Analysis of genomic-length HBV sequences to determine genotype and subgenotype reference sequences

Anna L McNaughton[1], Peter Revill[2,3], Margaret Littlejohn[2,3], Philippa C Matthews[1,4,5*], M Azim Ansari[6*]

[1] Nuffield Department of Medicine, Peter Medawar Building for Pathogen Research, South Parks Road, Oxford OX1 3SY, UK

[2] Victorian Infectious Diseases Reference Laboratory, Royal Melbourne Hospital at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia;

[3] Department of Microbiology and Immunology, University of Melbourne. Melbourne, Australia.

[4] Department of Infectious Diseases and Microbiology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headley Way, Oxford OX3 9DU, UK

[5] Oxford NIHR Biomedical Research Centre, John Radcliffe Hospital, Headley Way, Oxford OX3 9DU

[6] Wellcome Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

***Corresponding author:**

Please note that Philippa C Matthews and M. Azim Ansari are joint corresponding authors

Philippa Matthews; Address: Medawar Building for Pathogen Research, South Parks Road, Oxford OX1 3SY, UK.

Email: p.matthews@doctors.org.uk

M. Azim Ansari; The Wellcome Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

Email: ansari@well.ox.ac.uk

Key words

Hepatitis B virus, HBV, genotype, subgenotype, phylogeny, classification, reference, sequence

ABSTRACT

Hepatitis B virus (HBV) is a diverse, partially double-stranded DNA virus, with 9 genotypes (A-I), and a putative 10th genotype (J), thus far characterised. Given the broadening interest in HBV sequencing, there is an increasing requirement for a consistent, unified approach to HBV genotype and subgenotype classification. We set out to generate an updated resource of reference sequences using the diversity of all genomic-length HBV sequences available in public databases. We collated and aligned genomic-length HBV sequences from public databases and used maximum-likelihood phylogenetic analysis to identify genotype clusters. Within each genotype, we examined the phylogenetic support for currently defined subgenotypes, as well as identifying well-supported clades and deriving reference sequences for them. An alignment of these reference sequences and maximum-likelihood phylogenetic trees of the sequences are provided to simplify classification. Based on the phylogenies generated, we present a comprehensive set of HBV reference sequences at the genotype and subgenotype level.

INTRODUCTION

Hepatitis B virus (HBV) is the prototype virus of the *hepadnaviridae* family, an unusual family of partially double-stranded (ds)DNA viruses approximately 3.2kb in length encoding four genes, including a reverse transcriptase, encoded in overlapping reading frames on a circular genome (1,2). To date, 9 genotypes of HBV have been characterised (A-I), along with a putative 10th genotype (J) (3). The viruses display a relatively high amount of variation for a dsDNA virus (4), with the viral genotypes being further subdivided into upwards of 30 subgenotypes, many of which have distinct geographic and clinical associations (2).

At present, HBV genotyping is not routinely being performed in most clinical settings, as it has not been widely considered as relevant to patient management. However, as more treatment data become available and improvements are made in patient-stratified clinical care, guidelines may change to reflect different genotype-specific recommendations. This has been exemplified by the management of hepatitis C virus infection (5), and a similar approach for HBV is starting to emerge, with recent EASL guidelines for HBV treatment suggesting different stopping points for treatment non-response in genotypes A-D (6). Evidence increasingly supports a role for HBV genotype in influencing disease progression including risk of developing chronic infection, e-antigen seroconversion, transmission mode, and the development of hepatocellular carcinoma (7,8). Studies often refer to 'wild-type' virus (8,9), but 'wild-type' for one genotype may not reflect consensus for other genotypes (10,11). Understanding the diverse range of HBV strains circulating globally, and their associations with disease will allow us to move towards a more specific and precise approach to analysis.

A consistent approach to HBV subtyping is increasingly relevant, as interest in studying the associations between disease outcomes and viral sequence expands, and deep sequencing is used to investigate viral diversity in increasingly larger cohorts. A progressive body of sequencing data is emerging, and this expansion of available data is likely to increase over time (Suppl. Fig 1); however, the number of sequences covering the full genome lag a long way behind sequences for single genes or shorter fragments. To date, numerous subtyping misclassifications have been documented for HBV, predominantly driven by the use of partial genome sequences rather than full-length genomes (9,10), the inappropriate classification of recombinant strains (11), and publications re-designating same subgenotype classifications to distinctly different strains (12,13). Whilst a number of HBV sequence and analysis resources already exist, including HBVdb (14), HBVRegDB (15), HBVDR (16) and geno2pheno (https://hbv.geno2pheno.org/), each defines different sets of reference

sequences, and frequently only at the genotype level. Despite proposed criteria for defining HBV genotypes and subgenotypes (3), currently there is no single resource where all the necessary sequences and their designation is readily available for use.

With this dataset, we have set out to provide a unified resource, taking a pragmatic approach to describing the phylogeny of HBV genotypes A-I using all available genotyped sequences from the HBVdb repository (14). We removed highly similar sequences from downloaded alignments and performed maximum-likelihood phylogenetic analysis. We used the phylogeny of whole genome sequences to identify well-supported clades and selected representative reference sequences for each identifiable subgenotype.

RESULTS

Generation of a dataset of genomic-length HBV sequences

We downloaded a total of 7108 full length HBV sequences from the hepatitis B virus database (HBVdb, https://hbvdb.ibcp.fr/HBVdb/) on 31st of January 2019. Following clean-up of the dataset to remove identical sequences and recombinants (as described in detail in the methods section) we were left with a total of 5972 sequences. We aligned the sequences using MAFFT and calculated pairwise distances for hierarchical clustering, which was then used to remove highly similar isolates (see methods). At this stage we were left with 2839 sequences on which we based our phylogenetic analysis (all sequences shown in Fig. 1).

Classification of HBV genotypes, subgenotypes and designation of references

We selected appropriate reference sequences to represent each genotype and subgenotype groups; these sequences are shown in the context of the overall phylogeny highlighted in Figures 2-9 and listed in Table 1. In cases where we could identify multiple distinct clades within a single subgenotype (subgenotypes A1, B2, C2, D1; Figs 2-5), we selected a reference for each clade to ensure the diversity within the subgenotype was represented. In genotypes A-D, the most well-represented subgenotypes contained ≥45% of all the sequences within that genotype. Sequences were most evenly distributed between subgenotypes in genotype F, with the commonest subgenotype accounting for 33% of sequences.

Ambiguities and inconsistencies in existing data

There are a number of poorly characterised and disputed subgenotypes in the literature, particularly within genotype C, making accurate identification of the subgenotypes challenging. Strong evidence to support a number of previously described clades was also

difficult to ascertain, with specific difficulties in assessment of certain subgenotypes; D6, C7 and C9 were not easily distinguishable in this analysis. A search of Genbank indicated a single genomic-length sequence of subgenotypes A4 (KM606737), A6 (GQ331046), and D6 (KF170740) in the database, and no evidence of genomic-length sequences for subgenotype C7 or C9 isolates. Both A4 and A6 isolates clustered away from other sequences in the analysis, but few additional isolates of the subgenotypes were retained after removing highly similar sequences. In the case of subgenotype C7, previous publications on the phylogeny of genotype C used EU670263 as a C7 reference strain (17,18). However, in Genbank, EU670263 is listed as subgenotype C6 (19), and clusters within our phylogenetic analysis with GU721029, which is also designated subgenotype C6. A single provisional isolate of subgenotype C9 exists (AP011108), having been proposed by Mulyanto et al. in 2010 (17) but this designation is not present in the Genbank data associated with the sequence. Likewise, subgenotypes C13-C16 have been described (20), but we were unable to distinguish these as distinct subgenotypes in our analysis. The D6 isolate KF170740 was not retained in the sequences we selected for analysis, suggesting it is closely related to another genotype D subgenotype.

Comparative phylogenies and pairwise distance of HBV genotypes

We aligned our newly defined genotype and subgenotype reference sequences, and used these to generate a ML phylogenetic tree (Fig 10). Pairwise distance analysis for the majority of genotypes (Fig 11) revealed a bi-modal distribution of the distances, with one peak representing the relationship to sequences from the same (or closely related) clades (typically showing 1-3% divergence) and the other peak being characteristic of distance to sequences in other clades from other subgenotypes (typically 3-8% divergence). Genotype D differed from this distribution with a single peak where the majority of pairwise distances were between 2-4%. This indicates that genotype D phylogeny does not contain long evolutionary branches separating clades from each other. Genotypes E, F, H and I contained smaller number of sequences, which can impact the pairwise distances distribution. In this group genotypes E, H and I showed a uni-modal distribution of the pairwise distances. The lowest pairwise divergence was observed in genotype H, where the majority of pairwise distances were <3%. Genotype F demonstrated the greatest between-sequence diversity, where a high proportion of pairwise distances were >5%, likely a reflection of the long evolutionary distance between subgenotypes. Genotype G had only three sequences after our initial filtering which was too small for a meaningful pairwise distance analysis.

DISCUSSION

Summary of this resource

We have collated genomic-length HBV reference sequences into a repository, generating a resource that will help to underpin research. As clinical practice evolves to incorporate recommendations pertaining to HBV genotype, our dataset will also potentially become useful to support sequence-based insights for clinical practice.

Insights into diversity and geography

Genotype C has a particularly extensive phylogeny relative to the other genotypes. Genotype C is prevalent in the Asia-Pacific region and is thought to be the oldest extant genotype (21). The expansive diversity within this genotype is likely a consequence of its long evolutionary history in humans, although it should be noted that this genotype is also the most well represented genotype in the analysis, accounting for over 1/3 of all sequences in our dataset. Considerable diversity is also observed with genotypes B and D. Genotype D has a relatively broad geographic prevalence, being found in regions including the Mediterranean, North-Eastern Europe, India, Oceania and parts of southern Africa (22). It is possible that persistence of chronic HBV infection in these varied ethnic groups has contributed to the diversification of this genotype.

Genotype F has an atypical phylogeny compared to other genotypes, comprising four subgenotypes separated by long evolutionary distances, but with little intra-subgenotype diversity. It is unclear if the unusual phylogeny is a result of the elevated substitution rate documented for genotype F (23) or if there is a paucity of sampling relative to the other genotypes. The genotype is distributed throughout Latin America and the Arctic Circle (24,25), regions that are under-represented by clinical and research cohorts to date. Genotypes E, G and H, each of which has a single subgenotype, have relatively constrained phylogenies, showing low levels of genetic diversity in the currently sampled isolates. Genotype G has a particularly low evolutionary rate, likely resulting from the 36bp insertion in the core gene hampering viral replication (26). The genotype is typically observed in co-infection with either HIV or HBV genotype A2 (26).

A HBV strain isolated from a 16th century skeleton (LT992440) clustered between subgenotypes in genotype A. The isolate falls (with long branch lengths) together with two sequences both isolated from patients in Belgium, one of which has been proposed as the

hypothetical subgenotype A6 (GQ331046), isolated in 2006 (27). The clustering of these three sequences highlights the protracted association of HBV with human populations, and the marked lack of temporal population structure displayed by the viruses, frequently confounding phylogenetic attempts to understand the evolutionary history of HBV (28,29).

Limitations

(i) Errors and inconsistencies in existing classification system

There are a number of poorly characterised and disputed HBV subgenotypes in the literature, particularly within genotype C. A cause of these inconsistencies is the assigning of 'provisional' subgenotypes, based on a single isolate, as is the case with C7 and C9. Using the published literature to inform our analysis, we took an indication of strong bootstrap support, and a threshold of ≥2 sequences to define and identify the subgenotype clades. For some subgenotypes, including A4 and A6, multiple isolates have been reported in the literature but few isolates were retained in our analysis after removing highly similar sequences, suggesting divergence within the subgenotype is limited. Despite this approach, a number of clades grouped distinctly away from known subgenotypes but could not be categorically assigned to a subgenotype (thus we designated two 'unidentified clades' in genotype C).

We suggest that, in future, a minimum number of epidemiologically unlinked sequences and genomic coverage (e.g. ≥2 genomic-length unlinked sequences) should be required to designate a new HBV subgenotype. Bias introduced into databases from intra-host diversity studies, where multiple sequences are isolated from the same patient (such as clonal analysis) can confound such analyses, but should have been largely controlled for in this case by stripping the alignments of similar sequences.

The phylogenetic methods selected may affect the subgenotypes classified, as there may be variability in bootstrap support between methods. Both neighbour-joining and parsimony methods are considered less accurate than maximum-likelihood approaches (30,31), as used in this study. Future analysis to define HBV subgenotypes should be based on maximum-likelihood methods and the bootstrap support should be indicated.

(ii) Bias in published data

There is considerable variation in the numbers of sequences available for each genotype, with genotypes B and C particularly well represented and genotypes E-I relatively neglected. Whilst these differences may reflect genuine variation in the contribution of different genotypes to the total global pool of infection - with B and C coming from high-prevalence,

densely populated regions - significant under-sampling of HBV sequences from other regions is likely (especially from low-income regions). Similar trends in isolate bias have been reported in the HCV field (32). Consequently, the sequences available in publicly available databases are likely to under-represent the true extent of HBV diversity and, as more sequences are generated reference sets will need to be reviewed.

Several previous studies publishing HBV reference sequence sets have generated consensus sequences from alignments of full genome HBV sequences and submitted these new consensus sequences to Genbank (33,34), which leads to the potential inference that these are sequences of biological origin. A number of genotype B, D and F sequences generated with the same approach have also been submitted to Genbank but they are not supported by an associated publication, making it impossible to assess their provenance. The methods used to generate these sequences mean that, by definition, they occupy a basal position in the subgenotype clades after phylogenetic analysis and appear to be ideal reference sequences. However, these consensus sequences may differ from biologically derived sequences at key sites, and are therefore potentially misleading if used as reference sequences. Improved sequence metadata in the records of published sequences, clearly indicating the methods used to derive these sequences would be informative for researchers.

(iii) Recombination and 'quasi-subgenotypes'

Recombination is relatively common amongst HBV isolates (11,35) and a number of established HBV subgenotypes have been shown to have a recombinant origin, particularly in genotypes A, C, D and the putative genotype J (35–37). A number of sequences in our phylogenetic analysis were not easily classified, as they fall phylogenetically between two well-defined subgenotype clades. These sequences are possibly inter-subgenotype recombinants, combining genetic regions from two or more subgenotypes, making them challenging to classify in genomic-length genome phylogenies.

Previous work has indicated that some designated subgenotypes, including A3, A4, A5, C2 and B3 group as distinct, monophyletic clades in maximum-likelihood analysis but that they do not meet the required genetic distance to be classified as separate subgenotypes (3,9,38). It has been suggested these should instead be designated 'quasi-subgenotypes' to reflect their divergent nature but underscore that they do not meet the technical definition of a subgenotype (9). In our analysis, quasi-subgenotype A3 grouped into A3, A4 and A5 clades with strong bootstrap support. It was not possible to separate subgenotypes C2 and quasi-B3 (encompasses B3 and B5) into distinct subgenotypes.

Conclusion

Sequencing is increasingly utilised in the clinical management of HBV infection to inform on prognosis, treatment choice, drug resistance and to provide insight into complex cases (39,40). Whilst progress in applying new sequencing technologies to HBV has been slow to date, we are now entering an era of rapid change. A consistent, unified approach to classification will advance this field, improving consistency of curating, archiving and reporting sequence data. Recent work sampling poorly accessed populations has uncovered numerous new HCV isolates (41,42), and the diversity of sequences generated for HBV may expand in a similar way. The use of a robust and consistent classification system will also be important in screening new populations for potentially novel HBV isolates. All of the generated data, including the alignments, phylogenies, consensus sequences and chosen reference sequences are available online as a simple and open-access resource (https://figshare.com/account/projects/29779/articles/8851946). We have generated a new data resource for researchers and clinicians in the HBV field, providing a solid foundation for analysis, and a structure on which to build as more data emerge.

METHODS

Definitions

We based our analysis on previously agreed definitions regarding classification of HBV sequences on the basis of their nucleotide sequence diversity (3), as follows:

- ***Genotype:*** nucleotide divergence of >7.5% has been proposed as the threshold for the definition of distinct HBV genotypes (43).
- ***Subgenotype:*** Genotypes are further classified into subgenotypes, based on a divergence of 4-7.5%, monophyletic clustering, and strong bootstrap support for the clade (3).

In this study, we used strong bootstrap support (≥70) and monophyletic clustering to identify distinct clades and examined the pre-existing literature for the likely subgenotype designation. In instances when subgenotypes can be seen to split into distinct clades but genetic distance and poor bootstrap support do not suggest a unique subgenotype, such as in the case of subgenotypes A1 (Fig 2) and C2 (Fig 4), we have referred to these as ‘A1 (1)’ and ‘A1 (2)’.

HBV genome numbering convention

As HBV is a circular virus, there is technically no ‘start’ or ‘end’ to the genome. We have followed convention in the field (and HBVdb) for this study, defining nucleotide (nt)1 at an *EcoR1* restriction site in the Pol/Surface overlap (GAA/TTC, with nt1 starting at the first T) (2). The *EcoR1* site is hypothetical in many HBV sequences, with GAA/CTC being more frequent in many genotypes.

Sources of sequence data

We downloaded all available genomic-length genotyped HBV sequences (N=7108) for genotypes A-H from HBVdb (14) on 31st of January 2019, and removed sequences indicated as recombinant by the database (N=767), giving a total of 6341 sequences. We used HBVdb as the database runs sequences downloaded from Genbank through a genotyping algorithm prior to inclusion and all listed sequences are annotated to ensure they follow the HBV genome numbering (14), as outlined above.

We enhanced our dataset by excluding two sequences from the genotype G phylogeny, as they had been incorrectly genotyped with the HBVdb algorithm. Both sequences (FJ023674 and FR714503) were found to be partial genotype C recombinants. We also identified an additional 71 sequences belonging to genotype I and to genotype C and D subgenotypes (under-represented in the HBVdb database) on Genbank and added these to our data set, giving a total of 6412 sequences. Ancient HBV sequences, isolated from bodies hundreds of years old (28,44), are present in these databases and they were not excluded from this analysis, unless they grouped distantly from a known genotype.

Existing reference sequences

We included the genotype A strain X02763 (45) in the alignment with other reference sequences as this sequence (length 3221bp) is widely used as a numbering reference. Only Geno-G has insertions relative to genotype A. We suggest that the current NCBI HBV Reference Sequence, NC_003977.2 (46), a genotype D isolate, is less suitable as a numbering reference as genotype D strains have a 33-bp deletion in pre-S1, making them the shortest length HBV genotype at 3182bp. Nonetheless, we also included NC_003977.2 in the reference alignment as it is frequently used as a reference.

Phylogenetic analysis

We aligned all sequences using MAFFT (47) and calculated pairwise distances between all isolates. In order to reduce the possibility that multiple sequences from a single individual or closely related cluster could dominate the phylogeny, we kept only one sequence from each group of identical sequences, and used hierarchical clustering to group isolates based on

their pairwise distance and identified clusters of sequences that were within ≤1% of each other, retaining a single sequence from each cluster for the analysis (retaining the sequence with the lowest proportion of ambiguous sites). After stripping out these closely related sequences, we generated a maximum-likelihood (ML) phylogenetic tree of the remaining 2839 sequences using RAxML (48), including 1000 bootstrap replicates (Fig. 1). We used this tree to manually delineate sequences belonging to each genotype and excluded any remaining recombinant sequences that had not been identified earlier in the process.

We aligned the sequences for each genotype separately, and generated ML phylogenetic trees as previously described. We examined ML phylogenetic trees and generated bootstrap values to define distinct clades containing ≥2 sequences with strong phylogenetic support (Figures 2-9). Phylogenetic trees were examined in both traditional and radial layouts to ensure that the subgenotypes and clades were being delineated accurately. We excluded Genotype G from the phylogenetic analysis, as after removing highly similar sequences there were just three sequences remaining.

Definition of new reference sequences

For each well-defined clade, we generated a consensus sequence and selected the closest clinical isolate to the consensus sequence as the reference strain for that clade. We applied a process of quality assurance to selection of reference strains as follows:

- We verified that isolates selected as reference strains were associated with an associated publication indexed in PubMed;
- We checked that reference strains were genuine clinical isolates (rather than *in silico* reconstructions such as consensus sequences);
- We rejected isolates that had insertions or deletions relative to the majority of sequences in the clade;
- We avoided (where possible) sequences with many ambiguous sites;
- We avoided use of sequences containing known drug resistance mutations as reference strains. Thus in the situation in which a sequence containing an rtM204 substitution in the 'YMDD' motif (associated with escape from lamivudine therapy(49)) was initially selected as a reference sequence, we rejected this and selected the next closest sequence to consensus.
- Three subgenotype references were selected with Hamming distances >50 relative to the consensus sequence for the subgenotype clade (subgenotypes C4, C11 and C_UA2). Alternative, more closely related sequences were often present within the

subgenotype clade but indels in the sequences and a lack of supportive publications meant that other sequences were selected.

To define genotype reference sequences, we selected the reference sequence for the subgenotype that contributed the largest number of sequences to that genotype (grey boxes, Table 1). In cases where there were multiple clades for the subgenotype, we selected the sequence from the largest clade as the reference. Where there was strong evidence to support specific subgenotypes from both the literature and the phylogenetic analysis, we included the subgenotype in the analysis. Where subgenotype designation was unclear, we used BLAST to compare the sequence(s) to other sequences in online databases to see if the subgenotype could be identified. We also examined trees to find the accession numbers of subgenotypes that did not have an easily identifiable clade in the analysis using isolates listed in Genbank. If they were not present, these subgenotypes were most likely removed during the pairwise analysis for being highly similar to sequences already in the analysis, suggesting further examination of their designation as a unique subgenotype may be warranted. Where the subgenotype could not be identified, the clade was labelled 'unassigned'.

In situations in which we could not ascertain which subgenotype a clade belongs to, despite examination of Genbank records and BLAST analysis of sequences, we have described these as 'unassigned clades' (both in genotype C, Fig 4).

REFERENCES

1. Locarnini S, Zoulim F. Molecular genetics of HBV infection. Antiviral Therapy. 2010;15:3–14.
2. McNaughton AL, D'Arienzo V, Ansari MA, Lumley SF, Littlejohn M, Revill P, et al. Insights From Deep Sequencing of the HBV Genome—Unique, Tiny, and Misunderstood. Gastroenterology. 2019;156(2):384–99.
3. Kramvis A. Genotypes and genetic variability of hepatitis B virus. Intervirology. 2014;57(3–4):141–50.
4. Torres-Cornejo A, Lauer GM. Hurdles to the Development of Effective HBV Immunotherapies and HCV Vaccines. Pathogens and Immunity. 2017;2(1):102.
5. Pawlotsky JM, Negro F, Aghemo A, Berenguer M, Dalgard O, Dusheiko G, et al. EASL Recommendations on Treatment of Hepatitis C 2018. Journal of Hepatology. 2018;69(2):461–511.

6. European Association for Study of Liver. EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection. Journal of Hepatology. 2017;67(2):370–98.
7. Lin C-L, Kao J-H. The clinical implications of hepatitis B virus genotype: Recent advances. Journal of Gastroenterology and Hepatology. 2011;26:123–30.
8. Rajoriya N, Combet C, Zoulim F, Janssen HLA. How viral genetic variants and genotypes influence disease and treatment outcome of chronic hepatitis B. Time for an individualized approach? Journal of Hepatology. 2017;67(6):1281–97.
9. Pourkarim MR, Amini-Bavil-Olyaee S, Lemey P, Maes P, Van Ranst M. Are hepatitis B virus “subgenotypes” defined accurately? Journal of Clinical Virology. 2010;47(4):356–60.
10. Jia L, Hu F, Li H, Li L, Tang X, Liu Y, et al. Characterization of small genomic regions of the hepatitis B virus should be performed with more caution. 2018;24–6.
11. Shi W, Zhang Z, Ling C, Zheng W, Zhu C, Carr MJ, et al. Hepatitis B virus subgenotyping: History, effects of recombination, misclassifications, and corrections. Infection, Genetics and Evolution. 2013;16:355–61.
12. Schaefer S, Magnius L, Norder H. Under construction: Classification of hepatitis B virus genotypes and subgenotypes. Intervirology. 2009;52(6):323–5.
13. Ahn SH, Yuen L, Revill P. Clarification required for the definition of hepatitis B virus subgenotypes C1 and C2. Intervirology. 2009;52(6):321–2.
14. Hayer J, Jadeau F, Deléage G, Kay A, Zoulim F, Combet C. HBVdb: A knowledge database for Hepatitis B Virus. Nucleic Acids Research. 2013;41(D1):566–70.
15. Panjaworayan N, Roessner SK, Firth AE, Brown CM. HBVRegDB: Annotation, comparison, detection and visualization of regulatory elements in hepatitis B virus sequences. Virology Journal. 2007;4:1–10.
16. Bell TG, Yousif M, Kramvis A. Bioinformatic curation and alignment of genotyped hepatitis B virus (HBV) sequence data from the GenBank public database. SpringerPlus. 2016;5(1).
17. Sulaiman M, Depamede N, Surayah K. Identification and characterization of novel hepatitis B virus subgenotype C10 in Nusa Tenggara , Indonesia. 2010;705–15.
18. Mulyanto, Depamede SN, Wahyono A, Jitintai, Nagashima S, Takahashi M, et al. Analysis of the Full-Length Genomes of Novel Hepatitis B Virus Subgenotypes C11 and C12 in Papua , Indonesia. Journal of Medical Virology. 2011;64(July 2010):54–64.
19. Cavinta L, Sun J, May A, Yin J, Meltzer M Von, Radtke M, et al. A New Isolate of Hepatitis B Virus From the Philippines Possibly Representing a New Subgenotype C6. 2009;987:983–7.

20. Mulyanto, Pancawardani P, Depamede SN, Wahyono A, Jirintai S, Nagashima S, et al. Identification of four novel subgenotypes (C13-C16) and two inter-genotypic recombinants (C12/G and C13/B3) of hepatitis B virus in Papua province, Indonesia. Virus research. 2012 Jan;163(1):129–40.
21. Paraskevis D, Magiorkinis G, Magiorkinis E, Ho SYW, Belshaw R, Allain JP, et al. Dating the origin and dispersal of hepatitis B virus infection in humans and primates. Hepatology. 2013;57(3):908–16.
22. Zehender G, Ebranati E, Gabanelli E, Shkjezi R, Lai A, Sorrentino C, et al. Spatial and temporal dynamics of hepatitis B virus D genotype in Europe and the Mediterranean basin. PLoS ONE. 2012;7(5):1–8.
23. Paraskevis D, Angelis K, Magiorkinis G, Kostaki E, Ho SYW, Hatzakis A. Dating the origin of hepatitis B virus reveals higher substitution rate and adaptation on the branch leading to F/H genotypes. Molecular Phylogenetics and Evolution. 2015;93:44–54.
24. Livingston SE, Simonetti JP, McMahon BJ, Bulkow LR, Hurlburt KJ, Homan CE, et al. Hepatitis B virus genotypes in Alaska Native people with hepatocellular carcinoma: preponderance of genotype F. The Journal of infectious diseases. 2007;195:5–11.
25. Marciano S, Galdame OA, Gadano AC. HBV genotype F: Natural history and treatment. Antiviral Therapy. 2013;18(3 PARTB):485–8.
26. van der Kuyl AC, Zorgdrager F, Hogema B, Bakker M, Jurriaans S, Back NK, et al. High prevalence of hepatitis B virus dual infection with genotypes A and G in HIV-1 infected men in Amsterdam, the Netherlands, during 2000-2011. BMC Infectious Diseases. 2013;13(1):540.
27. Pourkarim MR, Lemey P, Amini-Bavil-Olyaee S, Maes P, Van Ranst M. Novel hepatitis B virus subgenotype A6 in African-Belgian patients. Journal of Clinical Virology. 2010;47(1):93–6.
28. Ross ZP, Klunk J, Fornaciari G, Giuffra V, Duchêne S, Duggan AT, et al. The paradox of HBV evolution as revealed from a 16th century mummy. PLOS Pathogens. 2018;14(2):e1006750.
29. Mühlemann B, Jones TC, Damgaard PDB, Allentoft ME, Shevnina I, Logvin A, et al. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. Nature. 2018;557(7705):418–23.
30. Hraber PT, Fischer W, Bruno WJ, Leitner T, Kuiken C. Comparative analysis of hepatitis C virus phylogenies from coding and non-coding regions: the 5’ untranslated region (UTR) fails to classify subtypes. Virology journal. 2006;3:103.
31. Ogden TH, Rosenberg MS. Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology. 2006;55(2):314–28.
32. Niebel M, Singer JB, Nickbakhsh S, Gifford RJ, Thomson EC. Hepatitis C and the

absence of genomic data in low-income countries: a barrier on the road to elimination? The Lancet Gastroenterology and Hepatology. 2017;2(10):700–1.

33. Cai Q, Zhu H, Zhang Y, Li X, Zhang Z. Hepatitis B virus genotype A: design of reference sequences for sub-genotypes. Virus Genes. 2016;52(3):325–33.

34. Zhang ZH, Zhang L, Lu MJ, Yang DL, Li X. Establishment of reference sequences of hepatitis B virus genotype B and C in China. Genetics and Molecular Research. 2015;14(4):16521–34.

35. Castelhano N, Araujo NM, Arenas M. Heterogeneous recombination among Hepatitis B virus genotypes. Infection, Genetics and Evolution. 2017;54(May):486–90.

36. Simmonds P, Midgley S. Recombination in the genesis and evolution of hepatitis B virus genotypes. Journal of virology. 2005;79(24):15467–76.

37. Tatematsu K, Tanaka Y, Kurbanov F, Sugauchi F, Mano S, Maeshiro T, et al. A Genetic Variant of Hepatitis B Virus Divergent from Known Human and Ape Genotypes Isolated from a Japanese Patient and Provisionally Assigned to New Genotype J. Journal of Virology. 2009;83(20):10538–47.

38. Shi W, Zhu C, Zheng W, Zheng W, Ling C, Carr MJ, et al. Subgenotyping of Genotype C Hepatitis B Virus: Correcting Misclassifications and Identifying a Novel Subgenotype. PLoS ONE. 2012;7(10):doi: 10.1371/journal.pone.0047271.

39. del Campo JA, Parra-Sánchez M, Figueruela B, García-Rey S, Quer J, Gregori J, et al. HCV deep-sequencing for sub-genotypes identification of mixed infections: a real life experience. International Journal of Infectious Diseases. 2017;67:114–7.

40. Fokam J, Bellocchi MC, Armenia D, Nanfack AJ, Carioti L, Continenza F, et al. Next-generation sequencing provides an added value in determining drug resistance and viral tropism in Cameroonian HIV-1 vertically infected children. Medicine. 2018;97(13).

41. Davis C, Mgomella GS, Filipe A da S, Frost EH, Giroux G, Hughes J, et al. New highly diverse hepatitis C strains detected in sub-Saharan Africa have unknown susceptibility to direct-acting antiviral treatments. Hepatology. 2018;0–3.

42. Borgia SM, Hedskog C, Parhy B, Hyland RH, Stamm LM, Brainard DM, et al. Identification of a Novel Hepatitis C Virus Genotype From Punjab, India: Expanding Classification of Hepatitis C Virus Into 8 Genotypes. The Journal of Infectious Diseases. 2018;(July):1–8.

43. Kramvis A, Arakawa K, Yu MC, Nogueira R, Stram DO, Kew MC. Relationship of Serological Subtype, Basic Core Promoter and Precore Mutations to Genotypes/Subgenotypes of Hepatitis B Virus. Journal of Medical Virology. 2008;80:27–46.

44. Mühlemann B, Jones TC, De Barros Damgaard P, Allentoft ME, Shevnina I, Logvin A, et al. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. Nature.

2018;557(7705):418–23.

45. Valenzuela P, Quiroga M, Zaldivar J, Gray P, Rutter WJ. The Nucleotide Sequence of the Hepatitis B Viral Genome and the Identification of the Major Viral Genes. Animal Virus Genetics. ACADEMIC PRESS, INC.; 1980. 57–70 p.

46. Galibert F, Mandart E, Fitoussi F, Tiollais P, Charnay P. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in E. coli. Nature. 1979 Oct;281(5733):646–50.

47. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics. 2017;doi: 10.1093/bib/bbx108.

48. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

49. Hatakeyama T, Noguchi C, Hiraga N, Mori N, Tsuge M, Imamura M, et al. Serum HBV RNA is a predictor of early emergence of the YMDD mutant in patients treated with lamivudine. Hepatology. 2007;45(5):1179–86.

50. Shi W, Zhu C, Zheng W, Carr MJ, Higgins DG, Zhang Z. Subgenotype reclassification of genotype B hepatitis B virus. BMC Gastroenterology. 2012;12:2–7.

51. Yousif M, Kramvis A. Genotype D of hepatitis B virus and its subgenotypes: An update. Hepatology Research. 2013;43(4):355–64.

FIGURE LEGENDS

Figure 1; Genomic-length maximum-likelihood phylogeny of all full length genotype A-I HBV sequences included in analysis (n=2839), indicating the number of sequences in each genotype analysed separately in Figures 2-9. Maximum-likelihood phylogenetic analysis of sequences used in this study after removing highly similar sequences, demonstrating the strong genotypic clustering of the sequences. Bootstrap support ≥70 after 1000 replicates is given for the deepest branches on the tree. The scale bar indicates the estimated nucleotide substitutions per site. *A strain known to be from a 14th century skeleton clustering distantly with genotype A, LT992441, was removed from the analysis. ** KU736915 was identified as a genotype D/E recombinant and removed from the subsequent analysis.

Figure 2; Genomic-length maximum-likelihood phylogeny of HBV genotype A sequences (n=259). Well-defined clades have been highlighted with coloured dotted lines and reference sequences for each clade indicated (red dots). Proposed reference strain for the genotype, FJ692557 is highlighted with a blue dot. The subgenotype is given where it could be reliably identified. Bootstrap support for branches ≥70 after 1000 replicates is indicated. The scale bar indicates the estimated nucleotide substitutions per site. Previous work has confirmed that there are at least 5 genotype A subgenotypes, although debate continues about whether or not A3, A4 and A5 should all be considered 'quasi-subgenotype A3' (9). Few sequences for subgenotypes *A4 (KM606737) and **A6 (GQ331046) were retained in the study after pairwise analysis. The putative subgenotype A6 has previously been identified in three African-Belgian patients (27).

Figure 3; Genomic-length maximum-likelihood phylogeny of HBV genotype B sequences. Well-defined clades have been highlighted with coloured dotted lines and reference sequences for each clade indicated (red dots). Proposed reference strain for the genotype, GU815637 is highlighted with a blue dot. The subgenotype is given where it could be reliably identified. Bootstrap support for branches ≥70 after 1000 replicates is indicated. The scale bar indicates the estimated nucleotide substitutions per site. An evaluation of the genotype B phylogeny reclassified a number of putative subgenotypes as quasi-B3, with debate continuing on whether or not this should also include B5 (50).

Figure 4; Genomic-length maximum-likelihood phylogeny of HBV genotype C sequences. Well-defined clades have been highlighted with coloured dotted lines and reference sequences for each clade indicated (red dots). Proposed reference strain for the

genotype, GQ377617 is highlighted with a blue dot. The subgenotype is given where it could be reliably identified. Bootstrap support for branches ≥70 after 1000 replicates is indicated. The scale bar indicates the estimated nucleotide substitutions per site. We were unable to verify the subgenotype of two genotype C clades, and these have been designated unassigned clade 1 and 2 (unassigned_C (1) and unassigned_C (2), respectively).

Figure 5; Genomic-length maximum-likelihood phylogeny of HBV genotype D sequences. Well-defined clades have been highlighted with coloured dotted lines and reference sequences for each clade indicated (red dots). Proposed reference strain for the genotype, KC875277is highlighted with a blue dot. The subgenotype is given where it could be reliably identified. Bootstrap support for branches ≥70 after 1000 replicates is indicated. The scale bar indicates the estimated nucleotide substitutions per site. Previous work has indicated D3 and D6 strains cluster together and should be classed as a single subgenotype (51).

Figure 6; Genomic-length maximum-likelihood phylogeny of HBV genotype E sequences. Genotype E sequences do not diverge into distinct subgenotypes. Bootstrap support for branches ≥70 after 1000 replicates is indicated. The scale bar indicates the estimated nucleotide substitutions per site. Proposed reference strain for the genotype, GQ161817is highlighted with a blue dot.

Figure 7; Genomic-length maximum-likelihood phylogeny of HBV genotype F sequences. Well-defined clades have been highlighted with coloured dotted lines and reference sequences for each clade indicated (red dots). Proposed reference strain for the genotype, HM585194 is highlighted with a blue dot. The subgenotype is given where it could be reliably identified. Bootstrap support ≥70 after 1000 replicates is given. The scale bar indicates the estimated nucleotide substitutions per site.

Figure 8; Genomic-length maximum-likelihood phylogeny of HBV genotype H sequences. Genotype H sequences do not diverge into distinct subgenotypes. Bootstrap support ≥70 after 1000 replicates is given. Proposed reference strain for the genotype, FJ356715 is highlighted with a blue dot. The scale bar indicates the estimated nucleotide substitutions per site.

Figure 9; Genomic-length maximum-likelihood phylogeny of HBV genotype I sequences. Well-defined clades have been highlighted with coloured dotted lines and reference sequences for each clade indicated (red dots). Proposed reference strain for the

genotype, AB562463is highlighted with a blue dot. Bootstrap support for branches ≥70 after 1000 replicates is indicated. The scale bar indicates the estimated nucleotide substitutions per site.

Figure 10; Genomic-length maximum-likelihood phylogenetic tree of HBV genotype, subgenotype and clade reference strains identified in Figures 2-9 and listed in Table 1. The genotype is given in each case and the subgenotype or clade identification is given where possible. Bootstrap support for branches ≥70 after 1000 replicates is indicated. The scale bar indicates the estimated nucleotide substitutions per site. In addition to the references identified in figures 2-9, genotype A isolate X02763 and genotype D isolate NC_003977.2 have been included in the tree.

Figure 11; Pairwise distances distribution for the genomic-length sequences of HBV genotypes A, B, C, D, E, F, H, I. Probability densities of pairwise distances for whole genome sequences of HBV genotypes. Genotypes E, F, H and I are shown on a separate plot from genotypes A-D as they contained smaller number of sequences. Too few sequences were available after filtering for genotypes G and (putative) genotype J to be analysed.

Declarations

Authors' contributions

AM and PM conceived the idea for this project. AM and MA identified and analysed the sequences in this study. AM, MA and PM wrote the manuscript and all authors read and approved the final manuscript.

Funding

PM is funded by the Wellcome Trust (Grant Number 110110).

Availability of data and material

The data described in this Data Note, including the genotype alignments (fasta format), consensus sequences and, phylogenetic trees (newick format) are available in an open access format on Figshare (https://figshare.com/account/projects/29779/articles/8851946).

Ethics approval and consent to participate

Not applicable

Acknowledgements

Not applicable

Table 1; Proposed reference sequences for HBV genotypes, subgenotypes and clades. The number of sequences in each clade is given for each subgenotype and clade identified. Note that the total sum of the subgenotype sequence clusters may not correspond to the total number of genotype sequences, as a number of sequences did not group within a specific clade. Reference sequences for the genotypes are highlighted in grey boxes. In Figures 2-9, genotype references are marked with blue dots and subgenotype reference sequences are marked with red dots. Subgenotypes A4, A6, B5, C7, C9 and D6 are not included as these sequences either did not cluster as monophyletic clades or were not retained in our analysis. Hamming distance indicates the number of nucleotide differences between the clade consensus and the chosen reference. The pairwise distance is the number of nucleotide differences between the clade consensus and the chosen reference normalised by length of the genome.

HBV genotype	Subgenotype	Number of sequences	Reference GenBank ID	Hamming distance	Pairwise distance	Journal link (DOI)	Collection year[a]	Country of origin
A	A1 (1)	50	KP168423	26	0.008	10.1186/s12879-016-1409-2	2012	Kenya
	A1 (2)	**87**	**FJ692557**	21	0.007	10.3201/eid1508.081642	2006	Haiti
	A2	80	EU594385	5	0.002	10.1002/jmv.20169	2004	Estonia
	A3	9	AM184126	29	0.009	10.1002/jmv.20678	2005	Gabon
	A4	2	KM606737	n/a	n/a	10.1371/journal.pone.0118959	2015	Cuba
	A5	25	FJ692601	10	0.003	10.3201/eid1508.081642	2006	Haiti
	A6	2	GQ331046	n/a	n/a	10.1016/j.jcv.2009.09.032.	2006	Belgium
B	B1	47	D23679	23	0.007	10.1002/jmv.1890440118	1993	Japan
	B2 (1)	131	JQ801514	15	0.005	10.1002/jmv.23430	2009	Thailand
	B2 (2)	**294**	**GU815637**	10	0.003	10.1128/JVI.02164-09	2010	Taiwan
	B3	106	AP011085	23	0.007	10.1007/s00705-009-0406-9	2001	Indonesia
	B4	69	AB073835	35	0.011	10.1128/jvi.76.12.5985-5992.2002	2001	Japan
	B6	36	AB287314	29	0.009	10.1086/523111	2006	Alaska
C	C1[b]	240	DQ089781	29	0.009	10.1086/430324	2005	China (Hong Kong)
	C2 (1)	261	KC774182	21	0.007	10.1186/1743-422X-11-101	2012	China
	C2 (2)	**280**	**GQ377617**	14	0.004	10.1016/j.jcv.2010.05.010	2007	China
	C2 (3)	157	AP011098	10	0.003	10.1007/s00705-009-0406-9	2009	Indonesia
	C4	21	KF873526	68	0.021	10.1002/jmv.23888	2011	Australia

	C5	15	AP011099	41	0.013	10.1007/s00705-009-0406-9	2009	Indonesia
	C6[c]	2	EU670263	28	0.009	10.1002/jmv.21475	2008	Philippines
	C8	15	AP011107	42	0.013	10.1007/s00705-009-0406-9	2009	Indonesia
	C10	21	KJ173333	33	0.010	10.1016/j.jcv.2014.06.010	2012	China
	C11[d]	28	AB554015	86	0.027	10.1002/jmv.21931	2010	Indonesia
	C_UA_1	31	KC774298	15	0.005	10.1186/1743-422X-11-101	2012	China
	C_UA_2	16	DQ089802_	55	0.018	10.1086/430324	2005	China (Hong Kong)
D	D1(1)	106	AB222711	11	0.003	10.1002/jmv.21035	2005	Uzbekistan
	D1(2)	**216**	**KC875277**	17	0.005	10.1371/journal.pone.0110012	2013	India
	D2	100	MF925358	21	0.007	10.1371/journal.pone.0188944	2015	Bangladesh
	D3	78	FJ692507	18	0.006	10.3201/eid1508.081642	2006	Haiti
	D4	15	FJ692533	17	0.008	10.3201/eid1508.081642	2006	Haiti
	D5	15	GQ205389	19	0.006	10.1128/JCM.01174-10	2008	India
	D7	15	FJ904435	44	0.014	10.1099/vir.0.009738-0	2006	Tunisia
E	**n/a**	**145**	**GQ161817**	19	0.006	10.1099/vir.0.012013-0	2006	Guinea
F	**F1**	**26**	**HM585194**	13	0.004	10.1002/jmv.22129	2007	Chile
	F2	13	DQ899143	26	0.008	10.1002/jmv.21024	2006	Venezuela
	F3	19	MH051986	17	0.005	10.1371/journal.pone.0197662	2011	Venezuela
	F4	18	KJ843175	17	0.005	10.1371/journal.pone.0121436	2012	Argentina
G	**n/a**	**3**	**AB056513**	n/a	n/a	10.1016/S0166-0934(01)00374-3	2001	United States of America
H	**n/a**	**11**	**FJ356715**	36	0.011	10.1128/JCM.01337-09	2008	Argentina
I	**I1**	**5**	**AB562463**	17	0.005	10.1002/jmv.21978	2007	Vietnam
	I2	4	FJ023669	17	0.005	10.3201/eid1411.080437	2008	Laos
J[e]	**n/a**	**1**	**AB486012**	n/a	n/a	10.1128/JVI.00462-09	2006	Japan/Borneo

n/a; not applicable, too few sequences identified belonging to the genotype/subgenotype to generate consensus sequences for selection of closest biological isolate.

[a] Collection date of sample or year submitted to Genbank (if collection date not given).

[b] C1 is a large clade that also contains sequences labelled as subgenotype C3, with no clear separation between the two sets of sequences.

[c] This sequence has been used previously as a subgenotype C7 reference in a number of publications (21,22). No other putative C7 sequences are proposed in the literature.

[d] C11, Large number of sequences were unpublished in this clade (>30 first closest seqs).

[e] Genotype J remains putative, with a single isolate identified in a Japanese patient. The isolate shows considerable divergence from other known HBV strains and is thought to be a recombinant of genotype C and a gibbon HBV isolate.

Figure 1; All sequences included in analysis

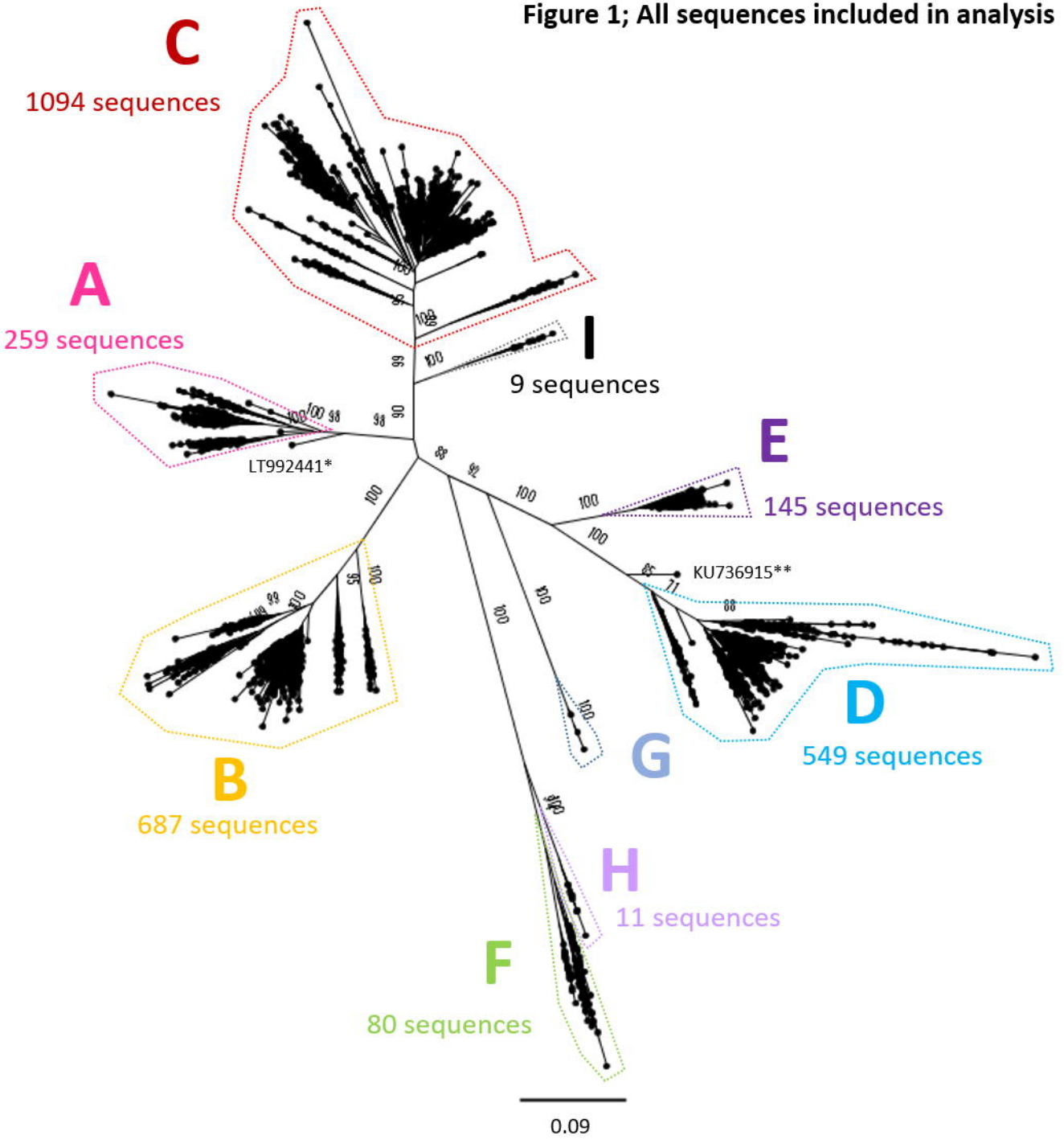


Figure 2; Genotype A

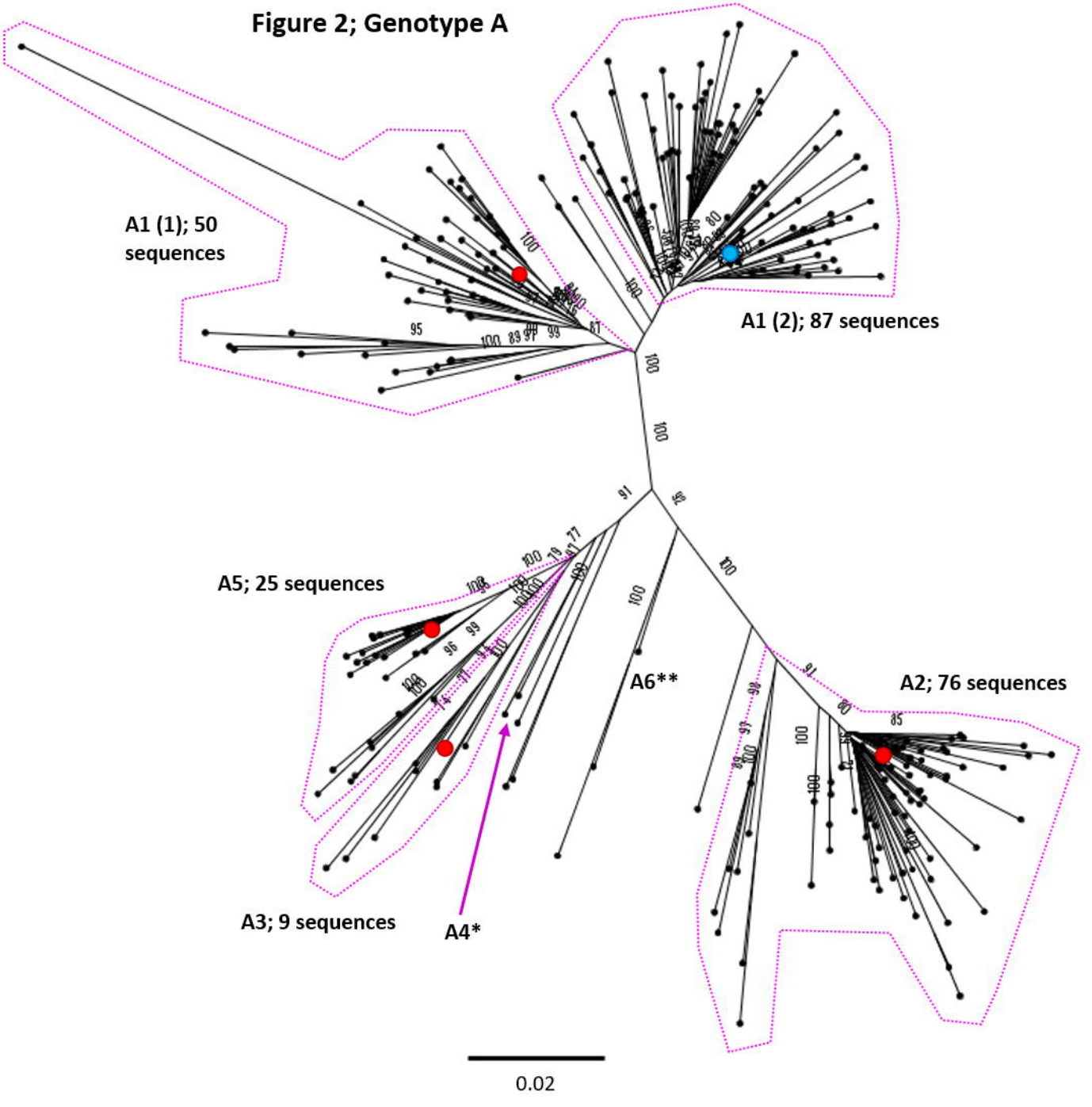


Figure 3; Genotype B

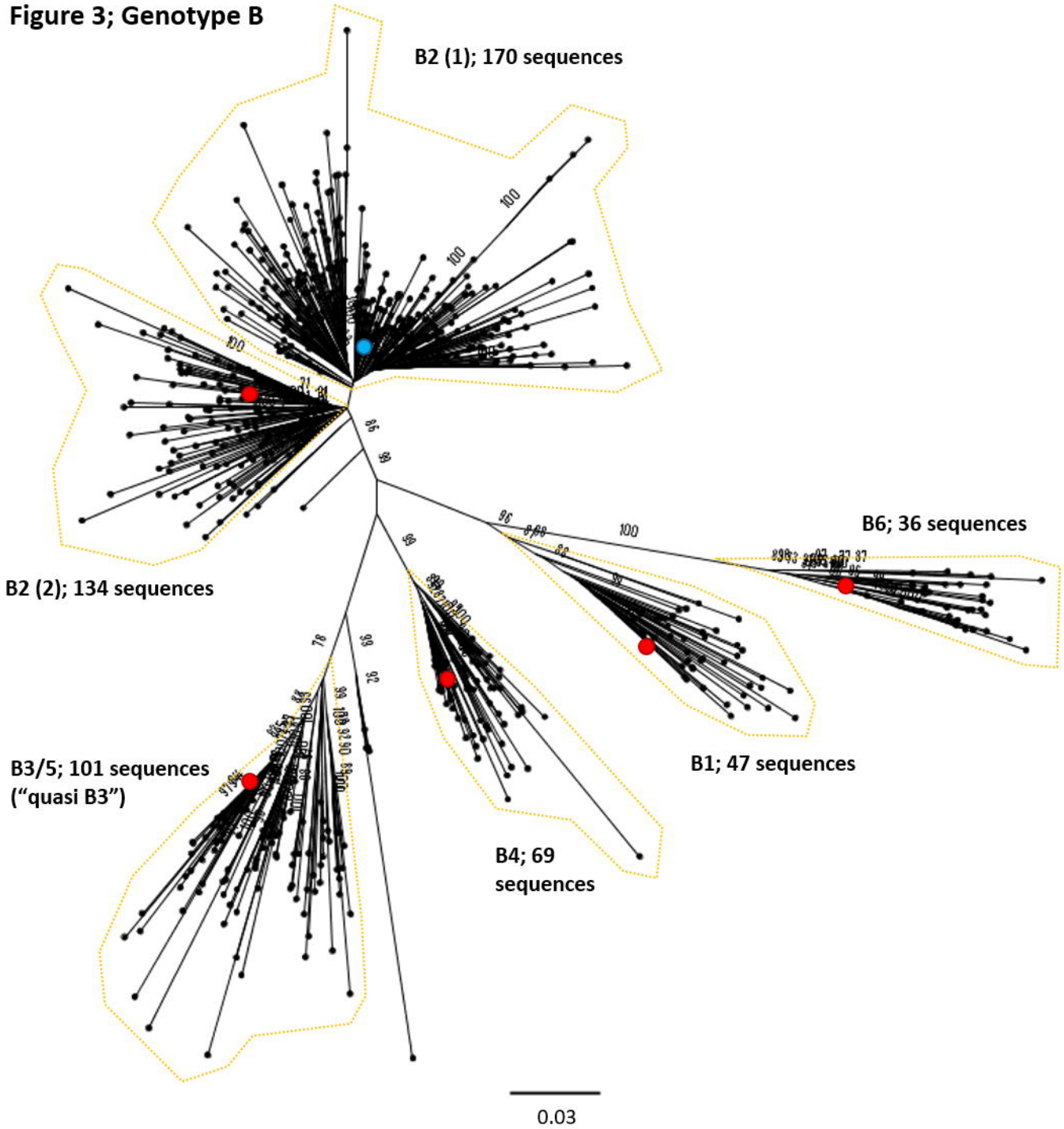


Figure 4; Genotype C

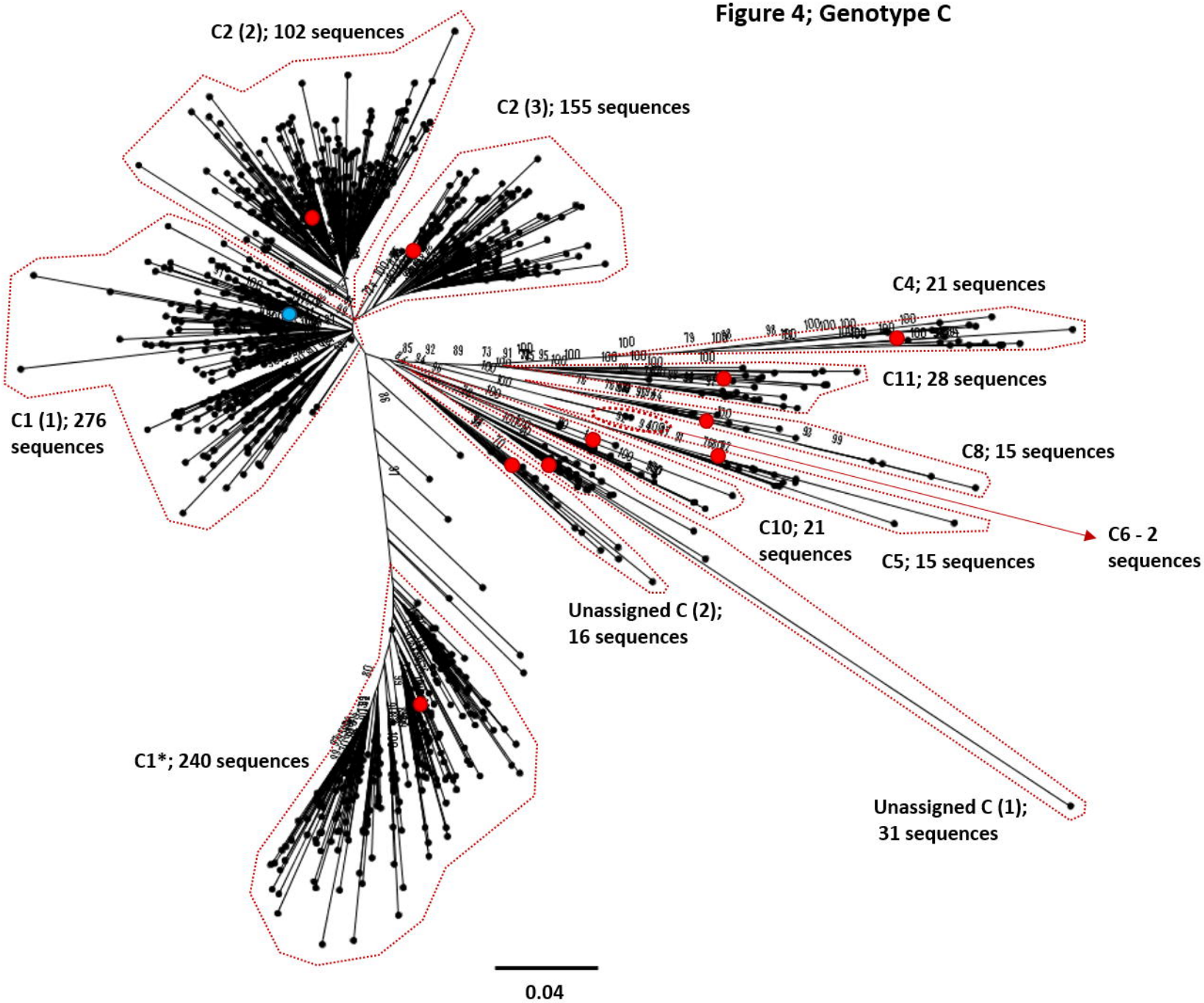


Figure 5; Genotype D

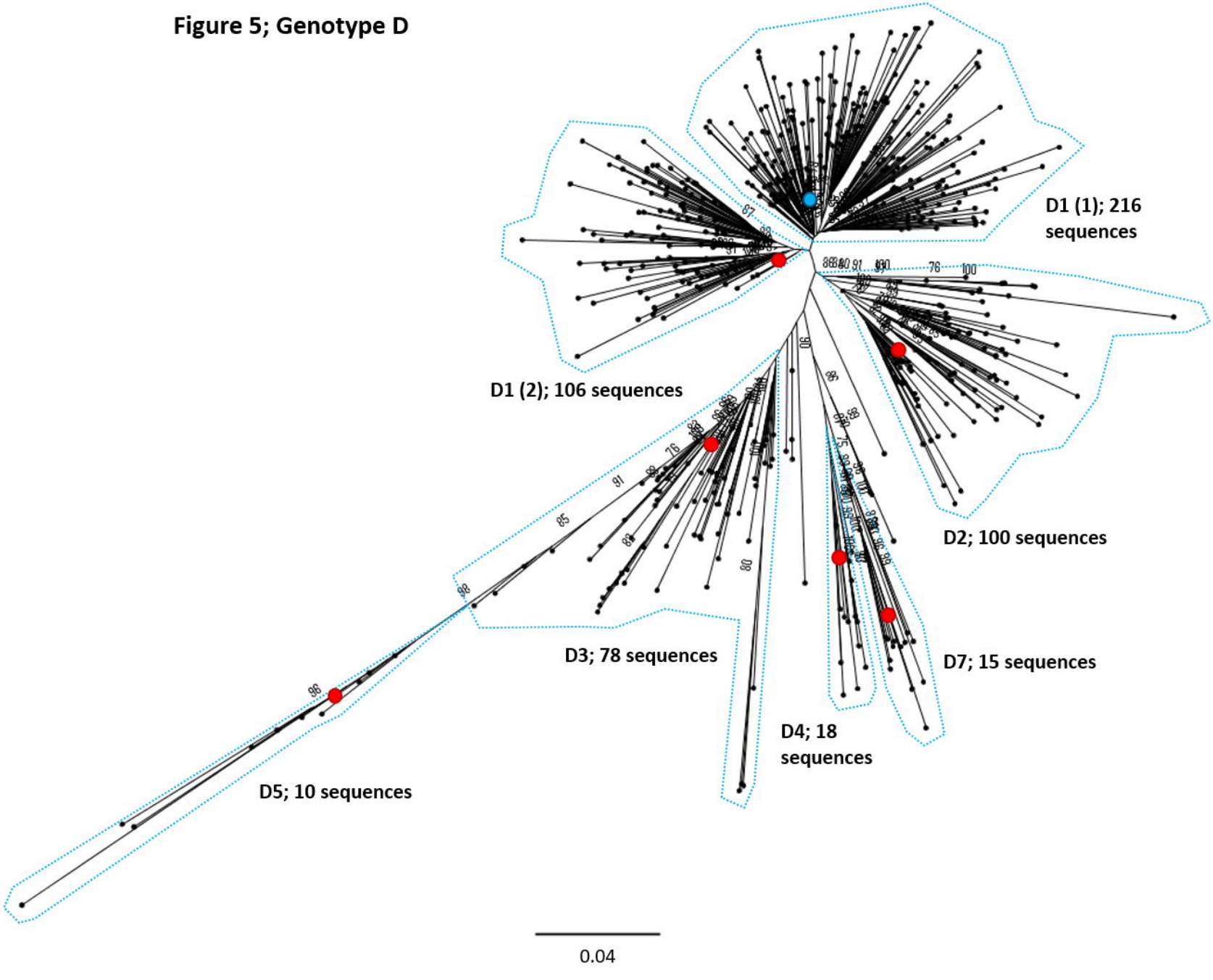
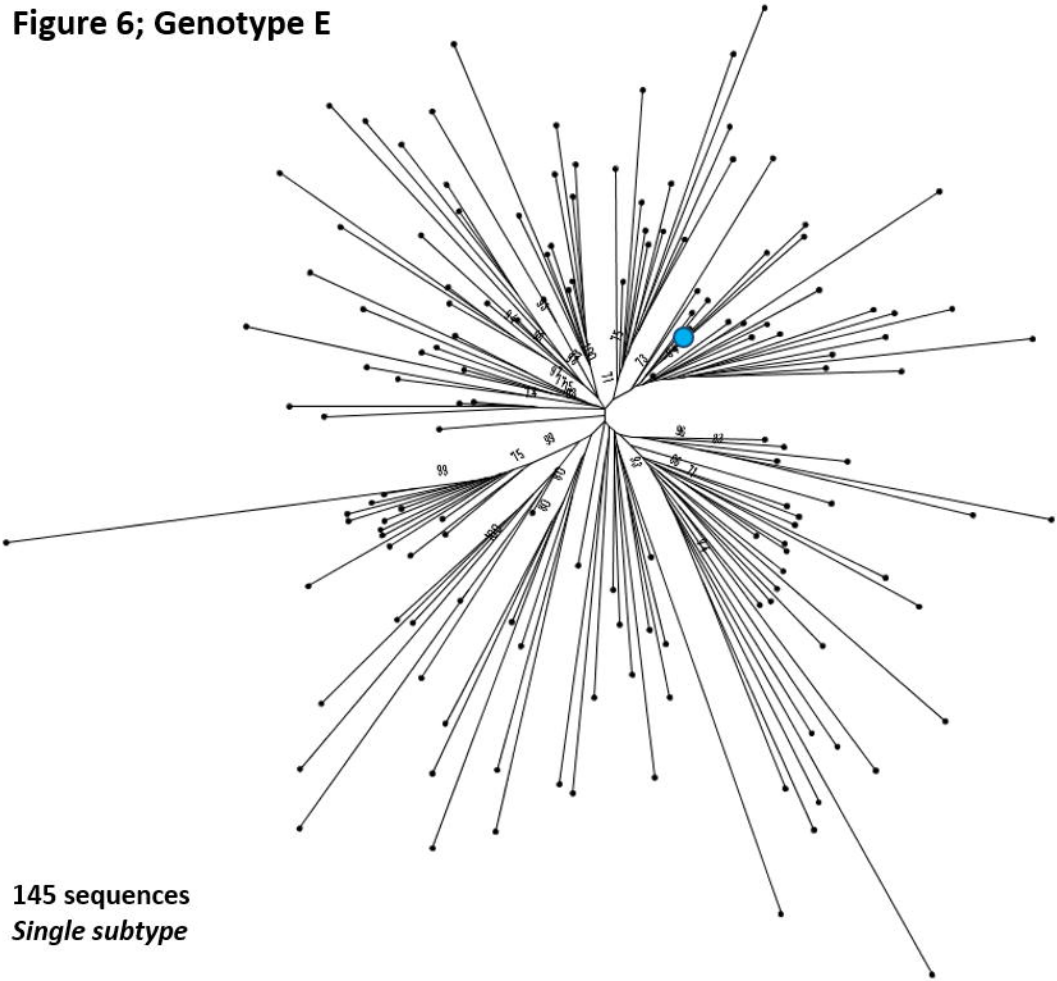


Figure 6; Genotype E



145 sequences
Single subtype

0.006

Figure 7; Genotype F

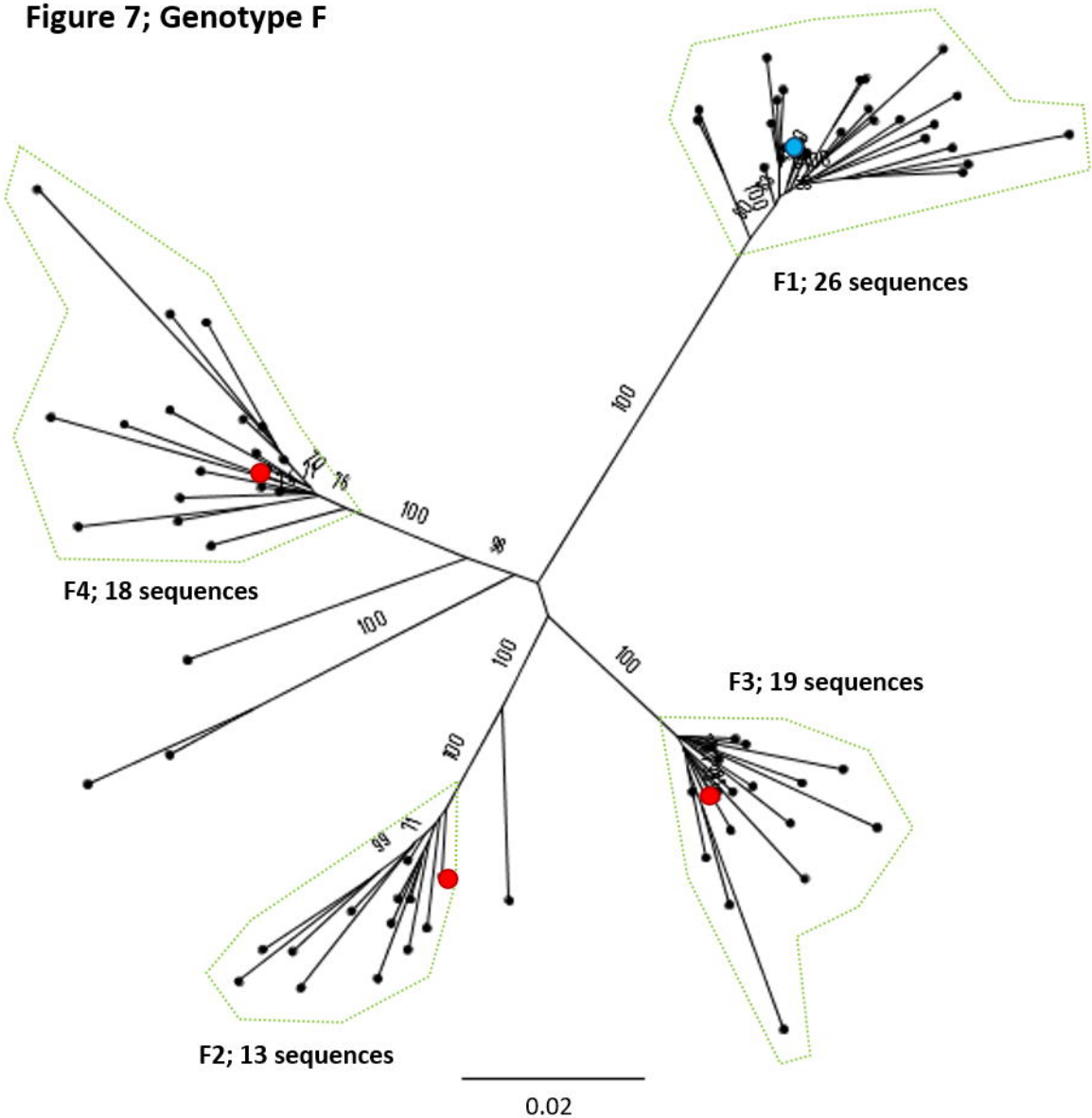


Figure 8; Genotype H

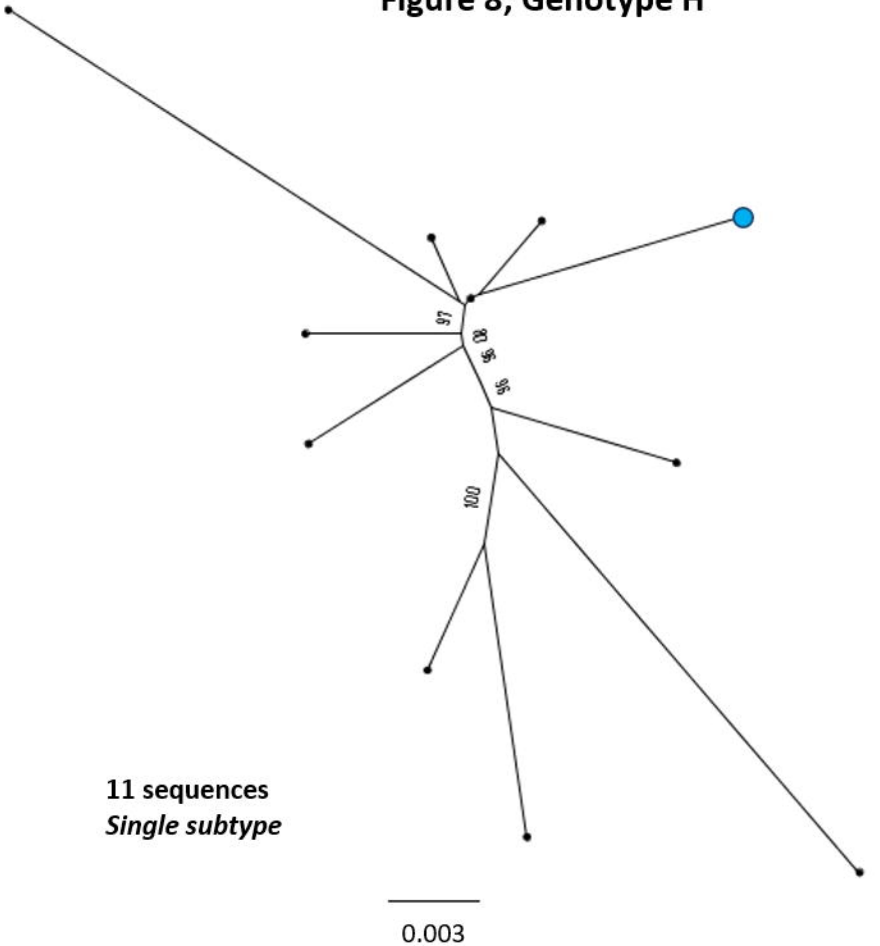


Figure 9; Genotype I

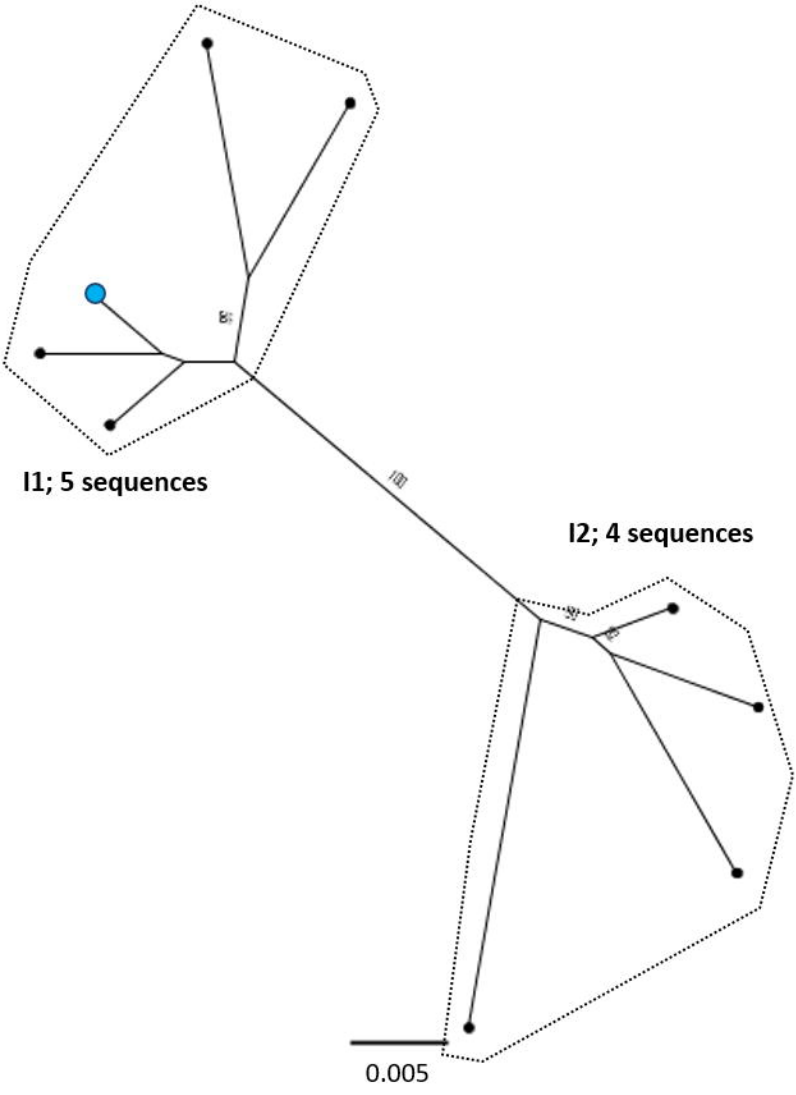


Figure 10; Reference sequences

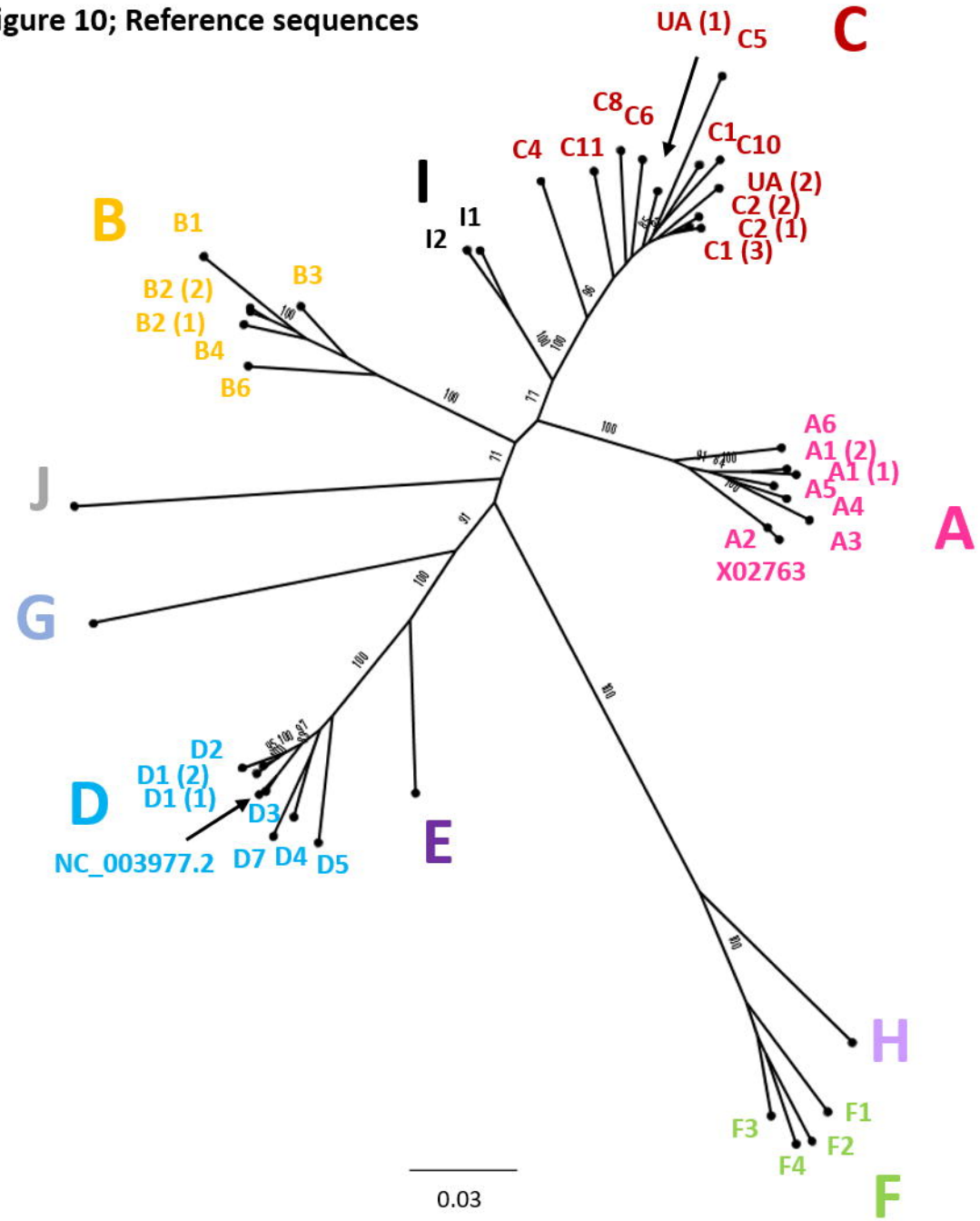
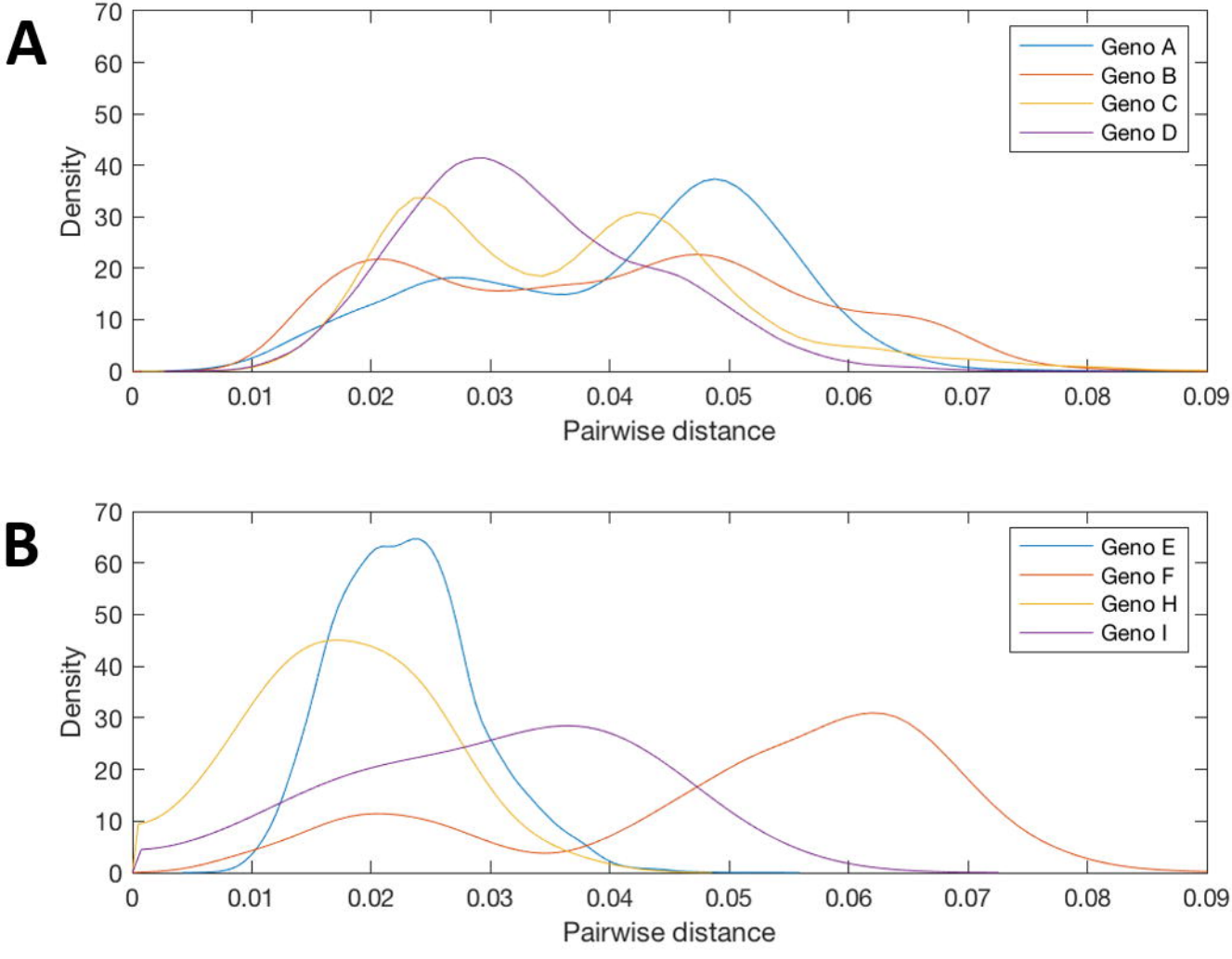



Figure 11; Pairwise analysis of genotypes