MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 1952–1963 doi: 10.1111/j.1365-294X.2011.05066.x

The performance of phylogenetic algorithms in estimating haplotype genealogies with migration

WALTER SALZBURGER,*§ GREG B. EWING†§ and ARNDT VON HAESELER‡

*Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland, †Mathematics and BioSciences Group, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Veterinary University of Vienna, Dr.-Bohr-Gasse 9, 1030 Vienna, Austria, ‡Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Veterinary University of Vienna, Dr.-Bohr-Gasse 9, 1030 Vienna, Austria

Abstract

Genealogies estimated from haplotypic genetic data play a prominent role in various biological disciplines in general and in phylogenetics, population genetics and phylogeography in particular. Several software packages have specifically been developed for the purpose of reconstructing genealogies from closely related, and hence, highly similar haplotype sequence data. Here, we use simulated data sets to test the performance of traditional phylogenetic algorithms, neighbour-joining, maximum parsimony and maximum likelihood in estimating genealogies from nonrecombining haplotypic genetic data. We demonstrate that these methods are suitable for constructing genealogies from sets of closely related DNA sequences with or without migration. As genealogies based on phylogenetic reconstructions are fully resolved, but not necessarily bifurcating, and without reticulations, these approaches outperform widespread 'network' constructing methods. In our simulations of coalescent scenarios involving panmictic, symmetric and asymmetric migration, we found that phylogenetic reconstruction methods performed well, while the statistical parsimony approach as implemented in TCS performed poorly. Overall, parsimony as implemented in the PHYLIP package performed slightly better than other methods. We further point out that we are not making the case that widespread 'network' constructing methods are bad, but that traditional phylogenetic tree finding methods are applicable to haplotypic data and exhibit reasonable performance with respect to accuracy and robustness. We also discuss some of the problems of converting a tree to a haplotype genealogy, in particular that it is nonunique.

Keywords: haplotype genealogy, maximum likelihood, maximum parsimony, neighbour-joining, networks

Received 1 June 2010; revision received 4 February 2011; accepted 9 February 2011

Introduction

Haplotypic genetic data play a prominent role in phylogenetic, population genetic, phylogeographic, molecular evolutionary, systematic biological, taxonomic (e.g. DNA barcoding), and nowadays also genomic research (see Brown *et al.* 1979; Savolainen *et al.* 2002; Hebert *et al.* 2003; Alonso & Armour 2004; The International HapMap Consortium, 2005, 2007; Gibb *et al.* 2007; Storey *et al.* 2007). Some fields, such as phylogeography, predominantly rely on highly similar haplotypic data, as the study objects are closely related and data are often comprised of population samples within species. But also when dealing with large data sets of closely related species (e.g. Verheyen *et al.* 2003; Barluenga *et al.* 2006) and when extrapolating species trees from gene trees using more than one locus (e.g. Maddison 1997), haplotypic genetic data are processed.

A *haplotype* is defined as a stretch of DNA on a single molecule, which is inherited as a single unit.

Correspondence: Greg B. Ewing, Fax: +43 1 4277 24098;
E-mail: gregory.ewing@univie.ac.at
§These authors contributed equally to this work.

Inheritance in units is intuitive for the relatively small haploid and usually nonrecombining cytoplasmic genomes of mitochondria and chloroplasts or for nonrecombinant sex chromosomes such as the Y-chromosome in e.g. mammals. However, haplotypic structures that act as single units over longer evolutionary times are also detected in diploid nuclear genomes (Lindblad-Toh *et al.* 2005; The International HapMap Consortium 2005, 2007). These nuclear haplotypes are stable when the stretch of DNA on a single chromosome is short enough so that recombination is unlikely to occur. New high-throughput sequencing techniques (e.g. Solexa, SOLiD and 454 methods) produce such nuclear haplotypic data in volume, and it is obvious that these will become increasingly important in phylogeography and population genetics (Gompert *et al.* 2010; Holsinger 2010; Tautz *et al.* 2010). The virtually haplotypic nature of these data eases analysis, as most commonly applied tree reconstruction methods assume absence of recombination (Posada & Crandall 2002; Felsenstein 2003). This is why, in phylogenetics, a suite of algorithms exists that ultimately aim for breaking up recombinant DNA molecules into haplotypes that can then be analysed separately (see e.g. Stephens & Donnelly 2003; Scheet & Stephens 2006; Sun *et al.* 2007). Hence, while recombination (in form of e.g. crossing over, horizontal gene transfer, or hybridization) is a real evolutionary event of its own significance, one seeks for nonrecombinant DNA sequences in phylogenetic inference (Posada & Crandall 2002; Felsenstein 2003).

Phylogeographic data sets have particular properties compared with conventional phylogenetic ones. First, they often contain more sequence data, as individuals from many locations are investigated. Second, and more importantly, genetic variation is typically rather limited as a consequence of the young age of the study group. Third, both ancestral and derived alleles are present in the data set. Finally, migration of individuals between populations may affect the composition of haplotypes in such data sets, which has consequences on the distribution of branch lengths (typically, migration leads to long internal branches).

In the last few years, several approaches have been put forward that specifically aim to reconstruct genealogies from haplotypic genetic data obtained from closely related taxa. These algorithms have been developed because it was assumed that 'traditional' phylogenetic methods such as neighbour-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML) would not accurately reconstruct relationships between closely related sequences or would pose analytical problems, for example because of the presence of ancestral and derived alleles (see e.g. Excoffier & Smouse 1994; Bandelt *et al.* 1995; Clement *et al.* 2000; Posada & Crandall 2001). Three commonly applied methods in the context of reconstructing haplotype genealogies are 'statistical parsimony' (SP) (Templeton *et al.* 1992) as implemented in TCS (Clement *et al.* 2000), the 'median-joining networks' (MJ) approach (Bandelt *et al.* 1999) as implemented in the computer program Network and the 'minimum-spanning network' method (MS) as implemented in Arlequin (Schneider *et al.* 2000; Excoffier *et al.* 2005). Two of these methods, SP and MJ, are based on MP algorithms, while MS is a distance-based approach. More recently, Cassens *et al.* (2005) presented a novel approach to unite the information provided by all equally most parsimonious trees (obtained by a tree search) into a single genealogy, which they termed the 'union of maximum parsimony trees' (UMP).

All four methods for constructing genealogies from closely related haplotypic data, SP, MJ, MS, and UMP may result in *reticulated genealogies*, which are often referred to as *networks*. This terminology is misleading, though, which is not least because of the ambivalent interpretation of reticulations (Cassens *et al.* 2005). Reticulations are typically illustrated as loops in the genealogy and are meant to indicate ambiguities and/or the presence of conflicting optimal topologies (e.g. in the set of most parsimonious trees). In these cases, a loop in a genealogy obtained from nonrecombining haplotypic genetic data merely illustrates the failure of the algorithm to decide for one of the alternative connections between haplotypes. Alternatively, a loop may also be intended to illustrate recombination events (which makes the data nonhaplotypic). Then, a loop would represent a real evolutionary scenario rather than an analytical problem. While some algorithms have explicitly been developed for nonrecombining haplotypic genetic data (e.g. MJ), others consider recombination at the population level (e.g. SP as implemented in TCS). However, there is no statistical way to interpret a reticulation, i.e. whether it is caused by recombination or migration or whether it is an analytical artefact.

The performance of SP, MJ, MS and UMP has previously been evaluated using simulated data sets (Cassens *et al.* 2005; Woolley *et al.* 2008). Cassens *et al.* (2005) simulated 100 sequence data sets each along four template topologies, which they subjected to individual analysis. They found that SP, MJ and UMP performed equally well, whereas MS shows poorer performance, in particular when internal haplotypes are not present in the data set. Woolley *et al.* (2008) simulated data sets of between 10 and 50 taxa under 18 conditions (with and without recombination), analysed them with the above-mentioned methods plus NJ and MP, and compared the outcome to the true tree. To this end, the networks were split into all embedded subtrees and the

frequency of correct subtrees was compared. They found that all methods (except MS) performed equally well in the absence of recombination; MP and UMP performed better at higher substitutions rates. However, the performance of network-construction methods has not yet been evaluated in data sets resembling those handled by empiricists (and with migration), and it remains unclear whether specifically designated network-construction methods are at all better suited in finding the 'true genealogy' than classic phylogenetic algorithms (NJ, MP, and ML) are.

Here, we use simulated data sets with known true genealogies to test the performance of classic phylogenetic algorithms such as NJ, MP and ML in the estimation of genealogies from nonrecombining haplotypic genetic data. These are widely used in phylogeographic and population genetic studies and constitute the basis for widely applied superimposed analyses such as the nested clade phylogeographic analysis (NCA) (Templeton *et al.* 1995; see Knowles & Maddison 2002; Knowles 2008; Beaumont & Panchal 2008; Beaumont *et al.* 2010 for known problems with NCA). We analysed 1000 simulated data sets with the phylogenetic reconstruction methods mentioned above and, additionally, with a commonly used algorithm for constructing haplotype genealogies. We mainly focus on one (TCS) out of the suite of approaches, as it has previously been shown that these perform equally well (see above). Still, we also applied MS to a representative set of simulations to evaluate whether also in our range of parameters MS is performing equally well or worse than TCS (Cassens *et al.* 2005; Woolley *et al.* 2008). Other methods were not considered for practical reasons, e.g. because they do not have a command line mode. The resulting genealogies from the different tree and network building algorithms were then compared with the known true trees. We tested whether classic algorithms are suited to infer genealogies from haplotypic data and then compared the performance of the different methods scoring (i) the percentage of correctly resolved topologies and (ii) the average number of errors introduced in the analysis. We find that, at least for the chosen set of parameters and simulation conditions, phylogenetic analyses do perform well in the re-construction of haplotype genealogies and that these actually outperform TCS (and MS).

Materials and methods

Definitions

We define a *haplotype genealogy* as a weighted acyclic graph with labelled leaves and some labelled internal nodes. Labels refer directly to haplotypes that have been sampled from the population. Weights of edges are integer valued and are greater than zero. Hence, a haplo-

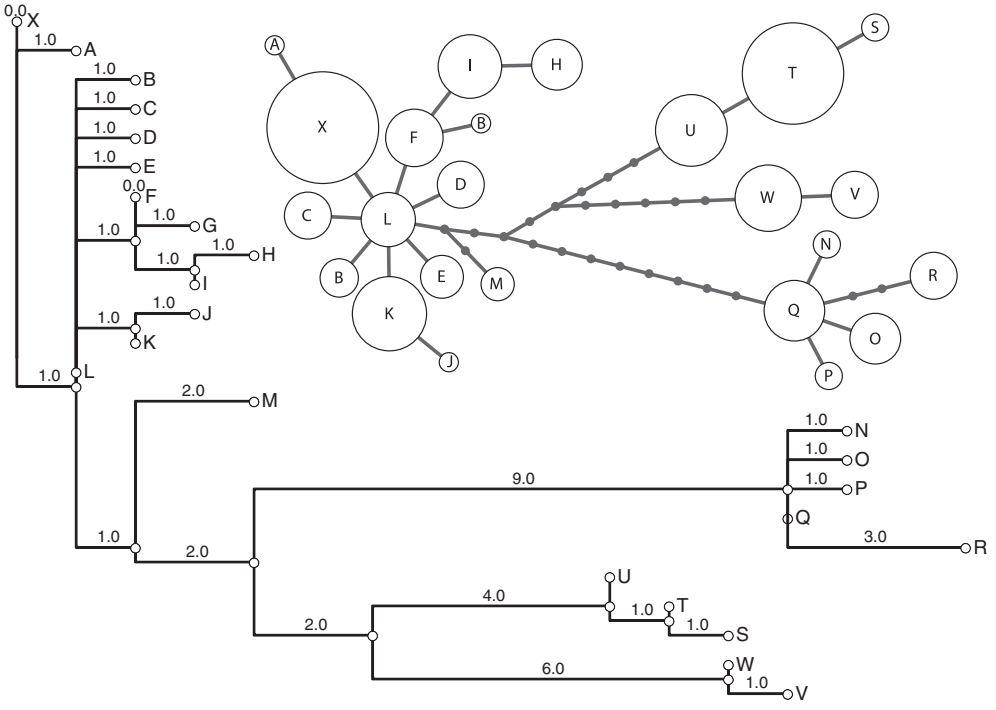

Fig. 1 Converting a tree to haplotype genealogy. The tree has Fitch branch lengths and zero length branches are collapsed. The size of the haplotype nodes in the haplotype genealogy denotes relative frequency.

type genealogy is an unrooted tree with the difference that internal nodes can be labelled and branch lengths are integers to represent discrete mutational steps.

Trees are defined in the usual manner but branch lengths are different in the way they map to haplotype genealogies. To avoid confusion in the following discussion, we will define two specific tree types. In particular, there is a distinction between trees with *natural* branch lengths and *Fitch* branch lengths defined below.

A tree with *natural* branch lengths is an undirected acyclic graph with labelled leaves and branch lengths, where the branch lengths are some measure of evolutionary distance from the method used to reconstruct the tree. If the tree is bifurcating, it is said to be fully resolved, otherwise it is unresolved. Here, we are not interested in the natural branch lengths.

In a *Fitch* tree, the branch lengths are given by the Hamming distance between the two sequences at the ends of the branch (Fitch branch lengths). That is, the number of sites that are different between the sequences at the ends of the branch. On a Fitch tree, all internal nodes have ancestral sequences derived from the leaf sequences using the Fitch algorithm (Fitch 1970). This provides the smallest number of changes or mutations for the given tree (where the number of changes is the parsimony score). It is noted that the parsimony score is unique for a given tree. But the Fitch tree is not unique, because Fitch branch lengths are not unique even though their sum is. For a given tree, there may be many possible branch length 'labellings'.

Any phylogenetic tree with sequence data can be converted into a haplotype genealogy in the following way: First, the natural branch lengths are replaced with the Fitch branch lengths derived from the original sequence data. Generally, there will be branches that have zero length, which represent multifurcations in the haplotype genealogy. A special case occurs when a terminal branch has a zero branch length. In this case, the leaf in the tree represents an internal node in the haplotype genealogy (Fig. 1). As a result, the different Fitch branch lengths derived from a given tree lead to alternative haplotype genealogies (Fig. 2).

We emphasize the reason why we are interested in the Fitch trees rather than phylogenetic trees: A plain phylogenetic tree is estimated with a number of different possible methods and generally gives an estimate of branch length usually in units of expected numbers of mutations per site. However, this is usually performed by integrating over all internal sequences rather than a particular subset thereof. The process of converting a tree with Fitch branch lengths to haplotype genealogies requires that we deal only with observable mutations. Thus, we must consider all possible Fitch tree labels and all the resultant haplotype genealogies. This has

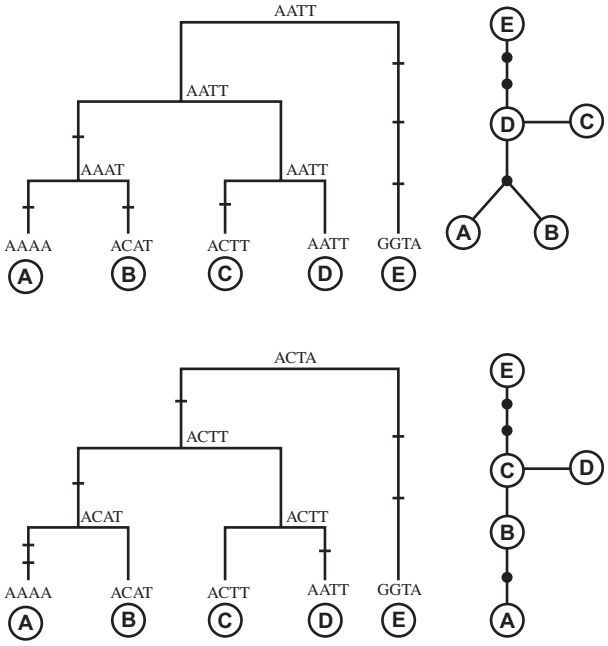

Fig. 2 Illustration of the nonuniqueness of Fitch trees. Mutations are denoted by the cutline on the trees on the left, while mutations are edges in the haplotype genealogy on the right. By making different choices with internal sequences, branches will also have different lengths. As can be seen, the resultant haplotype genealogy is altered. We also note the nonlocal behaviour of some choices of internal sequences. That is, by deciding what the internal sequence is at one location, it can affect a large section of the tree.

the important consequence that the only information used from any tree returned by any method is the topology, while branch length information is discarded at this stage (but later derived from the data).

We note that out of a set of different haplotype genealogies, no single genealogy offers a better description of the 'truth' than any other one does without considering external data such as the underlying DNA sequences (this is the same when dealing with a set of different MP trees with the same score). The question raised is how are we better off with a group of haplotype genealogies vs. a network that may not be tree-like. The existence of many haplotype genealogies is simply another way of representing ambiguity in the data.

However, the important difference between a network and a set of trees is the lack of independence of Fitch length labellings. We illustrate this in Fig. 2. We have the same initial tree with the same tip sequences, but the Fitch branch lengths and internal sequences are different. In the top figure we see that haplotype E connects to D, while haplotype A and B form a cherry also connecting to D. But an alternative is that haplotype E connects to C. This has the effect of changing the topology throughout the tree. So by making some choice in

one part of the Fitch tree, it can have topological consequences elsewhere in the tree. In the network case, each ambiguity is represented independently of each other.

Also, it is difficult to represent the same information in a graph compared to a set of trees. However, with suitable constraints such networks may be constructed and can, in fact, be informative, e.g. in cases where there is an expectation of a non-tree-like signal (Huson *et al.* 2005). Here, we are considering the case where the true signal is tree-like and that reticulations represent reconstruction ambiguity.

The Robinson-Foulds (RF) tree distance (Robinson & Foulds 1981) was used for comparing estimated and true haplotype genealogies, with the allowance that haplotype labels can be internal nodes. The RF distance of two genealogies is the number of splits (bipartitions of the labels induced by a branch) occurring in one tree only. The RF distance does not take into account branch lengths. This may seem disadvantageous, but recall that we do not use branch length information from the tree finding methods anyway. Rather, the branch lengths are completely determined by the DNA sequence alignment over an estimated tree. More importantly, the resultant branch lengths can and do change the resultant topology when this tree is converted into a haplotype genealogy.

The scaled mutation rate Θ is defined in the normal way as $\Theta = 2 * N_e\mu$, where μ is the mutation rate (per site and per locus) and N_e is the effective gene copy population size.

Simulation of data sets

A problem with regard to simulated data is: What should be considered the true haplotype genealogy when comparing with estimated genealogies? If one takes the simulated coalescent tree from a program like SIMCOAL (Excoffier *et al.* 2004) and then add Fitch branch lengths (see the definitions section above) with an alignment simulated over the same tree, we face the problem of many haplotype genealogies because of the nonuniqueness of Fitch branch lengths. This produces the situation that there can be multiple trees with an identical score as the true tree (and many inferred trees in total), which would give a false impression of the accuracy of the method. When dealing with simulations, we want the true haplotype genealogy to be known so as to not dilute the results. One way to do this, which is the method we have chosen, is to use the sequence simulation to give the sequences for internal nodes as well. Then, from the internal ancestral sequences, we can get unique true Fitch branch length labellings and a unique haplotype genealogy.

Because of the above-mentioned problems, the simulated data were generated with our own code that has been verified using various well-known population genetic results and against SIMCOAL (Ewing *et al.* 2004; Excoffier *et al.* 2004). Our method uses the Kingman coalescent (Kingman 1982a,b) with migration (Hudson 1990) to first simulate a genealogy under the given demographic model. In this case, we considered a panmictic population and a two-deme model with both symmetric and asymmetric migration. Sequence data were then generated over the coalescent genealogy using standard evolutionary models (Jukes & Cantor 1969; Felsenstein 1981), and a sequence length of 300 bp. Rate heterogeneity was included by assuming that each site has a different rate drawn from a *gamma* distribution (Ota & Nei 1994) with different shape parameters and a mean equal to 1. In practice, we simulated the sequence data using either a Jukes-Cantor (JC) (Jukes & Cantor 1969) or a GTR model (Felsenstein 1981; Lanave *et al.* 1984) of molecular evolution, with different shape parameters. Simulated trees had branch lengths measured in the number of *realized* mutations after sequence simulation. This tree was then pruned to include just one representative leaf for each haplotype for the phylogenetic analyses, where the representative is chosen randomly (in practice, haplotype frequency information would be added on the resulting genealogy). In all cases $\Theta = 0.01$ for both demes (or for the single deme in the panmictic case). The symmetric migration rates were 10 and 100, where for asymmetric migration one migration rate was set to zero. For example, the migration rates were 10 from deme *A* to *B* while it was zero from *B* to *A*.

Simulation parameters and statistics are summarized in Table 1. For convenience and compactness, we will denote the standard deviation of a mean value in brackets immediately after the value. In all simulations we assume the locus is 300 nucleotides long (thus resembling real data as e.g. generated by next generation sequencing). The panmictic simulations had 200 sampled individuals giving an average of 21.6 (3.7) distinct observed haplotypes. The longest branch had an average length of 5.7 (3.3) mutations, and the total average number of mutations was 34.7 (9.0) over the tree. In the second set of panmictic simulations, $\Theta = 0.0166$ with 500 sampled individuals. This resulted in an average number of distinct haplotypes of 39.3 (5.5), longest branch length of 9.2 (5.5) mutations and a total number of mutations of 67.0 (15.0). The number of sampled haplotypes in the first round was chosen based on the average number of sampled taxa of 201 (152) in 58 articles that appeared in 2006 in Molecular Ecology and were retrieved applying the search strings 'network' and 'haplotype' from the journal's homepage. The num-

Table 1 Simulation results

Simulation	% Correct	RF distance	Ave no. mutations	Ave longest branch	Ave no. taxa
$\Theta = 0.01$	92	0.12 (0.43)	35.5 (9.9)	5.9 (3.7)	21.7 (3.8)
Γ ($\alpha = 2.5$)	90.3	0.17 (0.58)	35.2 (9.8)	5.8 (3.3)	21.7 (3.9)
Γ ($\alpha = 0.5$)	81.1	0.32 (0.76)	34.6 (9.8)	5.7 (3.4)	21.6 (4.0)
$\Theta = 0.0166$	86.2	0.22 (0.6)	57.1 (13.5)	8.7 (5.1)	30.7 (4.6)
$\Theta = 0.0166$, 500 samples	85.8	0.24 (0.7)	67.9 (15.1)	9.0 (5.3)	39.9 (5.5)
$\Theta = 0.01$, $\lambda = 10$	73.3	0.41 (0.8)	102.6 (38.1)	24.4 (18.2)	35.4 (4.9)
$\Theta = 0.01$, $\lambda = 100$	83.8	0.27 (0.7)	74.3 (18.1)	11.8 (7)	35.3 (4.8)
$\lambda_{1 \to 2} = 10$, $\lambda_{2 \to 1} = 0$	72.4	0.44 (0.82)	118.6 (49.5)	32.1 (24.4)	35.3 (4.8)
Γ ($\alpha = 1$)	57.6	0.713 (1.0)	113.8 (43.7)	30.2 (21)	35.2 (4.0)
$\lambda_{1 \to 2} = 100$, $\lambda_{2 \to 1} = 0$	86.3	0.23 (0.68)	66.2 (14.3)	8.9 (4.7)	34.9 (4.9)

The perfect reconstructions were used to estimate the percentage of correct haplotype genealogies (% correct), the Robinson-Foulds (RF) distance between perfect reconstructions and true trees (RF distance; see section Definitions), as well as the average number of mutations (ave no. mutations), the average length of the longest branch (ave longest branch), and the average number of haplotypes (ave no. taxa); standard deviations are given in brackets. Discrepancies between perfect reconstructions and true trees arise from nonparsimonious truth, homoplasies, and back substitutions (see Materials and methods section). The first three rows are without migration, the remainder are with migration, where $\lambda_{i \to j}$ is the migration rate from deme i to j. Note that with low migration we obtained very large longest branches, which is a typical migration signature. The standard parameters, unless specified, are 200 samples, $\Theta = 0.01$, and 1000 simulations for each parameter set.

ber of 500 sampled haplotypes in the second round of simulations reflects the average number of sampled taxa in the 10% ($N = 6$) most taxon-rich studies in the same journal and year (average number of taxa = 535, SD = 135). Hence, our simulation conditions lie in the range encountered by empiricists.

Migration simulations have a distinctive pattern. The expected coalescent tree under migration tends to have a very long branch to the final coalescent event. This is caused by the fact that for two lineages to coalesce, they must be in the same deme. Thus, the waiting time to the last coalescent event is dominated by the time it takes for the last pair of lineages to migrate to a common deme. This results in a long internal branch as can be seen in the simulation results with a large increase in average length of the longest branch to 24.4 (18.2) mutations. This effect increased other internal branches as well, but it is stronger for the last coalescent event. When the migration rate was increased to 100, the average of the longest branch length decreased to 11.8 (7).

Selected rate heterogeneity results are also presented in Table 1 with a shape parameter of $\alpha = 0.5$ and $\alpha = 2.5$, respectively. The number of total mutations is expected to be the same under a *gamma* model. However, as some sites have a higher mutation rate, there should be more homoplasy. The effect was small though: For example, the average number of segregating sites (not shown) with no rate heterogeneity and no migration with $\Theta = 0.01$ was 33.2 sites, while with rate heterogeneity and a shape parameter of 0.5 (denoted $\alpha = 0.5$) was 29.3 sites.

We used various software for phylogenetic reconstruction to account for the diversity of methods available for tree searches, particularly with respect to heuristic search algorithms for MP and ML. NJ trees were reconstructed with PAUP* (Swofford 1993). PHYLIP (Felsenstein 1989) and PAUP* was used for MP with default settings (see below). For phylogenetic reconstruction using ML, the following software was utilized: PhyML (Guindon & Gascuel 2003), IQPNNI (Minh *et al.* 2005), PHYLIP and PAUP*. All were used with default options with the following exceptions: (i) All methods that had randomization of input options had them enabled. (ii) We also randomly shuffled the input with our own code to prevent any biases that may exist. (iii) If the method had a default that disabled global moves, these were enabled for a set of runs. That is, the method was used both with and without the option enabled. (iv) If we were dealing with data that were generated with rate heterogeneity, we also enabled estimating the shape parameter where available, using four categories. Again this was performed as an extra run. That is, the methods were run both with a *gamma* model and without a *gamma* model. (v) For IQPNNI, we set the maximum number of iterations to 20 000 to avoid excessive run times.

Generally, we tried a number of different parameters for each ML method. However, it was observed that different parameter settings had very little effect on the performance of the different methods. This is not surprising as we are estimating performance over an ensemble of coalescent trees rather than a fixed data set. In the following, we therefore present results from the simpler models plus some results from the *gamma* model of rate heterogeneity.

For comparison, we used TCS under standard parameters in command line mode with the gapped parame-

ter set to false and distances set to false. A major difference to the phylogenetic methods is that TCS can return disjoint networks. We interpreted these reticulated graphs as a way to represent ambiguity and considered the returned graph to be correct if at least one maximal spanning tree is correct. A maximal spanning tree is any embedded tree in the graph that includes all leaves. This was performed by enumerating all spanning trees of the graph (Shioura & Tamura 1995) and checking each tree in turn. We considered disjoint topologies incorrect, but we expand on this case in the results section. For comparative reasons, we also applied MS as implemented in Arlequin 3.0 (Excoffier *et al.* 2005) using standard parameters.

Comparison of haplotype genealogies

We first considered perfect reconstructions, i.e. we evaluated the genealogies constructed from the true trees using the simulation data. Possible errors (i.e. discrepancies between the perfect reconstructions and the true trees) arise from nonparsimonious truth, homoplasies and back substitutions. Nonparsimonious truth is where the true tree is not a parsimony tree. Simply put, there are more mutations than one would calculate from the data. Homoplasies, on the other hand, are when two identical haplotype sequences have different evolutionary histories (i.e. convergent evolution). Finally, back substitutions and multiple hits are where a single nucleotide mutates more than once and is hence not directly observable. Perfect reconstruction results give a bound to how good any reconstructions based on optimality can be.

One major disadvantage of many tree distance measures is that both trees must have the same set of taxa. That is, we cannot get a useful result if the number of leaves is different in one tree compared to the other, as can happen with TCS with disjoint graphs. A possible way around this is to prune the trees such that the label sets match in both. Unfortunately, this then leaves us with the problem of how to include the missing haplotypes in the metric. This should penalize TCS because we know that it should join the network somewhere. So, the normal method of pruning does not fully represent the cost of missing haplotypes. As an example, we can leave out one label and get a distance of 1. With the label present in the wrong place, the RF distance is at least 1 and will generally be much higher. This is only important when considering TCS, which can produce disjoint networks. Therefore, we pruned the larger tree until label sets matched and then calculated the RF distance and added the number of labels removed.

Many programs can return more than one tree. In these situations, we simply use the closest tree out of the returned tree for scoring. So we counted a method as producing the correct genealogy if one or more of the returned results are correct.

Results and discussions

Perfect reconstruction results

First, we consider perfect reconstruction results (Table 1). We find that reconstruction accuracy drops, when the longest branch gets longer or the total number of mutations on the tree increases. This is shown by the decrease in performance with migration, as the average length of the longest branch is much longer than without migration. This effect was expected, as the likelihood of homoplasies and back substitutions increases with longer branches and more mutations. We also note that rate heterogeneity degrades performance

Table 2 Percentiles of true trees found

Parameters	PhyML (%)	IQPNNI (%)	DNAML (%)	PAUPML (%)	DNAPARS (%)	PAUP* (%)	NJ (%)	TCS (%)
$\Theta = 0.01$	68.1	70	72.1	74.5	**79.5**	77.4	68.2	40.0
$\Gamma\ (\alpha = 2.5)$	55.8	55.9	54	56.3	**66.3**	63.4	55.4	32.2
$\Gamma\ (\alpha = 0.5)$	41	40.4	42	46.5	**51.8**	49.4	39.9	34.3
$\Theta = 0.0166$	53.7	54.9	54.4	59.2	**65.5**	55.7	52.2	11.3
$\Theta = 0.0166$, 500 samples	47.2	47.1	48	54	**60.2**	48	44.7	6.8
$\Theta = 0.01$, $\lambda = 10$	36	36	35.5	40.2	47.2	**47.6**	38.4	0.8
$\Theta = 0.01$, $\lambda = 100$	47.6	45.5	45.9	51.4	**61.6**	56.1	45.1	3.1
$\lambda_{1 \to 2} = 10$, $\lambda_{2 \to 1} = 0$	29.9	28.1	27.3	36.2	**40.9**	34.5	28.9	0.2
$\Gamma\ (\alpha = 1)$	14.9	14.9	14.7	17.3	**23.6**	19.8	17.7	0.5
$\lambda_{1 \to 2} = 100$, $\lambda_{2 \to 1} = 0$	50.1	49.1	51	57.4	**62.3**	59.5	49.2	5.1

Unless otherwise stated $\Theta = 0.01$ and there are 200 samples (and 1000 simulations for each parameter set). All ML methods perform equally well as does NJ. For parsimony we note that DNAPARS performs somewhat better than the base ML methods. TCS is the worst of all methods considered, which was due mainly to disjoint networks. As expected, larger Θ and low migration rates create difficulties for reconstruction methods, TCS in particular. ML, maximum likelihood. Values in bold indicate the best-scoring method.

with a smaller shape parameter. A smaller shape parameter results in a few sites with high mutation rate that can result in reconstruction problems. The effect is pronounced with the asymmetric migration case dropping performance from 72.4% to 57.6% caused by the occurrence of a long internal branch.

Estimation results without rate heterogeneity

The reconstruction results without rate heterogeneity are summarized in Tables 2 and 3. Most methods attain good reconstruction accuracy, and when they fail, the RF distance is small, which indicates minor misplacement errors of only one or two haplotypes. It is clear that TCS performs considerably worse than any of the other methods, while most of the ML, MP and NJ methods have approximately equal performance with the exception of DNAPARS (from PHYLIP), which performs slightly better than other methods (see also Woolley *et al.* 2008 who obtained similar results in their simulations). Also, PAUPML consistently performed better than any other ML method used, although not as well as DNAPARS. Just as reported in Woolley *et al.* (2008), MS performed equal to or worse than TCS (data not shown).

Increasing Θ somewhat reduces accuracy across all methods, as does increasing the number of samples. Both raise the total haplotype count and this increases the size of the valid tree space. So this trend is to be expected.

One problem with parsimony methods is, however, that these often return more than one tree, so that the scientist is left with the decision of which tree to choose. This renders parsimony methods impractical in some cases. A similar problem is that a single tree can give rise to more than one haplotype genealogy. For the basic case of $\Theta = 0.01$ and no migration, about 40% of the returned trees result in a unique haplotype topology for the ML methods (Table 4). Parsimony methods, in contrast, had a similar percentage of simulations with more than one tree, while fewer of the returned results led to unique haplotype genealogies. DNAPARS trees gave a unique haplotype genealogy in <5% of the cases and PAUP* scored 12% for the same metric. TCS had the least number of returned spanning trees that resulted in unique haplotype genealogies (2.4%) and the largest average number of returned genealogies [2.7 (1.2)]. However, recall that we decompose the graph into all possible spanning trees, and this may cause a bias.

Migration results are interesting from the point of view that often, one is building a haplotype genealogy as a visual aid to infer possible migration patterns. With low migration rates the accuracy suffers. From Table 1 we see that, generally, we can determine when this might be the case from some very long internal branches. The general trends are the same, however, with all methods performing approximately equally well, except for TCS, which performed considerably worse, and DNAPARS, which performed slightly better than other methods.

Asymmetric migration does not change the general trends. But for an asymmetric migration rate of 10, we find the largest mean RF distance for all methods, and the lowest reconstruction efficiency. With an asymmetric migration rate of 100, we observe that the results are similar to the symmetric migration case with reconstruction efficiency improving slightly for the asymmetric case. As noted above, migration gives, with high probability, longer internal branches. Also, the performance of reconstruction methods is highly correlated with these long branches [i.e. the average longest branch for asymmetric migration rate $\lambda = 10$ is 24.4 (18.2)], which can be considered predominately a reconstruction problem in the presence of long branches. This

Table 3 Mean RF distance from true tree

Parameters	PhyML	IQPNNI	DNAML	PAUPML	DNAPARS	PAUP*	NJ	TCS
$\Theta = 0.01$	0.7 (1.3)	0.7 (1.3)	0.6 (1.3)	0.5 (1.1)	0.5 (1.0)	0.7 (1.3)	0.7 (1.3)	5.2 (5.4)
Γ ($\alpha = 2.5$)	1.2 (1.7)	1.2 (1.7)	1.2 (1.7)	1.6 (1.5)	0.5 (0.6)	0.4 (0.7)	1.2 (1.7)	3.6 (4.0)
Γ ($\alpha = 0.5$)	1.7 (2.0)	1.8 (2.0)	1.7 (2.0)	1.6 (1.9)	0.77 (1.2)	0.6 (1.1)	1.8 (2.0)	3.6 (4.2)
$\Theta = 0.0166$	1.2 (1.7)	1.2 (1.7)	1.1 (1.7)	1.1 (1.7)	0.8 (1.3)	1.1 (1.6)	1.2 (1.6)	12 (8)
$\Theta = 0.0166$, 500 samples	1.6 (2)	1.6 (2)	1.5 (2)	1.4 (1.9)	1.0 (1.6)	1.6 (2)	1.5 (1.9)	15.6 (9.4)
$\Theta = 0.01$, $\lambda = 10$	1.9 (2.2)	2.0 (2.4)	2.1 (2.4)	2.1 (2.4)	1.3 (1.9)	1.7 (2.1)	1.7 (1.9)	21 (7)
$\Theta = 0.01$, $\lambda = 100$	1.4 (1.8)	1.4 (1.8)	1.4 (1.8)	1.5 (1.9)	0.9 (1.5)	1.5 (1.9)	1.4 (1.8)	17 (8)
$\lambda_{1 \to 2} = 10$, $\lambda_{2 \to 1} = 0$	2.48 (2.7)	2.5 (2.6)	2.7 (2.9)	2.4 (2.7)	1.7 (2.2)	2.2 (2.4)	2.1 (2.9)	22.4 (6.4)
Γ ($\alpha = 1$)	3.7 (3.1)	4.0 (3.2)	4.0 (3.3)	3.9 (3.4)	1.9 (2.1)	1.8 (2.1)	3.3 (2.8)	17.6 (5.3)
$\lambda_{1 \to 2} = 100$, $\lambda_{2 \to 1} = 0$	1.4 (1.8)	1.3 (1.8)	1.3 (1.8)	1.2 (1.7)	0.8 (1.4)	1.4 (1.8)	1.3 (1.7)	15 (9)

Unless otherwise stated $\Theta = 0.01$ and there are 200 samples (and 1000 simulations for each parameter set). The standard deviations are in parenthesis. The same trends from Table 2 are apparent. Note that for TCS the distance is not a pure RF distance (section Definitions) because so many returned results are disjoint graphs. See the text for details. RF, Robinson-Foulds.

Table 4 Number of haplotype genealogies per returned tree

Method	% unique	Average count
PhyML	38.4	1.9 (1.1)
IQPNNI	40.8	1.8 (1.1)
DNAML	41.3	1.9 (1.1)
DNAPARS	4.4	1.9 (0.4)
PAUP*	12.7	2.3 (1.1)
NJ	39.9	1.9 (1.1)
TCS	2.4	2.7 (1.2)

conclusion is supported by a similar deterioration of performance with increasing Θ and samples. Fortunately, when such a haplotype genealogy is reconstructed, the long branches are easily observed and we can assume that confidence may not be high. However, we still note that even in the worst case the RF distance has a mean of about 2, and so most of the genealogies did indeed reflect the true genealogies.

If we consider ML methods only, we note the differences in many options on the programs used made very little difference to the overall performance. Interestingly, however, all programs but PAUPML perform identically, with PAUPML performing slightly better. What could be the cause of PAUPML's improved performance? We propose that the heuristic search algorithm that PAUP uses by default is more exhaustive compared with other methods, therefore permitting a wider topology search space at the expense of run time. To test this, the wide moves for DNAML were turned on for some parameter values (not shown) and a slight performance increase was observed, although still lower than PAUPML. In practice, the difference is small and real world effects are likely to be a larger issue.

Rate heterogeneity

Although simulations for all parameter sets for a *gamma* model were carried out, we only show some results for clarity. Performance of most methods decreased slightly for the $\alpha = 2.5$ case and substantially for a shape parameter of $\alpha = 0.5$. However, TCS was not affected as much as other approaches, and even showed a slight improvement in performance. This 'enhancement' simply reflects the greater number of spanning trees per data set retrieved under these conditions. However, TCS still scores lower than any other method considered. A similar increase in the number of trees returned occurred with the parsimony methods. For example, PAUP parsimony returns an average of 2.1 trees per data set in the normal case, which increases to 6.8 trees per data set with a shape parameter of $\alpha = 0.5$.

When a *gamma* model of evolution was used for the case of asymmetric migration ($\lambda = 10$), the performance of all methods dropped dramatically. This is because, although the number of mutations is the same, the number of altered sites is not. So the total number of segregating sites is smaller, and more importantly, pairwise differences tend to be fewer than in the homogeneous mutation model. Consequently, the probability of back mutations and other confounding factors increases dramatically when there are some very long branches in the tree. It should be noted that even with 'perfect' reconstruction the accuracy drops to 57.6% (Table 1) and poor overall performance is to be expected.

The final observation with a *gamma* model is that MP still scored better compared with ML. This is at odds with the general view that, as models get more complicated, parsimony should do worse. Yet, both methods are somewhat similarly affected. First, long branch attraction is unlikely to play a role here, as all data are clocklike. The second detail to note is that we are not dealing with species data, but population data, and that these data are 'oversampled'. By oversampled we mean that we expect to see a haplotype more than once in our sample. This tends to reduce the expected length of the longest branch in the genealogy, as there is a high probability that a branch will be broken up by a coalescent event. Another consideration is that we convert a topology to a haplotype genealogy by using Fitch branch lengths, and that the parsimony score is the sum of Fitch branch lengths. Thus, the heuristic search algorithm used may favour conversion to relevant haplotype genealogies as it uses a more correct optimality criterion. It is also interesting to note that the two best ML scores are provided by DNAML and PAUPML. This gives strong evidence that both methods use a much more exhaustive heuristic search to finding a tree, regardless of the optimality criteria used. This further benefits these methods.

Poor TCS performance and its consequences for nested clade phylogeographic analyses

On further inspection of the results from TCS, we noted that almost half the time for $\Theta = 0.01$ disjoint graphs were returned. This trend gets worse for larger Θ and with migration, returning as little as 3.6% fully connected graphs for $\lambda = 10$. For the case of asymmetric migration with $\lambda = 100$, 5.1% returned graphs had the true tree embedded within it. A fraction of 9.4% of the returned results were trees, 81.5% were disjoint graphs; the average number of spanning trees from each graph was 8.8. The large number of disjoint graphs, which cannot possibly give a correct result, penalizes TCS substantially. Note, however, that MS performs even worse in most parameter sets.

One consideration is that the incomplete graph components may contain a pruned true tree. It was found

that for no migration and $\Theta = 0.01$ 61.2% of the returned incomplete graphs contained the pruned true tree, 4.0 (4.4) haplotypes were pruned on average, and there were 2.4 spanning trees per component. This compares to 68.1% of correctly resolved haplotype genealogies for the next worse method, PhyML (Table 2). For asymmetric migration with $\lambda = 100$, 22.9% of the incomplete graphs contained the pruned true tree with an average of 2.5 (5.5) labels pruned, and an average of 7.7 (49) spanning trees found per graph. Again the next worse method, IQPNNI in this case, performs considerably better with 49.1% of returned results giving the correct haplotype genealogy without the complications of incomplete haplotype labels and networks. Other parameter values had similar results and do not improve TCS results compared to other methods.

One possible criticism of the current study is that we are looking at a parameter region where TCS will perform very poorly. TCS performs better for data sets with much higher sequence similarity. To rule this out, we performed simulations with much smaller Θ values: in particular $\Theta = 0.005$ and $\Theta = 0.002$. With a small Θ and the same number of samples (200), the number of observed haplotypes falls to small values, for example when $\Theta = 0.002$ the average number of haplotypes is only 6.5. Under these conditions, TCS performance improves considerably as do the other methods, but TCS is never better than any other method (data not shown). If we change any parameter to increase the number of observed haplotypes, the performance degrades as shown above.

This has implications on an ongoing debate about the validity of some of the standard approaches used by many researchers in the field of phylogeography and population genetics (see commentary by Knowles 2008). Using simulated data sets, it has been shown that the widely applied nested clade phylogeographic analysis (NCPA) (Templeton *et al.* 1995; Templeton 2004, 2008) fails to infer the correct historical processes in about three quarters of the cases (Knowles & Maddison 2002; Panchal & Beaumont 2007). Here we show that already the standard input for NCPA networks derived from e.g. TCS are somewhat problematic and more error-prone than genealogies reconstructed with phylogenetic algorithms. Thus, our results seem to support the view that phylogeography's standard repertoire of methods are not accurate and that a more thorough evaluation of widely applied analyses is needed (Petit 2007; Knowles 2008).

Open questions and problems

A series of problems remain when reconstructing genealogies from closely related haplotypic sequence data, which should be addressed in future studies. The first problem is the evaluation of the reliability of connections in haplotype genealogies. Bootstrapping methods and posterior probabilities are highly inefficient when dealing with closely related taxa and hence highly similar sequence data. Standard methods for comparing alternative topologies, such as the Shimodaira–Hasegawa test (Shimodaira & Hasegawa 1999), are similarly impractical. To apply at least some quality criterion on branches in a haplotype genealogy, Salzburger *et al.* (2003) have mapped the consistency index (Kluge & Farris 1969) for each mutation responsible for a connection in the genealogy. By doing so, diagnostic mutations occurring only once can be highlighted. The branches defined by such diagnostic mutations should be considered more reliable compared to branches defined by homoplasious mutations. Other diagnostic characters, such as insertions or deletions, might also be informative at the intraspecific level (in particular when using noncoding haplotype sequences such as the mitochondrial control region). Yet, algorithms are lacking that would adequately take into account the phylogenetic information provided by 'gaps'. With growing numbers of taxa sampled and haplotypes sequenced, the graphical representation of haplotype genealogies also becomes problematic. Finally, much more effort should be devoted to hypothesis testing approaches and the modelling of phylogenetic, population genetic and phylogeographic scenarios on the basis of haplotype genealogies (Knowles & Maddison 2002).

Conclusion and outlook

Our comparative phylogenetic analysis of simulated data sets with known genealogies revealed that traditional phylogenetic algorithms perform well in estimating haplotype genealogies, and that the DNAPARS phylogeny most often leads to the correct tree. In cases where these methods failed, the RF distance was small, indicating only small errors. All ML methods worked approximately equally, with PAUPML performing slightly better in all cases. Surprisingly, the ML methods had similar performance to NJ. MP did perform better than ML methods in general, although the difference was not large. The good performance of parsimony in situations with low mutation count has been reported previously by DeBry & Abele (1995). In this regime, there is actually some equivalence between ML and MP (see Steel & Penny 2000 and references therein). The disadvantage of MP is that it suffers from the sometimes large amounts of equally good optimal trees, making practical application of these methods more difficult. Most importantly, we find that all traditional phylogenetic methods outperform TCS in recon-

structing the true haplotype genealogy. A main problem with TCS is that it frequently produces disjoint graphs. Even allowing for incomplete trees, TCS still did not perform as well as any other method tested. Although we have primarily considered SP as implemented in TCS here, we do not expect other network methods such as MS or MJ to perform significantly better, as it has previously been shown that these perform similarly (Cassens *et al.* 2005). This poses questions about empirical studies that employ TCS (Knowles 2008).

We have implemented a small program that creates high-quality haplotype geneologies from tree data and runs on Mac, Linux and Windows. It has a wide range of options and can output publication quality figures. The program is freely available at http://www.cibiv.at/~greg/haploviewer with instructions. This software is still under development.

Acknowledgements

We thank Jayne Ewing for the help provided for manuscript preparation and four anonymous reviewers and the Subject Editor Francois Rousset for valuable comments. WS is supported by the European Research Council (ERC), Starting Grant 'INTERGENADAPT', the University of Basel and the Swiss National Science Foundation (grant 3100A0_122458). GBE and AvH are supported by the Bioinformatics Integration Network II project and the WWTF (Vienna Science and Technology Fund).

References

Alonso S, Armour JAL (2004) Compound haplotypes at Xp11.23 and human population growth in Eurasia. *Annals of Human Genetics*, **68**, 428–437.

Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics*, **141**, 743–753.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A (2006) Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, **439**, 719–723.

Beaumont MA, Panchal M (2008) On the validity of nested clade phylogeographical analysis. *Molecular Ecology*, **17**, 2563–2565.

Beaumont MA, Nielsen R, Robert C *et al.* (2010) In defence of model-based inference in phylogeography. *Molecular Ecology*, **19**, 436–446.

Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences of the USA*, **76**, 1967–1971.

Cassens I, Mardulyn P, Milinkovitch MC (2005) Evaluating intraspecific 'network' construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? *Systematic Biology*, **54**, 363–372.

Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1660.

DeBry RW, Abele LG (1995) The relationship between parsimony and maximum–likelihood analyses: tree scores and confidence estimates for three real data sets. *Molecular Biology and Evolution*, **12**, 291–297.

Ewing GB, Nicholls GK, Rodrigo AG (2004) Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics*, **168**, 2407–2420.

Excoffier L, Smouse PE (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics*, **136**, 343–359.

Excoffier L, Novembre J, Schneider S (2004) Simcoal: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity*, **91**, 506–509.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (version 3.2). *Cladistics*, **5**, 164–166.

Felsenstein J (2003) *Inferring Phylogenies*, 2nd edn. Sinauer Associates, Sunderland, MA, USA.

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Systematic Biology*, **19**, 99–113.

Gibb GC, Kardailsky O, Kimball RT, Braun EL, Penny D (2007) Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. *Molecular Biology and Evolution*, **24**, 269–280.

Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle CA (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of lycaeides butterflies. *Molecular Ecology*, **19**, 2455–2473.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B*, **270**, 313–321.

Holsinger KE (2010) Next generation population genetics and phylogeography. *Molecular Ecology*, **19**, 2361–2363.

Hudson RR (1990) Gene genealogies and the coalescent process. vol. 7, pp. 1–44.

Huson DH, Klöpper T, Lockhart PJ, Steel MA (2005) Reconstruction of Reticulate Networks from Gene Trees. In *Research in Computational Molecular Biology*, (eds. Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner P, Waterman M), pp. 233–245.

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In *Evolution of Protein Molecules* (ed. Munro HN), vol. 3, pp. 21–132.

Kingman J (1982a) The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Kingman JFC (1982b) On the genealogy of large populations. *Journal of Applied Probability*, **19**, 27–43.

Kluge AG, Farris JS (1969) Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, **18**, 1–32.

Knowles LL (2008) Why does a method that fails continue to be used? *Evolution*, **62**, 2713–2717.

Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, **20**, 86–93.

Lindblad-Toh K, Wade CM, Mikkelsen TS *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.

Maddison WP (1997) Gene trees in species trees, pp. 523–536.

Minh BQ, Vinh LS, von Haeseler A, Schmidt HA (2005) pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics*, **21**, 3794–3796.

Ota T, Nei M (1994) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *Journal of Molecular Evolution*, **38**, 642–643.

Panchal M, Beaumont MA (2007) The automation and evaluation of nested clade phylogeographic analysis. *Evolution*, **61**, 1466–1480.

Petit RJ (2007) The coup de grce for the nested clade phylogeographic analysis? *Molecular Ecology*, **17**, 516–518.

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution*, **16**, 37–45.

Posada D, Crandall KA (2002) The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, **54**, 396–402.

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

Salzburger W, Brandstätter A, Gilles A *et al.* (2003) Phylogeography of the vairone (*Leuciscus souffia*, Risso 1826) in Central Europe. *Molecular Ecology*, **12**, 2371–2386.

Savolainen P, Zhang Y, Luo J, Lundeberg J, Leitner T (2002) Genetic evidence for an East Asian origin of domestic dogs. *Science*, **298**, 1610–1613.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629–644.

Schneider S, Roessli D, Excoffier L (2000) Arlequin: a software for population genetics data analysis. ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, **16**, 1114–1116.

Shioura A, Tamura A (1995) Efficiently scanning all spanning trees of an undirected graph. *Journal of the Operation Research Society of Japan*, **38**, 331–344.

Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution*, **17**, 839–850.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, **73**, 1162–1169.

Storey AA, Ramirez JM, Quiroz D *et al.* (2007) Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. *Proceedings of the National Academy of Sciences of the USA*, **104**, 10335–10339.

Sun S, Greenwood CMT, Neal RM (2007) Haplotype inference using a bayesian hidden markov model. *Genetic Epidemiology*, **31**, 937–948.

Swofford DL (1993) PAUP: phylogenetic analysis using parsimony, V3.1.1.

Tautz D, Ellegren H, Weigel D (2010) Next generation molecular ecology. *Molecular Ecology*, **19**, 1–3.

Templeton AR (2004) Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology*, **13**, 789–809.

Templeton AR (2008) Nested clade analysis: an extensively validated method for strong phylogeographic inference. *Molecular Ecology*, **17**, 1877–1880.

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**, 619–633.

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Verheyen E, Salzburger W, Snoeks J, Meyer A (2003) Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. *Science*, **300**, 325–329.

Woolley SM, Posada D, Crandall KA (2008) A comparison of phylogenetic network methods using computer simulation. *PLoS ONE*, **3**, e1913.

W.S. is Assistant Professor (with tenure-track) at the Zoological Institute of the University of Basel. The research of his team focuses on the understanding of the genetic basis of adaptation, evolutionary innovation and animal diversification. The main model systems in the group are the astonishingly diverse adaptive radiations of cichlid fishes in East Africa. The laboratory's homepage at http://www.evolution.unibas.ch/salzburger/ provides further details on the group's (research) activities. A.v.H. is Professor for Bioinformatics at the University of Vienna, the Medical University Vienna and the University of Veterinary Medicine, Vienna. His team works on a variety of bioinformatics topics among them computer methods for molecular evolution. The homepage (http://www.cibiv.at) provides further details on the research activities.
