# Inferring Population Phylogeny from Genetic Data

Senri Ethnological Studies 98: 25–37 ©2018
*Let's Talk about Trees: Genetic Relationships of Languages and Their Phylogenic Representation*
Edited by Kikusawa Ritsuko and Lawrence A. Reid

# 3. Inferring Population Phylogeny from Genetic Data

**Kimura Ryosuke**
*University of the Ryukyus, Japan*

## Abstract

In the field of evolutionary genetics, researchers estimate phylogenetic relationships among populations by reconstructing phylogenetic trees based on molecular data. However, a molecular phylogenetic tree can be accurate but have a topology that is not consistent with the population tree topology. Due to this phenomenon, called "incomplete lineage sorting", an analysis of a number of loci is required to accurately ascertain the phylogenetic relationships among populations. Another caution in phylogenetic reconstruction is that trees can only represent branching, not linkages between lineages (i.e., recombination in molecules and gene flow and admixture in populations). Phylogenetic networks are an effective method for representing the presence of reticulate events. However, complex demographic histories of populations are usually reconstructed by adopting model-based approaches where a small number of likely population topologies are specified in advance, and the best-fit demographic model and parameters are estimated.

## 3.1. Introduction

Phylogenetics is the field of biology concerned with reconstructing the evolutionary relationships among groups of organisms. The discipline of phylogenetics has been developed mainly in studies of taxonomy and molecular evolution, and it has been applied to other academic fields, such as linguistics. Over the past few decades, numerous algorithms for efficient and accurate reconstruction of phylogenetic diagrams have been proposed. Increased computer processing power and the development of generalized programs allow greater access for performing phylogenetic analysis. In this paper, I present considerations of the reconstruction and interpretation of phylogenetic diagrams.

## 3.2. Reconstruction of Molecular Phylogenetic Trees

Several books written by biologists and statisticians give detailed descriptions of methods for the efficient and accurate reconstruction of molecular phylogenetic trees (Felsenstein

2004; Nei and Kumar 2000). Here, I provide a brief introduction to these methods. The approaches for inferring molecular phylogenetic trees are classified by the data used to produce them and the principles and algorithms employed to arrive at the best tree topology.

### 3.2.1 Distance-based Methods

Distance-based methods, which are also classified as the phenetic approach, require determination of pairwise distances between operational taxonomic units (OTU*s*) prior to reconstruction of the phylogenetic tree.

UPGMA (unweighted pair-group method using arithmetic averages) is a simple agglomerative clustering method that assumes a constant rate of evolution, and thus produces a rooted tree (Sneath and Sokal 1973). Using UPGMA is fast, but caution is needed because incorrect trees will be produced when the assumption is violated.

The least squares (LS) approach depends on the principle of minimizing differences between the observed pairwise distances and distances over the entire phylogenetic tree. THE ORDINARY LS METHOD estimates branch lengths by minimizing the unweighted sum of the squared errors (Cavalli-Sforza and Edwards 1967). In THE WEIGHTED LS METHOD the squared errors are weighted several different ways (Fitch and Margoliash 1967). These methods assume independent estimates of pairwise distances. However, when the path between two OTUs shares any branches with a path between another two OTUs, the pairwise distances will be correlated. Covariances of pairwise distances can be taken into account using the GENERALIZED LS METHOD (Bulmer 1991). In the LS approach, all the possible branching patterns (TOPOLOGIES) are searched, and the FITCH AND MARGOLIASH METHOD algorithm is used to efficiently obtain the LS tree (Fitch and Margoliash 1967). This algorithm utilizes the fact that branch lengths between any three OTUs are unambiguously determined.

The MINIMUM EVOLUTION (ME) METHOD is based on the principle of seeking the tree having the minimum sum of branch lengths (Edwards and Cavalli-Sforza 1963; Rzhetsky and Nei 1993). In the same way as the LS method, the ME method requires an exhaustive search of all possible topologies, which results in a long computation time. THE NEIGHBOR-JOINING (NJ) METHOD is an algorithm that efficiently produces a phylogenetic tree based on the ME principle (Saitou and Nei 1987). The NJ method uses an agglomerative process that is like that of UPGMA, but it does not require the assumption of a constant rate of evolution. Since this method provides fast and accurate reconstruction of phylogenetic trees, it is the most widely used method among the distance-based methods.

### 3.2.2 Character-based Methods

Character-based methods, which are also called sequence-based methods and are classified as taking a cladistic approach, use the original information of characters, while distance-based methods discard the character data once the distance matrix has been generated.

The principle of THE MAXIMUM PARSIMONY (MP) METHOD is that the least complex

explanation for an observation is the preferred explanation (Henning 1966). Applied to phylogenetic analysis, the MP principle is geared to finding the tree with the minimum number of evolutionary changes for a given set of aligned sequences. When the number of OTUs is small, an exhaustive search for all the possible topologies is feasible. However, when the number of OTUs is ten or more, an exhaustive search carries with it a large computational load. There are two types of solutions for this problem: the BRANCH-AND-BOUND ALGORITHM and THE HEURISTIC SEARCH APPROACH. The branch-and-bound algorithm, in which large subsets of fruitless candidate trees are discarded, is guaranteed to find the minimal tree without having to evaluate all possible trees (Hendy and Penny 1982). As the following heuristic search algorithms have been developed, HEURISTIC SEARCH APPROACHES, STEPWISE ADDITION, BRANCH SWAPPING, and BRANCH-AND-BOUND-LIKE ALGORITHMS (Kumar et al. 1994; Maddison and Maddison 1992; Swofford 1998), they are also being utilized in distance-based methods.

With the use of THE MAXIMUM LIKELIHOOD (ML) METHOD, evolutionary events are described under a probabilistic model and the tree judged to have the maximum likelihood is chosen (Cavalli-Sforza and Edwards 1967; Felsenstein 1981). This method, which utilizes all of the original information of characters, usually provides the most robust result but is computationally very intensive and thus extremely slow.

THE BAYESIAN METHOD for phylogeny reconstruction produces a posterior probability distribution of trees produced from the data and a prior distribution of models and parameters (Huelsenbeck et al. 2001; Rannala and Yang 1996). This method has become possible due to advances in computational techniques of statistics, particularly, MARKOV CHAIN MONTE CARLO (MCMC) algorithms; the implementation of MCMC to estimate posterior distribution eliminates much of the complex summation and integration.

## 3.3. Inconsistencies between Molecular and Population Trees: Incomplete Lineage Sorting

The use of molecular phylogenetic trees, or GENE GENEALOGY, based on a single genomic region can produce an accurate reconstruction by the above-mentioned methods, if sequences in the region have enough informative sites and do not contain any recombination. To infer the history of multiple populations or species, molecular data is sampled from all of the populations or species that are examined. However, caution is observed in cases for which the molecular tree is not consistent with the phylogenetic relationships observed between the populations/species. Such systematic error is called "INCOMPLETE LINEAGE SORTING" (Maddison 1997; Maddison and Knowles 2006; Pamilo and Nei 1988; Rosenberg 2002; Takahata 1989). The transfer of genetic information from one generation to the next is a stochastic process. Some individuals leave multiple descendants and others do not leave any. In addition, of the two homologous chromosomes present in an individual, only one is passed to the offspring. The process of gene transmission, together with the occurrence of mutations resulting from occasional errors in DNA replication, generates patterns of genetic variation within a population (Figure 3-1a). Looking backward in time through the generations is an efficient way to

describe the probability that the observed gene genealogy will be realized under a certain demographic model. This approach is known as coalescent theory in population genetics (Hudson 1983; Kingman 1982; Tajima 1983). It is worth noting that as different genomic regions have different gene genealogies, the time to the most recent common ancestor (MRCA) differs among genomic regions in a population, and the expected time to MRCA depends on the effective population size. Incomplete lineage sorting can occur due to genetic variation (polymorphisms) in the ancestral population.

Let us consider a simple case in which one DNA sequence is sampled from each of three populations, X, Y, and Z, as shown in Figures 3-1b–3-1e. In contrast to the cases shown in Figures 3-1b and 3-1c, the topology of the molecular tree is inconsistent with that of the population tree in Figures 3-1d and 3-1e. Incomplete lineage sorting can be the result of sequences from X and Y being older than the population divergence time between Z and others (T1). For this situation, the occurrence probabilities for the three cases of Figures 3-1c–3-1e are equivalent and, consequently, the use of an appropriate

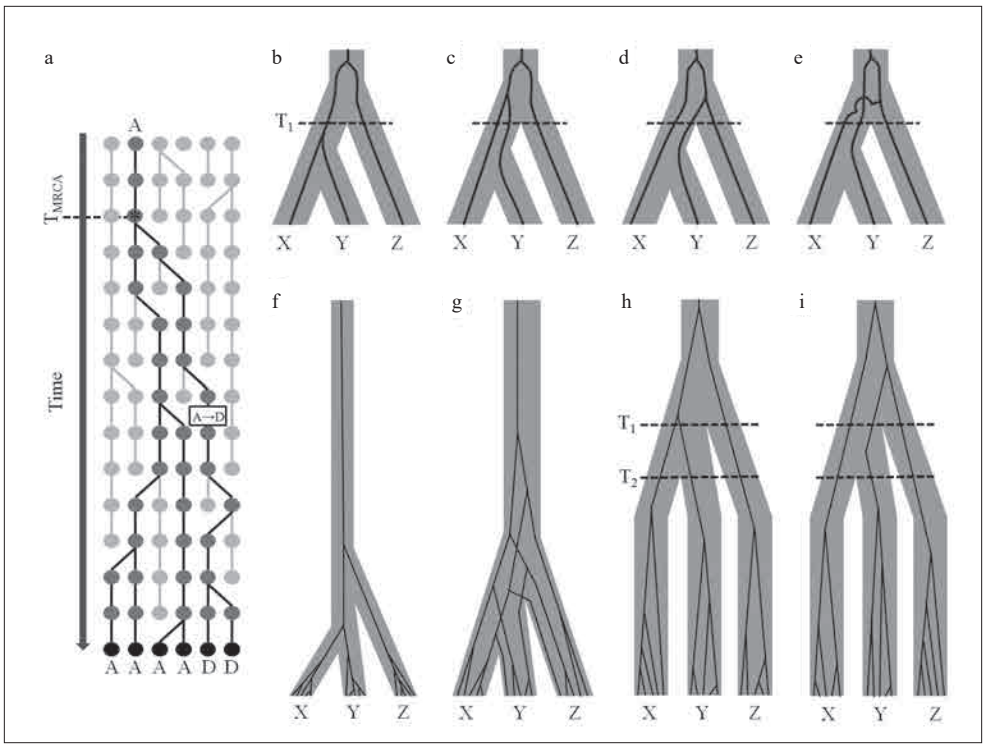


**Figure 3-1** Molecular and population trees. (a) Gene genealogy in a population. (b–e) One DNA sequence is sampled from each of three populations, X, Y, and Z. The topology of the molecular tree is consistent (b and c) or inconsistent (d and e) with that of the population tree. Occurrence probabilities for cases c, d, and e are equivalent. (f–i) Multiple sequences at a homologous locus are sampled from each population. (f) A case where the molecular tree exhibits population-specific clades. (g) A case where different populations share common variation. (h and i) Molecular divergence patterns in three species/populations (constructed by the author)

distance-based method and merging the data from a number of loci provides the reconstruction of an accurate population tree from molecular data, even with incomplete lineage sorting.

Next, we will consider the case of sampling multiple sequences at a homologous locus from each population. If the time to the MRCA in a particular population is more recent than the population divergence, the molecular tree clades will be population-specific (Figure 3-1f). In contrast, population divergence that is more recent than the time to the MRCA means that different populations will share the same variation (Figure 3-1g). In this case, it is possible that a sequence in a population will be more distant to another sequence from the same population than to a sequence from another population. Figures 3-1h and 3-1i show the molecular and species divergence patterns observed for humans, chimpanzees, and gorillas (Gibbs and Rogers 2012; Scally et al. 2012). Although the molecular tree shows population-specific clades, incomplete lineage sorting can occur. This phenomenon is due to the interval between species divergences (T1 and T2) being shorter than the molecular divergence time in the ancestral population of X and Y.

In summary, incomplete lineage sorting can be observed for polymorphic characters in the common ancestral population and for multiple divergences of the populations occurring within a short timeframe. In fields other than evolutionary genetics, such as linguistics, researchers may encounter cases in which different trees are produced for the examination of different characters (i.e., words in linguistics). These cases may be evidence not for the transfer of characters from one population to another (i.e., word borrowing) but for incomplete lineage sorting, if some of the characters used to produce the trees can be assumed to show a polymorphic state in the ancestral population. Although there are some differences between genes and words, as in how they are transmitted, incomplete lineage sorting must also be taken into account in linguistic studies.

## 3.4. Phylogenetic Networks Representing Hybrid Populations

One of the most critical problems for producing phylogenetic trees is that the tree topology can represent only branching patterns. However, making the assumption that evolution is a simple branching process is often unrealistic; evolution is thought to be better represented as a branching-and-joining pattern, or reticulate, process. Recently, in order to describe RETICULATE EVOLUTION, a number of methods for drawing PHYLOGENETIC *networks* have been proposed. A detailed overview of these methods is provided in a book by Huson *et al*. (2010). In the evolution of organisms, there are several levels of RETICULATE EVENTS: recombination for the molecular level, gene flow and admixture for the population level, and horizontal gene transfer and hybridization for the species level. The methods appropriate for drawing a network are dependent on what level is being examined.

When attention is focused on the phylogenetic relationships among populations, how can reticulate events be detected from among the data of multiple independent characters? As explained in the previous section, one cannot judge the presence of reticulate events

only by observing that different characters exhibit different tree topologies because incomplete lineage sorting alone can explain this observation. To show evidence of a reticulate event, one needs to detect an incompatibility among pairwise distances between populations WITH THE ASSUMPTION OF ONLY A SIMPLE BRANCHING PROCESS.

Figure 3-2 shows the results of a previous simulation study (Kimura 2013) in which phylogenetic trees and networks for different demographic models were reconstructed using the NJ method and the neighbor-net (NN) methods (Bryant and Moulton 2004), respectively, using the distance matrix. In a model lacking any reticulate events (Model S; Figure 3-2a), the phylogenetic tree well represents the original model (Figure 3-2d), and the network assumes a shape similar to that of the NJ tree, without any reticulate

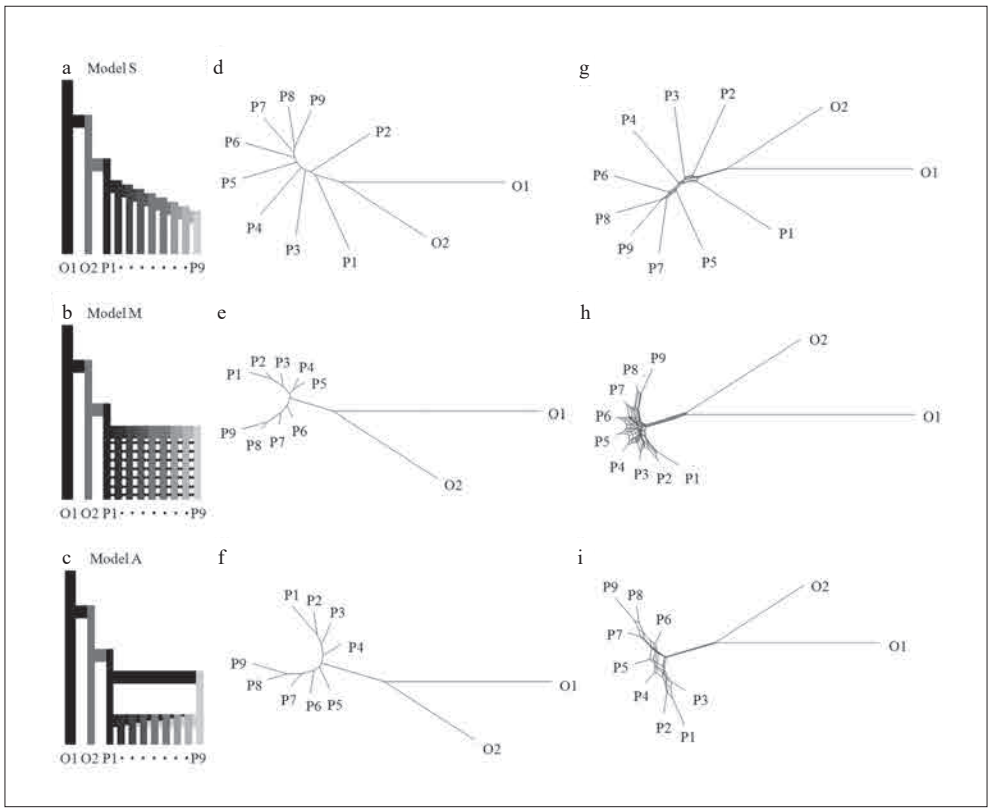


**Figure 3-2** Phylogenetic trees and networks drawn from simulated demographic histories (Kimura 2013). Nine populations and two out-groups were included in the simulations. (a) Serial splits model (S): in-group populations split off one-by-one in a temporal series. (b) Migration model (M): a population splits into nine populations simultaneously, and continuous gene flows occur at migration rate m between neighboring populations as in a one-dimensional stepping-stone model. (c) Admixture model (A): a population had split, and the two resulting populations (P1 and P9) were separated for a time; each of seven other populations was then generated by a distant single event of population admixture with different proportions of the two parental populations (7:1, 6:2, 5:3, 4:4, 3:5, 2:6, 1:7). (d–f) Neighbor-joining trees for the models S (d), M (e), and A (f). (g–i) Neighbor-net networks for the models S (g), M (h), and A (i) (constructed by the author)

structures (Figure 3-2g). In contrast, in a model with gene flows between populations (Model M; Figure 3-2b) and a model with population admixtures (Model A; Figure 3-2c), the networks with reticulation structures are reconstructed (Figures 3-2h and 3-2i), while the trees are distorted because of the discrepancies among the pairwise distances between populations (Figures 3-2e and 3-2f). When a tree analysis is applied to populations that have a history of admixture, caution regarding the distortion of the tree is warranted, such that the most mixed population is prone to earlier branching out from the clusters. Some previous studies employing phylogenetic tree analysis may have been misleading for this reason.

In an example of the analysis of real data (Figure 3-3), single nucleotide polymorphism (SNP) data of Indonesian and Melanesian populations from The HUGO Pan-Asian SNP Consortium (2009) were reanalyzed using an African population, the Yoruba, as an outgroup. The reconstructed NJ tree (Figure 3-3a) and NN network (Figure 3-3b), showing patterns similar to the tree in Figure 3-2f and the network in Figure 3-2i, indicated that the populations in the Lesser Sunda Islands (Alorese, Lembata, Lamaholot, Manggarai, and Kambera) are likely to have been formed by admixtures in the past between two different populations, known as Austronesians and non-Austronesians. Caution is needed in that, if only the tree is shown, different and incorrect interpretations may be made.

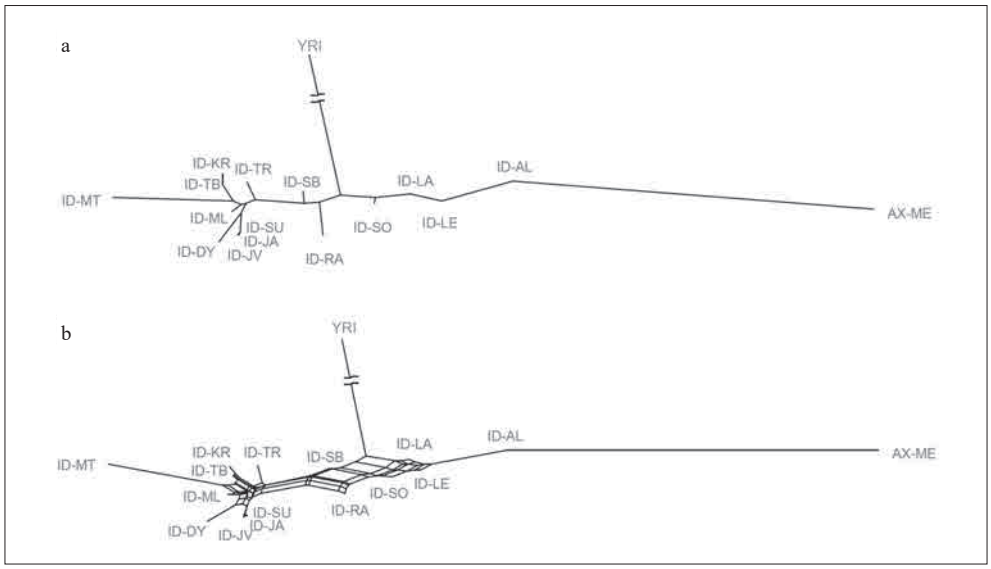


**Figure 3-3** Phylogenetic analysis of SNP data of Indonesian and Melanesian populations. Data were extracted from The HUGO Pan-Asian SNP Consortium (2009) and reanalyzed using Yoruba (YRI) as an out-group. (a) Neighbor-joining tree. (b) Neighbor-net network. AX-ME, Melanesian; ID-AL, Alorese; ID-LE, Lembata; ID-LA, Lamaholot; ID-SO and ID-RA, Manggarai; ID-SB, Kambera; ID-MT, Mentawai; ID-TR, Toraja; ID-ML, Malay; ID-KR, Batak Karo; ID-TB, Batak; ID-DY, Dayak; ID-JA and ID-JV, Javanese; ID-SU, Sundanese (constructed by the author)

### 3.5. Model-based Inference of Population Demographic History

If one assumes simple demographic histories of populations without reticulate events and with only a few unknown demographic parameters, the real population topology and parameters can be estimated using a relatively small data set and with relatively simple processes. However, under the assumption of much more complex histories with respect to reticulate events and variations in population size, large data sets including a number of genomic regions are required, and even then, it is difficult to analytically reconstruct the population topology. For example, it would be hard to precisely reconstruct the original demographic models in Figures 3-2b and 3-2c only from the networks shown in Figures 3-2h and 3-2i. Although network analysis can detect the presence of reticulate events, the network does not provide details of reticulation events, e.g., how many and direction of the migration events, proportions of admixture, and identification of parental populations (the real parental populations often remain unexamined). To reconstruct a complex demographic history, therefore, population geneticists use MODEL-BASED APPROACHES, in which a small number of likely population topologies must be specified in advance and, among them, the best-fitting demographic model and parameters are estimated.

In recent years, a great deal of effort has been put toward developing statistical approaches that use genomic data to estimate demographic parameters in complex demographic models, including migration. *Moment-based methods*, which use only means and variances of divergences in allele frequencies, are computationally feasible, but these have only relatively low statistical power and limited application (Lipson et al. 2013; Patterson et al. 2012; Pickrell and Pritchard 2012; Reich et al. 2009). *Likelihood-based methods* are a widely-applied approach that aims to compute the likelihood, i.e., the probability of generating the observed data over multiple genomic regions under a given model and set of parameters (Beerli and Felsenstein 2001; Hey and Nielsen 2004; Hey 2005; Nielsen and Wakeley 2001; Nielsen and Beaumont 2009; Stephens and Donnelly 2000). Although it is ideal to compute the likelihood of a demographic model using all observed gene genealogies, it is not easy to apply such "full-likelihood" approaches for reconstructing complex histories. Since the likelihood cannot be derived analytically in most demographic models, the computation of likelihood relies on simulations that explore highly dimensional parameter space. However, this strategy is computationally inefficient and, thus, expensive because the probability of exactly realizing each gene genealogy under a given model is very small. To overcome this problem, much focus has recently been placed on APPROXIMATE BAYESIAN COMPUTATION (ABC) *methods*, which use summary statistics instead of gene genealogies to find the parameter values that generate data sets with high similarity to the observed data sets (Balding and Nichols 1997; Beaumont et al. 2002; Beaumont 2010; Chikhi et al. 2001; Marjoram et al. 2003; Nielsen and Beaumont 2009; Wang 2003).

## 3.6. Conclusion

The population tree is not identical to trees based on characters. When the relationships among populations are estimated from their characters, researchers need statistical and probabilistic methods and must be careful of the presence of incomplete lineage sorting and reticulate events. Phylogenetic networks should be used in place of trees to elucidate relationships among populations because a history without reticulate events is unrealistic in many cases. Model-based approaches must be used to select the best-fitting model and to estimate parameters for populations having complex demographic history. Furthermore, the population history is sometimes too complicated to be represented as a simple diagram. Therefore, using phylogenetic analysis in combination with other statistical methods, such as the principle component analysis (Patterson et al. 2006; Price et al. 2006) and clustering analysis (Alexander et al. 2009; Pritchard et al. 2000; Tang et al. 2005), will be more effective for elucidating population history. Remarkable advances have been made in the previous two decades in statistical methods to efficiently and accurately infer the demographic history of and relationships among human populations from genomic variation. These methods may contribute to other fields of science, including linguistics.

## References

Alexander, D. H., J. Novembre, and K. Lange
2009 Fast Model-based Estimation of Ancestry in Unrelated Individuals. *Genome Research* 19(9): 1655–1664.

Balding, D. J. and R. A. Nichols
1997 Significant Genetic Correlations among Caucasians at Forensic DNA Loci. *Heredity* 78(6): 583–589.

Beaumont, M. A., W. Zhang, and D. J. Balding
2002 Approximate Bayesian Computation in Population Genetics. *Genetics* 162(4): 2025–2035.

Beaumont, M. A.
2010 Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution and Systematics* 41: 379–406.

Beerli, P. and J. Felsenstein
2001 Maximum Likelihood Estimation of a Migration Matrix and Effective Population Sizes in N Subpopulations by Using a Coalescent Approach. *Proceeding of the National Academy of Sciences USA* 98: 4563–4568.

Bryant, D. and V. Moulton
2004 Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution* 21: 255–265.

Bulmer, M.
1991 Use of the Method of Generalized Least Squares in Reconstructing Phylogenies From Sequence Data. *Molecular Biology and Evolution* 8: 868–883.

Cavalli-Sforza, L. L. and A. W. Edwards
1967 Phylogenetic Analysis. Models and Estimation Procedures. *American Journal of Human Genetics* 19(3): 233–257.
Chikhi, L., M. W. Bruford, and M. A. Beaumont
2001 Estimation of Admixture Proportions: A Likelihood-based Approach Using Markov Chain Monte Carlo. *Genetics* 158(3): 1347–1362.
Edwards, A. W. and L. Cavalli-Sforza
1963 The Reconstuction of Evolution. *Heredity* 18: 553.
Felsenstein, J.
1981 Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution* 17(6): 368–376.
2004 *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.
Fitch, W. M. and E. Margoliash
1967 Construction of Phylogenetic Trees. *Science* 155(3760): 279–284.
Gibbs, R. A. and J. Rogers
2012 Genomics: Gorilla Gorilla Gorilla. *Nature* 483: 164–165.
Hendy, M. D. and D. Penny
1982 Branch and Bound Algorithms to Determine Minimal Evolutionary Trees. *Mathematical Biosciences* 59(2): 277–290.
Henning, W.
1966 *Phylogenetic Systematics*. Urbana, IL: University of Illinois Press.
Hey, J.
2005 On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas. *PLoS Biology* 3(6): e193.
Hey, J. and R. Nielsen
2004 Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, with Applications to the Divergence of Drosophila Pseudoobscura and D. Persimilis. *Genetics* 167(2): 747–760.
Hudson, R. R.
1983 Testing the Constant-rate Neutral Allele Model with Protein Sequence Data. *Evolution* 37(1): 203–217.
Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback
2001 Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294(5550): 2310–2314.
Huson, D. H., R. Rupp, and C. Scornavacca
2010 *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge: Cambridge University Press.
Kimura, R.
2013 Interpretations of Practical Population Genetics: Analyses of Genome-wide SNP Data on Human Demography. In T. Akazawa, N. Ogihara, H. C. Tanabe, and H. Terashima (eds.) *Dynamics of Learning in Neanderthals and Modern Humans: Cognitive and Physical Perspectives*, vol. 2, pp. 105–117. Tokyo: Springer.

Kingman, J. F. C.
1982 On the Genealogy of Large Populations. *Journal of Applied Probability* 19A: 27–43.

Kumar, S., K. Tamura, and M. Nei
1994 MEGA: Molecular Evolutionary Genetics Analysis Software for Microcomputers. *Computer Applications in the Biosciences* 10(2): 189–191.

Lipson, M., P. R. Loh, A. Levin, D. Reich, N. Patterson, and B. Berger
2013 Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Molecular Biology and Evolution* 30(8): 1788–1802.

Maddison, W. P. and D. R. Maddison
1992 *MacClade: Analysis of Phylogeny and Caracter Evolution*. Sunderland, MA: Sinauer Associates.

Maddison, W. P.
1997 Gene Trees in Species Trees. *Systematic Biology* 46(3): 523–536.

Maddison, W. P. and L. L. Knowles
2006 Inferring Phylogeny Despite Incomplete Lineage Sorting. *Systematic Biology* 55(1): 21–30.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare
2003 Markov Chain Monte Carlo without Likelihoods. *Proceedings of the National Academy of Sciences USA* 100(26): 15324–15328.

Nei, M. and S. Kumar
2000 *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Nielsen, R. and J. Wakeley
2001 Distinguishing Migration from Isolation: A Markov Chain Monte Carlo Approach. *Genetics* 158(2): 885–896.

Nielsen, R. and M. A. Beaumont
2009 Statistical Inferences in Phylogeography. *Molecular Ecology* 18: 1034–1047.

Pamilo, P. and M. Nei
1988 Relationships Between Gene Trees and Species Trees. *Molecular Biology and Evolution* 5(5): 568–583.

Patterson, N., A. L. Price, and D. Reich
2006 Population Structure and Eigenanalysis. *PLoS Genetics* 2(12): e190.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich
2012 Ancient Admixture in Human History. *Genetics* 192(3): 1065–1093.

Pickrell, J. K. and J. K. Pritchard
2012 Inference of Population Splits and Mixtures from Genome-wide Allele Frequency Data. *PLoS Genetics* 8(11): e1002967.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich
2006 Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies. *Nature Genetics* 38(8): 904–909.

Pritchard, J. K., M. Stephens, and P. Donnelly
2000 Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155(2): 945–959.

Rannala, B. and Z. Yang
1996 Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference. *Journal of Molecular Evolution* 43(3): 304–311.
Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh
2009 Reconstructing Indian Population History. *Nature* 461: 489–494.
Rosenberg, N. A.
2002 The Probability of Topological Concordance of Gene Trees and Species Trees. *Theoretical Population Biology* 61(2): 225–247.
Rzhetsky, A. and M. Nei
1993 Theoretical Foundation of the Minimum-evolution Method of Phylogenetic Inference. *Molecular Biology and Evolution* 10(5): 1073–1095.
Saitou, N. and M. Nei
1987 The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* 4(4): 406–425.
Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G. Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Conner, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K. Wilson, M. H. Schierup, J. Rogers, C. Tyler-Smith, and R. Durbin
2012 Insights into Hominid Evolution from the Gorilla Genome Sequence. *Nature* 483: 169–175.
Sneath, P. H. A. and R. R. Sokal
1973 *Numerical Taxonomy*. San Francisco: Freeman.
Stephens, M. and P. Donnelly
2000 Inference in Molecular Population Genetics. *Journal of the Royal Statistical Society B* 62: 605–635.
Swofford, D.L.
1998 *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*. Sunderland, MA: Sinauer Associates.
Tajima, F.
1983 Evolutionary Relationship of DNA Sequences in Finite Populations. *Genetics* 105(2): 437–460.
Takahata, N.
1989 Gene Genealogy in Three Related Populations: Consistency Probability between Gene and Population Trees. *Genetics* 122(4): 957–966.
Tang, H., J. Peng, P. Wang, and N. J. Risch
2005 Estimation of Individual Admixture: Analytical and Study Design Considerations. *Genetic Epidemiology* 28(4): 289–301.

The HUGO Pan-Asian SNP Consortium
2009 Mapping Human Genetic Diversity in Asia. *Science* 326(5959): 1541–1545.
Wang, J. L.
2003 Maximum-likelihood Estimation of Admixture Proportions from Genetic Data. *Genetics* 164(2): 747–765.