1    Genotype-specific evolution of hepatitis E virus

2

3    Adam B. Brayne,[a] Bethany L. Dearlove,[a] James S. Lester,[a] Sergei L. Kosakovsky Pond,[b]

4    Simon D. W. Frost[a] #

5

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

7

8

9    Running Head: Genotype-specific evolution of hepatitis E virus

10

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

12

13

14

15

16

17

18

19    Abstract: 151 words. Other text: 5410 words

20

**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

42    first to detect positive selection in this overlap region. Genotype 1, which only

43    infects humans, appears to be evolving differently to genotypes 3 and 4, which infect

44    multiple species, possibly because genotypes 3 and 4 are unable to achieve the same

45    fitness due to repeated host jumps.

46    **Introduction**

47    Hepatitis E virus (HEV) is a non-enveloped, single stranded, positive sense RNA

48    virus, which infects around 20 million people globally each year (1). It causes large

49    propagated epidemics of acute hepatitis in Asia and Africa, and low level, sporadic

50    food-associated infections in the developed world (2, 3). Pathogenicity varies from

51    acute liver failure and up to 20% mortality in some sub-populations (for example in

52    pregnant women), to apparently asymptomatic infections in others (4). Acquired via

53    the fecal-oral route, HEV is associated with poor hygiene and living conditions. It

54    can also be acquired by eating contaminated food, including infected artiodactyls

55    (swine, deer and boar) and shellfish (4–6).

56    Mammalian HEV exists in four internationally recognised genotypes (7). Genotyping

57    is based on nucleotide divergence of the capsid open reading frame (8), and whole

58    genome phylogenetic analysis (9). Genotypes differ at epidemiological (distribution,

59    hosts) and virological (pathogenicity, translation mechanisms) levels.

60    In terms of epidemiology, there is a striking global distribution of autochthonous

61    genotypes whose origins are obscure (10): Genotype 1 is found in Asia and North

62    Africa; genotype 2 in Mexico and Southern Africa; genotype 3 in North and South

63    America, Europe and Asia; and genotype 4 almost exclusively in Japan and China. All

64 four genotypes infect humans, but only genotypes 3 and 4 infect other animals such

65 as artiodactyls. In the developed world infections are sporadic and the genotype is

66 usually the same as that in the native swine population, suggesting zoonotic

67 transmission by food or contact (2). Most likely this involves the consumption of

68 undercooked pork. In contrast in developing countries infections can be epidemic as

69 well as sporadic, with human and swine strains most often different. A recombinant

70 vaccine against HEV exists, based on its capsid protein, which has passed phase III

71 trials (11, 12). The vaccine is based on genotype 1 strains, and appears to provide

72 cross protection against at least genotype 3 (12).

73 Pathogenicity and molecular mechanisms vary between genotypes. In developed

74 countries clinical disease is rare, and seroprevalence vastly outweighs documented

75 incidence (13–15). In developing countries, the clinical presentation of HEV

76 infection tends to be more symptomatic than in the developed world. Symptoms are

77 shared with many viral illnesses and include fever, gastro-intestinal upset and

78 malaise, and liver function tests may be deranged (15). The natural history also

79 varies by demographic, with a strikingly high mortality amongst pregnant women in

80 the developing world (10-25%) and also more disease in children compared to the

81 developing world where it is elderly men that are most often symptomatic (15, 16).

82 Primate models suggest these differences in pathogenicity are associated with the

83 genotypes, as genotypes 3 and 4 produce less clinical disease in comparison to

84 genotypes 1 and 2 in rhesus monkeys (17). There are few known differences in

85 molecular mechanisms between genotypes, however genotype 4 viruses do have a

86    distinct mechanism for the translation of open reading frame 3 (ORF3), due to a

87    frame-disrupting single nucleotide insertion (18).

88    HEV has a c. ~7200 nucleotide genome comprising three partially overlapping open

89    reading frames. ORF1 encodes a nonstructural region, and ORF2 encodes the capsid

90    protein (19). The function of ORF3, which almost entirely overlaps ORF2, is not

91    totally clear. Interestingly ORF3 is not necessary for *in vitro* infection (20), but is

92    necessary for *in vivo* infection of macaques (21). It is most likely multifunctional

93    (19) and involved in pathogenesis (22–25). Most of the coding region in the HEV

94    genome is under purifying selection (26, 27), *i.e.* selection against change in the

95    amino acid sequence. Areas with an excess of amino acid substitutions, a signal of

96    positive selection, have been found in the N-terminus of ORF2 and the C-terminus of

97    ORF3, with another in the RNA dependent RNA polymerase (RdRp) region in ORF1

98    (26). Purdy *et. al.* (28) have described positive selection in the hypervariable region

99    (HVR) of ORF1; however, Smith *et. al.* (27) failed to reproduce these results with a

100    broader selection of statistical tests.

101    Phylogenetic analyses of HEV may help to shed light on evolutionary differences

102    between genotypes, which underlie the epidemiological and clinical disparities. In a

103    previous study, Chen *et. al.* (26) failed to discern any difference in selection

104    pressures between genotypes. Since 2012, the number of appropriate full genome

105    samples has increased by 150%. Using this expanded dataset, we revisit the

106    question of evolutionary differences between the genotypes of HEV, using state-of-

107    the-art methods. We focus on detecting natural selection, specifically investigating

108    regions of positive selection which stand out from a background of purifying

109    selection against non-synonymous substitutions. Our particular focus is on the

110    overlap region, making ours the first analysis of this region as a focus of positive

111    selection. We also carry out a detailed analysis of evolutionary rates, and link

112    phylogenetic findings to the virological characteristics of the genotypes.

113    **Methods**

114    **Sequence acquisition**

115    All available sequences of hepatitis E virus in Genbank (29) were obtained by

116    searching the NCBI Nucleotide Database using the taxonomic identifier (txid) 12461,

117    along with associated metadata on host, country, and date of sampling. As of August

118    6th, 2014 there were 10,041 sequences, of which 258 sequences were at least 7000

119    nucleotides long (i.e. near full length genomes).

120    **Sequence processing**

121    Open reading frames, corresponding to sequence regions between consecutive stop

122    codons, were identified for each sequence using `getorf`, part of the EMBOSS

123    package (30). ORFs 1, 2, and 3 for each sequence were identified by `blastp` (31),

124    with amino sequences of ORFs from the NCBI Reference Sequence NC_001434 as the

125    query, and translated ORF sequences as the reference. Multiple sequence alignments

126    (MSAs) for each ORF were generated using Clustal Omega (32), based on the

127    translated sequences. Nucleotide sequences were mapped on to the corresponding

128    aligned amino acid sequences using Seaview v. 4.5.0 (33). MSAs were trimmed,

129    based on the start and stop of ORFs in NC_001434, and checked manually. In order

130    to obtain a single in-frame sequence for the near-full length genome, we

131    concatenated ORFs 1 and 2. The alignments and associated inferred data are

132    available for download from github.com/veg/HEV-evolution-2015.

133    Sequences were screened for recombination using RDP4 (version 4.36 beta) (34),

134    using eight available methods; RDP (35), GENECONV (36), BootScan (37), MaxChi

135    (38), Chimaera (39), SiScan (40), PhylPro (41), LARD (42), and 3Seq (43), using

136    default settings. Following exploratory analyses to determine whether

137    recombination detection was simply an artifact of complex patterns of mutations, a

138    sequence was deemed recombinant if three or more methods had reported it as a

139    recombinant. Consistent with prior reports of recombination in HEV (26, 44, 45), we

140    identified 14 recombinant viruses, including novel recombinants (see Table 1).

141    Genotypes were assigned to each sequence by sequence similarity and phylogenetic

142    reconstruction. We used `tblastx` (from the BLAST 2.2.30+ software suite (31, 46))

143    to find the most similar sequences prototypical for each genotype; M73218

144    (genotype 1 (47)); M74506 (genotype 2 (48)); AF060668 (genotype 3 (49)); and

145    AJ272108 (genotype 4 (18)). Designations were further investigated by inspecting

146    phylogenetic reconstructions obtained using `FastTree` v2.1.8 (50). Of the 258 near-

147    full length genomes, 127 were isolated from humans, and were selected for further

148    analysis. Two sequences were excluded on the basis that they were abnormally

149    divergent from the other sequences: M74506, which is a genotype 2 virus, and

150    JQ013793, which is similar to a strain of HEV isolated from rabbits (51). Genotype-

151   specific alignments were generated and merged into a single master alignment

152   using MACSE v.1.01b (52). Sequences with a 100% identity to other isolates were

153   removed, resulting in a final dataset of 113 unique near full-length genomes isolated

154   from humans, comprised of concatenated ORF1 and ORF2 regions, with 26 genotype

155   1 sequences, 42 genotype 3 sequences, and 45 genotype 4 sequences. We split the

156   alignment into ORF1 and ORF2 regions, extracted the overlapping part of ORF3

157   from ORF2, and split ORF2 into the region overlapping ORF3, and the non-

158   overlapping region. We also identified 56 unique HEV genomes isolated from swine.

159   The swine HEV sequence alignment was merged with the human HEV dataset using

160   profile alignment in codon space using MACSE.

161   **Genome-level selection analyses**

162   Selection analyses employed a suite of phylogenetic methods, as implemented in

163   HyPhy(53) and Datamonkey (54, 55) using default settings. FUBAR (56) was used to

164   characterize pervasive selective pressures, i.e., those aggregated over all branches in

165   the phylogeny. Both an alignment-wide distribution of synonymous and non-

166   synonymous substitution rates, and site-level estimates were obtained using

167   FUBAR. MEME (57) was applied to identify individual sites subject to episodic

168   positive selection (i.e. operating along a subset of tree branches). aBSREL (58)

169   allowed us to estimate the complexity of evolutionary processes along individual

170   tree branches, and to determine which branches in the tree were subject to positive

171   selection along a subset of sites in the alignment. Finally, RELAX (59) was employed

172    to formally test whether or not the evolutionary pressures were relaxed or

173    intensified for HEV infecting human hosts relative to those infecting swine hosts.

174    So that we could formally test whether or not selection was relaxed or intensified in

175    the overlapping region of ORF3 relative to ORF2, we modified the RELAX method

176    (60) to accept two gene alignments as input. Briefly, we fit a 3-rate random effects

177    branch-site class model (61) with three $\omega$ classes to accommodate the variation in

178    selective forces across sites and branches in an unrestricted fashion jointly to both

179    alignments, while endowing each with its own branch lengths, equilibrium codon

180    frequencies, and nucleotide substitution biases. The RELAX test enforces a

181    functional relationship between the $\omega$ ratios in reference (ORF2) and test (ORF3)

182    alignments: $\omega$ ORF3 = ($\omega$ ORF2)K. The estimated value of K indicates whether

183    selection in the test frame is relaxed (K < 1) or intensified (K > 1) relative to the

184    reference frame. A likelihood ratio test of the null hypothesis (K=1), versus the

185    alternative hypothesis (K $\neq$ 1) establishes statistical significance of relaxation (or

186    intensification).

187    **Codon substitution model for overlapping regions**

188    We fitted three codon substitution models that explicitly consider whether

189    mutations are synonymous in just one of ORF2 and ORF3, or both. These models,

190    which have been previously used to screen for biologically meaningful alternative

191    reading frames in mammalian genomes (62), generate estimates of rates RXY, which

192    refer to the rates of substitutions which are synonymous (X = 0) or non-

193    synonymous (X=1) in the primary frame (ORF2), and synonymous (Y = 0) or non-

194    synonymous (Y=1) in the alternative frame (ORF3). R00 - the rate for substitutions

195    that are synonymous in both frames, is fixed at 1, and the other three rates are

196    estimated relative to R00. Maximum likelihood parameter estimates and associated

197    95% confidence intervals (profile likelihood) were calculated for a model in which

198    R01, R10, and R11 were allowed to vary freely. We also performed likelihood ratio

199    tests comparing the full model with two null models. The first null model assumes

200    that R11 is greater than one or both of R01 and R10; the expectation is that R11

201    (non-synonymous in both frames) should be less than either R01 or R10, because

202    changing both frames should be evolutionary constrained. The second null model

203    assumes that R01=R10; rejection of this null hypothesis suggests that one frame is

204    more constrained than the other.

205    **Molecular clock analyses**

206    Sequences were annotated by year of sampling. In many cases, these data were

207    obtained from Genbank records. In other cases, the primary reference was used. In

208    the cases where neither source gave the sampling year, we used the submission date

209    to Genbank as an upper bound for the sampling date, with the lower bound set as

210    the earliest known sampling year (March 1990, from (63)). To estimate the

211    evolutionary rate for genotype specific alignments whilst accommodating the

212    uncertainty in sampling times, we used a Bayesian phylogenetic approach, as

213    implemented in `MrBayes` v3.2.2 (64). A general time reversible (GTR) model was

214    fitted, with rate variation modelled as a discrete gamma distribution with 4

215    categories. Base frequencies were fixed at their empirical values, and a uniform

216    prior placed on topologies. A relaxed clock model was used, assuming that

217    evolutionary rates were drawn independently from a gamma distribution. Default

218    priors were used, with the exception of the clock rate, which was set to `lognorm(-`

219    `9,1)`. Two chains were run for 110 million generations with a burnin of 10 million,

220    thinned to give a sample of 1000 iterations. Results were processed using the `coda`

221    library (65) in R and the 95% upper and lower credible intervals were inferred from

222    the posterior distribution. Convergence was tested using manual inspection of

223    traces of parameter values, and calculation of the Gelman-Rubin statistic (66). The

224    `rv` library was used to generate credible intervals for the difference in clockrate

225    between genotypes in the same ORF. To validate the use of a relaxed clock we

226    analysed the parameter describing the variance of the rate distribution of the

227    relaxed clock, and found it to be distinct from zero with a median of 0.01914977

228    (95% credible interval=0.00115991-0.04887433), providing support for the use of a

229    relaxed clock over a strict clock. The `ggplot2` library (67) in R was used to create

230    rate plots.

231    **Host-specific patterns of evolution**

232    Human and swine HEV near full length genomes were split into genotype 3 and

233    genotype 4 alignments. Phylogenies for each genotype were reconstructed

234    separately using maximum likelihood with RAxML v.8 (68), assuming the GTR

235    model of nucleotide substitution with gamma distributed rate variation. Phylogenies

236    were rooted with `lsd` v.0.1 (69), using the median estimate of the sampling time for

237    each sequence. Terminal branches were classified as human or swine based on

238    which host they were isolated from. Interior branches were classified as

239    'human'/'swine' whenever all of their descendants were labelled as

240    'human'/'swine', following post-order tree traversal. Species-specific estimates of

241    the distribution of the $\omega$ ratio were obtained on the basis of the models

242    implemented in RELAX (59).

243    **Implementation**

244    Except where otherwise stated, selection analyses were performed using HyPhy

245    (53), using phylogenies of each region reconstructed using RAxML v.8 (68),

246    assuming the GTR model of nucleotide substitution with gamma distributed rate

247    variation, or the MG94xGTR model of codon substitution with analysis-defined

248    patterns of site-to-site and branch-to-branch rate variation. Tree visualisation was

249    carried out using the `phylotree.js` widget implemented as an extension of the D3

250    (D3js.org) JavaScript visualisation library (http://veg.github.io/hyphy-vision).

251    

252    **Results**

253    **Genome-wide patterns of selection**

254    To visually identify genomic regions under positive or purifying selection, we

255    estimated the number of non-synonymous (amino-acid changing, dN) and

256    synonymous (amino-acid preserving, dS) changes for each codon (Figure 1) using

257    the FUBAR method (56), which estimates these quantities for individual sites using

258    an Empirical Bayes procedure in the phylogenetic likelihood framework. Consistent

259    with previous findings, most of the genome was under purifying selection (dN < dS).

260    However, within each ORF, specific regions showed statistically significant evidence

261    of positive selection (dN > dS): the hypervariable region (HVR) in ORF1, the 5' end

262    of ORF2, and ORF3 (Figure 1). As the 5' end of ORF2 and ORF3 are overlapped, we

263    repeated FUBAR analysis of this area in each reading frame, finding a strong signal

264    of positive selection throughout the overlapped region of ORF2 and a weaker signal

265    in ORF3 (Figure 1).

266    **Rate variation amongst site and branches**

267    We fitted an adaptive branch-site model (58) to the alignment of 113 isolates with

268    near full length genomes. Overall, there was very strong evidence of variation in

269    selective pressure both over sites and lineages (Δ AIC = 1760 in favour of the model

270    which allows such variation), with 54 (24%) of branches supporting site-to-site

271    variation, with 2 rate classes per branch. The remaining 169 branches could be

272    adequately explained by a model where all sites evolve at a single rate. Eleven

273    branches were subject to statistically significant ($p < 0.05$ after Holm-Bonferroni

274    multiple testing correction) positive selection. Of the eleven, one belonged to

275    genotype 1 (M94177), 3 to genotype 3 (KJ701409, AF060669, and AF060668), and 7

276    to genotype 4 (AB220977, AB291964, AB291959, AB220979, AB220976,

277    AB220978, and AJ272108). In all cases, 98% or more sites were under strong

278    purifying selection ($\omega < 0.05$), and the remainder were under very strong positive

279    selection ($\omega > 50$). Interestingly, despite the fact that the estimated distribution for

280    all interior branches separating the individual genotypes had a component with $\omega >$

281    1, none rose to the level of statistical significance for positive selection, after

282    multiple test correction.

283    **Selection on individual sites in the ORF2/ORF3 overlap region**

284    We performed selection analyses on each genotype separately. Consistent with the

285    whole genome FUBAR analysis, signals of positive selection were found in the

286    overlap region. Using multiple methods for detecting selection, positive selection

287    was found in both frames of genotypes 3 and 4, whilst neither reading frame of

288    genotype 1 exhibited any significantly positively selected sites (see Table 2). This

289    trend is shown in Figure 2, which renders the genotype-specific FUBAR distribution

290    estimates for each reading frame, representing the proportion of sites evolving at

291    different nonsynonymous and synonymous rates. These selective 'fingerprints'

292    demonstrate that there are sites subject to positive selection in both reading frames

293    in enzoonotic genotypes 3 and 4, but none in the human-only genotype 1.

294    In order to further disentangle selection on different reading frames, we fitted a

295    codon substitution model (see Table 3) that considers whether mutations are non-

296    synonymous in ORF2, ORF3, or both ORF2 and ORF3. In all three genotypes, the rate

297    of substitutions that were non-synonymous at a codon level in both frames was

298    significantly lower than the rate of non-synonymous mutations in either of the

299    specific frames. This finding is consistent with a dual-coding region where both

300    frames are under purifying selection for functional conservation, on average. The

301    point estimates derived from Genotype 1 are lower than for genotypes 3 and 4,

302    hinting at stronger conservation for the former. For genotypes 1 and 3, ORF2 and

303   ORF3 are evolving at significantly different rates, when considering non-

304   synonymous substitutions affecting only one of the frames, with ORF2 experiencing

305   more of the latter. For genotype 4, the rates are statistically indistinguishable.

306   To formally test whether ORF3 is evolving differently from the overlapping region of

307   ORF2, we modified the RELAX method(59) to accept two gene alignments as input.

308   The RELAX test enforces a functional relationship between the $\omega$ ratios in reference

309   (ORF2) and test (ORF3) alignments: $\omega$ ORF3 = ($\omega$ ORF2)K. The estimated value of K

310   indicates whether selection in ORF3 is relaxed (K < 1) or intensified (K > 1) relative

311   to ORF2. A likelihood ratio test of the null hypothesis (K=1), versus the alternative

312   hypothesis (K ≠ 1) establishes statistical significance of relaxation (or

313   intensification). The application of the RELAX procedure (Table 4) suggests strong

314   relaxation of selection in ORF3 (namely, through the elimination of the positively

315   selected component) in genotypes 1 and 3, and a weak (non-significant)

316   intensification of selection in ORF3 in genotype 4. This finding of ORF 2 apparently

317   driving the signal of positive selection reproduces, by different means, the findings

318   in Table 3.

319   **Estimates of time-scaled synonymous and nonsynonymous substitution rates**

320   Differences in the rate of evolution between different genotypes could arise due to

321   different selection pressures on the genotypes (i.e. different ratios of

322   nonsynonymous to synonymous substitution), as suggested by the selection

323   pressure analyses, or could simply be due to differences in the substitution rate (i.e.

324   differences in synonymous rates), independent of selection pressure. To address

325    this question, we derived time-scaled estimates of synonymous and non-

326    synonymous rates, using the procedure described in (70). Briefly, a Maximum Clade

327    Credibility tree obtained using MrBayes was used as input to a codon analysis in

328    HyPhy, using the Muse-Gaut codon-substitution model with branch-specific α

329    (synonymous) and β (non-synonymous) rate parameters, which were used to

330    partition the fixed branch length into synonymous and non-synonymous

331    components. The conversion from expected substitutions per site to expected

332    substitutions / site / year was carried out under the assumption of a strict

333    molecular clock. The results are summarized in Table 5, and demonstrate that the

334    synonymous substitution rate of genotype 1 is approximately half that of genotypes

335    3 and 4. Whilst this is an important confounding factor, this effect merely adds to an

336    extant signal of positive selection in genotype 1 sequences, because lower dS would

337    work to elevate dN/dS for genotype 1 (for example, results in Table 3 are robust to

338    this confounding factor), it has not created the effect *de novo*.

339    **Analysis of evolutionary rates**

340    We estimated the evolutionary rate of each genotype, including information on the

341    estimated time of sampling (Table 6, Figure 3). The mean evolutionary rate was

342    similar across ORFs at approximately 0.003-0.005 substitutions per site per year.

343    Evolutionary rates of genotype 1 were significantly lower than those of genotypes 3

344    and 4 across all ORFs. Genotypes 3 and 4 demonstrate remarkably similar profiles,

345    with differences non significant across all ORFs. Table 5 shows that this is likely due

346    to both lower synonymous and non-synonymous substitution rates. Unsurprisingly,

347   the overlap region of ORF 2 appears very similar to ORF 3, as they overlap

348   extensively. More surprisingly the non-overlap region of ORF 2 has a similar

349   evolutionary rate profile to ORF 1, with which is does not overlap at all.

350   **Host-specific differences in evolution**

351   We investigated whether patterns of evolution differ not only by genotype but also

352   species of isolation. We constructed phylogenies of the human and swine lineages

353   for genotypes 3 and 4, and assigned branches as either human, swine or

354   indeterminate. For both genotypes 3 and 4, there was notable intermingling of

355   lineages (Figure 4), representing a continuous zoonotic process. Genome-wide

356   analyses of selective pressures using the null and alternative models in the RELAX

357   suite, found a slight, but statistically significant intensification of selection along

358   human branches relative to swine branches. For genotype 3, RELAX inferred the

359   intensification coefficient of $K = 1.09$ ($p = 0.013$ when compared to the null

360   hypothesis of $K = 1$). For genotype 4, the inferred values were $K = 1.12$ and $p <$

361   $0.001$. In brief, this test establishes that $\omega$ estimates on human branches are more

362   extreme (further away from $\omega = 1$, i.e., neutrality), than on swine branches. For

363   these analyses, indeterminate branches were endowed with their own $\omega$

364   distribution and branch-level relaxation/intensification coefficients, treated here as

365   nuisance parameters. For genotype 3, 91.5% of the bootstrapped trees supported p-

366   value of $<=0.05$ or less (count = 211, median p value = 0.019 (4E-5-0.0699), median

367   $K = 1.14987$ (1.0324-1.2159)). For G4 every single p-value for RELAX was $< 0.05$

368   (count = 352, median p value = 9E4 (7E-11-0.0051), median $K = 1.4059$ (1.0900-

369   1.8709)).

370 ⸻

371 **Discussion**

372 We have demonstrated differences in the evolution of hepatitis E virus (HEV)

373 between the three open reading frames, and quantified how evolutionary patterns

374 differ between genotypes. Using a high quality alignment comprising all available

375 near full length genomes, our analyses have identified and focused in on the main

376 genomic region of interest: the ORF2/ORF3 overlap region (Figure 1). Selection

377 analysis of the overlap region revealed multiple sites/regions undergoing positive

378 selection in genotypes 3 and 4, but a much weaker signal in genotype 1 (Figure 2).

379 This pattern is the same as that found in evolutionary rates, with significantly

380 reduced evolutionary rates in both ORF2 overlap and ORF3 of genotype 1 (Figure 3),

381 driven by differences in both synonymous and nonsynonymous rates. A genome-

382 wide analysis of genotype 3 and 4 isolates revealed a slight but statistically

383 significant intensification of selective pressures in human lineages compared to

384 swine lineages. We speculate, as genotype 1 viruses only infect one host and

385 genotypes 3 and 4 are enzoonotic, that genes in genotype 1 are subject to reduced

386 diversifying or balancing selection pressure as they have fine-tuned fitness by

387 specializing to their single host species. This functional constraint on amino acid

388 changes is particularly pertinent as this effect was found in both ORF2 and ORF3,

389 which are both believed to be important in the pathogen-host response as the

390 capsid protein and an immunomodulatory phosphoprotein, respectively (19). ORF1,

391 in contrast, contains housekeeping genes, which are less likely to be host-specific.

392    Cyclical host jumps seen in arboviruses, e.g., West-Nile virus, are associated with

393    purifying selection (71, 72). The concept behind this is that only substitutions

394    conferring a selection advantage in both hosts are preserved. However this

395    paradigm may not be globally applicable. *In silico* models of evolution under varying

396    selection pressure show that the rate of evolution and dN/dS can be either

397    suppressed or increased depending on how the timescale of the environmental

398    change compares to that of adaptation to the new environment (73, 74). In an

399    environment with very slow environmental fluctuations, each substitution will

400    either fix or go extinct during the epoch in which it arose, whilst in faster

401    oscillations a substitution will have the opportunity to be selected in both

402    environments (74). Therefore it may not be the case that all cyclical environments

403    induce stronger purifying selection. HEV may be an instance where the interaction

404    of oscillation period and time taken to reach a particular fitness level interact in

405    such a way as to promote diversity and a signal of positive selection. We therefore

406    postulate the signal of positive selection in those genomic regions which interact

407    with the host (ORF2 and ORF3, ORF1 contains housekeeping genes) represents a

408    cyclical but ultimately futile selection process in each species which results in a

409    phenotype which is sub-optimally fit in both. Although, interestingly, our host-

410    specific analysis provides evidence that the scales are currently tipped towards

411    optimizing for the human host.

412    Overlapping reading frames are not uncommon in RNA viruses (75), and have been

413    suggested as a mechanism of packing more genes in a limited genomic space (76).

414    Whilst other studies have found a scattering of positively selected codons in this

415   region (26), none have investigated the overlap region as a locus for positive

416   selection. Overlapping coding regions are often constrained as substitutions impact

417   two protein products instead of one, causing a reduction in evolutionary rates (77).

418   However, there is a precedent for rapid evolution in overlapped regions in both

419   viruses (e.g. PB1-F2 and PA-X in Influenza A virus (78, 79)) and mammalian

420   genomes (80), and statistical techniques designed for mammalian overlapping

421   regions (62) helped us to shed light on what is driving evolution in the overlap

422   region. Investigating selection in a region of overlapping reading frames requires

423   reading-frame aware models. Apart from a currently computationally infeasible full

424   Bayesian treatment of co-dependent evolution in multiple reading frames (e.g. see

425   (81)), two approximate approaches have been used in practice. Firstly, the

426   overlapping reading frames can be treated entirely independently, and analysed

427   using standard methods (e.g. (82)). When this approach is taken to estimate

428   synonymous and non-synonymous rates and carry out tests of selection, the

429   interpretation of results becomes difficult (e.g., how valid is the concept of a frame-

430   specific synonymous rate in this context?), and can lead to false positive results (83).

431   Secondly, codon-substitution models which correct for the "expected" context of a

432   codon in the alternative reading frame have been proposed (62, 84, 85). The benefit

433   of these models is that, while remaining computationally tractable, they directly

434   estimate frame-aware rates of synonymous and non-synonymous mutations. Such

435   models have been successfully used to perform genome-wide screens of ORFs with

436   multiple overlapping reading frames for functional constraint (62), and the

437   evolution of overlapping reading frames in Influenza A virus (84). Our analysis of

438   the overlapping reading frames shows a significant difference between the rate of

439   substitutions that were non-synonymous in both frames compared to only one

440   (Table 3), indicative of positive selection. Applying this model also allowed us to

441   find out which reading frame (and therefore likely gene product) was driving the

442   positive signal in this region. Although both reading frames are subject to positive

443   selection, the ORF2 overlap region appears to be driving selection (at least in

444   genotypes 1 and 3).

445   Evolutionary rates showed significant differences between the anthropotropic and

446   zoonotic genotypes. Across all genomic regions genotype 1 had significantly lower

447   evolutionary rates than genotypes 3 and 4, whilst genotypes 3 and 4 had

448   remarkably similar values. Evolutionary rates inferred from the posterior

449   distribution were typical of an RNA virus (86) and related viruses e.g. norovirus

450   (87). We, like Nakano *et. al.* (2012) and Purdy *et. al.* (2012), found a relaxed clock

451   most appropriate to reflect the variation in substitution rates between branches in

452   HEV, although our estimates of evolutionary rate are higher than those reported

453   previously (88–90). It should be noted that apparent evolutionary rates show time

454   dependency, with an elevation towards the present due to transient unfixed

455   substitutions, and apparent reduction in the past due to saturation (91).

456   Interestingly the ORF 2 overlap region has a very distinct profile of evolutionary

457   rates across all genotypes when compared with the non-overlap region. The non-

458   overlap region is strikingly similar to the ORF 1 profile, which is believed to contain

459   housekeeping genes.

460    Our analyses of the host specific patterns of evolution are important in showing that

461    the differences described above are largely genotype, not host species, dependent.

462    For genotypes 3 and 4, we detected a slight increase in selection intensity in human-

463    associated viral lineages compared to swine associated viral lineages, in contrast to

464    the large differences between genotypes. The construction of phylogenies

465    demonstrated intermingling of swine and human lineages, and suggest a high rate of

466    host jumps indicative of the frequent transmission between swine and humans and

467    back again. The transmission of HEV from humans to swine has been demonstrated

468    extensively in laboratory settings (92, 93), however its frequency and mechanism in

469    the wild remain unclear (94). As phylogenies do not contain independent

470    information on the direction of transmission, it is hard to demonstrate such 'reverse

471    zoonoses' from sequence data alone.

472    Our study represents the most comprehensive HEV sequence analysis to date. It is

473    important, however, to note the limitations in the publicly available data. Genotypes

474    3 and 4 dominate in the developed world, whilst genotypes 1 and 2 are found in the

475    developing world (4). This global differential distribution of genotypes may be an

476    important confounder, as in fact the virus is not interacting with a single

477    homogeneous human host, but rather different clades of virus are interacting with

478    specific groups of human hosts. These groups are likely to differ significantly, e.g. in

479    the population composition of Human Leukocyte Antigen alleles (95), which in turn

480    imposes differential selective pressures on the pathogen as part of host-pathogen

481    interactions. As is the case for most pathogens, sampling is heavily biased by

482    location. There are many samples from Europe and East Asia, but few from

483     Australasia and Africa, and there are many countries for which there are no

484     sequence data. Furthermore, little is known about genotype 2, with too few full

485     length viral genomes publicly available to build a reliable alignment, so studies

486     either omit it (26), or have low statistical power (96).

487     Hepatitis E virus is of increasing interest to public health officials and clinicians.

488     Attention in the developed world to date has been limited, partly due to the acute

489     nature of the infection in healthy individuals and the apparently asymptomatic

490     nature of infection in swine. However, the emergence of new strains of HEV, such as

491     one recently documented in the U.K. (97), emphasise the need for continuing

492     surveillance and characterisation of this pathogen.

493

504    provided technical assistance, including generation of figures, and revised the draft

505    paper. SF devised and initiated the collaborative project, designed the data

506    collection and analysis strategies and revised the draft paper.

507

508    **References**

509    1. **Rein DB**, **Stevens GA**, **Theaker J**, **Wittenborn JS**, **Wiersma ST**. 2012. The global

510    burden of hepatitis e virus genotypes 1 and 2 in 2005. Hepatology **55**:988–997.

511    2. **Meng X-J**. 2013. Zoonotic and foodborne transmission of hepatitis E virus. Semin

512    Liver Dis **33**:41–49.

513    3. **Berto A**, **Martelli F**, **Grierson S**, **Banks M**. 2012. Hepatitis E virus in pork food

514    chain, united kingdom, 2009-2010. Emerg Infect Dis **18**:1358–1360.

515    4. **Teshale EH**, **Hu DJ**, **Holmberg SD**. 2010. The two faces of hepatitis E virus. Clin

516    Infect Dis **51**:328–334.

517    5. **Lewis H**, **Wichmann O**, **Duizer E**. 2010. Transmission routes and risk factors for

518    autochthonous hepatitis E virus infection in europe: a systematic review. Epidemiol

519    Infect **138**:145–166.

520    6. **Banks M**, **Bendall R**, **Grierson S**, **Heath G**, **Mitchell J**, **Dalton H**. 2004. Human

521    and porcine hepatitis E virus strains, United Kingdom. Emerg Infect Dis **10**:953–955.

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

523    8. **Worm HC**, **Poel WHM van der**, **Brandstatter G**. 2002. Hepatitis E: an overview.

524    Microbes Infect **4**:657–666.

525    9. **Lu L**, **Li CH**, **Hagedorn CH**. 2006. Phylogenetic analysis of global hepatitis E virus

526    sequences: genetic diversity, subtypes and zoonosis. Rev Med Virol **16**:5–36.

527    10. **Schlauder GG**, **Mushahwar IK**. 2001. Genetic heterogeneity of hepatitis E virus.

528    J Med Virol **65**:282–292.

529    11. **Shrestha MP**, **Scott RM**, **Joshi DM**, **Mammen Jr MP**, **Thapa GB**, **Thapa N**, **Myint**

530    **KSA**, **Fourneau M**, **Kuschner RA**, **Shrestha SK, others**. 2007. Safety and efficacy of

531    a recombinant hepatitis e vaccine. New England Journal of Medicine **356**:895–903.

532    12. **Zhu F-C**, **Zhang J**, **Zhang X-F**, **Zhou C**, **Wang Z-Z**, **Huang S-J**, **Wang H**, **Yang C-L**,

533    **Jiang H-M**, **Cai J-P, others**. 2010. Efficacy and safety of a recombinant hepatitis e

534    vaccine in healthy adults: a large-scale, randomised, double-blind placebo-

535    controlled, phase 3 trial. The Lancet **376**:895–902.

536    13. **Kuniholm MH**, **Purcell RH**, **McQuillan GM**, **Engle RE**, **Wasley A**, **Nelson KE**.

537    2009. Epidemiology of hepatitis e virus in the united states: results from the third

538    national health and nutrition examination survey, 1988–1994. Journal of Infectious

539    Diseases **200**:48–56.

540    14. **Bendall R**, **Ellis V**, **Ijaz S**, **Ali R**, **Dalton H**. 2010. A comparison of two

541    commercially available anti-hEV igG kits and a re-evaluation of anti-hEV igG

542    seroprevalence data in developed countries. Journal of medical virology **82**:799–

543    805.

544  15. **Kamar N**, **Bendall R**, **Legrand-Abravanel F**, **Xia N-S**, **Ijaz S**, **Izopet J**, **Dalton**

545  **HR**. 2012. Hepatitis e. The Lancet **379**:2477–2488.

546  16. **Dalton HR**, **Stableforth W**, **Thurairajah P**, **Hazeldine S**, **Remnarace R**, **Usama**

547  **W**, **Farrington L**, **Hamad N**, **Sieberhagen C**, **Ellis V**, **others**. 2008. Autochthonous

548  hepatitis e in southwest england: natural history, complications and seasonal

549  variation, and hepatitis e virus igG seroprevalence in blood donors, the elderly and

550  patients with chronic liver disease. European journal of gastroenterology &

551  hepatology **20**:784–790.

552  17. **Purcell RH**, **Engle RE**, **Govindarajan S**, **Herbert R**, **St Claire M**, **Elkins WR**,

553  **Cook A**, **Shaver C**, **Beauregard S Michelle**, **Emerson S**. 2013. Pathobiology of

554  hepatitis e: lessons learned from primate models. Emerging Microbes & Infections

555  **2**:e9.

556  18. **Wang Y**, **Zhang H**, **Ling R**, **Li H**, **Harrison TJ**. 2000. The complete sequence of

557  hepatitis E virus genotype 4 reveals an alternative strategy for translation of open

558  reading frames 2 and 3. J Gen Virol **81**:1675–1686.

559  19. **Cao D**, **Meng X-J**. 2012. Molecular biology and replication of hepatitis e virus.

560  Emerging microbes & infections **1**:e17.

561  20. **Emerson SU**, **Nguyen H**, **Torian U**, **Purcell RH**. 2006. ORF3 protein of hepatitis

562  e virus is not required for replication, virion assembly, or infection of hepatoma cells

563  in vitro. Journal of virology **80**:10457–10464.

564    21. **Graff J**, **Nguyen H**, **Yu C**, **Elkins WR**, **Claire MS**, **Purcell RH**, **Emerson SU**. 2005.

565    The open reading frame 3 gene of hepatitis e virus contains a cis-reactive element

566    and encodes a protein required for infection of macaques. Journal of virology

567    **79**:6680–6689.

568    22. **Chandra V**, **Kar-Roy A**, **Kumari S**, **Mayor S**, **Jameel S**. 2008. The hepatitis E

569    virus ORF3 protein modulates epidermal growth factor receptor trafficking, STAT3

570    translocation, and the acute-phase response. J Virol **82**:7100–7110.

571    23. **Tyagi S**, **Korkaya H**, **Zafrullah M**, **Jameel S**, **Lal SK**. 2002. The phosphorylated

572    form of the oRF3 protein of hepatitis e virus interacts with its non-glycosylated form

573    of the major capsid protein, oRF2. Journal of Biological Chemistry **277**:22759–

574    22767.

575    24. **Tyagi S**, **Surjit M**, **Lal SK**. 2005. The 41-amino-acid c-terminal region of the

576    hepatitis e virus oRF3 protein interacts with bikunin, a kunitz-type serine protease

577    inhibitor. Journal of virology **79**:12081–12087.

578    25. **Tyagi S**, **Surjit M**, **Roy AK**, **Jameel S**, **Lal SK**. 2004. The oRF3 protein of hepatitis

579    e virus interacts with liver-specific $\alpha$1-microglobulin and its precursor $\alpha$1-

580    microglobulin/bikunin precursor (aMBP) and expedites their export from the

581    hepatocyte. Journal of Biological Chemistry **279**:29308–29319.

582    26. **Chen X**, **Zhang Q**, **He C**, **Zhang L**, **Li J**, **Zhang W**, **Cao W**, **Lv Y-G**, **Liu Z**, **Zhang J-X**,

583    **Shao Z-J**. 2012. Recombination and natural selection in hepatitis E virus genotypes. J

584    Med Virol **84**:1396–1407.

585    27. **Smith DB**, **Vanek J**, **Ramalingam S**, **Johannessen I**, **Templeton K**, **Simmonds**

586    **P**. 2012. Evolution of the hepatitis E virus hypervariable region. J Gen Virol

587    **93**:2408–2418.

588    28. **Purdy MA**, **Lara J**, **Khudyakov YE**. 2012. The hepatitis E virus polyproline

589    region is involved in viral adaptation. PLoS One **7**:e35974–e35974.

590    29. **Benson DA**, **Cavanaugh M**, **Clark K**, **Karsch-Mizrachi I**, **Lipman DJ**, **Ostell J**,

591    **Sayers EW**. 2013. GenBank. Nucleic Acids Res **41**:D36–D42.

592    30. **Rice P**, **Longden I**, **Bleasby A**. 2000. EMBOSS: the european molecular biology

593    open software suite. Trends in genetics **16**:276–277.

594    31. **Camacho C**, **Coulouris G**, **Avagyan V**, **Ma N**, **Papadopoulos J**, **Bealer K**,

595    **Madden TL**. 2009. BLAST+: architecture and applications. BMC Bioinformatics

596    **10**:421.

597    32. **Sievers F**, **Higgins DG**. 2014. Clustal omega, accurate alignment of very large

598    numbers of sequences. Methods Mol Biol **1079**:105–116.

599    33. **Gouy M**, **Guindon S**, **Gascuel O**. 2010. SeaView version 4: A multiplatform

600    graphical user interface for sequence alignment and phylogenetic tree building. Mol

601    Biol Evol **27**:221–224.

602    34. **Martin DP**, **Murrell B**, **Golden M**, **Khoosal A**, **Muhire B**. 2015. RDP4: Detection

603    and analysis of recombination patterns in virus genomes. Virus Evol **1**:vev003.

604  35. **Martin D**, **Rybicki E**. 2000. RDP: detection of recombination amongst aligned

605  sequences. Bioinformatics **16**:562–563.

606  36. **Padidam M**, **Sawyer S**, **Fauquet CM**. 1999. Possible emergence of new

607  geminiviruses by frequent recombination. Virology **265**:218–225.

608  37. **Martin DP**, **Posada D**, **Crandall KA**, **Williamson C**. 2005. A modified bootscan

609  algorithm for automated identification of recombinant sequences and

610  recombination breakpoints. AIDS Res Hum Retroviruses **21**:98–102.

611  38. **Maynard Smith J**. 1992. Analyzing the mosaic structure of genes. J Mol Evol

612  **34**:126–129.

613  39. **Posada D**, **Crandall KA**. 2001. Evaluation of methods for detecting

614  recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A

615  **98**:13757–13762.

616  40. **Gibbs MJ**, **Armstrong JS**, **Gibbs AJ**. 2000. Sister-scanning: a monte carlo

617  procedure for assessing signals in recombinant sequences. Bioinformatics **16**:573–

618  582.

619  41. **Weiller GF**. 1998. Phylogenetic profiles: a graphical method for detecting

620  genetic recombinations in homologous sequences. Mol Biol Evol **15**:326–335.

621  42. **Holmes EC**, **Worobey M**, **Rambaut A**. 1999. Phylogenetic evidence for

622  recombination in dengue virus. Mol Biol Evol **16**:405–409.

623    43. **Boni MF**, **Posada D**, **Feldman MW**. 2007. An exact nonparametric method for

624    inferring mosaic structure in sequence triplets. Genetics **176**:1035–1047.

625    44. **Cuyck H van**, **Fan J**, **Robertson DL**, **Roques P**. 2005. Evidence of recombination

626    between divergent hepatitis E viruses. J Virol **79**:9306–9314.

627    45. **Wang H**, **Zhang W**, **Ni B**, **Shen H**, **Song Y**, **Wang X**, **Shao S**, **Hua X**, **Cui L**. 2010.

628    Recombination analysis reveals a double recombination event in hepatitis E virus.

629    Virol J **7**.

630    46. **Altschul SF**, **Gish W**, **Miller W**, **Myers EW**, **Lipman DJ**. 1990. Basic local

631    alignment search tool. J Mol Biol **215**:403–410.

632    47. **Tam AW**, **Smith MM**, **Guerra ME**, **Huang CC**, **Bradley DW**, **Fry KE**, **Reyes GR**.

633    1991. Hepatitis e virus (hEV): molecular cloning and sequencing of the full-length

634    viral genome. Virology **185**:120–131.

635    48. **Huang CC**, **Nguyen D**, **Fernandez J**, **Yun KY**, **Fry KE**, **Bradley DW**, **Tam AW**,

636    **Reyes GR**. 1992. Molecular cloning and sequencing of the mexico isolate of hepatitis

637    E virus (HEV). Virology **191**:550–558.

638    49. **Schlauder GG**, **Dawson GJ**, **Erker JC**, **Kwo PY**, **Knigge MF**, **Smalley DL**,

639    **Rosenblatt JE**, **Desai SM**, **Mushahwar IK**. 1998. The sequence and phylogenetic

640    analysis of a novel hepatitis E virus isolated from a patient with acute hepatitis

641    reported in the united states. J Gen Virol **79 ( Pt 3)**:447–456.

642    50. **Price MN**, **Dehal PS**, **Arkin AP**. 2010. FastTree 2–approximately maximum-

643    likelihood trees for large alignments. PLoS One **5**:e9490.

644    51. **Izopet J**, **Dubois M**, **Bertagnoli S**, **Lhomme S**, **Marchandeau S**, **Boucher S**,

645    **Kamar N**, **Abravanel F**, **Guérin J-L**. 2012. Hepatitis E virus strains in rabbits and

646    evidence of a closely related strain in humans, France. Emerg Infect Dis **18**:1274–

647    1281.

648    52. **Ranwez V**, **Harispe S**, **Delsuc F**, **Douzery EJ**. 2011. MACSE: Multiple alignment

649    of coding SEquences accounting for frameshifts and stop codons. PLoS One

650    **6**:e22594.

651    53. **Pond SLK**, **Frost SDW**, **Muse SV**. 2005. HyPhy: hypothesis testing using

652    phylogenies. Bioinformatics **21**:676–679.

653    54. **Pond SLK, Frost SDW**. 2005. Datamonkey: rapid detection of selective pressure

654    on individual sites of codon alignments. Bioinformatics **21**:2531–2533.

655    55. **Delport W**, **Poon AFY**, **Frost SDW**, **Pond SLK**. 2010. Datamonkey 2010: a suite

656    of phylogenetic analysis tools for evolutionary biology. Bioinformatics **26**:2455–

657    2457.

658    56. **Murrell B**, **Moola S**, **Mabona A**, **Weighill T**, **Sheward D**, **Kosakovsky Pond SL**,

659    **Scheffler K**. 2013. FUBAR: a fast, unconstrained bayesian approximation for

660    inferring selection. Mol Biol Evol **30**:1196–205.

661    57. **Murrell B**, **Wertheim JO**, **Moola S**, **Weighill T**, **Scheffler K**, **Pond SK**. 2012.

662    Detecting individual sites subject to episodic diversifying selection. PLoS Genetics

663    **8**:e1002764–e1002764.

664    58. **Smith MD**, **Wertheim JO**, **Weaver S**, **Murrell B**, **Scheffler K**, **Kosakovsky Pond**

665    **SL**. 2015. Less is more: an adaptive branch-site random effects model for efficient

666    detection of episodic diversifying selection. Mol Biol Evol **32**:1342–53.

667    59. **Wertheim JO**, **Murrell B**, **Smith MD**, **Kosakovsky Pond SL**, **Scheffler K**. 2015.

668    RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol

669    **32**:820–32.

670    60. **Wertheim JO**, **Murrell B**, **Smith MD**, **Pond SLK**, **Scheffler K**. 2015. RELAX:

671    detecting relaxed selection in a phylogenetic framework. Molecular biology and

672    evolution **32**:820–832.

673    61. **Pond SLK**, **Murrell B**, **Fourment M**, **Frost SD**, **Delport W**, **Scheffler K**. 2011. A

674    random effects branch-site model for detecting episodic diversifying selection.

675    Molecular biology and evolution msr125.

676    62. **Chung W-Y**, **Wadhawan S**, **Szklarczyk R**, **Pond SK**, **Nekrutenko A**. 2007. A first

677    look at ARFome: Dual-coding genes in mammalian genomes. PLoS Comput Biol

678    **3**:855–861.

679    63. **Reyes GR**, **Purdy MA**, **Kim JP**, **Luk KC**, **Young LM**, **Fry KE**, **Bradley DW**. 1990.

680    Isolation of a cDNA from the virus responsible for enterically transmitted non-A,

681    non-B hepatitis. Science **247**:1335–1339.

682    64. **Ronquist F**, **Teslenko M**, **Mark P van der**, **Ayres DL**, **Darling A**, **Hohna S**,

683    **Larget B**, **Liu L**, **Suchard MA**, **Huelsenbeck JP**. 2012. MrBayes 3.2: Efficient

684     bayesian phylogenetic inference and model choice across a large model space. Syst

685     Biol **61**:539–542.

686     65. **Plummer M**, **Best N**, **Cowles K**, **Vines K**. 2006. CODA: Convergence diagnosis

687     and output analysis for mCMC. R News **6**:7–11.

688     66. **Gelman A**, **Goegebeur Y**, **Tuerlinckx F**, **Van Mechelen I**. 2000. Diagnostic

689     checks for discrete data regression models using posterior predictive simulations. J

690     R Stat Soc Ser C Appl Stat **49**:247–268.

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

692     York.

693     68. **Stamatakis A**. 2014. RAxML version 8: a tool for phylogenetic analysis and post-

694     analysis of large phylogenies. Bioinformatics **30**:1312–1313.

695     69. **To T-H**, **Jung M**, **Lycett S**, **Gascuel O**. 2015. Fast dating using least-squares

696     criteria and algorithms. Systematic biology syv068.

697     70. **Lemey P**, **Pond SLK**, **Drummond AJ**, **Pybus OG**, **Shapiro B**, **Barroso H**, **Taveira**

698     **N**, **Rambaut A**. 2007. Synonymous substitution rates predict hIV disease

699     progression as a result of underlying replication dynamics. PLoS Comput Biol **3**:e29.

700     71. **Coffey LL**, **Forrester N**, **Tsetsarkin K**, **Vasilakis N**, **Weaver SC**. 2013. Factors

701     shaping the adaptive landscape for arboviruses: implications for the emergence of

702     disease. Future microbiology **8**:155–176.

703 72. **Parameswaran P**, **Charlebois P**, **Tellez Y**, **Nunez A**, **Ryan EM**, **Malboeuf CM**,

704 **Levin JZ**, **Lennon NJ**, **Balmaseda A**, **Harris E**, **others**. 2012. Genome-wide patterns

705 of intrahuman dengue virus diversity reveal associations with viral phylogenetic

706 clade and interhost diversity. Journal of virology **86**:8546–8558.

707 73. **Kashtan N**, **Noor E**, **Alon U**. 2007. Varying environments can speed up

708 evolution. Proceedings of the National Academy of Sciences **104**:13711–13716.

709 74. **Cvijović I**, **Good BH**, **Jerison ER**, **Desai MM**. 2015. Fate of a mutation in a

710 fluctuating environment. Proc Natl Acad Sci U S A **112**:E5021–E5028.

711 75. **Neuhaus K**, **Oelke D**, **Fürst D**, **Scherer S**, **Keim DA**. 2010. Towards automatic

712 detecting of overlapping genes-clustered bLAST analysis of viral genomes. Springer.

713 76. **Chirico N**, **Vianelli A**, **Belshaw R**. 2010. Why genes overlap in viruses.

714 Proceedings of the Royal Society of London B: Biological Sciences **277**:3809–3817.

715 77. **Simon-Loriere E**, **Holmes EC**, **Pagan I**. 2013. The effect of gene overlapping on

716 the rate of RNA virus evol. Mol Biol Evol **30**:1916–1928.

717 78. **Suzuki Y**. 2006. Natural selection on the influenza virus genome. Molecular

718 biology and evolution **23**:1902–1911.

719 79. **Jagger B**, **Wise H**, **Kash J**, **Walters K-A**, **Wills N**, **Xiao Y-L**, **Dunfee R**,

720 **Schwartzman L**, **Ozinsky A**, **Bell G**, **others**. 2012. An overlapping protein-coding

721 region in influenza a virus segment 3 modulates the host response. Science

722 **337**:199–204.

723     80. **Szklarczyk R**, **Heringa J**, **Pond SK**, **Nekrutenko A**. 2007. Rapid asymmetric

724     evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its

725     function. Proc Natl Acad Sci U S A **104**:12807–12812.

726     81. **Pedersen A-MK**, **Jensen JL**. 2001. A dependent-rates model and an mCMC-

727     based methodology for the maximum-likelihood analysis of sequences with

728     overlapping reading frames. Molecular Biology and Evolution **18**:763–776.

729     82. **Obenauer JC**, **Denson J**, **Mehta PK**, **Su X**, **Mukatira S**, **Finkelstein DB**, **Xu X**,

730     **Wang J**, **Ma J**, **Fan Y**, **others**. 2006. Large-scale sequence analysis of avian influenza

731     isolates. Science **311**:1576–1580.

732     83. **Holmes EC**, **Lipman DJ**, **Zamarin D**, **Yewdell JW**. 2006. Comment on" large-

733     Scale sequence analysis of avian influenza isolates". Science **313**:1573–1573.

734     84. **Sabath N**, **Landan G**, **Graur D**. 2008. A method for the simultaneous estimation

735     of selection intensities in overlapping genes. PLoS One **3**:e3996.

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

737     one **9**.

738     86. **Jenkins GM**, **Rambaut A**, **Pybus OG**, **Holmes EC**. 2002. Rates of molecular

739     evolution in RNA viruses: A quantitative phylogenetic analysis. J Mol Evol **54**:156–

740     165.

741     87. **Cotten M**, **Petrova V**, **Phan MV**, **Rabaa MA**, **Watson SJ**, **Ong SH**, **Kellam P**,

742     **Baker S**. 2014. Deep sequencing of norovirus genomes defines evolutionary

743     patterns in an urban tropical setting. J Virol **88**:11056–11069.

744     88. **Takahashi K**, **Toyota J**, **Karino Y**, **Kang JH**, **Maekubo H**, **Abe N**, **Mishiro S**.

745     2004. Estimation of the mutation rate of hepatitis E virus based on a set of closely

746     related 7.5-year-apart isolates from sapporo, japan. Hepatol Res **29**:212–215.

747     89. **Purdy MA**, **Khudyakov YE**. 2010. Evolutionary history and population dynamics

748     of hepatitis E virus. Plos One **5**:9.

749     90. **Nakano T**, **Takahashi K**, **Pybus OG**, **Hashimoto N**, **Kato H**, **Okano H**,

750     **Kobayashi M**, **Fujita N**, **Shiraki K**, **Takei Y**, **Ayada M**, **Arai M**, **Okamoto H**, **Mishiro**

751     **S**. 2012. New findings regarding the epidemic history and population dynamics of

752     japan-indigenous genotype 3 hepatitis E virus inferred by molecular evolution. Liver

753     Int **32**:675–688.

754     91. **Ho SY**, **Shapiro B**, **Phillips MJ**, **Cooper A**, **Drummond AJ**. 2007. Evidence for

755     time dependency of molecular rate estimates. Systematic biology **56**:515–522.

756     92. **Meng X-J**, **Halbur PG**, **Shapiro MS**, **Govindarajan S**, **Bruna JD**, **Mushahwar IK**,

757     **Purcell RH**, **Emerson SU**. 1998. Genetic and experimental evidence for cross-

758     species infection by swine hepatitis E virus. J Virol **72**:9714–9721.

759     93. **Feagins A**, **Opriessnig T**, **Huang Y**, **Halbur P**, **Meng X**. 2008. Cross-species

760     infection of specific-pathogen-free pigs by a genotype 4 strain of human hepatitis E

761     virus. J Med Virol **80**:1379.

762     94. **Messenger AM**, **Barnes AN**, **Gray GC**. 2014. Reverse zoonotic disease

763     transmission (zooanthroponosis): a systematic review of seldom-documented

764     human biological threats to animals. PloS one **9**:e89055.

765     95. **Buhler S**, **Sanchez-Mazas A**. 2011. HLA DNA sequence variation among human

766     populations: molecular signatures of demographic and selective events. PLoS One

767     **6**:e14643.

768     96. **Okamoto H**. 2007. Genetic variability and evolution of hepatitis E virus. Virus

769     Res **127**:216–228.

770     97. **Ijaz S**, **Said B**, **Boxall E**, **Smit E**, **Morgan D**, **Tedder RS**. 2014. Indigenous

771     hepatitis E in england and wales from 2003 to 2012: evidence of an emerging novel

772     phylotype of viruses. J Infect Dis **209**:1212–1218.

773

**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 |

776 Table 1: Details of recombinants found. 14 recombinant HEV sequences were

777 identified in the 258 near full length genomes, generated by concatenating ORF1

778 and ORF2, by screening with RDP4 (version 4.36 beta) (34). With the exception of

779 KJ013414 and KJ013415, which shared a recombinant structure, all recombinants

780 were unique. Three recombinants had been previously described (see

781 Recombination reference column). The table also shows the genotype of the

782 recombinant, the host it was isolated from, and the putative major and minor

783 parents.

784

| 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) |

785  Table 2: The number of positively selected codon sites in each reading frame of each

786  genotype of the overlap region (numbers in parentheses show how many sites were

787  shared between MEME and FUBAR sets). Genotype 1 lacks any positively selected

788  sites, meanwhile genotypes 3 and 4 produce a consistent signal of positively

789  selected sites in both reading frames. Note that MEME is generally more sensitive,

790  because it can detect selection on a subset of viral lineages, whilst FUBAR pools the

791  signal of selection from all branches.

792

| 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) | | |

793   Table 3.  Estimates of substitution rates that result in non-synonymous changes in at

794   least one frame, relative to the rate of substitutions that are synonymous in both

795   frames. A dimensionless metric, based on the model from Chung *et al.* (62). The last

796   two columns show LRT-based p-values for rejecting the corresponding null

797   hypotheses. Genotypes 3 and 4 demonstrate highly significant reading frame

798   specific positive selection with ORF 2 convincingly driving the signal in genotype 3

799   but not 4 (rejection of null hypothesis). Genotype 1 has a lower background rate of

800   non synonymous substitutions although does achieve significance, with the ORF 2

801   rate again significantly higher than ORF 3 rate.

802

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

803 Table 4. Application of the RELAX procedure suggests strong relaxation of selection

804 (namely, through the elimination of the positive selected component) in ORF3 of

805 genotypes 1 and 3 relative to ORF2, and a weak, non-significant intensification of

806 selection in genotype 4 of ORF3 relative to ORF2. This suggests that ORF2 and ORF3

807 are evolving differently, and ORF 2 is more responsible than ORF 3 for the signal of

808 positive selection in genotypes 1 and 3.

809

| 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 |

810     Table 5. Estimation of genotype specific synonymous substitution rates and non-

811     synonymous substitution rates performed after Lemey *et. al.* (70). Synonymous

812     substitution rate of genotype 1 is approximately half that of genotypes 3 and 4,

813     which contributes to, but does not constitute, the signal of genotype specific positive

814     selection. The rate of substitutions in ORF 3 is consistently elevated in comparison

815     to other ORFs.

816

| 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 | |

817    Table 6. Assessing significance in differences in clockrates between genotypes for

818    each ORF. The credible intervals are significant if they do not include zero. This

819    shows genotype 1 has a significantly different clockrate from genotypes 3 & 4 across

820    all ORFs. This supports the clockrate data in Figure 3.

821

**Figure Legends**

823    Figure 1. FUBAR analysis of concatenated ORF1 and ORF2 sequences isolated from

824    humans (n=113). Genome-wide patterns of non-synonymous (β) and synonymous

825    (α) substitutions per site show that HEV has a background of purifying selection

826    with two discrete regions of elevated diversity corresponding to the hypervariable

827    region (HVR) and the overlap region between ORF2 and ORF3, as shown on the

828    genomic map. Sites subject to significant pervasive positive selection (FUBAR

829    posterior probability ≥ 0.95) are shown as black circles on the x-axis. FUBAR

830    analysis of the ORF2/3 overlap regions in their respective reading frames, showing

831    positive selection in both frames, but with ORF2 demonstrating a stronger signal

832    than ORF3, both in terms of the number of positively selected sites, and the

833    magnitude of β-α.

834    Figure 2. FUBAR Rate analysis of the ORF2/3 overlap region showing conserved

835    patterns of groups of selected sites across genotypes. The x axis represents

836    synonymous rates (α), while the y axis represents non-synonymous rates (β). As

837    labelled, all sites above the α=β line positively selected, and those below are

838    negatively selected. The plane is coloured by the weight assigned to each area by the

839    FUBAR algorithm. All six plots use the same colouring scale, so they are directly

840    comparable. Genotype 1 is unusual in having a very low proportion of positively

841    selected sites. In genotypes 1 and 3 both the codon substitution model (Table 3) and

842    RELAX procedure (Table 4) estimate that ORF 2 has significantly a stronger signal of

843    positive selection.

844    Figure 3. Estimates of evolutionary rates of HEV based on different genomic

845    regions. Anthropotropic genotype 1 has significantly reduced relative non-

846    synonymous evolutionary rates compared to their zoonotic counterparts across all

847    ORFs. Genotypes 3 and 4 demonstrate similar profiles, with non significant

848    differences across all ORFs. The overlap region of ORF 2 appears very similar to ORF

849    3, as they overlap extensively. Notably the non-overlap region of ORF 2 has a similar

850    evolutionary rate profile to ORF 1, with which is does not overlap at all. Asterisks

851    denote significance.

852    Figure 4. Maximum likelihood phylogenies of near-full-length sequences of HEV

853    isolated from humans and swine. Branch lengths are in expected substitutions per

854    nucleotide site estimated under the RELAX (59) general exploratory model. Swine

855    isolates are labelled using muted text, and all branches labelled as 'human' are

856    plotted using thicker lines. The k coefficients measures relaxation (k < 1) or

857    intensification (k > 1) of positive selective pressure relative to the phylogeny-wide

858    baseline (mean of k is constrained to be 1), represented by shades of grey. For G3,

859    91.5% of the bootstrapped trees supported p-value of <=0.05 or less. For G4 every

860    single p-value of the bootstrapped trees supported p-value of < 0.05.
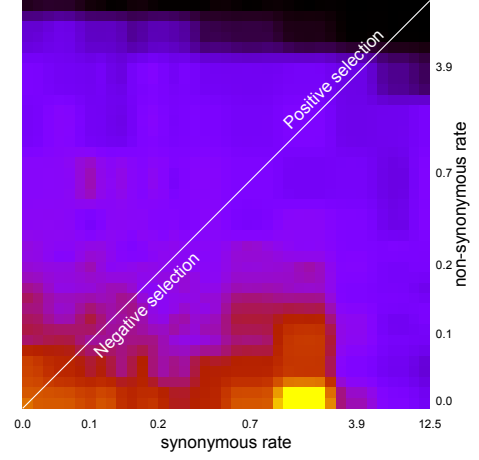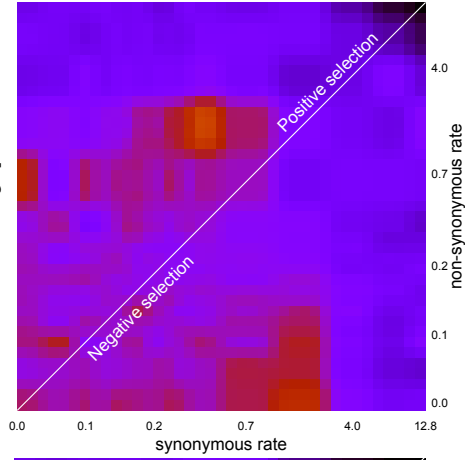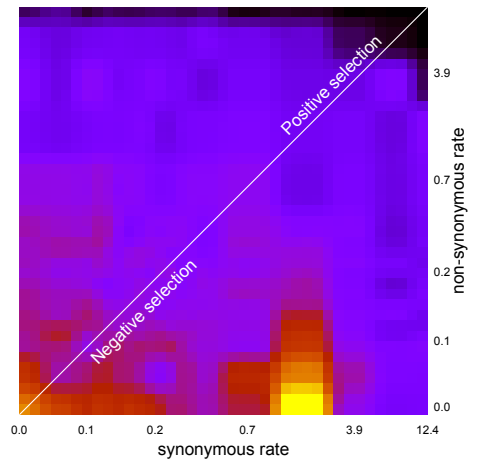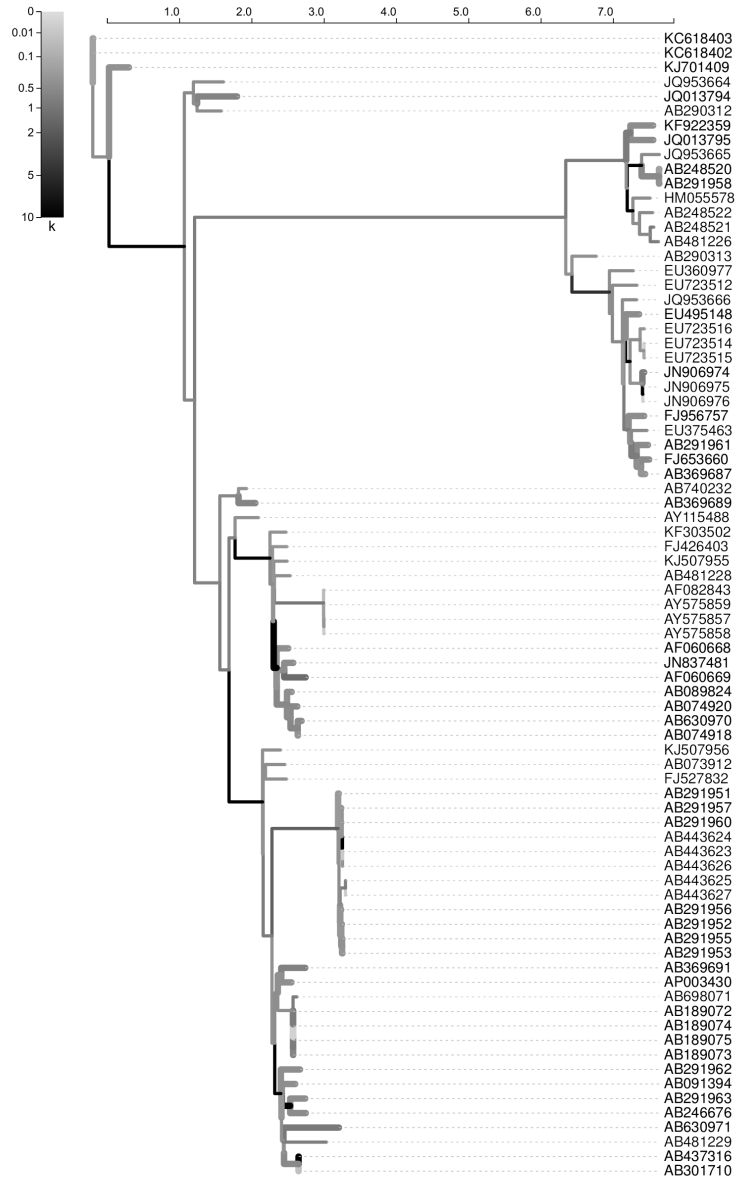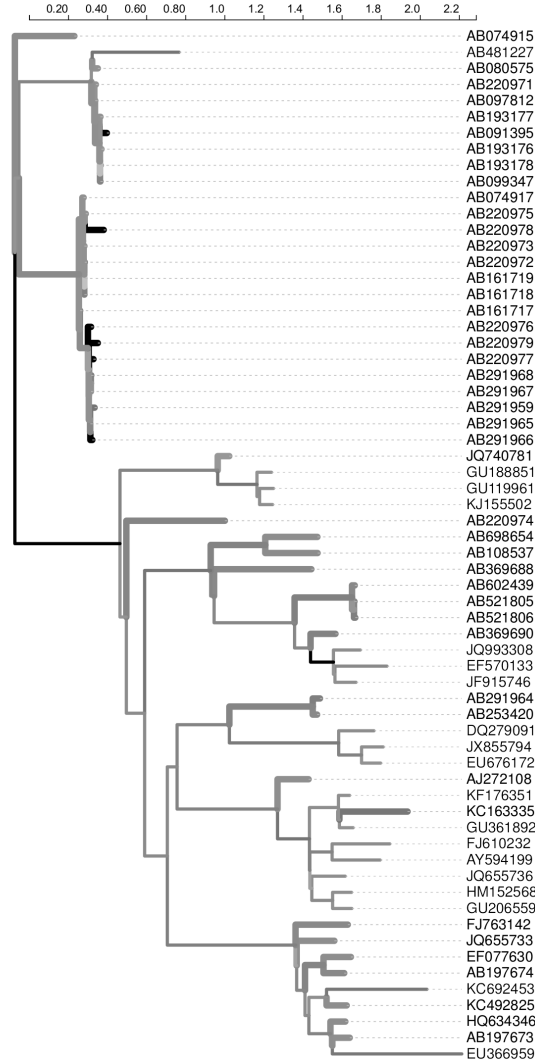
861
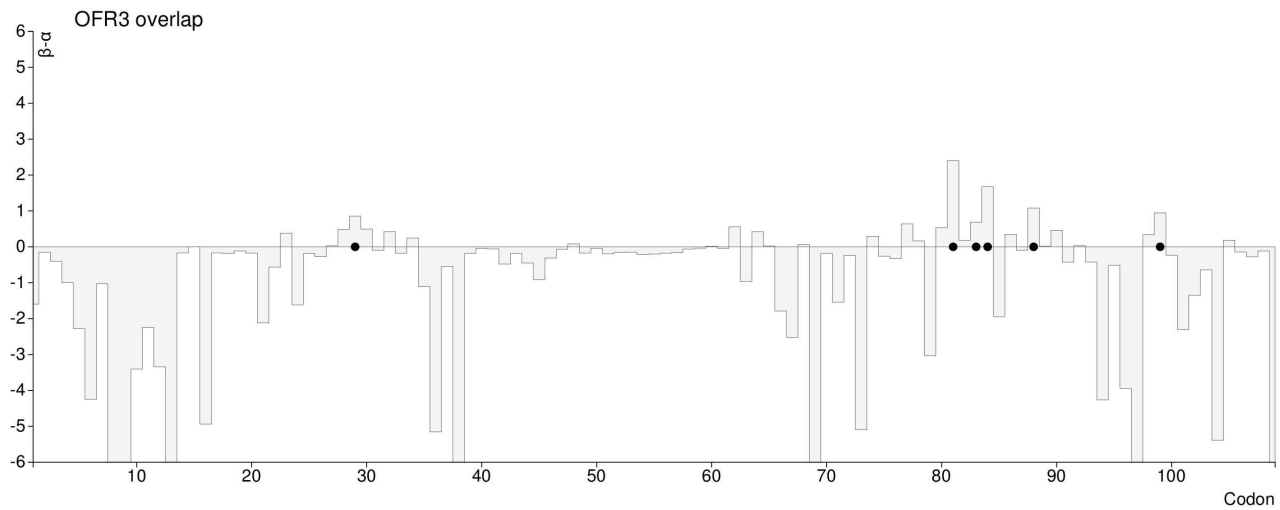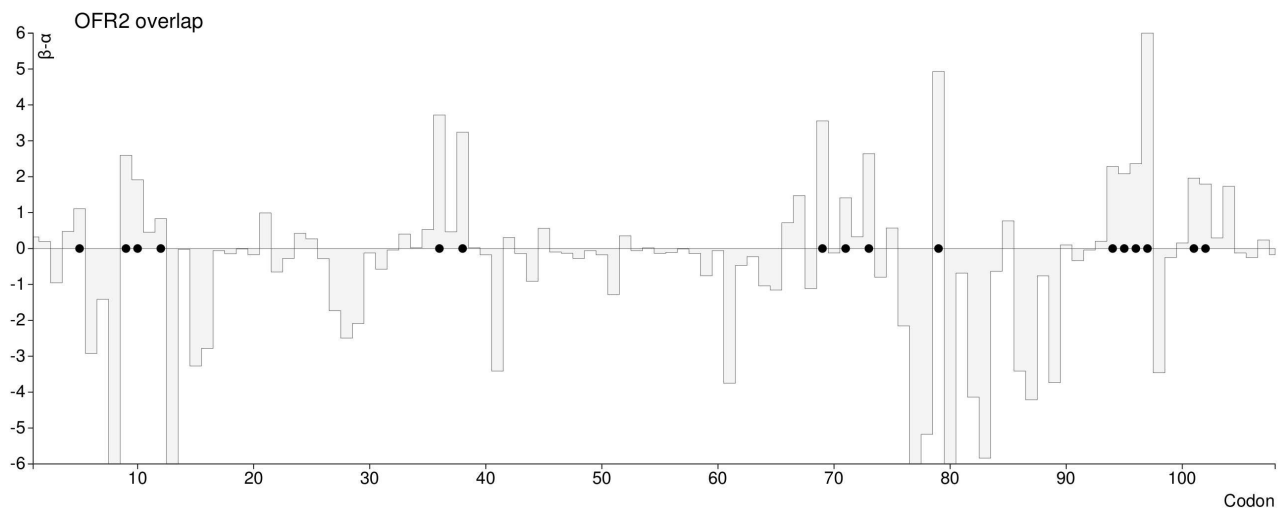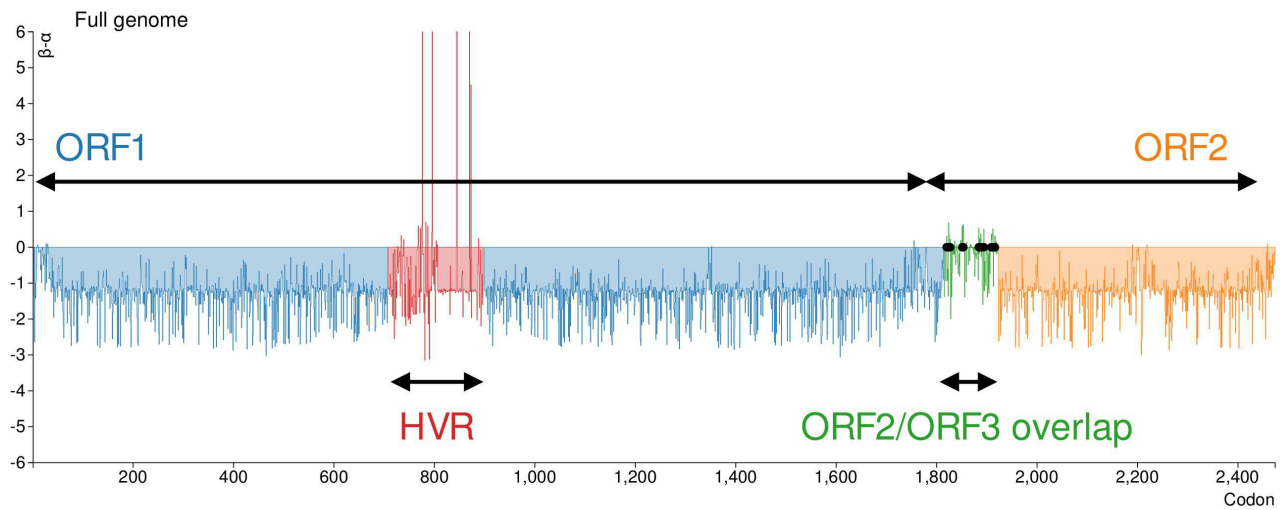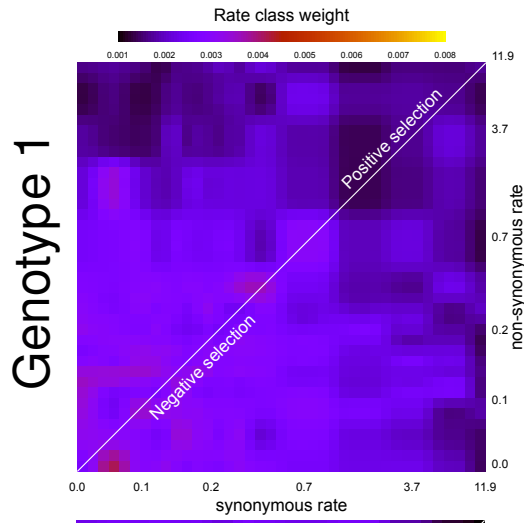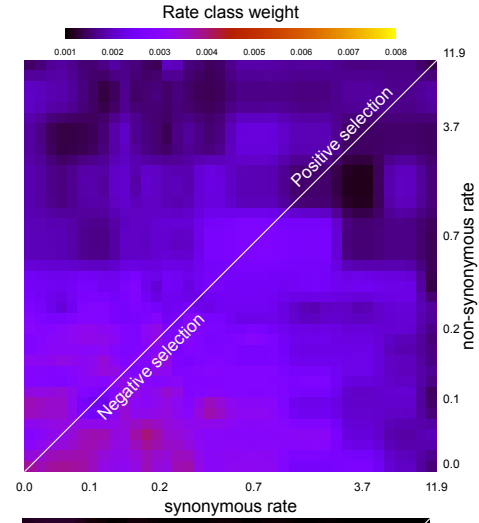
Genotype 3

Genotype 4

human isolates swine isolates