Available online at www.sciencedirect.com

ScienceDirect

Virology 359 (2007) 1–5

VIROLOGY

www.elsevier.com/locate/yviro

Rapid Communication

Lineage structures in the genome sequences of three Epstein–Barr virus strains

Duncan J. McGeoch *, Derek Gatherer

Medical Research Council Virology Unit, Institute of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, UK

Received 23 August 2006; returned to author for revision 29 September 2006; accepted 3 October 2006
Available online 13 November 2006

Abstract

Whole genome sequences for three Epstein–Barr virus strains (B95-8, GD1 and AG876) were aligned and compared. In addition to known variable loci (including type-specific alleles for the EBNA2, EBNA3A, EBNA3B and EBNA3C genes, plus the EBNA1 and LMP1 genes), seven large-scale regions of lower-level diversity were identified with strains at each in two major groupings. All three possible patterns of strain associations were represented across the seven loci. Tree-building studies supported the existence of two distinct lineages in each case, and occurrence of recombination between lineages therefore has to be invoked to account for the observed genotypes of virus strains.

Keywords: Herpesviridae; Gammaherpesvirinae; Lymphocryptovirus; Human herpesvirus 4; Viral evolution; Recombination

Introduction

Epstein–Barr virus (EBV) is highly prevalent in human populations worldwide and is associated with both non-malignant diseases and a number of cancers, including Burkitt's lymphoma, nasopharyngeal carcinoma and Hodgkin disease (Kieff and Rickinson, 2001). EBV belongs to the family *Herpesviridae* (subfamily *Gammaherpesvirinae*, genus *Lymphocryptovirus*). While most of EBV's genome is highly conserved among strains, 6 of the 80 protein-coding genes show striking and distinctive patterns of diversity (genomic locations of these six genes are indicated in Fig. 1A). All six are active in the latent cycle with roles in establishment and maintenance of the virus in B lymphocytes. Four of the variable genes define two 'types' of the virus, with diverged alleles in EBV type 1 and EBV type 2. Of these four, three form a contiguous genomic block, encoding proteins EBNA3A, EBNA3B and EBNA3C, while the fourth, encoding EBNA2, is at a separate genomic site. The two other notably variable genes encode proteins EBNA1 and LMP1, and diversity in these does not correlate with type 1 or type 2 status. The 172 kb genome sequence of EBV B95-8, a type 1 strain, was reported by Baer et al. (1984), and this stood for 21 years as the sole EBV sequence. Recently, however, the sequence of a second type 1 strain, GD1, was reported (Zeng et al., 2005), and our group has published the sequence of the type 2 strain AG876 (Dolan et al., 2006). B95-8 was isolated from a North American case of infectious mononucleosis (Miller and Lipman, 1973), GD1 from a Chinese case of nasopharyngeal carcinoma (Zeng et al., 2005) and AG876 from a West African case of Burkitt's lymphoma (Pizzo et al., 1978). Here we describe a genome-wide comparison of these three sequences, which has resulted in a significantly enhanced view of genomic relationships among EBV isolates.

Results

The analyses presented in this paper are primarily based on single nucleotide polymorphisms (SNPs). Insertion–deletion polymorphisms (indels) were found to be of lesser value for comparative exercises, being few in number and most prominently located at families of short repeat sequences, at least some of which are known to vary rapidly in their copy numbers. Indels not associated with repeat families were typically compatible with the SNP-based analyses.

The three EBV genome sequences were aligned using CLUSTAL W to give an overall alignment length of 173,848 residues, including gapping characters. Fig. 1B presents SNP

* Corresponding author. Fax: +44 141 337 2236.
E-mail address: d.mcgeoch@mrcvu.gla.ac.uk (D.J. McGeoch).

0042-6822/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2006.10.009

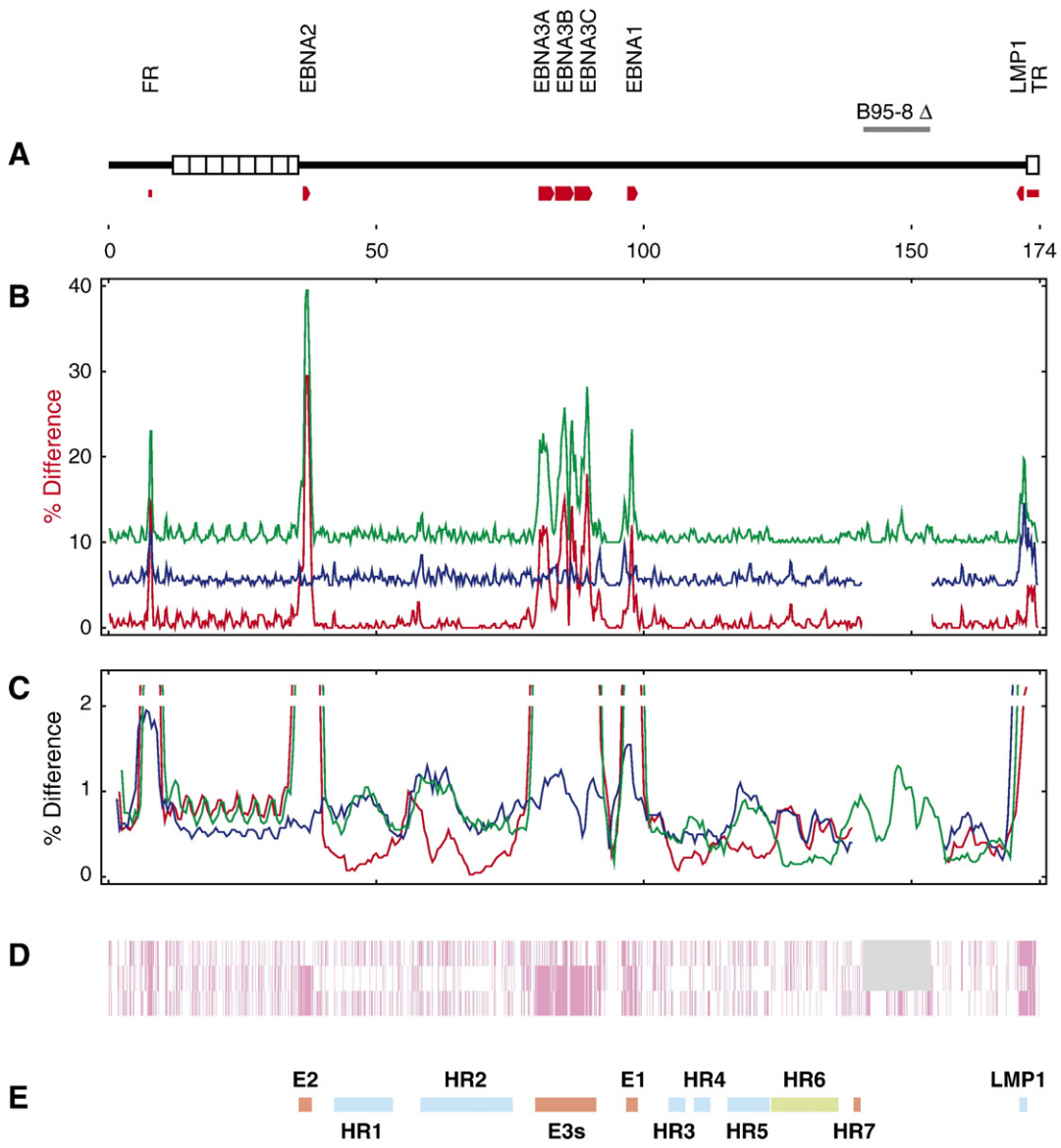

Fig. 1. Comparisons of three aligned EBV genomic sequences. Aspects of the genome sequences are presented in the five aligned panels. Panel A represents the EBV genome, with unique sequence as a heavy line and major internal repeats and terminal repeats as open boxes. Locations of genes and other elements mentioned in the text are indicated in red and annotated. The extent of the B95-8-specific deletion is indicated (see Materials and methods). Genome sequence numbers (kb) are shown. Panel B displays pairwise scans between aligned sequences, made by GCG program PLOTSIMILARITY and showing percentage substitution differences in a moving 400-nuc window; B95-8 v AG876 in red, B95-8 v GD1 in blue and AG876 v GD1 in green. To improve visibility, the blue and green plots are offset on the *y*-axis by 5 and 10% points respectively. Panel C shows an equivalent plot, with a 4000-nuc window and expanded *y*-axis scale. Here the blue and green plots have been given small offsets on the *x*-axis, negative and positive respectively. Panel D shows positions of SNP differences between pairs of sequences; B95-8 v GD1 in the top line, B95-8 v AG876 in the middle line and AG876 v GD1 at the foot. The positions affected by the B95-8-specific deletion are grayed out. Panel E shows the locations of diverged loci, including haplotype regions (HR1 to HR7), EBNA genes (abbreviated as E1, E2 and E3s) and LMP1 gene. Locations are colored according to associations of the sequences: pink, AG876 distinct; blue, GD1 distinct; and green, B95-8 distinct.

differences between genome pairs, as plotted over 400-nuc windows. As was previously seen in the comparison of B95-8 and AG876 (Dolan et al., 2006), the genomes appeared as very close except for the type-specific divergences at the EBNA2, EBNA3A, EBNA3B and EBNA3C loci, and also divergence at the EBNA1 and LMP1 loci, the family of repeats (FR) in the plasmid origin of DNA replication (*oriP*) and the terminal repeats (TRs); the diversity among strains of all these features is well known. However, by increasing the sensitivity of the scan with a larger window (of 4000 nuc) and expanded divergence scale, novel lower level patterns of similarity and dissimilarity were revealed which vary along the genome outside the strongly diverged loci (Fig. 1C). Fig. 1D presents an alternative view of variation with plots of positions of individual SNPs between pairs of sequences. Prominent novel diverged segments were taken as containing at least 10 defining SNPs, and their extremities were located conservatively by inspection of the sequence alignment. Seven such regions were identified, ranging in size from 1.3 kb to 17.2 kb, and designated as haplotype regions 1 to 7 (HR1 to HR7; see Fig. 1E). The sizes

Table 1
EBV haplotype regions

	Total length [a]	Compacted length [b]	SNPs, B95-8 distinct	SNPs, GD1 distinct	SNPs, AG876 distinct	SNPs, all distinct
EBNA2	2445	2175	7	5	382	3
HR1	10,948	10,944	10	80	8	0
HR2	17,245	17,189	26	128	13	0
EBNA3s	11,362	10,322	47	51	847	8
EBNA1	2167	2125	13	14	91	0
HR3	3137	3136	1	15	0	0
HR4	3109	3109	2	15	1	0
HR5	6914	6914	12	56	6	0
HR6	12,580	12,580	69	13	9	0
HR7	1327	1323	0	0	11	0
LMP1	1408	1407	5	83	6	2

[a] Residue length of alignment, including gapping introduced during the alignment process.
[b] Residue length of alignment after removal of any position with a gapping character.

and incidences of SNPs in all these variable regions are listed in Table 1. In the leftmost segments, HR1 and HR2, lying between the EBNA2 and EBNA3 loci, the B95-8 (type 1) and AG876 (type 2) sequences are very closely similar but are clearly distinct from that of GD1 (type 1). To the right of the EBNA3 and EBNA1 loci, HR3, HR4 and HR5 follow the same pattern, but in HR6 B95-8 is distinct from the other two, and in HR7 AG876 is distinct. The region labeled as LMP1 in Fig. 1E is 1.4 kb and contains about 60% of the LMP1 gene, with the 3′ portion of LMP1 coding sequence lying outside the left boundary of this haplotype locus. In the labeled LMP1 locus, the GD1 sequence is markedly distinct from those of B95-8 and AG876. We discounted for further analysis two variable regions which contain only repetitive sequences and were regarded as prone to rapid change, namely the TRs and the FR part of *oriP*.

We hypothesized that the three patterns of strain associations seen across HR1 to HR7, plus the EBNA1, LMP1, EBNA2 and EBNA3 regions, represent the occurrence of genomic segments that have arisen from different viral lineages, with recombinational mixing. In order to investigate this hypothesis, we wished to ascertain, based on the SNP distributions, whether one of the possible phylogenetic trees for each region was by statistical criteria clearly superior to the others. No suitable outgroup sequence was available to root the trees: the closest sequenced relative of EBV is rhesus lymphocryptovirus, which we judged to be too distant relative to the divergences among the EBV strains to act as a robustly informative outgroup for this purpose. With three EBV sequences, there is only one unrooted tree for any haplotype region, and it is necessary to assume the operation of a molecular clock in order to locate the root locus (on the longest branch). Each dataset was therefore tested for molecular clock compatibility by the method of Tajima (1993), which examines differences between the two shorter branches, with transitions and transversions treated separately. In no case was the hypothesis of a molecular clock rejected at the 5% significance level. The statistical tests of Felsenstein (1985), based on maximum parsimony, were then applied, and in every case, one tree was identified as better than the next best at well over the 95% confidence level. Furthermore, for each region, all possible maximum likelihood trees with global molecular clock were computed (Yang, 1997) and the RELL test of Kishino and Hasegawa (1989) applied. Again, in every case, a single tree emerged as highly supported, with RELL bootstrap proportion equal to or greater than 0.99, and in no instance was the tree with trifurcation at the root scored as top. Fig. 2 depicts the best such tree for four representative haplotype loci. Thus, at each of the segments examined, two major lineages were distinguished, and all three possible classes of such pairs of major lineages were present. In order to account for this pattern in the sequenced genomes, we must invoke the occurrence of homologous recombination events between EBV strains with distinct patterns of segment lineages.

Inferring possible recombinational histories for loci in the sequenced genomes on the basis of the observed SNP patterns is an exercise in model building, guided by the principle of parsimony. We proceeded by minimizing first the number of ancestral lineages invoked and then the number of crossover events needed to account for the observations. Since strains AG876 and GD1 show the greatest number of differing segment states, we treated them as representing non-recombinant and distinct lineages at all variable loci except HR6. Strain B95-8 is then GD1-like in EBNA2, the EBNA3s, EBNA1 and HR7, but is AG876-like in segments HR1, HR2, HR3, HR4, HR5 and LMP1. In segment HR6, B95-8 is distinct from the other two, so that either AG876 or GD1 was taken as recombinant at this locus. The inferred possible recombinational events needed to generate the three isolates are depicted in Fig. 3. Treating the genome as its circular episome form, six crossovers in B95-8 and two in either AG876 or GD1 were invoked to generate the observed genomic patterns from two ancestral lineages. It is noteworthy that, outside of the type-specific EBNA2 and EBNA3 loci, only the small (1.3 kb) HR7 region and the EBNA1 locus correlate with the ‘type’ designations, and this

Fig. 2. Trees for representative haplotype regions. The top-scoring trees are shown for HR2, HR5, HR6 and HR7, as computed by maximum likelihood under a global molecular clock. Divergence scale (substitutions per site) is at the foot.

Fig. 3. Modeling recombination events in histories of sequenced EBV genomes. Results are shown to account for lineage associations observed at sites in each EBV sequence, under a two-lineage model. Lineages at each site are marked as black or gray and invoked crossovers indicated by X. Panels A and B present unresolved alternative histories, in which either GD1 or AG876 has experienced recombination at HR6. Crossovers at the right end of the B95-8 lines are included to take account of the genome's circular episomal state in latently infected cells.

association between EBNA1 variant and type-specificity does not persist in comparisons of multiple isolates (e.g. MacKenzie et al., 1999). Our modeling treated the type-specific loci and the EBNA1 and LMP1 loci equivalently to the seven HR loci, but clearly additional factors are needed to explain the behavior of these highly diverse regions. At least some of these latent-cycle genes appear to be subject to forces of positive selection for amino acid variation (Midgley et al., 2003; Burrows et al., 2004; our unpublished data), and it seems likely that the EBNA1 and LMP1 loci have been evolving more rapidly than other genomic regions.

We emphasize that this analysis is based on large-scale trends in the isolates' sequences: not all of the genome was accounted for and we did not pursue assigning divergence patterns to smaller regions that showed local trends in differences between isolates. There is no assurance from the presently available data that the major lineages seen at each locus are in fact equivalent to major lineages at other loci in the sense of possessing a shared history. In principle, five classes of relationship of a given segment among three input sequences could occur, namely the three already discussed that have one strain distinct from the other two plus one with all strains definitely distinct and one with all strains very close. These latter features can also be discerned in Fig. 1: for instance, there is a region with all three sequences approximately equally diverged, of 4.5 kb centered at 55,950 between HR1 and HR2, and there are substantial invariant tracts, of 2.1 kb centered at residue 94,111 and 1.6 kb centered at 166,913. However, such features are not usefully interpretable for our purpose since we have no criteria in these cases to distinguish lineage structure from intrinsic high or low local variability. In all, our reconstruction of recombinational histories is thus likely to have incorporated a significant degree of simplification.

Discussion

The availability of three EBV genome sequences proved informative in improving our view of relationships among virus strains. We have demonstrated that two lineages can be distinguished at substantial segments of EBV strains' genomes and that differing associations of these lineages among strains at separate segments indicate the occurrence of recombination between different lineages. This is the first picture of recombining lineages at the level of the whole EBV genome, although there was previous evidence that recombination did occur, based on serological typing or local sequencing (e.g. Yao et al., 1996; Midgley et al., 2000). For a recombinant EBV strain to arise, two potential parental viruses must be present in the same cell of a single human host. This appears as a very stringent requirement, and it may be that productive recombination events are rare indeed. We note that the occurrence of recombination processes has been inferred for other human herpesviruses in natural populations (see Bowden et al., 2004).

A prominent mystery of EBV evolution concerns the genesis and relationship of the two EBV types (McGeoch, 2001). If the EBNA2 and EBNA3 genes have been evolving rapidly, why should there be just two major alleles seen for each gene? If the contemporary constellation of allele types arose by recombination between two distinct viruses, has one of these parents remained otherwise unobserved? We have now described the occurrence of seven other regions in EBV genomes, each roughly similar in size to the diverged EBNA2 and EBNA3 loci, that also exhibit two distinct lineages, although with divergences an order of magnitude lower than those between the type-specific alleles. These observations suggest the following hypothesis for the origins of the two EBV types: that, deep in human evolutionary history, two host populations became isolated and EBV developed in them into two distinct lineages; that in both lineages the EBNA2 and EBNA3 genes evolved rapidly relative to other parts of the genomes (for unidentified reasons presumed specific to the genes' functions); that the two host populations eventually regained contact; and that their EBV strains thereafter recombined to give the patterns now observed of both type-specific alleles and other haplotype regions. This hypothesis should be testable with accumulation of sequence data for more virus strains.

The classes of genomic variability observed for EBV now include: (1) large-scale regions of distinct lineage; (2) type 1 and type 2 alleles for the EBNA2 and EBNA3 genes; (3) marked diversity in the LMP1 gene; and (4) similar diversity in the EBNA1 gene. In addition, there are indications that positive selection effects may be operating in at least some of these cases. In all, the evolutionary processes acting on the EBV genome are evidently of considerable variety and complexity, and we are still far from constructing a comprehensive account of their action.

Materials and methods

EBV genome sequences

The current data library entry for the complete genome sequence of EBV B95-8 (Baer et al., 1984) was used, accession number NC_007605. This contains an 11.8 kb segment of another strain (Raji) inserted to make good a deletion specific to

B95-8, and the Raji segment was excluded from our analyses. The sequence of strain GD1 (Zeng et al., 2005) was obtained from accession number AY961628. The sequence of strain AG876, determined by our group (Dolan et al., 2006), has accession number DQ279927.

Computational analyses of sequences

The GCG (Accelrys Inc.) and EMBOSS packages were used. Genome sequences were aligned using CLUSTAL W (Thompson et al., 1994). The method of Tajima (1993) was used for evaluating molecular clock behavior of three-species trees followed by the maximum parsimony method of Felsenstein (1985) to identify best trees. Maximum likelihood trees were computed under global molecular clock for all three-species rooted tree topologies, including the trifurcated tree, using the HKY85 model of sequence substitution and such tree sets were evaluated by the RELL test of Kishino and Hasegawa (1989) (PAML package; Yang, 1997).

Acknowledgment

We thank Andrew Davison for critical review of the manuscript.

References

Baer, R., Bankier, A.T., Biggin, M.D., Deininger, P.I., Farrell, P.J., Gibson, T.G., Hatfull, G., Hudson, G.S., Satchwell, S.C., Seguin, C., Tuffnell, P.S., Barrell, B.G., 1984. DNA sequence and expression of the B95-8 Epstein–Barr virus genome. Nature 310, 207–211.

Bowden, R., Sakaoka, H., Donnelly, P., Ward, R., 2004. High recombination rate in herpes simplex virus type 1 natural populations suggests significant co-infection. Infect. Genet. Evol. 4, 115–123.

Burrows, J.M., Bromham, L., Woolfit, M., Piganeau, G., Tellam, J., Connolly, G., Webb, N., Poulsen, L., Cooper, L., Burrows, S.R., Moss, D.J., Haryana, S.M., Ng, M., Nicholls, J.M., Khanna, R., 2004. Selection pressure-driven evolution of the Epstein–Barr virus-encoded oncogene LMP1 in virus isolates from Southeast Asia. J. Virol. 78, 7131–7137.

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

Felsenstein, J., 1985. Confidence limits on phylogenies with a molecular clock. Syst. Zool. 34, 152–161.

Kieff, E., Rickinson, A.B., 2001. Epstein–Barr virus and its replication. In: Knipe, D.M., Howley, P.M. (Eds.), Fields Virology, vol. 2. Lippincott Williams and Wilkins, Philadelphia, pp. 2511–2573.

Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29, 170–179.

MacKenzie, J., Gray, D., Pinto-Pes, R., Barrezueta, L.F.M., Armstrong, A.A., Alexander, F.A., McGeoch, D.J., Jarrett, R.F., 1999. Analysis of Epstein–Barr virus (EBV) nuclear antigen 1 subtypes in EBV-associated lymphomas from Brazil and the United Kingdom. J. Gen. Virol. 80, 2741–2745.

McGeoch, D.J., 2001. Molecular evolution of the γ-Herpesvirinae. Philos. Trans. R. Soc. Lond., B 356, 421–435.

Midgley, R.S., Blake, N.W., Yao, Q.Y., Croom-Carter, D., Cheung, S.T., Leung, S.F., Chan, A.T.C., Johnson, P.J., Huang, D., Rickinson, A.B., Lee, S.P., 2000. Novel intertypic recombinants of Epstein–Barr virus in the Chinese population. J. Virol. 74, 1544–1548.

Midgley, R.S., Bell, A.I., McGeoch, D.J., Rickinson, A.B., 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.

Miller, G., Lipman, M., 1973. Release of infectious Epstein–Barr virus by transformed marmoset leukocytes. Proc. Natl. Acad. Sci. U.S.A. 70, 190–194.

Pizzo, P.A., Magrath, I.T., Chattopadhyay, S.K., Biggar, R.J., Gerber, P., 1978. A new tumour-derived transforming strain of Epstein–Barr virus. Nature 272, 629–631.

Tajima, F., 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135, 599–607.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555–556.

Yao, Q.Y., Tierney, R.J., Croom-Carter, D., Cooper, G.M., Ellis, C.J., Rowe, M., Rickinson, A.B., 1996. Isolation of intertypic recombinants of Epstein–Barr virus from T-cell-immunocompromised individuals. Virology 70, 4895–4903.

Zeng, M.S., Li, D.J., Liu, Q.L., Song, L.B., Li, M.Z., Zhang, R.H., Yu, X.J., Wang, H.M., Ernberg, I., Zeng, Y.X., 2005. Genomic sequence analysis of Epstein–Barr virus strain GD1 from a nasopharyngeal carcinoma patient. J. Virol. 79, 15323–15330.