DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Tangled Trees: The Challenge of Inferring Species Trees from Coalescent and Non-Coalescent Genes

Citation

Published version
https://doi.org/10.1007/978-1-61779-585-5_1

Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:33921640

Terms of use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at
https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility
https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story
The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

For the book *Evolutionary genomics: statistical and computational methods*

Tangled Trees: The challenge of inferring species trees from coalescent and non-coalescent genes

Christian Anderson[1], Liang Liu[2], Dennis Pearl[3] and Scott V. Edwards[1]

[1]Department of Organismic and Evolutionary Biology & Museum of Comparative Zoology,

Harvard University, Cambridge, MA 02138

[2]Department of Agriculture and Natural Resources, Delaware State University, Dover, DE 19901

[3]Department of Statistics, The Ohio State University, Columbus, OH 43210

Introduction

The concept of a "species tree", a bifurcating dendrogram graphically depicting the relationships of species to each other, is one of the oldest and most powerful icons in all of biology (Figures 1 and 2). After Charles Darwin sketched the first species tree (in *Transmutation of Species*, Notebook B, 1837), he remained fascinated by the image for 22 years, eventually including a species tree as the only figure in *On the Origin of Species* (1859). Though species trees reached their aesthetic apogee with Ernst Haeckel's *Tree of Life* in 1886, the pursuit of ever-more scientifically accurate trees has kept phylogenetics a vibrant discipline for the 150 years since.

Because the direct evolution of species is not observable (not even in the fossil record), relationships are often inferred by shared characteristics among extant taxa. Until the 1970s, this was done almost exclusively by using morphological characters. Although this approach had many successes, the paucity of characters and the challenges of comparing species with no obvious morphological homologies were persistent problems (Hillis 1986). When molecular techniques were developed in the late 1960s, it soon became clear that the sheer volume of molecular data that could be collected would represent a vast improvement. When DNA sequences became widely available for a range of species (Kocher et al. 1990), molecular comparisons quickly became *de rigueur* (Miyamoto and Cracraft 1990; Swofford, *et al.* 1996; Nei 1996; and Nei and Kumar 2000). Nonetheless, it was recognized early on that molecular phylogenies had their own suite of problems; the concept that not all gene tree topologies would

match the true species tree topology (i.e., would not be speciodendric *sensu* Rosenberg 2002) was implicit in studies as early as the 1960s (Cavalli-Sforza 1964; see also Avise et al. 1987). However, it was generally assumed that the idiosyncratic genealogical history of any one gene, as reconstructed from extant mutations, was an acceptable approximation for the true history of the species given the potentially overwhelming quantity and seductive utility of molecular data (Tajima 1983, Pamilo and Nei 1988; Takahata 1989; Avise 1994; Wollenberg and Avise 1998).

By and large, the ensuing decades of molecular phylogenetics has fulfilled much of this potential, revolutionizing taxonomies and resolving conundrums previously considered intractable (Gould 2001). However, as the amount of genetic data per species becomes ever-more voluminous, it has become clear that individual genes can conflict with each other and with the overarching species tree, both in topology and branch lengths (Maddison 1997; Jennings and Edwards 2005; Carstens and Knowles 2007; Wong *et al.* 2007). In the mean time, the term 'phylogeny' frequently became conflated with 'gene tree', the entity produced by many of the leading phylogenetics packages of the day. The term 'species tree', in use since the late 1970s to emphasize the distinction between lineage histories and gene histories (reviewed in Avise 1994; Maddison 1997), was only gradually acknowledged, despite the fact that species trees are the rightful heirs to the term 'phylogeny' and better encapsulate the true goals of molecular and morphological systematics (Edwards 2009; Figures 1 and 2).

At first, some researchers treated this phenomenon as though it were an information problem: when working with only a few mutations, you were bound to occasionally get unlucky and sequence a gene whose random signal of evolution did not match that of the taxa being studied.

The reasoning was, surely more and/or longer sequences would fix that problem and cause gene trees to converge. However, as more genes were sequenced, and as the properties of gene lineages within populations were studied in detail (e.g., Neigel and Avise 1986), the twin realities of gene tree heterogeneity and "incomplete lineage sorting" became clear (Figures 1 and 2). The probability of an event such as incomplete lineage sorting, which if considered alone would lead to inferring the wrong species tree, was worked out theoretically for the four individual/two species case first (Tajima 1983), followed by the three individual/three species case (Nei 1987; Takahata 1989), and then the generalized case (Pamilo and Nei 1988). This last study was among those that proposed that the solution was to simply acquire more gene sequences, after which the central tendency of this gene set would point to the correct relationships. On the empirical side, researchers adopted two general approaches. Pamilo and Nei (1988) suggested a "democratic vote" method, where each gene was allowed to propose its own tree, and the topology with the most "votes" was declared the winner, and therefore the true phylogeny. This method was used in theoretical and empirical work, particularly on primate data sets (Satta *et al.* 2000). However, though generally true for three-species cases, it can sometimes produce the wrong topology with four or more species. In fact, we now know that there is an "anomaly zone" for species trees with short branch lengths as measured in coalescence units, in which the addition of more genes is guaranteed to lead to the wrong species tree topology for the democratic vote method (Degnan and Rosenberg 2006, Rosenberg and Tao 2008). (Coalescent time units, equivalent to t/Ne where t is number of generations since divergence and Ne is the effective population size of the lineage. For a clear explanation, see Degnan and Rosenberg 2009). Though it is not clear that real species trees often display branch lengths short enough to

enter the anomaly zone (Huang and Knowles 2009), the possibility remains theoretically disconcerting. In addition, because the number of possible tree topologies increases as the double factorial of the number of tips, for species trees with more than four tips a very large number of genes is required to determine which gene tree is in fact the most frequent. Advanced consensus methods (Bryant 2003, Felsenstein 2004) have recently been introduced, and circumvent some of the problems of the democratic vote by using novel assembly methods, such as rooted triple consensus (Ewing *et al.* 2008), greedy consensus (Degnan *et al.* 2008), or supertree methods (eg Steel and Rodrigo 2008, Ranwez *et al.* 2010).

The second empirical approach to the problem of conflicting gene trees was to bypass it altogether. Concatenation methods appended one gene's sequence onto the next, to create long alignments or super-matrices (Weins 2003), a technique that in some situations was superior to standard consensus methods in resolving discordance or achieving statistical consistency (Gadagkar *et al.* 2005). But some researchers, including those who questioned the 'total evidence' approach to systematics (e.g., Bull,et al. 1993), advocated against concatenation when, for whatever reason, gene trees appeared to conflict with one another. One problem with the concatenation approach was that it assumed full linkage across the super-matrix, a situation that would obviously not be the case if genes were on different chromosomes. Even when the lineage lengths in a species tree are long in coalescent units, such that gene tree topologies are congruent, the branch lengths of trees of genes on different chromosomes will differ subtly from one another due to the stochasticity of the coalescent process. The early implementations of this method also assumed the same distribution of mutation rates across the sequence, which was clearly not the case if the matrix included coding and non-coding regions. Like democratic vote

methods, concatenation of many genes was sometimes defended as sufficient to override the conflicting signal across genes (Rokas *et al.* 2003; Driskell *et al.* 2004), despite widespread acknowledgement that gene tree heterogeneity is ubiquitous and that concatenation can sometimes give the wrong answer (Rokas 2006, Kubatko and Degnan 2007). Concatenation still remains popular by default (Wu and Eisen 2008), particularly among phylogenetic studies of higher taxa where incomplete lineage sorting is deemed rare. Another problem is that, in a strict sense, concatenation also does not generate species trees, in so far as the method treats all nucleotides as if they were part of a single non-recombining gene, and thus does not distinguish between gene and species trees (Edwards 2009). Some concatenation approaches also suffer from the same problem as democratic vote methods; in trees with short branches, more data can lead to the wrong answer with increasing confidence (Kubatko and Degnan 2007).

In the end the concatenation method will remain popular until there are software alternatives that are robust, efficient and easy-to-use. As a result, researchers are in something of a double-bind: either use just one gene and risk inferring the wrong species tree due to a lack of statistical power and incongruence with the underlying species tree, or use many genes and risk inferring the wrong species tree due to gene tree heterogeneity or short branches in some of the gene tree topologies. One solution is to use models for species trees that are consistent with what is known about biological heritability. One such model is the multispecies coalescent (Degnan and Salter 2005; Liu *et al.* 2008; Liu *et al.* 2009; Castillo *et al.* 2010). It is this model that provides the basis for a recent flurry of promising methods that permit efficient and consistent estimation of species trees under a variety of conditions.

The multispecies coalescent model

A plausible probabilistic model for analyzing multilocus sequences should involve not only the phylogenetic relationship of species (species tree), but also the genealogical history of each gene (gene tree), and allow different genes to have different histories. Unlike concatenation, such a model explains the evolutionary history of multilocus sequences through a two-stage process – from species tree to gene tree and from gene tree to sequences (Liu et al. 2009). Construction of the two-stage model requires an explicit description of how gene trees evolve in the species tree and how sequences evolve on gene trees. As the second question has been extensively studied in the traditional phylogenetic analyses for estimating gene trees, the key is to address the first question adequately. With a few exceptions (described below) the genealogical relationship (gene tree) of neutral alleles can be simply depicted by a coalescence process in which lineages randomly coalesce with each other backward in time. The coalescence model is simple in the sense that it assumes little or no effect of evolutionary forces such as selection, recombination, and gene flow, instead giving a prominent role to random genetic drift. Despite these seemingly oversimplified assumptions, the pure coalescent model is fundamental in explaining the gene tree – species tree relationship because it forms a baseline for incorporating additional evolutionary forces on top of random drift (Degnan and Rosenberg 2009). More importantly, the pure coalescent model provides an analytic tool to detect the evolutionary forces responsible for the deviation of the observed data (molecular sequences) from those expected from the model.

The coalescent process works, in effect, by randomly choosing ancestors with replacement from the population backwards through time for each sequence in the original sample. Eventually, two of these lineages will share a common ancestor, and the lineages are said to

"coalesce". The process continues until all lineages coalesce at the most recent common ancestor (MRCA). Multispecies coalescence works the same way but places constraints on how recently the coalescences occur, corresponding to the species' divergence times. Once a species tree has been proposed, the likelihood of each gene tree is evaluated; and these likelihoods are combined to evaluate the likelihood of the species tree. In this way, multispecies coalescent methods are the converse of consensus methods; rather than each locus proposing a potentially divergent species tree, a common species tree is assumed and evaluated, given the sometimes divergent patterns observed among multiple loci (Degnan and Rosenberg 2009).

A number of implementations of this idea have been developed (reviewed in Edwards 2009). The BATWING package (Wilson *et al.* 2003) was originally developed to generalize error estimates on a species tree from a single locus or group of 100% linked microsatellite loci (Wilson and Balding 1998). Several packages are available for moving from gene trees to species trees, including Minimization of Deep Coalescence (Maddison 1997; Maddison and Knowles 2006), STEM (Kubatko *et al.* 2009), JIST (O'Meara 2008 and 2010), GLASS (Mossel and Roch 2007), STAR, and STEAC (Liu *et al.* 2009). The MCMCcoal package (Rannala and Yang 2003) originally required a species tree topology *a priori* to approximate divergence times and population sizes, but now can infer species tree topologies as well (the "bpp" package, Yang and Rannala 2010). Several other full packages infer gene trees from DNA sequences, and then species trees from the inferred gene trees, given *a priori* assignment of the sequences to species groups. These include ESP-COAL (Carstens and Knowles 2007), AUGIST (within the Mesquite environment, Oliver 2008), BEST (Liu and Pearl 2007, Liu 2008), and *Beast (Heled and

Drummond 2010). Reviews describing these methods in more detail are available (Liu *et al.* 2009 and references therein).

The multispecies coalescent can under some circumstances be more efficient than concatenation (Edwards *et al.* 2007), and can recover the correct species tree even in the anomaly zone where concatenation methods fail (Liu & Edwards 2009). One drawback is that the estimation of larger numbers of parameters (population sizes and divergence times in addition to topologies) can slow computation and does not necessarily improve accuracy because of the many sources of error (Huang et al 2010). Another attractive aspect of species tree methods and multispecies coalescent models is that they appear to be less susceptible to the inflation of posterior probabilities that was early on attributed to Bayesian analyses (e.g., Suzuki et al. 2002). We have wondered (Edwards et al. 2007) whether such inflation is in fact due to traditional model misspecifications, such as incorrect substitution matrices for DNA sequences, or to concatenation, which of course can be viewed as a gross misspecification of the coalescent model. While the lower confidence values obtained from species trees are not deficiencies *per se*, they are also not conducive to the adoption of this new family of phylogenetic models by the empirical community!

Sources of gene tree / species tree discordance and violations of the multispecies coalescent model

Population processes

The "standard" and most common reason why gene trees are not speciodendritic is incomplete lineage sorting, i.e. lineages have not yet been reproductively isolated for long enough for drift to cause complete genetic divergence in the form of reciprocal monophyly of gene trees (Avise and Ball 1990; Figures 1 and 2). This source of gene tree heterogeneity is guaranteed to be ubiquitous, if only because it arises from the finite population sizes of all species that have ever come into existence. Almost all the techniques and software packages discussed above are designed to approximate uncertainties in species tree topology arising from this phenomenon.

Accurate delimitation of species and diverging lineages

For recent divergences, the definition of "species" can become problematic for species tree methods (O'Meara 2008, 2010), and the challenge of delimiting species has, if anything, increased now that the overly conservative strictures of gene tree monophyly as a delimiter of species have been mostly abandandoned. This fundamental issue in a phylogenetic study – whether the extent of divergence among lineages warrants species status – has not gone away in the species tree era. Researchers are often faced with a dilemma when deciding how deep a node must lie in a phylogeny in order to demonstrate genuine speciation. Each DNA sequence represents only one allele (and in some cases, only one mitochondrion within one cell of one individual, He *et al.* 2010), and because genetic diversity within species can be substantial, a few unfortunately selected representatives or an undersampling of a given species can lead to spurious species assignments. Simply avoiding the problem by calling groups of related individuals something else (such as operational taxonomic units (OTUs) or populations) does not

address the issue, because the key point is not so much whether the OTUs in a species tree study are genuine species, but whether or not gene flow has ceased (at least temporarily) at the time of sampling. Species trees need not use 'good' species as OTUs; they will work perfectly well on lineages that have recently diverged and ceased exchanging genes, but nonetheless are not sufficiently divergent as to be called species by other criteria. One common solution is to define speciation as occurring when both taxa in question are completely and reciprocally genetically isolated. However, this criterion is generally considered too conservative (de Querioz 2007), and fails to account for situations where genetic introgression occurs via a different mechanism than incomplete reproductive isolation.

The problem of species delimitation may ultimately be solved by data other than genetics, and today few species concepts use strictly genetic criteria (Hudson and Coyne 2002). Some have suggested that the line between a population-level difference and a species-level difference can be drawn empirically and with consistency in well-studied taxa such as birds, using morphological, environmental, and behavioral data simultaneously (Tobias *et al.* 2010). Thus, there is some hope that species delimitation can be performed rigorously *a priori* in some cases. Researchers who opt for delimiting species primarily with molecular data have a wide array of techniques and prior examples available to them (e.g., STRUCTURE, Pritchard et al. 2002; STRUCTURAMA, Huelsenbeck and Andalfatto 2007; Brownie, O'Meara 2010; rjMCMC, Yang and Rannala 2010, applied to geckos by Leaché and Fujita 2010; BEST-STEM approach, outlined by Knowles and Carstens 2007 and applied to Myotid bats by Carstens and Dewey 2010). Recent progress in species delimitation is motivated by the conceptual transition from "biological / reproductive isolation species" to the traditional "phylogenetic species" requiring

gene tree monphyly, and ultimately to the "lineage species concept", which defines species not in terms of monophyly of gene lineages but as population lineage segments in the species tree (de Queiroz 2007). Under that recently expanded concept, boundaries of species (i.e., lineages in the species tree) can be estimated from a collection of gene trees in the framework of the multispecies coalescent model (Knowles and Carstens 2007, Yang and Rannala 2010).

Gene flow

There are a number of other situations in which the assumptions of the coalescent are violated. A key assumption in most species tree methods developed thus far is whether or not gene flow is occurring between the taxa in a radiating clade. If some small amount of gene flow continues between species after divergence, then the multispecies coalescent can quickly destabilize, especially for a small number of loci and as the rate of genetic introgression increases (figure 6 in Wakeley 2000; Eckert and Carstens 2008). Further studies of the effect of gene flow on species tree inference are needed to determine the parameter space in which it is and is not a significant problem, and how sampling or analysis might ameliorate it.

Molecular processes

In addition to species delimitation and gene flow, there are at least three mechanisms that generate discordance on the molecular level (Figure 3). These include horizontal gene transfer (HGT), which poses a serious risk to phylogenetic analysis; gene duplication, whose risks can be avoided by certain models; and natural selection, which generally poses no direct threat but,

depending on its mode of action and consequences for DNA and protein sequences, can be the most challenging of all.

Horizontal gene transfer

HGT is now known to be so widespread in prokaryotes that a Tree of Life, even with reticulation, may not be an appropriate paradigm for these domains (Doolittle and Bapteste 2007; Boto 2010). Though generally ignored in eukaryotes, growing evidence shows that eukaryotic genomes contain substantial amounts of "uploaded" genetic material from bacteria, archea, viruses, and even fellow eukaryotes. Though gene sharing is most widespread between protists, it is also reasonably common between plant lineages (Andersson 2005), and has been documented in animals and fungi as well. For example, *Wolbachia* are able to insert their entire genomes (~1Mb) into the germlines of hosts, who still produced viable offspring in at least eight species of nematodes and insects (Hotopp *et al.* 2007). Transposable elements such as helitrons are continuously being shared among widely divergent eukaryotic lineages including fish and mammals, possibly using viruses as vectors (Thomas *et al.* 2010). Even though good techniques are not yet widely available for detecting HGT in eukaryotes, enough individual cases have been "accidentally" discovered that researchers have given up trying to list them all (Keeling and Palmer 2008).

The implications of HGT for species tree research are substantial. For example, following the standard assumption in coalescent theory that allelic divergences must occur earlier in time than the divergences of species harboring those alleles, many species tree techniques (Rannala and Yang 2003; Liu and Pearl 2007) assume that the gene tree exhibiting the most recent divergence

between taxon A and taxon B establishes a hard upper limit on the divergence time of those species in the species tree. For small sets of genes in taxa where HGT is rare, a researcher would need to be quite unlucky to choose a horizontally transferred gene for analysis. However, as the genomic era advances, it becomes likely that at least one of the thousands of genes studied will have been transferred horizontally and thus establish a spurious upper-bound for clade divergence at the species level. For example, if even one gene has ever been transferred between humans and fruit-flies in the last 910 million years (Blair 2009), or uploaded from the same pathogen, then the date of this transfer event will be taken as the maximum plausible divergence time for those species despite thousands of other genes in the continuant genomes implying a much deeper split. Although HGT is clearly a problem for some current methodologies, if transferred genes can first be identified, then they could be extremely useful as genomic markers for monophyletic groups that have inherited such genes and would otherwise be difficult to resolve (Huang and Gogarten 2006). Unfortunately, current methods to detect such events rely both on having the true species tree already in hand and also on the absence of other mechanisms causing gene tree discordance (Linz et al 2007; Rasmussen and Kellis 2007, 2010). For many types of comparisons, such as those among major groups of animals or vertebrates, the data show enough congruence to make identification of HGT events straightforward, and HGT appears to be infrequent among closely related species of eukaryotes, although data is sparse. HGT will pose particular challenges for phylogenetically understudied groups for which the expected shape of gene trees is not known.

Gene duplication

Gene duplication presents another violation of the coalescent model; like HGT, its potential problems are worst when they go unrecognized. Imagine a taxon where a gene of interest duplicated 10Mya into copy α and copy β; the taxon then split 5Mya into species 1 and 2. A researcher investigating the daughter species would therefore sequence four orthologous genes, with the potential to compare α1 to β2 and β1 to α2 and thus generate two gene trees where the estimated split time was 10Mya, rather than 5Mya. Such a situation will be easily recognized if copy α and β have diverged sufficiently by the time of their duplication, and a number of methods of phylogenetic analysis have incorporated gene duplication (e.g., Sanderson and McMahon 2007; Thomas 2010). Additionally, failure to recognize the situation may not have drastic consequences for phylogenetic analysis if the paralogs had not diverged much, in which case the estimated gene coalescence would be approximately correct no matter which comparison was made. However, if one of the copies has been lost and only one of the remaining copies is sequenced, then the chances of inferring an inappropriately long period of genetic isolation are larger, and will increase as the size of the family of paralogs increases. This problem will tend to overestimate gene coalescence times, and some species tree methods depend on minimum isolation times among a large set of genes. Still, these deep coalescences might spuriously increase inferred ancestral population sizes.

Natural selection

Natural selection causes yet another violation of the multispecies coalescent model. Selection can cause serious problems in some cases, although in other circumstances it is predicted not to cause problems of phylogenetic analysis (Edwards 2009). The usual stabilizing

selection can be helpful to taxonomists working at high levels because it slows the substitution rate; likewise selective sweeps, directional selection, and genetic surfing (Ray and Excoffier 2009) tend to clarify phylogenetic relationships by accelerating reciprocal monophyly for genes in rapidly diverging clades. However, challenges to phylogenetic inference are posed by convergent neutral mutations (homoplasy), balancing selection, and selection-driven convergent evolution. Given a finite number of sites at a neutral locus, occasional homoplasies will occur, and will be exacerbated by increased variation in mutation rate among sites. In the absence of other mechanisms, however, the addition of more informative and less noisy loci will often compensate for homoplasies at other loci. Because balancing selection tends to preserve beneficial alleles at a gene, two divergent taxa will appear interdigitated at that locus, and reticulated through time if ancient DNA is available. Again, including loci that are not under strong balancing selection, or removing loci influenced by balancing selection from the data set, should resolve this problem. Finally, convergent molecular evolution can occur across some genomic regions, at least in the mitochondrial genome, due to parallel selection on distantly related taxa (e.g., Castoe *et al.* 2009). This "insidious" form of evolution (Edwards 2009) is particularly difficult to resolve mathematically, entrapping tree-building algorithms on false topologies either because of strong support for local optima, or producing an excess of evidence favoring incorrect phylogenies. It can also be difficult to detect, since the synonymous/non-synonymous mutation ratio might suggest other types of selection, such as stabilizing selection, that do not pose problems for phylogenetic analysis.

Detecting phylogenetic outliers in the multispecies coalescent model

Detecting outliers in population genetics

Many of the instances of violations of the coalescent model will occur at individual genes, and usually will not dominate the signal of the entire suite of genes sampled for phylogenetic analysis. Thus we can think of such genes as phylogenetic outliers – genes whose phylogenetic signal differs significantly from that of the remainder of data set. This in turn raises the possibility of developing statistical tests to identify such outliers, prior to, during, or after phylogenetic analysis, so that they can ultimately be removed or down-weighted. There is a robust history of detecting outliers in phylogenetics, for example, detecting cases of incongruence (Swofford 1991) or genes subject to HGT (Linz et al. 2007; Roettger et al. 2009). However, there has been little work to our knowledge in detecting outliers while simultaneously accounting for the variation among genes introduced by the multispecies coalescent. In addition, with or without the context of the multispecies coalescent, there has been little work on detecting phylogenetic outliers due to forces other than HGT – for example, due to natural selection.

Detection of outliers has recently come to the fore in the field of population genomics, and recent years have seen a flurry of studies analyzing hundreds–if not thousands–of genetically independent loci, especially in surveys of model species such as humans and *Drosophila*. For example, there exist Bayesian methods to detect loci that differ significantly from the dominant signal as measured by Fst or some other metric of population divergence (Beaumont and Balding 2004). In the case of Fst, some means of correcting for the average heterozygosity among markers is necessary because the extent of differentiation of loci with higher average heterozygosity is expected to have a higher variance than markers with low variance. The variance in differentiation among loci is useful to set up a null hypothesis for the test statistic and

genes falling outside this expected variance are deemed outliers. In general, the construction of a valid null hypothesis for the average locus in a given multilocus data set–incorporating as many sources of variance as possible, including coalescent variance–can be useful in erecting statistical tests of outliers. We first mention some ways in which phylogenetic outliers can be identified using traditional methods in molecular evolution. We then outline several approaches that we suggest might be useful in identifying outliers in the multispecies coalescent model, and provide an example of a test that may prove useful to the community.

Genomic signals of phylogenetic outliers

Synonymous / Non-synonymous mutation ratio: One method of detecting potentially problematic forms of selection is to look for loci with unusual dn/ds ratios. According to neutral theory, most loci should be under stabilizing selection, and hence have many more mutations in the third codon position than in positions one and two. Regions under balancing selection should have higher non-synonymous mutation rates. However, using the dn/ds ratio as a means of detecting phylogenetic outliers presents some difficulties. Of course such a test would only be applicable to coding regions. Additionally, such genes may exhibit anomalous behavior at the amino acid level, they may not be anomalous in their phylogenetic signal, which is our primary concern. Finally, many coding loci may undergo substitutions more freely than expected due to canalization (*sensu* Waddington 1946) or genomic redundancy. Many genes exhibit a slight excess of non-synonymous ubstitutions within populations, because even strong directional selection rarely purges all such alleles from populations (Burke *et al.* 2010).

GC ratio and DNA word frequencies: Regions of the genome that have been acquired from another domain of life (such as a eukaryote with DNA from viruses, bacteria, or archea) will often have an unusual GC composition relative to the rest of the genome. Indeed, focusing on genomic regions with anomalous GC content is a common method for identifying genes that have undergone HGT. More complex consequences of base composition and mutation patterns, such as the frequencies of DNA oligonucleotides ('words') in coding or noncoding regions, have also been used to flag potential HGT genes, particularly in bacteria (Medrano-Sato et al. 2004; Dufraigne et al. 2005). Like the test above, the results of GC or DNA word frequency analysis should be considered suggestive, but not conclusive. There are other reasons for unusual GC content (e.g., leucine zipper motifs, a GC microsatellite, etc.), which are likely to occur by chance in a large genome. Again, the phylogenetic consequences of such deviations in evolutionary pattern are paramount. In this regard, high variation in GC content among genes can cause strong deviations in resulting phylogenies, although distinguishing the true gene tree from the tree suggested by the variation can be challenging (e.g., using LogDet distances; Lockhart et al. 1994).

Statistical tests to detect phylogenetic outliers in the multispecies coalescent model

When faced with a surprising or nonconvergent species tree, one possibility is that an unusual gene tree is to blame. Though techniques for dealing with violations of the coalescent model are in their infancy, researchers do have a few options. Below we list several ideas, some borrowed from classical phylogenetics or from methods used in bioinformatics. It is likely that the several tests constructed to detect phylogenetic outliers in classical phylogenetics can be extended

slightly to incorporate the additional variation among genes expected due to the coalescent process. Of course, with larger data sets, single anomalous genes may have little effect on the resulting species tree, particularly in species tree methods utilizing summary statistics (Liu et al. 2009, STAR/STEAC). However, as pointed out above, species tree methods such as BEST that rely on 'hard' boundaries for the species tree by individual genes could be de-railed due to the anomalous behavior of even a single gene.

Jackknifing: A straightforward approach to detecting phylogenetic outliers under the multispecies coalescent model is to rerun the analysis *n* times, where *n* is the number of loci in the study, leaving one locus out each time. An outlier can then be identified if the analysis that does not include that gene differs from the remaining analyses in which that gene is included. This approach has been applied successfully in fruit flies by Wong *et al.* (2007), who considered their problem resolved when the elimination of one of ten genes unambiguously resolved a polytomy. There may be other metrics of success that are more robust or sensitive, or do not depend as strongly on *a priori* beliefs about the relationships among taxa. Because some duplications or horizontal transfers may affect only one taxon, whole-tree topology summary statistics are unlikely to be sensitive enough to detect recent events. However, the cophenetic distance of each taxon to its nearest neighbor in the complete species tree could be compared across jackknife results. This procedure will produce a distribution of "typical" distances, and significance can therefore be assigned to highly divergent results. The drawback to such an approach is the computational demand. Species tree analyses on their own can be extremely time-consuming to run even once, so jackknifing may prove intractable for studies involving many species and loci.

Species tree methods accommodating anomalous loci

One attractive prospect is to develop algorithms for species tree construction that are less susceptible to the effects of single genes. STAR and STEAC are two approaches that use summary statistics (average ranks or coalescence times across genes) to reconstruct species trees. These methods are powerful and fast, yet they do not utilize all the information in the data, and hence can be less efficient than Bayesian or likelihood methods (Liu et al. 2009). A recently introduced likelihood method based on gene tree triples also seems relatively immune to events like HGT that compromise the signal in single genes (Liu et al. 2010). Nonetheless it would be desirable to have a fully Bayesian or likelihood method that can resist bias introduced by individual genes. For example, rather than basing clade divergence times on the minimum gene tree split times, as done in BEST, species divergence times could be chosen from the joint posterior distribution of divergence times produced across gene trees. This means that non-coalescent events would be incorporated into a coalescent analysis only as often as they actually occur in the data, given a sufficiently long MCMC run, and their effect on the final result would be diluted. However, an alternative method of evaluating the likelihood of the species tree would be necessary, as the Felsenstein likelihood (Felsenstein 1981; see also Heled and Drummond 2010) for genes converging post-isolation is essentially zero. It is possible to run MCMC chains without this likelihood using a summary statistic or epsilon-kernel approach (Marjoram *et al.* 2003), but software implementing the praxis is not yet available.

Outlier Analysis: One other option for multilocus studies is to construct either histograms of genetic distance, or regressions of molecular divergences between taxa in which each point

represents one locus, thereby allowing visual or statistical identification of outlier loci. From a pragmatic and computational point of view, this is an attractive option because distances between taxa already need to be calculated in most species tree software, thus a second step analyzing these distances would be computationally cheap. Such a method also has the benefit of being able to detect both duplication events and horizontal gene transfers. Below we provide a simulated example.

Example: We simulate a 10-species phylogeny (Figure 4) with normally distributed divergence times (since species trees generally do not exhibit the exponentially increasing divergence times of a coalescent model). We then "sprinkle" Jukes-Cantor mutations on this tree with mutation rates spanning two orders of magnitude (more than is commonly observed in nature to provide a rigorous model test), to generate 20 loci of 100 nucleotides each (a fairly modest total of 2000 base pairs). The key component of this test is the use of multiple loci to establish a pattern that can possibly be violated by HGT. Finally, a 21st gene is simulated on a species tree in which one taxon has acquired the gene laterally from another at some point in the past. We then need an appropriate statistic with which to quantify the phylogenetic patterns and divergences among gene trees. Though many statistics are available, here we simply count the number of variable sites displayed by a given pair of species for clarity. Regressing the number of variable sites across all 10 taxa versus the number of variable sites between pairs of taxa clearly demonstrates both the presence and the direction of HGT (Figure 4). The recipient taxon can be easily distinguished, because it will be anomalous in all pairwise comparisons. The donor taxon can be identified as the closest relative of the recipient in that gene tree, who is also a distant relative in all other gene trees. Since the HGT event should be detectable by pairing the

recipient taxon with any other taxon in the tree, one test that should provide substantial power is to count the number of times a locus lies outside the 95% confidence band for each pairwise comparison.

Future directions

Species tree methods are likely to continue to gain ascendancy as the strongest evidence of taxonomic relationship in phylogenetic research. As with any form of evidence, the conclusions of a species tree analysis are fallible, with each method susceptible to certain biases in exceptional cases. In the future, we hope that these biases and susceptibilities can be overcome, and that species tree methods will continue to multiply. Because the most robust techniques rely heavily on a coalescent paradigm, the field needs a method for detecting loci that violate the assumptions of coalescent theory. A few ideas for how to do this have been presented and outlined above, but need more rigorous theoretical, and certainly more empirical testing to establish their effectiveness in phylogenetic influence.

Detection is just the first step. Currently, when such loci are discovered, researchers have two options: they can use methods that are sufficiently robust (hopefully) to overcome the faulty assumptions of coalescence, or remove the loci from the analysis set. These solutions, though adequate, are not best-case scenarios. As discussed above, it would be preferable to develop methods that use the information contained in non-coalescent events to further support phylogenetic inference. Such a program, widely applied, would have the potential to not only solidify our understanding of the genetic relationships of all organisms, but also provide invaluable insight into the prevalence and significance non-standard evolutionary modes.

Appendix A: Simulating Gene Trees in Species Trees

At times, it may be desirable to simulate evolution within a species group. This can be done either to test hypotheses about isolation due to drift, or in the context of Bayesian analysis to infer other parameters regarding the demographic processes occurring at scales finer than the species group. A simple example of how this could be accomplished in Bayesian Serial SimCoal (Excoffier et al. 2000, Anderson et al. 2005) is described below. The suite of R-scripts in Phybase (Liu 2009

Imagine a species tree with 10 individuals, four species (with 4, 2, 3, and 1 representatives respectively) and with known (or previously inferred) split times among taxa. In addition, we will assume for this example that the effective population size N_e of each contemporary species is 1000, and that the size of ancestral populations are the sum of the sizes of their respective descendent population. This situation is analogous to that depicted in figure 5. The corresponding NEXUS formatted species tree is:

(D:1500,(C:800,(B:500,A:500):300):700);

Here, branch lengths are in units of generations, which is commensurate with using units of individuals for the population sizes (Other simulation methods use units of $\tau = \mu t$ and $\theta = 4N\mu$, in units of substitutions per site, instead of t and N_e, respectively).

A simple forward simulation can be run in any version of SimCoal using the following .par file:

```
Species tree input file; 10 taxa, 4 sp
4 demes
Deme sizes (arbitrary in this case)
1000
1000
1000
1000
Sample sizes: # samples, age, deme, stat group
4
2
3
1
Growth rates
0
0
0
0
Number of migration matrices
0
Historical event: date from to %mig new_N new_r migmat
3 events
500  1 0 1 2.00 0 0
800  2 0 1 1.50 0 0
1500 3 0 1 1.33 0 0
Mutations per generation for the whole sequence
.0001
Number of loci
10
Data type : either DNA, RFLP, or MICROSAT
DNA
//Mutation rates: gamma parameters, theta and k
0 0
```

In this case, the tree was perfectly ordered, so all populations could simply fuse with deme 0, readjusting the population size each time. Of course, there is no need to assume that all

populations have the same effective size, nor that N_e of ancestral populations was the sum of their N_e values of their descendants. If we wished to infer the size of clade AB at the time of the split, for example, we could replace the 2.00 in the first historical event with, for example {U:0.5,3.0}, which would allow the program to infer the most likely time-averaged N_e of clade AB in the range from 500 to 3,000 individuals. Similarly, if the mutation rate of the gene in question was unknown, or if a range of mutation rates would produce the simulated desiderata, then the mutation rate constant, set in the example above at .0001, could be replaced with {E:.0001}, creating an exponential distribution of mutation rates whose mean was 0.0001. Full documentation on the parameter files, and Bayesian inference using priors instead of constants, can be found at the BayeSSC website: http://www.stanford.edu/group/hadlylab/ssc/

Bibliography

Anderson CNK, Ramakrishnan U, Chan YL and Hadly EA (2005) Serial SimCoal: A population genetic model for data from multiple populations and points in time. Bioinformatics, 21, 1733-1734.

Avise, J. C. (1994) Molecular markers, natural history and evolution. New York: Chapman and Hall.

Avise, J. C. & Ball, R. M. (1990) Principles of genealogical concordance in species concepts and biological taxonomy. Oxford Surveys in Evolutionary Biology 7:45-67.

Beaumont, M. A. & Balding, D. J. (2004) Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology, 13:969-980.

Blair, J. E. (2009) Animals: metazoa. In S. B. Hedges & S. Kumar Eds, The Timetree of Life, Oxford University Press, 223-230.

Boto, L. (2010) Horizontal gene transfer in evolution: facts and challenges. Proceedings of the Royal Society B 277:819-827.

Bryant, D. (2003) A classification of consensus methods for phylogenetics. In: M. Janowitz *et al.*, Eds, BioConsensus, American Mathematical Society, 163–183.

Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L. & Waddell, P. J. (1993) Partitioning and Combining Data in Phylogenetic Analysis. Systematic Biology 43:384-397.

Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R. & Long, A. D. (2010) Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467:587-590.

Carstens, B. C. & Dewey, T. A. (2010) Species delimitation using a combined coalescent and information-theoretic approach: An example from North American *Myotis* bats. Systematic Biology, 59, 400-414.

Carstens, B. C. & Knowles, L. L. (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from melanoplus grasshoppers. Systematic Biology 56:400-411.

Castillo-Ramírez, S., Liu, L. Pearl, D. & Edwards, S. V. (2010) Bayesian estimation of species trees: a practical guide to optimal sampling and analysis. In: Estimating species trees: Practical and theoretical aspects. Knowles, L. L. & Kubatko L. S. eds. Hoboken, NJ: John Wiley & Sons.

Castoe, T. A., Koning, A. P. J. d., Kim, H., Gu, W., Noonan, B. P., Naylor, G., Jiang, Z. J., Parkinson, C. L. & Pollock, D. D. (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. Proceedings of the National Academy of Sciences, USA 106:8986-8991.

Cavalli-Sforza, L. L. (1964) Population structure and human evolution. Proceedings of the Royal Society of London, B 164:362-379.

de Queiroz, K. (2007) Species Concepts and Species Delimitation. Systematic Biology 56(6):879-886.

Degnan, J. H. & Rosenberg, N. A. (2006) Discordance of Species Trees with Their Most Likely Gene Trees. PLOS Genetics 2(5):e68.

Degnan, J. H. & Rosenberg, N. A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution 24:332-340.

Degnan, J. H. & Salter, L. A. (2005) Gene tree distributions under the coalescent process. Evolution 59:24-37.

Driskell, A. C., Ané, C., Burleigh, J. G., McMahon, M. M., O'Meara, B. C. & Sanderson, M. J. (2004) Prospects for building the tree of life from large sequence databases. Sciences, 306, 1172-1174.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. & Deschavanne, P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acid Research 33:e6.

Eckert, A. J. & Carstens, B. C. (2008) Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Molecular Phylogenetics and Evolution 49:832-842.

Edwards, S. V. (2009) Is a new and general theory of molecular systematics emerging? Evolution 63:1-19.

Edwards, S. V. (2009) Natural selection and phylogenetic analysis. Proceedings of the National Academy of Sciences, USA 106:8799-8800.

Edwards, S. V., Liu, L. & Pearl, D. K. (2007) High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences, USA 104:5936-5941.

Ewing, G. B., Ebersberger, I., Schmidt, H. A. & von Haeseler, A. (2008) Rooted triple consensus and anomalous gene trees. BMC Evolutionary Biology 8:118.

Excoffier L, Novembre J and Schneider S (2000) SIMCOAL: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. J. Hered., 91, 506-509.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Evolution 17(6):368-376.

Gadagkar, S. R., Rosenberg, M. S. & Kumar, S. (2005) Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. Journal of Experimental Zoology B 304:64-74.

Gould, S. J. (2001) The Book of Life: An illustrated history of the evolution of life on earth. New York: W. W. Norton & Co.

He, Y., Wu, J., Dressman, D. C., Iacobuzio-Donahue, C., Markowitz, S. D., Velculescu, V. E., Diaz Jr, L. A., Kinzler, K. W., Vogelstein, B. & Papadopoulos, N. (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 464: 610-614.

Heled, J. & Drummond, A. J. (2010) Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27:570-580.

Hillis, D. M. (1987) Molecular versus morphological approaches to systematics. Annual Review of Ecology and Systematics 18:23-42.

Miyamoto, M. M., and Cracraft, J. (1991). Phylogeny inference, DNA sequence analysis, and the future of molecular systematics. In ‘‘Phylogenetic Analysis of DNA Sequences’’ (M. M. Miyamoto and J. Cracraft, Eds.), pp. 3–17, Oxford Univ. Press, New York.

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Torres, M. C. M., Giebel, J. D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R. V., Shepard, J., Tomkins, J., Richards, S., Spiro, D. J., Ghedin, E., Slatko, B. E., Tettelin, H. & Werren, J. H. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Sciences 317:1753-1756.

Huang, H. & Knowles, L. L. (2009) What is the danger of the anomaly zone for empirical phylogenetics? Systematic Biology 58:527-536.

Huang, H., He, Q., Kubatko, L. S. & Knowles, L. L. (2010) Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Systematic Biology 59:573-583.

Huang, J. & Gogarten, J. P. (2006) Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends in Genetics 22:361-366.

Hudson, R. R. (1983) Testing the constant-rate neutral allele model with protein sequence data. International Journal of Organic Evolution 37: 203–217.

Hudson, R. R. & Coyne, J. A. (2002) Mathematical consequences of the genealogical species concept. Evolution 56:1557-1565.

Huelsenbeck, J. P. & Andolfatto, P. (2007) Inference of population structure under a dirichlet process model. Genetics 175:187-1802.

Keeling, P. J. & Palmer, J. D. (2008) Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9:605-618.

Knowles, L. L. & Carstens, B. C. (2007) Delimiting Species without Monophyletic Gene Trees. Systematic Biology 56(6):887-895.

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X. & Wilson, A. C. (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proceedings of the National Academy of Sciences, USA 86:6196-6200.

Kubatko L., Carstens, B. & Knowles, L.L. (2009) STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971-973.

Kubatko, L. S. & Degnan, J. H. (2007) Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Systematic Biology 56:17-24.

Leaché, A. D. & Fujita, M. K. (2010) Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). Proceedings of the National Academy of Sciences, USA 277:3071-3077.

Linz, S., Radtke, A., von Haesler, A. & Haeseler, A. v. (2007) A likelihood framework to measure horizontal gene transfer. Molecular Biology and Evolution 24:1312-1319.

Liu, L. (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24(21):2542-2543.

Liu, L. & Edwards, S. V. (2009) Phylogenetic analysis in the anomaly zone. Systematic Biology 58(4):452-460.

Liu, L. & Pearl, D. K. (2007) Species trees from gene trees: Reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Systematic Biology 56(3):504-514.

Liu, L., Pearl, D. K., Brumfield, R. T. & Edwards, S. V. (2008) Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080-2091.

Liu, L., Yu, L. (2010) Phybase: an R package for species tree analysis Bioinformatics (2010) 26: 962-963.

Liu, L., Yu, L., Kubatko, L., Pearl, D. K. & Edwards, S. V. (2009) Coalescent methods for estimating phylogenetic trees. Molecular Phylogenetics and Evolution 53:320-328.

Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Molecular Biology and Evolution 11:605-612.

Maddison, W. P. (1997) Gene trees in species trees. Systematic Biology, 46(14), 523-536.

Maddison, W. P. & Knowles, L. L. (2006) Inferring phylogeny despite incomplete lineage sorting. Systematic Biology 55:21-30.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences, USA 100:15324-15328.

Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A. & Collado-Vides, J. (2004) Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Molecular Biology and Evolution 21:1884-1894.

Nei, M. (1987). Molecular Evolutionary Genetics. New York: Columbia University Press.

Nei, M. & Kumar, S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Neigel, J. E. & Avise, J. C. (1986) Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. in Karlin, S. & Nevo, E. eds. Evolutionary processes and theory. Academic Press: New York, NY, 515-534

O'Meara, B. C. (2008) Using trees: *myrmecocystus* phylogeny and character evolution and new methods for investigating trait evolution and species delimitation (Ph.D. Dissertation). Available from Nature Proceedings http://dx.doi.org/10.1038/npre.2008.2261.1.

O'Meara, B. C. (2010) New heuristic methods for joint species delimitation and species tree inference. Systematic Biology 59: 59-73.

Oliver, J. C. (2008) AUGIST: inferring species trees while accommodating gene tree uncertainty. Bioinformatics 24:2932-2933.

Pamilo, P. & Nei, M. (1988) Relationships between gene trees and species trees. Molecular Biological Evolution 5:568-583.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Rannala, B. & Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. Genetics 164:1645-1656.

Ranwez, V., Criscuolo, A. & Douzery, E. J. (2010) SUPERTRIPLETS: a triplet-based supertree approach to phylogenomics. Bioinformatics 26:i115-i123.

Rasmussen, M. D. & Kellis, M. (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Research 17:1932-1942.

Rasmussen, M. D. & Kellis, M. (2010) A bayesian approach for fast and accurate gene tree reconstruction. Molecular Biology and Evolution, In press. Available online, DOI: 10.1093/molbev/msq189.

Ray, N. & Excoffier, L. (2009) inferring past demography using spatially explicit population genetic models. Human Biology 81:141-157.

Roettger, M., Martin, W. & Dagan, T. A. (2009) Machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. Molecular Biology and Evolution 26:1931-1939.

Rokas, A. (2006) Genomics and the tree of life. Sciences, 313, 1897-1899.

Rosenberg, N. A. & Tao, R. (2008) Discordance of species trees with their most likely gene trees: the case of five taxa. Sytematic Biology 57:131-140.

Sanderson, M. J., & McMahon, M. M. (2007) Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evolutionary Biology 7:S3.

Satta, Y., Klein, J. & Takahata, N. (2000) DNA archives and our nearest relative: The trichotomy problem revisited. Molecular Phylogenetics and Evolution, 14, 259-275.

Steel, M. & Rodrigo, A. (2008) Maximum likelihood supertrees. Systematic Biology 57(2):243-250.

Swofford, D. L. (1991) When are phylogeny estimates from molecular and morphological data incongruent? In Miyamoto, M. M. & Cracraft, J., eds. Phylogenetic analysis of DNA sequences. Oxford Univ. Press, New York: 295-333.

Swofford, D. L., Olsen, G. J., Waddell, P.J. & D.M. Hillis. (1996) Phylogenetic inference. Pp. 407-514 in Hillis, D. M., Moritz C. & Mable, B. K., eds. Molecular Systematics, Second Edition. Sinauer Associates: Sunderland, MA.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437-460.

Takahata, N. (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957-966.

Thomas, J., Schaack, S. & Pritham, E. J. (2010) Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biology and Evolution 2:656-664.

Thomas, P. D. (2010) GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. BMC Bioinformatics 11:312.

Tobias, J. A., Seddon, N., Spottiswoode, C. N., Pilgrim, J. D., Fishpool, L. D. C. & Collar, N. J. (2010) Quantitative criteria for species delimitation. Ibis 152(4): 724-746.

Wakeley, J. (2000) The effects of subdivision on the genetic divergence of populations and species, Evolution 54, 1092–1101.

Wiens, J. J. (2003) Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology 52:528-538.

Wilson, I. J. & Balding, D. J. (1998) Genealogical inference from microsatellite data. Genetics 150:499-510.

Wilson, I. J., Weale, M. E. & Balding, D. J. (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society: Series A, 166: 155-188

Wollenberg, K. & Avise, J. C. (1998) Sampling properties of genealogical pathways underlying population pedigrees. Evolution 52:957-966.

Wong, A., Jensen, J. D., Pool, J. E. & Aquadro, C. F. (2007) Phylogenetic incongruence in the Drosophila melanogaster species group. Molecular Phylogenetics and Evolution 43(3):1138-1150.

Wu, M., & Eisen, J. (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biology 9:R151

Yang, Z. & Rannala, B. (2010) Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences, USA 107:9264-9269.

Figure Captions

FIGURE 1 (color, for centerpiece of book): An example showing the utility of multiple gene trees in producing species tree topologies. (A) Nine unlinked loci are simulated (or inferred without error) from a species group with substantial amounts of incomplete lineage sorting. Note that no single gene recovers the correct relationship between clades. Furthermore, despite identical conditions for all nine simulations, no two genes agree on the correct topology, let alone the correct divergence times. (B) Superimposing the nine gene trees on top of each other clarifies the relationships. It can be (correctly) inferred that the true tree is perfectly ordered, with (ABC) diverging from D about 1500 generations ago, the (AB)-C split occurring at 800, and A diverging from B about 600 generations ago. Also, the amount of crossbreeding within the recently diverged taxa implies (correctly) that C has the effective smallest population size.

FIGURE 2: The relationship between gene trees and species trees. Lines within the species trees indicate gene lineages. Simplified gene trees are shown below each species tree. Whereas gene trees on the left vary due to deep coalescence, gene trees on the right are toplogically concordant but vary slightly in branch lengths due to the coalescent. Modified with permission from Edwards (2009).

FIGURE 3: Three examples of non-coalescent gene histories. (A) A duplication event that precedes a speciation event can lead to incorrect inference of divergence times in the species tree if copy 1 is compared to copy 2. This can be particularly difficult if one of the gene copies has been lost, or not sequenced by the researcher. (B) Convergent evolution can occur at the

molecular level, for example in certain genes under environmental selection if both taxa move into the same environment. It will tend to bring distantly related taxa into a jumbled polyphyletic clade, and is likely to be given additional false support by morphological data. (C) Horizontal gene transfer causes difficulties in current species tree methods, because it establishes a spurious lower-bound to divergence times. Though rare in eukaryotes, it is by no means unknown, and is likely to become a more difficult problem in the future when species trees are based on tens of thousands of loci.

FIGURE 4: HGT can be detected by comparing the diversity of genes in all taxa to the diversity of genes in pairs of taxa. Transfer events should appear as anomalies in regressions or histograms in each pair of species, in this case locus 21. In the example pair above, one of the 20 "normal" loci also lies outside the 95% confidence band as expected, but this locus would not be expected to lie outside the confidence band in all pairs. This particular locus highlights another hazard of such an analysis: the locus has saturated (100 segregating sites in a 100bp locus) and thus shows a positive deviation from expectation in closely related taxa.

FIGURE 5: The species tree simulated in the Appendix. Branch lengths are in units of generations and branch widths (population sizes) are in units of individuals. This particular tree has the constraint that ancestral population sizes are the sum of the population sizes of descendent lineages, but of course one can simulate without these constraints using either SerialSimcoal or Phybase.

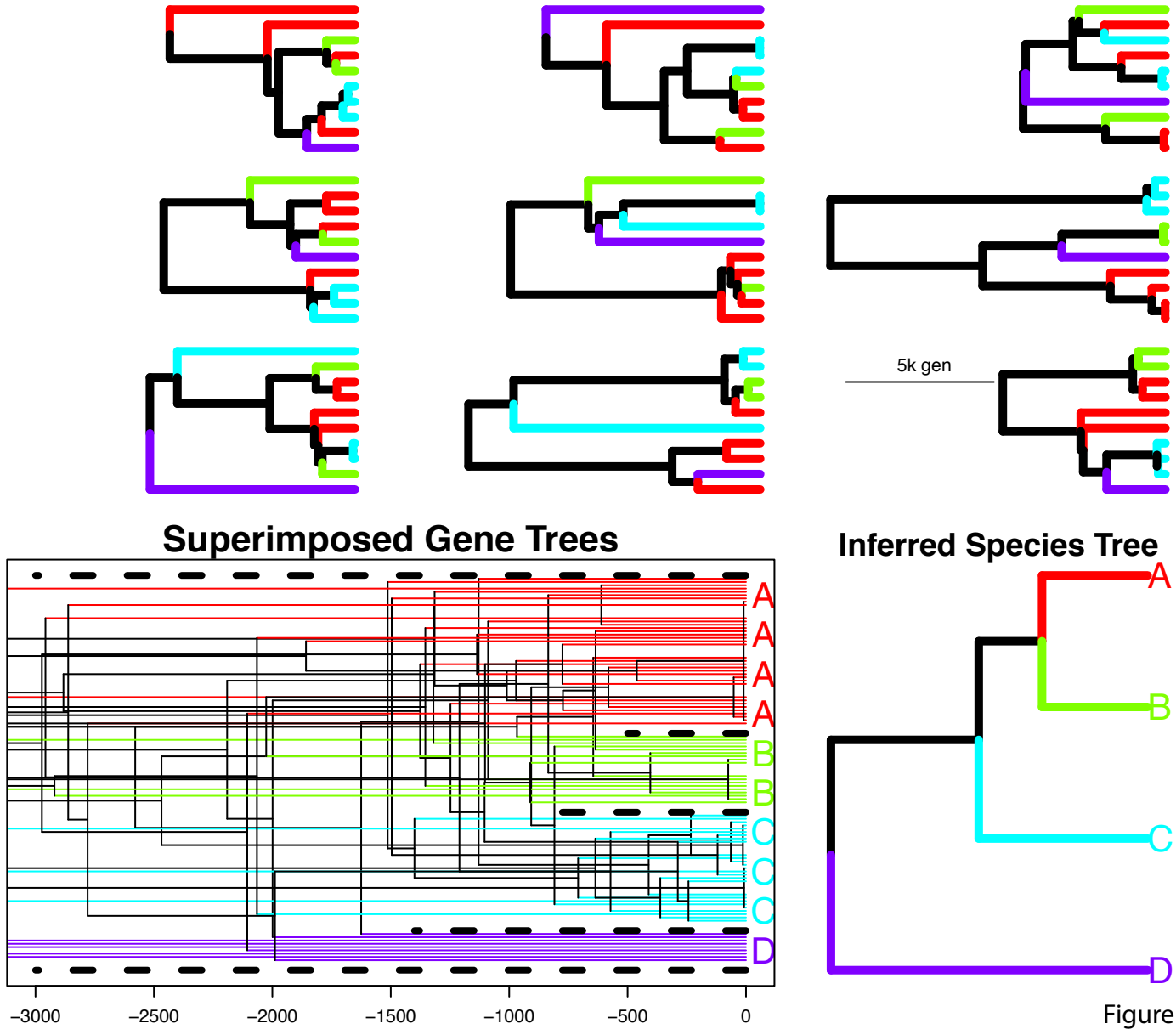

Set of 9 Gene Trees
5k gen
Superimposed Gene Trees
A
A
A
A
B
B
C
C
C
D
−3000
−2500
−2000
−1500
−1000
−500
0
Inferred Species Tree
A
B
C
D

Figure 1

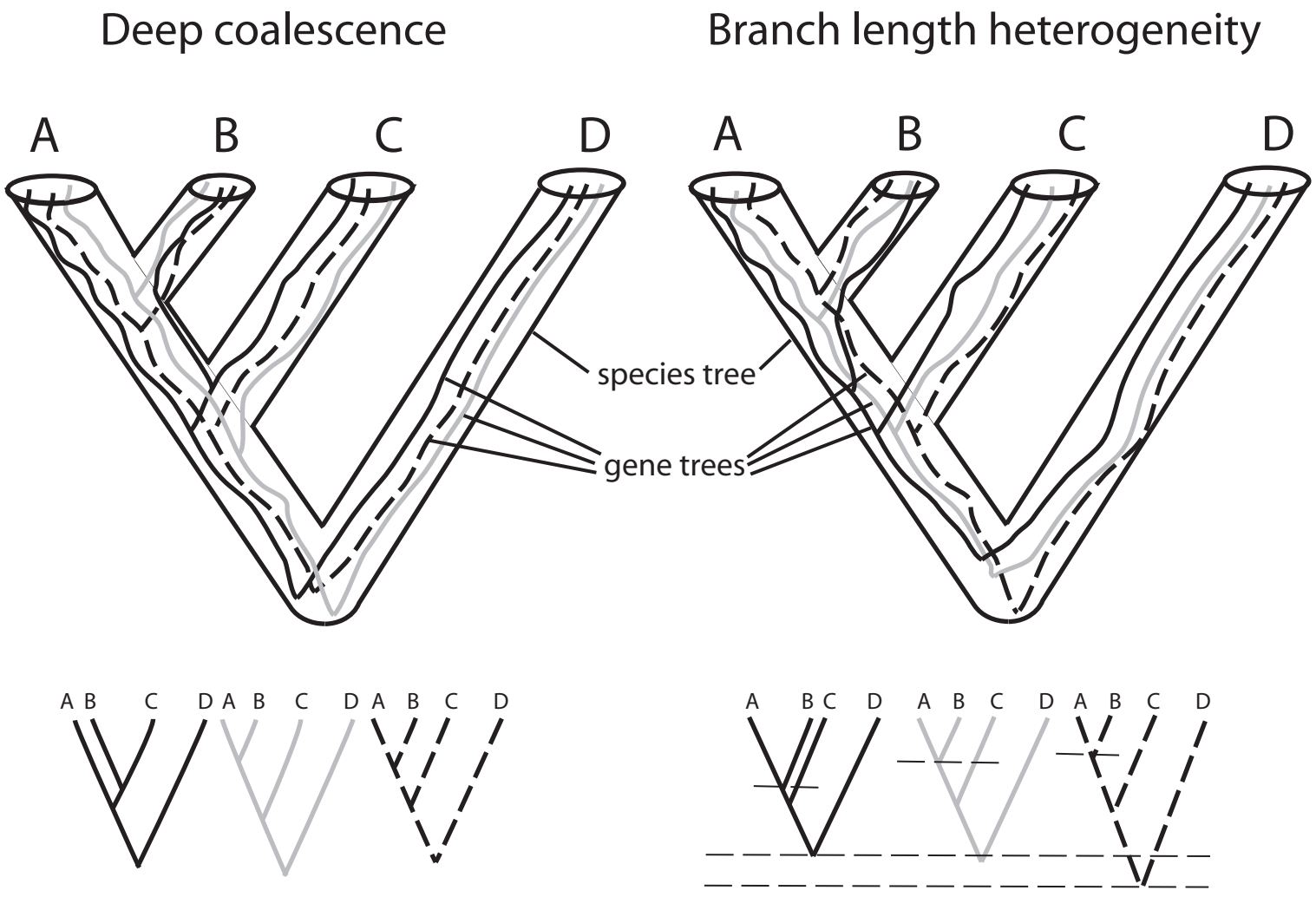
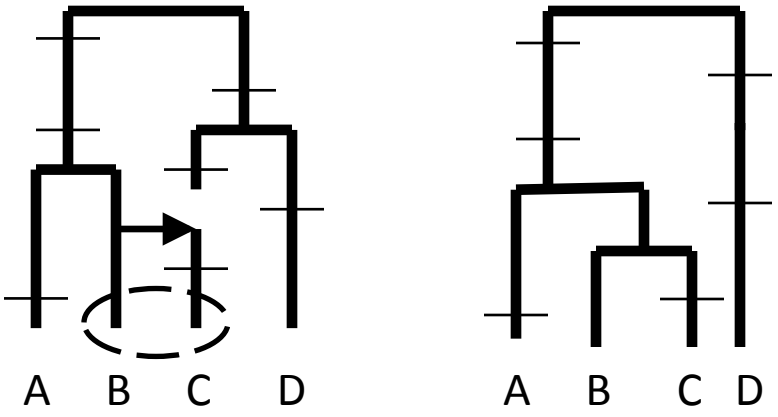

Deep coalescence
Branch length heterogeneity
A
B
C
D
species tree
gene trees
A B C D A B C D A B C D
A B C D A B C D A B C D

Figure 2

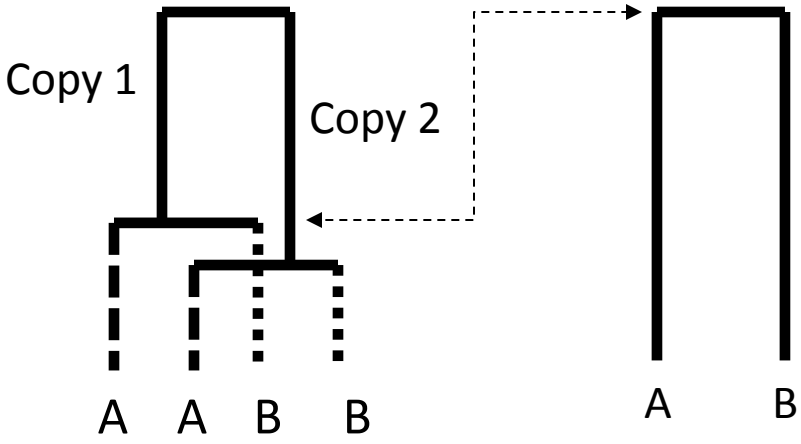
TRUE HISTORY

INFERRED HISTORY

Horizontal Gene Transfer



Gene Duplication



Convergent Evolution

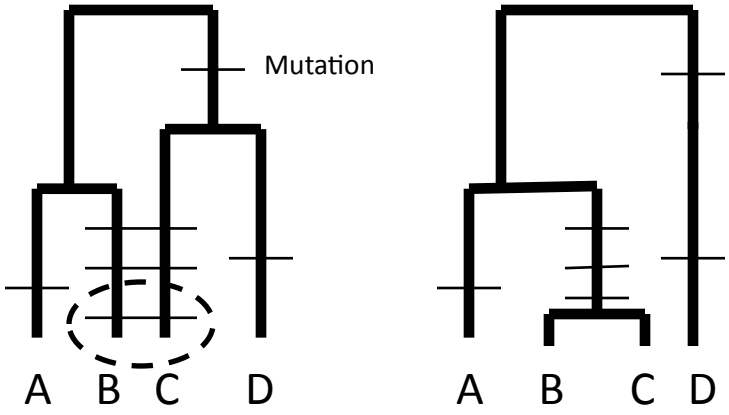


Fig. 3

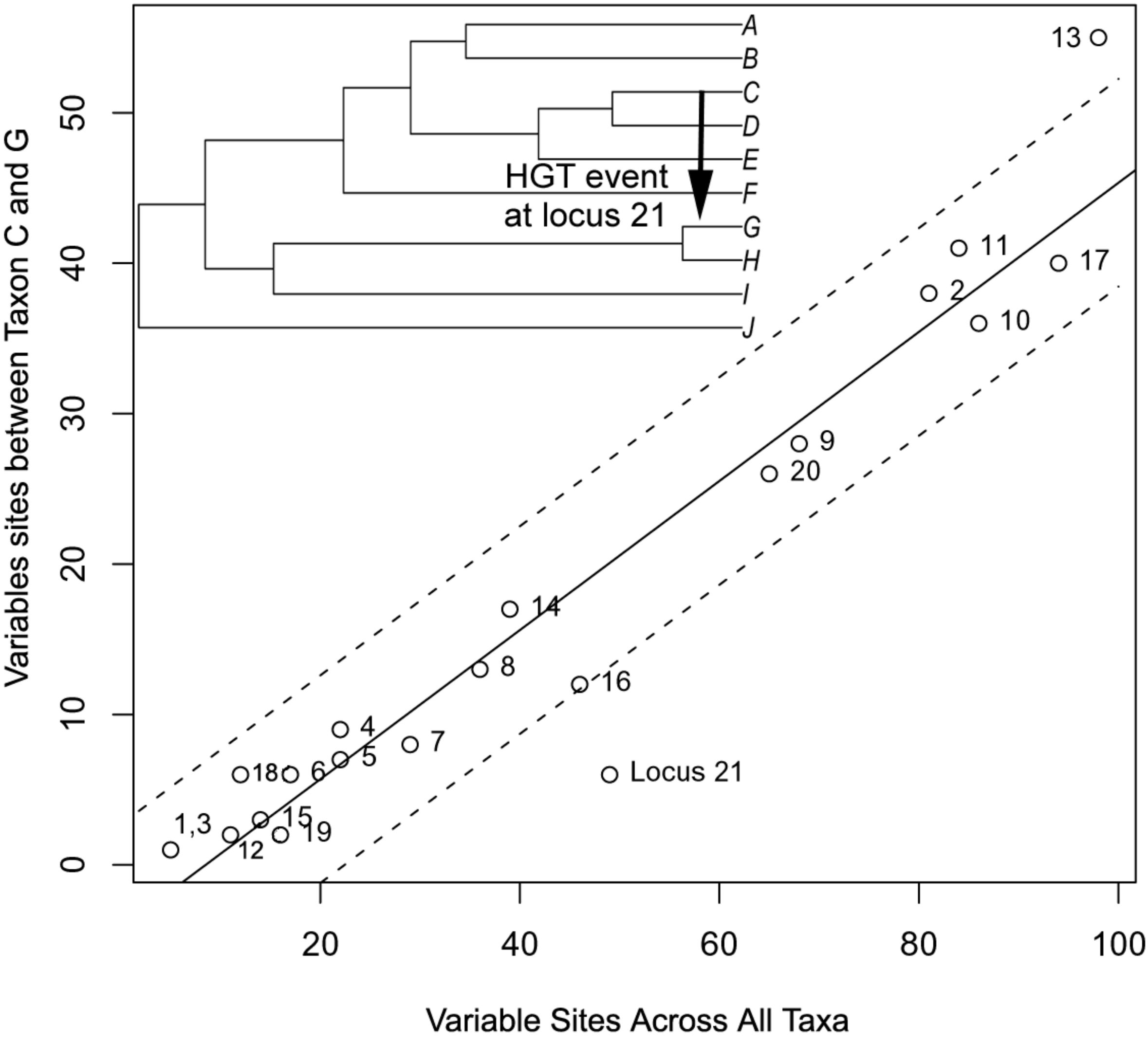

A
B
C
D
E
F
G
H
I
J
HGT event
at locus 21
13
11
17
2
10
9
20
14
8
16
4
7
5
18
6
Locus 21
1,3
15
19
12
Variables sites between Taxon C and G
0
10
20
30
40
50
20
40
60
80
100
Variable Sites Across All Taxa

Figure 4

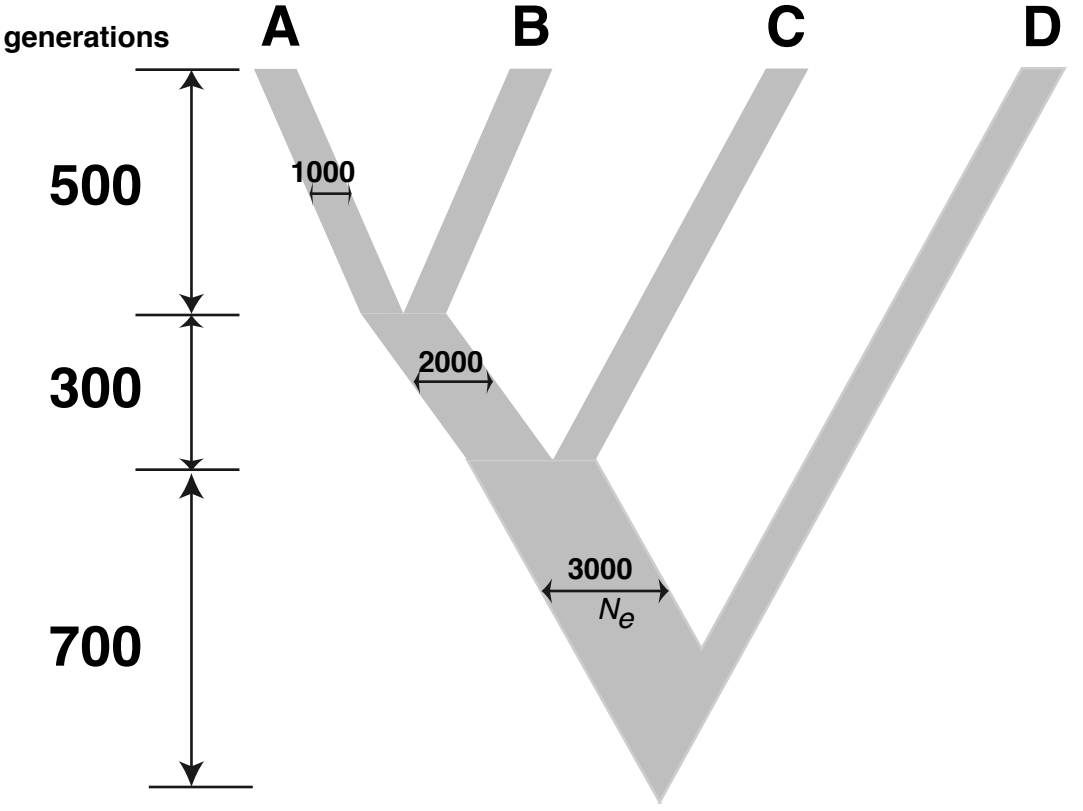


Figure 5