Genotype-specific evolution of hepatitis E virus

Adam B. Brayne,[a] Bethany L. Dearlove,[a] James S. Lester,[a] Sergei L. Kosakovsky Pond,[b] Simon D. W. Frost[a] #

University of Cambridge[a]; Temple University[b]

Running Head: Genotype-specific evolution of hepatitis E virus

#Address correspondence to Simon D.W. Frost, sdf22@cam.ac.uk

Abstract: 151 words. Other text: 5410 words

Abstract

Hepatitis E virus (HEV) is the most common cause of acute viral hepatitis globally. HEV comprises four genotypes with different geographic distributions and host ranges. We utilise this natural case-control study for investigating the evolution of zoonotic viruses compared to single host viruses, using 244 near full length HEV genomes. Genome wide estimates of dN/dS located a region of overlapping reading frames, which is subject to positive selection in genotypes 3 and 4. The open reading frames (ORFs) involved have functions related to host-pathogen interaction, so genotype specific evolution of these regions may reflect their fitness. Bayesian inference of evolutionary rates shows genotypes 3 and 4 have significantly elevated rates relative to genotypes 1 across all ORFs. Reconstructing phylogenies of zoonotic genotypes demonstrates significant intermingling of isolates between hosts. We speculate that the genotype specific differences may result from cyclical adaptation to different hosts in genotypes 3 and 4.

Importance:

Hepatitis E virus (HEV) is increasingly recognised as a pathogen which affects both the developing, and the developed world. While most often clinically mild, HEV can be severe or fatal in certain demographics, such as expectant mothers. Like many other viral pathogens, HEV has been grouped into several distinct genotypes. We show that most of the HEV genome is evolutionarily constrained. One locus of positive selection is unusual as it encodes two distinct protein products. We are the

first to detect positive selection in this overlap region. Genotype 1, which only infects humans, appears to be evolving differently to genotypes 3 and 4, which infect multiple species, possibly because genotypes 3 and 4 are unable to achieve the same fitness due to repeated host jumps.

Introduction

Hepatitis E virus (HEV) is a non-enveloped, single stranded, positive sense RNA virus, which infects around 20 million people globally each year (1). It causes large propagated epidemics of acute hepatitis in Asia and Africa, and low level, sporadic food-associated infections in the developed world (2, 3). Pathogenicity varies from acute liver failure and up to 20% mortality in some sub-populations (for example in pregnant women), to apparently asymptomatic infections in others (4). Acquired via the fecal-oral route, HEV is associated with poor hygiene and living conditions. It can also be acquired by eating contaminated food, including infected artiodactyls (swine, deer and boar) and shellfish (4–6).

Mammalian HEV exists in four internationally recognised genotypes (7). Genotyping is based on nucleotide divergence of the capsid open reading frame (8), and whole genome phylogenetic analysis (9). Genotypes differ at epidemiological (distribution, hosts) and virological (pathogenicity, translation mechanisms) levels.

In terms of epidemiology, there is a striking global distribution of autochthonous genotypes whose origins are obscure (10): Genotype 1 is found in Asia and North Africa; genotype 2 in Mexico and Southern Africa; genotype 3 in North and South America, Europe and Asia; and genotype 4 almost exclusively in Japan and China. All four genotypes infect humans, but only genotypes 3 and 4 infect other animals such as artiodactyls. In the developed world infections are sporadic and the genotype is usually the same as that in the native swine population, suggesting zoonotic transmission by food or contact (2). Most likely this involves the consumption of undercooked pork. In contrast in developing countries infections can be epidemic as well as sporadic, with human and swine strains most often different. A recombinant vaccine against HEV exists, based on its capsid protein, which has passed phase III trials (11, 12). The vaccine is based on genotype 1 strains, and appears to provide cross protection against at least genotype 3 (12).

Pathogenicity and molecular mechanisms vary between genotypes. In developed countries clinical disease is rare, and seroprevalence vastly outweighs documented incidence (13–15). In developing countries, the clinical presentation of HEV infection tends to be more symptomatic than in the developed world. Symptoms are shared with many viral illnesses and include fever, gastro-intestinal upset and malaise, and liver function tests may be deranged (15). The natural history also varies by demographic, with a strikingly high mortality amongst pregnant women in the developing world (10-25%) and also more disease in children compared to the developing world where it is elderly men that are most often symptomatic (15, 16). Primate models suggest these differences in pathogenicity are associated with the genotypes, as genotypes 3 and 4 produce less clinical disease in comparison to genotypes 1 and 2 in rhesus monkeys (17). There are few known differences in molecular mechanisms between genotypes, however genotype 4 viruses do have a distinct mechanism for the translation of open reading frame 3 (ORF3), due to a frame-disrupting single nucleotide insertion (18).

HEV has a c. ~7200 nucleotide genome comprising three partially overlapping open reading frames. ORF1 encodes a nonstructural region, and ORF2 encodes the capsid protein (19). The function of ORF3, which almost entirely overlaps ORF2, is not totally clear. Interestingly ORF3 is not necessary for *in vitro* infection (20), but is necessary for *in vivo* infection of macaques (21). It is most likely multifunctional (19) and involved in pathogenesis (22–25). Most of the coding region in the HEV genome is under purifying selection (26, 27), *i.e.* selection against change in the amino acid sequence. Areas with an excess of amino acid substitutions, a signal of positive selection, have been found in the N-terminus of ORF2 and the C-terminus of ORF3, with another in the RNA dependent RNA polymerase (RdRp) region in ORF1 (26). Purdy *et. al.* (28) have described positive selection in the hypervariable region (HVR) of ORF1; however, Smith *et. al.* (27) failed to reproduce these results with a broader selection of statistical tests.

Phylogenetic analyses of HEV may help to shed light on evolutionary differences between genotypes, which underlie the epidemiological and clinical disparities. In a previous study, Chen *et. al.* (26) failed to discern any difference in selection pressures between genotypes. Since 2012, the number of appropriate full genome samples has increased by 150%. Using this expanded dataset, we revisit the question of evolutionary differences between the genotypes of HEV, using state-of-the-art methods. We focus on detecting natural selection, specifically investigating regions of positive selection which stand out from a background of purifying selection against non-synonymous substitutions. Our particular focus is on the overlap region, making ours the first analysis of this region as a focus of positive selection. We also carry out a detailed analysis of evolutionary rates, and link phylogenetic findings to the virological characteristics of the genotypes.

Methods

Sequence acquisition

All available sequences of hepatitis E virus in Genbank (29) were obtained by searching the NCBI Nucleotide Database using the taxonomic identifier (txid) 12461, along with associated metadata on host, country, and date of sampling. As of August 6th, 2014 there were 10,041 sequences, of which 258 sequences were at least 7000 nucleotides long (i.e. near full length genomes).

Sequence processing

Open reading frames, corresponding to sequence regions between consecutive stop codons, were identified for each sequence using `getorf`, part of the EMBOSS package (30). ORFs 1, 2, and 3 for each sequence were identified by `blastp` (31), with amino sequences of ORFs from the NCBI Reference Sequence NC_001434 as the query, and translated ORF sequences as the reference. Multiple sequence alignments (MSAs) for each ORF were generated using Clustal Omega (32), based on the translated sequences. Nucleotide sequences were mapped on to the corresponding aligned amino acid sequences using Seaview v. 4.5.0 (33). MSAs were trimmed, based on the start and stop of ORFs in NC_001434, and checked manually. In order to obtain a single in-frame sequence for the near-full length genome, we concatenated ORFs 1 and 2. The alignments and associated inferred data are available for download from github.com/veg/HEV-evolution-2015.

Sequences were screened for recombination using RDP4 (version 4.36 beta) (34), using eight available methods; RDP (35), GENECONV (36), BootScan (37), MaxChi (38), Chimaera (39), SiScan (40), PhylPro (41), LARD (42), and 3Seq (43), using default settings. Following exploratory analyses to determine whether recombination detection was simply an artifact of complex patterns of mutations, a sequence was deemed recombinant if three or more methods had reported it as a recombinant. Consistent with prior reports of recombination in HEV (26, 44, 45), we identified 14 recombinant viruses, including novel recombinants (see Table 1).

Genotypes were assigned to each sequence by sequence similarity and phylogenetic reconstruction. We used `tblastx` (from the BLAST 2.2.30+ software suite (31, 46)) to find the most similar sequences prototypical for each genotype; M73218 (genotype 1 (47)); M74506 (genotype 2 (48)); AF060668 (genotype 3 (49)); and AJ272108 (genotype 4 (18)). Designations were further investigated by inspecting phylogenetic reconstructions obtained using `FastTree` v2.1.8 (50). Of the 258 near-full length genomes, 127 were isolated from humans, and were selected for further analysis. Two sequences were excluded on the basis that they were abnormally divergent from the other sequences: M74506, which is a genotype 2 virus, and JQ013793, which is similar to a strain of HEV isolated from rabbits (51). Genotype-specific alignments were generated and merged into a single master alignment using MACSE v.1.01b (52). Sequences with a 100% identity to other isolates were removed, resulting in a final dataset of 113 unique near full-length genomes isolated from humans, comprised of concatenated ORF1 and ORF2 regions, with 26 genotype 1 sequences, 42 genotype 3 sequences, and 45 genotype 4 sequences. We split the alignment into ORF1 and ORF2 regions, extracted the overlapping part of ORF3 from ORF2, and split ORF2 into the region overlapping ORF3, and the non-overlapping region. We also identified 56 unique HEV genomes isolated from swine. The swine HEV sequence alignment was merged with the human HEV dataset using profile alignment in codon space using MACSE.

Genome-level selection analyses

Selection analyses employed a suite of phylogenetic methods, as implemented in HyPhy(53) and Datamonkey (54, 55) using default settings. FUBAR (56) was used to characterize pervasive selective pressures, i.e., those aggregated over all branches in the phylogeny. Both an alignment-wide distribution of synonymous and non-synonymous substitution rates, and site-level estimates were obtained using FUBAR. MEME (57) was applied to identify individual sites subject to episodic positive selection (i.e. operating along a subset of tree branches). aBSREL (58) allowed us to estimate the complexity of evolutionary processes along individual tree branches, and to determine which branches in the tree were subject to positive selection along a subset of sites in the alignment. Finally, RELAX (59) was employed to formally test whether or not the evolutionary pressures were relaxed or intensified for HEV infecting human hosts relative to those infecting swine hosts.

So that we could formally test whether or not selection was relaxed or intensified in the overlapping region of ORF3 relative to ORF2, we modified the RELAX method (60) to accept two gene alignments as input. Briefly, we fit a 3-rate random effects branch-site class model (61) with three ω classes to accommodate the variation in selective forces across sites and branches in an unrestricted fashion jointly to both alignments, while endowing each with its own branch lengths, equilibrium codon frequencies, and nucleotide substitution biases. The RELAX test enforces a functional relationship between the ω ratios in reference (ORF2) and test (ORF3) alignments: ω ORF3 = (ω ORF2)K. The estimated value of K indicates whether selection in the test frame is relaxed ($K < 1$) or intensified ($K > 1$) relative to the reference frame. A likelihood ratio test of the null hypothesis ($K=1$), versus the alternative hypothesis ($K \neq 1$) establishes statistical significance of relaxation (or intensification).

Codon substitution model for overlapping regions

We fitted three codon substitution models that explicitly consider whether mutations are synonymous in just one of ORF2 and ORF3, or both. These models, which have been previously used to screen for biologically meaningful alternative reading frames in mammalian genomes (62), generate estimates of rates RXY, which refer to the rates of substitutions which are synonymous ($X = 0$) or non-synonymous ($X=1$) in the primary frame (ORF2), and synonymous ($Y = 0$) or non-synonymous (Y=1) in the alternative frame (ORF3). R00 - the rate for substitutions that are synonymous in both frames, is fixed at 1, and the other three rates are estimated relative to R00. Maximum likelihood parameter estimates and associated 95% confidence intervals (profile likelihood) were calculated for a model in which R01, R10, and R11 were allowed to vary freely. We also performed likelihood ratio tests comparing the full model with two null models. The first null model assumes that R11 is greater than one or both of R01 and R10; the expectation is that R11 (non-synonymous in both frames) should be less than either R01 or R10, because changing both frames should be evolutionary constrained. The second null model assumes that R01=R10; rejection of this null hypothesis suggests that one frame is more constrained than the other.

Molecular clock analyses

Sequences were annotated by year of sampling. In many cases, these data were obtained from Genbank records. In other cases, the primary reference was used. In the cases where neither source gave the sampling year, we used the submission date to Genbank as an upper bound for the sampling date, with the lower bound set as the earliest known sampling year (March 1990, from (63)). To estimate the evolutionary rate for genotype specific alignments whilst accommodating the uncertainty in sampling times, we used a Bayesian phylogenetic approach, as implemented in MrBayes v3.2.2 (64). A general time reversible (GTR) model was fitted, with rate variation modelled as a discrete gamma distribution with 4 categories. Base frequencies were fixed at their empirical values, and a uniform prior placed on topologies. A relaxed clock model was used, assuming that evolutionary rates were drawn independently from a gamma distribution. Default priors were used, with the exception of the clock rate, which was set to lognorm(-9,1). Two chains were run for 110 million generations with a burnin of 10 million, thinned to give a sample of 1000 iterations. Results were processed using the coda library (65) in R and the 95% upper and lower credible intervals were inferred from the posterior distribution. Convergence was tested using manual inspection of traces of parameter values, and calculation of the Gelman-Rubin statistic (66). The rv library was used to generate credible intervals for the difference in clockrate between genotypes in the same ORF. To validate the use of a relaxed clock we analysed the parameter describing the variance of the rate distribution of the relaxed clock, and found it to be distinct from zero with a median of 0.01914977 (95% credible interval=0.00115991-0.04887433), providing support for the use of a relaxed clock over a strict clock. The ggplot2 library (67) in R was used to create rate plots.

Host-specific patterns of evolution

Human and swine HEV near full length genomes were split into genotype 3 and genotype 4 alignments. Phylogenies for each genotype were reconstructed separately using maximum likelihood with RAxML v.8 (68), assuming the GTR model of nucleotide substitution with gamma distributed rate variation. Phylogenies were rooted with lsd v.0.1 (69), using the median estimate of the sampling time for each sequence. Terminal branches were classified as human or swine based on which host they were isolated from. Interior branches were classified as 'human'/'swine' whenever all of their descendants were labelled as 'human'/'swine', following post-order tree traversal. Species-specific estimates of the distribution of the ω ratio were obtained on the basis of the models implemented in RELAX (59).

Implementation

Except where otherwise stated, selection analyses were performed using HyPhy (53), using phylogenies of each region reconstructed using RAxML v.8 (68), assuming the GTR model of nucleotide substitution with gamma distributed rate variation, or the MG94xGTR model of codon substitution with analysis-defined patterns of site-to-site and branch-to-branch rate variation. Tree visualisation was carried out using the `phylotree.js` widget implemented as an extension of the D3 (D3js.org) JavaScript visualisation library (http://veg.github.io/hyphy-vision).

Results

Genome-wide patterns of selection

To visually identify genomic regions under positive or purifying selection, we estimated the number of non-synonymous (amino-acid changing, dN) and synonymous (amino-acid preserving, dS) changes for each codon (Figure 1) using the FUBAR method (56), which estimates these quantities for individual sites using an Empirical Bayes procedure in the phylogenetic likelihood framework. Consistent with previous findings, most of the genome was under purifying selection ($dN < dS$). However, within each ORF, specific regions showed statistically significant evidence of positive selection ($dN > dS$): the hypervariable region (HVR) in ORF1, the 5' end of ORF2, and ORF3 (Figure 1). As the 5' end of ORF2 and ORF3 are overlapped, we repeated FUBAR analysis of this area in each reading frame, finding a strong signal of positive selection throughout the overlapped region of ORF2 and a weaker signal in ORF3 (Figure 1).

Rate variation amongst site and branches

We fitted an adaptive branch-site model (58) to the alignment of 113 isolates with near full length genomes. Overall, there was very strong evidence of variation in selective pressure both over sites and lineages (Δ AIC = 1760 in favour of the model which allows such variation), with 54 (24%) of branches supporting site-to-site variation, with 2 rate classes per branch. The remaining 169 branches could be adequately explained by a model where all sites evolve at a single rate. Eleven branches were subject to statistically significant ($p < 0.05$ after Holm-Bonferroni multiple testing correction) positive selection. Of the eleven, one belonged to genotype 1 (M94177), 3 to genotype 3 (KJ701409, AF060669, and AF060668), and 7 to genotype 4 (AB220977, AB291964, AB291959, AB220979, AB220976, AB220978, and AJ272108). In all cases, 98% or more sites were under strong purifying selection ($\omega < 0.05$), and the remainder were under very strong positive selection ($\omega > 50$). Interestingly, despite the fact that the estimated distribution for all interior branches separating the individual genotypes had a component with $\omega >$

1, none rose to the level of statistical significance for positive selection, after multiple test correction.

Selection on individual sites in the ORF2/ORF3 overlap region

We performed selection analyses on each genotype separately. Consistent with the whole genome FUBAR analysis, signals of positive selection were found in the overlap region. Using multiple methods for detecting selection, positive selection was found in both frames of genotypes 3 and 4, whilst neither reading frame of genotype 1 exhibited any significantly positively selected sites (see Table 2). This trend is shown in Figure 2, which renders the genotype-specific FUBAR distribution estimates for each reading frame, representing the proportion of sites evolving at different nonsynonymous and synonymous rates. These selective 'fingerprints' demonstrate that there are sites subject to positive selection in both reading frames in enzoonotic genotypes 3 and 4, but none in the human-only genotype 1.

In order to further disentangle selection on different reading frames, we fitted a codon substitution model (see Table 3) that considers whether mutations are non-synonymous in ORF2, ORF3, or both ORF2 and ORF3. In all three genotypes, the rate of substitutions that were non-synonymous at a codon level in both frames was significantly lower than the rate of non-synonymous mutations in either of the specific frames. This finding is consistent with a dual-coding region where both frames are under purifying selection for functional conservation, on average. The point estimates derived from Genotype 1 are lower than for genotypes 3 and 4, hinting at stronger conservation for the former. For genotypes 1 and 3, ORF2 and

ORF3 are evolving at significantly different rates, when considering non-synonymous substitutions affecting only one of the frames, with ORF2 experiencing more of the latter. For genotype 4, the rates are statistically indistinguishable.

To formally test whether ORF3 is evolving differently from the overlapping region of ORF2, we modified the RELAX method(59) to accept two gene alignments as input. The RELAX test enforces a functional relationship between the ω ratios in reference (ORF2) and test (ORF3) alignments: ω ORF3 = (ω ORF2)K. The estimated value of K indicates whether selection in ORF3 is relaxed ($K < 1$) or intensified ($K > 1$) relative to ORF2. A likelihood ratio test of the null hypothesis ($K=1$), versus the alternative hypothesis ($K \neq 1$) establishes statistical significance of relaxation (or intensification). The application of the RELAX procedure (Table 4) suggests strong relaxation of selection in ORF3 (namely, through the elimination of the positively selected component) in genotypes 1 and 3, and a weak (non-significant) intensification of selection in ORF3 in genotype 4. This finding of ORF 2 apparently driving the signal of positive selection reproduces, by different means, the findings in Table 3.

Estimates of time-scaled synonymous and nonsynonymous substitution rates

Differences in the rate of evolution between different genotypes could arise due to different selection pressures on the genotypes (i.e. different ratios of nonsynonymous to synonymous substitution), as suggested by the selection pressure analyses, or could simply be due to differences in the substitution rate (i.e. differences in synonymous rates), independent of selection pressure. To address this question, we derived time-scaled estimates of synonymous and non-synonymous rates, using the procedure described in (70). Briefly, a Maximum Clade Credibility tree obtained using MrBayes was used as input to a codon analysis in HyPhy, using the Muse-Gaut codon-substitution model with branch-specific α (synonymous) and β (non-synonymous) rate parameters, which were used to partition the fixed branch length into synonymous and non-synonymous components. The conversion from expected substitutions per site to expected substitutions / site / year was carried out under the assumption of a strict molecular clock. The results are summarized in Table 5, and demonstrate that the synonymous substitution rate of genotype 1 is approximately half that of genotypes 3 and 4. Whilst this is an important confounding factor, this effect merely adds to an extant signal of positive selection in genotype 1 sequences, because lower dS would work to elevate dN/dS for genotype 1 (for example, results in Table 3 are robust to this confounding factor), it has not created the effect *de novo*.

Analysis of evolutionary rates

We estimated the evolutionary rate of each genotype, including information on the estimated time of sampling (Table 6, Figure 3). The mean evolutionary rate was similar across ORFs at approximately 0.003-0.005 substitutions per site per year. Evolutionary rates of genotype 1 were significantly lower than those of genotypes 3 and 4 across all ORFs. Genotypes 3 and 4 demonstrate remarkably similar profiles, with differences non significant across all ORFs. Table 5 shows that this is likely due to both lower synonymous and non-synonymous substitution rates. Unsurprisingly, the overlap region of ORF 2 appears very similar to ORF 3, as they overlap extensively. More surprisingly the non-overlap region of ORF 2 has a similar evolutionary rate profile to ORF 1, with which is does not overlap at all.

Host-specific differences in evolution

We investigated whether patterns of evolution differ not only by genotype but also species of isolation. We constructed phylogenies of the human and swine lineages for genotypes 3 and 4, and assigned branches as either human, swine or indeterminate. For both genotypes 3 and 4, there was notable intermingling of lineages (Figure 4), representing a continuous zoonotic process. Genome-wide analyses of selective pressures using the null and alternative models in the RELAX suite, found a slight, but statistically significant intensification of selection along human branches relative to swine branches. For genotype 3, RELAX inferred the intensification coefficient of $K = 1.09$ ($p = 0.013$ when compared to the null hypothesis of $K = 1$). For genotype 4, the inferred values were $K = 1.12$ and $p < 0.001$. In brief, this test establishes that ω estimates on human branches are more extreme (further away from $\omega = 1$, i.e., neutrality), than on swine branches. For these analyses, indeterminate branches were endowed with their own ω distribution and branch-level relaxation/intensification coefficients, treated here as nuisance parameters. For genotype 3, 91.5% of the bootstrapped trees supported p-value of <=0.05 or less (count = 211, median p value = 0.019 (4E-5-0.0699), median K = 1.14987 (1.0324-1.2159)). For G4 every single p-value for RELAX was < 0.05 (count = 352, median p value = 9E4 (7E-11-0.0051), median K = 1.4059 (1.0900-1.8709)).

Discussion

We have demonstrated differences in the evolution of hepatitis E virus (HEV) between the three open reading frames, and quantified how evolutionary patterns differ between genotypes. Using a high quality alignment comprising all available near full length genomes, our analyses have identified and focused in on the main genomic region of interest: the ORF2/ORF3 overlap region (Figure 1). Selection analysis of the overlap region revealed multiple sites/regions undergoing positive selection in genotypes 3 and 4, but a much weaker signal in genotype 1 (Figure 2). This pattern is the same as that found in evolutionary rates, with significantly reduced evolutionary rates in both ORF2 overlap and ORF3 of genotype 1 (Figure 3), driven by differences in both synonymous and nonsynonymous rates. A genome-wide analysis of genotype 3 and 4 isolates revealed a slight but statistically significant intensification of selective pressures in human lineages compared to swine lineages. We speculate, as genotype 1 viruses only infect one host and genotypes 3 and 4 are enzoonotic, that genes in genotype 1 are subject to reduced diversifying or balancing selection pressure as they have fine-tuned fitness by specializing to their single host species. This functional constraint on amino acid changes is particularly pertinent as this effect was found in both ORF2 and ORF3, which are both believed to be important in the pathogen-host response as the capsid protein and an immunomodulatory phosphoprotein, respectively (19). ORF1, in contrast, contains housekeeping genes, which are less likely to be host-specific.

Cyclical host jumps seen in arboviruses, e.g., West-Nile virus, are associated with purifying selection (71, 72). The concept behind this is that only substitutions conferring a selection advantage in both hosts are preserved. However this paradigm may not be globally applicable. *In silico* models of evolution under varying selection pressure show that the rate of evolution and dN/dS can be either suppressed or increased depending on how the timescale of the environmental change compares to that of adaptation to the new environment (73, 74). In an environment with very slow environmental fluctuations, each substitution will either fix or go extinct during the epoch in which it arose, whilst in faster oscillations a substitution will have the opportunity to be selected in both environments (74). Therefore it may not be the case that all cyclical environments induce stronger purifying selection. HEV may be an instance where the interaction of oscillation period and time taken to reach a particular fitness level interact in such a way as to promote diversity and a signal of positive selection. We therefore postulate the signal of positive selection in those genomic regions which interact with the host (ORF2 and ORF3, ORF1 contains housekeeping genes) represents a cyclical but ultimately futile selection process in each species which results in a phenotype which is sub-optimally fit in both. Although, interestingly, our host-specific analysis provides evidence that the scales are currently tipped towards optimizing for the human host.

Overlapping reading frames are not uncommon in RNA viruses (75), and have been suggested as a mechanism of packing more genes in a limited genomic space (76). Whilst other studies have found a scattering of positively selected codons in this region (26), none have investigated the overlap region as a locus for positive selection. Overlapping coding regions are often constrained as substitutions impact two protein products instead of one, causing a reduction in evolutionary rates (77). However, there is a precedent for rapid evolution in overlapped regions in both viruses (e.g. PB1-F2 and PA-X in Influenza A virus (78, 79)) and mammalian genomes (80), and statistical techniques designed for mammalian overlapping regions (62) helped us to shed light on what is driving evolution in the overlap region. Investigating selection in a region of overlapping reading frames requires reading-frame aware models. Apart from a currently computationally infeasible full Bayesian treatment of co-dependent evolution in multiple reading frames (e.g. see (81)), two approximate approaches have been used in practice. Firstly, the overlapping reading frames can be treated entirely independently, and analysed using standard methods (e.g. (82)). When this approach is taken to estimate synonymous and non-synonymous rates and carry out tests of selection, the interpretation of results becomes difficult (e.g., how valid is the concept of a frame-specific synonymous rate in this context?), and can lead to false positive results (83). Secondly, codon-substitution models which correct for the "expected" context of a codon in the alternative reading frame have been proposed (62, 84, 85). The benefit of these models is that, while remaining computationally tractable, they directly estimate frame-aware rates of synonymous and non-synonymous mutations. Such models have been successfully used to perform genome-wide screens of ORFs with multiple overlapping reading frames for functional constraint (62), and the evolution of overlapping reading frames in Influenza A virus (84). Our analysis of the overlapping reading frames shows a significant difference between the rate of substitutions that were non-synonymous in both frames compared to only one (Table 3), indicative of positive selection. Applying this model also allowed us to find out which reading frame (and therefore likely gene product) was driving the positive signal in this region. Although both reading frames are subject to positive selection, the ORF2 overlap region appears to be driving selection (at least in genotypes 1 and 3).

Evolutionary rates showed significant differences between the anthropotropic and zoonotic genotypes. Across all genomic regions genotype 1 had significantly lower evolutionary rates than genotypes 3 and 4, whilst genotypes 3 and 4 had remarkably similar values. Evolutionary rates inferred from the posterior distribution were typical of an RNA virus (86) and related viruses e.g. norovirus (87). We, like Nakano *et. al.* (2012) and Purdy *et. al.* (2012), found a relaxed clock most appropriate to reflect the variation in substitution rates between branches in HEV, although our estimates of evolutionary rate are higher than those reported previously (88–90). It should be noted that apparent evolutionary rates show time dependency, with an elevation towards the present due to transient unfixed substitutions, and apparent reduction in the past due to saturation (91). Interestingly the ORF 2 overlap region has a very distinct profile of evolutionary rates across all genotypes when compared with the non-overlap region. The non-overlap region is strikingly similar to the ORF 1 profile, which is believed to contain housekeeping genes.

Our analyses of the host specific patterns of evolution are important in showing that the differences described above are largely genotype, not host species, dependent. For genotypes 3 and 4, we detected a slight increase in selection intensity in human-associated viral lineages compared to swine associated viral lineages, in contrast to the large differences between genotypes. The construction of phylogenies demonstrated intermingling of swine and human lineages, and suggest a high rate of host jumps indicative of the frequent transmission between swine and humans and back again. The transmission of HEV from humans to swine has been demonstrated extensively in laboratory settings (92, 93), however its frequency and mechanism in the wild remain unclear (94). As phylogenies do not contain independent information on the direction of transmission, it is hard to demonstrate such 'reverse zoonoses' from sequence data alone.

Our study represents the most comprehensive HEV sequence analysis to date. It is important, however, to note the limitations in the publicly available data. Genotypes 3 and 4 dominate in the developed world, whilst genotypes 1 and 2 are found in the developing world (4). This global differential distribution of genotypes may be an important confounder, as in fact the virus is not interacting with a single homogeneous human host, but rather different clades of virus are interacting with specific groups of human hosts. These groups are likely to differ significantly, e.g. in the population composition of Human Leukocyte Antigen alleles (95), which in turn imposes differential selective pressures on the pathogen as part of host-pathogen interactions. As is the case for most pathogens, sampling is heavily biased by location. There are many samples from Europe and East Asia, but few from

Australasia and Africa, and there are many countries for which there are no sequence data. Furthermore, little is known about genotype 2, with too few full length viral genomes publicly available to build a reliable alignment, so studies either omit it (26), or have low statistical power (96).

Hepatitis E virus is of increasing interest to public health officials and clinicians. Attention in the developed world to date has been limited, partly due to the acute nature of the infection in healthy individuals and the apparently asymptomatic nature of infection in swine. However, the emergence of new strains of HEV, such as one recently documented in the U.K. (97), emphasise the need for continuing surveillance and characterisation of this pathogen.

Acknowledgements

This work was funded in part by a Medical Research Council Methodology Research Programme grant (MR/J013862/1) to SDWF, as well as by the Department of Pathology at the University of Cambridge. SLKP and HyPhy development was supported by the US National Institutes of Health (grants GM093939 and GM110749). Many thanks to Prof. Ian Goodfellow (University of Cambridge) for providing a genomic map of HEV.

AB performed data collection, annotation and interpretation, ran analyses, created figures and drafted the initial paper. SP devised overlapping reading frame analysis, ran further analyses, created figures, and revised the draft paper. BD and JL provided technical assistance, including generation of figures, and revised the draft paper. SF devised and initiated the collaborative project, designed the data collection and analysis strategies and revised the draft paper.

References

1. **Rein DB**, **Stevens GA**, **Theaker J**, **Wittenborn JS**, **Wiersma ST**. 2012. The global burden of hepatitis e virus genotypes 1 and 2 in 2005. Hepatology **55**:988–997.

2. **Meng X-J**. 2013. Zoonotic and foodborne transmission of hepatitis E virus. Semin Liver Dis **33**:41–49.

3. **Berto A**, **Martelli F**, **Grierson S**, **Banks M**. 2012. Hepatitis E virus in pork food chain, united kingdom, 2009-2010. Emerg Infect Dis **18**:1358–1360.

4. **Teshale EH**, **Hu DJ**, **Holmberg SD**. 2010. The two faces of hepatitis E virus. Clin Infect Dis **51**:328–334.

5. **Lewis H**, **Wichmann O**, **Duizer E**. 2010. Transmission routes and risk factors for autochthonous hepatitis E virus infection in europe: a systematic review. Epidemiol Infect **138**:145–166.

6. **Banks M**, **Bendall R**, **Grierson S**, **Heath G**, **Mitchell J**, **Dalton H**. 2004. Human and porcine hepatitis E virus strains, United Kingdom. Emerg Infect Dis **10**:953–955.

7. **Emerson S**, **Anderson D**, **Arankalle A**. 2004. VIIIth report of the ICTV. Report.

8. **Worm HC**, **Poel WHM van der**, **Brandstatter G**. 2002. Hepatitis E: an overview. Microbes Infect **4**:657–666.

9. **Lu L**, **Li CH**, **Hagedorn CH**. 2006. Phylogenetic analysis of global hepatitis E virus sequences: genetic diversity, subtypes and zoonosis. Rev Med Virol **16**:5–36.

10. **Schlauder GG**, **Mushahwar IK**. 2001. Genetic heterogeneity of hepatitis E virus. J Med Virol **65**:282–292.

11. **Shrestha MP**, **Scott RM**, **Joshi DM**, **Mammen Jr MP**, **Thapa GB**, **Thapa N**, **Myint KSA**, **Fourneau M**, **Kuschner RA**, **Shrestha SK**, **others**. 2007. Safety and efficacy of a recombinant hepatitis e vaccine. New England Journal of Medicine **356**:895–903.

12. **Zhu F-C**, **Zhang J**, **Zhang X-F**, **Zhou C**, **Wang Z-Z**, **Huang S-J**, **Wang H**, **Yang C-L**, **Jiang H-M**, **Cai J-P**, **others**. 2010. Efficacy and safety of a recombinant hepatitis e vaccine in healthy adults: a large-scale, randomised, double-blind placebo-controlled, phase 3 trial. The Lancet **376**:895–902.

13. **Kuniholm MH**, **Purcell RH**, **McQuillan GM**, **Engle RE**, **Wasley A**, **Nelson KE**. 2009. Epidemiology of hepatitis e virus in the united states: results from the third national health and nutrition examination survey, 1988–1994. Journal of Infectious Diseases **200**:48–56.

14. **Bendall R**, **Ellis V**, **Ijaz S**, **Ali R**, **Dalton H**. 2010. A comparison of two commercially available anti-hEV igG kits and a re-evaluation of anti-hEV igG seroprevalence data in developed countries. Journal of medical virology **82**:799–805.

15. **Kamar N**, **Bendall R**, **Legrand-Abravanel F**, **Xia N-S**, **Ijaz S**, **Izopet J**, **Dalton HR**. 2012. Hepatitis e. The Lancet **379**:2477–2488.

16. **Dalton HR**, **Stableforth W**, **Thurairajah P**, **Hazeldine S**, **Remnarace R**, **Usama W**, **Farrington L**, **Hamad N**, **Sieberhagen C**, **Ellis V**, **others**. 2008. Autochthonous hepatitis e in southwest england: natural history, complications and seasonal variation, and hepatitis e virus igG seroprevalence in blood donors, the elderly and patients with chronic liver disease. European journal of gastroenterology & hepatology **20**:784–790.

17. **Purcell RH**, **Engle RE**, **Govindarajan S**, **Herbert R**, **St Claire M**, **Elkins WR**, **Cook A**, **Shaver C**, **Beauregard S Michelle**, **Emerson S**. 2013. Pathobiology of hepatitis e: lessons learned from primate models. Emerging Microbes & Infections **2**:e9.

18. **Wang Y**, **Zhang H**, **Ling R**, **Li H**, **Harrison TJ**. 2000. The complete sequence of hepatitis E virus genotype 4 reveals an alternative strategy for translation of open reading frames 2 and 3. J Gen Virol **81**:1675–1686.

19. **Cao D**, **Meng X-J**. 2012. Molecular biology and replication of hepatitis e virus. Emerging microbes & infections **1**:e17.

20. **Emerson SU**, **Nguyen H**, **Torian U**, **Purcell RH**. 2006. ORF3 protein of hepatitis e virus is not required for replication, virion assembly, or infection of hepatoma cells in vitro. Journal of virology **80**:10457–10464.

21. **Graff J**, **Nguyen H**, **Yu C**, **Elkins WR**, **Claire MS**, **Purcell RH**, **Emerson SU**. 2005. The open reading frame 3 gene of hepatitis e virus contains a cis-reactive element and encodes a protein required for infection of macaques. Journal of virology **79**:6680–6689.

22. **Chandra V**, **Kar-Roy A**, **Kumari S**, **Mayor S**, **Jameel S**. 2008. The hepatitis E virus ORF3 protein modulates epidermal growth factor receptor trafficking, STAT3 translocation, and the acute-phase response. J Virol **82**:7100–7110.

23. **Tyagi S**, **Korkaya H**, **Zafrullah M**, **Jameel S**, **Lal SK**. 2002. The phosphorylated form of the oRF3 protein of hepatitis e virus interacts with its non-glycosylated form of the major capsid protein, oRF2. Journal of Biological Chemistry **277**:22759–22767.

24. **Tyagi S**, **Surjit M**, **Lal SK**. 2005. The 41-amino-acid c-terminal region of the hepatitis e virus oRF3 protein interacts with bikunin, a kunitz-type serine protease inhibitor. Journal of virology **79**:12081–12087.

25. **Tyagi S**, **Surjit M**, **Roy AK**, **Jameel S**, **Lal SK**. 2004. The oRF3 protein of hepatitis e virus interacts with liver-specific α1-microglobulin and its precursor α1-microglobulin/bikunin precursor (aMBP) and expedites their export from the hepatocyte. Journal of Biological Chemistry **279**:29308–29319.

26. **Chen X**, **Zhang Q**, **He C**, **Zhang L**, **Li J**, **Zhang W**, **Cao W**, **Lv Y-G**, **Liu Z**, **Zhang J-X**, **Shao Z-J**. 2012. Recombination and natural selection in hepatitis E virus genotypes. J Med Virol **84**:1396–1407.

27. **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.

28. **Purdy MA**, **Lara J**, **Khudyakov YE**. 2012. The hepatitis E virus polyproline region is involved in viral adaptation. PLoS One **7**:e35974–e35974.

29. **Benson DA**, **Cavanaugh M**, **Clark K**, **Karsch-Mizrachi I**, **Lipman DJ**, **Ostell J**, **Sayers EW**. 2013. GenBank. Nucleic Acids Res **41**:D36–D42.

30. **Rice P**, **Longden I**, **Bleasby A**. 2000. EMBOSS: the european molecular biology open software suite. Trends in genetics **16**:276–277.

31. **Camacho C**, **Coulouris G**, **Avagyan V**, **Ma N**, **Papadopoulos J**, **Bealer K**, **Madden TL**. 2009. BLAST+: architecture and applications. BMC Bioinformatics **10**:421.

32. **Sievers F**, **Higgins DG**. 2014. Clustal omega, accurate alignment of very large numbers of sequences. Methods Mol Biol **1079**:105–116.

33. **Gouy M**, **Guindon S**, **Gascuel O**. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol **27**:221–224.

34. **Martin DP**, **Murrell B**, **Golden M**, **Khoosal A**, **Muhire B**. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol **1**:vev003.

35. **Martin D**, **Rybicki E**. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics **16**:562–563.

36. **Padidam M**, **Sawyer S**, **Fauquet CM**. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology **265**:218–225.

37. **Martin DP**, **Posada D**, **Crandall KA**, **Williamson C**. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses **21**:98–102.

38. **Maynard Smith J**. 1992. Analyzing the mosaic structure of genes. J Mol Evol **34**:126–129.

39. **Posada D**, **Crandall KA**. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A **98**:13757–13762.

40. **Gibbs MJ**, **Armstrong JS**, **Gibbs AJ**. 2000. Sister-scanning: a monte carlo procedure for assessing signals in recombinant sequences. Bioinformatics **16**:573–582.

41. **Weiller GF**. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol **15**:326–335.

42. **Holmes EC**, **Worobey M**, **Rambaut A**. 1999. Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol **16**:405–409.

43. **Boni MF**, **Posada D**, **Feldman MW**. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics **176**:1035–1047.

44. **Cuyck H van**, **Fan J**, **Robertson DL**, **Roques P**. 2005. Evidence of recombination between divergent hepatitis E viruses. J Virol **79**:9306–9314.

45. **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**.

46. **Altschul SF**, **Gish W**, **Miller W**, **Myers EW**, **Lipman DJ**. 1990. Basic local alignment search tool. J Mol Biol **215**:403–410.

47. **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.

48. **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.

49. **Schlauder GG**, **Dawson GJ**, **Erker JC**, **Kwo PY**, **Knigge MF**, **Smalley DL**, **Rosenblatt JE**, **Desai SM**, **Mushahwar IK**. 1998. The sequence and phylogenetic analysis of a novel hepatitis E virus isolated from a patient with acute hepatitis reported in the united states. J Gen Virol **79 (Pt 3)**:447–456.

50. **Price MN**, **Dehal PS**, **Arkin AP**. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One **5**:e9490.

51. **Izopet J**, **Dubois M**, **Bertagnoli S**, **Lhomme S**, **Marchandeau S**, **Boucher S**, **Kamar N**, **Abravanel F**, **Guérin J-L**. 2012. Hepatitis E virus strains in rabbits and evidence of a closely related strain in humans, France. Emerg Infect Dis **18**:1274–1281.

52. **Ranwez V**, **Harispe S**, **Delsuc F**, **Douzery EJ**. 2011. MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. PLoS One **6**:e22594.

53. **Pond SLK**, **Frost SDW**, **Muse SV**. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**:676–679.

54. **Pond SLK**, **Frost SDW**. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics **21**:2531–2533.

55. **Delport W**, **Poon AFY**, **Frost SDW**, **Pond SLK**. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics **26**:2455–2457.

56. **Murrell B**, **Moola S**, **Mabona A**, **Weighill T**, **Sheward D**, **Kosakovsky Pond SL**, **Scheffler K**. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol **30**:1196–205.

57. **Murrell B**, **Wertheim JO**, **Moola S**, **Weighill T**, **Scheffler K**, **Pond SK**. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genetics **8**:e1002764–e1002764.

58. **Smith MD**, **Wertheim JO**, **Weaver S**, **Murrell B**, **Scheffler K**, **Kosakovsky Pond SL**. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol **32**:1342–53.

59. **Wertheim JO**, **Murrell B**, **Smith MD**, **Kosakovsky Pond SL**, **Scheffler K**. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol **32**:820–32.

60. **Wertheim JO**, **Murrell B**, **Smith MD**, **Pond SLK**, **Scheffler K**. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Molecular biology and evolution **32**:820–832.

61. **Pond SLK**, **Murrell B**, **Fourment M**, **Frost SD**, **Delport W**, **Scheffler K**. 2011. A random effects branch-site model for detecting episodic diversifying selection. Molecular biology and evolution msr125.

62. **Chung W-Y**, **Wadhawan S**, **Szklarczyk R**, **Pond SK**, **Nekrutenko A**. 2007. A first look at ARFome: Dual-coding genes in mammalian genomes. PLoS Comput Biol **3**:855–861.

63. **Reyes GR**, **Purdy MA**, **Kim JP**, **Luk KC**, **Young LM**, **Fry KE**, **Bradley DW**. 1990. Isolation of a cDNA from the virus responsible for enterically transmitted non-A, non-B hepatitis. Science **247**:1335–1339.

64. **Ronquist F**, **Teslenko M**, **Mark P van der**, **Ayres DL**, **Darling A**, **Hohna S**, **Larget B**, **Liu L**, **Suchard MA**, **Huelsenbeck JP**. 2012. MrBayes 3.2: Efficient

bayesian phylogenetic inference and model choice across a large model space. Syst Biol **61**:539–542.

65. **Plummer M**, **Best N**, **Cowles K**, **Vines K**. 2006. CODA: Convergence diagnosis and output analysis for mCMC. R News **6**:7–11.

66. **Gelman A**, **Goegebeur Y**, **Tuerlinckx F**, **Van Mechelen I**. 2000. Diagnostic checks for discrete data regression models using posterior predictive simulations. J R Stat Soc Ser C Appl Stat **49**:247–268.

67. **Wickham H**. 2009. ggplot2: elegant graphics for data analysis. Springer New York.

68. **Stamatakis A**. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313.

69. **To T-H**, **Jung M**, **Lycett S**, **Gascuel O**. 2015. Fast dating using least-squares criteria and algorithms. Systematic biology syv068.

70. **Lemey P**, **Pond SLK**, **Drummond AJ**, **Pybus OG**, **Shapiro B**, **Barroso H**, **Taveira N**, **Rambaut A**. 2007. Synonymous substitution rates predict hIV disease progression as a result of underlying replication dynamics. PLoS Comput Biol **3**:e29.

71. **Coffey LL**, **Forrester N**, **Tsetsarkin K**, **Vasilakis N**, **Weaver SC**. 2013. Factors shaping the adaptive landscape for arboviruses: implications for the emergence of disease. Future microbiology **8**:155–176.

72. **Parameswaran P**, **Charlebois P**, **Tellez Y**, **Nunez A**, **Ryan EM**, **Malboeuf CM**, **Levin JZ**, **Lennon NJ**, **Balmaseda A**, **Harris E**, **others**. 2012. Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity. Journal of virology **86**:8546–8558.

73. **Kashtan N**, **Noor E**, **Alon U**. 2007. Varying environments can speed up evolution. Proceedings of the National Academy of Sciences **104**:13711–13716.

74. **Cvijović I**, **Good BH**, **Jerison ER**, **Desai MM**. 2015. Fate of a mutation in a fluctuating environment. Proc Natl Acad Sci U S A **112**:E5021–E5028.

75. **Neuhaus K**, **Oelke D**, **Fürst D**, **Scherer S**, **Keim DA**. 2010. Towards automatic detecting of overlapping genes-clustered bLAST analysis of viral genomes. Springer.

76. **Chirico N**, **Vianelli A**, **Belshaw R**. 2010. Why genes overlap in viruses. Proceedings of the Royal Society of London B: Biological Sciences **277**:3809–3817.

77. **Simon-Loriere E**, **Holmes EC**, **Pagan I**. 2013. The effect of gene overlapping on the rate of RNA virus evol. Mol Biol Evol **30**:1916–1928.

78. **Suzuki Y**. 2006. Natural selection on the influenza virus genome. Molecular biology and evolution **23**:1902–1911.

79. **Jagger B**, **Wise H**, **Kash J**, **Walters K-A**, **Wills N**, **Xiao Y-L**, **Dunfee R**, **Schwartzman L**, **Ozinsky A**, **Bell G**, **others**. 2012. An overlapping protein-coding region in influenza a virus segment 3 modulates the host response. Science **337**:199–204.

80. **Szklarczyk R**, **Heringa J**, **Pond SK**, **Nekrutenko A**. 2007. Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. Proc Natl Acad Sci U S A **104**:12807–12812.

81. **Pedersen A-MK**, **Jensen JL**. 2001. A dependent-rates model and an mCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Molecular Biology and Evolution **18**:763–776.

82. **Obenauer JC**, **Denson J**, **Mehta PK**, **Su X**, **Mukatira S**, **Finkelstein DB**, **Xu X**, **Wang J**, **Ma J**, **Fan Y**, **others**. 2006. Large-scale sequence analysis of avian influenza isolates. Science **311**:1576–1580.

83. **Holmes EC**, **Lipman DJ**, **Zamarin D**, **Yewdell JW**. 2006. Comment on“ large-Scale sequence analysis of avian influenza isolates”. Science **313**:1573–1573.

84. **Sabath N**, **Landan G**, **Graur D**. 2008. A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS One **3**:e3996.

85. **Mir K**, **Schober S**. 2014. Selection pressure in alternative reading frames. PloS one **9**.

86. **Jenkins GM**, **Rambaut A**, **Pybus OG**, **Holmes EC**. 2002. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. J Mol Evol **54**:156–165.

87. **Cotten M**, **Petrova V**, **Phan MV**, **Rabaa MA**, **Watson SJ**, **Ong SH**, **Kellam P**, **Baker S**. 2014. Deep sequencing of norovirus genomes defines evolutionary patterns in an urban tropical setting. J Virol **88**:11056–11069.

88. **Takahashi K**, **Toyota J**, **Karino Y**, **Kang JH**, **Maekubo H**, **Abe N**, **Mishiro S**. 2004. Estimation of the mutation rate of hepatitis E virus based on a set of closely related 7.5-year-apart isolates from sapporo, japan. Hepatol Res **29**:212–215.

89. **Purdy MA**, **Khudyakov YE**. 2010. Evolutionary history and population dynamics of hepatitis E virus. Plos One **5**:9.

90. **Nakano T**, **Takahashi K**, **Pybus OG**, **Hashimoto N**, **Kato H**, **Okano H**, **Kobayashi M**, **Fujita N**, **Shiraki K**, **Takei Y**, **Ayada M**, **Arai M**, **Okamoto H**, **Mishiro S**. 2012. New findings regarding the epidemic history and population dynamics of japan-indigenous genotype 3 hepatitis E virus inferred by molecular evolution. Liver Int **32**:675–688.

91. **Ho SY**, **Shapiro B**, **Phillips MJ**, **Cooper A**, **Drummond AJ**. 2007. Evidence for time dependency of molecular rate estimates. Systematic biology **56**:515–522.

92. **Meng X-J**, **Halbur PG**, **Shapiro MS**, **Govindarajan S**, **Bruna JD**, **Mushahwar IK**, **Purcell RH**, **Emerson SU**. 1998. Genetic and experimental evidence for cross-species infection by swine hepatitis E virus. J Virol **72**:9714–9721.

93. **Feagins A**, **Opriessnig T**, **Huang Y**, **Halbur P**, **Meng X**. 2008. Cross-species infection of specific-pathogen-free pigs by a genotype 4 strain of human hepatitis E virus. J Med Virol **80**:1379.

94. **Messenger AM**, **Barnes AN**, **Gray GC**. 2014. Reverse zoonotic disease transmission (zooanthroponosis): a systematic review of seldom-documented human biological threats to animals. PloS one **9**:e89055.

95. **Buhler S**, **Sanchez-Mazas A**. 2011. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. PLoS One **6**:e14643.

96. **Okamoto H**. 2007. Genetic variability and evolution of hepatitis E virus. Virus Res **127**:216–228.

97. **Ijaz S**, **Said B**, **Boxall E**, **Smit E**, **Morgan D**, **Tedder RS**. 2014. Indigenous hepatitis E in england and wales from 2003 to 2012: evidence of an emerging novel phylotype of viruses. J Infect Dis **209**:1212–1218.

Tables

Recombination analysis

Accession	Recombination reference	Genotype	Host	Major parent	Minor parent
AB097811	Wang et al. (2010)	3	Swine	AB193177	AB481227
AB291954	NONE	3	Human	AB443626	AB291953
D11093	van Cuyck et al. (2005)	NA	Human	D11092	D10330
DQ450072	Wang et al. (2010)	4	Swine	JF915746	GU188851;AB091394
EU723513	NONE	3	Swine	EU723512	EU723515
FJ426404	NONE	3	Swine	Unknown	FJ426403
FJ457024	NONE	NA	Human	JF443725	AF459438
HM439284	NONE	4	Human	JQ993308	JX855794; EU676172
JF443720	NONE	1	Human	AF459438	JF443725
JN564006	NONE	3	Human	AB089824	JQ679014
JQ655735	NONE	4	Human	GU188851	JQ655733
JX565469	NONE	Rabbit	Rabbit	AB740222	GU937805
KJ013414	NONE	Rabbit	Rabbit	Unknown	JQ768461;JX121233
KJ013415	NONE	Rabbit	Rabbit	Unknown	JQ768461;JX121233

Table 1: Details of recombinants found. 14 recombinant HEV sequences were identified in the 258 near full length genomes, generated by concatenating ORF1 and ORF2, by screening with RDP4 (version 4.36 beta) (34). With the exception of KJ013414 and KJ013415, which shared a recombinant structure, all recombinants were unique. Three recombinants had been previously described (see

Recombination reference column). The table also shows the genotype of the recombinant, the host it was isolated from, and the putative major and minor parents.

Genotype	Reading frame (ORF)	FUBAR (posterior ≥ 0.95)	MEME (p ≤ 0.05)
1	2	0	0
1	3	0	0
3	2	7	4 (4)
3	3	2	4 (2)
4	2	6	7 (6)
4	3	1	6 (1)

Table 2: The number of positively selected codon sites in each reading frame of each genotype of the overlap region (numbers in parentheses show how many sites were shared between MEME and FUBAR sets). Genotype 1 lacks any positively selected sites, meanwhile genotypes 3 and 4 produce a consistent signal of positively selected sites in both reading frames. Note that MEME is generally more sensitive, because it can detect selection on a subset of viral lineages, whilst FUBAR pools the signal of selection from all branches.

Genotype	ORF2	ORF3	Both	Both < ORF2/ORF3 LRT p-value	ORF2 ≠ ORF3 LRT p-value
1	0.056	0.032	0.012	0.005	0.039
95% CI	(0.036,0.083)	(0.020,0.048)	(0.005,0.024)		
3	0.167	0.082	0.011	<0.001	<0.001
95% CI	(0.138,0.201)	(0.066,0.099)	(0.005, 0.018)		
4	0.113	0.092	0.015	<0.001	0.12
95% CI	(0.090,0.140)	(0.076,0.110)	(0.009, 0.024)		

Table 3. Estimates of substitution rates that result in non-synonymous changes in at least one frame, relative to the rate of substitutions that are synonymous in both frames. A dimensionless metric, based on the model from Chung *et al.* (62). The last two columns show LRT-based p-values for rejecting the corresponding null hypotheses. Genotypes 3 and 4 demonstrate highly significant reading frame specific positive selection with ORF 2 convincingly driving the signal in genotype 3 but not 4 (rejection of null hypothesis). Genotype 1 has a lower background rate of non synonymous substitutions although does achieve significance, with the ORF 2 rate again significantly higher than ORF 3 rate.

Genotype	Relaxation parameter (K)	RELAX test p-value
1	< 0.0001	0.002
3	< 0.0001	< 0.0001
4	1.42	0.16

Table 4. Application of the RELAX procedure suggests strong relaxation of selection (namely, through the elimination of the positive selected component) in ORF3 of genotypes 1 and 3 relative to ORF2, and a weak, non-significant intensification of selection in genotype 4 of ORF3 relative to ORF2. This suggests that ORF2 and ORF3 are evolving differently, and ORF 2 is more responsible than ORF 3 for the signal of positive selection in genotypes 1 and 3.

Region	Genotype	Expected Synonymous substitutions / site / year	Expected non-synonymous substitutions / site / year
ORF 1	1	0.0022	0.00034
ORF 1	3	0.0051	0.00039
ORF 1	4	0.0053	0.00054
ORF 2 (non-overlap)	1	0.0030	0.00022
ORF 2 (non-overlap)	3	0.0063	0.00022
ORF 2 (non-overlap)	4	0.0051	0.00024
ORF 2 (overlap)	1	0.0022	0.00027
ORF 2 (overlap)	3	0.0040	0.00029
ORF 2 (overlap)	4	0.0053	0.00054
ORF 3 (overlap)	1	0.00067	0.00016
ORF 3 (overlap)	3	0.0017	0.00092
ORF 3 (overlap)	4	0.0011	0.00075

Table 5. Estimation of genotype specific synonymous substitution rates and non-synonymous substitution rates performed after Lemey *et. al.* (70). Synonymous substitution rate of genotype 1 is approximately half that of genotypes 3 and 4, which contributes to, but does not constitute, the signal of genotype specific positive selection. The rate of substitutions in ORF 3 is consistently elevated in comparison to other ORFs.

ORF	Genotype	Genotype	2.5% CredibleInterval	97.5% CredibleInterval	Significance
1	1	3	-0.004568928	-0.0004705684	*
1	1	4	-0.004663462	-0.0003018292	*
1	3	4	-0.002376133	+0.0023648283	
2nonoverlap	1	3	-0.004737947	-0.0002768805	*
2nonoverlap	1	4	-0.004861239	-0.0001427095	*
2nonoverlap	3	4	-0.002609059	+0.0025491087	
2overlap	1	3	-0.003233479	-0.0005518198	*
2overlap	1	4	-0.005747935	-0.0014066999	*
2overlap	3	4	-0.004201915	+0.0006952047	
3overlap	1	3	-0.003521166	-0.0007273096	*
3overlap	1	4	-0.005043289	-0.0010736460	*
3overlap	3	4	-0.003172870	+0.0013850900	

Table 6. Assessing significance in differences in clockrates between genotypes for each ORF. The credible intervals are significant if they do not include zero. This shows genotype 1 has a significantly different clockrate from genotypes 3 & 4 across all ORFs. This supports the clockrate data in Figure 3.

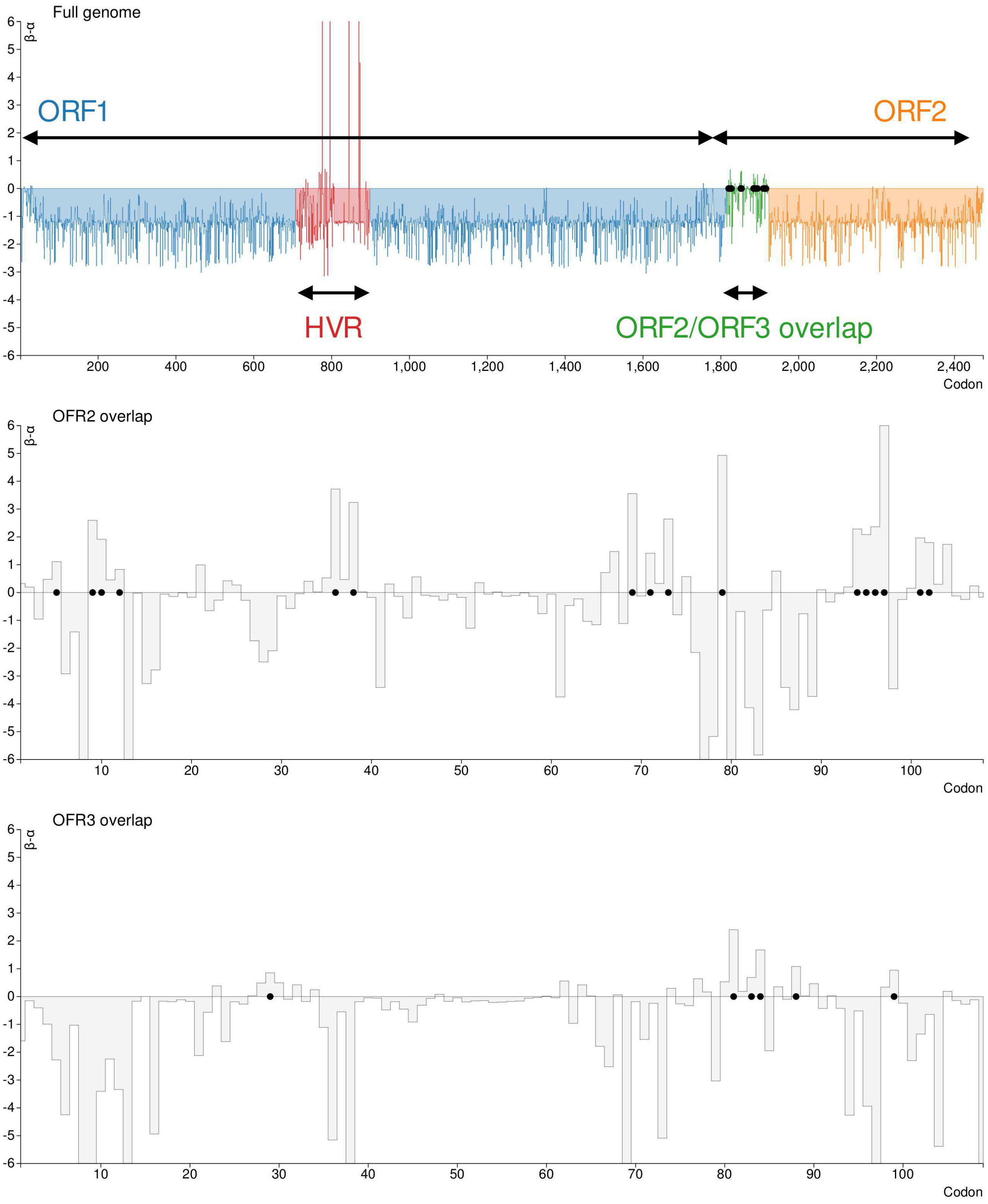
Figure Legends

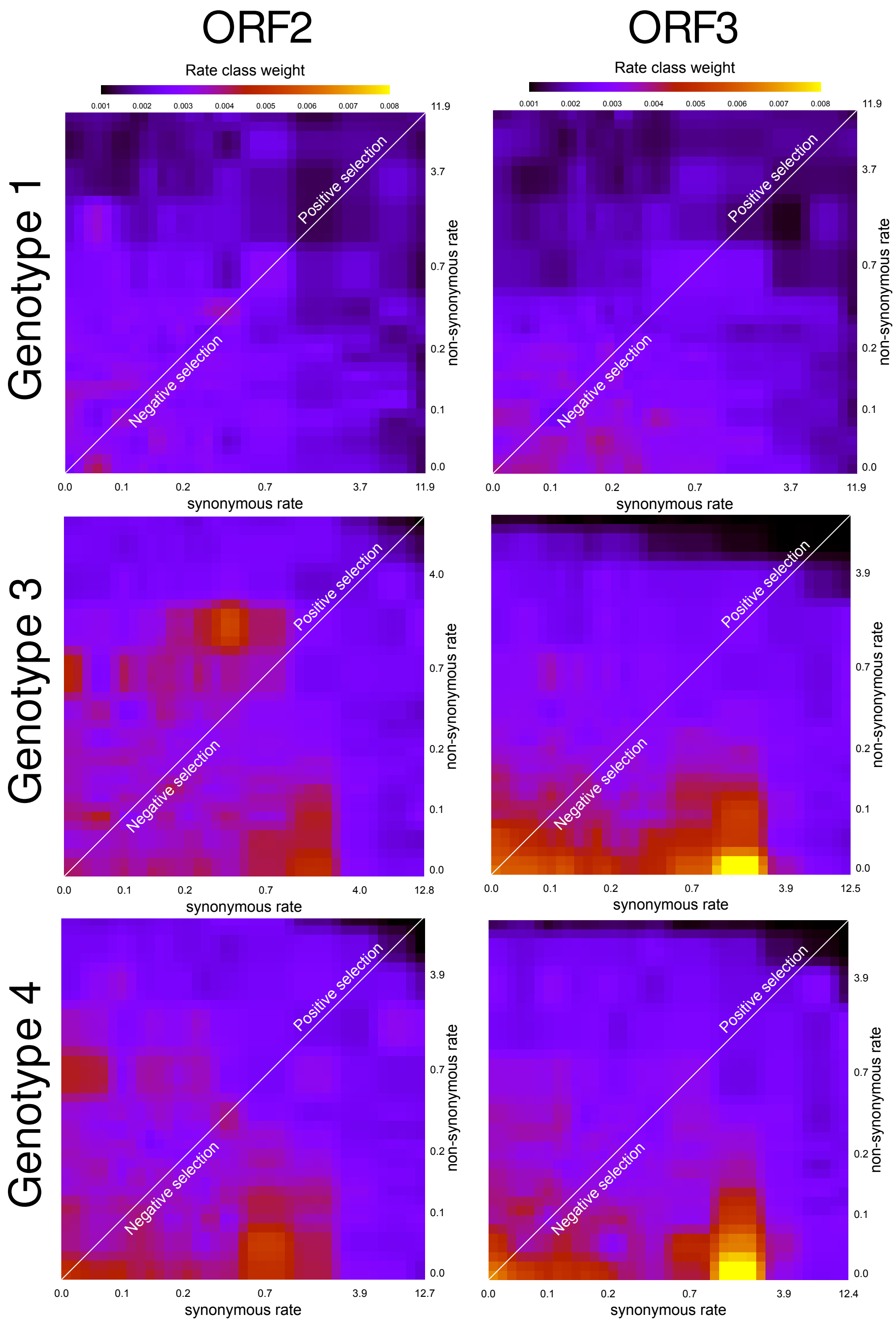
Figure 1. FUBAR analysis of concatenated ORF1 and ORF2 sequences isolated from humans (n=113). Genome-wide patterns of non-synonymous (β) and synonymous (α) substitutions per site show that HEV has a background of purifying selection with two discrete regions of elevated diversity corresponding to the hypervariable region (HVR) and the overlap region between ORF2 and ORF3, as shown on the genomic map. Sites subject to significant pervasive positive selection (FUBAR posterior probability ≥ 0.95) are shown as black circles on the x-axis. FUBAR analysis of the ORF2/3 overlap regions in their respective reading frames, showing positive selection in both frames, but with ORF2 demonstrating a stronger signal than ORF3, both in terms of the number of positively selected sites, and the magnitude of β-α.

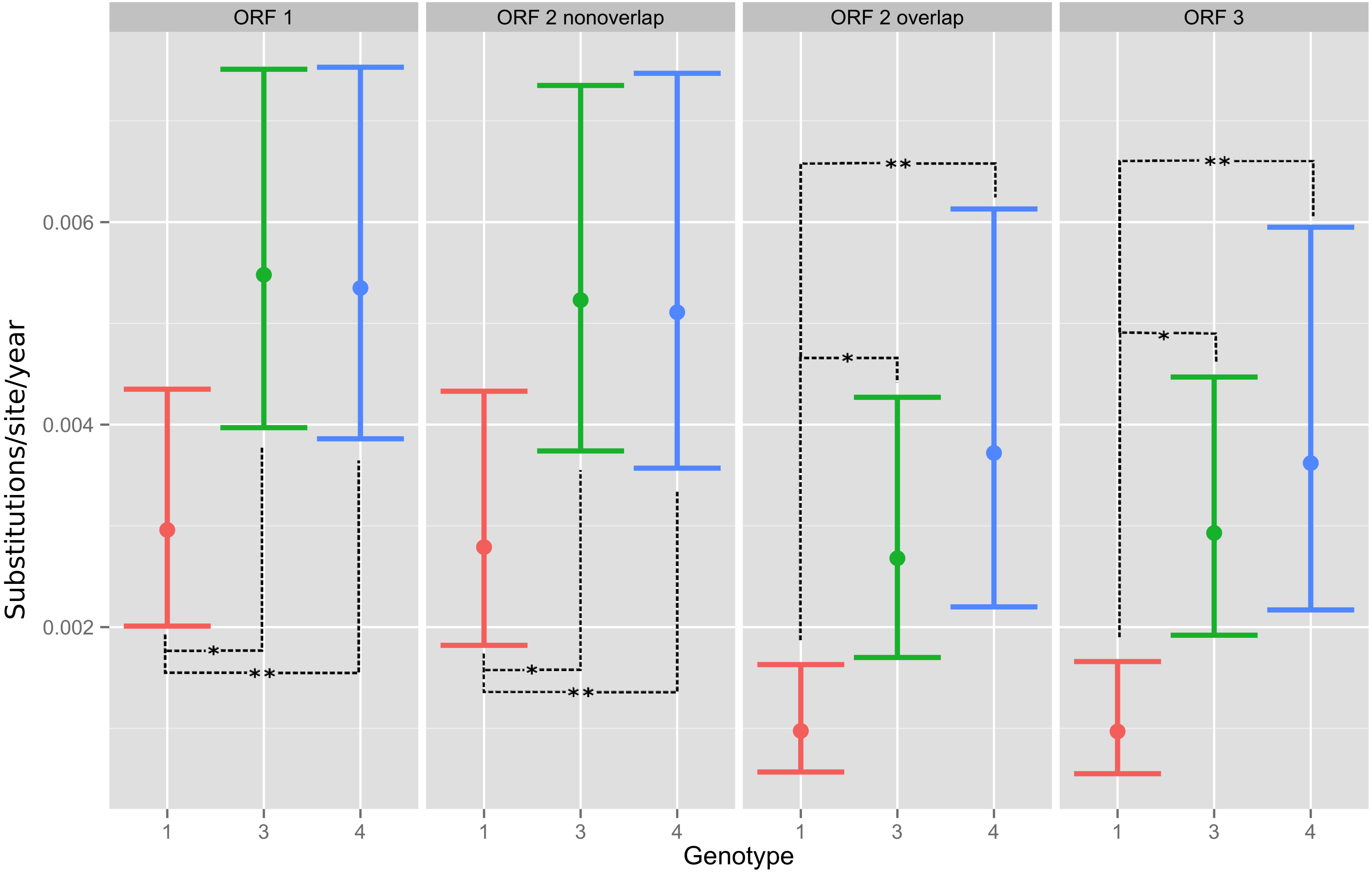
Figure 2. FUBAR Rate analysis of the ORF2/3 overlap region showing conserved patterns of groups of selected sites across genotypes. The x axis represents synonymous rates (α), while the y axis represents non-synonymous rates (β). As labelled, all sites above the α=β line positively selected, and those below are negatively selected. The plane is coloured by the weight assigned to each area by the FUBAR algorithm. All six plots use the same colouring scale, so they are directly comparable. Genotype 1 is unusual in having a very low proportion of positively selected sites. In genotypes 1 and 3 both the codon substitution model (Table 3) and RELAX procedure (Table 4) estimate that ORF 2 has significantly a stronger signal of positive selection.

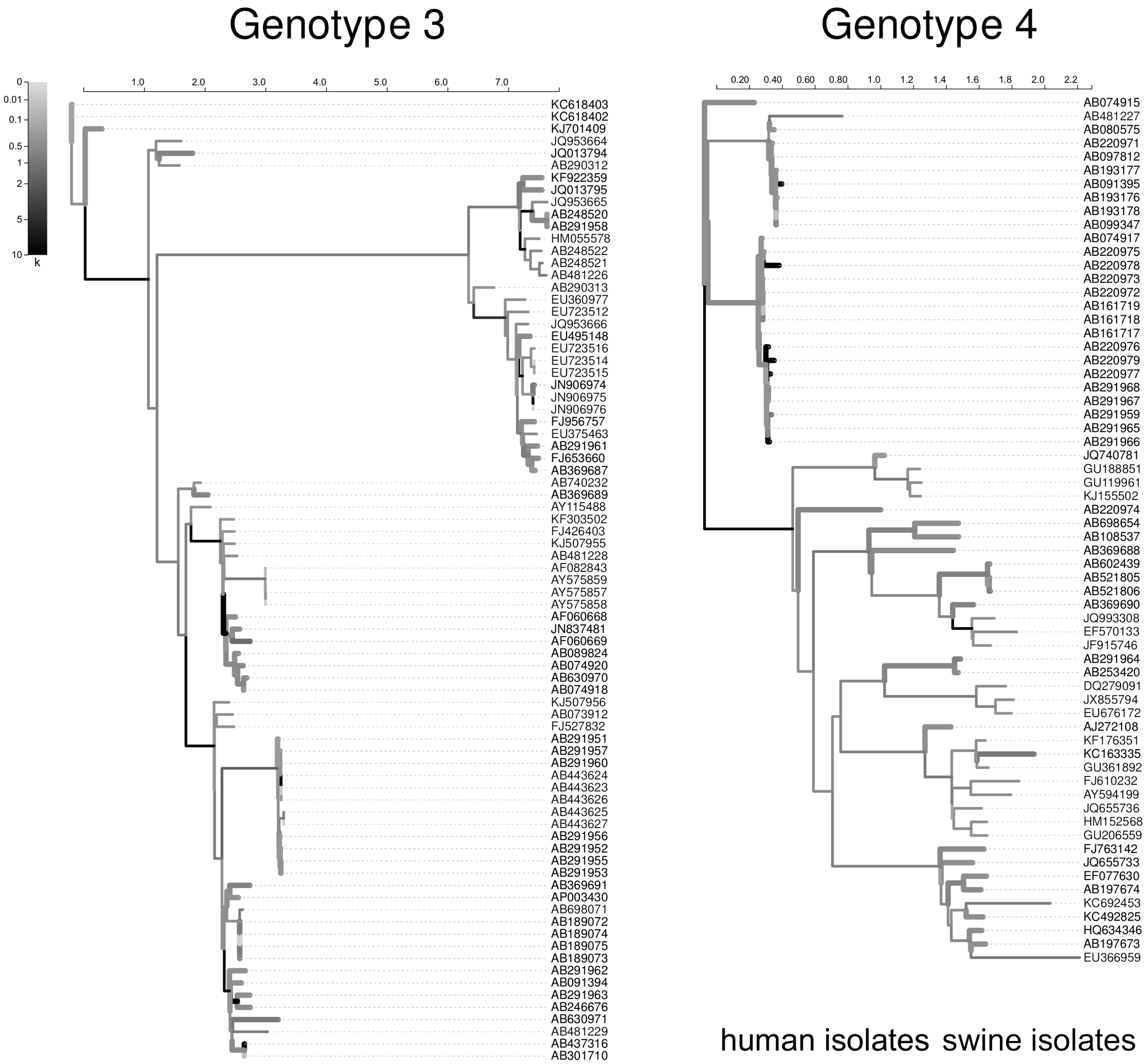
Figure 3. Estimates of evolutionary rates of HEV based on different genomic regions. Anthropotropic genotype 1 has significantly reduced relative non-synonymous evolutionary rates compared to their zoonotic counterparts across all ORFs. Genotypes 3 and 4 demonstrate similar profiles, with non significant differences across all ORFs. The overlap region of ORF 2 appears very similar to ORF 3, as they overlap extensively. Notably the non-overlap region of ORF 2 has a similar evolutionary rate profile to ORF 1, with which is does not overlap at all. Asterisks denote significance.

Figure 4. Maximum likelihood phylogenies of near-full-length sequences of HEV isolated from humans and swine. Branch lengths are in expected substitutions per nucleotide site estimated under the RELAX (59) general exploratory model. Swine isolates are labelled using muted text, and all branches labelled as 'human' are plotted using thicker lines. The k coefficients measures relaxation (k < 1) or intensification (k > 1) of positive selective pressure relative to the phylogeny-wide baseline (mean of k is constrained to be 1), represented by shades of grey. For G3, 91.5% of the bootstrapped trees supported p-value of <=0.05 or less. For G4 every single p-value of the bootstrapped trees supported p-value of < 0.05.


Full genome
β-α
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
ORF1
ORF2
HVR
ORF2/ORF3 overlap
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
Codon
OFR2 overlap
β-α
10
20
30
40
50
60
70
80
90
100
Codon
OFR3 overlap
β-α
10
20
30
40
50
60
70
80
90
100
Codon


ORF2
ORF3
Rate class weight
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
Genotype 1
Genotype 3
Genotype 4
Positive selection
Negative selection
synonymous rate
non-synonymous rate
0.0
0.1
0.2
0.7
3.7
11.9
4.0
12.8
3.9
12.5
12.7
12.4


ORF 1
ORF 2 nonoverlap
ORF 2 overlap
ORF 3
Substitutions/site/year
0.006
0.004
0.002
1
3
4
Genotype
*
**


Genotype 3
Genotype 4
0
0.01
0.1
0.5
1
2
5
10
k
1.0
2.0
3.0
4.0
5.0
6.0
7.0
KC618403
KC618402
KJ701409
JQ953664
JQ013794
AB290312
KF922359
JQ013795
JQ953665
AB248520
AB291958
HM055578
AB248522
AB248521
AB481226
AB290313
EU360977
EU723512
JQ953666
EU495148
EU723516
EU723514
EU723515
JN906974
JN906975
JN906976
FJ956757
EU375463
AB291961
FJ653660
AB369687
AB740232
AB369689
AY115488
KF303502
FJ426403
KJ507955
AB481228
AF082843
AY575859
AY575857
AY575858
AF060668
JN837481
AF060669
AB089824
AB074920
AB630970
AB074918
KJ507956
AB073912
FJ527832
AB291951
AB291957
AB291960
AB443624
AB443623
AB443626
AB443625
AB443627
AB291956
AB291952
AB291955
AB291953
AB369691
AP003430
AB698071
AB189072
AB189074
AB189075
AB189073
AB291962
AB091394
AB291963
AB246676
AB630971
AB481229
AB437316
AB301710
0.20
0.40
0.60
0.80
1.0
1.2
1.4
1.6
1.8
2.0
2.2
AB074915
AB481227
AB080575
AB220971
AB097812
AB193177
AB091395
AB193176
AB193178
AB099347
AB074917
AB220975
AB220978
AB220973
AB220972
AB161719
AB161718
AB161717
AB220976
AB220979
AB220977
AB291968
AB291967
AB291959
AB291965
AB291966
JQ740781
GU188851
GU119961
KJ155502
AB220974
AB698654
AB108537
AB369688
AB602439
AB521805
AB521806
AB369690
JQ993308
EF570133
JF915746
AB291964
AB253420
DQ279091
JX855794
EU676172
AJ272108
KF176351
KC163335
GU361892
FJ610232
AY594199
JQ655736
HM152568
GU206559
FJ763142
JQ655733
EF077630
AB197674
KC692453
KC492825
HQ634346
AB197673
EU366959
human isolates
swine isolates