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

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

Affiliations

[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

[3]Karlsruhe Institute of Technology, Institute for Theoretical Informatics, 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

*Corresponding Author:
Paschalia Kapli
Darwin Building, Gower Street, London, WC1E 6BT, UK
Email address: p.kapli@ucl.ac.uk

Deleted: ic

Formatted: Portuguese (Brazil)

Deleted: Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Deleted: Centre for Life's Origins and Evolution, Department of Genetics Evolution and Environment, University College 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 find 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 acknowledged clusters among methods and taxonomically genotypes implies that virus classification can be successfully automated.

Deleted: -
Deleted: -
Deleted: ,
Formatted: Font: Times, 12 pt
Deleted:
Formatted: Font: Times, 12 pt
Deleted: well informed parameter
Deleted:
Formatted: Font: Times, 12 pt
Deleted: acknowledge

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; Savoleinen et al., 2005; Waugh, 2007; Bucklin et al., 2010; Valentini et al., 2009; Li et al., 2014*), while it also provided a new means of analysis in several fields (*Galimberti et al., 2013; Mishra et al., 2015; Lewray & Knowlton, 2015; Bell et al., 2016; Batovska et al., 2017*). 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., 2012a; Stadler et al., 2014b; 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 et al., 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

Deleted: provide

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 et al., 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

Deleted: distance based

Deleted: ,

Deleted: a

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 10kb (*Radziwill et al., 1990; Tang et al., 2001; Martell et al., 1992*). Both virus types comprise at least two taxonomic levels (HBV: genotypes, sub-genotypes; 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*).

Materials 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 Suppl. Appendix).

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 et al., 2014*). The inter-genotypic and inter-subgenotypic divergence exceeds 8% and 4%-8%, respectively across the genome. No sub-genotypes 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, 2013; Biswas, 2013; Moura, 2013; Shi, 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

Deleted: obtained

Deleted: two previously published

Deleted: (*Kramvis, 2014; Smith et al., 2014*, respectively)

Deleted: .

Deleted: differ

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, Suppl. Fig. 1 and 2, Suppl. Appendix). 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 (Suppl. Fig. 1). 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; 6h, 6i, 6j, 6k, 6l, 6m, 6n, 6xb, 6xe; 6xa; 6v (Suppl. Fig. 1, Suppl. Appendix). Similarly, for genotype 3, the delimited clusters were i) 3g, 3b, 3i, 3a, 3e, 3d, ii) 3k, and iii) 3h 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

Deleted: '

Deleted: lack of adequate

Deleted: for the subtype

Deleted: 3

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.

The so far classification of HCV into genotypes and subtypes has been defined mostly by visual identification of clades in phylogenetic inference of HCV sequences (*Simmonds 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., 2015*) 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., 2011; 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 et al., 2007*) of the virus that comprises eight genotypes, except for subgenotype C4 which formed a new cluster (Fig. 3, Suppl. Fig. 3 and 4, Suppl. Appendix). 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

Deleted: 3

Deleted: distance based

Deleted: ying

Deleted: is

Deleted: for

Deleted: 2

Deleted: 2

matched the eight genotypes of the HBV classification (Fig. 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., subtype C4 was recovered as a distinct cluster from the remaining seven subtypes of genotype.

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.

Acknowledgments

The authors gratefully acknowledge the support of the Klaus Tschira Foundation.

References

Deleted: 2

Deleted: The application of mPTP to the HCV and HBV data sets shows that automated viral strain clustering using phylogeny-aware methods yields clusters that largely agree with the current standard taxonomy

Deleted: The major advantage of mPTP over distance-based approaches is that it 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. On the contrary, 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. …

Deleted: Acknowledgements

Formatted: Portuguese (Brazil)

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, *106*(51), 21484-21489.

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.

Batovska, J., Lynch, S. E., Cogan, N. O. I., Brown, K., Darbro, J. M., Kho, E. A., & Blacket, M. J. (2018). Effective mosquito and arbovirus surveillance using metabarcoding. *Molecular ecology resources*, *18*(1), 32-40.

Bell, K. L., Burgess, K. S., Okamoto, K. C., Aranda, R., & Brosi, B. J. (2016). Review and future prospects for DNA barcoding methods in forensic palynology. *Forensic Science International: Genetics*, *21*, 110-116.

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.

Bolotov, I. N., Vikhrev, I. V., Kondakov, A. V., Konopleva, E. S., Gofarov, M. Y., Aksenova, O. V., & Tumpeesuwan, S. (2017). New taxa of freshwater mussels (Unionidae) from a species-rich but overlooked evolutionary hotspot in Southeast Asia. *Scientific Reports*, *7*(1), 11573.

Breitbart, M., & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, *13*(6), 278-284.

Bucklin, A., Hopcroft, R. R., Kosobokova, K. N., Nigro, L. M., Ortman, B. D., Jennings, R. M., & Sweetman, C. J. (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.

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

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.

Formatted: Portuguese (Brazil)

Ge, D., Fellay, J., Thompson, A. J., Simon, J. S., Shianna, K. V., Urban, T. J., Heinzen, E.L., Qiu, P., Bertelsen, A.H., Muir, A.J., & Sulkowski, M. (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature*, *461*(7262), 399.

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., 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*, *111*(22), 8007-8012.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C.B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnie, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L. B., Chapman, S.B., Bochichio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheiffelin, J. S., Lander, E. S., Happi, C., Gevao, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H., & Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, *345*(6202), 1369-1372.

Huang, C. C., Kuo, T. M., Yeh, C. T., Hu, C. P., Chen, Y. L., Tsai, Y. L., Chen, M. L., 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.

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. *bioRxiv*, 447110.

Kramvis, A., & Kew, M. C. (2007). Epidemiology of hepatitis B virus in Africa, its genotypes and clinical associations of genotypes. *Hepatology Research*, *37*, S9-S19.

Kramvis, A. (2014). Genotypes and genetic variability of hepatitis B virus. *Intervirology*, *57*(3-4), 141-150.

Lauber, C., & Gorbalenya, A. E. (2012). Genetics-based classification of filoviruses calls for expanded sampling of genomic sequences. *Viruses*, *4*(9), 1425-1437.

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*, *112*(7), 2076-2081.

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: from gene to genome. *Biological Reviews*, *90*(1), 157-166.

Lu, L., Li, C., Yuan, J., Lu, T., Okamoto, H., & Murphy, D. G. (2013). Full-length genome sequences of five hepatitis C virus isolates representing subtypes 3g, 3h, 3i and 3k, and a unique genotype 3 variant. *The Journal of general virology*, *94*(Pt 3), 543.

Martell, M., Esteban, J. I., 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.

Messina, J. P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G. S., Pybus, O. G., & Barnes, E. (2015). Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology*, *61*(1), 77-87.

Mishra, P., Kumar, A., Nagireddy, A., Mani, D. N., Shukla, A. K., 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.

Modha, S., Thanki, A. S., Cotmore, S. F., Davison, A. J., & Hughes, J. (2018). ViCTree: an automated framework for taxonomic classification from protein sequences. *Bioinformatics*, *34*(13), 2195-2200.

Moritz, C., & Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS biology*, *2*(10), e354.

Moura, I. F., Lopes, E. P., Alvarado-Mora, M. V., Pinho, J. R., & Carrilho, F. J. (2013). Phylogenetic analysis and subgenotypic distribution of the hepatitis B virus in Recife, Brazil. *Infection, Genetics and Evolution*, *14*, 195-199.

Nakano, T., Lau, G. M., Lau, G. M., 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.

Formatted: Portuguese (Brazil)

Norder, H., Couroucé, A. M., Coursaget, P., Echevarria, J. M., Shou-Dong, L., Mushahwar, I. K., & Magnius, L. O. (2004). Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. *Intervirology*, *47*(6), 289.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N. & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, *536*(7617), 425.

Paraskevis, D., Magiorkinis, G., Magiorkinis, E., Ho, S. Y., Belshaw, R., Allain, J. P., & Hatzakis, A. (2013). Dating the origin and dispersal of hepatitis B virus infection in humans and primates. *Hepatology*, *57*(3), 908-916.

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular ecology*, *21*(8), 1864-1877.

Pybus, O. G., Tatem, A. J., & Lemey, P. (2015). Virus evolution and transmission in an ever more connected world. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1821), 20142878.

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, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, *61*(3), 539-542.

Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., Apfel, W., Weigel, A., Kopetz, A., Köhler, J., Fritylar, F., Hartmann, M., Hadulla, K., Schmidt, J., Hörren, T., Krebs, D., Theves, F., Eulity, U., Skale, A., Rohwedder, D., Kleeberg, A., Astrin, I. I., Geiger, M. F., Wägele, W., Grobe, P., & Ahrens, D. (2017). Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes. *Methods in Ecology and Evolution*, *8*(12), 1878-1887.

Savolainen, V., Cowan, R. S., Vogler, A. P., Roderick, G. K., & 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.

Schaefer, S. (2007). Hepatitis B virus genotypes in Europe. *Hepatology Research*, *37*, S20-S26.

Shi, W., Zhang, Z., Ling, C., Zheng, W., Zhu, C., Carr, M. J., & Higgins, D. G. (2013). Hepatitis B virus subgenotyping: history, effects of recombination, misclassifications, and corrections. *Infection, genetics and Evolution*, *16*, 355-361.

Simmonds, P. (2015). Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology*, *96*(6), 1193-1206.

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davidson, A.J., Delwart, E., Gorbalenya A. E., Harrach, B., Hull, R., King, A. M. Q., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowity E. J., Nibert, M. L., Orton, R., Roossinck, M. J., Sabanadyovic, S., Sullivan, M. B., Suttle, C. A., Tesh, R. B., van der Vlugt, R. A., Varsani, A., & Zerbini, F. M. (2017). Consensus statement: virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161.

Simmonds, P., Bukh, J., Combet, C., Deléage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspe, G., Kuiken, C., Maertens, G., Miyokami, M., Murphy, D. G., Okamoto, H., Pawlotsky, J. M., Penin, F., Sablon, E., Shin-I, T., Stuyver, L. J., Thiel, H. J., Viayov, S., Weiner, A. J., & Widell, A. (2005). Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology*, *42*(4), 962-973.

Smith, D. B., Bukh, J., Kuiken, C., Muerhoff, A. S., Rice, C. M., Stapleton, J. T., & 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.

Strader, D. B., Wright, T., Thomas, D. L., & Seeff, L. B. (2004). Diagnosis, management, and treatment of hepatitis C. *Hepatology*, *39*(4), 1147-1171.

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., Günthard, H. F., Drummond, A. J., & 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, D. A., & du Plessis, L. (2014). Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS currents*, *6*.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Stanaway, J. D., Flaxman, A. D., Naghavi, M., Fitzmaurice, C., Vos, T., Abubakar, I., Abu-Raddad, L., Assadi, R., Bhala, N., Cowie, B., Forouzanfour, M. H., Groeger, J., Hanafiah, M., Jacobsen, K. H., James, S. L., MacLachlan, J., Malekyadeh, R., Martin, N. K., Mokhad, A. A., Mokdad, A. H., Murray, C. J. L., Plass, D., Rana, S., Rein, D. B., Richardus, J. H., Sanabria, J., Sazyan, M., Shahraz, S., So, S., Vlassov, V. V., Weiderpass, E., Wiersma, S. T., Younis, M., Yu, C., El Sayed Zaki, M., & Cooke, G. S.(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.

Formatted: Portuguese (Brazil)

Sulbaran, M. Z., Di Lello, F. A., Sulbaran, Y., Cosson, C., Loureiro, C. L., Rangel, H. R.,Cantaloube, J.F., Campos, R.H., Moratorio, G., Cristina, H.,& Pujol, F. H. (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.

Suttle, C. A. (2013). Viruses: unlocking the greatest biodiversity on Earth. *Genome*, *56*(10), 542-544.

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.

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*, *98*(4), 1841-1846.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in ecology & evolution*, *18*(2), 70-74.

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.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. *Biological conservation*, *183*, 4-18.

Tong, Y. Q., Liu, B., Liu, H., Zheng, H. Y., Gu, J., Song, E. J., 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.

Waugh, J. (2007). DNA barcoding in animal species: progress, potential and pitfalls. *BioEssays*, *29*(2), 188-197.

Yu, C., Hernandez, T., Zheng, H., Yau, S. C., Huang, H. H., He, R. L., Yang, J., & Yau, S. S. T. (2013). Real time classification of viruses in 12 dimensions. *PloS one*, *8*(5), e64328.

Yusim, K., Korber, B. T., Brander, C., Barouch, D., de Boer, R., Haynes, B. F., & Watkins, D. (2016). *HIV Molecular Immunology 2015* (No. LA-UR-16-22283). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Deleted: Thézé, J., Lopez-Vaamonde, C., Cory, J. S., & Herniou, E. A. (2018). *Viruses*, *10*(7), 366

Deleted: ...

Formatted: Portuguese (Brazil)

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, *29*(22), 2869-2876.