Geographic Population Structure in Epstein-Barr Virus Revealed by Comparative Genomics

Matteo Chiara[1,†], Caterina Manzari[2,†], Claudia Lionetti[2,†], Rosella Mechelli[3], Eleni Anastasiadou[4], Maria Chiara Buscarinu[3], Giovanni Ristori[3], Marco Salvetti[3], Ernesto Picardi[2,5], Anna Maria D'Erchia[2,5], Graziano Pesole[2,5], and David S. Horner[1,2,*]

[1]Department of Biosciences, University of Milan, Milan, Italy

[2]Institute of Biomembranes and Bioenergetics, Consiglio Nazionale delle Ricerche, Bari, Italy

[3]Centre for Experimental Neurological Therapies, S. Andrea Hospital-site, Department of Neuroscience, Mental Health and Sensory Organs (NESMOS), Faculty of Medicine and Psychology, Sapienza University, Rome, Italy

[4]Department of Pathology, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, Massachusetts

[5]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, "A. Moro", Bari, Italy

*Corresponding author: E-mail: david.horner@unimi.it.

†These authors contributed equally to this work.

Accepted: September 9, 2016

Abstract

Epstein-Barr virus (EBV) latently infects the majority of the human population and is implicated as a causal or contributory factor in numerous diseases. We sequenced 27 complete EBV genomes from a cohort of Multiple Sclerosis (MS) patients and healthy controls from Italy, although no variants showed a statistically significant association with MS. Taking advantage of the availability of ~130 EBV genomes with known geographical origins, we reveal a striking geographic distribution of EBV sub-populations with distinct allele frequency distributions. We discuss mechanisms that potentially explain these observations, and their implications for understanding the association of EBV with human disease.

Key words: comparative genomics, genome sequence, population structure, Epstein-Barr virus.

Introduction

Epstein-Barr virus (EBV), a member of the gamma subdivision of the herpesviruses, is capable of both lytic and latent infection of human B-cells (and occasionally epithelial cells) and infects the majority of the global human population. Lytic infection can cause infectious mononucleosis, and while latent infection is essentially asymptomatic for the majority of carriers, EBV shows a causal association with a variety of human malignancies including Burkitt lymphoma (BL), nasopharangeal carcinoma (NPC), and Hodgkin lymphoma, and has been tentatively associated with other clinical conditions. The double-strand DNA genome encodes over 100 proteins as well as non-coding transcripts. The expression of various combinations of latent genes (EBNA1, EBNA2, EBNA3a,3b, and 3c, LP and LMP1 and 2 as well as several ncRNAs) are associated with different forms of viral latency. The remaining genes are expressed in a coordinated program during the lytic cycle.

The complete sequence of the c. 172 Kb circular genome of the prototypical EBV strain B95-8 was determined in 1984 (Baer et al. 1984). Patterns of genetic divergence of the EBNA2 and EBNA3a,b,c alleles allow the division of EBV isolates into types 1 and 2 that have differing geographical distributions, with the latter more widespread in Africa. Single locus genotyping studies have also tentatively associated specific variants with particular clinical manifestations and/or geographic regions (reviewed in Tzellos and Farrell 2012). However, EBV is subject to recombination (Walling and Raab-Traub 1994) which complicates such assignments, and comparisons of complete genome sequences suggest that recombination is widespread (McGeoch and Gatherer 2007; Palser et al. 2015).

The advent of high-throughput sequencing and DNA capture technologies has revolutionized EBV genomics. Palser et al. (2015) sequenced 71 viral genomes, confirming frequent

recombination, and demonstrating a strong geographical clustering of similar strains. Other studies (Kwok et al. 2014; Lei et al. 2015) have increased the number of distinct EBV genome sequences available to around 100.

Multiple lines of evidence suggest an association between EBV and multiple sclerosis (MS) (Salvetti et al. 2009). For example, we previously used direct amplification and sequencing from blood samples to demonstrate a statistically significant association between the presence of the EBNA2 1.2 variant and MS in a cohort of 53 MS patients and 38 healthy donors from central and southern Italy (Mechelli et al. 2015).

To explore the diversity of EBV genomes from MS patients and controls, to increase our understanding of the relationship between virus genetic diversity and host geography, and to validate a novel EBV capture array, we sequenced the complete genomes of viruses isolated from spontaneous lymphoblastoid cell lines (LCLs) obtained from 18 MS patients and 9 healthy donors (of known EBNA2 genotype) drawn from individuals participating in our previous study (Mechelli et al. 2015). Although, no significant association of viral variants with MS was detected, population structure analyses suggest that EBV allele pools and associated allele frequencies vary between geographic locations, with partially overlapping geographic distributions for some sub-populations. We discuss the implications of these findings for our understanding of EBV phylogeography and for studies directed towards the identification of EBV variants associated with human disease. We suggest that our findings provide a new paradigm for understanding EBV diversity.

Methods

Sampling and Isolation of Genomic DNA

Spontaneously outgrowing lymphoblastoid cell lines (spLCL)—transformed by EBV carried by the donor, were generated from 18 subjects with relapsing remitting MS (Polman et al. 2011) and 9 healthy donors matched for age, sex and geographic origin. These individuals represent a subset of a cohort previously subjected to partial sequencing of EBNA2. All subjects with MS were sampled during the stable phase of the disease and were free of disease modifying therapies. The local institutional review board approved the study and all participating subjects gave written informed consent.

Peripheral blood mononuclear cells (PBMCs) were obtained by density centrifugation over Ficoll–Hypaque according to standard procedures. PBMCs (5×10^4/well) were seeded in 96-flat well plates and cultured in 200 μl/well RPMI1640 medium, supplemented with 20% FBS (HyClone), 1% L-Glutamine, 100 IU penicillin, and 100 μg/ml streptomycin. Cyclosporin A (1 μg/ml, Calbiochem) was added to the medium to inhibit T-cell activation and cultures were fed twice a week. Genomic DNA was extracted from c. five million cells using the QIAamp DNA mini kit (Qiagen, Venlo, the Netherlands).

EBV Episomal DNA Enrichment and Deep-Sequencing

In total, 3,642 overlapping 120-mer RNA baits spanning the length of the EBV reference genome (NCBI Reference Sequence: NC_007605.1) were designed using the Agilent SureDesign software. The specificity of all baits was verified by BLASTn searches against the Human Genomic + Transcript database. Libraries were prepared according to the SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library protocol (Version B.1, December 2014) (Agilent Technologies, Santa Clara, CA). A B95-8 cell line was used as a control for the capture system. Around 6M pairs of reads (2×121 bp) were generated for each library on the Illumina MiSeq platform.

Bioinformatics Analyses

Raw sequence reads were trimmed using Trimmomatic (Bolger et al. 2014), and *de novo* assembly was performed using SPAdes (Bankevich et al. 2012). Reads were mapped to the hybrid reference genome of B95-8 strain with the known deletion complemented with the appropriate region of the Rajii genome (NC_007605.1) using Bowtie2 (Langmead and Salzberg 2012), SNP and indel calling were performed using VarScan (Koboldt et al. 2009) with default parameters (mean QS > 15, min coverage = 8, $P < 0.01$, min variant freq 0.2), considering only pairs of reads with a unique best mapping solution on the reference genome. *De novo* assemblies were aligned to the reference genome using Nucmer (Delcher et al. 2002) and a custom Perl script was employed to compare variant calls from *de novo* and reference mapping approaches.

For estimation of BamW repeat copy numbers, sequencing reads from the current work (or downloaded from the ENA archive) were aligned to the reference EBV B95-8 genome complemented with the Rajii deletion. A smoothing curve was calculated for every sample, by fitting a robust linear model (Wang et al. 2009), contrasting local coverage with local GC composition and average insert size, on unique (present in one copy) genomic windows of 400 bp overlapping by 200 bp. Genome wide coverage profiles were normalized according to the respective smoothing curves. BamW copy numbers were estimated as the ratio between the (normalized) read density calculated on a single copy of the repeat and the corresponding value for unique genomic regions.

Population structure analyses were performed using the Structure program (Hubisz et al. 2009), considering 1–20 clusters. Because character-based methods such as maximum-likelihood do not account for extensive recombination, we employed a simpler phenetic clustering procedure using the Bio-NJ algorithm as implemented in the software SeaView (Gouy et al. 2010) with uncorrected distances to reveal overall

genome similarity. To infer the extent and pattern of recombination, NeighborNets were produced using SplitsTree (Huson and Bryant 2006) and population splits were estimated using Treemix (Pickrell and Protchard 2012), both utilizing default parameters. Principal component analysis (PCA) analyses were performed using the prcomp function in the R stats package. Dunn indexes were calculated using the clv R package (Nieweglowski 2013) on sliding windows of 200 parsimony informative SNVs, overlapped by 100.

Results

Viral DNA was enriched from 27 spontaneous LCLs obtained from 18 MS patients and 9 healthy donors of known EBNA2 genotype, from a cohort considered in a previous study (Mechelii et al. 2015), and sequenced using the Illumina MiSeq platform. In all cases, *de novo* genome assembly yielded three to four 4 fragments which were alignable in a collinear manner with the reference genome assembly (NC_007605.1), gaps being associated with annotated sequence repeats. In addition, all reads were mapped to the reference genome using Bowtie2 (Langmead and Salzberg 2012) and variants called using Varscan (Koboldt et al. 2009).

Genome coverage, ENA accession numbers, number and type of variants detected with respect to the reference genome (conclusions were identical between genome alignment and read mapping approaches, and the B95-8 control showed no variation with respect to the reference sequence) are reported in table 1. A total of 3,405 SNPs, and 413 short INDELS were detected, of which 2,184 and 278, respectively, are shared by at least 2 genomes and 1,314 variants are unique to genomes sequenced here.

The EBNA2 1.2 allele (Tierney et al. 2006) is carried by 8 of the 27 genomes sequenced here and is otherwise observed only in two other African genomes. The LMP1 "med" haplotypes (Edwards et al. 1999) are present in 13/27 novel Italian genomes. In the subset of genomes sequenced here, no variants showed significant association with MS according to Fisher's exact test. (EBNA2 and LMP1 types of all genomes analyzed are shown in supplementary table S3, Supplementary Material online, for details).

BamW Repeats in EBV Genomes

The Internal Repeat 1 (IR1) region of the EBV genome typically contains, between 5 and 8 (Allan and Rowe 1989) tandem repeats of a c. 3 kb sequence (broadly corresponding to the BamW fragment of the viral genome). Each iteration of the repeat contains a copy of the major early latent "W" promoter and two short coding exons of the latent Leader Protein (LP). It is not possible to unambiguously map the majority of read pairs derived from IR1 owing to the length of the BamW repeat unit and, accordingly, the BamW repeats were excluded from all subsequent analyses. However, the number of repeats present in each isolate was inferred (supplementary table S1, Supplementary Material online, for details) using a method to correct for library-specific and composition-derived coverage biases and evidence recovered for concerted evolution (Liao 1999) of the BamW repeat unit (supplementary fig. 1*A* and *B*, Supplementary Material online, for details). All variants detected in the BamW repeat are described in supplementary table S2, Supplementary Material online, for details.

Population Genetics of EBV

Structure (Hubisz et al. 2009) uses multilocus genotype data to infer population stratification, optimizing the number of sub-populations under a model whereby each sub-population is characterized by a set of allele frequencies at each locus. Individuals are assigned to populations, or jointly to two or more populations where admixing is inferred.

The variation within the 127—27 new and 100 other complete—available EBV genomes is best explained by partitioning the isolates among 10 sub-populations (fig. 1*A*, probabilities assigned by Structure are shown in supplementary table S3, Supplementary Material online). Although numerous individual genomes are inferred to derive from admixing between distinct sub-populations, we can identify "pure" individuals—that is, with low (<10%) probability of admixing—in eight populations. Considering only these "non-admixed" individuals, three populations—G1, G2, and G4 (G4 is associated with the presence of type 2 EBNA2/3 genes) are only found in Kenya. Two additional populations (G5 and G6) were both predominantly associated with northern Europe, the United Kingdom, and Australia. The B95-8 genome was assigned to a population containing North and South American isolates as well as additional African genomes (G10). All Asian genomes derived from a single population (G7), while two genomes from this study as well as an African BL isolate were inferred to represent non-admixed exemplars of a provisionally labeled "Mediterranean" population (G9). Patterns can also be inferred among genomes that are inferred to be mosaic, with some admixes [e.g., between the Mediterranean (G9) and a northern European (G5) population, the Asian (G7) and G3 populations as well as between the American (G10) and G8 groups] being predominant in distinct areas (fig. 1*A*). The observed degree of geographic clustering for all genomes was significantly greater than expected by chance ($P \leq 1 \times 10^{-5}$) using the test proposed by Chen and Holmes (2009). Consideration of suboptimal solutions identified by Structure (supplementary fig. S2 and table S4, Supplementary Material online, for details) reveals that when the number of sub-populations is constrained, overall patterns remain consistent, with changes principally seen in the partitioning of the American population (G10), the occasional merging of a European (G5) and African (G2) population, and the merging of the Mediterranean (G9) and other European groups (G5, G6) when low numbers of sub-populations are used.

Table 1

Accession numbers of the novel EBV genomes assembled in this study and basic statistics concerning the number of variants detected

Genome	Acc no	Mean coverage (X)	Common variants	Private variants
CAR	ERS1100719	2,048.89	852	59
PP	ERS1100731	2,227.2	741	126
MV	ERS1100733	1,947.52	856	67
BL	ERS1100735	2,116.65	904	2
VL	ERS1100730	1,815.83	911	9
NM	ERS1100715	1,975.69	900	0
MC	ERS1100718	1,680.85	883	37
MFA	ERS1100723	1,968	1,031	27
GV	ERS1100726	2,059.6	949	97
GIOVS	ERS1100717	2,159.79	1,033	29
CS	ERS1100724	2,167.5	903	0
GR	ERS1100714	2,001.78	971	110
PT	ERS1100716	2,107.06	997	78
BA	ERS1100728	1,735.74	1,134	7
MM	ERS1100734	2,331.46	1,056	69
LOL	ERS1100725	1,956.2	899	0
IM	ERS1100727	2,078.37	874	108
GF	ERS1100729	1,974.4	909	28
BR	ERS1100721	1,952.53	1,130	2
CM	ERS1100732	1,742.88	1,043	8
LUL	ERS1100722	1,883.03	894	0
TM	ERS1100713	2,018.29	891	7
SC	ERS1100710	1,882.56	1,131	194
SA	ERS1100711	2,027.31	803	0
RT	ERS1100712	2,199.21	1,013	61
MST	ERS1100720	2,031.73	897	43
CAS	ERS1100709	1,849.45	960	70

PCA of the non-admixed genomes (excluding type II isolates whose high divergence tends to obscure resolution) separates the populations inferred by Structure; while the first principal component represents the division of European and Asian groups, the second component separates the European and Asian from the American genomes (fig. 1*B*). Importantly, phenetic clustering of non-mosaic genomes on the basis of genetic distances (fig. 1*C*) recovers identical groups with the exception of the placement of the sequence "NM", which is likely to represent an admixed genome which was not detected by Structure.

Split Decomposition (SplitsTree; Huson and Bryant 2006) analysis of the non-admixed genomes (fig. 2*A*) suggests equivalent clusters with reticulate relationships between sub-populations, consistent with past admixing or phylogenetic relationships between sub-populations. As with phenetic clustering, the sequence "NM" fails to cluster within its assigned population. Equivalent results were obtained using TreeMix (Pickrell and Pritchard 2012), which reconstructs possible patterns of population splits from allele frequency data, also suggesting a possible affinity between Asian (G7), Mediterranean (G9), and one of the United Kingdom/Australian populations (G8) dependent on the position of the root.

To explore possible partitioning of variation between sub-populations along the viral genome, we employed the Dunn index (Dunn 1973) to evaluate the coherence of signal in sliding windows (defined by 200 parsimony informative sites, shifting by 100 informative sites) with the sub-populations defined by Structure. The results (supplementary fig. S3*A*, Supplementary Material online, for details) indicate that the genomic region around EBNA-1 best discriminates between sub-populations, although this trait does not seem to correlate with latent cycle genes in general. Interestingly, the Dunn index shows a significant negative correlation with the proportion of sites that are polymorphic in more than one sub-population. Consistent with this, phenetic trees for individual latent genes (supplementary fig. S3*B–H*, supplementary Material online, for details) suggest that the least intra-population allelic diversity is observed in EBNA1. We note that where more than one allele type is present within single populations, each of the subgroups is typically monophyletic, consistent with long-term presence of distinct allele groups in various sub-populations.

Discussion

We have expanded, by over 25%, the number of complete EBV genome sequences available and leveraged geographic

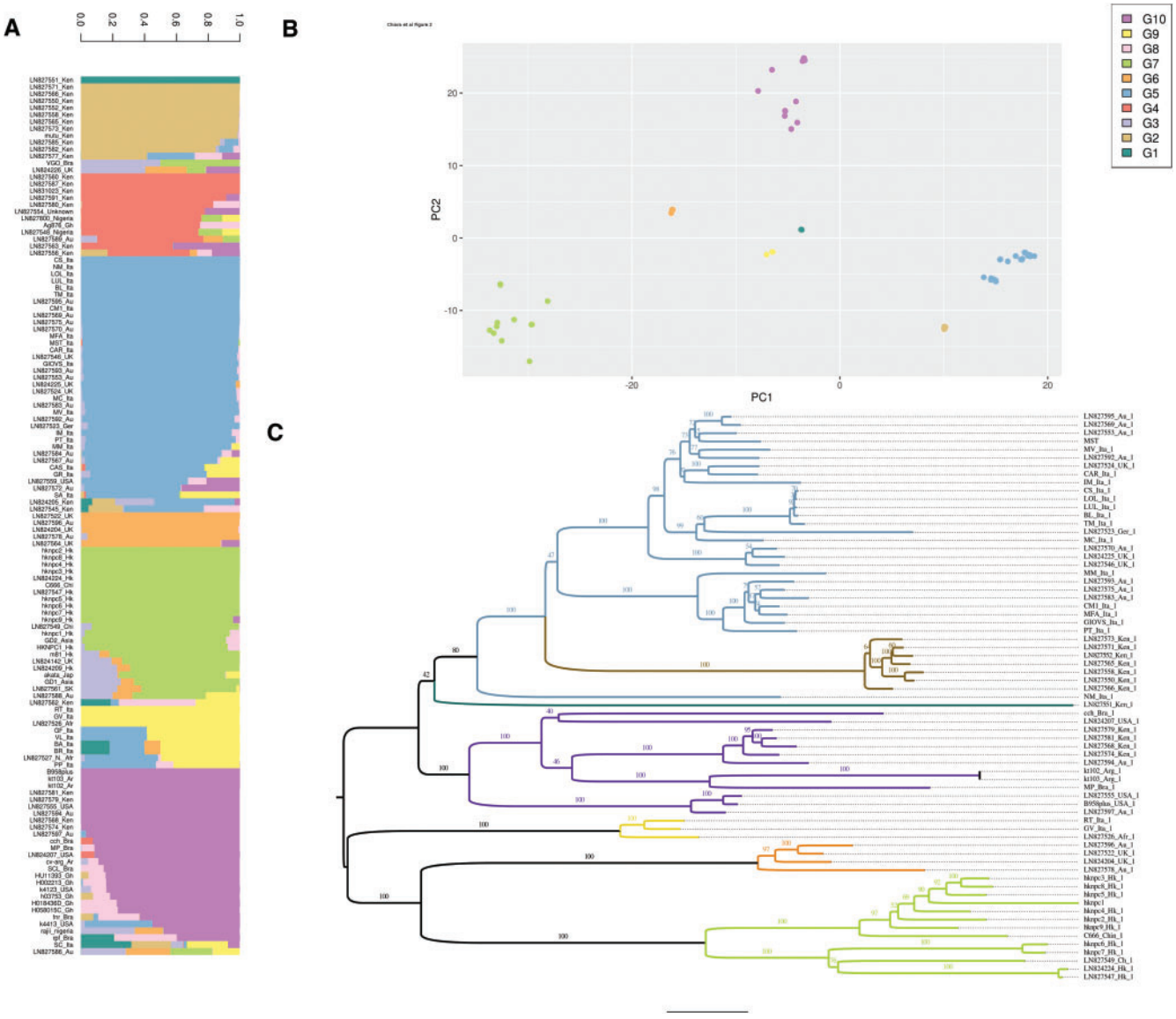

Fig. 1.—Population structure, PCA and phenetic clustering of EBV genome sequences. (*A*) Barplot displaying the probability of provenance as inferred by Structure for all the 127 EBV genomes considered in this study, geographic origins are shown for each isolate. *(B)* Scatterplot of the PCA of 75 "pure" genomes. (*C*) Phenetic tree of the "pure" genomes. Different groups are indicated by colors and the root position is arbitrary. "Pure" genomes are defined as those where Structure assigned a $\geq$ 90% probability of provenance from a single population. Colors are consistent between panels A, B and C.

origin information to infer patterns of population stratification. Both the geographical partitioning of "pure" members of sub-populations, and the observation of frequent admixing between geographically (or culturally) overlapping sub-populations are consistent with geographically isolated viral sub-populations whose structure has been partially disrupted by recent migration and admixing. Although biased sampling (e.g., most Asian samples derive from NPC isolates, whereas South American samples are restricted to EBV positive BL biopsies) complicates the interpretation of our results, it is interesting to consider factors that could account for the presence of distinct and apparently relatively isolated historical viral gene pools in various localities.

Under one scenario, local gene pools result from regular historical hybridization with "immigrant" genotypes, leading, under the influence of genetic drift and possibly episodic selection/virus host co-evolution, to present day allele frequency distributions. Alternatively, gene pools of geographically distinct viral sub-populations might reflect viral pools carried by ancestral human colonizers, potentially as far back as early migrations of modern humans out of Africa. Although selection for particular viral genotypes in conjunction with HLA or

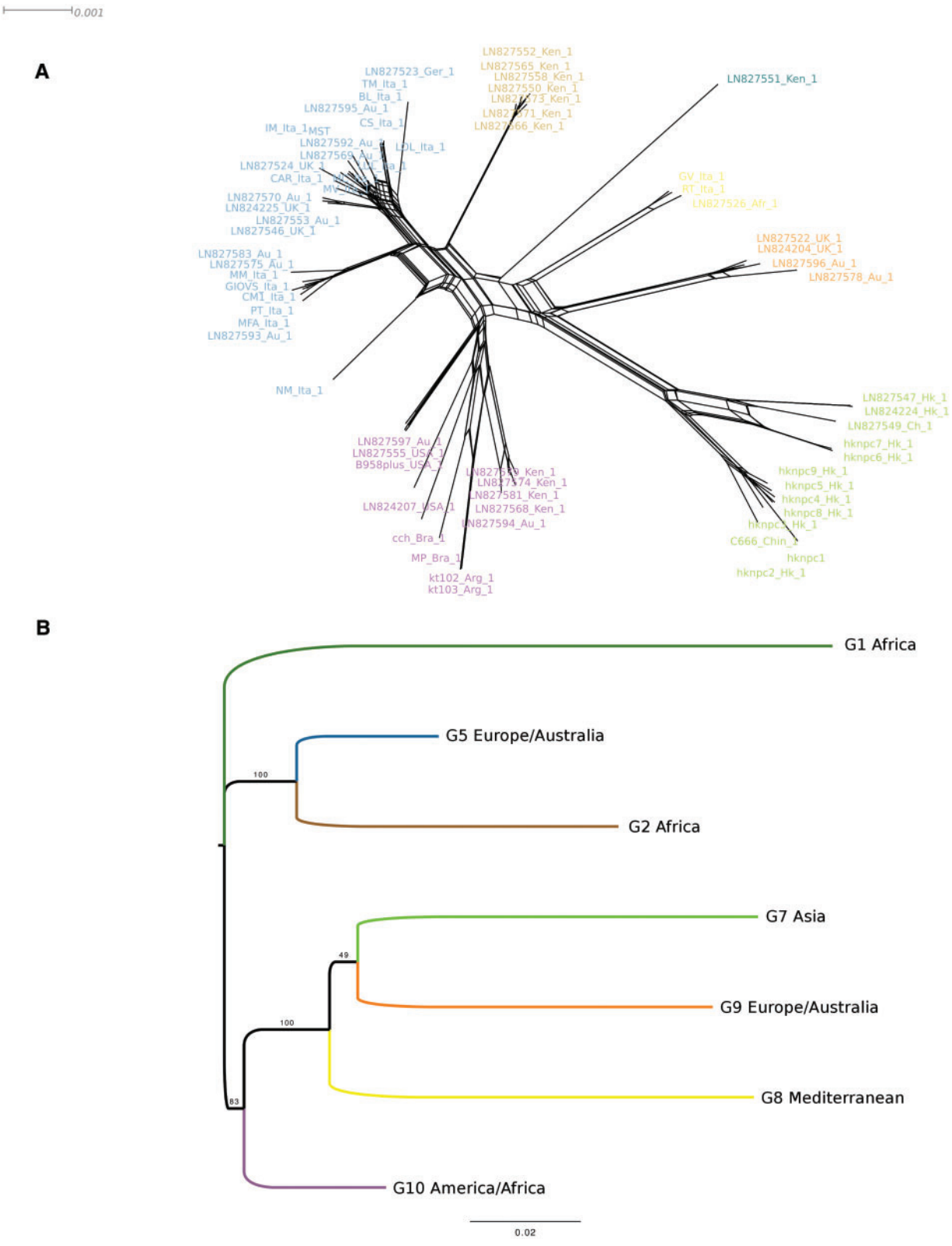

Fig. 2.—NeighborNet and Population allele frequency relationships. *(A)* NeighborNet analysis of 69 non-admixed representatives of inferred EBV subpopulations. NeighborNet resolves the same clusters of non-admixed genomes as Structure and phenetic clustering, highlighting conflicts that correspond to allele types shared between sub-populations. *(B)* Allele frequency bootstrap tree of possible relationships between inferred EBV subpopulations. The tree was estimated using Treemix (Pickrell and Pritchard 2012) using default parameters (no migration, no linkage disequilibrium).

other host haplotypes is known to occur (de Campos-Lima et al. 1993), these interactions may derive from ancient allele frequency distributions among humans. The presence of various, frequently admixed, apparently partially spatially super-imposed and yet still recognizable EBV sub-populations in Europe, as well as the observation that Africa appears to carry the greatest number of sub-populations, may be consistent with a capacity for long-term maintenance of geographic population identity. This is potentially consistent with the fact that the EBV is typically acquired in early childhood, presumably from local individuals and the observed topologies of EBNA gene phylogenetic trees.

Reconciliation of the inferred population relationships with current hypothesis of ancestral human migrations is complicated by several factors. First, it is possible that current sampling does not include representatives of the earliest human colonization of Australia. A similar situation is plausible for the Americas, where viral gene pools might be expected to show characteristics of various European colonizing populations combined with African populations that arrived through slavery. Additional considerations include potential "back to Africa" migrations and the likelihood that archaic humans, with who admixing is known to have occurred, carried EBV (Abi-Rached et al. 2011) and may have contributed alleles to contemporary viral populations. Strategic sampling and sequencing of viral genomes from carefully chosen locations and ethnic groups will be required to assist in resolving the identities and histories of EBV sub-populations worldwide more conclusively.

Population stratification can have a non-trivial impact on Genome Wide Association Studies and contemporary approaches in humans routinely employ statistical methods that attempt to account for the effects of population structure (Price et al. 2010). Indeed, an apparent association an LMP1 variant carrying a 30 bp deletion with Asian NPC was later shown to likely be a consequence of geographic biases in allele frequencies (Zhang et al. 2002). One of the most compelling associations of an EBV genotype with a particular human disease is the finding that particular variants in the non-coding EBER RNAs are consistently present in NPC isolates from both high and intermediate incidence areas in China, but at lower proportions in the background population in intermediate incidence areas (Shen et al. 2015). However, the lack of complete genome data for these isolates does not exclude the potential presence of additional tightly linked variants that may show higher association with the disease. We note that the EBER haplotype associated with Chinese NPC is not present in a single, admixed North African genome from an NPC biopsy (LN827527). However, 16/16 NPC derived viral genomes (including LN827527), all three of the non-mosaic Mediterranean genomes and two other Italian isolates (but only seven other isolates, four of which are Asian), share a series of substitutions in non-coding regions immediately upstream of the Cp latency promoter and between the DS and FR elements of the latent replication origin OriP (supplementary table S5, Supplementary Material online, for details). This observation might be of interest as, in the Mediterranean basin, NPC shows an incidence intermediate between northern Europe (<1 per 100,000 per year) and endemic areas of China (>20 per 100,000 per year) (Abdel-Hamid et al. 1992). These observations, and analogous patterns among other EBV related pathologies, underline both the importance of complete genome sequencing and the need incorporation of population level allele frequency data in association studies.

Although capture sequencing of EBV genomes from spontaneous LCLs has some limitations, particularly insensitivity to multiple infections, which may be of particular relevance in the case of MS (Santón et al. 2011), it provides an unprecedented opportunity to study the genetics and epidemiology of EBV, to address disease associations, and to reveal long-term virus-host evolutionary interactions.

Supplementary Material

Supplementary figures S1-S3 and tables S1-S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Eddie Holmes for critical reading and constructive comments on the article. This study was supported by grants from Fondazione Italiana Sclerosi Multipla (2011/R/31 to M.S. and G.P.), from Ministero della Salute (Progetto Strategico 2013 to M.S.) and by the Italian Ministero dell'Istruzione, Università e Ricerca (MIUR): PRIN 2009, 2010, and 2012, and the Consiglio Nazionale delle Ricerche: Medicina Personalizzata and Aging Program 2012–2014.

Literature Cited

Abdel-Hamid M, Chen JJ, Constantine N, Massoud M, Raab-Traub NJ. 1992. EBV strain variation: geographical distribution and relation to disease state. Virology 190(1):168–175.

Abi-Rached L, et al. 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334(6052):89–94.

Allan GJ, Rowe DT. 1989. Size and stability of the Epstein-Barr virus major internal repeat (IR-1) in Burkitt's lymphoma and lymphoblastoid cell lines. Virology 173(2):489–498.

Baer R, et al. 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310(5974):207–211.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.
Dunn JC. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybernet. 3 (3):32–57.
Chen R, Holmes EC. 2009. Frequent inter-species transmission and geographic subdivision in avian influenza viruses from wild birds. Virology 383(1):156–161.
de Campos-Lima PO, et al. 1993. HLA-A11 epitope loss isolates of Epstein-Barr virus from a highly A11+ population. Science 260(5104):98–100.
Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30(11):2478–2483.
Edwards RH, Seillier-Moiseiwitsch F, Raab-Traub N. 1999. Signature amino acid changes in latent membrane protein 1 distinguish Epstein-Barr virus strains. Virology 261(1):79–95.
Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27(2):221–224.
Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. Mol Ecol Res. 9(5):1322–1332.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254–267.
Koboldt DC, et al. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25(17):2283–2285.
Kwok H, et al. 2014. Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. J Virol 88(18):10662–10672.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359.
Lei H, et al. 2015. Epstein-Barr virus from Burkitt Lymphoma biopsies from Africa and South America share novel LMP-1 promoter and gene variations. Sci Rep. 23(5):16706.
Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet. 64(1):24–30.
McGeoch DJ, Gatherer D. 2007. Lineage structures in the genome sequences of three Epstein-Barr virus strains. Virology 359(1):1–5.
Mechelli R, et al. 2015. Epstein-Barr virus genetic variants are associated with multiple sclerosis. Neurology 84:1362–1368.
Nieweglowski L. 2013. Clv: Cluster Validation Techniques. CRAN
Palser AL, et al. 2015. Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. J Virol. 89(10):5222–5237.Pickrell J, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8(11):e1002967.
Polman CH, et al. 2011. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol. 69(2):292–302.
Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 11(7):459–463.
Salvetti M, Giovannoni G, Aloisi F. 2009. Epstein-Barr virus and multiple sclerosis. Curr Opin Neurol. 22(3):201–206.
Santón A, et al. 2011. High frequency of co-infection by Epstein-Barr virus types 1 and 2 in patients with multiple sclerosis. Mult Scler. 17(11):1295–1300.
Shen ZC, et al. 2015. High prevalence of the EBER variant EB-8m in endemic nasopharyngeal carcinomas. Plos One 25(10):e0121420.
Tierney RJ, et al. 2006. Multiple Epstein-Barr virus strains in patients with infectious mononucleosis: comparison of ex vivo samples with in vitro isolates by use of heteroduplex tracking assays. J Infect Dis. 193(2):287–297.
Tzellos S, Farrell PJ. 2012. Epstein-Barr virus sequence variation—biology and disease. Pathogens 1:156–175.
Walling DM, Raab-Traub N. 1994. Epstein-Barr virus intrastrain recombination in oral hairy leukoplakia. J Virol. 68(12):7909–7917.
Wang J, et al. 2009. Package "robust". CRAN
Zhang XS, et al. 2002. The 30-bp deletion variant: a polymorphism of latent membrane protein 1 prevalent in endemic and non-endemic areas of nasopharyngeal carcinomas in China. Cancer Lett. 176(1):65–73.

Associate Editor: Patricia Wittkopp