15

Introduction to recombination detection

Philippe Lemey and David Posada

15.1 Introduction

Genetic exchange, henceforth recombination, is a widespread evolutionary force. Natural populations of many organisms experience significant amounts of recombination, sometimes with markedly different underlying molecular mechanisms. Although mutation is the ultimate source of genetic variation, recombination can easily introduce allelic variation into different chromosomal backgrounds, generating new genetic combinations. This genetic mixing can result in major evolutionary leaps, for example, allowing pathogens to acquire resistance against drugs, and ensures, as part of meiosis in eukaryotes, that offspring inherit different combinations of alleles from their parents. Indeed, when organisms exchange genetic information they are also swapping their evolutionary histories, and because of this, ignoring the presence of recombination can also mislead the evolutionary analysis. Because of its fundamental role in genomic evolution and its potential confounding effect in many evolutionary inferences, it is not surprising that numerous bioinformatic approaches to detect the molecular footprint of recombination have been developed. While the next chapter discusses methods to detect and characterize individual recombination events with a particular focus on genetically diverse viral populations, the aim of this chapter is to briefly introduce different concepts in recombination analysis and to position the approaches discussed in Chapter 16 in the large array of currently available recombination detection tools.

15.2 Mechanisms of recombination

Before we briefly review genetic recombination in different organisms, it should be pointed out that there are several scenarios for genetic exchange. On the one hand, the donor nucleotide sequence can neatly replace a homologous region in

The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme (eds.). Published by Cambridge University Press.

the acceptor molecule (***homologous recombination***). On the other hand, recombination can also occur as a result of crossovers at non-homologous sites or between unrelated nucleotide sequences (***non-homologous recombination***). From another perspective, the exchange can occur in both directions (*symmetrical recombination*) or there can be a donor organism and an acceptor (*non-symmetrical recombination*).

Different organisms have evolved and integrated various processes of genetic exchange in their life cycle. Biologists are generally most familiar with the process of genetic recombination that is part of the sexual reproduction cycle in eukaryotes. In males and females, a special type of cell division called ***meiosis*** produces *gametes* or sex cells that have halved the number of complete sets of chromosomes. During meiosis, crossing over between homologous chromosomes occurs as part of the segregation process. Whether genetic recombination occurs in a particular chromosome depends on how *chromatides* or chromosome arms are cut and ligated before the chromosome homologues are pulled apart. Half of the time, the original parental chromosome arms can be rejoined. This mechanism is not present in all eukaryotes; in a few systems, crossing does not occur during meiosis, and asexual reproduction is frequent. It is noteworthy that ***recombination rate*** variation in some human genome locations has been directly estimated from the analysis of sperm samples (Jeffreys *et al.*, 1998; Jeffreys *et al.*, 2000). Although the difficulty of carrying out such experiments prohibits a detailed and comprehensive picture of genomic recombination rates, this information can be extremely valuable in evaluating recombination rate profiles estimated from population genetic data (Stumpf & McVean, 2003), and in stimulating the development of more realistic statistical models of recombination (e.g. Wiuf & Posada, 2003).

In bacteria, unidirectional genetic exchange can be accomplished in at least three ways. A bacterium can pass DNA to another through a tube – the *sex pilus* – that temporarily joins to bacterial cells. This process of *lateral gene transfer* is called ***conjugation*** and only occurs between closely related bacteria. The second process, ***transduction***, occurs when *bacteriophages* transfer portions of bacterial DNA from one bacterial cell to another. In the third process, referred to as ***transformation***, a bacterium takes up free pieces of DNA from the surrounding environment. These mechanisms for lateral transfer and recombination are the bacterial equivalent of sexual reproduction in eukaryotes. Comparative analysis has revealed that many bacteria are genomic chimaeras, and that foreign genomic regions can be often associated with the acquisition of pathogenicity (Ochman *et al.*, 2000; Ochman & Moran, 2001). For bacteria, recombination rates can also be estimated in the laboratory and, interestingly, these are in general agreement with population genetic estimates (Awadalla, 2003; Ochman & Moran, 2001).

Viruses can exchange genetic material when at least two viral genomes co-infect the same host cell, and in the process of infection, they can even acquire genes

from their hosts. Indeed, the physical process of recombination can occur between identical genomes, but the evolutionary impact of genetic recombination is only noticeable when these two genomes are genetically different. Physically exchanging genetic segments within nucleotide molecules can occur in both segmented and non-segmented genomes. Genetic mixing among viruses is greatly facilitated when their genomes are segmented (*multipartite* viruses). In these cases, these segments can simply be reshuffled during co-infection, a process called ***reassortment***. Antigenic shift in influenza A is an important example of the evolutionary significance of reassortment. For DNA viruses, recombination is probably similar to that seen in other DNA genomes involving breakage and rejoining of DNA strands. For RNA genomes, it was long thought that recombination was absent until recombinant polioviruses were detected (Cooper *et al.*, 1974). The most supported model of RNA virus recombination nowadays is a copy-choice mechanism during replication, which involves mid replication switches of the RNA-dependent RNA polymerase between RNA molecules. A similar template-switching mechanism during reverse transcription has been invoked for retroviruses (discussed in more detail in the next chapter). However, alternatives to the copy-choice model of RNA virus recombination have been suggested (e.g. Negroni & Buc, 2001).

Viral recombination can have important biological implications. Recombination events have been demonstrated to be associated with viruses expanding their host range (Gibbs & Weiller, 1999; Vennema *et al.*, 1998) or increasing their virulence (Suarez *et al.*, 2004). In addition, for many virus examples, the theoretical advantages of recombination (discussed below) have been experimentally put to test (Chao *et al.*, 1992; Chao *et al.*, 1997), analyzed through simulation (e.g. Carvajal-Rodriguez *et al.*, 2007), or demonstrated by naturally occurring examples (Georgescu *et al.*, 1994). There appears to be a considerable variation in recombination rate estimates for different viruses (Chare *et al.*, 2003; Chare & Holmes, 2006). Different constraints in viral recombination will be, at least partially, responsible for this observation (for review see Worobey & Holmes, 1999). In addition, genome architecture and networks of interactions will shape the evolutionary consequences of recombination along the genome (Martin *et al.*, 2005).

15.3 Linkage disequilibrium, substitution patterns, and evolutionary inference

In population genetic data sets, the extent to which recombination breaks up linkage between loci is generally reflected in the pattern of ***linkage disequilibrium***. Linkage disequilibrium is observed when the frequency of a particular multilocus haplotype is significantly different from that expected from the product of the observed allelic frequencies at each locus. Consider, for example, two loci, *A* and *B*,

with alleles *A/a* and *B/b* respectively. Let P_A denote the frequency of allele *A*, and so forth. Similarly, let P_{AB} stand for the frequency of the *AB* haplotype. The classical linkage disequilibrium coefficient is then $D = P_{AB} - P_A P_B$. Recombination breaks up linkage disequilibrium. When a population is recombining "freely," segregating sites become independent, and the population is in *linkage equilibrium*. Measuring linkage disequilibrium as function of the distance between loci is the cornerstone of population genetic methods to estimate recombination rates, $4N_e r$, where N_e is the ***effective population size*** and r is the recombination rate per site per generation. It is also interesting to note that intermediate levels of linkage disequilibrium, and thus intermediate levels of recombination, assist us in identifying disease markers or drug resistance-genes in genome-wide association studies. Recombination ensures that only particular genome regions will be associated with a particular phenotype. Full independence among sites, however, will break up all the associations between markers and the phenotype (Anderson *et al.*, 2000).

When measuring linkage disequilibrium in sequence alignments, it is important to realize that haplotype structure can be shaped in a similar way by recombination and recurrent substitution (see Box 15.1). Because population genetic analyses have traditionally been performed on closely related genotypes, analytical methods were initially developed under an ***infinite sites*** model (Kimura, 1969), which constrains each mutation to occur at a different nucleotide site. Under this assumption, the possibility of recurrent substitution does not exist. For most populations, however, infinite site models are not appropriate and recurrent substitution needs to be accounted before attributing *incompatible sites* – sites for which the character evolution cannot be reconciled on a single tree, see Box 15.1 – to recombination. In particular, ***homoplasies*** as a result of convergent changes at the same site can be relatively frequent in sites under positive selection (especially under ***directional selection***, but also under ***diversifying selection***), even more than in neutral finite site models. In sequence analysis, the contribution of recombination and selection has been notoriously difficult to separate (Grassly & Holmes, 1997).

15.4 Evolutionary implications of recombination

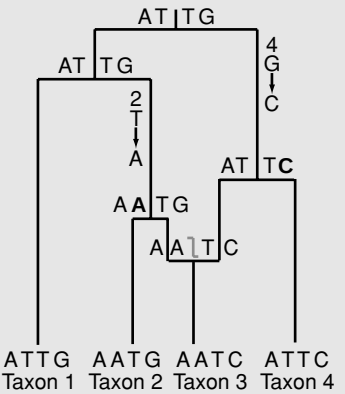
Shuffling genes or parts of genes through recombination inevitably results in mosaic genomes that are composed of regions with different evolutionary histories, and in the extreme case, this will result in the complete independence of ***segregating sites***. Conversely, sites in which allelic combinations are inherited together across generations, because recombination never separated them, are said to be linked, and share a single, common evolutionary history. In this way, recombination can be considered as a process that releases allelic variation at neutral loci from the action of selection at nearby, linked sites.

Box 15.1 Recombination and recurrent substitution

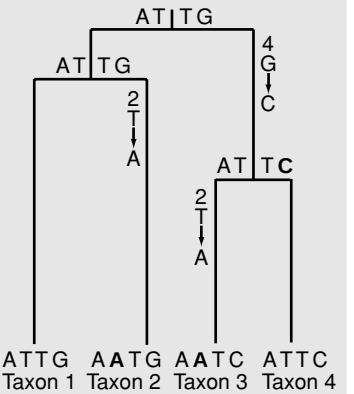
Consider four taxa with the following nucleotide sequences:

```
Taxon1 ATTG
Taxon2 AATG
Taxon3 AATC
Taxon4 ATTC
```

Both sites 2 and 4 are ***parsimony informative sites***: they contain at least two types of nucleotides, and at least two of them occur with a frequency of two. However, these two site patterns are *incompatible* because there is no single tree that can represent the evolutionary history of these taxa. In order to reconcile the evolutionary history of two incompatible sites, either recombination or recurrent substitution needs to be invoked:



Recombination



Recurrent mutation

In this figure sequence characters are indicated at the internal and terminal nodes. Sites that have mutated along a branch are indicated in bold. In the left graph, the site patterns are explained by a *recombination event*, represented by the joining of two branches into the lineage of taxon 3, which carries a recombinant sequence. In the right graph, character evolution is explained by *recurrent substitution*: change T→A has evolved independently and in parallel at site 2 in taxa 2 and 3. Both recurrent substitution and recombination result in ***homoplasies***.

To fully understand the implications of this, let us consider a population that recently experienced a severe population bottleneck or a founder effect. Such populations will only show limited genetic diversity because much ancient diversity has been lost due to the bottleneck event or through the founder effect. Similarly, a beneficial mutation can be rapidly fixed in the population through a ***selective sweep***. In non-recombining populations, neutral mutations at all other linked sites in the variant harboring the beneficial mutation will reach a high frequency in the population (a process known as ***genetic hitchhiking***), which will lead to a

genome-wide low degree of genetic diversity. When recombination occurs, however, the fate of neutral variation in the genome can become independent from the action of selection at another site. And this is increasingly so for neutral variation that occurs at genomic sites more distantly located from the site where selective pressure operates because of the higher probability of a recombination event between them. This can be of major importance when, for example, a beneficial mutation occurs in an individual that has a less favorable genetic background. The presence of (slightly) deleterious mutations in this individual will seriously delay the time to fixation for the beneficial mutation in the population (if it will become fixed at all). However, recombination can remove deleterious mutations in the genetic background and thus speed up the fixation of the beneficial mutation. It has been predicted that in finite, asexual populations deleterious alleles can gradually accumulate because of the random loss of individuals with the fewest deleterious alleles (Felsenstein, 1974; Muller, 1932). Purging a population from lower fitness mutations has therefore been a longstanding theoretical explanation for the evolution of recombination and some aspects of sexual reproduction.

15.5 Impact on phylogenetic analyses

A first step in understanding the impact of recombination on phylogenetic inference is to consider the result of a single recombination event in the evolutionary history of a small sample of sequences. Figure 15.1 illustrates how the recombination event, described in Box 15.1 results in different phylogenetic histories for the left and right part of the alignment.

The impact of a single recombination event on the genetic make-up of the sampled sequences, and hence on further evolutionary inference, will strongly depend on how fast substitutions are accumulated in the evolutionary history and on which lineages recombine at what time point in the history. We illustrate this in Fig. 15.2 for three different evolutionary scenarios. Note that the top three diagrams represent the processes of recombination and coalescence of lineages as we go back in time; mutation will be added later. For simplicity, we will assume that all breakpoints occur in the middle of the sequences, represented by the blocks under the diagrams. In Fig. 15.2a, a recombination event occurred shortly after a lineage split into two lineages. None of the "pure" parental lineages persist and only the recombinant taxon B is sampled. In the second example, an additional splitting event has occurred before two lineages recombine (Fig. 15.2b). Taxon or sequence C represents a descendant of one of the parental lineages of recombinant D. In the last example in this row, two relatively distinct lineages recombined, after which the recombinant lineage split into two lineages leading to recombinant individuals B and C that share the same breakpoint (Fig. 15.2c). Note that we have described

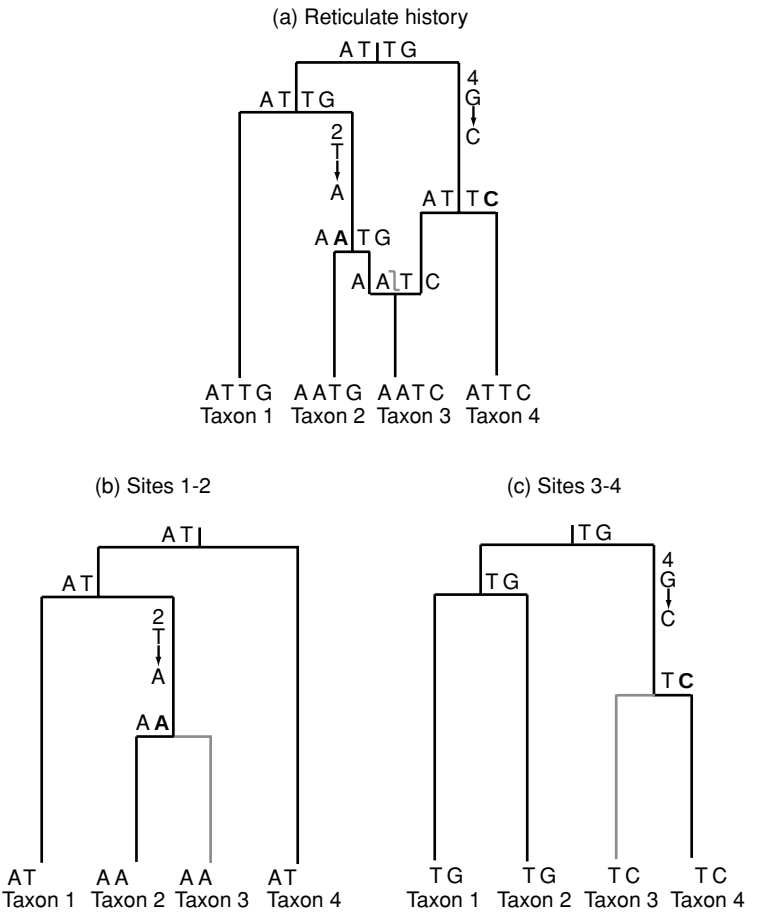


Fig. 15.1 Recombination generates different phylogenetic histories for different regions in the genome. (a) Reticulated history for the alignment discussed in Box 15.1. A recombination event between lineages 2 and 4 results in recombinant lineage 3. The recombination breakpoint was places between sites 2 and 3. (b) Phylogenetic tree for sites 1–2. (c) Phylogenetic tree for sites 3–4.

these histories of ancestral and descendent lineages without referring to mutation. In fact, these could represent extreme cases where no mutations have occurred. In this case, none of the recombination events resulted in any observable effect on the DNA sequences.

To illustrate an observable effect on the genetic sequences, we have added 8 (low substitution rate) and 20 (moderate substitution rate) non-recurrent substitution events to these histories (Fig. 15.2d–f, and Fig. 15.2g–i respectively). Horizontal lines in the evolutionary history represent the substitutions; a small circle on either side of the line indicates whether the substitution occurs on the left or right side of the genome, which is important to assess the presence and absence

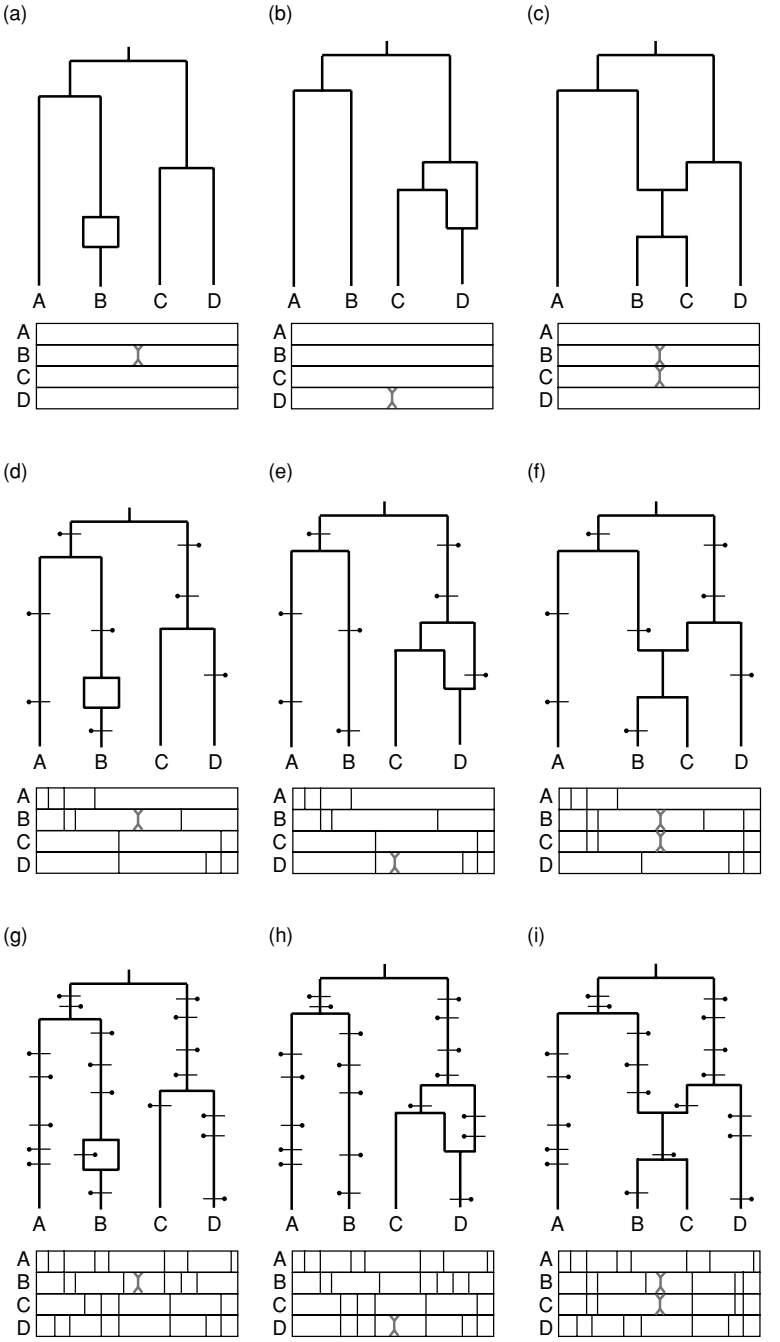


Fig. 15.2 Effect of different recombination events on the genetic make-up of molecular sequences. Three evolutionary histories with a different recombination event are represented in (a), (d), (g); (b), (e), (h), and (c), (f), (i), respectively. Different substitution processes are superimposed onto these histories resulting in different amount of substitutions in the sampled sequences: no mutation in (a), (b) and (c), a low mutation rate in (d), (e), and (f), a moderate mutation rate in (g), (h), and (i). A circle on either side of a horizontal line, representing a mutation in the evolutionary history, indicates whether a mutation occurs in the left or right part of the sequences, which are represented by boxes underneath the diagrams. Vertical lines in the sequence boxes represent inherited mutations.

of substitutions in recombinant lineages. For a recombinant, we assume that the left part of the genome was inherited from the leftmost pathway in the diagram and the right part from the rightmost pathway in the diagram (if the diagram could be disentangled into two evolutionary histories, similar to Fig. 15.1, then the left part of the genome evolved according to the left history, whereas the right part of the genome evolved according to the right history). For the first example (Fig. 15.2d and g), both low and moderate mutation rate leave no trace of incompatible sites or mosaic patterns and hence it will not impact our evolutionary analyses. Mutations that occur after the splitting event and before the recombination will be inherited by the recombinant depending on where in the genome they occur (Fig. 15.2g). In the second example, and under low substitution rate (Fig. 15.2e), the sequences have the same genetic make-up as in the first example (Fig. 15.2d). This is not the case for a higher substitution rate; sequence C and D have become more similar due to the recombination event (Fig. 15.2h). In this case, although it will impact the estimation of branch lengths, the recombination event does not generate incompatible sites, and estimates of the tree topology will not be affected. In the last example, incompatible sites are generated for both low and moderate substitution rates (Fig. 15.2h and i). Some mutations that were shared for A–B and C–D are now also shared for both recombinant lineages. The relative proportions of sites shared among A–B–C and B–C–D determine whether the recombinants will cluster with A or D. As a consequence, this recombination event will affect estimates of both tree topology and branch lengths.

The examples above illustrate the different impact of a single recombination event in the middle of the genome. The complexity of how recombination shapes genetic variation increases enormously as multiple recombination events occur during evolutionary history, each shuffling different parts of the genome. This is illustrated in Fig. 15.3, where different histories were simulated using a population genetic model with increasing recombination rates. The shaded boxes represented the sequence alignments, and alternate white and shaded areas indicate different phylogenetic histories. Recombination events similar to the three types we described in Fig. 15.3 can be observed, but now overlapping recombination events occur. If not only random substitutions, but also recurrent substitutions and selection processes generating convergent and parallel changes occur in these histories, it is not difficult to imagine how complex the impact on evolutionary inferences can be and how challenging the task of recombination detection will be.

Generalizing our arguments above, no single strictly bifurcating tree can accurately capture the true evolutionary relationships if different genome regions have evolved according to different phylogenetic histories. In this respect, phylogenetic network methods can be very useful to visualize complex evolutionary relationships (Posada & Crandall, 2001b)(Chapter 21). We illustrate some of the consequences

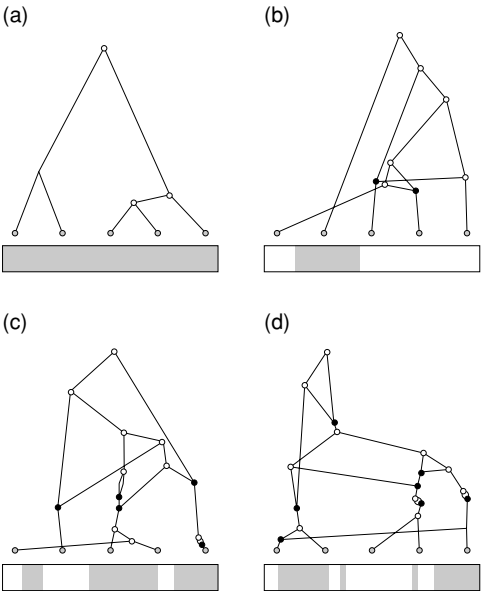
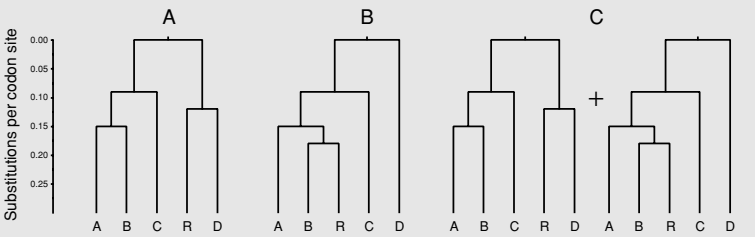


Fig. 15.3 Different population histories simulated for five sampled individuals using a population genetic model with increasing recombination rates. Each recombination event results in differently positioned breakpoint; *n* recombination events generate $n + 1$ different phylogenetic histories, represented by alternating white and shaded boxes. The histories were simulated using the HUDSON RECOMBINATION ANIMATOR (*http://www.coalescent.dk/*) with increasing population recombination rate: 0, 1, 2, 3 for a, b, c, and d, respectively.

of including a mosaic sequence in phylogenetic inference in Box 15.2. The first one is the effect on branch lengths and tree shape; phylogenetic simulation of recombinant data sets show a reduced ratio of internal branch lengths relative to external branch lengths. This agrees with population genetic simulations showing that relatively frequent recombination in the evolutionary history of larger data sets can create star-like trees (Schierup & Hein, 2000a). This effect on tree-shape is important to keep in mind when applying, for example, variable population size models that relate demography to tree shape (see the chapters on population genetic inference in this book). Phylogenetic simulations have also demonstrated the confounding effect of recombination on tree topology inference (Posada & Crandall, 2002), especially when the recombining lineages were divergent and the breakpoints divide the sequences in half (Fig. 15.2i). In some cases, a phylogeny can be inferred that is very different from any of the true histories underlying the data. Figure 15.2 suggests that recombination homogenizes the alignments; this is also evident from the simulations showing a reduced variance in pairwise distances in Box 15.2. Not surprisingly, the variance of pairwise differences has been interpreted

Box 15.2 The impact of recombination on phylogenetic inference

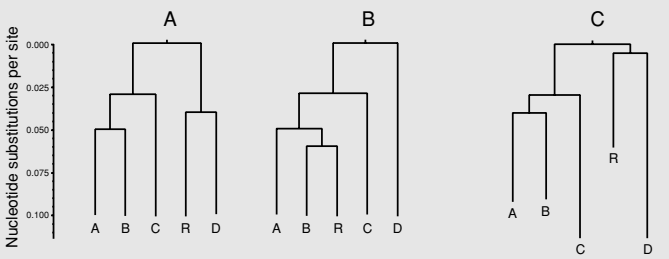
To demonstrate some of the effects of recombination on phylogenetic analyses, 100 data sets of five sequences were simulated under three different scenarios A, B and C:



In A and B, sequences were simulated according to phylogenetic trees that only differ with respect to the position of taxon R. One hundred alignments of sequences encompassing 500 codons were simulated using a codon model of evolution with three discrete site classes: 40% with $d_N/d_S = 0.1$, 30% with $d_N/d_S = 0.5$ and 30% with $d_N/d_S = 1.0$, representing strong negative selection, moderate negative selection and neutral evolution (for more information on codon models and d_N/d_S, see Chapter 14). In the case of C, 100 data sets were constructed by concatenating the alignments obtained in A and B. As a consequence, these data sets contain a mosaic sequence R that has a different evolutionary history in the first and second 500 codons.

Tree shape and branch lengths

Maximum likelihood trees were inferred using the best-fit nucleotide substitution model. The inferred tree topologies were the same for all data sets within their specific simulation scenarios, but branch lengths varied. Topologies with branch lengths averaged over all data sets are shown below:

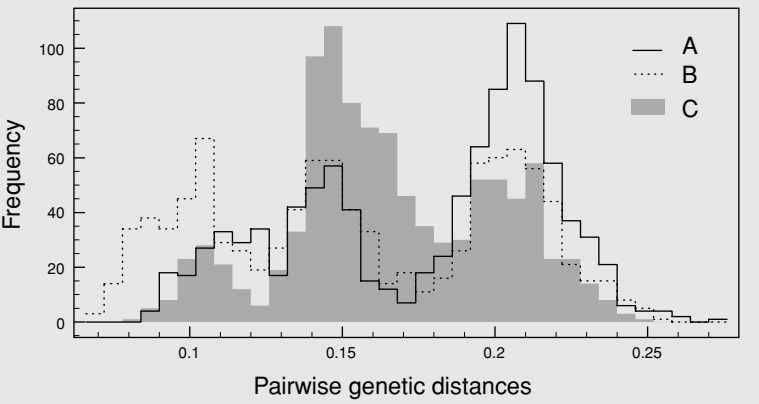


For the non-recombinant data sets in A and B, tree topologies were correctly reconstructed with mean branch lengths in nucleotide substitutions being almost exactly one third of the branch lengths in codon substitution units used for simulation. In the case of C, the best "compromise" tree is similar to the one used for simulation in A, however, with remarkably different branch lengths. Although R clusters with D, the branch lengths indicate a much larger evolutionary distance between R and D, but a lower evolutionary distance between R and A/B/C. This markedly reduces the ratio of internal branch lengths relative to external branch lengths; the mean ratio of external/internal branch lengths is 3.26 and 6.89 for A and C, respectively.

Box 15.2 (*cont.*)

Genetic distances

The effect on tree shape results from the homogenizing of sequences induced by recombination. This should also be apparent from pairwise genetic distances; histograms for the pairwise distances of all data sets are shown below:



In general, the pairwise distance distributions appear to be trimodal, with data sets from A having a higher frequency of large pairwise genetic distances. For the data sets including the recombinant R sequence (C), the distances regress to the average value. This has little effect on the average pairwise genetic distance (0.174, 0.153, and 0.162 substitutions per site for A, B, and C, respectively), but it reduces the variance to some extent in the recombinant data sets (0.00175, 0.0022 and 0.00118 for A, B, and C, respectively).

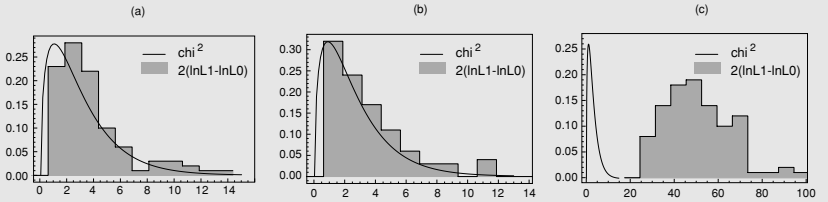
Rate variation among sites

In addition to branch lengths, also parameters in the evolutionary model are affected when including mosaic gene sequences. In particular, the shape parameter of the gamma distribution to model rate heterogeneity among sites is lower in the recombinant data sets (median value of 1.96, 2.10, and 1.17 for A, B, and C, respectively), indicating higher rate variation in this case (see Chapter 4). This is because the tree with the mosaic sequence will require extra changes at some sites to account for the homoplasies introduced by recombination. So, although R clusters with D, assuming extra changes at some sites in the region where R is most closely related to sequence B will be required.

Rate variation among lineages

Sequences were simulated according to an ***ultrametric*** tree meaning that each tip is exactly at the same distance from the root. The inferred trees with mean branch lengths for A and B are perfect reproductions of such trees. Because the tree with mean branch lengths in C is far from ultrametric, the presence of a recombinant might be erroneously interpreted as rate variation among branches (Schierup & Hein, 2000). To demonstrate

this more formally, we have performed a ***likelihood ratio test*** (*LRT*) for the ***molecular clock*** in each simulated data set and compared the distribution of the LRT statistic with the expected χ^2 distribution with three degrees of freedom:



Both simulated and expected distributions should be very similar under conditions of ***asymptotic normality***. This appears to be the case for the non-recombinant data sets. In fact, the 95% cut-off values are very similar for simulated (9.1 and 7.6 for A and B, respectively) and expected distributions (7.8). The picture for the recombinant data sets is remarkably different. In this case, the LRT would strongly reject the molecular clock for each simulated data set based on the χ^2 distribution cut-off.

Positive selection

The codon sequences were simulated under a selection regime that did not allow for positively selected sites. Although a sample size of five sequences is not very powerful to estimate site-specific selection parameters, we can still compare the fit of a codon model representing neutral evolution (M7; Yang *et al.*, 2000) and a model that allows an extra class of positively selected sites (M8) (for codon models, see also Chapter 14). These models can be compared using a LRT, similar to the molecular clock test, and significance can be assessed using a χ^2 distribution with two degrees of freedom. For a 95% cut-off value of 5.99, a proportion of 0.06 and 0.02 simulated data sets in A and B, respectively rejects the neutral model in favor of a model that allows for positively selected sites, which are very close to the expected 5%. For the recombinant simulation, however, there is a proportion of 0.10 data sets that reject the neutral model. This tendency confirms a more elaborate simulation study showing that the LRT often mistakes recombination as evidence for positive selection (Anisimova *et al.*, 2003).

as a measure of linkage disequilibrium and this and other summary statistics have been employed to estimate recombination rates (Hudson, 1985). The simulations in Box 15.2 also indicate inflated apparent rate heterogeneity, an important parameter in phylogenetic reconstruction (see Chapters 4 and 5). This observation has motivated the development of a specific test for recombination (Worobey, 2001).

Finally, any evolutionary analysis that models shared ancestry using a single underlying bifurcating tree will be affected by recombination. The simulations in Box 15.2 clearly indicate that recombination results in apparent rate heterogeneity among lineages. This corroborates the claim that even small levels of recombination invalidate the ***likelihood ratio test*** of the ***molecular clock*** (Schierup & Hein, 2000b) (but see Posada, 2001). Likewise, likelihood ratio tests for detecting selection using

codon models are also invalid in the presence of recombination (see also Anisimova *et al.*, 2003). Importantly, also ***non-synonymous/synonymous rate ratios*** (d_N/d_S) can be overestimated and high number of sites can be falsely identified as positively selected (Shriner *et al.*, 2003) (see Chapter 14 for inferences of this type). Because of this, several approaches have been recently developed to estimate this ratio and recombination rate simultaneously (e.g. OMEGAMAP: Wilson & McVean, 2006), or to accommodate the presence of obvious mosaic sequences when estimating d_N/d_S (Scheffler *et al.*, 2006). In summary, it is important to be aware that recombination affects many of the evolutionary analyses discussed in this book.

15.6 Recombination analysis as a multifaceted discipline

The myriad of available recombination analysis tools probably reflects the difficulty of evaluating recombination in molecular sequence data and the different questions we can pose about it (Posada *et al.*, 2002). Different bioinformatics strategies can be explored to tackle the problem, and very importantly, with different objectives. Identifying specific questions prior to data analyses will therefore be an important guidance in choosing among several approaches. In this chapter, we will distinguish four major goals: (i) detecting evidence of recombination in a data set, (ii) identifying the mosaic sequences, (iii) delineating their breakpoints, and (iv) quantifying recombination.

15.6.1 Detecting recombination

The first goal is determining whether recombination has occurred during the evolution of the sampled sequences. Although a judgment can be made based on different approaches, including graphical exploration tools, statistical tests are required to appropriately evaluate any hypothesis. Typically, *substitution distribution* and *compatibility* methods have been developed for this purpose and were among the earliest available recombination detection tools. These methods examine the uniformity of relatedness across gene sequences, measure the similarity or compatibility between closely linked sites or measure their composition in terms of homoplasies or two-state ***parsimony informative sites***. The null distribution of the test statistic, which determines the level of significance, can be obtained using specific statistical models; it can be generated by randomly reshuffling sites, or by simulating data sets according to a strictly bifurcating tree. The latter approach to determine *p*-values is often referred to as ***Monte Carlo simulation*** (or ***parametric bootstrapping***). Both permutation and simulation procedures have made different approaches amenable to statistical evaluation (including, for example, population genetic methods). Substitution distribution and compatibility methods are generally relatively powerful compared to methods that measure phylogenetic discordance (Posada & Crandall, 2001a).

15.6.2 Recombinant identification and breakpoint detection

If significant evidence for recombination can be detected in a data set, the question naturally arises which sequences are responsible for this signal. The methods that try to answer this question are usually, but not necessarily, methods that also attempt to localize recombination breakpoints. Obviously, if a mosaic pattern can be demonstrated for a particular sequence, then this also provides evidence for the recombinant nature of the sequence. Scanning methods employing distance or phylogenetic methods are generally well suited for this purpose, but also substitution distribution methods that examine the uniformity of relatedness across gene sequences can be used in this context. The next chapter expands on detecting and characterizing individual recombination events using these types of methods. It is, however, important to be aware of some caveats in using sliding window approaches for detecting recombinants and delineating breakpoints. First, scanning methods usually require *a priori* specification of a query sequence and putative parental sequences (or rather, the progeny thereof). So, if this approach is used in an attempt to identify recombinant sequences in a data set, the analysis needs to iterate through every sequence as possible recombinant, which obviously generates multiple testing problems. Second, Suchard *et al.* (2002) pointed out that scanning approaches fall into a "sequential testing trap," by first using the data to determine optimal breakpoints and parental sequences for a particular query sequence and then using the same data to assess significance conditional on the optimal solution. To overcome this, a Bayesian approach was developed to simultaneously infer recombination, breakpoints, and parental representatives, avoiding the sequential testing trap (Suchard *et al.*, 2002). Interestingly, such probabilistic models have also been extended to map recombination hotspots in multiple recombinants (Minin *et al.*, 2007).

15.6.3 Recombination rate

A different goal would be to quantify the extent to which recombination has shaped the set of sequences under study. Simply estimating the proportion of mosaic sequences in the data set does not provide adequate information because different sequences might share the same recombinant history and/or different recombinants might have different complexity resulting from a different number of recombinant events in their past. Some substitution distribution methods that measure homoplasies or two-state parsimony informative sites also determine the expected value for this compositional measure under complete linkage equilibrium. The ratio of the observed value over the expected value for complete independence of sites therefore provides some assessment of how pervasive recombination is (e.g. the *homoplasy ratio* and the *informative sites index*, Maynard Smith & Smith, 1998; Worobey, 2001). However, this is not always suitable for comparative analysis because the number of recombination events also depends on the time to The

Most Recent Common Ancestor (TMRCA) for a particular sample of sequences (and related to this, the effective population size of the population from which the sequences were sampled). The same criticism is true for methods that simply count the number of recombination events that have occurred in the history of a sample. This can be achieved by counting the occurrences where all allelic combinations are observed for two bi-allelic loci (AB, Ab, aB, ab). If an infinite sites model is assumed (see above), such an occurrence can only be explained by a recombination event. Because this *four-gamete test* only scores a recombination event if all four possible two-locus haplotypes can be observed in the sample, it is only a conservative estimate of the minimum number of recombination events that have occurred in the absence of recurrent mutation.

A quantitative estimate of recombination that takes into account the evolutionary time scale of the sampled sequences would be an estimate of the ***population recombination rate***. Very importantly, this implies a completely different scenario. Now we are trying to estimate a population parameter from a sample (*parametric* methods), while before we were trying to characterize the history of a given alignment (non-*parametric* methods). In this setting it is very important that the random sample provides a good estimate of the frequency of the different alleles present in the population. Note that now we will work with all the sequences, when before, we just used the different haplotypes. In Chapter 17, a population genetics model is introduced that describes the genealogical process for a sample of individuals from a population (***the coalescent***). Recombination can easily be incorporated into this model, in which case the coalescent process is represented using an ***ancestral recombination graph*** (**ARG**) instead of a single ***genealogy***. (In Figs. 15.2 and 15.3, most graphs represent ARGs, whereas the graph in Fig. 15.3a represents a genealogy.) Parametric methods based on linkage disequilibrium attempt to estimate the population recombination rate, ρ, which is the product of the per-generation recombination rate, r, and the effective population size (N_e): $\rho = 4N_e r$ (for diploid populations). The ARGs in Fig. 15.3 were simulated using increasing population recombination rates. If the ***population mutation rate*** Θ, a fundamental population genetics quantity equal to $4N_e m$ in diploid populations, can also be estimated, then the recombination rate estimate can be expressed as $\rho/\Theta = r/m$, which represents importance of per generation recombination relative to per generation mutation. As discussed above (Box 15.1), applying such estimators on diverse sequences, like many bacterial and viral data sets, requires relaxing the infinite site assumption. Other assumptions that we need to make under the parametric approaches are also important to keep in mind. The standard, neutral coalescent process operates under a constant population size, no selection, random mating, and no population structure. Although this ideal might be far from biological reality, the model can still be useful for comparing recombination rates among

genes and for the purpose of prediction (Stumpf & McVean, 2003). Nevertheless, it is clear that factors like population structure and demographic history may affect the ability of coalescent methods to correctly infer the rate of recombination (Carvajal-Rodriguez *et al.*, 2006). Despite the disadvantage of having to make simplifying assumptions, population genetic methods offer a powerful approach to estimating recombination rates across the genome, which can lead to a better understanding of the molecular basis of recombination, its evolutionary significance, and the distribution of linkage disequilibrium in natural populations (Stumpf & McVean, 2003).

15.7 Overview of recombination detection tools

Although we have attempted to associate different objectives to different recombination detection approaches, this does not result in an unequivocal classification of the available bioinformatics tools. Many tools share particular computational approaches or algorithmic details, but none of these are ideal classifiers. Both in terms of algorithms and objectives, the line has become blurred in many software packages. To provide a relatively detailed overview of different methods and software packages available, we have compiled a table that includes relevant features from a user perspective and particular algorithmic details (Table 15.1). To provide an approximate chronological overview, the methods are roughly ordered according to the year of publication. An updated version of this table will be maintained at www.thephylogenetichandbook.org and URLs for the software packages will be provided at *www.bioinf.manchester.ac.uk/recombination/*.

The criteria we list as important from a user perspective are statistical support, the ability to identify recombinant and parental sequences, the ability to locate breakpoints, and the speed of the application. A method is considered to provide statistical support if a *p*-value can support the rejection of the null hypothesis of clonal evolution. Methods that are designed to detect recombination signal in sequence data based on some test statistic are generally well suited to provide statistical support (e.g. through a Z-test, a chi^2-test, a binomial *p*-value or parametric, and non-parametric bootstrapping; see Section 15.5 and the next chapter). Statistical support can also be obtained in the form of a ***posterior probability*** (Suchard *et al.*, 2002) or using an incongruence test (e.g. the Shimodaira–Hasegawa test in GARD; Kosakovsky Pond *et al.*, 2006). Methods that are more graphical in nature do not always provide such support. For example, we do not consider bootscanning to be statistically supported, even though bootstrap values might give a good indication of the robustness of different clustering patterns in different gene regions. Bootscanning might not only fall into the sequential testing trap, but it also does

Table 15.1 Overview of different recombination detection tools

		User perspective criteria				Algorithmic criteria			
Method	Program	Statistical support	Identifies recombinants/ parentals	Locates breakpoints	Speed	Sliding window	Phylogenetic Incongruence	Run mode	Method/Program reference
Sawyer's statistical test for gene conversion	GENECONV	yes	yes	yes	fast	no	no	pairs (against average)	(Sawyer, 1989)
	RDP3	.	.	.	.	.	.	triplet/query vs. reference[a]	(Martin *et al.*, 2005)
	START	.	.	.	.	.	.	.	(Jolley *et al.*, 2001)
Maximum chi-squared	MAXIMUM CHI-SQUARED	yes	no	yes	fast	yes	no	triplets	(Maynard Smith, 1992)
	RDP3 (maxchi)	.	yes	.	.	.	.	triplets/query vs. references[a]	(Martin *et al.*, 2005)
	RDP3 (Chimaera)	.	yes	.	.	.	.	triplet	(Posada & Crandall, 2001)
	PHIPACK	.	.	no	.	.	.	.	(Bruen *et al.*, 2006)
	START	.	.	.	.	.	.	.	(Jolley *et al.*, 2001)
	4ISIS	.	.	.	.	no	.	quartet	(Robertson *et al.*, 1995)
Heuristic parsimony analysis of recombination	RECPARS	no	no	yes	fast	no	yes	all sequences	(Hein, 1993)
Similarity/distance plot	SIMPLOT	no	no	yes	fast	yes	no	query vs. references	(Lole *et al.*, 1999)
	RDP3	.	.	.	.	.	.	.	(Martin *et al.*, 2005)
	RIP	.	.	.	.	.	.	.	(Siepel *et al.*, 1995)
	FRAGDIST	.	.	.	.	.	.	.	(Gao *et al.*, 1998; van Cuyck *et al.*, 2005)
	BLAST genotyping	.	.	.	.	.	.	.	n.a.

Bootscanning	SIMPLOT	no	no	yes	fast	yes	yes	query vs. references[b]	(Lole *et al.*, 1999)
	RDP3	.	.	.	.	.	.	.	(Martin *et al.*, 2005)
	REGA	.	.	.	.	.	.	.	(de Oliveira *et al.*, 2005)
Compatibility matrix and neighbor similarity score	RETICULATE	yes	no	no	fast	no	no	all sequences	(Jakobsen & Easteal, 1996)
Partition matrices	PARTIMATRIX	no	no	yes	fast	no	no	all sequences	(Jakobsen *et al.*, 1997)
Difference in Sums of Squares method	TOPALI/RDSS	yes	yes[c]	yes	slow[d]	yes	yes	all sequences	(McGuire *et al.*, 1997)
	RDP3	.	.	.	.	.	.	quartets	(Martin *et al.*, 2005)
Likelihood detection of Spatial Phylogenetic Variation	PLATO	yes	no	yes	fast	yes	yes	all sequences	(Grassly & Holmes, 1997)
Homoplasy test	(qbasic programs)	yes	no	no	fast	no	no	all sequences	(Maynard Smith & Smith, 1998)
	START	.	.	.	.	.	.	.	(Jolley *et al.*, 2001)
	PHIPACK	.	.	.	.	.	.	.	(Bruen *et al.*, 2006)
Modified Sherman Test	SNEATHST	yes	yes	no	fast	no	no	pairs	(Sneath, 1998)
Graphical recombination detection using Phylogenetic Profiles	PHYLPRO	no	yes	yes	fast	yes	no	query vs. references[e]	(Weiller, 1998)
	RDP3	.	.	.	.	.	.	.	(Martin *et al.*, 2005)
Likelihood Analysis of Recombination in DNA	LARD	yes	no	yes	slow	yes	yes	triplet	(Holmes *et al.*, 1999)
	RDP3	.	yes	.	.	.	.	.	(Martin *et al.*, 2005)

(*cont.*)

Table 15.1 (*cont.*)

		User perspective criteria				Algorithmic criteria			
Method	Program	Statistical support	Identifies recombinants/ parentals	Locates breakpoints	Speed	Sliding window	Phylogenetic Incongruence	Run mode	Method/Program reference
Sister scanning method	SISCAN	no	no	yes	fast	yes	no	quartet/triplet	(Gibbs *et al.*, 2000)
	RDP3	yes	yes	.	.	.	.	quartet/triplet/ query vs. references[a]	(Martin *et al.*, 2005)
the RDP method	RDP3	yes	yes	yes	fast	yes	no	triplet	(Martin & Rybicki, 2000)
Informative sites test	PIST	yes	no	no	fast	no	no	all sequences	(Worobey, 2001)
Phylogenetic hidden Markov model with Bayesian inference	SERAD (Matlab)[f]	yes	no	yes	slow	no	yes	quartet	(Husmeier & Wright, 2001a)
	BARCE[f]	.	.	.	.	.	.	.	(Husmeier & McGuire, 2003)
	TOPALi/rHMM[f]	.	no[c]	.	slow[d]	.	.	.	(Milne *et al.*, 2004)
Probabilistic divergence method using MCMC	JAMBE[f]	yes	no	yes	slow	yes	yes	all sequences	(Husmeier & Wright, 2001b)
TOPALi/rPDM		.	no[c]	.	slow[d]	.	.	.	(Milne *et al.*, 2004)
Bayesian multiple change-point modelling	DUALBROTHERS	yes	no	yes	slow	no	yes	query vs. reference	(Suchard *et al.*, 2002)
	cBROTHER	.	.	.	.	.	.	.	(Fang *et al.*, 2007)
Distance-matrix calculation across breakpoints	BELLEROPHON	no	yes	yes	fast	yes	no	query vs. references[e]	(Huber *et al.*, 2004)

Visual recombination detection using quartet scanning	VisRD	no	no	yes	fast	yes	yes	quartets	(Strimmer *et al.*, 2003)
Distance-based recombination analysis tool	Rat	no	yes	yes	fast	yes	no	query vs. references[e]	(Etherington *et al.*, 2005)
Automated bootscanning (Recscan)	Rdp3	yes	yes	yes	fast	yes	yes	triplets	(Martin *et al.*, 2005)
Recombination detection using multiple approaches	Rdp3	yes	yes	yes	method dependent	yes	yes	triplets & quartets	(Martin *et al.*, 2005)
Stepwise recombination detection	Stepwise (R-package)	———		method dependent[g]	———				(Graham *et al.*, 2005)
Pairwise homoplasy index	PhiPack	yes	no	no	fast	no	no	all sequences	(Bruen *et al.*, 2006)
	SplitsTree								(Huson, 1998)
Phylogenetic compatibility method	Simmonics	no	yes	yes	slow	yes	yes	all sequences	(Simmonds & Welch, 2006)
Recombination analysis using cost optimization	Recco	yes	yes	yes	slow	no	no	all sequences	(Maydt & Lengauer, 2006)
Genetic algorithm for recombination detection	Gard	yes	no	yes	slow	no	yes	all sequences	(Kosakovsky Pond *et al.*, 2006)
Jumping profile hidden markov models	jpHMM	no	no	yes	slow	no	yes	query vs. reference	(Schultz *et al.*, 2006)
Recombination detection using hyper-geometric random walks	3Seq	yes	yes	yes	fast	no	no	triplets	(Boni *et al.*, 2007)
	Rdp3	.	.	.	.	.	.	triplets/query vs. references[a]	(Martin *et al.*, 2005)

(*cont.*)

Table 15.1 (*cont.*)

		User perspective criteria				Algorithmic criteria			
Method	Program	Statistical support	Identifies recombinants/ parentals	Locates breakpoints	Speed	Sliding window	Phylogenetic Incongruence	Run mode	Method/Program reference
Building evolutionary networks of serial samples	Sliding MinPD	no	yes	yes	fast	yes	method dependent[h]	all sequences	(Buendia & Narasimhan, 2007)
Comparing trees using likelihood ratio testing	TOPALi/rLRT	yes	no[c]	yes	slow [d]	yes	yes	all sequences	(Milne *et al.*, 2004)

[+] A dot in the cells refers to the same content as the first software package implementing the same approach.

[a] For these approaches, **RDP3** allows the user to perform a "manual" analysis by assigning a query and parental sequences or to perform an "automated" analysis that evaluates every possible quartet or triplet. The latter setting provides a useful approach to identify putative recombinant sequences in the data set.

[b] The bootscan approach analyzes all sequences simultaneously (phylogenetic tree inference), but uses the "query vs. reference" scheme *a posteriori* to trace the clustering of a particular query sequence.

[c] In **TOPALi**, a modified difference in sums of squares (DSS) method can be used to find which sequence(s) appear to be responsible for the recombination breakpoints. This "leave-one-out" method uses as windows the homogeneous regions between the breakpoints, identified using any method. The DSS scores for each breakpoint are calculated, leaving out one sequence at a time. To assess significance, 100 alignments are simulated. A sequence is a putative recombinant if removing it results in a non-significant recombination breakpoint. This algorithm can be applied after a recombinant pattern is identified using any method implemented in **TOPALi**.

[d] The methods in **TOPALi** are generally slow when run on a single processor, but when spread on multiple CPUs analyses will run significantly faster.

[e] Although these software packages compare a query sequence against all the other sequences, they can perform this comparison for every sequence in the data set being assigned as a query. In PHYLPRO, phylogenetic correlations, which are based on pairwise genetic distances, are computed for each individual sequence in the alignment at every position using sliding-window techniques. BELLEROPHON evaluates for each sequence the contribution to the absolute deviation between two distance matrices for two adjacent windows. RAT has an "auto search" option to evaluate the similarity of every sequence to all other sequences. Therefore these approaches can be useful in identifying putative recombinants.

[f] SERAD is the MATLAB precursor of BARCE (C++ program). Both BARCE and JAMBE have been integrated into TOPALI, which provides a user-friendly GUI and several on-line monitoring diagnostic tools. Note that the BARCE method may predict erroneous recombination events when a DNA sequence alignment shows strong rate heterogeneity. An improved method that addresses this problem via a factorial hidden Markov model has been described in Husmeier (2005). The method has been implemented in a MATLAB program (*http://www.bioss.ac.uk/∼dirk/Supplements/phyloFHMM/*), but unfortunately, this implementation is slow and computationally inefficient for the time being.

[g] The stepwise approach can be applied to any recombination detection method that uses a permutation test and provides estimates of breakpoints. The criteria depend on the method that is used in the stepwise approach.

[h] SLIDING MINPD implements three different methods to identify recombinants: a percentage identity method (as implemented in the `Recombination Detection Program`, RIP), a standard bootscanning method and a distance bootscanning method. Only the standard bootscanning method infers trees for each alignment window.

not assess how much topological variability can be expected as a result of chance alone (under the null hypothesis).

The ability to detect recombinant sequences and breakpoint locations has been discussed above (see Section 15.5). It should be noted that methods examining every possible triple or quartet combination in a data set (e.g. RDP3 and 3SEQ), are generally designed to identify (a) combination(s) of taxa for which the null hypothesis can be rejected, with appropriate multiple testing correction (see Section 15.5 and next chapter), but cannot always discriminate between the recombinant and parental sequences within that triplet/quartet. In terms of speed, we have only used a slow/fast classification based on user-experience or input from the original authors. If the analysis of a moderate size data set was thought to take no more than a coffee break, we classified it as "fast." Of course, coffee breaks are stretchable, software speed can heavily depend on the settings involved, and a more objective evaluation on benchmark data sets is required to provide a more accurate and quantitative classification.

The algorithmic criteria in Table 15.1 include the use of a sliding window approach, "phylogenetic incongruence" and "run mode." Methods are classified as using a phylogenetic incongruence criterion if phylogenetic trees form the cornerstone of the inference. Whereas sliding window refers to a particular form of alignment partitioning, the "run mode" refers to a taxa partition used by the method. Several methods do not partition the taxa in their analysis strategy and detect recombination in the complete data set ("all sequences"). Other methods focus on subsets like pairs, triplets or quartets. In some cases, these methods analyze all possible combinations of this partitioning scheme in an attempt to pinpoint putative recombinants and their parental sequences (e.g. RDP3 and 3SEQ) or in an attempt to provide a graphical visualization of phylogenetic incongruence (e.g. VISRD). In other software programs, the application of the algorithm is simply restricted to such subsets (e.g. a quartets in BARCE and TOPALI/RHMM and a triplet in the original MAXIMUM CHI-SQUARED program). Methods that use the "query vs. reference" setup focus on the relationship – frequently inferred from pairwise genetic distances (e.g. PHYLPRO, SISCAN, SIMPLOT, BELLEROPHON and RAT) – between one particular sequence and all the other sequences in the alignment. Some of these programs iterate through all sequences when evaluating these relationships (see footnote *e* in Table 15.1), while others restrict themselves to the *a priori* assigned query sequence.

Because population genetic inferences are considered as a different class of methods with the primary objective of quantifying recombination (see Section 15.5), they are not included in Table 15.1. A list of population genetic methods is provided in Stumpf and McVean (2003), which includes software packages like SEQUENCELD and FINS (Fearnhead & Donnelly, 2001), LDHAT (McVean

et al., 2002) and its four-allele extension that implements more complex evolutionary models (Carvajal-Rodriguez *et al.*, 2006) (*http://darwin.uvigo.es*), RecMin (Myers & Griffiths, 2003) and Lamarc (Kuhner *et al.*, 2000) (see Chapter 19).

15.8 Performance of recombination detection tools

Probably the most important user criterion to make an objective choice is lacking in Table 15.1; how good are these methods at achieving their goal? Two aspects are important in their evaluation: *power* (or false negative rate) and *false positive rate*. The power and false positive rate of detecting recombination signal in molecular data has been evaluated for different methods (Brown *et al.*, 2001; Posada & Crandall, 2001a; Smith, 1999; Wiuf *et al.*, 2001). For this purpose, coalescent-based simulations appear to be very useful. To evaluate power, two variables are important in the simulation procedures: recombination rate and mutation rate. Increasing recombination rates result in an increasing number of recombination events in the history of the sampled sequences (Fig. 15.3). Although not all events leave a molecular footprint of recombination (Fig. 15.2), the frequency of simulated data sets in which recombination can be detected should increase towards 100% as higher recombination rates are used in the simulation. However, also the mutation/substitution process that is superimposed onto the ancestral recombination graphs to simulate the sequences will impact the power of recombination detection (Fig. 15.2). The higher the mutation rate per site per generation, the larger the genetic difference between two randomly drawn sequences and the higher the frequency of incompatible sites and conflicting phylogenetic information (compare Fig. 15.2c, f, and i). For every tool, it is hoped that the sensitivity in detecting recombination is not at the expense of the rate of false positive detection (Type I error). To evaluate this, genealogies are simulated without recombination events. In the mutational process, not only increasing genetic divergence is now important, but also increasing rate heterogeneity among sites (Posada & Crandall, 2001a) (see Chapter 4 on how this is modeled in the nucleotide substitution process). The latter increases the probability of recurrent substitution, which also increases the frequency of incompatible sites in the generated sequences (see Box 15.1).

The most comprehensive study comparing 14 different methods using such simulations revealed that recombination detection tools are generally not very powerful, but they do not seem to infer many false positives either (Posada & Crandall, 2001a). Methods that examine substitution patterns or incompatibility among sites appeared more powerful than phylogenetic methods, a conclusion also shared by smaller-scale studies (Brown *et al.*, 2001; Wiuf *et al.*, 2001). This might not be surprising in case of low sequence diversity. If the boxes in Fig. 15.2 would be translated to real sequences, no well-supported trees could probably be inferred

from either side of the breakpoint, even for those representing "moderate" mutation rates. However, the substitution patterns and the amount of incompatible sites can still exhibit significant deviations from clonality. In real molecular data, such deviations might also result from other evolutionary process. So comparisons on empirical data sets (Posada, 2002), and those for which recombination is well characterized in particular (Drouin *et al.*, 1999), can provide important insights. Particular processes that can lead to false positive results are now also being implemented in simulation procedures (e.g. substitution rate correlation, Bruen *et al.*, 2006). Finally, it is worth noting that methods have only been systematically evaluated for their performance in revealing the presence of recombination. Similar studies to evaluate the accuracy of identifying recombinant sequences within a data set and locating breakpoints will greatly assist the selection among available methods (e.g. Chan *et al.*, 2006).

Acknowledgment

We thank David Robertson and Darren Martin for comments and suggestions on how to classify recombination detection methods.