THE EVOLUTIONARY GENOMICS OF PATHOGEN RECOMBINATION

Philip Awadalla

A pressing problem in studying the evolution of microbial pathogens is to determine the extent to which these genomes recombine. This information is essential for locating pathogenicity loci by using association studies or population genetic approaches. Recombination also complicates the use of phylogenetic approaches to estimate evolutionary parameters such as selection pressures. Reliable methods that detect and estimate the rate of recombination are, therefore, vital. This article reviews the approaches that are available for detecting and estimating recombination in microbial pathogens and how they can be used to understand pathogen evolution and to identify medically relevant loci.

LATERAL GENE TRANSFER
The transfer of DNA, frequently cassettes of genes, between organisms.

EFFECTIVE POPULATION SIZE
(N_e). The theoretical number of organisms or copies of a locus for which the genetic variation in a given sample of the organisms or copies can be explained solely by mutation and genetic drift. N_e is related to, but never exceeds, the actual population size (N).

GENETIC DRIFT
Random changes in allele frequency that result from the sampling of gametes from generation to generation.

Section of Evolution and Ecology, University of California at Davis, California 95616, USA.
e-mail: pawadalla@ucdavis.edu

doi:10.1038/nrg964

Almost all organisms engage in some form of genetic mixing, even if only occasionally. One of the more striking findings from inter-specific analyses of whole-genome sequences is the extent to which many bacterial genomes are chimaeras[1], even though these taxa are predominantly asexual. Comparative analyses indicate that, in some cases, the presence of foreign genomic material in bacteria can be attributed to ancient LATERAL TRANSFER events from other bacterial species[1]. However, intraspecific genomic analyses have also shown that recombination events also happen frequently in many bacterial, as well as viral and parasitic, protozoan species[2–7]. Not only is recombination ubiquitous in these three taxa, but also the underlying mechanisms of recombination are extremely varied across species (BOX 1; FIG. 1).

The extent to which recombination occurs in natural populations is either unknown or controversial. Population geneticists have argued that it is important to know the presence and frequency of recombination for three reasons. First, the basic population parameters of a species, including the EFFECTIVE POPULATION SIZE (N_e), and the mutation, recombination and migration rates, might help us to predict the extent to which genes are exchanged among genomes in the same population and between geographically separated populations. For pathogens in particular, this information helps to explain the dynamics of drug resistance and pathogenicity, and indicates which epidemiological processes should be targeted for disease control (for example, see REF. 8). Second, determining the extent or rate of genetic rearrangement that occurs through recombination in natural populations is crucial if we are to use genome and genetic mapping information to locate the genes that underlie important phenotypes (for example, genes that are associated with virulence, transmission and immune evasion). In medical genetics, associations between virulence and genetic markers that have built up through GENETIC DRIFT and that are broken down by recombination are central for mapping mutations that aid immune evasion[9]. Finally, although nearly all organisms engage in some form of recombination, our understanding of why recombination occurs and is maintained remains controversial (for example, see REFS 10–15).

This article outlines the importance of detecting and estimating recombination rates. I describe various phylogenetic and population genetic methods to detect and estimate population rates of recombination and show how these approaches have been applied to various pathogenic unicellular taxa. The effect that recombination has on inferences of key epidemiological and population genetic parameters are described. I also report the estimated population rate of recombination for several of these species and finally discuss the relevance of recombination to pathogenicity, and to identifying targets for disease control.

Box 1 | The mechanisms and frequency of recombination in pathogens

Bacteria

In bacteria, recombination occurs through the asymmetrical transfer of genetic material from a donor to a recipient cell by TRANSDUCTION, TRANSFORMATION or CONJUGATION. The molecular machinery that is associated with recombination in many eukaryotes might have evolved from mechanisms that are associated with DNA repair in prokaryotes, such as the Rec proteins in *Escherichia coli* (for example, see REFS 89–91; for a review on recombination-related repair, see REF. 92). Recombination modifiers that increase recombination rates might derive a fitness advantage from the immediate benefit of ensuring cell survival in the face of DNA damage[89].

Viruses

Many DNA and RNA viral genomes recombine by standard means — they generate a new nucleotide strand from two parental strands. However, in some RNA viruses, such as the human influenza virus and rotaviruses, standard recombination does not occur[93], but rather exchange between parental genomes is mediated by the breakdown of their genomes into smaller fragments. If different genotypes invade the same host, the shuffling of these fragments will produce recombinant genotypes[94,95]. Other viruses, such as the human immunodeficiency virus type 1 (HIV1), recombine through a TEMPLATE-SWITCHING PROCESS during reverse transcription; hot spots of recombination have been identified in the regions that control this reverse transcription[96]. These modes of viral recombination are not associated with nucleotide repair; proofreading is absent in these genomes, and they consequently have high rates of mutation[29,30,97].

Eukaryotes

In most eukaryotic species, including *Plasmodium falciparum* and yeast, symmetrical recombination is prevalent, even though asexual reproduction (by binary fission) is also frequent. During symmetrical transfer, two gametes fuse (in a process called syngamy) and divide (by meiosis) to produce new genome combinations. Crossing over or chiasmata are necessary for proper chromosomal segregation during meiosis and, as a result, there is usually at least one chiasma per chromosome arm[98]. However, achiasmate meiosis is known in a few systems, such as in *Drosophila melanogaster* males. In *Plasmodium*, recombination seems to occur uniformly over the genome and at a frequency that allows many double crossovers to occur[99]. In these species, the recombination rates per base are high and are conditional on meiosis. For example, the recombination rates that are inferred from genetic map data for *P. falciparum*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* are much higher (up to a hundred-fold) than those of *D. melanogaster* and humans (FIG. 1). However, it is clear that the need for chiasma formation during chromosomal segregation does not account for all of the variation in recombination rates.

Recombination and evolutionary inference

Recombination is one of the main forces that underlie local patterns of genome diversity in a species. By breaking up linkage between loci, recombination allows different regions to have different evolutionary histories. As a result, the action of selection and mutation on a given genomic region will be independent of similar forces acting at other regions. For example, if positive selection is acting at a locus, alleles that contribute to relatively greater fitness will eventually be FIXED. Linked neutral variants at other loci will also reach a high frequency in the population (in a process known as hitchhiking), and heterozygosity at these nearby neutral loci will consequently be reduced. Recombination releases allelic variation at neutral loci from the action of selection at nearby sites. Therefore, low levels of neutral variation are footprints of past selection or hitchhiking events[16]. Finally, because recombination creates independence among segregating alleles, the variance that is associated with estimates of population genetic parameters, such as the effective population size, are reduced, as are the results of statistical tests of natural selection (for example, TAJIMA'S *D*).

Because recombination allows genomic regions to have independent histories, no single phylogeny can describe the ancestry of a length of nucleotide sequence, therefore complicating phylogenetic analyses and inferences. For example, epidemiologists routinely use phylogenies to make inferences about routes of disease transmission for viral and bacterial species, and to estimate evolutionary parameters (for example, rates of MOLECULAR CLOCKS, rate heterogeneity, mutation bias and selection targets) or demographic processes[17–24]. Epidemiologists are particularly interested in timing events, such as the acquisition of pathogenicity in the evolutionary history of an organism, and phylogenetic approaches are often used for this purpose. For example, the main cause of the present-day HIV pandemic has been attributed to viruses from the M group of HIV type 1, which are thought to have originated in chimpanzees (*Pan troglodytes troglodytes*)[25]. Phylogenetic analyses on population data have dated the common ancestor of present-day diversity in HIV1 to ~1930 (REF. 23), with the proliferation and expansion of most of the common variants occurring soon after. The presence of recombination affects the phylogenetic predictions that can be made from population data, and some examples of this are outlined below.

Demographic and/or mutational processes alter the shapes and sizes of phylogenies or gene trees relative to trees for sequences that are generated under the assumption of neutrality and constant population size. For example, in any random neutral genealogy, branch lengths have an expected distribution[26]. A tree pattern in which most terminal branches (tips of the tree) appear long is often interpreted as evidence for population expansion[26]. However, population geneticists[27] know that recombination has a similar effect on gene trees: because recombination makes sequences more homogeneous (reduces the variance), it creates more star-like trees, or trees with longer terminal branches relative to internal branches, compared with those expected if the population size were constant (FIG. 2). It is therefore difficult to separate demographic explanations from recombination. Many viruses, such as HIV1 (REF. 28) or the virus that is associated with foot and mouth disease (FMDV) have star-like trees[24]. The shape of the tree might indicate a recent population expansion, but it could also have arisen as a consequence of recombination.

In a non-recombining environment, homoplasies on a tree result from several or recurrent mutations at a site (FIG. 3). For many pathogens and unicellular taxa, which are known to have high mutation rates per base relative to multicellular taxa (FIG. 1), the probability of recurrent mutation, and therefore of homoplasy, is high[29,30]. However, recombination also generates phylogenetic homoplasies[31,32] by moving mutations onto different genetic backgrounds (FIG. 3). Therefore, some sites will look as though they have been subject to recurrent mutations relative to other sites and appear more mutable than they truly are[33]. These 'recombination-induced' homoplasies will affect phylogenetic tests of mutation and selection, including

TRANSDUCTION
The introduction of a gene into a target cell by a viral vector.

TRANSFORMATION
The uptake of DNA by a bacterium from the surrounding environment.

CONJUGATION
The transfer of DNA from a donor cell to a recipient cell that is mediated by direct cell–cell contact.

TEMPLATE-SWITCHING PROCESS
A process by which the reverse transcriptase will switch templates during the replication process. If two viral haplotypes are present in the host, this will result in a recombinant product.

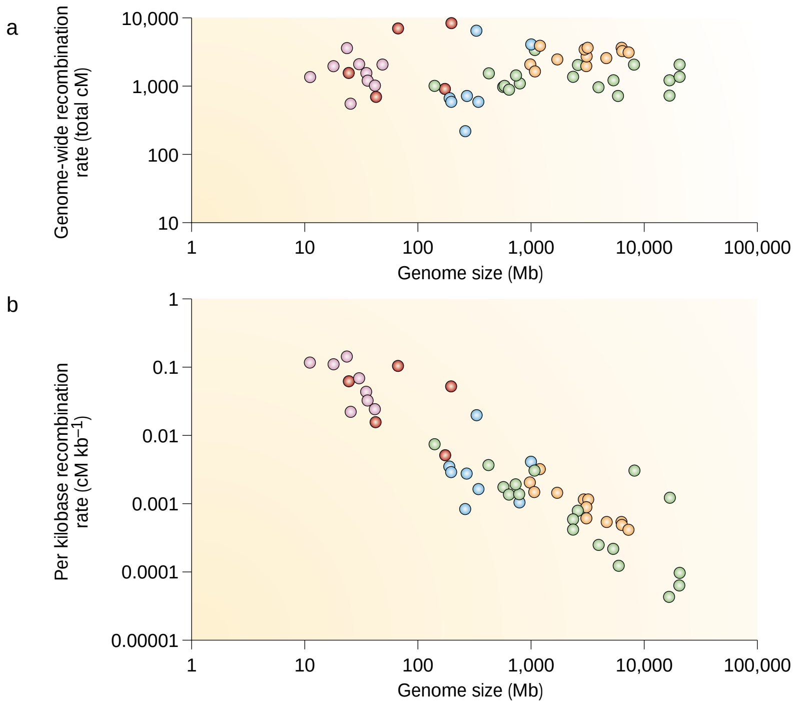


Figure 1 | **Recombination rates across eukaryotic taxa. a** | There is remarkable uniformity across eukaryotes, both unicellular and multicellular, in the overall recombination rate. This has been represented in the figure by relating the genetic map lengths (in centiMorgans, cM) for fungi, protists, invertebrates, vertebrates and plants to genome size (in megabases). **b** | Unicellular pathogenic taxa have much larger recombination rates per base than multicellular taxa. Here, the recombination rate per base ($cM\ kb^{-1}$) is plotted against genome size. Estimates for each species are averaged across sexes. Data from various sources. Protists (red circles): *Dictyostelium discoideum*, *Eimeria tenella*, *Plasmodium falciparum*, *Tetrahymena thermophila*, *Toxoplasma gondii*. Fungi (pink circles): *Agaricus bisporus*, *Aspergillus nidulans*, *Bremia lactucae*, *Cochliobolus heterostrophus*, *Cryptococcus neoformans* var. *neoformans*, *Gibberella fujikuro*, *Neurospora crassa*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*. Invertebrates (blue circles): *Aedes aegypti*, *Anopheles gambiae*, *Apis mellifera*, *Bombyx mori*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *D. simulans*, *Tribolium castaneum*, *Trichogramma brassicae*. Vertebrates (orange circles): *Bos taurus*, *Canis domesticus*, *Capra hircus*, *Danio rerio*, *Felix domesticus*, *Fugu rubripes*, *Gallus gallus*, *Homo sapiens*, *Ictalarus punctatus*, *Mus musculus*, *Oryzias javanicus*, *Ovis aries*, wild boar × domesticated hybrid species. Plants (green circles): *Arabidopsis thaliana*, *Brassica oleraceae*, *Chlamydomonas reinhardtii*, *Coffea canephora*, *Eucalyptus globulus*, *E. tereticornus*, *E. grandis*, *E. urophylla*, *Glycine max*, *Helianthus annus*, *Hordeum vulgare*, *Lactuca sativa*, *Lycopersicon esculentum*, *Oryza sativa*, *Pinus taeda*, *Pisum sativum*, *Solanum tuberosum*, *Sorghum bicolor*, *Triticum monococcum*, *T. aesitivum*, *Vigna inguiculata*, *Zea mays*.

phylogenetic tests of the molecular clock[34] (but see REF. 35). Recombination probably also affects approaches that attempt to identify regions or sites targeted by natural selection, which rely on accurately estimating the number of amino-acid and synonymous mutations in a given phylogeny (for example, see REF. 36). Recombination creates longer trees (FIG. 2), and therefore, more mutations — whether they be synonymous or non-synonymous — might be incorrectly inferred from the tree itself. Future research should explore the effects of recombination on these approaches to detect selection.

In summary, if recombination has occurred, but is ignored, and phylogenetic methods are applied, the following might result: mutation-rate heterogeneity across a given length of nucleotide or amino-acid sequence might be overestimated; the molecular clock hypothesis might be falsely rejected; and the distribution of terminal branch lengths of a tree will be overestimated, such that the timing of events are misinterpreted or underestimated.

It is therefore imperative that we have reliable tests for the presence of recombination in pathogens. In the past, inferences about the extent of genetic exchange have come either from laboratory-based approaches, such as LINKAGE MAPPING, or have been restricted to population genetic analyses of a few experimental organisms that have relatively low mutation rates. This is mainly because population genetic methods that assume an INFINITE-SITES MODEL[37] of evolution do not deal with violations of key assumptions. In the past ten years, several comparative or phylogenetic approaches have come to the fore. These approaches have been applied to highly polymorphic genomes, such as those of viruses (HIV, dengue, FMDV and others[5,33,38,39,121]) and bacteria (for example, see REFS 40,41), and have clearly shown that different genomic regions have different evolutionary histories (for example, see REFS 38,40), probably as a result of recombination.

More recently, new model-based approaches have been developed that do not rely on an underlying phylogeny and can also estimate the population rate of recombination in microbes and pathogens, rather than simply detecting the presence of recombination. All of these approaches depend on an assumed underlying mutation model and on assumptions about demography and population structure.

Methods to detect and estimate recombination. Although it is possible to identify recombination events by visually examining sequences for breakpoints in patterns of identity (for example, see REFS 41–44), other approaches are required if either mutation or recombination occur frequently. During the past 30 years, many methods have been developed to detect and estimate recombination rates. However, the high mutation rates of many unicellular taxa[29–30] make it difficult to analyse haplotype structure for the presence or absence of recombination; these high rates can mask the haplotype structures that have evolved over time. Recurrent mutation events at a single site could be attributed to variants at different sites being swapped by recombination onto different backgrounds. Consequently, numerous NON-PARAMETRIC and PARAMETRIC METHODS have been developed to detect the presence or absence of recombination and that take into account these mutation properties.

Non-parametric approaches. The non-parametric methods fall into two main categories: comparative and phylogenetic. Some comparative non-parametric approaches statistically evaluate various genome-wide characteristics or properties among divergent taxa. Regions that are more similar in base identity, codon usage and base composition might be footprints of ancestral recombination events. Tracts of nucleotide similarity can be statistically evaluated by testing whether individual sites that differ between taxa are more clustered in the genome than expected by chance.

FIXATION
The accumulation of a mutation to a frequency of 100% in a gene pool.

TAJIMA'S *D*
Summary statistic of the spectrum of allelic frequencies at different sites. An excess of rare variants indicates a recent reduction in variation either due to a selective sweep or an expanding population.

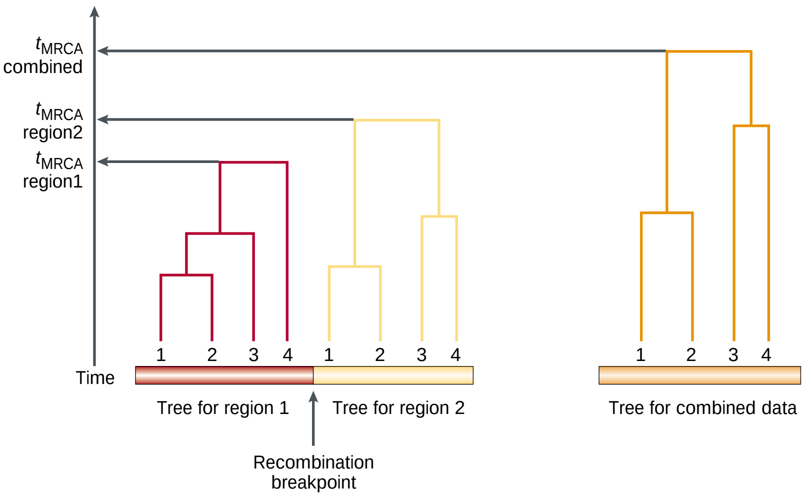


Figure 2 | **The effect of recombination on branch lengths.** Scheirup and Hein[127] were the first to describe how recombination affects the shapes of phylogenies. Even under the simplest conditions, such as when population sizes are at equilibrium, recombination can have marked effects on tree shape. The figure depicts a sequence that has been interrupted by a recombination event. A phylogeny for region 1 might resolve differently from that of region 2 because they are now allowed to have different evolutionary histories (trees on the left). If the total length of sequence is used to reconstruct phylogenetic relationships (tree on the right), two things will occur. First, the total length of the tree becomes upwardly biased. This is because recombination created incongruent trees for the two regions and, to accommodate this incongruence, more mutation events had to be inferred. Second, the terminal branches (tips of the tree) become longer relative to both the case in which no recombination occurs (not shown) and the internal branches in the tree. t_{MRCA}, time to most recent common ancestor.

MOLECULAR CLOCK
The principle that any sequence has a near-constant rate of evolution in all branches of a clade, which means that the amount of sequence divergence between two sequences will be proportional to the amount of time elapsed since their shared ancestor existed.

Such methods have been applied to ORTHOLOGOUS and PARALOGOUS comparisons[41,45]. Similarities in base or codon composition between more divergent taxa[1,46,47] have indicated frequent lateral transfers between bacterial species. By relying on codon bias and G+C composition, these methods distinguish native versus foreign DNA in host genomes by identifying regions that have an anomalous base composition relative to the rest of the host genome. For example, the genomes of numerous non-pathogenic and pathogenic species seem to have acquired foreign genes or regions (ranging in size from 5 kb to 500 kb) in the recent past. These foreign sequences have targeted integration sites at or near transfer RNA loci (see later section on the 'Evolutionary significance of recombination').

Many non-parametric methods (TABLE 1) use phylogenetic tools or some property of the phylogenetic tree on intraspecific data to identify recombinants. In principle, these methods are similar to identifying incongruent topologies among the phylogenies of different species. Recombination is implicated when regions or genes have different phylogenetic histories[2,49]; for example, when the topologies of the gene trees for one genomic region differ significantly from those of another[50]. Significant differences in topology are often ascertained by comparing the likelihoods of individual topologies of different genomic regions[50]. For example, methodologies have been developed to assess 'spatial variation' along a contiguous stretch of DNA using a 'sliding window' approach[49]. If different 'windows' of sequence have significantly different topologies this might indicate alternative evolutionary histories. A similar procedure compares the phylogenies of allelic variation at several coding loci[2,40]. Finally, other approaches (such as the 'informative sites test'[33]) take only a single tree and assess whether the properties of the tree deviate significantly from that inferred under the assumption that no recombination is occurring[31,33].

All phylogenetic and comparative methods vary in their ability to detect recombination. Posada and Crandall[48] recently assessed the ability of 14 non-parametric methods to detect recombination by using simulations. The parameters they varied in their simulations were recombination rates (from no recombination between loci to effectively unlinked sites), mutation rates and mutation-rate heterogeneity. The methods were examined with respect to power (to detect recombination) and to their sensitivity to the inferred underlying model of sequence evolution (false positives). The result is defined as being a false positive when the test falsely attributes a homoplasy to a recombination event, even when recombination is absent from the simulation. The tests varied greatly in power, especially when recombination rates were 'low'. When recombination was high, some tests, such as the informative sites test, only detected recombination 0–25% of the time, depending on the mutation rate. Other tests, such as the homoplasy test[32,42,48], were powerful but sensitive to false positives, and the informative sites test was neither powerful nor insensitive to false positives. It is clear that reliance on one phylogenetic method might therefore be problematic[51]. In fact, many of the phylogenetic methods often give incongruent results for the same data sets[51]. Finally, it is not helpful to apply a large battery of methods to test for recombination and allow the majority to rule. It is better to have one good test that is both powerful and reliable given the conditions, or preferably one that is based on a population genetic model of recombination.

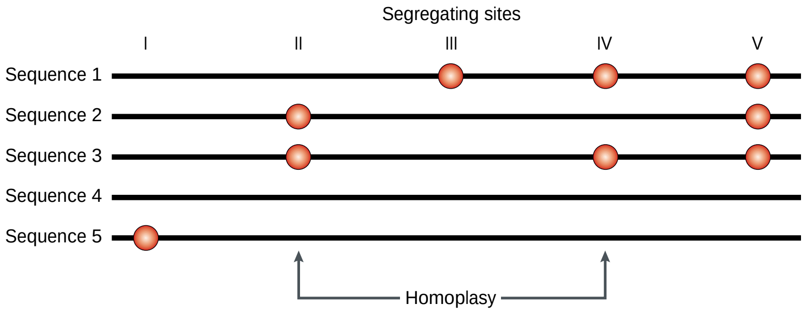


Figure 3 | **Homoplasy in a data set of five sequences.** Recombination and recurrent mutation can have similar effects on HAPLOTYPE variation among segregating sites in the genome by generating parallel patterns of evolution (homoplasy). The example in the diagram illustrates the point. The red circles are mutations that have arisen on either the same or different genetic backgrounds (sequences 1–5). If recombination occurs between a pair of sites, then the two products of crossing over, plus the two parental types — in total, four haplotypes — will be observed. Of the ten possible pairwise comparisons, only the presence of recombination between sites II and IV can explain the presence of all five haplotypes represented here. Note that this same haplotype configuration could have been generated by recurrent mutation at either site in the five sequences. The probability of this increases when mutation rates are high or variable across regions. In this example, homoplasy is said to be present between sites II and IV.

HAPLOTYPE
The combination of alleles at several loci on a single chromosome.

LINKAGE MAPPING
Markers that are physically close to a locus of interest segregate 'tightly' with the locus and will statistically be more closely associated with the observed variance of a trait. This property can be used to detect association in a population between a genetic marker and a locus that contributes to a particular phenotype.

INFINITE SITES MODEL
A simple model of the inheritance of quantitative traits that assumes an infinite number of unlinked loci. As a result, it is possible to assume that mutations occur only once at a particular locus, and that the probability of a mutation occurring at a site has a Poisson distribution.

NON-PARAMETRIC METHODS
Statistical procedures that are not based on models or assumptions pertaining to the distribution of the variable.

PARAMETRIC METHODS
Estimators and procedures that are based on models.

ORTHOLOGOUS GENES
Homologous genes in different species, the lineage of which derives from a common ancestral gene without gene duplication or horizontal transmission.

PARALOGOUS GENES
Homologous genes that originated by gene duplication (for example, human α-globin and β-globin).

LINKAGE DISEQUILIBRIUM
(LD). The condition in which the frequency of a particular haplotype for two loci is significantly greater than that expected from the product of the observed allelic frequencies at each locus.

MONTE CARLO METHOD
The use of randomly generated or sampled data and computer simulations to obtain approximate solutions to complex mathematical and statistical problems.

Perhaps the greatest criticism of the above approaches has less to do with power, false detection or even the reliance on an inferred phylogeny, than with the fact that not one of these methods can estimate the 'rate' of recombination. The inability of phylogenetic methods to estimate the population rate of recombination is a serious limitation: comparisons between genes or species become difficult, evolutionary models remain untested and population rates of recombination for many species remain unknown.

Parametric approaches. The parametric methods are model based and deal specifically with haplotype structure — LINKAGE DISEQUILIBRIUM (LD) — in population genetic data sets. In population genetics, measuring LD and assessing its significance has become standard practice. However, acquiring an estimate of LD alone is not satisfying as it merely tells us whether two sites are in significant LD or not; it does not tell us how often recombination occurs nor how it varies from region to region or from species to species.

To estimate population recombination rates, a population model that addresses the information embedded in haplotype variation is required. Several existing parametric methods are independent of any inferred phylogeny and use statistics that summarize LD between segregating sites to detect recombination events. The more distant two sites are from each other, the greater the probability that a recombination event will occur between them; recombination therefore creates a decay of LD between nucleotide sites. How LD decays with respect to the physical distance between sites is both an analytical and theoretical description of how recombination breaks up pairwise associations and is often used as a test of recombination[3,52–54].

The main advantage of LD-based methods over phylogenetic methods is that LD statistics have direct relationships both with the data and with population genetic theory[55]. As a result, LD can be used with model-based (parametric) approaches to estimate the rate of recombination[56–60]. However, a simplification of the population history of a sample, and of the mutation and recombination models, is almost always required for these models when estimating population parameters.

An appropriate tool for modelling the effects of recombination on a population sample of gene sequences is the coalescent. The coalescent is a statistical, genealogical description of a set of alleles, haplotypes or sequences that are sampled randomly from a population. It differs from classical population genetics in that it focuses on the genealogical structure of a sample, rather than on the properties of the population as a whole[61–63]. Because the genealogy is unknown, it is treated as an unknown parameter; therefore several potential genealogies exist for a data set, unlike in a phylogeny where a single history is assumed (BOX 2).

Several coalescent and LD-based methods have been proposed to estimate the population recombination rate[56,58,59,64,65]. For example, the minimum number of recombination events[60] that have taken place in a sample can be obtained using the four-gamete test: an estimate of the recombination rate is derived from the number of pairwise comparisons that show all four possible gametes (FIG. 3). Hudson[58] also derived an estimator based on the variance in pairwise differences. Generally, the estimator is a summary of the degree of association between all pairs of segregrating sites. This and other so-called 'method of moments' estimators are known to overestimate the population recombination rate and to have large confidence limits[66]. Likelihood estimators of the population recombination rate also exist that use computationally intensive MONTE CARLO METHODS[64,65,67]. Because the need for high computational power is a limitation, Hudson[59] developed a flexible, *ad hoc* method for estimating the population recombination rate by combining the coalescent likelihoods of each haplotype configuration for all pairwise comparisons of segregating sites. The resulting composite likelihood estimate (CLE) is the estimate that is associated with the largest sum of probabilities over all pairs of sites (BOX 3). The method does well in terms of minimizing bias and variance[58] compared with Hudson's earlier moment estimator[58] and with other *ad hoc* approaches. However, because the method sums over all pairwise comparisons, it introduces non-independence and, as a result, the true variance of the single estimate of the CLE recombination rate is unknown[59]. A simplification of this approach is presented in BOX 3.

Table 1 | Non-parametric tests of exchange*

Method	Implementation	Reference
Bootscanning	SIMPLOT	6,104
Geneconv	GENECONV	41
Homoplasy test	HOMOPLASY TEST	31
Informative sites	PIST	33
Phylogenetic profiles	PHYPRO	105
Partial likelihood	PLATO	106
Rdp	RDP	107
Recombination	RECPARS	108
Reticulate	RETICULATE	109
Runs test	RUNS TEST	44
Sneath test	SNEATH TEST	110
Triple	TRIPLE	111

*Only a representative sample is listed. All the methods listed in the table are designed to detect the presence of recombination or some form of genetic exchange (such as gene conversion). Each test was evaluated by Posada and Crandall[48] with respect to accuracy and power (see REF. 48 either for URLs at which the programs can be found or for the source code).

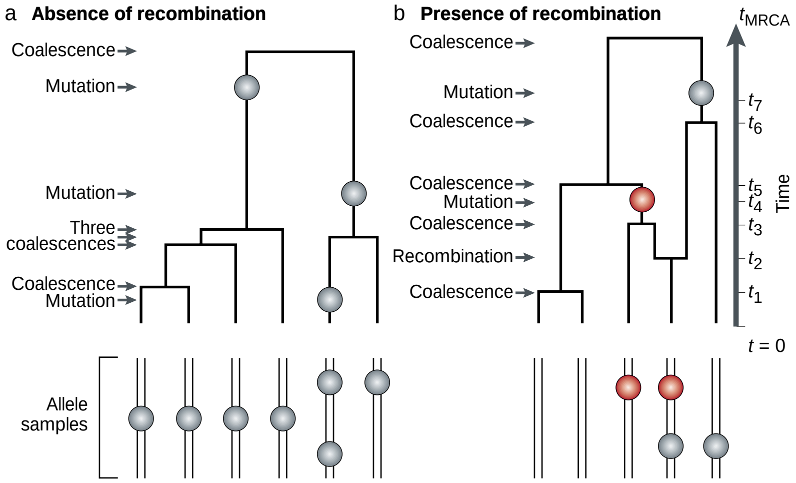
Because of the flexibility of the CLE method, it can potentially be expanded to incorporate deviations from the standard coalescent. Recently, an extension of Hudson's CLE estimator incorporated two main changes[53]: relaxation of the assumptions of the infinite sites model (see below), which allows for recurrent mutation and rate variation among sites; and a model of GENE CONVERSION (BOX 3). A model of gene conversion is necessary for circular genomes. Recombination requires two breakpoints for a circular genome, otherwise genome products of different sizes are created.

Box 2 | Using the coalescent to estimate recombination

The coalescent is a statistical genealogical description of a sample of alleles that is useful for modelling population parameters, such as the effects of recombination. A simulation of the coalescent can be carried out by allowing pairs of lineages to 'coalesce' by going backwards in time (panel **a**). The point in time at which all alleles have coalesced is known as "the mean time (across genealogies) to the most recent common ancestor" (t_{MRCA}). The rate at which samples or branches coalesce in the genealogy depends on the size of the sample and the size of the population. The more individuals sampled, the faster a pair of lineages will be found that can coalesce. Afterwards, neutral mutations can be placed randomly onto branches of the genealogy at a constant rate $\theta/2$ (for haploids, $\theta/2 = 2N_e m$, the population mutation parameter, where N_e is the effective population size and m is the mutation rate per generation per locus[61–63]). In the absence of recombination (panel **a**), a genealogy reflects the ancestral relationships of a sample of alleles (the terminal branches). Circles are mutations that have been mapped onto branches (shown below the genealogy). If mutations occur at a constant rate, then larger genealogies will be expected to have a greater number of mutations, and the t_{MRCA} of all sampled alleles from a single population will be directly related to the level of polymorphism observed in the population. This shows how the effects of genealogical history and mutation are kept separate.

Similarly, in the presence of recombination, the number of recombination events that is observed in a population also depends on the time to the MRCA (panel **b**). The older the population, the greater the probability that recombination events occur either among individuals or in the ancestral history of the sample. The effects of recombination on sample history are therefore a function not of the absolute recombination rate, but of the product of the per-gene per generation rate of crossing over (genetic map length), r, and the effective population size, N_e.

Panel **b** shows how recombination breaks down linkage relationships; before the recombination event occurred, the mutations (grey and red circles) were on different lineages, but after the event, they are on the same haplotype as well as different haplotypes. As shown below the genealogy in the figure, different parts of an allele have different ancestral relationships and, as in FIG. 3, all four possible haplotypes are shown. The times for each coalescent and recombination event are shown. Using the coalescent, the probability of a recombination event occurring can be modelled.



GENE CONVERSION
The non-reciprocal transfer of information between homologous genes as a consequence of heteroduplex formation, followed by repair of mismatches in the heteroduplex. In this context, conversion is associated with two crossovers.

Recombination rates of pathogens. Allowing for recurrent mutation, rate heterogeneity and a model of gene conversion, we provide some estimates of the population recombination rates for several pathogens in TABLE 2 and FIG. 4. In some cases, many loci for the same species, or the same loci sampled for different populations, are included. Regardless, these results show that recombination rates can be substantial in these organisms (TABLE 2), especially relative to the population mutation rate (FIG. 4), and that LD approaches, such as the ones used in TABLE 2, have sufficient power to detect recombination. Even if the mutation processes for many of the taxa listed in TABLE 2 are more complex than anticipated, the method used to estimate the population recombination rate (LDhat) seems to be robust even when the mutation model is incorrect[53]. Many of the estimates should be considered to be rough approximations as the confidence limits are unknown; however, the sample sizes and mutation rates are substantial and contribute sufficiently to reasonable estimates of the recombination rate. Furthermore, a randomization approach can be used to determine which estimates are significantly different from zero (BOX 3).

Although some genes and species show substantial recombination rates per gene, some data sets either give no sign of recombination or yield estimates that are not significantly different from zero, even when homoplasy is present in the data (TABLE 2). These homoplasies probably result from recurrent mutation rather than recombination. The existence of a population structure might also reduce the composite-likelihood estimate by elevating LD; nevertheless, recombination will still be greater than zero, even if not significantly, if all four gametes (the product of recombination) between any pair of sites are present. Different population histories, such as a recent colonization or population expansion, will also affect the probability of observing all four haplotypes, or the overall pattern of LD[68], and so will bias the estimate downwards. Finally, in one reported case — *Trypanosoma cruzi* loci — does the result contrast with that previously published for the same data set, in that recombination was not detected using the LDhat approach shown in TABLE 2. Note also that Machado and Ayala[69] observed incongruencies between the phylogenies of two separate genes — *trp* and *dhi* — of *T. cruzi* (TABLE 2). It could be that recombination is actually occurring; however, it might be occurring at a rate low enough that it is not detectable within loci, but is detectable between loci.

Are unicellular pathogens freely recombining or are the estimates shown in TABLE 2 more consistent with clonal (asexual) life histories? Do specific phylogenetic groups freely exchange genetic material between individuals more so than others? To answer these questions, it is necessary to investigate both the population mutation rates and the recombination rates of various unicellular pathogens. Species certainly vary, but there seem to be some phylogenetic consistencies between recombination rates, relative to the population mutation rates, across broad phylogenetic groupings of taxa (FIG. 4). For bacteria, the estimated number of recombination events per mutation is in fairly good agreement with direct estimates obtained from laboratory experiments (reviewed in REF. 70). The wide range of values that is observed in protozoans is clearly an artefact of sampling and arises as a consequence of different geographic populations having different levels of nucleotide polymorphism. The viral estimates are smaller than those of the bacteria and protozoans that were sampled. This is probably because viruses have high population mutation rates (relative to those estimated for humans, *Drosophila* and *Arabidopsis*), as well as high recombination rates.

Box 3 | Composite likelihood estimation of recombination

Hudson[59] proposed a linkage disequilibrium (LD)-based parametric method for estimating the population recombination rate, known as the composite likelihood estimate (CLE). This approach involves applying LIKELIHOOD METHODS to analyse a pair of polymorphic sites and is based on the probability that certain allelic configurations will be present at pairs of sites[60,100,101]. McVean *et al.*[53] extended this rationale by allowing for either recurrent mutations or more complex mutations to be incorporated (this is called the LDhat method here[53]). To estimate the recombination rates of sequences that are subject to recurrent mutation, the population mutation rate θ must first be estimated, taking into account the possibility of recurrent mutation ($\theta = 4N_e\mu$, where N_e is the effective population size and μ is the per-locus mutation rate per generation). This is done using a FINITE-SITES version of the Watterson estimate of θ. Given θ, the probability for each pairwise configuration of sites is then calculated using the approach of Fearnhead and Donnelly[65] (not shown).

The composite likelihood (ℓ_c) of the population recombination ($4N_e r$, where r is the rate of recombination) is the summation of the likelihoods across all pairs of segregating sites:

$$\ell_c(4N_e r) = \Sigma_{i,j}(X_{ij} \mid 4N_e r_{ij}),$$

where $\ell(X_{ij} \mid 4N_e r_{ij})$ is the logarithm of the likelihood of the data for segregating sites i and j given that

$$r_{ij} = \frac{rd_{ij}}{L-1}$$

where d_{ij} is the distance between pairs of sites i and j, and L is the total length of the sampled sequence (REFS 53,59).

For recombination to occur in a circular genome, two recombination events, rather than one, are required to maintain the integrity of the genome. Therefore, in this situation, recombination must be estimated by using a model that is based on gene conversion — which effectively corresponds to two crossovers occurring in a sampled sequence. Using a fixed average tract length (t, defined as the distance between the two recombination breakpoints), the population rate of gene conversion, γ, can be estimated, where

$$y = 8N_e ct$$

and c is the probability of gene conversion[53]. For the estimates calculated in TABLE 2 and FIG. 4, t is assumed to be 500 bases. For sample gene lengths that are 500 bases or less, the model is no different than that for the single crossover case.

Sometimes the data is of dubious quality, for example owing to sequencing errors. In these cases, the test can still be used: sequencing errors often manifest themselves as rare variants, so a 'frequency sieve' can be used to remove all rare variants that are present at less than 10% frequency.

Significant deviations from the null model — which assumes that no recombination takes place — can also be assessed by using a randomization of likelihoods[52]. In this test, the genomic positions of the sites are randomly reassigned, and then the likelihoods are recalculated. A result that deviates significantly from zero is obtained when the likelihood of the estimate for the observed data is greater than the randomized estimates at least 95% of the time. For example, the LDhat approach revealed that human mitochondrial DNA, which is largely maternally inherited[102,103] and was thought never to recombine, had estimates of population recombination rate that were larger than zero for two complete genome data sets. However, the estimates were found to not be significantly different from zero when the randomization test was applied.

LIKELIHOOD METHOD
The use of a model to determine the most probable estimate of a parameter that best fits the observed data.

FINITE SITES MODEL
By contrast to the infinite sites model of evolution, in this model, multiple mutation events can occur at the same site.

Maynard Smith and colleagues[71] argued that, if the recombination rate were 20 times that of the mutation rate, then a population should be in linkage equilibrium (that is, freely recombining). Under this criterion, the populations described in FIG. 4 are not at linkage equilibrium, as the overall recombination rate relative to the mutation rate is less than 20:1 (REF. 71). Therefore, the clonal life history of these unicellular taxa, as well as their population structure, probably contributes to the LD that is observed.

The evolutionary significance of recombination

Bacteria. Recombinant genomes are known to be associated with changes in phenotype or fitness, including heightened or reduced pathogenicity or virulence. For example, comparative analyses of bacterial genomes have shown that 'foreign' genomic regions seem to be associated with the acquisition of pathogenicity among some taxa[1,46,47,72]. These 'pathogenic islands' are either not found or are disrupted in related non-pathogenic species (for example, see REF. 72). 'Hot spots' of pathogenicity in species such as *Neisseria meningitidis*, uropathogenic and enteropathogenic *Escherichia coli*, *Shigella*, *Salmonella*, *Haemophilus influenzae* and *Helicobacter pylori* have been acquired through lateral transfers[1,46–48,72]. Recombination through lateral transfers of genetic elements often confer changes in susceptibility to immunity or drugs. For example, epidemic strains of *Staphylococcus aureus* in the United Kingdom have recently acquired methicillin resistance through repeated horizontal transfer of genes that confer resistance from ancestral methicillin-resistant strains[73].

Multi-locus genotypes have been used extensively to investigate the genetic structure of bacterial pathogens in the past 20 years (REFS 74–78). Phylogenetic incongruencies between housekeeping genes in the genomes of *Neisseria*, *Staphylococcus*, *Campylobacter*, *E. coli*, *Salmonella* and *Helicobacter* species (for example, see REF. 4) reveals the relevance of recombination in the evolutionary histories of these species. In particular, this approach showed that, although some bacteria, such as *N. gonorrhoeae* show high levels of recombination, others, such as *E. coli* and *Salmonella*, have a predominantly clonal population structure. It has been argued that, in *Neisseria*, recombination generates a pool of variation from which resistant clones can arise and proliferate, having evolved to a specific ecological niche (for example, see REFS 70,71). However, the clonal propagation of a successful clone or haplotype will also contribute to extensive LD in the population[70] and might obscure the estimate of the frequency with which recombination actually occurs.

Parasitic protozoans. Genetic exchange is now known to occur during the life cycle of many parasitic protozoa, including malaria parasites, coccidia and trypanosomes. Even in *Toxoplasma gondii*, in which variation falls into only three distinct clonal lines, recombinants are occasionally found. The three strains vary in their degree of virulence, with one strain clearly being the most virulent. Grigg *et al.*[79] have argued that virulent genotypes have arisen from the 'mixing' of two out of the three lines and have supported this argument with experimental assays showing recombinant genotypes that seem to be associated with much higher virulence.

A large proportion of the *Plasmodium* and *Trypanosoma* genomes are devoted to variant antigen genes or encode surface proteins, such as merozoite (MSP) and circumsporozoite (CSP) proteins, and other variant surface glycoproteins (VSGs). This indicates a

Table 2 | Examples of recombination and gene-conversion rates for pathogen loci

Species	Gene	Population sample size, gene length (bp)	Population mutation rate (θ)	Population recombination rate (γ)	Reference
Haemophilus Influenzae	*recA* *adk*	37, 428 37, 428	0.0191 0.021, 0.0045	22* 5*	40 40
Streptococcus Pneumoniae	*arc1* *recA* *ldh* *hexA*	25, 401 36, 424 13, 475 12, 443	0.0112 0.0137 0.0109 0.00897	>100 77 22 NS >100	112 112 112 112
S. dysgalactiae ssp. *equisimilis*	*recA*	16, 459	0.145	11*	113
Neisseria meningitis	*adk*	28, 648	0.0196	11.11*	76
Trypanosome cruzi	*trp* *dhfr*	38, 1290 38, 1473	0.01476 0.00688	0 0	69 69
Plasmodium falciparum	*ama* *glurp* *pfs48/45*	23, 1578 41, 1248 44, 1374	0.016 0.0047 0.007	>100 10.32 14.14	114 115 116
P. vivax	*msp2*	175, 249	0.0264	0	117
Measles Khartoum	Nucleoprotein	40, 456	0.011	0	118
Measles USA	Nucleoprotein	35, 456	0.069	25.25	118
Measles Vietnam	Nucleoprotein	25, 456	0.02	0	119
Measles	Nucleoprotein	50, 1830	0.089	3 NS	120
HCV	CC	6, 8922	0.325	0.84 NS	121
Helicobacter pylori	*flaA* *bab*	33, 471 39, 506	0.045 0.021	41 33*	122 123
HIV2	*env3* *env12*	21, 682 21, 1364	0.302 0.102	26 43	124 124
Dengue	*C*, *prM*/*M*, *E*	7, 2322	0.053	60	38
HIV1B	*env3* *env12*	93, 658 93, 1316	0.333 0.144	100* 100*	124
FMDV	capsid gene	22, 2404	0.15	>100	39
Human rhinovirus	VP4/VP2 region	72, 420	0.092	100	125
TT	DNA virus unknown gene	18, 222	0.110	0	126

For bacterial and viral data sets, the gene conversion model was used, assuming a mean tract length of 500 bases. These population recombination rate estimates include a subset of estimates calculated in McVean *et al.*[53]. For protozoan parasites, the standard recombination model was used. NS indicates an estimate that is not significant from zero using the permutation test[53]. *Estimates that were significantly different from zero when the frequency sieve (the exclusion of sites for which the rarer segregating variant is less than 10% in frequency) was used. *adk*, adenylate kinase; *ama*, apical membrane a; *arc1*, anoxic reductase control 1; *bab*, blood group antigen binding; bp, base pair; *C*, capsid genes; CC, complete coding sequence; *dhfr*, dihydrofolate reductase; *E*, envelope genes; *env*, envelope-protein-encoding gene; *flaA*, flagellin A; FMDV, foot and mouth disease virus; *glurp*, glutamine-rich protein-encoding gene; HCV, hepatitis C virus; *hexA*, hexokinase A; HIV1B, human immunodeficiency virus type 1B; *ldh*, lactose dehydrogenase; *msp*, merozoite-surface-protein-encoding gene; *pfs*, *Plasmodium falciparum* surface-protein-encoding gene; *prM*/*M*, pre-membrane/membrane genes; *trp*, tryptophan; VP4/2, variable protein-encoding genes 4/2.

role for genetic exchange in enhancing the diversity of these species' repertoire. The rate at which recombination breaks down the associations between such genes might influence the maintenance of antigenically distinct 'strains'[80,81], as well as the spread of drug resistance[82,83]. The frequency of sex in trypanosomes in nature has been a matter for speculation and controversy, with conflicting results arising from population genetics analyses (for example, see REF. 69). However, the underlying mechanisms of genetic exchange in *T. brucei* are becoming clearer. It seems that recombination has an active role in generating antigenic diversity; for example, a gene that encodes a eukaryotic homologue of RecA is associated with homologous and non-homologous recombination of VSG loci[84].

The mechanisms that underlie genetic exchange and sex in *Plasmodium* spp. are much clearer. Malaria parasites are hermaphroditic and haploid for most of their life cycle; asexual replication occurs in the primate host, whereas zygote formation and meiosis occur during the mosquito phase of development. In populations in which endemicity is low, the fusion of male and female gametes from the same clone (selfing) is more likely and results in no effective recombination. However, in populations with high endemicity, individual mosquito hosts are infected with several *Plasmodium* genotypes, and fusion of gametes from different clones (outcrossing) can result in recombination.

Recombination clearly affects nucleotide haplotype structure in *P. falciparum* genomes. Conway *et al.*[3] have shown that haplotype structure breaks down readily at

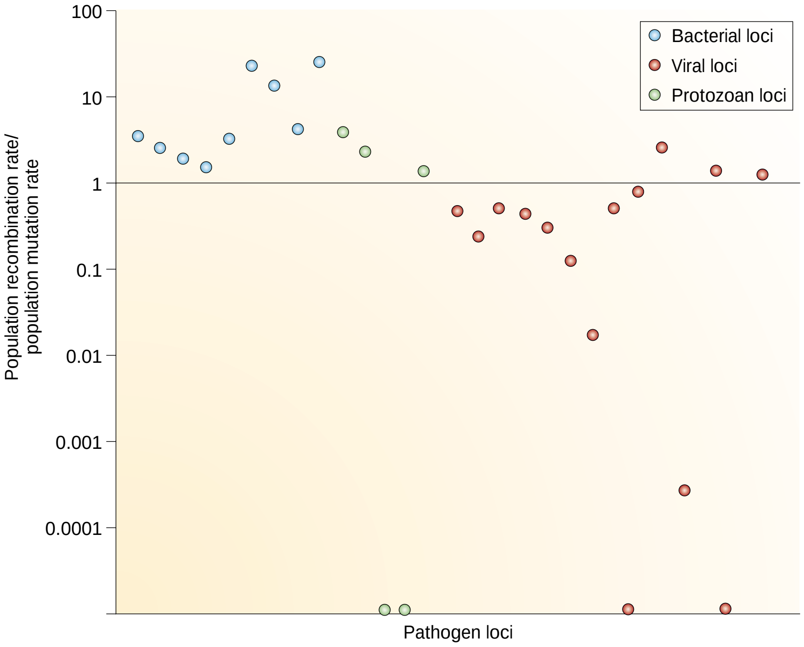


Figure 4 | **Recombination rates relative to mutation rates for bacterial, protozoan and viral loci.** The loci analysed are those listed in TABLE 2. As the graph shows, the three groupings of taxa have similar recombination rates relative to mutation rates. The values of the ratio of the population recombination rate and the mutation rate are smaller for viruses than for bacteria (1.5 to >25) or protozoans (0 to >100), although note the large variance among the viral estimates. This is mainly due to recombination being undetected in some data sets. In addition, these estimates are by no means independent of each other as some estimates are for different loci from the same taxa. The mean viral recombination rates per site are not necessarily lower than that of bacteria (indeed, for many viruses, the population rate of recombination is >100); the low ratios probably arise because of the extremely large mutation rate of many viruses. The large variance observed among protozoan and viral loci reflects not only reduced recombination (or increased mutation rate, as argued above for the virus data), but also demographic effects on polymorphism and recombination rate estimates. For example, endemicity is likely to be important as host individuals that are infected by several genotypes will contribute to increased recombination frequencies. The straight line indicates a 1:1 relationship between population recombination rates and mutation rates.

the merozoite surface protein 1 (*msp1*) locus in samples taken from populations in Africa, which indicates extensive recombination among these parasites in some populations. MSP1 is the most-abundant surface component of the erythrocyte-invading stage of *P. falciparum* and has highly divergent alleles with stable frequencies in endemic populations. Recombination might contribute to the diversity that is maintained as an arsenal against host immunity. LD among 12 microsatellite loci in each of 9 populations, 3 each from Africa, South America and Asia[9] revealed considerable global variation in the amount of LD among loci within and between populations. It was clear that populations with higher variation, higher endemicity and hence higher values of N_e had the least amount of LD.

Applications of LD approaches

Identifying drug-resistance genes. In medical genetics, associations between phenotypes and genetic markers that are created through genetic drift are important for mapping mutations that are associated with pathogenicity or virulence. Although recombination might complicate phylogenetic approaches to identifying regions that contribute to fitness differences in pathogens (for example, see REFS 22,36), the action of both genetic drift and recombination are necessary and beneficial when using population genetic approaches. However, too little or too much recombination can also be detrimental. For example, recombination is necessary to identify regions that are important for immunogenicity: it generates independence among sites, which ensures that not all markers will be linked to the phenotype. By contrast, if recombination rates are too high, it might break up all associations between markers and the phenotype[9], and a larger, more densely distributed number of markers will be required to pinpoint the region of interest.

Two examples of how population genetic analysis of LD has helped to identify potential drug-resistant targets come from studies of *P. falciparum*, the malaria parasite. In the first study, genomic regions that are likely to contain drug-resistance targets were identified owing to these regions having greater haplotype structure relative to other genomic regions. Chloroquine-resistant (CQR) *P. falciparum* parasites were initially reported ~45 years ago in Southeast Asia and South America[85]. An analysis of genome-wide microsatellite variation of *P. falciparum*[86] showed that the level of genetic diversity varied substantially among regions of the parasite genome, and that extensive LD surrounds the key CQR gene (*pfcrt*) on chromosome 7. This disequilibrium and its decay rate in the *pfcrt*-flanking region are consistent with strong directional SELECTIVE SWEEPS[87] in this region occurring over only 20–80 sexual generations. This disequilibrium was particularly pronounced in a single resistant haplotype that has spread to high frequencies throughout most of Asia and Africa.

In the second example, the evolutionary strategy used by the parasites to evade host immunity was used to identify the most important targets of protective immunity from the host. Natural selection maintains allelic variation in some antigens of *P. falciparum*. To identify the region of *msp1* under strongest selection to maintain alleles in populations, Conway *et al.*[88] identified the region with the lowest inter-population variance in allele frequencies. Serum IgG antibodies against each of the two most frequent allelic types of the variable block 2 in the *msp1*-encoded protein were strongly associated with protection from malaria caused by *P. falciparum*. The above analysis depended on the ability of recombination to generate independence among sites in the gene. If all sites were completely linked, then they might show the same (high) population variation due to linkage and it would be impossible to identify the crucial region.

Conclusions

Recombination occurs at substantial frequencies in natural populations of many pathogenic species. It seems reasonable to interpret this as evidence that recombination has an important evolutionary role. Whether the fitness of obligately sexual species is greater than that of asexual or FACULTATIVE ASEXUAL SPECIES is still not known; however, it seems clear that facultative sex endows these taxa with an alternative mechanism, other than mutation, to generate new multilocus genotypes with different fitnesses. Although recombination can complicate

SELECTIVE SWEEPS
As a positively selected allele rises to fixation, linked alleles will be maintained in the population; by contrast, alleles that are linked to the non-selected allele are lost from the population. The consequence of this selective sweep is usually a reduced variation around the selected locus.

FACULTATIVE ASEXUAL SPECIES
Species in which reproduction is known to occur asexually or sexually.

some forms of analysis, and needs to be reckoned with before further investment is made, it might facilitate other (population genetic) modes of analysis.

It is common for epidemiologists to use traditional phylogenetic methods to study the intraspecific population sequence data of pathogens, even though recombination occurs frequently. However, the phylogenetic approach is only applicable under the assumption that recombination does not occur. Violation of this assumption means that a single tree cannot describe the evolutionary history of a genomic region. If recombination has indeed occurred in the data, methods that model the recombination process explicitly should be used. Such methods might not model accurately the mutation process of many of these taxa (for example, viruses), which might involve a potentially large number of parameters. However, tests based on simulations of more complex models indicate that the method used might be surprisingly reliable even in the face of this complexity[53]. In the future, when computation is less of a hindrance, full-likelihood approaches will be more amenable, and therefore, confidence limits for estimates will be available.

Better estimates of the rate of recombination will facilitate the development of association strategies for identifying regions of interest in pathogen genomes. The number and density of markers used in a mapping study depends on the rate of recombination between markers and phenotypes. High rates of recombination require many more markers because linkage relationships between markers and phenotypes will be broken down. Furthermore, to pinpoint the most important sites or regions of a protein for vaccine target development, the population genetic footprints of selection on genetic diversity can be identified only in the context of the independence generated by recombination. It seems clear that a reliable estimator and test of recombination is necessary to assess the significance of recombination in natural populations, especially if a particular haplotype is thought to be associated with a fitness change. We have shown how recombination facilitates these studies in *Plasmodium*. Similar approaches in other systems will aid in the development of effective means by which to control pathogenic species.

1. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
Reviews work of the authors and other researchers (for example, see references 46,47,73) who have used interspecific comparative approaches to identify the foreign transfers between species that correlate with a change in pathogenicity and fitness.
2. Guttman, D. S. & Dykhuizen, D. E. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**, 1380–1383 (1994).
3. Conway, D. J. *et al.* High recombination rate in natural populations of *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **96**, 4506–4511 (1999).
4. Feil, E. J. & Spratt, B. G. Recombination and the population structures of bacterial pathogens. *Annu. Rev. Microbiol.* **55**, 561–590 (2001).
5. Robertson, D. L., Hahn, B. H. & Sharp, P. M. Recombination in AIDS viruses. *J. Mol. Evol.* **40**, 249–259 (1995).
6. Lole, K. S. *et al.* Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* **73**, 152–160 (1999).
7. Motomura, K. *et al.* Emergence of new forms of human immunodeficiency virus type 1 intersubtype recombinants in central Myanmar. *AIDS Res. Hum. Retroviruses* **16**, 1831–1843 (2000).
8. Conway, D. J. *et al.* A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. *Nature Med.* **6**, 689–692 (2000).
By using a population genetic approach, this paper reports the identification of highly variable regions in the *P. falciparum msp1* gene that seem to be under diversifying selection. The authors also developed a vaccine based on the translated portion of the *msp1*-encoded protein.
9. Anderson, T. J. *et al.* Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol. Biol. Evol.* **17**, 1467–1482 (2000).
A microsatellite study that pre-dates some of the recent malaria single-nucleotide polymorphism surveys and that reveals extensive diversity and recombination in *P. falciparum*.
10. Michod, R. E. & Levin, B. R. *The Evolution of Sex: An Examination of Current Ideas* (Sinauer, Sunderland, Massachusetts, 1988).
11. Barton, N. H. & Charlesworth, B. Why sex and recombination? *Science* **281**, 1986–1990 (1998).
12. Otto, S. P. & Lenormand, T. Resolving the paradox of sex and recombination. *Nature Rev. Genet.* **3**, 252–261 (2002).
A balanced and up-to-date review of the evolutionary significance of sex and recombination, and of the current status of the field.
13. Muller, H. J. Some genetic aspects of sex. *Am. Nat.* **66**, 118–138 (1932).
14. Muller, H. J. The relation of recombination to mutation advance. *Mutat. Res.* **1**, 2–9 (1964).
15. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
16. Kim, Y. & Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777 (2002).
17. 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).
18. Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569 (1992).
19. Holmes, E. C. *et al.* The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *J. Infect. Dis.* **17**, 45–53 (1995).
20. Leigh-Brown, A. J. Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population. *Proc. Natl Acad. Sci. USA* **94**, 1862–1865 (1997).
21. Yang, Z. PAML: a program for package for phylogenetic analysis by maximum likelihood. *Cabios* **15**, 555–556 (1997).
22. Yang, Z. & Bielawski, B. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).
23. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796 (2000).
24. Haydon, D. T., Bastos, A., Samuel, A. & Knowles, N. Evidence for positive selection in foot-and-mouth-disease-virus capsid genes from field isolates. *Genetics* **157**, 7–15 (2001).
25. Gao, F. *et al.* Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441 (1999).
26. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562 (1991).
27. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
28. Frost, S. D., Dumaurier, M. J., Wain-Hobson, S. & Leigh-Brown, A. J. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. *Proc. Natl Acad. Sci. USA* **98**, 6975–6980 (2001).
29. Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA* **88**, 7160–7164 (1991).
30. Drake, J. W. & Holland, J. J. Mutation rates among RNA viruses. *Proc. Natl Acad. Sci. USA* **96**, 13910–13913 (1999).
31. Maynard Smith, J. & Smith, N. H. Detecting recombination from gene trees. *Mol. Biol. Evol.* **15**, 590–599 (1998).
32. Eyre-Walker, A., Smith, N. H. & Smith, J. M. How clonal are human mitochondria? *Proc. R. Soc. Lond. B Biol. Sci.* **266**, 477–483 (1999).
33. Worobey, M. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria and mitochondria. *Mol. Biol. Evol.* **18**, 1425–1434 (2001).
34. Schierup, M. H. & Hein, J. Recombination and the molecular clock. *Mol. Biol. Evol.* **17**, 1578–1579 (2000).
35. Posada, D. Unveiling the molecular clock in the presence of recombination. *Mol. Biol. Evol.* **18**, 1976–1978 (2001).
36. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002).
37. Kimura, M. Theoretical foundation of population genetics at the molecular level. *Theor. Popul. Biol.* **2**, 174–208 (1971).
38. Worobey, M., Rambaut, A. & Holmes, E. C. Widespread intraserotype recombination in natural populations of dengue virus. *Proc. Natl Acad. Sci. USA* **96**, 7352–7357 (1999).
39. Bastos, A. D. *et al.* Genetic heterogeneity of SAT-1 type foot-and-mouth disease viruses in southern Africa. *Arch. Virol.* **146**, 1537–1551 (2001).
40. Feil, E. J. *et al.* Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl Acad. Sci. USA* **98**, 182–187 (2001).
A thorough analysis of allelic variation at several loci in six bacterial pathogens. Four of the species seem to recombine extensively; however, all show some sign of recombination.
41. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526–538 (1989).
Describes a 'runs' test to compare clustering of segregating variants between species. The principle on which this test is based underlies several non-parametric tests.
42. Maynard Smith, J. The detection and measurement of recombination from sequence data. *Genetics* **153**, 1021–1027 (1999).
43. Maynard Smith, J. Analysing the mosaic structure of genes. *J. Mol. Evol.* **34**, 126–129 (1992).
44. Takahata, N. Comments on the detection of reciprocal recombination or gene conversion. *Immunogenetics* **39**, 146–149 (1994).
45. Betran, E., Rozas, J., Navarro, A. & Barbadilla, A. 1997 The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics* **146**, 89–99 (1994).
46. Lawrence, J. G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA* **95**, 9413–9417 (1998).
47. Ochman, H. & Moran, N. A. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099 (2001).
48. Posada, D. & Crandall, K. A. Evaluation of methods for detecting recombination from DNA sequences: computer

simulations. *Proc. Natl Acad. Sci. USA* **98**, 13757–13762 (2001).

49. Grassly, N. C. & Holmes, E. C. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**, 239–247 (1997).
50. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
51. Posada, D. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.* **19**, 708–717 (2002).
52. Awadalla, P. & Charlesworth, D. Recombination and selection at *Brassica* self-incompatibility loci. *Genetics* **152**, 413–425 (1999).
53. McVean, G. A., Awadalla, P. & Fearnhead, P. A coalescent approach to detecting and estimating the population recombination rate. *Genetics* **160**, 1231–1241 (2002).
54. Miyashita, N. T., Aguade, M. & Langley, C. H. Linkage disequilibrium in the *white* locus region of *Drosophila melanogaster*. *Genet. Res.* **62**, 101–109 (1993).
55. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **33**, 54–78 (1968).
56. Hey, J. & Wakeley, J. A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846 (1997).
57. Hudson, R. R. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631 (1985).
58. Hudson, R. R. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**, 245–250 (1987).
Describes a method of moments estimator of recombination that is widely used by population geneticists.
59. Hudson, R. R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).
60. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
61. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27–43 (1982).
62. Kingman, J. F. C. The coalescent. *Stochastic Process. Appl.* **13**, 235–248 (1982).
63. Hudson, R. R. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44 (1990).
An excellent review of the coalescent, its applications, and some source code in C language for carrying out simulations.
64. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401 (2000).
65. Fearnhead, P. & Donnelly, P. J. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
66. Wall, J. D. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17**, 156–163 (2000).
67. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502 (1996).
68. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
Critically assesses the various approaches used to estimate LD. The authors examined how various demographic models affect the relationship of LD with physical distance.
69. Machado, C. A. & Ayala, F. J. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of *Trypanosoma cruzi*. *Proc. Natl Acad. Sci. USA* **98**, 7396–7401 (2001).
70. Spratt, B. G., Hanage, W. P. & Feil, E. J. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr. Opin. Microbiol.* **4**, 602–606 (2001).
71. Maynard Smith, J. & Smith, N. H., O'Rourke, M. & Spratt, B. G. How clonal are bacteria? *Proc. Natl Acad. Sci. USA* **90**, 4383–4388 (1993).
72. Covacci, A., Falkow, S., Berg, D. E. & Rappuoli, R. Did the inheritance of a pathogenicity island modify the virulence of *Helicobacter pylori*? *Trends Microbiol.* **5**, 205–208 (1997).
73. Enright, M. C. *et al.* The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc. Natl Acad. Sci. USA* **99**, 7687–7692 (2002).
74. Caugant, D. A. *et al.* Intercontinental spread of a genetically distinctive complex of clones of *Neisseria meningitidis* causing epidemic disease. *Proc. Natl Acad. Sci. USA* **83**, 4927–4931 (1986).
75. Haubold, B., Travisano, M., Rainey, P. B. & Hudson, R. R. Detecting linkage disequilibrium in bacterial populations. *Genetics* **150**, 1341–1348 (1998).
76. Feil, E., Carpenter, G. & Spratt, B. G. Electrophoretic variation in adenylate kinase of *Neisseria meningitidis* is due to inter- and intraspecies recombination. *Proc. Natl Acad. Sci. USA* **92**, 10535–10539 (1995).
77. McGee, L., Koornhof, H. J. & Caugant, D. A. Epidemic spread of subgroup III of *Neisseria meningitidis* serogroup A to South Africa in 1996. *Clin. Infect. Dis.* **27**, 1214–1220 (1998).
78. Souza, V., Rocha, M., Valera, A. & Eguiarte, L. E. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. *Appl. Environ. Microbiol.* **65**, 3373–3385 (1999).
79. Grigg, M. E., Bonnefoy, S., Hehl, A. B., Suzuki, Y. & Boothroyd, J. C. Success and virulence in *Toxoplasma* as the result of sexual recombination between two distinct ancestries. *Science* **294**, 161–165 (2001).
80. Gupta, S. *et al.* The maintenance of strain structure in populations of recombining infectious agents. *Nature Med.* **2**, 437–442 (1996).
81. Hastings, I. M. & Wedgewood-Oppenheim, B. Sex, strains and virulence. *Parasitol. Today* **13**, 375–383 (1997).
82. Hastings, I. M. & Mackinnon, M. J. The emergence of drug-resistant malaria. *Parasitology* **117**, 411–417 (1998).
83. Dye, C. & Williams, B. G. Multigenic drug resistance among inbred malaria parasites. *Proc. R. Soc. Lond. B Biol. Sci.* **264**, 61–67 (1997).
84. McCulloch, R. & Barry, J. D. A role for RAD51 and homologous recombination in *Trypanosoma brucei* antigenic variation. *Genes Dev.* **13**, 2875–2888 (1999).
85. Payne, D. Spread of chloroquine resistance in *Plasmodium falciparum*. *Parasitol. Today* **3**, 241–246 (1987).
86. Wootton, J. C. *et al.* Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* **418**, 320–323 (2002).
Describes the microsatellite diversity along chromosome 3 of *P. falciparum*. Extensive LD is described around the *pfcrt* locus, which indicates that recent selection might be acting at this locus.
87. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189 (2002).
88. Conway, D. J. *et al.* A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. *Nature Med.* **6**, 689–692 (2002).
89. Bernstein, H., Byerly, H. C., Hopf, F. A. & Michod, R. E. Genetic damage, mutation, and the evolution of sex. *Science* **229**, 1277–1281 (1985).
90. Cavalier-Smith, T. *The Evolution of Genome Size* (John Wiley & Sons, New York, 1985).
91. Cavalier-Smith, T. Origins of the machinery of recombination and sex. *Heredity* **88**, 125–141 (2002).
92. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2**, 360–369 (2001).
93. Kilbourne, E. D. Molecular epidemiology — influenza as archetype. *Harvey Lect.* **73**, 225–228 (1979).
94. Basler, C. F. *et al.* Sequence of the 1918 pandemic influenza virus nonstructural gene (NS) segment and characterization of recombinant viruses bearing the 1918 NS genes. *Proc. Natl Acad. Sci. USA* **98**, 2746–2751 (2001).
95. Iturriza-Gomara, M., Isherwood, B., Desselberger, U. & Gray, J. Reassortment *in vivo*. Driving force for diversity of human rotavirus strains isolated in the United Kingdom between 1995 and 1999. *J. Virol.* **75**, 3696–3705 (2001).
96. Moumen, A., Polomack, L., Roques, B., Buc, H. & Negroni, M. The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucleic Acids Res.* **15**, 3814–3821 (2001).
97. Sniegowski, P. D., Gerrish, P. J., Johnson, T. & Shaver, A. The evolution of mutation rates: separating causes from consequences. *Bioessays* **22**, 1057–1066 (2000).
98. Bell, G. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality* (Univ. California Press, Berkeley, California, 1982).
99. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **12**, 1351–1353 (1999).
100. Ethier, S. N. & Griffiths, R. C. On the two-locus sampling distribution. *J. Math. Biol.* **29**, 131–159 (1990).
101. Golding, G. B. The sampling distribution of linkage disequilibrium. *Genetics* **108**, 257–274 (1984).
102. Schwartz, M. & Vissing, J. The paternal inheritance of mitochondrial DNA. *N. Engl. J. Med.* **347**, 576–580 (2002).
103. Wong, L.-J. C., Wong, H. & Liu, A. Intergenerational transmission of pathogenic heteroplasmic mitochondrial DNA. *Genet. Med.* **4**, 78–83 (2002).
104. Salminen, M. O., Carr, J. K., Burke, D. S. & McCutchan, F. E. Identification of breakpoints in intergenotypic recombinants of HIV type-1 by bootscanning. *Aids Res. Hum. Retroviruses* **11**, 1423–1425 (1995).
105. Weiller, G. F. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol.* **15**, 326–335 (1998).
106. Grassly, N. C. & Holmes, E. C. A likelihood method for the detection of selection and recombination using sequence data. *Mol. Biol. Evol.* **14**, 239–247 (1997).
107. Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563 (2000).
108. Hein, J. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.* **98**, 185–200 (1990).
109. Jakobsen, I. B. & Easteal, S. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.* **12**, 291–295 (1996).
110. Sneath, P. H. A. The effect of evenly spaced constant sites on the distribution of the random division of a molecular sequence. *Bioinformatics* **14**, 608–616 (1998).
111. Kuhner, M. K., Lawlor, D. A., Ennis, P. D. & Parham, P. Gene conversion in the evolution of the human and chimpanzee MHC class I loci. *Tissue Antigens* **38**, 152–164 (1991).
112. Robinson, D. A. *et al.* Molecular characterization of a globally distributed lineage of serotype 12F *Streptococcus pneumoniae* causing invasive disease. *J. Infect. Dis.* **179**, 414–422 (1999).
113. Kalia, A., Enright, M. C., Spratt, B. G. & Bessen, D. E. Directional gene movement from human-pathogenic to commensal-like streptococci. *Infect. Immun.* **69**, 4858–4869 (2001).
114. Kocken, C. H. *et al.* Molecular characterisation of *Plasmodium reichenowi* apical membrane antigen-1 (AMA-1), comparison with *P. falciparum* AMA-1, and antibody-mediated inhibition of red cell invasion. *Mol. Biochem. Parasitol.* **109**, 147–156 (2000).
115. de Stricker, K., Vuust, J., Jepsen, S., Oeuvray, C. & Theisen, M. Conservation and heterogeneity of the glutamate-rich protein (GLURP) among field isolates and laboratory lines of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **111**, 123–130 (2000).
116. Escalante, A. A. *et al.* Polymorphism in the gene encoding the Pfs48/45 antigen of *Plasmodium falciparum*. XI. Asembo Bay Cohort Project. *Mol. Biochem. Parasitol.* **119**, 17–22 (2002).
117. Figtree, M. *et al.* *Plasmodium vivax* synonymous substitution frequencies, evolution and population structure deduced from diversity in *AMA* 1 and *MSP* 1 genes. *Mol. Biochem. Parasitol.* **108**, 53–66 (2000).
118. Rota, P. A. Molecular epidemiology of measles viruses in the United States, 1997–2001. *Emerg. Infect. Dis.* **8**, 902–908 (2002).
119. Liffick, S. L. *et al.* Genetic characterization of contemporary wild-type measles viruses from Vietnam and the People's Republic of China: identification of two genotypes within clade H. *Virus Res.* **77**, 81–87 (2001).
120. Woelk, C. H., Li, J., Holmes, E. C. & Brown, D. W. G. Immune and artificial selection in the hemagglutinin (h) glycoprotein of measles virus. *J. Gen. Virol.* **82**, 2463–2474 (2001).
121. Worobey, M. & Holmes, E. C. Homologous recombination in GB virus C/hepatitis G virus. *Mol. Biol. Evol.* **18**, 254–261 (2001).
122. Suerbaum, S. *et al.* Free recombination within *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA* **95**, 12619–12624 (1998).
123. Pride, D. T., Meinersmann, R. J. & Blaser, M. J. Allelic variation within *Helicobacter pylori babA* and *babB*. *Infect. Immun.* **69**, 1160–1171 (2001).
124. Kuiken, C., Thakallapali, R., Esklid, A. & de Ronde, A. Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus. *Am. J. Epidemiol.* **152**, 814–822 (2000).
125. Savolainen, C., Blomqvist, S., Mulders, M. N. & Hovi, T. Genetic clustering of all 102 human rhinovirus prototype strains: serotype 87 is close to human enterovirus 70. *J. Gen. Virol.* **83**, 333–340 (2002).
126. Tagger, A. *et al.* A case–control study on a novel DNA virus (TT virus) infection and hepatocellular carcinoma. The Brescia HCC Study. *Hepatology* **30**, 294–299 (1999).
127. Schierup, M. H. & Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891 (2000).
The first paper to show how recombination affects the shape of phylogenies and how it can be confounded with other demographic or mutation properties.

Acknowledgements

We thank J. Anderson, D. Haydon, C. Langley, G. McVean, K. Thorton and three anonymous reviewers for discussions and readings of this manuscript. P.A. is funded by a Wellcome Trust Fellowship.

Online links

FURTHER INFORMATION
PlasmoDB: http://www.tigr.org/tdb/e2k1/pfa1
Access to this interactive links box is free online.