Proposed reference sequences for hepatitis E virus subtypes

Donald B. Smith,1 Peter Simmonds,1 Jacques Izopet,2 Edmilson F. Oliveira-Filho,3 Rainer G. Ulrich,4,5 Reimar Johne,6 Matthias Koenig,7 Shahid Jameel,8 Tim J. Harrison,9 Xiang-Jin Meng,10 Hiroaki Okamoto,11 Wim H. M. Van der Poel12† and Michael A. Purdy13†

1University of Edinburgh, Centre for Immunity, Infection and Evolution, Edinburgh, Scotland, UK
2INSERM, UMR1043, Toulouse F-31300, France
3Veterinary Virology and Animal Viral Diseases, Department of Infectious and Parasitic Diseases, Fundamental and Applied Research for Animal Health Centre, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium
4Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Institute for Novel and Emerging Infectious Diseases, Greifswald – Insel Riems, Germany
5German Center for Infection Research (DZIF), partner site Hamburg-Luebeck-Borstel-Insel Riems, Germany
6Federal Institute for Risk Assessment, Berlin, Germany
7Justus-Liebig-University, Giessen, Germany
8The Wellcome Trust/DBT India Alliance, Hyderabad, India
9University College London, London, UK
10College of Veterinary Medicine, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
11Division of Virology, Department of Infection and Immunity, Jichi Medical University School of Medicine, Tochigi-ken, Japan
12Central Veterinary Institute, Wageningen University and Research Centre, Lelystad, The Netherlands
13Centers for Disease Control and Prevention, National Center for HIV/Hepatitis/STD/TB Prevention, Division of Viral Hepatitis, Atlanta, GA, USA

The nomenclature of hepatitis E virus (HEV) subtypes is inconsistent and makes comparison of different studies problematic. We have provided a table of proposed complete genome reference sequences for each subtype. The criteria for subtype assignment vary between different genotypes and methodologies, and so a conservative pragmatic approach has been favoured. Updates to this table will be posted on the International Committee on Taxonomy of Viruses website (http://talk.ictvonline.org/r.ashx?C). The use of common reference sequences will facilitate communication between researchers and help clarify the epidemiology of this important human pathogen. This subtyping procedure might be adopted for other taxa of the genus Orthohepevirus.

The current literature contains several inconsistencies in the naming of hepatitis E virus (HEV) subtypes, which often creates confusion in the scientific community. HEV is a member of the family Hepeviridae within the genus Orthohepevirus. The genus has three species that infect birds (Orthohepevirus B), rodents, soricomorphs and carnivores (Orthohepevirus C) or bats (Orthohepevirus D), and one species, Orthohepevirus A, comprising seven genotypes that infect humans (HEV-1, -2, -3, -4 and -7), pigs (HEV-3 and -4), rabbit (HEV-3), wild boar (HEV-3, -4, -5 and -6),

†Members of the ICTV Hepeviridae study group.
mongooses (HEV-3), deer (HEV-3), yak (HEV-4) and camel (HEV-7) (Lee et al., 2015; Smith et al., 2014).

This division of HEV into seven genotypes and criteria for their assignment and identification are based on a demarcation p-distance threshold between genotypes of 0.088 for amino acid distances of concatenated ORF1- and ORF2-encoded proteins (lacking hypervariable regions between ORF1 protein amino acid residues 706–778 and 928–929, numbered with reference to GenBank accession no. M73218) (Smith et al., 2014). However, the criteria by which HEV sequences can be assigned to subtypes within genotypes are less consistent and sometimes confusing. When HEV subtypes were first comprehensively tabulated a decade ago, only 49 complete genome sequences were available and many subtype assignments were based on the analysis of subgenomic regions (Lu et al., 2006). Since then, the number of complete genome sequences has increased to almost 300 and most of the subtypes defined by Lu et al. (2006) are now represented by at least one complete genome sequence. However, there is currently no agreed list of reference sequences for these subtypes, although attempts at standardization have been made for HEV-3 (Smith et al., 2015; Vina-Rodriguez et al., 2015). One problem that is encountered in assigning sequences to particular subtypes is that no consistent criteria have been identified that define intra- and inter-subtype distances (Oliveira-Filho et al., 2013; Smith et al., 2013). For example, nucleotide p-distances between subtypes of HEV-1 are all less than 0.12, while those between subtypes of HEV-3 range from 0.12 to 0.26 and from 0.13 to 0.18 for subtypes of HEV-4. In addition, within these genotypes, the ranges of within- and between-subtype distances overlap. As a result, some complete genome sequences have been given conflicting subtype assignments.

An example comes from a recent paper (Lhomme et al., 2015) in which strain TR19 (GenBank accession no. JQ013794) was used as the reference sequence for subtype 3c. The frequency of subtype 3c infections has increased over the last decade in France, similar to the increase in subtype 3c documented previously in England and Wales (Ijaz et al., 2014). However, the ‘subtype 3c’ strains from the UK actually correspond to the subtype 3i reference sequence used in the French study. In other cases, subgenomic sequences used as reference sequences (Thiry et al., 2015) derive from strains for which no further sequence information is available. As a result, it has become difficult to compare the results of phylogenetic analyses carried out using different subgenomic regions or even using the same region in different studies.

To address these issues, we propose a standard reference set of complete genome sequences (Table 1). This table is available online on the International Committee on Taxonomy of Viruses (ICTV) website (http://talk.ictvonline.org/r.ashx?C) and will be updated as new information becomes available. In producing this table, we considered using the most central sequence (the medoid) in each subtype group as the reference sequence. Although this method avoids the possibility of choosing a divergent member of a subtype as the reference sequence, it would also mean that subtype reference sequences might not be stable because the medoid could change as more sequences are obtained or as the structure of the subtype is redefined by the addition or exclusion of divergent strains. In addition, our decision to use the designations of Lu et al. (2006), with priority given to strains with the earliest date of accession, will be minimally disruptive to the existing literature.

The criteria used were as follows:

1. To minimize disruption of previous subgenotype assignments, priority was given to the subtype assignments given by Lu et al. (2006).

2. To enable phylogenetic analyses to be carried out on different fragments of the genome, subtype reference sequences must comprise both the ORF1 and ORF2 coding regions and not be a recombinant between previously assigned subtypes.

3. If more than one complete genome sequence was available for a subtype, priority was given to the first sequence to be submitted to GenBank or, where submission dates were identical, the lowest alphabetic/numeric accession number.

4. If a subtype was assigned by Lu et al. (2006) based on the analysis of subgenomic fragments, these fragments were used to identify potential reference sequences by performing a BLAST search against GenBank. The highest-scoring complete genome sequences were considered as potential reference sequences if nucleotide sequence identities were >90% and formed a discontinuous distribution compared with identities with previously named complete genome sequences.

5. Complete genome sequences that were phylogenetically distinct from previously assigned complete genome sequences, and not related to any of the subtypes described by Lu et al. (2006), were only assigned as a new subtype if at least three complete ORF1 and ORF2 sequences were available that were epidemiologically unrelated (from different studies or localities). Unassigned complete genome sequences were labelled ‘genotype_accession number’ (e.g.’3_AB369689’).

**PHYLOGENETIC AND SEQUENCE ANALYSES**

HEV sequences >7000 nt were downloaded from GenBank on 27 October 2015 and aligned using sse v.1.2 (Simmonds, 2012). Sequences differing by <1% (HEV-1 and HEV-3) or 2% (HEV-4) of nucleotide positions were analysed by producing neighbour-joining trees, based on maximum composite likelihood distances using MEGA6 (Tamura et al., 2013), or by analysing the distribution of nucleotide p-distances using sse. Analyses in sequence sets lacking
hypervariable regions or lacking the overlapping ORF2/3 region produced similar results (data not shown).

**GENOTYPE 1**

Subtypes 1a–1e were all originally assigned on the basis of an analysis of complete genome sequences (Lu et al., 2006).

A group of sequences that share a common branch with subtype 1a (JF443721–JF443726 and AB720035, Fig. 1a) are more divergent from subtype 1a (nucleotide p-distances 0.052–0.075, apart from M73218 to JF443726, with a distance of 0.046), than sequences of subtype 1a are from each other (<0.056), these distances being comparable to those between subtypes 1b and 1c (<0.058–0.065). We propose that this phylogenetically distinct group of sequences
be considered as subtype 1f, although no discontinuity exists in the distribution of pairwise nucleotide \( p \)-distances within HEV-1 sequences that distinguishes within- and between-subtype distances. Sequence FJ457024 is intermediate between subtypes 1a and 1f, but bootstrap analysis using sse suggests that it is a recombinant between these two subtypes (data not shown). All \( p \)-distances >0.087 derive from comparisons between subtypes 1a, 1b, 1c and 1f and either subtype 1d or 1e ( >0.101), or between subtypes 1d and 1e (0.096), supporting the division of HEV-1 into two clades: 1abcf (comprising subtypes 1a, 1b, 1c and 1f) and 1de (comprising subtypes 1d and 1e).

**GENOTYPE 2**

Only a single complete genome sequence has been reported for genotype 2a; genotype 2b was identified from the analysis of a 318 nt ORF2 fragment (Lu et al., 2006).

**GENOTYPE 3**

The distribution of nucleotide \( p \)-distances among HEV-3 subtypes shows a complex pattern with multiple hierarchies of relatedness, even if the more divergent rabbit-derived strains are excluded. Subtypes 3a, 3b, 3c, 3h, 3i and 3j (3abchij) form one major clade (Fig. 1b), while subtypes 3e, 3f and 3g form another (3efg) (Hewitt et al., 2014; Ijaz et al., 2014; Oliveira-Filho et al., 2013; Smith et al., 2015; Widen et al., 2011). The reference sequences for subtypes 3e and 3f were assigned according to their date of accession to GenBank. Five strains belonging to subtype 3c were listed by Lu et al. (2006); their partial ORF1 and ORF2 sequences group with the corresponding regions of the complete genome sequence FJ705359, and separately from IJQ13794, previously described as subtype 3c (Izopet et al., 2012). The latter sequence becomes the subtype 3h reference sequence since it groups with the ORF1 and ORF2 sequences of a subtype 3h strain listed by Lu et al. (2006) (AF110390 and AF110387). The other 3h strain (swNZ) listed by Lu et al. (2006) groups separately from all complete genome sequences. Four strains of subtype 3i are listed by Lu et al. (2006); sequences of one of these strains groups with FJ998008 for both the ORF1 and ORF2 regions. The other three strains have sequences only loosely (ORF1) or not associated (ORF2) with this sequence. Accordingly, we have assigned FJ998008 as the 3i reference sequence. Nucleotide \( p \)-distances between these subtypes ( >0.120) overlap distances within subtypes ( <0.123), making it difficult to unambiguously assign some subtypes. For example, nucleotide \( p \)-distances between subtype 3f and EU360977 (0.116–0.125) and between 3f and KJ873911 (0.116–0.125) span this range, as does between 3h and AB290312 (0.120), while AB369689 and AB740232 are equally related to subtypes 3a (nucleotide \( p \)-distances 0.124–0.134) and 3b (0.126–0.137). We have chosen not to assign a subtype to these sequences, or to more divergent sequences such as IJQ53664 and AB290313. Divergence among the HEV-3 rabbit-derived strains ranges up to 0.255, again with multiple levels of sequence divergence; assignment of these strains into named subtypes within the 3a clade awaits the availability of further complete genome sequences.

**GENOTYPE 4**

Seven HEV-4 subtypes were defined by Lu et al. (2006) (subtypes 4a–4 g). The distribution of nucleotide sequence distances between and within HEV-4 subtypes is nearly continuous with distances between subtypes (>0.133) overlapping those within subtypes (<0.139), although a peak from 0.15 to 0.18 consists only of distances between subtypes. Phylogenetic analysis also reveals multiple levels of branching (Fig. 1c) but without higher-level groupings akin to those observed for HEV-1 and HEV-3. Consequently we used a pragmatic approach, adopting previous designations and avoiding the proliferation of new subtype names. One of the 4f subgenomic ORF1 GenBank accession numbers given by Lu et al. (2006) (AY427953) should be AY684253. However, both this sequence and another subtype 4f ORF1 sequence (AB075970) group with the subtype 4a reference sequence (data not shown). Two additional subtype 4f sequences given by Lu et al. (2006) (AB082547 and AB082558) derive from the HE-JA2 strain for which a complete genome sequence is now available (AB220974) and which is distinct from previously named subtypes, so this becomes the subtype 4f reference sequence. Two additional subtypes (4h and 4i) follow the assignments given in a previous publication (Liu et al., 2012). Sequence AB369688, although distinct from other subtypes, is represented by a single complete genome sequence and therefore remains unassigned.

**GENOTYPES 5–7**

The distance between the two complete genome sequences of HEV-6 (AB602441 and AB856243) is 0.198, and between the three complete genome sequences of HEV-7 (KJ496143, KJ496144 and KT818608) is 0.06–0.147. Comparison with distances between subtypes of HEV-3 and between subtypes of HEV-4 would suggest that both HEV-6 and HEV-7 could also be divided into two subtypes. However, as fewer than three complete genome sequences are currently available for each of these clades, we have not made any subtype assignments except to designate the first sequence of each genotype as subtype ‘a’.

**CONCLUDING REMARKS**

A perennial problem in classifying virus diversity is that discrete, man-made categories used for classification become arbitrary as their genetic distinctness blurs into a continuum of variability with the description of additional novel lineages or recombinants. This problem has hindered the assignment of subtypes of HEV because of different levels of diversity within different HEV genotypes, and because both distance-based and phylogenetic methods do not provide clear criteria for demarcation between
groups. Despite this problem, it is important that researchers have a common set of named reference sequences so that results from different studies can be compared. We hope that our table of subtype reference sequences will assist the interpretation of epidemiological and evolutionary studies of HEV. However, an important caveat is that researchers should use these reference sequences as way-markers in a complex landscape and be cautious about treating subtypes as stable biological or epidemiological entities.

ACKNOWLEDGEMENTS

This work was supported by The Wellcome (grant 095831/Z/11/Z) to the Centre for Immunity, Infection and Evolution at the University of Edinburgh, UK.

REFERENCES


