HHS Public Access

Author manuscript

Arch Virol. Author manuscript; available in PMC 2018 March 01.

Published in final edited form as:

Arch Virol. 2017 March ; 162(3): 645–656. doi:10.1007/s00705-016-3135-x.

The effect of phylogenetic signal reduction on hepatitis E virus Orthohepevirus A genotyping

Michael A. Purdy[1] and **Amanda Sue**[1]

[1]Division of Viral Hepatitis, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA 30333, USA

Abstract

Commonly hepatitis E virus (HEV) sequences are genotyped phylogenetically using sub genomic sequences. This paper examines this practice with *Orthohepevirus A* sequences. As the length of sequences becomes progressively shorter the number of identical sequences in an alignment tends to increase; however these sequences don't lose their genotypic identity down to 100 nucleotides in length. The best substitution models tend to become less parametrized, bootstrap support decreases and trees created from short sub genomic fragments are less likely to be isomorphic with trees from longer sub genomic fragments or complete genome sequences. However, it is still possible to correctly genotype sequences using fragments as small as 200 nucleotides. While it is possible to correctly genotype sequences with short sub genomic sequences, the estimates of evolutionary relationships between genotypes degrade to such extent that sequences below 1600 nucleotides long cannot be used to reliably study these relationships, and comparisons of trees from different sub genomic regions with little or no sequence overlap can be problematic. Subtyping may be done but requires a careful examination of the region to be used to ensure it correctly resolves subtypes.

Introduction

Hepatitis E virus (HEV) is enterically transmitted, and is the causal agent for a self-limiting acute hepatitis with mortality rates <2% among immunocompetent individuals [8]. However, the infection may become chronic in immunocompromised individuals and high mortality rates (10–30%) are seen among pregnant women [7, 21]. Initially HEV was only isolated from humans but over the past two decades HEV and HEV-like viruses have been isolated from a number of hosts. These discoveries have resulted in a recent restructuring of HEV

Corresponding author: Michael A. Purdy, phone: 404-639-2332, fax: 404-235-1892, mup3@cdc.gov.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. It 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. Use of trade names is for identification only and does not imply endorsement by the U.S. Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention.

Compliance with Ethical Standards

Conflict of Interest: Michael A. Purdy declares that he has no conflict of interest, and Amanda Sue declares she has no conflict of interest.

Ethical approval: This article does not contain any studies with animals or human participants performed by either of the authors.

taxonomy [6, 17]. Two genera are now recognized; *Piscihepevirus* containing cutthroat trout virus (CTV) and *Orthohepevirus* containing all known avian and mammalian HEVs. *Orthohepevirus* contains four species; *Orthohepevirus A*, *Orthohepevirus B*, *Orthohepevirus C* and *Orthohepevirus D*. *Orthohepevirus A* is important because the only HEV strains presently known to infect humans belong to this species. There are seven recognized genotypes in this species. Genotypes 1 and 2 infect humans anthroponotically. Genotypes 3 and 4 infect humans zoonotically, usually through consumption of infected meats that are improperly prepared. At present it is not known whether genotypes 5 or 6 can infect humans. A recent report details the infection of a transplant patient who became chronically infected with genotype 7 after consuming meat and milk from camels [10].

A phylogenetic tree is a graph showing the inferred evolutionary relationships among a set of taxa. This relationship is estimated through an algorithm that calculates the degree of similarity/dissimilarity among the taxa. The taxa occupy the external nodes (leaves) of the tree. Internal nodes represent hypothetical ancestor states among the taxa. The edges between pairs of nodes can represent evolutionary time or genetic distance between the nodes. Two trees containing the same taxa and showing identical inference of evolutionary relationships among these taxa are said to be isomorphic.

It is important to be able to genotype HEV isolates for molecular epidemiologic research, and diagnostics, for example among pregnant women. Sequencing full length genome sequences can be expensive, time consuming and difficult [2]. Because of this, many researchers opt to isolate and sequence subgenomic regions [26]. Because these subgenomic regions contain less phylogenetic signal than the complete genome there is a question as to whether subgenomic regions can be used to faithfully reproduce the evolutionary relationships inferred among complete genome sequences. Can a specific subgenomic region be used to genotype or subtype a set of taxa? Is the subgenomic tree isomorphic with the complete genome tree?

To examine these questions researchers have compared phylogenetic trees from various subgenomic regions versus the complete HEV genome to determine the best subgenomic region to genotype HEV (*Orthohepevirus A*) [1, 14, 24]. Statistical methods have also been applied to determine the best subgenomic region to use for genotyping HEV [22, 23, 25, 26]. Most of these analyses have been conducted with genotypes 1–4 from the *Orthohepevirus A* species because of the impact of these genotypes on human health. With the characterization of additional *Orthohepevirus A* genotypes and a restructuring of *Hepeviridae* taxonomy the use of subgenomic region for genotyping of HEV needs to be re-examined. The ability to accurately genotype and subtype HEV has practical applications for gaining information about the evolutionary history of HEV, and epidemiological and clinically relevant information may be associated with specific subtypes. For example, delineating an outbreak by the clustering of sequences, confirming transmission from a suspected source and associating the connection between genotype/subtype and phenotype, e.g. differences in transmissibility between susceptible species and disease severity.

As sequence length shortens, the number of segregating sites in that sequence is reduced in a linear fashion. With the loss of segregating sites, the amount of phylogenetic information

available to a phylogenetic algorithm is reduced. This paper examines how this loss of information affects genotyping of HEV with respect to the genomic region used and the length of sequence required to unambiguously genotype *Orthohepevirus A* sequences.

Materials and Methods

Sequences

Only sequences from the *Orthohepevirus A* genus were examined because few full length representatives exist for other *Orthohepevirus* species and this species is important to human health. Full length genome sequences from *Orthohepevirus A* were collected from GenBank. Heterogenotypic recombinants were removed from the collection. Open reading frame 1 (ORF1) and ORF2 sequences were concatenated. This included all nucleotides from position 1 to 5106 for ORF1 and positions 5147 to 7126 for ORF2 (reference M73218). The polyproline region was removed from these sequences because the indel (insertion/deletion) structure of this region suggests that the evolution of this region is complex and the assumptions under which phylogenetics is conducted cannot be applied to the polyproline region [15]. The bases removed were 2119 to 2358. The rabbit HEV insertion starting at position 2814 was removed from the rabbit HEV sequences. Because most of ORF3 overlaps ORF2 it was not concatenated to ORF1/ORF2 as this would duplicate this sequence information. This resulted in a sequence 6846 nucleotides long. Sequences with gaps and/or ambiguous bases were not removed. The percentage of gaps and ambiguous bases did not exceed 0.022%. Duplicate sequences were removed (Table S1). For analysis of subgenomic regions the sequences were divided into three sub regions; ORF1N, 2118 nt long (positions 1 to 2118, corresponding to positions 1 to 2118 in M73218), ORF1C, 2748 nt long (positions 2119 to 4866, corresponding to positions 2359 to 5106 in M73218) and ORF2, 1980 nt long (positions 4867 to 6846, corresponding to positions 5147 to 7126 in M73218). Sub regions were further subdivided using a sliding window strategy (Table 1). Sequence fragments longer than 1800, 2200 and 1600 nt for the ORF1N, ORF1C and ORF2 regions, respectively, were created as a single alignment from the center of each subgenomic region. Sequences were segregated by genotype. Genotype 3 was further segregated into three subtype clusters; 3A (containing subtypes a, b, c, h, I and j), 3B (containing subtypes e, f, and g) and 3R (HEV from rabbits) based on a maximum likelihood phylogenetic tree created using the concatenated ORF1/ORF2 sequence alignment (Fig. S1) [5, 18]. A total of 182 sequences were examined in each data set. The number of data sets examined in each subgenomic region is listed in Table S2.

A child sequence, or a sub-genomic region, is a sequence that contains a subset of bases from a parent sequence such that all the base positions found in the child are contained in the parent and the total number of bases in the child are less than the number of bases in the parent.

Identical Sequences

Identical sequences were identified as taxa having identical character strings in a sequence fragment data set and were found using a regular expression in a Perl script. Although the

number of identical sequences and the number of clusters formed by these identical sequences were determined only the total number of taxa found is reported here.

Phylogenetics

Phylogenetic analysis was done using MEGA-CC (ver. 7.00-beta) [9]. MEGA-CC was used to determine the best substitution model for each sequence set. The best substitution model was used to create a maximum likelihood tree with 200 bootstrap iterations unless otherwise noted [12]. Information on the best substitution model and bootstrap values were saved for each tree. Segregating sites were calculated using Tajima's test of neutrality as implemented in MEGA-CC [20]. The number of segregating sites was calculated by subgenomic region by progressively truncating each sequence set by 50 nt from the 3′-end. Pairwise distances were calculated for the concatenated ORF1/ORF2 sequences and the three subgenomic sequence data sets using the p-distance model with default parameters.

Isomorphism

TOPD-FMTS (version 3.3) was used to compare trees with the split method (a normalized Robinson-Foulds metric) [13]. The Robinson-Foulds metric calculates the distance between two unrooted trees (A and B), where the distance is calculated as the sum of the number of leaf partitions found in A but not in B and the number of partitions in B but not in A [13]. Isomorphic trees will have a normalized distance of zero and trees that do not share any partitions will have a normalized distance of one. Trees from each sequence window were compared to their parent sequence data set to determine whether any tree generated from a child sequence data set was isomorphic with its parent sequence data set. Child sequences from each subgenomic region of the same length were compared to determine whether child sequences were isomorphic with each other.

Results

The concatenated ORF1/ORF2 sequence was subdivided into three regions. The first, ORF1N, containing sequences from the 5′ end of ORF1 to the polyproline region. The polyproline region was not included because of its evolutionary history [15]. The second, ORF1C, containing sequences from the polyproline region to the 3′ end of ORF1. The third, ORF2, contained ORF2 sequences. This was done because of the computational time required to do some of the analyses with the concatenated ORF1/ORF2, and it also allowed an examination of the behavior of these three regions.

Segregating Sites

It is expected that as the length of a sequence fragment decreases the number of segregating sites will also decrease. This can be seen in the three *Orthohepevirus A* subgenomic regions, which were progressively shortened from their 3′ ends. (Fig 1). As the length of the sequence fragments are decreased the number of segregating sites decreased in a linear fashion. This indicates that the number of segregating sites in an oligonucleotide is proportional to the length of the fragment. This loss of phylogenetic information should result in sequences tending to become more similar; leading to the question of how this information loss affects intragenotypic or intergenotypic clustering of sequences.

Identical Sequences

As phylogenetic information is lost some sequences may contain identical information. The original ORF1/ORF2 sequences data set did not contain any identical sequences. This is not true of the sequences from the subgenomic regions. There were 6, 3 and 2 identical sequences in the ORF1N, ORF1C and ORF2 sequence alignments, respectively. Using a sliding window strategy (Table 1) these three regions were fragmented into sub alignments to see how the number of identical sequences changed with fragment length. Unlike the loss of segregation sites the decrease in fragment length and increase in the number of identical sequences appears to follow a power law curve (Fig. 2). Interestingly, even down to a fragment length of 100 nt identical sequences were always found among homogenotypic sequences. None of the identical sequences found in a cluster of identical sequences was from different genotypes. So even though more identical sequences are found as the fragment length decreases, this did not lead to a situation where a sequence would have been misidentified as belonging to an incorrect genotype.

Substitution Models

The best substitution model for each sequence alignment was determined in MEGA-CC using the find best model option. The best model for the parent sequences was the GTR+G +I model. As the length of the fragments being analyzed decreases the best models tend to be less highly parameterized (Table 2). The best model for alignments below 400 nt is not uniform across all fragments, and suggests that modeltest, a program that selects the best nucleotide substitution model for a set of aligned sequences, should be run to determine the best substitution model for shorter length sequence alignments.

Branch Support

To test branch support as a function of fragment length bootstrap analysis was used. Because of the computational time involved in using 1000 bootstrap replicates for longer fragment lengths, this analysis was done using 200 replicates [12]. To test the assumption that 200 replicates would be sufficient, bootstrapping was conducted on ORF2 analyzing the results of 200 versus 1000 replicates for the sub genomic 200 nt fragment and the parent ORF2 alignments. The p-value for the comparison of two samples assuming equal variances was 0.46 and 0.45 for the 200 nt fragments and the parent ORF2 alignment, respectively, indicating that the 1000 replicate bootstrapping was not significantly better that the 200 replicate bootstraps.

Bootstrap values only yield useful information for branches within a tree, and not for the tree as a whole [3, 4]; however, the mean bootstrap value for a tree is still an overall measure of reliability when comparing multiple trees. The more branches with high bootstrap values the higher the mean bootstrap and the higher the number of credible clades within a tree. Thus, a tree with a higher mean bootstrap value probably has a higher number of branches with credible bootstrap values. To test this the number of branches with minimum bootstrap values of 0.9 were examined by fragment length. This showed that as the fragment length decreased the number of branches with minimum values of at least 0.9 decreased (Fig S2). As the length of a sequence fragment decreases the mean bootstrap value for those

fragments also decrease indicating that the overall support for the branches within the tree decrease (Fig. 3 and Table S3).

To use phylogenetics to genotype a set of sequences, the bootstrap value for a clade with sequences from a single genotype is important to ensure that a genotypic clade is well supported. To examine branch support two types of branches were examined. The first was the parent branch of a genotypic clade. The second was the branch connecting different genotypic clades. For the parent ORF1N, ORF1C and ORF2 maximum likelihood trees the mean bootstrap values for genotypic clades as 0.98±0.036, 1.0 and 0.99±0.018, respectively. The bootstrap for the connecting branches was 0.98±0.026, 0.96±0.085 and 0.93±0.054, respectively. These data were also examined for well-formed genotypic clades from the 200 nt fragment trees for the three subgenomic regions. For ORF1N the mean bootstrap value for the genotypic clades was 0.80±0.22, for the connecting branches the mean bootstrap value was 0.44±0.26 and the p-value between these two sets of data was <0.001. For ORF1C these values were 0.77±0.23 and 0.41±0.21, respectively ($p < 0.001$). For ORF2 these values were 0.72±0.28 and 0.41±0.28, respectively ($p < 0.001$). This indicates that the support for the branches connecting genotypic clades decreases more rapidly than the support for the genotypic clades. This suggests that as fragment length decreases the information about the evolution relationship between genotypic clades deteriorates more quickly than the information about the sequences contained within a genotypic clade (see also Genotyping with Sub Fragments, below).

A similar trend is seen when comparing the 1600 nt trees to 200 nt trees (Fig. S3). Support for genotypic clades is higher than for the branches connecting the genotypic clades and the 1600 nt fragments have better support than the 200 nt fragments. Outliers are seen for the ORF1N and ORF1C 1600 nt fragments (Fig. S3). With respect to the genotypic outliers seen for the ORF1N 1600 nt fragments, these outliers are all due to the genotype 6 sequences. With respect to the connecting branch outliers seen for the ORF1C 1600 nt fragments, these outliers are all due to the branch connecting genotypes 4 and 7. These types of outliers are not seen with ORF2.

Isomorphic trees

While the main thrust of sub genomic fragment genotyping has been to genotype and subtype HEV sequences [1, 14, 26] the statistical methods are based on finding a specific tree from a sub genomic region that is most nearly isomorphic with the parent tree [22, 23, 26]. To determine which sub genomic regions generate trees that are the most isomorphic with their parent tree, trees from sub genomic fragments were compared with their parent tree using the Robinson-Foulds metric [13]. Fig. 4 shows that each sub genomic region has a range of values with some fragments being more nearly isomorphic than others (closer to a value of zero), but overall as the fragment length becomes shorter the trees created from these fragments are less likely to be isomorphic with their parent. This indicates that fragment length is more important than the region chosen for isomorphism. These data also show that even the removal of as little as 18 nt (Fig. 4; ORF1N, 2100 nt) results in a tree that may no longer be isomorphic with its parent tree.

Next the trees within a sub genomic length were compared to each other. Because a sliding window was used to create the fragments, there is overlap among some of the sequences. Comparing the tree from a sliding window allows for comparison of the effect of sequence overlap. Fig. S4 shows the results for trees created from 1000 nt long sequences. As the amount of overlap between the child alignments increases the trees tend to become more nearly isomorphic. This supports the results from Fig. 4 where the longer the common length between two alignments the more nearly isomorphic the trees. These results are similar for other fragment lengths (data not shown). In addition the 1000 nt overlaps in this figure compare the structural similarity between the child trees to the parent tree. Child trees with no overlap tend to be no more isomorphic than they are with the parent tree, and in many cases are less isomorphic. Additionally, as the amount of overlap increases among sub fragment trees, the more isomorphic they become with each other than with their parent. This indicates that not only do trees tend to be more nearly isomorphic as the length of the fragment increases, but they tend to be more nearly isomorphic as the fraction of shared sequence increases with respect to the total number of bases being compared.

Genotyping with Sub Fragments

To this point isomorphism has been defined as tip by tip comparison of tree structures. However, for the purposes of genotyping trees may be considered to be isomorphic if genotype specific sequences cluster into genotypic clades, and these clades are isomorphic to the genotypic clades in the parent tree. In other words, the comparison is not tip by tip, but genotypic clade by genotypic clade. All genotype sequences are clustered into genotype specific clades, and the branching among genotype clades is the same as seen in the parent tree. The parent trees for ORF1N and ORF1C are genotypically isomorphic with the ORF1/ORF2 tree, but the parent ORF2 tree is not genotypically isomorphic. Examination of trees from sub genomic alignments versus their parent sequence shows that as the fragment lengths decrease the corresponding trees are less likely to be genotypically isomorphic with their parent (Table 3, column I). As the fraction of isomorphic trees decreases there is an increase in the number of anisomorphic trees (Table 3, column A). These are trees in which all genotype sequences are clustered into genotype specific clades, but the branching relationship between genotype clades is not maintained. This may be due to the more rapid decrease in support for connections between genotypic clades versus the support for genotypic clades (Fig. S2). All sequences genotype correctly in the anisomorphic trees. These data indicate there is a sub genomic fragment of 200 nt in each of these three sub regions in which sequences can be genotyped correctly. There is also an increase in the number of trees in which not all sequences will genotype correctly (Table 3, column D). This disruption involves a single genotype until the fragment length is 200 nt where multiple genotypes may be disrupted (Table 3D).

ORF1N—This region is more likely to exhibit disruptions to genotypic clades. The disruption of genotypes seen in this region for fragments >= 400 nt in length is due to genotype 5 merging with genotype 6 (Fig. S2; ORF1N, 1600 nt). Sequences from genotypes 1–4 and 7 cluster correctly into their respective genotypes.

ORF1C—This sub region is the most stable with respect to genotyping with disruption of some genotypes only occurring with the 200 nt fragments.

ORF2—The sub genomic sequences in this region tend to be more isomorphic with the parent ORF2 parent alignment than in the other two regions for fragments >=600 nt in length.

Rabbit HEV—The segregation of genotype 3 sequences into three clades allows for an examination of whether HEV from rabbits should be a genotype unto itself (Fig S1). Currently rabbit HEV is classified as genotype 3. Table 3 shows there is sequence mixing among genotype 3 clades (Table 3, column 3). For the ORF1C region this is due primarily to sequence AF455784 (This is HEV genotype 3 isolated from a piglet experimentally infected with virus from a human stool collected during an outbreak in Osh, Kyrgyzstan between 1987–1989.) for fragments greater than 400 nt. This is not the case with ORF1N or ORF2. The intermixing of the genotype 3 clades occurs before disruption of genotype 3 sequences is seen among genotypes (Table 3, column 3 vs. column G) except for ORF1N where genotype 5 merges with genotype 6. The reason for this is seen when p distance values for maximum and minimum inter-clade distances are compared. The minimum p-distance between the 3A/3B clade and rabbit HEV (R) is smaller than the maximum within clade p-distance for 3A/3B and R, respectively, in the ORF1N and ORF2 sub regions (Table 4). This indicates there is overlap in the range of distances between some of the 3A/3B and R sequences, and the rabbit HEV sequences belong to genotype 3. For ORF1C the minimum p-distance between the 3A/3B clade and rabbit HEV (R) is smaller than the maximum distance within 3A/3B but not within the R clade. Still this shows that there is overlap in the range of distances between some rabbit HEV sequences and 3A/3B sequences. An examination of minimum p-distances between all other clades shows that these distances are larger than the corresponding within clade maximum distances, indicating there is no overlap in the range of distances between any other clade (genotype).

Subtyping

There is controversy surrounding HEV subtyping. One example is the recent question over whether sequences found in France belonged to subtype 3i or 3c [11, 19]. An attempt has been made to delineate HEV subtypes [19]; however, an examination of Smith et al., 2016 will show that some genotype 3 and 4 subtypes are represented by a single sequence and some sequences could not be unambiguously subtyped. This raises the question of whether an analysis of subtypes can be done using sequences that have not been classified or whether it is meaningful to do such an analysis with only sequences that have been assigned. It was decided to do an analysis of subtyping with genotype 1 as a proxy for other genotypes to see if a lower limit could be established. But even using genotype 1 doesn’t get around the problem of unassigned subtypes as FJ547024 has not been assigned to a specific genotype. Another way to estimate how well subtyping can be done is to look at the behavior of the genotype 3 clades 3A, 3B and 3R to see how well the sequences within each clade maintain their clade integrity.

Figure 5 shows the results of these analyses. The gray blocks in this figure show the beginning of windows that could not be used to unambiguously subtype genotype 1 or maintain the clade structure of genotype 3. The beginning of these regions is shown because of the ambiguity in resolving subtype or clade structure. For example in the ORF1N region using an alignment of 1000 nt genotype 1 subtypes cannot be completely resolved using an alignment starting at position 401 but if the start of the alignment is shifted 50 bases upstream or downstream the subtypes can be completely resolved phylogenetically. This figure shows that while the regions in which neither genotype 1 subtypes nor genotype 3 clades can be resolved overlap, there are regions where they don't overlap. Additionally the punctuated regions where the start of alignments do not result in resolution of genotype 1 subtypes or genotype 3 clades suggest that the bases contributing to the loss of resolution are not uniformly distributed across the genome or concentrated to specific regions. This suggests that the best strategy for subtyping is to create an alignment of as many full length sequences across the ORF1N, ORF1C or ORF2 where subtyping is to be done and create a sub genomic alignment of the specific region to be used and determine how similar the sub genomic tree is to the parent tree. Alignments equal to or greater than 1400 nt appear to resolve subtypes correctly.

The results for the ORF1N region cannot be compared with either ORF1C or ORF2 because sequence D11093, initially identified as subtype 1b, was found to be a 1b/1a recombinant. The crossover points from 1b to 1a and back are about 860 and 2020 nt based on a bootscan analysis using Simplot (ver. 3.5.1)(data not shown). Additionally FJ547024 appears to be recombinant with subtype 1a in the same region, but this could not be confirmed using Simplot. Because of the recombinant nature of these sequences they were removed from the analysis.

Discussion

Genotyping and subtyping of HEV sequences is important for the discovery of evolutionary, epidemiological and clinical significant information. The most valuable information will come from complete genome sequences, but it is not always possible to fully sequence a genome from a specimen. For this reason it is important to be able to determine whether the information obtained from a sub genomic region contains relevant and significant information, and reflects results that are representative of the information obtained from genomic sequences. Several statistical analyses have attempted to delineate the best subgenomic region for genotyping and subtyping [22, 23, 25]. These analyses use mathematical models to fine the best region for analysis but do not describe the features that result in the determination of a specific region. In addition each analysis has described a different region. This study attempts to examine this question from the standpoint of the factors that may affect the use of a subgenomic region and determine how these factors change from region to region along the genome. These factors include segregating sites, polytomes, bootstrap support, substitution models and isomorphism. We also examined whether different subgenomic regions can be compared to each other as multiple subgenomic regions are used for genotyping and subtyping of HEV.

One problem with using sub genomic fragments for analysis of genotypes and subtypes is the loss of segregating sites (Fig. 1). One effect of using sub genomic fragments is that such fragments are more likely to become more similar to each other with the loss of phylogenetic information. This can be seen with the increase in the number of identical sequences present in an alignment as the size of the fragment decreases (Fig. 2). However, as fragment length decreases and the number of identical sequences rises, the genotypic identity of these sequences is not lost and their genotypic identity is maintained down to fragments as short as 100 nt (Fig. 2).

Another effect of sequence reduction is that the best substitution model as determined by modeltest tends to become less parametrized than the best model for the parent sequence (Table 2). However, this change is not uniform across all fragments of the same size and each fragment needs to be analyzed to find the best substitution model.

As fragment length is reduced the number of well supported branches in the resulting trees as estimated by bootstrapping tends to decrease (Fig. 3 and S1). This decrease occurs more rapidly among the branches connecting genotypic clades than for the genotypic clades themselves (Fig. 3 and S2). If a researcher is careful they can find fragments as short as 200 nt, which will allow unambiguous genotyping of sequences, but the evolutionary relationship between the genotypes will be lost (Fig. S3, Table 3 and Table S3).

While sub genomic fragments can be found that are statistically better for genotyping HEV than similar sized fragments, this study indicates that the length of a subgenomic sequence and the fraction of bases in common between two subgenomic regions are more important variables for genotyping HEV sequences with subgenomic fragments than the region chosen.

As subgenomic fragments become shorter it is more likely that some of these sequences will have identical sequences and form a polytomy. As soon as a polytomy forms the subgenomic region containing the tree created from these child sequences cannot be isomorphic with their parent.

An evaluation of subtyping is complicated by the fact that there is differing genetic diversity within genotypes, in some genotypes genetic distinctiveness blurs into a continuum of variability and it is difficult to reliably cluster sequences into subtypes even with complete genome sequences [16, 19]. Because of this, distance based and phylogenetic methods cannot always provide clear criteria for demarcation of sequences into subtypes [19]. Using genotype 1 subtypes and genotype 3 subclades an attempt was made to determine the effectiveness of subtyping using subgenomic regions. This analysis suggests that the diversity in these sequences that interferes with resolving subtypes in subgenomic regions is neither uniformly distributed nor confined to specific subregions along the genome. Additionally, a shift in the position of a subgenomic alignment as small as 50 bases from one position to another may result in one region that will resolve subtypes while the small shift may result in an alignment that cannot resolve subtypes unambiguously even in subgenomic alignments as long as 1000 nt (Fig 5).

One limitation with this study is the low number of sequences available for genotypes 2, 5, 6 and 7. The true nucleotide diversity within these genotypes is unknown, and the behavior of these genotypes cannot be fully examined. This means that most of the results for genotypic isomorphism come from genotypes 1, 3 and 4. Additionally, data presented here shows that genotype 5 merges with genotype 6 in ORF1N in fragments equal to or shorter than 1000 nt. This may indicate that genotypes 5 and 6 are not separate genotypes, but more sequences will need to be isolated from these genotypes before such a determination can be made.

The present study suggests that researchers can reliably genotype HEV *Orthohepevirus A* sequences using sub genomic fragments, even down to 200 nt, if they realize the pros and cons of using sub genomic alignments. These results also suggest that the evolutionary relationships between *Orthohepevirus A* genotypes cannot be reliably correlated between alignments that do not share a high fraction of sites in common. The best option would be to use full length genome sequences or complete gene sequences depending on what is being researched. If complete genome sequencing is not an option then the longer the sub genomic region being analyzed the more reliable the genotyping and the evolutionary relationships between the genotypes. On the other hand subtyping is more complex and may require comparisons between parent ORF1N, ORF1C and ORF2 sequences and the subgenomic region to be used to determine the best region for subtyping.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This study was funded completely with internal departmental funds.

References

1. Arankalle VA, Paranjape S, Emerson SU, Purcell RH, Walimbe AM. Phylogenetic analysis of hepatitis E virus isolates from India (1976–1993). Journal of General Virology. 1999; 80:1691–1700. [PubMed: 10423137]
2. Bartholomeusz A, Schaefer S. Hepatitis B virus genotypes: comparison of genotyping methods. Rev Med Virol. 2004; 14:3–16. [PubMed: 14716688]
3. Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Science, USA. 1996; 93:7085–7090.
4. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution. 1985; 39:783–791. [PubMed: 28561359]
5. Hewitt PE, Ijaz S, Brailsford SR, Brett R, Dicks S, Haywood B, Kennedy ITR, Kitchen A, Patel P, Poh J, Russell K, Tettmar KI, Tossell J, Ushiro-Lumb I, Tedder RS. Hepatitis E virus in blood components: a prevalence and transmission study in southeast England. Lancet. 2014; 394:1766–1773.
6. Johne R, Dremsek P, Reetz J, Heckel G, Hess M, Ulrich RG. Hepeviridae: An expanding family of vertebrate viruses. Infect Genet Evol. 2014; 27:212–229. [PubMed: 25050488]
7. Kamar N, Rostaing L, Izopet J. Hepatitis E Virus Infection in Immunosuppressed Patients: Natural History and Therapy. Semin Liver Dis. 2013; 33:62–70. [PubMed: 23564390]
8. Khuroo MS. Discovery of hepatitis E: The epidemic non-A, non-B hepatitis 30 years down the memory lane. Virus Res. 2011; 161:3–14. [PubMed: 21320558]

9. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics. 2012; 28:2685–2686. [PubMed: 22923298]
10. Lee G-H, Tan B-H, Teo EC-Y, Lim S-G, Dan Y-Y, Wee A, Aw PPK, Zhu Y, Hibberd ML, Tan C-K, Purdy MA, Teo C-G. Chronic Infection With Camelid Hepatitis E Virus in a Liver-transplant Recipient Who Regularly Consumes Camel Meat and Milk. Gastroenterology. 2016; 150:355–357. [PubMed: 26551551]
11. Moal V, Gerolami R, Ferretti A, Purgus R, Devichi P, Burteya S, Colson P. Hepatitis E Virus of Subtype 3i in Chronically Infected Kidney Transplant Recipients in Southeastern France. J Clin Micro. 2014; 52:3967–9372.
12. Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. How many bootstrap replicates are necessary? J Comput Biol. 2010; 17:337–354. [PubMed: 20377449]
13. Puigbo P, Garcia-Vallve S, McInerney JO. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics. 2007; 23:1556–1558. [PubMed: 17459965]
14. Schlauder GG, Mushahwar IK. Genetic heterogeneity of hepatitis E virus. J Med Virol. 2001; 65:282–292. [PubMed: 11536234]
15. Smith DB, Vanek J, Ramalingam S, Johannnessen I, Templeton K, Simmonds P. Evolution of the Hepatitis E virus hypervariable region. J Gen Virol. 2012; 93:2408–2418. [PubMed: 22837418]
16. Smith DB, Purdy MA, Simmonds P. Genetic variability and the classification of hepatitis E virus. J Virol. 2013; 87:4161–4169. [PubMed: 23388713]
17. Smith DB, Simmonds P, Jameel S, Harrison TJ, Meng X-J, Okamoto H, Van der Poel WHM, Purdy MA. Consensus Proposals for Classification of the Family Hepeviridae. J Gen Virol. 2014; 95:2223–2232. [PubMed: 24989172]
18. Smith DB, Ijaz S, Tedder R, Hogema B, Zaaijer H, Izopet J, Bradley-Stewart A, Gunson R, Harvala H, Kokki I, Simmonds P. Variability and pathogenicity of Hepatitis E virus genotype 3 variants. J Gen Virol. 2015; 96:3255–3264. [PubMed: 26282123]
19. Smith DB, Simmonds P, Izopet J, Oliveira-Filho EF, Ulrich RG, Johne R, Koenig M, Jameel S, Harrison TJ, Meng XJ, Okamoto H, Van dP WH, Purdy MA. Proposed reference sequences for hepatitis E virus subtypes. J Gen Virol. 2016; 97:537–542. [PubMed: 26743685]
20. Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics. 1989; 123:585–595. [PubMed: 2513255]
21. Teshale EH, Hu DJ, Holmberg SD. The Two Faces of Hepatitis E Virus. Clin Infect Dis. 2010; 51:328–334. [PubMed: 20572761]
22. Wang S, Luo X, Wei W, Y Z, Dou Y, X C. Calculation of Evolutionary Correlation between Individual Genes and Full-Length Genome: A Method Useful for Choosing Phylogenetic Markers for Molecular Epidemiology. PLoS ONE. 2013; 8:e81106. [PubMed: 24312527]
23. Wang S, Wei W, Luo X, Cai X. Genome-Wide Comparisons of Phylogenetic Similarities between Partial Genomic Regions and the Full-Length Genome in Hepatitis E Virus Genotyping. PLoS ONE. 2014; 9:e115785. [PubMed: 25542033]
24. Wang Y, Ling R, Erker JC, Zhang H, Li H, Desai S, Mushahwar IK, Harrison TJ. A divergent genotype of hepatitis E virus in Chinese patients with acute hepatitis. J Gen Virol. 1999; 80:169–177. [PubMed: 9934699]
25. Xun PC, Chen F, Dong C, Qian GH, Lai DJ, Meng J. A score method for comparison of partial genomic regions in their representatives of full-length genome of hepatitis E virus for genotyping. Intervirology. 2007; 50:328–335. [PubMed: 17687190]
26. Zhai L, Dai X, Meng J. Hepatitis E virus genotyping based on full-length genome and partial genomic regions. Virus Res. 2006; 120:57–69. [PubMed: 16472882]

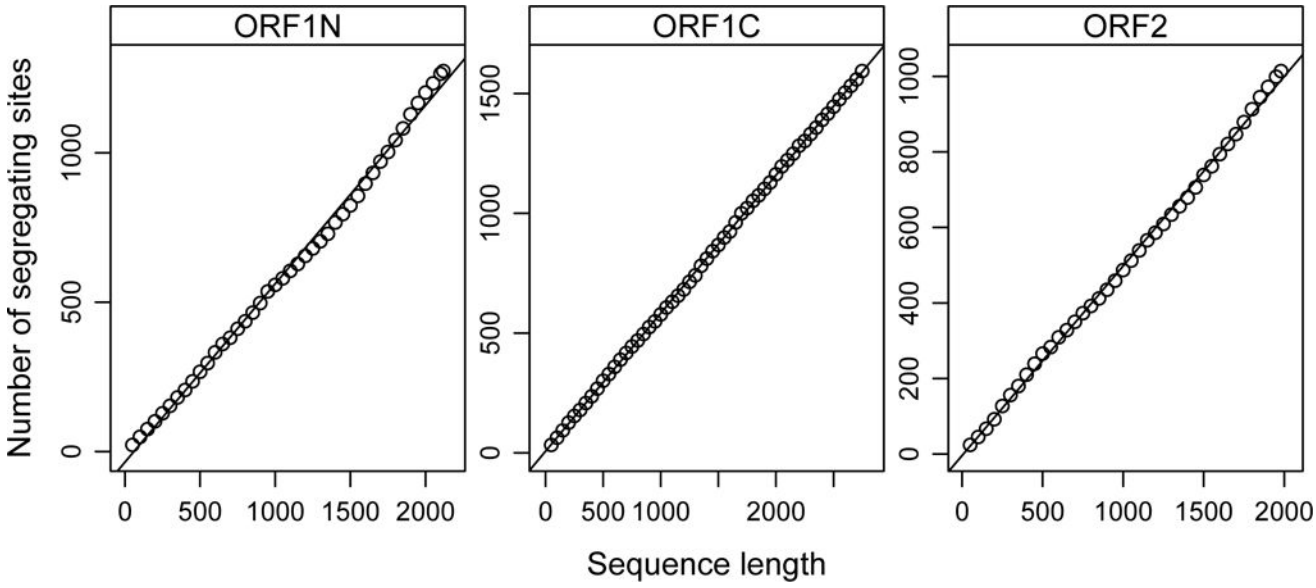

Fig. 1. Number of segregating sites by sub region as sequence length is decreased

Parent sequences for each sub region were progressive shortened from the 3′-end of the sequence. The black line through the data points is the regression line through the data.

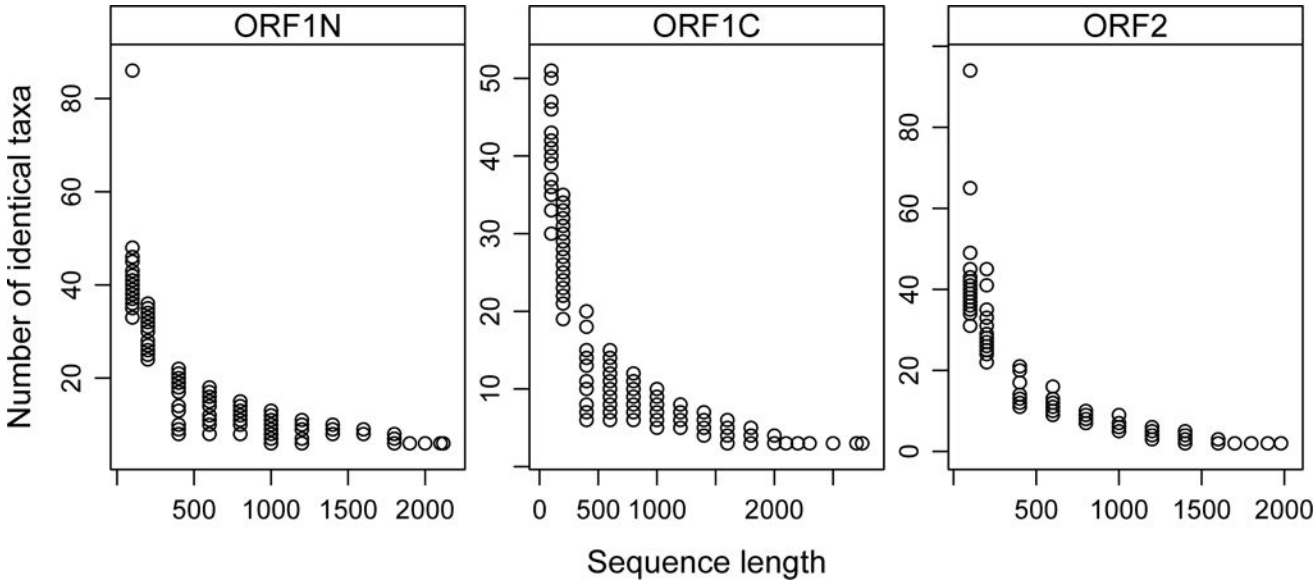

Fig. 2.
Change in the number of identical sequences with respect to sequence length by sub region

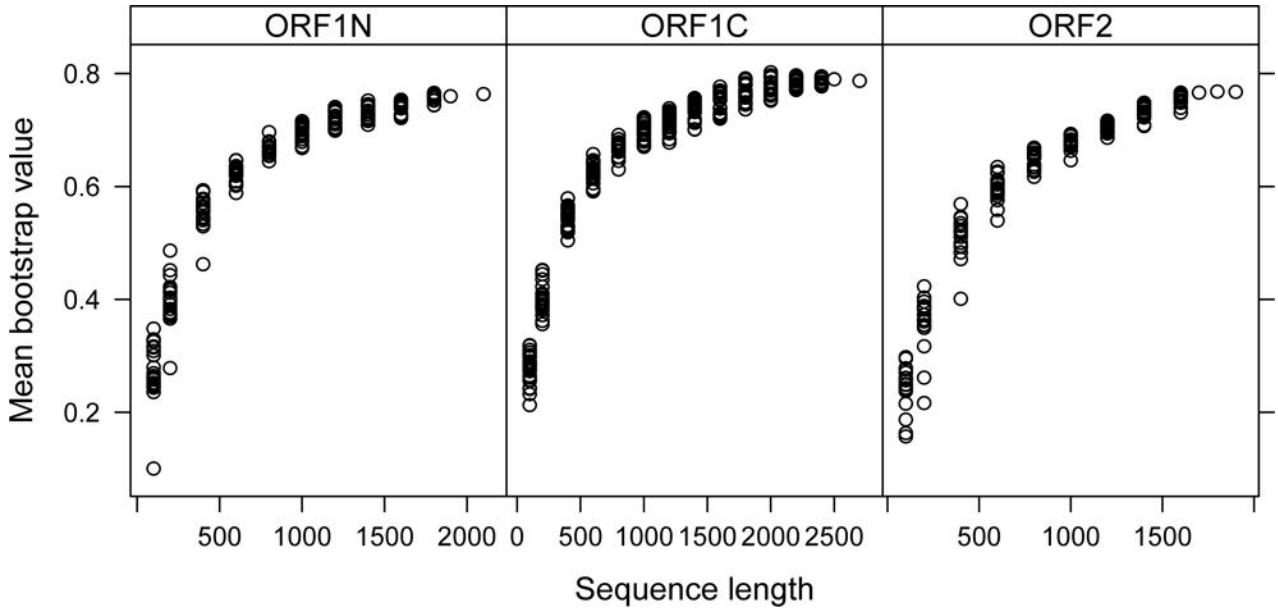

Fig. 3.

Change in the mean of bootstrap support for all branches in a tree with respect to sequence length by sub region

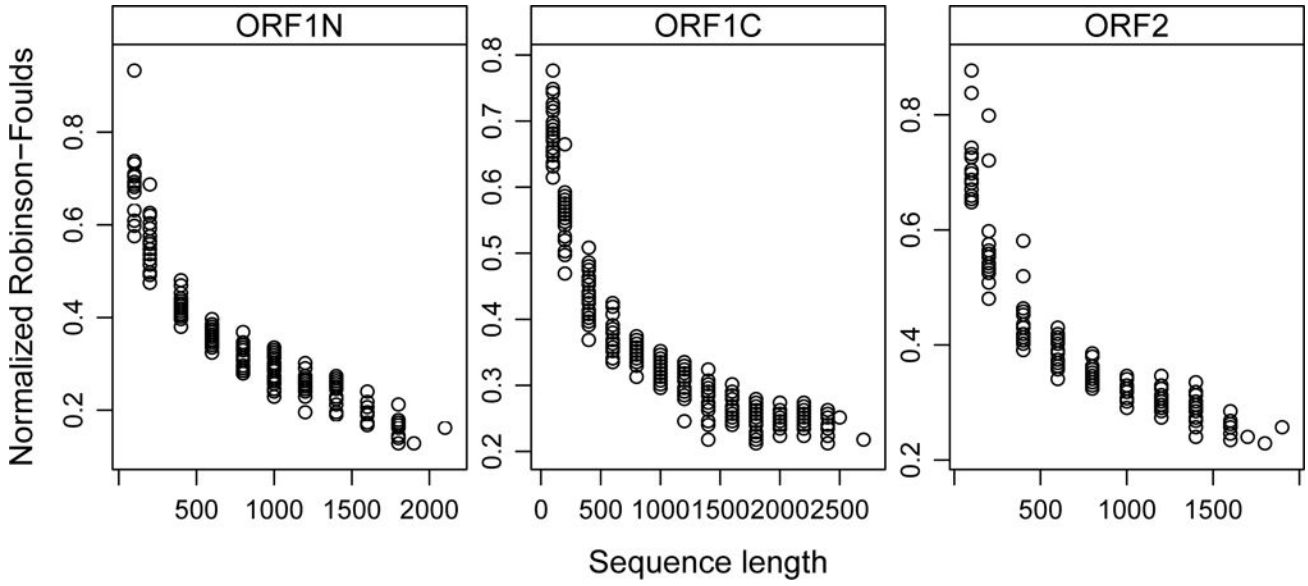

Fig. 4.

Change in the normalized Robinson-Foulds metric with respect to sequence length by sub region

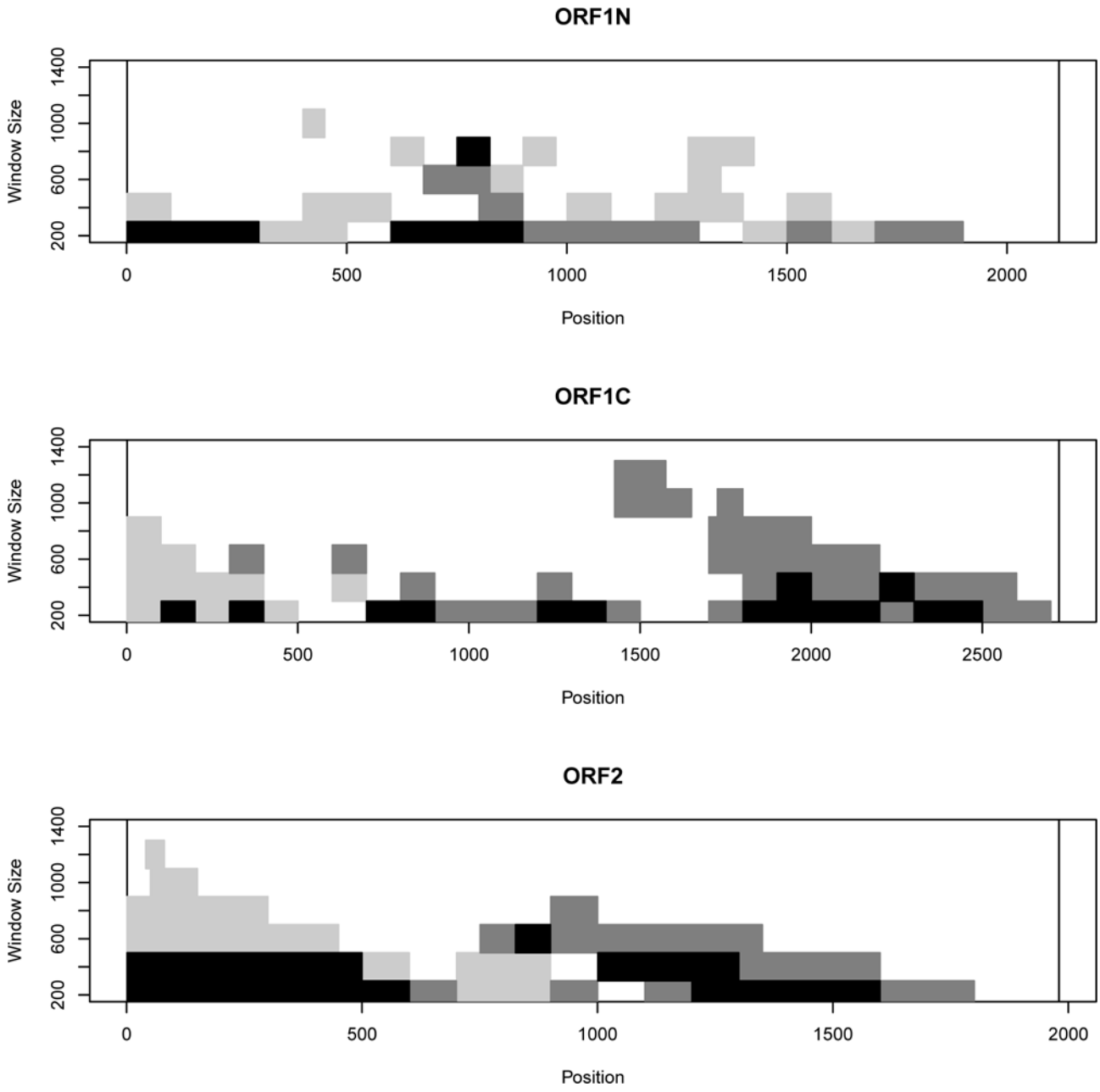

Fig. 5.

Starting positions of sliding window alignments in which genotype 1 subtype and/or genotype 3 clade structure is not conserved in the alignment. Light gray – genotype 1 subtypes, dark gray – genotype 3 clades, black – both genotypes 1 and 3. The vertical lines demark the start and stop of the parent alignment. See the text for more information.

Table 1

Step size (nt) used for each sliding window by genomic sub region.

	Window Size (nt)												
	100	**200**	**400**	**600**	**800**	**1000**	**1200**	**1400**	**1600**	**1800**	**2000**	**2200**	**2400**
ORF1N	100	100	100	75	75	50	50	50	40	30	ND[a]	ND	ND
ORF1C	100	100	100	100	100	75	75	75	50	40	40	30	30
ORF2	100	100	100	75	75	50	40	40	30	ND	ND	ND	ND

[a]ND – not done

Table 2

Best nucleotide substitution model by sequence fragment length.

	Model[a]	100	200	400	600	800	1000	1200	1400	1600
ORF1N	GTR+G+I		0.55	0.89	1.00	1.00	1.00	1.00	1.00	1.00
	GTR+G	0.10	0.05	0.11						
	HKY+G+I	0.24	0.05							
	HKY+G		0.05							
	K2+G+I	0.38	0.05							
	K2+G	0.05								
	K2+I	0.05								
	T92+G	0.10								
	TN93+G+I	0.10	0.20							
	TN93+G		0.05							
ORF1C	GTR+G+I	0.07	0.38	0.88	1.00	1.00	1.00	1.00	1.00	1.00
	GTR+G			0.08						
	HKY+G+I	0.07	0.08							
	HKY+G	0.07	0.08							
	K2+G+I	0.37	0.23							
	K2+G	0.04								
	T92+G+I	0.07	0.04							
	T92+G	0.04								
	TN93+G+I	0.19	0.19	0.04						
	TN93+G	0.07								
ORF2	GTR+G+I	0.05	0.50	0.94	0.95	1.00	1.00	1.00	0.93	1.00
	GTR+G			0.06						
	GTR+I				0.05				0.07	
	HKY+G+I	0.20	0.22							
	K2+G+I	0.50	0.11							
	K2+G	0.05	0.11							

	Model[a]	100	200	400	600	800	1000	1200	1400	1600
	T92+G	0.05								
	TN93+G+I	0.15	0.06							

[a]Cells list the fraction of sequences having a specified substitution model. Empty cells indicate that the model was not selected for any sequence having the specified length. The best substitution model for all sequence lengths greater than 1600 nt was the GTR+G+I model. GTR – general time reversal, HKY – Hasegawa-Kishino-Yano, K2 – Kimura 2-parameter, T92 – Tamura 3-parameter, T93 – Tamura-Nei, G - gamma rate categories, I – invariant sites.

Table 3

Comparison of trees by sequence fragment length.

	L[a]	I	A	D	3	1
ORF1N	100		0.10	0.90	0.95	0.95
	200		0.25	0.75	0.60	0.65
	400	0.22	0.22	0.56	0.11	0.28
	600	0.29	0.42	0.29	0.10	0.24
	800	0.33	0.39	0.28		0.05
	1000	0.43	0.35	0.22		
	1200	0.50	0.39	0.11		
	1400	0.73	0.27			
	1600	1.00				
ORF1C	100		0.22	0.78	0.93	0.78
	200	0.12	0.68	0.23	0.62	0.58
	400	0.25	0.75		0.33	0.29
	600	0.23	0.77		0.32	0.09
	800	0.35	0.65		0.15	0.02
	1000	0.46	0.54		0.17	
	1200	0.57	0.43		0.10	
	1400	0.78	0.22			
	1600	1.00				
ORF2	100			1.00	0.95	0.89
	200		0.37	0.63	0.84	0.67
	400		0.69	0.31	0.69	0.69
	600	0.26	0.79	0.05	0.42	0.39
	800	0.69	0.31[b]		0.13	0.25
	1000	0.65	0.35		0.10	0.10
	1200	1.00				0.05
	1400	1.00				

Arch Virol. Author manuscript; available in PMC 2018 March 01.

	L[a]	I	A	D	3	1
	1600	1.00				

[a]L – sequence length (nt), I – fraction of trees with genotypic isomorphism, A – fraction of anisomorphic trees, D – fraction of trees with disrupted genotypes, 3 – fraction of trees with anisomorphism within genotype 3 clades (3A, 3B and 3R), 1 – fraction of trees with disrupted genotype 1 subtypes. Empty cells indicate that the altered tree structure was not found among any tree generated from the specified length.

[b]~40% of these trees were anisomorphic with respect to the parent ORF2 tree, but were isomorphic with respect to the parent ORF1N and ORF1C trees.

Table 4

Nucleotide p-distance matrix by clade.

ORF1N[a]	1	2	3	4	5	6	7	R
1	0.132	0.273	0.300	0.290	0.281	0.284	0.289	0.306
2	0.253	0.000	0.289	0.285	0.281	0.275	0.283	0.302
3	0.264	0.260	0.205	0.282	0.293	0.281	0.269	0.242
4	0.254	0.267	0.245	0.188	0.256	0.244	0.277	0.297
5	0.268	0.281	0.266	0.218	0.000	0.221	0.278	0.289
6	0.264	0.274	0.260	0.209	0.218	0.194	0.269	0.278
7	0.269	0.280	0.248	0.253	0.265	0.269	0.146	0.282
R	0.267	0.272	**0.200**[b]	0.256	0.270	0.262	0.254	0.217
ORF1C	1	2	3	4	5	6	7	R
1	0.118	0.255	0.273	0.268	0.268	0.266	0.259	0.273
2	0.239	0.000	0.266	0.271	0.261	0.267	0.259	0.279
3	0.236	0.251	0.204	0.261	0.255	0.263	0.248	0.230
4	0.242	0.251	0.229	0.175	0.239	0.247	0.255	0.269
5	0.257	0.261	0.235	0.221	0.000	0.219	0.250	0.267
6	0.249	0.266	0.242	0.220	0.216	0.195	0.258	0.280
7	0.242	0.254	0.223	0.236	0.245	0.249	0.135	0.253
R	0.252	0.257	**0.191**	0.236	0.255	0.251	0.223	0.187
ORF2	1	2	3	4	5	6	7	R
1	0.106	0.195	0.220	0.217	0.218	0.225	0.212	0.225
2	0.186	0.000	0.223	0.220	0.210	0.237	0.214	0.231
3	0.186	0.201	0.173	0.219	0.214	0.225	0.221	0.198
4	0.181	0.202	0.175	0.158	0.201	0.214	0.213	0.225
5	0.198	0.210	0.194	0.181	0.000	0.178	0.208	0.219
6	0.210	0.226	0.201	0.182	0.171	0.167	0.229	0.225
7	0.194	0.207	0.195	0.184	0.206	0.215	0.115	0.219
R	0.201	0.204	**0.160**	0.184	0.201	0.211	0.206	0.171

[a] For each sequence region the boxed diagonal is the maximum p-distance within a clade. The upper right of the matrix contains the maximum p-distances between clades. The lower left of the matrix contains the minimum p-distances between clades. Matrix column and row headers indicate genotype except for 3 (3A and 3B sequences) and R (3R, HEV from rabbits).

[b] The bold value in the lower left of the matrices shows the minimum distance between 3 (3A and 3B) and R (HEV from rabbits).