

## THE POPULATION GENETICS AND EVOLUTIONARY EPIDEMIOLOGY OF RNA VIRUSES

Andrés Moya\*, Edward C. Holmes<sup>‡</sup> and Fernando González-Candelas\*

RNA viruses are ubiquitous intracellular parasites that are responsible for many emerging diseases, including AIDS and SARS. Here, we discuss the principal mechanisms of RNA virus evolution and highlight areas where future research is required. The rapidity of sequence change in RNA viruses means that they are useful experimental models for the study of evolution in general and it enables us to watch them change in 'real time', and retrace the spread through populations with molecular phylogenies. An understanding of the mechanisms of RNA virus sequence change is also crucial to predicting important aspects of their emergence and long-term evolution.

### GENETIC DRIFT

Stochastic changes in allele frequencies in a finite population due to the random sampling of genes at reproduction.

### FITNESS

The contribution of a genotype to the next generation, relative to that of other genotypes in the population, reflecting its probability of survival and its reproductive output.

### ERROR THRESHOLD

Maximum mutation rate that is tolerable for a given genome size. Crossing the error threshold leads to dramatic fitness losses.

\**Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, Apartado Postal 22085, 46071 Valencia, Spain.*  
<sup>‡</sup>*Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.*  
 Correspondence to A.M.  
 e-mail: andres.moya@uv.es  
 doi:10.1038/nrmicro863

RNA viruses have become an important area of study for epidemiologists and evolutionary biologists alike<sup>1–3</sup>. Much of this research is centred on two main themes; understanding the mechanisms of RNA virus evolution, often through experimental analyses, and reconstruction of the epidemiological history of a given virus, namely its origin and spread through populations and the forces that promote its emergence. Here, we review the progress and current state of both these research topics.

Mechanisms of viral evolution

**Processes of evolutionary change.** Central to population genetics is understanding how the five main forces of evolutionary change — mutation, recombination, natural selection, GENETIC DRIFT and migration — interact to shape the genetic structure of populations. These same forces are also central to understanding RNA virus evolution, although their relative strengths differ to those observed for DNA-based organisms.

For RNA viruses, most attention has been directed towards mutation, selection and genetic drift. We can understand their importance and interaction by considering four basic properties of RNA virus populations. First, RNA viruses often have very large population sizes, such that the number of viral particles in a given organism might be as high as 10<sup>12</sup>. Second, such immense population sizes, which are several orders of

magnitude larger than those observed for cellular organisms, are a product of explosive replication. For example, a single infectious particle can produce an average of 100,000 viral copies in 10 hours. As natural selection is most efficient with large populations, it is no surprise that experiments using RNA viruses have shown that selection is of fundamental importance in controlling their evolutionary dynamics, such that new mutants with increased FITNESS (as measured by their selection coefficient, *s*) continually appear and out-compete older, inferior alleles<sup>4</sup>. Third, owing to the lack of proofreading activity in their polymerase proteins, RNA viruses exhibit the highest mutation rates of any group of organisms, approximately one mutation per genome, per replication<sup>5,6</sup>. Finally, the genome sizes of RNA viruses are typically small, ranging from only 3 kb to ~30 kb, with a median size of ~9 kb. These last two properties are intimately related because high-mutation rates are theoretically expected to limit genome size. In particular, a mutation rate that exceeds a notional ERROR THRESHOLD (set at approximately the reciprocal of the genome size) generates so many deleterious mutations in each replication cycle that even the fittest viral genomes are unable to reproduce, and population size decreases to extinction<sup>7,8</sup>. However, RNA viruses that exist close to (but below) the error threshold are also

able to produce many beneficial mutations in a short time, thereby enhancing adaptability, provided that their populations are sufficiently large.

In the simple situation outlined above, RNA viruses should evolve in a highly deterministic manner, with the process of natural selection working efficiently on a vast array of mutational variants. Although it is true that RNA virus populations are often highly diverse, this is not sufficient to explain the entirety of RNA virus evolution. In particular, deterministic approaches assume that population sizes are universally large, such that the fate of a given mutation can be predicted if its frequency and fitness are known. Although the population sizes of RNA viruses are often very large, factors such as variation in replication potential among variants, differences in generation time among infected cells and POPULATION BOTTLENECKS, most notably during transmission between hosts, might lead to an EFFECTIVE POPULATION SIZE (denoted  $N_e$ ) that is much smaller than the actual number of infected cells. Theory predicts that in populations where  $N_e$  is small (such that the compound parameter  $N_e s < 1$ ), genetic drift has an important role in determining the frequency and fate of mutations<sup>9</sup>.

Recombination might also have an important role in RNA virus evolution. Although most studies indicate that recombination rates in many RNA viruses are often lower than those in other organisms<sup>10</sup>, there are notable exceptions. Perhaps the most dramatic is HIV, in which the genomic recombination rate exceeds the genomic mutation rate<sup>11</sup>. Frequent recombination seems advantageous because it can create high fitness genotypes more rapidly than by mutation alone. Moreover, recombination might also purge deleterious mutations from virus populations, thereby preventing a dramatic decrease in fitness (see below). However, simulation studies have indicated that frequent recombination is more likely to reduce fitness when mutation rates are close to the error threshold<sup>12</sup>. Finally, recombination rates in RNA viruses might not be set by natural selection at all. Rather, they could simply be a passive function of the replication machinery or ecological circumstances of the virus in question. For example, recombination rates seem to be particularly low in negative-sense RNA viruses<sup>13</sup>, which might be a result of the RNA-packaging mechanism. Understanding the causes of variation in recombination rate among RNA viruses is a key area for future study.

A final factor to consider in RNA virus evolution is migration. Migration (also referred to as gene flow) must not only be understood at a macroscopic level (that is, among hosts within a population, among populations or between host species), but also within a single infected individual. From the site of inoculation, viruses can be transported to several tissues, generating intra-host spatial variation<sup>14</sup>. However, the effect of a non-uniform population distribution on the spread, fitness and variability of virus populations has been much less studied than other evolutionary factors, although in some experiments a positive correlation between migration rate and the average fitness of the population has been observed<sup>15</sup>.

**The quasispecies as a model of RNA virus evolution.**

The remarkable mutational power of RNA viruses has meant that their evolution has often been considered to be different to DNA-based organisms<sup>1</sup>. Key to this is the concept of the quasispecies, which was first developed by Eigen and Schuster<sup>16</sup> to understand the dynamics of primitive evolutionary systems. RNA viruses are of particular importance in this respect as they might represent biological entities that evolve according to the rules of quasispecies theory. The basis of quasispecies theory is the notion that the target of natural selection is not simply the fastest growing replicator, but rather a broad spectrum of mutants that are produced by erroneous copying of the fittest (or master) sequence<sup>16,17</sup>. Natural selection acts on the entire quasispecies because mutation rates close to the error threshold mean that individual viral genomes are linked by a mutational coupling — all the possible mutational links between viral genomes are established — so that the whole population evolves as a single unit. One particularly important implication of this special form of group selection is that the fastest replicating RNA viral genomes could be out-competed by those with lower replication rates if the latter have a high probability of being generated by mutation from closely related variants.

An important question is whether the quasispecies model is an accurate description of RNA virus evolution. Experimental evidence for the quasispecies was first reported for the bacteriophage Q $\beta$ <sup>18</sup>. Subsequent experiments with mammalian vesicular stomatitis virus (VSV) provided one of the most important supporting observations for the quasispecies — a high-fitness viral variant was suppressed by one of lower fitness<sup>19</sup>. However, this can also be explained by genetic drift; the probability that any variant achieves fixation in a population is partially dependent on its initial frequency, so most rare, albeit advantageous, variants are lost by drift in small populations. Indeed, a generic problem of quasispecies theory is that genetic drift is expected to be extremely restricted<sup>17</sup>, which might not be the case for viruses in Nature<sup>20,21</sup>. More recently, *in vitro* studies of the evolutionary dynamics of bacteriophage  $\phi$ 6 provided evidence for one aspect of quasispecies theory — that viral genomes differ in their mutational spectra and that this affects fitness<sup>22</sup>. However, because these experiments used small populations and RUGGED FITNESS LANDSCAPES, and because little is known about fitness landscapes in Nature, the generality of these results is uncertain.

Although the quasispecies has a firm theoretical foundation, and there is some evidence for it in laboratory populations, whether it applies to RNA viruses in nature is less clear. For example, simple observations of high levels of genetic variation in RNA viruses are not sufficient to prove the existence of quasispecies, although this is often the only evidence presented, nor is the existence of an error threshold, which can easily be explained using evolutionary models<sup>8</sup>. Rather, to demonstrate that RNA viruses form quasispecies it is necessary to show that natural selection acts on viral

**POPULATION BOTTLENECK**

A severe reduction in population size that causes a loss of genetic variation. Under strong bottlenecks, genetic drift can be more important in evolution than natural selection.

**EFFECTIVE POPULATION SIZE**

( $N_e$ ). The size of an idealized population that would experience genetic drift in the same way as the actual population.  $N_e$  can be lower than the census population size ( $N$ ) owing to various factors, including a history of population bottlenecks and variance in reproductive rates.

**RUGGED FITNESS LANDSCAPE**

A rugged fitness landscape is one with multiple fitness peaks. The more rugged a landscape the lower the average fitness correlations between neighbouring points.

**CONVERGENT EVOLUTION**

The independent evolution of similar traits in two or more unrelated or distantly related lineages.

**CLONAL INTERFERENCE**

In asexual populations beneficial mutations compete (or 'interfere') with each other as they go to fixation. Therefore, the fixation of advantageous mutations is sequential. This has important implications for the rate of adaptive evolution.

**COMPLEMENTATION**

The cooperative interaction of mutant genes in viral populations resulting in a phenotype closer to the wild type. In a broader sense, it refers to the use of genetic information belonging to another member of the population.

**MULLER'S RATCHET**

The successive build-up of deleterious mutations in finite asexual populations. It has been proposed to be an important reason for the evolution of sexual reproduction.

## Box 1 | Experiments in RNA virus evolution

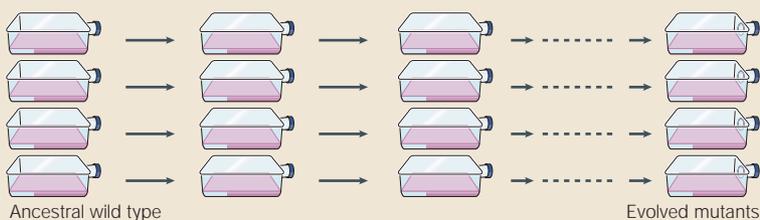
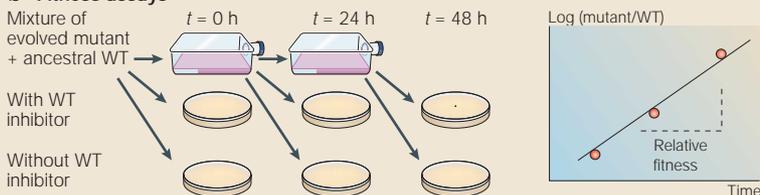
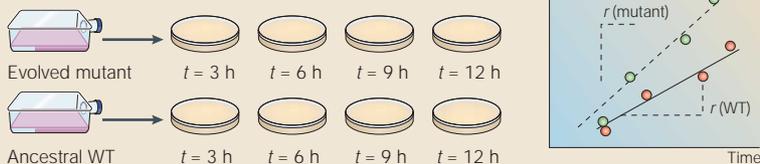
At the start of an experiment (see figure panel a),  $n$  flasks containing a cell monolayer are infected with the original mutant (resistant to the wild-type (WT) inhibitor) population. Subsequently, a sample is transferred between flasks until the end of the experiment, when evolved mutants are collected. The final samples can then be analysed in two different ways.

**Fitness assays**

This technique (shown in panel b of the figure) requires a mixture of the evolved mutant and the ancestral WT, which is then used to infect cell monolayers in a flask and two additional dishes (with and without the WT inhibitor, respectively). On the following day, a sample is taken from the first flask to infect cell monolayers in a new flask and two new dishes (under the same conditions as used previously). The sample from the second flask is then used to infect two new dishes, with and without the WT inhibitor. Plaque numbers are counted as a measure of virus particle numbers. Calculating the logarithm of the mutant/WT quotient against time gives the relative fitness of the evolved mutant.

**Growth curves**

This procedure (shown in panel c of the figure) requires the infection of the evolved mutants into cell monolayers in separate flasks, taking several samples from each flask at different times, and using the ancestral WT clone as a control. Plaque numbers are counted in the plate assay to measure viral growth. Calculation of the logarithm of the viral titre against time ( $t$ ) shows the growth rate for the evolved mutant populations.

**a Experimental evolution****b Fitness assays****c Growth curves**

populations as a unit. Testing this prediction in Nature will be one of the most important future areas of study for those investigating the mechanics of viral evolution.

**Experimental evolution of RNA viruses**

Experimental evolution constitutes a powerful tool for simulating natural evolution<sup>23</sup> and is frequently used to test the basic principles of evolutionary theory<sup>24</sup>. BOX 1 provides a schematic overview of experiments of this type. The strength of the experimental approach is that the phenotypic and molecular changes of RNA viral populations can be monitored in real time. More importantly, under an appropriate experimental setting,

it is also possible to directly estimate changes in fitness, one of the main goals of modern population genetics. A panoply of experiments have highlighted both the mechanisms of RNA virus evolution and what RNA viruses can tell us about evolution in general.

**What have experiments told us about viral evolution?**

Experimental studies have made a substantial contribution towards understanding the processes that govern RNA virus evolution<sup>25</sup>. The main findings of these experiments can be grouped into three general types. First, the molecular basis of adaptive evolution in viruses, including the occurrence and frequency of CONVERGENT EVOLUTION<sup>26,27</sup>, viral attenuation<sup>28</sup> and compensatory mutations<sup>29</sup>. Second, the role of population bottlenecks and the accumulation of deleterious mutations, and how they affect fitness<sup>30–33</sup>. Finally, the importance of CLONAL INTERFERENCE<sup>34</sup> and COMPLEMENTATION<sup>35</sup> in determining rates of viral adaptation. In brief, these studies make eight conclusions<sup>36</sup>. First, there is extensive convergent and parallel evolution (both in genotype and in phenotype) across lineages replicating in the same host, perhaps reflecting the fact that relatively few sites are free to vary when genome sizes are small. Second, advantageous mutations that are fixed early on when viruses are challenged with new environments confer the largest fitness benefit. Third, phenotypic evolution tends to mirror the evolution of fitness increments, with large changes occurring early in new environments. Fourth, rates of nucleotide substitution remain approximately constant through time. Fifth, overall genetic diversity remains low during the phase of maximum fitness increase and rises once fitness becomes asymptotic. Sixth, evolutionary changes that increase fitness in one host often, although not always, reduce fitness in an alternative host. Seventh, population bottlenecks and spatial heterogeneity lead to an increase in unique nucleotide substitutions. Finally, severe reductions in population size can lead to the accumulation of deleterious mutations and consequent fitness losses.

**What have experiments told us about evolution in general?**

Many fundamental questions in evolutionary biology have been addressed using RNA viruses as model experimental systems. One of the first questions addressed concerned a severe consequence of deleterious mutation accumulation known as MULLER'S RATCHET<sup>37</sup>. Studies have explored how clonally evolving RNA viruses prevent the excessive build-up of deleterious mutations in populations that are experiencing strong bottlenecks or small effective sizes<sup>38</sup>. One popular suggestion is that sexual reproduction has evolved in RNA viruses, because it allows them to escape Muller's ratchet when effective population sizes are small<sup>39–41</sup>. In this model, the accumulation of deleterious mutations is expected to be less of a problem for RNA viruses that either recombine or undergo reassortment in the case of viruses with segmented genomes. However, recombination rates are often so low in RNA viruses that it is difficult to hypothesize that they have a direct fitness benefit. In the face of low

recombination rates, RNA viruses could escape Muller's ratchet because their long-term effective population sizes might not be small enough to allow deleterious mutations to accumulate, or perhaps compensatory mutations are sufficiently frequent to counteract fitness losses<sup>42</sup>.

Experimental evolution with RNA viruses has also been crucial for studying the dynamics of natural selection. This has been the case, for instance, in studies of COMPETITIVE EXCLUSION<sup>43</sup>, the RED QUEEN HYPOTHESIS<sup>43</sup>, convergent evolution<sup>26,27,44</sup> and the rhythm of adaptive change<sup>45</sup>. We expect RNA viruses to continue to have an important role in this area for many years to come. Finally, one promising area involves using RNA viruses to test theories for the evolution of cooperation. For example, Turner and Chao conducted a series of experiments with bacteriophage  $\phi 6$  in which they demonstrated that RNA viruses could evolve under PRISONER'S DILEMMA conditions<sup>46</sup>, and also escape from it<sup>47</sup>.

Despite the power of experiments, there are still difficulties in estimating some of the important parameters in evolutionary biology, such as the rate of deleterious mutation and fitness values. In recent years, a number of *in silico* approaches have been used to answer these questions, most notably with computer-generated genomes ('digital organisms') that are designed to behave as living systems. Digital organisms have the ability to create a copy of their own genome, but are subject to copying errors, so that populations of programs evolve in, and adapt to, their environments<sup>48</sup>. Although digital organisms are not as sophisticated as viruses, they are useful study tools because experiments can be easily controlled and repeated. Digital organisms have been useful in studying predicted adaptive evolution over short and long timescales<sup>49,50</sup>, the role of epistasis in evolution<sup>51</sup> and testing key aspects of quasispecies theory, particularly whether fast replicators can be less fit than slower replicators at high-mutation rates<sup>52</sup>.

#### COMPETITIVE EXCLUSION

In the absence of ecological niche differentiation, only one of a set of competing species can occupy a particular niche, leading to the elimination of other species.

#### RED QUEEN HYPOTHESIS

A dynamic equilibrium between competing species, in which no species can ever win, and new adversaries continually replace the losers. It therefore depicts an ongoing arms race.

#### PRISONER'S DILEMMA

A theoretical game that highlights the costs and benefits of the evolution of cooperation. For example, games can be played when the population is composed of 'defectors' and 'cooperators'.

#### MOLECULAR CLOCK

The theory that the evolution of gene or protein sequences proceeds at a constant rate.

#### Molecular epidemiology of RNA viruses

Molecular epidemiology was first introduced to the study of infectious disease in the early 1970s<sup>53</sup>. Since this time, the analysis of gene sequence variation has become a standard practice for virologists with an interest in epidemiology, especially with the advent of new, high-throughput sequencing technologies. From the typing of viral populations to the study of the origins of a new virus, viral gene sequence variation has been used to answer a wide variety of questions, increasing both the quantity and quality of the epidemiological data available. In this section, we discuss how particular aspects of RNA virus evolution affect the reconstruction of their epidemiological history. In this context, it is important to note that the observed epidemiological patterns of viruses result from their evolution at two different levels: within individual hosts<sup>54</sup> (and vectors<sup>55</sup>) and among hosts at the population level. RNA viruses differ greatly in their patterns and processes of intra- and inter-host evolution, as well as in the duration of the infection caused and the type of immune response that is induced. Such factors must be considered when discussing their epidemiological dynamics in a comparative setting<sup>3</sup>.

**Inferring epidemiological history.** Many aspects of the epidemiological history of viruses can be graphically summarized in a phylogenetic tree. The timescale of these trees can vary from a few weeks to many centuries, and depends on the rate of accumulation of variation in the sequences under study and the timescale of sampling. Although graphically very similar, there are important differences between phylogenetic trees and gene genealogies<sup>56</sup>. The former are used to analyse the evolutionary history of distinct viral species or genes, usually by sampling one representative of each unit under study. By contrast, genealogies depict the history of genetic polymorphisms segregating in contemporaneous populations. Gene genealogies have been used extensively in the study of RNA viruses

#### Box 2 | Molecular epidemiology and forensics

Establishing the source of a viral outbreak is one of the many applications of molecular epidemiology. Occasionally, this work has implications other than epidemiological. For instance, Ou *et al.*<sup>68</sup> were able to identify an HIV-infected dentist as the inadvertent donor of the virus to some of his patients, and to simultaneously discard him as the source of viruses infecting others. More recently, a similar analysis identified a medical doctor as the source of HIV infection in a former lover<sup>133</sup>. This is the first case in which a molecular epidemiological analysis has been accepted as evidence in a criminal court in the United States.

Owing to the rapid rate of RNA virus evolution, they can be useful in forensic medicine, given a rigorous statistical framework. For this, it is necessary to translate the proposals of both the defendant and the prosecutor into testable phylogenetic hypotheses. González-Candelas *et al.*<sup>134</sup> used this approach to evaluate the individual likelihoods of a number of hepatitis C virus patients having been infected in a hospital from a common source. In this case, the prosecutor's hypothesis was that each patient was infected at the hospital. Conversely, the defendant's hypothesis was that each patient became separately infected from the general population. These two proposals translate into different phylogenetic tree topologies, the likelihoods of which can be compared directly. The prosecutor's proposal is equivalent to observing a monophyletic group comprising the infected patients, which is significantly distinct from the background population, whereas the defendant's proposal corresponds to the grouping of the sequences from each infected patient with those of the background population. This molecular analysis proved that all the individuals were from a single outbreak, and was the first report of the use of molecular phylogenies to determine the likelihood of a patient sharing the source of infection with other infected patients.

because their rapid evolution means that sequence variation increases over very short periods of time. Indeed, RNA viruses constitute the most important class of measurably evolving populations<sup>57</sup>, with evolution even occurring during the infection of a single individual<sup>58</sup>. More accurate methods of evolutionary inference have already been designed for these rapidly evolving populations<sup>59–61</sup>. In particular, whereas phylogenetic trees generally depict strictly bifurcating patterns of relationships, gene genealogies can take into account recombination, which will lead to interconnected networks of lineages. Moreover, using a statistical approach called coalescent theory<sup>62–64</sup>, it is possible to infer demographic processes from genetic polymorphism data, most notably, rates of viral population growth and decline. Coalescent methods can operate under several population genetic models and use gene genealogies as key analytical tools. Coalescent theory therefore provides a crucial conceptual link between phylogenetics and population genetics.

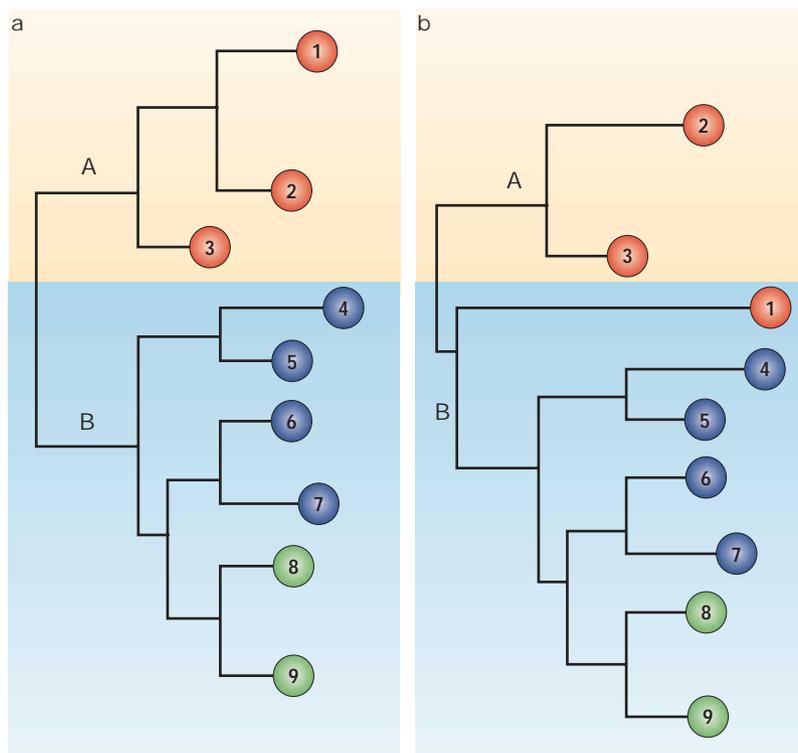
Both phylogenies and gene genealogies are relevant for the epidemiological analysis of RNA viruses and are useful for investigating the origin of new viruses<sup>65–67</sup> or identifying the source of an outbreak<sup>68–71</sup>. Used in conjunction with population genetic theory, they are also

essential for the identification of positive<sup>72,73</sup>, or purifying<sup>74</sup>, natural selection at nucleotide sites, the presence and extent of recombination<sup>13,75,76</sup> and for dating important points in the history of epidemics<sup>67,77,78</sup>. Genealogical analysis is especially relevant for reconstructing the recent epidemiological history of viral populations<sup>66,79–81</sup>, such as in forensic studies (BOX 2, FIG. 1), and therefore have important implications for determining public health policies.

**Rates of evolutionary change in RNA viruses.** Although it is fair to assume that frequent mutation means that long-term rates of nucleotide substitution are usually high in RNA viruses, in reality these rates might vary widely, both within and among genes in the same species and among viral species<sup>82</sup>. Indeed, present data indicate that viral substitution rates are much more variable than their underlying mutation rates<sup>5,6</sup>, which is most likely a reflection of important differences in replication dynamics. For example, the nucleotide substitution rate in human T-lymphotropic virus type II (HTLV-II) varies from  $\sim 1 \times 10^{-4}$  substitutions per site per year in epidemics with high rates of transmission and where replication is rapid, such as those in injecting drug users, to  $\sim 1 \times 10^{-7}$  substitutions per site per year in endemic situations, where viruses are maintained within hosts through the clonal expansion of infected cells rather than by active replication<sup>83</sup>. However, in many RNA viruses, substitution rates of  $1 \times 10^{-3}$  to  $1 \times 10^{-4}$  substitutions per site per year are observed<sup>82</sup>. The variation in substitution rates across viral genomes has benefits, because it allows different epidemiological questions to be addressed, relating to different temporal scales. So, rapidly evolving (hypervariable) gene regions are informative for studying viral evolution within individuals, or for identifying the source of a particular disease outbreak. More conserved regions are better suited for in-depth phylogenetic inference, from analysing viral genotypes at the species to family levels.

In a number of cases it has been proposed that molecular evolution within specific RNA virus genes proceeds at a constant rate. Such MOLECULAR CLOCKS have been proposed for human influenza A virus<sup>84</sup>, although the constancy of the evolutionary rate does not hold in many other cases<sup>82,85–87</sup>. Non-clocklike evolution can result from a number of evolutionary forces, such as changes in host species, changes in structural and functional constraints<sup>88,89</sup>, and the occurrence of positive selection. Although most modern methods of phylogenetic analysis incorporate such rate variation — so that it is unlikely to cause significant error in the reconstruction of tree topologies — it can have an important impact on the analysis of divergence times.

The emergence of RNA viruses  
The past 25 years have seen the emergence of several RNA viruses, which are either new to medical science or have increased in prevalence to the extent that they are now a major concern for public health. Agents that fall into this category include HIV, hepatitis C virus and, most recently, severe acute respiratory syndrome



**Figure 1 | Two alternative epidemiological scenarios translate into different phylogenetic tree topologies, the statistical support for which can be compared directly.** The tree in panel **a** depicts a common and close origin for samples 1–3 (node A), which is separate from the control samples 4–7 (node B). Node A might correspond to a single outbreak or a suspected transmission among these patients, whereas node B includes samples suspected, but not related to, the outbreak (4–7) and unrelated population controls (8, 9). Panel **b** represents the alternative proposal for sample 1, which is now separated from the former cluster and instead groups with the control samples. Similar proposals can be separately formulated for each of the samples 1–3.

coronavirus (SARS-CoV). Given the continuing threat that is posed by viral diseases, it is essential that we determine the factors underlying viral emergence.

**The evolutionary genetics of viral emergence.** Hosts acquire RNA viruses by two different mechanisms. First, owing to host–virus co-speciation, host populations might have carried a specific RNA virus for their entire evolutionary history. Although co-speciation has been proposed in some RNA viruses<sup>90</sup>, the process seems to be rare. This is most likely a result of the short infectious periods of many RNA viruses, so that they have limited opportunity for the sustained transmission that is probably needed for co-speciation. By contrast, many DNA viruses establish persistent infections and are therefore expected to be able to follow long-term patterns of host speciation<sup>91</sup>.

A more common method by which RNA viruses could enter new host species is through lateral transfer from different reservoir species. Both ecology and genetics seem to be important in determining whether a virus is able to successfully cross species boundaries. In many cases, ecological factors are the most important. Although such factors are diverse, and have been reviewed in detail elsewhere<sup>92,93</sup>, they usually reflect changes in either the proximity or density of the host and/or reservoir species, which increase the likelihood that humans are exposed to new pathogens and that sustained transmission networks will be established.

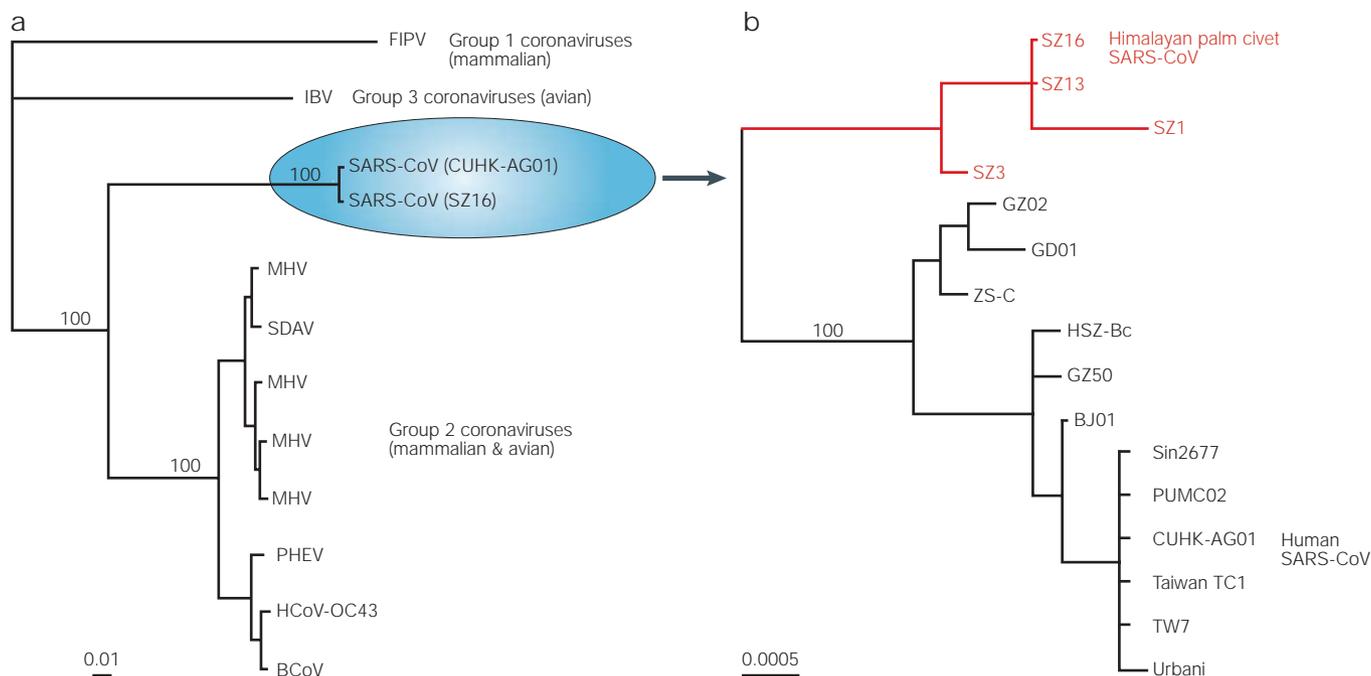
Far less is known about the possible genetic factors that might affect the ability of viruses to cross species boundaries. Although RNA viruses are the group of pathogens that seem most able to cross species boundaries<sup>94</sup>, perhaps because high mutation rates provide them with an increased capacity to adapt to new hosts<sup>95</sup>, not all RNA viruses are equally equipped in this respect. For example, in many cases, RNA viruses (such as rabies virus infection in humans) establish ‘dead-end’ infections in specific hosts, without subsequent transmission, which reflects imperfect adaptation. This indicates that there are constraints that inhibit viral adaptation to new hosts, perhaps owing to the fitness trade-offs that seem commonplace in viruses that need to infect different hosts or cell types<sup>25,96</sup>. Therefore, infecting different hosts is likely to represent a major adaptive challenge for RNA viruses, despite their mutation rate. Examples are animal vector-borne viruses, which are less subject to adaptive evolution than their non-vector-borne counterparts, presumably owing to the difficulties that are associated with simultaneous replication in hosts as divergent as invertebrates and vertebrates<sup>97</sup>. If extended over longer periods of time, this will lead to a phylogenetic rule of cross-species transmission, such that the greater the evolutionary distance between hosts, the lower the probability of viral transfer among them<sup>98</sup>.

A fundamental aspect of the mechanistic basis of viral emergence is the relationship between virus and host cell receptors<sup>99</sup>. Unless a virus has sequences that are able to recognize the cellular receptors of a potential host species, successful cross-species transmission will not be

possible. Therefore, jumping species boundaries might only be a problem for an RNA virus if it has to adapt to different cellular receptors, although this still does not guarantee that sustained transmission will be established in the new host. An informative example is provided by influenza A virus. Birds are the main species reservoir, and avian influenza A viruses are usually unable to jump directly into humans because they lack the necessary mutations in the haemagglutinin (HA) gene to infect human cells<sup>100</sup>. Even when avian influenza A viruses do infect humans, human-to-human transmission might not be established. More generally, the relationship between the virus and the host cell receptor predicts an association between the number of cells a virus infects and its host range, thereby explaining whether a virus is a host ‘specialist’ or ‘generalist’. Determining whether such a relationship exists should be a key goal in understanding the genetic basis of viral emergence.

**Case studies in viral emergence.** The complex interplay between ecology and genetics in viral emergence can be seen in HIV. An important ecological factor in HIV emergence involves the bushmeat trade in west Africa. Not only have a wide range of related simian immunodeficiency viruses (SIVs) been isolated from animal carcasses<sup>101</sup>, but the bushmeat trade has increased owing to encroachment by humans on the ranges occupied by non-human primates. The SIV that is found in chimpanzees (SIVcpz) is the closest relative, and therefore the most likely ancestor, of human HIV-1 (REF. 102), whereas SIVsm from sooty mangabeys seems to be the reservoir population for HIV-2 (REF. 103). For both HIV-1 and HIV-2, there have been multiple transfers of virus from their reservoirs into humans, with these viruses most likely establishing themselves in humans during the last century<sup>67,77</sup>. Also of importance, was the movement of individuals infected early in the epidemic from small, isolated rural populations to cities in Africa<sup>103</sup>, which enabled incipient epidemic strains to reach a large number of susceptible hosts. Yet, genetics is also likely to have been important in the emergence of HIV. In particular, phylogenetic studies indicate that SIVs are most easily transmissible among related primate hosts<sup>104</sup>, implying that not all possible instances of cross-species viral transmission that could occur do occur, and that adaptive constraints might exist.

A more recent and highly publicized example of viral emergence is provided by the SARS-CoV, the agent of a severe form of pneumonia that has killed more than 700 people worldwide since its appearance in China in November 2002. It is unclear whether the epidemic of 2002–2003 was the first appearance of SARS, or whether the virus had sporadically entered human populations previously, but without detrimental consequences. The animal reservoir for SARS-CoV is also a subject for debate. Phylogenetic analysis reveals that SARS-CoV is equidistant between coronavirus groups 1 and 2, which are usually isolated from mammalian species, and coronavirus group 3, which is currently confined to birds<sup>65,105,106</sup> (FIG. 2). Moreover, the sequence divergence between these three groups



**Figure 2 | The phylogenetic relationships of SARS coronavirus (SARS-CoV) inferred using sequences of the spike glycoprotein. a** | Phylogenetic relationship of SARS-CoV to the known coronaviruses. Owing to the highly divergent nature of these viruses, the analysis was conducted using an alignment of 12 amino acid sequences that are 1,270 residues in length. The tree was inferred using the maximum likelihood (ML) method available in TREE-PUZZLE<sup>135</sup>. Numbers next to some branches represent quartet puzzling support values, which give an indication of the reliability of that branch. SARS-CoV appears as a distinct lineage. **b** | Magnified phylogeny of representative SARS-CoV strains isolated from humans and the Himalayan palm civet (*Paguma larvata*), a putative reservoir species. The tree was constructed using the same region as in part **a** but using nucleotide sequences (16 sequences, 3,765 bp). The tree was inferred using the ML method available in PAUP<sup>+</sup><sup>136</sup>. Maximum-likelihood bootstrap values are shown for the main branches. Both trees are mid-point rooted and all horizontal branches are drawn to a scale of the number of substitutions per site (note the difference in scale between the two trees). All parameter settings used in the phylogenetic analysis are available from the authors on request. The following sequences were analysed (abbreviated viral names and GenBank accession numbers are given in parentheses); Group 1 coronaviruses: feline infectious peritonitis virus (FIPV; CAA29535); Group 2 coronaviruses: bovine coronavirus (BCoV; AF220295), human coronavirus OC43 (HCoV-OC43; S44241), murine hepatitis virus (MHV; AF029248, AF201929, AF208066, CAA28484), rat sialodacryoadenitis coronavirus (SDAV; AAF97738); porcine haemagglutinating encephalomyelitis virus (PHEV; AF481863); Group 3 coronaviruses: infectious bronchitis virus (IBV; AJ311317); SARS coronaviruses: Himalayan palm civet SARS-CoV, strains SZ1 (AY304489), SZ3 (AY304486), SZ13 (AY304487) and SZ16 (AY304488), and human SARS-CoV, strains Sin2677 (AY283795), BJ01 (AY278488), CUHK-AG01 (AY345986), GD01 (AY278489), GZ02 (AY390556), GZ50 (AY304495), HSZ-Bc (AY394994), PUMCO2 (AY357075), Taiwan TC1 (AY338174), TW7 (AY502930), Urbani (AY278741) and ZS-C (AY395003).

and SARS-CoV is so large that SARS-CoV has clearly experienced a long period of independent evolution. Studies of animals sold at Chinese markets have detected antibodies in a number of mammalian species<sup>107</sup>. Most notably, viruses obtained from the Himalayan palm civet (a member of the *Viverridae*) are closely related to human strains of SARS-CoV. Whether this species represents the main reservoir for SARS-CoV is still unclear. Finally, there have also been suggestions that SARS-CoV is a recombinant of mammalian and avian coronaviruses and that this genetic event might have triggered viral emergence<sup>108</sup>. However, because the sequences involved are so divergent, the phylogenetic incongruence in trees of SARS-CoV seems more likely to be due to variation in the molecular clock than inter-coronavirus recombination.

**Predicting viral emergence.** The 'holy grail' for studies of emerging diseases is to predict which infectious agents are likely to infect human populations in the future. Although we are a long way from making accurate predictions, evolutionary genetics does allow

some basic rules to be established, and phylogenetic methods have been used to successfully predict the future population survival of strains of influenza A virus<sup>109</sup>. First, the larger the population size of the reservoir species, the more viruses it can harbour, including those with shorter durations of infection and increased virulence<sup>110</sup>. Consequently, animal species that live at high densities, such as some bats, rodents and birds, are most likely to be reservoirs, particularly those animal populations that already live in close proximity to humans. Less intuitively, if there is a relationship between the breadth of cell tropism and the number of species infected, most attention should be given to those viruses that infect several cell types. More importantly, a comprehensive survey of RNA virus diversity should be undertaken in appropriate animal species. This can be done through the use of degenerate PCR primers that have been designed for several RNA virus families, followed by studies to determine whether the viruses will grow in human cells. Similar approaches have already uncovered a plethora of new virus families from marine environments<sup>111</sup>.

RNA virus evolution in the long term

One aim of studies of RNA virus evolution is to use our understanding of evolutionary processes in the short term, which have often been acquired from experiments, to predict what evolution will do in the long term. Although evolutionary biologists are rightly nervous about predicting future change, the rapid pace of RNA virus evolution means that these predictions can be tested quickly. Of most immediate interest are patterns of drug resistance and viral virulence.

**The evolution of drug resistance.** Understanding drug resistance is one area in which population biology has a direct impact on public health<sup>112,113</sup>. In the case of RNA viruses, most interest has focused on the potential of drugs to control HIV infection. Despite the optimism that initially surrounded the deployment of highly active antiretroviral therapy (HAART), which involves combinations of drugs<sup>114</sup>, antiviral therapy is unlikely to provide a cure for HIV. There are several reasons for this, not least of which is that despite our inability to detect viruses in some patients receiving HAART, viral replication is ongoing, although at greatly reduced levels<sup>115</sup>. Early studies predicting the length of time it would take for resistance to arise under multiple drug therapy also underestimated the importance of recombination in HIV, which we now know is extensive<sup>116</sup>. Frequent recombination could allow drug resistance to be acquired more rapidly than acquisition through mutation alone.

There are many factors that influence the evolution of drug resistance, and important results have been obtained — for example, regarding the probability of resistance mutations arising before and during treatment, and the optimal time for the onset of drug treatment<sup>117</sup>. One important question, which also relates to the mechanics of viral evolution in general, is whether drug-resistance mutations have a fitness cost compared with wild-type alleles in the absence of the drug. If there is a fitness cost, we would hypothesize that resistance mutations would not reach high frequencies in populations, despite their benefit to the virus in hosts. There is evidence that, in the absence of the drug, HIV strains harbouring drug-resistant mutations are less fit than wild-type HIV strains<sup>118</sup>. Unfortunately, in other cases, drug-resistant HIV mutants seem to have greater infectivity, and even replication capacity, than wild-type viruses<sup>119</sup>. Not surprisingly, these mutations are increasingly sampled from drug-naïve patients<sup>120</sup>.

Even if drug mutations are universally advantageous, their long-term success depends on more than their individual fitness. An important mediating factor is the strength of genetic drift at the population level. If drift is strong, which will be the case if effective population sizes are small, the frequencies that mutations eventually attain in populations has a large stochastic component. For RNA viruses, effective population size reflects the mode of transmission. This can be shown by comparing HIV with influenza A virus<sup>3</sup>. In the case of the respiratory transmitted influenza A virus, the

host population is large and extensively mixed. Consequently, advantageous mutations, most notably those that confer antigenic escape, are able to be fixed in the virus population in a regular manner<sup>121</sup>. By contrast, population-level selection seems to be considerably weaker in the sexually transmitted HIV, although some evidence for long-term cytotoxic T lymphocyte (CTL)-mediated selection has been found<sup>122</sup>. This contrasts with intra-host HIV evolution, in which immune-driven natural selection is the dominant evolutionary process<sup>72,123,124</sup>. The reduced impact of selection at the population level is most likely to be caused by extensive variation in rates of partner exchange, which in turn reduces the effective population size<sup>125</sup>, and because there is a large bottleneck at transmission<sup>126</sup>. Therefore, for HIV, intra-host and inter-host evolution seem to be largely decoupled. Although fewer studies have compared intra- and inter-host evolution in acute RNA virus infections, a recent analysis of dengue virus indicated that most amino acid changes that arise within hosts are deleterious in the long term<sup>74</sup>.

**The evolution of viral virulence.** The evolution of virulence has long been of interest to population biologists. A common view is that if virulence is a selected trait at all, then it is often involved in a trade-off with transmissibility; the balance of these factors that maximizes the BASIC REPRODUCTIVE RATE ( $R_0$ ) of the pathogen is favoured by natural selection<sup>127</sup>, although this has recently been questioned<sup>128</sup>. For RNA viruses it is therefore important to determine whether virulence is optimized and, if so, how it is linked to transmissibility. Complexities arise because virulence is also likely to vary according to the transmission mode<sup>129</sup>, and whether there is a long period of intra-host evolution, including superinfection by other strains, which increases intra-strain competition and therefore virulence<sup>130</sup>. In short, predictions about the long-term evolution of virulence in RNA viruses need to be made on a case-specific basis.

However, some aspects of the evolution of virulence reflect those that are associated with drug resistance. For example, if particular mutations confer virulence, then whether they become fixed in populations also depends on the strength of genetic drift, even if they are advantageous. Consequently, the optimal level of virulence might not be acquired by chance in small populations. Similarly, if the evolutionary process differs greatly within and among hosts, a selectively favoured level of virulence within hosts might be disadvantageous among hosts. The intra-host evolution of HIV tends to result in the production of high-virulence viral strains that preferentially use the CXCR4 chemokine receptor, infect cells faster and cause AIDS to develop more rapidly<sup>131</sup>. However, these strains seem to be transmitted less often, indicating that they are selectively disadvantageous in new hosts<sup>132</sup>. Understanding the interplay between virulence and transmissibility is clearly central to understanding the evolution of virulence of RNA virus in diseases.

**BASIC REPRODUCTIVE RATE ( $R_0$ ).** For viruses, this is defined as the number of secondary infections that are caused by a single index case in an entirely susceptible population. For an epidemic to proceed,  $R_0$  needs to exceed 1.

## Conclusions

Establishing the rules of RNA virus evolution is important: not only will this provide information that is essential for understanding the basic mechanisms of evolutionary change, but it will assist in the design of strategies for the control, treatment and eradication of RNA viruses, and perhaps for predicting their emergence. Although it is clear that RNA viruses are unique in the rapidity with

which they mutate, their evolution cannot be described in full without a consideration of all the processes of evolutionary change. A particular challenge for the future is to determine whether viral evolution in nature is similar to that established *in vitro*. The beauty of RNA viruses is that the link between experimental and natural systems can be made simply — few other organisms are as well suited for studying evolutionary processes.

1. Domingo, E. & Holland, J. J. RNA virus mutations and fitness for survival. *Ann. Rev. Microbiol.* **51**, 151–178 (1997).
2. Domingo, E., Webster, R. & Holland J. (eds). *Origin and Evolution of Viruses* (Academic Press, London, 1999).
3. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
4. Elena, S. F., Miralles, R. & Moya, A. Frequency-dependent selection in a mammalian RNA virus. *Evolution* **5**, 984–987 (1997).
5. Drake, J. W. Rates of spontaneous mutation among RNA viruses. *Proc. Natl Acad. Sci. USA* **90**, 4171–4175 (1993).
- A key reference giving figures for the spontaneous rate of mutation in RNA viruses and the difficulties in making these measurements.**
6. Malpica, J. M. *et al.* The rate and character of spontaneous mutation in an RNA virus. *Genetics* **162**, 1505–1511 (2002).
7. Eigen, M. Self-organisation of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465 (1971).
8. Maynard Smith, J. & Szathmáry, E. *The Major Transitions of Evolution*. (W. H. Freeman & Co., Oxford, 1995).
9. Rouzine, I. M., Rodrigo, A. & Coffin, J. M. Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiol. Mol. Biol. Rev.* **65**, 151–185 (2001).
10. Osada, D., Crandall, K. A. & Holmes, E. C. Recombination in evolutionary genomics. *Ann. Rev. Genet.* **36**, 75–97 (2002).
11. Rhodes, T., Wargo, H. & Hu, W. S. High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication. *J. Virol.* **77**, 11193–11200 (2003).
12. Boerlijst, M. C., Bonhoeffer, S. & Nowak, M. A. Viral quasi-species and recombination. *Proc. R. Soc. Lond. B Biol. Sci.* **263**, 1577–1584 (1996).
13. Chare, E. R., Gould, E. A. & Holmes, E. C. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *J. Gen. Virol.* **84**, 2691–2703 (2003).
14. Navas, S., Martin, J., Quiroga, J. A., Castillo, I. & Carreño, V. Genetic diversity and tissue compartmentalization of the hepatitis C virus genome in blood mononuclear cells, liver, and serum from chronic hepatitis C patients. *J. Virol.* **72**, 1640–1646 (1998).
15. Miralles, R., Moya, A. & Elena, S. F. Effect of population patchiness and migration rates on the adaptation and divergence of vesicular stomatitis virus quasispecies populations. *J. Gen. Virol.* **80**, 2051–2059 (1999).
16. Eigen, M. & Schuster, P. The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften* **64**, 541–565 (1977).
17. Eigen, M. *Steps Towards Life* (Oxford Univ. Press, New York, 1996).
18. Domingo, E., Sabo, D., Taniguchi, T. & Weissman, C. Nucleotide sequence heterogeneity of an RNA phage population. *Cell* **13**, 735–744 (1978).
19. de la Torre, J. C. & Holland, J. J. RNA virus quasispecies populations can suppress vastly superior mutant progeny. *J. Virol.* **64**, 6278–6281 (1990).
- An important experimental study showing that a high-fitness variant was suppressed in a low-fitness population. Often interpreted as indicating that RNA viruses form quasispecies.**
20. Jenkins, G. M., Worobey, M., Woelk, C. H. & Holmes, E. C. Evidence for the nonquasispecies evolution of RNA viruses. *Mol. Biol. Evol.* **18**, 987–994 (2001).
21. Holmes, E. C. & Moya, A. Is the quasispecies concept relevant to RNA viruses? *J. Virol.* **76**, 460–462 (2002).
- A paper questioning the relevance of the quasispecies model to the evolution of RNA viruses in Nature. The accompanying paper by Domingo provides a counter argument.**
22. Burch, C. L. & Chao, L. Evolvability of an RNA virus is determined by its mutational neighborhood. *Nature* **406**, 625–628 (2000).
23. Wichman, H. A., Scott, L. A., Yarber, C. D. & Bull, J. J. Experimental evolution recapitulates natural evolution. *Philos. Trans R. Soc. Lond. B Biol. Sci.* **355**, 1677–1684 (2000).
24. Ebert, D. Experimental evolution of parasites. *Science* **282**, 1432–1435 (1998).
25. Elena, S. F. Restrictions to RNA virus adaptation: an experimental approach. *Antonie van Leeuwenhoek* **81**, 135–142 (2002).
26. Bull, J. J. *et al.* Exceptional convergent evolution in a virus. *Genetics* **147**, 1497–1507 (1997).
27. Cuevas, J. M., Elena, S. F. & Moya, A. A molecular basis of adaptive convergence in experimental populations of RNA viruses. *Genetics* **162**, 533–542 (2002).
28. Badgett, M. R., Auer, A., Carmichael, L. E., Parrish, C. R. & Bull, J. J. Evolutionary dynamics of viral attenuation. *J. Virol.* **76**, 10524–10529 (2002).
29. Rokyta, D., Badgett, M. R., Molineux, I. J. & Bull, J. J. Experimental genomic evolution: extensive compensation for loss of DNA ligase activity in a virus. *Mol. Biol. Evol.* **19**, 230–238 (2002).
30. Elena, S. F. & Moya, A. Rate of deleterious mutation and the distribution of its effects on fitness in vesicular stomatitis virus. *J. Evol. Biol.* **12**, 1078–1088 (1999).
31. Chao, L., Rang, C. U. & Wong, L. E. Distribution of spontaneous mutants and inferences about the replication mode of the RNA bacteriophage  $\phi 6$ . *J. Virol.* **76**, 3276–3281 (2002).
32. Wahl, L. M., Gerrish, P. J. & Saika-Voivod, I. Evaluating the impact of population bottlenecks in experimental evolution. *Genetics* **162**, 961–971 (2002).
33. Lazaro, E., Escarmis, C., Perez-Mercader, J., Manrubia, S. C. & Domingo, E. Resistance of virus to extinction on bottleneck passages: study of a decaying and fluctuating pattern of fitness loss. *Proc. Natl Acad. Sci. USA* **100**, 10830–10835 (2003).
34. Miralles, R., Gerrish, P. J., Moya, A. & Elena, S. F. Clonal interference and the evolution of RNA viruses. *Science* **285**, 1745–1747 (1999).
- An important experimental demonstration that clonal interference slows down the rate of adaptive evolution in RNA viruses.**
35. Novella, I. S. Contributions of vesicular stomatitis virus to the understanding of RNA virus evolution. *Curr. Opin. Microbiol.* **6**, 399–405 (2003).
36. Wahl, L. M. & Krakauer, D. C. Models of experimental evolution: the role of genetic chance and selective necessity. *Genetics* **156**, 1437–1448 (2000).
37. Chao, L. Fitness of RNA virus decreased by Muller's ratchet. *Nature* **348**, 454–455 (1990).
- A seminal paper on experimental evolution with RNA viruses demonstrating the action of Muller's ratchet.**
38. de la Peña, M., Elena, S. F. & Moya, A. Effect of deleterious mutation-accumulation on the fitness of RNA bacteriophage MS2. *Evolution* **54**, 686–691 (2000).
39. Fralle, A. *et al.* Genetic exchange by recombination or reassortment is infrequent in natural populations of a tripartite RNA plant virus. *J. Virol.* **71**, 934–940 (1997).
40. Chao, L., Tran, T. T. & Tran, T. T. The advantage of sex in the RNA virus  $\phi 6$ . *Genetics* **147**, 953–959 (1997).
41. Elena, S. F. *et al.* Evolutionary dynamics of fitness recovery from the debilitating effects of Muller's ratchet. *Evolution* **52**, 309–314 (1998).
42. Chao, L. Evolution of sex and the molecular clock in RNA viruses. *Gene* **205**, 301–308 (1997).
43. Clarke, D. K. *et al.* The red queen reigns in the kingdom of RNA viruses. *Proc. Natl Acad. Sci. USA* **91**, 4821–4824 (1994).
44. Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. Different trajectories of parallel evolution during viral adaptation. *Science* **285**, 422–424 (1999).
- A key experimental paper that provides evidence for both frequent adaptive and parallel evolution in RNA viruses.**
45. Gerrish, P. J. The rhythm of microbial adaptation. *Nature* **413**, 299–302 (2001).
46. Turner, P. E. & Chao, L. Prisoner's dilemma in an RNA virus. *Nature* **398**, 441–443 (1999).
47. Turner, P. E. & Chao, L. Escape from prisoner's dilemma in RNA phage  $\phi$ . *Am. Nat.* **161**, 497–505 (2003).
48. Wilke, C. O. & Adami, C. The biology of digital organisms. *Trends Ecol. Evol.* **17**, 528–532 (2002).
49. Yedid, G. & Bell, G. Microevolution in an electronic microcosm. *Am. Nat.* **157**, 465–487 (2001).
50. Yedid, G. & Bell, G. Macroevolution simulated with autonomously replicating computer programs. *Nature* **420**, 810–812 (2002).
51. Lenski, R. E., Ofria, C., Collier, T. C. & Adami, C. Genome complexity, robustness and genetic interactions in digital organisms. *Nature* **400**, 661–664 (1999).
52. Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E. & Adami, C. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* **412**, 331–333 (2001).
53. Kilbourne, E. D. The molecular epidemiology of influenza. *J. Infect. Dis.* **127**, 478–487 (1973).
54. Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A. & Brown, A. J. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc. Natl Acad. Sci. USA* **89**, 4835–4839 (1992).
55. Scott, T. W., Weaver, S. C. & Mallampalli, V. L. In *The Evolutionary Biology of Viruses* (ed. Morse, S. S.) 293–324 (Raven Press, New York, 1994).
56. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380–390 (2002).
57. Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. Measurably evolving populations. *Trends Ecol. Evol.* **18**, 481–488 (2003).
- Theoretical and practical introduction to the analysis of serially sampled gene sequences, including those from RNA viruses.**
58. Shankarappa, R. *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10489–10502 (1999).
59. Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320 (2002).
60. Drummond, A. J., Forsberg, R. & Rodrigo, A. G. The inference of stepwise changes in substitution rates using serial sequence samples. *Mol. Biol. Evol.* **18**, 1365–1371 (2001).
61. Seo, T. K., Thorne, J. L., Hasegawa, M. & Kishino, H. A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics* **18**, 115–123 (2002).
62. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Prob.* **19**, 27–43 (1982).
63. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
64. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* **23**, 183–201 (1983).
65. Marra, M. A. *et al.* The genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399–1404 (2003).
66. Robbins, K. E. *et al.* US human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J. Virol.* **77**, 6359–6366 (2003).
67. Lemey, P. *et al.* Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl Acad. Sci. USA* **100**, 6588–6592 (2003).

68. Ou, C. Y. *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**, 1165–1171 (1992).
69. Albert, J., Wahlberg, J., Leitner, T., Escanilla, D. & Uhlen, M. Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 *pol* and *gag* genes. *J. Virol.* **68**, 5918–5924 (1994).
70. Power, J. P. *et al.* Molecular epidemiology of an outbreak of infection with hepatitis C virus in recipients of anti-D immunoglobulin. *Lancet* **345**, 1211–1213 (1995).
71. Esteban, J. I. *et al.* Transmission of hepatitis C virus by a cardiac surgeon. *N. Engl. J. Med.* **334**, 555–560 (1996).
72. Ross, H. A. & Rodrigo, A. G. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J. Virol.* **76**, 11715–11720 (2002).
73. Nielsen, R. & Yang, Z. Likelihood model for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936 (1998).  
**Develops a rigorous, statistical methodology to determine selection pressures at individual amino acid sites.**
74. Holmes, E. C. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *J. Virol.* **77**, 11296–11298 (2003).
75. Costa-Mattoli, M. *et al.* Evidence of recombination in natural populations of hepatitis A virus. *Virology* **311**, 51–59 (2003).
76. Kalinina, O., Norder, H., Mukomolov, S. & Magnus, L. O. A natural intergenotypic recombinant of hepatitis C virus identified in St Petersburg. *J. Virol.* **76**, 4034–4043 (2002).
77. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796 (2000).
78. Pavesi, A. Origin and evolution of GBV-C/hepatitis G virus and relationships with ancient human migrations. *J. Mol. Evol.* **53**, 104–113 (2001).
79. Pybus, O. G., Rambaut, A. & Harvey, P. H. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437 (2000).
80. Pybus, O. G. *et al.* The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323–2325 (2001).  
**Establishes the link between population genetics and mathematical epidemiology through the use of molecular phylogenies to study the population dynamics of infectious diseases.**
81. Tanaka, Y. *et al.* A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. *Proc. Natl Acad. Sci. USA* **99**, 15584–15589 (2002).
82. Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* **54**, 156–165 (2002).
83. Salemi, M. *et al.* Different population dynamics of human T-cell lymphotropic virus type 1 in intravenous drug users compared with endemically infected tribes. *Proc. Natl Acad. Sci. USA* **96**, 13253–13258 (1999).
84. Buonagurio, D. A. *et al.* Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. *Science* **232**, 930–932 (1986).  
**A pioneering paper describing the molecular epidemiological analysis of RNA viruses.**
85. Gillespie, J. H. Episodic evolution of RNA viruses. *Proc. Natl Acad. Sci. USA* **90**, 10411–11041 (1993).
86. Nichol, S. T., Rowe, J. E. & Fitch, W. M. Punctuated equilibrium and positive Darwinian evolution in vesicular stomatitis virus. *Proc. Natl Acad. Sci. USA* **90**, 10424–11042 (1993).
87. Li, W. H., Tanimura, M. & Sharp, P. M. Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.* **5**, 313–330 (1988).
88. Overbaugh, J. & Bangham, C. R. M. Selection forces and constraints on retroviral sequence variation. *Science* **292**, 1106–1109 (2001).  
**Illustrates some of the key factors that constrain the adaptability of RNA viruses.**
89. Holmes, E. C. Molecular clocks and the puzzle of RNA virus origins. *J. Virol.* **77**, 3893–3897 (2003).
90. Plyusnin, A. & Morzunov, S. P. Virus evolution and genetic diversity of hantaviruses and their rodent hosts. *Curr. Top. Microbiol. Immunol.* **256**, 47–75 (2001).
91. McGeoch, D. J., Dolan, A. & Ralph, A. C. Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J. Virol.* **74**, 10401–10406 (2000).
92. Daszak, P., Cunningham, A. A. & Hyatt, A. D. Emerging infectious diseases of wildlife: threats to biodiversity and human health. *Science* **287**, 443–449 (2000).
93. Morse, S. S. Factors in the emergence of infectious diseases. *Emerg. Infect. Dis.* **1**, 7–15 (1995).
94. Cleaveland, S., Laurenson, M. K. & Taylor, L. H. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 991–999 (2001).
95. Woolhouse, M. E. J., Taylor, L. H. & Haydon, D. T. Population biology of multihost pathogens. *Science* **292**, 1109–1112 (2001).
96. Holmes, E. C. Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* **11**, 543–546 (2003).
97. Woelk, C. H. & Holmes, E. C. Reduced positive selection in vector-borne RNA viruses. *Mol. Biol. Evol.* **19**, 2333–2336 (2002).
98. DeFilippis, V. R. & Villarreal, L. P. In *Viral Ecology* (ed. Hurst, C. J.) 126–208 (Academic Press, New York, 2000).
99. Baranowski, E., Ruiz-Jarabo, C. M. & Domingo, E. Evolution of cell recognition by viruses. *Science* **292**, 1102–1105 (2001).
100. Schollissek, C., Ludwig, S. & Fitch, W. M. Analysis of influenza A virus nucleoproteins for the assessment of molecular genetic mechanisms leading to new phylogenetic virus lineages. *Arch. Virol.* **131**, 237–250 (1993).
101. Peeters, M. *et al.* Risk to human health from a plethora of simian immunodeficiency viruses in primate bushmeat. *Emerg. Infect. Dis.* **8**, 451–457 (2002).  
**A remarkable demonstration of the frequency with which primate bushmeat is infected with SIVs.**
102. Gao, F. *et al.* Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441 (1999).
103. Hahn, B. H., Shaw, G. M., de Cock, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–614 (2000).
104. Charleston, M. A. & Robertson, D. L. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst. Biol.* **51**, 528–535 (2002).
105. Rota, P. A. *et al.* Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**, 1394–1399 (2003).
106. Eickmann, M. *et al.* Phylogeny of the SARS coronavirus. *Science* **302**, 1504–1505 (2003).
107. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
108. Stavrinides, J. & Guttman, D. S. Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J. Virol.* **78**, 76–82 (2004).
109. Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. Predicting the evolution of human influenza A. *Science* **286**, 1921–1925 (1999).
110. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans* (Oxford Univ. Press, Oxford, 1991).
111. Culley, A. I., Lang, A. S. & Suttle, C. A. High diversity of unknown picorna-like viruses in the sea. *Nature* **424**, 1054–1057 (2003).  
**A PCR-based analysis of viral diversity in marine environments, which revealed a remarkable number of previously unknown viruses.**
112. Levin, S. A. & Andreasen, V. Disease transmission dynamics and the evolution of antibiotic resistance in hospitals and communal settings. *Proc. Natl Acad. Sci. USA* **96**, 800–801 (1999).
113. Levin, B. R., Lipsitch, M., & Bonhoeffer, S. Population biology, evolution, and infectious disease: convergence and synthesis. *Science* **283**, 806–809 (1999).
114. Wain-Hobson, S. Down or out in blood and lymph? *Nature* **387**, 123–124 (1997).
115. Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
116. Jung, A. *et al.* Recombination: multiply infected spleen cells in HIV patients. *Nature* **418**, 144 (2002).
117. Bonhoeffer, S., May, R. M., Shaw, G. M. & Nowak, M. A. Virus dynamics and drug therapy. *Proc. Natl Acad. Sci. USA* **94**, 6971–6976 (1997).
118. Back, N. K. *et al.* Reduced replication of 3TC-resistant HIV-1 variants in primary cells due to a processivity defect of the reverse transcriptase enzyme. *EMBO J.* **15**, 4040–4049 (1996).
119. Simon, V. *et al.* Infectivity and replication capacity of drug-resistant human immunodeficiency virus type 1 variants isolated during primary infection. *J. Virol.* **77**, 7736–7745 (2003).
120. Pitisuttithum, P. *et al.* Safety and immunogenicity of combinations of recombinant subtype E and B human immunodeficiency virus type 1 envelope glycoprotein 120 vaccines in healthy Thai adults. *J. Infect. Dis.* **188**, 219–227 (2003).
121. Fitch, W. M., Leiter, J. M. E., Li, X. Q. & Palese, P. Positive Darwinian evolution in human influenza-A viruses. *Proc. Natl Acad. Sci. USA* **88**, 4270–4274 (1991).
122. Yusim, K. *et al.* Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *J. Virol.* **76**, 8757–8768 (2002).
123. Moore, C. B. *et al.* Evidence for HIV-1 adaptation to HLA-restricted immune responses at the population level. *Science* **296**, 1439–1443 (2002).
124. Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422**, 307–312 (2003).  
**Recent paper showing that the intra-host evolution of HIV-1 is dominated by immune-driven positive selection.**
125. Grassly, N. C., Harvey, P. H. & Holmes, E. C. Population dynamics of HIV-1 inferred from gene sequences. *Genetics* **151**, 427–438 (1999).
126. Delwart, E. *et al.* Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection. *AIDS* **16**, 189–195 (2002).
127. Lipsitch, M. & Moxon, E. R. Virulence and transmissibility of pathogens: what is the relationship? *Trends Microbiol.* **5**, 31–37 (1997).
128. Ebert, D. & Bull, J. J. Challenging the trade-off model for the evolution of virulence: is virulence management feasible? *Trends Microbiol.* **11**, 15–20 (2003).
129. Bergstrom, C. T., McElhany, P. & Real, L. A. Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proc. Natl Acad. Sci. USA* **96**, 5095–5100 (1999).
130. Nowak, M. A. & May, R. M. Superinfection and the evolution of parasite virulence. *Proc. R. Soc. Lond. B Biol. Sci.* **255**, 81–89 (1994).
131. Zhang, L. *et al.* Chemokine coreceptor usage by diverse primary isolates of human immunodeficiency virus type 1. *J. Virol.* **72**, 9307–9312 (1998).
132. Zhu, T. *et al.* Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* **261**, 1179–1181 (1993).
133. Metzker, M. L. *et al.* Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl Acad. Sci. USA* **99**, 14292–14297 (2002).
134. González-Candelas, F., Bracho, M. A. & Moya, A. Molecular epidemiology and forensic genetics: application to a hepatitis C virus transmission event at a hemodialysis unit. *J. Infect. Dis.* **187**, 352–358 (2003).
135. Strimmer, K. & von Haeseler, A. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969 (1996).
136. Swofford, D. L. *PAUP\*. Phylogenetic Analysis Using Parsimony (and other methods). Version 4.* (Sinauer Associates, Sunderland, Massachusetts, 2002).

**Acknowledgements**  
This work has been funded by The Wellcome Trust, the Spanish Ministerio de Ciencia y Tecnología and Generalitat Valenciana. We would like to thank J. M. Cuevas and R. Sanjuan for help in preparing Box 1. We also thank four referees for useful comments.

**Competing interests statement**  
The authors declare that they have no competing financial interests.

 **Online links**

**FURTHER INFORMATION**  
**Unidad de Investigación de Genética Evolutiva:**  
<http://www.uv.es/cavanilles/general/>  
**Evolutionary Biology Group:**  
<http://evolve.200.ox.ac.uk>  
Access to this interactive links box is free online.