Annu. Rev. Genet. 2002. 36:75–97
doi: 10.1146/annurev.genet.36.040202.111115

RECOMBINATION IN EVOLUTIONARY GENOMICS

David Posada[1,2], Keith A. Crandall[3,4], and
Edward C. Holmes[5]

[1]Variagenics Inc. Cambridge, Massachusetts 02139, [2]Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, [3]Department of Integrative Biology, [4]Department of Microbiology and Molecular Biology, Brigham Young University, Provo, Utah 84602, and [5]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom; e-mail: dposada@variagenics.com

Key Words phylogeny, incongruence, maximum likelihood, bioinformatics, linkage disequilibrium

■ **Abstract** Recombination can be a dominant force in shaping genomes and associated phenotypes. To better understand the impact of recombination on genomic evolution, we need to be able to identify recombination in aligned sequences. We review bioinformatic approaches for detecting recombination and measuring recombination rates. We also examine the impact of recombination on the reconstruction of evolutionary histories and the estimation of population genetic parameters. Finally, we review the role of recombination in the evolutionary history of bacteria, viruses, and human mitochondria. We conclude by highlighting a number of areas for future development of tools to help quantify the role of recombination in genomic evolution.

CONTENTS

0066-4197/02/1215-0075$14.00

INTRODUCTION

The comparative analysis of genome sequence data is transforming evolutionary biology. Not only does genomic analysis allow us to reconstruct phylogenetic patterns and processes with more accuracy than ever before, but it also provides new insights to the fundamental mechanisms of evolutionary change. One such mechanism is recombination. Already, the bioinformatic analysis of genome sequence data has revolutionized our understanding of this central evolutionary process, including its impact on genome structure (104) and on phenotypic variation (146), and its relationship to the study of genetic disease (12). Further, there is now a greater understanding of how recombination confounds our attempts to infer phylogenetic history and other key evolutionary parameters, and that lateral gene transfer has been a common occurrence in the evolutionary history of many species, so that taxa cannot always be related by single phylogenetic tree (73).

Given the central importance of recombination in evolutionary biology, it is crucial that we have bioinformatic tools that are able to accurately detect its occurrence and understand how it affects the inference of phylogenetic relationships. Our review covers current tools available for detecting recombination and discusses the impact of recombination on phylogeny estimation. For this purpose, it is important to distinguish between homologous recombination, which affects related gene sequences, from nonhomologous recombination, which does not. Although both conform to a broad definition of recombination—an evolutionary event that has as a consequence the horizontal exchange of genetic material—our discussion of the phylogenetic impact and detection of recombination implicitly assumes that we are dealing with homologous sequences. We also consider recombination in both prokaryotes and eukaryotes, where traditionally the process of recombination is thought to act differently. This distinction is significant because the concept of recombination prevalent in evolutionary genetics is based on meiosis in eukaryotic organisms, where recombination is a complex molecular process by which a fragment of DNA is reciprocally exchanged between homologous chromosomes. On the other hand, prokaryotes provide several possible pathways of recombination—conjugation, transformation, and transduction—that are more accurately denominated lateral gene transfer or gene conversion, as they involve the nonreciprocal replacement or addition of sequences rather than their exchange, involving either homologous or nonhomologous sequences (although gene conversion is also a frequent process in eukaryotic multigene families). It is also important to distinguish between recombinational events that occur between different genes (intergenic recombination) or between alleles of the same gene (intragenic recombination). Hence, whereas there are many different mechanisms to generate recombinant genomes (in our broadly defined sense), the evolutionary outcomes of recombination are largely the same in whichever system is analyzed. It is the impact of these outcomes that we address here.

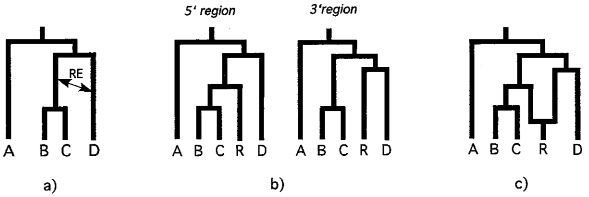

Figure 1 Recombination may generate different phylogenies for different regions of a gene or alignment. (*a*) A recombinational event between the ancestor of B&C and D generates a recombinant R that is present in the sample. (*b*) This recombinant will cluster with the ancestor of B&C in the region 3′ to the recombination breakpoint, whereas in the 5′ region it will cluster with D. (*c*) The fact that there is more than one history underlying the data is often represented as loops or reticulations (and therefore the term "reticulate evolution" is frequently used).

A central theme of our review is the impact of recombination on phylogenetic inference. The reconstruction of phylogenies has been the subject of considerable and often intemperate debate for many years, and more recently, the accumulation of molecular data has added a new level of interest and analytical power (47). Although there are many examples of the myriad uses of molecular phylogenies (40), most of these applications rely on an accurate estimation of the phylogenies themselves. Traditional methods of phylogeny estimation, such as maximum parsimony (MP), minimum evolution (ME), or maximum likelihood (ML) [see (125)], assume that only one evolutionary history underlies the sample under study. However, this assumption is violated by the occurrence of recombination, which can lead to samples with several underlying phylogenies, in which case it is more accurate to describe relationships in terms of reticulate evolution (Figure 1). Indeed, it is important to remember that a bifurcating tree is a hypothesis about how taxa are related, not a truism. In those studies that have explored the possibility of recombination, it has had a significant impact in our understanding of the history of gene genealogies and arguments based on these phylogenies (14, 16, 38, 48, 49, 97, 108, 110, 143, 147). Technological advances have allowed for even larger regions of DNA to be sequenced, thereby increasing the chances for recombination to have occurred in the sample under study. A clear understanding of how we can detect and estimate the rate at which recombination occurs is therefore essential.

THE DETECTION OF RECOMBINATION

Given the importance of recombination in the evolutionary analysis of sequence data and as a potentially dominant force in the rearrangement of genetic variation, it is essential to be able to identify whether a given set of sequences has been affected by recombination, to identify the boundaries of the recombinational units, and to evaluate the impact of recombination on our ability to reconstruct evolutionary histories and estimate population genetic parameters. In the following sections we summarize different methods for detecting the presence of recombination and their relative performance. By detecting recombination we mean just to answer the question of whether recombination has occurred or not. How to measure the amount of recombination is discussed in the next section.

Statistical Methods for Detecting Recombination

During the past 15 years numerous methods have been developed to test for the occurrence of recombination, to identify the parental and recombinant individuals, and to determine the location of the recombinational break-points. These techniques differ greatly in approach and applicability, but may be (tentatively) classified into five nonexclusive general categories: similarity, distance, phylogenetic, compatibility, and nucleotide substitution distribution methods. Here we provide a brief overview of current methods within each of these categories. For a more detailed review of these methods see Crandall & Templeton (11), or the supplementary material in Posada & Crandall (101). David Robertson (Department of Zoology, University of Oxford) also offers a web site with links to the implementations of these methods at http://grinch.zoo.ox.ac.uk/RAP_links.html.

(*a*) *Similarity Methods* These methods infer gene conversion when synonymous substitutions at variable regions exceed those at conserved regions (85, 94). However, they have not been used extensively, in part because they are most useful for detecting gene conversion in multigene families and can be applied only to coding regions.

(*b*) *Distance Methods* Several methods look for inversions of distance patterns among the sequences (138). In general, they use a sliding window approach and the estimation of some statistic based on the genetic distances among the sequences. Because the phylogeny does not need to be known, these methods are highly computationally efficient.

(*c*) *Phylogenetic Methods* Other methods infer recombination when phylogenies from different parts of the genome result in discordant topologies or when orthologous genes from different species are clustered. When comparisons of adjacent sequences yield topological incongruence, there is good reason to suspect the involvement of recombination (3, 29, 35, 41, 42, 49, 55, 56, 64, 75, 79–82, 107–109, 116, 117). Such phylogeny-based methods are currently the most common in use to detect recombination.

(*d*) *Compatibility Methods* Compatibility methods test for partition phylogenetic incongruence on a site-by-site basis. These methods do not require a phylogeny of the sequences under study (14, 17, 59, 60, 120).

(*e*) *Substitution Distribution* This family of methods include strategies that examine sequences for a significant clustering of substitutions or fit to an expected statistical distribution (5, 11, 16, 32, 57, 77, 83, 111, 112, 118, 119, 121, 127, 131, 142).

Performance of Recombination Detection Methods

The performance of several methods for detecting recombination has been evaluated through the analysis of simulated (7, 76, 101, 141) and empirical data (15, 99). These studies have focused on the detection of the presence of recombination rather than on the identification of parentals and recombinant individuals, or on the location of the specific break-points, and hence give an incomplete picture of our ability to accurately detect every aspect of recombination.

Recombination detection methods differ in performance depending on the amount of recombination, the genetic diversity of the data, and the degree of rate variation among sites. Most methods are efficient, showing more power with increasing recombination rates, although some methods are more efficient than others. Most methods also show better performance at higher levels of divergence, most likely because of an increase in the amount of signal for recombination present in the data. For the majority of methods, a minimum nucleotide diversity of 5% seems necessary to obtain substantial power, and several recombination events are needed to infer the presence of recombination. Recombination is also difficult to detect when the phylogeny has long terminal, and short internal, branches (141). Rate variation among sites (145) can also be confounded with recombination, and in some cases it leads to false positives (99, 101, 114, 142). Perhaps the most interesting consensus result from these studies is that methods that use the substitution patterns or incompatibility among sites seem to be more powerful than methods based on phylogenetic incongruence. This might be partially explained by the fact that, in general, phylogenetic methods can only detect recombination events that change the topology of the tree, and at high recombination rates there should be many such events.

Note also that there are two different contexts in which we may wish to detect recombination: rare, sporadic recombination or frequent, repeated recombination (76). Not surprisingly, most methods have trouble detecting rare recombinational events, especially when sequence divergence is low. Indeed, recent events should be more easily identifiable than older events, as the latter may be obscured by subsequent mutation. On the other hand, when recombination rates are extremely high, leading to situations close to linkage equilibrium, we would expect substitution methods to have difficulty in identifying site patterns (76), although this is not what is observed with real data sets (99). Indeed, we are interested in maximizing the chances of detecting recombination while minimizing the chances of false positives. In order to do so, we need to take into account levels of variation. For example, for data sets with very low divergence (1%), the homoplasy test (77) appears to be a reasonable method, as long as there is little among-site rate variation. For higher levels of divergence the homoplasy test is not adequate, and methods like the modified maximum chi-square (101, 141), GENECONV (113) or RDP (75) are more powerful. However, perhaps the key conclusion from simulation and analytical studies is that one should not rely on a single method to detect recombination (101, 141).

The Proportion of Undetectable Recombination

In general, our ability to detect recombination depends on the amount of genetic variation in the population. If recombination occurs between two identical strands of DNA, then this event is undetectable. Therefore, estimates of recombination will always be underestimates due to our inability to detect recombination between identical or nearly identical sequences. Hudson & Kaplan (54) studied the theoretical sampling distribution of the number of recombination events that have occurred during the history of a sample of DNA sequences. Through computer simulation they compared the known number of recombination events with the number inferred by a detection technique based on parsimony (four-gamete test), and found that only a small fraction of known recombination events were detected. In this context, recombination events can be divided in two categories (122): those that do not result in any observable effect on the DNA sequences and hence are undetectable with any analytical method, and those that do affect the DNA sequence and here are potentially detectable. The first category includes recombination events between identical sequences (Figure 2*a*), between sequences that differ at a single site (Figure 2*b*,*c*) and between sequences that differ at several sites with the crossover point flanking the segregating sites (Figure 2*d*,*e*). The number of nucleotide differences, *d*, between a random pair of DNA sequences is related to the quantity $\theta = 4N_e\mu$, where N_e is the effective population size and μ the mutation rate. Stephens (122) has shown that even for relatively high values of θ, a substantial fraction of the recombination events cannot be detected, even with $d \geq 2$. Undetected recombinant events occur mainly because of inefficiency in the detection, and less commonly because of redundant recombination events. Quite clearly, attempts to estimate recombination rates should take into account the fraction of undetectable events (122). Similarly, and from a phylogenetic point of view, recombination events can be classified with respect to their effect on tree topology, as (*a*) events that do not change the branch lengths, (*b*) events that do change the branch lengths, but do not change the tree topology, and (*c*) events that change the tree topology (Figure 3). Wiuf et al. (141) give theoretical expectations for the tree types of events.

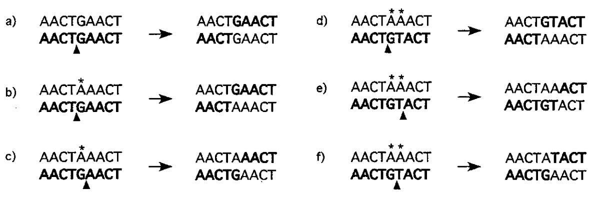

Figure 2 Different types of recombination events depicted in substitution patterns. Events *a*–*e* belong to the category of undetectable events. Event *f* is a detectable event and so belongs to the category of detectable events. Sites with the * symbol are variable sites. A triangle indicates the location of the recombination break-point.

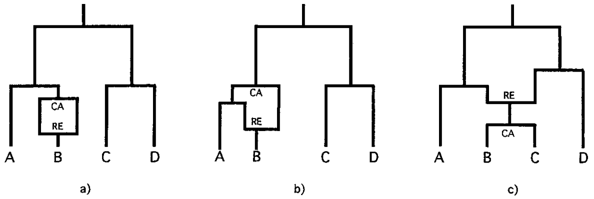

Figure 3 Different types of recombination events and how they affect tree topology and branch lengths. CA indicates a coalescent event, while RE indicates a recombination event. (*a*) No change in topology nor in branch lengths. (*b*) No change in topology but a change in branch lengths. (*c*) A change both in topology and branch lengths. Events of type (*a*) and (*b*) can occur in samples of any size, but events of type (*c*) can only occur in samples of size ≥4. Most events in samples of large size are of type (*c*) (141).

Recombination in Polymerase Chain Reaction

Once recombination is detected, questions of the validity of the recombination event in vivo remain because recombination can also be produced in vitro during the polymerase chain reaction (PCR) used to amplify the desired region of DNA (6, 87, 93). Recombination is a particular concern when attempting to amplify long products by PCR, as has been demonstrated with HIV-1 sequences. Phylogenetic analyses offer a way to distinguish between PCR-induced recombination events and actual in vivo recombination events. If mutational events are mapped along the branches of the reconstructed histories, recent and historical recombination events can be differentiated by the number of accumulated mutations after the recombination event. In PCR recombination, we do not expect to accumulate additional substitutions, so that if such additional substitutions are present along the recombinant branch, then the inferred event most likely occurred in vivo.

ESTIMATING RECOMBINATION RATES

Recombination can play a dominant role in the generation of novel genetic variants through the rearrangement of existing genetic variation generated through mutation. Recombination also plays a role in the dissipation of linkage disequilibrium. Hence, when coupled with selection, recombination can be a key evolutionary force (26, 46). To understand the role of recombination in the generation of genetic diversity relative to the role of mutation we need to be able to accurately estimate recombination rates. Whereas in the previous section we focused on detection of the presence of recombination, here we focus on the rate of recombination. Indeed, recombination rate estimators can be used to build tests for the presence of recombination (7).

The population recombination parameter is defined as $\rho = 4N_e r$, where N_e is the effective population size and r is the per-locus (or per-site) recombination rate per generation. The population mutation parameter (genetic diversity) can be similarly defined as $\theta = 4N_e\mu$, where μ is the per-locus (or per-site) mutation rate per generation. If we can accurately estimate both ρ and θ, we can then define the relative rate of recombination compared to mutation as

$$\varepsilon = \frac{\rho}{\theta} = \frac{4N_e r}{4N_e \mu} = \frac{r}{\mu}. \qquad 1.$$

On the rare occasion that this quantity has been estimated from nucleotide sequence data, it has provided keen insights into the population dynamics of the organism under study (e.g., 23, 103).

Recombination Rate Estimators

As there are a variety of methods to estimate θ (19, 28, 31, 66, 126, 132, 136), there are also a number of methods to estimate recombination rates in populations. Like recent methods for estimating θ, most approaches for estimating ρ are based on neutral coalescent theory (62) [reviewed in (52)] with recombination (37, 50). The estimators of recombination generally take one of two approaches, either quantifying recombination as a summary statistic or estimating recombination rates by considering all the data. Hudson (51) took the former approach to derive an estimator of ρ based on the observed variance of the number of pairwise differences. The expected variance in pairwise differences decreases with decreasing amounts of linkage disequilibrium between segregating sites as a result of increasing recombination. Therefore, this observed variance in pairwise differences is a measure of the amount of linkage disequilibrium and hence also of the recombination rate (134). Wakeley (133) improved this estimator by considering only nonidentical pairs of sequences, which reduces the bias and standard error on this estimator of recombination rate.

The second approach is to use a maximum likelihood framework to provide a joint estimate of mutation and recombination rate that uses the maximal information in the sample, rather than a summary statistic. The first such estimator was developed by Griffiths & Marjoram (36); they used a coalescent process with recombination resulting in a genealogy with reticulations that they termed an "ancestral recombination graph." Kuhner et al. (67) also used a "recombinant genealogy" to co-estimate the recombination rate and mutation rate (our ε; Equation 1) using a Metropolis-Hastings sampling strategy across genealogies. Fearnhead & Donnelly (24) similarly present a full-likelihood–based approach to the joint estimate of recombination and mutation rates. Their method develops an improved importance sampling scheme, which should result in more accurate estimates. The advantage of these methods is that they use all the data to estimate recombination rate instead of a summary statistic. However, they accomplish this at the expense of computational efficiency. A compromise solution was proposed by Hey & Wakeley (45) whose approach averages likelihood estimates of recombination rate for subsets of sequences. An alternative approach proposed by Wall (134) uses the number of distinct haplotypes to estimate ρ by using maximum likelihood on summary statistics. A Bayesian approach to estimating recombination rate was recently proposed by Falush et al. (23), but the statistical properties of this method are unexplored. An alternative Bayesian approach was suggested by Nielsen (91), and the statistical foundation of this approach is much better laid out. This method is similar to that of Kuhner et al. except that a Bayesian approach is used in parameter estimation instead of importance sampling. Hudson (53) recently proposed an alternative approach that considers polymorphic sites in pairs and then utilizes likelihood methods appropriate for analyzing a pair of polymorphic sites. This composite-likelihood estimator has the advantage of being more computationally efficient relative to the full-likelihood methods but without summarizing the data in a single statistic. McVean et al. (83) extended this to accommodate different models of evolution (including, importantly, rate variation) and to relax the infinite-sites assumption (typically violated by many empirical data sets). All of these approaches assume constant population sizes, independence of sites, neutral evolution, and an infinite-sites model of evolution (with the exception of McVean et al.'s method). However, they differ considerably in terms of the required population sampling, level of nucleotide polymorphism, and number and type of nucleotide positions surveyed.

Performance of Estimators of Recombination Rate

At least two studies have extensively compared different estimators of recombination rate. The first compared ten estimators and found that their relative performance depended greatly on the amount of genetic diversity (θ), with most methods performing poorly at low levels of genetic diversity (134). The best performing estimator in these simulations was that of Kuhner et al., which had the smallest mean squared error, the greatest proportion of estimates within a factor of two of the actual value, and the second smallest bias (134). The second study compared the relative performance of the full-maximum–likelihood methods. In this simulation study, the authors distinguish between two possible comparisons. One can either compare how accurately the methods approximate the likelihood surface or the properties of the methods' ability to estimate ρ and θ (24). While the former study based comparisons on the second criterion, these authors argue that the first is more fundamental and therefore report results from this approach. They show that their new sampling method is up to four orders of magnitude more efficient than the previous method of Griffiths & Marjoram (36). In addition, they showed their approach outperformed Kuhner et al.'s method and also that this method often gave misleading results. The discrepancy between these results and those obtained by Wall and Kuhner et al. are presumably due to the different criterion of assessment (likelihood surface instead of the parameter estimates themselves) and the difference in the relative amount of recombination to mutation (Kuhner et al. simulated data with mutation rates much higher than recombination rates, whereas Fearnhead & Donnelly simulated under the opposite conditions) (24). Clearly, the comparisons of these methods have just begun. As new methods emerge from our better understanding of existing methods, further research is needed to discern how robust such estimators are to violations of the standard coalescent assumptions (135).

RECOMBINATION AND PHYLOGENETIC INFERENCE

Phylogenetic studies typically ignore the potential occurrence of recombination, which may produce sequence regions with different evolutionary histories. An accurate history of such mosaic sequences cannot be estimated by traditional phylogenetic methods that assume a single nonreticulate tree. If recombination is present and we have ignored it, can we expect the inferred phylogeny to represent any of the underlying evolutionary histories? Furthermore, what happens if we then use these trees to estimate relevant evolutionary parameters? Partial answers to these questions have been only recently investigated, and in this section we outline our current understanding of the impact of recombination in phylogenetic studies.

The Effect of Recombination on Phylogeny Estimation

Recombination has long been recognized as a serious confounding factor for phylogeny estimation. However, only a few studies have explicitly addressed this question. Wiens (140) carried out a simulation study to explore the effect of combining data sets with different phylogenetic histories. This problem is identical to the problem of recombination, only that when we combine data sets we have already defined the potential partitions of the data. Wiens explicitly investigated the effect of combining genes generated under different genealogies on the estimation of the true "species tree." The main conclusion was that a combined analysis provides a poor estimate of the species tree in areas where the gene genealogies are very different, but an improved estimate in regions where the gene genealogies agree. Wiens also provides a simple strategy to deal with such situations, consisting of (*a*) defining the data partitions (e.g., by gene), (*b*) performing a separate analysis on each partition, and (*c*) undertaking a combined analysis, with caution directed toward nodes not supported in the analyses of separate partitions (i.e., step *b*).

However, in many cases we only have a single data set with no obvious partitions and the question then becomes what happens when recombination has occurred, but is ignored? In such a case can we expect the inferred phylogeny to represent any of the underlying evolutionary histories? Posada & Crandall (102) examined this question by applying traditional phylogenetic reconstruction methods to mosaic sequence alignments. Their results suggest that the effect of recombination on phylogeny estimation is dependent upon the relatedness of the sequences involved in the recombination event and on the relative size of the regions with different phylogenetic histories. When recombination occurred between closely related taxa, or when recombination was ancient, one of the histories underlying the data was inferred. In these cases, the phylogeny under which the majority of sites were evolved was generally recovered. On the other hand, when recombination occurred recently among divergent taxa and the recombinational break-point divided the alignment in two regions of similar length, a phylogeny that was very different from any of the true phylogenies underlying the data was inferred. Hence, recombination can be very misleading, resulting in the inference of wrong topologies, but only in some circumstances. More extensive simulations are needed to determine the generality of these conclusions.

Estimating Parameters from Recombinant Trees

While recombination can have a major impact on phylogenetic trees, the tree is seldom the endpoint of a phylogenetic analysis. Indeed, trees are now used to infer many relevant evolutionary parameters and to test different evolutionary hypotheses. Schierup & Hein (114) characterized some of the consequences of ignoring recombination when using phylogenies to make demographical, chronological, or substitutional inferences. Long terminal branches appear in a more star-shaped phylogeny, which suggests apparent exponential growth when the population size is actually constant. Further, parallel mutations are postulated to fit the data to a single tree and the extent of rate heterogeneity among-sites is wrongly inferred. Crucially, however, recombination affects different phylogenetic methods in different ways. While distance methods underestimate the time to the most common ancestor, maximum likelihood leads to an overestimate of the total number of mutations. The amount of recombination needed for these effects to be evident is not high, and such effects were found with just 100 bp in *Drosophila*, or 2000 bp in humans, although obviously the recombination rate varies extensively over the genome.

Recombination and the Molecular Clock

Ignoring recombination may lead to the false rejection of the molecular clock if phylogenetic methods like the likelihood ratio test (27) are used (115). To appropriately test the clock hypothesis in the presence of recombination, we need to use a test that is independent of tree topology. Muse & Weir (89) proposed a triplet likelihood ratio test to test for equality of evolutionary rates for two species at a time using a third species as an outgroup. Posada (98) has shown that this can be used as a conservative test for recombinant sequences if an outgroup is selected that did not recombine with the ingroup.

REPRESENTING RETICULATE EVOLUTION

The presence of recombination in an evolutionary history presents a significant problem for the representation of that history. A typical representation of an evolutionary history consists of a bifurcating evolutionary tree. However, with recombination, the true underlying history is reticulate in nature. Therefore, a bifurcating tree is, at best, only a partial representation of the actual evolutionary history. To better represent the actual reticulate evolutionary history, researchers have developed network approaches for estimating phylogenetic trees. Similar to standard tree estimation approaches, network approaches employ a variety of optimality criteria, including parsimony, distance, and likelihood approaches. Many combine parsimony, distance, and/or likelihood approaches into a single method. One of the first methods developed for comparing sequences in a network fashion was statistical geometry (18). This approach considers quartet combinations of nucleotide sequences and then develops a geometric configuration to represent the combination of quartets. Statistical parsimony (129), implemented in the software package TCS (10), makes minimum pairwise connections among sequence variants up to a point determined by the calculation of a probability of parsimonious connections. Netting (30) is an approach that represents all the most parsimonious trees in a single network by presenting homoplasies as networked connections in different dimensions. Molecular variance parsimony (20) takes into account haplotype frequencies and their geographic distributions to estimate network relationships. Split decomposition (4) takes sequence characters and divides them into partitions of mutually exclusive sets and then compares these splits across characters. When the splits are incompatible, loops are formed in the graphical representation of genealogical relationships. Finally, a likelihood network procedure (123, 124) allows for a directed graphical model (where nodes are stochastic variables and branches indicate correlation between these variables) to represent the evolutionary history of sequences along a network. These methods and their theoretical advantages over standard bifurcating approaches for the representation of gene genealogies have been recently reviewed in detail (100). However, much work remains in terms of testing the accuracy of these methods in reconstructing evolutionary histories and their relative performances.

THE IMPACT OF RECOMBINATION: EMPIRICAL EXAMPLES

The detection and estimation of recombination has led to major biological insights in a variety of cases. These studies have been particularly important in microbiology where the application of the bioinformatic tools described in this review to the growing data base of gene sequences has radically changed our perspective on how frequently recombination occurs in both bacteria and viruses. Such findings have wide-ranging implications, from the successful reconstruction of evolutionary and epidemiology history, to preventing the development of drug resistance and the evolution of virulence. Moreover, the high levels of genetic variation in many microbial species, particularly RNA viruses, also mean that they constitute ideal model organisms to assess the reliability of different estimators of the presence and rate of recombination.

Recombination in Bacteria

Because bacteria reproduce by binary fission, it was generally assumed that they evolved entirely by asexual mechanisms. Such a view was initially confirmed by studies of *Escherichia coli* using multilocus enzyme electrophoresis (MLEE), which indicated that populations were characterized by high levels of linkage disequilibrium and phylogenetic congruence [reviewed in (78)]. However, the growing availability of genetic data gradually shifted opinion toward the view that recombination could occur, sometimes frequently, among bacterial species other than

E. coli (78). A more radical overhaul came with the availability of large-scale nucleotide sequence data, either in the guise of multilocus sequence typing [MLST (74)] or large regions of bacterial genomes. The bioinformatic analyses of these data indicated that recombination rates in bacterial species can be both extremely high and extremely variable, and that lateral gene transfer can occur among very distantly related species, even between bacteria and eukaryotes (43). This last observation has important implications for reconstructing the evolutionary history of cellular life forms (W. F. Doolittle, personal communication).

MLST data provide a genome-wide subsample of housekeeping genes, but from a very large number of isolates. This makes it possible to measure a range of evolutionary parameters, including recombination frequency. This can be done indirectly by assessing the degree of incongruence between phylogenies of different genes within MLST data sets. Using a maximum likelihood method, in which the differences in log likelihood between the trees estimated for each gene are compared to a null distribution generated using random tree topologies, very high rates of recombination were inferred in *Neisseria meningitidis*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Streptococcus pyogenes*, as tree topologies from different genes were no more similar than random (25, 48). As expected, less incongruence was found in *E. coli* (25, 105), indicating that the rate of recombination is far lower in this species. In some cases, it has also been possible to estimate ε (see above) from MLST data, $\sim$5 for *N. meningitidis* (25), and $\sim$1 for both *N. gonorrheae* (103) and *Helicobacter pylori* (23), that are broadly similar to the recombination rates estimated for human genes. In contrast, the evidence that lateral gene transfer can occur between very distantly related bacterial species has more often been obtained through studies of aberrant $G + C$ content than incongruence (71, 92). Such studies have revealed lateral gene transfer to be a recurrent evolutionary event; in *E. coli*, for example, foreign DNA is estimated to have been imported at a frequency of up to16 Kb per million years (71).

Despite the growing evidence that recombination is a fundamental process in bacterial evolution, the precise mechanisms by which it occurs, and why rates vary so extensively, are less clear. In species that are competent for the uptake of naked DNA from the environment, such as *Neisseria* sp., transformation clearly plays a major role. In many other cases, conjugation, usually involving the transfer of plasmids between bacteria that have come into physical contact, has been described. A role for bacteriophage-mediated transduction is also a frequent suggestion (92). At present, the best evidence for this latter process is that known attachment sites of bacteriophage integrases are frequently found next to imported regions of bacterial DNA, such as the LEE pathogenicity island of *E. coli* (39). However, as bacteriophages from natural environments have received little study, their overall role in bacterial recombination remains uncertain.

Recombination in Viruses

Recombination in DNA and RNA viruses occurs by very different processes. In DNA viruses recombination is likely to take place in the same manner as in other

DNA genomes, i.e., involving an enzyme-mediated breakage-reunion mechanism. This appears to be a relatively common process and can also result in the capture of host genes, which may allow the virus to mimic or block host proteins, thereby assisting in the development of persistent infection (9). Recombination in RNA viruses can occur through either reassortment or "copy-choice" replication. Reassortment describes the process by which viruses with segmented genomes shuffle those segments during mixed infection. This has been described in detail in influenza A virus, where it is associated with the production of novel strains that can evade pre-existing immunity through an "antigenic shift" (137). In copy-choice replication, the viral RNA-dependent RNA polymerase switches from one RNA molecule to another during replication, generating mosaic genomes (90). This process is now thought to occur in a wide variety of positive-sense RNA viruses and retroviruses (143).

As with bacteria, much of the evidence for homologous RNA virus recombination involves the detection of phylogenetic incongruence (143). Such topological mismatching has been documented at a variety of phylogenetic levels, from within single species, to different viral families, in one case between a RNA and a DNA virus (33). However, it is equally clear that RNA viruses vary greatly in their ability to undergo recombination, although, to date, rates of recombination relative to that of mutation have not been estimated through sequence comparisons. For example, hepatitis C virus (HCV) and GBV-C are members of the same viral family (the *Flaviviridae*), yet recombination in HCV is rare, whereas GBV-C appears to recombine at high frequency (144).

It is possible that recombination rate in RNA viruses is a selectively determined trait. Indeed, recombination has been shown to result in direct fitness increases, for example by bringing together different genomes carrying individual drug-resistance mutations in HIV (88), and there is also evidence that reassortment can allow viruses to escape from deleterious mutation accumulation (8). Conversely, it is also possible that recombination rate is simply an outcome of mechanistic constraints set by genome structure and is not a selected entity at all. For example, the highest rates of recombination are found in retroviruses, which carry two copies of their genome within each mature virion, so making recombination easier, and in viruses that can frequently reassort their segmented genomes. Far lower rates of recombination are found in positive-sense, single-strand RNA viruses, where it is easily detected as mosaic sequences, and more so in negative-strand RNA viruses, which have genomes packaged into filamentous ribonucleoprotein (RNP) structures, greatly limiting their ability to recombine. Furthermore, most recombinants, unless they are very similar in sequence, will be deleterious and hence removed by purifying selection. This further emphasizes how estimates of recombination rate based on sequence comparisons are likely to be underestimates of the actual number of recombination events. Determining the basis for the variation in viral recombination rates, and whether they correlate with other biological features such as virulence, is clearly an important area for future study.

Recombination in Human Mitochondrial DNA

One of the most controversial claims in evolutionary genetics in recent years is that, contrary to mainstream opinion, the human mitochondrial genome may undergo recombination. Although signals of genetic exchange have been found in other animal mitochondria (68, 69), the claim that it can occur in human mtDNA has provoked intense debate (21, 44), not least because it has serious implications for our attempts to infer the origin and migration of modern humans.

Two pieces of evidence have been cited in support of mtDNA recombination: that there is excessive homoplasy at polymorphic sites, as revealed in the homoplasy test (22), and that there is a decrease in linkage equilibrium (LD) with physical distance in the mtDNA genome (2). Because recombination creates convergent/parallel evolutionary change, the occurrence of widespread homoplasy at face value represents strong evidence for this process. However, homoplasy can also occur through excessive multiple substitution at single sites, such as those that are known to be hypervariable in mtDNA. Indeed, if the complex pattern of among-site rate heterogeneity is taken into account through the use of the gamma distribution, the evidence for recombination in mtDNA seemingly disappears (142).

The evidence for some degree of linkage equilibrium in the mitochondrial genome has also been questioned. In particular, LD values appear to be highly dependent on the analytical method used, with different estimates obtained with r^2 and $|D'|$ (86), and also the particular data in question. Most notably, a recent survey of 53 complete mitochondrial genome sequences from a variety of geographical regions provided no evidence for any decline in LD with physical distance (58). What causes these conflicting signals is unknown but clearly requires explanation (84). More fundamentally, there is as yet no evidence for incongruence in mtDNA phylogenies. In sum, the evidence for recombination in human mtDNA appears to be weak on current data.

CONCLUSIONS

Several important messages regarding the detection of recombination stem from our review. First, the fact that many recombinational events cannot be detected implies that current methods detect less recombination than is possible. As a consequence, we are consistently underestimating the number of recombination events that have occurred among sequences, and therefore also the overall recombination rate. In addition, we should keep in mind that the power to detect recombination decreases with the degree of genetic variation. Significantly, no single recombination detection strategy seems to perform optimally under all scenarios, so that using a combination of methods currently appears to be the best strategy.

It is also important to distinguish between frequent recombination happening within a population and rare recombination generating mosaic sequences. In the first case, nonrecombinant regions may be very difficult to identify, and therefore genealogies or phylogenies will be difficult to reconstruct. In such cases, network approaches may offer a general idea of the (reticulate) evolutionary history. Indeed, population genetic estimates should be interpreted with care in the light of assumptions made regarding the presence of recombination. When recombination is rare, mosaic and parental sequences can be identified, as well as recombination break-points. In this case, independent phylogenies can be reconstructed for the nonrecombinant regions and then compared to decipher the recombinant history. Alternatively, recombinants can be "peeled off" the tree to reveal the underlying phylogenetic structure (1, 128, 130).

Still outstanding are issues regarding our ability to measure and depict the action of recombination. First, it is not known how robust the estimators for recombination rate are to violations in the key assumptions, particularly that real populations are subdivided and not panmictic, and that natural selection, as well as genetic drift, may have shaped patterns of genetic diversity. The impact of natural selection may be particularly important because most estimators of ε available at present make use of θ, a measure of genetic diversity that assumes exclusively neutral evolution. Whether natural selection can seriously bias estimates of recombination rate is clearly an area that needs urgent attention. Second, although there is a growing appreciation that network methods are a more appropriate representation of evolutionary relationships when recombination is relatively frequent, the accuracy and power of the network methods proposed to date has yet to be tested. Studies using both simulated and real data, such as those used to determine the accuracy of recombination detection models, are clearly a goal for the immediate future.

Recombination clearly plays a significant role in shaping the genetic architecture of organisms. As a case in point, recombination within introns allows for the shuffling of exons and domains (63, 72, 96), and provides a powerful mechanism for the evolution and adaptation of genomes. Over the next few years, we will see an increasing application of "genome shuffling" techniques to rapidly generate "improved" organisms (146). Another of the most exciting promises of the genomics era is the mapping of genes for common human diseases. Here, again, studies of recombination have a key role to play. Whole-genome association studies have been proposed as an indirect strategy to find genes for disease (70, 106) and within these strategies, whole-genome linkage disequilibrium (LD) scans seem to be the most feasible approach (65). Understandably, this has raised considerable interest in revealing the patterns of LD in human populations. Recently, it has been suggested that human haplotypes are structured into discreet blocks of high LD and low diversity, separated by hot spots of recombination (12, 34, 61, 95), although whether this block structure is a general property of human genomes across populations is still to be demonstrated. In any event, the usefulness and completion of a haplotype map (or maps) [see (139)] of the human genome will be dependent upon a good understanding and description of recombination at every level.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (DP, KAC), the National Institutes of Health (DP, KAC), the US-UK Fulbright Commission (KAC), the Wellcome Trust (KAC, ECH), the Royal Society (ECH), and the Brigham Young University International Studies Office (KAC).

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

LITERATURE CITED

1. Antunes A, Templeton AR, Guyomard R, Alexandrino P. 2002. The role of nuclear genes in intraspecific evolutionary inference: genealogy of the *transferrin* gene in the brown trout. *Mol. Biol. Evol.* In press
2. Awadalla P, Eyre-Walker A, Maynard Smith J. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* 286:2524–25
3. Balding DJ, Nichols RA, Hunt DM. 1992. Detecting gene conversion: primate visual pigment genes. *Proc. R. Soc. London Ser. B* 1992:275–80
4. Bandelt H-J, Dress AWM. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* 1:242–52
5. Betrán 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
6. Bradley RD, Hillis DM. 1997. Recombinant DNA sequences generated by PCR amplification. *Mol. Biol. Evol.* 14:592–93
7. Brown CJ, Garner EC, Dunker KA, Joyce P. 2001. The power to detect recombination using the coalescent. *Mol. Biol. Evol.* 18:1421–24
8. Chao L, Tran TT. 1997. The advantage of sex in the RNA virus phi6. *Genetics* 147:953–59
9. Chaston TB, Lidbury BA. 2001. Genetic ‘budget’ of viruses and the cost to the infected host: a theory on the relationship between the genetic capacity of viruses, immune evasion, persistence and disease. *Immunol. Cell Biol.* 79:62–66
10. Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9:1657–60
11. Crandall KA, Templeton AR. 1999. Statistical methods for detecting recombination. In *The Evolution of HIV*, ed. KA Crandall, pp. 153–76. Baltimore, MD: Johns Hopkins Univ. Press
12. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* 29:229–32
13. Deleted in proof
14. Drouin G, Dover GA. 1990. Independent gene evolution in the potato actin gene family demonstrated by phylogenetic procedures for resolving gene conversions and the phylogeny of angisperm actin genes. *J. Mol. Evol.* 31:132–50
15. Drouin G, Prat F, Ell M, Clarke GDP. 1999. Detecting and characterizing gene conversions between multigene family members. *Mol. Biol. Evol.* 16:1639–90
16. DuBose RF, Dykhuizen DE, Hartl DL. 1988. Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 85:7036–40
17. Eastbrook G. 1978. Some concepts for the estimation of evolutionary relationships in systematic biology. *Syst. Bot.* 3:146–58
18. Eigen M, Winkler-Oswatitsch R, Dress A. 1988. Statistical geometry in sequence space: a method of quantitative sequence analysis. *Proc. Natl. Acad. Sci. USA* 85: 5917
19. Ewens WJ. 1979. *Mathematical Population Genetics*. Berlin: Springer-Verlag. 325 pp.
20. Excoffier L, Smouse PE. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a

species: molecular variance parsimony. *Genetics* 136:343–59

21. Eyre-Walker A, Awadalla P. 2001. Does human mtDNA recombine? *J. Mol. Evol.* 53:430–35
22. Eyre-Walker A, Smith NH, Maynard Smith J. 1999. How clonal are human mitochondria? *Proc. R. Soc. London Ser. B* 266:477–83
23. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, et al. 2001. Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. *Proc. Natl. Acad. Sci. USA* 98:15056–61
24. Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. *Genetics* 159:1299–318
25. Feil EJ, Holmes EC, Bessen DE, Chan M-S, Day NPJ, et al. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. USA* 98:182–87
26. Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* 78:737–56
27. Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–76
28. Felsenstein J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res. Cambridge* 59: 139–47
29. Fitch DHA, Goodman M. 1991. Phylogenetic scanning: a computer assisted algorithm for mapping gene conversions and other recombinational events. *CABIOS* 7:207–15
30. Fitch WM. 1997. Networks and viral evolution. *J. Mol. Evol.* 44:S65–S75
31. Fu Y-X. 1994. A phylogenetic estimator of effective population size or mutation rate. *Genetics* 136:685–92
32. Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16:573–82
33. Gibbs MJ, Weiller GF. 1999. Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. *Proc. Natl. Acad. Sci. USA* 96:8022–27
34. Goldstein DB. 2001. Islands of linkage disequilibrium. *Nat. Genet.* 29:109–11
35. Grassly NC, Holmes EC. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* 14:239–47
36. Griffiths RC, Marjoram P. 1996. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* 3:479–502
37. Griffiths RC, Tavare S. 1994. Ancestral inference in population genetics. *Stat. Sci.* 9:307–19
38. Guttman DS, Dykhuizen DE. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266:1380–83
39. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* 23:1089–97
40. Harvey PH, Holmes EC, Mooers AO, Nee S. 1994. Inferring evolutionary processes from molecular phylogenies. In *Models in Phylogeny Reconstruction*, ed. RW Scotland, DJ Siebert, DM Williams, pp. 313–33. Oxford: Clarendon
41. Hein J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.* 98:185–200
42. Hein J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.* 36:396–405
43. Heinemann JA, Sprague GFJ. 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* 340:205–9
44. Hey J. 2000. Human mitochondrial DNA

recombination: Can it be true? *Trends Ecol. Evol.* 15:181–82

45. Hey J, Wakeley J. 1997. A coalescent estimator of the population recombination rate. *Genetics* 145:833–46
46. Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* 8:269–94
47. Hillis DM, Huelsenbeck JP, Cunningham CW. 1994. Application and accuracy of molecular phylogenies. *Science* 264:671–77
48. Holmes EC, Urwin R, Maiden MCJ. 1999. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol. Biol. Evol.* 16:741–49
49. Holmes EC, Worobey M, Rambaut A. 1999. Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.* 16:405–9
50. Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23:183–201
51. Hudson RR. 1987. Estimating the recombination parameter of a finite population model without selection. *Genet. Res. Cambridge* 50:245–50
52. Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxford Surveys Evol. Biol.* 7:1–44
53. Hudson RR. 2001. Two-locus sampling distributions and their application. *Genetics* 159:1805–17
54. Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111:147–64
55. Husmeier D, Wright F. 2001. Detection of recombination in DNA alignments with hidden Markov models. *J. Comput. Biol.* 8:401–27
56. Husmeier D, Wright F. 2001. Probabilistic divergence measure for detecting interspecies recombination. *Bioinformatics* 17:S123–S31
57. Imanishi T. 1996. DNA polymorphisms shared among different loci of the major histocompatibility complex genes. In *Current Issues in Molecular Evolution*, ed. M Nei, N Takahata, pp. 89–96. Hayama, Jpn.: Inst. Mol. Evol. Genet., Penn. State Univ. and Grad. Sch. Adv. Stud.
58. Ingman M, Kaessmann H, Paabo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708–13
59. Jakobsen IB, Easteal S. 1996. A program for calculating and displaying compatibility matrices as an aid to determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.* 12:291–95
60. Jakobsen IB, Wilson SE, Easteal S. 1997. The partition matrix: exploring variable phylogenetic signals along nucleotide sequences alignments. *Mol. Biol. Evol.* 14: 474–84
61. Jeffreys AJ, Kauppi L, Neummann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibilty complex. *Nat. Genet.* 29:217–22
62. Kingman JFC. 1982. The coalescent. *Stoch. Process. Appl.* 13:235–48
63. Kolkman JA, Stemmer WP. 2001. Directed evolution of proteins by exon shuffling. *Nat. Biotechnol.* 19:423–28
64. Koop BF, Siemieniak D, Slightom JL, Goodman M, Dunbar J, et al. 1989. Tarsius delta- and beta-globin genes: conversions, evolution and systematic implications. *J. Biol. Chem.* 264:68–79
65. Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22:139–44
66. Kuhner MK, Yamato J, Felsenstein J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149:429–34
67. Kuhner MK, Yamato J, Felsenstein J. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* 156:1393–401
68. Ladoukakis ED, Zouros E. 2001.

Direct evidence for homologous recombination in mussel (*Mytilus galloprovincialis*) mitochondrial DNA. *Mol. Biol. Evol.* 18:1168–75

69. Ladoukakis ED, Zouros E. 2001. Recombination in animal mitochondrial DNA: evidence from published sequences. *Mol. Biol. Evol.* 18:2127–31
70. Lander ES. 1996. The new genomics: global views of biology. *Science* 274:536–39
71. Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* 95: 9413–17
72. Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. *Nature* 409:847–49
73. Maddison W. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–36
74. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, et al. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* 95:3140–45
75. Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16:562–63
76. Maynard Smith J. 1999. The detection and measurement of recombination from sequence data. *Genetics* 153:1021–27
77. Maynard Smith J, Smith NH. 1998. Detecting recombination from gene trees. *Mol. Biol. Evol.* 15:590–99
78. Maynard Smith J, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? *Proc. Natl. Acad. Sci. USA* 90:4384–88
79. McGuire G, Wright F. 1998. TOPAL: recombination detection in DNA and protein sequences. *Bioinformatics* 14:219–20
80. McGuire G, Wright F. 2000. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* 16:130–34
81. McGuire G, Wright F, Prentice MJ. 1997. A graphical method for detecting recombination in phylogenetic data sets. *Mol. Biol. Evol.* 14:1125–31
82. McGuire G, Wright F, Prentice MJ. 2000. A Bayesian model for detecting past recombination events in DNA multiple alignments. *J. Comput. Biol.* 7:159–70
83. McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–41
84. McVean GAT. 2001. What do patterns of genetics variability reveal about mitochondrial recombination? *Heredity* 87:613–20
85. Menotti-Raymond M, Starmer WT, Sullivan DT. 1991. Characterization of the structure and evolution of the Adh region of *Drosophila hydei*. *Genetics* 127:355–66
86. Meunier J, Eyre-Walker A. 2001. The correlation between linkage disequilibrium and distance: implications for recombination in hominid mitochondria. *Mol. Biol. Evol.* 18:2132–35
87. Meyerhans A, Vartanian J-P, Wain-Hobson S. 1990. DNA recombination during PCR. *Nucleic Acids Res.* 18:1687–91
88. Moutouh L, Corbeil J, Richman DD. 1996. Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proc. Natl. Acad. Sci. USA* 93:6106–11
89. Muse SV, Weir BS. 1992. Testing for equality of evolutionary rates. *Genetics* 132:269–76
90. Nagy PD, Simon AE. 1997. New insights into the mechanisms of RNA recombination. *Virology* 235:1–9
91. Nielsen R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931–42
92. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304
93. Odelberg SJ, Weiss RB, Hata A, White

R. 1995. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res.* 23:2049–57

94. Ohta T, Basten CJ. 1992. Gene conversion generates hypervariability at the variable regions of kallikreins and their inhibitors. *Mol. Phylogenet. Evol.* 1:87–90

95. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–23

96. Patthy L. 1999. Genome evolution and the evolution of exon-shuffling—a review. *Gene* 238:103–14

97. Popadic A, Anderson WW. 1995. Evidence for gene conversion in the amylase multigene family of *Drosophila pseudobscura*. *Mol. Biol. Evol.* 12:564–72

98. Posada D. 2001. Unveiling the molecular clock in the presence of recombination. *Mol. Biol. Evol.* 18:1976–78

99. Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.* 19:708–17

100. Posada D, Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16:37–45

101. Posada D, Crandall KA. 2001. Performance of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA* 98:13757–62

102. Posada D, Crandall KA. 2002. The effect of recombination in phylogeny reconstruction. *J. Mol. Evol.* 54:396–402

103. Posada D, Crandall KA, Nguyen M, Demma JC, Viscidi JC. 2000. Population genetics of the *porB* gene of *Neisseria gonorrheae*: different dynamics in different homology groups. *Mol. Biol. Evol.* 17:423–36

104. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199–204

105. Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* 406:64–67

106. Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516–17

107. Robertson DL, Hahn BH, Sharp PM. 1995. Recombination in AIDS viruses. *J. Mol. Evol.* 40:249–59

108. Robertson DL, Sharp PM, McCutchan FE, Hahn BH. 1995. Recombination in HIV-1. *Nature* 374:124–26

109. Salminen MO, Carr JK, Burke DS, McCutchan FE. 1996. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res. Hum. Retrovir.* 11:1423–25

110. Sanderson MJ, Doyle JJ. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy, and confidence. *Syst. Biol.* 41:4–17

111. Satta Y. 1992. Balancing selection at *HLA* loci. In *Population Paleo-Genetics*, ed. N Takahata, pp. 129–49. Tokyo: Jpn. Sci. Soc. Press

112. Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6:526–38

113. Sawyer SA. 1999. GENECONV: a computer package for the statistical detection of gene conversion. *Distributed by the author, Dep. Math., Wash. Univ. St. Louis, available at* http://www.math.wustl.edu/∼sawyer

114. Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156:879–91

115. Schierup MH, Hein J. 2000. Recombination and the molecular clock. *Mol. Biol. Evol.* 17:1578–79

116. Siepel AC, Halpern AL, Macken C, Korber BTM. 1995. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retrovir.* 11:1413–16

117. Siepel AC, Korber BK. 1995. Scanning the data base for recombinant HIV-1 genomes. In *Human Retroviruses and AIDS 1995: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences,* ed. G Myers, B Korber, B Hahn, K-T Jeang, J Mellors, et al. Los Alamos, NM: Theor. Biol. Biophys. Group, Los Alamos Natl. Lab.
118. Sneath PHA. 1995. The distribution of the random division of a molecular sequence. *Binary* 7:148–52
119. Sneath PHA. 1998. The effect of evenly spaced constant sites on the distribution of the random division of a molecular sequence. *Bioinformatics* 14:608–16
120. Sneath PHA, Sackin MJ, Ambler RP. 1975. Detecting evolutionary incompatibilities from protein sequences. *Syst. Zool.* 24:311–22
121. Stephens JC. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.* 2:539–56
122. Stephens JC. 1986. Of the frequency of undetectable recombination events. *Genetics* 112:923–26
123. Strimmer K, Moulton V. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.* 17:875–81
124. Strimmer K, Wiuf C, Moulton V. 2001. Recombination analysis using directed graphical models. *Mol. Biol. Evol.* 18:97–99
125. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In *Molecular Systematics*, ed. DM Hillis, C Moritz, BK Mable, pp. 407–514. Sunderland, MA: Sinauer
126. Tajima F. 1983. Evolutionary relationships of DNA sequences in finite populations. *Genetics* 105:437–60
127. Takahata N. 1994. Comments on the detection of reciprocal recombination or gene conversion. *Immunogenetics* 39: 146–49
128. Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. 2000. Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* 66:69–83
129. Templeton AR, Crandall KA, Sing CF. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132:619–33
130. Templeton AR, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. 2000. Cladistic structure within the human *Lipoprotein Lipase* gene and its implications for phenotyopic association studies. *Genetics* 156:1259–75
131. Valdés AM, Piñero D. 1992. Phylogenetic estimation of plasmid exchange in bacteria. *Evolution* 46:641–56
132. Vasco D, Crandall KA, Fu Y-X. 2000. Molecular population genetics: coalescent methods based on summary statistics. In *Computational and Evolutionary Analysis of HIV Molecular Sequences,* ed. AG Rodrigo, GH Learn, pp. 173–216. Dordrecht: Kluwer
133. Wakeley J. 1997. Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res.* :45–458
134. Wall JD. 2000. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* 17:156–63
135. Wall JD. 2001. Insights from linked single nucleotide polymorphisms: What we can learn from linkage disequilibrium. *Curr. Opin. Genet. Dev.* 11:647–51
136. Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7: 256–76
137. Webster RG, Laver WG, Air GM, Schild GC. 1982. Molecular mechanisms of variation in influenza viruses. *Nature* 296: 115–21
138. Weiller GF. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombination in homologous sequences. *Mol. Biol. Evol.* 15:326–35

139. Weiss KM, Clark AG. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18:19–24
140. Wiens JJ. 1998. Combining data sets with different phylogenetic histories. *Syst. Biol.* 47:568–81
141. Wiuf C, Christensen T, Hein J. 2001. A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol.* 18:1929–39
142. Worobey M. 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Mol. Biol. Evol.* 18:1425–34
143. Worobey M, Holmes EC. 1999. Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.* 80:2535–43
144. Worobey M, Holmes EC. 2001. Homologous recombination in GB virus C/hepatitis G virus. *Mol. Biol. Evol.* 18:254–61
145. Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11:367–72
146. Zhang YX, Perry K, Vinci VA, Powell K, Stemmer WP, del Cardayre SB. 2002. Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415:644–46
147. Zhou J, Bowler LD, Spratt BG. 1997. Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and skikimate dehydrogenase genes of *Neisseria meningitidis* and commensal *Neisseria* species. *Mol. Microbiol.* 23:799–812