PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Saxenhofer M, Weber de Melo V, Ulrich RG, Heckel G. 2017 Revised time scales of RNA virus evolution based on spatial information. *Proc. R. Soc. B* **284**: 20170857. http://dx.doi.org/10.1098/rspb.2017.0857

Received: 21 April 2017
Accepted: 6 July 2017

Subject Category:
Evolution

Subject Areas:
evolution

Keywords:
virus evolution, emergence dating, mutational saturation, phylogenetics, hantavirus

Author for correspondence:
Gerald Heckel
e-mail: gerald.heckel@iee.unibe.ch

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3832525.

THE ROYAL SOCIETY PUBLISHING

Revised time scales of RNA virus evolution based on spatial information

Moritz Saxenhofer[1,2], Vanessa Weber de Melo[1,3], Rainer G. Ulrich[4,5] and Gerald Heckel[1,2]

[1]Computational and Molecular Population Genetics, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland
[2]Swiss Institute of Bioinformatics, Quartier Sorge – Bâtiment Génopode, Lausanne, Switzerland
[3]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[4]Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Institute of Novel and Emerging Infectious Diseases, Greifswald – Insel Riems, Germany
[5]German Center for Infection Research (DZIF), partner site Hamburg – Luebeck – Borstel – Insel Riems, Germany

GH, 0000-0002-0162-323X

The time scales of pathogen evolution are of major concern in the context of public and veterinary health, epidemiology and evolutionary biology. Dating the emergence of a pathogen often relies on estimates of evolutionary rates derived from nucleotide sequence data. For many viruses, this has yielded estimates of evolutionary origins only a few hundred years in the past. Here we demonstrate through the incorporation of geographical information from virus sampling that evolutionary age estimates of two European hantaviruses are severely underestimated because of pervasive mutational saturation of nucleotide sequences. We detected very strong relationships between spatial distance and genetic divergence for both Puumala and Tula hantavirus—irrespective of whether nucleotide or derived amino acid sequences were analysed. Extrapolations from these relationships dated the emergence of these viruses most conservatively to at least 3700 and 2500 years ago, respectively. Our minimum estimates for the age of these hantaviruses are ten to a hundred times older than results from current non-spatial methods, and in much better accordance with the biogeography of these viruses and their respective hosts. Spatial information can thus provide valuable insights on the deeper time scales of pathogen evolution and improve our understanding of disease emergence.

1. Introduction

Rapidly evolving pathogens cause a majority of emerging diseases in humans and livestock. There is often little consensus about their origin and evolutionary history [1,2], but understanding the process of disease emergence and spread is crucial for the development of preventive strategies and to combat epidemics. Important information such as the time of emergence and evolutionary rates can be derived from time-calibrated phylogenies of pathogen sequences using the dates of sample collection during an outbreak [3–5]. For example, for recent epidemics of Ebola virus, influenza virus, human immunodeficiency virus (HIV) and others, the combination of nucleotide sequence data of the pathogens with epidemiological reports has enabled detailed reconstructions of spatial and temporal transmission patterns of the infectious agents [4,6,7].

Serially sampled sequence data are often also used to infer the dates of much older events in the evolutionary history [8], but the absolute age of many pathogens remains contentious due to the lack of fossils, ancient DNA or other information that could provide support for the accuracy of estimated time scales [9]. Moreover, estimates based on heterochronous sequences (tip-dating *sensu* [10]) are often in conflict with biogeographic evidence or the evolutionary history of co-evolving host species. For example, tip-dated phylogenies of simian immunodeficiency viruses—which gave rise to HIV—indicated an

emergence less than 2000 years ago [11]. By contrast, the biogeography of these ubiquitously distributed viruses of African non-human primates and their association with the host phylogeny clearly revealed an ancient history of co-divergence over tens of thousands of years [12].

Temporal reconstruction of virus evolutionary history based on heterochronous sequence data has been questioned on grounds of insufficient temporal structure in phylogenetic trees [13] and underestimated genetic divergence due to mutational saturation [14]. The latter is a consequence of the high evolutionary rates of many pathogen species where multiple substitutions can occur at the same sequence position already over relatively short absolute time scales [15]. Sophisticated models of sequence evolution are designed to consider multiple substitutions at the same position, but the full extent of divergence in rapidly evolving sequences might still not be captured [16].

In this study, we exploit spatial distance as a hitherto unused source of information about genetic divergence of rapidly evolving pathogens and their time scales of emergence. We capitalize on the long-established realization that genetic similarity representing evolutionary divergence tends to be higher among spatially close organisms in systems with low dispersal, a pattern called isolation by distance (IBD) in population genetics [17,18]. With this correlation, the spatial distance between individuals can be informative about their genetic divergence. Over larger distances, this relationship is more likely to be affected by intrinsic (e.g. mutational saturation) and extrinsic factors (e.g. habitat availability, colonization history). A pattern of IBD at short spatial distances that decays over larger distances is thus indicative of such factors impacting on genetic divergence estimates. IBD has been detected in many organisms, including humans [19,20], with only relatively few formal reports from viruses [21–25]. However, many studies report the spatial clustering of nucleotide sequences in phylogenetic trees [26–30] which might indicate an IBD pattern.

Here we revise evolutionary time scales by incorporating spatial data for the case of hantaviruses (family Hantaviridae)—RNA viruses with often enigmatic evolutionary history and phenology [31]. Tip-dated phylogenies indicated the origin of rodent-borne hantaviruses sampled across North America, Asia and Europe only 849 [32] or 1915 years ago [33]. Such a recent emergence is unlikely in the light of, for example, the separation of the American and the Eurasian continent by the Bering Sea preventing rodent migration for more than 10 000 years [34,35]. Further, the specialization of hantaviruses to different widely co-distributed rodent hosts (e.g. in Europe Puumala virus, PUUV—*Myodes glareolus*; Tula virus, TULV—*Microtus arvalis*; Dobrava-Belgrade virus—*Apodemus* spec.) would have taken place within very short time. By stark contrast, studies focusing on the association of hantaviruses with their mostly Muroidea hosts place their origin millions of years ago based on phylogenetic relationships of the hosts dated with molecular and fossil data [36–38], but the actual phylogenetic evidence from virus data suggesting long-time coevolution is debated [39,40]. We focus our analyses on the hantaviruses PUUV and TULV because large-sequence datasets with initial evidence of spatial clustering are available, and their mostly European distribution range is relatively well covered [41–44]. Furthermore, their sedentary arvicoline rodent hosts display evidence of IBD at larger geographical scales [45–47]. We demonstrate that time estimates of hantavirus origins can be strongly biased by excessive mutational saturation in divergent sequence data, which can be detected and improved by taking spatial information about virus sampling locations into account.

2. Material and methods

(a) Phylogenetic analysis

Nucleotide sequence data from the PUUV and TULV nucleocapsid protein-encoding region of the small genomic segment (S-segment) and from complete PUUV genomes were collected from GenBank. Sequences of short overlap in the alignment, those originating from humans or cell lines instead of rodent hosts or with unknown place or date of sampling, as well as identical sequences from the same location, were excluded. Final alignments consisted for PUUV S-segment of 97 sequences of 504 bp length and for TULV of 115 sequences of 543 bp length. Sampling dates ranged between 1987 and 2012 for PUUV, and between 1987 and 2013 for TULV. The PUUV full genome alignment contained concatenated coding sequences from all three genomic segments (total: 11 228 bp). While PUUV sampling localities were heterogeneously distributed over a large region of Europe, TULV sequences mainly originated from Central Europe (figure 1*a*,*b*). Both datasets contained additional distant sequences from Russia and Kazakhstan (figure 1*c*). For phylogenetic inference, TULV was used as outgroup for PUUV and vice versa. Bayesian reconstruction of phylogenetic relationships was done with MrBayes v. 3.2.2 [48] on the CIPRES platform [49]. Nucleotide data were partitioned into two groups: combined 1st + 2nd and 3rd codon position, with the evolutionary rate unlinked across partitions. Reversible-jump sampling across the entire general time-reversible (GTR) substitution model space was implemented for each partition [50]. Metropolis-coupled Markov chain Monte Carlo (MCMC) sampling was performed for 10^7 generations in four independent runs of four chains for all datasets. A sample was recorded every 10^3 generation with a burn-in fraction of 25%. For amino acid sequence analyses, nucleotide sequences were translated in MrBayes using the protein substitution model. MCMC analyses were run as described above implementing a mixed amino acid model prior to averaging over different amino acid rate matrices. FigTree v. 1.4.2 [51] was used to draw consensus trees.

(b) Association between geographical distance and genetic divergence

Information on the geographical origin of virus sequences was obtained from the original publications or their authors. Geographic Distance Matrix Generator [52] was used to calculate pairwise Euclidean geographical distances between localities of origin. A GTR model with four substitution classes, among-site rate variation and proportion of invariable sites (TIM2 + G + I) was the best model of nucleotide substitution for all datasets according to Bayesian information criterion inferred with jModelTest v. 2.1.6 [53]. Pairwise genetic divergence was calculated in MEGA 6 [54] as (i) p-distance (percentage of variable sites) and (ii) under the Tamura and Nei (TrN + G + I) model [55], as it best corresponds to TIM2 + I + G. From the Bayesian phylogenetic analyses described above, trees with highest likelihood scores were used to infer (iii) pairwise genetic distance along branches (sum of branch lengths) between samples, representing phylogenetically interdependent divergence estimates [16]. For amino acid sequences, p-distance and tree distance was inferred analogously. Statistical significance of the correlation between the matrices of geographical distance and genetic divergence was determined with a Mantel test using 10^5 permutations performed in Arlequin v. 3.5.1.3 [56].

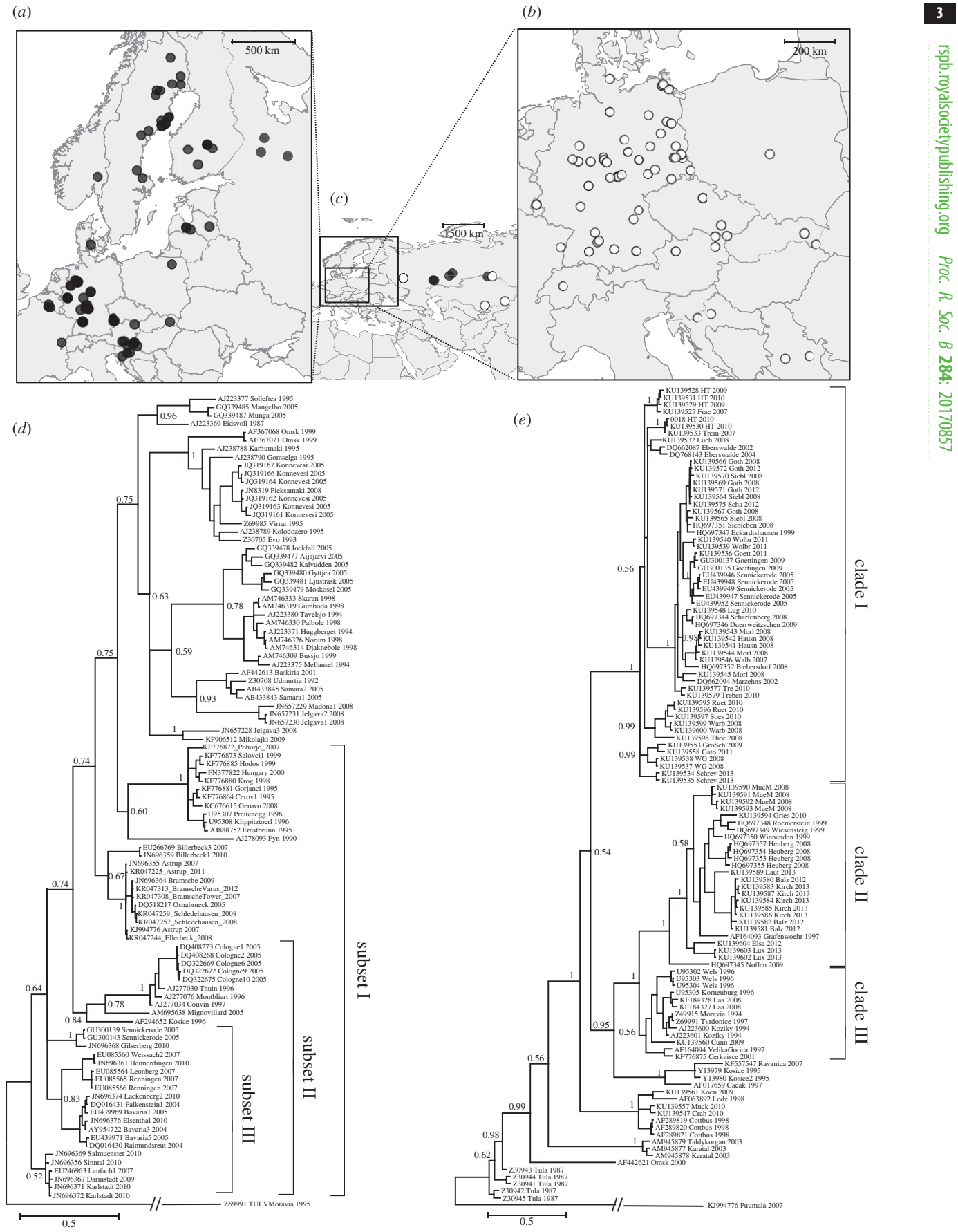

Figure 1. Geographical origin and phylogenetic relationships of Puumala (PUUV) and Tula (TULV) hantavirus sequences. (*a*,*b*) Maps show European sampling locations of (*a*) PUUV and (*b*) TULV sequence data (note the difference in scale). (*c*) Additional distant sampling locations in Russia and Kazakhstan are indicated by black and white symbols for PUUV and TULV sequence origins, respectively. (*d*,*e*) Maximum clade credibility phylogenetic trees based on partial S-segment sequences of (*d*) PUUV and (*e*) TULV with posterior node probabilities shown for major branches only. Subsets of different evolutionary levels in PUUV and evolutionary clades in TULV (see text) are indicated with brackets. Leaves are labelled with GenBank sequence accession number, sampling location and year of sample collection.

(c) Testing for mutational saturation

To examine mutational saturation, we plotted pairwise transition to transversion ratios (Ti/Tv) against genetic divergence estimated with the TrN + I + G model. Saturation at synonymous and non-synonymous sites was analysed by plotting pairwise genetic divergence based on the combined 1st + 2nd and the separated 3rd codon positions against geographical distance. Signals of mutational saturation among different evolutionary levels were investigated following Duchêne *et al.* [16] by removing basal sequences from the fixed topology of a phylogenetic tree to obtain sequence subsets of reduced evolutionary age. Tree topologies for this saturation test were estimated with Maximum-likelihood (ML) in GARLI [57] on CIPRES. For the PUUV dataset, basal branches of the ML topology were removed to form an initial subset 1, and younger subsets 2 and 3 were obtained by further removing basal branches (electronic supplementary material, figure S1). For TULV, the evolutionary clades described in [44] were defined as reduced-age subsets if represented by at least ten sequences in our dataset (figure 1*e*). For all trees and subsets, we calculated Ti/Tv, which is expected to decay with evolutionary time when divergence is underrepresented by the evolutionary model [58]. Analogously, we inferred the ratio of non-synonymous to synonymous substitutions (dN/dS), the GC contents and the shape parameter of a gamma-distributed among-site rate variation (α) to examine potential effects of natural selection and base composition on Ti/Tv. α is expected to scale negatively with increasing divergence when mutational saturation is accurately accounted for [16]. All parameters were estimated using CODEML implemented in PAML v. 4.8 [59].

(d) Virus divergence and age estimates

We used the observed correlation between genetic divergence and geographical distance among hantavirus sequences to deduce the per-distance increase of genetic divergence (slope). While mutational saturation among spatially distant sequences is expected to weaken this correlation (and flatten the slope), for recently diverged sequence pairs, the impact of mutational saturation on genetic divergence is low because the overall number of accumulated mutations during short time spans is small. To infer slopes mostly from sequence data at short distances that were less affected by mutational saturation, two different approaches were applied. (i) We simultaneously fitted two linear regressions with minimal overall residuals on either side of a separation point to the relationships between geographical distance and genetic divergence using R [60]. This resulted in a short-distance partition of more recently diverged sequence pairs of low mutational saturation characterized by a positive slope, and a partition at longer geographical distances representing a plateau of saturated sequence pairs (see Results). (ii) We first determined the saturation growth rate (SGR) model ($y = ax/(b + x)$) with CurveExpert v. 1.3 [61] as the best single model with regard to parametrization and standard error to fit the geographical distance to genetic divergence relationships of all data points. Then, we calculated the slope as the derivation for zero geographical distance where no effect of mutational saturation on the SGR curve is expected and determined confidence intervals with 1000 bootstrap replicates.

The inferred slopes from both approaches were then applied to linear models of unsaturated growth to provide estimates of revised genetic divergence across the full distribution range of virus species and clades. The minimum age of PUUV and TULV was assessed by dividing the expected unsaturated genetic divergence at maximum geographical distance by twice the substitution rate. We used a median substitution rate of 2.70×10^{-4} substitutions per site per year as assessed on recently diverged PUUV sequences sampled over eight years from the same geographical region [43] so that effects of mutational saturation should be insignificant. To compare our estimates incorporating geographical information with results of frequently used time-calibrated phylogenetic methods, we ran analyses in BEAST v. 2.3.0 [62] on CIPRES. Bayesian model averaging was performed using the bModelTest package [63] on partitioned data as described above. A Bayesian skyline coalescent tree prior with relaxed lognormal molecular clock was implemented in MCMC sampling of 10^8 generations with samples recorded every 5000 generations. TempEst v. 1.5 [64] showed that the temporal signal in our sequence datasets was very low as indicated by weak correlations of sampling dates with root-to-tip distances for both viruses (PUUV: $R^2 = 0.00022$ and TULV: $R^2 = 0.053$). The resulting age estimates are thus given for comparative purposes only.

3. Results

(a) Phylogeographic structure of PUUV and TULV

We detected very high sequence variability in both hantaviruses across their large distribution ranges (PUUV S-segment: 44.2% variable sites; PUUV full genomes: 32.7%; TULV S-segment: 43.5%), but all phylogenies were informative about the geographical origin of sequences irrespective of the year of sample collection (figure 1). Local phylogeographic clusters were highly supported in both viruses despite exclusion of identical sequences, whereas posterior probabilities of basal tree relationships were generally lower (figure 1*d*,*e*). For PUUV sequences from a large geographical area with heterogeneously distributed sampling localities, no deep phylogeographic structure could be determined (figure 1*a*,*d*), while TULV formed geographically segregated genetic clades (figure 1*b*,*e*).

(b) Association of geographical distance, divergence and mutational saturation

For both hantaviruses, we found an extremely strong association between geographical distance and genetic divergence (all $p < 0.001$; figure 2), irrespective of the estimator of divergence used for S-segment, full genome (in PUUV; electronic supplementary material, figure S2) or amino acid sequences (electronic supplementary material, figure S3). There were strong positive relationships at local scale and all slopes flattened at larger geographical distances, demonstrating strong mutational saturation for synonymous and non-synonymous codon positions likewise (electronic supplementary material, figure S4). This evidence of mutational saturation was corroborated by a clear negative correlation between Ti/Tv and genetic divergence (electronic supplementary material, figure S5) as well as between Ti/Tv and evolutionary age of phylogenetic relationships (table 1). dN/dS, α and GC contents remained constant among evolutionary levels, providing no evidence for influences on the decay of Ti/Tv other than mutational saturation (table 1). As expected under mutational saturation, pairwise genetic divergence based on p-distances was lower than estimated with the TrN + I + G model, which accounts for multiple substitutions per site (figure 2*a*,*b*,*d*,*e*; electronic supplementary material, table S1). Tree distances—representing branch lengths extracted from Bayesian analyses—were considerably larger than both pairwise sequence divergence estimates. Branch lengths in trees were sensitive to branch length priors, but the pattern of mutational saturation at larger geographical distances was unaffected (figure 2*c*,*f*; electronic supplementary material, table S1).

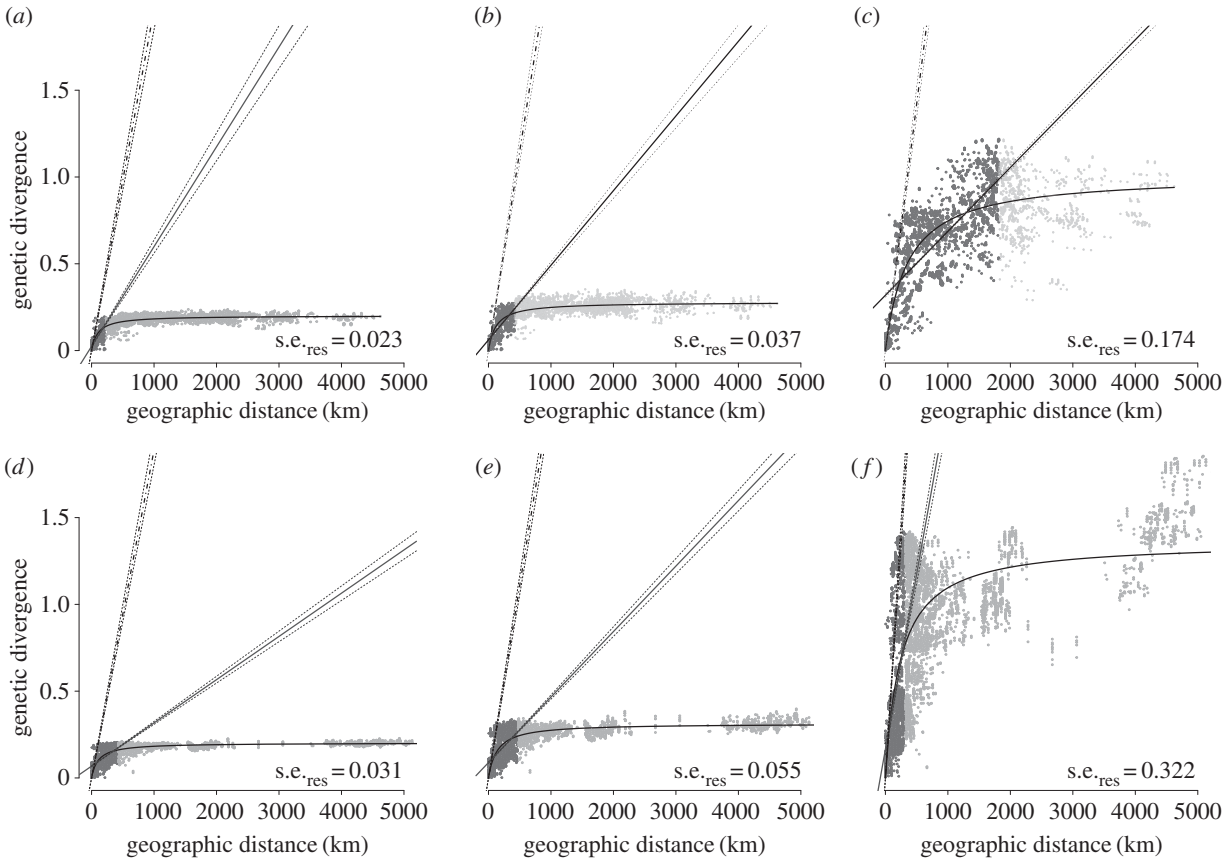

Figure 2. Relationships between geographical distance and genetic divergence among European hantaviruses. For PUUV (*a*–*c*) and TULV (*d*–*f*) genetic divergence is inferred either as p-distance (*a*,*d*), under the TrN + I + G model of nucleotide evolution (*b*,*e*), or as branch lengths in the Bayesian tree reconstruction (figure 1) (*c*,*f*). Dark grey points and the solid line (dotted lines: 95% CIs) represent the linear regression over the short-distance partition derived from fitting two linear regressions simultaneously (see text). Light grey data points represent the long distance partition with flat slope. Curves represent the saturation growth rate model (SGR; see text) with residual standard errors ($s.e._{res}$). The slope of the derivation of the SGR at zero distance is indicated by the hatched line with 1% and 99% quantiles from 1000 bootstraps.

Extrapolation of genetic divergence at maximum geographical distance in our datasets (PUUV: 4623 km; TULV: 5296 km) led to 16- to 52-fold higher estimates based on the derivation of the SGR curve compared with the fitted SGR curve (figure 2; electronic supplementary material, table S1). Extrapolation from the regression slope over short geographical distances resulted in two to 14 times higher genetic divergence at maximum geographical distance compared with the SGR curve (figure 2; electronic supplementary material, table S1). For PUUV, slopes of the short-distance regressions and SGR curves were consistent no matter how genetic divergence was inferred (figure 2*a*–*c*; electronic supplementary material, table S1). Phylogenetic clades in TULV (figure 1*e*) appear to affect the slope most noticeably for tree distances (figure 2*d*–*f*; electronic supplementary material, table S1) as highly diverged sequences from different clades can be found at short geographical distances.

Table 1. Sample sizes and substitution parameters determined for PUUV and TULV S-segment phylogenies and several subsets or clades of different evolutionary age (see text). *n*, number of sequences; Ti/Tv, transition/transversion ratio; d*N*/d*S*, ratio of non-synonymous to synonymous substitutions per site; GC, GC-content; α, shape parameter of the gamma distribution of among-site rate variation.

	n	Ti/Tv	d*N*/d*S*	GC	α
PUUV full tree	97	5.570	0.019	0.405	3.010
subset I	55	6.055	0.018	0.405	2.455
subset II	30	7.908	0.017	0.407	2.275
subset III	21	9.885	0.023	0.406	3.369
TULV full tree	115	6.025	0.010	0.426	2.237
clade I	56	8.987	0.007	0.428	2.365
clade II	26	9.263	0.005	0.422	1.804
clade III	13	13.033	0.009	0.424	1.830

(c) Hantavirus emergence dating using spatial information

Age estimates of both hantaviruses were considerably older based on spatially informed genetic divergence than those from tip-dated phylogenies. Analyses with BEAST estimated the age of PUUV overall as 346 years (95% highest posterior density interval, HPD: 201–571 years) and for TULV as 254 years (95% HPD: 145–392 years) (table 2; electronic supplementary material, figure S6). Geographically coherent sequence clusters in both viruses spanning vast areas of Europe were estimated to be very young. For example, the time to the most recent common ancestor (TMRCA) of PUUV subset I covering much of Central and Eastern Europe was estimated to 296 years (95% HPD: 177–487 years; electronic

Table 2. Minimum age estimates in years for full PUUV and TULV S-segment datasets and for several subsets or clades of younger evolutionary age. Calculations are based on extrapolation of linear regression slopes over short-distance data or derived slopes from the fitted saturation growth rate (SGR) curve. Both methods were applied on data with genetic distances calculated as p-distance, applying the TrN + I + G model of nucleotide substitution or as sum of branch lengths in a phylogenetic tree (tree). Numbers in brackets indicate 95% CIs. The last column contains results of the tip-dating analysis in BEAST with 95% highest posterior density (HPD) intervals indicated.

	extrapolation linear			extrapolation SGR			
	p-distance	TrN + I + G	tree	p-distance	TrN + I + G	tree	phylogenetics dating
PUUV							
full dataset	4966 (4610–5320)	3795 (3591–3998)	3741 (3665–3815)	16 841 (15 582–18 372)	19 697 (17 912–21 892)	24 354 (22 592–26 175)	346 (201–571)
subset I	1360 (1270–1447)	1098 (1049–1143)	1432 (1417–1444)	4501 (4165–4910)	5264 (4787–5851)	6509 (6038–6996)	296 (177–487)
subset II	781 (734–825)	664 (641–685)	1061 (1048–1072)	2519 (2331–2748)	2946 (2679–3274)	3643 (3379–3915)	209 (119–341)
subset III	469 (445–490)	431 (420–440)	861 (846–875)	1451 (1342–1583)	1698 (1544–1887)	2099 (1947–2256)	130 (475–217)
TULV							
full dataset	2569 (2465–2673)	3865 (3716–4014)	20 102 (18 600–21 603)	18 551 (17 364–19 671)	21 808 (20 551–23 086)	54 969 (51 349–58 574)	254 (145–392)
clade I	338 (333–343)	486 (478–492)	2026 (1936–2107)	1612 (1509–1709)	1895 (1786–2006)	4776 (4462–5090)	81 (46–128)
clade II	338 (332–342)	485 (477–490)	2022 (1929–2099)	1608 (1505–1704)	1890 (1781–2000)	4763 (4450–5076)	89 (51–138)
clade III	403 (395–410)	584 (572–593)	2551 (2420–2670)	2103 (1969–2230)	2473 (2330–2618)	6233 (5822–6642)	73 (45–109)

supplementary material, figures S1, S6). Similarly, the TMRCA of TULV clade III with sequences from Germany, Czech Republic, Slovakia, Austria, Slovenia and Croatia was 73 years (95% HPD: 45–109 years), and the split between clades II and III covering the area from Luxembourg to Croatia was estimated to only 126 years (95% HPD: 74–194 years) (figure 1*b*; electronic supplementary material, S6). By contrast, divergence time estimates based on linear regression slopes suggested a minimum evolutionary age of PUUV overall between 3700 and 5000 years, and of TULV between 2600 and 20 100 years (table 2). The minimum age estimates based on derivations from SGR slopes ranged between 16 800 and 24 300 years for the full PUUV dataset and 18 600–55 000 years for all TULV sequences. Based on Bayesian branch lengths, PUUV subset I was at least 6500 years and TULV clade III at least 6200 years old (table 2). When we alternatively applied the extremely high substitution rates inferred with BEAST (9.38×10^{-4} substitutions per site per year for PUUV and 1.51×10^{-3} substitutions per site per year for TULV), our age estimates incorporating spatial information were still 2 to 38 times higher than the tip-dated estimates (electronic supplementary material, table S2).

4. Discussion

Consistent spatial structure in PUUV and TULV allows for the first time to describe the extent of mutational saturation in two RNA virus species. We showed that models of sequence evolution severely underestimate the genetic divergence in our datasets (electronic supplementary material, table S1), which corroborates concerns about the accuracy of evolutionary time scales inferred with time-calibrated phylogenies for many rapidly evolving pathogen species [15,16]. For two RNA viruses, we demonstrated that combining sequence data with spatial information can lead to highly improved age estimates. The strong correlation between geographical distance and genetic divergence in PUUV and TULV makes the sampling locations informative about evolutionary differences even in highly saturated sequence data. While local disturbance of the IBD pattern is expected (e.g. through geographical barriers like larger water bodies), the general association between spatial distance and genetic divergence in these hantaviruses remains relatively unaffected. Our approach relies on predictable geographical structuring that may not necessarily be found among far-dispersing pathogens or may break down with extensive host migration. These factors differ widely between different study systems, but it is generally straightforward to test for IBD patterns in a sequence dataset and fit a saturation model. In more complex systems, simple spatial distances are unlikely to describe evolutionary relationships sufficiently well, but it might be possible to characterize pathogen diffusion patterns, for example through network connections [65]. The use of network distances that consider host mobility, migration history [66] and landscape factors [12] could thus potentially enable distance-based divergence estimation even for highly dispersed infectious agents.

We found that phylogenetic relationships of serially sampled PUUV and TULV sequences are determined rather by spatial proximity (figure 1) than by the time-points of sample collection. This suggests a long-lasting stationary presence of both hantaviruses across Europe, where genetic divergence between sequences is rather the result of long-term IBD than due to the collection of samples through time.

The observed loss of phylogeographic resolution even for whole-genome and amino acid sequences over geographical distances of less than 1000 km due to strong mutational saturation (figure 2; electronic supplementary material, S2, S3) supports a scenario of ancient colonization and low dispersal for both hantaviruses. Ancient emergence times as inferred in this study (table 2) are therefore in much better accordance with phylogeographic insights than results of tip-dated analyses [32,33], and they are in line with hypotheses of virus–host co-divergence at least in recent evolutionary periods [36–38]. Still, our results represent minimum estimates for the emergence of these hantaviruses, and their true age might be even higher: the data analysed here may not cover the full spatial distribution of PUUV and TULV and larger geographical distances would indicate higher overall evolutionary divergence. Higher divergence estimates deduced from the derivation of the SGR curve might be closer to true values than the ones from linear regressions because slopes over short-distance data are expected to be affected by gradually increasing mutational saturation (figure 2). While the variance of evolutionary time scales of hantaviruses inferred in this study remains large, all estimates are decisively older than tip-dated results (table 2).

Unlike in recent virus outbreaks, where the place of origin and direction of spread could be recovered from serially sampled sequence data [21,23,28], high mutational saturation probably impedes better temporal calibration of PUUV and TULV phylogenies using the dates of sample collection [13]. We found strong signals of mutational saturation at different levels of genetic divergence indicated by a clear decline in Ti/Tv ratios towards more diverged evolutionary levels (table 1). We tested for the influence of other factors on phylogenetic dating studying further evolutionary parameters, namely time-dependency of evolutionary rates [67]. Recently diverged sequences may contain slightly deleterious non-synonymous mutations, which have not yet been removed by purifying selection. Substitution rates of younger evolutionary groups would therefore appear to be higher than those of groups that have diverged a long time ago. Rates inferred for recently diverged sequences may thus not be applicable to subsequently date deeper phylogenetic relationships. However, we did not find evidence for changes in the evolutionary rates over time [10,68] as dN/dS ratios were consistently low among evolutionary levels (table 1). The number of pairwise synonymous and non-synonymous substitutions both increased with spatial distance and displayed mutational saturation over large distances (electronic supplementary material, figure S4). While we cannot exclude some effects of rate variability, our results demonstrate that for highly diverged hantavirus sequences, mutational saturation is the predominant cause of biases in molecular clock dating.

5. Conclusion

Evolutionary models, time-calibrated phylogenies and sampling data provide valuable information on pathogen transmission dynamics and instant processes of sequence change during disease outbreaks. However, the investigation of deeper evolutionary processes in rapidly evolving pathogens benefits also from taking spatial information into account. With a better understanding of the particular ecology of a pathogen–host system, it might be possible to define more reliably evolutionary time scales and reconstruct key processes of host jumps and long-term adaptive change. Learning about the evolutionary past of rapidly evolving pathogens is crucial to understand their fundamental biology, and to prevent and control future disease outbreaks.

Data accessibility. Sequence alignments including accession numbers, sampling dates and geographical coordinates were submitted to Dryad (http://dx.doi.org/10.5061/dryad.dc770) [69].

Authors' contributions. G.H. conceived the study. M.S. and V.W.d.M. collected and analysed the data guided by G.H. R.G.U. provided advice and discussion. M.S. and G.H. wrote the manuscript that was revised and approved by all authors.

Competing interests. We declare no competing interests.

Funding. This study was supported by a grant from the Swiss National Science Foundation (31003A-149585) to G.H. and Deutsche Forschungsgemeinschaft (SPP 1596 'Ecology and Species Barriers in Emerging Viral Diseases', UL 405/1-1) to R.G.U.

Acknowledgements. We are very grateful to Seraina Klopfstein for helpful suggestions, discussion and support with phylogenetic analyses, Stephan Peischl for mathematical assistance with the extrapolation methods and the authors of original TULV and PUUV publications for sampling location information.

References

1. Duffy S, Shackelton LA, Holmes EC. 2008 Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**, 267–276. (doi:10.1038/nrg2323)
2. Holmes EC. 2009 The evolutionary genetics of emerging viruses. *Annu. Rev. Ecol. Evol. Syst.* **40**, 353–372. (doi:10.1146/annurev.ecolsys.110308.120248)
3. Lam TT-Y *et al.* 2012 Phylodynamics of H5N1 avian influenza virus in Indonesia. *Mol. Ecol.* **21**, 3062–3077. (doi:10.1111/j.1365-294X.2012.05577.x)
4. Faria NR *et al.* 2014 The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56–61. (doi:10.1126/science.1256739)
5. Carroll MW *et al.* 2015 Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature* **524**, 97–101. (doi:10.1038/nature14594)
6. Verhagen JH, Herfst S, Fouchier RAM. 2015 How a virus travels the world. *Science* **347**, 616–617. (doi:10.1126/science.aaa6724)
7. Attar N. 2015 Viral evolution: keeping a watchful eye on Ebola. *Nat. Rev. Genet.* **16**, 437. (doi:10.1038/nrg3983)
8. Stadler T, Yang Z. 2013 Dating phylogenies with sequentially sampled tips. *Syst. Biol.* **62**, 674–688. (doi:10.1093/sysbio/syt030)
9. O'Reilly JE, dos Reis M, Donoghue PCJ. 2015 Dating tips for divergence-time estimation. *Trends Genet.* **31**, 637–650. (doi:10.1016/j.tig.2015.08.001)
10. Duchêne S, Lanfear R, Ho SYW. 2014 The impact of calibration and clock-model choice on molecular estimates of divergence times. *Mol. Phylogenet. Evol.* **78**, 277–289. (doi:10.1016/j.ympev.2014.05.032)
11. Wertheim JO, Worobey M. 2009 Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput. Biol.* **5**, e1000377. (doi:10.1371/journal.pcbi.1000377)
12. Worobey M *et al.* 2010 Island biogeography reveals the deep history of SIV. *Science* **329**, 1487. (doi:10.1126/science.1193550)
13. Zhang Y-Z, Holmes EC. 2014 What is the time-scale of hantavirus evolution? *Infect. Genet. Evol.* **25**, 144–145. (doi:10.1016/j.meegid.2014.04.017)
14. Duchêne S, Holmes EC, Ho SYW. 2014 Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. B* **281**, 20140732. (doi:10.1098/rspb.2014.0732)

15. Holmes EC. 2003 Molecular clocks and the puzzle of RNA virus origins. *J. Virol.* **77**, 3893–3897. (doi:10.1128/JVI.77.7.3893-3897.2003)
16. Duchêne S, Ho SY, Holmes EC. 2015 Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evol. Biol.* **15**, 36. (doi:10.1186/s12862-015-0312-6)
17. Wright S. 1943 Isolation by distance. *Genetics* **28**, 114–138.
18. Slatkin M. 1993 Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279. (doi:10.2307/2410134)
19. Novembre J *et al.* 2008 Genes mirror geography within Europe. *Nature* **456**, 98–101. (doi:10.1038/nature07331)
20. Elhaik E *et al.* 2014 Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat. Commun.* **5**, 3513. (doi:10.1038/ncomms4513)
21. Fargette D *et al.* 2004 Inferring the evolutionary history of rice yellow mottle virus from genomic, phylogenetic, and phylogeographic studies. *J. Virol.* **78**, 3252–3261. (doi:10.1128/JVI.78.7.3252-3261.2004)
22. Real LA *et al.* 2005 Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *Proc. Natl Acad. Sci. USA* **102**, 12 107–12 111. (doi:10.1073/pnas.0500057102)
23. Walsh PD, Biek R, Real LA. 2005 Wave-like spread of Ebola Zaire. *PLoS Biol.* **3**, e371. (doi:10.1371/journal.pbio.0030371)
24. Carrel MA, Emch M, Jobe RT, Moody A, Wan X-F. 2010 Spatiotemporal structure of molecular evolution of H5N1 highly pathogenic avian influenza viruses in Vietnam. *PLoS ONE* **5**, e8631. (doi:10.1371/journal.pone.0008631)
25. Padhi A, Moore AT, Brown MB, Foster JE, Pfeffer M, Brown CR. 2010 Isolation by distance explains genetic structure of Buggy Creek virus, a bird-associated arbovirus. *Evol. Ecol.* **25**, 403–416. (doi:10.1007/s10682-010-9419-9)
26. Grenfell BT, Bjørnstad ON, Kappey J. 2001 Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414**, 716–723. (doi:10.1038/414716a)
27. Song J-W, Baek LJ, Song K-J, Skrok A, Markowski J, Bratosiewicz-Wasik J, Kordek R, Liberski PP, Yanagihara R. 2004 Characterization of Tula virus from common voles (*Microtus arvalis*) in Poland: evidence for geographic-specific phylogenetic clustering. *Virus Genes* **29**, 239–247. (doi:10.1023/B:VIRU.0000036384.50102.cf)
28. Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA. 2007 A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proc. Natl Acad. Sci. USA* **104**, 7993–7998. (doi:10.1073/pnas.0700741104)
29. Johne R *et al.* 2012 Rat hepatitis E virus: geographical clustering within Germany and serological detection in wild Norway rats (*Rattus norvegicus*). *Infect. Genet. Evol.* **12**, 947–956. (doi:10.1016/j.meegid.2012.02.021)
30. Drewes S *et al.* 2017 Host-associated absence of human Puumala virus infections in Northern and Eastern Germany. *Emerg. Infect. Dis.* **23**, 83–86. (doi:10.3201/eid2301.160224)
31. Vaheri A, Strandin T, Hepojoki J, Sironen T, Henttonen H, Mäkelä S, Mustonen J. 2013 Uncovering the mysteries of hantavirus infections. *Nat. Rev. Microbiol.* **11**, 539–550. (doi:10.1038/nrmicro3066)
32. Ramsden C, Holmes EC, Charleston MA. 2009 Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol. Biol. Evol.* **26**, 143–153. (doi:10.1093/molbev/msn234)
33. Souza WM, Bello G, Amarilla AA, Alfonso HL, Aquino VH, Figueiredo LTM. 2014 Phylogeography and evolutionary history of rodent-borne hantaviruses. *Infect. Genet. Evol.* **21**, 198–204. (doi:10.1016/j.meegid.2013.11.015)
34. Dixon EJ. 2013 Late pleistocene colonization of North America from Northeast Asia: new insights from large-scale paleogeographic reconstructions. *Quat. Int.* **285**, 57–67. (doi:10.1016/j.quaint.2011.02.027)
35. Fink S, Fischer MC, Excoffier L, Heckel G. 2010 Genomic scans support repetitive continental colonization events during the rapid radiation of voles (Rodentia: *Microtus*): the utility of AFLPs versus mitochondrial and nuclear sequence markers. *Syst. Biol.* **59**, 548–572. (doi:10.1093/sysbio/syq042)
36. Hughes AL, Friedman R. 2000 Evolutionary diversification of protein-coding genes of hantaviruses. *Mol. Biol. Evol.* **17**, 1558–1568. (doi:10.1093/oxfordjournals.molbev.a026254)
37. Sironen T, Vaheri A, Plyusnin A. 2001 Molecular evolution of Puumala hantavirus. *J. Virol.* **75**, 11 803–11 810. (doi:10.1128/JVI.75.23.11803-11810.2001)
38. Plyusnin A, Sironen T. 2014 Evolution of hantaviruses: co-speciation with reservoir hosts for more than 100 MYR. *Virus Res.* **187**, 22–26. (doi:10.1016/j.virusres.2014.01.008)
39. De Vienne DM, Giraud T, Shykoff JA. 2007 When can host shifts produce congruent host and parasite phylogenies? A simulation approach. *J. Evol. Biol.* **20**, 1428–1438. (doi:10.1111/j.1420-9101.2007.01340.x)
40. Geoghegan JL, Duchêne S, Holmes EC. 2017 Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* **13**, e1006215. (doi:10.1371/journal.ppat.1006215)
41. Schmidt-Chanasit J *et al.* 2010 Extensive host sharing of Central European Tula virus. *J. Virol.* **84**, 459–474. (doi:10.1128/JVI.01226-09)
42. Mertens M *et al.* 2011 Phylogenetic analysis of Puumala virus subtype Bavaria, characterization and diagnostic use of its recombinant nucleocapsid protein. *Virus Genes* **43**, 177–191. (doi:10.1007/s11262-011-0620-x)
43. Weber de Melo V *et al.* 2015 Spatiotemporal dynamics of Puumala hantavirus associated with its rodent host, *Myodes glareolus*. *Evol. Appl.* **8**, 545–559. (doi:10.1111/eva.12263)
44. Schmidt S *et al.* 2016 High genetic structuring of Tula hantavirus. *Arch. Virol.* **161**, 1135–1149. (doi:10.1007/s00705-016-2762-6)
45. Heckel G, Burri R, Fink S, Desmet J-F, Excoffier L. 2005 Genetic structure and colonization processes in European populations of the common vole, *Microtus arvalis*. *Evolution* **59**, 2231–2242. (doi:10.1111/j.0014-3820.2005.tb00931.x)
46. White TA, Perkins SE, Heckel G, Searle JB. 2013 Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. *Mol. Ecol.* **22**, 2971–2985. (doi:10.1111/mec.12343)
47. Fischer MC, Foll M, Heckel G, Excoffier L. 2014 Continental-scale footprint of balancing and positive selection in a small rodent (*Microtus arvalis*). *PLoS ONE* **9**, e112332. (doi:10.1371/journal.pone.0112332)
48. Ronquist F *et al.* 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi:10.1093/sysbio/sys029)
49. Miller MA, Pfeiffer W, Schwartz T. 2010 Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Proceedings of the Gateway Computing Environments Workshop* (*GCE*), *14 November 2010, New Orleans, LA*, pp. 1–8. Piscataway, NJ: IEEE.
50. Huelsenbeck JP, Larget B, Alfaro ME. 2004 Bayesian phylogenetic model selection using reversible jump Markov Chain Monte Carlo. *Mol. Biol. Evol.* **21**, 1123–1133. (doi:10.1093/molbev/msh123)
51. Rambaut A. 2012 FigTree version 1.4. See http://tree.bio.ed.ac.uk/software/figtree/.
52. Ersts P. 2012 Geographic Distance Matrix Generator (version 1.2.3). See http://biodiversityinformatics.amnh.org/open_source/gdmg.
53. Darriba D, Taboada GL, Doallo R, Posada D. 2012 jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772. (doi:10.1038/nmeth.2109)
54. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013 MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729. (doi:10.1093/molbev/mst197)
55. Tamura K, Nei M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.
56. Excoffier L, Lischer HEL. 2010 Arlequin suite version 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567. (doi:10.1111/j.1755-0998.2010.02847.x)
57. Zwickl D. 2008 GARLI (Genetic algorithm for rapid likelihood inference), version 0.96. See http://www.bio.utexas.edu/faculty/antisense/garli/garli.html.
58. Purvis A, Bromham L. 1997 Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny.

J. Mol. Evol. **44**, 112–119. (doi:10.1007/PL00006117)
59. Yang Z. 2007 PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)
60. R Core Team. 2014 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
61. Hyams DG. 2010 CurveExpert software version 1.3. See https://www.curveexpert.net/.
62. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014 BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537. (doi:10.1371/journal.pcbi.1003537)
63. Bouckaert RR, Drummond AJ. 2017 bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* **17**, 42. (doi:10.1186/s12862-017-0890-6)
64. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016 Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*. **2**, vew007. (doi:10.1093/ve/vew007)
65. Brockmann D, Helbing D. 2013 The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342. (doi:10.1126/science.1245200)
66. Pybus OG, Tatem AJ, Lemey P. 2015 Virus evolution and transmission in an ever more connected world. *Proc. R. Soc. B* **282**, 20142878. (doi:10.1098/rspb.2014.2878)
67. Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005 Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* **22**, 1561–1568. (doi:10.1093/molbev/msi145)
68. Wertheim JO, Pond SLK. 2011 Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.* **28**, 3355–3365. (doi:10.1093/molbev/msr170)
69. Saxenhofer M, Weber de Melo V, Ulrich RG, Heckel G. 2017 Data from: Revised time scales of RNA virus evolution based on spatial information. Dryad Digital Repository. (http://dx.doi.org/10.5061/dryad.dc770)