Epstein-Barr Virus Latent Membrane Protein 1 Genetic Variability in Peripheral Blood B Cells and Oropharyngeal Fluids

Nicholas Renzette,^a Mohan Somasundaran,^{b,c} Frank Brewster,^b James Coderre,^{b,c} Eric R. Weiss,^{b,c} Margaret McManus,^{b,c} Thomas Greenough,^b Barbara Tabak,^{c,d} Manuel Garber,^{c,d} Timothy F. Kowalik,^a Katherine Luzuriaga^{b,c}

Department of Microbiology and Physiological Systems,^a Department of Pediatrics,^b Program in Molecular Medicine,^c and Department of Bioinformatics and Integrative Biology,^d University of Massachusetts Medical School, Worcester, Massachusetts, USA

ABSTRACT

We report the diversity of latent membrane protein 1 (LMP1) gene founder sequences and the level of Epstein-Barr virus (EBV) genome variability over time and across anatomic compartments by using virus genomes amplified directly from oropharyngeal wash specimens and peripheral blood B cells during acute infection and convalescence. The intrahost nucleotide variability of the founder virus was 0.02% across the region sequences, and diversity increased significantly over time in the oropharyngeal compartment ($P = 0.004$). The LMP1 region showing the greatest level of variability in both compartments, and over time, was concentrated within the functional carboxyl-terminal activating regions 2 and 3 (CTAR2 and CTAR3). Interestingly, a deletion in a proline-rich repeat region (amino acids 274 to 289) of EBV commonly reported in EBV sequenced from cancer specimens was not observed in acute infectious mononucleosis (AIM) patients. Taken together, these data highlight the diversity in circulating EBV genomes and its potential importance in disease pathogenesis and vaccine design.

IMPORTANCE

This study is among the first to leverage an improved high-throughput deep-sequencing methodology to investigate directly from patient samples the degree of diversity in Epstein-Barr virus (EBV) populations and the extent to which viral genome diversity develops over time in the infected host. Significant variability of circulating EBV latent membrane protein 1 (LMP1) gene sequences was observed between cellular and oral wash samples, and this variability increased over time in oral wash samples. The significance of EBV genetic diversity in transmission and disease pathogenesis are discussed.

Epstein-Barr virus (EBV), a double-stranded DNA gammaherpesvirus, infects more than 95% of the world's population by the 4th decade of life (1). Transmission occurs predominantly through exposure to infected saliva. EBV persistence is lifelong and is characterized by infection of epithelial tissues of the oropharynx and circulating memory B cells (2, 3). Lytic infection of epithelial and B cells results in the production of progeny viruses, which are chronically shed at high levels (average, $>10^4$ particles/ml) in saliva (4, 5). Latently infected memory B cells serve as a viral reservoir in which EBV evades immune surveillance through the transcription of a restricted subset of viral genes and limited protein expression (6).

While most EBV infections are asymptomatic, primary infection in late childhood or after the 2nd decade of life may result in acute infectious mononucleosis (AIM). In addition, EBV has been associated with epithelial and lymphoid malignancies. Characterization of the sequences of infecting strains of EBV, and of the diversity and stability of the viral genome over time, would contribute to the development of a prophylactic vaccine (7). Examination of EBV genetic diversity over time in healthy and diseased hosts may also improve insights into the transforming properties of EBV (6, 8) that could allow the development of improved therapies for EBV-associated malignancies.

Prior studies have relied primarily on analyses of restriction fragment length polymorphisms or EBV nuclear antigen (EBNA) protein length variations ("EBNotype") of EBV strains to evaluate the genetic diversity of individual EBV genes in oropharyngeal secretions, peripheral blood, and tumor tissues. Collectively, these data reveal diversity in 6 genes (EBNA 1, EBNA 2, EBNA 3A, EBNA 3B, EBNA 3C, and latent membrane protein 1 [LMP1])

important in the establishment and maintenance of the virus in B lymphocytes (9–12). However, these assays are more qualitative than quantitative. More-recent studies have employed PCR-based analyses (heteroduplex tracking assays or cloning and Sanger sequencing) to demonstrate particular heterogeneities of LMP1 sequences (13, 14).

While EBV polymorphisms have been described, the diversity of circulating EBV strains remains poorly defined. Whole-genome sequences are available for only a limited number of primary viruses (15). We and others have previously shown significant differences in sequence and *in vitro* functional properties between viruses propagated *in vitro* and those sequenced directly from peripheral blood or body fluids (e.g., HIV and human cytomegalovirus [HCMV] [16–19]). Moreover, sequencing of tumor-associated viruses may select EBV strains with enhanced transforming properties for analysis (20–27).

High-throughput sequencing is a powerful tool for examining viral diversity directly from patient samples. These methods offer the ability to accurately quantify viral variants at the initiation of infection and to track the evolution of the quasispecies over time

Received 15 November 2013 Accepted 9 January 2014

Published ahead of print 15 January 2014

Editor: R. M. Sandri-Goldin

Address correspondence to Katherine Luzuriaga, Katherine.Luzuriaga@umassmed.edu.

Copyright © 2014, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JVI.03378-13

(19). They also offer unique challenges, most notably distinguishing between true viral diversity and sequencing error (28, 29) and correctly characterizing complex population structures (30, 31). We have previously developed Illumina deep-sequencing protocols to perform genomewide sequencing of HCMV (18). Those studies showed that HCMV populations are highly diverse, with levels of nucleotide variation comparable to those previously described for RNA viruses, such as dengue virus and HIV (18). Others have shown that viral diversity is correlated with pathogenesis in both DNA and RNA viruses (16, 17, 32), while we have shown that viral diversity can contribute to rapid evolution in human hosts (19).

In the present study, we have applied high-throughput sequencing to investigate directly from patient samples the degree of diversity in founder virus populations and the extent to which EBV genome diversity develops over time in the infected host. Since whole-genome sequencing requires relatively large volumes of blood (especially during chronic infection, when circulating memory B cell frequencies are low), we targeted LMP1 for amplification and sequencing. LMP1 is an integral transmembrane protein with relatively high genetic diversity, functional importance, and a large number of known CD8⁺ T cell epitopes (33–39). It consists of a short amino-terminal cytoplasmic tail (amino acids [aa] 1 to 25), six transmembrane domains (aa 26 to 196), and a carboxyl-terminal tail (aa 197 to 386). The C terminus contains 3 functional domains (carboxyl-terminal activating regions 1, 2, and 3 [CTAR1, -2, and -3]) that interact with cellular proteins to activate a wide array of signaling pathways, including nuclear factor κ B (NF- κ B), and cell cycle-regulatory molecules (20, 22, 40–42).

In this study, we specifically analyzed the EBV LMP1 gene present in latently infected B cells (BC) as well as in virus shed from infected oropharyngeal epithelial cells (oral wash [OW] samples) during AIM and 6 months postinfection (convalescence [CONV]). Significant interhost and intrahost variability of circulating EBV LMP1 sequences was observed over time and between anatomic compartments. The potential importance of the genetic variability observed for EBV transmission and disease pathogenesis is discussed.

MATERIALS AND METHODS

Study population and specimens. The Institutional Review Board of the University of Massachusetts Medical School approved these studies, and all participants provided written informed consent. Blood and oropharyngeal samples were obtained from a previously described cohort of young adults (44, 45) who presented with symptoms of AIM, which was further confirmed by a positive monospot (heterophile antibody) assay. Primary EBV infection was confirmed by the detection of serum IgM antibodies at presentation, along with the development of serum IgG antibodies in convalescence, to the EBV viral capsid antigen. Paired blood and oropharyngeal samples were again collected 6 months after presentation with AIM (i.e., during convalescence [CONV]).

B cell purification and DNA isolation. B cells were purified (>95%) from whole blood using the RosetteSep human B-cell enrichment cocktail and the manufacturer's protocol (StemCell Technologies, Vancouver, BC, Canada). DNA was extracted using Qiagen (Valencia, CA) DNeasy Mini spin 150 columns. DNA from oral wash samples was isolated using the Roche (Indianapolis, IN) High Pure viral nucleic acid kit, and viral loads were measured using the Roche (Indianapolis, IN) LightCycler EBV quantitation kit by quantitative PCR (qPCR) on the Roche LightCycler system (44, 45).

Amplification of EBV genomic DNA. Overlapping PCRs were performed using primer pairs that were designed specifically to optimize the amplification of EBV LMP1 (primers 165342F [5'-ACGGCACTACCT GTGCCG-3'] and 167810R [5'-ATGGCGGCGGTGATCCACAC-3'] for reaction A and primers 166820F [5'-TGG TCTACCCGGTGCATG G-3'] and 169603R [5'-ACACTCGCACAGCCACAC-3'] for reaction B). These primers were also used to amplify the identical region of the type I strain of EBV, B95-8, from a previously described bacterial artificial chromosome (BAC) construct (46) (kindly provided by Fred Wang).

Library construction and Illumina sequencing. The amplicon DNA was processed for analysis by Illumina sequencing as described previously (18). The ends of the Bioruptor-fragmented DNA segments were repaired (Epicentre end repair kit); adapters with appropriate bar codes were ligated; and a library with a median size of 300 bp $\pm$ 50 bp was produced. The library was amplified with Illumina paired-end primers and was then sequenced (UMMS Deep Sequencing Core). The reads were sorted by bar code, after which the bar code sequence was removed. Raw Illumina sequence reads were deposited to the NCBI Sequence Read Archive (SRA) and are associated with BioProject PRJNA231919.

Sequence data analysis: alignment, coverage, filtering process, and variant selection. An analysis pipeline, similar to that reported previously (18, 47), provided a computationally expedient framework for the extraction of information from the large data set. The development and validation of the pipeline depended on resequencing of the LMP1 region of an EBV B95-8 BAC. After removal of the adapter sequences, a consensus sequence was called for each sample on the basis of *de novo* contig assembly with the SSAKE program (48), followed by full-length sequence assembly with the assembly algorithm of the Geneious software suite (Geneious, version 4.8; created by Biomatters). The full-length consensus sequences were then used to identify polymorphisms and insertions/deletions (indels) between patient samples. The full-length consensus sequence was also used for reference-guided alignment to call intrahost variants (49). Variants from the reference sequence were called using a variant-filtering algorithm described previously (18, 47). This algorithm determines those putative polymorphisms that are likely to be "true" variants on the basis of base call quality, mapping quality, local depth, variant frequency, and variant count. The threshold values for these parameters were selected empirically to balance sensitivity and specificity based on our sequence data for the B95-8 BAC. Variants included in downstream analysis have base call qualities of ≥ 30 , mapping qualities of ≥ 89 , a local depth of ≥ 15 , variant counts of ≥ 3 , and a variant frequency of $\geq 1.86\%$. The false-positive rate (the rate of inclusion of errors in the sequence data) is highly dependent on the frequency threshold used in the filtering algorithm as well as on the sequencing depth. In this study, we have achieved an average depth of 52,000 times for the samples, with a minimum depth of $\sim 6,000$ times. We used a minimum variant frequency threshold of 1.86%, leading to a maximum false-positive rate of 1% across all samples. This threshold is similar to those that we and others have reported previously (18, 19).

Data from the variant filter were then further processed to calculate nucleotide diversity (π), mean diversity, inferred amino acid diversity, and the ratio of nonsynonymous to synonymous substitutions (dN/dS). JalView (50) was used to visualize the amino acid sequence alignment and to determine conservation histograms, columns with mutations, and columns with conserved physicochemical properties. Quantitative analysis of the alignment yielded a numerical index score based on conserved physicochemical properties for each column as described by Livingstone and Barton (51).

Nucleotide sequence accession numbers. Consensus sequence data from this study have been deposited in the NCBI database under accession numbers [KF871337](#) to [KF871363](#).

RESULTS

False-positive rate in deep-sequencing data as a function of depth and the variant frequency threshold. EBV BAC B95-8

TABLE 1 Sequence reads, sequencing depths and coverage, and intrahost diversity

Patient	Sample source and time	No. of reads (10 ⁷)	No. of sequences generated (10 ⁸)	Avg sequencing depth (times)	% Coverage	No. of variable sites	Nucleotide diversity (π) (%)	No. of nonsynonymous variants	π_{AA} ^a (%)
B95-8	BAC	3.95	13	72,174	100				
E1347	BC.AIM	1.04	3.44	9,782	100	23	0.287	18	0.242
	OW.CONV	0.786	2.59	5,071	99.96	71	0.205	48	0.145
E1369	OW.AIM	1.60	5.28	90,111	100	6	0.066	6	0.066
	BC.AIM	1.20	3.96	46,850	100	0	0.000	0	0.000
	BC.CONV	1.04	3.42	8,369	100	5	0.064	5	0.064
E1376	OW.AIM	0.976	3.22	38,612	100	0	0.000	0	0.000
	BC.AIM	2.90	9.57	75,689	100	0	0.000	0	0.000
E1405	OW.AIM	1.23	4.08	93,015	100	0	0.000	0	0.000
	BC.AIM	1.16	3.84	48,731	100	0	0.000	0	0.000
	OW.CONV	0.768	2.53	57,191	100	0	0.000	0	0.000
	BC.CONV	1.02	3.38	3,562	100	1	0.016	0	0.000
E1420	OW.AIM	1.33	4.39	88,987	100	1	0.020	0	0.000
	BC.AIM	1.83	6.02	101,346	100	1	0.013	0	0.000
	OW.CONV	0.889	2.94	67,005	100	0	0.000	0	0.000
	BC.CONV	0.115	0.379	4,849	100	1	0.008	0	0.000
E1435	OW.AIM	0.946	3.12	55,378	99.96	1	0.009	1	0.009
	BC.AIM	0.452	1.49	7,619	100	5	0.052	5	0.052
	OW.CONV	0.099	0.327	3,959	100	6	0.052	5	0.050
E1458	OW.AIM	0.471	1.55	28,433	100	0	0.000	0	0.000
E1471	OW.AIM	1.75	5.78	49,849	100	1	0.006	0	0.000
	BC.AIM	1.45	4.79	15,625	100	0	0.000	0	0.000
	OW.CONV	2.03	6.71	56,322	99.96	1	0.007	0	0.000
	BC.CONV	0.929	3.07	15,183	100	1	0.012	0	0.000
E1475	OW.AIM	2.02	6.68	129,393	100	1	0.002	0	0.000
	BC.AIM	1.29	4.27	38,843	100	0	0.000	0	0.000
	OW.CONV	2.30	7.58	119,234	100	31	0.140	23	0.098
	BC.CONV	0.956	3.15	8,083	100	1	0.002	0	0.000

^a π_{AA} , amino acid diversity.

DNA was PCR amplified and was processed as a template for amplification and paired-end sequencing on the Illumina GAI system in order to estimate the background error rate (see Materials and Methods). The output was 10⁸ Mb of EBV sequence, or the equivalent of approximately 466 EBV genomes. A key parameter of the filtering is a frequency threshold, such that all variants below the frequency threshold are excluded from the analysis. In this study, we achieved nearly 100% coverage for all samples, with a minimum depth of ~3,500 times, an average depth of 52,000 times, and a minimum variant frequency threshold of 1.86%, leading to a maximum false-positive rate of 1% across all samples (Table 1).

Intrahost nucleotide and amino acid diversity. The intrapatient nucleotide diversity (π) of LMP1 populations was calculated using the formula described previously (52). The π value measures the average pairwise difference between any two sequences within a sample and allows for the assessment of genetic variation within a set of samples, since increasing π values correspond to increasing genetic variation of a population. Samples were grouped by compartment (BC, OW) or time (AIM, CONV) in

order to analyze diversity patterns. The average nucleotide π value of the *lmp1* open reading frames of all samples present in our study was 0.02% (Table 1 and Fig. 1). Overall, nucleotide diversity in samples during primary infection (AIM) (π , 0.01%) was lower than that in convalescent-phase samples (CONV) (π , 0.03%); greater diversity was observed in the OW samples (π , 0.025%) than in BC (π , 0.012%).

The intrahost nucleotide diversity of all B cell and OW samples was low in AIM (π , ~0.015% and <0.01%, respectively), suggesting that the founder virus was relatively homogenous. Diversity increased significantly in oral wash samples during CONV (π , >0.04%), while it remained low in B cells (π , <0.01%) (Fig. 1). Indeed, the diversity of EBV genomes in oral wash samples during CONV was significantly greater than that in all other samples (P , 0.004 by the Kruskal-Wallis test with the Bonferroni-Dunn *post hoc* test). Kruskal-Wallis analysis of the diversity across all 4 combinations of compartments and times (B cells during AIM, B cells during CONV, OW samples during AIM, and OW samples during CONV) showed a significant difference (P , 0.004), but with only 3 combinations (B cells during AIM, B cells during CONV, and OW

FIG 1 Intrahost nucleotide diversity (π) by sample timing and compartment. Samples were grouped by time or compartment in order to analyze patterns of diversity, and results were plotted as points. Black horizontal lines indicate mean values. The diversity of OW samples taken during convalescence (OW.CONV) is statistically significantly higher than the diversities of all other samples (P , 0.004 by the Kruskal-Wallis test with the Bonferroni-Dunn *post hoc* test). Data on all individual samples can be found in [Table 1](#). (Inset) The mean nucleotide diversity for each patient sample taken at the indicated time (AIM or CONV) from the indicated compartment (B cells, OW) was plotted in order to analyze patterns. Horizontal lines indicate mean nucleotide diversity for each timepoint (AIM or CONV) and compartment (B cells, OW) from matched patient samples. Filled circles, B cells in AIM; upright triangles, B cells in CONV; filled squares, oral wash samples in AIM; inverted triangles, oral wash samples in CONV.

samples during AIM), the difference was not significant (P , 0.52). However, when these combinations were pooled (B cells during AIM, B cells during CONV, and OW samples during AIM), the difference from OW samples during CONV was significant by the Mann-Whitney test (P , 0.0004). Furthermore, instead of grouping the samples according to sampling time or source, the median nucleotide diversity (π score) for each patient sample was plotted in order to analyze patterns, albeit with less power to detect differences ([Fig. 1](#), inset). A Mann-Whitney analysis of the diversity of samples with time and compartment suggested a trend (P , 0.086) toward greater diversity in OW samples during CONV than in OW samples during AIM.

Population differentiation across time and compartment.

To measure the extent of divergence among populations relative to the net genetic diversity, F_{ST} (fixation index) was calculated for all pairs of samples across time or compartments for each patient ([Fig. 2](#)). F_{ST} , a commonly used metric, is a measure of population differentiation bounded by 0 and 1; higher values indicate greater variation in single nucleotide polymorphism (SNP) frequency between populations and, thus, higher levels

FIG 2 Population differentiation across times and compartments. F_{ST} was calculated for all pairs of samples across times or compartments for each patient. The values of F_{ST} were then grouped by time or compartment and were plotted as points. Horizontal lines indicate means.

FIG 3 Intrahost amino acid variability in LMP1 sequences from the samples of four AIM patients. Shown are the frequencies of sequences from four patients that exhibited intrahost LMP1 amino acid variability of $>1\%$. Three patterns of variability were observed. In the first, high-frequency variants are clustered between amino acids 260 and 360. This pattern is seen in samples from patients E1435 and E1369. The second pattern features low-frequency variation found throughout the primary sequence of the protein, as demonstrated by the sample from patient E1475. The third pattern, a hybrid of patterns 1 and 2, is evident in patient sample E1347. Open circles, BC during AIM; filled circles, BC during CONV; open squares, OW samples during AIM; filled squares, OW samples during CONV.

of population differentiation. F_{ST} was calculated, as described previously (47), for all longitudinal or compartmental pairs of samples. The values of F_{ST} were then grouped by time or compartment and are presented as scatter plots, with the mean values shown as black lines (Fig. 2). The average F_{ST} value across all samples was 0.10 (zero indicates no divergence), suggestive of a low to moderate degree of population differentiation between samples. However, there was little difference in values when they were compared across time or across compartments, suggesting that differentiation is not dependent on these parameters. No significant evidence of positive selection in gene sequences was observed by the method of Nielsen et al. (53) (data not shown). These data, along with those presented in Fig. 1, show that while intrahost LMP1 sequences are diverse, the specific collection of mutations that occur within a single host is established during primary infection and is fairly stable over the sampling period (6 months) in this study.

The average amino acid π value of LMP1 for all samples present in our study was 0.03% (Table 1). To determine the intrahost amino acid variability found throughout the primary sequence of LMP1, the frequency of variation was charted for each amino acid position in LMP1. High ($>1\%$) sequence diversity was observed in four AIM patients (Fig. 3). Overall, three distinct patterns were observed in these patients. First, in patients E1435 and E1369, high-frequency variants clustered between amino acids 260 and 360 (Fig. 3, top). Second, a pattern of low-frequency variation was found throughout the primary sequence of the protein, as dem-

onstrated by samples from patient E1475 (Fig. 3, bottom). The third pattern, which is a hybrid of patterns 1 and 2, is evident in samples from patient E1347. We used previously reported genotyping information (27, 41, 54) to determine whether these sites were informative of the presence of one or several previously identified LMP1 genotypes. Only a very small proportion of the amino acid variable sites were genotype informative (6 of 66 for E1347, 1 of 11 for E1435, 1 of 23 for E1475, and none for the other patients), suggesting that genotype classification schemes poorly characterize the variability that exists within *in vivo* EBV populations.

Interhost nucleotide and amino acid diversity: phylogeny and protein alignment. To examine interhost genetic variability and evolutionary relationships, patient LMP1 gene sequences were analyzed by constructing a maximum likelihood tree (Fig. 4). LMP1 gene sequences from type I (B95-8, GD1, GD2, Akata, and Mutu) and type II (AG876) strains, as well as LMP1 sequences obtained from specific geographical regions (China 1, China 2, Mediterranean + [Med+], Med-, North Carolina [NC], and Alaskan sequences), were also included. All samples from the same patient/host were highly similar and clustered together, regardless of time point or compartment of origin. Samples from three patients (E1376, E1471, and E1420) were most closely related to the prototypical type I strain B95-8, isolated from a patient with infectious mononucleosis. The remainder of the cohort demonstrated greater similarity to other reference LMP1 sequences. For example, E1347 samples appeared to be most closely related to a cluster that contained both type I (Med, Mutu) and type II

FIG 4 Phylogenetic analysis of patient LMP1 sequences. Shown is a maximum likelihood tree constructed using patient LMP1 consensus sequences as well as sequences from a prototypic EBV type I strain (B95-8) and from type I (Akata, Mutu, GD1, GD2) and type II (AG876) reference sequences associated with distinct geographical locations. Nodes with >90% bootstrap support are labeled.

(AG876) LMP1 sequences. Samples from four other patients (E1458, E1369, E1405, and E1475) aligned better with a cluster containing Akata, GD1, GD2, and China 1 sequences.

A principal component scatterplot of patient samples and B95-8 was created with data specifying the nucleotide contents of variant positions relative to B95-8, as also called by the Genome Analysis Toolkit (GATK) (55) (data not shown). Samples from the same patient clustered together but failed to indicate a relationship between the patient LMP1 sequence and the calendar time of infection, HLA type, or clinically available information.

Amino acid and protein alignments of LMP1 sequences of patient samples with LMP1 sequences of reference strains. To identify potential selective pressures that might result in LMP1 diversity in our patient cohort, we looked for specific mutations at the amino acid level in regions of functional significance. To that end, a full-length LMP1 protein consensus sequence from each patient was aligned with a consensus sequence generated from all reference and patient sequences (Fig. 5). The alignment showed areas of interpatient variation where residues differ from the total consensus sequence, along with regions of high conservation, i.e., either with no change in residue or with conservation of polarity/charge. The N-terminal cytoplasmic tail, to which no specific function has been attributed, displayed almost complete conservation across our patient samples. A notable exception was the consensus sequence of patient E1475, which differed from the other patient samples at four positions in the cytoplasmic region (specifically, aa 3, 13, 17, and 25); however, all four variant residues were present in at least half of the reference samples (Akata, Alaskan, China 1, China 2, GD1, and GD2), and three were observed to be present in LMP1 sequenced from a nasopharyngeal

carcinoma (NPC) patient from North Carolina (NC). Similarly, analysis of the transmembrane domain of LMP1, contained within residues 25 to 186 of the protein, demonstrated less diversity in patient samples than in reference samples. The majority of the amino acid variation in the transmembrane domain of patient LMP1 sequences was restricted to only 14 total positions spread across the 160-aa region, compared to 29 positions in reference samples. In patient samples, amino acid substitutions at these positions were functionally conserved and were evenly represented in the samples, e.g., the Leu residue at 122 was observed in half of the patient samples (E1369, E1405, E1435, E1458, and E1475) or was replaced with Ile (E1347, E1376, E1420, and E1471). As with the LMP1 cytoplasmic tail region, almost all variants present in the patient samples were observed in at least one of the reference samples, with two notable exceptions. The amino acid sequence of patient E1376 displayed a functionally conserved Leu substitution at Ile78, which was unique among all samples. An additional, more dramatic substitution (Gly at Arg133) was observed in three patient samples (E1369, E1405, and E1458); this amino acid was not present in any of the reference sequences.

In contrast to the findings for the cytoplasmic tail and transmembrane regions of LMP1, analysis of the CTARs revealed substantial variability in amino acids between patient and reference samples. Reference samples showed remarkable conservation in CTAR1 (aa 192 to 234), with only two positions, Ser212 and Ser229, displaying variation from the consensus sequence, although these variants maintained either size (Gly) or side chain reactivity (Thr). Variants at both of these residues were also observed, separately, in patient samples E1347 and E1376. Additional alternative amino acids were determined to be present in the CTAR1 of several patient samples (E1347, E1420, E1458, and E1475) and not in any of the reference samples. The C-terminal region of the protein, containing a portion of CTAR2 (aa 351 to 386), displayed a high degree of conservation, as demonstrated by the presence of only a few, sporadic variants in both reference and patient samples. A greater degree of variation was observed in the N-terminal portion of this domain; indeed, the greatest levels of LMP1 variability in our patient samples were determined to be concentrated between the C terminus of CTAR1 and the N terminus of CTAR2, in the span containing CTAR3 (Fig. 5 and 6).

As noted above and as seen in Fig. 3 and 5, the region where high-frequency variants clustered was between amino acids 260 and 360 and corresponded to the signaling domains overlapping CTAR3 and CTAR2. To identify specific variations within this region, the protein alignment of LMP1 sequences from study patients was compared with reference EBV LMP1 sequences and with the consensus sequence derived from all reference and patient samples (Fig. 6). Comparison of CTAR3 and a significant portion of CTAR2 of the consensus LMP1 sequences from our patient samples demonstrated considerable variability; the nine patient samples segregated into 6 distinct groups. Three groups consisted of two patient sequences (E1369 and E1405; E1420 and E1471; E1458 and E1475), while the remaining three patient samples each formed a distinct group (E1347, E1376, and E1435). Five of these groups were unique and did not match any of the reference samples used in our analysis, while the remaining patient sample (E1376) aligned with the LMP1 reference sequence from B95-8. The LMP1 consensus sequence from patient E1435 displayed the greatest variability among all patient samples, containing substitutions observed in other patient and reference samples

FIG 6 Protein alignment of LMP1 sequences from study patients with reference EBV LMP1 sequences. A conservation histogram, with numerical index scores reflecting conserved physicochemical property scores as described in the legend to Fig. 5, is shown below the top alignment. The top alignment includes both reference sequences and a single representative for each patient, with any positions of intrapatient variability marked by a caret both above the alignment and within the applicable sequence. The bottom alignments include the sequences for all samples of patients with any intrapatient variability in the corresponding regions. A filled triangle marks each of the three positions where multiple patients have intrapatient variability. At two of these positions, two patients exhibit the same deviation from B95-8 in one or both oral wash (OW) samples.

(e.g., Gln276, present in AG876, Alaskan, China 2, Med+, Med–, and NC sequences; Gln 322, present in E1376, E1420, and E1471 as well as in B95-8 and China 2 sequences). This patient's consensus sequence also displayed several unique amino acids substituting Asn for Asp at position 317 of the protein, Asp for Gly (at aa 319), and Cys for Gly (at aa 355); usage of these amino acids was not observed in any other patient or reference sequences used in our analysis.

In addition to alternative amino acid usage, our patient samples also displayed variability in overall LMP1 sequence length in

this region, particularly in comparison to LMP1 reference sequences. With the notable exception of B95-8, all reference samples displayed a previously described deletion of 5 to 15 amino acids in CTAR3 (aa 274 to 289) (11). In contrast, consensus sequences from our patient samples lacked this deletion in CTAR3, although two patient samples (E1347 and E1435) did contain a minimal deletion of 2 residues in this region (Fig. 6, bottom alignment). A separate sequence length variant, resulting from the insertion of a 10-amino-acid repeat between residues 342 and 355, spanning the N-terminal region of CTAR2, was present in about

half of the reference samples (B95-8, Alaskan, China 2, Med–, Mutu, and NC samples) as well as in 7 of 9 patient samples (E1347, E1376, E1420, E1471, E1435, E1458, and E1475).

Two alleles differing within the highly variable CTAR2–CTAR3 sequence were of particular interest, since they showed similar compartmental patterns in two different patients. Two patients (E1405, E1435) showed a His (major allele in BC samples)-to-Glu (major allele in OW samples) substitution at aa 276. Similarly, we observed an allele change in two different patients (E1347, E1435) involving a His (BC samples)-to-Arg (OW samples) substitution at aa 352. We also noted a significant amino acid substitution in this region for patient E1435 samples, in which the conserved hydrophobic residue at position 355 (either Gly, Ala, or Val) had been replaced with Cys.

DISCUSSION

We sequenced the EBV LMP1 gene present in latently infected B cells (BC), as well as in virus shed from the oropharyngeal compartments (oral wash [OW] samples), of infected individuals during acute infectious mononucleosis (AIM) and 6 months postinfection (during convalescence [CONV]) in order to determine the interhost and intrahost variability of circulating EBV LMP1 sequences. Our results suggest the presence of considerable interpatient viral genome diversity in circulating EBV strains. Viral diversity in BC and OW samples was lower during acute infection; diversity increased over the course of EBV infection, particularly in infected epithelial tissue. In addition, our results suggest qualitative differences between sequences of viral proteins produced during the EBV infectious cycle and reported sequences from tissues of individuals with EBV-associated disease. To our knowledge, this is among the first studies to leverage deep-sequencing technology to quantitatively measure the genomic diversity of circulating EBV strains.

Intrahost genetic variability was observed in EBV-infected individuals. Our previous work has shown that the diversity of HCMV, a DNA virus, is comparable to that observed in populations of RNA viruses, such as dengue virus (18). As Table 1 and Fig. 1 demonstrate, the average overall intrahost variability of EBV sampled from human hosts in our study was 0.02%. This value is comparable to the whole-genome variability seen in early infection with RNA viruses such as West Nile virus (0.03%) and HIV-1 (0.04%) (17, 21, 30, 31), although it is approximately 10-fold lower than that of HCMV (0.21%) or dengue virus (0.3%) populations characterized in early infection. RNA viruses, such as hepatitis C virus (HCV) or HIV-1, have been shown to harbor high levels of genetic diversity (56–59), presumably due to high mutation rates intrinsically associated with RNA replication or reverse transcription (60). It is becoming increasingly apparent that DNA viruses may also exhibit significant intrahost diversity. For example, significant population variability has been observed in pseudorabies virus (61), HCMV (18), and Kaposi's sarcoma-associated herpesvirus (KSHV) (62), members of the alpha-, beta-, and gammaherpesviruses, respectively. Our results extend these findings and suggest that this is a general phenomenon of the *Herpesviridae* family, although the factors contributing to the diversity remain unclear. Population bottlenecks associated with transmission and dissemination have been described in RNA viruses, such as HIV-1 (63), and, to a lesser extent, in DNA viruses (64), including the herpesvirus HCMV (19). The mutation rates

of DNA viruses, and of herpesviruses specifically, are significantly lower than those of RNA viruses (65).

The relatively limited overall genomic diversity of EBV in acute infection suggests the possible selection of variants for transmission or for amplification posttransmission. In this study, we also documented a distinct increase in the genomic diversity of EBV in the OW samples of our cohort at time points between 6 months and a year after initial infection. In contrast, the viral genomic variability present in the B cell compartment of our cohort was determined to be very low. These findings are compatible with the notion that LMP1 variance is introduced during the replication of the viral genome during particle production in the oropharynx. This observation is consistent with our current understanding of the EBV replication cycle and of the requirement for LMP1 functionality in establishing latent EBV infection in B cells. The low level of genomic variation measured in the B cell compartment at both time points further suggests that seeding of B cells by viruses produced by the host is infrequent, at least over the limited time course (6 months to 1 year) under investigation.

Although the overall level of EBV genomic nucleotide variation in B cells was determined to be low, analysis of LMP1 at the amino acid level indicated the presence of allelic variants not present in virus from oral wash samples in several patients. Samples from patients E1347, E1405, and E1435 displayed differences in the major alleles present in the two compartments, notably at His276 and His352, as well as at Gly355. In all cases, the substitution was relatively modest (Glu for His at aa 276, Arg for His at aa 352, and Cys for Gly at aa 355), however, and these alleles did appear to segregate by compartment, with Glu, Arg, and Cys substitutions appearing in samples derived from oral washes. Whether these substitutions provide further evidence for an increase in genome diversity generated during transcription or whether they indicate selected colonization of compartments during infection remains to be determined.

To date, only a few full-length EBV genomic sequences are available, and most EBV genotyping was done prior to the availability of next-generation sequencing. Moreover, many available sequences were derived from patients diagnosed with diseases associated with EBV infection (i.e., Burkitt's lymphoma or NPC) and may not adequately reflect viral genomes in circulation. These sequences have also been identified as potentially endemic to specific geographic regions (e.g., GD1, GD2, China 1, and China 2). In contrast to this notion, our maximum likelihood tree demonstrated that while the sequences of samples from each individual patient in our cohort taken at different time points or from different compartments clustered with each other, their sequences clustered with LMP1 sequences from a range of geographic isolates. This finding suggests that the overall diversity of EBV circulating within a population may be sizeable.

While polymorphisms in various EBV genes have been identified, the factors contributing to genetic variability are not well understood. Low but detectable genomic diversity at presentation with symptoms of AIM suggests the transmission of more than one variant. Alternatively, diversity may result from the transmission of a single variant followed by early inpatient diversification. While EBV polymerase has proofreading activity, the detection of sequence diversity in AIM (i.e., several weeks postexposure) could have resulted from high-level ongoing viral replication coupled with inpatient selective pressures (e.g., viral tropism, viral replication fitness, host immune responses). Re-

peated infection (“reinfection”) over time may also contribute to EBV diversity, as reported in an LMP1 genotype study (23, 24, 27). The data presented in Fig. 1 and 2 (particularly the OW sample data) support a model of inpatient diversification due to viral replication or other mechanisms and are not consistent with a reinfection model. However, our data are based on limited (6 months) longitudinal follow-up samples. Additional sampling time points and the analysis of larger genomic loci are necessary. Altogether, however, the observed interhost and early intrahost variability may pose a significant challenge for the development of prophylactic vaccines against EBV.

Better understanding of factors contributing to genetic diversity is important. Of interest is that we detected previously described CD8⁺ T cell escape mutations in AIM patient samples (54, 66), specifically the peptide YLL (YLLEMLWRL), which was identified previously as the major HLA-A2-restricted cytotoxic T lymphocyte (CTL) epitope in LMP1. Three of the nine patients (E1376, E1420, and E1471) had the wild-type (prototypic) B95-8 sequences for the YLLEMLWRL epitope at early and late time points, both in BC and in OW samples. Sequences from four of the remaining patients (E1369, E1405, E1458, E1475) showed a substitution in the A2 anchor residue at position 2 (L126F), expected to alter major histocompatibility complex (MHC) class I binding. Duraiswamy et al. (67) reported that this substitution is present in LMP1 genes of healthy individuals and in NPC biopsy specimens across a broad range of ethnicities at frequencies similar to what we observed with our largely white cohort. Further evaluation of the role of CD8 T cell selective pressure in sequence diversity is necessary.

We observed a notable difference between the sequences of LMP1 CTAR3 isolated from nasopharyngeal carcinoma tissue and from our patient samples. This region exhibits the only amino acid changes that are compartment specific in two different patients, suggesting that CTAR3 may be under different pressures in different compartments. The cytoplasmic region of LMP1 contains 2 JAK3 binding sites (PHDPLP) and a stretch of 11 amino acids encoded by the so-called 33-bp repeat (aa 275 to 307) (68, 69). Recent evidence has indicated that both of these regions are required for the recruitment of Ubc9, the E2 component of the sumoylation cascade, as demonstrated by a general increase in the levels of sumoylated proteins following LMP1 expression (70). Additional research has shown that expression of a variant of LMP1 lacking the Ubc9 interaction domain results in a decrease in the level of sumoylation of interferon regulatory factor 7 (IRF7), a transcription factor of type I interferon genes (71).

Interestingly, EBV LMP1 is capable of promoting multiple modifications to IRF7, including ubiquitylation (both K48- and K63-polyubiquitin chains) through ubiquitin ligase recruitment, deubiquitylation through interaction with A20, and sumoylation through interaction with Ubc9 (reviewed in reference 68). These modifications can affect IRF7 activity by stimulating protein degradation, nuclear or cytoplasmic localization, and chromatin binding, thus providing LMP1 with a high degree of regulatory control over IRF7 function (68). It has been proposed that LMP1 activity is critical for maintaining EBV latency, and disruption of LMP1 function may lead to oncogenesis (71). This theory seems to be supported by the results of our sequencing analyses, which contrast the amino acid sequences of LMP1 from samples derived from patients with NPC or lymphoproliferative disorders with the LMP1 se-

quences from our own AIM cohort. As Fig. 6 illustrates, consensus LMP1 sequences from AIM patients and from strain B95-8 contain complete CTAR3, possessing both JAK3 binding site motifs and a complete 11-amino-acid repeat. These LMP1 variants would be expected to regulate IRF7 activity tightly to establish and maintain viral latency. In contrast, reference sequences assembled from NPC patient samples lack both features and would thus be expected to have altered ability to regulate IRF7. Determining the role of LMP1 mutations in EBV-associated pathologies will require additional investigation into the full EBV genome, as well as the establishment of longitudinal studies to ascertain whether the differences observed in the LMP1 sequence predate the development of the carcinoma or arise after a separate oncogenic event.

The observation that LMP1 is essential for EBV-induced proliferation and transformation *in vitro* (40), along with the fact that it induces nuclear factor κ B (NF- κ B) activity, has led to the hypothesis that LMP1 polymorphisms may be associated with the development of EBV-associated tumors. Zuercher et al. (72) have reported several specific LMP1 polymorphisms in blood and tumor viral sequences from HIV-infected patients with Hodgkin lymphoma that are possibly associated with increased NF- κ B activation. These included the single amino acid changes F106Y, I124V, and F144I. I124V was not observed in any of our study patients, but F106Y variants were found in all nine study patients, and F144I variants were observed in five of the nine study patients. The common detection of variants associated with enhanced NF- κ B activation in the peripheral blood of healthy individuals was somewhat surprising, although it is possible that net NF- κ B activation may be modulated by other polymorphisms.

For RNA viruses, next-generation sequencing has suggested that minor viral variants, together with the diverse major variants, evolve and impact disease progression (29, 32, 73). Likewise, the minor variants may have biological relevance in DNA virus infections and pathogenesis. Here we have demonstrated an improved sequencing methodology that has near 100% coverage and a depth of as much as 130,000 times, allowing the detection of minor variants with a $\leq 1\%$ error rate. Defining virus population structures based on genetic variation is advantageous in determining the founder viruses and coinfecting or coresident viral variants and in defining transmission chains. Additional studies using recently reported methodologies to sequence whole viral genomes directly from clinical samples (74, 75) will be critically important for identifying genetic variants associated with disease.

ACKNOWLEDGMENTS

This work was supported by NIH/NIAID grants P01 AI049320 (to K.L.) and F32 AI084437 (to N.R.).

We are grateful to the study subjects for their participation. For clinical coordination, we thank Rose Ciccarelli and Gail Gnatek. We are grateful to Fred Wang (Brigham and Women's Hospital, Harvard Medical School, Boston, MA) for generously providing the B95-8 EBV BAC construct. For the initial PCR trials to amplify LMP1, we thank Surbhi Jain. For the initial genetic analyses, we thank Jeffrey Jensen (currently at the School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne [EPFL], Swiss Institute of Bioinformatics [SIB], Lausanne, Switzerland). For excellent laboratory support, we thank Robin Brody, Joyce Pepe, Linda Lambrecht, Kristina McNeal, and Richard Hudson. We are grateful to Mindy Donovan for preparation of the manuscript.

REFERENCES

- Luzuriaga K, Sullivan JL. 2010. Infectious mononucleosis. *N. Engl. J. Med.* 362:1993–2000. <http://dx.doi.org/10.1056/NEJMc1001116>.
- Thorley-Lawson DA. 2001. Epstein-Barr virus: exploiting the immune system. *Nat. Rev. Immunol.* 1:75–82. <http://dx.doi.org/10.1038/35095584>.
- Thorley-Lawson DA, Allday MJ. 2008. The curious case of the tumour virus: 50 years of Burkitt's lymphoma. *Nat. Rev. Microbiol.* 6:913–924. <http://dx.doi.org/10.1038/nrmicro2015>.
- Hadinoto V, Shapiro M, Greenough TC, Sullivan JL, Luzuriaga K, Thorley-Lawson DA. 2008. On the dynamics of acute EBV infection and the pathogenesis of infectious mononucleosis. *Blood* 111:1420–1427. <http://dx.doi.org/10.1182/blood-2007-06-093278>.
- Hadinoto V, Shapiro M, Sun CC, Thorley-Lawson DA. 2009. The dynamics of EBV shedding implicate a central role for epithelial cells in amplifying viral output. *PLoS Pathog.* 5:e1000496. <http://dx.doi.org/10.1371/journal.ppat.1000496>.
- Rickinson AB, Kieff E. 2007. Epstein-Barr virus, p 2655–2700. *In* Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE (ed), *Fields virology*, 5th ed. Lippincott Williams & Wilkins, Philadelphia, PA.
- Cohen JL, Fauci AS, Varmus H, Nabel GJ. 2011. Epstein-Barr virus: an important vaccine target for cancer prevention. *Sci. Transl. Med.* 3:107fs7. <http://dx.doi.org/10.1126/scitranslmed.3002878>.
- Luzuriaga K, Sullivan JL. 2009. Epstein-Barr virus, p 521–536. *In* Richman DD, Whitley RJ, Hayden FG (ed), *Clinical virology*, 3rd ed. ASM Press, Washington, DC.
- Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, Seguin C, Tuffnell PS, Barrell BG. 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature* 310:207–211. <http://dx.doi.org/10.1038/310207a0>.
- Busson P, Edwards RH, Tursz T, Raab-Traub N. 1995. Sequence polymorphism in the Epstein-Barr virus latent membrane protein (LMP)-2 gene. *J. Gen. Virol.* 76(Part 1):139–145. <http://dx.doi.org/10.1099/0022-1317-76-1-139>.
- 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. <http://dx.doi.org/10.1016/j.virusres.2009.07.005>.
- Wang WY, Chien YC, Jan JS, Chueh CM, Lin JC. 2002. Consistent sequence variation of Epstein-Barr virus nuclear antigen 1 in primary tumor and peripheral blood cells of patients with nasopharyngeal carcinoma. *Clin. Cancer Res.* 8:2586–2590. <http://clincancerres.aacrjournals.org/content/8/8/2586.long>.
- Sitki-Green DL, Edwards RH, Covington MM, Raab-Traub N. 2004. Biology of Epstein-Barr virus during infectious mononucleosis. *J. Infect. Dis.* 189:483–492. <http://dx.doi.org/10.1086/380800>.
- Walling DM, Brown AL, Etienne W, Keitel WA, Ling PD. 2003. Multiple Epstein-Barr virus infections in healthy individuals. *J. Virol.* 77:6546–6550. <http://dx.doi.org/10.1128/JVI.77.11.6546-6550.2003>.
- McGeoch DJ, Gatherer D. 2007. Lineage structures in the genome sequences of three Epstein-Barr virus strains. *Virology* 359:1–5. <http://dx.doi.org/10.1016/j.virol.2006.10.009>.
- Coquette A, Bourgeois A, Dirand C, Varin A, Chen W, Herbein G. 2004. Mixed cytomegalovirus glycoprotein B genotypes in immunocompromised patients. *Clin. Infect. Dis.* 39:155–161. <http://dx.doi.org/10.1086/421496>.
- Pfeiffer JK, Kirkegaard K. 2005. Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathog.* 1:e11. <http://dx.doi.org/10.1371/journal.ppat.0010011>.
- Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF. 2011. Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. *PLoS Pathog.* 7:e1001344. <http://dx.doi.org/10.1371/journal.ppat.1001344>.
- Renzette N, Gibson L, Bhattacharjee B, Fisher D, Schleiss MR, Jensen JD, Kowalik TF. 2013. Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. *PLoS Genet.* 9:e1003735. <http://dx.doi.org/10.1371/journal.pgen.1003735>.
- Edwards RH, Seillier-Moisewitsch F, Raab-Traub N. 1999. Signature amino acid changes in latent membrane protein 1 distinguish Epstein-Barr virus strains. *Virology* 261:79–95. <http://dx.doi.org/10.1006/viro.1999.9855>.
- Edwards RH, Sitki-Green D, Moore DT, Raab-Traub N. 2004. Potential selection of LMP1 variants in nasopharyngeal carcinoma. *J. Virol.* 78:868–881. <http://dx.doi.org/10.1128/JVI.78.2.868-881.2004>.
- Lorenzetti MA, Gantuz M, Altchek J, De Matteo E, Chabay PA, Preciado MV. 2012. Distinctive Epstein-Barr virus variants associated with benign and malignant pediatric pathologies: LMP1 sequence characterization and linkage with other viral gene polymorphisms. *J. Clin. Microbiol.* 50:609–618. <http://dx.doi.org/10.1128/JCM.05778-11>.
- Sitki-Green D, Covington M, Raab-Traub N. 2003. Compartmentalization and transmission of multiple Epstein-Barr virus strains in asymptomatic carriers. *J. Virol.* 77:1840–1847. <http://dx.doi.org/10.1128/JVI.77.3.1840-1847.2003>.
- Sitki-Green D, Edwards RH, Webster-Cyriaque J, Raab-Traub N. 2002. Identification of Epstein-Barr virus strain variants in hairy leukoplakia and peripheral blood by use of a heteroduplex tracking assay. *J. Virol.* 76:9645–9656. <http://dx.doi.org/10.1128/JVI.76.19.9645-9656.2002>.
- Tang YL, Lu JH, Cao L, Wu MH, Peng SP, Zhou HD, Huang C, Yang YX, Zhou YH, Chen Q, Li XL, Zhou M, Li GY. 2008. Genetic variations of EBV-LMP1 from nasopharyngeal carcinoma biopsies: potential loss of T cell epitopes. *Braz. J. Med. Biol. Res.* 41:110–116. <http://dx.doi.org/10.1590/S0100-879X2008000200006>.
- Walling DM, Shebib N, Weaver SC, Nichols CM, Flaitz CM, Webster-Cyriaque J. 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. <http://dx.doi.org/10.1086/314672>.
- Tierney RJ, Edwards RH, Sitki-Green D, Croom-Carter D, Roy S, Yao QY, Raab-Traub N, Rickinson AB. 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. <http://dx.doi.org/10.1086/498913>.
- Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* 3:329. <http://dx.doi.org/10.3389/fmicb.2012.00329>.
- Zagordi O, Klein R, Daumer M, Beerenwinkel N. 2010. Error correction of next-generation sequencing data and reliable estimation of HIV quasi-species. *Nucleic Acids Res.* 38:7400–7409. <http://dx.doi.org/10.1093/nar/gkq655>.
- Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. 2011. ShorAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12:119. <http://dx.doi.org/10.1186/1471-2105-12-119>.
- Prosperi MCF, Salemi M. 2012. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* 28:132–133. <http://dx.doi.org/10.1093/bioinformatics/btr627>.
- Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439:344–348. <http://dx.doi.org/10.1038/nature04388>.
- Miller WE, Earp HS, Raab-Traub N. 1995. The Epstein-Barr virus latent membrane protein 1 induces expression of the epidermal growth factor receptor. *J. Virol.* 69:4390–4398.
- Morris MA, Dawson CW, Young LS. 2009. Role of the Epstein-Barr virus-encoded latent membrane protein-1, LMP1, in the pathogenesis of nasopharyngeal carcinoma. *Future Oncol.* 5:811–825. <http://dx.doi.org/10.2217/fon.09.53>.
- Raab-Traub N. 2002. Epstein-Barr virus in the pathogenesis of NPC. *Semin. Cancer Biol.* 12:431–441. <http://dx.doi.org/10.1016/S1044579X0200086X>.
- Tsao SW, Tramoutanis G, Dawson CW, Lo AK, Huang DP. 2002. The significance of LMP1 expression in nasopharyngeal carcinoma. *Semin. Cancer Biol.* 12:473–487. <http://dx.doi.org/10.1016/S1044579X02000901>.
- Faumont N, Chanut A, Benard A, Cogne N, Delsol G, Feuillard J, Meggetto F. 2009. Comparative analysis of oncogenic properties and nuclear factor- κ B activity of latent membrane protein 1 natural variants from Hodgkin's lymphoma's Reed-Sternberg cells and normal B-lymphocytes. *Haematologica* 94:355–363. <http://dx.doi.org/10.3324/haematol.13269>.
- Coffin WF, III, Erickson KD, Hoedt-Miller M, Martin JM. 2001. The cytoplasmic amino-terminus of the latent membrane protein-1 of Epstein-Barr virus: relationship between transmembrane orientation and effector functions of the carboxy-terminus and transmembrane domain. *Oncogene* 20:5313–5330. <http://dx.doi.org/10.1038/sj.onc.1204689>.
- Middeldorp JM, Pegtel DM. 2008. Multiple roles of LMP1 in Epstein-

- Barr virus induced immune escape. *Semin. Cancer Biol.* 18:388–396. <http://dx.doi.org/10.1016/j.semcancer.2008.10.004>.
40. Kaye KM, Izumi KM, Kieff E. 1993. Epstein-Barr virus latent membrane protein 1 is essential for B-lymphocyte growth transformation. *Proc. Natl. Acad. Sci. U. S. A.* 90:9150–9154. <http://dx.doi.org/10.1073/pnas.90.19.9150>.
 41. Miller WE, Edwards RH, Walling DM, Raab-Traub N. 1994. Sequence variation in the Epstein-Barr virus latent membrane protein 1. *J. Gen. Virol.* 75(Part 10):2729–2740. <http://dx.doi.org/10.1099/0022-1317-75-10-2729>.
 42. Rowe M, Khanna R, Jacob CA, Argat V, Kelly A, Powis S, Belich M, Croom-Carter D, Lee S, Burrows SR, Trowsdale J, Moss DJ, Rickinson AB. 1995. Restoration of endogenous antigen processing in Burkitt's lymphoma cells by Epstein-Barr virus latent membrane protein-1: coordinate up-regulation of peptide transporters and HLA-class I antigen expression. *Eur. J. Immunol.* 25:1374–1384. <http://dx.doi.org/10.1002/eji.1830250536>.
 43. Reference deleted.
 44. Greenough TC, Campellone SC, Brody R, Jain S, Sanchez-Merino V, Somasundaran M, Luzuriaga K. 2010. Programmed Death-1 expression on Epstein Barr virus specific CD8⁺ T cells varies by stage of infection, epitope specificity, and T-cell receptor usage. *PLoS One* 5:e12926. <http://dx.doi.org/10.1371/journal.pone.0012926>.
 45. Precopio ML, Sullivan JL, Willard C, Somasundaran M, Luzuriaga K. 2003. Differential kinetics and specificity of EBV-specific CD4⁺ and CD8⁺ T cells during primary infection. *J. Immunol.* 170:2590–2598. <http://www.jimmunol.org/content/170/5/2590.long>.
 46. Chen A, Divisconte M, Jiang X, Quink C, Wang F. 2005. Epstein-Barr virus with the latent infection nuclear antigen 3B completely deleted is still competent for B-cell growth transformation in vitro. *J. Virol.* 79:4506–4509. <http://dx.doi.org/10.1128/JVI.79.7.4506-4509.2005>.
 47. Cockerham CC, Weir BS. 1984. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* 40: 157–164. <http://dx.doi.org/10.2307/2530754>.
 48. Warren RL, Sutton GG, Jones SJ, Holt RA. 2007. Assembling millions of short DNA sequences using SNAKE. *Bioinformatics* 23:500–501. <http://dx.doi.org/10.1093/bioinformatics/btl629>.
 49. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18:1851–1858. <http://dx.doi.org/10.1101/gr.078212.108>.
 50. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189–1191. <http://dx.doi.org/10.1093/bioinformatics/btp033>.
 51. Livingstone CD, Barton GJ. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* 9:745–756.
 52. Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76: 5269–5273. <http://dx.doi.org/10.1073/pnas.76.10.5269>.
 53. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566–1575. <http://dx.doi.org/10.1101/gr.4252305>.
 54. Brooks JM, Croom-Carter DS, Leese AM, Tierney RJ, Habeshaw G, Rickinson AB. 2000. Cytotoxic T-lymphocyte responses to a polymorphic Epstein-Barr virus epitope identify healthy carriers with coresident viral strains. *J. Virol.* 74:1801–1809. <http://dx.doi.org/10.1128/JVI.74.4.1801-1809.2000>.
 55. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491–498. <http://dx.doi.org/10.1038/ng.806>.
 56. Sakai A, Kaneko S, Honda M, Matsushita E, Kobayashi K. 1999. Quasispecies of hepatitis C virus in serum and in three different parts of the liver of patients with chronic hepatitis. *Hepatology* 30:556–561. <http://dx.doi.org/10.1002/hep.510300234>.
 57. Troesch M, Meunier I, Lapierre P, Lapointe N, Alvarez F, Boucher M, Soudeyns H. 2006. Study of a novel hypervariable region in hepatitis C virus (HCV) E2 envelope glycoprotein. *Virology* 352:357–367. <http://dx.doi.org/10.1016/j.virol.2006.05.015>.
 58. Saag MS, Hahn BH, Gibbons J, Li Y, Parks ES, Parks WP, Shaw GM. 1988. Extensive variation of human immunodeficiency virus type-1 in vivo. *Nature* 334:440–444. <http://dx.doi.org/10.1038/334440a0>.
 59. Zhu T, Mo H, Wang N, Nam DS, Cao Y, Koup RA, Ho DD. 1993. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* 261:1179–1181. <http://dx.doi.org/10.1126/science.8356453>.
 60. Domingo E, Martínez-Salas E, Sobrino F, de la Torre JC, Portela A, Ortín J, López-Galindez C, Pérez-Breña P, Villanueva N, Nájera R, VandePol S, Steinhauer D, DePollo N, Holland J. 1985. The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance—a review. *Gene* 40:1–8. [http://dx.doi.org/10.1016/0378-1119\(85\)90017-4](http://dx.doi.org/10.1016/0378-1119(85)90017-4).
 61. Szpara ML, Tafuri YR, Parsons L, Shamim SR, Verstrepen KJ, Legendre M, Enquist LW. 2011. A wide extent of inter-strain diversity in virulent and vaccine strains of alphaherpesviruses. *PLoS Pathog.* 7:e1002282. <http://dx.doi.org/10.1371/journal.ppat.1002282>.
 62. Leao JC, de Faria ABS, Fonseca DDD, Gueiros LAM, Silva IHM, Porter SR. 2013. Intrahost genetic variability of human herpes virus-8. *J. Med. Virol.* 85:636–645. <http://dx.doi.org/10.1002/jmv.23491>.
 63. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, Han CS, Gleason CD, Green L, Lo CC, Nag A, Wallstrom TC, Wang S, McMichael AJ, Haynes BF, Hahn BH, Perelson AS, Borrow P, Shaw GM, Bhattacharya T, Korber BT. 2010. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* 5:e12303. <http://dx.doi.org/10.1371/journal.pone.0012303>.
 64. Monsion B, Froissart R, Michalakos Y, Blanc S. 2008. Large bottleneck size in *Cauliflower mosaic virus* populations during host plant colonization. *PLoS Pathog.* 4:e1000174. <http://dx.doi.org/10.1371/journal.ppat.1000174>.
 65. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J. Virol.* 84:9733–9748. <http://dx.doi.org/10.1128/JVI.00694-10>.
 66. de Campos-Lima PO, Levitsky V, Brooks J, Lee SP, Hu LF, Rickinson AB, Masucci MG. 1994. T cell responses and virus evolution: loss of HLA A11-restricted CTL epitopes in Epstein-Barr virus isolates from highly A11-positive populations by selective mutation of anchor residues. *J. Exp. Med.* 179:1297–1305. <http://dx.doi.org/10.1084/jem.179.4.1297>.
 67. Duraiswamy J, Burrows JM, Bharadwaj M, Burrows SR, Cooper L, Pimthanotai N, Khanna R. 2003. Ex vivo analysis of T-cell responses to Epstein-Barr virus-encoded oncogene latent membrane protein 1 reveals highly conserved epitope sequences in virus isolates from diverse geographic regions. *J. Virol.* 77:7401–7410. <http://dx.doi.org/10.1128/JVI.77.13.7401-7410.2003>.
 68. Ersing I, Bernhardt K, Gewurz BE. 2013. NF-κB and IRF7 pathway activation by Epstein-Barr virus latent membrane protein 1. *Viruses* 5:1587–1606. <http://dx.doi.org/10.3390/v5061587>.
 69. Gires O, Kohlhuber F, Kilger E, Baumann M, Kieser A, Kaiser C, Zeidler R, Scheffer B, Ueffing M, Hammerschmidt W. 1999. Latent membrane protein 1 of Epstein-Barr virus interacts with JAK3 and activates STAT proteins. *EMBO J.* 18:3064–3073. <http://dx.doi.org/10.1093/emboj/18.11.3064>.
 70. Bentz GL, Whitehurst CB, Pagano JS. 2011. Epstein-Barr virus latent membrane protein 1 (LMP1) C-terminal-activating region 3 contributes to LMP1-mediated cellular migration via its interaction with Ubc9. *J. Virol.* 85:10144–10153. <http://dx.doi.org/10.1128/JVI.05035-11>.
 71. Bentz GL, Shackelford J, Pagano JS. 2012. Epstein-Barr virus latent membrane protein 1 regulates the function of interferon regulatory factor 7 by inducing its sumoylation. *J. Virol.* 86:12251–12261. <http://dx.doi.org/10.1128/JVI.01407-12>.
 72. Zuercher E, Buttica C, Wyniger J, Martinez R, Battegay M, Boffi El Amari E, Dang T, Egger JF, Fehr J, Mueller-Garmavogly E, Parini A, Schaefer SC, Schoeni-Affolter F, Thurnheer C, Tinguely M, Telenti A, Rothenberger S. 2012. Genetic diversity of EBV-encoded LMP1 in the Swiss HIV Cohort Study and implication for NF-κB activation. *PLoS One* 7:e32168. <http://dx.doi.org/10.1371/journal.pone.0032168>.
 73. Djikeng A, Spiro D. 2009. Advancing full length genome sequencing for human RNA viral pathogens. *Future Virol.* 4:47–53. <http://dx.doi.org/10.2217/17460794.4.1.47>.
 74. Depledge DP, Palser AL, Watson SJ, Lai IY, Gray ER, Grant P, Kanda RK, Leproust E, Kellam P, Breuer J. 2011. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One* 6:e27805. <http://dx.doi.org/10.1371/journal.pone.0027805>.
 75. Lin Z, Wang X, Strong MJ, Concha M, Baddoo M, Xu G, Baribault C, Fewell C, Hulme W, Hedges D, Taylor CM, Flemington EK. 2013. Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. *J. Virol.* 87:1172–1182. <http://dx.doi.org/10.1128/JVI.02517-12>.