GENOMIC STRUCTURE OF EPSTEIN-BARR VIRUS AS IDENTIFIED BY HIDDEN MARKOV MODEL CLUSTERING

Anastasia Lankina[1,2], Cristina Venturini[1], Oscar Charles[1], Richard Goldstein[3], Judith Breuer[1,4]

[1] Infection, Immunity and Inflammation Research Department, University College London, Institute for Child Health, London, United Kingdom; [2] University College London, Institute for Immunity and Transplantation, London, United Kingdom; [3] Division of Infection and Immunity, University College London, Cruciform Building, London, United Kingdom; [4] UCL Great Ormond Street Institute for Child Health, Infection, Immunity and Inflammation Research Department, London, United Kingdom

anastasia.lankina.20@ucl.ac.uk

Epstein-Barr virus (EBV) infects about 90% of the world's adult population, but the incidence of EBV-associated diseases varies greatly worldwide. Aside from human and environmental factors, viral variation is likely to play a role in contributing to disease. EBV has a highly conserved genome and so far variation has been mostly found between EBV types 1 and 2, as well as some geographic variation within type 1 sequences.

Here we explore the EBV genomic variation using hidden Markov model clustering (HMM) as a gene-independent approach to identify regions of population structure encoding multiple alleles. Across 241 publicly available globally representative EBV genomes [1], we identified 33 multi-allelic regions in both type 1 and type 2, 17 of which were also present within type-1-only sequences. In addition to known variable regions (such as those in latency-associated EBNA genes), we observed the separation of EBV type 1 and type 2 in alleles overlapping the lytic transcription activator gene BZLF-1 (F_{st}=0.510, $p<0.001$) and envelope glycoprotein gene BLLF-1 (F_{st}=0.759, $p<0.001$). Within type 1, multi-allelic regions overlapped with the latent origin of replication OriP (F_{st}=0.299, $p=0.008$) and late gene transcription activator BcRF-1 (F_{st}=0.301, $p<0.001$). Although genomic structure can explain some of the geographic segregation (especially Asian versus non-Asian clustering), much of the geographic variation remains present within the relatively conserved portions of the genome. Notably, at least two identified allelic regions, present in both analyses, do not correspond with any known open reading frame or regulatory transcript, raising the question of the evolutionary pressure behind their multi-allelicity.

In summary, we confirmed that type 1 and type 2 separation is the biggest contributing factor to EBV genome variation. We identified novel variable regions in the EBV genome that represent type 1 and type 2 population structure differences, and found geographically informative regions which may be associated with the different incidence of EBV-related conditions across the globe. This might provide an insight into the association between EBV variation and disease, and may provide a useful model for EBV genomic variability in future disease association studies.

1. Correia S, Bridges R, Wegner F, Venturini C, Palser A, Middeldorp JM, et al; 2018 Nov 15; *Sequence Variation of Epstein-Barr Virus: Viral Types, Geography, Codon Usage, and Diseases.* Journal of Virology