Genetic diversity in pathogen species contains information about evolutionary and epidemiological processes, including the origins and history of disease, the nature of the selective forces acting on pathogen genes and the role of recombination in generating genetic novelty. Here, we review recent developments in these fields and compare the use of population genetic, or population-model based, approaches to phylogenetic, or population-model free, methodologies. We show how simple epidemiological models can be related to the ancestral, or coalescent, process underlying samples from pathogen species, enabling detailed inference about pathogen biology from patterns of molecular variation.

A pathogen, like any other organism, has an evolutionary history that is reflected in the distribution of genetic diversity within the species. What makes a pathogen special is that this evolutionary history is dominated by the successful and ongoing colonization of a host. Therefore, analyses of pathogen genomes can not only tell us things about the history of disease (when did the epidemic begin?), but can also inform efforts to understand (which genetic changes made the ancestral organism pathogenic?) and control the disease (which is the best target for a vaccine, will vaccines be effective in different populations?).

The statistical and analytical tools available for comparing molecular sequences (DNA, RNA or protein) from representative pathogen isolates are becoming increasingly sophisticated. Our first aim here is to summarize recent research where molecular sequences alone have been used to understand pathogen biology. We focus on the reconstruction of the origin and history of a pathogen, the nature of immune-mediated selection acting on pathogen genomes, and the role of recombination (see Glossary) in generating genetic novelty. Our second aim is to discuss the different methodologies that can be applied to molecular sequence data. In particular, we contrast the use of phylogenetic methods with that of population genetic methods. Phylogenetic methods were originally developed for the analysis of sequences from different species, and make no assumptions about how population-level processes, such as genetic drift, natural selection, changes in population size or geographical structure, influence the shape of underlying gene trees. Population genetic approaches can be used to understand such factors by explicitly modelling their effects on gene tree shape. We discuss the gains and potential pitfalls of using population genetics models and outline how they can be integrated into existing epidemiological models.

The origin and history of pathogens
Tracing the origins and history of pathogen species informs our understanding of what causes new epidemics and how they spread. Phylogenies constructed from samples of contemporary pathogen diversity reconstruct the history of those ancestors that have left descendants, the depth and shape of which can tell us about the size and structure of historical populations. For example, explosive growth generates characteristic ‘star-like’ phylogenies as seen in the HIV viruses and subtypes [1–3]. Historical changes in the pathogen population size might also be detected; for example, the major increase in population size of the hepatitis C virus during the first half of the 20th century [4]. Dating events in phylogenies constructed from contemporary genetic diversity requires an independent...
immunity, and also at the population level, through the dynamics of herd immunity and cross immunity. How such factors influence patterns of genetic variation within pathogen populations depends on the relative timescales of host and pathogen adaptation. In species such as HIV-1, where rates of adaptation in the pathogen are high [11], immune-escape mutants will arise and be selected for within hosts. The effect of such selection is to distort transiently patterns of pathogen genetic variation within the host through the hitch-hiking effect (Box 2), a pattern detected in longitudinal samples from HIV-1-infected patients [12]. However, immune-escape mutations do not generally provide an advantage to viruses infecting other hosts, who are unlikely to have encountered a virus with the same antigen type. Instead, diversifying selection within infected individuals results in pathogen species characterized by diverse and rapidly changing antigenic variation, the hallmark of which is an excess of protein-changing variation (relative to putatively neutral, non-protein changing variation) at antigenic genes during the course of the infection. Such a pattern is seen in HIV-1, particularly in the env gene, by the use of codon-based phylogenetic methods [13,14].

Another route to detecting diversifying selection comes from comparison of within-species variation with between-species divergence. Because immune-escape mutants are unlikely to ever become fixed within a species, high levels of protein-changing variation at antigenic genes do not necessarily translate into high rates of change between species. For example, in a study of the gene encoding the erythrocyte-binding antigen EBA-175 in Plasmodium falciparum and the corresponding gene in P. reichenowi, there is an excess of within-species variation relative to between-species divergence. The effect is not seen in the related gene eba-140, which suggests that eba-175 is under within-host diversifying selection, probably as a result of interaction with the human immune system [15].

When cross-immunity is strong and rates of pathogen adaptation are slower, the pathogen population can theoretically become structured into different antigenic types [16–18]. Such types are maintained, or ‘balanced’, over time by frequency-dependent selection. Structuring can be detected by comparing patterns of genetic variation to those expected under simple mathematical models of genetic variation, such as the neutral coalescent theory. In particular, balancing selection can result in genes with elevated levels of genetic diversity, changes in the distribution of allele frequencies and can also inhibit drift by maintaining genetic variation within multiple populations in spite of geographical isolation. Such patterns are observed at ama1 [19], a gene of P. falciparum that encodes an antigen that represents a potential vaccine target.

Genome-wide structuring of genetic variation (in the sense that the population is clustered into groups of closely related individuals) is found in many pathogen
Box 2. Shifting timescales in adaptive evolution

Adaptive evolution results from the fixation of beneficial mutations in populations, which will typically arise on a single genetic background. As the beneficial mutation increases in frequency, it drags along with it the genetic background on which it occurred, a phenomenon known as hitch-hiking [72]. This generates patterns of genetic variation that differ in characteristic ways from those expected in the absence of selection [62,73]. These patterns are erased by mutation and genetic drift after fixation of the beneficial mutation [74]. However, if further beneficial mutations appear and become fixed within the gene, adaptive evolution can be detected over longer periods of time through an elevated rate of change (typically in the protein sequence) relative to neutral sites (typically silent codon positions) [50,53]. The timescale over which evolution can be observed (either directly with sequences sampled longitudinally over time, or indirectly through the depth of the genealogy; Figure I) is therefore crucial in determining which statistical tools a researcher interested in detecting adaptive evolution should use.

Both short-term hitch-hiking and long-term recurrent adaptation are observed within patients infected by HIV-1. Longitudinal sequence data show the repeated fixation of adaptive mutations in the \textit{env} gene [75], generating strong, but intermittent, signals of hitch-hiking [12]. Over time, certain codon positions accumulate multiple changes [50]. Importantly, differences between patients in how such changes accumulate in different parts of the gene can be associated with the rate of disease progression [75,76].

The relevance of recombination

The tools available for inferring evolutionary history depend considerably on the biology of the pathogen. If recombination is rare, or if hosts are only ever infected by a single pathogen strain, reconstruction of a single phylogeny is the natural starting point for any analysis. By contrast, if recombination between different strains is common, different parts of the genome will have different phylogenetic histories, thus limiting the use of phylogenetic methods. In recombinating species, instead of reconstructing a phylogenetic tree when a single tree might not exist, data sets can be described by summaries of the data, such as the frequency distribution of polymorphisms, levels of linkage disequilibrium and measures of differentiation between populations (these summaries are also applicable to non-recombining species). Such summaries are the starting point for making inferences about the evolutionary history of the pathogen species, so knowing whether a species is recombing is crucial in the choice of appropriate analyses.

Recombination also has major implications in studies that attempt to map phenotypically important genes by association, or through the hitch-hiking effect of adaptive mutations [23], because the rate of recombination determines the density of markers required to detect causative mutations reliably. Furthermore, estimates of important quantities, such as mutation rates or the age of the most recent common ancestor (MRCA) of a species are strongly biased if data from a recombing species are treated as having come from a clonal species [24–26].

The simplest way of detecting recombination from gene sequences is the identification of mosaic sequences. For example, in an alignment of sequences from avian influenza A, a highly pathogenic strain was shown to have a 30-nucleotide insert in the gene encoding hemagglutinin relative to the low pathogenic strains, which is 100% identical to part of the gene encoding neuraminidase [27]. More-sophisticated approaches to detecting mosaic structures have recently been developed; for example, scanning methods that detect recombinant forms such as those observed in HIV-1 among characterized subtypes [28], and methods for weakly linked markers that detect admixture between subpopulations as found in \textit{Helicobacter pylori} [8,9].

Mosaic identification assumes that all recombination events are recent, and that genomes can be separated into ‘pure’ and ‘mosaic’. In unstructured (panmictic) recombing species, such a distinction is invalid, in which case an alternative is to try to identify the positions along the molecular sequence at which the phylogenetic tree changes. Many methodologies for detecting shifts in phylogeny have been developed, with recent work focusing on methods that aim to accommodate uncertainty about the tree reconstructions [29,30]. These methods work well at
detecting a low number of recombination breakpoints along a sequence; for example, in an alignment of the entire 3.2-kb genome of four strains of hepatitis B, two changes in topology were detected [30]. Yet for many pathogens, the rate of recombination is sufficient that changes in phylogeny are expected to occur every few base pairs [31,32].

For most species, the rate of recombination (relative to mutation) is high enough that there is little information about the underlying tree at any given position in the genome and, therefore, little chance of detecting recombination breakpoints exactly. Under such circumstances, the impact of recombination can be summarized either by a nonparametric estimate of the minimum number of recombination events in the history of the gene samples [33] (although this is only possible if recurrent or back mutations are assumed not to occur), or by a model-based estimate of the rate of recombination relative to genetic drift [34]. Coalescent methods can estimate recombination rates under models with recurrent and back mutations [35,36], and have demonstrated high levels of recombination in various pathogens, including HIV-1 [35] and P. falciparum [15]. Because genetic exchange can only occur between pathogen genomes in the same host, coalescent approaches measure the effective recombination rate, which can provide an indication of the rate of multiple infection [37]. Genomes with high intrinsic recombination rates, such as P. falciparum [38] and HIV-1 [39,40], can therefore exhibit either high or low levels of historical recombination depending on the wider pathogen epidemiology [35,41].

Recombination has important biological, as well as methodological, consequences. Recombination (both homologous and nonhomologous) is an important source of genetic novelty, particularly at antigenic loci such as the hemagglutinin- and neuraminidase-encoding genes of influenza [42,43] (where the origin of novel strains by recombination is known as antigenic shift) or the tbpB and opa genes of the meningococcus, Neisseria meningitidis [44,45].

Phylogenetic and population genetic approaches to inference

Diverse biological questions in diverse pathogen species require a variety of approaches to analyzing molecular sequence data. However, we can define a broad distinction between those approaches that derive from the phylogenetic background and those that are rooted in population genetics modelling. The key distinction is that phylogenetic models make no assumptions about how population-level processes (e.g. genetic drift, natural selection, inbreeding or restricted gene flow) influence the shape of genealogies (or gene trees) underlying samples of genetic material from within populations, whereas population genetic approaches model such factors explicitly.

Phylogenetic approaches were first developed for the analysis of molecular sequences sampled from different species and have become widespread in the analysis of pathogen species diversity [1,3,11,46–51]. In addition to estimating phylogenetic trees, such approaches can be used to date epidemics [1,3,52], detect recombination events [30] and identify sites of diversifying selection [50,53]. However, because phylogenetic approaches were originally designed to analyze sequences from different species, they naturally assume that the shape of the tree itself is not informative about the quantities of interest. Post-hoc interpretation of tree shape has, however, been important in the analysis of pathogen diversity; for example, the observation of ladder-like trees for influenza has shaped theories of antigenic drift and shift [46,54,55].

Population genetic methods, by contrast, are based on mathematical models of populations; initially the ‘beanbag’ genetics of Fisher, Wright and Haldane and, more recently, the coalescent theory of Kingman [56] and Hudson [57]. Coalescent models describe in a probabilistic manner how population-level processes influence the shape of genealogies underlying samples of gene sequences from within a population, and the resulting patterns of genetic variation. The standard neutral model (which underlies coalescent theory) assumes selective neutrality, constant population size and random mating, but can be extended to consider complexities such as population growth, inbreeding, geographical subdivision and different forms of natural selection (reviewed in [58]).

The difference between phylogenetic and population genetic approaches leads to conceptual differences in how data are analyzed. Where phylogenetic approaches make statements about the tree and the substitutions mapped on to it, population genetic approaches use the same genealogy to make statements about parameters of the coalescent model. For example, phylogenetic methods summarize variability among sequences by the branch lengths of the estimated tree, whereas population genetic methods estimate the ‘population mutation rate’ parameter, which is the product of the per generation mutation rate and the effective population size of a species, $N_e$. Likewise, phylogenetic methods detect adaptive evolution by the relative rate of protein-changing and silent substitutions on the tree, whereas population genetic methods estimate the selection coefficient of individual mutations from their effect on the shape of the genealogy [59].

Advantages and disadvantages of population genetics

The benefit of fitting an explicit population genetic model is that it gives extra power to detect phenomena of interest and to test specific hypotheses. For example, phylogenetic methods cannot test for population growth, because only in a model-based context do the star-like genealogies that population growth generates differ from those expected without growth (without a model, all genealogy shapes are equally probable). Similarly, phylogenetic methods cannot detect single adaptive substitutions because distortion to allele frequencies caused by the hitch-hiking effect is only quantifiable by comparison to the standard neutral model (again, without a model, all allele frequency distributions are equally probable). More generally, comparison of data to the expectations of the standard neutral model is a route to learning about which biological processes have been important in shaping genetic diversity. Many statistical methods for testing the (null) standard neutral model are available. These are either goodness of fit tests that aim to reject the null model (e.g. Tajima’s D test [60], Fu and Li’s $D^*$ test [61], Fay and Wu’s $H$ test [62], the MK test [63] and the HKA test [64]; discussed in [65,66]), or
likelihood-based approaches that compare models with and without parameters of interest [59,67].

The problem of fitting a population-model to the data is that the biological simplifications required to make the model tractable might also render it meaningless. The coalescent process derives from a simplification of reproduction in natural populations. For pathogens, where successful reproduction requires both replication within, and transmission between, hosts, population genetics must either incorporate epidemiological parameters explicitly in models of ancestry, or demonstrate that ignoring epidemiology still provides useful and meaningful inferences.

Both tasks are very much in their infancy. There is hope that the dynamics of simple epidemiological models, such as the susceptible–infectious–susceptible (SIS) model, might give rise to genealogical models that are identical to those in well characterized non-pathogen population genetic models, such as metapopulations (Box 3). However, where multiple strains with different epidemiological characteristics are considered, for example, the epidemic-clone model for bacterial populations (in which competition and selection maintain clonal structure in spite of frequent recombination) [68], it seems probable that novel population genetic models will be required.

Conclusions

Integrating epidemiological models into phylogenetic and population genetic approaches to inference presents many challenges. Which aspects of data are informative about key epidemiological parameters? Can epidemiological dynamics be incorporated into coalescent theory? Is it possible to discriminate between alternative epidemiological models simply given molecular sequence information? Such theoretical challenges can only make the computational difficulties involved in likelihood-based inference worse, hence progress might only be possible through the use of approximate statistical methods [69], or simplifications of the coalescent process, such as those developed for structured populations [8] or recombination [70]. The exciting prospect remains of using the large-scale analyses of pathogen diversity that are currently under way for many species to provide insight into the evolutionary history and biological mechanisms of infectious disease.

Acknowledgements

We thank Rosalind Harding, Chris Spencer and the reviewers for comments on the article. D.J.W. is funded by the BBSRC and D.F. by The Wellcome Trust.

References

10 Carbone, I. et al. (2004) Recombination and migration of Cryptonectria hypovirus 1 as inferred from gene genealogies and the coalescent. Genetics 166, 1611–1629


www.sciencedirect.com

**Forthcoming Conferences**

Are you organizing a conference, workshop or meeting that would be of interest to *TREE* readers? If so, please e-mail the details to us at TREE@elsevier.com and we will feature it in our Forthcoming Conference filler.

**30 March–1 April 2005**
Annual BES Symposium, Ecological Consequences of Extreme Events, Manchester, UK.
http://www.britishecologicalsociety.org/articles/meetings/current/annualsymposium2005/

**8–13 May 2005**
8th International Workshop on Seeds, Brisbane, Australia

**9–10 May 2005**
Royal Society Discussion Meeting: Sexual Conflict: A New Paradigm? London, UK
http://www.royalsoc.ac.uk/events/

**10–14 June 2005**
Evolution 2005 (jointly sponsored conference of the Society for the Study of Evolution, the Society of Systematic Biologists, and the American Society of Naturalists), University of Alaska Fairbanks, AL, USA
http://www.evolution05.uaf.edu/

**19–24 June 2005**
Gordon Research Conference: Structural, Functional and Evolutionary Genomics, Lewiston, ME, USA
http://www.grc.uri.edu/05sced.htm#GRC

**1–5 July 2005**
International Society for Molecular Biology and Evolution Annual Meeting, Auckland, New Zealand
http://www.mbe05.com/

**12–13 July 2005**
BES/IEMM Conference on Ecological Impact Assessments: Science and Best Practice, Bath, UK
http://www.britishecologicalsociety.org/articles/groups/conservation/bes_iemm_conf/

**15–19 July 2005**
The 19th Annual Meeting of the Society for Conservation Biology, Brasilia, Brazil.
http://www.unb.br/ib/zoo/scb2005/

**19–21 July 2005**
ASAB Summer Meeting, Behavioural Interactions: Visions for the Future, University of Lancaster, UK
http://www.societies.ncl.ac.uk/asab/meetings.html

**31 July–5 August 2005**
Gordon Research Conference: Evolutionary and Ecological Functional Genomics, Oxford, UK
http://www.grc.uri.edu/05sced.htm#GRC

**6–10 August 2005**
The Animal Behavior Society 2005 Annual Meeting, Snowbird, UT, USA
http://www.animalbehavior.org/ABS/Program/

**7–12 August 2005**
90th Annual Meeting of the Ecological Society of America, Montréal, Canada
http://www.esa.org/montreal/

**15–20 August 2005**
10th Congress of the European Society for Evolutionary Biology, Krakow, Poland

**2–4 September 2005**
Special BES Symposium, Body size and the organisation and function of aquatic ecosystems, University of Hertfordshire, UK
http://www.britishecologicalsociety.org/articles/meetings/current/specialsymposium2005/

**21–23 September 2005**
9th Evolutionary Biology Meeting at Marseilles, Marseilles, France
http://www.up.univ-mrs.fr/evol/congres/

**25–27 September 2005**
http://www.royalsoc.ac.uk/events/

**17–18 October 2005**
Royal Society Discussion Meeting: Evolution of the Antarctic Ice Sheet: New Understandings and Challenges, London, UK
http://www.royalsoc.ac.uk/events/

**19–22 October 2005**
The Society of Vertebrate Paleontology, Mesa, Arizona, USA
http://www.vertpaleo.org/meetings/futuremeetings.html