Open reading frame structure analysis as a novel genotyping tool for hepatitis E virus and the subsequent discovery of an inter-genotype recombinant

Jun Fan

Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK

Correspondence
Jun Fan
jun.fan-rres@bbsrc.ac.uk

Hepatitis E virus (HEV) is a virus which has been characterized recently (Bradley, 1990a, b; Feinstone & Purcell, 1978). HEV was first classified as a member of the family Caliciviridae (Cubitt et al., 1995), then put into the ‘unassigned virus’ status and is now in the new genus Hepevirus (Emerson et al., 2005).

Cases caused by HEV have been found worldwide. Outbreaks of HEV have been recorded in Asia, Africa and Mexico (Emerson & Purcell, 2003; Worm et al., 2002). Cases of acute, sporadic hepatitis have been found in the USA, Europe and Africa (Coursaget et al., 1998; Erker et al., 1999; Schlauder et al., 1999; Zanetti et al., 1999).

The HEV genome has three partially overlapped open reading frames (ORFs). ORF1 is located at the 5' terminus of the genome and encodes non-structural proteins (Erker et al., 1999; Koonin et al., 1992). ORF2 is at the 3' terminus of the HEV genome and encodes the viral capsid protein which has three glycosylation sites (Worm et al., 2002). ORF3 overlaps with either ORF1 or ORF2 (Wang et al., 2000).

Genotyping in HEV can have diagnostic and clinical implications. Several genotyping schemes for HEV have been proposed by different research groups. Four genotypes are agreed upon from all schemes: genotype 1 includes Asian and African pandemic strains (Aye et al., 1992; Chatterjee et al., 1997; Coursaget et al., 1998; Tam et al., 1991), genotype 2 includes Mexican and Nigerian strains (Buisson et al., 2000; Huang et al., 1992), genotype 3 includes the USA, Japanese and European strains (Erker et al., 1999; Schlauder et al., 1998, 1999; Takahashi et al., 2002) and genotype 4 includes the strains from China and Japan (Takahashi et al., 2002; Wang et al., 1999, 2000).

The genotyping groups used the following methods: nucleotide identities/genetic distances, deduced amino acid similarities, phylogenetic trees and restriction endonuclease sites, in order to determine genotypes based on either partial genome or full-length sequences. Comparing nucleotide identities/genetic distance is the most commonly used method to define HEV genotypes. Worm et al. (2002) defined an HEV genotype as viruses having nucleotide divergence of less than 20% in the ORF2 region. Another scheme defines a genotype as viruses having nucleotide divergence of less than 15% in the RNA polymerase region (Arankalle et al., 1999). It was also suggested that the genetic distance of the ORF1 region should be greater than 0.235 between genotypes (Wang et al., 1999).

Amino acid and phylogenetic analyses can be treated as the extension of genetic distance analysis. Amino acids are deduced from DNA sequences and phylogenetic trees are constructed based on genetic distances. HEV-US1 and HEV-US2 sequences were demonstrated to be only 74% identical to the Burmese and Mexican strains, which combining with amino acid and phylogenetic analyses indicates the presence of the third genotype in HEV (Erker et al., 1999).

It can be concluded that analyses of genetic distance, amino acid similarity and phylogenetic relationships have become
a common practice in defining HEV genotypes (Erker et al., 1999; Hsieh et al., 1998; Schlauder et al., 1999; Schlauder & Mushahwar, 2001; Wang et al., 1999).

Restriction endonuclease analysis (REA) has been much less used. Different restriction endonucleases are applied to strains. Different genotypes are supposed to give different restriction site patterns. PCR-restriction fragment length polymorphism (PCR-RFLP) can be observed from the different enzyme digestion patterns. Gouvea et al. (1998) compared ten complete and five partial HEV sequences from genotype 1 and 2 using REA. Another case was that genotypes can be distinguished among Taiwanese, Asian and Mexican strains by PCR-RFLP (Hsieh et al., 1998).

At the early stage of defining HEV genotypes, partial regions were commonly used to define genotypes because the majority of available HEV strains were only sequenced partially. It had also been demonstrated that an ORF1 region generates trees consistent with the trees generated with full-length sequences (Schlauder et al., 1998, 1999; Wang et al., 1999; Zanetti et al., 1999).

However, Zhai et al. (2006) found that in 23 out of the 24 cases, the partial genome region can not sufficiently determine the genotype of the complete genome by using one-way and two-way ANOVA to compare 24 partial genomic regions statistically.

Here I provide a new approach based on full-length genome sequences and the ORF structure (the lengths of three ORFs and the overlapping relationships among the ORFs), to investigate genotypes in HEV.

All available complete HEV sequences were retrieved from GenBank. The details of the retrieved sequences are listed in Supplementary Table S1, available in JGV Online.

Since there is only one available complete sequence for genotype 2, three genotype alignments were built using CLUSTAL W (Thompson et al., 1994) for genotype 1, 3 and 4, respectively, followed by manual curation referring to the corresponding genotype amino acid alignments. The four-genotype alignment was constructed by the profile alignment. Three sequences (GenBank accession nos AB091395, AB189071 and L08816) were removed from the alignment because they are identical to AB200239, AB189074 and NC_001434, respectively. The final alignment contains 73 sequences of up to 7169 bp (including gaps) and is available in JGV Online.

There are obvious differences in the ORF structure among genotypes. The average and sd values of the ORF structure are listed in Table 1. Genotype 1 and genotype 2 have a similar ORF structure. In genotype 1, 2 and 3, ORF2 overlaps with ORF3, whilst in genotype 4 ORF2 overlaps with ORF1. ORF1 in genotypes 3 and 4 is on average 40 nt longer than in genotypes 1 and 2.

As shown in Table 1, DQ450072 was observed to have a mosaic ORF structure. DQ450072 is a swine strain which was initially reported to belong to genotype 4. It has a longer ORF1 than other genotype 4 sequences. However, it has a shorter ORF2 and a longer ORF3, which is more similar to genotypes 1, 2 and 3. The mosaic ORF structure indicates that DQ450072 could be a putative inter-genotype recombinant.

To test this hypothesis, an alignment was built which includes DQ450072 sequence and the four genotype-consensus sequences. Parti-on (Fan & Robertson, unpublished tool) and Recco (Maydt & Lengauer, 2006) (\(\alpha\) is set to 0.2) were used to detect breakpoints.

Parti-on located two optimal breakpoints between genotypes 3 and 4 at position 4749–4758 and 5226–5270 with P-values lower than 0.001 for both breakpoints (Supplementary Fig. S1, available in JGV Online). Recco detected two breakpoints at similar positions. The two breakpoints were located in the RNA-dependent RNA polymerase (RDRP) region in ORF1 and the overlapping region of ORF2 and ORF3, respectively.

To verify the statistical support for the detected recombination events and to infer the phylogenetic relationship for the recombinant, the alignment was divided into three genomic regions according to the detected breakpoints. For each region, a maximum-likelihood (ML) tree was constructed based on the best fitted evolutionary model determined by MODELTEST (Posada & Crandall, 1998). The reliability of the branching order was verified by performing 1000 neighbour-joining bootstrapping. To apply the one-tailed Kishino–Hasegawa (KH) test and the

### Table 1. ORF structures of four HEV genotypes and one putative inter-genotype recombinant

<table>
<thead>
<tr>
<th>Genotype</th>
<th>ORF1 length</th>
<th>ORF2 length</th>
<th>ORF3 length</th>
<th>Total length</th>
<th>Gap12</th>
<th>Gap13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5079.59 (6.53)</td>
<td>1982.81 (0.75)</td>
<td>371.81 (0.75)</td>
<td>7099.35 (6.48)</td>
<td>39 (0)</td>
<td>−2 (0)</td>
</tr>
<tr>
<td>2</td>
<td>5076</td>
<td>1980</td>
<td>372</td>
<td>7093</td>
<td>39</td>
<td>−2</td>
</tr>
<tr>
<td>3</td>
<td>5116.56 (12.00)</td>
<td>1983 (0)</td>
<td>368.64 (1.8)</td>
<td>7133.56 (12.00)</td>
<td>36 (0)</td>
<td>−1.64 (1.8)</td>
</tr>
<tr>
<td>4</td>
<td>5121.29 (12.38)</td>
<td>2024.5 (1.94)</td>
<td>344.8 (1.10)</td>
<td>7142 (12.58)</td>
<td>−1.7 (1.64)</td>
<td>26 (0)</td>
</tr>
<tr>
<td>3/4 recombinant</td>
<td>5124</td>
<td>1983</td>
<td>369</td>
<td>7141</td>
<td>36</td>
<td>−2</td>
</tr>
</tbody>
</table>
Shimodaira–Hasegawa (SH) test, the alternative tree for each region was constructed by manually relocating the recombinant into the phylogenetic position of the ML tree(s) inferred from adjacent regions. These phylogenetic analyses were implemented with PAUP* (Swofford, 1999).

The phylogenetic analyses indicated that there are two recombination events between genotype 3 and genotype 4 (Fig. 1), which is statistically supported by both the KH test and the SH test \((P<0.001)\). The presence of recombination events caused the discordant phylogenetic relationships in different regions.

A bootscanning analysis was performed on the 73-sequence alignment with a window size of 500 bp and increment size of 50 bp to verify whether any other inter-genotype recombinant exists, which cannot be revealed by observing the ORF structure.

The bootscanning result indicated that the only available complete sequence in genotype 2, the Mexican strain, could be another inter-genotype recombinant.

The same procedure of detecting and verifying breakpoints in DQ450072 was applied to the Mexican strain. Two recombination events were detected between genotypes 1 and 3 with strong bootstrapping support, which were located at positions 1847–1856 and 2069–2072 (Fig. 2). However, the KH test and the SH test rejected the hypothesis that genotype 2 is clustered with genotype 3.

Fig. 1. Maximum-likelihood phylogenetic trees inferred for the three regions split by two detected breakpoints located at positions 4749–4758 and 5226–5270 in the inter-genotype recombinant with GenBank accession no. DQ450072. Each HEV sequence is represented by the name in the format of accession number _ genotype _ host _ sequence name. Only bootstrap values for the clustering of genotypes are indicated. The recombinant is highlighted in yellow. Scale bar, number of nucleotide substitutions per site. b, Boar; h, human; mng, mongoose; sw, swine; u, unknown; wb, wild boar.
rather than genotype 1 in the region of 1857–2069 \((P=0.097 \text{ and } 0.104 \text{ for the KH test and the SH test, respectively})\), which rejects the hypothesis.

As a recently characterized virus, many efforts have been made on the taxonomy and genotyping of HEV. These efforts have led to some widely accepted results, including the four-genotype nomenclature. Among these efforts, different methods have been used to genotype HEV strains based on partial genome regions. However, the insufficiency of statistical support for determining genotypes for HEV strains based on genomic regions has been reported (Zhai et al., 2006). In this study, the strain DQ450072, which was previously classified as a genotype 4 sequence, has been found to be a recombinant of genotype 3 and genotype 4. The discovery of this first inter-genotype recombinant means that genotyping HEV strains should be based on full-length genome sequences. Genotyping based on genomic regions could give misleading results to inter-genotype recombinant sequences.

The recombinant was revealed by observing the ORF structure of HEV sequences. The ORF structure has drawn our attention for several reasons. Firstly, the ORF structure is highly related to the commonly used genotyping methods based on nucleotide identities/genetic distances, deduced amino acid similarities and phylogenetic trees.

**Fig. 2.** Maximum-likelihood phylogenetic trees inferred for the three regions split by two debatable breakpoints located at positions 1847–1856 and 2069–2072 in genotype 2 GenBank accession no. M74506. Each HEV sequence is represented by the name in the format of accession number _ genotype _ host _ sequence name. Only bootstrap values for the clustering of genotypes are indicated. The recombinant is highlighted in yellow. Scale bar, number of nucleotide substitutions per site. b, Boar; h, human; mng, mongoose; sw, swine; u, unknown; wb, wild boar.
genotypes generate gaps which increase the distance and further affect the inferred phylogenetic trees. The gap region between genotypes also decreases the nucleotide and amino acid identities. It has been shown that several new genotypes suggested by genetic distances, sequence identities and phylogenetic trees are not sufficient to be a separate genotype (Schlauder et al., 1999; Schlauder & Mushahwar, 2001; Wang et al., 2000). Therefore the ORF structure can be a new criterion to define HEV genotypes. The small SD values in Table 1 mean that there is less variance among ORF structures of strains from the same genotype, which indicates the feasibility of using the ORF structure as a criterion for genotyping.

Secondly, due to the different ORF structures in different genotypes, the regions of genotype-specific gaps can be used to design genotype-specific primers. There are two regions containing the most genotype-specific primer variance, one is in the hypervariable region and the other is located at the overlapping region of ORF1, 2 and 3. Compared with the restriction endonuclease genotyping method, primers can provide more flexibility, reliability and accuracy, considering that the usual lengths of restriction endonuclease sites and primers are 6–8 bp and 25–30 bp, respectively.

Finally the target organisms which HEVs can infect differ among genotypes (Supplementary Table S1). Genotypes 1 and 2 infect only primates (Meng et al., 1998), genotype 4 can infect swine and humans and genotype 3 has the widest target range. In addition, genotypes 3 and 4 are believed to be attenuated (Emerson & Purcell, 2003), which is the reason that genotype 3 and 4 strains normally have sporadic cases whereas genotype 1 usually causes epidemic infections. Considering the different ORF structures observed among genotypes, several hypotheses can be made. For example, (i) does the longer ORF1 give the ability to infect organisms other than primates, but with the sacrifice of losing the ability to cause epidemic infections? (ii) Does the longer ORF3 provide genotype 3 with the ability to infect more types of animals other than swine?

In this research, the ORF structure has been shown to be helpful in indicating a putative inter-genotype recombinant which was confirmed by phylogenetic analyses. However, this indication is limited to recombinants where recombination events change the ORF structure.

In the bootscanning analysis, the Mexican strain was indicated to be another recombinant. The phylogenetic analyses supported this hypothesis with high bootstrapping value (Fig. 2). It is predictable that HEV has a low inter-genotype recombination rate because there has only been one confirmed recombinant since the first full-length HEV sequence became available in 1991 (Tam et al., 1991). The recombinant hypothesis for the Mexican strain somehow explains the phenomenon that none of the recently fully sequenced strains belongs to genotype 2. The fact that the Mexican strain, which could be a recombinant of genotypes 1 and 3, was allocated to a separate genotype before the US strains (genotype 3) were defined as the prototype sequences of the third genotype further supports this recombinant hypothesis.

However, the statistical tests do not support that the Mexican strain clusters with genotype 3 in the region of 1857–2069 rather than genotype 1. The putative breakpoints are located in the hyper-variable region in ORF1. Recombination can be disguised by high mutational rate regions (Anderson et al., 2000), and vice versa. The long branch-lengths of the middle region in Fig. 2 indicate that it is a region with a high mutational rate. However, there is still a possibility of genotype 2 being a recombinant because both KH and SH tests are conservative. To determine whether genotype 2 is a pure genotype, more full-length sequences in genotype 2 are required in order to obtain more reliable results.

The detection of positive selections in HEV was also carried out in this study by using the CodeML program in the PAML package (Yang et al., 2005). There are two, five and four sites under positive selection across the genome for genotypes 1, 3 and 4, respectively. Both estimated ω value (marginally greater than 1) and the number of selective sites indicate that the evolution in HEV is mainly due to point mutations. Antibody against HEV and vaccine of HEV are effective in the prevention of hepatitis E (Bryan et al., 1994; Shrestha et al., 2007), which indicates neither positive selection nor recombination has generated vaccine escape mutations. Recombination is not necessary, which can be the plausible explanation as to why only a few HEV recombinants have been observed.

In conclusion, the ORF structure can be a good criterion for genotyping HEV. An inter-genotype HEV recombinant, DQ450072, was identified by the ORF scheme. The genotype 2 Mexican strain is a suspicious inter-genotype recombinant, which requires more HEV sequences to confirm.

Acknowledgements

Many thanks to David Robertson and James Eales (University of Manchester, Manchester, UK) for their constructive discussions. J. F. is supported by a University of Manchester OSS award.

References


