Infection, Genetics and Evolution 65 (2018) 96–103

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Recombination rates along the entire Epstein Barr virus genome display a highly heterogeneous landscape

Ariel José Berenstein[1], Mario Alejandro Lorenzetti[1], María Victoria Preciado*

Instituto Multidisciplinario de Investigaciones en Patologías Pediátricas (IMIPP), CONICET-GCBA, Laboratorio de Biología Molecular, División Patología, Hospital de Niños Ricardo Gutiérrez, Buenos Aires, Argentina

ARTICLE INFO

Keywords:
EBV genome
EBV recombination
Recombination rate
Recombination motifs

ABSTRACT

Epstein Barr virus (EBV) has a large DNA genome assumed to be stable, but also subject to mutational processes such as nucleotide substitution and recombination, the latter explored to a lesser extent. Moreover, differences in the extent of recombination events across herpes sub-families were recently reported. Given the relevance of recombination in viral evolution and its possible impact in pathogenesis, we aimed to fully characterize and quantify its extension in all available EBV complete genome by assessing global and local recombination rate values (ρ/bp).

Our results provide the first EBV recombination map based on recombination rates assessment, both at a global and gene by gene level, where the mean value for the entire genome was 0.035 (HPDI 0.020–0.062) ρ/bp. We quantified how this evolutionary process changes along the EBV genome, and proved it to be non-homogeneous, since regulatory regions depicted the lowest recombination rate values while repetitive regions the highest signal. Moreover, GC content rich regions seem not to be linked to high recombination rates as previously reported.

At an intragenic level, four genes (EBNA3C, EBNA3B, BRRF2 and BBLF2-BBLF3) presented a recombination rate above genome average. We specifically quantified the signal strength among different recombination-initiators previously described features and concluded that those which elicited the greatest amount of changes in ρ/bp, TGGAG and CCCAG, were two well characterized recombination inducing motifs in eukaryotic cells. Strikingly, although TGGAG was not the most frequently detected DNA motif across the EBV genome (697 hits), it still induced a significantly greater proportion of initiation events (0.025 events/hits) than other more represented motifs, p-value = 0.04; one tailed proportion test.

Present results support the idea that diversity and evolution of herpesviruses are impacted by mechanisms, such as recombination, which extends beyond the usual consideration of point mutations.

1. Introduction

Epstein Barr virus (EBV) or *Human gammaherpesvirus 4* is one of the most successful human viruses, which has co-evolved with its host since the origins of mankind (Ba Abdullah et al., 2017). Viral evolution is driven by the joint processes of mutation and recombination which, given the case of an improvement in fitness, may be later fixed by immune selective pressure and genetic drift. While mutations may introduce single nucleotide substitution or even result in small insertions and deletions, recombination is a major source of genomic rearrangement (Ba Abdullah et al., 2017; Arenas et al., 2017). Recombination is a frequent phenomenon with high impact in viral evolution. Concerning DNA viruses, this phenomenon is one of the major events participating in the emergence of new viral variants, alterations in virulence and pathogenesis; modulation of immunosurveillance evasion; tissue tropism or antiviral resistance (Martin et al., 2011a) and even poses a challenge for diagnostic purposes (Martin et al., 2011b).

For recombination to occur, at least two viral variants must co-infect the same cell. In this way recombination may be homologous (between the same loci in both parental strands) or non-homologous (occurring in different regions of the genetic fragments involved, frequently leading to aberrant structures (Galli and Bukh, 2014). Recombination does not create new mutations at the nucleotide level but introduces new combinations of the existing ones, which would be

* Corresponding author at: Instituto Multidisciplinario de Investigaciones en Patologías Pediátricas (IMIPP), CONICET-GCBA, Laboratorio de Biología Molecular, División Patología. Hospital de Niños Ricardo Gutiérrez. Gallo 1330, C1425EFD Buenos Aires, Argentina.
E-mail address: preciado@conicet.gov.ar (M.V. Preciado).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.meegid.2018.07.022
Received 17 May 2018; Received in revised form 10 July 2018; Accepted 18 July 2018
Available online 19 July 2018
1567-1348/

hardly attainable by mere substitution, even for fast evolving large virus populations. In this way linked mutations previously incorporated in a genomic region can be transferred to another region by a single recombination event (Pérez-Losada et al., 2015).

Recombination frequency varies among viral families and even throughout subfamilies. Extensive recombination events have been described across the genomes of alpha-herpesviruses such as VZV (Varicella Zoster virus or *Human alphaherpesvirus 3*) and HSV 1 (Herpes Simplex virus 1 or *Human alphaherpesvirus 1*) as well as for beta-herpesviruses such as HCMV (Human Cytomegalovirus or *Human betaherpesvirus 5*) (Renner and Szpara, 2018). However, to date, no genome-wide studies concerning recombination in gamma-herpesviruses were performed.

EBV, the gamma-herpesvirus type specie, is transmitted among individuals through saliva. Following primary infection, which may develop into infectious mononucleosis (IM), latent infection is established in memory B cells. Although latency supposes no risk for the immunocompetent individual, it is this type of infection that is usually associated with neoplastic pathologies, including Hodgkin lymphoma (HL), Burkitt lymphoma (BL), nasopharyngeal carcinoma (NPC), gastric carcinoma, T-cell lymphoma and lymphoproliferative disorders in immunocompromised individuals (PTLD) (Young et al., 2016). Yet, the involvement of EBV in the etiology, progression and/or outcome of these malignancies is not completely understood. Association of EBV with these pathologies varies dramatically in different geographic regions (extensively reviewed in (Chang et al., 2009)). Even though differences in viral association with malignancies may be accounted for by other factors, there is also the possibility that disease-specific genetic variation in EBV in different parts of the world might account for them (Tzellos and Farrell, 2012). In this regard, the major source of variation in EBV is given by the EBNA2 and EBNA3 genes, which allow for a broad classification into type 1 and type 2 EBV; where type 1 EBV has a worldwide prevalence, while type 2 EBV is prevalent in sub-Saharan Africa and in Papua-New Guinea (Tzellos and Farrell, 2012). Moreover, recombination between type 1 and type 2 EBV was described (Palser et al., 2015). Evenmore, when considering single nucleotide polymorphisms in specific genes, EBV may be further classified into viral variants which also display geographic distribution patterns, mainly to Asia, the Americas, Europe and Africa (Ba Abdullah et al., 2017; Palser et al., 2015; Chiara et al., 2016; Correia et al., 2017). However, to date no study has undoubtedly addressed sequence variation and its geographic distribution pattern considering the whole extent of the viral genome.

Given the relevance of the process of recombination in viral evolution and its possible impact in viral pathogenesis, it is important not only to detect but also to fully characterized and quantify its extension in various viral isolates from different sources. Regarding quantification, recombination rate is often used as a measure of the frequency of recombination events in a data-set (Castelhano et al., 2017).

Since we previously described a recombinant variant of LMP1, the major EBV oncoprotein, as the predominant genetic variant circulating in our region (Gantuz et al., 2017); we sought to assess global (entire genome) and local (gene level) recombination rates in a data-set containing all available whole genome EBV sequences. We quantified how this evolutionary process changes along different genomic regions, namely CDS, exons, introns, and different repetitive regions among others. Moreover, we specifically quantified the signal strength among different recombination-initiators features reported in bibliography (Brown, 2014).

Our results agree with previous literature in that repetitive regions in EBV genome are key features for the recombination process. Nevertheless, we observed no correlation between fold changes in recombination rate signal and GC content.

2. Methods

2.1. Data, sequence alignment and position mapping to reference genome

All available EBV complete or nearly complete genome sequences were downloaded from NCBI nucleotide database during October 2017. Accession numbers to sequences included in this work are shown in Supplementary Table 1. All 171 sequences, including the reference genome sequence (NCBI accession number NC_007605) were aligned with MAFFT alignment tool (Katoh and Standley, 2013), with default parameters, and the resulting multiple sequence alignment (MSA) was trimmed to remove poorly aligned positions (those having > 95% of gaps) with trimAl software (Capella-Gutierrez et al., 2009). The resulting MSA mapped the reference to 96.5%, and was only longer by 3.06%, given open gaps within the alignment. Finally, each position in the reference genome was mapped to the corresponding position within the trimmed MSA by means of a personalized python script and a default “colnumbering” parameter in trimAl algorithm. This new genomic position map allowed us to identify the precise position within the reference genome of those recombination rate signal values inferred in the MSA.

2.2. Recombination analysis

Local population recombination rates per site ρ/bp (4Ne.r/bp) were estimated with a LDhat's stable version implemented in RDP4 software (Martin et al., 2015). The program informs recombination rates for each nucleotide. However, a first attempt to run RDP4/LDhat software with the entire genomic MSA resulted in a data overflow that caused the program to crash. In order to overcome this limitation, the MSA was split in six sliding windows of 60Kbp, each of which overlapped the preceding one by 50% in order to reduce border effects. Because of the circular nature of viral DNA inside the infected cell, and the fact that LMP2 gene encompasses coding sequences in both 5′ and 3′ terminal regions of the genome, we considered the last sliding windows as a circular structure, including both ends of linear alignment (MSA positions between 150,000 to 30,000 nt positions).or each sliding windows, LDhat was run with ten million iterations of Monte Carlo Markov Chains (MCMC) were implemented. Given the construction of the sliding windows (50% overlapping), two ρ/bp values were obtained for each nucleotide position, and the final reported ρ/bp values was the mean value for each nucleotide position. Finally convergence was checked by considering the 95% highest posterior density interval (HPDI) using the ‘coalescent’ Rscript provided by LDHat authors (Auton and McVean, 2007) … As a sanity check, an overall recombination test, PHI test (Bruen et al., 2006), was performed for each analyzed windows with RDP4 software as established by Castelhano et al. (2017).

2.3. Genomic feature annotation

Annotated genomic features were extracted from the reference genome (NCBI accession number NC_007065). Gene positions, codifying regions, introns, as well as regulatory and repeat motifs were considered. A personalized python script that takes advantage of ‘biopython’ modules to extract the entire metadata contained in GenBank files was used.

2.4. Repeats and DNA motif extraction

In order to analyze inverted and tandem repeats reported by Brown, embos scripts *‘einverted’* (default parameters) and *‘etandem’* (min repeat size = 10, max repeat size = 50, threshold score = 30) were used (Brown, 2014).

Another personalized python script was used to localize recombination-inducing motifs and their reverse-complementary sequences. In all cases, the same number of DNA motifs found by our

script was exactly the same as the numbers reported by Brown. These DNA motifs were: i) a core chi-like recombination sequence (TGGTGG) (Chuzhanova et al., 2009); ii) a canonical meiotic recombination motif (CCTCCCCT) (Myers et al., 2005); iii) two sequences found to initiate recombination events (TGGAG and CCCAG) (Bengesser et al., 2010); iv) a human Ig class switch sequence (GGGCT) (Chuzhanova et al., 2009), and v), a sequence found to arrest DNA synthesis by DNA polymerase α (AGGAG) (Cullen et al., 2002).

2.5. Detection of recombination rate changes downstream of recombination-inducing motifs

The mean recombination rate (ρ/bp) spanning 10 nucleotide downstream and 10 nucleotides upstream of each recombination-inducing motif was assessed and the ratio between them was calculated with Eq. [1], in order to detect abrupt increases in recombination rate signal. A bootstrapping test, considering 1000 randomly selected k-mers (k = 5) along the genome, was performed on each computed $\rho^{rel}_{(motif)}$ value so as to check for statistical significance (cut-off value was 5%).

$$\rho \text{ rel motif} = \frac{\rho/\text{bp}((10\text{nt downstream of motif}))}{\rho/\text{bp (10nt upstream of motif)}} \quad (1)$$

3. Results

3.1. Repeat regions and GC content in the recombination process

Segregating sites, non-conservative positions which display polymorphisms between related genomic sequences in an alignment, were detected across 57.9% of EBV genome (Fig. 1A). Interestingly, segregating sites were found in 65 of 91 protein coding genes (73.6%), and provided the context for recombination rates estimation. In this way, a recombination rate map across the entire EBV genome was constructed (Fig. 1B). This map evidenced the heterogeneous and highly variable landscape of EBV recombination rate along its genome, denoted by sharp peaks along specific genomic positions, which included viral genes (depicted in Fig. 1B) and other regions. As a sanity check, a PHI test (Bruen et al., 2006) was also computed for testing the overall evidence of recombination in each analyzed windows, and significant signals were found in all reported cases (p-value = 0.001, in all cases).

The mean recombination rate per base pair ($\rho/\text{bp} = 4N_e r/\text{bp}$) for the entire genome was 0.035 (HPDI 0.020 to 0.062, gray zone in Fig. 1C). However, when recombination rates were grouped into genomic categories (repeat region, mRNA, CDS, genes, exon, misc-feature or regulatory regions), values fluctuated across > 4 orders of magnitude; where repetitive regions displayed the highest recombination rate and regulatory regions the lowest. The remaining categories that displayed recombination rates below genome average were other coding regions and genes, particularly their exons (Fig. 1C). The finding of repetitive regions displaying the highest recombination rate signals was not surprising and consistent with the idea that repetitive regions provide an appropriate genomic context for homologous recombination (Brown, 2014; Guiretti et al., 2007). On the other hand, it is worth to mention that even though these regions displayed a recombination rate one order of magnitude above other regions, they still presented low nucleotide diversity (π), which suggests that recombination rate is not given by nucleotide diversity (Table 1). Moreover, the correlation between per site mutation rate and recombination rate was not significant ($\rho_{spearman} = 0.17$, p-value = 0.13) (Fig. S1). As an additional sanity check, in order to prove that the depicted signals are not due to artefacts derived from alignment difficulties, no correlation was observed between the mean recombination rate per region and the mean alignment quality, the latter calculated as the fraction of well aligned sequences (absence of gaps) in each position ($\rho_{spearman} = -0.035$, p-value = 0.94).

To further explore the impact of repetitive regions in more detail, all those annotated in NC_007605 reference genome, including type A-D repeat families, tandem, inverted repeats and terminal repeats, as well as internal repeats (IR1-4 annotated in AG876 genome), were considered for recombination rate estimation. Repetitions in tandem, that showed a recombination signal above the mean genome value, were located in EBNA family of genes, genes (genome position 87834 to 88043). As well, another class of repeats, namely type A, B, C and D repeat families within genome positions 81920–82781 bp are also among the most exposed regions to recombination. Tandem repeats in LMP1 gene also displayed a recombination signal, but below genome mean ρ/bp (168036 to 168134 bp). Otherwise than expected, no recombination signal was found when assessing inverted or internal repeats; however, it is important to highlight that the lack of signal in the these type of repeats may be due to the fact that they fall within genomic regions that display insufficient segregating sites to render them conclusive or just be the consequence of repetitive sequences having been trimmed out during genome *de novo* assembly in a proportion of the studied data set (Fig. 1A, nt [7500–25000]) (Ba Abdullah et al., 2017).

Further analysis, correlating GC content in the mentioned genomic regions and their mean ρ/bp value, failed to achieve a statistical significance between them ($\rho_{spearman} = 0$; p-value = 1; Spearman's rank correlation). In this way, the cases of the repeated regions and miscellaneous features categories may be highlighted; which despite having the greatest GC content, displayed the highest and lowest mean ρ/bp values, respectively (Table 1). Furthermore, there was no correlation between the GC content in each of the individually analyzed genes and its observed recombination rate ($\rho_{spearman} = -0.02$, p-value = 0.86; Spearman's rank correlation, data not shown). However, a word of caution should be spoken regarding to the role of high GC content in the recombination process, since it was previously suggested that high GC content and repetitive regions could be present in viral genomes with the putative function of increasing viral potential to recombine (Brown, 2014).

3.2. Recombination rate in viral protein-coding genes

Next, the effect of recombination signal distribution among all protein-coding genes was explored. Fig. 2 showed the mean recombination rate per bp for all 65 protein-coding genes that displayed a measurable signal. It is worth to mention that the mean ρ/bp value, allows for comparison among genes of different lengths, while the corresponding HPDI (error bars shown if Fig. 2) accounts for intragenic heterogeneity of ρ/bp signal. This means that in a gene with a large HPDI interval, more variability in ρ/bp signal is expected.

It could be noticed from Fig. 2, that EBNA3C and EBNA3B (members of the EBNA3 gene family which participates in attenuating DNA damage responses during EBV infection and transformation of naive B cells) (Wang et al., 2015), BRRF2 (tegument protein) (Lindsey, 2017) and BBLF2-BBLF3 (an accessory protein to the viral helicase-primase complex) (Thierry et al., 2015) displayed a recombination signal statistically higher than that observed for the entire genome. Statistical significance was considered when the gene's HPDI interval did not overlap with the whole genome's HPDI (gray area) in Fig. 2. Other genes such as BZLF2 (codes for viral glycoprotein gp42) (Rowe et al., 2011) and BZLF1 (viral master regulator protein involved in latent to lytic cycle switch) (Lorenzetti et al., 2014) also displayed mean ρ/bp values above genome average, but with overlapping HPDI. In a similar way, other 25 genes had a recombination rate similar to that of the entire genome. On the contrary, the remaining 34 genes displayed a lower recombination signal than that observed for the entire genome. Hence, the fact that the majority of genes had a mean ρ/bp value below genome average without an overlapping HPDI suggested that viral genes were relatively stable to recombination events, consistent with the genomic stability of herpesviruses. However, caution should be

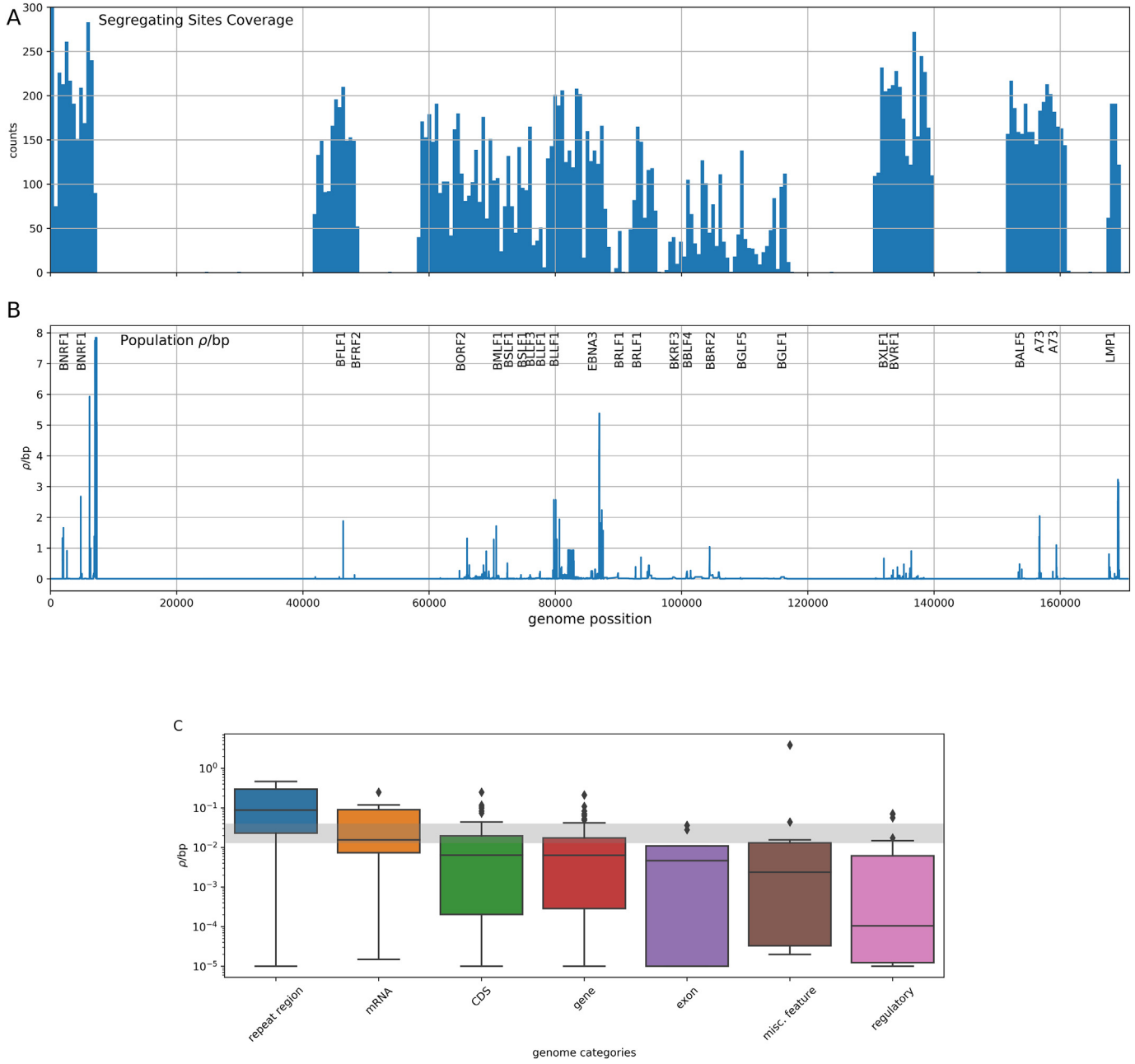

Fig. 1. EBV recombination map. (A) Positioning of segregating site along the entire EBV genome. (B) Recombination signal map along the entire EBV genome. Only regions with sufficient segregating sites were used to compute recombination signal. Only genes with signal spikes above genome average are shown over signal's peak. Nucleotide positions relative to NC_007605 reference genome. (C) Recombination rate distribution in EBV genomic categories; only those categories with > 10 detected events are shown. Gray stripe denotes the 95% HPDI for genome's mean ρ/bp value.

applied in cases such as for BKRF2 (virion glycoprotein) (Dolan et al., 2006), BFRF2 (early lytic protein and late promoter activator) (McKenzie et al., 2016) and BFLF1 (crucial for cleavage and packaging of viral particles) (McKenzie et al., 2016; Pavlova et al., 2013), which despite of a low mean ρ/bp value, have a wide HPDI. The large HPDI could mean that these genes have at least one region with a high ρ/bp value (*e.g.* BFLF1 peak in Fig. 1), while low or null in the rest of the gene.

Overall, our results demonstrated that some genes within the viral genome were more prone to suffer recombination than others (intergenomic heterogeneity). Moreover, the high variability of HPDI intervals across individual genes suggested that recombination signal was more heterogeneous in some genes than in others (intragenic heterogeneity).

3.3. Recombination-inducing DNA sequences

Given that Brown (2014) suggested the existence of specific DNA sequences related to the initiation of the recombination process, we extended our study to assess the occurrence of these DNA features in the present data set and their relation with each gene's recombination rate. Six recombination-initiating features were tested. Initiation events were considered when the fold-change in the recombination rate signal, (downstream/upstream, see methods for details), of each of the selected DNA motifs was greater than expected by chance. Moreover, this analysis was executed in a sense and anti-sense direction separately, so as to consider genetic elements coding in both DNA strands. Fig. 3A represents an example of a detected recombination signal change downstream of the CCCAG recombination-initiating sequence within

Table 1
Basic statistics by genomic category.

Category	π	Mean GC	Mean ρ/bp	Mean length
Repeat region	0.0327	0.6548	0.1613 [0.1052–0.219]	869
mRNA	0.056	0.5566	0.0488 [0.259–0.0788]	939
CDS	0.046	0.5807	0.0163 [0.0106–0.0255]	1000
Gene	0.0417	0.4486	0.0148 [0.0086–0.0309]	1290
Exon	0.0288	0.5844	0.0105 [0.006–0.0237]	233
Regulatory	0.025	0.1599	0.0089 [0.0043–0.0296]	5.58
Miscellaneous feature	0.0257	0.768	0.0078 [0.015–0.011]	524

π: Mean nucleotide diversity; Mean ρ/bp: Mean recombination rate per base pair and HPDI (highest posterior density interval) between brackets. CDS: Coding region. In order to avoid bias, outliers were assessed and removed. Only one outlier value was detected in CDS, one in miscellaneous features and one in the gene category, (*p*-value < 10 E-5, Smirnov Grubbs test).

EBNA3C gene.

Among the studied DNA sequences, CCCAG and TGGAG, which were well known recombination initiator sequences, preceded the greatest amount of changes in recombination rate signal (20 and 18 events, respectively); followed by GGGCT, a human immunoglobulin class-switch inducing sequence (14 events); and finally by AGGAG and TGGTGG (9 and 5 events, respectively) (Fig. 3B). Strikingly, although TGGAG was not the most frequently detected DNA motif across the EBV genome (697 hits), it still induced a significantly greater proportion of initiation events (0.025 events/hits) than other more represented motif, *e.g.* AGGAG (0.012 events/hits), *p*-value = 0.04; one tailed proportion test.

On the contrary, a classical meiotic recombination sequence in humans, CCTCCCT, was barely detected in the EBV genome (28 hits) and was not associated with changes in recombination rate signal.

When focusing the analysis in individual genes, and taking into consideration the direction of transcription, no significant differences in recombination rate signals were detected among the forward and reverse transcribing genes; not in the amount of genes presenting changes in ρ/bp (16 forward genes *vs* 13 reverse genes), nor in the distribution of signal change magnitude (*p*-value = 0.11, Kolmogorov-Smirnov one side test) (Fig. 3C and D).

In accordance with our previous results the CCCAG sequences was the most widely distributed, in both, forward (9/16) and reverse transcribing genes (7/13), followed by TGGAG sequence in (7/16) forward genes and (6/13) reverse genes (Fig. 3 C and D).

Considering each gene as a separate coding element, BBLF2-BBLF3, one of the genes with a recombination signal statistically higher than that observed for the entire genome (Fig. 2), contained the 5 recombination inducing sequences. Other two genes with high recombination rates were BRRF2 and EBNA3B, interestingly they both contained CCCAG sequence; moreover, EBNA3B additionally contained the GGGCT sequence. Even though EBNA3C, the last gene to show a recombination rate above genome average, did not contain any recombination-initiating sequence, its recombination potential may be contributed to by its internal repeats.

4. Discussion

Given the recent advances in NGS technologies it is now possible to obtain complete sequences of viruses with large genomes, like EBV. This technology has provided a new layer of analysis, facilitating the study of polymorphisms and structural alterations in a global genomic scale. In fact, recombination has been studied in herpesviruses such as HSV-1, VZV and HCMV (reviewed in (Renner and Szpara, 2018)). In particular, Lasalle et al. found that recombination is a widespread phenomenon in HCMV (Lassalle et al., 2016). On the other hand, this issue is more controversial in HSV-1. While Szpara et al. described

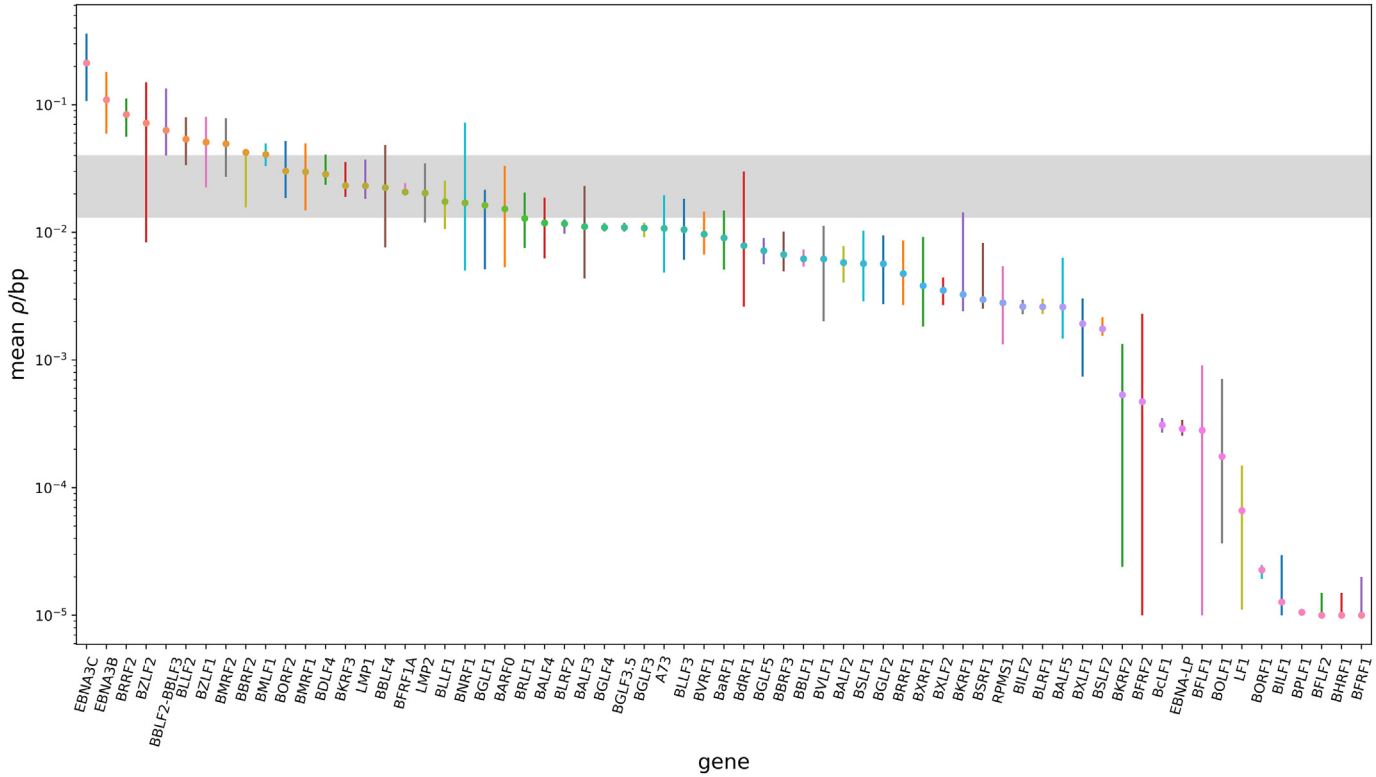

Fig. 2. Mean recombination rate per bp (ρ/bp) among protein-coding genes. Error bars indicate 95% HPDI (highest posterior density interval), Gray stripe denotes the 95% HPDI for genome's mean ρ/bp value.

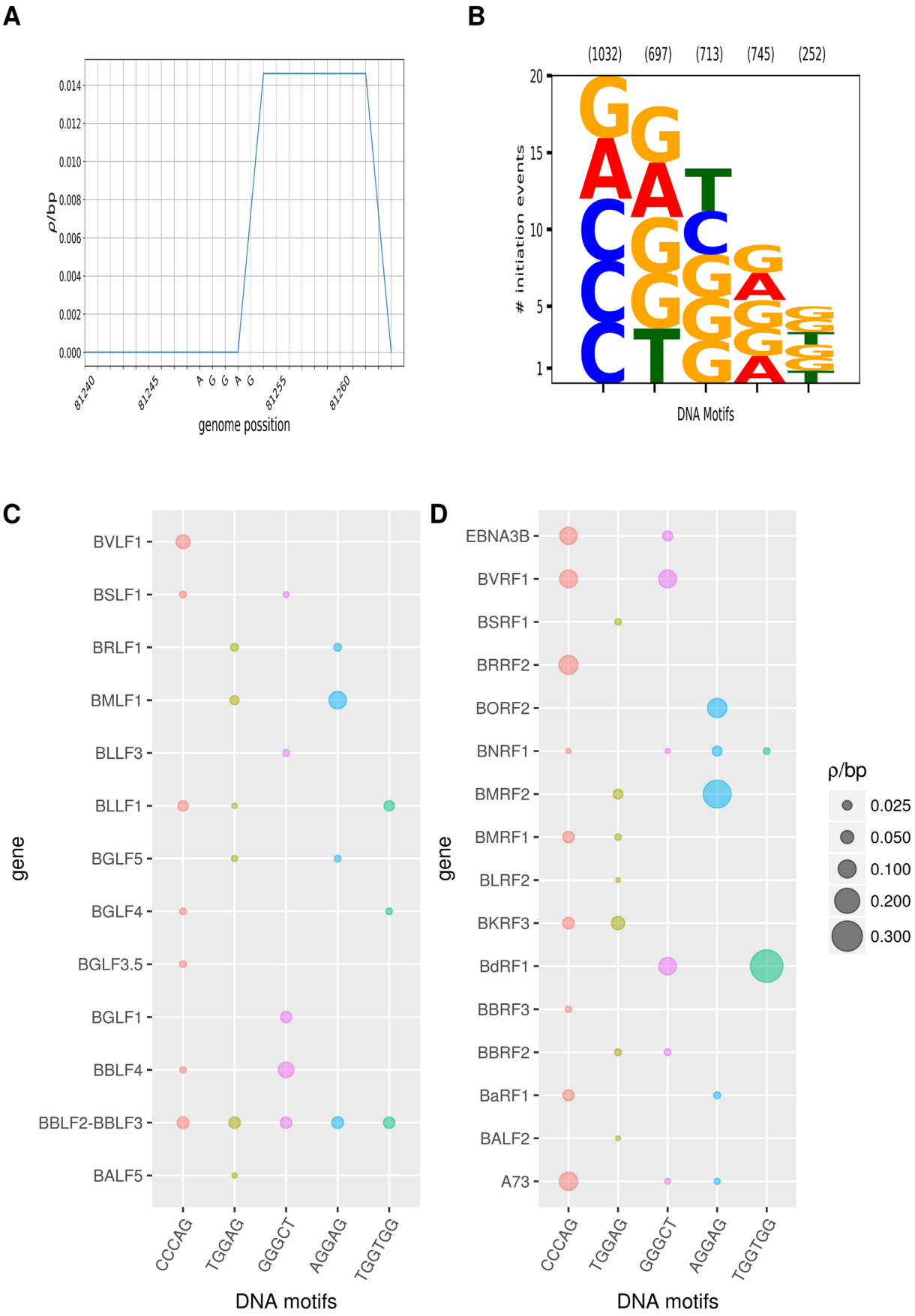

Fig. 3. Assessment of recombination-inducing DNA motifs on EBV genome. (A) Change in ρ/bp downstream of AGGAG motif in EBNA3C gene. Nucleotide positions are relative to NC_007605 reference genome. (B) Recombination rate change counts for the analyzed DNA motifs. Between parenthesis: number of detected hits across the EBV genome. Change in ρ/bp downstream of each DNA motif in reverse (C) and forward (D) transcribed genes. Circle size depicts the mean value of signal change around the DNA motif. Only statistically significant events are shown (p-value < 0.05, bootstrap two tailed test).

recombination as a genome widespread evolutionary event; Lee et al. demonstrated that breakpoints in HSV-1 are biased towards repetitive sequences, higher GC content areas and intergenic regions (Lee et al., 2015; Szpara et al., 2014). Recombination among clades was reported in VZV (Norberg et al., 2015), but a detailed analysis of genomic regions prone to recombination is still pending. In a similar fashion Palser et al. and Chiara et al. characterized the pattern of recombination in EBV by means of phylogenetic analysis, either by split decomposition or network analysis (Palser et al., 2015; Chiara et al., 2016). One first approach to calculate recombination rates in herpesviruses was performed by Bowden et al. on small segments of HSV-1 genome (Bowden et al., 2004); however, to date no study has quantified global and local recombination rates in the entire viral genome. Here we computed recombination rates along the entire EBV genome and performed a gene by gene comparison taking advantage of the increasing number of complete genomes generated by NGS.

As expected, only segregating sites provided a reliable signal to quantify recombination rates. In contrast to what would be expected by chance, the signal was not homogeneously distributed along the EBV genome, but appear as highly concentrated in certain genomic regions, in particular those with sequence repeats. This observation is consistent with the notion that repeated regions provide unstable genomic areas that are prone to undergo homologous recombination. On the other hand, regulatory regions such as “TATA boxes” or polyA sequences, which are vital for gene expression and mRNA translation, displayed low recombination rates, a fact that highlights the importance of these sequence conservation. Brown suggested that GC content and repeat regions might play a relevant role in recombination processes in herpesviruses (Brown, 2014); however GC content did not correlate with higher recombination rate values in the present study. Our result is consistent with that presented by Castelhano et al. who also failed to detect a relation between GC content and recombination in HBV, another DNA virus (Castelhano et al., 2017). Taken together our findings suggest that, besides GC rich sequences being proposed to increase recombination probability, features such as repeated regions are greater contributors to recombination events in EBV. It should be pointed out that about 20 repeated regions within the EBV genome, in a non-neglectable subset of 71 sequences used in our study were originally deleted by the submitting authors (Palser et al., 2015), given the major challenge that repetitive sequences suppose for the *de novo* assembly algorithms (Ba Abdullah et al., 2017). These repetitive sequences could have further contributed to changes in recombination rate signal and hence, increase the power of our observations; however, the lack these repetitive regions does not invalidate our findings. This could explain the absence of variation in recombination rate in long repetitive regions such as the internal repeat 1 region (BamW repeats). However, given that long repetitive regions are inherently difficult to assemble, our findings regarding these problematic features operate under the assumption that available genbank records were correctly assembled; supposing a possible limitation to our retrospective approach.

Another way to evidence the heterogeneous nature of the recombination process could be that only 4 genes presented a recombination rate signal above that of genome average, while the remaining 61 displayed either a similar or below-average recombination rate. Interestingly, two of the genes with recombination rate above average, EBNA3C and EBNA3B, both contain tandem repeats. Notably, these two genes were also shown to be prone to recombination by Palser et al. (Palser et al., 2015).

The other two genes with recombination rate above genome average, BRRF2 and BBLF2-BBLF3, contained recombination inducing DNA motifs precisely upstream of the recombination rate fold-change increase. These motifs were previously described by Brown in all herpesvirus but statistically overrepresented, as solely expected by chance, in gammaherpesviruses such as EBV (Brown, 2014). Combined, our results suggest that both tandem repeats and recombination-inducer DNA motifs could induce recombination events in EBV. Out of the six motifs described by Brown, only five were related to recombination induction in our study. Of notice, those two (CCCAG and TGGAG) to induce the greater amount of recombination signals were well characterized recombination initiator motifs in eukaryotic cells (Bengesser et al., 2010). The most frequent recombination-inducer DNA motif throughout the genome, CCCAG, induced the highest amount of overall changes in recombination rate and also elicited recombination rate changes in a broader amount of genes. Even though TGGAG was not the most frequent motif in the entire EBV genome, it still provided the greatest ratio of changes in recombination rate signal (recombination inducing events/hits along the genome). On the contrary, the meiotic recombination inducer CCTCCCT motif, although present, did not precede any changes in recombination rate. Given that this is a meiotic recombination inducer (Myers et al., 2005), it may be plausible to hypothesize that the motif was incorporated into the viral genome from the host cell DNA at some point during EBV and human co-evolution; but during viral replication in B lymphocytes, specific meiotic factors involved in DNA binding may be lacking. This idea could be further supported by the fact that the virus's principle recombinase, coded by the BALF2 gene, is regulated at the transcriptional level by Sp1, CREB and AP-1, the same cellular factors that regulate RAG-1, a cellular recombinase, and other viral genes. Both EBV replication and V(D)J recombination occur synchronously, during the G0/G1 phase in the nucleus of EBV infected lymphocytes when these undergo proliferation and differentiation (Dreyfus, 2009). These results highlight that it is not only important to characterize the presence of these DNA motifs in viral genomes, but also to assess their relation with a recombination-related estimator.

Our results provide the first EBV recombination map based on the assessments of recombination rates, both at a global and gene by gene level and support the theory proposed by Renner et al. (Renner and Szpara, 2018) that the diversity and evolution of herpesviruses are impacted by mechanisms, such as recombination, which extend beyond the usual consideration of point mutations introduced by viral polymerases. Increasing the knowledge on recombination as a driving force in EBV evolution may aid future analyses of genomic variations and their relation to geographical distribution, linkage to diseases and their outcomes as well as impact on clinical practice and vaccine development.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.07.022.

Acknowledgments

This study was supported in part by a grant from National Agency for Science and Technology Promotion (PICT 2016 N°0548) and (PBIT 2013 N°12). A.J.B., M.A.L, and M.V·P are members of the CONICET Research Career Program.

References

Arenas, M., Araujo, N.M., Branco, C., Castelhano, N., Castro-Nallar, E., Pérez-Losada, M., 2017. Mutation and recombination in pathogen evolution: relevance, methods and controversies. Infect. Genet. Evol. https://doi.org/10.1016/j.meegid.2017.09.029.

Auton, Adam, McVean, Gil, 2007. Recombination rate estimation in the presence of hotspots. Genome Res. 17 (8), 1219–1227.

Ba Abdullah, M.M., Palermo, R.D., Palser, A.L., Grayson, N.E., Kellam, P., Correia, S., et al., 2017. Heterogeneity of the Epstein-Barr Virus (EBV) major internal repeat reveals evolutionary mechanisms of EBV and a functional defect in the prototype EBV strain B95-8. J. Virol. 91 (e00920–17).

Bengesser, K., Cooper, D.N., Steinmann, K., Kluwe, L., Chuzhanova, N.A., Wimmer, K., et al., 2010. A novel third type of recurrent NF1 microdeletion mediated by nonallelic homologous recombination between LRRC37B-containing low-copy repeats in 17q11.2. Hum. Mutat. 31, 742–751 Wiley subscription services, Inc., A Wiley Company.

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.

Brown, J.C., 2014. The role of DNA repair in herpesvirus pathogenesis. Genomics 104, 287–294.

Bruen, T.C., Philippe, H., Bryant, D., 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172, 2665–2681.
Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.
Castelhano, N., Araujo, N.M., Arenas, M., 2017. Heterogeneous recombination among hepatitis B virus genotypes. Infect. Genet. Evol. 54, 486–490.
Chang, C.M., Yu, K.J., Mbulaiteye, S.M., 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.
Chiara, M., Manzari, C., Lionetti, C., Mechelli, R., Anastasiadou, E., Chiara Buscarinu, M., et al., 2016. Geographic population structure in Epstein-Barr virus revealed by comparative genomics. Genome Biol Evol. 8, 3284–3291.
Chuzhanova, N., Chen, J., Bacolla, A., Patrinos, G.P., Férec, C., Wells, R.D., et al., 2009. Gene conversion causing human inherited disease: evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. Hum. Mutat. 30, 1189–1198 Wiley Subscription Services, Inc., A Wiley Company.
Correia, S., Palser, A., Elgueta Karstegl, C., Middeldorp, J.M., Ramayanti, O., Cohen, J.I., et al., 2017. Natural variation of Epstein-Barr virus genes, proteins, and primary MicroRNA. J. Virol. 91. https://doi.org/10.1128/JVI.00375-17.
Cullen, M., Perfetto, S.P., Klitz, W., Nelson, G., Carrington, M., 2002. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am. J. Hum. Genet. 71, 759–776.
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.
Dreyfus, D.H., 2009. Paleo-immunology: evidence consistent with insertion of a primordial herpes virus-like element in the origins of acquired immunity. PLoS ONE 4, e5778.
Galli, A., Bukh, J., 2014. Comparative analysis of the molecular mechanisms of recombination in hepatitis C virus. Trends Microbiol. 22, 354–364.
Gantuz, M., Lorenzetti, M.A., Chabay, P.A., Preciado, M.V., 2017. A novel recombinant variant of latent membrane protein 1 from Epstein Barr virus in Argentina denotes phylogeographical association. PLoS ONE 12, e0174221.
Guiretti, D.M., Chabay, P.A., Valva, P., Stefanoff, C.G., Barros, M.H.M., De Matteo, E., et al., 2007. Structural variability of the carboxy-terminus of Epstein–Barr virus encoded latent membrane protein 1 gene in Hodgkin's lymphomas. J. Med. Virol. 79, 1722–1730.
Katoh, K., Standley, D.M., 2013. MAFFT: Iterative Refinement and Additional Methods. Methods in Molecular Biology. pp. 131–146.
Lassalle, F., Depledge, D.P., Reeves, M.B., Brown, A.C., Christiansen, M.T., Tutill, H.J., et al., 2016. Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus Evol. 2 (vew017).
Lee, K., Kolb, A.W., Sverchkov, Y., Cuellar, J.A., Craven, M., Brandt, C.R., 2015. Recombination analysis of herpes simplex virus 1 reveals a Bias toward GC content and the inverted repeat regions. J. Virol. 89, 7214–7223.
Lindsey, J.W., 2017. Antibodies to the Epstein-Barr virus proteins BFRF3 and BRRF2 cross-react with human proteins. J. Neuroimmunol. 310, 131–134.
Lorenzetti, M.A., Gantuz, M., Altcheh, J., De Matteo, E., Chabay, P.A., Preciado, M.V., 2014. Epstein–Barr virus BZLF1 gene polymorphisms: malignancy related or geographically distributed variants? Clin. Microbiol. Infect. 20, O861–O869.
Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., Varsani, A., 2011a. Recombination in eukaryotic single stranded DNA viruses. Viruses 3, 1699–1738.
Martin, D.P., Lemey, P., Posada, D., 2011b. Analysing recombination in nucleotide sequences. Mol. Ecol. Resour. 11, 943–955.
Martin, D.P., Murrell, B., Golden, M., Khoosal, A., Muhire, B., 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 1 (vev003).
McKenzie, J., Lopez-Giraldez, F., Delecluse, H.-J., Walsh, A., El-Guindy, A., 2016. The Epstein-Barr virus Immunoevasins BCRF1 and BPLF1 are expressed by a mechanism independent of the canonical late pre-initiation complex. PLoS Pathog. 12, e1006008.
Myers, S., Bottolo, L., Freeman, C., McVean, G., Donnelly, P., 2005. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome Science. 310. American Association for the Advancement of Science, pp. 321–324.
Norberg, P., Depledge, D.P., Kundu, S., Atkinson, C., Brown, J., Haque, T., et al., 2015. Recombination of globally circulating varicella-zoster virus. J. Virol. 89, 7133–7146.
Palser, A.L., Grayson, N.E., White, R.E., Corton, C., Correia, S., Ba Abdullah, M.M., et al., 2015. Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. J. Virol. 89, 5222–5237.
Pavlova, S., Feederle, R., Gärtner, K., Fuchs, W., Granzow, H., Delecluse, H.-J., 2013. An Epstein-Barr virus mutant produces immunogenic defective particles devoid of viral DNA. J. Virol. 87, 2011–2022.
Pérez-Losada, M., Arenas, M., Galán, J.C., Palero, F., González-Candelas, F., 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. Infect. Genet. Evol. 30, 296–307.
Renner, D.W., Szpara, M.L., 2018. Impacts of genome-wide analyses on our understanding of human herpesvirus diversity and evolution. J. Virol. 92. https://doi.org/10.1128/JVI.00908-17.
Rowe, C.L., Matsuura, H., Jardetzky, T.S., Longnecker, R., 2011. Investigation of the function of the putative self-association site of Epstein-Barr virus (EBV) glycoprotein 42 (gp42). Virology 415, 122–131.
Szpara, M.L., Gatherer, D., Ochoa, A., Greenbaum, B., Dolan, A., Bowden, R.J., et al., 2014. Evolution and diversity in human herpes simplex virus genomes. J. Virol. 88, 1209–1227.
Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.
Thierry, E., Brennich, M., Round, A., Buisson, M., Burmeister, W.P., Hutin, S., 2015. Production and characterisation of Epstein-Barr virus helicase-primase complex and its accessory protein BBLF2/3. Virus Genes 51, 171–181.
Tzellos, S., Farrell, P., 2012. Epstein-Barr virus sequence variation—biology and disease. Pathogens 1, 156–174.
Wang, A., Welch, R., Zhao, B., Ta, T., Keleş, S., Johannsen, E., 2015. Epstein-Barr virus nuclear antigen 3 (EBNA3) proteins regulate EBNA2 binding to distinct RBPJ genomic sites. J. Virol. 90, 2906–2919.
Young, L.S., Yap, L.F., Murray, P.G., 2016. Epstein–Barr virus: more than 50 years old and still providing surprises. Nat. Rev. Cancer 16, 789–802.