of genes: those expressed during the lytic phase and those expressed during latency. The average ω value for lytic genes (0.22) indicates that most EBV genes are under similar constraints than average human genes (~0.25) (Kosiol et al. 2008). In contrast, ω is significantly increased in latency genes, with an average of 0.84 (Wilcoxon test *P* value 0.000015) (supplementary table S1.4, Supplementary Material online).

We also studied dN/dS at the individual gene level in order to identify particular genes under strong purifying or positive selection. In general terms, genes that are very conserved at the protein sequence level tend to be considered under strong purifying selection, while, in contrast, genes with relatively larger rates of nonsynonymous than synonymous changes may have achieved that through the effect of adaptive evolution). We had estimated ω values for each gene and obtained likelihoods for every ω value under the M0 model, considering one ω ratio for the whole tree. We then fixed ω at 0.22 (the average ω value in lytic genes) obtaining a second likelihood value, always under Model 0. Finally, we performed likelihood ratio tests (LRTs) to determine what were the genes whose ω values were significantly different from 0.22. We performed multiple test correction with a FDR *q*-value cut-off of 5% (supplementary material S7, Supplementary Material online).

A total of 11 genes are under strong purifying selection, as indicated by their significantly lower ω values. Five of them are related to DNA replication and nucleotide metabolism, including *BALF5* (the catalytic subunit of the viral DNA pol), *BALF2* (part of the replication fork machinery), *BBLF4* (helicase), *BGLF4* (kinase involved in many functions required for efficient lytic DNA replication), and *BORF2* (a ribonucleosidereductase). One of the most conserved genes, *BcLF1*, encodes the major capsid protein (MCP), which self-assembles in pentamers and hexamers to form the virus capsid. Three other very constrained genes are *BALF3*, *BGRF1*, and *BBRF1*, which are involved in DNA packaging inside the capsid during the lytic phase. Interestingly, the other two genes are related to tegument and cell attachment (*BSRF1*) and/or interaction with host proteins (LF2, which interacts with IRF7).

We then repeated the codeml Model 0 analysis fixing ω to 1 (the expected value of ω under neutrality) and compared with the same model with free ω. Only *EBNA-1*, which encodes for essential proteins for the establishment and maintenance of EBV latency, presented a ω value significantly greater than one ($\omega = 1.78$). This constitutes evidence that the *EBNA-1* gene has been under positive selection.

We found six nonlatency genes to be accelerated relative to average ω values found in lytic genes. *BRRF2* and *BFRF2* encode for two viral tegument proteins with unknown functions. *BRRF2* is known to be expressed and translated because peptides BRRF2 were abundantly detected in LCLs (Johannsen et al. 2004). BFRF2 has only clear evidence at transcript level (Concha et al. 2012). BARF0 has evidence at transcript level but may not be translated in vivo. Finally and more interestingly, we found *BKRF2*, *BLLF1*, and *BLLF2* to be accelerated. *BKRF2* encodes for the envelope glycoprotein L required for the fusion of viral and plasma membranes for a virus to be able to enter into the host cell and interact with the host's integrins (Connolly et al. 2011). *BLLF2* ORF comprised *BLLF1* coding region. *BLLF1* encodes for gp350, which initiates the virus attachment to the host's cells via interaction with B cells CR2 (Young et al. 2008).

One potential problem can arise from forcing a single tree topology produced by whole-genome alignment to genes that have different genealogies, as we did in the calculations above, because assuming the wrong tree may lead to misestimation of ω. To control for that source of error, we estimated ω under all possible tree topologies produced by eight sequences (supplementary material S8, Supplementary Material online). In general, ω estimates for latency genes are robust to different tree topologies. In particular, ω is significantly higher than 1 ($P < 0.05$) in 88.3% of the possible tree topologies of *EBNA-1*. Similarly robust ω values are observed also in lytic genes *BLLF1* and *BRRF2*. However, ω estimates for *BLLF2*, *BFRF2*, and *BKRF2* seem to be more affected by tree topology, making these measures less reliable (supplementary material S8, Supplementary Material online).

It is possible that, even if a gene as a whole is under strong purifying selection, some of its codons or groups of codons have been the targets of positive selection. In order to test specifically for this hypothesis while controlling for contribution of a relaxation of selective constraints, we performed another codeml analysis that compares two different models that allow variation of ω values between codons. The first model (M8) allows a subset of codons to have ω greater than one. That model is usually tested against a null model that does not permit any codons to have ω greater than 1 (M7). Even when using a 5% *q*-value cut-off, the LTRs showed that M8 explained our data better than M7 in eight of the latency genes: *EBNA-1*, *-3A*, *-3B*, -3C, -LP, *LMP-1*, and *LMP-2A*, which seem to have undergone accelerated evolution under the influence of positive selection. *BLLF1*, *BLLF2*, *BARF0*, *BFRF2*, and *BRRF2* but not BKRF2 also show this signature at codon level. In addition, 20 other genes whose ω values are not particularly high and, thus, did not show up in the previous analyses also bear the signature of adaptive evolution (supplementary material S7, Supplementary Material online). Two examples are *BNRF1*, which encodes the major tegument protein (MTP) involved in virus entry, attachment, and membrane fusion, and *BVRF2*, encoding one capsid protein and *BFRF1A* and *BFLF1*, which are involved in packaging; the later being also responsible for placing capsids in the nuclear replication compartments. Finally, we detected signatures of positive selection at some codons in genes whose complete sequence appears particularly constrained such as *BALF2* and *BALF3*, indicating putative functional sites.

Discussion

We have carried out the most complete comparative genomics study of EBV so far. To do so, we combined published EBV genomes from diseased donors with our own survey of sequence data from the 1kG Project, which provided EBV genomes from healthy individuals. In total, we deeply analyzed eight EBV genome sequences.

One major obstacle that we had to overcome to carry out our study was the presence of the B95-8 transforming strain in most of the 1kG Project LCLs (~92%). However, we could obtain clear evidence of co-occurring natural EBV strains in 10 out of 946 individuals. Given the high prevalence of EBV infection, the number of individuals in which we observe co-occurrence of natural EBV may seem surprisingly low. To understand that apparent paradox, we need to consider that in healthy individuals (such as those from the 1kG Project), EBV persists in peripheral blood B cells at a frequency of about 1–50 infected cells per million B cells (Babcock et al. 1998). Given that around 2 million B cells are contained in 1 ml of blood, we would expect to find low counts of infected lymphocytes per blood extraction. Moreover, only ~3% of the peripheral leukocytes of adult donors become immortalized by EBV (Sugden and Mark 1977). This clearly reduces the chances of a B95-8 transformation of an already naturally infected B cell. In addition, although initially EBV-immortalized B-cell lines are polyclonal, after prolonged culture in vitro, cell lines become oligoclonal or monoclonal (Ryan et al. 2006). Except for a small subset of immune-compromised individuals, most EBV-associated tumors and LCLs contain monoclonal EBV (Gulley et al. 1994). This monoclonality may be indicative either of genetic drift or of natural selection among cell clones. At any rate, the most common scenario is that a few clones quickly become dominant (Ryan et al. 2006), reducing the probability of finding clones infected with the natural EBV relative to clones infected only with the B95-8 strain.

Nevertheless, it is likely that by increasing sequencing coverage in 1kG Project samples, it would be possible to detect more LCLs with some wild-type EBV. Actually, a recent study provides analyses of publicly available RNA-seq data from 319 LCLs derived from 143 individuals (Arvey et al. 2012), some of them overlapping our set of samples. By looking at their published RPKM values (Reads per Kilo Base per Million) of the LF3 gene, which is located in the deletion of the B95-8 transforming strain, we can deduce that at least 76 LCLs (~50%) contain some load of natural EBV strains.

Infections by multiple EBV strains are most commonly observed in immunocompromised individuals (Tierney et al. 2006), but also happen in healthy individuals (Lung et al. 1988; Walling et al. 2003). Regarding only the two major EBV types, rates of coinfection of EBV type 1 and type 2 range from 0% to 53% (Kunimoto et al. 1992; Srivastava et al. 2000; Walling et al. 2003). High frequency of coinfection with both type 1 and type 2 EBV has been reported in MS patients (Santón et al. 2011). Given that we do not find evidence of coinfection by different natural EBV types, it is worth noting that we might have underestimated intraindividual EBV diversity if our variants came only from a dominant isolate or the LCLs that were sequenced within the 1kG Project were the result of a selection bias toward transformation-competent EBV isolates (Tierney et al. 2006). By amplifying and cloning fragments of EBV of our three LCLs with strong evidence of natural coinfection, we have been able to prove the existence and accuracy of our predicted natural strains, as well as the existence of major dominant isolates of natural EBV coexisting in these studied LCLs.

Are the Variants Determined Really Natural?

Several facts support that the variants we have studied were wild-type variants present in the B cells of 1kG Project individuals before the establishment of LCLs. The strongest piece of evidence is that LCLs with no sign of coinfection by natural EBV contain very low levels of polymorphism, which given their low Ti/Tv ratio and extreme values of allele balance are likely to be sequencing errors. In this subset of 61 LCLs, we observed an average of only five low-quality variants per LCL, so it is unlikely that B95-8 variants are biasing our results. In

fact, the opposite is more likely: we cannot rule out that traces of natural EBV could account for some of the variants that we attribute to B95-8. Cloning and Sanger sequencing of fragments of EBV provide conclusive evidence that the B98-5 version of these fragments is identical to the published reference sequence. In short, we have determined that the B95-8 strain possesses a very stable genome that does not present any relevant contribution to the pool of variants that we determined in our three natural strains. The same experiment also validated the confidence of our variant calling in the evaluated EBV fragments. However, an inherent limitation in our data remains. Given the contribution of reads of both natural and transforming strain, however the small the later may be, it is not possible to guarantee the total absence of false-negative variant calls. As explained in supplementary material S3, Supplementary Material online, this is particularly true for NA19315, but we estimate that even in the worst case scenario in which all very low frequency variants were true positives, in NA19315 we would still have a ~94% of sensitivity.

Phylogeny and Recombination

Phylogeny is not always consistent in different regions of the EBV genome, allowing the inference of recombination events, as previously reported by McGeoch and Gatherer (2007). By studying genomic blocks delimited by predicted recombination breakpoints, we observed that current haplotypes are mixed inter- and intratypically. EBV type2 (AG876) is in fact a composite of high-identity blocks already present in all known type 1 strains. It has been proposed that the EBV type 2 strains come from a type 1 prototype strain that incorporated *EBNA-2* and *EBNA-3* blocks from a closely related lymphocryptovirus (McGeoch 2001). Given the crucial role of these genes in latency establishment, it has been hypothesized that such acquisitions increased the virus' fitness as EBNA-2 and EBNA-3 blocks from type 1 and type 2 were maintained in the present combinations relative to a possible recombinant version between type 1 and type 2 (McGeoch 2001). Another possibility would be that intensive substitution processes in those genes were driven by positive selection (McGeoch 2001). Having said that, inter-typic recombinants showing recombination breakpoints in the *EBNA-3* gene block have been reported in healthy and immunecompromised individuals from different geographical regions (Midgley et al. 2000, 2003; Görzer et al. 2006). During productive EBV replication, intrastrain homologous and nonhomologous recombination events can occur and have been observed in repetitive regions of the genome (Walling et al. 1992; Walling and Raab-Traub 1994; Walling et al. 1999). Thus, our results warrant caution in classifying EBV strains: recombination events may confound relatively simple classifications based on single-gene polymorphisms such as those found in the *EBNA-1* gene.

Rates of Protein Evolution in EBV Genes

We analyzed SNV and INDEL polymorphism in our three strains in relation to the B95-8/Raji reference genome and observed that diversity found in latency related genes is the highest compared with any other genomic element. EBV genome comparison between simian homologs of EBV had revealed that some latency genes diverged faster (Wang et al. 2001). Confirming and expanding these results, we found that diversity in latency genes doubles the level of nucleotide variation found in lytic genes and is even higher than diversity found in introns. We calculated ω values, studied the signatures of selection for all coding regions in EBV, and compared different selection models using a LRT. We acknowledge the well-known fact that accuracy and power of the LRT for detecting selection on genes depend on the number of sequences, with power being 100% at ~15 sequences (Anisimova et al. 2001). By adding our three new sequences to five already reported EBV genomes, we increase the power to detect positive selection acting on genes; however, we still might be underpowered for some low-divergence genes. In particular, the limitations of site-prediction methods are greater when assessing protein evolution rates at the codon level (Nozawa et al. 2009) and thus it is clear that experimental validation of putatively selected residues would be needed.

Our analysis provides evidence for the action of both purifying and positive selection. The average ω value of lytic genes resembles the average of ~0.25 found in primates (Kosiol et al. 2008), indicating similar evolutionary constraints in viral and host genes. We reported evidence for strong purifying selection acting upon 11 EBV genes. These include genes encoding proteins that work in crucial enzymatic activities such as the viral DNA polymerase (BALF5) or the helicase function (BBLF4). One host-interacting element has also been the target of purifying selection: LF2. LF2 binds the central inhibitory association domain of IRF7, inhibiting the interferon-mediated immunity, which is the major defense raised by the host against viral infections (Wu et al. 2009).

Latency genes, in contrast, present higher ω values, indicating that positive selection may be the cause of their high diversity levels. Using various models we observed that *EBNA-1*, *-3A*, *-3B*, -3C, -LP, *LMP-1*, and *LMP-2A* contain a subset of codons, which have experienced positive selection. The action of positive selection inferred from an increased proportion of nonsynonymous variation has already been described in *LMP-1* and *EBNA-1* (McGeoch and Davison 1999; Walling et al. 1999). In addition to these genes, other coding regions related to virus attachment and entry in host cells (*BLLF1* and *BKRF2*) show signatures of accelerated protein evolution rates. *BRRF2* and *BFRF2* present a higher ω value and, together with *BNRF1* and three other capsid and packaging-related genes, also show evidence of positive selection at codon level. *BNRF1* encodes for the MTP. The BRRF2 protein has been abundantly detected in LCLs by means of mass spectrometry (Johannsen

et al. 2004), and it is likely that both constitute viral tegument proteins. As these kinds of proteins can regulate viral attachment and immune evasion of viruses, evidence of selection in *BLLF1*, *BKRF2*, *MTP*, *BRRF2*, and *BFRF2* suggests that knowledge may be gained from a detailed analysis of genetic variation in these genes in different populations and pathology groups and from the study of functional implications of these variants in the biology and pathogenesis of EBV.

Supplementary Material

Supplementary materials S1–S9 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Belen Lorente-Galdos for generously reading and improving the manuscript and Amelia Hunter for correcting the manuscript. They also thank Damien de Vienne for his help in the phylogenetic analysis. This work was supported by the Spanish Multiple Sclerosis Network (REEM), of the Instituto de Salud Carlos III (RD07/0060 and RD12/0032/0011) to A.N., A.A., and P.V.; by the Spanish Government Grants BFU2009-13409-C02-02 and BFU2012-38236 to A.N.; and by FEDER.

Literature Cited

Aitken C, Sengupta SK, Aedes C, Moss DJ, Sculley TB. 1994. Heterogeneity within the Epstein-Barr virus nuclear antigen 2 gene in different strains of Epstein-Barr virus. J Gen Virol. 75(Pt 1):95–100.

Al-Mozaini M, et al. 2009. Epstein-Barr virus BART gene expression. J Gen Virol 90:307–316.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 18:1585–15892.

Arvey A, et al. 2012. An atlas of the Epstein-Barr virus transcriptome and epigenome reveals host-virus regulatory interactions. Cell Host Microbe. 12:233–245.

Ascherio A, Munger KL. 2010. Epstein-barr virus infection and multiple sclerosis: a review. J Neuroimmune Pharmacol. 5:271–277.

Babcock GJ, Decker LL, Volk M, Thorley-Lawson DA. 1998. EBV persistence in memory B cells in vivo. Immunity 9:395–404.

Baer R, et al. 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310:207–211.

Borza CM, Hutt-Fletcher LM. 2002. Alternate replication in B cells and epithelial cells switches tropism of Epstein-Barr virus. Nat Med. 8: 594–599.

Brennan RM, et al. 2010. Strains of Epstein-Barr virus infecting multiple sclerosis patients. Mult Scler. 16:643–651.

Chang CM, Yu KJ, Mbulaiteye SM, Hildesheim A, Bhatia K. 2009. The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: a need for reappraisal. Virus Res. 143:209–221.

Charif D, Lobry JR. 2007. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations. New York: Springer Verlag. p. 207–232.

Concha M, et al. 2012. Identification of new viral genes and transcript isoforms during Epstein-Barr virus reactivation using RNA-Seq. J Virol. 86:1458–1467.

Connolly SA, Jackson JO, Jardetzky TS, Longnecker R. 2011. Fusing structure and function: a structural view of the herpesvirus entry machinery. Nat Rev Microbiol. 9:369–381.

Connors J, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43: 491–498.

Dolan A, Addison C, Gatherer D, Davison AJ, McGeoch DJ. 2006. The genome of Epstein-Barr virus type 2 strain AG876. Virology 350: 164–170.

Durbin RM, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Edwards RH, Seillier-Moiseiwitsch F, Raab-Traub N. 1999. Signature amino acid changes in latent membrane protein 1 distinguish Epstein-Barr virus strains. Virology 261:79–95.

Edwards RH, Sitki-green D, Moore DT, Raab-traub N. 2004. Potential selection of LMP1 variants in nasopharyngeal carcinoma. J Virol. 78: 868–881.

Görzer I, Niesters HGM, Cornelissen JJ, Puchhammer-Stöckl E. 2006. Characterization of Epstein-Barr virus Type I variants based on linked polymorphism among EBNA3A, -3B, and -3C genes. Virus Res. 118: 105–114.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Gulley BML, et al. 1994. Epstein-Barr virus DNA is abundant and monoclonal in the reed-sternberg and cells of Hodgkin's disease: association with mixed cellularity subtype Hispanic American ethnicity. Blood 83: 1595–1602.

Gutiérrez MI, et al. 2000. Association of EBV strains, defined by multiple loci analyses, in non-Hodgkin lymphomas and reactive tissues from HIV positive and HIV negative patients. Leuk Lymphoma. 37:425–429.

Gutiérrez MI, et al. 2002. Discrete alterations in the BZLF1 promoter in tumor and non-tumor-associated Epstein-Barr virus. J Natl Cancer Inst. 94:1757–1763.

Habeshaw G, Yao QY, Bell AI, Morton D, Rickinson AB. 1999. Epstein-barr virus nuclear antigen 1 sequences in endemic and sporadic Burkitt's lymphoma reflect virus strains prevalent in different geographic areas. J Virol. 73:965–975.

Johannsen E, et al. 2004. Proteins of purified Epstein-Barr virus. Proc Natl Acad Sci U S A. 101:16286–16291.

Johnstone IM. 2001. On the distribution of the largest eigenvalue in principal component analysis. Ann Stat. 29:295–327.

Koboldt DC, et al. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25: 2283–2285.

Kosiol C, et al. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4:e1000144.

Kunimoto M, Tamura S, Tabata T, Yoshie O. 1992. One-step typing of Epstein-Barr virus by polymerase chain reaction: predominance of type 1 virus in Japan. J Gen Virol. 73(Pt 2):455–461.

Kwok H, et al. 2012. Genomic sequencing and comparative analysis of Epstein-Barr virus genome isolated from primary nasopharyngeal carcinoma biopsy. PLoS One 7:e36939.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Lin Z, et al. 2013. Whole genome sequencing of the Akata and Mutu Epstein-Barr virus (EBV) strains. J Virol. 87:1172–1182.

Liu P, et al. 2011. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. J Virol. 85: 11291–11299.
Lorenzetti MA, et al. 2010. EBNA1 sequences in Argentinean pediatric acute and latent Epstein-Barr virus infection reflect circulation of novel South American variants. J Med Virol. 82:1730–1738.
Lung MLI, Chang RS, Jones JH. 1988. Genetic polymorphism of natural Epstein-Barr virus isolates from infectious. J Virol. 62:3862–3866.
Martini M, et al. 2007. Characterization of variants in the promoter of EBV gene BZLF1 in normal donors, HIV-positive patients and in AIDS-related lymphomas. J Infect. 54:298–306.
Maydt J, Lengauer T. 2006. Recco: recombination analysis using cost optimization. Bioinformatics 22:1064–1071.
McGeoch D, Davison AJ. 1999. Chapter 17—The molecular evolutionary history of the herpesviruses. In: Domingo E, Webster R, Holland J, editors. Origin and evolution of viruses. London: Academic Press. p. 441–465.
McGeoch DJ. 2001. Molecular evolution of the gamma-Herpesvirinae. Philos Trans R Soc Lond B Biol Sci. 356:421–435.
McGeoch DJ, Gatherer D. 2007. Lineage structures in the genome sequences of three Epstein-Barr virus strains. Virology 359:1–5.
McVean G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5:e1000686.
Midgley RS, Bell AI, Mcgeoch DJ, Rickinson AB. 2003. Latent gene sequencing reveals familial relationships among Chinese Epstein-Barr virus strains and evidence for positive selection of A11 epitope changes. J Virol. 77:11517–11530.
Midgley RS, et al. 2000. Novel intertypic recombinants of Epstein-Barr virus in the Chinese population. J Virol. 74:1544–1548.
Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A. 106:6700–6705.
Pages H, Aboyoun P, Lawrence M. IRanges: infrastructure for manipulating intervals on sequences. R package version 1.4.2 [data unknown].
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.
R Development Core Team. 2012. R: a language and environment for statistical computing. Available from: http://www.r-project.org/.
Ryan JL, et al. 2006. Clonal evolution of lymphoblastoid cell lines. Lab Invest. 86:1193–1200.
Santón A, et al. 2011. High frequency of co-infection by Epstein-Barr virus types 1 and 2 in patients with multiple sclerosis. Mult Scler. 17: 1295–1300.
Sawada A, et al. 2011. Epstein-Barr virus latent gene sequences as geographical markers of viral origin: unique EBNA3 gene signatures identify Japanese viruses as distinct members of the Asian virus family. J Gen Virol. 92:1032–1043.
Skare J, Edson C, Farley J, Strominger JL. 1982. The B95-8 isolate of Epstein-Barr virus arose from an isolate with a standard genome. J Virol. 44:1088–1091.
Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker. Available from: http://repeatmasker.org.
Srivastava G, Wong KY, Chiang AKS, Lam KY, Tao Q. 2000. Coinfection of multiple strains of Epstein-Barr virus in immunocompetent normal individuals: reassessment of the viral carrier state. Blood 95: 2443–2445.
Sugden B, Mark W. 1977. Clonal transformation of adult human leukocytes by Epstein-Barr virus. J Virol. 23:503–508.
Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.
Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178–192.
Tierney RJ, et al. 2006. Multiple Epstein-Barr virus strains in patients with infectious mononucleosis: comparison of ex vivo samples with in vitro isolates by use of heteroduplex tracking assays. J Infect Dis. 193: 287–297.
Walling DM, Brown AL, Etienne W, Keitel WA, Ling PD. 2003. Multiple Epstein-Barr virus infections in healthy individuals. J Virol. 77: 6546–6550.
Walling DM, Raab-Traub N. 1994. Epstein-Barr virus intrastrain recombination in oral hairy leukoplakia. J Virol. 68:7909–7917.
Walling DM, et al. 1992. Coinfection with multiple strains of the Epstein-Barr virus in human immunodeficiency virus-associated hairy leukoplakia. Proc Natl Acad Sci U S A. 89:6560–6564.
Walling DM, et al. 1999. The molecular epidemiology and evolution of Epstein-Barr virus: sequence variation and genetic recombination in the latent membrane protein-1 gene. J Infect Dis. 179:763–774.
Wang F, Rivailler P, Rao P, Cho Y. 2001. Simian homologues of Epstein-Barr virus. Philos Trans R Soc Lond B Biol Sci. 356:489–497.
Wang X, Liu X, Jia Y, et al. 2010. Widespread sequence variation in the Epstein-Barr virus latent membrane protein 2A gene among northern Chinese isolates. J Gen Virol. 91:2564–2573.
Wang Y, Liu X, Xing X, et al. 2010. Variations of Epstein-Barr virus nuclear antigen 1 gene in gastric carcinomas and nasopharyngeal carcinomas from Northern China. Virus Res. 147:258–264.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 276: 256–276.
Wickham H. 2009. ggplot2: elegant graphics for data analysis. New York: Springer. Available from: http://had.co.nz/ggplot2/book.
Wu L, et al. 2009. Epstein-Barr virus LF2: an antagonist to type I interferon. J Virol. 83:1140–1146.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Young KA, Herbert AP, Barlow PN, Holers VM, Hannan JP. 2008. Molecular basis of the interaction between complement receptor type 2 (CR2/CD21) and Epstein-Barr virus glycoprotein gp350. J Virol. 82:11217–11227.
Young LS, Rickinson AB. 2004. Epstein-Barr virus: 40 years on. Nat Rev Cancer. 4:757–768.
Zeng M-S, et al. 2005. Genomic sequence analysis of Epstein-Barr virus strain GD1 from a nasopharyngeal carcinoma patient. J Virol. 79: 15323–15330.

Associate editor: Ya-Ping Zhang