THE ROYAL SOCIETY FirstCite® e-publishing

Published online

Viral evolution and the emergence of SARS coronavirus

Edward C. Holmes* **and Andrew Rambaut**

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

The recent appearance of severe acute respiratory syndrome coronavirus (SARS-CoV) highlights the continual threat to human health posed by emerging viruses. However, the central processes in the evolution of emerging viruses are unclear, particularly the selection pressures faced by viruses in new host species. We outline some of the key evolutionary genetic aspects of viral emergence. We emphasize that, although the high mutation rates of RNA viruses provide them with great adaptability and explain why they are the main cause of emerging diseases, their limited genome size means that they are also subject to major evolutionary constraints. Understanding the mechanistic basis of these constraints, particularly the roles played by epistasis and pleiotropy, is likely to be central in explaining why some RNA viruses are more able than others to cross species boundaries. Viral genetic factors have also been implicated in the emergence of SARS-CoV, with the suggestion that this virus is a recombinant between mammalian and avian coronaviruses. We show, however, that the phylogenetic patterns cited as evidence for recombination are more probably caused by a variation in substitution rate among lineages and that recombination is unlikely to explain the appearance of SARS in humans.

Keywords: emerging viruses; SARS coronavirus; evolution; phylogeny; recombination

1. INTRODUCTION

Since the first descriptions of AIDS in the early 1980s, much has been written about the causes and consequences of emerging viral diseases. Yet, despite an extensive research effort, viral infections continue to appear in human and wildlife populations, as demonstrated by the 'new' viruses identified since the rise of AIDS, such as HCV, Sin Nombre, Nipah, Hendra and most recently SARS-CoV, and more venerable pathogens, including West Nile virus and dengue virus, that have recently expanded their global prevalence. Even less progress has been made in what is perhaps the ultimate goal of research into emerging viruses, predicting what viruses are likely to emerge in the future.

Defining emerging viruses as those that are newly appeared or have recently increased in prevalence and/or geographical range reveals some important general patterns. First, almost all emerging viruses have RNA rather than DNA genomes. Although RNA viruses are ordinarily more commonplace than DNA viruses, we will argue later that this difference primarily reflects the differing evolutionary rates of these two types of infectious agent. Second, almost all emerging viruses have an animal reservoir, such that the process of viral emergence can usually be categorized as cross-species transmission (Cleaveland *et al.* 2001). For example, HIV type 1 (HIV-1), the major cause of AIDS, has its origins in the related SIV found in chimpanzees (Gao *et al.* 1999), while SARS-CoV has close relatives in Himalayan palm civets (Guan *et al.* 2003), although it is not yet established that these are the source population for the human form of the virus. The most significant exception to the rule that cross-species transmission is central to viral emergence is HCV, which was first identified in 1989 but which is likely to have a much longer history in human populations (Simmonds 1995). Despite surveying a number of animal species, the ultimate reservoir species for HCV remains a mystery, although this is more likely to reflect the fact that the wrong species have been surveyed or that the reservoir viruses are too divergent in sequence to be recognized, than the absence of an animal reservoir altogether.

In many cases, the specific cause of emergence—why the virus has crossed from animals into humans—can be assigned to ecological factors, often relating to changes in land use and deforestation (Morse 1995). Although a multitude of such factors exist, they can often be placed into one of two general groups; either changes in the proximity of the donor and recipient populations, so that humans have an increased chance of exposure to animal pathogens, or changes in the size and density of the donor and recipient populations, which increases both the exposure and the likelihood that sustained networks of transmission will be established once a virus has entered a new species. Moreover, it is clear that, as human ecology has changed through time with, for example, the rise of farming and later urbanization, so our burden of infectious disease has increased (Dobson & Carper 1996). Accepting the general importance of ecology, it is also possible that genetic factors, in either the host or more probably the virus, contribute to the process of disease emergence. As these genetic factors have been considered only in a cursory manner up until now, we will outline in general evolutionary terms

* Author for correspondence (edward.holmes@zoo.ox.ac.uk).

One contribution of 15 to a Discussion Meeting Issue 'Emerging infections: what have we learnt from SARS?'.

the genetic basis of viral emergence before considering the specific case of SARS-CoV.

2. THE EVOLUTIONARY GENETICS OF VIRAL EMERGENCE

The elemental nature of the evolutionary interaction between host and pathogen is critical to understanding the mechanics of viral emergence. On the host side, different species, or individuals within a species, may have differing susceptibilities to a specific viral infection. However, as host evolution obviously occurs on a different temporal scale to viral evolution (Schliekelman *et al.* 2001), it is more profitable to consider the differing abilities of viruses to cross species boundaries. Most fundamentally, different viruses or strains within a particular virus may differ in their ability to recognize the cellular receptors of a new host species (Baranowski *et al.* 2001) or in their ability to transmit successfully between individuals in the new host species (with unsuccessful transmission resulting in 'dead-end' infections). Given that viruses may need to adapt to replicate in different host species, it is also likely that the more genetically variable and hence adaptable the virus in question, the more likely it will be to be able to jump species boundaries and establish productive infections in new host species (Woolhouse *et al.* 2001). It is this fact that gives RNA viruses the edge in emergence: the mutation rates of RNA viruses are many orders of magnitude greater than those of their DNA counterparts (although some populations of DNA viruses are highly variable, which hints at higher mutation rates; Sanz *et al.* 1999). On average, RNA polymerases produce almost one error in each replication cycle (Drake *et al.* 1998; Malpica *et al.* 2002), thus when populations of RNA viruses are large, they will produce a myriad of potentially adaptively useful genetic variation. Similarly, while many DNA viruses lead to persistent infections in their hosts, many RNA viruses (with the notable exception of retroviruses) generate acute infections. This is critical to viral emergence because a short duration of infection means that the most likely way for RNA viruses to infect new host species is through cross-species transmission, rather than long-term cospeciation, which is usually associated with persistence (Holmes 2004).

Although it is clear that the high mutation rates of RNA viruses enhance their adaptability, a more compelling question is whether all RNA viruses are equally equipped in this respect? Put another way, given the same amount of exposure, are all RNA viruses equally likely to jump species boundaries? This question is at the heart of understanding the evolutionary genetics of viral emergence, and, although we are a long way from a complete answer, there is growing evidence that specific evolutionary constraints make host switching more likely in some RNA viruses than in others. An important idea in this context is that there is a general phylogenetic rule regarding the ability of a virus to jump hosts: the more phylogenetically distant the host species in question, the less likely it is that their viruses will be able to jump between them (DeFilippis & Villarreal 2000). For example, the reservoir populations for most human viruses are other mammalian species, and while we probably eat virally infected plant matter on a regular basis we do not suffer the viruses experienced by plants (Hull *et al.* 2000). On a more localized scale, studies of HIVs and SIVs suggest that the ability of these viruses to jump species to some extent reflects the phylogenetic relationships of the hosts (Charleston & Robertson 2002), although the fact that HIV-2 jumped from sooty mangabey monkeys to humans shows that exceptions are possible (Hahn *et al.* 2000). Such a phylogenetic trend is compatible with a simple evolutionary rule: the more specialized species become in one environment (in this case a particular host species), the less likely it is that they will be able to adapt to new environments (Bell 1997). The rapid pace of RNA virus evolution means that these host specificities are likely to be established quickly, as observed in experimental systems (Turner & Elena 2000). Testing the extent of the relationship between phylogenetic distance and the ability to jump hosts should be one of the key areas for future research into viral emergence.

At face value, it may seem strange that with their remarkable power of mutation RNA viruses are not able to exploit every adaptive solution. Ironically, the adaptive constraints faced by RNA viruses may be a function of their high mutation rates, as this may limit their genome size, which in turn hinders their ability to increase complexity. The causal link between mutation rate and genome size can be made by invoking the concept of the 'error threshold'. This theory was first introduced by Eigen as a crucial element in the evolution of the first RNA replicators (Eigen 1987), though in reality it can be extended to any living system (Maynard Smith & Szathmáry 1995). Put simply, the theory states that there is a maximum error rate that is tolerable for a genome of a particular size: as most mutations are deleterious, longer genomes than those imposed by the threshold would be over-burdened with deleterious mutations leading to dramatic fitness losses and eventual extinction. Hence, for RNA viruses that have high mutation rates because of their intrinsically error-prone RNA polymerases, genome sizes must be small to prevent the accumulation of lethal numbers of deleterious mutations.

Evidence for an error threshold in RNA virus evolution comes from a number of sources. First, RNA viruses occupy a very narrow range of genome sizes, with a median size of only *ca*. 9 kb and a maximum size of *ca*. 32 kb, as exhibited by some coronaviruses. This is in dramatic contrast to DNA viruses, which can range in size from only a few thousand bases to *ca*. 400 kb (figure 1), and suggests that the upper limit on genome size cannot be the result of common virological factors, such as capsid size or packaging requirements. Second, in experiments in which mutagens are applied to populations of RNA viruses as a form of antiviral therapy, thereby increasing the mutation rate, fitness declines dramatically, as expected if the manipulated mutation rate has breached the error-threshold (Sierra *et al.* 2000; Crotty *et al.* 2001). Finally, although the observed substitution rates of RNA viruses (which can be regarded as markers of the background mutation rate if most mutations are neutral) fall into a fairly narrow distribution, there is a significant negative correlation between substitution rate and genome size, exactly as predicted under the error-threshold model (Jenkins *et al.* 2002). The relationship between genome size and mutation rate is of particular importance for coronaviruses such as SARS-CoV: as their genomes are

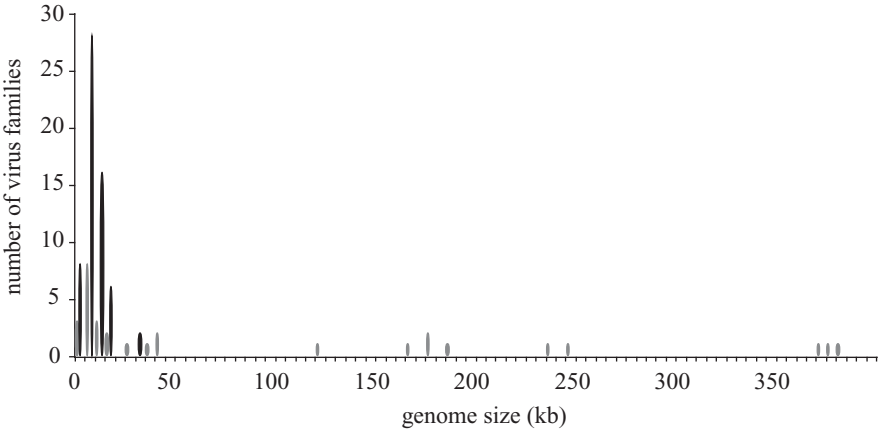

Figure 1. Distribution of genome sizes of families of RNA (black bars) and DNA (grey bars) viruses. All data taken from http://www.ncbi.nlm.nih.gov/ICTV/.

relatively long, their mutation rates should be correspondingly low, although, to our knowledge, this has yet to be formally tested.

By limiting genome size, high mutation rates act to constrain RNA virus evolution (Eigen 1996; Holmes 2003*a*). In particular, small genome sizes mean that sequence regions may sometimes encode multiple functions and that individual mutations will be subject to rather complex fitness trade-offs. Briefly, pleiotropy and epistasis will be major players in RNA virus evolution. Although this theory has yet to be widely tested, sequence analyses are starting to provide evidence for evolutionary constraints in RNA virus evolution (some of which may also be found in small DNA viruses). For example, many RNA viruses use overlapping reading frames as this increases the information content in small genomes. Likewise, because the limit to genome size means that there are relatively few nucleotide sites that are free to vary, convergent evolution, in which the same solution evolves on multiple occasions, appears to be relatively common in RNA viruses (Cuevas *et al.* 2002). On a larger scale, there is growing evidence that intricate fitness trade-offs are important in shaping RNA virus evolution. A well-studied example concerns the arthropod viruses of vertebrates ('arboviruses'), which are unusual in that they replicate in hosts that are phylogenetically very different. From the argument presented earlier, such a life-history strategy might be expected to be extremely difficult, given the very different natures of the host cells in each case. Analyses of both selection pressures (Woelk & Holmes 2002) and substitution rates (Jenkins *et al.* 2002) now reveal that arboviruses are subject to weaker positive selection pressure, suggesting that they are indeed subject to complex fitness trade-offs. This has been particularly well documented in dengue virus, in which the analysis of sequence diversity revealed a very strong force of purifying selection (Holmes 2003*b*), presumably because amino acid changes that work well in primates lower fitness in mosquitoes (and vice versa). Finally, the upper bound on the genome sizes of RNA viruses means that they will be less subject to the processes of gene duplication and lateral gene transfer (from either hosts or other viruses) that appear to be common in the evolution of bacteria (Daubin *et al.* 2003), eukaryotes (McLysaght *et al.* 2002) and large DNA viruses (McLysaght *et al.* 2003). Interestingly, coronaviruses represent one of the few examples of lateral gene transfer in an RNA virus, as viruses assigned to the mammalian group 2 coronaviruses have seemingly acquired the haemagluttinin-esterase gene from the influenza C virus (Marra *et al.* 2003). Whether lateral gene transfer has played a more extensive role in generating the large genomes of coronaviruses remains to be seen.

Finally, it is a simple matter to predict that constraints on viral evolution in general will also affect their ability to emerge in new host species in particular. Hence, although all RNA viruses are likely to mutate rapidly, it is not necessarily the case that they will always be able to adapt to replicate and transmit in a new host species: the intricate epistatic and pleiotropic environments experienced by a particular virus may mean that the mutations required to infect a new host species will lower some other component of fitness, even if they are relatively simple to produce by mutation. As such, documenting the details of pleiotropy and epistasis in viral evolution is likely to be crucial to understanding viral emergence.

3. IS THE EMERGENCE OF SARS-CoV THE RESULT OF RECOMBINATION?

There is growing evidence that some RNA viruses are able to generate adaptively useful genotypic variation through recombination as well as through mutation (Worobey & Holmes 1999). One such group are the coronaviruses, within which recombination in the spike glycoprotein has been extensively described (reviewed in Lai 1996). More controversially, it has also been suggested that SARS-CoV may be a recombinant of different coronaviruses (Rest & Mindell 2003; Stanhope *et al.* 2004; Stavrinides & Guttman 2004), an event that may even have been central to its recent emergence in humans (Stanhope *et al.* 2004; Stavrinides & Guttman 2004). Although a variety of different recombination analyses have been undertaken and different recombination events proposed, they follow the same general scheme in which a mammalian coronavirus assigned to group 1 (such as human coronavirus 229E) or group 2 (such as human coronavirus OC43 or mouse hepatitis virus) has recombined with an avian coronavirus assigned to group 3 (such as infectious bronchitis virus), giving rise to the distinct evolutionary lineage represented by SARS-CoV. Given

the obvious importance of SARS-CoV, particularly the desire to predict how it might re-emerge in the future, it is important to assess the evidence that it is a recombinant. Moreover, if true, it would represent one of the few cases in which a distinct genetic event has played a key role in viral emergence.

Despite the evidence for incongruence in the evolutionary history of SARS-CoV, in which different gene regions produce different phylogenetic trees, which is often taken as strong evidence for the action of recombination, there are a number of reasons why this theory should be treated with caution. First, SARS-CoV very clearly represents a distinct evolutionary lineage, roughly equidistant from the coronaviruses previously assigned to groups 1, 2 and 3, irrespective of what genes are analysed (Marra *et al*. 2003; Rota *et al*. 2003; figure 2). Indeed, SARS-CoV is so phylogenetically distinct that it can reasonably be regarded as a fourth lineage of the coronaviruses, although some authors have suggested that it is most likely to share a common ancestry with the group 2 mammalian coronaviruses (Eickmann *et al.* 2003). Although this tree structure does not in itself rule out ancient recombination events among the different groups of coronaviruses, it effectively eliminates any role for recombination in the emergence of SARS in humans. This point is particularly well illustrated by the phylogenetic position of the SARS-CoV strains from humans relative to those isolated from the Himalayan palm civet. These viruses are almost indistinguishable at the amino acid level (and appear as almost identical in figure 2) and are much more closely related to each other than they are to the other groups of coronaviruses. If the civet is indeed the reservoir species for SARS-CoV, then the jump to humans occurred a very long time after any putative recombination events involving different coronaviruses.

It is also probable that the main evidence presented for recombination—phylogenetic incongruence—is in fact caused by lineages of coronaviruses evolving at very different rates. Coronavirus phylogenies are distinctive for two reasons. First, sequences from different groups of coronaviruses are highly divergent, with average amino acid distances between them indicating that at least one amino acid replacement has been fixed at each site. Distances are even greater when the arteriviruses are included in the analysis as outgroups to the coronaviruses. Although using outgroups to root the tree of coronaviruses greatly assists in the analysis of recombination (Rest & Mindell 2003; Stavrinides & Guttman 2004), the sequences in this case are so divergent that accurate assessment of positional homology is difficult, seriously compromising the analysis. Even if there were no recombination events in the history of the three coronavirus lineages and SARS-CoV, when dividing the genome into short fragments and estimating the phylogenetic relationships for each, we would expect incongruence among the resulting phylogenies owing to the stochastic nature of molecular evolution. This will be particularly true in this case because the tree linking SARS-CoV to the other coronaviruses is highly distinctive, in that it comprises variable-length external branches (representing the four major lineages of coronaviruses), linked by a short internal branch (figure 2). Tree topologies with this general structure can be subject to long-branch attraction (Felsenstein 1978), in which the same changes evolve on unrelated long branches, such that rapidly evolving lineages tend to group together. Moreover, rate variation among lineages will also bias both network methods of phylogenetic analysis, such as split decomposition (Worobey *et al.* 2002), which have also been used to provide evidence for recombination in SARS-CoV (Stavrinides & Guttman 2004), and analyses of phylogenetic robustness, such as the bootstrap or quartet puzzling.

To test the possibility that rate variation among lineages has produced a false-positive signal for recombination, we reanalysed the data provided in one of the studies suggesting that SARS-CoV has a recombinant history—that of Stavrinides & Guttman (2004). Based on a phylogenetic analysis, these authors suggested that the replicase polyprotein 1a and the spike glycoprotein are of mammalian origin, since SARS-CoV is a sister-group to the group 2 coronaviruses in trees of these proteins, whereas the membrane glycoprotein and the nucleocapsid protein are of avian origin, as SARS-CoV is most closely related to the group 3 coronaviruses in these proteins. These conflicting phylogenetic positions are shown in figure 2. We undertook a series of likelihood-ratio tests to determine whether these four gene trees had significantly different tree topologies (unlike bootstrap and quartet puzzling, likelihood-ratio tests are not biased by rate variation). First, model tree topologies depicting each possible phylogenetic position of SARS-CoV were constructed in which it was related, in turn, to coronavirus groups 1, 2 and 3. Next, the likelihoods of each of the competing trees were compared using a maximum-likelihood method (full details of the procedure used are given in the legend to figure 2). The results of this analysis are striking as in three out of the four proteins analysed the likelihoods of the competing trees are so similar that none can be significantly favoured over any other, strongly arguing against the hypothesis of incongruence and hence recombination (table 1). In only a single comparison—that involving the highly variable spike glycoprotein—can one phylogenetic hypothesis, in this case that of SARS-CoV being most closely related to the group 2 mammalian coronaviruses, significantly reject the competing trees. As such, there is in fact little major difference in the tree topologies reconstructed using the different genes of SARS-CoV, with any differences in branching order more probably reflecting rate variation than distant recombination; the localized differences in evolutionary rate among genes have produced tree topologies that show minor differences at the base of the coronavirus tree, giving a false impression of ancient recombination.

One way in which the problem of long-branch attraction can be reduced is by including more taxa in the analysis, especially those that break up long branches, as this tends to distribute the convergent and parallel mutations more evenly across the tree, thereby reducing their influence (Hillis 1996). Indeed, it is inevitable that the sample of coronaviruses currently available is only a subset, and perhaps a tiny subset, of those actually present in nature. With such a potentially small sample of lineages it is difficult, if not dangerous, to reach firm conclusions about the evolutionary history of SARS-CoV, particularly whether its ancestry lies with mammalian or avian coronaviruses. A key task for the future should therefore be a more extensive sampling of the genetic diversity of the coronaviruses in

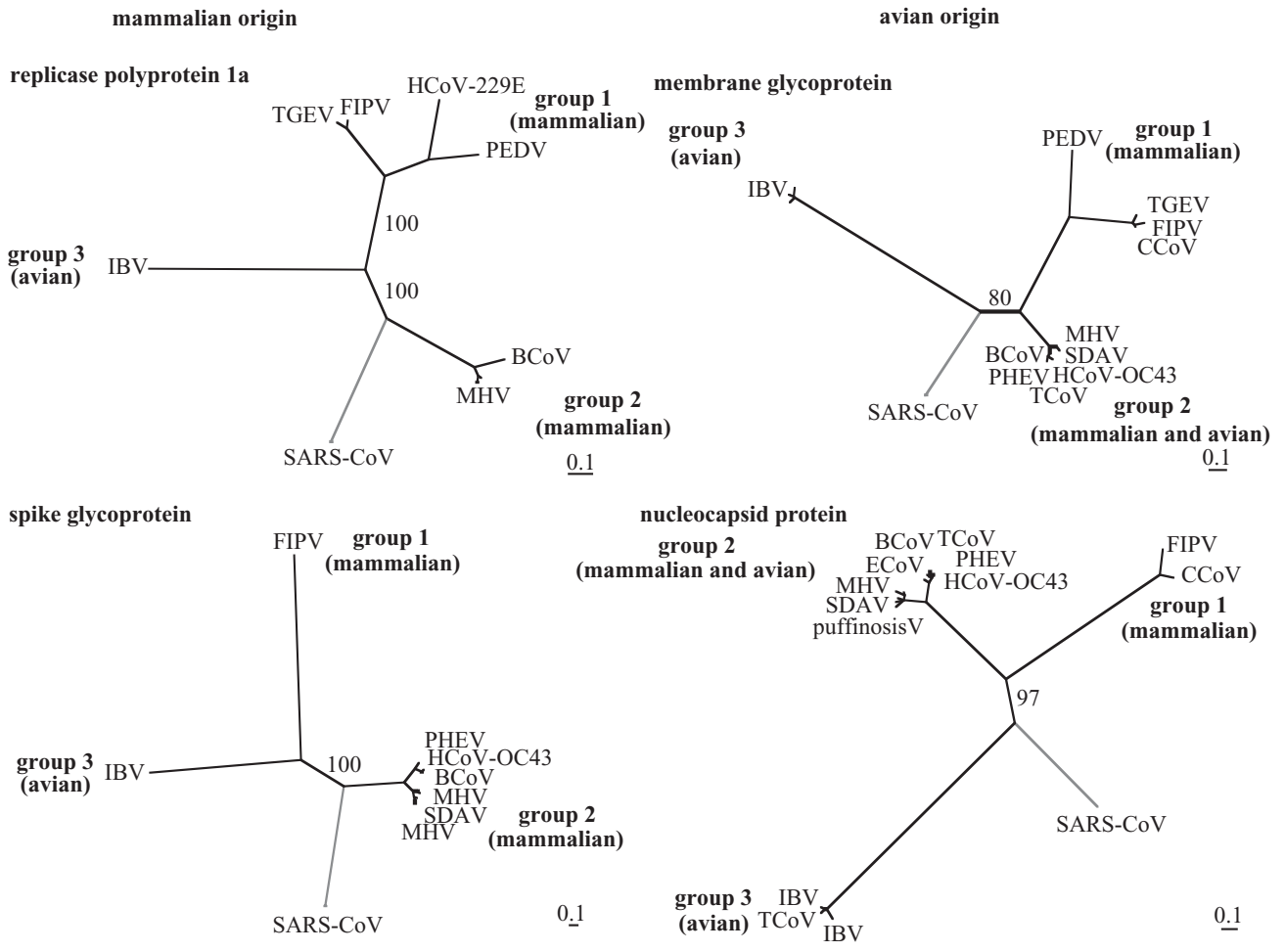

Figure 2. Phylogenetic relationships of SARS-CoV within the coronaviruses. Phylogenetic trees were inferred using the maximum-likelihood option in TREE-PUZZLE (Strimmer & von Haeseler 1996) for the following gene regions and alignments kindly made available by Stavrinides & Guttman (2004): replicase polyprotein 1a (11 sequences, 338 amino acids), spike glycoprotein (12 sequences, 1270 amino acids), membrane glycoprotein (15 sequences, 214 amino acids) and nucleoprotein (17 sequences, 415 amino acids). In all cases the Whelan and Goldman model of amino acid replacement was used (Whelan & Goldman 2001). A gamma distribution of rate heterogeneity was also incorporated, with the value of the shape parameter (α) estimated from the empirical data (parameter values available from the authors on request). Numbers next to the main branches of the tree depict quartet puzzling support values, which give an indication of the reliability of each branch (with 100 signifying maximum support for the branch in question). All trees are unrooted, with branches drawn to a scale of the number of amino acid replacements per site. The following sequences were analysed (abbreviated viral names, where applicable, and NCBI accession numbers given in parentheses): group 1 coronaviruses: canine coronavirus (CCoV; BAC65328, AAO33711), feline infectious peritonitis virus (FIPV; AAK09095, CAA29535, CAA39850, CAA39851), human coronavirus 229E (HCoV-229E; NP_073549), porcine epidemic diarrhoea virus (PEDV; AF353511) and transmissible gastroenteritis virus (TGEV; AJ271965); group 2 coronaviruses: bovine coronavirus (BCoV; AF220295, P26020), equine coronavirus (ECoV; AAG39339), human coronavirus OC43 (HCoV-OC43; P33469, Q01455, S44241), murine hepatitis virus (MHV; AF029248, AF201929, AF208066, CAA28484, NP_068668, P18446), porcine haemagglutinating encephalomyelitis virus (PHEV; AAL80031, AAM77004, AAM77005), puffinosis virus (CAD67607) and rat sialodacryoadenitis coronavirus (SDAV; AAF97738, AAF97742); group 3 coronaviruses: infectious bronchitis virus (IBV; AAF35863, AAk83027, AJ311317) and turkey coronavirus (TCoV; P26021, AAF23872); SARS coronaviruses: Himalayan palm civet (SARS-CoV), strain SZ16 (AY304488) and human SARS coronavirus (SARS-CoV), strain CUHK-AG01 (AY345986).

nature. However, on the basis of the current data, the evidence that SARS-CoV has a recombinant history is weak at best, and there is nothing to suggest that recombination has played a role in the emergence of SARS in humans.

4. CONCLUSIONS

Viral diseases pose a continual threat to human populations. As we live in ever-increasing populations and become increasingly mobile so it is inevitable that new viruses, such as SARS-CoV, will appear. Although in most cases it is ecological changes that trigger viral emergence, as is most probably the case with SARS, it is evident that some viruses are more predisposed than others to jump species barriers. The study of viral evolutionary genetics is therefore critical to understanding fundamental aspects of viral emergence. Genetics may also play a role in predicting what diseases might emerge in the future. In particular, it will soon be possible to survey those animal populations that are most likely to harbour potentially emergent viruses. First, it is a relatively simple matter to make predictions about what sorts of animals are most

Table 1. Maximum-likelihood analysis of tree topologies depicting the possible phylogenetic positions of SARS-CoV within the coronaviruses.
(Trees were compared using the Kishino–Hasegawa test (Kishino & Hasegawa 1989). In all cases, the competing topologies were compared with the maximum-likelihood (ML) tree, with $p < 0.05$ deemed to be a significant difference. Only in the spike glycoprotein could one tree (SARS-CoV with the group 2 coronaviruses) be significantly favoured over another.)

protein/tree topology	−ln L[a]	difference to ML tree	significantly worse?
membrane glycoprotein			
SARS-CoV + group 1	2727.04	2.91	no
SARS-CoV + group 2	2725.69	1.56	no
SARS-CoV + group 3	2724.13	ML tree	—
nucleocapsid			
SARS-CoV + group 1	5566.40	3.34	no
SARS-CoV + group 2	5567.00	3.93	no
SARS-CoV + group 3	5563.06	ML tree	—
replicase polyprotein 1a			
SARS-CoV + group 1	4170.34	6.11	no
SARS-CoV + group 2	4164.23	ML tree	—
SARS-CoV + group 3	4171.44	7.21	no
spike glycoprotein			
SARS-CoV + group 1	16901.50	12.20	yes
SARS-CoV + group 2	16889.30	ML tree	—
SARS-CoV + group 3	16901.89	12.60	yes

[a] −ln L, log likelihood.

likely to carry viruses new to humans. As with humans, species with large and/or dense populations are the most likely to carry a wide variety of pathogens as well as the most virulent diseases. This includes some species of rodents, bats and birds, particularly those that live in close proximity to humans and are closely related to us. A more significant advance perhaps comes from modern molecular biology. It is now possible to develop degenerate PCR primers for families of RNA viruses (based on conserved genes), which can be used to survey animal populations, or even specific environments, for new viruses. The recent application of these methods in a marine setting uncovered an enormous diversity of previously unknown RNA viruses (Culley *et al.* 2003). In addition to simply surveying biodiversity, it may be possible to isolate any viruses detected and determine whether they are able to grow in human cells. If such viruses are found, continually monitoring their spread and even attempting vaccination in reservoir species are reasonable strategies. Although these techniques are only just beginning to be developed and clearly represent a long-term research programme, they have the potential to provide efficient tools to survey the natural world for those pathogens that could have devastating effects on our health.

The authors thank the Wellcome Trust (grant 071979) and The Royal Society for financial support, and Robin Weiss and Angela McLean for extremely useful comments.

REFERENCES

Baranowski, E., Ruiz-Jarabo, C. M. & Domingo, E. 2001 Evolution of cell recognition by viruses. *Science* **292**, 1102–1105.

Bell, G. 1997 *Selection: the mechanism of evolution*. New York: Chapman & Hall.

Charleston, M. A. & Robertson, D. L. 2002 Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst. Biol.* **51**, 528–535.

Cleaveland, S., Laurenson, M. K. & Taylor, L. H. 2001 Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Phil. Trans. R. Soc. Lond.* B **356**, 991–999. (DOI 10.1098/rstb.2001.0889.)

Crotty, S., Cameron, C. E. & Andino, R. 2001 RNA virus error catastrophe: direct test by using ribavirin. *Proc. Natl Acad. Sci. USA* **98**, 6895–6900.

Cuevas, J. M., Elena, S. F. & Moya, A. 2002 Molecular basis of adaptive convergence in experimental populations of RNA viruses. *Genetics* **162**, 533–542.

Culley, A. I., Lang, A. S. & Suttle, C. A. 2003 High diversity of unknown picorna-like viruses in the sea. *Nature* **424**, 1054–1057.

Daubin, V., Moran, N. A. & Ochman, H. 2003 Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832.

DeFilippis, V. R. & Villarreal, L. P. 2000 An introduction to the evolutionary ecology of viruses. In *Viral ecology* (ed. C. J. Hurst), pp. 126–208. San Diego, CA: Academic Press.

Dobson, A. P. & Carper, E. R. 1996 Infectious diseases and human population history. *Bioscience* **46**, 115–126.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. 1998 Rates of spontaneous mutation. *Genetics* **148**, 1667–1686.

Eickmann, M. (and 15 others) 2003 Phylogeny of the SARS coronavirus. *Science* **302**, 1504–1505.

Eigen, M. 1987 New concepts for dealing with the evolution of nucleic acids. *Cold Spring Harbor Symp. Quantitative Biol.* **52**, 307–320.

Eigen, M. 1996 On the nature of viral quasispecies. *Trends Microbiol.* **4**, 216–218.

Felsenstein, J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.

Gao, F. (and 11 others) 1999 Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441.

Guan, Y. (and 17 others) 2003 Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278.

Hahn, B. H., Shaw, G. M., de Cock, K. M. & Sharp, P. M. 2000 AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–614.

Hillis, D. M. 1996 Inferring complex phylogenies. *Nature* **383**, 130–131.

Holmes, E. C. 2003*a* Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* **11**, 543–546.

Holmes, E. C. 2003*b* Patterns of intra- and inter-host nonsynonymous variation reveal strong purifying selection in dengue virus. *J. Virol.* **77**, 11 296–11 298.

Holmes, E. C. 2004 The phylogeography of human viruses. *Mol. Ecol.* **13**, 745–756.

Hull, R., Covey, S. & Dale, P. 2000 Genetically modified plants and the 35S promoter: assessing the risks and enhancing the debate. *Microb. Ecol. Hlth Dis.* **12**, 1–5.

Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. 2002 Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* **54**, 152–161.

Kishino, H. & Hasegawa, M. 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the hominoidea. *J. Mol. Evol.* **29**, 170–179.

Lai, M. M. C. 1996 Recombination in large RNA viruses: coronaviruses. *Seminars Virol.* **7**, 381–388.

McLysaght, A., Hokamp, K. & Wolfe, K. H. 2002 Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31**, 200–204.

McLysaght, A., Baldi, P. F. & Gaut, B. S. 2003 Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl Acad. Sci. USA* **100**, 14 960–14 965.

Malpica, M. J., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W. & García-Arenal, F. 2002 The rate and character of spontaneous mutation in an RNA virus. *Genetics* **162**, 1505–1511.

Marra, M. A. (and 57 others) 2003 The genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399–1404.

Maynard Smith, J. & Szathmáry, E. 1995 *The major transitions of evolution*. Oxford: Freeman.

Morse, S. S. 1995 Factors in the emergence of infectious diseases. *Emerg. Infect. Dis.* **1**, 7–15.

Rest, J. S. & Mindell, D. P. 2003 SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting. *Infect. Genet. Evol.* **3**, 219–225.

Rota, P. A. (and 34 others) 2003 Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**, 1394–1399.

Sanz, A. I., Fraile, A., Gallego, J. M., Malpica, J. M. & García-Arenal, F. 1999 Genetic variability of natural populations of cotton leaf curl geminivirus, a single-stranded DNA virus. *J. Mol. Evol.* **49**, 672–681.

Schliekelman, P., Garner, C. & Slatkin, S. 2001 Natural selection and resistance to HIV. *Nature* **411**, 545–546.

Sierra, S., Dávila, M., Lowenstein, P. R. & Domingo, E. 2000 Response of foot-and-mouth disease virus to increased mutagenesis. Influence of viral load and fitness in loss of infectivity. *J. Virol.* **74**, 8316–8323.

Simmonds, P. 1995 Variability of hepatitis C virus. *Hepatology* **21**, 570–583.

Stanhope, M. J., Brown, J. R. & Amrine-Madsen, H. 2004 Evidence from the evolutionary analysis of nucleotide sequences for a recombinant history of SARS-CoV. *Infect. Genet. Evol.* **4**, 15–19.

Stavrinides, J. & Guttman, D. S. 2004 Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J. Virol.* **78**, 76–82.

Strimmer, K. & von Haeseler, A. 1996 Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969.

Turner, P. E. & Elena, S. F. 2000 Cost of host radiation in an RNA virus. *Genetics* **156**, 1465–1470.

Whelan, S. & Goldman, N. 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.

Woelk, C. H. & Holmes, E. C. 2002 Reduced positive selection in vector-borne RNA viruses. *Mol. Biol. Evol.* **19**, 2333–2336.

Woolhouse, M. E. J., Taylor, L. H. & Haydon, D. T. 2001 Population biology of multihost pathogens. *Science* **292**, 1109–1112.

Worobey, M. & Holmes, E. C. 1999 Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.* **80**, 2535–2544.

Worobey, M., Rambaut, A., Pybus, O. P. & Robertson, D. L. 2002 Questioning the evidence for genetic recombination in the 1918 'Spanish flu' virus. *Science* **296**, 211.

GLOSSARY

HCV: hepatitis C virus

HIV: human immunodeficiency virus

SARS: severe acute respiratory syndrome

SARS-CoV: severe acute respiratory syndrome coronavirus

SIV: simian immunodeficiency virus