Submission: Article

Title: Co-evolution of sites under immune selection shapes Epstein-Barr Virus population structure

Authors: Fanny Wegner[1,2+], Florent Lassalle[3,4], Daniel P. Depledge[1*], François Balloux[3], Judith Breuer[1+]

Affiliations:

[1]Division of Infection & Immunity, University College London, London, UK

[2]Microbial Evolutionary Genomics, Institut Pasteur, Paris, France

[3]UCL Genetic Institute, University College London, London, UK

[4]MRC Centre for Outbreak Analysis and Modelling, Imperial College, London, UK.

[*]present address: Department of Microbiology, New York University School of Medicine, New York, USA

[+]Corresponding authors: j.breuer@ucl.ac.uk, fanny.wegner@pasteur.fr

Abstract:

Epstein-Barr virus (EBV) is one of the most common viral infections in humans and persists within its host for life. EBV therefore represents an extremely successful virus that has evolved complex strategies to evade the host's innate and adaptive immune response during both initial and persistent stages of infection. Here, we conducted a comparative genomics analysis on 223 whole genome sequences of world-wide EBV strains. We recover extensive genome-wide linkage disequilibrium (LD) despite pervasive genetic recombination. This pattern is explained by the global EBV population being subdivided into three main sub-populations, one primarily found in East Asia, one in Southeast Asia and Oceania, and the third including most of the other globally distributed genomes we analyzed. Additionally, sites in LD were overrepresented in immunogenic genes. Taken together, our results suggest that host immune selection and local adaptation to different human host populations has shaped the genome-wide patterns of genetic diversity in EBV.

Introduction:

Epstein-Barr virus (EBV) is a member of the gamma-herpesviruses family, present in most humans worldwide. Primary infection is either symptomless or causes infectious mononucleosis (IM), and is followed by lifelong latent infection within the memory B cell pool. EBV has been associated with a variety of cancerous diseases of B cell origin, such as endemic Burkitt's lymphoma (BL), Hodgkin's lymphoma (HL) and post-transplant lymphoproliferative disorder (PTLD), cancers of epithelial origin including nasopharyngeal carcinoma (NPC) and gastric carcinoma and even in rare cases NK- and T-cell tumours (Young et al. 2016). Recent evidence also points towards an involvement in autoimmune diseases such as multiple sclerosis and systemic lupus erythematosus (Lossius et al. 2012; Pender and Burrows 2014). Development of disease is associated with various factors, such as immune status (e.g. PTLD, HIV-related lymphoma), co-infections (e.g. endemic BL and Malaria, HL and HIV) and geography with EBV-positive NPC particularly prevalent in adults from Southern China and Northern Africa, and endemic BL in children from equatorial Africa.

The double-stranded DNA genome of EBV has a length of around 172 kb and contains at least 94 annotated open reading frames (ORF). It usually resides as a circular, double-stranded DNA molecule in the nucleus. Previous whole genome sequencing analyses have focused on geographically related strains and provided evidence for extensive recombination (Kwok et al. 2014; Palser et al. 2015). Palser et al. (2015) reported two cases of inter-typic recombinants, but also presented evidence for multiple recombination events throughout the genome. This latter observation crucially impacts how EBV ancestry can be studied as recombination events are expected to affect tree topology and can render inference derived from phylogenetic approaches largely meaningless (Rieux and Balloux 2016).

Because pervasive recombination in EBV prevents us from inferring a single, genome-wide evolutionary history, we dissected the population structure of EBV genomes at the level of single nucleotides or genes. We adapted and applied a recently developed approach (Lassalle et al. 2016) based on genome-wide patterns of linkage disequilibrium to a dataset comprising 223 EBV genomes collected from all around the world. Specifically, our analysis focuses on how sequence variation, recombination and linkage have shaped the global population structure of EBV and may have influenced its evolution.

Results

EBV genome sequences are highly recombinant

We utilised a dataset comprising 223 type 1 EBV whole genome sequences that have previously been published (de Jesus 2003; Zeng et al. 2005; Liu et al. 2011; Kwok et al. 2012; Lin et al. 2012 Nov 14; Lei et al. 2013; Tsai et al. 2013; Kwok et al. 2014; Palser et al. 2015; Chen et al. 2018; Hui et al. 2018; Correia et al. 2018; Bridges et al. 2019). These samples are representative of diverse geographical regions, body compartments, and malignancies (suppl. table 1).

We first examined our dataset for evidence of recombination. A PHI-test (Bruen et al. 2006) found global evidence for recombination ($p < 0.05$) and a genome-wide PHI-profile scan revealed areas of significant recombination throughout the genome (suppl. fig. 1). The presence of numerous reticulations in a recombination network confirms this (suppl. fig. 2). Consequently, genetic recombination cannot be ignored, thus precluding the use of phylogenetic inference from whole genome sequences.

Evidence of genome-wide linkage disequilibrium despite widespread recombination

A useful way to assess recombination on a larger scale is to consider linkage disequilibrium (LD), i.e. the correlation between the occurrence of polymorphisms at different loci in the genome (Haydon et al. 2004). Two loci are considered to be in LD when they occur together more often than would be expected by chance under a uniform distribution of allele combinations given their respective frequencies. There are several factors influencing LD, including physical proximity, the rate of recombination, natural selection, and population structure. For a given recombination rate, the likelihood of a recombination event is inversely proportional to the physical distance between a pair of loci, a negative relationship is expected between the LD between two bi-allelic sites and the physical distance separating them.

Genome-wide LD was assessed for all combinations of bi-allelic sites using Fisher's Exact test (significant if $p < 0.05$ after Bonferroni correction). Fig. 1A shows a map of LD between bi-allelic sites. In total, 253,935 pairs of sites were found to be in LD, which represent 2,752 individual sites out of the 9,822 bi-allelic sites analysed. There are 242,372 pairs with at least one site being located in an open reading frame (ORF) and 165,407 pairs where both sites fall within ORFs (2,229 unique sites). Of these, there are 41,587 pairs where allelic variation at both sites is synonymous, 81,582 pairs with at least one nonsynonymous site and 42,238 pairs where allelic variation at both sites correspond to nonsynonymous changes.

LD was essentially independent of physical distance in the genome between linked sites (suppl. fig. 3), i.e. with similar distributions of p-values over many distance classes, although proximal sites showed a lower distribution of p-values. The detection of LD throughout the genome even when sites are distal (fig. 1A) is somewhat counterintuitive given the evidence of pervasive recombination. Focusing on subsets of sites that are in LD with at least one other site (all sites in LD, nonsynonymous sites in LD, and synonymous sites in LD), recombination networks (suppl. fig. 4) and PHI-test still gave evidence for recombination occurring within all subsets ($p < 0.05$). Fig. 1B shows an association network of sites linked with each other. Each site is represented by a circle, while connections are drawn between them if they are in LD. We recovered 62 components, i.e. subnetworks in which every pair of sites (nodes) is connected by a path. The components consisted of a large network comprising the majority of sites (2,501 out of 2,752, fig. 1C), and smaller sets and pairs of independently linked SNPs (fig. 1B-C). However, even the largest component displays evidence for recombination (suppl. fig. 5). For the subsequent analyses we focussed on the largest component of the network, referred to as the major component, to provide a majority, non-chimaeric representation of the genome's linkage structure.

Genome-wide LD can be explained by population structure

One possible explanation for the pattern of linkage is the influence of population structure whereby apparent co-inheritance of bi-allelic sites simply reflects the independent segregation of different alleles in isolated populations. We used the program Admixture (Alexander et al. 2009) to cluster individuals into populations. We found no evidence for genetic subdivision by body compartment and malignancy (results not shown).

Using all bi-allelic sites (fig. 2A), a very striking top-level structure could be uncovered when assuming three subpopulations. Almost all sequences originating from Asia and Oceania were assigned to two clusters (C2 and C3, dark red and orange), while the majority of African, European, North and South American as well as Australian isolates belong to a third cluster (C1, blue), suggesting the existence of two separate virus populations in East Asia and in the Pacific, as well as a third separate population of viruses spread throughout the rest of the world (as represented by the data set). A number of sequences could not be unambiguously assigned to a single cluster and are likely of admixed ancestry.

Restricting the dataset to all bi-allelic sites in LD (major component) (fig. 2B) did not change the proportional assignment of isolates to populations. The same remains true when the analysis was restricted to the subset of pairs on nonsynonymous and synonymous sites in LD (fig. 2C-D). By contrast, bi-allelic sites not in LD with any other site do not show evidence for a clear population structure corresponding to any geographic pattern, as the assignment to C3 and C1 is noisy across genomes (fig. 2E), with the exception of three samples from Oceania (all from Papua New Guinea). These three sequences are on very long branches in the recombination network (suppl. fig. 2) indicating they are very distant from the rest of the dataset. The sites responsible for these long branches were probably not detected as being in LD due to the low number of sequences in which they occur.

This top-level view does not capture the full complexity of the EBV population structure. We inferred that clustering the isolates into 20 subpopulations best describes the data (suppl. fig. 6). This leads to an increasingly finer structure in each of the geographical groups. There is a large overlap in assignment between sequences from Europe and Australia as well as some from North America. By contrast, there are a few clusters formed only by African sequences, highlighting the subtler differences between genomes from Africa and Europe/Australia. Similarly, a finer structure within the Southeast Asian, East Asian and Oceanian sequences emerges. For example, there are distinct subpopulations within China and Hong Kong, and one subpopulation frequently associated with Japan. While several of these additional subpopulations are comprised of unadmixed individuals, a large number of sequences are strongly admixed, confirming the persistent signal of recombination even within the subset of sites in LD (suppl. fig. 4-5).

The population structure of EBV is linked to immune genes

Selection acts primarily on polymorphisms resulting in amino acid changes, and selection for co-functional substitutions in proteins could explain the observed LD pattern. In this context, it is interesting to note that the proportion of nonsynonymous sites in linkage increases with the strength of LD (fig. 3A). To determine which genes are preferentially linked to each other, the data was restricted to nonsynonymous sites. When examining which genes contained sites that are most often found in LD (number of SNPs in LD), the genes in the top 1% of gene pairs were *BPLF1*, *BOLF1*, *BLLF1*, *EBNA3A-C*, *EBNA1*, *BcRF1*, and *LMP1* (suppl. fig. 7). Interestingly, seven of these nine ORFs are known to encode antigens. The other two genes (*BcRF1* and *BPLF1*) do not fulfil our conservative criterion of carrying at least two independent records of experimentally confirmed epitopes.

This led us to hypothesise that adaptation to the host immune system and maintenance of a variation at a specific subset of sites might have played a role in shaping the global population structure of EBV. To test for this, the dataset of genes was divided in two: genes that are known to code for immunogenic (IG) proteins and non-immunogenic (NIG), for which there is no current record of experimentally confirmed epitopes (table 1). Nonsynonymous sites within these ORFs belonging to IG are more often in LD with each other than would be expected if a uniform distribution of links across all genes is assumed ($p < 2.2e-16$, Chi-square test), even when excluding links between proximal SNPs (fig. 3B-C). Conversely, genes belonging to NIG are less often linked with each other than expected by chance.

As sites are linked with each other across the whole genome, i.e. SNPs (and ORFs) are not only linked to one but to several other SNPs (and ORFs), we sought to study this interconnectedness with a graph theoretical approach at the level of individual genes. The resulting gene network consisted of 74 genes, 32 of them belonging to IG and 41 belonging to NIG, respectively. Edges were weighted based on a linkage score. This linkage score was significantly higher for edges between genes both belonging to IG (Mann-Whitney, $p = 0.005$ for IG-IG vs. NIG-NIG, and $p = 3.6e-5$ for IG-IG vs. NIG-IG, respectively; suppl. fig. 8). Structurally, this network is connected, i.e. any node can be reached by any other node through one or more edges (there are no disconnected components). However, structure within the network in terms of clustering seems to be low.

Identifying the most important ORFs in a network can be done by ranking nodes based on their properties. Eigenvector centrality does this by measuring the influence of a node, i.e. a node's score is higher if it is connected to other high-scoring nodes (fig. 4). Of the top 25 highest ranked genes, 11 belong to the IG group (table 2). An additional five genes within the top 25 also appear in the IEDB database as antigens, but did not meet our conservative criterion of a minimum of two independent records as experimentally confirmed antigens. In total, 36 of the 94 annotated ORFs in the EBV genome contain experimentally confirmed epitopes that fulfil this criterion. Of those, 32 are represented in the linked gene network (table 1), and 11/32 IG nodes are within the top 25 highest ranked genes.

Discussion

Recombination plays an important role in viral evolution as a source of genetic diversity. By combining mutations that previously appeared separately in different genomes, recombination allows the creation of new haplotypes. Genome-wide recombination has been described in detail for all well-studied herpesviruses (Norberg et al. 2007; Smith et al. 2013; Norberg et al. 2015; Lassalle et al. 2016; Koelle et al. 2017). It has been best studied at the molecular level in Herpes simplex virus 1 (HSV-1) where it is part of the replication process (Wilkinson and Weller 2003). However, recombination has also been proposed as a mechanism for herpesviruses to maintain the integrity of viral genomes during latency (Wilkinson and Weller 2004; Brown 2014).

Here we show that despite extensive recombination, EBV retains considerable population structure with evidence for extensive genome-wide linkage disequilibrium. These findings differ from those observed for Human cytomegalovirus (HCMV) (Lassalle et al. 2016) and HSV-1 (Lassalle, Beale et al., unpublished data) both of which appear to be essentially freely recombining with only localised areas of linkage disequilibrium. The extent to which free recombination can occur in different herpesviruses might be due to their employment of the host's homologous recombination (HR) systems (Brown 2014).

In alpha-herpesviruses such as HSV-1 and Varicella zoster virus, inverted and tandem repeats are the most prevalent HR initiating sequences and these are enriched in the unique short (U_S or S) segment of the genome. By contrast, HR initiating sequences in gamma-herpesviruses like EBV have been shown to be specific short GC-rich sequences that are evenly distributed across the genome (Brown 2014). Additional studies are required to better understand these apparent differences in LD and recombination across different herpesvirus families. Other potential causes for disparities in the inferred recombination intensity between these viral species may be driven by their opportunity to recombine and the distribution of the genetic divergence of recombining viral strains.

In the presence of extensive genetic recombination, LD is generally driven by an underlying population structure, which in turn might reflect biological or environmental constraints (McVean et al. 2002), or genetic drift within geographically segregated populations. In the case of EBV, population structure analyses identified a complex structure which largely corresponds on the top-level to three geographic groups: East Asia, Oceania (represented by Papua New Guinea) & Southeast Asia (represented by Indonesia), and the rest of the world.

This structure is primarily supported by the sites in LD, with the exception of three very divergent sequences from Oceania (suppl. fig. 2) for which there is a signal of structure even in the set of SNPs not in LD. On a finer scale, EBV population structure is naturally more complex. In our analysis, we found 20 subpopulations to best describe our data. Similarly, a previous study with a smaller, partly different data set (including Mediterranean and type 2 sequences) described ten subpopulations (Chiara et al. 2016). Both works show a correlation between geography and subpopulation, in particular with distinct subpopulations present in Africa, as well as overlapping subpopulations across the world, e.g. Europe/Australia and America/Africa. However, our study also highlights the presence of distinct subpopulations within Asia (and Oceania), which was previously not described.

Our data confirm previous findings where Asian and Indonesian sequences cluster quite distinctly from other genomes in a PCA (Palser et al. 2015; Correia et al. 2018). A few samples have been assigned to different clusters than the majority of genomes from the same region. For example, one genome from Hong Kong and one genome from Indonesia have been completely assigned to C1 (associated with non-Asian and non-Oceanian sequences). Similarly, four sequences from Indonesia, one sequence each from the US and from Brazil, as well as five UK sequences seem to belong to C2 (associated with East-Asia). However, the geographic label of the sequences is based on where they have been isolated and does not necessarily reflect the actual evolutionary origin of the virus genotype. There are only a handful of sequences isolated in the UK, for which the geographic origin of the donor is given (mentioned in brackets in fig. 2 and suppl. table 1) and where the assignment to different clusters than C1 is traceable. It is easily imaginable that a sequence isolated in the UK or the US, countries with mixed ethnic populations, could originally be an Asian strain, as primary infection often occurs through close family members in early age (Hjalgrim et al. 2007).

Recombination is also occurring within these subpopulations, with evidence for recombination within the subset of sites in LD (PHI-test, $p < 0.05$). Within-population recombination is also supported by the presence of admixed individuals in the population structure assignment, both for the top-level structure as well as the fine-resolution population structure analysis into 20 subpopulations (suppl. fig. 7). HKNPC2, for example, has been described as recombinant of HKNPC7 and -9 (Kwok et al. 2014), all isolates from Hong Kong that have been clearly assigned to the East Asian population (Palser et al. 2015).

In contrast to other herpesviruses, and despite this ever-present signal of recombination between and within subpopulations, EBV largely maintains its population structure. Interestingly, the proportion of nonsynonymous sites increases with the strength of LD (fig. 3A). This finding suggests that the pattern of genetic linkage is driven by natural selection on encoded proteins, notably preserving combination of residues determining how proteins functionally interact with each other. This hypothesis is compatible with the correlation of higher strength of LD with the increase in fraction of non-synonymous bi-allelic sites and could indicate that synonymous sites are less likely to be constrained to co-evolve; instead, the existence of stronger LD within the 2-kb range (suppl. fig. 3) suggests that synonymous sites may be hitch-hiking with physically linked sites, i.e. only contingently segregating with closely located non-synonymous sites under selection.

The association of linked polymorphism distribution with geography or ethnicity might reflect different biological constraints in each viral subpopulation. We thus investigated the possibility that important protein-protein interactions (PPI) could be determining EBV population structure. By representing the genes with the strongest LD in a linked gene network (fig. 4), we were able to identify the 25 most important nodes via an Eigenvector centrality-based ranking and compare them with data on PPIs previously described for EBV (Calderwood et al. 2007; Fossum et al. 2009) (suppl. fig. 9). While a few recovered interactions relate to known interactions (e.g. interaction between tegument and envelope proteins BPLF1, BALF4 and BOLF1), no straightforward hypothesis of biological cause can be proposed for others, due to their primary expression occurring in different stages of the life cycle as well as the current evidence for their localisation within the cell or virion, for example between the proteins encoded by BDLF3 (a glycoprotein expressed late in lytic cycle) and EBNA3A (which is located in the nucleus and expressed during latency). Nevertheless, PPI might explain some of the sites in LD observed in non-immunogenic genes.

Since PPI data based on yeast two hybrid screen is known to have a high false positive rate (Deane et al. 2002), we looked for other explanations for gene associations and considered a possible role of variation in host genetic makeup. Host genetics has been previously suggested to shape pathogen population structure, for example in HIV (Kløverpris et al. 2016)*, Mycobacterium tuberculosis* (Gagneux et al. 2006; Gagneux 2012) or *Helicobacter pylori* (Thorell et al. 2017). However, it is challenging to disentangle the effect of the demography of the pathogen and its host(s) from local adaption of a pathogen to its different host populations and their wider environment. It is probably fair to state that so far there is no case where host genetics could be uncontroversially identified as the driver for the apportionment of genetic diversity in a pathogen. Moreover, none of the previous putative cases of within-species host adaptations invoked selective pressures at such a large number of variants spread throughout the genome of a pathogen. Our results on EBV are also to the best of our knowledge the first case of such widespread gene-by-gene epistasis.

In EBV, the number of LD links between genes encoding proteins that are targets of adaptive immunity was significantly higher than between proteins with no adaptive immune function (fig. 3B-C). Moreover, 11 out of the 25 most important linked genes code for protein sequences that are the target of adaptive immunity. However, some of the genes we classify as non-immunogenic might in fact also contain epitopes that have not been described so far. The results raise the possibility that the EBV population structure may to a large extent have been shaped by host immunity, perhaps because the virus has adapted to HLA alleles common in the subpopulation in which it is circulating. The data is in concordance with a model of non-overlapping combinations of epitope regions, that are being held in LD despite genetic exchange via recombination between pathogens in other parts of the genome (Gupta et al. 1996).

In summary, we find that EBV retains a strong population structure in the face of considerable recombination and that this population structure is geographically stratified. The maintenance of the viral population structure may be partly driven by intrinsic viral co-adaptation of genes. Though, the evidence that genes in strongest LD are enriched in immunogenic genes suggests that adaptive immune selection, likely HLA mediated, has played a significant role in the maintenance of epistasis within this population. This raises the intriguing possibility that host genetic factors of the human populations in which the virus subpopulations have been circulating have been shaping the global population structure of EBV through local adaptation to its local human host populations. Irrespective of the underlying evolutionary forces, our findings starkly distinguish EBV from other human herpesviruses such as HCMV, which has been shown to be essentially freely recombining (Lassalle et al. 2016).

Methods

Dataset

The dataset consisted of 223 type 1 EBV whole genome sequences available from Genbank (suppl. table 1), comprising samples from various geographical regions and malignancies. The selection of sequences is explained in the supplementary methods.

Sequence analysis

Multiple sequence alignments were obtained using mafft v7.407 (Katoh and Standley 2013) and manually corrected to reduce unnecessary gaps generated around short tandem repeat sequences. SNPs were called based on differences to the reference genome NC_007605 (B95-8) using the R packages adegenet (Jombart 2008) and ape (Paradis et al. 2004).

Recombination and Linkage Disequilibrium (LD)

A test for the presence of recombination (PHI-test) was performed on the whole genome alignment excluding gapped sites with PhiPack under default parameters (Bruen et al. 2006). Additionally, a profile of PHI test p-values was computed along the genomes with sliding window of 1,000bp and a step size of 25bp. Split networks were generated using Splitstree4 (Huson and Bryant 2006).

For the genome wide analysis of linkage, the whole genome alignment of type 1 sequences was restricted to its 9,822 bi-allelic sites where maximally five sequence were missing. Linkage between all possible combinations of bi-allelic SNPs was tested with Fisher's Exact test, with a pair of SNPs being significantly associated if $p < 0.05$ after Bonferroni correction. Based on all linked sites, an association network was constructed with igraph (Csárdi and Nepusz 2006).

Split networks were generated from subsamples of aligned sites: all SNPs, all linked SNPs (sites in LD with at least one other site), all linked sites of the biggest component of the association network, all nonsynonymous linked SNPs, and all synonymous linked SNPs.

Population structure

The population structure was analysed with Admixture 1.3.0 (Alexander et al. 2009). It was run ten times for every *k* ranging from 1 to 40 with the cross-validation option to infer the number of clusters *k* that best fits the data. Replicate runs were further processed using CLUMPAK (Kopelman et al. 2015) and results for the major modes were visualised with

Distruct (Rosenberg 2004).

Gene linkage network

A network was constructed and analysed with the R package igraph (Csárdi and Nepusz 2006), where each node represents an ORF. An edge between two nodes was drawn if there was at least one pair of nonsynonymous SNPs in LD between them. The edges were weighted by a linkage score based on the number of linked nonsynonymous SNPs between two ORFs, normalised by the sum of ORF lengths (expressed as a fraction of the genome length).

The nodes of the network were divided into two sets of nodes, a) those ORFs known to encode antigens (immunogenic, IG) and b) those that do not (non-immunogenic, NIG). This classification was based on the experimentally confirmed antigens found in the IEDB database (http://www.iedb.org, April 2019)(Vita et al. 2018), with restriction to those antigens whose epitopes had been confirmed by at least two studies.

Funding

This work was supported by Agilent as part of the UCL Impact PhD stipend (FW) and the Rosetrees Trust (A501), the European Research Council (ERC) (grant ERC260801—BIG_IDEA to FB). JB receives funding from UCL/UCLH Biomedical Research Centre. We thank the BRC/MRC Pathogen Genomics Unit.

Bibliography

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. doi:10.1101/gr.094052.109.

Bridges R, Correia S, Wegner F, Venturini C, Palser A, White RE, Kellam P, Breuer J, Farrell PJ. 2019. Essential role of inverted repeat in Epstein-Barr virus IR-1 in B cell transformation; geographical variation of the viral genome. Philos Trans R Soc B Biol Sci. doi:10.1098/rstb.2018.0299.

Brown JC. 2014. The role of DNA repair in herpesvirus pathogenesis. Genomics. doi:10.1016/j.ygeno.2014.08.005.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics. 172(4):2665–81. doi:10.1534/genetics.105.048975.

Calderwood MA, Venkatesan K, Xing L, Chase MR, Vazquez A, Holthaus AM, Ewence AE, Li N, Hirozane-Kishikawa T, Hill DE, et al. 2007. Epstein-Barr virus and virus human protein interaction maps. Proc Natl Acad Sci U S A. doi:10.1073/pnas.0702332104.

Chen JN, Zhou L, Qiu XM, Yang RH, Liang J, Pan YH, Li HF, Peng GR, Shao CK. 2018. Determination and genome-wide analysis of Epstein-Barr virus (EBV) sequences in EBV-associated gastric carcinoma from Guangdong, an endemic area of nasopharyngeal carcinoma. J Med Microbiol. 67(11):1614–1627. doi:10.1099/jmm.0.000839.

Chiara M, Manzari C, Lionetti C, Mechelli R, Anastasiadou E, Buscarinu MC, Ristori G, Salvetti M, Picardi E, D'Erchia AM, et al. 2016. Geographic population structure in Epstein-Barr Virus revealed by comparative genomics. Genome Biol Evol.(1):evw226. doi:10.1093/gbe/evw226.

Correia S, Bridges R, Wegner F, Venturini C, Palser A, Middeldorp JM, Cohen JI, Lorenzetti MA, Bassano I, White RE, et al. 2018. Sequence variation of Epstein-Barr virus: viral types, geography, codon usage and diseases. J Virol. doi:10.1128/JVI.01132-18.

Csárdi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Syst. doi:10.3724/SP.J.1087.2009.02191.

Deane CM, Salwiński Ł, Xenarios I, Eisenberg D. 2002. Protein Interactions Two Methods for Assessment of the Reliability of High Throughput Observations. Mol Cell Proteomics. doi:10.1074/mcp.M100037-MCP200.

Fossum E, Friedel CC, Rajagopala S V., Titz B, Baiker A, Schmidt T, Kraus T, Stellberger T, Rutenberg C, Suthram S, et al. 2009. Evolutionarily conserved herpesviral protein interaction networks. PLoS Pathog. doi:10.1371/journal.ppat.1000570.

Gagneux S. 2012. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc B Biol Sci. doi:10.1098/rstb.2011.0316.

Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, et al. 2006. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci. doi:10.1073/pnas.0511240103.

Gupta S, Maiden MCJ, Feavers IM, Nee S, May RM, Anderson RM. 1996. The maintenance of strain structure in populations of recombining infectious agents. Nat Med. doi:10.1038/nm0496-437.

Haydon DT, Bastos ADS, Awadalla P. 2004. Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments. J Gen Virol. 85:1095–1100. doi:10.1099/vir.0.19588-0.

Hjalgrim H, Friborg J, Melbye M. 2007. Chapter 53: The epidemiology of EBV and its association with malignant disease. Arvin A, Campadelli-Fiume G, Mocarski E, Moore PS, Roizman B, Whitley R, Yamanishi K, editors. Cambridge University Press.

Hui KF, Chan TF, Yang W, Shen JJ, Lam KP, Kwok H, Sham PC, Tsao SW, Kwong DL, Lung ML, et al. 2018. High risk Epstein-Barr virus variants characterized by distinct polymorphisms in the EBER locus are strongly associated with nasopharyngeal carcinoma. Int J Cancer. 144(12):3031–3042. doi:10.1002/ijc.32049.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254–67. doi:10.1093/molbev/msj030.

de Jesus O. 2003. Updated Epstein-Barr virus (EBV) DNA sequence and analysis of a promoter for the BART (CST, BARF0) RNAs of EBV. J Gen Virol. 84(6):1443–1450. doi:10.1099/vir.0.19054-0.

Jombart T. 2008. Adegenet: A R package for the multivariate analysis of genetic markers. Bioinformatics. 24(11):1403–1405. doi:10.1093/bioinformatics/btn129.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–80. doi:10.1093/molbev/mst010.

Kløverpris HN, Leslie A, Goulder P. 2016. Role of HLA adaptation in HIV evolution. Front Immunol. doi:10.3389/fimmu.2015.00665.

Koelle DM, Norberg P, Fitzgibbon MP, Russell RM, Greninger AL, Huang ML, Stensland L, Jing

L, Magaret AS, Diem K, et al. 2017. Worldwide circulation of HSV-2 × HSV-1 recombinant strains. Sci Rep. doi:10.1038/srep44084.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour. doi:10.1111/1755-0998.12387.

Kwok H, Tong AHY, Lin CH, Lok S, Farrell PJ, Kwong DLW, Chiang AKS. 2012. Genomic sequencing and comparative analysis of Epstein-Barr virus genome isolated from primary nasopharyngeal carcinoma biopsy. PLoS One. 7(5):e36939. doi:10.1371/journal.pone.0036939.

Kwok H, Wu CW, Palser a L, Kellam P, Sham PC, Kwong DLW, Chiang a KS. 2014. Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsies. J Virol. doi:10.1128/JVI.01665-14.

Lassalle F, Depledge DP, Reeves MB, Brown AC, Christiansen MT, Tutill HJ, Williams RJ, Einer-Jensen K, Holdstock J, Atkinson C, et al. 2016. Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus Evol. doi:10.1093/ve/vew017.

Lei H, Li T, Hung G-C, Li B, Tsai S, Lo S-C. 2013. Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology. BMC Genomics. 14:804. doi:10.1186/1471-2164-14-804.

Lin Z, Wang X, Strong MJ, Concha M, Baddoo M, Xu G, Baribault C, Fewell C, Hulme W, Hedges D, et al. 2012 Nov 14. Whole genome sequencing of the Akata and Mutu Epstein-Barr virus (EBV) strains. J Virol. doi:10.1128/JVI.02517-12.

Liu P, Fang X, Feng Z, Guo Y-M, Peng R-J, Liu T, Huang Z, Feng Y, Sun X, Xiong Z, et al. 2011. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. J Virol. 85(21):11291–9. doi:10.1128/JVI.00823-11.

Lossius A, Johansen JN, Torkildsen Ø, Vartdal F, Holmoy T. 2012. Epstein-barr virus in systemic lupus erythematosus, rheumatoid arthritis and multiple sclerosis-association and causation. Viruses. doi:10.3390/v4123701.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics.

Norberg P, Depledge DP, Kundu S, Atkinson C, Brown J, Haque T, Hussaini Y, MacMahon E, Molyneaux P, Papaevangelou V, et al. 2015. Recombination of Globally Circulating Varicella Zoster Virus. J Virol. doi:10.1128/JVI.00437-15.

Norberg P, Kasubi MJ, Haarr L, Bergstrom T, Liljeqvist J-A. 2007. Divergence and Recombination of Clinical Herpes Simplex Virus Type 2 Isolates. J Virol. doi:10.1128/JVI.01310-07.

Palser AL, Grayson NE, White RE, Corton C, Correia S, Ba abdullah MM, Watson SJ, Cotten M, Arrand JR, Murray PG, et al. 2015. Genome Diversity of Epstein-Barr Virus from Multiple Tumour Types and Normal Infection. J Virol.(March):JVI.03614-14. doi:10.1128/JVI.03614-14.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 20(2):289–290. doi:10.1093/bioinformatics/btg412.

Pender MP, Burrows SR. 2014. Epstein–Barr virus and multiple sclerosis: potential opportunities for immunotherapy. Clin Transl Immunol. doi:10.1038/cti.2014.25.

Rieux A, Balloux F. 2016. Inferences from tip-calibrated phylogenies: A review and a practical guide. Mol Ecol. doi:10.1111/mec.13586.

Rosenberg NA. 2004. DISTRUCT: A program for the graphical display of population structure. Mol Ecol Notes. doi:10.1046/j.1471-8286.2003.00566.x.

Smith LM, McWhorter AR, Shellam GR, Redwood AJ. 2013. The genome of murine cytomegalovirus is shaped by purifying selection and extensive recombination. Virology. doi:10.1016/j.virol.2012.08.041.

Thorell K, Yahara K, Berthenet E, Lawson DJ, Mikhail J, Kato I, Mendez A, Rizzato C, Bravo MM, Suzuki R, et al. 2017. Rapid evolution of distinct Helicobacter pylori subpopulations in the Americas. PLoS Genet. doi:10.1371/journal.pgen.1006546.

Tsai M-H, Raykova A, Klinke O, Bernhardt K, Gärtner K, Leung CS, Geletneky K, Sertel S, Münz C, Feederle R, et al. 2013. Spontaneous lytic replication and epitheliotropism define an Epstein-Barr virus strain found in carcinomas. Cell Rep. 5(2):458–70. doi:10.1016/j.celrep.2013.09.012.

Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. 2018. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. doi:10.1093/nar/gky1006.

Wilkinson DE, Weller SK. 2003. The Role of DNA Recombination in Herpes Simplex Virus DNA Replication. IUBMB Life. doi:10.1080/15216540310001612237.

Wilkinson DE, Weller SK. 2004. Recruitment of Cellular Recombination and Repair Proteins to

Sites of Herpes Simplex Virus Type 1 DNA Replication Is Dependent on the Composition of Viral Proteins within Prereplicative Sites and Correlates with the Induction of the DNA Damage Response. J Virol. doi:10.1128/jvi.78.9.4783-4796.2004.

Young LS, Yap LF, Murray PG. 2016. Epstein-Barr virus: More than 50 years old and still providing surprises. Nat Rev Cancer. doi:10.1038/nrc.2016.92.

Zeng M, Li D, Liu Q, Song L, Li M, Zhang R, Yu X, Wang H, Ernberg I, Zeng Y. 2005. Genomic Sequence Analysis of Epstein-Barr Virus Strain GD1 from a Nasopharyngeal Carcinoma Patient †. Society. 79(24):15323–15330. doi:10.1128/JVI.79.24.15323.

Figures

Figure 1: A) Heatmap of SNPs which are in significant LD. Darker colours indicate lower p-values, with insignificant pairs being colours white. Numbers denote the genome positions (sampled uniformly), with rows and columns that did not contain any significant pair of SNPs in LD removed. B) Association network of all sites in LD with at least one other site. Each node represents a bi-allelic site, each link between two nodes signifies they are in LD with each other. C) Frequency plot of all connected components in the associated network.

Figure 2: Population assignment for all genome sequences assuming a population number of $k = 3$ for different subsets of sites. Every bar represents a strain that has been preassigned to either "Africa", "Asia" or "Western" (comprised of American, European and Australian isolates). The colouring of the bars represents the proportion of the input sites that have been assigned to a certain population. A) all bi-allelic sites; B) all sites in LD in the largest component; C) nonsynonymous pairs of sites in LD; D) synonymous pairs of sites in LD; E) sites not in LD. B-D refer to the subset of sites in LD that are in the largest component in the association network (figure 1B).

Figure 3: A) Proportion of pairs of nonsynonymous sites in LD with each other over LD strength. B-C) Number of links between nonsynonymous sites between different categories of genes. B) All sites (Chi-square test, $p < 2.2e-16$). C) Sites with a minimal distance of 1 kb (Chi-square test, $p < 2.2e-16$).

Figure 4: Whole gene network, coloured based on Eigenvector centrality, with warm colours indicating higher and cooler colours lower scores, respectively. Square node symbols denote genes belonging to IG, circular nodes denote genes belonging to NIG.

Tables

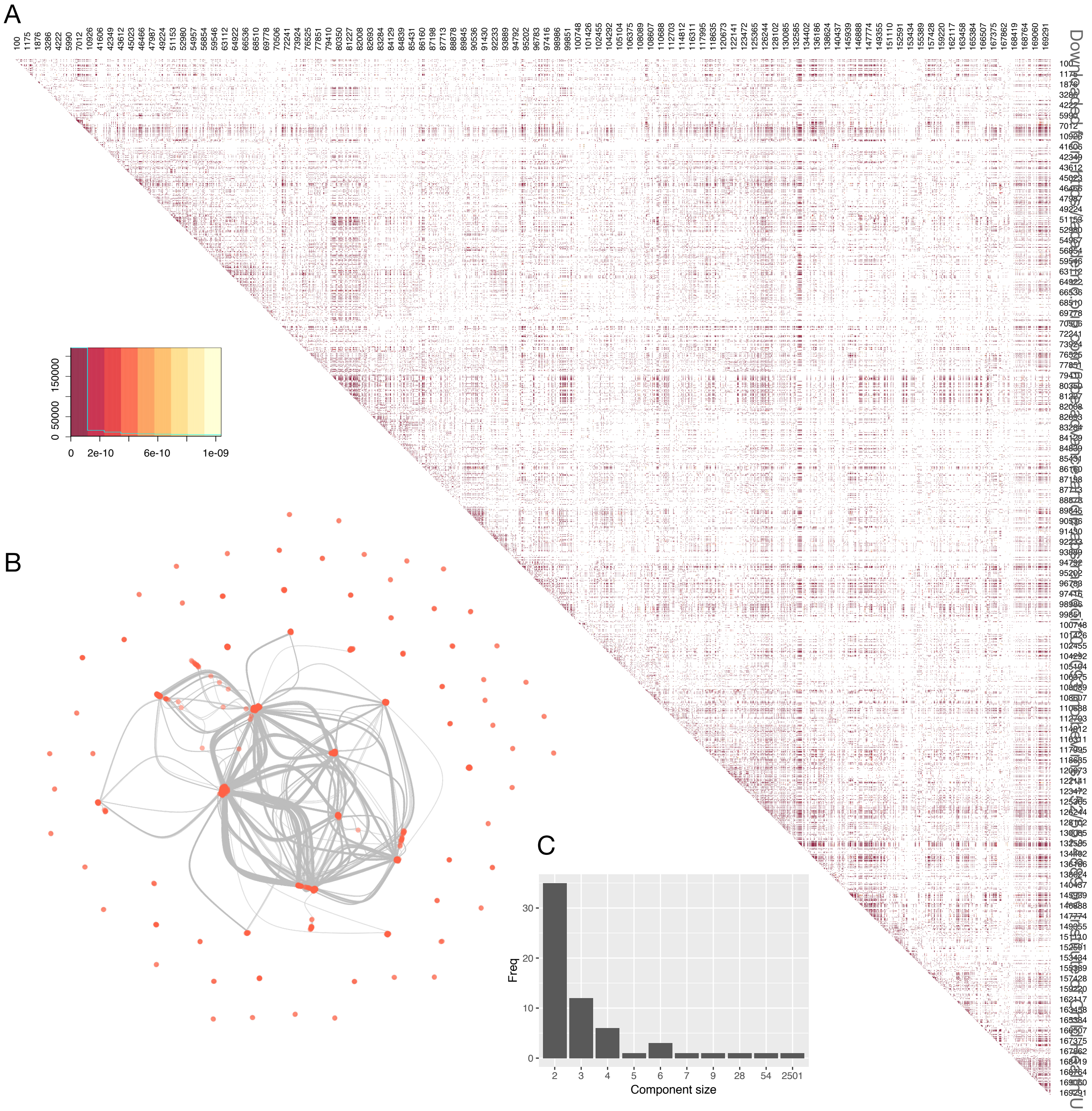
Table 1: **List of 32 genes present in the gene network considered to code for immunogenic proteins. Each epitope must have at least two references listed in IEDB.**

Protein	ORF	number of epitopes
Major DNA-binding protein	BALF2	2
Tripartite terminase subunit UL28 homolog	BALF3	1
Envelope glycoprotein B	BALF4	16
DNA polymerase catalytic subunit	BALF5	1
Ribonucleoside-diphosphate reductase small chain	BaRF1	1
Portal protein UL6 homolog	BBRF1	1
Major capsid protein	BcLF1	3
Triplex capsid protein VP23 homolog	BDLF1	1
Capsid protein VP26	BFRF3	13
Protein BGLF3	BGLF3	1
Apoptosis regulator BHRF1	BHRF1	3
Envelope glycoprotein GP350	BLLF1	2
Deoxyuridine 5'-triphosphate nucleotidohydrolase	BLLF3	3
DNA polymerase processivity factor BMRF1	BMRF1	8
Protein BMRF2	BMRF2	1
Major tegument protein	BNRF1	4
Protein BOLF1	BOLF1	2
Ribonucleoside-diphosphate reductase large subunit	BORF2	1
Replication and transcription activator	BRLF1	8
Tegument protein BRRF2	BRRF2	2
DNA primase	BSLF1	1
Envelope glycoprotein H	BXLF2	11
Trans-activator protein BZLF1	BZLF1	24
Epstein-Barr nuclear antigen 1	EBNA1	82
Epstein-Barr nuclear antigen 2	EBNA2	12
Epstein-Barr nuclear antigen 3	EBNA3A	30
Epstein-Barr nuclear antigen 4	EBNA3B	23
Epstein-Barr nuclear antigen 6	EBNA3C	33
Protein LF2	LF2	1
Uncharacterised protein LF3	LF3	1
Latent membrane protein 1	LMP1	17
Latent membrane protein 2	LMP2	32

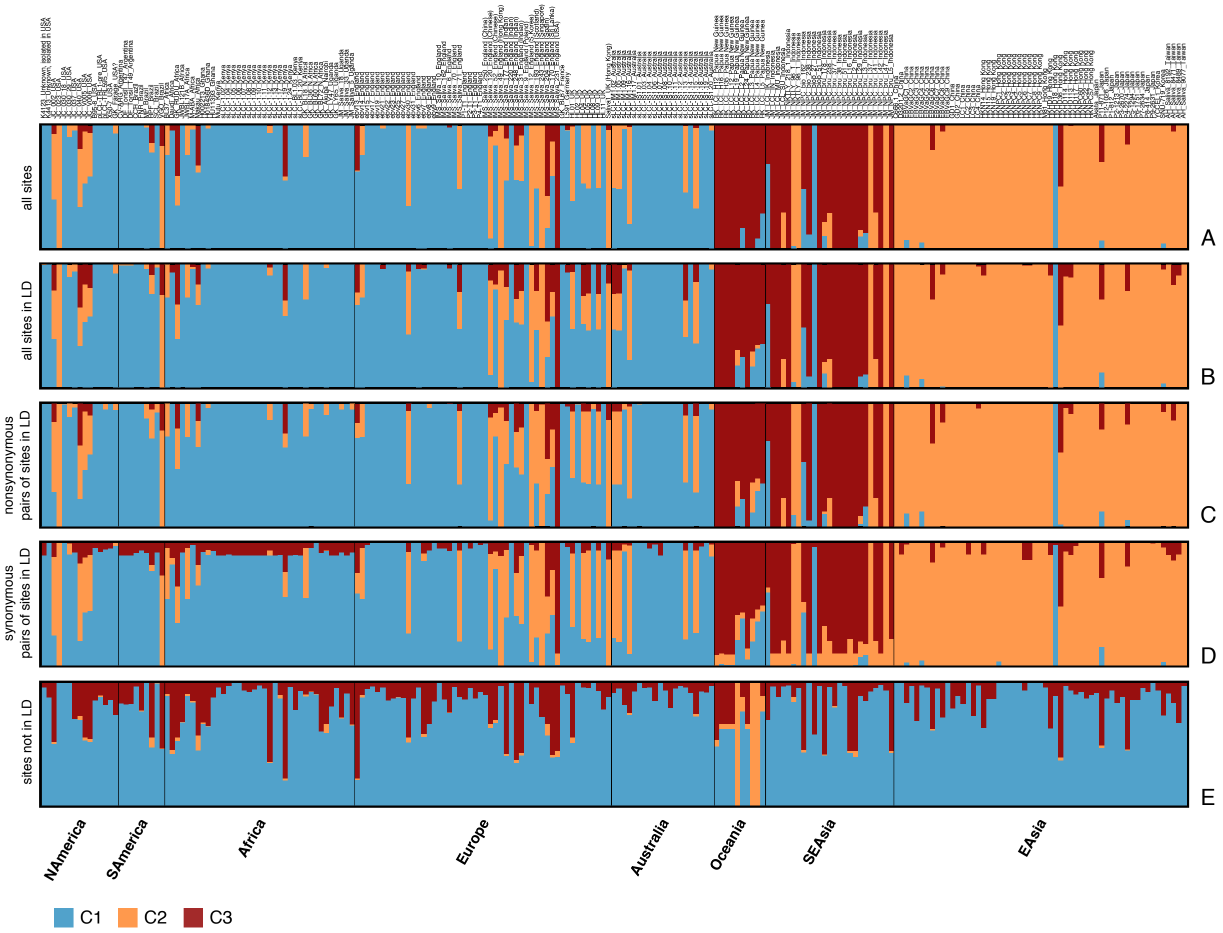
Table 2: **Most influential nodes in the network.**

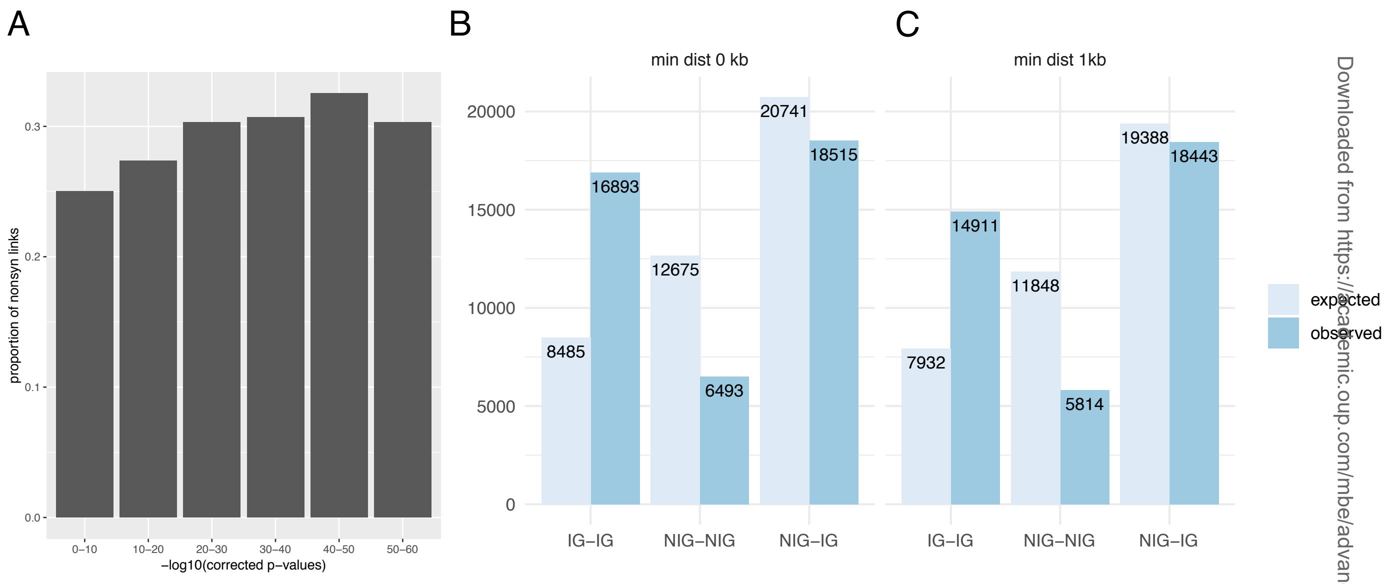
Eigenvector rank	ORF	Protein	IG
1	BPLF1	Large tegument protein deneddylase	○
2	BcRF1	TBP-like protein	
3	BBLF4	DNA replication helicase	
4	EBNA3B	Epstein-Barr nuclear antigen 4	●
5	BLLF1	Envelope glycoprotein GP350	●
6	BNRF1	Major tegument protein	●
7	EBNA1	Epstein-Barr nuclear antigen 1	●
8	BALF3	Tripartite terminase subunit 1	
9	BGLF1	Capsid vertex component 1	○
10	BKRF4	Tegument protein	
11	EBNA3A	Epstein-Barr nuclear antigen 3	●
12	LMP1	Latent membrane protein 1	●
13	EBNA3C	Epstein-Barr nuclear antigen 6	●
14	BOLF1	Protein BOLF1	●
15	BRRF2	Tegument protein	○
16	BALF2	Major DNA-binding protein	
17	BDLF3	BDLF3 (Glycoprotein)	
18	BSLF1	DNA primase	
19	BVRF1	Capsid vertex component 2	○
20	BRLF1	Replication and transcription activator	●
21	BXLF2	Envelope glycoprotein H	●
22	BDLF4	Uncharacterised protein	
23	LF1	Uncharacterised protein	
24	BBLF2-BBLF3	DNA helicase/primase complex-associated protein	○
25	BALF4	Envelope glycoprotein B	●

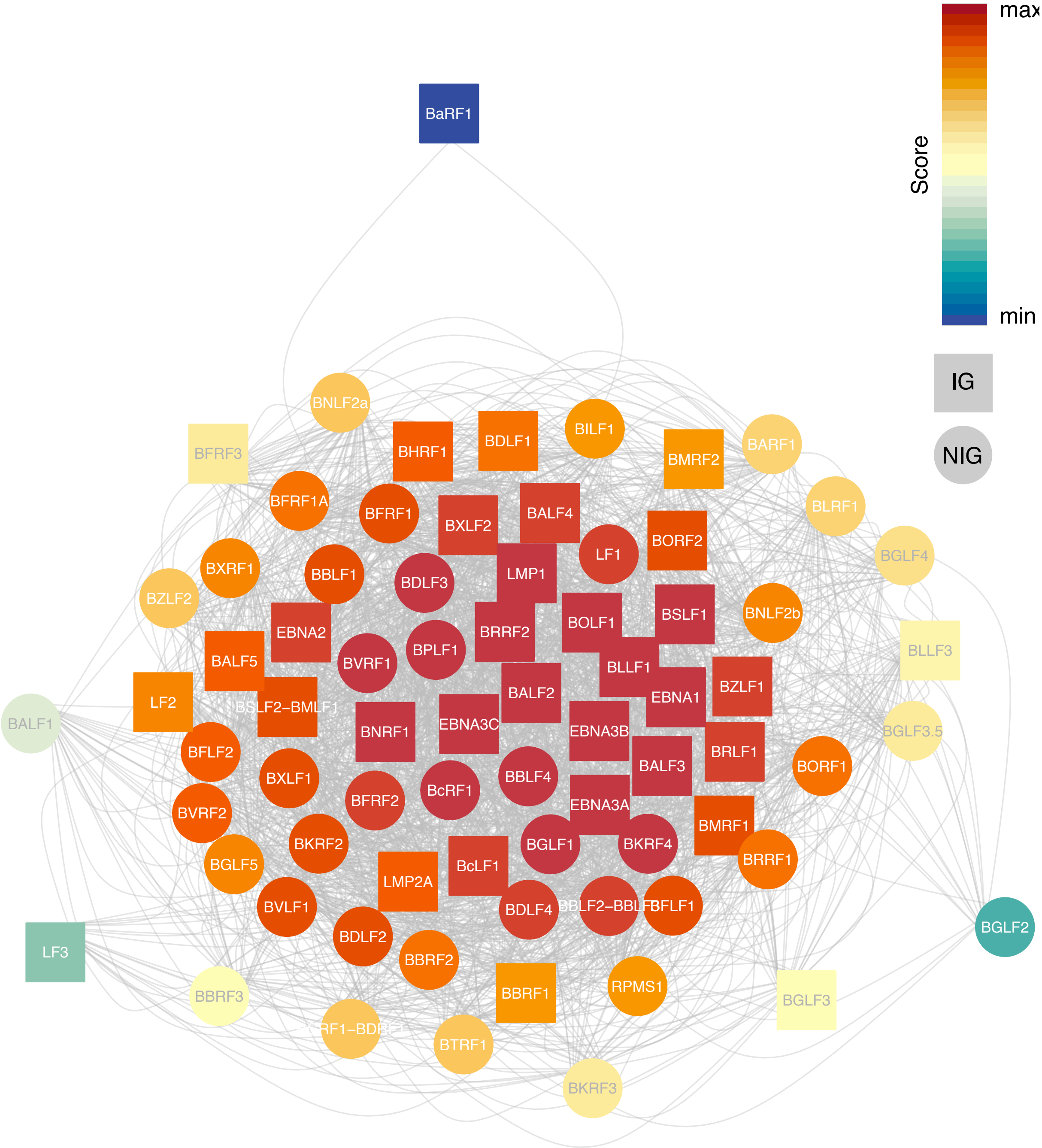
Circles in the column labelled IG mark proteins for which an immune response has been reported, with filled circles fulfilling the criterium of having at least two references and empty circles having fewer than two.



A
100
1175
1876
3286
4222
5990
7012
10926
41606
42349
43612
45023
46466
47987
49224
51153
52980
54957
56854
59546
63112
64922
66536
68510
69778
70506
72241
73924
76525
77851
79410
80350
81227
82008
82693
83284
84129
84839
85431
86160
87198
87713
88878
89845
90536
91430
92233
93889
94792
95202
96783
97416
98986
99851
100748
101426
102455
104292
105104
106375
108089
108607
110688
112703
114812
116311
117995
118635
120673
122141
123472
125365
126244
128102
130085
132585
134402
136186
138624
140437
145939
146888
147774
149355
151110
152591
153434
155389
157428
159220
162117
163458
165384
166507
167375
167862
168419
168764
169060
169291
0
50000
150000
0
2e-10
6e-10
1e-09
B
C
Freq
0
10
20
30
2
3
4
5
6
7
9
28
54
2501
Component size


all sites
all sites in LD
nonsynonymous pairs of sites in LD
synonymous pairs of sites in LD
sites not in LD
A
B
C
D
E
NAmerica
SAmerica
Africa
Europe
Australia
Oceania
SEAsia
EAsia
C1
C2
C3


A
proportion of nonsyn links
0.0
0.1
0.2
0.3
0−10
10−20
20−30
30−40
40−50
50−60
−log10(corrected p−values)
B
min dist 0 kb
20000
15000
10000
5000
0
8485
16893
12675
6493
20741
18515
IG−IG
NIG−NIG
NIG−IG
C
min dist 1kb
7932
14911
11848
5814
19388
18443
IG−IG
NIG−NIG
NIG−IG
expected
observed


BaRF1
BNLF2a
BFRF3
BHRF1
BDLF1
BILF1
BMRF2
BARF1
BFRF1A
BFRF1
BXLF2
BALF4
BORF2
BLRF1
LF1
BGLF4
BXRF1
BBLF1
BDLF3
LMP1
BZLF2
BOLF1
BSLF1
BNLF2b
EBNA2
BRRF2
BLLF3
BPLF1
BALF5
BVRF1
BLLF1
LF2
BZLF1
BALF1
BSLF2–BMLF1
BALF2
EBNA1
BNRF1
EBNA3C
EBNA3B
BGLF3.5
BFLF2
BRLF1
BORF1
BXLF1
BALF3
BBLF4
BcRF1
BFRF2
EBNA3A
BVRF2
BMRF1
BKRF2
BGLF1
BKRF4
BRRF1
BGLF5
BcLF1
LMP2A
BVLF1
BDLF4
BBLF2–BBLF3
BFLF1
BGLF2
LF3
BDLF2
BBRF2
RPMS1
BBRF1
BGLF3
BBRF3
RF1–BDRF1
BTRF1
BKRF3
Score
max
min
IG
NIG