Journal Title:

Hepatitis Monthly

MANUSCRIPT INFORMATION	
Manuscript ID:	31951
Manuscript Title:	Hepatitis E Virus circulation in Italy: phylogenetic and evolutionary analysis
Manuscript Type:	Research Article
Submission Date:	Friday, July 31, 2015
Submitted By:	Prof. Massimo Ciccozzi
Running title:	Phylogeny of HEV in Italy
Keywords:	Hepatitis E virus, phylogeny, evolution.
Classification:	epidemiology ; infectious diseases ; osteoporosis; virology;
No. of Revision:	0
Implication of your manuscript:	The genetic diversity of HEV in Italy between human and swine was investigated. This study should contribute to improve the knolewdge of this virus in order to prevent the spread of HEV infection in Italy.
Acknowledgment:	Authors declare they have no financial interests related to the material in the manuscript. | None funding support | None sponsor support | Study concept and design: Marta Giovanetti, Marco Ciotti, Massimo Ciccozzi. Acquisition of data: Marta Giovanetti, Marco Ciotti. Analysis and interpretation of data: Marta Giovanetti, Eleonora Cella, Alessandra Lo Presti. Drafting of the manuscript: Marta Giovanetti, Marco Ciotti, Carla Montesano, Silvia Angeletti, Gianguglielmo Zehender, Massimo Ciccozzi. Critical revision of the manuscript for important intellectual content: Marco Ciotti, Gianguglielmo Zehender, Massimo Ciccozzi. | The authors would like to thank Dr Valerio Ciccozzi for the English revision of the manuscript.

FAQs

- **How can we solve the issues/errors which can be seen in PDF?** All contents of this manuscript has been produced from the files that authors submitted to the online system. Text of this PDF is generated especially from the word file that is uploaded as a simple text. If you saw any issue, simply go to the submission wizard and edit the word file. Consequently by re-generating the pdf file, the issue should be solved.
- **Is there any subsitution for PDF files?** If you, as the reviewer, have any problem in the PDF, simply use other attachments like word file or contact your technical editor.
- **Can we make any comment in PDF?** Adding Comments in PDF files is not possible. All reviewer comments should be added to the reviewer text box.


Published by Kowsar Corporation

Hepatitis E Virus circulation in Italy: phylogenetic and evolutionary analysis

Carla Montesano [2#]Marta Giovanetti [2 #], Marco Ciotti [3], Eleonora Cella [1,4], Alessandra Lo Presti [1], Alba Grifoni [7], Gianguglielmo Zehender [5], Silvia Angeletti [6], and Massimo Ciccozzi [1, 6 *].

[1] Epidemiology Unit, Department of Infectious, Parasitic and Immune-Mediated Diseases, Istituto Superiore di Sanità, Rome, Italy.

[2] Department of Biology, University of Rome "Tor Vergata", Rome, Italy.

[3] Laboratory of Molecular Virology, Polyclinic "Tor Vergata" Foundation, Rome, Italy.

[4] Department of Public Health and Infectious Diseases, Sapienza University of Rome, Rome, Italy

[5] **Laboratory of Infectious Diseases and Tropical Medicine, University of Milan.**

[6] Clinical Pathology and Microbiology Laboratory, University Hospital Campus Bio-Medico.

[7] ProxAgen Ltd, Sofia, Bulgaria.

[#] These authors equally contributed

Running title: Phylogeny of HEV in Italy

[*] **Corresponding author**: Massimo Ciccozzi
Department of Infectious Parasitic and Immunomediated Diseases, Reference Centre on Phylogeny, Molecular Epidemiology and Microbial Evolution (FEMEM)/ Epidemiology Unit, National Institute of Health, Rome, Italy
e-mail: ciccozzi@iss.it ; Phone number: +390649903187

ABSTRACT

Background: Hepatitis E virus (HEV) is a major cause of acute viral hepatitis in developing countries; it has been classified into four main genotypes and a number of subtypes. New genotypes have been recently identified in various mammals. HEV genotype 3 has a worldwide distribution and it is widespread among pigs in developed countries.

Objectives: In this study, the genetic diversity of HEV in Italy among human and swine was investigated. The estimation of the date of origin and the demographic history of the HEV was also performed.

Material and Methods: Three-hundred-twenty-seven Italian HEV sequences of swine and human origins were downloaded from the National Centre for Biotechnology Information. Three different data set were built. The first and the second data set were used to confirm the genotype of the sequences analyzed. The third data set was used to estimate the mean evolutionary rate, to perform the time-scaled phylogeny and demographic history.

Results: The Bayesian maximum clade credibility tree and the time of the most common recent ancestor estimates, showed that the root of the tree dated back to the year 1907 (95% HPD: 1811-1975). Two main clade were found, divided in two sub-clades.Skyline plot analysis performed separately for human and swine sequences, demonstrated the presence of the bottleneck only in the skyline plot from the swine sequences.Selective pressure analysis revealed only negatively selected sites.

Conclusions: This study could contribute to the hypothesis that human is probably infected after contact with swine sources. It is important to check swine country of origin and to improve sanitary control measures by the veterinary point of view, in order to prevent the spread of HEV infection in Italy.

Keywords: Hepatitis E virus, phylogeny, evolution.

1. Background

Hepatitis E virus (HEV) is a major cause of acute viral hepatitis in developing countries .

HEV is a non-enveloped virus that belongs to the Hepevirus genus of the Hepeviridae family. The viral genome, which is represented by a positive-sense single-stranded RNA of about 7.2 kb, contains three partially overlapping open reading frames (ORFs): ORF1, which encodes a non-structural protein with different enzymatic activities (RNA-dependent RNA polymerase, RNA helicase, protease); ORF-2, which encodes the capsid protein, and ORF-3 (overlapping ORF-2), which encodes a viral protein used in virion morphogenesis and release . HEV has been classified into four main genotypes (1–4) and a number of subtypes , and other new genotypes have been recently identified in various mammals . The geographical and host-range distribution of HEV genotypes are different. Genotypes 1 and 2 are restricted to humans and are typically faecal-orally transmitted. Genotype 1 is the main cause of sporadic and epidemic hepatitis E in developing regions of Asia, Africa, and South America, while genotype 2 has so far been identified in patients in Mexico, Chad, and Nigeria . Genotypes 3 and 4 have been recovered from humans, pigs and other species and are responsible for sporadic cases of HEV in humans . HEV genotype 3 has a worldwide distribution and is widespread among pigs in developed countries . There is evidence indicating that HEV-3 infection can be transmitted through the ingestion of raw or undercooked meat from infected animals highlighting the zoonotic nature of this infection. Genotype 4 is mainly found in eastern Asia .

In Italy, HEV has been found in pigs, boars and human being . The relatively high seroprevalence of HEV in domestic pigs indicates that there is active circulation of HEV in Italy . Genotype 3 is the only one that has been reported in Italian pigs and also in wild boars . Caruso et al. performed a serological and virological survey of HEV in wild boar populations in northwestern Italy. Overall, the seroprevalence was 4.9 %, and HEV RNA was detected in 3.7 % of liver samples, while no serum samples resulted positive for HEV RNA. By phylogenetic analysis of the ORF2 region the

isolates were clustered within genotype 3, subtypes 3e and 3f, and were closely related to HEV strains previously described in domestic pigs from the same geographic area.

Genotyping of HEV virus in association with the evolutionary rate estimate by phylogenetic analysis can help to follow the circulation of the virus and to understand the viral evolution and mechanism of infection.

2. Objectives

In this study the genetic diversity of HEV in Italy was investigated, to get deeper insight into the phylogenetic relationships existing among different strains of genotype 3, the most frequent circulating in Italy, among human and swine and to estimate the date of origin and the demographic history of the HEV circulation in Italy.

3. Materials and Methods

3.1 Data set

Three-hundred-twenty-seven Italian HEV sequences of swine and human origins were downloaded from the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

Three different data set were built. The first data set contained 11 Italian sequences of HEV from human ORF2 capsid gene, genotype 3, plus 13 genotype specific reference sequences.

The second data set contained 65 Italian sequences of HEV from swine ORF2 capsid gene, genotype 3, plus 13 genotype specific reference sequences.

The third data set contained the 76 ORF2 capsid gene Italian human and swine sequences of genotype 3.

The first and the second data set were used to confirm the genotype of the sequences analyzed.

For the selective pressure analysis and the demographic history both in human and swine Italian sequences, the first and the second dataset separately, without reference sequences, were used.

The third data set was used to estimate the mean evolutionary rate, to perform the time-scaled phylogeny and demographic history.

All the reference sequences were downloaded from the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). The reference sequences were selected on the basis of the following inclusion criteria: 1) sequences already published in peer-reviewed journals; 2) no uncertainty about genotype/subtype assignment; 3) sampling date were known and clearly established in the original publication.

3.2 Likelihood mapping

The phylogenetic signal in a data set of aligned DNA or amino acid sequences can be investigated with the likelihood mapping method by analyzing groups of four sequences, randomly chosen, called quartets . For a quartet, three unrooted tree topologies are possible. The likelihood of each

topology is estimated with the maximum likelihood method and the three likelihoods are reported as a dot in an equilateral triangle (the likelihood map).

Three main areas can be distinguished in the map: the three corners representing fully resolved tree topologies, i.e. the presence of treelike phylogenetic signal in the data; the center, which represents star-like phylogeny, and the three areas on the sides indicating network-like phylogeny, i.e. presence of recombination or conflicting phylogenetic signals. Extensive simulation studies have shown that > 33% dots falling within the central area indicate substantial star-like signal, i.e. a star-like outburst of multiple phylogenetic lineages. Likelihood mapping analyses was performed with the program TREE-PUZZLE by analyzing 10,000 random quartets .

3.3 Phylogenetic analysis

The sequences of all dataset were aligned using Clustal X and manually edited by Bioedit . Genotype of the Italian sequences was determined by phylogenetic analysis.

The maximum likelihood (ML) phylogenetic tree was generated with the HKY + I + G model of nucleotide substitution, by using Phyml v 3.0 . The evolutionary model was chosen as the best-fitting nucleotide substitution model in accordance with the results of the hierarchical likelihood ratio test (HLRT) implemented in Modeltest software version 3.7 . The statistical robustness and reliability of the branching order within the phylogenetic trees was confirmed by the bootstrap analysis, considering as significant statistical support a bootstrap value > 70%.

3.4 Bayesian phylogenetic analysis: evolutionary rate estimate, dated tree and demographic history

The evolutionary rate and the dated tree for the third data set was co-estimated by using a Bayesian Monte Carlo Markov Chain (MCMC) approach implementing the HKY+I+G model using both a strict and an uncorrelated log-normal relaxed clock model. As coalescent priors, were compared three parametric demographic models of population growth (constant size, exponential, expansion)

and a Bayesian skyline plot (BSP, a non-parametric piecewise-constant model). The best fitting models were selected by means of a Bayes factor (BF, using marginal likelihoods) implemented in Beast v. 1.7.4 .

In accordance with Kass and Raftery the strength of the evidence against H_0 (null hypothesis) was evaluated as follows: 2lnBF < 2 = no evidence; 2–6 = weak evidence; 6–10 = strong evidence; and >10 = very strong evidence. A negative 2lnBF indicates evidence in favor of H_0. Only values ≥ 6 were considered significant. The MCMC chains were run for at least 50 million generations, and sampled every 5,000 steps.

Convergence was assessed by estimating the effective sampling size (ESS) after a 10% burn-in, using Tracer software, version 1.5 (http://tree.bio.ed.ac.uk/software/tracer/), and accepting ESS values of 250 or more.

Uncertainty in the estimates was indicated by 95% highest posterior density (95% HPD) intervals. Statistical support for specific clades was obtained by calculating the posterior probability of each monophyletic clade.

The obtained tree was summarised by Tree Annotator (included in the Beast package) by choosing the tree with the maximum product of posterior probabilities (maximum clade credibility or MCC) after a 10% burn-in.

The demographic history was analyzed on the first and second dataset, by performing the Bayesian skyline plot.

3.5 Selective pressure analysis

The CODEML program implemented in the PAML 3.14 software package (http://abacus.gene.ucl.ac.uk/software/ paml.html) was used to investigate the adaptive evolution of HEV capsid gene. The sequences alignments of the first and second dataset were used to test whether they were under positive selection.

Six models of codon substitution: M0 (one-ratio), M1a (nearly neutral), M2a (positive selection), M3 (discrete), M7 (beta), and M8 (beta and omega) were used in this analysis . Since these models are nested, we used codon-substitution models to fit the model to the data using the likelihood ratio test (LRT) (28). The discrete model (M3), with three dn/ds (ω) classes, allows ω to vary among sites by defining a set number of discrete site categories, each with its own ω value. Through maximum-likelihood optimization, it is possible to estimate the ω and P values and the fraction of sites in the aligned data set that falls into a given category. Finally, the algorithm calculates the a posteriori probability of each codon belonging to a particular site category. Using the M3 model, sites with a posterior probability exceeding 90% and a ω value > 1.0 were designated as being “positive selection sites”. The site rate variation was evaluated comparing M0 with M3, while positive selection was evaluated comparing M1 with M2. The Bayes empirical Bayes (BEB) approach implemented in M2a and M8 was used instead to determine the positively selected sites by calculating the posterior probabilities (P) of ω classes for each site . It is worth noting that PAML LRTs have been reported to be conservative for short sequences (e.g. positive selection could be underestimated), although the Bayesian prediction of sites under positive selection is largely unaffected by sequence length . The dN/dS rate (ω) was also estimated by the ML approach implemented in the program HyPhy . In particular, the global (assuming a single selective pressure for all branches) and the local (allowing the selective pressure to change along every branch) models were compared by likelihood ratio test (LRT). Site-specific positive and negative selection were estimated by two different algorithms: the fixed-effects likelihood (FEL), which fits an ω rate to every site and uses the likelihood ratio to test if dN = dS; and random effect likelihood (REL), a variant of the Nielsen–Yang approach , which assumes that a discrete distribution of rates exists across sites and allows both dS and dN to vary independently site by site. The three methods have been described in more detail elsewhere . In order to select sites under selective pressure and keep our test conservative, a P value of ≤ 0.1 or a posterior probability of ≥ 0.9 as relaxed critical values

was assumed. Part of the analysis was conducted by using the web-based interface Datamonkey (http://www.datamonkey.org/) .

For the evolutionary analysis, the reference sequence with accession number AB369687.1 (complete genome cds), was used to trace the exact position of the amino acids under selection.

4. Results

4.1 Likelihood mapping

The phylogenetic noise of each data set, was investigated by means of likelihood mapping (Fig. S1). The percentage of dots falling in the central area of the triangles were 0.9% (panel a), 3.6% (panel b) and 2.8% (panel c) for the first, second and third data set, respectively. As none of the dataset showed more than 33% of noise, they contained sufficient phylogenetic signal.

4.2 Phylogenetic analysis

Maximum Likelihood phylogenetic trees of the first and second data set showed that all the sequences analyzed in this study are classified as genotype 3 (data not shown). The phylogenetic relationships among the different sequences of HEV were supported by bootstrap analysis with values >70%.

4.3 Bayesian phylogenetic analysis: evolutionary rate estimate, dated tree and demographic history.

The BF analysis showed that the relaxed clock fitted the data significantly better than the strict clock (2 lnBF =56.49 for relaxed clock). Under the relaxed clock the BF analysis showed that the exponential growth model was better than the other models (2lnBF > 15.736). The estimated mean value of HEV capsid gene evolutionary rate was 3.9×10^{-3} substitution/site/year (95% HPD: $1.3 \times 10^{-3} – 7.0 \times 10^{-3}$).

The Bayesian maximum clade credibility tree and the time of the most common recent ancestor (tMRCA) estimates performed on the third data set are shown in Fig. 1. The root of the tree had a tMRCA of 106 years corresponding to the year 1907 (95% HPD: 1811-1975). Two main clade (clade I and II) were found.

Clade I, includes 40 sequences, 35 from swine and 5 from human divided in two sub-clades (Ia and Ib). Sub-clade Ia, dated back to the year 1973 (95% HPD: 1943-2001) and includes two other clusters statistically supported: the first one dated back to the year 1993 (95% HPD: 1972-2005) and the other one dated back to the year 1997 (95% HPD: 1978-2010) whose sequences from swine and human appeared closely related. Sub-clade Ib dated back to the year 1981 (95% HPD: 1951-2001) includes only one cluster, statistically supported, which dated back to the year 2010 (95% HPD: 2007-2013) and consists of only swine sequences.

Clade II, includes 36 sequences, 30 from swine and 6 from human, divided in sub-clades (IIa and IIb). Sub-clade IIa, dated back to the year 1966 (95% HPD: 1929-1993) includes four clusters statistically supported. The first one dated back to the year 1995 (95% HPD: 1985-2002); the second one to the year 1993 (95% HPD: 1999-2008) and includes sequences from swine and human closely related; the third and the fourth clusters dated back, respectively, to 1996 (95% HPD: 1975-2009) and 1985 (95% HPD: 1963-1999) and include only sequences of swine origin.

Sub-clade IIb, dated back to the year 1972 (95% HPD: 1938-1996) including two clusters statistically supported: the first one dated back to the year 1987 (95% HPD: 1970-1988) and the second one to the year 2006 (95% HPD: 2003-2007); in these two clusters, only three sequences were from human.

4.4 Population dynamics

Analysis of the skyline plot (Fig. 2 panel A) showed that the effective number (Ne) of HEV infections from human and swine analysed together started to growth approximatively in the early 1900s and reached a plateau in 2000. At the plateau, the epidemic stopped growing but remained at

a level higher than it was at the beginning even if a decreasing phase, showing a typical "bottleneck" is evident immediately after 2000.

Skyline plot analysis was also performed separately for human and swine sequences (Fig. 2 panel B and C) to determine if the bottleneck evidenced in early 2000 was present in human as well as swine infections or not. The separated analysis demonstrated the presence of this bottleneck only in the skyline plot from the swine sequences (Fig. 2 panel C).

4.5 Evolutionary analysis

Selection pressure analysis performed on the first data set (sequences isolated from *Human*), revealed that there were not positively selected sites statistically supported (using both HYPHY and PAML). The Alfa parameter of the gamma distribution was <1, showing a characteristic L-shape which suggests a nucleotide substitution rate heterogeneity across sites but with most sites highly conserved. Regarding the selective pressure analysis on the first data set, the average ω ratio ranges from 0.0083 to 0.0086 among all models, suggesting that a non-synonymous mutation has around 0.83-0.86% as much chance as a synonymous mutation of being fixed in the population. Negative selection analysis identified 87 statistically supported sites using FEL, Table 1.

Similarly, selective pressure analysis performed on the second data set (sequences isolated from swine) showed that the Alfa parameter of the gamma distribution was <1, indicating that also this distribution has a characteristic L-shape, suggesting a nucleotide substitution rate heterogeneity across sites. The average ω ratio ranges from 0.0339 to 0.0389 among all models, suggesting that a non-synonymous mutation has only around 3.39-3.89% as much chance as a synonymous mutation of being fixed in the population. Negative selection was also shown for the capsid protein of the second data set: specifically 80, statistically supported, negatively selected sites were identified by HYPHY, Table 2.

5. Discussion

HEV infection is a global cause for morbidity and mortality. Besides endemic infections, autochthonous infections in developed countries are frequent .

Transmission route is one of the most discussed issue about HEV, with marked differences between different geographical areas. In developed countries, two main transmission routes are described, the fecal-oral route associated with genotypes 1 and 2 transmission, and the transmission route through the ingestion of raw meat of infected animals associated with genotypes 3 and 4 transmission .

In this study the phylogenetic relationships existing among different strains of HEV genotype 3, circulating in Italy, between human and swine were analyzed to estimate the date of origin, the spread and the demographic history of the HEV epidemic in Italy.

Our estimate of capsid gene evolutionary rate was 1.8×10^{-3} substitution/site/year with a broad credibility interval (between 1.2×10^{-3} and 5.0×10^{-3}) that is similar to that reported by other authors . Based on this temporal reconstruction, we suggest that HEV genotype 3 strains circulating in Italy in the first decade of 1900 diverged into the two main clades I and II which include the subclades Ia, Ib and IIa, IIb, respectively, originated between 1966 and 1981.

Interestingly, all swine sequences cluster together except in some cases where they are intermixed with human sequences as expected.

Skyline plot analysis was performed globally as well as separately for human and swine sequences. By separated analysis, the presence of a bottleneck after the year 2000 was detected only in the swine sequences data set. This could due to swine slaughter that consequently induced a decrease in the number of swine infections registered. These data could suggest that the control of HEV infection depends on adequate measure of prevention to avoid farmers' infection through swine meat and the consequent spread of the virus among humans.

The selective pressure analysis was also used to investigate the presence of sites under negative and positive selection. An average ω ratio < 1 was found, both in human and swine Italian ORF2 capsid gene sequences dataset, and only statistically supported negatively selected sites were identified; this finding confirms the stability of this viral protein. The HEV evolution has until now been characterized by neutral genetic drift. More studies are needed to examine zoonotic transmission and subsequent spillover into human populations, which would better explain the spread and the bottlenecks observed in swine in different HEV epidemics.

Overall, continued genomic surveillance of the HEV human and animal infection is required to assess adaptability and selection, which is increasingly important on the verge of an eventual vaccine deployment.

In conclusion, this study contributes to the hypothesis that human is probably infected after contact with swine sources. On these bases, it is important to check the swine country of origin and to improve sanitary control measures in order to prevent the spread of HEV infection in Italy.

References

1. **Labrique AB**, **Zaman K**, **Hossain Z**, **Saha P**, **Yunus M**, **Hossain A**, et al. An exploratory case control study of risk factors for hepatitis E in rural Bangladesh. ***PLoS One***. 2013; **8**, 61351.

2. Panda SK, Thakral D, Rehman S. Hepatitis E virus. *Rev Med Virol*. 2007; **17**, 151-180.

3. **Kamar N**, **Bendall R**, **Legrand-Abravanel F**, **Xia NS**, **Ijaz S**, **Izopet J**, et al. Hepatitis E. ***Lancet***. 2012; **379**, 2477-88.

4. Tam AW, Smith MM, Guerra ME, Huang CC, Bradley DW, Fry KE, et al. Hepatitis E virus (HEV): molecular cloning and sequencing of the full-length viral genome. *Virology*. 1991; **185**, 120-131.

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

6. Johne R, Plenge-Bönig A, Hess M, Ulrich RG, Reetz J, Schielke A. Detection of a novel hepatitis E-like virus in faeces of wild rats using a nested broad-spectrum RT-PCR. *J Gen Virol*. 2010; **91**, 750-758.

7. Takahashi M, Nishizawa T, Sato H, Sato Y, Jirintai NS, Okamoto H. Analysis of the full-length genome of a hepatitis E virus isolate obtained from a wild boar in Japan that is classifiable into a novel genotype. *J Gen Virol*. 2011; **92**, 902-908.

8. Zhao C, Ma Z, Harrison TJ, Feng R, Zhang C, Qiao Z, et al. A novel genotype of hepatitis E virus prevalent among farmed rabbits in China. *J Med Virol*. 2009; **81**, 1371-1379.

9. Buisson Y, Grandadam M, Nicand E, Cheval P, Van Cuyck-Gandre H, Innis B, et al. Identification of a novel hepatitis E virus in Nigeria. *J Gen Virol*. 2000; **81**, 903-909.

10. Van Cuyck-Gandré H, Zhang HY, Tsarev SA, Clements NJ, Cohen SJ, Caudill JD, et al. Characterization of hepatitis E virus (HEV) from Algeria and Chad by partial genome sequence. *J Med Virol*. 1997; **53**, 340-347.

11. Tam AW, White R, Reed E, Short M, Zhang Y, Fuerst TR, et al. In vitro propagation and production of hepatitis E virus from in vivo-infected primary macaque hepatocytes. *Virology*. 1996; **215**, 1-9.

12. Garbuglia AR, Scognamiglio P, Petrosillo N, Mastroianni CM, Sordillo P, Gentile D, et al. Hepatitis E virus genotype 4 outbreak, Italy, 2011. *Emerg Infect Dis*. 2013; **19**, 110-114.

13. Giordani MT, Fabris P, Brunetti E, Goblirsch S, Romanò L. Hepatitis E and lymphocytic leukemia in man, Italy. *Emerg Infect Dis*. 2013; **19**, 2054-2056.

14. Caruso C, Modesto P, BertolinI S, Peletto S, Acutis PL, Dondo A, et al. Serological and virological survey of hepatitis E virus in wild boar populations in northwestern Italy: detection of HEV subtypes 3e and 3f. *Arch Virol.* 2015; **160**, 153-160.

15. Colson P, Borentain P, Queyriaux B, Kaba M, Moal V, Gallian P, et al. Pig liver sausage as asource of hepatitis E virus transmission to humans. *J Infect Dis*. 2010; **202**, 825-834.

16. Hoofnagle JH, Kenrad EN, Purcell RH. Hepatitis E. *N Engl J Med.* 2012; **367**, 1237-1244.

17. Wedemeyer H, Pischke S, Manns MP. Pathogenesis and treatment of hepatitis E virus infection. *Gastroenterol*. 2012; **142**, 1388-1397.

18. Martinelli N, Pavoni E, Filogari D, Ferrari N, Chiari M, Canelli E, et al. Hepatitis E virus in wild boar in the central northern part of Italy. *Transboundary Emerg Dis*. 2015; **62**, 217-222.

19. Romanò L, Paladini S, Tagliacarne C, Canuti M, Bianchi S, Zanetti AR. Hepatitis E in Italy: A long-term prospective study. *J Hepatol*. 2011; **54**, 34-40.

20. Martinelli N, Luppi A, Cordioli P, Lombardi G, Lavazza A. Prevalence of hepatitis E virus antibodies in pigs in Northern Italy. *Infect Ecol Epidemiol*. 2011; **1**, 7331-7333.

21. Martelli F, Caprioli A, Zengarini M, Marata A, Fiegna C, Di Bartolo I, et al. Detection of hepatitis E virus (HEV) in a demographic managed wild boar (Sus scrofa scrofa) population in Italy. *Vet Microbiol*. 2008; **126**, 74-81.

22. Lo Presti A, Ciccozzi M, Cella E, Lai A, Simonetti FR, Galli M, et al. Origin, evolution, and phylogeography of recent epidemic CHIKV strains. *Infect Genet Evol.* 2012; 12, 392-8.

23. Ciccozzi M, Lo Presti A, Cella E, Giovanetti M, Lai A, El-Sawaf G, et al. Phylogeny of Dengue and Chikungunya viruses in Al Hudayda governorate, Yemen. Infect. *Genet Evol.* 2014; **27**, 395-401.

24. Lo Presti A, Ciccozzi M, Cella E, Giovanetti M, Zehender G, Valenti D, et al. Migration patterns of HIV-1 subtype B virus in Northern Italy. *New Microbiol.* 2013; **36,** 75-9.

25. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol.* 2005; **53**, 793-808.

26. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 2012; **90**, 773-795.

27. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 2000; **155**, 431- 449.

28. Anisimova M, Bielawsky JP, Yang Z. Accuracy and power of likelihood ratio test in detecting adaptive evolution. *Mol Biol Evol*. 2011; **8**, 1585-1592.

29. Yang Z, Wong WS, Nielsen R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 2005; **22**, 1107-1118.

30. Anisimova M, Bielawsky JP, Yang Z. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*. 2002; **19**, 950-958.

31. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. 2005; **21**, 676-679.

32. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*. 2005; **22**, 1208-1222.

33. Lo Presti A, Cammarota R, Apostoli P, Cella E, Fiorentini S, Babakir-Mina M, et al. Genetic variability and circulation pattern of human metapneumovirus isolated in Italy over five epidemic seasons. *New Microbiol.* 2011; **34**, 337-44.

34. Pérez-Gracia MT, Suay B, Mateos-Lindemann ML. Hepatitis E: An emerging disease. Infect. Genet. Evol. 2014; 22, 40-59.

35. Zehender G, Ebranati E, Lai A, Luzzago C, Paladini S, Tagliacarne C, et al. Phylogeography and phylodynamics of European genotype 3 hepatitis E virus. *Infect Genet Evol*. 2014; **25**, 138-143.

Figure Legends

Fig. 1. Bayesian maximum clade credibility tree including 76 Hepatitis E Virus capsid gene sequences. The asterisk (*) along the branches represents significant statistical support for the clade subtending that branch (posterior probability > 0.98). The scale at the bottom of the tree represents time in years. Main clades and clusters are indicated. Human and swine sequences are indicated with different symbols next to the tip of the sequences.

Fig. 2. A) Bayesian skyline plot (BSP) of the HEV Human and swine ORF2 capsid gene sequences from Italy. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. The coloured area corresponds to the credibility interval based on 95% highest posterior density interval (HPD). **B)** Bayesian skyline plot (BSP) of the HEV human ORF2 capsid gene sequences from Italy. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. The coloured area corresponds to the credibility interval based on 95% highest posterior density interval (HPD). **C)** Bayesian skyline plot (BSP) of the HEV swine ORF2 capsid gene sequences from Italy. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. The coloured area corresponds to the credibility interval based on 95% highest posterior density interval (HPD).

Supplementary fig. S1. Likelihood mapping of the first (panel a), second (panel b) and third data set (panel c). Each dot represents the likelihoods of the three possible unrooted trees for a set of four sequences (quartets) selected randomly from the data set: dots close to the corners or the sides represent, respectively, tree-like, or network-like phylogenetic signal in the data. The central area of the likelihood map represents star-like signal. The percentage of dots in the central area is given at the basis of each map.

TABLE 1 - ***Selection analysis for the capsid protein of HEV (sequences isolated from human).***

Negatively selected sites* (w for sites <1) HYPHY software	1994 (V); 1996(L,F); 1997(C,F); 1998(I); 2000(G); 2001(S); 2002(P); 2003(N); 2004(S); 2005(Y); 2006(T); 2007(N,Y); 2008(T); 2009(P); 2010(Y); 2011(T); 2014(A); 2015(L); 2016(G); 2017(L); 2019(D); 2020(A); 2021(L); 2022(E,G);2023(L);2024(E);2025(R); 2026(N); 2027(L);2028(P); 2034(T);2036(T); 2038(V); 2040(R); 2042(T,S); 2043(S); 2044(T); 2045(A); 2046(R); 2047(H);2048(R); 2049(L); 2050 (R);2051(R);2052(G); 2053(A); 2054(D); 2055(G); 2056(T); 2057(A); 2058(E); 2059(L); 2060(T); 2061(T); 2062(T); 2063(A);2065(T); 2066(R); 2067(F);2069(K);2070(D);2071(L); 2072(H); 2073(F); 2075(G); 2076(T,M); 2077(N);2078(G); 2079(V);2080(G); 2082(V); 2085(G); 2086(I);2087(A);2088(L);2089(T,I);2090(L);2091(F);2092(N);2095D);2096(T); 2098(L); 2099(G);2100(G);2101 (L);2105(L);2108(S).

*Negatively selected sites are numbered according to amino acid position of capsid protein of HEV isolate accession number AB369687.1

TABLE 2 - *Selection analysis for the capsid protein of HEV (sequences isolated from swine).*

Negatively selected sites*	2011 (Y,H); 2012(T,P); 2013(G); 2014(A); 2015(L); 2016(G); 2017(L); 2018(L); 2019(D); 2020(F); 2021(A); 2022(L); 2023(E); 2024(L); 2025(E); 2026(F); 2027(R,I); 2028(N); 2029(L,Q); 2030(T); 2031(P); 2032(G); 2033(N); 2034(T); 2035(N);
(w for sites <1)	2036(T); 2037(R);2040(R); 2041(Y); 2042(T); 2043(S); 2044(T); 2045(A,T); 2046(R); 2047(H); 2049(L); 2050(R,G); 2051(R); 2052(G); 2053(A); 2054(D);
HYPHY software	2055(G); 2056(T); 2057(A); 2058(E); 2059(L); 2060(T); 2061(T); 2062(T); 2063(A); 2065(T); 2066(R,H); 2067(F); 2070(D;E); 2071(L); 2072(H,L); 2073(F); 2074(T,I); 2075(G,E); 2076(T,M); 2077(N,K); 2078(G); 2079(V,I); 2080(G,R); 2081(E,K); 2082(V,G); 2083(G); 2084(R,L); 2085(G); 2086(I,V,L,R); 2088(L); 2089(T); 2090(L,P,K); 2091(F); 2092(N); 2095(D); 2096(T); 2099(G); 2100(G); 2101(L).

*Negatively selected sites are numbered according to amino acid position of capsid protein of HEV isolate accession number AB369687.1

Financial disclosure

Authors declare they have no financial interests related to the material in the manuscript.

Acknowledgements

The authors would like to thank Dr Valerio Ciccozzi for the English revision of the manuscript.

Authors' Contributions

Study concept and design: Marta Giovanetti, Marco Ciotti, Massimo Ciccozzi. Acquisition of data: Marta Giovanetti, Marco Ciotti. Analysis and interpretation of data: Marta Giovanetti, Eleonora Cella, Alessandra Lo Presti. Drafting of the manuscript: Marta Giovanetti, Marco Ciotti, Carla Montesano, Silvia Angeletti, Gianguglielmo Zehender, Massimo Ciccozzi. Critical revision of the manuscript for important intellectual content: Marco Ciotti, Gianguglielmo Zehender, Massimo Ciccozzi.

FIGURES

Fig. 1. Bayesian maximum clade credibility tree including 76 Hepatitis E Virus capsid gene sequences. The asterisk (*) along the branches represents significant statistical support for the clade subtending that branch (posterior probability > 0.98). The scale at the bottom of the tree represents time in years. Main clades and clusters are indicated. Human and swine sequences are indicated with different symbols next to the tip of the sequences

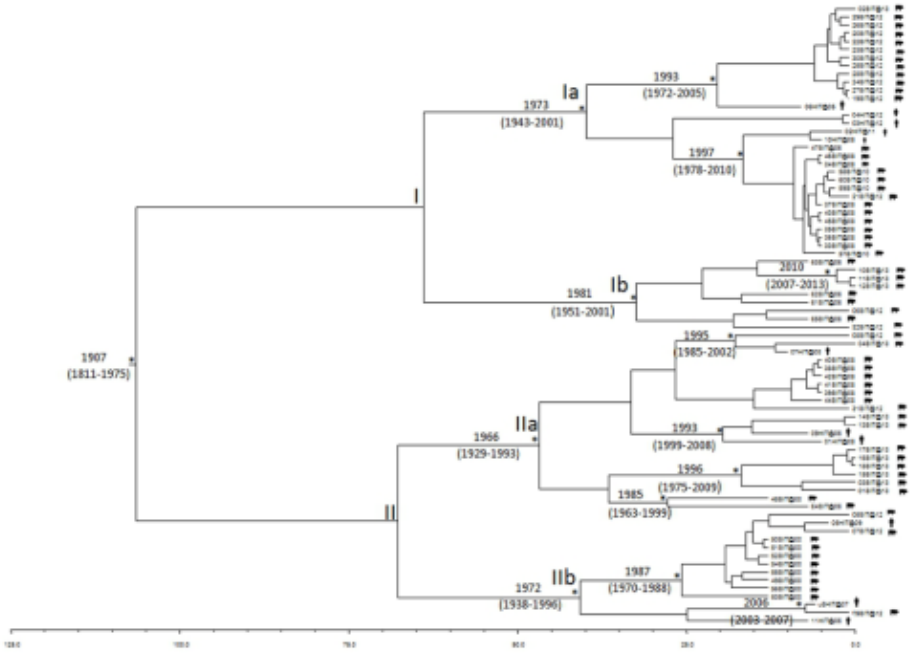
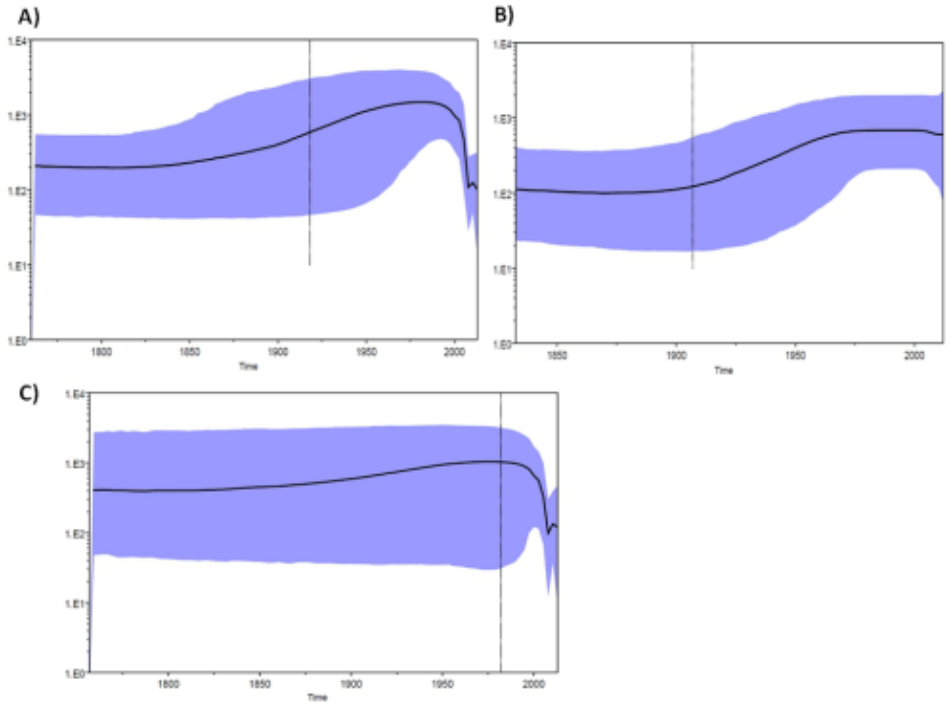


Fig. 2. A) Bayesian skyline plot (BSP) of the HEV Human and swine ORF2 capsid gene sequences from Italy. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. The coloured area corresponds to the credibility interval based on 95% highest posterior density interval (HPD). B) Bayesian skyline plot (BSP) of the HEV human ORF2 capsid gene sequences from Italy. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. The coloured area corresponds to the credibility interval based on 95% highest posterior density interval (HPD). C) Bayesian skyline plot (BSP) of the HEV swine ORF2 capsid gene sequences from Italy. The effective number of infections is reported on the Y-axis. Time is reported in the X-axis. The coloured area corresponds to the credibility interval based on 95% highest posterior density interval (HPD).



Supplementary fig. S1. Likelihood mapping of the first (panel a), second (panel b) and third data set (panel c). Each dot represents the likelihoods of the three possible unrooted trees for a set of four sequences (quartets) selected randomly from the data set: dots close to the corners or the sides represent, respectively, tree-like, or network-like phylogenetic signal in the data. The central area of the likelihood map represents star-like signal. The percentage of dots in the central area is given at the basis of each map