Genetic variability and the classification of hepatitis E virus

Journal of Virology

Genetic Variability and the Classification of Hepatitis E Virus

Donald B. Smith, Michael A. Purdy and Peter Simmonds
J. Virol. **2013, 87(8):4161. DOI: 10.1128/JVI.02762-12.**
Published Ahead of Print 6 February 2013.

JVI Journals.ASM.org

Genetic Variability and the Classification of Hepatitis E Virus

Donald B. Smith,[a] Michael A. Purdy,[b] Peter Simmonds[a]

University of Edinburgh, Centre for Immunology, Infection and Evolution, Ashworth Laboratories, Edinburgh, United Kingdom[a]; Centers for Disease Control and Prevention, National Center for HIV/Hepatitis/STD/TB Prevention, Division of Viral Hepatitis, Atlanta, Georgia, USA[b]

The classification of hepatitis E virus (HEV) variants is currently in transition without agreed definitions for genotypes and subtypes or for deeper taxonomic groupings into species and genera that could incorporate more recently characterized viruses assigned to the *Hepeviridae* family that infect birds, bats, rodents, and fish. These conflicts arise because of differences in the viruses and genomic regions compared and in the methodology used. We have reexamined published sequences and found that synonymous substitutions were saturated in comparisons between and within virus genotypes. Analysis of complete genome sequences or concatenated ORF1/ORF2 amino acid sequences indicated that HEV variants most closely related to those infecting humans can be consistently divided into six genotypes (types 1 to 4 and two additional genotypes from wild boar). Variants isolated from rabbits, closely related to genotype 3, occupy an intermediate position. No consistent criteria could be defined for the assignment of virus subtypes. Analysis of amino acid sequences from these viruses with the more divergent variants from chickens, bats, and rodents in three conserved subgenomic regions (residues 1 to 452 or 974 to 1534 of ORF1 or residues 105 to 458 of ORF2) provided consistent support for a division into 4 groups, corresponding to HEV variants infecting humans and pigs, those infecting rats and ferrets, those from bats, and those from chickens. This approach may form the basis for a future genetic classification of HEV into four species, with the more divergent HEV-like virus from fish (cutthroat trout virus) representing a second genus.

The classification of hepatitis E virus (HEV) is currently in transition. Following its first recognition as an enterically transmitted virus that causes acute self-limiting hepatitis in large epidemics and sporadic cases, the genome was characterized as having a single-stranded positive-sense RNA genome that encodes three open reading frames (1–3). Additional sequences were quickly obtained from isolates sampled around the world, and a hierarchy of relatedness became clear, with variants classified as further genotypes, subtypes, or isolates depending upon their degree of sequence relatedness to existing variants. Early classification schemes were based on partial genome sequences; for example, there was a suggestion that variants differing in nucleotide sequence by >20% in the ORF2 region should be classified into different genotypes (4). A more comprehensive analysis of the bootstrap support for phylogenetic groupings and the nucleotide distances between these groupings in comparisons of complete virus genomes or a variety of subgenomic regions led to the recognition of 4 genotypes and 24 subtypes (5). Subsequent studies have differed in their assignment of isolates to particular virus subtypes (6) or in their recognition of subtypes as a taxonomic grouping (7). More recently, there has been controversy about whether isolates from rabbits and wild boar should be considered new genotypes or subtypes (8–12).

These problems have arisen in part because of the use of different genomic regions (complete genomes or subgenomic regions) and phylogenetic methods (pairwise distances or bootstrap values of the resulting neighbor-joining or maximum likelihood trees). There are also no commonly agreed criteria to define different taxonomic categories. The most recent statement from the International Committee on Taxonomy of Viruses (ICTV) *Hepeviridae* study group (13) considers the *Hepeviridae* to consist of HEV genotypes 1 to 4, together with a number of divergent isolates from chickens (14) and rats (15), whose taxonomic status is undecided. The situation has been further compounded by the recent identification of HEV isolates infecting bats (16) and ferrets (17). An example of the type of confusion that has arisen is that HEV "genotype 5" has variously been assigned to avian HEV (14), to variants found in rabbits (9) and rats (18), and to one (11) or two (16) variants found in wild boar.

In this study, we have undertaken a reanalysis of HEV phylogenetic relationships using the wider data set now available, taking into account the variable diversity observed between coding and noncoding regions, between synonymous and nonsynonymous sites, and between hypervariable and constrained regions. We have used a variety of methods in order to describe the phylogenetic relationships among HEV and related viruses isolated from nonhuman species. This analysis may be of value in developing a consensus and evidence-based classification of HEV species and genotypes.

MATERIALS AND METHODS

One hundred eighty complete genome sequences were downloaded from GenBank on 13 March 2012. Sequences were removed from the data set if there was evidence that they were recombinant (GenBank accession numbers DQ450072, D11093, and AF051830) (19–21), if they differed from any other sequence in the data set by <2% of nucleotide positions (excluding the hypervariable region [HVR]). This left a total of 108 sequences (accession numbers FJ906896, JQ013793, JQ013791, JQ013792, FJ906895, GU937805, AB291960, AB291953, JN564006, JQ013795, JQ013794, AF060669, AY575857, AB074920, AB481228, AB630970,

Received 4 October 2012 **Accepted** 24 January 2013

Published ahead of print 6 February 2013

Address correspondence to Donald B. Smith, D.B.Smith@ed.ac.uk.

doi:10.1128/JVI.02762-12

The authors have paid a fee to allow immediate free access to this article.

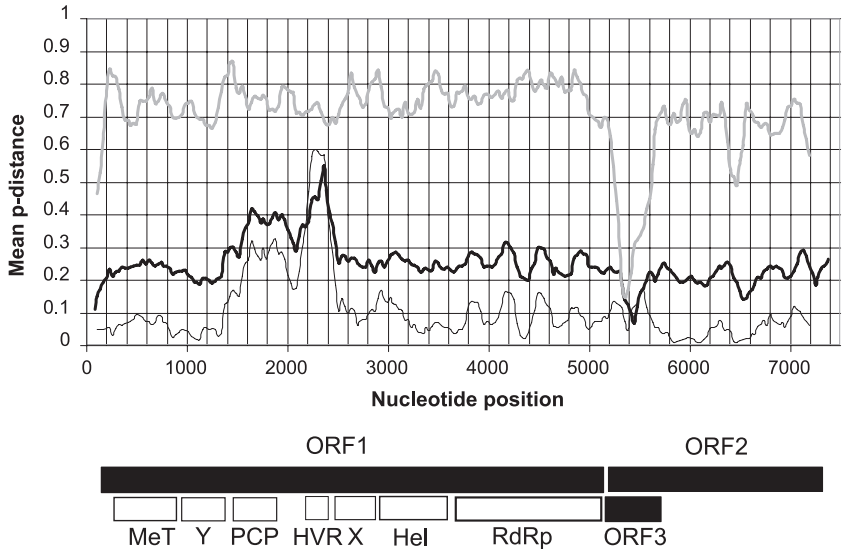

FIG 1 Sliding-window analysis of *p*-distance between HEV genomes. Nucleotide sequences of single representatives of genotypes 1 (GenBank accession number M80581), 2 (accession number M74506), 3 (accession number AF060669), and 4 (accession number GU119960); a rabbit isolate (accession number FJ906895); and the divergent wild boar isolates reported under GenBank accession numbers AB573435 and AB602441 were aligned, and mean distances were calculated for overlapping windows of 150 nucleotides shifted by 30 nucleotides and plotted against the midpoint of the window. Plots are shown for all sites (dark line), synonymous sites (gray line), and nonsynonymous sites (thin line) for concatenated ORF1 and ORF2 regions. The positions of the three open reading frames are shown along with the approximate positions within ORF1 of the methyltransferase (MeT), Y domain (Y), papain-like cysteine protease (PCP), hypervariable region (HVR), X domain (X), helicase (Hel), and RNA-dependent RNA polymerase (RdRp).

AB091394, AB369691, AB591734, AB089824, AB073912, AB290312, FJ705359, AB222184, AP003430, FJ998008, AB481229, AB630971, AB291962, AB291963, AB246676, AB591733, AB189070, AB222182, AY115488, AB236320, AB222183, FJ527832, AF455784, FJ426403, FJ426404, AF060668, AB369689, AB290313, FJ653660, FJ956757, AB248522, EU360977, EU723512, AB481226, AB291961, EU375463, EU495148, EU723514, AB248520, EU723516, AB369687, EU723513, HM055578, FJ998015, AY594199, AB253420, AJ272108, HM152568, GU206559, AB074915, FJ610232, GU361892, EU366959, HQ634346, EF077630, AB197674, GU119960, AB197673, FJ763142, GU119961, GU188851, AY723745, EU676172, HM439284, DQ279091, AB108537, AB220974, AB369688, AB602441, AB573435, EF570133, JF915746, AB369690, AB602440, AB481227, AB193176, AB080575, AB161719, AB291967, M80581, M94177, X98292, M73218, AF459438, FJ457024, AF076239, X99441, AF185822, DQ459342, AY204877, AY230202, and M74506), to which were added sequences of HEV-like variants from chickens (accession numbers JN997392, AM943646, GU954430, AY535004, EF206691, JN597006, and AM943647), rats (accession numbers GU345042, GU345043, JN167537, and JN167538), bats (accession numbers JQ001749 and NC_018382), ferrets (accession numbers JN998606 and JN998607), and cutthroat trout (accession number HQ731075).

Nucleotide sequences were aligned by using the SSE v1.0 package (22), with reference to amino acid alignments obtained by using MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/). Phylogenetic trees were produced by using MEGA 5.1 (23).

RESULTS

Pattern of diversity in the HEV genome. The 7.2-kb genome of HEV consists of a 5′ cap, a 35-nucleotide (nt) 5′-noncoding region, three open reading frames, and a 3′-noncoding region of about 79 nt followed by a poly(A) tail.

ORF1 (~5,000 nt) encodes a nonstructural polyprotein, and ORF2 (1,920 nt) encodes the virus capsid protein, while ORF3 (~340 nt) overlaps the 5′ end of ORF2 by ~300 nt and encodes a phosphoprotein that interacts with cellular signaling proteins. Coding sequences therefore comprise more than 95% of the HEV genome. We investigated the distribution of sequence variation in the HEV genome by carrying out sliding-window analysis of nucleotide differences between single representatives of the major phylogenetic groupings known to infect humans (genotypes 1 to 4) together with a representative of the genotype 3-related group isolated from rabbits and two more distant isolates from wild boar (Fig. 1). Mean nucleotide distances were relatively homogenous over the genome except for two peaks of variability in ORF1. The first region, between positions 1500 and 2000, includes the papain-like cysteine protease domain and a downstream region of unknown function. The second region of increased variability occurred near position 2200 and corresponds to the previously described hypervariable region of ORF1. This proline- and serine-rich region has been predicted to be intrinsically disordered (24), with no satisfactory alignment of sequences from different genotypes (25, 26). There were also two regions of reduced variability, one near the beginning of the genome and the other centered at position 5350, corresponding to the region where ORF2 and ORF3 are encoded by overlapping reading frames. Similar findings were obtained if the rabbit or wild boar sequences were omitted from the analysis (data not shown).

We also investigated the distribution of synonymous and nonsynonymous variation across the genome on concatenated ORF1/ORF2 sequences from which noncoding regions and termination codons had been removed. As observed for all nucleotide sites, mean nonsynonymous *p*-distances peaked in the protease and HVR regions of ORF1 (Fig. 1). Nonsynonymous distances also varied elsewhere in the genome, although in none of these regions did mean sequence distances exceed 0.2. In contrast, synonymous *p*-distances were close to saturation (≈0.75) throughout the HEV

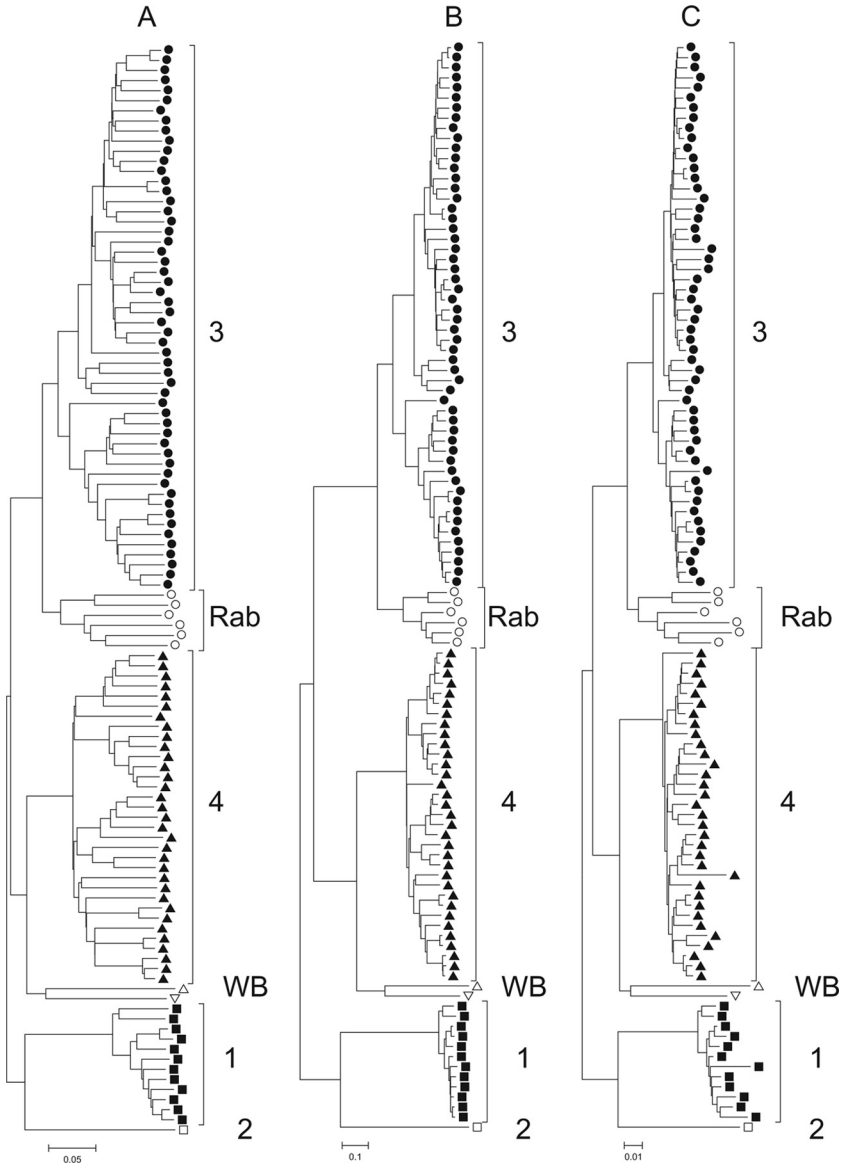

FIG 2 Phylogenetic analysis of HEV complete genome sequences. Neighbor-joining trees were produced by using nucleotide *p*-distances of complete genome sequences (A), maximum likelihood analysis of complete genome sequences (B), or amino acid *p*-distances of concatenated ORF1/ORF2 from which the HVR had been removed (C). Genotypes are indicated as follows: ■, genotype 1; □, genotype 2; ●, genotype 3; ○, rabbit variants (Rab); ▲, genotype 4; △ and ▽, wild boar (WB). Branches leading to these groups of sequences were supported by 100% of bootstrap replications in all cases.

genome (Fig. 1). The only exceptions were the 5′ terminus of ORF1, presumably reflecting constraints imposed by RNA structures proposed for this region (27), which may be involved in RNA replication and translation, and the ORF2/3 overlap region centered at position 5350. Suppression of synonymous substitutions in this region is expected as a consequence of the constraint imposed by the overlapping ORF2 and ORF3 reading frames but possibly also by the presence of RNA structures in this region required for the initiation of translation (28). A more modest suppression of synonymous substitutions was observed in a region of ORF2 centered at position 6400, possibly indicating the presence of a currently uncharacterized RNA structure in this region.

A similar analysis of sequences from isolates grouping with genotype 3 or genotype 4 revealed that synonymous substitutions also approached saturation (*p*-distance > 0.5) in comparisons between isolates that differed in nucleotide sequence by a distance of >0.14.

Phylogenetic analysis. From the complete genome sequences available in GenBank, we obtained 108 primary sequences that differed from each other at >2% of nucleotide positions (excluding the HVR) and that were not recombinants between different viruses. Phylogenetic analysis based on distances between these nucleotide sequences revealed three groupings: group A, genotypes 1 and 2; group B, genotype 3 and isolates from rabbits; and group C, genotype 4 and two sequences isolated from wild boar (Fig. 2A). The same groupings, albeit with much longer basal branch lengths, were observed by maximum likelihood (Fig. 2B).

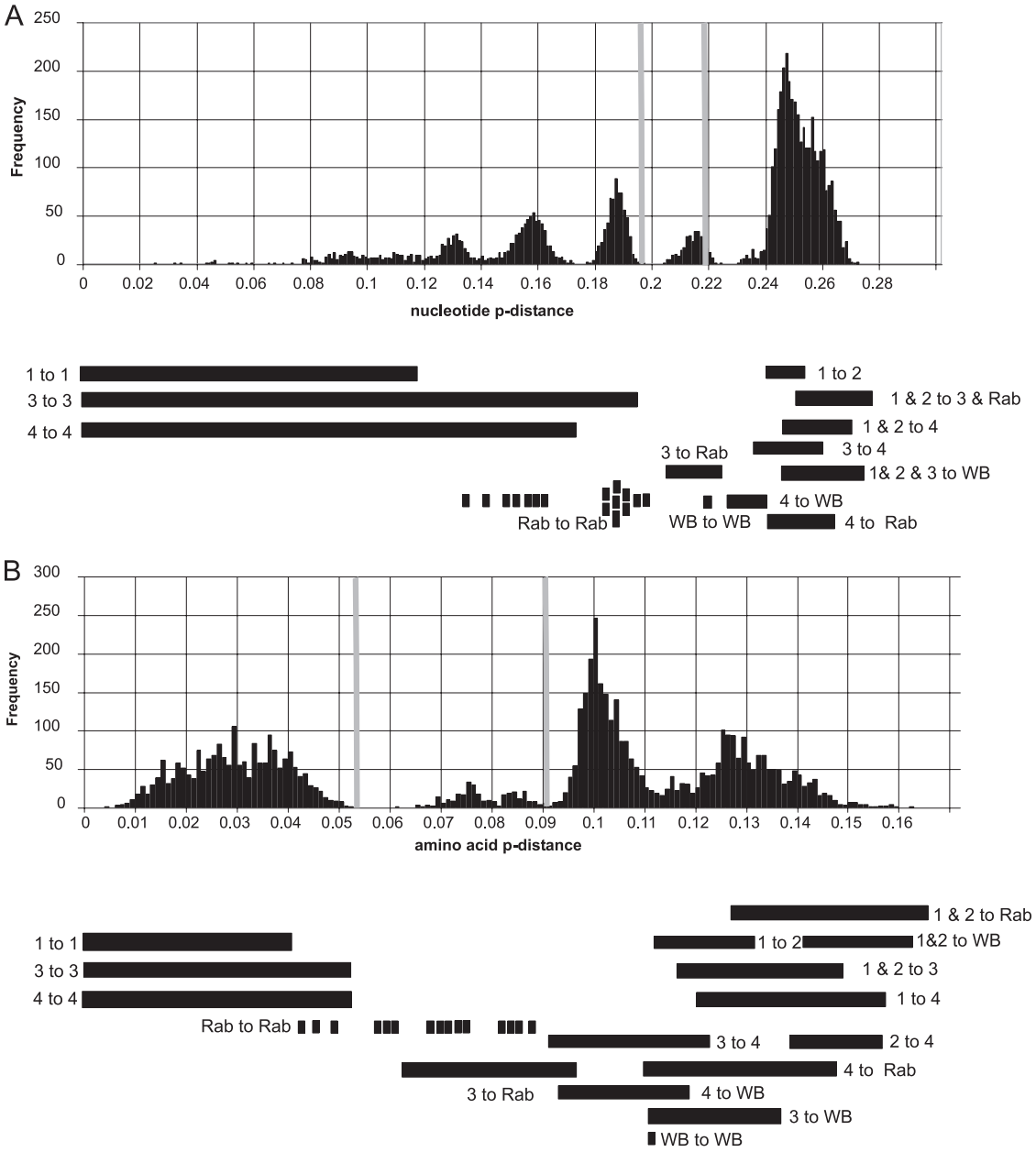

FIG 3 Distribution of nucleotide and amino acid *p*-distances. Nucleotide distances between 108 complete genome sequences (A) or amino acid distances between corresponding ORF1/ORF2 concatenated sequences from which the HVR had been removed (B) were calculated. Vertical bars indicate the upper and lower limits, respectively, of within-genotype and between-genotype distances for genotypes 1, 2, 3, and 4 and the two variants from wild boar. The ranges in which values fall for particular within-and between-group comparisons are shown at the bottom. Values for comparisons within the rabbit group are shown as individual points.

Since sites of synonymous substitutions approached saturation in intergenotypic and some intragenotype sequence comparisons (Fig. 1) and since the HVR could not be aligned for different genotypes (25, 26), we also analyzed sequences consisting of ORF1 joined directly to ORF2 and from which the HVR had been removed. The rationale for this was to remove sites that are uninformative for comparisons between genotypes and subtypes. Phylogenetic analysis of these concatenated sequences produced trees with the same three groupings of sequences whether we compared nonsynonymous sites (data not shown) or amino acid sequences using distance-based (Fig. 2C) or maximum likelihood (data not shown) methods. Again, basal branch lengths were much longer than observed for nucleotide distances at all sites.

In order to test the interrelationships of these groupings, we also examined the frequency distributions of nucleotide and amino acid distances (Fig. 3A and B). Several distinct peaks of sequence distances were observed, compressed toward large or small distances, respectively. For genotypes 1 to 4, distances within each genotype were less than 0.19 (nucleotide) or 0.052 (amino acid), while those between genotypes ranged from 0.23 to 0.27 (nucleotide) or 0.09 to 0.156 (amino acid). The two wild boar isolates differed from each other by a distance of 0.22 (nucleotide)

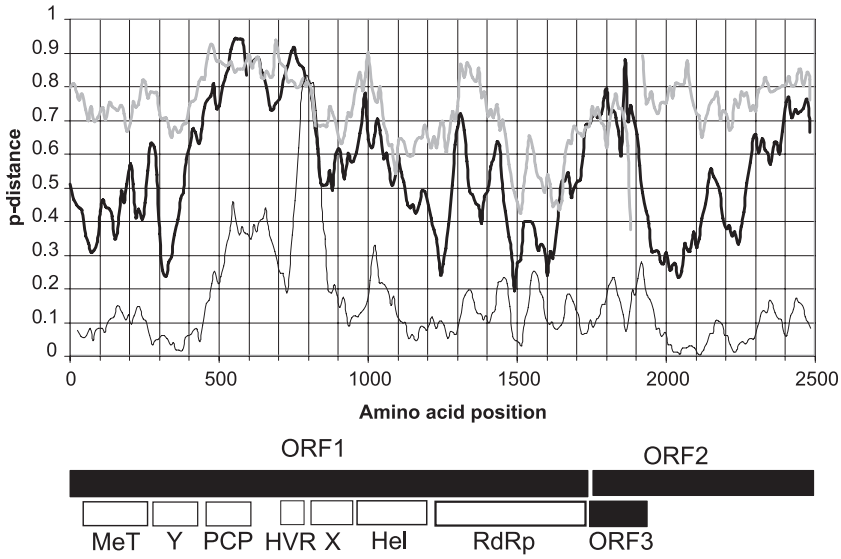

FIG 4 Sliding-window analysis of mean amino acid *p*-distances between representative divergent HEV sequences. Concatenated ORF1/ORF2 sequences of HEV type 1 reported under GenBank accession number D10330, avian HEV accession number AM943646, rat HEV accession number GU345042, and bat HEV accession number JQ001749 were aligned by using MUSCLE. Mean amino acid *p*-distances were calculated with SSE v1.0, using a sliding window of 50 amino acids shifted by 10 amino acids (dark line); between cutthroat trout virus and these sequences (gray line); or among HEV genotypes 1, 2, 3, and 4 and the two variants from wild boar (thin line). The positions of the open reading frames and domains within ORF1 are shown (labeled as described in the legend of Fig. 1).

or 0.11 (amino acid) and from genotype 4 isolates by a distance of 0.23 to 0.24 (nucleotide) or 0.092 to 0.117 (amino acid), overlapping the range of distances observed between genotypes 3 and 4 (0.23 to 0.26 [nucleotide] or 0.09 to 0.122 [amino acid]). These values imply that the two wild boar sequences should be considered additional genotypes, namely, genotypes 5 (GenBank accession number AB602441) and 6 (accession number AB573435), as previously suggested (10).

The situation is less clear with regard to the genotype 3-related sequences isolated from rabbits (and a closely related sequence isolated from a human [29]). These sequences differed from genotype 3 isolates by a distance of 0.20 to 0.22 (nucleotide) or 0.062 to 0.095 (amino acid), ranges intermediate between that of the intragenotypic distributions and that of the intergenotypic distributions. In addition, some of the rabbit sequences differed from each other as much as they did from genotype 3 sequences and could themselves be divided into two groups. Hence, the rabbit sequences could be considered two additional genotypes that are more closely related to genotype 3 than are other genotypes or unusually divergent subtypes of genotype 3. For genotype 3, two barely overlapping distributions were observed: a major peak centered at a distance of 0.187 (nucleotide) or 0.038 (amino acid) and a second distribution of distances of less than 0.175 (nucleotide) or 0.032 (amino acid). Fewer sequences are available for genotype 1; sequences reported under GenBank accession numbers AY204877 and AY23002 had nucleotide distances from other sequences of >0.1, compared to <0.09 among the other sequences, but this distinction was not observed for amino acid distances. Within genotype 4, nucleotide and amino acid *p*-distances formed single broad distributions centered on a distance of 0.158 (nucleotide) or 0.03 (amino acid). Hence, any boundaries used to define HEV subtypes that are based on nucleotide or amino acid distances would have to be specific for individual types and not reflect discontinuities in the distribution of distances within a genotype.

Nonhuman HEV-like variants. We have also analyzed sequence distances and phylogenetic relationships of the more divergent HEV-like variants isolated from chickens, rats, bats, and ferrets. Since synonymous substitutions were also saturated between these even more distantly related sequences, we aligned the amino acid sequences of concatenated ORF1/ORF2 using MUSCLE. Sliding-window analysis of amino acid *p*-distances between representative avian, rat, bat, and genotype 1 sequences revealed that the average of the pairwise distance between these four sequences fell below 0.5 in only three regions (Fig. 4). These more conserved regions correspond to amino acids 1 to 452 and 974 to 1534 of ORF1 and residues 105 to 458 of ORF2 (numbered according to positions in the amino acid sequence of HEV genotype 1 reported under GenBank accession number M80581). Phylogenetic analysis of amino acid *p*-distances in these three regions produced trees that were very similar to each other, with four main groupings: group A, HEV genotypes 1 to 4 and the rabbit and wild boar isolates; group B, avian HEV isolates; group C, bat HEV isolates; and group D, rat and ferret HEV isolates (Fig. 5). Expressed as within- and between-group ranges, amino acid *p*-distances were less than 0.12 or greater than 0.38 for ORF1 amino acids 1 to 452, less than 0.14 or greater than 0.36 for ORF1 amino acids 974 to 1534, and less than 0.07 or greater than 0.25 for ORF2 amino acids 105 to 458 (Fig. 6). The only exceptions were comparisons between the rat and ferret sequences, which were intermediate in all three regions, with amino acid *p*-distances of 0.14 to 0.16, 0.2 to 0.22, and 0.1 to 0.12, respectively.

Alignment of HEV and HEV-like sequences with that of cutthroat trout virus, a virus with a similar genome organization, was more difficult. Amino acid sequence *p*-distances were greater than 0.6 over most of the genome, the only exception being residues 1480 to 1770 of ORF1, where most distances were in the range of 0.4 to 0.6 (Fig. 4). Phylogenetic analysis of this region of the genome produced a tree in which cutthroat trout virus branched separately and in which the human-related and rat/ferret se-

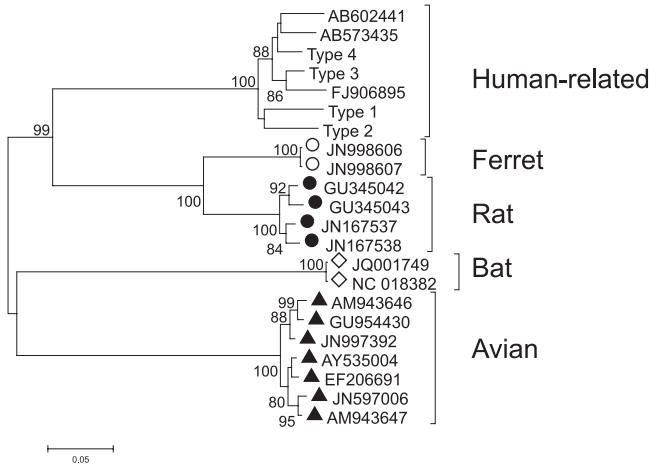

FIG 5 Phylogenetic analysis of zoonotic HEV sequences. Amino acid sequences of human and related HEV isolates (types 1 to 4 and isolates reported under GenBank accession numbers AB602441, AB573435, and FJ906895) and rat (●), ferret (○), bat (◇), and avian (▲) HEV isolates were aligned, and *p*-distances for the region of ORF1 at residues 1 to 452 were used to construct neighbor-joining trees. Bootstrap values (1,000 replicates) and a scale bar are indicated.

quences were more closely related to each other than they were to the bat or avian sequences (data not shown). The observed phylogeny does not support the proposal that avian HEV forms a separate phylogenetic grouping from the mammalian HEV isolates (11).

DISCUSSION

Our analysis of the pattern and extent of diversity among HEV genome sequences is relevant to current discussions about the classification of HEV variants. We report here that comparisons between the different HEV genotypes that infect humans approach saturation at synonymous sites throughout the genome and also at nonsynonymous sites in the HVR (Fig. 1). These observations lead us to suggest that the classification of HEV and its variants should be based on methods that exclude these sites. For simplicity, we have generally used phylogenetic relationships and distance distributions of amino acid sequences of concatenated ORF1/ORF2 coding regions, excluding the HVR, for analysis of genetic relationships between HEV genotypes and more divergent HEV-like viruses. However, similar phylogenetic trees can be produced by using distances at nonsynonymous sites or maximum likelihood analysis of complete genome sequences.

This approach is not unprecedented: distances between amino acid sequences of individual proteins are specified in the ICTV recommendations for the classification of genera and species within many different RNA and DNA virus families. Further classification into virus types and subtypes is usually based on nucleotide distances, for example, in the cases of hepatitis B virus (HBV) and hepatitis C virus (HCV). However, these viruses differ from HEV in the extensive overlapping reading frames of the HBV genome that suppress synonymous substitutions, while for HCV, divergence at nonsynonymous sites is greater. For these reasons, the phylogenetic analysis of HEV genomes that include synonymous sites is compressed relative to maximum likelihood or amino acid distance trees (Fig. 2). Similarly, although a reasonable phylogeny of HEV can be obtained by an analysis including the HVR (30), much of the phylogenetic signal derives from conserved regions flanking the HVR; more compressed trees are obtained if analysis is limited to the HVR alone (data not shown).

Following this rationale, those isolates of HEV most closely related to those infecting humans would be divided into six genotypes, as follows: genotypes 1 to 4, as defined by the ICTV (13), with two additional genotypes represented by two isolates from wild boar (10). If the rabbit-derived isolates are excluded, all intergenotypic *p*-distances between genotypes are greater than 0.22 (nucleotide) or 0.09 (amino acid), while distances within these groups are no more than 0.19 (nucleotide) or 0.052 (amino acid). This conclusion extends that of previous analyses of wild boar isolates based upon the nucleotide sequences of a subgenomic region or of complete genome sequences (10, 12). Our analysis also revealed a higher-level grouping of genotypes into three groups: group A, genotypes 1 and 2; group B, genotype 4 and the two wild boar-derived isolates; and group C, genotype 3 and the "rabbit" isolates.

Our analysis was less definitive about the group of six sequences isolated from rabbits and a closely related variant from a

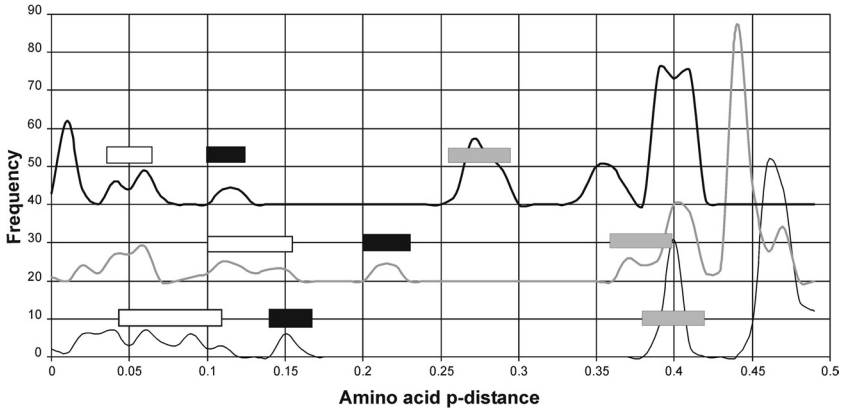

FIG 6 Distribution of amino acid *p*-distance among divergent HEV variants. The frequency of amino acid *p*-distances among the sequences used in Fig. 5 were plotted for ORF1 residues 1 to 452 (thin line), ORF1 residues 974 to 1534 (gray line; frequency, +20), and ORF2 residues 105 to 458 (solid line; frequency, +40). Values corresponding to distances between rat/ferret sequences and other sequences are indicated by gray bars, between rat and ferret sequences are indicated by black bars, and between variants of HEV (single representatives of HEV genotypes 1 to 4 and each of the two wild boar variants) are indicated by open bars.

human. Reflecting this difficulty, some studies have concluded that these isolates represent an additional genotype (8, 9, 29), while others considered them to be a subtype of genotype 3 (10, 13, 18). We found that *p*-distances between the rabbit isolates and genotype 3 isolates ranged from 0.2 to 0.22 (nucleotide) or 0.062 to 0.095 (amino acid) and were therefore intermediate between distances within genotypes (<0.19 [nucleotide] or <0.052 [amino acid]) and among different genotypes (>0.23 [nucleotide] or >0.09 [amino acid]) (Fig. 3). Other information about these isolates is also ambiguous: all seven isolates share a 93-nt insertion in the X domain of ORF1 encoding a proline-rich peptide (25, 29) that appears to be hypervariable (amino acid *p*-distances of 0.37 to 0.73 between isolates, compared to 0.04 to 0.1 for ORF1 as a whole), but the lineage does not appear to be host specific, since one isolate is of human origin (29). Such difficulties are likely to increase as additional sequences become available for analysis; analysis of a 189-nt fragment of ORF2 of 37 variants from rabbits revealed even greater diversity (29) than observed among the seven available complete genome sequences.

Similar ambiguities apply to the assignment of intragenotypic variants as subtypes, assignments previously based upon phylogenetic analysis and pairwise distances between complete or subgenomic nucleotide sequences (5, 6). We observed broad distributions of nucleotide and amino acid distances within genotypes 1 and 4, while two barely overlapping distributions were observed for genotype 3, excluding the rabbit isolates (data not shown). Although a division of genotype 3 into two subgroups has been proposed (12), some of the distances between these two genotype 3 lineages were less than those observed within the single distributions observed for genotypes 1 and 4. Hence, the currently defined subtype designations are not supported by boundaries that are consistent between genotypes or that reflect discontinuities in the distribution of distances.

Another reason for questioning the assignment of virus subtypes is the lack of evidence for host range or pathogenic differences between isolates belonging to some of these intragenotypic phylogenetic clusters. HEV isolates from humans and pigs are interspersed within intragenotypic lineages, while clinical correlates of phylogenetic differences between such variants have yet to be reported. In this context, the categorization of virus isolates at levels below genotype loses much of its importance. Epidemiological studies of virus outbreaks (31, 32), infectious sources (33, 34), transmission events (35, 36), or mixed infections (37) can demonstrate phylogenetic relationships without requiring the assignment of isolates to particular subtypes.

These conclusions differ from those reached in several previous studies for a number of reasons. In most cases, previous taxonomic proposals were based upon the phylogenetic comparison of nucleotide sequences. For example, an analysis of 37 complete genome sequences led to a description of 4 genotypes and 12 subtypes (6), while a more comprehensive study of 49 complete genomes and a variety of subgenomic sequences concluded that HEV could be divided into 4 genotypes and 24 subtypes (5). A more cautious analysis of 75 complete genome sequences identified 4 genotypes and at least seven subclusters or subgroups (7). Unfortunately, the designations given in those studies were sometimes contradictory and still give rise to confusion (12). The most recent ICTV statement on the taxonomy of HEV recognizes four genotypes of HEV (13) but does not explain the methodology used to produce the phylogenetic trees presented. While all of those studies identified the same four genotypes, they each utilized different subgenomic regions, and none provided criteria for the designation of subtypes. Those analyses all included synonymous substitutions and in some cases the HVR and therefore include many sites at which substitutions approach saturation and potentially obscure sequence relationships. Our analysis of amino acid differences in concatenated ORF1/ORF2 sequences that exclude the HVR avoids many of these difficulties.

We applied similar reasoning to extend the analysis to the more divergent isolates from chicken, rat, and bat. The alignment of concatenated ORF1/ORF2 amino acid sequences was more convincing in some parts of the genome than in others (Fig. 5), as previously reported for comparisons of bat HEV (16). In order to remove the spurious phylogenetic signal derived from potentially misaligned regions, we limited our analysis to three regions (ORF1 residues 1 to 452, ORF1 residues 974 to 1534, and ORF2 residues 105 to 458) where amino acid *p*-differences were less than 0.5. The phylogenetic trees produced from these three regions were congruent and consistent with four groupings. These groupings, extending a recent analysis based upon the 108-residue amino acid sequence from ORF1 (residues 1419 to 1526) or the entire ORF1 or ORF2 (16), were as follows: group A, HEV isolates that infect humans or are closely related to such isolates (genotypes 1 to 4, the two wild boar isolates, and the rabbit isolates); group B, avian HEV; group C, bat HEV; and group D, rat HEV and ferret HEV. The latter two viruses are only slightly more different from each other than are the genotypes of human-related HEV (Fig. 6) and could be considered species-specific genotypes of rat HEV. A taxonomic distinction between mammalian and avian HEV isolates (11) is not supported by phylogenetic analysis of distance comparisons for either of the ORF1 regions, whether or not cutthroat trout virus is included as an outgroup. The ICTV study group suggested that avian HEV may comprise a distinct genus within the *Hepeviridae* family (13), while additional genera have been proposed for rat HEV (12) and bat HEV (16). However, amino acid *p*-distances between all these HEV-like variants were in the range of 0.25 to 0.5 (Fig. 6). For comparison, the currently recognized genera within the *Picornaviridae* differ by a distance of more than 0.58 in polyprotein amino acid sequence, show minor differences in gene complements and, in some genera, insertion of nonhomologous genes (such as 2A), and possess several different classes of internal ribosomal entry sites that prevent alignment of complete genome sequences. In contrast, members of the same genus typically differ by a *p*-distance of 0.3 to 0.5 and almost invariably possess alignable, colinear genomes (38). If these criteria were applied to HEV and HEV-like variants, the *Hepevirus* genus would include three additional species, namely, avian HEV, bat HEV, and rat/ferret HEV.

Cutthroat trout virus, which differs in amino acid sequence over almost all of its genome by an average distance of 0.6 to 0.9 from representatives of these four HEV groupings (Fig. 4) but which has small regions of lower divergence and similarities in overall genome organization, would therefore represent a plausible candidate member of a second genus within the *Hepeviridae*. In the longer term, it is possible that the *Hepeviridae* family may be further expanded. For example, a partial genome sequence that differs by an amino acid distance of >0.7 in the RNA-dependent RNA polymerase region of ORF1 was recently described from deep sequencing of untreated sewage from Nepal. Although

awaiting further sequence data to confirm whether or not it possesses an HEV-like genome organization, its divergence potentially justifies its potential addition to the virus family as an additional new genus (39).

One of the factors that has complicated the classification of HEV variants is the variety of methods that have been used in order to make taxonomic assignments. Most previous studies produced phylogenetic trees based upon nucleotide distances based upon complete or subgenomic regions (5–8, 10–12, 17, 29). However, phylogenetic relationships are obscured by this approach, since they include sites where substitutions have become saturated. This problem can be partly avoided by using maximum likelihood (10, 19) and Bayesian (16) methods to examine phylogenetic relationships or, as shown here, by comparing reliably aligned amino acid sequences. This approach has the advantage of being more transparent as well as making analysis simpler.

Another issue that has clouded the definition of taxonomic groupings within the *Hepeviridae* is that these have often been defined based on the presence of phylogenetic branches that have substantial bootstrap support. However, in many instances, there is a hierarchy of such branches so that it becomes difficult to decide which branches are taxonomically informative. We suggest that the distribution of amino acid sequence *p*-distances provides a simple method for assessing the relationship of different taxonomic groupings. Isolates with intermediate distances in these distributions, such as the rabbit isolates related to genotype 3 and the ferret HEV related to rat HEV, should perhaps be considered ambiguous. Although robust classification schemes based on nucleotide sequences have been developed for viruses such as hepatitis C virus (40, 41) and the *Picornaviridae* (38), there is no biological reason why patterns of virus variation and phylogenetic relationships (or optimal methods of phylogenetic analysis) should necessarily be the same between different virus families or fit into discrete categories such as species, type, and subtype, as defined for other virus families. Complications can also arise where a high frequency of recombination dilutes the concept of distinct types as, for example, with HIV and enteroviruses. While there is no reason to expect naturally occurring patterns of sequence diversity of hepeviruses to be fitted perfectly to what is essentially a human-made taxonomic hierarchy of genera, species, and types, the comprehensive analysis presented in the current study does provide the starting point for their future classification. Future agreement on the criteria that define these taxonomic ranks is important not only for future ICTV descriptions of the virus family but also for a further understanding the species specificity and zoonotic sources of infection that characterize the epidemiology of HEV worldwide.

ACKNOWLEDGMENTS

This work was supported by a grant from the Wellcome Trust to the CIIE at the University of Edinburgh.

This information has not been formally disseminated by the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry. It does not represent and should not be construed to represent any agency determination or policy.

REFERENCES

1. **Tam AW, Smith MM, Guerra ME, Huang CC, Bradley DW, Fry KE, Reyes GR.** 1991. Hepatitis E virus (HEV): molecular cloning and sequencing of the full-length viral genome. Virology **185**:120–131.
2. **Koonin EV, Gorbalenya AE, Purdy MA, Rozanov MN, Reyes GR, Bradley DW.** 1992. Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. Proc. Natl. Acad. Sci. U. S. A. **89**:8259–8263.
3. **Tsarev SA, Emerson SU, Reyes GR, Tsareva TS, Legters LJ, Malik IA, Iqbal M, Purcell RH.** 1992. Characterization of a prototype strain of hepatitis E virus. Proc. Natl. Acad. Sci. U. S. A. **89**:559–563.
4. **Worm HC, van der Poel WH, Brandstatter G.** 2002. Hepatitis E: an overview. Microbes Infect. **4**:657–666.
5. **Lu L, Li C, Hagedorn CH.** 2006. Phylogenetic analysis of global hepatitis E virus sequences: genetic diversity, subtypes and zoonosis. Rev. Med. Virol. **16**:5–36.
6. **Zhai L, Dai X, Meng J.** 2006. Hepatitis E virus genotyping based on full-length genome and partial genomic regions. Virus Res. **120**:57–69.
7. **Okamoto H.** 2007. Genetic variability and evolution of hepatitis E virus. Virus Res. **127**:216–228.
8. **Zhao C, Ma Z, Harrison TJ, Feng R, Zhang C, Qiao Z, Fan J, Ma H, Li M, Song A, Wang Y.** 2009. A novel genotype of hepatitis E virus prevalent among farmed rabbits in China. J. Med. Virol. **81**:1371–1379.
9. **Geng Y, Zhao C, Song A, Wang J, Zhang X, Harrison TJ, Zhou Y, Wang W, Wang Y.** 2011. The serological prevalence and genetic diversity of hepatitis E virus in farmed rabbits in China. Infect. Genet. Evol. **11**:476–482.
10. **Takahashi M, Nishizawa T, Sato H, Sato Y, Jirintai, Nagashima S, Okamoto H.** 2011. Analysis of the full-length genome of a hepatitis E virus isolate obtained from a wild boar in Japan that is classifiable into a novel genotype. J. Gen. Virol. **92**:902–908.
11. **Cao D, Meng X-J.** 2012. Molecular biology and replication of hepatitis E virus. Emerg. Microbes Infect. **1**:e17. doi:10.1038/emi.2012.7.
12. **Bouquet J, Cherel P, Pavio N.** 2012. Genetic characterization and codon usage bias of full-length hepatitis E virus sequences shed new lights on genotypic distribution, host restriction and genome evolution. Infect. Genet. Evol. **12**:1842–1853.
13. **Meng XJ, Anderson DA, Arankalle VA, Emerson SU, Harrison TJ, Jameel S, Okamoto H.** 2012. Hepeviridae, p 1021–1028. *In* King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses. Academic Press, London, United Kingdom.
14. **Huang FF, Sun ZF, Emerson SU, Purcell RH, Shivaprasad HL, Pierson FW, Toth TE, Meng XJ.** 2004. Determination and analysis of the complete genomic sequence of avian hepatitis E virus (avian HEV) and attempts to infect rhesus monkeys with avian HEV. J. Gen. Virol. **85**:1609–1618.
15. **Johne R, Plenge-Bonig A, Hess M, Ulrich RG, Reetz J, Schielke A.** 2010. Detection of a novel hepatitis E-like virus in faeces of wild rats using a nested broad-spectrum RT-PCR. J. Gen. Virol. **91**:750–758.
16. **Drexler JF, Seelen A, Corman VM, Fumie Tateno A, Cottontail V, Melim Zerbinati R, Gloza-Rausch F, Klose SM, Adu-Sarkodie Y, Oppong SK, Kalko EK, Osterman A, Rasche A, Adam A, Muller MA, Ulrich RG, Leroy EM, Lukashev AN, Drosten C.** 2012. Bats worldwide carry hepatitis E virus-related viruses that form a putative novel genus within the family Hepeviridae. J. Virol. **86**:9134–9147.
17. **Raj VS, Smits SL, Pas SD, Provacia LB, Moorman-Roest H, Osterhaus AD, Haagmans BL.** 2012. Novel hepatitis E virus in ferrets, the Netherlands. Emerg. Infect. Dis. **18**:1369–1370.
18. **Cossaboom CM, Cordoba L, Dryman BA, Meng XJ.** 2011. Hepatitis E virus in rabbits, Virginia, USA. Emerg. Infect. Dis. **17**:2047–2049.
19. **Fan J.** 2009. Open reading frame structure analysis as a novel genotyping tool for hepatitis E virus and the subsequent discovery of an intergenotype recombinant. J. Gen. Virol. **90**:1353–1358.
20. **van Cuyck H, Fan J, Robertson DL, Roques P.** 2005. Evidence of recombination between divergent hepatitis E viruses. J. Virol. **79**:9306–9314.
21. **Wang H, Zhang W, Ni B, Shen H, Song Y, Wang X, Shao S, Hua X, Cui L.** 2010. Recombination analysis reveals a double recombination event in hepatitis E virus. Virol. J. **7**:129. doi:10.1186/1743-422X-7-129.
22. **Simmonds P.** 2012. SSE: a nucleotide and amino acid sequence analysis platform. BMC Res. Notes **5**:50. doi:10.1186/1756-0500-5-50.
23. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**:2731–2739.

24. **Purdy MA, Lara J, Khudyakov YE.** 2012. The hepatitis E virus polyproline region is involved in viral adaptation. PLoS One **7**:e35974. doi:10.1371/journal.pone.0035974.
25. **Smith DB, Vanek J, Ramalingam S, Johannessen I, Templeton K, Simmonds P.** 2012. Evolution of the hepatitis E virus hypervariable region. J. Gen. Virol. **93**:2408–2418.
26. **Purdy MA.** 2012. Evolution of the hepatitis E virus polyproline region: order from disorder. J. Virol. **86**:10186–10193.
27. **Huang CC, Nguyen D, Fernandez J, Yun KY, Fry KE, Bradley DW, Tam AW, Reyes GR.** 1992. Molecular cloning and sequencing of the Mexico isolate of hepatitis E virus (HEV). Virology **191**:550–558.
28. **Huang YW, Opriessnig T, Halbur PG, Meng XJ.** 2007. Initiation at the third in-frame AUG codon of open reading frame 3 of the hepatitis E virus is essential for viral infectivity in vivo. J. Virol. **81**:3018–3026.
29. **Izopet J, Dubois M, Bertagnoli S, Lhomme S, Marchandeau S, Boucher S, Kamar N, Abravanel F, Guerin JL.** 2012. Hepatitis E virus strains in rabbits and evidence of a closely related strain in humans, France. Emerg. Infect. Dis. **18**:1274–1281.
30. **Legrand-Abravanel F, Mansuy JM, Dubois M, Kamar N, Peron JM, Rostaing L, Izopet J.** 2009. Hepatitis E virus genotype 3 diversity, France. Emerg. Infect. Dis. **15**:110–114.
31. **Grandadam M, Tebbal S, Caron M, Siriwardana M, Larouze B, Koeck JL, Buisson Y, Enouf V, Nicand E.** 2004. Evidence for hepatitis E virus quasispecies. J. Gen. Virol. **85**:3189–3194.
32. **Widen F, Sundqvist L, Matyi-Toth A, Metreveli G, Belak S, Hallgren G, Norder H.** 2011. Molecular epidemiology of hepatitis E virus in humans, pigs and wild boars in Sweden. Epidemiol. Infect. **139**:361–371.
33. **Feagins AR, Opriessnig T, Guenette DK, Halbur PG, Meng XJ.** 2007. Detection and characterization of infectious hepatitis E virus from commercial pig livers sold in local grocery stores in the USA. J. Gen. Virol. **88**:912–917.
34. **Colson P, Borentain P, Queyriaux B, Kaba M, Moal V, Gallian P, Heyries L, Raoult D, Gerolami R.** 2010. Pig liver sausage as a source of hepatitis E virus transmission to humans. J. Infect. Dis. **202**:825–834.
35. **Takahashi K, Kitajima N, Abe N, Mishiro S.** 2004. Complete or near-complete nucleotide sequences of hepatitis E virus genome recovered from a wild boar, a deer, and four patients who ate the deer. Virology **330**:501–505.
36. **Matsubayashi K, Kang JH, Sakata H, Takahashi K, Shindo M, Kato M, Sato S, Kato T, Nishimori H, Tsuji K, Maguchi H, Yoshida J, Maekubo H, Mishiro S, Ikeda H.** 2008. A case of transfusion-transmitted hepatitis E caused by blood from a donor infected with hepatitis E virus via zoonotic food-borne route. Transfusion **48**:1368–1375.
37. **Moal V, Gerolami R, Colson P.** 2012. First human case of co-infection with two different subtypes of hepatitis E virus. Intervirology **55**:484–487.
38. **Knowles NJ, Hovi T, Hyypia T, King AMQ, Lindberg AM, Pallansch MA, Palmenberg AC, Simmonds P, Skern T, Stanway G, Yamashita T, Zell R.** 2012. Picornaviridae, p 855–880. *In* King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses. Academic Press, London, United Kingdom.
39. **Ng TF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E.** 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J. Virol. **86**:12161–12175.
40. **Simmonds P, Bukh J, Combet C, Deleage G, Enomoto N, Feinstone S, Halfon P, Inchauspe G, Kuiken C, Maertens G, Mizokami M, Murphy DG, Okamoto H, Pawlotsky JM, Penin F, Sablon E, Shin I, Stuyver LJ, Thiel HJ, Viazov S, Weiner AJ, Widell A.** 2005. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology **42**:962–973.
41. **Kuiken C, Simmonds P.** 2009. Nomenclature and numbering of the hepatitis C virus. Methods Mol. Biol. **510**:33–53.

JVI Journals.ASM.org

ERRATUM

Genetic Variability and the Classification of Hepatitis E Virus

Donald B. Smith, Michael A. Purdy, Peter Simmonds

University of Edinburgh, Centre for Immunology, Infection and Evolution, Ashworth Laboratories, Edinburgh, United Kingdom; Centers for Disease Control and Prevention, National Center for HIV/Hepatitis/STD/TB Prevention, Division of Viral Hepatitis, Atlanta, Georgia, USA

Volume 87, no. 8, p. 4161–4169, 2013. Page 4165, column 1, line 10 from the bottom: "AY23002" should read "AY230202."

doi:10.1128/JVI.01092-13