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], **Paschalia Kapli** [Corresp. 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

Corresponding Author: Paschalia Kapli
Email address: p.kapli@ucl.ac.uk

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 clusters with taxonomically acknowledged genotypes implies that virus classification can be successfully automated.

1 **Title:** Automated, phylogeny-based genotype delimitation of the Hepatitis Viruses HBV and
2 HCV

4 **Authors:** Dora Serdari[1], Evangelia-Georgia Kostaki[2], Dimitrios Paraskevis[2], Alexandros
5 Stamatakis[1,3], Paschalia Kapli[4*]
6
7 **Affiliations**
8 [1]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies,
9 Heidelberg, Germany
10 [2]Department of Hygiene Epidemiology and Medical Statistics, School of Medicine, National and
11 Kapodistrian University of Athens
12 [3]Karlsruhe Institute of Technology, Institute for Theoretical Informatics, Karlsruhe, Germany
13 [4]Centre for Life's Origins and Evolution, Department of Genetics Evolution and Environment,
14 University College London, University of London, London, United Kingdom
15
16
17
18
19
20
21
22
23
24
25 *Corresponding Author:
26 Paschalia Kapli
27 Darwin Building, Gower Street, London, WC1E 6BT, UK
28 **Email address:** p.kapli@ucl.ac.uk
29

30   **Abstract**

31

32   **Background:** The classification of hepatitis viruses still predominantly relies on ad hoc criteria,

33   i.e., phenotypic traits and arbitrary genetic distance thresholds. Given the subjectivity of such

34   practices coupled with the constant sequencing of samples and discovery of new strains, this

35   manual approach to virus classification becomes cumbersome and impossible to generalize.

36

37   **Methods:**

38   Using two well-studied hepatitis virus datasets, HBV and HCV, we assess if computational

39   methods for molecular species delimitation that are typically applied to barcoding biodiversity

40   studies can also be successfully deployed for hepatitis virus classification. For comparison, we

41   also used ABGD, a tool that in contrast to other distance methods attempts to automatically

42   identify the barcoding gap using pairwise genetic distances for a set of aligned input sequences.

43

44   **Results - Discussion:**

45   We find that, the mPTP species delimitation tool identified even without adapting its default

46   parameters taxonomic clusters that either correspond to the currently acknowledged genotypes or

47   to known subdivision of genotypes (subtypes or subgenotypes). In the cases where the delimited

48   cluster corresponded to subtype or subgenotype, there were previous concerns that their status

49   may be underestimated. The clusters obtained from the ABGD analysis differed depending on

50   the parameters used. However, under certain values the results were very similar to the taxonomy

51   and mPTP which indicates the usefulness of distance based methods in virus taxonomy under

52   appropriate parameter settings . The overlap of predicted clusters with taxonomically

53   acknowledged genotypes implies that virus classification can be successfully automated.

54

55

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

87   used for other life forms, such as plants or animals. An analogous hierarchical classification

88   system containing orders, families, subfamilies, genera, and species is being applied (*Simmonds,*

89   *2015*). The ICTV is typically based on phenotypic criteria, such as morphology, nucleic acid

90   type (i.e., DNA or RNA), hosts, symptoms, mode of replication, geographical data, or presence

91   of antigenic epitopes, to name a few. Generally, such criteria, despite being informative, can be

92   subjective, require highly specialized knowledge, and are time consuming to apply. In contrast,

93   sequence evolution takes into account the evolutionary history of life forms and, thus, may offer

94   a more objective source of information for taxonomic classification. An important difference in

95   viruses compared to other organisms is that they lack a common set of universal genes such as

96   the 18S rRNA in eukaryotes or the 16S rRNA in prokaryotes. Therefore, we cannot infer a

97   comprehensive virus tree of life (*Simmonds et al., 2017*), and, more importantly for species

98   delimitation, we cannot rely upon barcoding markers that are universally suitable for all viruses.

99   We can nonetheless gain valuable insights for their systematics by utilizing phylogenetic

100   information at lower taxonomic ranks (e.g., families, genera, species), using appropriate genes

101   for each dataset. In this context, methods using genetic-distance thresholds (*Bao et al., 2014,*

102   *Lauber & Gorbalenya, 2012, Yu et al., 2013*) have been suggested as a complementary method

103   to the traditional virus classification for accelerating new species identification.

104   In this study, we explore whether a recently developed algorithm for molecular species

105   delimitation on barcoding or marker gene phylogenies can be deployed for ICTV. In contrast to

106   genetic distance-based methods the multi-rate Poisson Tree Processes (mPTP, *Kapli et al., 2017*)

107   infers the number of genetic clusters given a phylogenetic input tree. Such trees can easily be

108   inferred using both, Maximum Likelihood (*Stamatakis, 2014*), or Bayesian approaches (*Ronquist*

109   *et al., 2012*) on single-gene or multi-gene multiple sequence alignments. The fundamental

110   assumption of the model is that variance in the data, as represented by the phylogeny, is greater

111   among species than within a species (*Zhang et al., 2013*). The additional assumption of mPTP,

112   that the genetic variation may differ substantially among species allows to accurately delimit

113   species in large (meta-) barcoding datasets comprising multiple species of diverse life histories

114   (*Kapli et al., 2017*). Experiments using empirical data for several animal phyla (*Kapli et al.,*

115   *2017*) and recently also viruses (*Thézé et al., 2018; Modha et al, 2018*) show that the method

116   consistently provides extremely fast and sensible species estimates on 'classic' phylogenetic

117   marker and barcoding genes.

118   To assess whether mPTP can be deployed as a quantitative ICTV method we analyze two

119   medically important viruses, Hepatitis B (HBV) and Hepatitis C (HCV) that are leading global

120   causes of human mortality (*Stanaway et al., 2016*). Both viruses cause liver inflammation, but

121   are substantially different from each other. HBV has a partially double-stranded circular DNA

122   genome with a length of about 3.2 kb while HCV is a single-stranded, positive-sense RNA virus,

123   with a genome length of approximately 10kb (*Radziwill et al., 1990; Tang et al., 2001; Martell*

124   *et al., 1992*). Both virus types comprise at least two taxonomic levels (HBV: genotypes, sub-

125   genotypes; HCV: genotypes, subtypes). Besides the significance of the two viruses for human

126   health, we selected them as test cases since due to the substantial amount of taxonomic research

127   that has been conducted and that we can hence use to assess the efficiency of genetic clustering

128   (e.g., *Simmonds et al., 2005; Schaefer, 2007; Smith et al., 2014; Messina et al., 2015*).

129

130

131   **Materials and methods**

132

133   *Datasets*

134   We generated multiple sequence alignments (MSAs) corresponding to two virus types: HBV and

135   HCV from two sets of full-length genomic sequences downloaded from publicly available

136   databases (NCBI: http://www.ncbi.nlm.nih.gov, Accession Numbers provided in Suppl.

137   Appendix)

138   The HBV dataset comprises 110 sequences corresponding to eight genotypes (i.e., A-H) and 31

139   subgenotypes. The genotypes (A through D, F, and H) have been further divided into

140   subgenotypes indexed by numbers for the corresponding genotype (e.g., A_1, A_2, B_1, B_2, B_3, etc.;

141   *Kramvis et al., 2014*). The inter-genotypic and inter-subgenotypic divergence exceeds 8% and

142   4%-8%, respectively across the genome. No sub-genotypes have been reported for genotypes E,

143   G and H which shows that they are of lower levels of genetic divergence than the rest. The

144   distribution of HBV genotypes differs greatly with respect to the geographical origin. Moreover,

145   they differ in their natural history, response to treatment and disease progression (*Huang, 2013;*

146   *Biswas, 2013; Moura, 2013; Shi, 2013*). For our study we included the sequences of the eight

147   genotypes (A-H) that form part of the oldest identified HBV groups.

148   The HCV dataset I) comprises 213 sequences corresponding to seven major taxonomic

149   units named after genotypes (1, 2, 3, 4, 5, 6, and 7) and numerous subtypes (*Smith et al., 2014*).

150   The HCV classification into genotypes and subtypes was based on genetic-distance thresholds

151   that were verified by the fact that they formed monophyletic clades in an inferred phylogeny

152   (*Smith et al., 2014*). Therefore, the HCV classification serves as an appropriate test case for

153   assessing whether a similar clustering can be identified with a more objective and automated

154   method, such as mPTP, that does not require any user input apart from a phylogeny.

155

156   *Genetic Cluster delimitation*

157   To delimit the putative species, additionally to mPTP, we used the distance-based "*Automatic*

158   *Barcode Gap Discovery*" tool (ABGD, *Puillandre et al., 2012*). ABGD is a popular distance-

159   based barcoding method that, compared to other distance-based methods attempts to

160   automatically identify the threshold value for the transition from intra-specific variation to inter-

161   specific divergence (*Puillandre et al., 2012*).

162   For the mPTP delimitation, a fully binary (bifurcating) rooted phylogeny is required.

163   Therefore, using the aligned sequences we inferred the phylogenetic relationships under the

164   GTR+Γ model of nucleotide substitution using RAxML-NG (*Kozlov et al., 2018*). We rooted the

165   phylogenetic trees according to the originally published phylogenies (i.e., using the branch

166   leading to genotypes F/H for HBV and genotype 7 for HCV). Using heuristic search algorithms

167   for finding the 'best' delimitation given the rooted phylogeny and without any further prior

168   assumptions. We performed the mPTP delimitation under Maximum Likelihood (ML) and

169   calculated the support of the delimited clusters using Markov-chain Monte Carlo (MCMC)

170   sampling (*Kapli et al., 2017*). We conducted the MCMC sampling twice for 10^6 generations, to

171   identify potential lack of convergence with a sampling frequency of 0.1.

172   For ABGD, the user has to define two important parameters, i) the prior maximum

173   divergence of intraspecific diversity (*P*), which implies that the barcode gap is expected to

174   exceed this value and should not be confused with the genetic thresholds assumed to define the

175   inter-specific relationships, ii) a proxy for the minimum gap width (*X*), which indicates that the

176   barcoding gap is expected to be *X* times larger than any intraspecific gap *(Puillandre et al.,*

177   *2012*). For both, HBV, and HCV, we used 10 prior maximum thresholds in the range of p =

178   0.001 and P = 0.05. The proxy for the minimum gap width (*X*) was set to the default value (X =

179 1.5) for HCV, while for HBV the default value did not yield any delimitation and we therefore

180 set it to a lower value (X = 0.5).

181

182

183 **Results & Discussion**

184

185 The biodiversity of viruses is tremendous and it is broadly accepted that our understanding of

186 their ecology and evolution is constrained to a small fraction of species (*Paez-Espino et al.,*

187 *2016)*. In just a kilo of marine sediment there can be a million of different viral genotypes

188 (*Breitbart & Rohwer, 2005*), while on a global scale the number of viruses is 10 million-fold

189 higher than the number of stars in the universe (*Suttle, 2013*). The classification of such a diverse

190 set of organisms constitutes a challenging task and is impossible to accomplish within reasonable

191 time using phenotypic characters. Quantitative computational methods could provide a viable

192 alternative, particularly for large scale clustering and fast identification of viral strains

193 (*Simmonds et al., 2017; Modha et al., 2018*). Using empirical data of the HBV and HCV viruses

194 we show that by applying phylogeny-aware and distance-based tools to classify the strains of the

195 two virus types, the corresponding genetic clustering closely recovers their currently accepted

196 taxonomy.

197

198 **HCV Clustering**

199 The current taxonomy of HCV comprises seven genotypes, while mPTP yielded 16 genetic

200 clusters (Fig. 1, Suppl. Fig. 1 and 2, Suppl. Appendix). From the 16 clusters, five were congruent

201 with the current taxonomy, i.e, genotypes 1, 2, 4, 5 and 7. On the contrary, genotype 3 and

202 genotype 6 were further split into three and eight sub-clusters correspondingly (Fig. 1), which

203 corroborates former views that divergent variants of these genotypes may qualify as separate

204 major genotypes (*Simmonds et al., 2005*, *Smith et al, 2014*). In particular, the additional clusters

205 identified by mPTP correspond to previously identified groups of subtypes (Suppl. Fig. 1). For

206 genotype 6, these clusters consisted of the following subtype groups: 6a; 6b and 6xd; 6c, 6d, 6e,

207 6f, 6g, 6o, 6p, 6q, 6r, 6s, 6t, 6u, 6w, 6xc and 6xf; 6h, 6i, 6j, 6k, 6l, 6m, 6n, 6xb, 6xe; 6xa; 6v

208 (Suppl. Fig. 1, Suppl. Appendix). Similarly, for genotype 3, the delimited clusters were i) 3g, 3b,

209 3i, 3a, 3e, 3d, ii) 3k, and iii) 3h and 3. All clusters were substantially supported by the MCMC

210   sampling, except the split of 3k subtype from its sister group (Fig. 1), which may be due to the

211   limited amount of corresponding sequences.

212   The number of clusters inferred with ABGD ranged from 1 to 208 depending on the

213   value of the maximum intraspecific divergence threshold (Fig. 2). The most reasonable result

214   (i.e., the one closest to the current standard taxonomy) comprised 19 clusters and was obtained

215   for a minimum of intraspecific genetic diversity of 5.99% (i.e., p=0.0599). Under this threshold,

216   the delimitation is largely identical to the delimitation obtained with mPTP (Fig. 1), with three

217   differences: i) that genotype 3 was split into four clusters, instead of three, ii) genotype six was

218   divided into nine clusters instead of eight, and, iii) genotype 7 is divided into two clusters. When

219   the prior intraspecific divergence was increased to a higher minimum of 10%, all sequences were

220   grouped in a single cluster. When the threshold was set to a lower value (3.6%) the number of

221   clusters increased to 135 (Fig. 2). Nevertheless, the delimitation with the 5.99% threshold is

222   largely congruent to current taxonomy and the clusters obtained with mPTP, thus indicating the

223   usefulness of distance-based methods in virus taxonomy under well informed parameters.

224   The so far classification of HCV into genotypes and subtypes has been defined mostly by

225   visual identification of clades in phylogenetic inference of HCV sequences (*Simmonds et al.,*

226   *2005; Smith et al, 2014*). Specifically, the genotypes correspond to the seven major highly-

227   supported phylogenetic HCV clusters while subtypes were defined as the secondary hierarchical

228   clusters found within each genotype (*Smith et al, 2014*). This classification scheme has been

229   widely adopted (*Combet et al., 2007*; *Yusim et al., 2015*) and has been shown to be robust (in

230   terms of stability of the HCV phylogeny) and relevant for clinical practice, since response rates

231   to immunomodulatory treatment for the chronic hepatitis C differs across genotypes.

232   Nevertheless, new, unassigned lineages are often discovered from understudied areas (*Sulbaran*

233   *et al., 2010; Nakano et al., 2011; Lu et al., 2013; Tong et al., 2015*) and it is challenging to

234   assign them a taxonomic status, given that the genetic distance cut-off among intra and inter-

235   specific relationships is arbitrary and variable for different parts of the HCV phylogeny

236   (*Simmonds et al., 2005*). The greatly overlapping mPTP and ABGD clusters with the HCV

237   genotypes shows that the classification, and, consequently, the identification, of the genotypes

238   can be easily automated utilizing objective, transparent, and unifying approaches. Embracing

239   such alternatives can be crucial for viruses like HCV, taking into account that the correct

240   identification of the HCV genotypes could be of clinical importance in providing the appropriate

241   medical treatment (*Strader et al., 2004; Ge et al., 2009*).

242

243   **HBV clustering**

244   In the case of HBV, the mPTP clustering is almost identical to the current classification (*Norder*

245   *et al., 2004; Kramvis et al., 2007*) of the virus that comprises eight genotypes, except for

246   subgenotype C4 which formed a new cluster (Fig. 3, Suppl. Fig. 3 and 4, Suppl. Appendix). This

247   is in line with the greater genetic divergence of C4 compared to the other subgenotypes due to its

248   ancient origin in native populations in Oceania (*Paraskevis et al, 2013*). However, the split of C4

249   from its sister cluster (genotype C) is not supported by the MCMC sampling, potentially

250   reflecting the lack of adequate sampling. On the other hand, the number of clusters identified by

251   ABGD varied from 1 to 85 under different thresholds of minimum intraspecific divergence,

252   while the delimitation for a threshold of 1.29% exactly matched the eight genotypes of the HBV

253   classification (Fig. 2 and 3). Both ABGD and mPTP identified seven of the genotypes (A-F) as

254   distinct genetic clusters. The only difference was that mPTP split genotype C into two distinct

255   clusters (Fig. 3), i.e., subtype C4 was recovered as a distinct cluster from the remaining seven

256   subtypes of genotype.

257

258   **Conclusions**

259   The application of mPTP to the HCV and HBV data sets shows that automated viral species

260   delimitation using phylogeny-aware methods yields clusters that largely agree with the current

261   standard taxonomy. The additional clusters identified for HCV by mPTP is not surprising as they

262   have been previously considered divergent sub-clusters within the genotypes 3 and 6.

263   Analogously, for HBV, mPTP yielded almost identical results to the current nomenclature

264   system with the exception of a single sub-genotype, C4, that was previously mentioned to be

265   more genetically divergent within genotype C (Paraskevis et al, 2013). In both cases, these new

266   clusters indicate the potential need for taxonomic revision. However, given the wide use of the

267   current nomenclature in the medical field, and the lack of other sources of information such as

268   recombination, particularly for HBV, and, response to treatment, we wouldn't suggest taxonomic

269   changes at present. Regarding distance methods, the example of HCV and HBV, shows that

270   meaningful parameter values for distance- based methods may differ substantially among

271   datasets, and, therefore, establishing global thresholds is impossible. On the contrary, mPTP can

272   be seamlessly applied to taxa of substantially different life histories (e.g., variable population

273   sizes, evolution rates), as it does not require any input parameters except a phylogeny. Overall,

274   the ease-of-use of mPTP in conjunction with its computational efficiency on phylogenies with

275   hundreds of samples render it a useful tool for viral biodiversity estimates, initial classification of

276   understudied taxa, and accelerating the viral species identification process.

277

278

279

280

281

282   **Acknowledgments**

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

284

285   **References**

286

287   Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009).

288   Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the*

289   *National Academy of Sciences*, *106*(51), 21484-21489.

290

291   Bao, Y., Chetvernin, V., & Tatusova, T. (2014). Improvements to pairwise sequence comparison

292   (PASC): a genome-based web tool for virus classification. *Archives of virology*, *159*(12), 3293-

293   3304.

294

295   Batovska, J., Lynch, S. E., Cogan, N. O. I., Brown, K., Darbro, J. M., Kho, E. A., & Blacket, M.

296   J. (2018). Effective mosquito and arbovirus surveillance using metabarcoding. *Molecular*

297   *ecology resources*, *18*(1), 32-40.

298

299   Bell, K. L., Burgess, K. S., Okamoto, K. C., Aranda, R., & Brosi, B. J. (2016). Review and

300   future prospects for DNA barcoding methods in forensic palynology. *Forensic Science*

301   *International: Genetics*, *21*, 110-116.

302

303   Biswas, A., Panigrahi, R., Pal, M., Chakraborty, S., Bhattacharya, P., Chakrabarti, S., &

304   Chakravarty, R. (2013). Shift in the hepatitis B virus genotype distribution in the last decade

305   among the HBV carriers from eastern India: possible effects on the disease status and HBV

306   epidemiology. *Journal of medical virology*, *85*(8), 1340-1347.

307

308 Bolotov, I. N., Vikhrev, I. V., Kondakov, A. V., Konopleva, E. S., Gofarov, M. Y., Aksenova, O.
309 V., & Tumpeesuwan, S. (2017). New taxa of freshwater mussels (Unionidae) from a species-rich
310 but overlooked evolutionary hotspot in Southeast Asia. *Scientific Reports*, *7*(1), 11573.
311

312 Breitbart, M., & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus?
313 *Trends in microbiology*, *13*(6), 278-284.
314

315 Bucklin, A., Hopcroft, R. R., Kosobokova, K. N., Nigro, L. M., Ortman, B. D., Jennings, R. M.,
316 & Sweetman, C. J. (2010). DNA barcoding of Arctic Ocean holozooplankton for species
317 identification and recognition. *Deep Sea Research Part II: Topical Studies in Oceanography*,
318 *57*(1-2), 40-48.
319

320 Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., A., Dehne-Garcia, C.,
321 Geourjon, E., Bettler, C., H., P., Le Mercier, R., Bartenschlager, H., Diepolder, D., Moradpour,
322 J.-M., Pawlotsky, C., M., Rice, C., Trépo, F., Penin, G., Deléage. euHCVdb: the European
323 hepatitis C virus database. *Nucleic Acids Res* 2007; 35: D363- D366.
324

325 Galimberti, A., De Mattia, F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., Martelos, S., &
326 Labra, M. (2013). DNA barcoding as a new tool for food traceability. Food Research
327 International, 50(1), 55-63.
328

329 Ge, D., Fellay, J., Thompson, A. J., Simon, J. S., Shianna, K. V., Urban, T. J., Heinzen, E.L.,
330 Qiu, P., Bertelsen, A.H., Muir, A.J., & Sulkowski, M. (2009). Genetic variation in IL28B
331 predicts hepatitis C treatment-induced viral clearance. *Nature*, *461*(7262), 399.
332

333 Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., Hallwachs,
334 W., & Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a
335 bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National
336 Academy of Sciences*, *111*(22), 8007-8012.
337

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

348

349 Huang, C. C., Kuo, T. M., Yeh, C. T., Hu, C. P., Chen, Y. L., Tsai, Y. L., Chen, M. L., Chang,
350 C., & Chang, C. (2013). One single nucleotide difference alters the differential expression of
351 spliced RNAs between HBV genotypes A and D. *Virus research*, *174*(1-2), 18-26.

352

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

356

357 Kozlov, A., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2018). RAxML-NG: A fast,
358 scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *bioRxiv*,
359 447110.

360

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

363

364 Kramvis, A. (2014). Genotypes and genetic variability of hepatitis B virus. *Intervirology*, *57*(3-
365 4), 141-150.

366

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

369

370 Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples
371 reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*,
372 *112*(7), 2076-2081.

373

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

376

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

380

381 Martell, M., Esteban, J. I., Quer, J., Genesca, J., Weiner, A., Esteban, R., Guardia, J., & Gomez,
382 J. (1992). Hepatitis C virus (HCV) circulates as a population of different but closely related
383 genomes: quasispecies nature of HCV genome distribution. *Journal of virology*, *66*(5), 3225-
384 3229.

385

386 Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., & Vespignani, A. (2011). Modeling
387 human mobility responses to the large-scale spreading of infectious diseases. *Scientific reports*,
388 *1*, 62.

389

390 Messina, J. P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G. S., Pybus, O. G., & Barnes, E.
391 (2015). Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology*, *61*(1), 77-
392 87.

393

394 Mishra, P., Kumar, A., Nagireddy, A., Mani, D. N., Shukla, A. K., Tiwari, R., & Sundaresan, V.
395 (2016). DNA barcoding: an efficient tool to overcome authentication challenges in the herbal
396 market. *Plant biotechnology journal*, *14*(1), 8-21.

397

398 Modha, S., Thanki, A. S., Cotmore, S. F., Davison, A. J., & Hughes, J. (2018). ViCTree: an
399 automated framework for taxonomic classification from protein sequences. *Bioinformatics*,
400 *34*(13), 2195-2200.

401

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

404

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

408

409 Nakano, T., Lau, G. M., Lau, G. M., Sugiyama, M., & Mizokami, M. (2012). An updated
410 analysis of hepatitis C virus genotypes and subtypes based on the complete coding region. *Liver*
411 *international*, *32*(2), 339-345.

412

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

416

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

420

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

424

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

427

428 Pybus, O. G., Tatem, A. J., & Lemey, P. (2015). Virus evolution and transmission in an ever
429 more connected world. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1821),
430 20142878.
431
432 Radziwill, G., Tucker, W., & Schaller, H. (1990). Mutational analysis of the hepatitis B virus P
433 gene product: domain structure and RNase H activity. *Journal of virology*, *64*(2), 613-620.
434
435 Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B.,
436 Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian
437 phylogenetic inference and model choice across a large model space. *Systematic biology*, *61*(3),
438 539-542.
439
440 Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., Apfel, W., Weigel,
441 A., Kopetz, A., Köhler, J., Fritylar, F., Hartmann, M., Hadulla, K., Schmidt, J., Hörren, T.,
442 Krebs, D., Theves, F., Eulity, U., Skale, A., Rohwedder, D., Kleeberg, A., Astrin, I. I., Geiger,
443 M. F., Wägele, W., Grobe, P., & Ahrens, D. (2017). Using taxonomic consistency with
444 semi-automated data pre-processing for high quality DNA barcodes. *Methods in Ecology and*
445 *Evolution*, *8*(12), 1878-1887.
446
447 Savolainen, V., Cowan, R. S., Vogler, A. P., Roderick, G. K., & Lane, R. (2005). Towards
448 writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions*
449 *of the Royal Society B: Biological Sciences*, *360*(1462), 1805-1811.
450
451 Schaefer, S. (2007). Hepatitis B virus genotypes in Europe. *Hepatology Research*, *37*, S20-S26.
452
453 Shi, W., Zhang, Z., Ling, C., Zheng, W., Zhu, C., Carr, M. J., & Higgins, D. G. (2013). Hepatitis
454 B virus subgenotyping: history, effects of recombination, misclassifications, and corrections.
455 *Infection, genetics and Evolution*, *16*, 355-361.
456
457 Simmonds, P. (2015). Methods for virus classification and the challenge of incorporating
458 metagenomic sequence data. *Journal of General Virology*, *96*(6), 1193-1206.
459
460 Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davidson,
461 A.J., Delwart, E., Gorbalenya A. E., Harrach, B., Hull, R., King, A. M. Q., Koonin, E. V.,
462 Krupovic, M., Kuhn, J. H., Lefkowity E. J., Nibert, M. L., Orton, R., Roossinck, M. J.,
463 Sabanadyovic, S., Sullivan, M. B., Suttle, C. A., Tesh, R. B., van der Vlugt, R. A., Varsani, A.,
464 & Zerbini, F. M. (2017). Consensus statement: virus taxonomy in the age of metagenomics.
465 *Nature Reviews Microbiology*, *15*(3), 161.
466
467 Simmonds, P., Bukh, J., Combet, C., Deléage, G., Enomoto, N., Feinstone, S., Halfon, P.,
468 Inchauspe, G., Kuiken, C., Maertens, G., Miyokami, M., Murphy, D. G., Okamoto, H.,

469 Pawlotsky, J. M., Penin, F., Sablon, E., Shin-I, T., Stuyver, L. J., Thiel, H. J., Viayov, S.,
470 Weiner, A. J., & Widell, A. (2005). Consensus proposals for a unified system of nomenclature of
471 hepatitis C virus genotypes. *Hepatology*, *42*(4), 962-973.

473 Smith, D. B., Bukh, J., Kuiken, C., Muerhoff, A. S., Rice, C. M., Stapleton, J. T., & Simmonds,
474 P. (2014). Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated
475 criteria and genotype assignment web resource. *Hepatology*, *59*(1), 318-327.

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

480 Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B.,
481 Rieder, P., Xie, D., Günthard, H. F., Drummond, A. J., & Bonhoeffer, S., (2011). Estimating the
482 basic reproductive number from viral sequence data. *Molecular biology and evolution*, *29*(1),
483 347-357.

485 Stadler, T., Kühnert, D., Rasmussen, D. A., & du Plessis, L. (2014). Insights into the early
486 epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS currents*, *6*.

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

491 Stanaway, J. D., Flaxman, A. D., Naghavi, M., Fitzmaurice, C., Vos, T., Abubakar, I., Abu-
492 Raddad, L., Assadi, R., Bhala, N., Cowie, B., Forouzanfour, M. H., Groeger, J., Hanafiah, M.,
493 Jacobsen, K. H., James, S. L., MacLachlan, J., Malekyadeh, R., Martin, N. K., Mokhad, A. A.,
494 Mokdad, A. H., Murray, C. J. L., Plass, D., Rana, S., Rein, D. B., Richardus, J. H., Sanabria, J.,
495 Sazyan, M., Shahraz, S., So, S., Vlassov, V. V., Weiderpass, E., Wiersma, S. T., Younis, M., Yu,
496 C., El Sayed Zaki, M., & Cooke, G. S.(2016). The global burden of viral hepatitis from 1990 to
497 2013: findings from the Global Burden of Disease Study 2013. *The Lancet*, *388*(10049), 1081-
498 1088.

500 Sulbaran, M. Z., Di Lello, F. A., Sulbaran, Y., Cosson, C., Loureiro, C. L., Rangel, H.
501 R.,Cantaloube, J.F., Campos, R.H., Moratorio, G., Cristina, H.,& Pujol, F. H. (2010). Genetic
502 history of hepatitis C virus in Venezuela: high diversity and long time of evolution of HCV
503 genotype 2. *PloS one*, *5*(12), e14315.

505 Suttle, C. A. (2013). Viruses: unlocking the greatest biodiversity on Earth. *Genome*, *56*(10), 542-
506 544.

507

508 Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards
509 next-generation biodiversity assessment using DNA metabarcoding. *Molecular ecology*, *21*(8),
510 2045-2050.
511
512 Tang, H., & McLachlan, A. (2001). Transcriptional regulation of hepatitis B virus by nuclear
513 hormone receptors is a critical determinant of viral tropism. *Proceedings of the National*
514 *Academy of Sciences*, *98*(4), 1841-1846.
515
516 Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2003). A plea for DNA
517 taxonomy. *Trends in ecology & evolution*, *18*(2), 70-74.
518
519
520 Thézé, J., Lopez-Vaamonde, C., Cory, J., & Herniou, E. (2018). Biodiversity, evolution and
521 ecological specialization of baculoviruses: a treasure trove for future applied research. *Viruses*,
522 *10*(7), 366.
523
524 Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA–An emerging tool in conservation
525 for monitoring past and present biodiversity. *Biological conservation*, *183*, 4-18.
526
527 Tong, Y. Q., Liu, B., Liu, H., Zheng, H. Y., Gu, J., Song, E. J., Song, C., & Li, Y. (2015).
528 Accurate genotyping of hepatitis C virus through nucleotide sequencing and identification of
529 new HCV subtypes in China population. *Clinical microbiology and infection*, *21*(9), 874-e9.
530
531 Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in*
532 *ecology & evolution*, *24*(2), 110-117.
533
534 Waugh, J. (2007). DNA barcoding in animal species: progress, potential and pitfalls. *BioEssays*,
535 *29*(2), 188-197.
536
537 Yu, C., Hernandez, T., Zheng, H., Yau, S. C., Huang, H. H., He, R. L., Yang, J., & Yau, S. S. T.
538 (2013). Real time classification of viruses in 12 dimensions. *PloS one*, *8*(5), e64328.
539
540 Yusim, K., Korber, B. T., Brander, C., Barouch, D., de Boer, R., Haynes, B. F., & Watkins, D.
541 (2016). *HIV Molecular Immunology 2015* (No. LA-UR-16-22283). Los Alamos National
542 Lab.(LANL), Los Alamos, NM (United States).
543
544 Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method
545 with applications to phylogenetic placements. *Bioinformatics*, *29*(22), 2869-2876.
546

Figure 1

Clustering of the HCV samples into genotypes

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 Suppl. Fig. 2). The phylogenetic relationships were inferred using RAxML under the GTR+Γ model.
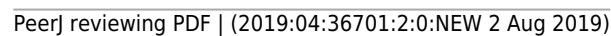
Figure 2

Clustering of the HBV samples into genotypes

Figure 2: 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 Suppl. Fig. 2). The phylogenetic relationships were inferred using RAxML under the GTR+Γ model.
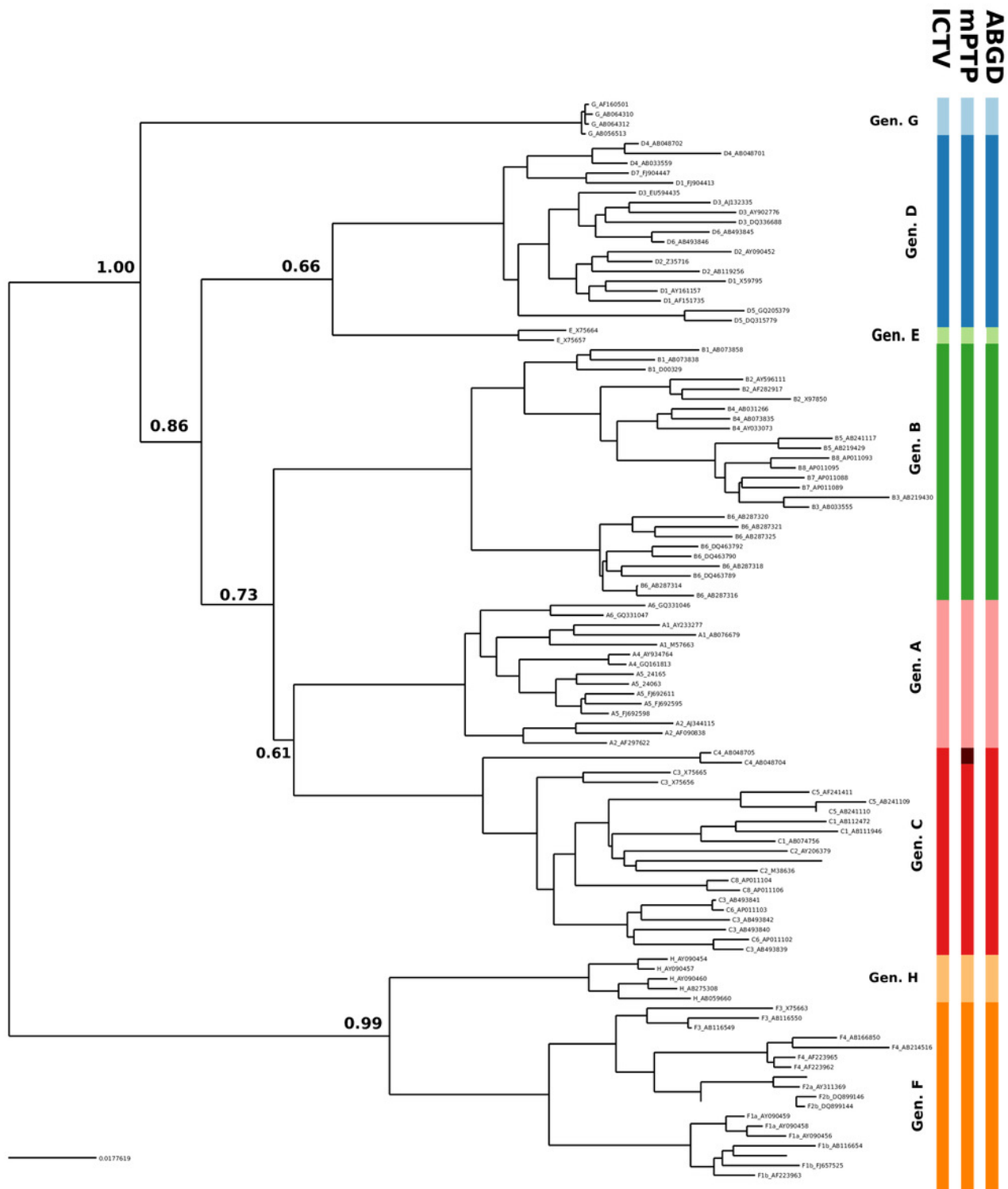
Figure 3

Number of delimited clusters for ABGD with respect to input parameters

Figure 3: 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.