Automated, phylogeny-based genotype delimitation of the Hepatitis Viruses HBV and HCV

Dora Serdari[1], Evangelia-Georgia Kostaki[2], Dimitrios Paraskevis[2], Alexandros Stamatakis[1,3] and Paschalia Kapli[4]

[1] The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[2] Department of Hygiene Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece
[3] Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[4] Centre for Life's Origins and Evolution, Department of Genetics Evolution and Environment, University College London, University of London, London, United Kingdom

ABSTRACT

Background. The classification of hepatitis viruses still predominantly relies on ad hoc criteria, i.e., phenotypic traits and arbitrary genetic distance thresholds. Given the subjectivity of such practices coupled with the constant sequencing of samples and discovery of new strains, this manual approach to virus classification becomes cumbersome and impossible to generalize.

Methods. Using two well-studied hepatitis virus datasets, HBV and HCV, we assess if computational methods for molecular species delimitation that are typically applied to barcoding biodiversity studies can also be successfully deployed for hepatitis virus classification. For comparison, we also used ABGD, a tool that in contrast to other distance methods attempts to automatically identify the barcoding gap using pairwise genetic distances for a set of aligned input sequences.

Results—Discussion. We found that the mPTP species delimitation tool identified even without adapting its default parameters taxonomic clusters that either correspond to the currently acknowledged genotypes or to known subdivision of genotypes (subtypes or subgenotypes). In the cases where the delimited cluster corresponded to subtype or subgenotype, there were previous concerns that their status may be underestimated. The clusters obtained from the ABGD analysis differed depending on the parameters used. However, under certain values the results were very similar to the taxonomy and mPTP which indicates the usefulness of distance based methods in virus taxonomy under appropriate parameter settings. The overlap of predicted clusters with taxonomically acknowledged genotypes implies that virus classification can be successfully automated.

INTRODUCTION

The continuous advances in next generation sequencing technologies lead to an increasingly easier and inexpensive production of genome and metabarcoding data. The wealth of

available data has triggered the development of new models of molecular evolution, algorithms, and software, that aim to improve molecular sequence analyses in terms of biological realism, computational efficiency, or a trade-off between the two. In response to such technological and technical advancements, several fields of biology have undergone a substantial transformation. Sequence-based species delimitation and identification, in the framework of DNA-(meta)barcoding constitutes a representative example that revived taxonomy and systematics (*Tautz et al., 2003*; *Moritz & Cicero, 2004*; *Savolainen et al., 2005*; *Waugh, 2007*; *Bucklin et al., 2010*; *Valentini, Pompanon & Taberlet, 2009*; *Li et al., 2015*), while it also provided a new means of analysis in several fields (*Galimberti et al., 2013*; *Mishra et al., 2016*; *Leray & Knowlton, 2015*; *Bell et al., 2016*; *Batovska et al., 2018*). Among others, the development of novel species delimitation tools has substantially advanced the study of biodiversity of microorganism that are often hard to isolate and study (*Taberlet et al., 2012*; *Gibson et al., 2014*; *Thomsen & Willerslev, 2015*). The sequencing of environmental samples in conjunction with algorithms for genetic clustering has led to the identification of a plethora of previously unknown organisms and a re-assessment of the microbial biodiversity in several settings.

In a similar context, genetic information has been a rich source of information for viral species. Several studies show how phylogenetic information can be deployed for identifying the spatial and temporal origin of a virus, potential factors that trigger its dispersal, and other key epidemiological parameters (*Stadler et al., 2011*; *Stadler et al., 2014*; *Gire et al., 2014*). In an era of high human mobility, such methods are important, as the increase of emerging and re-emerging epidemics is even more prominent than in the past (*Balcan et al., 2009*; *Meloni et al., 2011*; *Pybus, Tatem & Lemey, 2015*). Nevertheless, phylogenetic information is still not used in the context of virus species classification or identification. As we have witnessed for other microorganisms, using or adapting already available methods for fast and automated delimitation or identification of virus species can greatly contribute to better understand their evolution.

To date, the official taxonomy of viruses (ICTV, i.e., International Committee on Taxonomy of Viruses) has mainly been based on established biological classification criteria as used for other life forms, such as plants or animals. An analogous hierarchical classification system containing orders, families, subfamilies, genera, and species is being applied (*Simmonds, 2015*). The ICTV is typically based on phenotypic criteria, such as morphology, nucleic acid type (i.e., DNA or RNA), hosts, symptoms, mode of replication, geographical data, or presence of antigenic epitopes, to name a few. Generally, such criteria, despite being informative, can be subjective, require highly specialized knowledge, and are time consuming to apply. In contrast, sequence evolution takes into account the evolutionary history of life forms and, thus, may offer a more objective source of information for taxonomic classification. An important difference in viruses compared to other organisms is that they lack a common set of universal genes such as the 18S rRNA in eukaryotes or the 16S rRNA in prokaryotes. Therefore, we cannot infer a comprehensive virus tree of life (*Simmonds et al., 2017*), and, more importantly for species delimitation, we cannot rely upon barcoding markers that are universally suitable for all viruses. We can nonetheless gain valuable insights for their systematics by

utilizing phylogenetic information at lower taxonomic ranks (e.g., families, genera, species), using appropriate genes for each dataset. In this context, methods using genetic-distance thresholds (*Bao, Chetvernin & Tatusova, 2014*; *Lauber & Gorbalenya, 2012*; *Yu et al., 2013*) have been suggested as a complementary method to the traditional virus classification for accelerating new species identification.

In this study, we explore whether a recently developed algorithm for molecular species delimitation on barcoding or marker gene phylogenies can be deployed for ICTV. In contrast to genetic distance-based methods the multi-rate Poisson Tree Processes (mPTP, *Kapli et al., 2017*) infers the number of genetic clusters given a phylogenetic input tree. Such trees can easily be inferred using both, Maximum Likelihood (*Stamatakis, 2014*), or Bayesian approaches (*Ronquist et al., 2012*) on single-gene or multi-gene multiple sequence alignments. The fundamental assumption of the model is that variance in the data, as represented by the phylogeny, is greater among species than within a species (*Zhang et al., 2013*). The additional assumption of mPTP, that the genetic variation may differ substantially among species allows to accurately delimit species in large (meta-) barcoding datasets comprising multiple species of diverse life histories (*Kapli et al., 2017*). Experiments using empirical data for several animal phyla (*Kapli et al., 2017*) and recently also viruses (*Thézé et al., 2018*; *Modha et al., 2018*) show that the method consistently provides extremely fast and sensible species estimates on 'classic' phylogenetic marker and barcoding genes.

To assess whether mPTP can be deployed as a quantitative ICTV method we analyze two medically important viruses, Hepatitis B (HBV) and Hepatitis C (HCV), that are leading global causes of human mortality (*Stanaway et al., 2016*). Both viruses cause liver inflammation, but are substantially different from each other. HBV has a partially double-stranded circular DNA genome with a length of about 3.2 kb while HCV is a single-stranded, positive-sense RNA virus, with a genome length of approximately 10 kb (*Radziwill, Tucker & Schaller, 1990*; *Tang & McLachlan, 2001*; *Martell et al., 1992*). Both virus types comprise at least two taxonomic levels (HBV: genotypes, subgenotypes; HCV: genotypes, subtypes). Besides the significance of the two viruses for human health, we selected them as test cases since due to the substantial amount of taxonomic research that has been conducted and that we can hence use to assess the efficiency of genetic clustering (e.g., *Simmonds et al., 2005*; *Schaefer, 2007*; *Smith et al., 2014*; *Messina et al., 2015*).

MATERIAL AND METHODS

Datasets
We generated multiple sequence alignments (MSAs) corresponding to two virus types: HBV and HCV from two sets of full-length genomic sequences downloaded from publicly available databases (NCBI: http://www.ncbi.nlm.nih.gov/, accession numbers provided in Appendix S1).

The HBV dataset comprises 110 sequences corresponding to eight genotypes (i.e., A-H) and 31 subgenotypes. The genotypes (A through D, F, and H) have been further divided into subgenotypes indexed by numbers for the corresponding genotype (e.g., A_1, A_2, B_1, B_2,

B_3, etc.; *Kramvis, 2014*). The inter-genotypic and inter-subgenotypic divergence exceeds 8% and 4%–8%, respectively across the genome. No subgenotypes have been reported for genotypes E, G and H which shows that they are of lower levels of genetic divergence than the rest. The distribution of HBV genotypes differs greatly with respect to the geographical origin. Moreover, they differ in their natural history, response to treatment and disease progression (*Huang et al., 2013*; *Biswas et al., 2013*; *Moura et al., 2013*; *Shi et al., 2013*). For our study we included the sequences of the eight genotypes (A–H) that form part of the oldest identified HBV groups.

The HCV dataset (I) comprises 213 sequences corresponding to seven major taxonomic units named after genotypes (1, 2, 3, 4, 5, 6, and 7) and numerous subtypes (*Smith et al., 2014*). The HCV classification into genotypes and subtypes was based on genetic-distance thresholds that were verified by the fact that they formed monophyletic clades in an inferred phylogeny (*Smith et al., 2014*). Therefore, the HCV classification serves as an appropriate test case for assessing whether a similar clustering can be identified with a more objective and automated method, such as mPTP, that does not require any user input apart from a phylogeny.

Genetic cluster delimitation

To delimit the putative species, additionally to mPTP, we used the distance-based "*Automatic Barcode Gap Discovery*" tool (ABGD, *Puillandre et al., 2012*). ABGD is a popular distance-based barcoding method that, compared to other distance-based methods attempts to automatically identify the threshold value for the transition from intra-specific variation to inter-specific divergence (*Puillandre et al., 2012*).

For the mPTP delimitation, a fully binary (bifurcating) rooted phylogeny is required. Therefore, using the aligned sequences we inferred the phylogenetic relationships under the GTR+Γ model of nucleotide substitution using RAxML-NG (*Kozlov et al., 2018*). We rooted the phylogenetic trees according to the originally published phylogenies (i.e., using the branch leading to genotypes F/H for HBV and genotype 7 for HCV). Using heuristic search algorithms for finding the 'best' delimitation given the rooted phylogeny and without any further prior assumptions. We performed the mPTP delimitation under Maximum Likelihood (ML) and calculated the support of the delimited clusters using Markov-chain Monte Carlo (MCMC) sampling (*Kapli et al., 2017*). We conducted the MCMC sampling twice for 10^6 generations, to identify potential lack of convergence with a sampling frequency of 0.1.

For ABGD, the user has to define two important parameters, (i) the prior maximum divergence of intraspecific diversity (P), which implies that the barcode gap is expected to exceed this value and should not be confused with the genetic thresholds assumed to define the inter-specific relationships, (ii) a proxy for the minimum gap width (X), which indicates that the barcoding gap is expected to be X times larger than any intraspecific gap (*Puillandre et al., 2012*). For both, HBV, and HCV, we used 10 prior maximum thresholds in the range of $p = 0.001$ and $P = 0.05$. The proxy for the minimum gap width (X) was set to the default value ($X = 1.5$) for HCV, while for HBV the default value did not yield any delimitation and we therefore set it to a lower value ($X = 0.5$).

RESULTS & DISCUSSION

The biodiversity of viruses is tremendous and it is broadly accepted that our understanding of their ecology and evolution is constrained to a small fraction of species (*Paez-Espino et al., 2016*). In just a kilo of marine sediment there can be a million of different viral genotypes (*Breitbart & Rohwer, 2005*), while on a global scale the number of viruses is 10 million-fold higher than the number of stars in the universe (*Suttle, 2013*). The classification of such a diverse set of organisms constitutes a challenging task and is impossible to accomplish within reasonable time using phenotypic characters. Quantitative computational methods could provide a viable alternative, particularly for large scale clustering and fast identification of viral strains (*Simmonds et al., 2017*; *Modha et al., 2018*). Using empirical data of the HBV and HCV viruses we show that by applying phylogeny-aware and distance-based tools to classify the strains of the two virus types, the corresponding genetic clustering closely recovers their currently accepted taxonomy.

HCV Clustering

The current taxonomy of HCV comprises seven genotypes, while mPTP yielded 16 genetic clusters (Fig. 1, Figs. S1 and S2 and Appendix S1). From the 16 clusters, five were congruent with the current taxonomy, i.e, genotypes 1, 2, 4, 5 and 7. On the contrary, genotype 3 and genotype 6 were further split into three and eight sub-clusters correspondingly (Fig. 1), which corroborates former views that divergent variants of these genotypes may qualify as separate major genotypes (*Simmonds et al., 2005*; *Smith et al., 2014*). In particular, the additional clusters identified by mPTP correspond to previously identified groups of subtypes (Fig. S1). For genotype 6, these clusters consisted of the following subtype groups: 6a; 6b and 6xd; 6c, 6d, 6e, 6f, 6g, 6o, 6p, 6q, 6r, 6s, 6t, 6u, 6w, 6xc and 6xf; 6 h, 6i, 6j, 6k, 6l, 6m, 6n, 6xb, 6xe; 6xa; 6v (Fig. S1 and Appendix S1). Similarly, for genotype 3, the delimited clusters were (i) 3g, 3b, 3i, 3a, 3e, 3d, (ii) 3k, and (iii) 3 h and 3. All clusters were substantially supported by the MCMC sampling, except the split of 3k subtype from its sister group (Fig. 1), which may be due to the limited amount of corresponding sequences.

The number of clusters inferred with ABGD ranged from 1 to 208 depending on the value of the maximum intraspecific divergence threshold (Fig. 2). The most reasonable result (i.e., the one closest to the current standard taxonomy) comprised 19 clusters and was obtained for a minimum of intraspecific genetic diversity of 5.99% (i.e., $p = 0.0599$). Under this threshold, the delimitation is largely identical to the delimitation obtained with mPTP (Fig. 1), with three differences: (i) that genotype 3 was split into four clusters, instead of three, (ii) genotype six was divided into nine clusters instead of eight, and, (iii) genotype 7 is divided into two clusters. When the prior intraspecific divergence was increased to a higher minimum of 10%, all sequences were grouped in a single cluster. When the threshold was set to a lower value (3.6%) the number of clusters increased to 135 (Fig. 2). Nevertheless, the delimitation with the 5.99% threshold is largely congruent to current taxonomy and the clusters obtained with mPTP, thus indicating the usefulness of distance-based methods in virus taxonomy under well informed parameters.

So far, the classification of HCV into genotypes and subtypes has been defined mostly by visual identification of clades in phylogenetic inference of HCV sequences (*Simmonds*
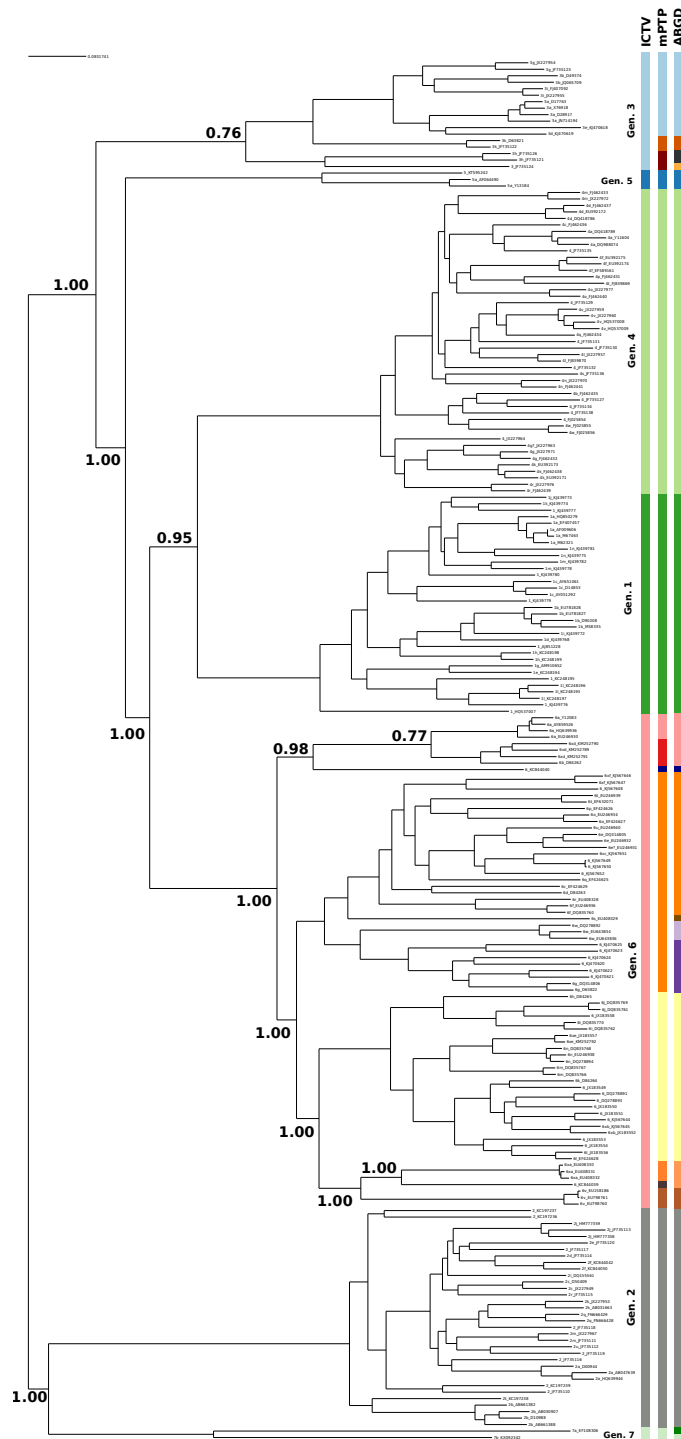
Figure 1 Clustering of the HCV samples into genotypes; the first bar of colors corresponds to the genotypes currently acknowledged by ICTV, the second to the mPTP ML clustering and the third to the ABGD clustering ($p = 0.0599$, $X = 1.5$). The numbers indicate the support for a particular node being a speciation node obtained by the MCMC sampling under the mPTP model (support <0.5 not shown, but see Fig. S2). The phylogenetic relationships were inferred using RAxML under the GTR+Γ model.

Full-size ☑ DOI: 10.7717/peerj.7754/fig-1
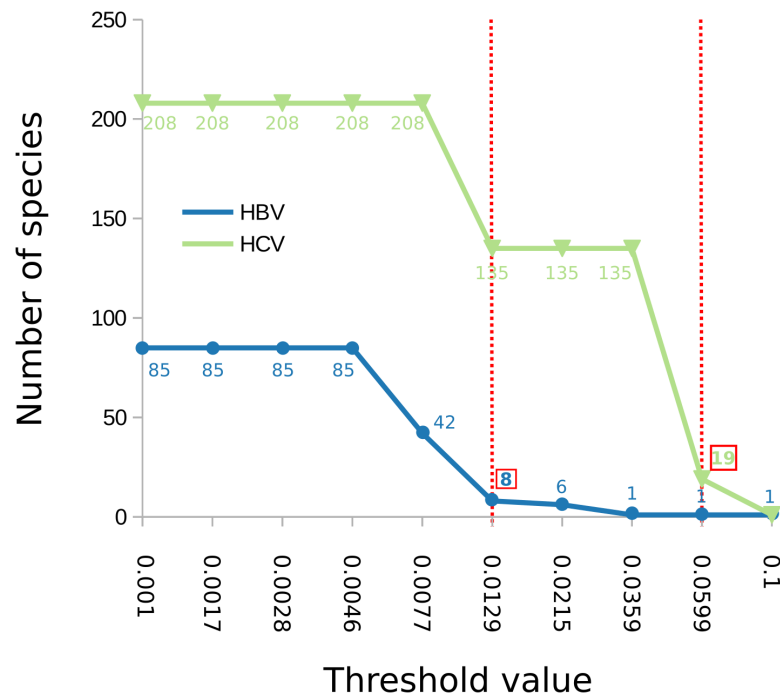
Figure 2 **Number of delimited clusters for ABGD with respect to input parameters.** The graph shows the change of the number of delimited clusters (y axis) with respect to the minimum intraspecific threshold ("p") assumed by ABGD (x axis). The threshold that yielded the most sensible clustering for HBV was $p = 0.0129$ while for HCV was $p = 0.0599$, both are shown with a dotted red line in the figure; the corresponding number of clusters is indicated in a red box.

Full-size 🖼 DOI: 10.7717/peerj.7754/fig-2

et al., 2005; *Smith et al., 2014*). Specifically, the genotypes correspond to the seven major highly-supported phylogenetic HCV clusters while subtypes were defined as the secondary hierarchical clusters found within each genotype (*Smith et al., 2014*). This classification scheme has been widely adopted (*Combet et al., 2007*; *Yusim et al., 2016*) and has been shown to be robust (in terms of stability of the HCV phylogeny) and relevant for clinical practice, since response rates to immunomodulatory treatment for the chronic hepatitis C differs across genotypes. Nevertheless, new, unassigned lineages are often discovered from understudied areas (*Sulbaran et al., 2010*; *Nakano et al., 2012*; *Lu et al., 2013*; *Tong et al., 2015*) and it is challenging to assign them a taxonomic status, given that the genetic distance cut-off among intra and inter-specific relationships is arbitrary and variable for different parts of the HCV phylogeny (*Simmonds et al., 2005*). The greatly overlapping mPTP and ABGD clusters with the HCV genotypes shows that the classification, and, consequently, the identification, of the genotypes can be easily automated utilizing objective, transparent, and unifying approaches. Embracing such alternatives can be crucial for viruses like HCV, taking into account that the correct identification of the HCV genotypes could be of clinical importance in providing the appropriate medical treatment (*Strader et al., 2004*; *Ge et al., 2009*).

HBV clustering

In the case of HBV, the mPTP clustering is almost identical to the current classification (*Norder et al., 2004*; *Kramvis & Kew, 2007*) of the virus that comprises eight genotypes, except for subgenotype C4 which formed a new cluster (Fig. 3, Figs. S3 and S4 and Appendix S1). This is in line with the greater genetic divergence of C4 compared to the other subgenotypes due to its ancient origin in native populations in Oceania (*Paraskevis et al., 2013*). However, the split of C4 from its sister cluster (genotype C) is not supported by the MCMC sampling, potentially reflecting the lack of adequate sampling. On the other hand, the number of clusters identified by ABGD varied from 1 to 85 under different thresholds of minimum intraspecific divergence, while the delimitation for a threshold of 1.29% exactly matched the eight genotypes of the HBV classification (Figs. 2 and 3). Both ABGD and mPTP identified seven of the genotypes (A–F) as distinct genetic clusters. The only difference was that mPTP split genotype C into two distinct clusters (Fig. 3), i.e., subgenotype C4 was recovered as a distinct cluster from the remaining seven subgenotype.

CONCLUSIONS

The application of mPTP to the HCV and HBV data sets shows that automated viral species delimitation using phylogeny-aware methods yields clusters that largely agree with the current standard taxonomy. The additional clusters identified for HCV by mPTP is not surprising as they have been previously considered divergent sub-clusters within the genotypes 3 and 6. Analogously, for HBV, mPTP yielded almost identical results to the current nomenclature system with the exception of a single sub-genotype, C4, that was previously mentioned to be more genetically divergent within genotype C (*Paraskevis et al., 2013*). In both cases, these new clusters indicate the potential need for taxonomic revision. However, given the wide use of the current nomenclature in the medical field, and the lack of other sources of information such as recombination, particularly for HBV, and, response to treatment, we wouldn't suggest taxonomic changes at present. Regarding distance methods, the example of HCV and HBV, shows that meaningful parameter values for distance-based methods may differ substantially among datasets, and, therefore, establishing global thresholds is impossible. On the contrary, mPTP can be seamlessly applied to taxa of substantially different life histories (e.g., variable population sizes, evolution rates), as it does not require any input parameters except a phylogeny. Overall, the ease-of-use of mPTP in conjunction with its computational efficiency on phylogenies with hundreds of samples render it a useful tool for viral biodiversity estimates, initial classification of understudied taxa, and accelerating the viral species identification process.
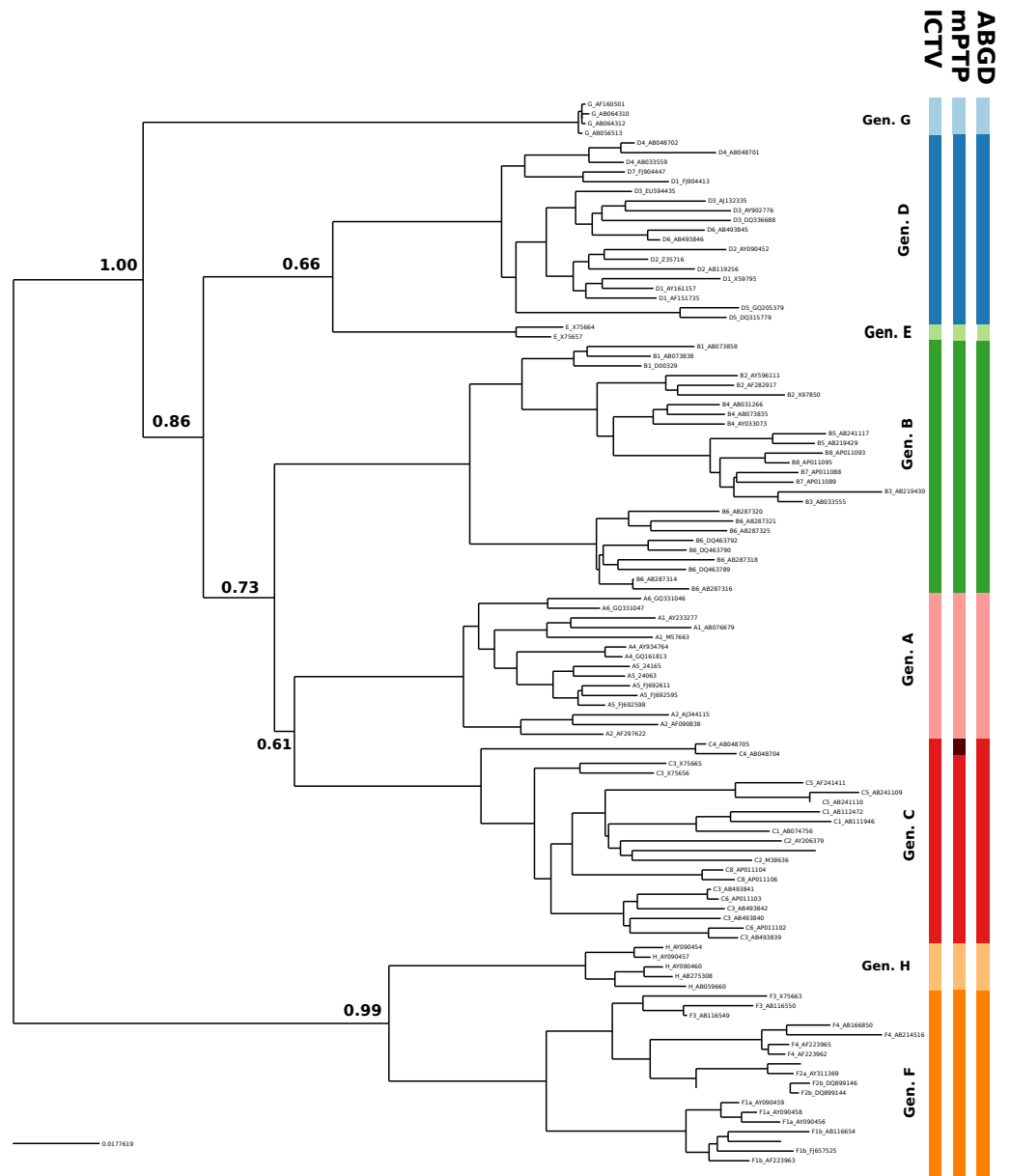
ACKNOWLEDGEMENTS

Figure 3 Clustering of the HBV samples into genotypes; the first colored bar corresponds to the genotypes currently acknowledged by ICTV, the second to the mPTP ML clustering and the third to the ABGD clustering ($p = 0.0129$, $X = 0.5$). The numbers indicate the support for a particular node being a speciation node obtained by the MCMC sampling under the mPTP model (support <0.5 not shown, but see Fig. S4). The phylogenetic relationships were inferred using RAxML under the GTR+Γ model.

Full-size 🖼 DOI: 10.7717/peerj.7754/fig-3

Serdari et al. (2019), *PeerJ*, DOI 10.7717/peerj.7754

9/15

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Dora Serdari analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Evangelia-Georgia Kostaki analyzed the data, authored or reviewed drafts of the paper, approved the final draft.
- Dimitrios Paraskevis authored or reviewed drafts of the paper, approved the final draft.
- Alexandros Stamatakis conceived and designed the experiments, authored or reviewed drafts of the paper, approved the final draft.
- Paschalia Kapli conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The sequence data used in this study were downloaded from NCBI. The relevant Accession Numbers are available in the Supplementary Appendix. The resulting data from the phylogenetic and clustering analyses are available in the GitHub repository: https://github.com/Pas-Kapli/peerj-mptp-HBV-HCV-data.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.7754#supplemental-information.

REFERENCES

Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America* **106**(51):21484–21489 DOI 10.1073/pnas.0906910106.

Bao Y, Chetvernin V, Tatusova T. 2014. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. *Archives of Virology* **159**(12):3293–3304 DOI 10.1007/s00705-014-2197-x.

Batovska J, Lynch SE, Cogan NOI, Brown K, Darbro JM, Kho EA, Blacket MJ. 2018. Effective mosquito and arbovirus surveillance using metabarcoding. *Molecular Ecology Resources* **18(1)**:32–40 DOI 10.1111/1755-0998.12682.

Bell KL, Burgess KS, Okamoto KC, Aranda R, Brosi BJ. 2016. Review and future prospects for DNA barcoding methods in forensic palynology. *Forensic Science International: Genetics* **21**:110–116 DOI 10.1016/j.fsigen.2015.12.010.

Biswas A, Panigrahi R, Pal M, Chakraborty S, Bhattacharya P, Chakrabarti S, Chakravarty R. 2013. Shift in the hepatitis B virus genotype distribution in the last decade among the HBV carriers from eastern India: possible effects on the disease status and HBV epidemiology. *Journal of Medical Virology* **85(8)**:1340–1347 DOI 10.1002/jmv.23628.

Breitbart M, Rohwer F. 2005. Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology* **13(6)**:278–284 DOI 10.1016/j.tim.2005.04.003.

Bucklin A, Hopcroft RR, Kosobokova KN, Nigro LM, Ortman BD, Jennings RM, Sweetman CJ. 2010. DNA barcoding of Arctic Ocean holozooplankton for species identification and recognition. *Deep Sea Research Part II: Topical Studies in Oceanography* **57(1–2)**:40–48 DOI 10.1016/j.dsr2.2009.08.005.

Combet C, Garnier N, Charavay C, Grando D, Crisan D, Lopez J, Dehne-Garcia A, Geourjon C, Bettler E, Hulo C, Le Mercier P, Bartenschlager R, Diepolder H, Moradpour D, Pawlotsky J-M, Rice CM, Trépo C, Penin F, Deléage G. 2007. euHCVdb: the European hepatitis C virus database. *Nucleic Acids Research* **35**:D363–D366 DOI 10.1093/nar/gkl970.

Galimberti A, De Mattia F, Losa A, Bruni I, Federici S, Casiraghi M, Martelos S, Labra M. 2013. DNA barcoding as a new tool for food traceability. *Food Research International* **50(1)**:55–63 DOI 10.1016/j.foodres.2012.09.036.

Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, Qiu P, Bertelsen AH, Muir AJ, Sulkowski M. 2009. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461(7262)**:399–401 DOI 10.1038/nature08309.

Gibson J, Shokralla S, Porter TM, King I, Van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America* **111(22)**:8007–8012 DOI 10.1073/pnas.1406468111.

Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang P, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JLB, Chapman SB, Bochichio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC.

2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345(6202)**:1369–1372 DOI 10.1126/science.1259657.

Huang CC, Kuo TM, Yeh CT, Hu CP, Chen YL, Tsai YL, Chen ML, Chang C, Chang C. 2013. One single nucleotide difference alters the differential expression of spliced RNAs between HBV genotypes A and D. *Virus Research* **174(1–2)**:18–26 DOI 10.1016/j.virusres.2013.02.004.

Kapli P, Lutteropp S, Zhang J, Kobert K, Pavlidis P, Stamatakis A, Flouri T. 2017. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. *Bioinformatics* **33(11)**:1630–1638.

Kozlov A, Darriba D, Flouri T, Morel B, Stamatakis A. 2018. RAxML-NG: a fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* Epub ahead of print May 9 2019 DOI 10.1093/bioinformatics/btz305.

Kramvis A. 2014. Genotypes and genetic variability of hepatitis B virus. *Intervirology* **57(3–4)**:141–150 DOI 10.1159/000360947.

Kramvis A, Kew MC. 2007. Epidemiology of hepatitis B virus in Africa, its genotypes and clinical associations of genotypes. *Hepatology Research* **37**:S9–S19 DOI 10.1111/j.1872-034X.2007.00098.x.

Lauber C, Gorbalenya AE. 2012. Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences. *Viruses* **4(9)**:1425–1437 DOI 10.3390/v4091425.

Leray M, Knowlton N. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America* **112(7)**:2076–2081 DOI 10.1073/pnas.1424997112.

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. 2015. Plant DNA barcoding: from gene to genome. *Biological Reviews* **90(1)**:157–166 DOI 10.1111/brv.12104.

Lu L, Li C, Yuan J, Lu T, Okamoto H, Murphy DG. 2013. Full-length genome sequences of five hepatitis C virus isolates representing subtypes 3g, 3 h, 3i and 3k, and a unique genotype 3 variant. *The Journal of General Virology* **94(Pt 3)**:543–548 DOI 10.1099/vir.0.049668-0.

Martell M, Esteban JI, Quer J, Genesca J, Weiner A, Esteban R, Guardia J, Gomez J. 1992. Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *Journal of Virology* **66(5)**:3225–3229.

Meloni S, Perra N, Arenas A, Gómez S, Moreno Y, Vespignani A. 2011. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports* **1**:62 DOI 10.1038/srep00062.

Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, Barnes E. 2015. Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology* **61(1)**:77–87 DOI 10.1002/hep.27259.

Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, Sundaresan V. 2016. DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnology Journal* **14(1)**:8–21 DOI 10.1111/pbi.12419.

Modha S, Thanki AS, Cotmore SF, Davison AJ, Hughes J. 2018. ViCTree: an automated framework for taxonomic classification from protein sequences. *Bioinformatics* **34(13)**:2195–2200 DOI 10.1093/bioinformatics/bty099.

Moritz C, Cicero C. 2004. DNA barcoding: promise and pitfalls. *PLOS Biology* **2(10)**:e354 DOI 10.1371/journal.pbio.0020354.

Moura IF, Lopes EP, Alvarado-Mora MV, Pinho JR, Carrilho FJ. 2013. Phylogenetic analysis and subgenotypic distribution of the hepatitis B virus in Recife, Brazil. *Infection, Genetics and Evolution* **14**:195–199 DOI 10.1016/j.meegid.2012.11.022.

Nakano T, Lau GM, Lau GM, Sugiyama M, Mizokami M. 2012. An updated analysis of hepatitis C virus genotypes and subtypes based on the complete coding region. *Liver International* **32(2)**:339–345 DOI 10.1111/j.1478-3231.2011.02684.x.

Norder H, Couroucé AM, Coursaget P, Echevarria JM, Shou-Dong L, Mushahwar IK, Magnius LO. 2004. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. *Intervirology* **47(6)**:289–309 DOI 10.1159/000080872.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. *Nature* **536(7617)**:425–430 DOI 10.1038/nature19094.

Paraskevis D, Magiorkinis G, Magiorkinis E, Ho SY, Belshaw R, Allain JP, Hatzakis A. 2013. Dating the origin and dispersal of hepatitis B virus infection in humans and primates. *Hepatology* **57(3)**:908–916 DOI 10.1002/hep.26079.

Puillandre N, Lambert A, Brouillet S, Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology* **21(8)**:1864–1877 DOI 10.1111/j.1365-294X.2011.05239.x.

Pybus OG, Tatem AJ, Lemey P. 2015. Virus evolution and transmission in an ever more connected world. *Proceedings of the Royal Society B: Biological Sciences* **282(1821)**:20142878 DOI 10.1098/rspb.2014.2878.

Radziwill G, Tucker W, Schaller H. 1990. Mutational analysis of the hepatitis B virus P gene product: domain structure and RNase H activity. *Journal of Virology* **64(2)**:613–620.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck J. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61(3)**:539–542 DOI 10.1093/sysbio/sys029.

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R. 2005. Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360(1462)**:1805–1811 DOI 10.1098/rstb.2005.1730.

Schaefer S. 2007. Hepatitis B virus genotypes in Europe. *Hepatology Research* **37**:S20–S26 DOI 10.1111/j.1872-034X.2007.00099.x.

Shi W, Zhang Z, Ling C, Zheng W, Zhu C, Carr MJ, Higgins DG. 2013. Hepatitis B virus subgenotyping: history, effects of recombination, misclassifications, and corrections. *Infection, Genetics and Evolution* **16**:355–361 DOI 10.1016/j.meegid.2013.03.021.

Simmonds P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology* **96(6)**:1193–1206 DOI 10.1099/vir.0.000016.

Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davidson AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, Van der Vlugt RA, Varsani A, Zerbini FM. 2017. Consensus statement: virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* **15(3)**:161 DOI 10.1038/nrmicro.2016.177.

Simmonds P, Bukh J, Combet C, Deléage G, Enomoto N, Feinstone S, Halfon P, Inchauspe G, Kuiken C, Maertens G, Miyokami M, Murphy DG, Okamoto H, Pawlotsky JM, Penin F, Sablon E, Shin IT, Stuyver LJ, Thiel HJ, Viazov S, Weiner AJ, Widell A. 2005. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology* **42(4)**:962–973 DOI 10.1002/hep.20819.

Smith DB, Bukh J, Kuiken C, Muerhoff AS, Rice CM, Stapleton JT, Simmonds P. 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology* **59(1)**:318–327 DOI 10.1002/hep.26744.

Stadler T, Kouyos R, Von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D, Günthard HF, Drummond AJ, Bonhoeffer S. 2011. Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution* **29(1)**:347–357.

Stadler T, Kühnert D, Rasmussen DA, Du Plessis L. 2014. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLOS Currents Outbreaks* **6**: Edition 1 DOI 10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30(9)**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Stanaway JD, Flaxman AD, Naghavi M, Fitzmaurice C, Vos T, Abubakar I, Abu-Raddad L, Assadi R, Bhala N, Cowie B, Forouzanfour MH, Groeger J, Hanafiah M, Jacobsen KH, James SL, MacLachlan J, Malekyadeh R, Martin NK, Mokhad AA, Mokdad AH, Murray CJL, Plass D, Rana S, Rein DB, Richardus JH, Sanabria J, Sazyan M, Shahraz S, So S, Vlassov VV, Weiderpass E, Wiersma ST, Younis M, Yu C, El Sayed Zaki M, Cooke GS. 2016. The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. *The Lancet* **388(10049)**:1081–1088 DOI 10.1016/S0140-6736(16)30579-7.

Strader DB, Wright T, Thomas DL, Seeff LB. 2004. Diagnosis, management, and treatment of hepatitis C. *Hepatology* **39(4)**:1147–1171 DOI 10.1002/hep.20119.

Sulbaran MZ, Di Lello FA, Sulbaran Y, Cosson C, Loureiro CL, Rangel HR, Cantaloube JF, Campos RH, Moratorio G, Cristina H, Pujol FH. 2010. Genetic history of hepatitis C virus in Venezuela: high diversity and long time of evolution of HCV genotype 2. *PLOS ONE* **5(12)**:e14315 DOI 10.1371/journal.pone.0014315.

Suttle CA. 2013. Viruses: unlocking the greatest biodiversity on Earth. *Genome* **56(10)**:542–544 DOI 10.1139/gen-2013-0152.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* **21(8)**:2045–2050 DOI 10.1111/j.1365-294X.2012.05470.x.

Tang H, McLachlan A. 2001. Transcriptional regulation of hepatitis B virus by nuclear hormone receptors is a critical determinant of viral tropism. *Proceedings of the National Academy of Sciences of the United States of America* **98(4)**:1841–1846 DOI 10.1073/pnas.98.4.1841.

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler A. 2003. A plea for DNA taxonomy. *Trends in Ecology & Evolution* **18(2)**:70–74 DOI 10.1016/S0169-5347(02)00041-1.

Thézé J, Lopez-Vaamonde C, Cory J, Herniou E. 2018. Biodiversity, evolution and ecological specialization of baculoviruses: a treasure trove for future applied research. *Viruses* **10(7)**:366 DOI 10.3390/v10070366.

Thomsen PF, Willerslev E. 2015. Environmental DNA—an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation* **183**:4–18 DOI 10.1016/j.biocon.2014.11.019.

Tong YQ, Liu B, Liu H, Zheng HY, Gu J, Song EJ, Song C, Li Y. 2015. Accurate genotyping of hepatitis C virus through nucleotide sequencing and identification of new HCV subtypes in China population. *Clinical Microbiology and Infection* **21(9)**:874–e9.

Valentini A, Pompanon F, Taberlet P. 2009. DNA barcoding for ecologists. *Trends in Ecology & Evolution* **24(2)**:110–117 DOI 10.1016/j.tree.2008.09.011.

Waugh J. 2007. DNA barcoding in animal species: progress, potential and pitfalls. *BioEssays* **29(2)**:188–197 DOI 10.1002/bies.20529.

Yu C, Hernandez T, Zheng H, Yau SC, Huang HH, He RL, Yang J, Yau SST. 2013. Real time classification of viruses in 12 dimensions. *PLOS ONE* **8(5)**:e64328 DOI 10.1371/journal.pone.0064328.

Yusim K, Korber BT, Brander C, Barouch D, De Boer R, Haynes BF, Watkins D. 2016. *HIV Molecular Immunology 2015 (No. LA-UR-16-22283)*. Los Alamos: Los Alamos National Lab.(LANL).

Zhang J, Kapli P, Pavlidis P, Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* **29(22)**:2869–2876 DOI 10.1093/bioinformatics/btt499.