Understanding How Stochasticity Impacts Reconstructions of Recent Species Divergent History

by

Huateng Huang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in The University of Michigan
2012

Doctoral Committee:

Associate Professor L. Lacey Knowles, Chair
Assistant Professor Timothy Y. James,
Assistant Professor Sebastian K. Zöllner
Associate Professor Laura S. Kubatko, Ohio State University

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Molecular phylogenetic studies are complicated by the fact that differentiation between orthologous gene copies is determined by two stochastic processes, namely, the mutational and lineage sorting (coalescent) processes. The stochasticity associated with the mutational process has been extensively examined for its effect on gene-tree estimation in past decades, whereas only recently has the idea of incorporating the coalescent process into species-tree estimation been applied in empirical phylogenetics. Because of the stochastic lineage sorting process, the divergence time between genes is always longer than the divergence time between species and the gene-tree topology can differ from the species-tree topology (Figure 1.1). The histories of gene lineages are not equivalent to the histories of species divergence. Variations between loci on a genome and between individuals in a population should be considered in the study of recent species divergence. My thesis focuses on examining the impacts of these two processes on the species-tree estimates with both simulated and empirical data, and also answering relevant questions for empirical phylogenetic studies. The following provides a short overview of the five projects in this thesis.

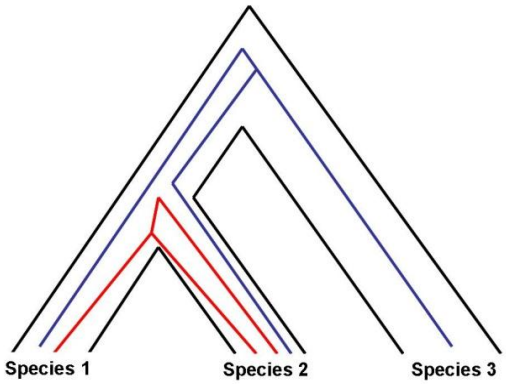


Figure 1.1 Possible incongruences of genealogies and species tree.
Blue line indicates a genealogy with a different topology from the species tree. Red line indicates a genealogy showing a paraphyletic structure between species 1 and 2

Realized Anomalous Gene Trees (AGTs). Recent theoretical work on the coalescent process has identified a very ominous situation in which the most probable gene trees do not match the underlying species tree. That is, "democratic voting"—a simple and intuitive way to solve the problem of gene-tree discordance might not work. However, empirical data contain another important stochastic component – mutational variance. Hence, my work takes a simulation approach to investigate the prevalence of AGTs, among estimated gene trees, thereby characterizing the boundaries of the anomaly zone taking into account both coalescent and mutational variance. The frequency of realized AGTs is also determined, which is critical to putting the theoretical work on AGTs into a realistic biological context. The result shows that mutational variance can indeed expand the parameter space (i.e., the relative branch lengths in a species tree) where AGTs might be observed in empirical data, because the mutation process and gene-tree estimation procedure also bias towards the anomalous gene trees. Using loci with more informative sites and rooting the trees are the strategies that could reduce this bias. Moreover, for the empirical species histories where AGTs are possible, unresolved trees – not AGTs – predominate the pool of estimated gene trees. This result suggests that the risk of AGTs, while they exist in theory, may rarely be realized in practice, and the challenge for empirical studies is how to extract information from unresolved gene trees.

Relative effects of coalescent and mutational variance on species-tree estimation. Current species-tree estimation methods differ considerably in dealing with the variance generated by mutation process. For example, Minimizing Deep Coalescent method only took the gene tree topology (i.e., handling the mutation variance causing topological changes, while ignore those changing the gene-tree branch lengths); Species Tree Estimation using Maximum-likelihood includes the branch length information but assuming it is correctly estimated. To assess the adequacy of these methods, it is necessary to re-evaluate the effect of mutational variance in the species-tree estimation context. In this part of my work, I partitioned the effects of mutational and coalescent processes on accuracy of species-tree estimates by comparing species trees estimated from gene trees (i.e., the actual coalescent genealogies) to those estimated from estimated gene trees (i.e. trees estimated from nucleotide sequences, which contain both coalescent and mutational variance). The result shows that the relative magnitude of the effects

differs systematically depending on the timing of divergence, the sampling design, and the method used for species-tree estimation, which explains why using more information on gene trees does not necessarily translate into more accurate species-tree estimates. The performance of a method depends not only on the method per se, but also on the compatibilities between the input genetic data and the method (e.g., methods that are insensitive to mutation variance would be more accurate for a set of loci with fewer informative sites than methods that relies on accurate estimation of gene trees).

Assessing the impact of coalescent variance in previous empirical studies. Many previously published phylogenetic estimates are based on single loci or the concatenation of multilocus data despite discord in the gene trees of individual loci. Future phylogenetic studies will no doubt benefit from the increased availability of genomic data, coupled with new computation methods. What is not clear is the extent to which species relationships estimated with data and methods that predate these developments are robust given a fundamental assumption of past analyses is now known to be untrue. In this part of my work, I proposed a parametric bootstrap species tree (PBST) approach to assess how the reliability of past phylogenetic studies have been affected by overlooking the stochastic lineage sorting process. An assumption is first made that previously published single-locus gene trees represent the "true" species trees, which is followed by simulation of the lineage sorting process under these "true" species trees. The robustness of the published single-gene trees is assessed by comparing each published gene tree with the corresponding simulated gene trees. This PBST approach is applied as a meta-analysis of east African cichlid phylogenies – a model system for evolutionary radiations. While some inferences are robust, many gene-tree based phylogenetic analyses of cichlids have a high probability of being misleading, and this difference is concordant with the radiation history of cichlids. This approach is also used to assess the likely clade-specific performance of species-tree estimation methods given different sampling strategies.

Estimating Species Tree with Next Generation Sequencing (NGS) data. Next-generation sequencing (NGS) combined with Reduce Representation Library (RRL) technique has the premise of generating multilocus sequence data for non-model organisms in a quick and low cost way. Nevertheless, this technique is mainly used as means for marker development, and concerns exist about whether NGS data with high

error probability are amenable for direct use for phylogenetic analysis. In this study, I explored the use of NGS as primary data for reconstructing the divergent history of four montane grasshopper species. NGS data was obtained from 1/13 of a multiplexed 454 sequencing run, a model was developed to jointly estimate the genotypes and haplotypes. Twenty-five highly variable and phylogenetic informative loci with sequences from all species were found. A Bayesian species-tree estimate was obtained, and the effect of including loci with low variation and adding additional filters for sequencing errors were assessed by comparing the estimated species trees. Parametric simulation was used to examine three possible sources of uncertainty in the estimated species tree: the true species divergent history, sequencing errors and error correction method. Possible improvement on sampling design and the methodological developments needed for future studies are discussed.

Inferring the process of speciation from geographic distribution of molecular genetic variation. When species divergence is relatively recent, the footprint of the demographic history during speciation might be preserved and used to reconstruct the biogeography of species divergence. In this study, patterns of genetic variation were examined throughout the geographical range of two cryptic sister taxa of field crickets, *Gryllus texensis* and *G. rubens*. Despite significant molecular divergence between the species, they were not reciprocally monophyletic. Several analyses were devised to statistically explore what historical processes might have given rise to this genealogical structure. The analyses indicated that the bio- geographical pattern of genetic variation does not support a model of recent gene flow between species. Instead, coalescent simulations suggested that the genealogical structure within *G. texensis*, namely a deep split between two geographically overlapping clades, reflects historical substructure within *G. texensis*. Additional tests that consider the concentration of *G. rubens* haplotypes in one of the two *G. texensis* genetic clusters suggest a model of speciation in which *G. rubens* was derived from one lineage of a geographically subdivided ancestor (i.e., an peripatric origin in which *G. rubens* was derived from one of the lineages in the geographically subdivided ancestor). This proposed model of species divergence suggests how the interplay of geography and selection may give rise to new species, although this requires testing with multilocus data.

Chapter 2 Anomaly Zone Dangers for Empirical Phylogenetics

Incongruence between gene trees and species trees has long been acknowledged as a serious challenge for phylogenetic studies (Pamilo and Nei, 1988; Takahata, 1989; Maddison, 1997). However, it is only recently that the potential magnitude of the problem has become apparent. Discordant gene trees are routinely encountered in multilocus studies (e.g. Jennings and Edwards, 2005; Wong et al., 2007; Carstens and Knowles, 2007; Carling and Brumfield, 2008; Knowles and Carstens, 2007; Sanderson et al., 2008). Moreover, recent theoretical work has also identified a very ominous situation in which the most probable gene trees do not match the underlying species tree – anomalous gene trees, AGTs (Degnan and Rosenberg, 2006; Rosenberg and Tao 2008). Under such species histories, the gene trees can lead to incorrect conclusions about the history of species divergence with current methodologies. Moreover, within the anomaly zone increased sampling of loci, by itself, will not increase the accuracy of phylogenetic inference because the most frequent gene trees will provide positively misleading information about species relationships.

While the theoretical proof of the existence of AGTs is alarming, the actual risk that AGTs pose to empirical phylogenetic study is far from clear. First, establishing the conditions (i.e., the branch-lengths in a species tree) for which AGTs are possible (Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008) does not address the critical issue of how prevalent they might be. For example, if AGTs are possible, but not probable, then even for those species histories where AGTs can theoretically occur (i.e., species trees within the anomaly zone), they may not represent a significant danger. On the other hand, if the frequency of AGTs for a given history of species divergence is high, then they may very well result in misleading phylogenetic inferences. Second, theoretical characterization of the species trees for which AGTs may pose a problem is based on consideration of just one source of variance that contributes to species tree and gene tree discordance – gene lineage coalescence. Yet, empirical data contain another inherent

stochastic component – mutational variance. Estimated gene trees will differ from the underlying gene tree produced by the coalescent process because of the random process of mutation (Figure 2.1). The impact of this mutational variance on the zone (i.e., species tree branch lengths) for which AGTs will be realized in empirical data remains to be investigated; therefore, unlike previous theoretical investigations (Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008), are study focuses on estimated gene trees that are AGTs – that is, the most frequent estimated gene tree does not match the species tree.

Here we take a simulation approach to investigate the prevalence of AGTs to determine how significant a threat they actually represent for empirical phylogenetic investigation. We focus on a four-taxon species tree with an asymmetric topology (Figure 2.2), which is the simplest tree that can produce AGTs. Moreover, the relative branch lengths of the species tree defining the boundary of the anomaly zone (i.e., the conditions under which AGTs are theoretically possible) has also been solved analytically (Degnan and Rosenberg, 2006). In this study we focus on both: i) the frequency of AGTs across the zone where AGTs are possible (i.e., the prevalence of AGTs for species trees with differing branch lengths), and ii) the impact of mutational variance on the boundary of the anomaly zone (i.e., what are the relative branch lengths in a species tree where AGTs are possible, and does this differ depending on the mutation rate).

Unlike the coalescent (Kingman, 1982), the complex properties of mutational variance make investigating its effect on species tree estimation difficult. Coalescence variance, under assumptions of a Fisher-Wright population, can be expressed via analytical equations (Takahata and Nei, 1985; Pamilo and Nei, 1988). That is, given a species history, the frequency spectrum of different gene tree topologies can be calculated (Degnan and Salter, 2005). In contrast, no such treatment of mutational variance exists, which no doubt reflects the difficulties associated with characterizing the multifarious effects of mutation. The difference between a single estimated gene tree (i.e., topology and branch lengths) from its underlying gene tree (e.g. Saitou and Nei, 1987) cannot simply be ascribed to the nucleotide substitution process (i.e., the model of molecular evolution, which includes parameters such as the transition-transversion ratio, frequencies of nucleotides, and heterogeneity in mutation rate across sites; Ripplinger and Sullivan, 2008). Differences between the actual genealogical history of a locus and the

estimated gene tree (Figure 2.1) can also arise from the criteria used for evaluating trees and tree-space-searching algorithms (e.g. Zhou et al., 2007). Lastly, when considered in the context of multiple independent loci, because estimated gene trees may differ from their underlying gene trees, mutation can also cause a deviation from the expected frequency spectrum of topologies for any given species tree (Degnan and Salter, 2005; Kubatko and Degnan 2007). Given that AGTs are defined by the frequency spectrum of gene tree topologies (i.e., AGTs are the most frequent tree topology, which nevertheless do not match the underlying species tree), any mutational-induced shift in the frequency spectrum of topologies needs to be examined. In terms of evaluating the threat that AGTs pose for empirical phylogenetic study, shifts in the frequency spectrum of particular interest would be those that could produce an expanded zone (i.e., species tree branch lengths) where AGTs are possible (i.e., a greater range of species histories might be subject to AGTs).

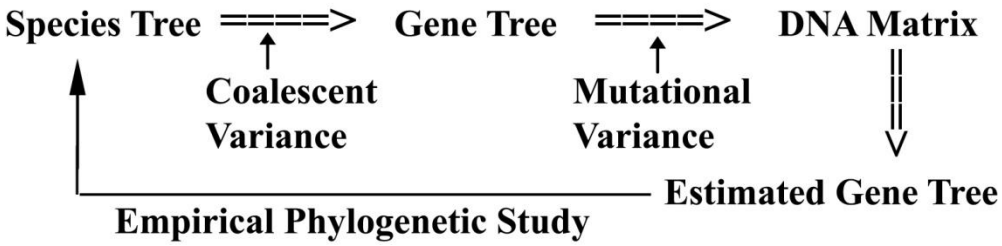


Figure 2.1 Concepts and terminology related to genetic source of variance in phylogeny.
The schematic shows the concepts and terminology relevant to investigating the genetic sources of variance in an empirical phylogenetic study. A species tree represents the actual history of divergence among species. A (true) gene tree in this context refers to the actual evolutionary history of orthologous genes across species, where the coalescent process introduces a variance in the gene trees across loci that evolve within the branches of different species. Estimated gene tree refers to the evolutionary history estimated from DNA sequences; because of the mutational process, and errors with estimating the gene tree, the topology of the estimated gene tree may not match the underlying gene tree for a locus. In contrast to previous treatments of the anomaly zone (i.e., Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008), this study considers both coalescent and mutational sources of variances for empirical phylogenetic study, focusing on estimated gene trees rather than gene trees

Our results show that mutational variance can indeed expand the parameter space (i.e., the relative branch lengths in a species tree) where AGTs are possible. The cause of the expansion is discussed along with detailed analysis of several divergence scenarios. Yet, when we examine the frequency of AGTs among the estimated gene trees for those species histories where AGTs are possible, we show that they are improbable for biologically realistic values of mutation rate (i.e., $\theta = 0.01$ to 0.001; Drost and Lee, 1995).

For these conditions, a polytomy (i.e., an unresolved internode) is more likely than an AGT. Therefore, while theoretically possible, there is insufficient mutation for AGTs to be realized in practice. As discussed, the minimum species tree branch lengths wherein estimated gene trees faithfully reflect the topology of the underlying gene tree actually exceeds the zone where AGTs are theoretically possible, meaning that AGTs themselves are unlikely to pose a significant danger to empirical phylogenetic study.

METHODS

General Simulation Procedure

When considering coalescent variance, two internal branch lengths on a four-taxon, asymmetric bifurcating species tree with no migration after speciation (species A, B, C, and D), *x* and *y* (see Figure 2.2), determine whether there are AGTs, as well as the number of AGTs (i.e. the number of topologies that occur with higher frequency than gene trees matching the species tree topology; see Degnan and Roseberg, 2006). To characterize the impact of mutational variance on the prevalence of AGTs for estimated gene tree, genealogies for diploid loci were simulated using the program Ms (Hudson, 2002) under a neutral coalescent model, with 1 individual per species, for different species trees with specific *x* and *y* branch lengths. For each gene tree, one set of DNA sequences with a length of 1000 base-pairs was simulated with the program Seq-gen (Rambaut and Grassly, 1997), using a HKY85 mutation model (discrete gamma distribution with a shape parameter of 0.8 and four categories, transition-transversion ratio of 0.3 with nucleotide probabilities set to 0.3 A, 0.2 C, 0.2 G, 0.3 T). A gene tree was estimated for each set of sequences with PAUP* version 4.0b10 (Swofford, 2003) and one gene tree was selected by the maximum-likelihood criteria (ML) from an exhaustive search. The frequency of the 16 tree topologies (i.e., 12 asymmetric topologies, 3 symmetric topologies, and any estimated gene tree with a polytomy, which was considered as one topological category) was calculated from the replicated data sets per species tree (see below for details about the number of replicates used).

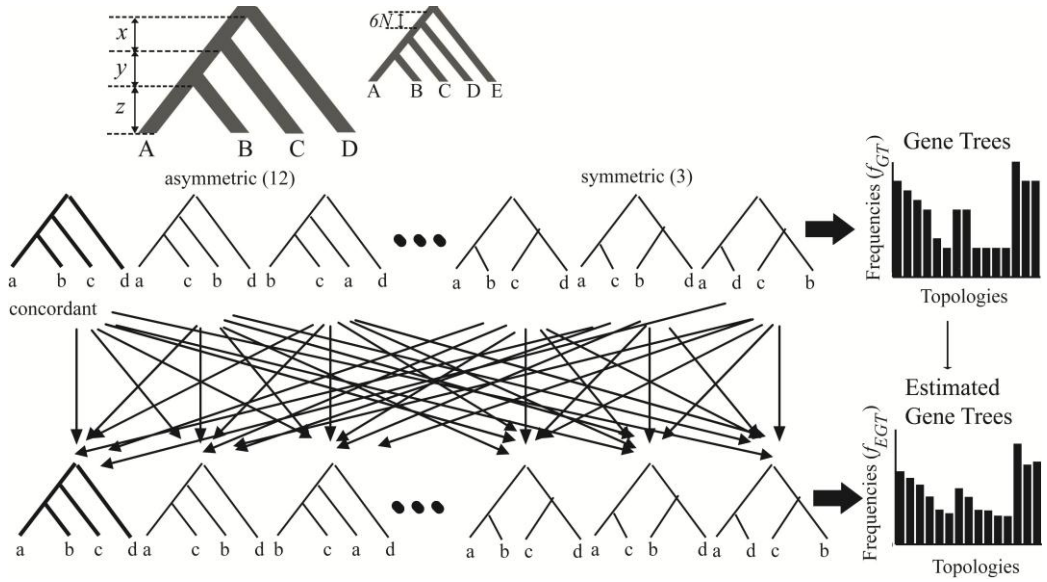


Figure 2.2 The probability distribution of gene trees and estimated gene trees.
For the four-taxon (species A, B, C, and D), asymmetric species tree used in the study, there are fifteen possible tree topologies: 12 asymmetric trees and 3 symmetric trees; an outgroup, species E, was used as the root. Variation in the species tree branch lengths *x* and *y* was explored to characterize the anomaly zone (i.e., species tree where the most frequent gene tree and/or estimated gene tree topology does not match the species tree). The frequency distribution of gene tree topologies (f_{GT}) may differ from the frequency distribution of estimated gene tree topologies (f_{EGT}), as illustrated with the two histograms, because of the mutation process (i.e., the estimated gene tree topology may differ from the actual underlying gene tree topology). Hence, the boundaries of the anomaly zone characterized with gene trees may differ when estimated gene trees are analyzed.

An outgroup species (E) that diverged 6*N* generations prior to the common ancestor of species A, B, C, and D (where *N*= effective population size) was used to root the estimated gene trees (Figure 2.2). With this branch length, species A, B, C and D formed a monophyletic group in most of the gene trees (i.e., in about 95% of the replicate data sets). In those few cases where a deep coalescence occurred between species A, B, C and D, and the outgroup, species E, the gene tree was excluded from further analysis to avoid additional coalescent variance being mistaken as mutational variance (i.e., the estimated gene tree differed from its underlying gene tree because species E was not an outgroup because of coalescent stochasticity); this does not affect conclusions about the prevalence of AGTs (i.e., the anomaly zone for the simulated data and based on the coalescent variance is comparable to that defined by Degnan and Rosenberg, 2006).

Characterization of the Frequency Spectrum of Estimated Gene Trees for Different Species Trees

Only symmetric gene tree topologies can have a higher probability than the topology that is concordant with a four-taxon species tree (Degnan and Rosenberg, 2006). The frequency of each of the three topologies (((a, b), (c, d)), ((a, c), (b, d)) and ((a, d), (b, c))), relative to the concordant topology (((a, b), c), d), was tallied for each species tree, under two simulation strategies designed to characterize the realized anomaly zone with mutational variance (i.e., the relative branch lengths in a species tree where AGTs are possible for estimated gene trees, specifically the branches x and y in the species tree; Figure 2.2).

In principle, the effect of mutational variance on the anomaly zones could be examined by simulating replicate data sets for each branch-length combination of x and y in the species tree. However, this approach has two potential problems. The number of replicate data sets needed for accurately assessing the existence of AGTs is not clear. For values of x and y near the boundaries of the anomaly zone, the probabilities of the different estimated gene tree topologies are so similar that the number of AGTs (i.e., 0 or > 0) may not be inferred accurately with a limited number of simulations. To combat this problem, we used a method of incrementally increasing the number of simulated data sets for each set of x and y branch lengths. For each species tree, the number of AGTs was assessed for every additional set of 1,000 replicate data sets. When there was no change in the number of observed AGTs with an additional two sets of replicates (in increments of 1,000), no additional data were simulated. In other words, the number of AGTs calculated was invariant across a minimum range of 3,000 estimated gene trees, and therefore it is unlikely that the estimated frequency of AGTs would change with additional data. To deal with the second problem of computational inefficiency in finding the boundaries of the anomaly zone, we used a bisection approach. For a given x, the y value starts from $0.01N$ and was continually updated to y' (= $y + 1N$) until reaching a point where (x, y') have 0 AGTs, thereby defining a parameter space where the boundary of the anomaly zone was crossed. The next set of simulations were conducted based on

the average value y'' ($=\frac{y+y'}{2}$), and depending on whether the number of AGTs at (x, y'') is bigger than 0, the next set of branch lengths explored in the parameter space was 0.25N distance either above or below y''. This bisecting step was repeated until the resolution (i.e., minimal distance between parameter values) was ± 0.015625 N (or ± 1,562 years for a species with a population size of 10^5 and one generation per year), which is a very small increment for phylogenetic study. This fine scale mapping of the boundary zone was conducted across a range of x from 0.11 to 1.91N, and 0.01 to 11N for y (these values span almost the entire anomaly zone, see results for details).

This procedure was carried out on data sets simulated under three different mutation rates: θ ($4N\mu$, μ=mutation rate per site) of 0.005, 0.01 and 0.05, which not only span the range observed in empirical data (i.e., θ = 0.01 to 0.001), but also includes an artificially inflated mutation rate (i.e., θ = 0.05). All these simulations were also explored under two differing external branch lengths of the species tree (i.e., z in Figure 2.2): z = 5N and 15N. While the external branch lengths have no effect on the frequency spectrum of gene tree resulting from the coalescent process (i.e., f_{GT} in Figure 2.2), z might have an effect on the frequency spectrum of estimated gene trees (i.e., f_{EGT} in Figure 2.2) through its effect on the absolute genetic distances between taxa.

Exploring the Cause for an Expansion of the Anomaly Zone

While the expansion of the anomaly zone with gene trees estimated from the simulated nucleotide data sets indicates that mutational variance has induced a shift in the frequency distribution of tree topologies (Figure 2.2), the cause for this shift is not intuitively obvious. Two different factors could contribute to this shift: i) the percent of estimated gene trees that match the topologies of the underlying gene trees, herein referred to as the percent of correct reconstruction (P_C), may differ across the fifteen possible topologies, and ii) the estimated gene trees that do not match the underlying gene trees, herein referred to as the percent of misidentification (P_M), may not be equally distributed among the fifteen possible topologies. To investigate the relative contributions of these two factors in causing a deviation from the expected frequency distribution of tree topologies, P_C and P_M were calculated for each of the 15 possible topologies. For

example, for a gene tree with the *ith* topology, P_C^i is the percentage of replicate data sets with an estimated gene tree with the *ith* topology, whereas P_M^i is the percentage of estimated gene trees with the *ith* topology representing misidentified topologies. Therefore, the frequency of the *ith* topology among the estimated gene trees (f_{EGT}^i) can be calculated as:

$$f_{EGT}^i = f_{GT}^i \times P_C^i + \{\sum_{i=1}^{15}[f_{GT}^i \times (1 - P_C^i)]\} \times P_M^i ,$$

where f_{GT}^i denotes the frequency of the *ith* topology in gene trees.

These analyses are performed on eight species trees located along the boundaries of the anomaly zone, namely for species trees with an *x* branch length of 0.05, 0.10, 0.15, and 0.20, and the corresponding *y* values, as calculated according to equations (4) and (5) in Degnan and Rosenberg (2006), where *z* was set to 5*N*. On average, 4729 replicate estimated gene trees were examined for each species tree from simulated data with θ = 0.01; the number of replicates for each species tree differed slightly because only those gene trees for which species A, B, C, and D formed a monophyletic group relative to the outgroup, species E, were considered from the 5,000 simulated data sets.

RESULTS

Impact of Mutational Variance on the Size of the Anomaly Zone

Compared to species trees with AGTs when just the coalescent variance is considered, there is an obvious expansion of the anomaly zone based on analysis of estimated gene trees, which also incorporate mutational variance (Figure 2.3). The expansion is apparent at all mutation rates and for different lengths of the external species tree branch, z, although the magnitude of the effect differs. It is worth noting that the degree of expansion shown here should also be considered conservative given that the mutational model of evolution was known, whereas as the substitution model for an empirical study would have to be estimated.

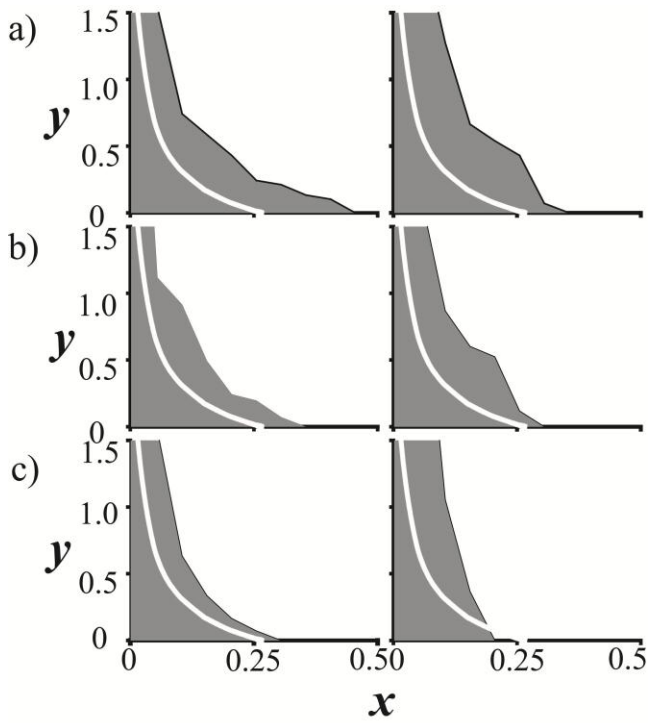


Figure 2.3 Anormaly zone with estimated gene trees.
Distribution of species tree branch lengths (i.e., x and y, in units of $2N$) that define the boundary of the anomaly zone, which shows an expansion of species trees characterized by AGTs when mutational variance is considered (shaded area). The anomaly zone below the white line delimits the area with AGTs based on consideration of just the stochasticity of the coalescent (i.e., characterization of the anomaly zone based on gene trees rather than estimated gene trees, Figure 2.1). Results from the different simulations conditions are shown, with plots on the left versus right reflecting species trees with $z = 5N$ and $z = 15N$, respectively, under three different mutation rates (a) $\theta = 0.005$, (b) $\theta = 0.01$, and (c) $\theta = 0.05$.

Cause for the Expansion of the Anomaly Zone

Expansion of the species tree branch lengths defining the boundary of the anomaly zone (Figure 2.3) reflects an increased number of estimated gene trees with anomalous topologies caused by mutation variance. This shift in the expected frequency distribution of topologies appears to be caused by a concomitant increase in the frequencies of estimated gene trees with symmetric topologies ($f_{GT} < f_{EGT}$) and a decrease in the frequency of the concordant tree topology ($f_{GT} > f_{EGT}$). The percentage of gene trees that are correctly reconstructed (P_C) reveals two distinct groups (Figure 2.4a), with asymmetric gene trees showing a significantly lower frequency of correct estimation compared to symmetric gene trees. This pattern identifies one mechanism underlying the shift in the anomaly zone – a deficit of correctly estimated asymmetric gene trees, relative to symmetric ones. When the estimated gene trees do not match the underlying

gene tree, examination of the percent of the misidentified gene trees (P_M) shows that both asymmetric and symmetric topologies are represented in similar proportions among the misidentified gene trees (Figure 2.4b). Although there is a slightly lower average P_M for symmetric gene trees, which is consistent with the inherent bias associated with phylogenetic estimation procedures (e.g., Huelsenbeck and Kirkpatrick, 1996), the relatively small effect suggests a minimal contribution to the shift in the frequency distribution of tree topologies. This suggests the expanding anomaly zone is not caused by an inflation of the frequency of symmetric trees from misidentified gene trees.

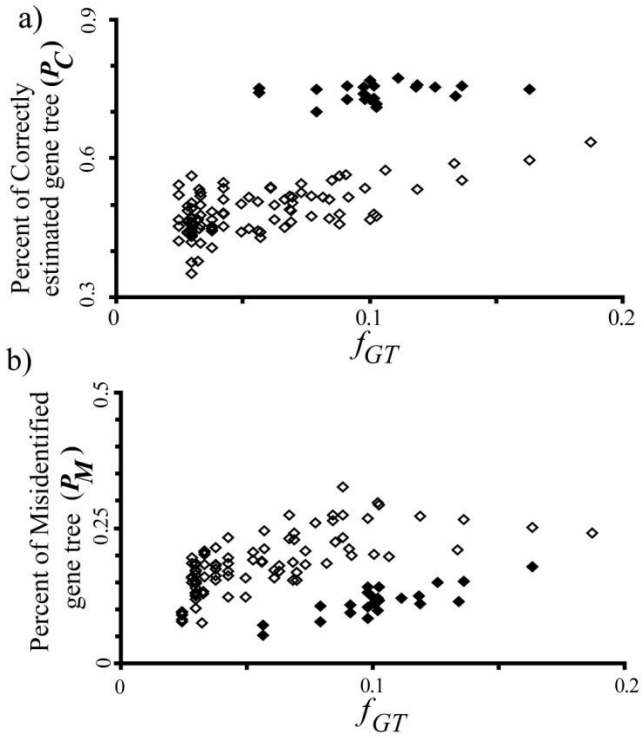


Figure 2.4 Comparative correct and incorrect estimations.
Comparison of the percentage of (a) correctly estimated (Pc) and (b) misidentified (P_M) gene trees for symmetric (shown in solid diamonds) and asymmetric topologies (shown as open diamonds) for species trees along the anomaly-zone boundary; the frequency of the respective gene tree topologies (f_{GT}) are shown along the x-axis.

Prevalence of AGTs based on Estimated Gene Tree

The expansion of the anomaly zone (Figure 2.3) when estimated gene trees are analyzed, relative to their underlying gene trees (Figure 2.1), suggests that AGTs might represent a bigger problem for empirical phylogenetic study than initially identified (Rosenberg and Degnan, 2006). In fact the increase in the anomaly zone caused by

mutational variance is comparable to or greater than the amount of expansion observed with increasing the number of taxa (Rosenberg and Tao, 2008). However, with estimated gene trees, there is another important class of topologies that doesn't exist with coalescent gene trees – estimated gene trees with polytomies (i.e., estimated gene trees with unresolved branches, as identified by a maximum-likelihood tree with zero branch length). This polytomy zone actually predominates species trees with very short internodes (Figure 2.5).

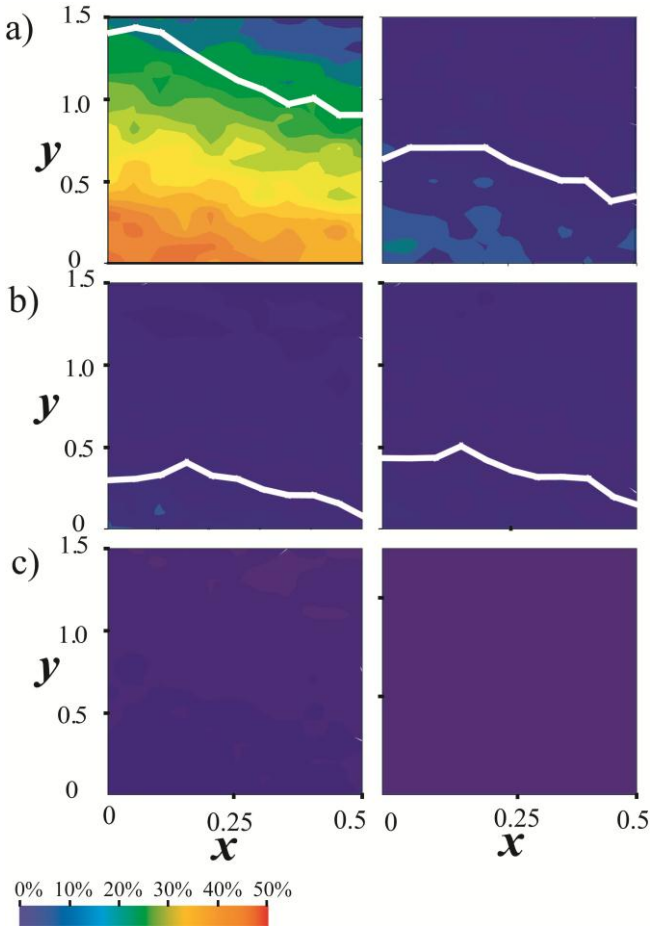


Figure 2.5 Frequency of polytomy gene trees.
Frequencies of estimated gene trees with a polytomy for the different species trees (i.e., differing x and y branch lengths, in units of $2N$) for different simulation conditions: specifically, for three different mutation rates (a) $\theta = 0.005$, (b) $\theta = 0.01$, and (c) $\theta = 0.05$, and a branch length of $z = 5N$ (shown on the left) and $z = 15N$ (shown on the right). The white lines demark the boundary of the polytomy zone – for all species trees under the white line, a polytomy is the most frequent topology of the estimated gene trees.

For this region of species tree parameter space, the polytomy zone overlaps broadly with most of the region where AGTs are possible (Figure 2.3). These analyses indicate that the most probable topology is a polytomy (Figure 2.5a, b), not an AGT; AGTs are

only realized for a confined parameter space (i.e., species trees with very short x-branch lengths and a y greater than 0.5), with $\theta = 0.01$ (Figure 2.5b). As the branch lengths (i.e., *x* and y) in the species tree increase, the polytomy zone is replaced by a region where estimated gene trees are resolved (i.e., area above the white line, Figure 2.5a, b). However, at these species tree branch lengths (i.e., areas where resolved estimated gene trees, not polytomies, predominate), the most frequent topology observed in the estimated gene tree is likely to be the one that is concordant with the species tree (i.e., the species tree branch lengths fall outside the anomaly zone – above the grey area Figure 2.3). Only with artificially high mutation rates (i.e., $\theta = 0.05$, Figure 2.5c), does the frequency of AGTs exceed the frequency of polytomies. However, loci with this high of a mutation rate, or non-recombining DNA fragments greater than the 1000 base-pairs used here, generally are rarely seen in phylogeny studies.

DISCUSSION

Given a sample of estimated gene trees from multiple independent loci, one might intuit that the actual species tree could be accurately identified using a democratic consensus procedure (e.g., Jennings and Edwards, 2005). However, the discovery of AGTs (Degnan and Rosenberg, 2006) indicated that even with unlimited data, the democratic consensus would not identify the correct species tree, which is alarming for phylogenetic studies. Moreover, recent study on a five-taxon species tree (Rosenberg and Tao, 2008) revealed that the anomaly zone expanded with the addition of taxa, again signaling an inherent danger for estimating species relationships in groups that have radiated recently (i.e., short internal branch lengths in the species tree). Notwithstanding the virtues of these theoretical studies, for empirical studies the question is how frequently will AGTs be represented among a set of estimated gene trees. Our analysis of the anomaly zone based on estimated gene trees (Figure 2.1), in contrast to previous treatments based on coalescent gene trees (Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008), provides this much-needed context. Two salient results emerge from this investigation into the effects of mutational variance on the anomaly zone. First, the documented expansion of the anomaly zone with estimated gene trees (as opposed to gene trees), and its underlying cause, identifies aspects of empirical data relevant to

avoiding the problems that AGTs pose for species tree inference from multi-locus data. Second, with realistic mutation rates (i.e., $\theta \leq 0.01$) the predominance of unresolved estimated gene trees, rather than AGTs, within the anomaly zone suggests that the risk of AGTs, while they exist in theory, may rarely be realized in practice.

Why are asymmetric gene trees less likely to be correctly estimated?

The analyses suggest that the cause of the expanded anomaly zone with estimated gene trees (Figure 2.3) is a deficit of correctly estimated asymmetric gene trees (Figure 2.4a), as opposed to a significant increase in the representation of symmetric topologies among the misidentified gene trees (Figure 2.4b). Why would asymmetric gene trees be less likely to be correctly reconstructed for a given species tree and mutation rate?

The answer appears to lie in the differing branch lengths of asymmetric and symmetric gene trees, with the shortest branch length in asymmetric gene trees being considerably shorter than the shortest branch length in a symmetric gene tree (Figure 2.6). Since the length of a gene tree branch determines the probability density function for the number of mutations, the greatest effects of mutational variance will be manifest with the shortest branches. A similar pattern is observed when the second-shortest branch of the gene tree is considered – that is, the average branch length is shorter for asymmetric compared to symmetric topologies, though the effect is much less dramatic (Figure 2.6b).

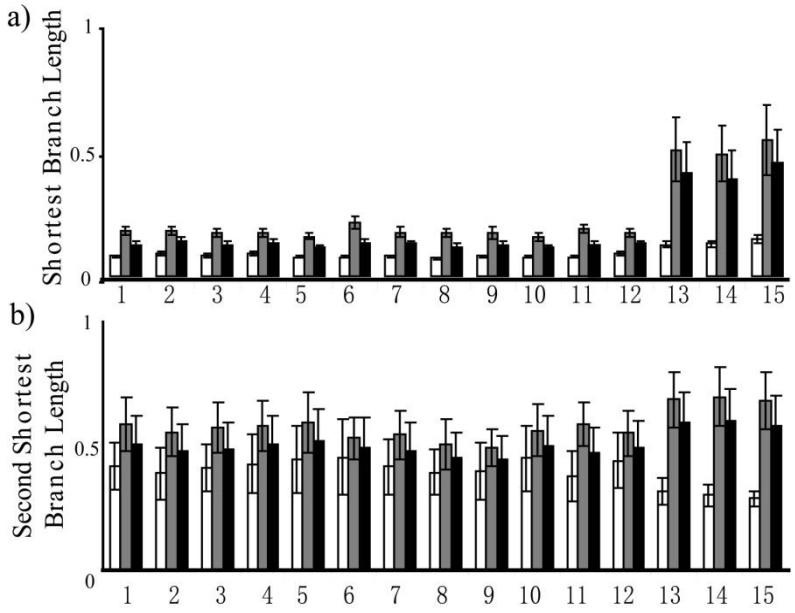


Figure 2.6 Difference in branch lengths for asymmetric and symmetric gene trees.
An example of how the average lengths (in units of *2N*) for the (a) shortest and (b) second shortest branch differ between asymmetric and symmetric topologies, based on 10,000 simulated data sets. The fifteen different gene tree topologies are shown along the x-axis, identifying the average length of the shortest and second shortest branch (shown in black), as well as the length of the shortest and second shortest branch for incorrectly estimated topologies (shown in white) and correctly estimated topologies (shown in grey). Asymmetric topologies are shown as 1 through 12, with 1 representing the one concordant with the species tree, and 13 through 15 are symmetric trees. The species tree is characterized by branch length $x = 0.10$ and $y = 0.088$.

To confirm that branch length, and not some other factor related to topology per se, is responsible for the shift in the frequency distribution of topologies of estimated gene trees (relative to the frequency distribution of gene trees, Figure 2.2), we examined the percentage of correctly estimated gene trees (P_C) for asymmetric and symmetric gene trees with the same average shortest branch (Table 2.1). When controlling for the differences in branch lengths (i.e., selecting species trees where asymmetric and symmetric gene trees had the same average shortest branch), there was no difference in the percentage of correctly estimated gene trees between asymmetric and symmetric topologies. Thereby confirming that it is not topology per se (Table 2.1), but the length of the shortest branch, and specifically the relatively shorter branches of asymmetric compared to symmetric gene trees (Figure 2.6), that results in a deficit of correctly estimated asymmetric gene trees, and hence an expansion of the anomaly zone.

By identifying the underlying cause for the expansion of the anomaly zone that occurs with estimated gene trees, there are several strategies empiricist can apply to avoid

potential problems with ATGs. The first is a simple one – choose loci with fast mutation rates. This will minimize the effects of mutational variance, which leads to a smaller realized anomaly zone (see Figure 2.3). Second, use outgroups to estimate species relationships, not a distance based procedure like mid-point rooting, where the effects of mutational variance may be further amplified. Indeed, a preliminary analysis shows a dramatic expansion of the anomaly zone with mid-point rooted trees (Figure 2.7).

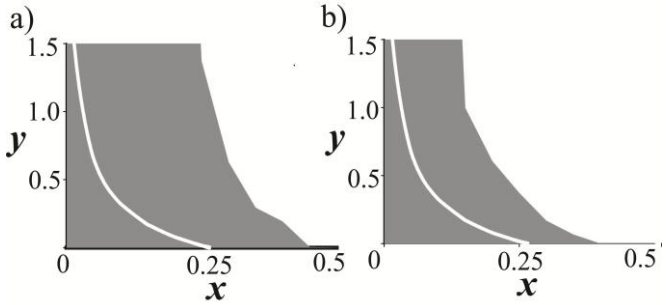


Figure 2.7 Anomaly zone with middle-point rooted gene trees.
Distribution of species tree branch lengths (x and y are shown in units of $2N$) that define the boundary of the anomaly zone for mid-point rooted estimated gene trees (shaded area) with $z = 5N$ (a) and $z = 15N$ (b), under mutation rate $\theta = 0.01$. The white dashed line marks the boundary for anomaly zones without mutational variance (i.e., based on analysis of gene trees).

Evaluating the Danger of AGTs for Empirical Phylogenetic Study

Despite the occurrence of AGTs, and an expansion of the anomaly zone when mutational variance is considered (Figure 2.3), our results also indicate that the danger AGTs actually pose for empirical phylogenetic study is limited. In contrast to coalescent genealogies, where the probability that more than two lineages will coalesce in the same generation is extremely small (Hudson 1990), polytomies dominate the estimated gene trees for the class of species trees located within the anomaly zone (Figure 2.5a, b). Therefore, instead of encountering AGTs, the most probable estimated gene tree an empiricist is likely to recover (at least for typical mutation rates) is one that is uninformative about the species tree. In other words, by focusing on the estimated gene trees, as opposed to coalescent gene trees, the results show that when resolved estimated gene trees are likely (and hence there is the potential for AGTs), AGTs are no longer a threat because the species trees are not located within the anomaly zone (i.e., the species tree branch lengths are too long to generate AGTs). Moreover, the flatness of the frequency distribution of estimated gene trees within the anomaly zone, as revealed by

the low maximum frequency of estimated gene tree topologies (Figure 2.8), also lessens any real danger of AGTs for empirical phylogenetic study. For example, even if the frequency of one anomalous topology is 15%, and other topologies have equal frequencies (i.e. 6% for the other 14 possible topologies), the chance that the anomalous topology will be the most frequent one in a sample of 20 loci is only 50%. Moreover, since the frequency of AGTs is actually much lower than 15% across most of the anomaly zone (typically slightly less than 7%; Figure 2.8), the chance that the most frequent topology among 20 loci will be the anomalous topology is indeed very, very low. Placing the danger posed by AGTs in an empirical context is important for highlighting the true challenges for phylogenetic study. For example, recently developed methods based on triplets offer one way that AGTs might be overcome (e.g., Degnan and Rosenberg, 2008; Ewing et al., 2008). Nevertheless, if the actual problem with the anomaly zone is the lack of resolution, not AGTs (as shown here; also see Ewing et al., 2008), then these methods will do little to address the challenges facing empirical phylogenetic study.

General Lessons from the Impact of Mutational Variance on Estimated Gene Trees

Investigation of the anomaly zone was motivated by the suggestion that the inherent mismatch between the most frequent estimated gene tree and the actual history of species divergence would pose significant (and perhaps insurmountable) challenges to obtaining an accurate estimate of species relationships (Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008). However, our study indicates that the danger of AGTs in practice is not what it is in theory, once mutational variance is taken into account. This finding does not mean that the difficulties with estimating species relationships (i.e., the underlying species tree) have gone away. The predominance of the polytomy zone, coupled with the low frequency of estimated gene trees with a topology matching the species tree (Figure 2.8) in adjoining regions of parameter space that define the polytomy (Figure 2.5) and anomaly zone (Figure 2.3) indicate that the primary focus should be developing a method that accurately extracts the gene tree signal to infer the species tree. By considering the biological realities of both mutational and coalescent variance, the study has refined, and

perhaps redefined, the problem by identifying what those challenges actually are for empirical phylogenetic study. Therefore, it is informative to consider the implications of our results for the procedures we might decide to use to estimate species trees (Maddison 1997).

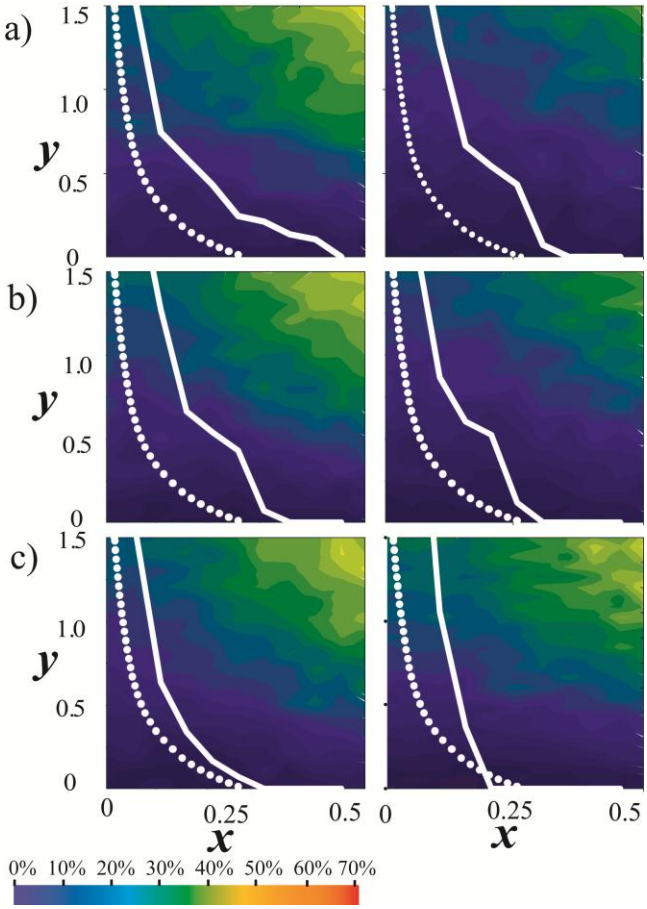


Figure 2.8 Highest frequency of resolved topologies in estimated gene trees for different species trees. x and y branch lengths are both expressed in units of $2N$, with $z = 5N$ (shown on the left) and $z = 15N$ (shown on the right), under the three different mutation rates: (a) $\theta = 0.005$, (b) $\theta = 0.01$, and (c) $\theta = 0.05$. The white solid and dashed lines mark the boundary of the anomaly zone based on estimated gene tree and true gene trees, respectively.

With respect to estimating a species tree, the available methodological procedures differ in how they extract information about the underlying history of species divergence, as well as the type of information they utilize. Studies have shown that depending on the approach, the accuracy of the estimated species trees can be similar across methods for some species histories (namely, older species divergences), but may differ considerably, especially for recently diverged species (e.g., Brumfield et al., 2008; McCormack et al.,

2009). Such studies highlight the potential gains that complex procedures, which extract more information from the estimated gene tree data, can offer for estimating species trees. While it may indeed be desirable to fully utilize the information contained in the estimated gene trees, this study suggests some caution is warranted. More information will be better than less, but only as long as the information being extracted from estimated gene trees is accurate. For example, in the case of AGTs, more loci will not necessarily provide a more accurate estimate of the species tree – the most probable gene trees (and therefore, the most frequent topology) will not match the underlying species tree (Degnan and Rosenberg, 2006). Likewise, for recent species divergence where the effects of mutational variance are exaggerated, an estimated gene tree will not faithfully reflect the genealogical history, differing not only in branch lengths, but also in topology (Figure 2.4). Even for species trees outside the anomaly zone, the percentage of correctly estimated gene trees (P_C) was below 75% (Table 2.1). These estimates should also be considered conservative, given that in this case the DNA substitution model used to obtain the estimated gene trees was known (i.e., it matched exactly the conditions under which the data were generated) and that an exhaustive search was performed to estimate the gene trees. In other words, the effects of mutational variance may be greater with empirical data than what is documented here. Consequently, it may not be possible to accurately estimate species relationships for such histories (i.e., those characterized by rapid and recent speciation), even with recent developed methods for directly inferring the underlying species tree (e.g., Maddison and Knowles, 2006; Carling and Brumfield, 2007; Liu and Pearl, 2007; Ewing et al., 2008; Knowles and Chan 2008; Kubatko et al., 2009; McCormack et al. 2009).

Table 2.1 The percentage of correctly estimated gene trees for different coalescent gene trees.
The percentages of correctly estimated gene trees (*Pc*) for the asymmetric topology that is concordant with the species tree and the symmetric coalescent gene trees when they have similar average shortest branch lengths; results are shown for four different species trees (branch lengths *x* and *y* are shown in units of *2N*), based on 10,000 simulated replicate data sets for each species tree.

		P_C	
x	***y***		
		concordant	symmetric

0.075	6.297	0.645	0.653
0.125	4.828	0.695	0.675
0.175	4.860	0.689	0.664
0.225	4.328	0.698	0.677

Another important issue and open question is how many loci and how much variation in the loci is needed to obtain an accurate estimate of the species tree, irrespective of what approach might be used for estimating species trees (e.g., minimizing the number of deep coalescences, estimating the most-likely species tree, or a Bayesian analysis for a species tree)(Maddison and Knowles, 2006; Liu and Pearl, 2007; Kubatko et al. 2008). Studies have shown that sampling effort and design have a significant impact (e.g., Maddison and Knowles, 2006; Edwards et al., 2007; McCormack et al., 2009); however, systematic investigation is lacking for recently developed methods, especially with regards to the mutational process. As shown here, mutational variance can cause not only a mismatch between an estimated gene tree and the underlying gene tree for any single locus, but it also results in a flatter frequency distribution of tree topologies than expected (Degnan and Salter, 2005). This is expected to increase the number of sampled loci needed to obtain an accurate characterization of the underlying probability distribution of estimated gene tree topologies for a given species tree. It remains to be determined if the realized anomaly zone is as intractable as the classic Felsenstein zone, another example where phylogenetic accuracy is compromised by the mutational process. Mutational variance, as an inevitable part in empirical sequence data, obviously needs to be investigated in the context of species tree estimation.

REFERENCES

Brumfield, R. T., L. Liu, D. E. Lum, and S. V. Edwards. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, Manacus) from multilocus sequence data. Syst Biol 57:719-31.

Carling, M. D., and R. T. Brumfield. 2007. Gene sampling strategies for multi-locus population estimates of genetic diversity (theta). PLoS ONE 2:e160.

Carling, M. D., and R. T. Brumfield. 2008. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in Passerina buntings. Genetics 178:363-77.

Carstens, B. C., and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. Syst Biol 56:400-11.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet 2:e68.

Degnan, J. H., and N. A. Rosenberg. 2008. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol., in press.

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

Drost, J. B., and W. R. Lee. 1995. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. Environ Mol Mutagen 25 Suppl 26:48-64.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proc Natl Acad Sci U S A 104:5936-41.

Ewing, G. B., I. Ebersberger, H. A. Schmidt, and A. von Haeseler. 2008. Rooted triple consensus and anomalous gene trees. BMC Evol Biol 8:118.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 in D. Futuyma and J. Antonovics, eds. Oxford surveys in evolutionary biology. Vol. 7. Oxford University Press, New York.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-8.

Huelsenbeck, J. P., and M. Kirkpatrick. 1996. Do phylogenetic methods produce trees with biased shapes? Evolution 50:1418-1424.

Jennings, W. B., and S. V. Edwards. 2005. Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. Evolution 59:2033-47.

Kingman, J. F. C. 1982. The coalescent. Stochastic Process. Appl. 13:235-248.

Knowles, L. L., and B. C. Carstens. 2007. Estimating a geographically explicit model of population divergence. Evolution 61:477-493.

Knowles, L. L, and Y-H. Chan. 2008. Resolving species phylogenies of recent evolutionary radiations. Ann. Missouri Bot. Gard. 95:224-231.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971-973.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56:17-24.

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

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

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55:21-30.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol Biol Evol 5:568-83.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13:235-8.

Ripplinger, J., and J. Sullivan. 2008. Does choice in model selection affect maximum likelihood analysis? Syst Biol 57:76-85.

Rosenberg, N. A., and R. Tao. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst Biol 57:131-40.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-25.

Sanderson, M. J. 2008. Phylogenetic signal in the eukaryotic tree of life. Science 321:121-3.

Swofford, D. L. 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*And Other Methods). Ver 4.0b.Sinauer Associates, Sunderland, Massachusetts.

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957-66.

Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325-44.

Wong, A., J. D. Jensen, J. E. Pool, and C. F. Aquadro. 2007. Phylogenetic incongruence in the Drosophila melanogaster species group. Mol Phylogenet Evol 43:1138-50.

Zhou, H., J. Gu, S. J. Lamont, and X. Gu. 2007. Evolutionary analysis for functional divergence of the toll-like receptor gene family and altered functional constraints. J Mol Evol 65:119-23.

Chapter 3 Sources of Error for Species-Tree Estimation: Impact of Mutational and Coalescent Effects on Accuracy and Implications for Choosing Among Different Methods

Recent developments in phylogenetic reconstruction methods for the direct estimation of species trees emphasize the considerable challenge of recovering species divergence histories from patterns of molecular genetic divergence. Molecular phylogenetic studies are complicated by the fact that differentiation between orthologous gene copies is influenced by two major sources of stochastic genetic variance – mutational and coalescent variance (Figure 3.1) (Maddison 1997). The first one has been extensively examined for its effect on gene-tree estimation in past decades. For example, sophisticated models have been developed to capture the heterogeneous substitution process across the genome and along the branches of a gene tree (e.g., Singh et al. 2009). The relative merits of collecting more base pairs from one fragment (e.g., whole mitochondrial genome) versus concatenating data from multiple independent loci in terms of attaining higher nodal support values has been evaluated (e.g., Rokas et al. 2003; Robins et al. 2008; although higher support may simply be an artifact of constraining the data to fit a single tree when in reality there is a mixture of trees, as described by Mossel and Vigoda 2005; Cranston et al. 2009). Statistical tests have also been designed to investigate whether gene trees differ significantly (e.g., likelihood-ratio test, Huelsenbeck et al., 1996). However, all of these efforts focus exclusively on the mutation process, and thereby share the problematic null hypothesis that there is only one "true" tree for every gene. Only recently has the idea of independent gene trees been revived in empirical phylogenetics— each gene tree is a different realization of a stochastic lineage sorting process, with the mutation process subsequently acting upon each realized gene tree (Pamilo and Nei 1988; Takahata 1989; Maddison 1997; Edwards 2009; Degnan and Rosenberg 2009; Knowles 2009a). Therefore, it is necessary to re-evaluate the effect of

mutational variance in this new context where differences in the genealogical history of loci are also explicitly acknowledged.

The question of how to account for differences in the genealogical history of loci in phylogenetics has triggered the development over the past few years of several new methods (e.g., Maddison and Knowles 2006; Ané et al. 2007; Mossel and Roch 2007; Kubatko et al. 2009; Liu 2008; Liu et al. 2009; Knowles and Kubatko 2010). However, in contrast to the extensive evaluation of methods in traditional phylogenetics that has focused on mutational processes while ignoring the coalescent process (e.g. Ripplinger and Sullivan 2008), the effects of mutational variance on the accuracy of species-tree estimates in this emerging field has yet to be thoroughly explored. For example, theoretical studies (Degnan and Rosenberg 2006) based solely on the properties of the coalescent process have revealed a counterintuitive result in which the most frequent gene tree will not match the underlying species tree (i.e., anomalous gene trees, AGTs; but see Huang and Knowles 2009). Gene tree probabilities have been used for species tree estimation with the assumption of correctly reconstructed gene trees (Carstens and Knowles 2007), and not all coalescent-based methods for species tree estimation take into account the contributions of mutational variance (Maddison and Knowles 2006; Kubatko et al. 2009). The difficulties with examining the specific impact of mutational variance in the context of species-tree estimation no doubt contributes to the lack of thorough investigation. For example, accounting for the mutation process requires simulation approaches that are time intensive, especially in contrast to the analytical tractability of the coalescent process (Huang and Knowles 2009). Likewise, for species-tree estimation procedures where the input data are DNA sequences (e.g., the program BEST; Liu 2008), the effects of mutation and coalescent variance cannot be disentangled, and again would require time-consuming simulation approaches to examine different population-mutation parameters. Nevertheless, understanding the effects of mutation relative to the coalescent on the accuracy of estimated species trees is fundamental to the development of this nascent area of phylogenetic study, as it was for obtaining accurate estimates of gene trees (e.g., Kimura 1980; Huelsenbeck and Hillis 1993; Gaut and Lewis 1995; Sullivan and Swofford 1997, 2001).

Here, we use a simulation approach to study the relative effects of mutational and coalescent variance on the accuracy of estimated species trees. The impact of these two stochastic genetic processes is investigated under two methods: STEM (Species Tree Estimation using Maximum-likelihood; Kubatko et al. 2009) and MDC (Minimizing Deep coalescent; Maddison and Knowles 2006; Than and Nakhleh 2009). These particular methods were chosen because both methods use gene trees as input. This provides a unique opportunity for partitioning the errors in species-tree estimates between those arising from mutation versus gene lineage coalescence. By using gene trees (as opposed to DNA sequences) as input, the accuracy of species-tree estimates obtained from estimated gene trees versus coalescent gene trees can be compared, thereby providing a measure of the error associated with coalescent and mutation variance versus coalescent variance alone (Figure 3.1).

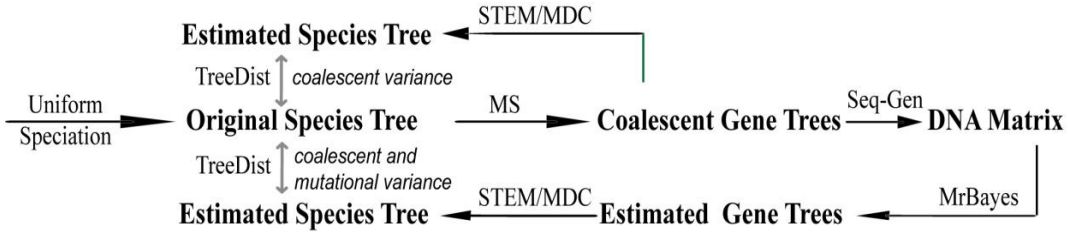


Figure 3.1 Error contribution from mutual and coalescent variance.
The contribution of mutational and coalescent variance to errors in species-tree estimates can be evaluated by comparing species trees estimated from coalescent gene trees (i.e., the actual genealogy of independent loci) to those estimated from estimated gene trees. The discordance between the true species tree and a species tree estimated from gene trees reflects the combined effects of two variances, whereas only coalescent variance is represented in the discordance between the true species tree and a species tree estimated from coalescent gene trees.

Furthermore, contrasting the results from the two methods could also be informative with regard to how different methodological simplifications might influence the sensitivity of species-tree estimates to mutational versus coalescent variance specifically. Although MDC and STEM both use gene trees as the input (as opposed to nucleotide sequences), they extract different types of information from the gene trees for estimating species trees. The MDC approach uses only information contained in the gene-tree topologies, whereas STEM incorporates information contained in both the topologies and branch-lengths of gene trees. The two methods also differ in their treatment of coalescent variance (i.e., how they incorporate conflicting genealogical information into the species-tree estimation procedure). STEM evaluates the likelihood of a species tree based on a

full probabilistic model of gene lineage coalescence. MDC is a summary-statistic approach (i.e., it is based on the minimal number of deep coalescent events) instead of explicitly modeling the coalescent process. Because the computational time of both methods is manageable, we can also test the generality of the relative impact of mutational and coalescent processes on species-tree accuracy by examining a large number of replicates for each simulated history for a diverse array of species phylogenies, as well as different sampling configurations.

In addition to interest in understanding how mutational and coalescent variance affect a methods' ability to recover the actual history of species divergence, the aim of the simulation study is also to develop guidelines for empirical studies in two key aspects: (i) choosing among different methods for estimating species trees given specific data configurations (i.e., total sampling effort and how it is divided across sampled individuals versus loci), and (ii) choosing sampling strategies that minimize errors in species-tree estimates. As the available genetic markers and number of individuals sampled per species varies from study to study, species tree estimation methods are likely to result in differing levels of errors among empirical datasets. Despite having several species-tree estimation methods to choose from, the robustness of the species-tree estimates to different data set properties is a topic that has yet to gain the attention it deserves. We lack a basic understanding of what aspects of the species divergence histories and data set properties make certain aspects of species-tree estimation procedures particularly reliable (or unreliable). Rather than comparing the performance across a limited set of histories (which would be a constraint imposed by computationally intensive methods like the program BEST; Liu 2008), this work is focused on developing a more general understanding of how the basics of mutation and gene lineage coalescence impact our efforts to estimate species trees, and how these principles can help us make informed decisions for our phylogenetic studies.

METHODS

Two discordance scores were used to quantify the inaccuracies of species-tree estimates attributable to mutational and coalescent processes. The first quantifies the discordance between the true species tree and the species trees estimated from coalescent

gene trees. Given that gene trees are generated by the coalescent process (Fig 1), this discord (D_C) represents the effect of coalescent variance on species-tree inference (i.e., the discordance due to coalescent variance alone). The second discordance score is based on the difference between the true species trees and the species tree estimated from estimated gene trees. This discord (D_{CM}) contains both coalescent and mutational sources of variance. Thus, the difference between the two scores (D_{CM}-D_C) should correspond to the effect of mutational variance in species-tree estimates, or the discordance due to mutational variance (denoted D_M). All discordances were measured using Robinson-Foulds distance (Robinson and Foulds. 1981), which quantifies the differences between two different tree topologies, implemented in the program TreeDist (which is part of the PHYLIP statistical package; Felsenstein 1993). This symmetric distance for rooted trees assesses the number of clades found in one tree, but not the other. Larger values correspond to lower accuracy, with a score of zero indicating no topological discord (i.e., all clades are the same between the true and estimated species trees) and a maximum discord of 12 (i.e., twice the number of internal branches for a rooted tree with 8 terminal taxa). We also calculate $P = (D_{CM}\text{-}D_C)/ D_{CM}$ as the percentage of discordance due to mutational variance.

The general steps of the simulation (see Figure 3.1) involve: (1) generating a species tree under a uniform speciation model, (2) simulating coalescent gene trees for each species tree, (3) simulating DNA sequences under a specified model of nucleotide evolution along the branches of each gene tree, (4) estimating gene trees from the simulated DNA matrix, (5) estimating species trees from the coalescent gene trees and estimated gene trees, and (6) calculating the discordance score between the true species tree and the two species-tree estimates (i.e., D_{CM} and D_C). These steps were repeated for 50 different species trees at two different times of divergence and a range of sampling designs (i.e., different numbers of loci and individuals), as discussed in detail below.

Eight-taxon species trees were generated in the Mesquite software package (Maddison and Maddison 2009), and coalescent gene trees were generated using the program *ms* (Hudson 2002). Gene trees were simulated under a neutral coalescent model with constant population size and no migration after speciation. For each individual DNA sequence, 1000 base pairs were generated with the program Seq-Gen (Rambaut and

Grassly, 1997) under an HKY85 model of nucleotide substitution with a transition-transversion ratio of 3.0, a gamma mutation rate distribution with shape parameter of 0.8, and nucleotide frequencies of A = 0.3, C = 0.2, T = 0.3, and G = 0.2 for the ancestral sequence. The HKY model is a commonly used model in phylogeny literature, with a moderate level of complexity and flexibility in terms of the number of estimated parameters. Different loci were independently simulated under a coalescent model, representing loci with free recombination between loci, but no recombination within each locus. From the simulated DNA sequences, gene trees were estimated using MrBayes version 3.1.2 (Huelsenbeck et al., 1996) with a molecular clock (a requisite assumption for all species-tree estimation procedures) and Dirichlet distribution as the prior for nucleotide frequencies. For the estimated gene trees, the actual parameter values for the HKY85 model were estimated for each set of sequences. By estimating the model that generated the sequence data for the procedure of estimating gene trees, model misspecification is not an additional source of error in the estimated gene trees (see Ripplinger and Sullivan, 2008), which allows us to focus specifically on the contribution of mutational variance alone. Mrbayes was stopped after the standard deviation of two independent runs dropped to less than 0.01. A consensus tree was calculated after discarding the first 25% of the total number of generations as burn-in.

Species trees were estimated from coalescent gene trees and estimated gene trees with the two methods: MDC (Minimizing deep coalescent, implemented in Mesquite; Maddison and Maddison 2009), and STEM (Species Tree Estimation using Maximum likelihood; Kubatko et al. 2009). MDC heuristically searches tree space for the specific species-tree topology that minimizes the number of deep coalescences (i.e., ancestral coalescent events prior to speciation; Maddison and Knowles 2006). We used STEM to derive analytically the maximum likelihood (ML) species tree (both branch lengths and topology) for a set of gene trees with branch-length information (see Liu et al. 2010 for details), where θ ($4N\mu$, where μ is the mutation rate per site per generation and N is the effective population size) was set to 0.01, matching the conditions under which the data were simulated.

Data were collected from 50 species trees for a recent divergence (total tree depth of $1N$ generations) and deeper divergence (total tree depth of $10N$ generations). For both

tree depths, nine sets of data were created with different numbers of individuals and loci sampled (i.e., a ratio of individuals to loci = 1:1, 3:1, 9:1, 27:1, 1:3, 3:3, 9:3, 1:9, 3:9). For each species tree at the two total tree depths and each sampling design, 20 independent replicates were generated; these replicates are used to calculate average D_C and D_M scores for each species tree under the different sampling strategies. To characterize the general effects of different methods, tree depths, and sampling strategies on the accuracy of species-tree estimates, D_C and D_M scores were averaged across the fifty species trees. A detailed examination of the effect of sampling strategy was also performed where gains in accuracy were calculated for each species tree per the addition of a single locus across a broad range of loci sampled in each taxon (1 to 50 loci for dataset with 1 sampled individual per species, 1 to 30 loci for 3 sampled individuals, and 1 to 10 loci for 9 individuals). Incremental gains in accuracy were evaluated from average discordance scores calculated from 10 replicates for each species tree under each sampling configuration and time depth, where gene trees were sampled randomly from the total pool of gene trees for each species tree (i.e., from each species-tree specific gene-tree pools, which contained 540, 180, and 60 gene trees with 1, 3, or 9 sampled individuals per taxon, respectively).

RESULTS

Mutational and Coalescent Effects on the Accuracy of MDC Species-Tree Estimates

Coalescent variance (DC) contributes disproportionately to the discordance between species trees estimated by MDC and the true species trees relative to mutational variance (DM, Figure 3.2). With increased sampling (both adding loci and individuals) the effect of mutational variance relative to coalescent variance increases (P, Figure 3.3a). However, mutational variance always has a minor impact on the accuracy of species-tree estimates (P < 35%, Figure 3.3a). Improvements in the accuracy of species-trees estimates with increased sampling using the MDC approach reflect the lower contribution of coalescent variance (Figure 3.2).

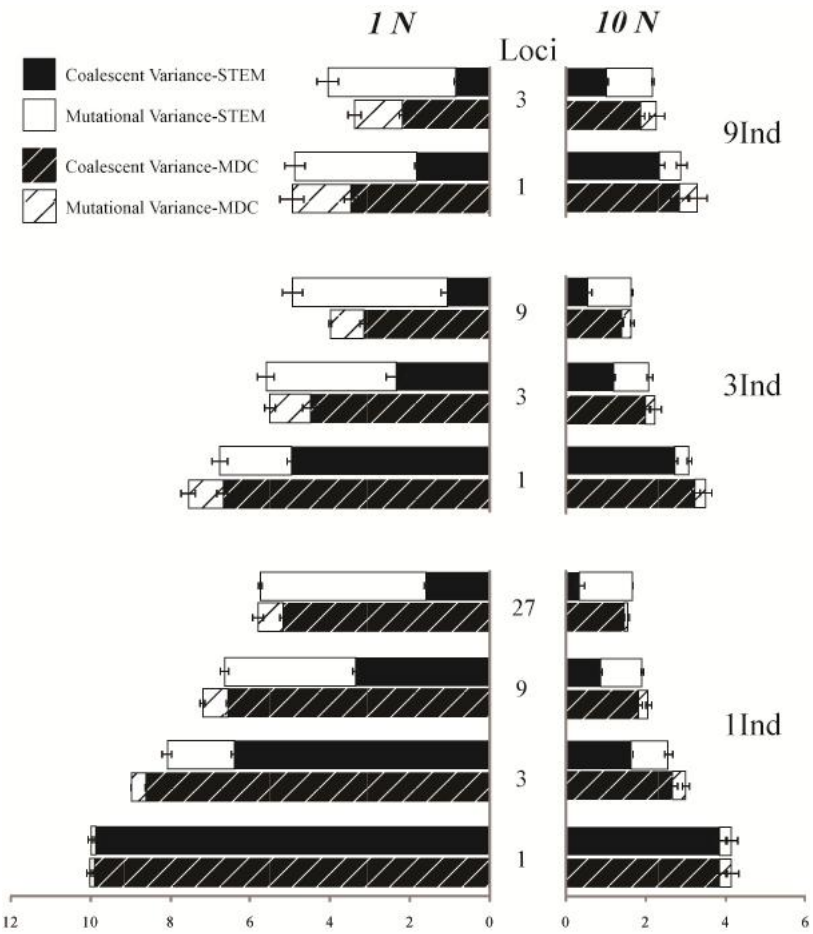


Figure 3.2 Dicordance between actual and estimated species trees.
Comparisons of the accuracy of species-trees estimated by the MDC and STEM methods with different true species-tree depths (1*N* and 10*N* generations, *N* is the effective population size, different number of sampled individuals per species (1Ind, 3Ind, 9Ind and 27Ind) and different number of sampled loci (1, 3 and 9 loci). The horizontal axis is the discordance (Robinson-Foulds distance) between estimated species trees and their true species trees averaged across the fifty original true species trees, with the error bars showing the standard error.

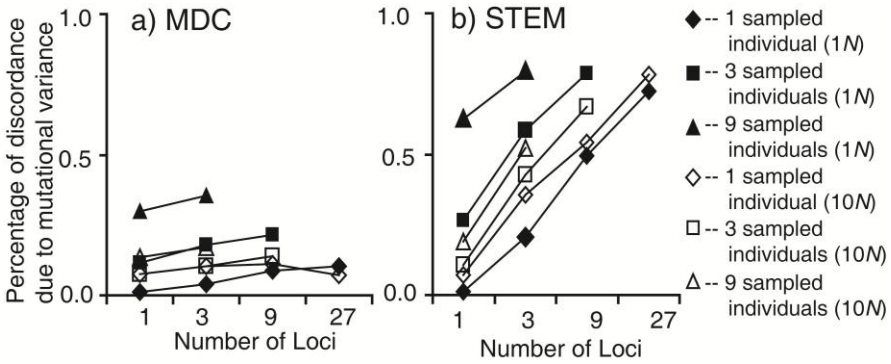


Figure 3.3 True and species tree estimation discordance.
The percent of the total discord between the true and species tree estimated with a) MDC and b) STEM due to mutational variance. Closed and open symbols represent results from the recent versus deeper divergence

histories, respectively, and lines connecting the symbols identify sampling strategies with the same number of individuals sampled per species.

The total tree depth of the species tree (i.e., recent versus deeper divergences) impacts the incremental gains in accuracy of species-tree estimates achieved by adding loci versus adding individuals. For shallow species trees (1*N* total tree depth), adding individuals results in greater gains in accuracy than adding loci (Figure 3.2). The increased accuracy achieved by adding individuals cannot be compensated for by adding loci instead of individuals (Figure 3.4). For deeper divergences (10N total tree depth), the addition of loci is more efficient in reducing species-tree estimation errors. Although there is a negligible effect on the errors in species-tree estimates for deeper species trees when sampling 9 versus 3 individuals (Figure 3.4), there is a notable increase in the accuracy of species-tree estimates with the MDC approach when 3 versus 1 individual are sampled per species (Figure 3.4).

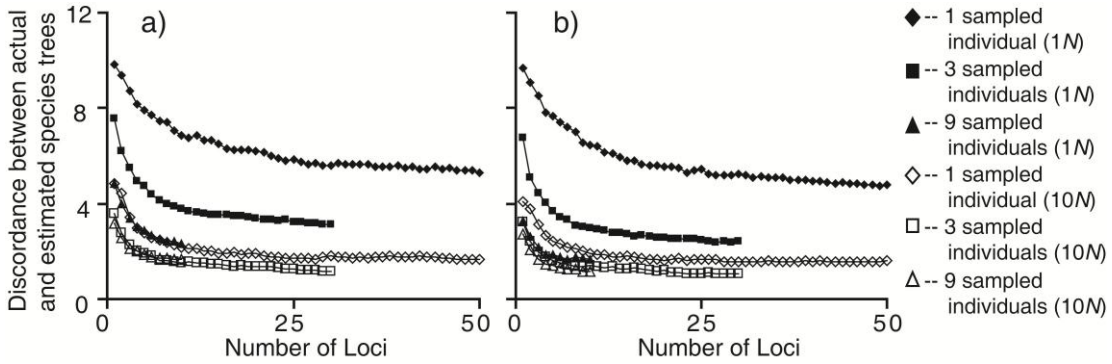


Figure 3.4 Accuracy with incremental loci addition – MDC approach.
The effect of incrementally adding loci on the accuracy of species tree estimated with the MDC approach when (a) both coalescent and mutational variances, DCM, versus (b) only coalescent variance, DC, are considered. Closed and open symbols represent results from the recent versus deeper divergence histories, respectively.

Mutational and Coalescent Effects on the Accuracy of ML Species-Tree Estimates

The errors for species-tree estimates due to coalescent variance can be reduced to exceedingly low levels with sufficient sampling using maximum likelihood (ML) species-tree estimates from STEM (Figure 3.2). Yet, with increasing sampling efforts the discord between the estimated and actual species tree persists because of the effects of mutational variance. As much as 75% of the errors in ML species-tree estimates are

attributable to mutational variance (Figure 3.3b). Moreover, the differing contribution of mutational variance for recent versus deeper divergences also explains why the accuracy of ML species-tree estimates depends on the species-tree depth (Figure 3.2). The reconstructability of shallow divergence histories does not actually differ from that of older divergent histories when the effect of mutational variance is excluded. In fact, with 9 individuals sampled per species, the shallow species trees (1N total tree depth) are more accurately estimated than deeper species trees (10N total tree depth) (Figure 3.5), but the disproportionate effect of mutational variance on recent divergence times leads to the opposite pattern.

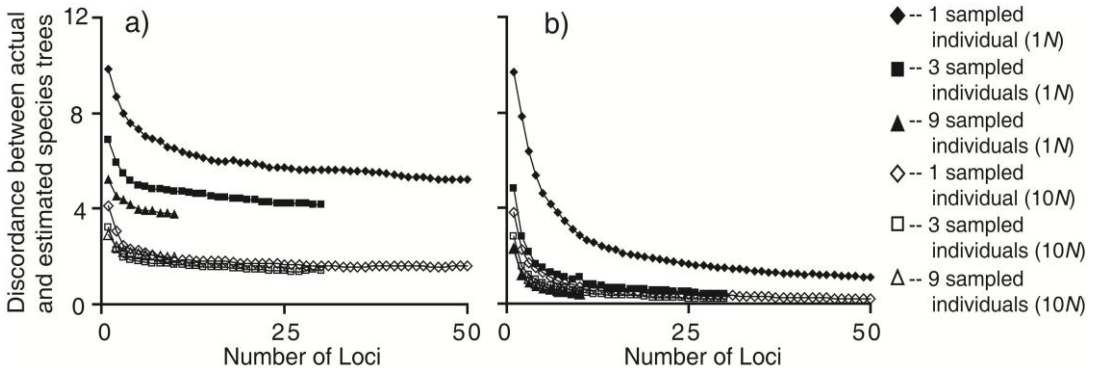


Figure 3.5 Accuracy with incremental loci addition – STEM approach.
The effect of incrementally adding loci on the accuracy of species tree estimation of STEM when (a) both coalescent and mutational variances, D_{CM}, versus (b) only coalescent variance, D_C, are considered. Closed and open symbols represent results from the recent versus deeper divergence histories, respectively.

Analysis of the incremental gains in accuracy associated with adding loci shows the errors with ML species-tree estimates decreases as the number of individuals increases (Figure 3.5). Although adding loci reduced the discordance contributed by coalescent variance, this gain is quickly offset by elevated contributions of mutational variance with the addition of more loci used for the ML species-tree estimate. At about 10 loci, the accuracy scores reach a plateau (i.e., there are negligible increases in accuracy with increased sampling; Figure 3.5). As more individuals are sampled per locus, there is a greater contribution of mutational variance per locus and the plateau in accuracy is reached with fewer sampled loci. Nonetheless, it is worth noting that when this plateau is reached, the species-tree estimates are more accurate when multiple individuals (as opposed to 1 individual) are sampled in the multi-locus data sets (Figure 3.5).

Comparison of the Accuracy of Species Trees Estimated with MDC versus STEM

The accuracy of species tree estimated with MDC and STEM is similar (Figure 3.2). However, the comparable discordant scores (i.e., levels of inaccuracy) between the two methods, irrespective of sampling effort and design, arise from very different causes. Errors associated with the MDC approach reflect this method's ineffectiveness at accommodating the effects of coalescent variance (i.e., the discord generated between a species tree and gene trees from the stochasticity of the coalescent process), although it has the advantage of being fairly robust to mutational variance. In contrast, increasing sampling lowers the errors in species-tree estimates due to coalescent variance in STEM, but mutational variance contributes to inaccurate species-tree estimates. In other words, the MDC method is sensitive to coalescent variance and robust to mutational variance, whereas STEM has the opposite pattern. With limited sampling of individuals and loci, the errors associated with both methods are primarily due to coalescent variance. In particular, with the sampling of one individual and one locus, both methods will estimate the same species-tree topology – namely, the gene-tree topology – as it is the maximum-likelihood tree and the topology with minimal deep coalescent events.

Comparison of the species-tree specific errors in estimation (as opposed to averaging across all species-trees; Figure 3.6) from the MDC and STEM methods confirms two key issues. First, the high correlation between the accuracy of species-trees estimates from the two methods suggests that the actual species tree itself has a large effect on the absolute amount of error of the species-tree estimate, irrespective of the method used (Figure 3.6). Second, the sampling strategy is a main determinant of the relative contribution of mutational and coalescent variance to errors in species-tree estimates, as it determines the covariance of discordance scores from the MDC and STEM methods. With regard to the total sampling effort required in empirical studies to reach the level of accuracy reported here, the number of sampled individuals and loci may be larger to achieve accurate estimates of population size used in species-tree estimation procedures (i.e., θ was known in the simulated data sets), and if larger numbers of taxa are considered (i.e., more than the 8 species studied here).

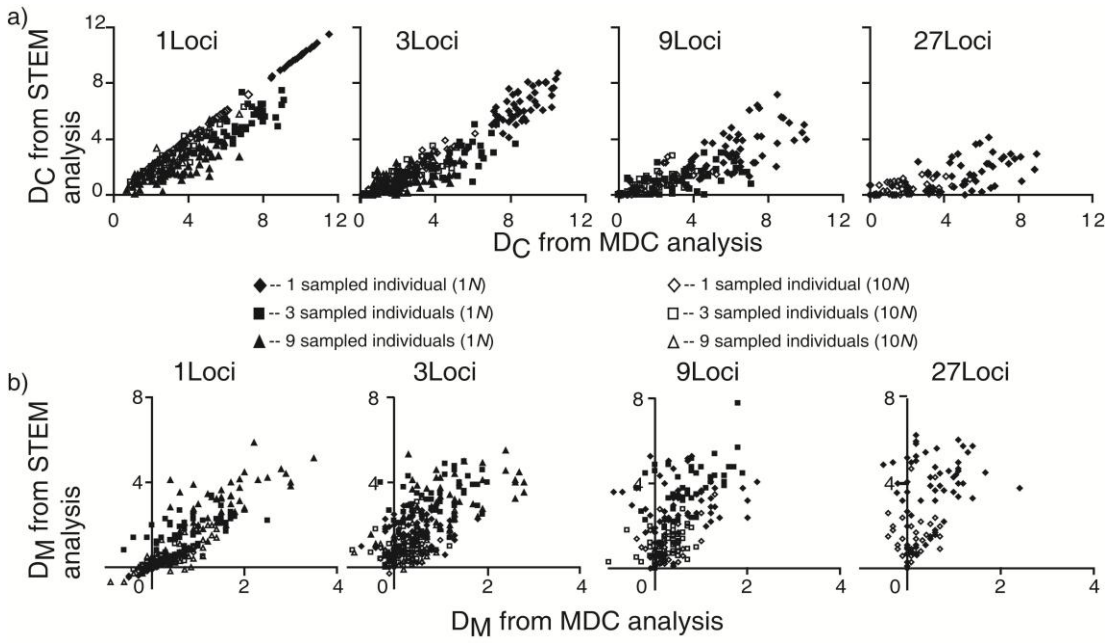


Figure 3.6 MDC and STEM method species tree error correlation.
Correlation between the errors in species trees estimated with the MDC and STEM methods due to (a) coalescent variances, DC, versus (b) mutational variance, DM. Closed and open symbols represent results from the recent versus deeper divergence histories, respectively.

DISCUSSION

Understanding the contribution of mutational and coalescent variance to errors in species-tree estimation is fundamental to increasing the accuracy of phylogenetic inference. The determination that coalescent and mutational variance have disproportionate impacts on the accuracy of species-tree estimates depending on the method of inference has two important implications. First, it highlights how method choice and sampling strategy can significantly impact the results from empirical studies (see also Maddison and Knowles 2006; Knowles 2009b; Liu et al. 2009; McCormack et al. 2009). Second, the finding emphasizes that data quality, not simply quantity, may be an important determinant of the accuracy of species-tree inference. These issues are likely to be a common challenge with all methods of species-tree estimation; we reiterate that the methods used here were chosen because the impact of the discord generated by mutational versus coalescent processes on the accuracy of species-tree estimates could be quantified (i.e., discord arising from differences between the species tree and coalescent gene trees versus differences between the coalescent gene trees and estimated gene trees; Figure 3.1), not because they necessarily would represent an ideal method for analysis.

This partitioning of the variance is essential to understanding how the species-divergence histories interact with aspects of data sampling, and thereby affect the accuracy of species-tree estimates. We discuss how data properties (i.e., levels of genetic variation, number of individuals and loci sampled, as well as taxon sampling) should be considered when choosing among methods for empirical investigations, and possible future developments of species-tree estimation methods relevant to improving phylogenetic inference given these empirical considerations.

Variation in the Sources of Error when Estimating Species Tree

Dependence on the history of species divergence. — The accuracy of species-tree estimates is known to be dependent upon the specific details of the divergence history (see also Maddison and Knowles 2006; Eckert and Carstens 2008; Knowles and Chan 2008; McCormack et al. 2009). This study highlights that decrease in the accuracy of species-tree estimates for very recent, as opposed to older divergence times (i.e., total species-tree depths of 1N and 10N, respectively) arises from both a decrease in mutational and coalescent variance (Figure 3.2). In fact, the improved concordance between estimated gene trees and their underlying genealogies as the time of divergence increases consistently accounts disproportionately for the gains in accuracy at the older species divergence times (i.e., irrespective of sampling design, or method of analysis), with just a few exceptions (see details below). The fact that the accuracy scores of species tree estimated from the two methods are highly correlated (Figure 3.6) further confirmed the importance of the specific history of divergence, albeit the correlation between other species-tree estimation methods remains to be studied.

It is worth noting that the species-tree depths reported here are measured in units of *N* generations, which is the product of the effective population size and the generation time. Hence, for organisms with larger population size (e.g., cosmopolitan *Drosophila* species), or with long generation time (e.g., trees), the time of divergence as measured in years will be correspondingly much longer than species with small populations and short generations.

Dependence on the sampling strategy. — Previous investigations into the trade-off of sampling more individuals versus genes (e.g., McCormack et al. 2009) revealed several

aspects of sampling strategy that impact the accuracy of species trees. Among these are increased accuracy with increases in total sampling effort and shifting towards sampling more individuals for recent divergences, as we document here. However, by partitioning the sources of errors associated with mutational and coalescent processes, we are able to interpret some enigmatic effects of sampling strategy noted, but not explained, in previous work.

For recent species divergence, adding more loci does not achieve the same high level of accuracy of species-tree estimates when individuals are added for a given total sampling effort (Figure 3.3 and Figure 3.5; see also Maddison and Knowles 2006; McCormack et al. 2009). Our study reveals that this observation reflects the lower information content contained in the pattern of coalescence among loci about species relationships with recent divergence – that is, the impact of the coalescent variance on species-tree accuracy continues to be quite high when sampling more loci (Figure 3.4b and Figure 3.5b). In contrast, there are significant declines in errors attributable to coalescent variance when species trees are estimated with multiple individuals sampled per species and locus (Figure 3.4b and Figure 3.5b). This result confirms the proposal of Maddison and Knowles (2006) proposal that the pattern of deep coalescence itself contains significant phylogenetic signal when there is widespread incomplete lineage sorting. The average accuracy of species-tree estimates nevertheless does plateau with increased sampling of individuals (Figure 3.4a and Figure 3.5a). Examining the relative contribution of mutational and coalescent processes shows that there is a notable increase in the proportional effect of the mutational variance on species-tree accuracy when more than 3 individuals are sampled. For older divergence (i.e., tree depth of $10N$) our results confirm predictions derived from coalescent theory (Takahata 1989; Hudson 1990) that with increased intraspecific sampling these individuals tend to have rather shallow coalescence times (i.e., short branches) rather than adding significant genealogical depth (Figure 3.4b and Figure 3.5b). Moreover, with very short times to coalescence, and hence shorter branch lengths in the underling genealogy, it becomes less likely that there will be sufficient mutations for reconstructing the gene tree (i.e., mutations are proportional to branch lengths). Consequently, the potential information contained in the pattern of gene-lineage coalescence with additional sampling is never realized.

Dependence on the method of analysis. — Despite the similar levels of accuracy achieved with estimating species trees with the two different methods, irrespective of sampling strategy and divergence history, the cause of the errors associated with the MDC (Maddison and Knowles 2006) and STEM (Kubatko et al. 2009) approaches differ (Figure 3.2). Partitioning the errors into those associated with mutational and coalescent processes reveals how the accuracy is inextricably linked to the procedural details, and specifically, the way in which information about the coalescence of gene lineages is incorporated in a method.

Species trees estimated using the MDC approach are relatively insensitive to mutational processes because the method relies on gene-tree topology, not branch lengths, to the extent that estimated gene trees differ in branch lengths, but not topology, from their underlying coalescent genealogies. However, the MDC suffers from a loss of information by relying on a summary statistic (i.e., minimizing the number of deep coalescences), as opposed a full probabilistic model of gene-lineage coalescence, to estimate species trees. This tradeoff is apparent in the partitioning of errors associated with the mutational and coalescent processes (Figure 3.2), in which the coalescent contributes disproportionately to errors in species trees estimated by the MDC approach. Consequently, estimates from MDC may be compromised if too few loci and individuals are sampled, although it is less sensitive to loci with limited genetic variance (i.e., low contribution of mutational variance).

The incorporation of a stochastic model of gene-lineage coalescence in STEM (Kubatko et al. 2009) means that the method is very efficient in extracting phylogenetic signal from coalescent gene trees, despite widespread incomplete lineage sorting (Figure 3.2). However, these potential gains in accuracy are offset by errors attributable to the discord mutation induces between estimated gene trees and coalescent gene trees (Figure 3.3b). Hence, STEM can achieve accurate estimates with less sampling efforts than a summary statistic based approach that does not fully utilize the information content in the gene trees, but requires high quality data (i.e., sufficient genetic variation for accurate estimation of gene trees, including their topology and branch lengths).

The partitioning of errors associated with coalescent and mutational processes has yet to be explored for any other method of species tree inference. Neither the MDC nor

STEM approaches implemented here modeled errors in the estimation of gene trees. Such consideration might result in significant gains in the accuracy of species-tree estimates (see discussion below). However, the basic findings of this study highlight three factors (i.e., sampling strategy, details of the history of divergence, and data properties) that make it difficult to generalize about the performance of methods or predict how robust other methods might (or might not) be to mutational variance, the implications of which are discussed below.

Implications of the Partitioned Effects of Mutational and Coalescent Processes

Methods for species-tree estimation. — The context-dependent effect of the mutational process on the accuracy of species-tree estimates apparent in this work suggests that generalizations about the likely impact of mutational variance will be difficult. For example, the benefits gained by computational approaches that invest significant effort (i.e., computational time) into incorporating a model of mutational process into the species-tree estimation procedure will vary. That is, our analyses show that the mutational variance may contribute significantly or very little to errors with a species-tree estimate depending on the sampling strategy, the species-tree estimation method, the timing of divergence (Figure 3.2), and the specific details of the underlying species tree itself (Figure 3.6). It is not clear at this point how much the effect of mutational variance on the accuracy of species-tree estimates will be reduced by incorporating the errors associated with gene-tree estimation into the species-tree estimation procedure when there is limited genetic variation (e.g., see Cranston et al. 2009; Kubatko and Gibbs 2010; Linnen 2010). This important point awaits investigation, but is hampered by computational constraints that severely limit such investigation to a few specific species trees. Consequently, the utility of methods should not be evaluated simply on whether they employ complicated algorithms that explicitly model both mutational and coalescent processes when accurate information about branch lengths is not forthcoming (see also Liu et al., 2009, for an example when mutations do not accumulate in a clock-like manner).

The partitioning of the sources of error in species-tree estimates also provides basic information about factors affecting the accuracy of species-tree estimates that is only possible because of the simplicity of the approaches. This baseline is important for identifying areas that need further exploration and development, as well as revealing the specific sensitivities of the MDC and STEM approaches. With regard to the MDC method, even when coalescent genealogies are analyzed for fairly considerable sampling efforts (e.g., sampling 30 loci and 3 individuals per species), there are still species trees that are not estimated accurately (Figure 3.4b). This plateau in the accuracy could simply reflect very small gains in accuracy with increased sampling (i.e., perhaps with infinite sampling the species-trees would be estimated accurately, again excluding the effects of mutation). However, it may reflect that the heuristic searching algorithms might not always find the tree with the fewest deep coalescent event (Than and Nakhleh 2009), or that this summary statistic is inconsistent for some histories. For example, the MDC approach may be sensitive to the anomaly zone based on analysis of the coalescent gene trees (Degnan and Rosenberg 2006), even though this danger may not be realized in empirical data once mutational processes are considered (see Huang and Knowles 2009). The high sensitivity of the STEM method to mutational variance in contrast suggests that developing a way to consider uncertainty in gene-tree estimates (e.g., using a consensus tree as input), while avoiding the computational burden of modeling the mutational process during the species-tree estimation procedure, could improve the performance and make the method especially useful for large data sets, given that large data sets exceed the computational capacity of sophisticated algorithms (e.g., Cranston et al. 2009).

Empirical investigations. — One critical implication of the results, which has not received any attention in the context of estimating species trees, is how the quality of the data collected per locus (i.e., the amount of genetic variation that influences the accuracy of gene-tree estimates), not simply the number of loci or individuals sampled, may impact the accuracy of inferred species relationships. Our results highlight that this is especially important as sampling effort increases. For example, the disproportionate increase in mutational variance as more loci are sampled significantly offsets the gains in accuracy achieved by increased sampling effort (Figure 3.2). It remains to be determined the extent to which data quality also impacts the potential gains of adding loci and

individuals with more sophisticated methods that incorporate error in the estimated gene trees (e.g., the program BEST; Liu 2009). In addition to the problems with achieving convergence, which often thwart analysis with Markov Chain Monte Carlo-based algorithms (e.g., with programs such as BEST; see Cranston et al. 2009; Kubatko and Gibbs 2010; Linnen 2010), a non-tree based analysis may be more appropriate when there is limited genetic variation in loci (e.g., Rannala and Yang 2003; Hobolth et al. 2007).

Marker development (e.g., Carstens and Knowles 2006; Hahn et al. 2009), especially as it relates to identifying variable loci, needs to be treated carefully (Knowles 2010). Although the typical view is that gains in phylogenetic information will be made simply through the collection of more data, this generalization appears to be more nuanced in the context of estimating species trees (Figure 3.4 and Figure 3.5). Moreover, removing invariant individuals or loci from the analyses in an attempt to obtain a species-tree estimate more efficiently or to avoid problems with limited variation is not advisable (Knowles 2010). Because these methods rely on expectations from coalescent theory to define the relationship between sampled gene trees and a species tree (reviewed in Degnan and Rosenberg 2009), only using data with a minimal amount of variation introduces an ascertainment bias (Wakeley et al. 2001) that may affect the reliability of the analysis. However, this too is an issue that has not yet been explored in the context of species-tree estimation.

Because the sensitivity of methods to various aspects of the specific history of divergence and properties of the genetic data collected differs (see also Maddison and Knowles 2006; Cranston et al. 2009; Eckert and Carstens 2008; Knowles 2009b; Liu et al. 2009), the appropriateness of a method will differ among empirical investigations. Such decisions require a thorough consideration of the strengths and limitations of the study (e.g., number of loci and their levels of variation) and of the way in which a method extracts information (e.g., relies only on topology, uses branch length information, or considers uncertainty in the estimated gene trees) and characterizes the coalescent process (e.g., uses a full probabilistic model, versus a summary statistic based on either average or minimal coalescence times). When a species-tree estimation method is used, the reliability of its result should therefore be considered based on the compatibility of

the method to the data. Different species-tree estimation methods might give conflicting inference about the species divergence history because of violation of the methods assumptions (e.g., Eckert and Carstens 2008; Cranston et al. 2009; Liu et al. 2009), or as we have shown here, differences in how the method extracts information from the data (Figure 3.2).

CONCLUSIONS

The direct estimation of species trees is a nascent prosperous field in phylogenetics (Knowles and Kubatko 2010). With fairly modest numbers of loci, applications of these methods demonstrate their promise for resolving species relationships despite widespread incomplete lineage sorting (e.g., Carstens and Knowles 2007; Edwards et al. 2007; Brumfield et al. 2008; Liu et al. 2008). Nevertheless, there is still a shortage of analyses that examine the performance of these methods across variety of species divergent histories, sampling configurations, and data quality (Maddison and Knowles 2006; Eckert and Carstens 2008; Liu et al. 2009; McCormack et al. 2009). The findings from this study show that without such information, decisions made in empirical studies might constrain the potential insights gained from species-tree estimates. By examining the sources of error contributed by the basic genetic processes underlying patterns of genetic variation – mutational and coalescent processes – we reveal how estimates of species-trees can be improved by considering the complex interactions between the data set properties, the history of divergence, and the method of analysis. The detailed analyses also show why the estimation of species trees is subject to these complex interactions. Such information highlights the importance of data quality (not simply quantity) and the selection of methods according to data-specific features, issues that are particularly relevant given present-day advances in sequencing technologies. The significant contribution of mutational variance to the errors in species-tree estimates described in the paper emphasize that the impact of mutation on the accuracy of species-tree estimates is an area that needs immediate attention. For example, with the increasing amount of multi-locus data for non-model organisms being generated, the impact that different mutation rates and lengths of sequence used to estimate gene trees, or recombination within the loci, will have on species-tree estimates in not known. In addition to the issues arising from

mutational variance when errors in the estimated gene trees are not considered (as with both MDC and STEM), we note that data set properties, and in particular issues associated with mutation can also complicate methods that actually consider uncertainty in gene-tree estimates (as with BEST; see Cranston et al. 2009; Liu et al. 2009; Linnen 2010; Kubatko and Gibbs 2010). Together, the studies emphasize how important it is to consider the properties of empirical data when estimating species trees and that these practicalities need to be considered when the performance of methods for such inferences are evaluated.

REFERENCES

Ane, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412-426.

Brumfield, R. T., L. Liu, D. E. Lum, and S. V. Edwards. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: *Pipridae, Manacus*) from multi-locus sequence data. Syst. Biol. 57:719-731.

Carstens, B. C., and L. L. Knowles. 2006. Variable nuclear markers for *Melanoplus oregonensis* identified from the screening of a genomic library. Mol. Ecol. Notes 6:683-685.

Carstens, B. C., and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanoplus* grasshoppers. Syst. Biol. 56:400-411.

Cranston, C., B. Hurwitz, D. Ware, L. Stein, and R. A. Wing. 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58:489-500.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:762-768.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332-340.

Eckert, A. J., and B. C. Carstens. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol. 49:832-842.

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

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U. S. A. 104:5936-5941.

Felsenstein, J. 1993. Phylip (phylogeny inference package) version 3.5c.Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.URL: http://evolution.genetics.washington.edu/phylip.html

Gaut, B. S., and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

Hahn, D. A., G. J. Ragland, D. D. Shoemaker, and D. L. Denlinger. 2009. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. BMC Genomics 10:9

Hobolth, A., O. F. Christensen, T. Mailund, and M. H. Schierup. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. PLoS Genet. 3:e7.

Huang, H., and L. L. Knowles. 2009. What is the danger of the anomaly zone for empirical phylogenetics?. Syst. Biol. 58:527-536.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology Vol 7: 1-44.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

Huelsenbeck, J. P., D. M. Hillis, and R. Nielsen. 1996. A likelihood-ratio test of monophyly. Syst. Biol. 45:546-558.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. of Mol. Evol. 16:111-120.

Knowles, L. L. 2010. Sampling Strategies for Species-Tree Estimation. In: Estimating Species Trees: Practical and Theoretical Aspects (L. L. Knowles and L. S. Kubatko, eds.). Wiley-Blackwell.

Knowles, L. L. 2009a. Statistical phylogeography. Annul. Rev. Ecol. Syst. 40:593-612.

Knowles, L. L. 2009b. Species tree estimation: Methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58:463-467.

Knowles, L. L., and Y.-H. Chan. 2008. Resolving species phylogenies of recent evolutionary radiations. Ann. Missouri Bot. Gard. 95:224-231.

Knowles, L. L., and L. S. Kubatko. 2010. Estimating Species Trees: An Introduction to Concepts and Models. In: Estimating Species Trees: Practical and Theoretical Aspects (L. L. Knowles and L. S. Kubatko, eds.). Wiley-Blackwell.

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

Kubatko, L. S., and H. L. Gibbs. 2010. Estimating Species Relationships and Taxon Distinctiveness in *Sistrurus* Rattlesnakes Using Multi-locus Data. In: Estimating Species Trees: Practical and Theoretical Aspects (L. L. Knowles and L. S. Kubatko, eds.). Wiley-Blackwell.

Linnen, C. 2010. Species-Tree Estimation for Complex Divergence Histories: A Case Study in *Neodiprion* Sawflies. In: Estimating Species Trees: Practical and Theoretical Aspects (L. L. Knowles and L. S. Kubatko, eds.). Wiley-Blackwell.

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

Liu L., Yu L. Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–477.

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

Liu, L., L. Yu, and D. K. Pearl. 2010. Maximum Tree: A consistent estimator of the species tree. J. Math. Biol. 60:95-106.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523-536.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21-30.

Maddison, W. P., and D. R. Maddison. 2009. MESQUITE: A modular system for evolutionary analysis., version 2.6. Version 2.71 http://mesquiteproject.org

McCormack, J. P., H. Huang, and L. L. Knowles. 2009. Maximum-likelihood estimates of species trees: How accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58:501-508

Mossel, E., and S. Roch. 2007. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. IEEE/ACM Trans Comput. Biol. Bioinform. 7: 166–171.

Mossel, E., and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309:2207-2209.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568-83.

Rambaut, A., and N. C. Grassly. 1997. SEQ-GEN: An application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235-238.

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

Ripplinger, J., and J. Sullivan. 2008. Does choice in model selection affect maximum likelihood analysis? Syst Biol 57:76-85.

Robins, J. H., P. A. McLenachan, M. J. Phillips, L. Craig, H. A. Ross, and E. Matisoo-Smith. 2008. Dating of divergences within the *Rattus* genus phylogeny using whole mitochondrial genomes. Mol. Phylogenet. Evol. 49:460-466.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131-147.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Singh, N. D., P. F. Arndt, A. G. Clark, and C. F. Aquadro. 2009. Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. Mol. Biol. Evol. 26:1591-1605.

Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. 4:77–86.

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inferencewhenweknowthat assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50:723–729.

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

Than, C., and L. Nakhleh. 2009. Species tree inference by minimizing deep coalescences. PLoS Comput. Biol., 5:5.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero, and K. Ardlie. 2001. The discovery of single-nucleotide polymorphisms and inferences about human demographic history. Am. J. Hum. Gen. 69:1332-1347.

Chapter 4 Do Estimated and Actual Species Phylogenies Match?: Evaluation of African Cichlid Radiations using a Parametric Bootstrap Species Tree (PBST) Approach

Having accurate and reliable estimates of species diversifying histories (i.e., species trees) is fundamental to studying evolutionary biology. Aiming to enhance the reliability of estimated trees, most of the efforts in previous molecular phylogeny studies were spent on searching for better genetic markers (e.g., longer sequences, more appropriate mutation rates) and applying improved gene-tree estimation methods (e.g., more realistic DNA substitution models). However, the implicit assumption of these effects—that an accurately estimated gene tree (e.g., trees with high nodal supports) is equivalent to the actual species tree—is untrue, because gene trees can differ from the actual species trees because of the random lineage sorting process, even for trees estimated without error and complication from gene flow, gene duplication and gene horizontal transfer (Maddison 1997). Hence, the reliabilities of many previous phylogenetic estimates based on single-locus (or a concatenated dataset) are in question: gene trees estimated with an alternative locus might be different (Figure 4.1a). For those studies that also only sampled a single individual (or a limited number of individuals) per species, the estimates of tree branch lengthes (i.e., divergence time) or even the tree topology might be different with an alternative individual in the species (Figure 4.1b). It would be rash to deny the merit of all these previous single-locus studies because of the flawed assumptions; nevertheless, it is necessary to assess the robustness of these previous "species-tree" estimates in terms of random lineage sorting before building further studies upon them.

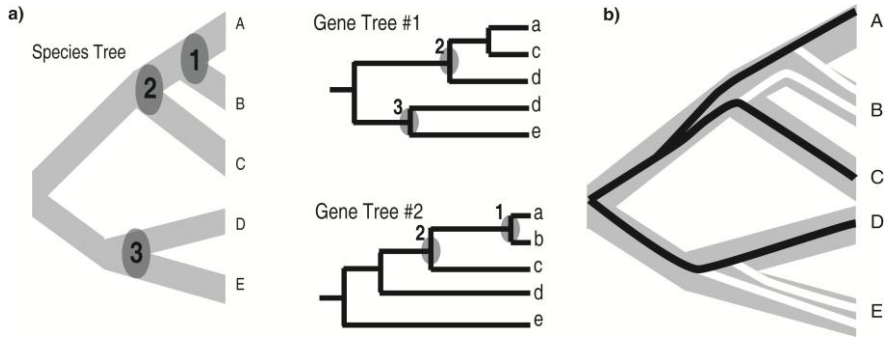


Figure 4.1 Consequences of incomplete lineage sorting.
a) The incongruence between gene trees and their species tree. The hypothetical 5-taxon species tree can be seen as a collection of three internodes, 1, 2 and 3 (with its depth indicated). Each of them corresponds to a monophyletic clade—internodes 1, 2 and 3 correspond to the clade of species A and B, the clade of A, B and C, and the clade of D and E, respectively. These internodes might not be found on the gene trees—both gene tree #1 and #2 recovered only two of the three internodes. The lower case letters a, b, c, d and e represent the sampled lineages from species A, B, C, D and E, respectively. b) Possible problems with non-monophyletic species. The solid black or white lines indicate sampled gene lineages. For species E, sampling different individuals (i.e., different white lines) gives a different estimation of the divergence time of species D and E. For species B, sampling different individuals changes the inference about which species (A or C) is its sister species.

Assessing the robustness of a single-locus species-tree estimate is different from looking at the values of nodal support on the gene tree, rather, it is asking: what is the chance of finding the same tree if a different locus is sequenced or a different individual is sampled? The exact chance would depend on specific parameters of the actual species tree, which determine the variation among gene trees generated by the random lineage sorting process (Pamilo and Nei, 1988; Takahata, 1989). To circumvent the problem of unknown species trees in empirical studies, we propose a parametric bootstrapping approach, called PBST. An assumption is first made that previously published single-locus gene trees represent the "true" species trees, which is followed by simulation of the lineage sorting process under these "true" species trees. The robustness of the published single-gene trees would be inversely correlated with the amount of difference between the published and the simulated gene trees. The difference was quantified by two percentages. First, the percentage of innernodes on the published gene tree that can be recovered by simulated gene trees is used to describe the chance of having an incongruent gene tree from another locus (Figure 4.1a). Second, since sampling a different individual can lead to a different tree estimate (branch-lengths or topology) only when species are non-monophyletic, the percentage of non-monophyletic species on the simulated gene trees is used to proximate the chance of

recovering the published gene trees with a different individual (Figure 4.1b). A single-locus phylogenetic inference is more likely to be robust with lower values of these two percentages.

As an example of this PBST approach, it was applied as a meta-analysis method to a classic example of vertebrate radiation—East African cichlid fish (Perciformes: Cichlidae), a challenging system for reconstructing the species tree because of its exceptionally fast speciation rate (e.g., many incomplete lineage sorting observed in (Genner et al., 2007a). Despite having long been an active research topic (Fryre and Iles, 1972; Greenwood, 1974) and the fact that as many as 46 gene-tree based phylogenetic studies can be found in the literature, this cichlid group is not exceptional in terms of having the majority of its phylogenetic studies to be based on a single locus. Over 75% of these studies only sequenced a single locus and sampled less than 3 individuals per species (Figure 4.2). In addition to calculating the two percentages to assess the robustness of these previously published trees, we explored the possibility of minimizing the difference between the estimated species trees and the "true" species trees by increasing sampling effort and applying methods to incorporate lineage sorting process into species-tree estimation.

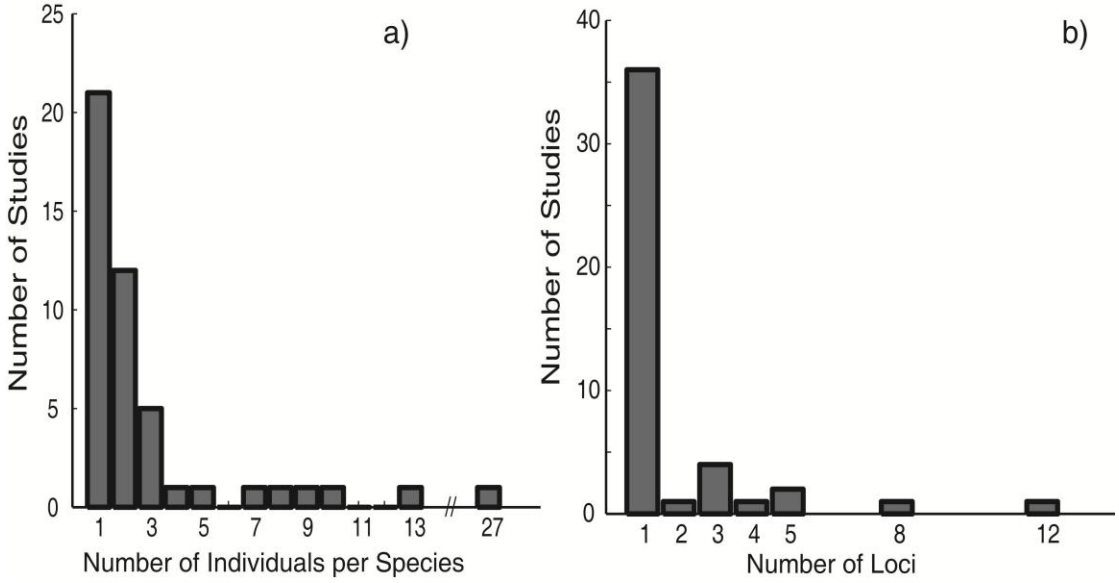


Figure 4.2 Summary statistics of published gene trees on East African cichlids.
a) Histogram of the average number of sampled individuals per species across the 46 published studies; b) Histogram of the number of loci used in the studies.

METHODS

Identifying cichlid phylogenetic studies

Published cichlid phylogenetic studies were identified by a search of the ISI Web of Science using the topic words "cichlids" and "phylogeny", as well as from citations of these papers (see Table 4A.1 for the list of studies). We excluded studies that did not include molecular phylogenetic analyses and review studies that re-analyzed previously published datasets were excluded so as not to introduce a bias in our analyses by the repeated representation of certain phylogenetic groups. The focus here is on East African species (i.e., species within clade A1 from Schwarzer et al., 2009) because the goal was to investigate a system that should theoretically represent one of the most difficult historical scenarios for phylogenetic estimation given the rapid rates of speciation. Other cichlids were not included in the analyses; in the few cases in which the monophyletic hypothesis of the East African cichlid radiation was not supported, the portion of gene trees containing the largest inclusive group of East African cichlids (i.e., taxa forming a monophyletic clade containing species exclusively from clade A1 from Schwarzer et al., 2009) were used.

Phylogenies used to evaluate the match between estimated and actual phylogenies

For the 46 studies included for analysis, new estimates of the gene trees were made from sequences downloaded from GenBank or that were provided by the authors. This was necessary to provide a standardization to assure that any difference observed between the estimated species tree and the published gene tree across the studies does not reflect differences in the procedures used to estimate the gene trees. Moreover, this also provided branch length information that was then used to investigate the association between the relative timing of species divergence times and whether phylogenetic relationships are likely to be estimated accurately. Gene trees were also estimated using a single individual per species for studies that included more than one representative per species; this is necessary because gene trees represent the relationships among haplotypes as opposed to species per se (i.e., each taxon in the species tree is the operational taxonomic unit, OTU, see Appendix Figure 4.A1). The first gene lineage to coalesce with a gene lineage from a sister taxon was

retained because species divergence always post-dates the latest gene divergence time, assuming no gene flow after speciation.

Sequences were aligned in ClustalX2.0 (Larkin et al., 2007) and gene trees were estimated using Genetic Algorithm for Rapid Likelihood Inference (GARLI) version 1.0 (Zwickl, 2006) with the same substitution model either reported in the original studies (Table SI1) or the best fit model estimated using DT-Modsel (Minin et al., 2003) for studies that did not identify a substitution model. The maximum-likelihood (ML) gene tree was identified in each analysis after 10,000 generations if no significant likelihood improvement (<0.05) and no significant topological improvement (<0.01) were observed. Because the branches in a species tree are ultrametric, gene trees were re-estimated with a molecular clock and rooted with outgroups identified in the respective studies using PAUP* 4.0 Beta (Swofford, 2002). To avoid confounding interpretations about the cause of discord between estimated and actual phylogenies, species without sequence for a locus were excluded to assure that errors in estimating the species tree did not simply reflect missing data.

For multilocus studies in which there was discord among the estimated gene trees, and hence ambiguities in what constitutes the actual species tree, each gene tree was considered as a possible representation of the diversification history. Similarly, the actual species tree was ambiguous for studies in which more than one representative was sequenced per species and the species were not monophyletic. To account for this uncertainty, multiple alternative species trees were also considered. A set of alternative species trees were identified by considering all the different phylogenetic positions of each individual, except for studies involving more than 50 alternative diversification histories. Because of the difficulties with enumeration across such a large set of alternative species trees, for these cases, an individual from a species was randomly selected to represent a potential phylogenetic position; this was repeated to generate 50 alternative species trees. Results were averaged across the analyses of alternative species trees for any one case study to avoid introducing a bias in the meta-analysis. In other words, 925 species trees were evaluated, but the results on the match between estimated and actual phylogenetic histories are based on the 46 independent diversification histories collated from the different published studies (Appendix Table 4.A1).

Parametric bootstrap of species tree (PBST) approach

Based on the extracted species tree, gene trees were simulated with the program ms (Hudson, 2002). The ratio between gene-tree branch length to actual time length was either obtained from the paper or two recent papers (Genner et al., 2007b; Schwarzer et al., 2009), generation time was set to 3 years (Charlesworth, 1980), and the effective population size was set to 50,000. This population size is higher than some species with extremely small populations (e.g., the population size of *Tropheops gracilior* is estimated to be 1,500-4,900 (Won et al., 2005)); however, it is likely to be an underestimate for cichlids in general, as various surveys have shown that many species have population sizes on the order of 10^5 in all three major lakes: LT(Sefc et al., 2007), LV (Nagl et al., 1998; Samonte et al., 2007) and LM (Parker and Kornfield, 1997; Shaw et al., 2000), see Appendix Table 4.A2 for a complete list, and Appendix Figure 4.A2 for further analysis on the population size setting). A larger population size (100,000) and the lower confidence intervals of the divergence time estimates (referred to as setting L) were also used for simulation to explore the possible range of the effect of coalescent variance (see Appendix Figure 4.A3-A5 for result).

RESULTS

The chances of recovering the published gene tree with a different locus

Approximately 40% of studies (18 out of 46) have more than a quarter of innernodes on the "species trees" that cannot be recovered by simulated gene trees (Figure 4.3a). On the other hand, 12 studies have less than 5% unrecoverable innernodes, suggesting considerable amount of variation among studies (Figure 4.3a). The percentage of non-recoverable innernodes for each study is based upon one hundred gene trees simulated by sampling one individual per species, which also provides the chance of recovery for every innernode on the published gene trees. Meta-analysis of these innernodes' chances across studies allowed us to investigate which periods of the cichlids' divergent history—younger or older diversifications—contribute more to the incongruence. Although a significant negative correlation exists between the probabilities of incongruence and innernodes' depths, the low R-square value indicates that only a small part of the variances is explained by the depth (Figure 4.3b).

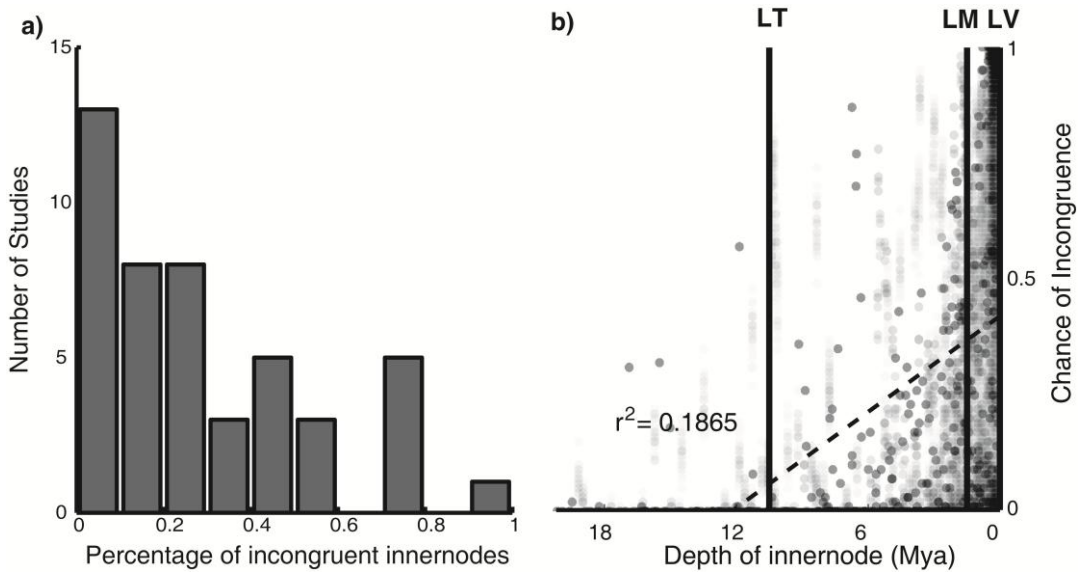


Figure 4.3 The chance of incongruence between "true" species trees and gene trees.
a) The histogram of the percentage of incongruent innernodes of all the studies.b) The correlation between the chance of incongruence and the depth of the innernodes on the time scale. The dash line indicates the relationship predicted by linear regression, with the R^2 value shown by the side. The three vertical lines indicate the median estimates of lake age. LT: Lake Tanganyika; LM: Lake Malawi; LV: Lake Victoria. The intensity of the grey dots indicates the weight, which is the inverse of the number of divergent histories corresponding to one study.

The chances of recovering the published gene tree with a different individual

Fourty-five percent of studies (21 out of 46) have more than a quarter of non-monophyletic species if 5 individuals are sampled per species (Figure 4.4a). For these published single-individual gene trees, the branch length leading to a non-monophyletic species or even the related topology might not be recovered if a different individual in the species is sampled. Variation among studies also exists—12 (13) studies have less than 5% non-monophyletic species when sampling 5(10) individuals per species. The percentage has an overall correlation with the percentage observed on published gene trees (Figure 4.4b), but is generally higher (Figure 4.4b). The simulated datasets also provide the chance of being non-monophyletic for every species. In contrast to the chances of incongruence, meta-analysis shows that the chance of being a non-monophyletic species quickly approaches zero with increasing divergence time between sister species pairs (Figure 4.4c).

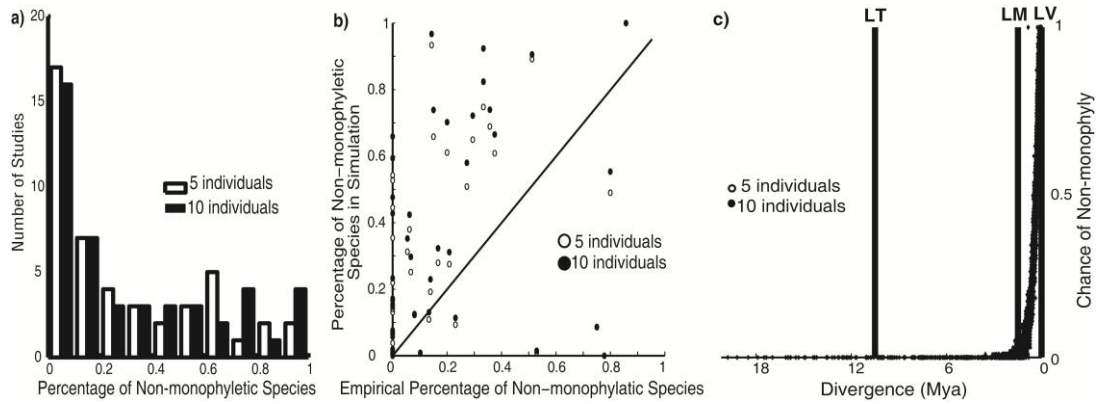


Figure 4.4 The chance of being non-monophyletic.
a) Histogram of the percentage of non-monophyletic species across the 46 studies. The open (filled) bars represent results with sampling 5 (10) individuals per species;b) The correlation between the percentage of non-monophyletic species in simulated datasets and the percentage observed in the published studies. The solid line is a 1:1 line; c) the correlation between a species' chance of being non-monophyletic and the divergence time between it and its sister species/clade. The three vertical lines indicate the median estimates of lake age. LT: Lake Tanganyika; LM: Lake Malawi; LV: Lake Victoria. In both b) and c), the open (filled) dots represents simulations with sampling 5 (10) individuals per species.

The chance of recovering the published gene tree with a species-tree estimation method

The low chance of recovering the published gene tree suggests that even though these trees might be accurate by themselves (i.e., high nodal supports); they might not represent the true species tree. As one step further, we applied one of the species-tree estimation methods, MDC (Minimizing Deep Coalescences, (Maddison and Knowles, 2006)), to incorporate the incongruence among the simulated gene trees into species-tree estimation process, and examine its chance in recovering the "true" trees with different multi-locus and multi-individual datasets (individuals-to-loci ratio = 1:1, 1:5, 1:9, 5:1, 5:5, 5:9, 10:1, 10:5, 10:9).

Adding more individuals per species considerably decreases the percentage of incongruent innernodes, and an even larger shift was caused by adding more loci (i.e., the distribution shifts toward the left, Figure 4.5a). Meta-analysis across studies also revealed two shifts—less old innernodes' contribution to the percentages of incongruence with more loci and less young innernodes' contribution when sampling more individuals (Figure 4.5b). With the maximum sampling effort tested in this study—9 loci and 10 individuals, the ideal situation of zero incongruence is not reached for a few innernodes at very shallow depths; yet, the percent of incongruent innernodes is much lower than that of one locus and one individual.

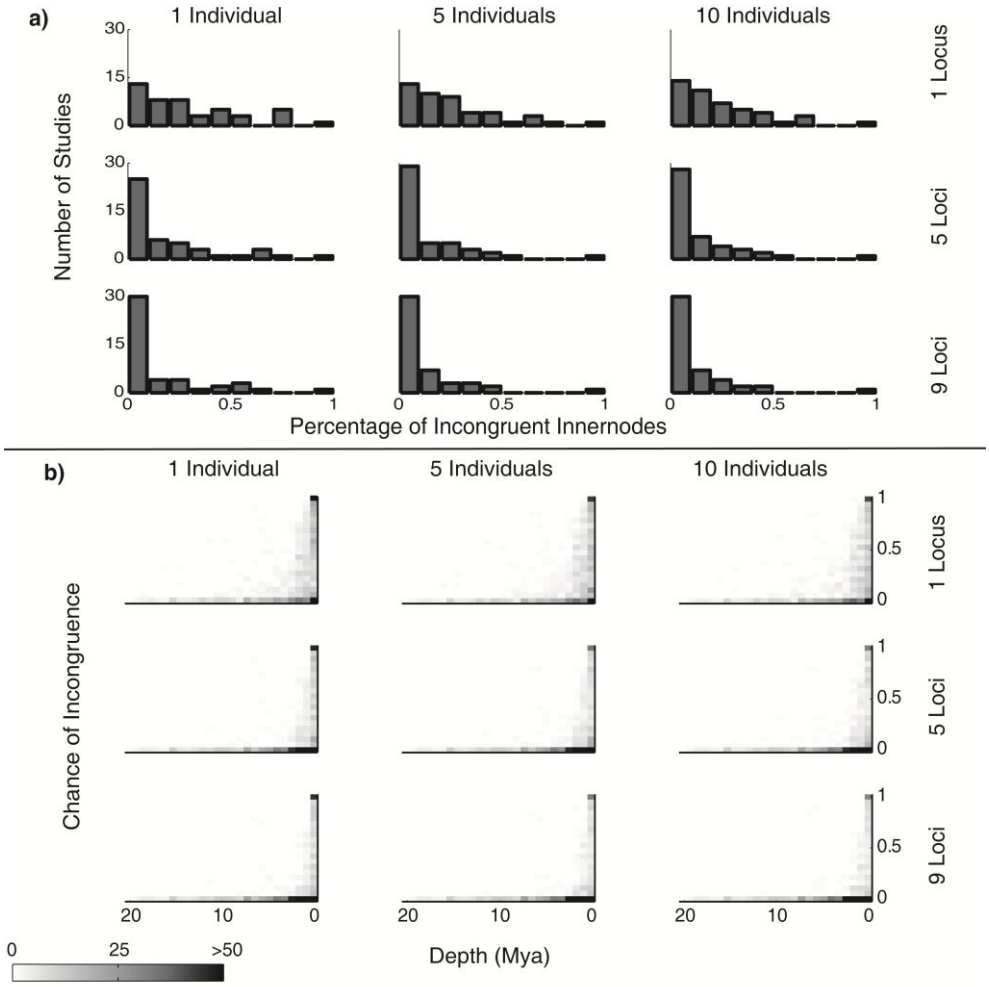


Figure 4.5 Effects of increased sampling effort and MDC use to estimate the species tree despite discordance.

a) Histograms of the percentage of incongruent innernodes (i.e., the innernodes that cannot be found on the estimated species tree by MDC) under nine different sampling designs. b) Grey-scale plots of the correlation between the chance of incongruence and the depths of the innernodes on the time scale with different sampling designs. A darker color means more points in the block, as indicated by the scale in the lower left corner

DISCUSSION

Searching for a reliable estimate of an actual species tree is not merely seeking an accurate gene tree. A well-supported gene tree does not necessarily represent a well-supported species tree unless the gene-tree species-tree difference generated by lineage sorting process is negligible. The PBST approach, as a way to assess the difference between single-locus gene trees and their unknown species trees, provides many insights for future cichlids' study, which will be discussed below, and it can be expected that applying this approach to other organismal groups will gain equivalently helpful insights.

Do previous phylogenetic estimates need to be re-evaluated?

The answer depends on the goal of the study and its species sampling. The robustness of species-tree estimates has a wide range among studies as represented by the amount of variation in the two percentages (Figure 4.3b and Figure 4.4b). At one end of the extreme, two studies (Farias et al., 2000; Sparks and Smith, 2004) have zero percent of non-monophyletic species as well as zero percent of incongruent innernodes. Aiming to resolve the family-level phylogeny of cichlid fishes, the two studies had sparsely sampled taxa in the clade, which represents major lineages that already emerged before the radiation. At the other end of the extreme, both of the percentages exceed one quarter in 16 studies. Aiming to resolve the divergent history within subsets of East African cichlids (e.g., Lake Malawi species,(Won et al., 2006)), these studies have denser sampling of species, with many divergences happening during the "radiation" phase. Although all studies are based on single locus, or artificial "single" locus concatenated from multiple loci, the reliability of species tree estimates varies. Some apparently need to be re-evaluated with more data; some could be seen as faithful representations of the species trees. As phylogenetic studies of the same species group usually aim to resolve relationships at different levels, the variation among studies can be expected in many other organismal systems. Identifying the robust estimates to avoid re-evaluating all the studies is a more efficient way to make progress.

Need more loci or more individuals?

For the studies with low chances of obtaining the same gene tree with a different locus or individual, an accurate estimate of cichlids' phylogenetic history would require datasets with multiple sequenced loci and sampled individuals. Futhermore, the reliabilities of different parts of the gene tree assessed by the PBST approach by checking individual species or innernodes could aid in the design of an efficient sampling strategy for future re-evaluation work. Furthermore, the two chances—the chance of recovery for every innernodes and the chance of being non-monophyletic for every species—describe different aspects of robustness, and reducing them requires different kinds of additional data. Species with high chances of being non-monophyletic would require more field work to collect speciemens, while laboratory work for developing more genetic makers can target those groups with

innernodes with high chance of incongruence when finding nuclear DNA makers that work across distantly related species is difficult.

For future cichlids' phylogenetic studies in general, the PBST results from meta-analysis provide some guidelines as well. Having multiple loci would be necessary for solving both old and young divergent events (i.e., innernodes). Even though the decline of the chance of incongruence is statistically significant, the large variance and the young ages of the three major lakes, in which the majority of the radiation took place, even the oldest divergent event in the oldest lake would still have the chance of not being recovered by single-locus gene tree (Figure 4.3b). On the other hand, having multiple individuals would be less important for relatively older species (i.e., those species generated in old radiations in Lake Tanganyika, Figure 4.4c), because having additional lineages in a monophyletic species would not add much information to the species-tree estimation (Maddison and Knowles, 2006). Nevertheless, one point to notice is that the percentage of non-monophyletic species is higher in the simulated gene trees than that in the published gene trees (Figure 4.3b), the reason might be that the simulated gene trees are of diploid loci, while most of the empirical studies use haploid mtDNA markers. This suggests that sampling multiple individuals would be necessary even if the species was found to be monophyletic in previous mtDNA analyses.

Can the phylogenetic history of the cichlid radiations be accurately estimated?

Our results show that MDC, a relatively simple parsimony species-tree estimation method, can resolve most of the incongruence in the simulated gene trees for cichlid groups with a reasonable amount of sampling (Figure 4.5a). Only sampling 5 loci and 5 individuals per species could solve most of the discordances between estimated and "true" species trees. Adding more individuals primarily helps resolving younger divergent events, while adding more loci can attack the problem of incongruence at a much deeper time depth (Figure 4.5b). The very few innernodes at extreme shallow depth that still have non-zero chance of incongruence with 9 loci and 10 individuals, more data is probably needed.

These results not only confirmed the conclusions from previous simulation studies on species-tree estimation methods (Huang et al., 2010; Kubatko et al., 2009; McCormack et al., 2009), but also complement the simulation approach as an empirically informed evaluation

of the species-tree estimation method. It has been shown in previous studies that divergent history itself plays an important role in determining a species-tree estimation method's performance (McCormack et al. 2009). The use of fixed species trees, or trees generated by models in the simulation approach put a limit on the generality of the result, since they not necessarily represent the cladogenetic process in nature, notwithstanding that no model has been found to be universally suitable for all organismal groups (Barraclough, 2010; Bininda-Emonds et al., 2002; Rabosky, 2009). Here, by obtaining the "species trees" from published gene trees, we provided the evaluation process with an approximation of the real speciation model in a specific clade, and the result—the assessment of the method's performance and the requirement on sampling effort—would be more accurate for and relevant to organismal groups of interest.

SUMMARY

Reconstructing species trees is one of the primary goals of evolutionary biology, and often a prerequisite for other studies. The difference between gene trees and the species trees, has gained an increasing amount of recognition in empirical studies as phylogenetics entering the phylogenomic era. While future studies will have huge benefits from the increased availability of data and the development of new species-tree estimation methodd, it is necessary to examine how much of our knowledge based upon single locus are robust despite this flawed assumption. Here, using the PBST approach, we revealed a considerable amount of risk in making phylogenetic inferences from single gene trees, while showcasing the variation among studies. The "species trees" extracted during the procedure also offer a biologically-realistic, system-specific approximation of the cladogenetic model to evaluate species-tree estimation methods and sampling designs. For its ability to assess the robustness of previous phylogenetic estimates and help to find an efficient sampling design for future studies, this PBST approach would be applicable and helpful in many other organismal groups.

REFERENCES

Ane, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24:1575-1575.

Banister, K. E., and M. A. Clarke. 1980. A Revision of the Large Barbus (Pisces, Cyprinidae) of Lake Malawi with a Reconstruction of the History of the Southern African Rift-Valley Lakes. Journal of Natural History 14:483-542.

Barraclough, T. G. 2010. Evolving entities: towards a unified framework for understanding diversity at the species and higher levels. Philosophical Transactions of the Royal Society B-Biological Sciences 365:1801-1813.

Bininda-Emonds, O. R. P., J. L. Gittleman, and M. A. Steel. 2002. The (Super)tree of life: Procedures, problems, and prospects. Annual Review of Ecology and Systematics 33:265-289.

Cameron, S. L., C. L. Lambkin, S. C. Barker, and M. F. Whiting. 2007. A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. Systematic Entomology 32:40-59.

Charlesworth, B. 1980. Evolution in Age-Structured Populations. Cambridge University Press, Cambridge, U.K.

Cohen, A. S., M. J. Soreghan, and C. A. Scholz. 1993. Estimating the Age of Formation of Lakes - an Example from Lake Tanganyika, East-African Rift System. Geology 21:511-514.

Day, J. J., J. A. Cotton, and T. G. Barraclough. 2008. Tempo and Mode of Diversification of Lake Tanganyika Cichlid Fishes. Plos One 3.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. Plos Genetics 2:762-768.

Farias, I. P., G. Orti, and A. Meyer. 2000. Total evidence: molecules, morphology, and the phylogenetics of cichlid fishes. J Exp Zool 288:76-92.

Fryre, G., and T. D. Iles. 1972. The cichlid fishes of the great lakes of Africa: their biology and evolution. Oliver and Boyd, London, UK.

Gee, H. 2003. Evolution - Ending incongruence. Nature 425:782-782.

Genner, M. J., P. Nichols, G. Carvalho, R. L. Robinson, P. W. Shaw, A. Smith, and G. F. Turner. 2007a. Evolution of a cichlid fish in a Lake Malawi satellite lake. Proceedings of the Royal Society B-Biological Sciences 274:2249-2257.

Genner, M. J., O. Seehausen, D. H. Lunt, D. A. Joyce, P. W. Shaw, G. R. Carvalho, and G. F. Turner. 2007b. Age of cichlids: New dates for ancient lake fish radiations. Molecular Biology and Evolution 24:1269-1282.

Greenwood, P. H. 1974. Cichlid fishes of Lake Victoria, East Africa: biology and evolution of a species flock. Bulletin of the British Museum Natural History Zoology Suppl. 6:1-134.

Harrison, R. G. 1989. Animal Mitochondrial-DNA as a Genetic-Marker in Population and Evolutionary Biology. Trends in Ecology & Evolution 4:6-11.

Heled, J., and A. J. Drummond. 2010. Bayesian Inference of Species Trees from Multilocus Data. Molecular Biology and Evolution 27:570-580.

Hey, J. 1992. Using Phylogenetic Trees to Study Speciation and Extinction. Evolution 46:627-640.

Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: Traditional and Bayesian approaches. Nature Reviews Genetics 4:275-284.

Huang, H., Q. He, and L. L. Knowles. 2010. Sources of Error for Species-Tree Estimation: Impact of Mutational and Coalescent Effects on Accuracy and Implications for Choosing Among Different Methods. Systematic Biology.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics 22:225-231.

Kocher, T. D. 2004. Adaptive evolution and explosive speciation: The cichlid fish model. Nature Reviews Genetics 5:288-298.

Knowles L. L. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Systematic Biology. 2009. 58(5):463–467.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971-973.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology 56:17-24.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and clustal X version 2.0. Bioinformatics 23:2947-2948.

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

Liu, L., L. L. Yu, and D. K. Pearl. 2010. Maximum tree: a consistent estimator of the species tree. Journal of Mathematical Biology 60:95-106.

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

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

McCormack, J. P., H. Huang, and L. L. Knowles. 2009. Maximum-likelihood estimates of species trees: How accuracy of phylogenetic inference depends upon the divergence history and sampling design. Systematic Biology:501-508.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Systematic Biology 52:674-683.

Nagl, S., H. Tichy, W. E. Mayer, N. Takahata, and J. Klein. 1998. Persistence of neutral polymorphisms in Lake Victoria cichlid fish. Proceedings of the National Academy of Sciences of the United States of America 95:14238-14243.

Pamilo, P., and M. Nei. 1988. Relationships between Gene Trees and Species Trees. Molecular Biology and Evolution 5:568-583.

Parker, A., and I. Kornfield. 1997. Evolution of the mitochondrial DNA control region in the mbuna (Cichlidae) species flock of Lake Malawi, East Africa. Journal of Molecular Evolution 45:70-83.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Salzburger, W., A. Meyer, S. Baric, E. Verheyen, and C. Sturmbauer. 2002. Phylogeny of the Lake Tanganyika Cichlid species flock and its relationship to the Central and East African Haplochromine Cichlid fish faunas. Systematic Biology 51:113-135.

Samonte, I. E., Y. Satta, A. Sato, H. Tichy, N. Takahata, and J. Klein. 2007. Gene flow between species of Lake Victoria haplochromine fishes. Molecular Biology and Evolution 24:2069-2080.

Schwarzer, J., B. Misof, D. Tautz, and U. K. Schliewen. 2009. The root of the East African cichlid radiations. Bmc Evolutionary Biology 9:-.

Seehausen, O. 2006. African cichlid fish: a model system in adaptive radiation research. Proceedings of the Royal Society B-Biological Sciences 273:1987-1998.

Sefc, K. M., S. Baric, W. Salzburger, and C. Sturmbauer. 2007. Species-specific population structure in rock-specialized sympatric cichlid species in Lake Tanganyika, East Africa. J Mol Evol 64:33-49.

Shaw, P. W., G. F. Turner, M. R. Idid, R. L. Robinson, and G. R. Carvalho. 2000. Genetic population structure indicates sympatric speciation of Lake Malawi pelagic cichlids. Proceedings of the Royal Society of London Series B-Biological Sciences 267:2273-2280.

Sparks, J. S., and W. L. Smith. 2004. Phylogeny and biogeography of cichlid fishes (Teleostei : Perciformes : Cichlidae). Cladistics 20:501-517.

Swofford, D. L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). version 4. Sinauer Associates.

Takahashi, K., Y. Terai, M. Nishida, and N. Okada. 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Molecular Biology and Evolution 18:2057-2066.

Takahata, N. 1989. Gene Genealogy in 3 Related Populations - Consistency Probability between Gene and Population Trees. Genetics 122:957-966.

Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. Bmc Bioinformatics 9:-.

Verheyen, E., W. Salzburger, J. Snoeks, and A. Meyer. 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science 300:325-329.

Won, Y. J., A. Sivasundar, Y. Wang, and J. Hey. 2005. On the origin of Lake Malawi cichlid species: A population genetic analysis of divergence. Proceedings of the National Academy of Sciences of the United States of America 102:6581-6586.

Won, Y. J., Y. Wang, A. Sivasundar, J. Raincrow, and J. Hey. 2006. Nuclear gene variation and molecular dating of the cichlid species flock of Lake Malawi. Molecular Biology and Evolution 23:828-837.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.The University of Texas at Austin.

APPENDIX

Table 4 A. 1 The phylogenetic studies used in meta-analysis.

Studies	#loci	Loci	Length	#taxa
Booton1999	1	ITS	605	17
Brandstatter2005	3	CR,cytb,ND2	441,402,1047	48
Clabaut2005	3	ND2,RAGexon3,RAGintron2	1047,1117,691	37
Day2007	1	NADHCR	1930	107
Duftner2005	2	CR,NADH	972,1047	55
Egger2007	1	CR	444	117
Farias2000	3	16s,TMO4C4,TMOM27	533,511,401	9
Farias2001	1	cytb	1138	11
Genner2007(1)	1	CR	483	378
Genner2007(2)	3	16s,cytb,TMO	576,1212,511	15
Joyce2005	1	CR	897	89
Koblmuller2004	3	CR,cytb,ND2	445,402,1047	29
Koblmuller2005	3	CR,cytb,ND2	362,402,1047	27
Koblmuller2007(1)	2	Dloop,ND2	363,1048	42
Koblmuller2007(2)	1	ND2	1047	63
Koblmuller2008	2	cytb,ND2	402,1047	65
Koblmuller2010	2	CR,ND2	958,1047	109
Koch2007	1	CR	361	14
Kocher1995	1	ND2	1047	31
Mayer1998	1	DXTU1	500	81
Meyer1990	1	CR, cytb	803	5
Nagl2000	1	CR	845	114
Nagl2001	1	CR	454	70
Nevado2009_mtDNA	2	CR, cytb	849,795	201
Nevado2009_nDNA	5	ATP,ITS,LSU,RAG-i2,RAG-e3	948,676,1306,928,820	75
Rubber1999	2	CR,cytb	482,402	93
Salzburger2002_Fig4	2	cytb,ND2	402, 1047	42
Salzburger2002_Fig6	1	CR	969	46
Salzburger2005	2	CR, ND2	1000, 1477	103
Schelly2006	2	CR, ND2	369, 1047	58
Schwarzer2009	8	12S_16s, Nd2,exons, (ENCI,PTR, SH3PX3,Tmo4c4), introns	892, 1008, 2522, 493	63
Shaw2000	2	CR, ND2	981, 865	81
Sparks2004	4	COI, 16s, H3, TMO4c4	649, 580, 335, 507	9
Streelman1998	2	TMOM27, TMO4c4	511, 511	13
Sturmbauer1992	1	Cytb	400	15
Sturmbauer1993	2	CR,cytb	488, 402	30
Sturmbauer1994	1	CR	444	35
Sturmbauer2003	2	CR,cytb	1,004,402	44
Sultmann1995	3	DXTU1,DXTU2,DXTU3	499,360,174	58
Verheyen1996	1	CR	441	44
Verheyen2003	1	CR	843	287
Won2006_ND2	1	ND2	1048	52
Won2006_nuDNA	12	AIM1,DXTUCA3,EDNRB1,	263,419,773,732,556,	13

		MITFB,Ppun7,PZMSAT2,U14396, U66814,U66815, UNH001,UNH130,UNH143-II	642,442,633,200,357, 374,594	
Zardoya1996	1	Tmo27	467	67
Samonte2007_mtDNA	1	CR		10
Samonte2007_nuDNA	5	Hag, Opn1lws, MC1R,Tyr, SINE1357		10

Table 4 A . 2 Population genetic studies on cichlids fish having estimates of population size.

Paper	**Species**	**Direct or indirect size estimates**	**Estimates of population genetic parameter**	**mutation rate used**
Kolbmuller et al. 2007 Genetica	*Neolamprologus caudopunctatus*	east, Ne = 300,000–410,000; west, Ne = 820,000–1,100,000		
Samonte et al. 2007 MBE	Lake Victoria *Haplochromine*	41,000–165,000		
Crispo and Chapman. 2008 ME	*Pseudocrenilabrus multicolor victoriae*	17,500-23,000	average π of populations is 0.00154	6.5-8.8*10^-8 (from Baric 2003)
Maeda et al. 2009 Gene	*Haplochromis pyrrhocephalus* and *H. laparogramma*	173,000	theta old=0.001-0.004	2.3*10^-8
Mzighani et al.2010 Gene	*Haplochromis laparogramma*	50,000-8,400	theta=1.013	0.2–1.2 × 10^4
Elmer et al. 2010 BMC Biology	*Amphilophus cf. citrinellus* two morphs	27,258-36,108		
Sato et al. 2003 MBE	*Haplochromis* "Nshere" and *Haplochromis* "Lutoto."	10,000-100,000		
van Oppen et al. 2000 MBE	related species of mbuna	2,500-18,000		
Hey et al. 2004 ME	*Ropheops tropheops* and *T. gracilior*	15,900 and 5,000-8,000	theta=0.53	10^-9 for nuclear point mutation rate
Won et al. 2005 PNAS		T. gracilior 1,500-4,900, T. tropheops and T. broad mouth15,400-19,000		
Duftner et al. 2006 ME	*Variabilichromis moorii*	105,000-142,000	average theta=0.009245	6.5-8.8*10^-8(from Baric 2003)
Danley et al.2000 Evolution	*Metriaclim*	12,000	Hoo ~ 0.9	10^-4,OHTA and KIMURA 1973
Shaw et al. 2000 Proc Roy soc London B	Diplotaxodon species	27,000	ho~0.79	
Taylor et al.2001 Proc Rsof london	*Eretmodus cyanostictus*	36,000	ho~0.5, 9 populations	
Ruber et al.	*Eretmodus, Tanganicodus,*	20,000-100,000;		

2001 ME				
Nagl et al. 1998 PNAS	Lake Victoria *Haplochromine*	100,000		
Parker and Kornfield 1997 JME	mbuna species	110,000 -237,000		
Meyer et al. 1996 Me	*Simochromis babaulti* and *S. diagramma*	29,000-56,000	Pi=0.0026, 0.0037	6.5-8.8*10^-8(from Baric 2003)
sefc et al. 2007 JME	*Eretmodus cyanostictus*, *Tropheus moorii*, and *Ophthalmotilapia ventralis*	297,000-633,000;319,000-600,000;99,000-300,000		

Figure 4A.1 Diagrammatic illustration of the gene-tree trimming process. a) A hypothetical gene tree. Different colors of external branches represent lineages from different species. In b) and c), the branches with lighter color and an "X" sign represent the trimmed branches. b) Only one tip was preserved for a monophyletic clade of tips from the same species (the orange species), and the lineages first coalescing with a lineage from the other species was retained for the paraphyletic clade (the blue species). c) All possible combinations of species positions were considered for species that still had multiple lineages on the gene tree after step b) (the red species).

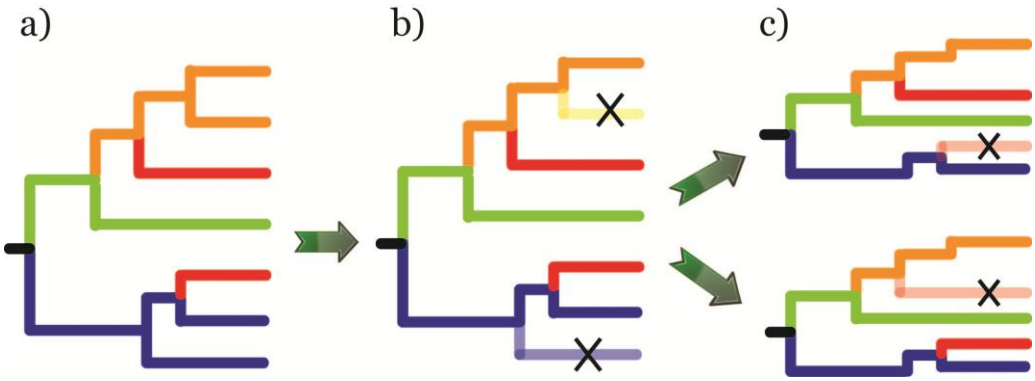
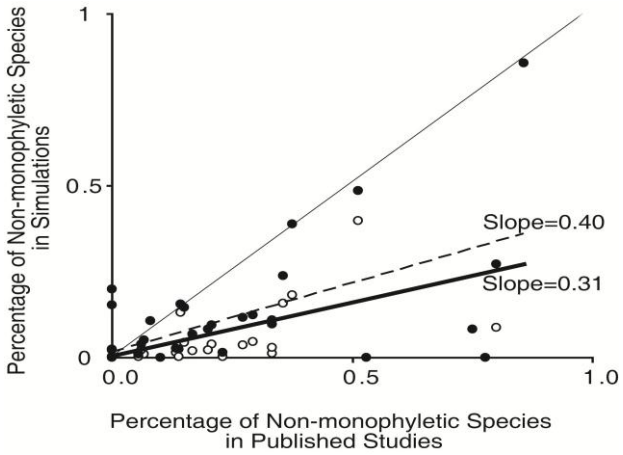


Figure 4A.2. The correlation between the percentages of non-monophyletic species in simulation and the percentage in published studies. The simulation was performed with the exact sampling configuration as in empirical studies and use of ¼ of the effective population size for mtDNA studies. The open (close) dots represent results with the population size set as 50,000 (100,000), and the thick solid (dash) line is the regression line with the slope shown besides it. ‡ The thin solid is the 45 degree line.



‡Note: Since the sampling configurations are the same, if the set population size is a good approximation to real population sizes, the percentages of non-monophyletic species in simulation should be similar to the ones seen in published studies (i.e., located on the 45 degree line). If the population size set in simulation is too high, the percentages in simulation will be higher, and vice versa. However, another problem intervenes with the calculation of non-monophyletic species—species delimitation. Species delimitation is a subject of debate for many recently diverged groups, and the cichlid group—447-535 species in LM, 451-600 species in LV (Genner et al., 2004) and 200 in LT (Day et al., 2008)—is no exception. That is, species might be non-monophyletic in empirical studies but not in simulation because specimens of different species were grouped into one "species." Therefore, we choose to report the conservative result with smaller population size, even though the larger value (100,000) provides a better fit to the empirical percentages of non-monophyletic species.

Figure 4A.3. Chance of non-monophyly under simulation setting L. a) Histogram of the percentage of non-monophyletic species across the 46 studies. The open (filled) bars represent result with sampling 5(10) individuals per species; b) The correlation between the percentage of non-monophyletic species in simulated datasets and the percentage observed in the published studies. The solid line is a 45 degree line; c) The correlation between a species' chance of being non-monophyletic and the divergence time between it and its sister species/clade. The three vertical lines indicate the median estimates of lake age. LT: Lake Tanganyika; LM: Lake Malawi; LV: Lake Victoria. In both b) and c), the open (filled) dots represents simulations with sampling 5 (10) individuals per species.

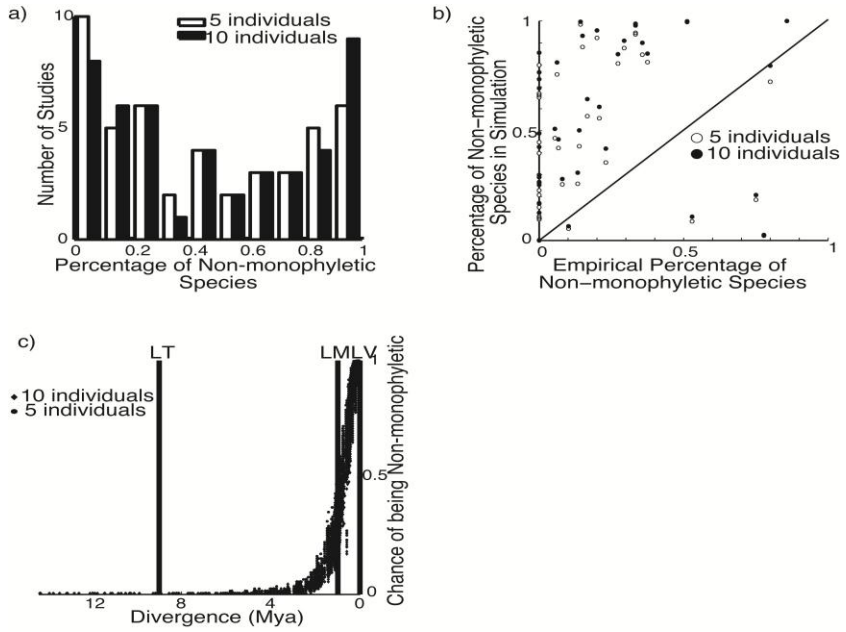


Figure 4A.4. Chance of incongruence between the "true" species trees and gene trees under setting L. a) Histogram of the percentage of incongruent innernodes of all the studies. b) Correlation between the chance of incongruence and the depth of the innernodes on the time scale. The dash line indicates the relationship predicted by linear regression, with the R^2 value shown by the side. The three vertical lines indicate the median estimates of lake age. LT: Lake Tanganyika; LM: Lake Malawi; LV: Lake Victoria. The intensity of the grey dots indicates the weight, which is the inverse of the number of divergent histories corresponding to one study.

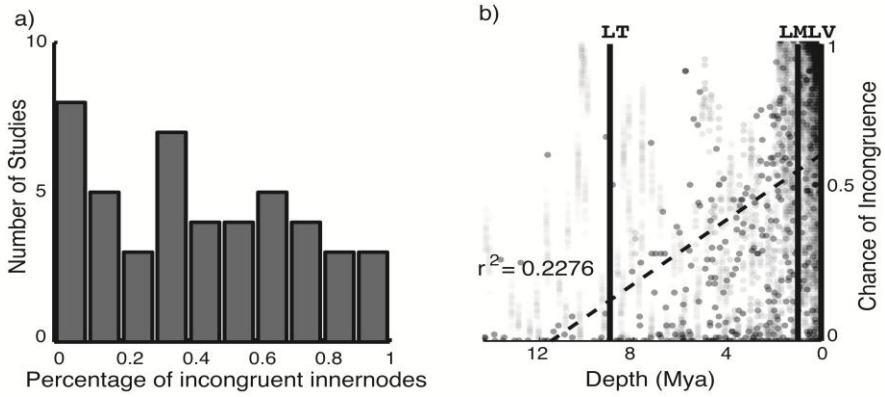
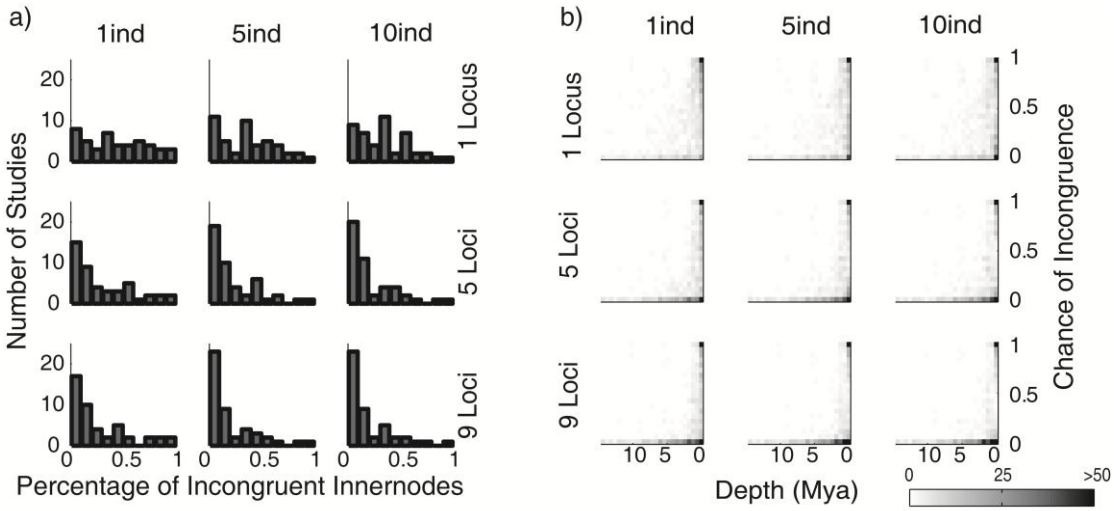


Figure 4A.5 Effects of increasing sampling effort and using MDC to estimate species tree despite of discordance with setting L. a) Histograms of the percentage of incongruent innernodes (i.e., the innernodes that cannot be found on the estimated species tree by MDC) under nine different sampling designs. b) Grey-scale plots of the correlation between the chance of incongruence and the depths of the innernodes on the time scale with different sampling designs. A darker color means more points in the block, as indicated by the scale in the lower left corner.



Chapter 5 Exploring the use of Next-generation sequencing in Species-tree Estimation

Finding suitable genetic markers is a well-known obstacle when studying recent species divergent history in non-model organisms. Most studies have to rely on mitochondrial genes or one of a few universal nuclear markers. Many times, these universal markers do not have high enough mutation rate to resolve shallow divergent histories. Even if the mitochondrial genes could give well-supported gene trees, they do not give a robust estimation of the species-tree topology as a single realization of the lineage sorting process. The practice of concatenating multiple genes to obtain enough informative sites could give positively misleading results under certain divergent histories (Degnan and Rosenberg, 2006; Kubatko and Degnan 2007). Recent work on species-tree estimation methods has shown promise in resolving these shallow divergent histories despite conflicting gene trees of different locus (e.g., Maddison and Knowles 2006; Ané et al. 2007; Mossel and Roch 2007; Kubatko et al. 2009; Liu 2008; Liu and Pearl 2010; Heled and Drummond 2010), but applying these methods requires sequencing data from multiple loci, which is critical in decreasing the estimation errors associated with coalescent variance (Huang et al. 2010). Hence, the old problem for systematics re-emerges in a new context—how can we obtain enough independent and variable loci for species-tree estimation?

Next-generation sequencing (NGS) offers efficient and affordable ways for obtaining nuclear markers in non-model organisms. This technique can generate a large amount of shotgun sequences from non-model organisms' genomes, which eliminates the lengthy process of cloning and screening for variable loci in the traditional approach of BAC cloning and Sanger sequencing. Several Reduced Representation Library (RRL) methods were explored to select a certain subset of the genome to increase the coverage of the loci being sequenced (for review of the methods see Davey et al. 2011). Techniques for labeling sequences allow multiplexed NGS runs, further reducing the cost of generating multi-locus data for multiple species (Binladen et al., 2007; Meyer et al., 2008). Despite all these benefits,

the use of NGS in studying population demographic history is still limited comparing to its popularity in studies of disease genetics and community metagenomics (Mardis 2008). To date, the majority of research papers about NGS in non-model organisms have focused on using the Single Nucleotide Polymorphisms (SNPs) found in NGS data, or treat NGS merely as the first step to screening for variable loci which need to be confirmed and re-sequenced in follow up studies (e.g., Wiedmann et al., 2008; Van Tassell et al., 2008; Amaral et al., 2009; Hyten et al., 2010; Williams et al., 2010).

The potential of directly using NGS data for studying evolutionary history has not been fully explored. The main concern when using the raw data is the high error rate associated with pyrosequencing. Moreover, the errors in the PCR steps prior to the sequencing run are amplified. There are also errors in assembling and mapping the short NGS reads, resulting in false positive or negative SNP calls. However, as reviewed in Pool et al. (2011), many effects of these sequencing errors could be corrected or mitigated by appropriate statistical methods, and accurate population genetic inferences were achievable. A recent study (Luca et al. 2011) used Illumina data combined with a simple reduced representation technique to collect genome-wide variation data in human populations, showing that even with a limited sample size (only 19 individuals), NGS data could recapitulate many of the demographic features known from previous studies. This further demonstrates that the errors in NGS data are amendable for estimating population demorgraphic history.

Obtaining species-tree estimates that are robust to NGS errors should be possible, especially given that the primary interest of phylogeny—tree topology—is likely to be less sensitive to sequencing errors. For instance, random sequencing errors can greatly inflate the estimates of nucleotide diversity by "creating" singleton alleles, but for estimating species trees, purely random errors are likely to increase the estimates of tip branch lengths or current population sizes, but the occurrence of topological change would need a more profound pattern of errors. Nevertheless, directly using NGS data for species-tree estimation is not without its own challenges. As most non-model organisms do not have a reference genome, distinguishing paralogous and othologous contigs would be a crucial step as species trees estimated from or jointly estimated with gene trees of paralogous genes would be erroneous. Since NGS provides shotgun sequencing data, it would also include many loci with little

variation. Given the size of NGS datasets it would be wasteful to discard these loci, but the impact of including these markers is unknown.

With these questions in mind, we explored the direct use of NGS data from the 454 platform to estimate the species tree for four grasshopper species in the genus *Melanoplus*, a group with high species diversity and recent divergence time (Knowles and Otte 2000), using a similar reduced-representation strategy as in Gompert et al. (2010). Different stringency levels for error correction were used, and the corresponding species-tree estimates were compared to assess the robustness of species-tree estimates to NGS errors. The effects of including low variation loci were investigated by excluding loci without a minimum number of informative sites (detailed below). Parametric simulations were also conducted to examine whether the effects of the true divergent history itself versus errors associated with the NGS and filters are the primary determinants of uncertainty for species-tree estimation.

METHODS

AFLP-based Next-generation sequencing

Aiming to obtain orthologous sequences from four species of *Melanoplus*: *M. oregonensis*, *M.marshalli, M. triangularis* and *M.montanus* using the 454 next-generation sequencing platform, we applied the AFLP (amplified fragment length polymorphism) technique prior to the next-generation sequencing run to reduce the complexity of the genomic DNA using a protocol adapted from Gompert et al. (2010). By digesting the genomic DNA template with endonucleases, and selectively amplifying fragments of 350-600 base pairs, the chance of capturing orthologous sequences among species was increased.

One grasshopper was sampled from each of the four species, and leg muscle tissues were isolated for DNA extraction using the Qiagen's DNeasy Blood and Tissue Kit (Qiagen Inc.) following the recommended protocols. Extracted genomic DNA were digested with three combinations of two restriction endonucleases—*EcoRI* and *MseI*. Each digestion used ~5ng genomic DNA, was achieved by incubating the enzyme together with DNA at 37ºC for 2 hours, and followed by 70ºC of 15 min to deactivate the enzymes. Subsequently, these digested fragments were linked to *EcoRI and MseI* adaptor oligos at 20ºC for 2 hours. All the reagents used in the digestion and ligation steps except the enzyme solutions were from the

Invitrogen's AFLP core Reagent Kit (Invitrogen Inc.). The ligation products were amplified with pre-selective AFLP primers (EcoRI and MseI) in PCR machine with 2min at 74ºC, 20s at 98ºC, 30 cycles of 20s at 98ºC, 30s at 56ºC and 2 minutes at 72ºC, followed by 10min at 72ºC for final extension. To minimize the PCR error, iProof high fidelity polymerase was used for all the PCR steps. The PCR products were separated on a 2% agorose gel, and fragments between 350-650bp were manually excised and purified using GENECLEAN Turbo DNA purification kit. Those fragments with additional tags and primers (~ 50bp) would be in the recommend length range of the 454 GS FLX Titanum platform for amplicons. A secondary PCR amplification was performed to attach species-specific MID barcodes to amplicons. That is, the PCR primers were composed of two segments: the EcoRI or MseI sequence as described above and a species-specific 10 base pairs attached to the 5' end (MID barcodes; Roche). These MIDs enable us to differentiate the reads of different species from the 454 sequencing run. The resulting PCR products were purified, and samples of the same species with different combinations of digestion enzymes were combined.

The concentration of the pooled samples for each species was quantified at the University of Michigan sequencing core facility using pico green analysis. Samples of the four grasshopper species were then combined with 52 other species (which were the subject of other studies) in equal concentrations as the template for one 454 sequencing run. 454 sequencing was performed by the University of Michigan sequencing core using the 454 GS FLX Titanum platform, following the protocols provided by Roche. The sequencing principles and procedures are described in Margulies et al. (2005). Briefly, amplicons were blunted-ended and ligated to adaptors with sequences (primer A and primer B), which could later attach to capture beads. Emulsion PCR was performed on the immobilized single-stranded fragment on each bead.

Filtering reads, assembling sequence and finding phylogentic informative loci

Because of the multiplexed samples, reads were first assigned to their original sample according to their MID barcodes. To avoid false recognition of MIDs (e.g., the last 8bp of MID1 only has one base pair difference from the first eight base pairs of the EcoRI primer), the location of EcoRI or MseI adaptor the strategy was identified first, and then the MID

barcodes were searched in in the 5' sequences. Specifically, the first and last 60 base pairs of the reads were used for adaptor searching (reads shorter than 100 base pairs were discarded. No more than two errors (including point errors, insertions and deletions) were allowed in the matching. Only those 5'sequences with reads longer than or equal to 5bp before the adaptors were extracted were aligned to MID barcodes (i.e., sequences with less than 5bp were considered as too short to accurately identify MIDs). If the sequence length is less than or equal to 6bp, an exact match was required. One error was allowed for 7-8bp sequences, and two errors for 9~10bp sequences. After sorting, the reads with repetitive sequences – which are difficult to align between species and estimate evolution models – were excluded using program TRFv3.21 (tandem repeats finder; Benson 1999). The scoring parameters were set to recommended values, and the minimal required score for a tandem repeat was set to 50 without any limit on repeat size (i.e., set to 2000).

Reads for the four grasshopper species were assembled into contigs in the SeqMan NGen sequence assembler v2.0.0 program (DNA*, Madison, WI, USA), excluding the sequences corresponding to the MIDs and adaptors. As reads were from different species, the assembly settings allowed larger variation within contigs – a mer match size of 19, and mer spacing of 20 bp and a minimum march percentage of 85%. The match score was set to 10, while mismatch and gap penalties were set to 15 and 75, respectively. Quality scores from 454 runs were used by the assembler to trim off ends with low quality—specifically, regions over 10bp with average quality score lower than 15 were flagged and removed.

Two types of possible assembly errors might interfer with phylogenetic analysis: (i) splitting reads from the same loci into different contigs, and (ii) grouping reads from different loci into one contig (e.g., failure to detect paralogs). The first error type would result in "duplicate" loci in the dataset, while the second type of error will leads to gene trees that cannot be modeled by lineage sorting process. To minimize the effect of assembly errors, consensus sequences of contigs were aligned to each other in the CAP3 program (Huang and Madan 1999). According to the contigs of consensus sequences, reads of different contigs were pooled and reassembled in CAP3 if their consensus sequences have more than 85% similarities. The contigs were excluded if their consensus sequences have high similarity but the reads can not be assembled into one contig. These contigs might represent loci with high similarities or some reads with high error rates, and either reason means high uncertainty in

assembling. We also assumed a maximum fixed coverage of 50× as in Gompert et al., (2010), as copy number variants were associated with high coverage loci (Alkan et al., 2009).

Given the aim was to identify genes for reconstructing species tree, two criteria were applied to the assembled contigs. First, contigs that do not have reads from all species were excluded. Second, contigs that only have species-specific alleles were removed since they would not contain information for reconstructing genealogical history. Customize Perl scripts were used to remove contigs.

Characterizing error pattern and haplotype estimation

Next-generation sequencing technologies are known to have higher error rates than Sanger sequencing. Such errors would lead to an overestimation of genetic variation. However, with overly stringent criteria, true polymorphisms would mis-assigned as errors, thereby reducing the phylogenetic signal in the data. It is important to characterizing the error pattern in NGS data. Two approaches were used to characterize error rates: (i) using the known adaptor sequences, and (ii) considering the proportion of variable sites in different sites categories.

For the first approach, error rates, as well as the different types of errors (i.e., point errors, deletion and insertions) were calculated from the 15-16bp sequences of EcoR and Mse adaptors added to every genomic DNA pieces after digesting with restriction enzyme by comparing reads for a given contig. The error rate is calculated as the number of total mismatches divided by the total length of adaptors. The point error, deletion and insertion rates were calculated without the first and last nucleotides because of the difficulty in distinguishing these three types of errors at these two sites. 454 sequencing technique also provide a quality score along with each base, but the correlation between these quality scores to actual error rates is not well studied, in contrast with Sanger sequencing (Ewing and Green 1998). Groups of sites in the adaptor sequences were created according to quality scores. The error rates in each quality-score groups were calculated, and the relationship between the machine-giving scores and actual point error rates (proportion of mismatching) was examined.

For the second approach, error rates were quantified by comparing the pattern of variable sites in contigs to determine whether certain types of sites should be included or not. Specifically, potential sequencing errors were labelled as one of four types: “within

homopolymers", "adjacent to homopolymers", "adjacent to gaps" and "other" (i.e., all other sites not encompassed by the other 3 categories). Pyrosequencing techniques are known to have high error rate for homopolymers stretches (Huse et al. 2007). We therefore focused on homopolymers (i.e.,three consecutive identical nucleotides in the consensus sequence). Errors in the homopolymers might also lead to assembly errors in the adjacents sites, and are recognized as polymorphisms. Gaps in the assembled contigs could be a sign for error-prone area. Even though they could represent true insertion/deletion variation, given the high insertion and deletion rates observed in the adaptor sequences, gaps are more likely to reflect sequencing errors. Hence, gaps were treated conservatively in the analysis—a gap in the reads is kept and coded as 'N' only when there is only one read have the gap, and at least two other reads from the same species with non-gap alleles. Yet, gaps could still be associated with assembly errors that make adjacent sites being falsely identified as variable sites. Therefore, the proportion of variable sites for the sites besides gaps is also calculated.

With an estimate of the probability of error, the probability a variable site is a true polymorphic site can be calculated based on the coverage (i.e., the number of reads at the site) and the number of the major and minor alleles. Despite the variety of statistical genotype-calling methods (e.g., Hellmannn et al., 2008; Lynch 2008), the majority of them examine variation one site at a time. This approach wastes valuable information in the data—namely, the linkage between sites. If minor alleles on two sites were linked across reads, the probability they reflect true polymorphic loci would be higher than the probability when treating them independently (Figure 5.1). Furthermore, most of species-tree methods are based on gene trees, which require haplotypes as input data. Here, we developed a method to obtain the joint estimates of the haplotypes and genotype. Specifically, given a locus with n reads and m number of variable sites, the task is to find the configuration of dividing these reads into two haplotypes—C, and the genotype of m sites—G that have the maximum probability given the data—D (Figure 5.1).

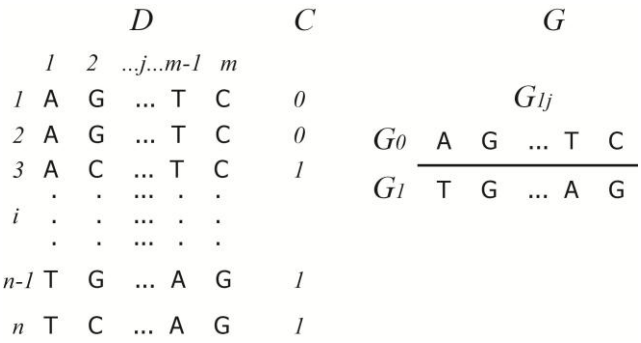


Figure 5.1Hypothetical data example.
One hypothetical example of data (D), configuration of dividing reads into two haplotypes (C) and the genotype of sites (G), where i is used to denote different reads, and j for different variable sites

According to Bayes' theorem:

$$P(C,G \mid D) \propto P(D \mid C,G)P(C \mid G)P(G)$$

C could be expressed as a vector of length *n*, where

$$C_i = \begin{cases} 0, when \quad ith \quad read \quad grouped \quad to \quad haplotype \quad 0 \\ 1, when \quad ith \quad read \quad grouped \quad to \quad haplotype \quad 1 \end{cases}$$

D could be deemed as a matrix with *n* rows and *m* columns, where D_{ij} represents the reading at the *j*th variable site from the *i*th read. Accordingly, G_{0j} and G_{1j} would be the genotype of the two haplotypes at site *j* (Figure 5.1). Then, when $\sum_{i=1}^{n} C_i \neq 0$, that is, there are two haplotypes in the configuration

$$P(D \mid C,G) = (1-\varepsilon)^{\sum_{i=1}^{n}\sum_{j=1}^{m} I(D_{ij}=G_{c_i j})} \left(\frac{\varepsilon}{3}\right)^{\sum_{i=1}^{n}\sum_{j=1}^{m} I(D_{ij}\neq G_{c_i j})}$$

where ε is the error rate.

$$P(C \mid G) = \left(\begin{matrix} n \\ \sum_{i=1}^{n} C_i \end{matrix} \right) \left(\frac{1}{2}\right)^n$$

This is the probability of sampling $\sum_{i=1}^{n} C_i$ reads of one haplotype when two haloptypes have equal chance of being sequences. Assuming that there is g number of variable sites with the inferred genotype *G*, where *g* could be different than *m* if some variable sites were considered as sequencing errors,

$$P(G) = \left(\frac{1}{2}\right)^{g-1} \times P(g)$$

The first part is the probability of having the genotypes of the *g* variable sites distributed between the two haplotypes as in *G*, and *P(g)* is the probability of have *g* segregating sites when sampling two haplotypes from the population. *P(g)* depends on the effective population size (*Ne*), mutation rate (μ), and the locus length *L*, and could be calculated using Ewen's formula (Ewen 1972):

$$P(g) = (\frac{1}{1+\theta})(\frac{\theta}{1+\theta})^g, \theta = 4N_e\mu \times L$$

Given previous estimates of population sizes in Montane grasshoppers (~ 10^6), we set θ to be 0.01 (assuming the average mutation rate for nuclear loci to be 2.5×10^{-9}). The calculation differs a little when $\sum_{i=1}^{n} C_i = 0$. That is, all the reads are considered either coming from one haplotype or there is no true variable site among this reads:

$$P(G,C \mid D) = (1-\varepsilon)^{\sum_{i=1}^{n}\sum_{j=1}^{m} I(D_{ij}=G_{c_i j})} (\frac{\varepsilon}{3})^{\sum_{i=1}^{n}\sum_{j=1}^{m} I(D_{ij}\neq G_{c_i j})} \times [P(g=0) + P(g\neq 0) \times \binom{2}{1} (\frac{1}{2})^n]$$

For all the contigs, perl scripts were written to go through all the possible haplotype configurations, and consensus sequences of the two haplotypes were used as genotype, as non-consensus genotypes would always have lower probability given the small error rate (3×10^{-4}, calculated based on adaptor sequences). To simplify the problem, only the most and the second most common alleles were considered for each variable site. The configuration and genotype having the highest probability were saved as a point estimate of the true genotypes.

Haplotype construction itself is one way of filtering sites, as it prefers the linked variable sites and labels unlinked variations as sequencing errors. To assess the effect of additional filtering, we also applied a "twice" filtering criterion to the data before inferring haplotypes. That is, the minor allele in each species has to appear in at least two reads. The species-tree estimation results with or without this additional layer of filtering were compared.

Species tree estimation and parametric simulation

Species tree were estimated using a Bayesian method in *BEAST (Heled and Drummond 2010). Gene trees and the species trees were jointly estimated in the Bayesian framework. This method is chosen in particular for it easy way to obtain variance estimates from the

posterior distribution of trees. *BEAST was applied with a separate relaxed molecular clock model and unlinked HKY substitution model for each gene. Random starting gene trees under coalescent model were used, and the Yule process was used as the species tree's prior. Population sizes were set to be piecewise constant, with a gamma distribution as prior. Analysis were run for 100 million generations and sampled every 4,000 generations with other default operator settings. After a 1000 generation burn-in phase, sampled generations were used for estimating the majority-rule consensus species tree, the posterior probability for different species-tree topologies (see Table 5.1for the list of all possible topologies and their corresponding number used in this paper), and the 95% credible set of topologies was determined based on the cumulative frequencies. Multiple independent runs were performed to ensure the convergence on the distribution of posterior probabilities.To evaluate the effects of including loci of low variability, species-tree estimates were obtained from datasets with and without loci with less than seven informative sites, and the posterior distributions were compared. Species-tree estimates were also compared when different error filters were employed.

Table 5.1 Topologies for a four-taxon tree.
The fifteen possible topologies for a four-taxon tree, and their assigned number in this paper. M.oregonensis, M.marshalli, M.triangularis, and M.montanus were shorten as ore, mar, tri and mon, respectively.

Topology	Number
(mon,(tri,(ore,mar)));	1
(tri,(mon,(ore,mar)));	2
(mon,(mar,(tri,ore)));	3
(mar,(mon,(tri,ore)));	4
(mar,(tri,(ore,mon)));	5
(tri,(mar,(ore,mon)));	6
(mon,(ore,(tri,mar)));	7
(ore,(mon,(tri,mar)));	8
(ore,(tri,(mon,mar)));	9
(tri,(ore,(mon,mar)));	10
(ore,(mar,(tri,mon)));	11
(mar,(ore,(tri,mon)));	12

((tri,mon),(ore,mar)); 13

((mon,mar),(tri,ore)); 14

((tri,mar),(ore,mon)); 15

Parametric simulation were used to examine which factors could contribute to the uncertainty in specie-tree estimation—the divergent history itself, the NGS errors that passed through haplotype estimation, and the stringent "twice" filter (i.e., the minor allele needs to be in at least two reads). The estimated species tree from the NGS data without additional filter and excluding low-variability loci was used to simulate ten sets of genealogies using the program *ms* (Hudson 2002). The number of sampled haplotypes in each species was set according to the estimated number of haplotypes for the empirical data. Seq-gen (Rambaut and Grassly 1997) was used to simulate error-free sequencing data for each locus using the corresponding mutation rates and substitution models estimated in *BEAST. To simulate the NGS data with errors, the same number of reads were generated with the estimated haplotype configuration for each locus, bases were masked as 'N' if the corresponding bases in the empirical dataset are 'N' and errors were randomly assigned to the sequences by flipping bases into any of three other types of nucleotides with equal chances. To test the robustness of the species-tree estimates to NGS errors, an error probability ten times higher than the empirical rate (i.e., 3×10^{-3}) was used in the simulations. Haplotypes were reconstructed for these simulated NGS datasets with and without the "twice" filtering, and species trees were estimated in *BEAST with the same setting for the empirical datasets.

RESULTS

Sequence assembly and patterns of sequencing errors

The multiplexed 454 run produced 1,142,629 reads in total, 138,905 (12%) of which can be attributed to grasshoppers by MID identification. Comparing the number of reads among the four *Melanoplus* species, *M.montanus* were over-represented (51,664 reads, 37%), and *M.triangularis* were under-represented (13,849 reads, 10%), while *M.orengonensis* and *M.marshalli* have 36,069 (26%) and 37,323 (27%) reads, respectively. The average length of reads after removing MIDs and EcoR/Mse adaptors are around 350 base pairs, which is at the

very low end of the intended length range. This might be caused by co-migration of shorter AFLP fragments on agarose gel and preference in binding shorter fragments in the 454 sequencing run. TRF (Tandem Repeat Finder, Benson 1999) identified tandem repeats in 453 (0.33%) reads, leaving an average of 34,613 reads per species (see Figure 5.2 for summary statistics and length distribution for each species). NGen sequence assembler assembled 135,288 (98%) reads into 4,584 contigs. After reassembling in CAP3 (Huang and Madan 1999), there are 405 contigs which have distinct consensus sequences (i.e., could not be aligned to other consensus sequences at 85% similarity level in CAP3), and coverage lower than the assumed maximum coverage. 119 (29.4%) contigs have sites with shared polymorphisms across species, 33 out of which have reads from all four species.

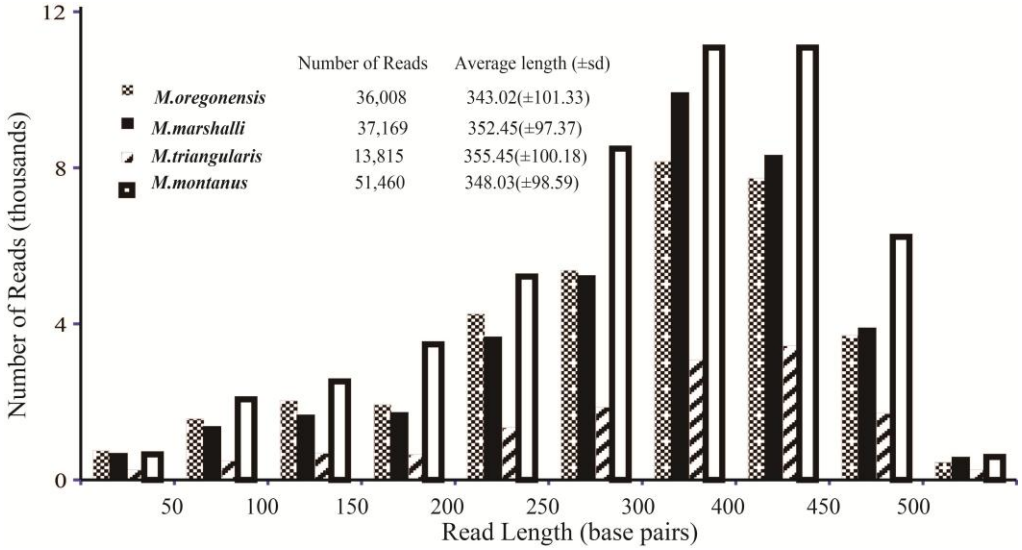


Figure 5.2 The number of reads and the length distribution in the four grasshopper species.

Based upon the adaptors' sequences, the average probability of sequencing error per base is 2.1×10^{-3}. Among the three types of errors, the rate of insertion is the highest—1.3×10^{-3}, and the rate of deletion and point errors are similar—3×10^{-4}. Although overall the error rate given by the quality scores is close to the rate of point errors, the scores have a weak correlation with the actual error rates (Figure 5.3a), which indicates that quality scores might not be a good indicator of base-specific error rates for this particular NGS dataset. Grouping sites into different categories revealed that the proportion of variable sites among sites adjacent to assembly gaps is significantly higher than the proportions in other categories

(Figure 5.3b). Hence, gaps and sites besides gaps were excluded from further analysis to minimize the influence from high error-prone regions of the reads.

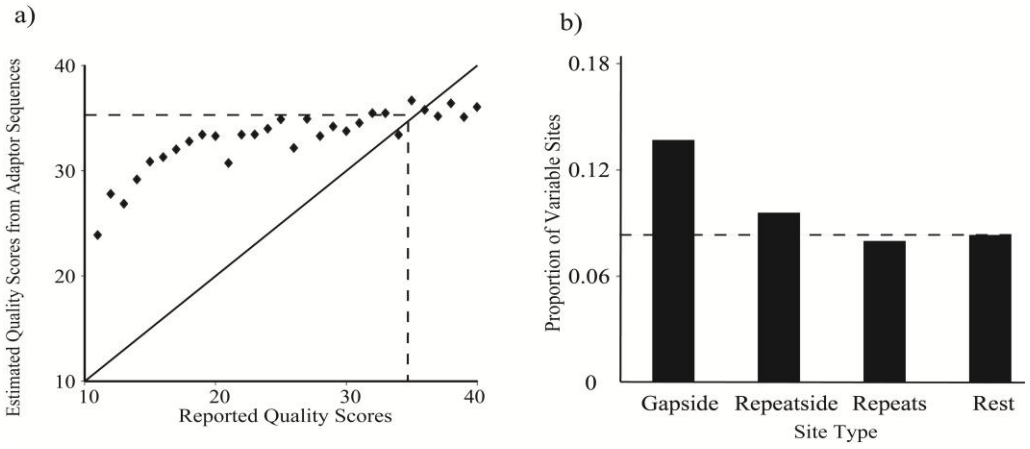


Figure 5.3 The pattern of sequencing errors.
a) The weak correlation between the reported quality score and the estimated quality scores from Adaptor sequences, the solid diagonal line indicate the expected 1:1 relationship, and the dash lines indicate the average reported and estimated quality scores of the adaptor sequences. b) The proportions of variable sites in sites of different categories.

Species tree estimation

As all previously developed genetic markers for this species group showed incomplete lineage sorting among species, a useful contig for species-tree estimation is expected to have not only variable sites, but also sites with shared polymorphisms (i.e., at least two alleles appeared in more than two species). This leaves 33 contigs that have at least one site with informative polymorphism and at least one read from each species. The number of contigs, the total number of sites, variable sites and sites with informative polymorphisms were summarized in Table 5.2. Haplotype estimation itself could function as a filter of sequencing errors—randomly occurred errors are unlikely to be linked across reads. It filtered out around 33% percent of variable sites. The distributions of length, variable sites, sites with informative polymorphisms and coverage in each species among contigs were summarized in the histograms in Figure 5.4.

Table 5.2 Summary of variable sites before and after haplotype estimation with different filters.
[1]The number of contigs having at least one site with shared polymorphism—at least two alleles appear in at least two species.[2]Sites with at least two alleles appear in at least two species

	No filter	Twice Filtered
Before Haplotype Estimation		

Contigs[1]	33	32
Total Sites	13,884	13,410
Variable Sites	2,112	1,260
Sites with shared polymorphisms[2]	381	185
After Haplotype Estimation		
Contigs[2]	33	30
Total Sites	13,884	12,991
Variable Sites	1,428	9,11
Informative Sites	408	139
Sites with shared polymorphisms[2]	220	100

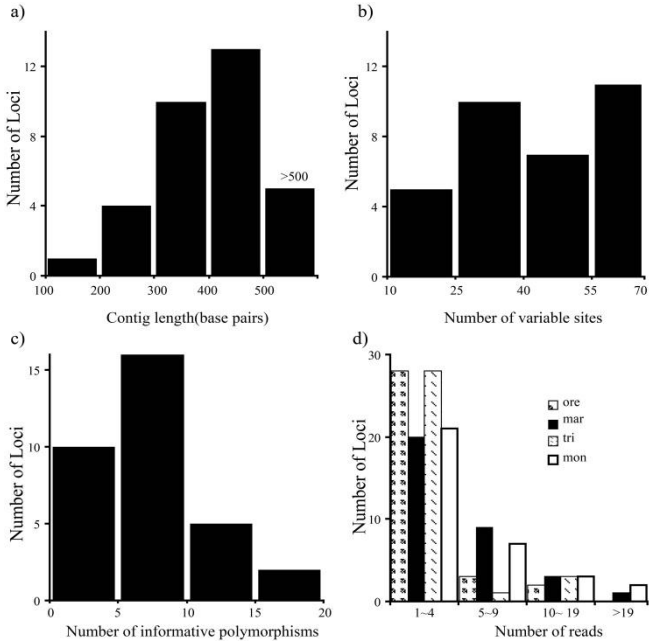


Figure 5.4 Distributions of length, variable sites, and informative polymorphisms.
The distribution of a) length, b) number of variable sites, c) number of sites with informative polymorphisms and d) number of reads for each species of loci that have at least one site with informative polymorphisms. The numbers in b) and c) are calculated based on reconstructed haplotypes.

The topology of majority-rule consensus tree is identical regardless of whether loci of low variability are included in the estimation (Figure 5.5a&b). Compared to the estimates based upon previously developed five genetic markers (Carstens and Knowles 2007), which were

sequenced in multiple individuals using ABI-Sanger sequencing, both NGS and Sanger datasets identified *M.triangularis* as the most distantly related species in this group (Figure 5.5c). The trees differ on which two species are the sister species, but neither type of datasets has high posterior support on this innernode. Checking the posterior probabilities of the nodes on majority-consensus tree is only one way to summarize the posterior distributions of species tree topologies. The distribution of posterior probabilities across topologies, as well as the size of 95% credible set of trees—the set of topologies whose summed probability reach 0.95—are informative in terms of assessing whether a dataset has enough information to distinguish alternative tree topologies. Multiple independent runs of *BEAST was examined to ensure that the posteriors of species-tree topologies converged given the generations of MCMC chain. With the NGS data, the number of topologies in the 95% credible set decreased, and there are much larger differences between the first and the second most likely topologies (Figure 5.5d-f). Including loci of low variation increases the number of loci by almost 1/3 (33 loci versus 25 loci), but flattens the posterior distribution (Figure 5.5e).

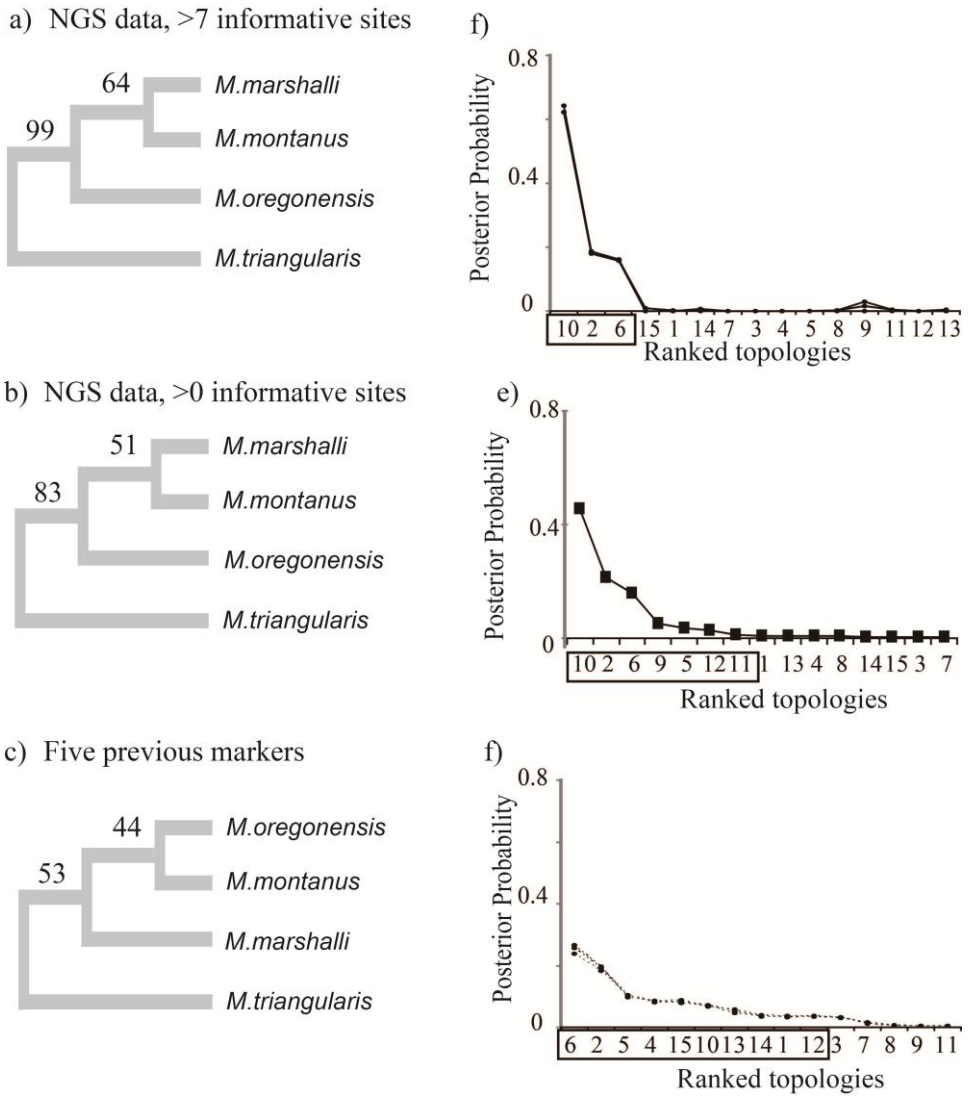


Figure 5.5 Species trees estimated by *BEAST.
Majority-consensus tree given by a) NGS data without loci of low variation; b) b) NGS data with loci of low variation and c) five markers developed in previous studies. Numbers are the posterior probabilities of the trees' innernodes. d) –f) Corresponding posterior probability of different species-tree topologies. Overlayed lines are from independent *BEAST runs. Topologies are ranked according to their posterior probabilities and the topologies in box are in the 95% credible set.

Applying the "twice" filter—requiring the minor allele to be found in at least two reads in each species— almost halved the number of variable sites (Table 5.2). For abbreviation, the dataset gone through the twice filter is referred to as "T dataset", and the unfiltered dataset is referred to as "N dataset". This reduction in the number of variable sites is not limited to subsets of contigs, leaving only nine contigs with more than seven informative sites. The distribution of posterior probabilities are flat across different topologies, eleven of which are in the 95% credible set. There are also dramatic changes in the estimated species-tree

topology—*M.triangularis* is no longer the most distantly related species in this group; rather, its divergence from *M.oregoensis* becomes the most recent speciation events in this group, but both nodes on the tree have low posterior probabilities (Figure 5.6).

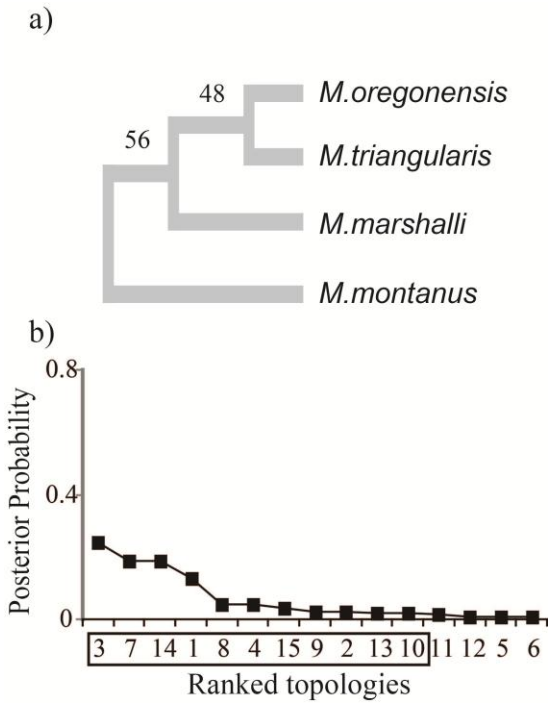


Figure 5.6 Estimated species tree posterior probability distributions.
The estimated species tree with twice-filtered NGS dataset and the posterior probability distribution across different species-tree topologies. Numbers on branch representing the posterior probabilities of innernodes and the topologies in the box are those in 95% credible sets.

Parametric simulation

Given that there were as many as twenty-five highly informative loci in the NGS dataset, why the second innernode on the species tree does not have high nodal support? Is it because the divergent history itself is too recent to resolve, or because the amount of errors—sequencing, assembly and haplotype-estimation errors—in the NGS data, which obscures the phylogenetic signal in the data? Parametric simulation approach was adopted to address this question. All ten replicates identify *M.triangularis* as the out-group species, but the nodal support for the more recent innernode has a wide range (Figure 5.7). This suggests that the species divergent history itself partially determines the difficulty in species-tree estimation.

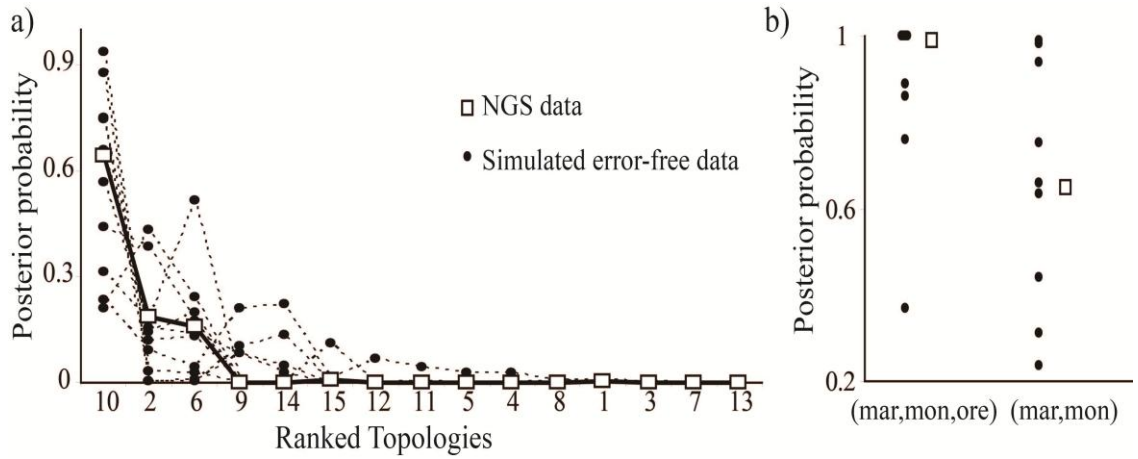


Figure 5.7 Posterior probabilities.
The posterior probabilities of a) fifteen different species-tree topologies and b) species-tree innernodes obtained from simulated error-free data (filled circles), comparing with those obtained from empirical NGS data (open squares).

Another question is why a stringent filter, expected to be eliminating more sequencing errors, change the species-tree estimates? Datasets with errors were simulated by randomly assign errors to the error-free sequences simulated in the previous step. With a known species tree, the probability of topological change caused by additional filter can be assessed. The result showed many differences in terms of estimated species trees between the simulated T and N datasets (Figure 5.8). First, the 95% credible sets of topologies are much larger with T datasets than N datasets, implying loss of phylogenetic signals with a more stringent filter (Figure 5.8a). Second, there is more uncertainty regarding which species or clade will be the most distantly related group on the majority-rule consensus tree (Figure 5.8b). In two out of the ten replicates, *M.montanus* becomes the out-group as observed in the empirical T dataset.

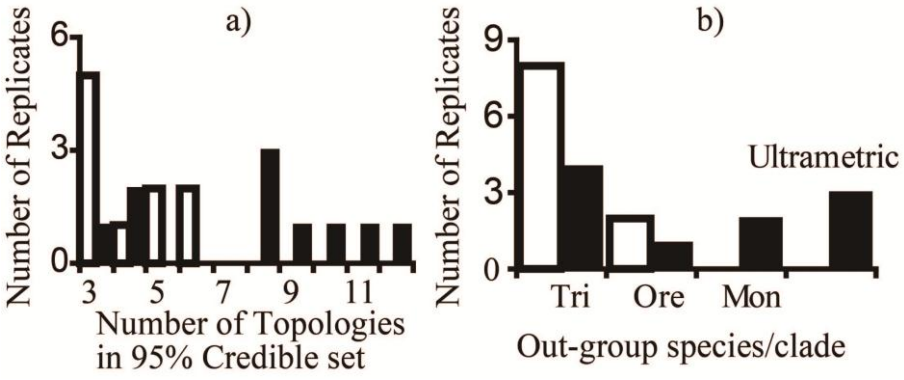


Figure 5.8 Simulated dataset comparisons.
Comparisons between the simulated N datasets and T datasets. a) The number of topologies in 95% credible sets; b) the most distantly related species or clade on the majority-rule consensus species trees. Open bars represent results from simulated N datasets, and filled bars are results from simulated T datasets.

DISCUSSION

In this study, NGS data was directly used for species-tree estimation for four grasshopper species, and the estimates from *BEAST with/without loci of low variability and additional error filter were compared. The majority-rule consensus species tree seems to be robust to loci with low information content, but adding these loci does not help distinguishing the most likely tree topology from other topologies. Strikingly, an additional filter changed the estimated species topology. Parametric simulation was used to separately examine the effects of true species divergent history, NGS sequencing errors and counting-based filters on species-tree estimation. These effects are discussed in different sections, followed by a discussion of possible improvements with better experimental design and data analyzing methods.

The shallow divergent history

The true species divergent history itself is one of the primary factors that can influence the accuracy of species-tree estimates. As shown in many previous simulation studies, the accuracy of species-tree estimates improves quickly when the depth of the true species tree increases (Maddison and Knowles 2006; McCormack et al. 2009; Huang et al. 2010). Moreover, the accuracy is also correlated with the topology of the true tree. Even when the total tree depth and the number of species are fixed, trees with a more even distribution of internal branch lengths and more symmetric topologies are easier to reconstruct from genetic data (McCormack et al. 2009). The divergent history of the four grasshopper species investigated in this paper belongs to the most difficult scenarios. The estimated total species-tree depth is around ~0.6*Ne* generations (*Ne* the effective population size) regardless of the filters and whether low variant contigs were used. Given the estimated effective population sizes in a previous study (~ 10^6, Knowles 2001), this suggested that the first speciation event in this species group is around 0.6 Mya. Furthermore, the topologies of the estimated species trees are asymmetric in this study, as well as previous studies (Knowles 2000; Carstens and Knowles 2007), which means shorter intervals between speciation events comparing to a symmetric species tree. The rough estimation of speciation rate (~ 200,000 years per speciation event) falls into the range of previous estimates of the whole *Melanoplus* genus given by a mitochondrial gene (10,000—200,000 yrs per speciation, Knowles and Otte 2000).

The fast rate of diversification is comparable to the extreme rates used in simulation studies (seven speciation events in 1*Ne* generations, McCormack et al. 2009; Huang et al. 2010). This makes the conclusion of this paper comparable to previous conclusions derived from simulated data. According to simulation results, the accuracy of maximal likelihood species-tree estimates, as well as the estimates based on the criterion minimal deep coalescent events, is low even when sampling fifty loci (Huang et al. 2010). In this study, with twenty five highly variable loci extracted from NGS data, the quality of species-tree estimates greatly improved compared to only using five previously developed markers, as shown by the larger difference in posterior probability between the most likely topology to other topologies and the shrinked 95% credible set of topologies (Figure 5.5). Yet, there is still uncertainty associated with the more recent internode on the tree. Our parametric bootstrap simulation showed difficulty in obtaining high nodal support for this internode even without any NGS related errors, indicating the difficulty of obtaining well-supported species-tree estimations for extreme shallow histories.

The effects of NGS errors

To filter sequencing errors, we developed a probabilistic model to obtain a joint estimation of genotypes and haplotypes from NGS raw data. Most of the species-tree estimation methods are gene-tree based, which requires haplotype sequences as input. However, majority of error-correction methods are geared towards extracting SNP data. Not only are these methods not ideal for phylogenetic analysis (at least not for methods that are gene-tree based; see Holboth et al. (2007) for an example of phylogenetic estimation based on SNPs), but they also ignore the information harbored in the linkage patterns among variable sites for evaluating potential sequencing errors. The idea of using haplotype estimation as methods for error correction has been proposed elsewhere (method *ShoRAH*, Zagordi et al. 2011) and experiments showed that haplotype reconstruction is more precise than simple counting-base calling method (Zagordi 2010). However, these methods and tests were developed for estimating haplotype frequencies of pooled samples. Here we modified the model because with only one specimen per species sampled, we do not have more than two haplotypes, and therefore the probability of haplotypes was calculated with a prior of population genetic diversity.

NGS data with sequencing errors was simulated to test the robustness of species-tree estimates to errors. Comparing to the error-free datasets, the number of simulated replicates with a correctly estimated majority-rule consensus species trees decreased from eight to five. Another three replicates with errors identified the correct outgroup species. Although the number of replication (10) is too small to assess the statistical significance, this means that the species-tree estimates are robust to the errors to certain extent, especially these dataset were simulated with an error rate that is 10 times of the empirical rate. In the simulated dataset, the exact number of true variable sites and sequencing errors can be traced by comparing the error-free sequences to the sequences with errors after haplotype estimation. Given the amount of partial reads in dataset, only around 80% of the true variable sites and sites with shared polymorphism across species will be preserved in the simulated datasets (Figure 5.9a). Even though sequencing errors generate around 12% spurious variable sites, but on average only lead to 3.8 (1.7%) spurious sites with shared polymorphism (Figure 5.9b). Sites with shared polymorphisms were crucial to infer the incomplete lineage sorting among species. The negligible proportion of spurious sites with shared polymorphism could be the reason why estimated species-tree topology is robust to the unfiltered errors.

The effects of stringent counting-based filters

Why we applying the twice-filter (requiring the minor allele to be found in at least two reads in each species) change the topology of the estimated species tree? An additional filter could remove more sequencing errors from NGS dataset. Even though some true variable sites might be excluded along with errors, there has to be systematic bias among species to change the estimated topology. Using the simulated dataset, the exact number of true variable sites and sequencing errors can be traced by comparing the error-free sequences to the estimated haplotypes with errors. Applying the twice filter reduces the proportion of spurious variable sites by 5%, but excludes another 24% true variable sites and 42% true shared-polymorphic sites (Figure 5.9a&b), which might explain why the twice filter excludes more sequencing errors but reduced the phylogenetic signal in the dataset. Moreover, the twice filter preferentially discards variable sites in *M.oregonensis* and *M.triangularis* (Figure 5.9c&d), and the proportion of kept true variable sites is significantly associated with the number of reads in each contigs only when the twice filter is applied (Figure 5.9e&f). Given

that *M.montane* and *M.marshalli* has almost twice as many reads (146 and 157 respectively) than the other two species (86 for *M.oregonensis* and 84 from *M.triangularis*), the relative proportion of variable sites in *M.oregonenesis* and *M.triangularis* considerably decreased after applying the twice filter – only 19% and 15% of variable sites of *M.oregonenesis* and *M.triangularis* pass the twice filter respectively, while 31% variable sites pass the additional filter for *M.montane* and *M.marshalli*.

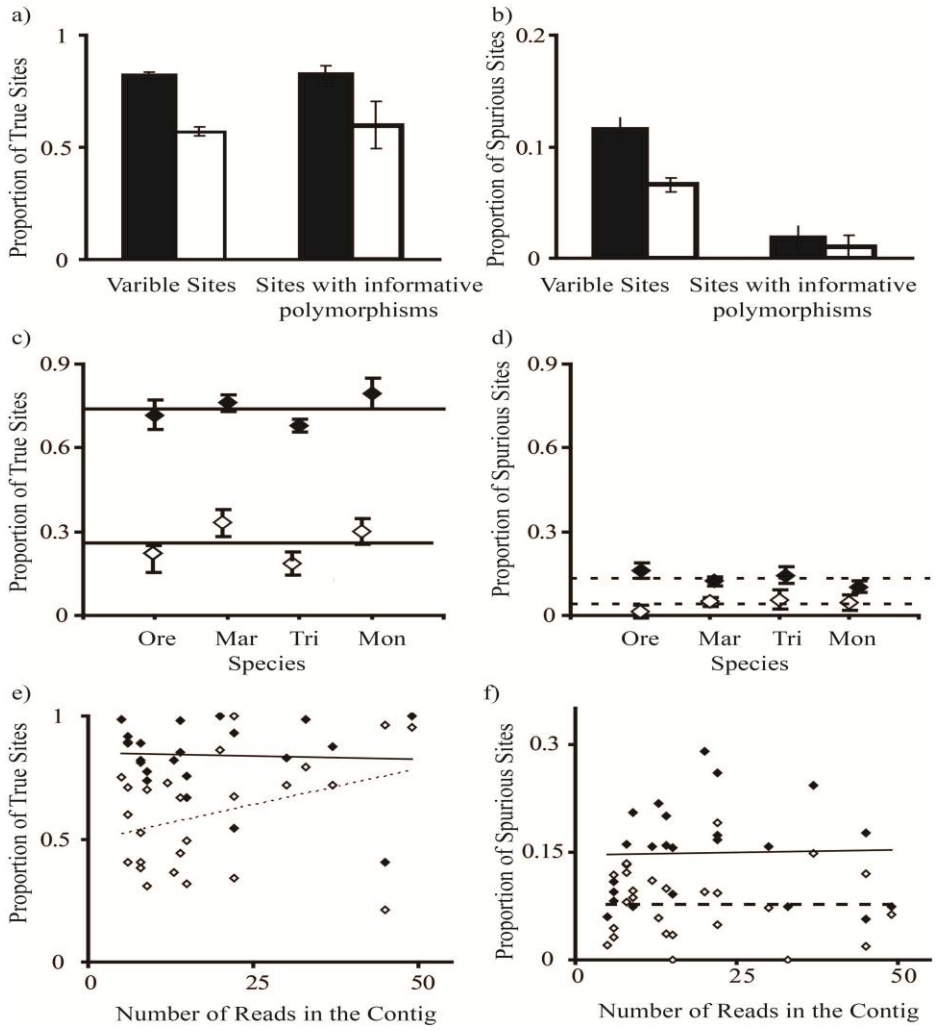


Figure 5.9 Kept and spurious site proportions.
The proportions of kept true sites and the proportions of spurious sites for simulated N and T datasets. a) and b) Overall proportions for variable sites and sites with informative polymorphisms, lines indicate the average across species; c) and d) the proportions for each species; e) and f) the correlation between the proportions of sites in contigs with the number of reads, lines indicate the linear regression lines. Filled bars and points represents proportions calculated with simulated N datasets, while open bars and points are for T datasets. Error bars indicate the standard deviation of the ten replicates.

Both the simulated and empirical data suggests that applying a stringent counting-based filter for errors could not only weaken the phylogenetic signal in NGS datasets, but also produce systematic bias when the number of reads is uneven among species. A strict filter could preferentially remove variable sites in species with smaller number of reads, which might lead to inaccuracies with species-tree estimates.

Caveats and future direcions

While twenty-five highly variable and phylogenetic informative loci were successfully obtained for the four grasshopper species with 1/13 of a 454 sequencing run, and the estimated species tree were shown to be robust to errors in NGS data to certain extent, this study could be improved with better experimental designs, and future studies would benefit from advances in data analyzing methods.

Experimental designs that consider the random shot-gun sequencing feature of NGS and RRL techniques are important. Dataset generated by each NGS run has considerable uncertainty. There is no precise control over which region would be sequenced, as well as the exact number of species in which a locus would be sequenced. How many loci can be obtained from a run would depend on coverage—a higher coverage means that each piece of digested genomic DNA from RRL has more chance of being sequenced, then more pieces would have orthologous copies sequenced in different species, resulting more usable contigs for phylogenetic analysis. Yet, most of the non-model organisms do not have whole genome sequenced. Estimating the number of digested genomic DNA that would be in the selected length range is difficult. Some of the non-model organisms can use genomic sequence of closely related species, but for species groups such as *Melanoplus*, assumptions about the genomic nucleotide composition would be needed. Moreover, bias in lab work (e.g., the amplification bias in both pre-NGS PCR and in NGS runs) and in data processing steps (e.g., eliminating contigs with high coverage) could easily increase or decrease the number of usable contigs. In retrospect, this study would be improved by increasing the coverage (on average 4.7 reads per species for the twenty-five locus). Higher coverage not only leads to more loci, but also more accurate estimation of the genotypes and haplotypes—not only a point estimation can be obtained, but also the confidence in the estimation can be assessed.

This could be achieved by only using one restriction enzyme or using enzymes with longer recognition sequences to reduce the number of digested genomic pieces.

The problem of modeling the coverage would be even more important for larger species trees. More species means the coverage needs to be higher to have the same number of loci sequenced in all the species. Although many species tree estimation methods could handle missing data (i.e., STEM, Kubatko et al. 2009), the effects of missing data on species-tree estimation still awaits further investigation (Kubatko and Pearl 2011). Moreover, for NGS dataset generate from RRL, finding orthologous loci relies on the same cutting sites of restriction enzymes across species. Species that are more distantly related are more likely to have mutations at the enzyme's cutting sites, resulting in missing data (McCormack et al., in press; Althoff et al. 2007). Therefore the missing pattern itself is informative about the species relationship. Future studies are needed to explore the effects of these informative missings.

Comparing the sampling design to previous simulation studies (reviewed in Knowles 2010) also point out a caveat that worth considering for future studies. As shown in many simulation studies, increasing the number of sampled individuals is a more efficient way of increasing the accuracy of species-tree estimation for shallow species trees (e.g., Maddison and Knowles 2006). Because only one specimen per species was used, the maximum number of chromosomes sampled per species is two. This restricts the accuracy of species-tree estimation. However, adding more individuals per species means higher cost in labeling individuals in the library preparation step, which could be even more costly than the sequencing itself (Glenn 2011). More individuals also mean lower coverage per individuals, which would lower the number of orthologous loci. One solution would be pooling multiple individuals from the same species together (Cutler and Jensen 2010), and extracting the haplotype sequences from the mixed reads. However, the effect of this approach on species-tree estimation is not clear. Moreover, as shown in our result, uneven coverage among different species has profound interaction with data processing step (i.e., error correction). With pooled individuals, slight difference in concentration, caused by difference in genomic quality or library preparation steps, could be manifested, and how to keep comparable coverage among species would be a challenge. Hence, the pros and cons of not having individually labeled sequences still need more investigation.

NGS combined with RRL techniques condensed the previous lengthy process of finding and amplifying multiple loci into a time-efficient process, but orthologous loci can only be determined according to assembly. Errors in assembly can violate the assumption of independent locus in most of the species-tree estimation methods in both directions. Contigs that actually have reads from paralogous loci would have many spurious incomplete lineage sorting, which could not be modeled by the coalescent process. In addition, pseudo independent loci will be generated if reads from the same locus are assembled into multiple contigs. In this study, reads with repetitive sequences were precluded from assembly, contigs were reassembled and only those with distinct consensus sequence were kept to avoid to the second type of error. To reduce the chance of including paralogous loci, a maxium coverage is assumed in the assembly step. Nevertheless, preferential amplifications in PCR and NGS might lead to overrepresentation of some amplicons in the result. That is, the strategy of eliminating high coverage contigs adopted in this study and others (Emerson et al., 2010; Gompert et al., 2010) might be too conservation. A method that could give a measure of the probability of being paralogs given error rate, locus coverage, and configurations of alleles at variable sites would be helpful for future studies.

Another aspect that needs improvement is a better estimate of the error probability. The quality of the reads from NGS is not directly comparable to the reads given by Sanger sequencing. In NGS with RRL, each read is an imperfect representation of one PCR product, which might have PCR errors itself; while a read given by Sanger sequencing is an average across many reactions on many PCR products. Although NGS platforms offer quality scores along each read, the correlation between these scores to the actually error rates is less well studied. Here the known sequences of adaptors were used to estimate the error rate and examine the relationship between reported quality score with the actual error rate. The overall error rate is much higher than the one calculated from quality scores, and the quality scores has very weak correlation with the rate of point errors (Figure 5.3a). Therefore, these provided quality scores might be misleading (i.e., high quality scores are not always corresponding to high reliability). Yet, as the techniques of NGS is quickly upgrading in recent years, along with the progress in base-calling algorithms (see Ledergerber 2011 for review), future studies would benefit from better estimations of base-specific error rates.

CONCLUSIONS

With the paradigm shift to species-tree estimation from gene-tree estimation, the task of empirical phylogenetic studies is no longer finding enough informative sites to reconstruct a highly supported gene tree, but rather, finding enough independent loci to account for the stochasticity in lineage inheritance. The combination of multiplexed NGS with RRL (reduced representation library) has the potential to obtain multiple loci from multiple species almost instantaneously with reasonable costs even in non-model organisms (Ekblom and Galindo 2010). In recent years, this approach has been applied to various organisms (e.g., Van Tassell et al., 2008; Wiedmann et al., 2008; Amaral et al., 2009; Kerstens et al., 2009; Sanchez et al., 2009). Yet the use of the NGS datasets is mostly limited to marker development or analyzing the extracted SNPs. In this study, we show that despite the errors in NGS data, such data are amendable for estimating species trees with coalescent-based methods. Loci with low information content will not necessarily change the estimated species tree, but could flatten the posterior distributions. Stringent filters, which are commonly applied to NGS data to avoid false-positive calling of variable sites, however, could generate systematic bias that alter the estimated species tree when the coverage significantly differs among species. Consequently, phylogenetic studies with NGS data should guard against possible bias introduced by error correction methods. While the species-tree estimates are robust to sequencing errors, many aspects of data processing still needs to be improved, including better estimates of base-specific error probability and probabilistic modeling to distinguish orthologous versus paralogous loci. Simulation studies are needed for investigating the effects of missing data, and the trade-off of having pooled individuals. Lastly, progress in both analytical methods and experimental design are needed to improve the utility of this new sequencing platform for phylogenetic studies in non-model organisms.

REFERENCES

Alkan, C., J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs, and E. E. Eichler. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics 41:1061-U29.

Althoff, D. M., M. A. Gitzendanner, and K. A. Segraves. 2007. The utility of amplified fragment length polymorphisms in phylogenetics: a comparison of homology within and between genomes. Syst Biol 56:477-84.

Amaral, A. J., H. J. Megens, H. H. D. Kerstens, H. C. M. Heuven, B. Dibbits, R. P. M. A. Crooijmans, J. T. Den Dunnen, and M. A. M. Groenen. 2009. Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. Bmc Genomics 10.

Ane, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24:412-426.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27:573-580.

Binladen, J., M. T. Gilbert, J. P. Bollback, F. Panitz, C. Bendixen, R. Nielsen, and E. Willerslev. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2:e197.

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

Cutler, D. J., and J. D. Jensen. 2010. To pool, or not to pool? Genetics 186:41-3.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499-510.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. Plos Genetics 2:762-768.

Ekblom, R., and J. Galindo. 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1-15.

Emerson, K. J., C. R. Merz, J. M. Catchen, P. A. Hohenlohe, W. A. Cresko, W. E. Bradshaw, and C. M. Holzapfel. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc Natl Acad Sci U S A 107:16196-200.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87-112.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. Mol Ecol Resour 11:759-69.

Gompert, Z., M. L. Forister, J. A. Fordyce, C. C. Nice, R. J. Williamson, and C. A. Buerkle. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. Molecular Ecology 19:2455-2473.

Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol 27:570-80.

Hellmann, I., Y. Mang, Z. P. Gu, P. Li, F. M. de la Vega, A. G. Clark, and R. Nielsen. 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Research 18:1020-1029.

Hobolth A, Christensen OF, Mailund T, Schierup MH, 2007 Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. PLoS Genet 3(2): e7. doi:10.1371/journal.pgen.0030007

Huang, H. T., Q. I. He, L. S. Kubatko, and L. L. Knowles. 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, X. Q., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Research 9:868-877.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-8.

Huse, S. M., J. A. Huber, H. G. Morrison, M. L. Sogin, and D. Mark Welch. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8.

Hyten, D. L., S. B. Cannon, Q. J. Song, N. Weeks, E. W. Fickus, R. C. Shoemaker, J. E. Specht, A. D. Farmer, G. D. May, and P. B. Cregan. 2010. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. Bmc Genomics 11.

Kerstens, H. H., R. P. Crooijmans, A. Veenendaal, B. W. Dibbits, A. W. T. F. Chin, J. T. den Dunnen, and M. A. Groenen. 2009. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. Bmc Genomics 10:479.

Knowles, L. L. 2000. Tests of Pleistocene speciation in montane grasshoppers (genus *Melanoplus*) from the sky islands of western North America. Evolution 54:1337-1348.

Knowles, L. L. 2001. Genealogical portraits of speciation in montane grasshoppers (genus Melanoplus) from the sky islands of the Rocky Mountains. Proceedings of the Royal Society of London Series B-Biological Sciences 268:319-324.

Knowles, L. L. 2010 Sampling strategies for species tree estimation. In L. Knowles & L.Kubatko (Eds.), Estimating species trees: *practical* and theoretical aspects （pp. 163-172）. Hoboken, New Jersy: John Wiley &Sons, Inc.

Knowles, L. L., and D. Otte. 2000. Phylogenetic analysis of montane grasshoppers from western North America (Genus *Melanoplus*, Acrididae : Melanoplinae). Annals of the Entomological Society of America 93:421-431.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971-973.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology 56:17-24.

Kubatko, L. S., and D. K. Pearl. 2011. Seeing the Trees in Your Terrace. Science 333:411-412.

Ledergerber, C., and C. Dessimoz. 2011. Base-calling for next-generation sequencing platforms. Brief Bioinform 12:489-97.

Liu, L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542-3.

Liu, L., L. Yu, and D. K. Pearl. 2010. Maximum tree: a consistent estimator of the species tree. J Math Biol 60:95-106.

Luca, F., R. R. Hudson, D. B. Witonsky, and A. Di Rienzo. 2011. A reduced representation approach to population genetic analyses and applications to human evolution. Genome Res 21:1087-98.

Lynch, M. 2008. Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects. Molecular Biology and Evolution 25:2409-2419.

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

Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet 24:133-41.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. T. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. G. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

McCormack, J. P., H. Huang, and L. L. Knowles. 2009. Maximum-likelihood Estimates of Species Trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Systematic Biology (2009) 58 (5): 501-508.

McCormack, J. P., J. M. Maley, S. M. Hird, Elizabeth P. Derryberry, Gary R. Graves, and R. T. Brumfield. 2011(In press).. Next-generation sequencing reveals phylogeographic 1 structure and a species tree for recent bird divergences. Molecular Phylogenetics and Evolution

Meyer, M., U. Stenzel, and M. Hofreiter. 2008. Parallel tagged sequencing on the 454 platform. Nat Protoc 3:267-78.

Mossel, E., and S. Roch. 2007. Incomplete Lineage Sorting: Consistent Phylogeny Estimation From Multiple Loci. arXiv:0710.0262v2.

Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen. 2010. Population genetic inference from genomic sequence variation. Genome Res 20:291-300.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13:235-8.

Sanchez, C. C., T. P. Smith, R. T. Wiedmann, R. L. Vallejo, M. Salem, J. Yao, and C. E. Rexroad, 3rd. 2009. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. Bmc Genomics 10:559.

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5:247-252.

Wiedmann, R. T., T. P. Smith, and D. J. Nonneman. 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet 9:81.

Williams, L. M., X. Ma, A. R. Boyko, C. D. Bustamante, and M. F. Oleksiak. 2010. SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genet 11:32.

Zagordi, O., A. Bhattacharya, N. Eriksson, and N. Beerenwinkel. 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. Bmc Bioinformatics 12.

Zagordi, O., R. Klein, M. Daumer, and N. Beerenwinkel. 2010. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Research 38:7400-7409.

Chapter 6 Molecular evidence of a peripatric origin for two sympatric species of field crickets (*Gryllus rubens* and *G. texensis*) revealed from coalescent simulations and population genetic tests

Research on species pairs with slight morphological differences poses intriguing questions about the origin and maintenance of species distinctiveness. What are the reproductive isolating factors and what is the geographic context under which those barriers evolved? Much of the focus on speciation in cryptic species has been on the former, with an emphasis on experiments that examine the consequences of differences in mating signals. For example, empirical work now complements theoretical models that rely upon strong assortative mating for species divergence in sympatry involving strong divergent sexual selection (e.g., Turner & Burrows 1995; Payne & Krakauer 1997; Higashi et al. 1999; Higgie et al. 2000; Takimoto et al. 2000) or a by-product of ecological divergence (e.g., Doebeli et al. 2005; Vines & Schluter 2006; Duffy et al. 2007; Gavrilets et al. 2007). However, understanding the geographic context of species divergence is critical to establishing what initiated divergence, and specifically, whether the contemporary sympatry of species faithfully reflects how the barrier evolved (Perret et al. 2007).

Phylogeographic study provides the context for deciphering that geographic history (Avise 2000), but evaluation of alternative explanations for observed patterns of genetic variation requires a framework where the impact of different processes can be considered (Knowles 2004; Petit 2007). This task becomes especially challenging for recent divergence (Wakeley 2003), for which a pattern of incomplete lineage sorting might reflect the retention of ancestral polymorphism or gene flow (e.g., Kliman et al. 2002; Knowles 2001; Masta & Maddison 2002; Buckley et al. 2006; Carstens & Knowles 2007; Linnen & Farrell 2007; Richards & Knowles 2007; Peters et al. 2007), and statistical approaches for directly computing the likelihood of these alternative scenarios (e.g., Hey 2005) are rendered

unsuitable by their simplifying assumptions (Voight et al. 2005; Leaché et al. 2007; Knowles & Carstens, 2007; Fagundes et al. 2007).

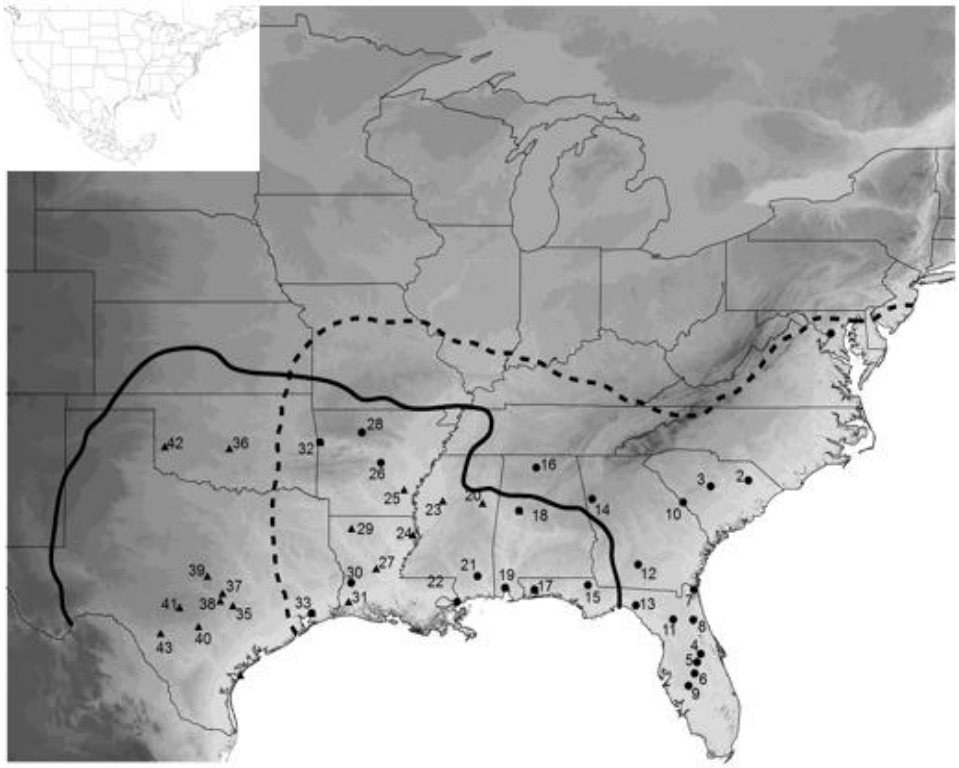


Figure 6.1 Distribution and range of sampled populations.
G. texensis and *G. rubens* population distributions identified by triangles and circles, respectively, and approximate range of *G. texensis* and *G. rubens* delimited with solid and dashed-line, respectively.

Here we use molecular data to address the biogeography of speciation in a cryptic pair of sister species of field crickets, *Gryllus rubens* and *Gryllus texensis*. These crickets are distributed across the southern US gulf states, with *G. rubens* ranging from Florida and the southern Atlantic states westward to eastern Texas and *G. texensis* ranging from central-western Texas eastward across the southern gulf states to far western Florida (Figure 6.1). Thus each species is broadly sympatric from western Florida to eastern Texas and also has a sizable area of allopatry. Prior work with these species has revealed (1) morphological divergence in females but not males (Gray et al. 2001), (2) strong divergence in the long-range male calling song used to attract females for mating (Walker 1998; Gray & Cade 2000; Walker 2000; Izzo & Gray 2004) with no evidence of reproductive character displacement (Gray & Cade 2000; Izzo & Gray 2004), (3) female preference for conspecific calling song and for conspecific close-range courtship song (Gray & Cade 2000; Gray 2005), and (4) in *G.*

texensis, heritable genetic variation in both male calling song and female preference for song, two genetically correlated traits (Gray & Cade 2000). Such conditions are predicted to give rise to reproductive isolation via the assortative mating that results from the rapid co-evolution of male signals and female preferences (Lande 1981; West-Eberhard 1983; Higashi et al. 1999). Previous molecular work (Harrison 1979; Gray et al. 2006) had suggested that *G. rubens* harbors relatively little genetic variation, and that *G. rubens* and *G. texensis* mitochondrial DNA sequences produce a paraphyletic gene tree (based on mitochondrial sequences from a sample of 20 individuals). However, tests of the historical biogeographic context of divergence were limited by insufficient sampling. Here we dramatically increase the scale of sampling to provide tests of (1) whether the lack of reciprocal monophyly reflects gene flow between the species, and (2) whether the genealogical structure supports an peripatric mode of speciation involving the partitioning and retention of ancestral variation in the descendant taxa *G. rubens* and *G. texensis*. These hypotheses are tested using a combination of biogeographic analyses and coalescent simulations we devised for exploring specific historical scenarios relevant to the origin of these taxa.

MATERIALS AND METHODS

Throughout the range of *G. texensis* and *G. rubens*, 48 populations were sampled for a total of 177 individuals from 25 populations of *G. rubens* and 188 individuals from 23 populations of *G. texensis*. Species identity was confirmed by analysis of male calling song (for male specimens) or by a combination of female ovipositor length and analysis of the calling songs of laboratory reared sons (for female specimens). Together these characters are diagnostic of species identity with little to no overlap (Gray & Cade 2000; Gray et al. 2001; Izzo & Gray 2004). A 724 bp fragment of the mitochondrial gene *Cytochrome Oxidase C subunit I* (COI) was amplified using primers C1-J-2183 and TL2-N-3014 (Simon et al. 1994). Amplification was by Polymerase Chain Reaction (30 cycles, annealing temperature 52 C). Negative controls were employed with each reaction. Sequencing was done on an ABI Prism 377 DNA Sequencer platform with BigDye v.3.1 chemistry. Consensus sequences for each sample were obtained by manual alignment of forward and reverse sequences using BioEdit (Hall 1999).

Data analyses

Standard measures of population genetic diversity were performed with programs DnaSP (Rozas et al. 2003), including measures of haplotype diversity, the average number of pairwise differences per nucleotide site, π (Nei 1987), and Tajima's D as a measure of historical demography assuming COI is evolving neutrally (Tajima 1989). The population structure within each species was examined using the IBD (isolation by distance) program (Jensen et al. 2005). A maximum likelihood gene tree was generated using PAUP* 4.0b10 with midpoint rooting (Swofford 2002) with the PaupUp graphical interface (Calendini & Martin 2005). An analysis of molecular variance (AMOVA) using Arlequin (Schneider et al. 2000) was used to estimate genetic differentiation, including the proportion of genetic variance attributable to different hierarchical levels (i.e., between species, within species between populations and within populations).

To investigate gene flow between the species, populations were designated as either sympatric or allopatric according to collection location. The amount of genetic divergence (D; average number of substitutions per site) between allopatric and sympatric populations was calculated with DnaSP (Rozas et al. 2003). If gene flow contributed to the distribution of *G. rubens* among *G. texensis* haplotypes in the gene tree, then genetic distance between allopatric *G. texensis* and *G. rubens* would be expected to be greater than sympatric *G. texensis* and *G. rubens* (i.e., gene flow would erode any genetic differences in sympatry, whereas genetic differences would be maintained in allopatry where there is no opportunity for gene flow). An AMOVA was also used to determine whether geography contributed significantly to patterns of genetic differentiation i.e., an AMOVA with three levels: between allopatric and sympatric groups, within groups between populations, and within populations.

Figure 6.2 Gene tree of COI alleles

Alleles from G. rubens and G.texensis estimated by maximum likelihood, with midpoint rooting and a model of evolution estimated from the data (HKY +I + □□□□□A = 0.314, □C = 0.219, □G = 0.136, □T = 0.331; ti/tv = 6.35; PINV = 0.769; □ = 0.155). The different colored vertical bars indicate the location of alleles from G. rubens and G.texensis cluster I and II. Each haplotype label indicates, in order: the species (marked as either r, t, or rt for G. rubens, G. texensis, and both species, respectively), the geographic location of the haplotype identified by a population number (see Fig. 1 for distribution of populations), and haplotype number. Haplotypes that are distributed across multiple populations are identified with each of the respective populations; three haplotypes (rt-1-1, rt-2-1 and rt4-1) occurred in numerous populations.

For recently diverged species, shared polymorphism might result from ancestral lineage sorting or gene flow. While these two factors might in principle be distinguished with a robust estimate of migration, in our study, the complexities of the species history precluded such an approach (i.e., the data did not fit the assumptions of the population genetic models employed in the program IM; Hey 2005, e.g., lack of convergence of the posterior probability distributions indicated that migration estimates were not reliable for this dataset, nor were estimates of the time of divergence). Instead we employed coalescent simulations to explore whether the structure in the mitochondrial gene tree reflects the biogeography of species divergence. Specifically, we used two separate analyses to test whether (a) the differences in the levels of lineage sorting observed between *G. rubens* and each of two genetic clusters within *G. texensis* (referred to here after as *G. texensis* I and *G. texensis* II – see Figure 6.2), and (b) the specific genealogical structure within *G. texensis* I and *G. texensis* II was informative about how the species diverged. We use different summary statistics, each summarizing different aspects of genetic variation in the data, to explore the history of divergence of the species. While summary statistics do not utilize fully all the information contained in DNA sequences, they nonetheless provide a computationally tractable framework with demonstrated utility for exploring demographic and biogeographic scenarios (Knowles & Maddison 2002; Voight et al. 2005; Hickerson et al. 2006; Fagundes et al. 2007). We use a variety of summary statistics (described in detail in the following sections), each summarizing different aspects of genetic variation in the data, to maximize the information content for the given data (i.e., multiple summaries of the data capture more information than any single summary statistic).

Calculation and evaluation of the summary statistics dwII/dwI and exII/exI

Coalescent simulations and summary statistics were used to determine whether the unusual genetic structure within *G. texensis* – two genetic clusters: one with comb-like structure and the other with long external branches – was consistent with a single panmictic population that was suggested by the geographical overlap of individuals from the two genetic groups (i.e., *G. texensis* I and *G. texensis* II, Figure 6.2). Genealogies were simulated using the program **MS** (Hudson 2002). For each simulated genealogy, the average of pair-wise distances within clusters, dw_I and dw_{II} for cluster I and cluster II, respectively (Figure

6.3), and the ratio (dw_{II} / dw_{I}) were calculated. Since this ratio might be affected by how many haplotypes are distributed in two clusters, the analyses were constrained to simulated genealogies with similar proportions of haplotypes as observed in the empirical data (i.e., 24 haplotypes in the smaller cluster), where the two genetic clusters were identified using a root that corresponded to the deepest coalescent time between lineages and the larger of the two groups was designated as cluster I (as observed in the empirical data); because of the correspondence between the number of nucleotide differences between sequences and the order of coalescences (Takahata and Nei 1985) this rooting scheme is consistent the with midpoint rooting used for the empirical data. The summary statistics were calculated on 1000 replicate genealogies to generate an expected distribution for the ratio (dw_{II} / dw_{I}); the hypothesis that the geographically overlapping genetic groups in *G. texensis* reflects a history without past structure would be rejected if the observed value (i.e., the ratio calculated for the empirical data) exceeds the values observed in 95% of the simulated data (i.e., $P < 0.05$ under the null model). This test is not sensitive to assumptions about effective population size, N_e, since different population sizes affect the total depth of a gene tree (i.e., the time to coalescence), but not the shape of the genealogy – the relevant feature in testing whether the empirical data departs from expectations for the ratio dw_{II} / dw_{I}.

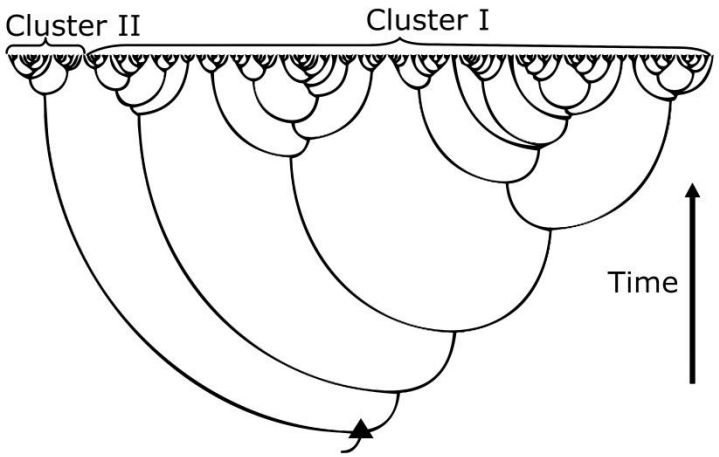


Figure 6.3 An example simulated genealogy.
An example genealogy simulated by a neutral coalescent showing the two genetic clusters (corresponding to *G. texensis* I and II, Figure 6.2) for which summary statistics were calculated (see text for details); the black triangle indicates the root of this genealogy.

The robustness of the conclusions to different demographic histories was also evaluated by conducting the simulations with changes in population size. Four different demographic

scenarios with a range of rates of expansion and decreases in population size were considered (see Table 6.4 for details), and include a model of (a) exponential increase, (b) exponential decrease, and bottlenecks involving a (c) 10-fold and (d) 100-fold decrease in population size. Since the shape of genealogies generated under scenarios involving increases and decreases in population size differ from those under constant population size (Wakeley 2003), a second statistic was used to evaluate the probability of the gene tree structure in the empirical data under models of population expansion and bottlenecks. Moreover, the use of multiple summary statistics provides more statistic power for statistical phylogeographic tests (Knowles 2004; Voight et al. 2005). This additional statistic (ex_{II}/ex_{I}) is the ratio of average external branch lengths (i.e., the length of singleton branches) for each of the two genetic clusters in *G. texensis* (see Figure 6.3). The numbers of genealogies with both higher dw_{II}/dw_{I} and higher ex_{II}/ex_{I} than the empirical data were recorded for the 1000 genealogies simulated for each historical scenario (Table 6.4), where the alpha-level for significance was determined with a Bonferroni correction because of the multiple tests conducted for each model.

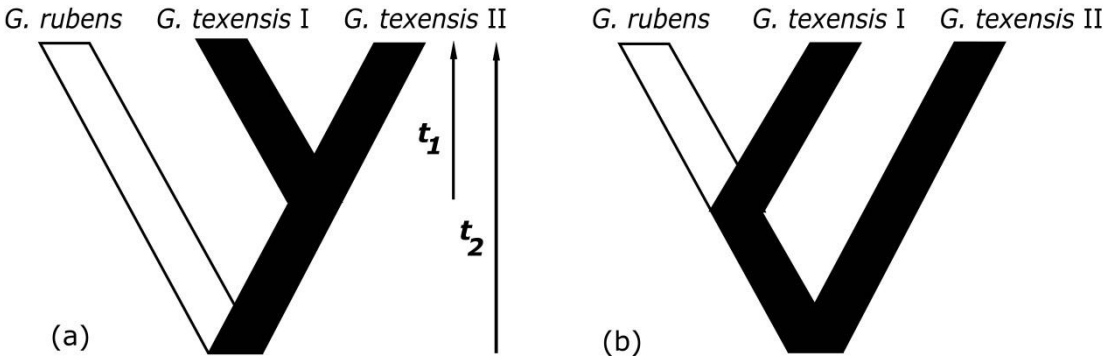


Figure 6.4 The historical substructure within G. texensis lineages.
The historical substructure may have (a) occurred after the divergence of *G. texensis* and *G. rubens*, or (b) predate the divergence of *G. rubens*, indicating that *G. rubens* originated from a subset of variation in a subdivided ancestor.

Calculation and evaluation of the summary statistics drtII / drtI

An additional coalescent analysis was performed to address whether the genealogical split observed within *G. texensis* (a) occurred after the divergence between *G. texensis* and *G. rubens*, or (b) may reflect a historical substructure in which *G. rubens* was derived from a subset of the ancestral variation present in *G. texensis* (i.e., by parapatric speciation; Harrison 1991). This test involves computing a ratio of the genetic distance between haplotypes of *G.*

rubens to *G. texensis* I (d_{rtI}) and *G. texensis II* (d_{rtII}). This ratio (i.e. d_{rtII} / d_{rtI}) would be greater than one if *G. texensis* I shares a more recent common ancestor with *G. rubens* than *G. texensis* II (i.e., the genetic distance between *G. rubens* and *G. texensis* I haplotypes is expected to be less than that between haplotypes of *G. rubens* and *G. texensis II*; Figure 6.4b), whereas d_{rtII} / d_{rtI} is expected to be equal to one, on average, if the historical substructure did not play a role in the origin of *G. rubens* (Figure 6.4a). To determine whether a d_{rtII} / d_{rtI} = 6.07 (the value for the empirical data) differs significantly from what is expected if the historical substructure did not play a role in the origin of *G. rubens* (i.e., whether the model depicted in Figure 6.4a could be statistically rejected), an expected distribution for the range of d_{rtII} / d_{rtI} values was generated from simulated data that takes into account the variance on the expectation arising from the stochasticity of genetic drift. The sample sizes used in the coalescent simulations matched those from the empirical gene tree (i.e., 177, 164, and 24 for *G. rubens*, *G. texensis* I and *G. texensis* II, respectively), and the population sizes were all scaled to the same value given similar estimates of θ of the three populations (Table 6.1). Without a reliable estimate of when the substructure occurred (t_1 in Figure 6.4a), a conservative test with $t_1 = t_2$ was used; this test of the null hypothesis is conservative because d_{rtI} and d_{rtII} is expected to be more similar (i.e. d_{rtII} / d_{rtI} will approach one) as the time of divergence between the two *G. texensis* lineages from a common ancestor is shorter (i.e., $t_1 < t_2$, Figure 6.4a).

Given that *G. rubens* shows the genetic signature of population expansion (Table 6.1), the coalescent simulations were conducted for both a constant of changing population size for *G. rubens*. Specifically, a model of exponential population growth was considered (i.e., $N_t = N_e e^{(-\alpha \times t/4Ne)}$, where N_t is the population size t generations ago and N_e is the current population size) with rates of change (i.e., an $\alpha = 1$, $\alpha = 4$, and $\alpha = 7$). The robustness of these results were examined over a range of differing divergence times of divergence. 1000 genealogies each were simulated under a range of divergence times of $0.5N_e$ to $4N_e$ at 0.5 intervals; only relatively recent divergence times were considered since they are the conditions in which non-monophyly of the species is expected. With an N_e of 5×10^5, this translates into divergence times ranging 0.25 Mya to 2 Mya, with one generation per year. It is worth noting that these conditions also encompass a range of different population sizes, for a given

divergence time (measured in N generations). For example, the results also scale to a population size that ranges from 250,000 to 2 million for a divergence time of 0.5 Mya.

RESULTS

Nucleotide polymorphism

Of the 365 sequenced individuals, 170 haplotypes were identified: 27 haplotypes from the 177 individuals of *G. rubens* and 164 haplotypes from the 188 individuals of *G. texensis* sampled (Table 6.1). An AMOVA detected significant divergence between species (

Table 6.2). The estimated gene tree suggests the species are not reciprocally monophyletic (Figure 6.2), and two unusual features characterize the gene tree. First, there are two genetic clusters within *G. texensis* that differ in structure, namely a portion of the *G. texensis* I cluster includes a comb-like section with very closely related haplotypes compared to the relatively longer internal branches of *G. texensis* II (Figure 6.2). Second, most of the *G. rubens* haplotypes are nested within *G. texensis*, but are primarily limited to just the *G. texensis* I cluster (all but one of the *G. rubens* haplotypes that nest within *G. texensis* occur within the *G. texensis* I cluster). This structure may reflect the geography of divergence in which *G. rubens* was derived from a subset of a *G. texensis* like ancestor (see Figure 6.4) or may reflect gene flow between *G. rubens* and *G. texensis* (see results for each hypothesis below).

With respect to the demographic history of each species, there was no evidence indicating that *G. texensis* had experienced a recent expansion (Table 6.1), contrasting with the demographic history of *G. rubens*. Likewise, there was no relationship between genetic distance and the geographic distribution of individuals in *G. texensis,* whereas significant isolation by distance was detected in *G. rubens* ($r = 0.243$, $P < 0.005$; the IBD test). Despite these apparent differences in the demographic history between *G. rubens* and *G. texensis,* estimates of genetic diversity were similar between the species based on π and θ (Table 6.1), indicating similar effective population sizes (i.e., π and $\theta = 4N_e\mu$; Tajima 1983); population-level estimates of diversity does differ between *G. texensis* and *G. rubens*, possibly reflecting the demographic expansion detected in *G. rubens* (Table 6.1). This genetic signature of population growth in *G. rubens* is incorporated into the coalescent simulations to avoid

misinterpretations based on inappropriate assumptions about constant population size, if *G. rubens* had indeed undergone a change in population size (for thoroughness, the possibility of changes in the population size of *G. texensis* was also considered, Table 6.4). While it is not possible to definitively rule out a selective sweep as causing the significant negative Tajima's *D* (Table 6.1), the context for such selection being limited to just one of these two species is not obvious, especially given their geographic overlap in large portions of their ranges (Figure 6.1) and ecological similarity.

Tests of gene flow between species

Two separate biogeographic analyses indicate that recent gene flow is most likely not the underlying cause for the lack of reciprocal monophyly of *G. rubens* and *G. texensis*. First, if gene flow was homogenizing *G. texensis* I and *G. rubens*, *G. texensis* I (but not *G. texensis* II) is expected to be codistributed with *G. rubens*, since the *G. texensis* II cluster is generally distinct from *G. rubens* (i.e., haplotype mixing occurs between *G. rubens* and *G. texensis* I; Figure 6.2). However, the individuals from the two *G. texensis* genetic clusters are broadly overlapping geographically. Moreover, there was no significant differentiation between sympatric and allopatric *G. texensis* (Table 6.3) as expected if the genetic composition of *G. texensis* I reflected gene flow with *G. rubens*. Comparison of the genetic distance between *G. rubens* and sympatric versus allopatric *G. texensis* also showed a pattern contrary to that expected under a hypothesis of introgressive gene flow. If gene flow was the underlying cause for the lack of reciprocal monophyly, the genetic distance between sympatric *G. rubens* and *G. texensis* should be smaller than the genetic distance between sympatric *G. rubens* and allopatric *G. texensis*. In our data, however, the former is significantly larger than the later (0.0129 versus 0.0099, $p < 0.01$, Z-test). In summary, each of the tests - the geographic distribution of haplotypes, the genetic variation within the two *G. texensis* clusters, and the distances between *G. rubens* and sympatric versus allopatric *G. texensis* – indicate that recent gene flow is not a likely explanation for the lack of reciprocal monophyly between the species.

Table 6.1 Description of genetic variation in *G. rubens* and *G. texensis*.
Shown, from left to right, are the sample sizes (n), the number of segregating sites (s), and the number of haplotypes (k), Waterson's theta (θ_w) and nucleotide diversity (π) (population averages are also shown in parentheses for each species), and the values of Tajima's D and Fu and Li's D and F (significant values are marked with an asterisk).

	n	s	K	π	θ	Tajima's D	P	Fu and Li's D	P	Fu and Li's F	P
Gryllus rubens	177	42	27	0.00176 (0.001338)	0.01009 (0.00168)	-2.45	<0.01*	-6.79	<0.02*	-5.94	<0.02*
Gryllus texensis	188	49	147	0.01459 (0.012274)	0.01188 (0.0124)	0.68	>0.10	-0.36	>0.10	0.11	>0.10

Table 6.2 Analysis of molecular variance (AMOVA) of the data of the two species.
Populations are grouped by species, and F-statistics are computed from haplotype frequency.

Source of Variation	d.f.	Sum of squares	Variance Components	*F*-statistics	% total	*P*-value
Among Species	1	9.341	0.048	F_{ct} = 0.10	10.22	<0.0001
Among populations within species	46	22.242	0.009	F_{sc} = 0.02	1.90	<0.0001
within populations	317	131.985	0.416	F_{st} = 0.12	87.88	<0.0001
total	364	163.567	0.4738			

Table 6.3 Analysis of molecular variance (AMOVA).
Analysis of the Gryllus texensis where the sympatry versus allopatry of the G. texensis with G. rubens was used to group populations (i.e., the among groups term).

Source of Variation	d.f.	Sum of squares	Variance Components	*F*-statistics	% total	*P*-value
Among sympatric versus allopatric *G. texensis*	1	0.537	0.00014	F_{ct} = 0.00029	0.03	>0.291
Among populations within groups	21	10.818	0.00313	F_{sc} = 0.00633	0.63	>0.0635
within populations	165	80.911	0.49037	F_{st} = 0.00662	99.34	<0.0635
total	187	92.266	0.49364			

Table 6.4 Statistical models of genealogical structure.
Test of whether the genealogical structure observed in G. texensis (Figure 6.2) is probable under different models of population size change. These models include: (a) a constant population size, (b) population expansion with differing amounts of size change (i.e., different α-values, under a model of exponential change Nt = Nee-α t/4Ne, where Nt is the population size t generations ago and Ne is the current population size), (c) exponential decreases in population size with different rates of decrease (i.e., different α-values; a constant the population size was assumed prior to 16Ne generations to avoid the problem of infinite time waiting for lineage coalescence), and a population bottleneck in which the contemporary population is either (d) 10 times or (e) 100 times smaller than the ancestral population size, for a range of different bottleneck times (t). The data is not consistent with most models (i.e., the values for the simulated data are higher than those for the empirical data, specifically dwII / dwI = 2.05 and exII / exI = 2.59). Those models that can be statistically rejected are shown in bold, after a Bonferroni correction for multiple tests for each model (i.e., P > 0.025).

Demographic Models	dw_{II} / dw_I	dw_{II} / dw_I and ex_{II} / ex_I
(a) constant population size	**0.009**	**0.001**
(b) exponential change on population size expansion		
α = 0.5	**0.008**	**0.003**
α = 1	**0.004**	**0.001**
α = 4	**0.003**	**0.001**
α = 7	**0.001**	**0.000**

(c) exponential decline in population size		
α = -0.5	**0.012**	**0.003**
α = -1	**0.020**	**0.006**
α = -4	0.072	**0.013**
α = -7	0.044	**0.006**
(d) 10-fold decrease in population size		
t = 0.5 N_e	0.043	**0.008**
t = 1 N_e	0.058	**0.007**
t = 2 N_e	**0.023**	**0.004**
t = 3 N_e	**0.014**	**0.003**
t = 4 N_e	**0.006**	**0.001**
(d) 100-fold decrease in population size		
t = 0.5 N_e	0.030	**0.007**
t = 1 N_e	0.085	**0.013**
t = 2 N_e	0.053	**0.007**
t = 3 N_e	**0.010**	**0.003**
t = 4 N_e	**0.013**	**0.003**

Tests that the gene tree structure reflects the biogeography of species divergence

To determine whether the genealogical structure of two genetic clusters within *G. texensis* (i.e., *G. texensis* I and *G. texensis* II) is indicative of ancestral substructure, as opposed to a single species lineage, coalescent simulations were used to evaluate the probability of observing a ratio of dw_{II} / dw_I $\geq$ the observed empirical ratio of average pairwise distance between cluster I and II (see Figure 6.3). Out of 1000 simulated genealogies with the same distribution of haplotypes across two clusters, less than 5% (p= 0.009) of the genealogies exhibited a dw_{II} / dw_I $\geq$ the observed empirical ratio of 2.05 (Figure 6.5), indicating that the genealogical structure is inconsistent with the unstructured species lineage. Coalescent simulations also confirm that these results are

generally robust to changes in population size (even though there was no significant genetic signature of population expansion in *G. texensis*, Table 6.1). The probability of observing a gene tree with the structure observed in the empirical data is very low under both models of exponential increase and decrease, as well as population bottlenecks (Table 6.4). A few of the models cannot be rejected after adjusting the level of significance for multiple comparisons involving different rates of expansion or different times for the bottleneck based on a characterization of the gene tree involving just one summary statistic (i.e., dw_{II} / dw_I). However, when multiple aspects of the gene tree structure are considered jointly (i.e., dw_{II} / dw_I and ex_{II} / ex_I) everyone of the different scenarios of population change are significantly rejected – that is, less than 1.3 % of the genealogies exhibited characteristics observed in the empirical data (i.e., dw_{II} / $dw_I \geq$ 2.05 and ex_{II} / $ex_I \geq$ 2.59)(Table 6.4). These results all indicate that it is highly unlikely that the two genetic clusters observed in *G. texensis* could have been derived from an unstructured species lineage, suggesting instead that despite the contemporary overlap of *G. texensis* I and *G. texensis* II, in the past the *G. texensis* lineage was subdivided.

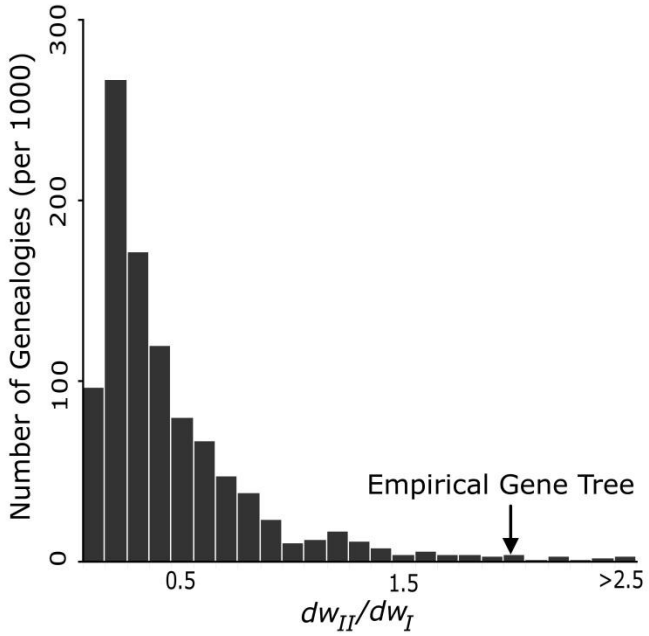


Figure 6.5 Lineage model frequency distributions.
Frequency distribution of dwII/dwI generated by a single lineage model in G. texensis, where dwI and dwII are the average pair-wise distances within genetic clusters I and II (see Figure 6.3). Note that the deviation of the expected distribution from a general expectation of 1 is because the test involved simulated genealogies where the distribution of haplotype numbers between the two genetic clusters also matched the empirical data.

To determine whether such historical substructure might have played a role in speciation (Figure 6.4), the timing of this substructure relative to the origin of *G. texensis* was investigated with coalescent simulations. The observed average genetic distance between *G. rubens* and *G. texensis* I, d_{rtI}, relative to that between haplotypes of *G. rubens* and *G. texensis II*, d_{rtII}, is significantly greater than null expectation as generated from the coalescent simulations (Figure 6.6). This suggests that the structure in the observed gene tree (Figure 6.2) is more consistent with a history in which *G. rubens* and *G. texensis I* shared a more recent common ancestor (i.e., supports the model in Figure 6.4b) than did *G. texensis I* and *G. texensis II* (i.e., the model in Figure 6.4a is rejected). This result is consistent across the range of divergence times considered, and is robust to possible changes in population size in *G. rubens* (Figure 6.6).

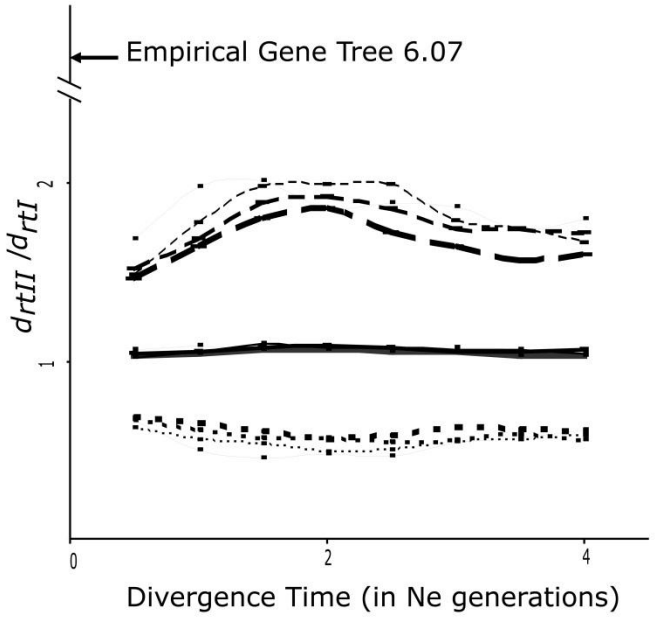


Figure 6.6 Genetic distance.
The average genetic distance between G. rubens and G. texensis I, drtI, relative to that between haplotypes of G. rubens and G. texensis II, drtII, is significantly greater than expected under a model where the historical substructure played no role in speciation (i.e., the model in Figure 6.4a can be rejected). The solid line represents the mean of drtII / drtI (i.e., the average genetic distance between G. rubens and G. texensis I, drtI, relative to that between haplotypes of G. rubens and G. texensis II, drtII, at different divergence times for a range of demographic conditions. These models include a constant population size and exponential population growth for G. rubens, Nt = Nee-α t/4Ne, where Nt is the population size t generations ago and Ne is the current population size, and α represents different rates of change. The different rates of population change are identified by different line widths ranging from the thickest to thinnest lines, for α = 1, 4, and 7, respectively. Note that the mean expectations (shown by solid lines) overlap between the constant population size and those involving population expansion (i.e., a drtII / drtI of

approximately 1); the dashed lines represent the 95% confidence interval for each demographic model. The arrow indicates the drtII / drtI value calculated from empirical data.

DISCUSSION

While multiple processes might underlie the lack of reciprocal monophyly between taxa, when such messy gene trees are examined in a predictive framework they can yield valuable insights about species divergence (Knowles 2004). In this study, the data clearly show that *G. rubens* and *G. texensis* are very closely related, corroborating behavioral data (Gray & Cade 2000; Izzo & Gray 2004; Gray 2005), but that they are not reciprocally monophyletic based on the mitochondrial gene tree, as suggested by previous molecular studies with limited sampling (Harrison 1979; Gray et al. 2006). However, by considering how the processes involved in speciation would affect the geographic distribution of haplotypes, as well as the patterns of relationships among haplotypes (i.e., the gene tree structure), we were able to statistically explore different hypotheses about how divergence in these cricket species might have proceeded. The sister taxa, *G. rubens* and *G. texensis*, show very different phylogeographic patterns. *G. texensis* has an abundance of singleton haplotypes (Table 6.1), and shows no evidence of isolation by distance. *G. rubens* appears to have undergone a recent population expansion (Table 6.1), and shows significant isolation by distance among subpopulations. While the genetic variation within *G. texensis* is characterized by an abundance of singleton haplotypes, the data are not consistent with a simple demographic model of expansion (i.e., non-significant Fu and Li's D and Tajima's D; Table 6.1). The coalescent simulations also indicate that an unstructured population, even considering an array of different demographic scenarios involving increases or decreases in population size, is significantly unlikely (Table 6.4). Instead, the history of this species appears to be quite complex. *G. texensis* shows no significant isolation by distance; rather, the haplotypes are distributed into two broadly distributed and geographically overlapping clusters (*G. texensis I* and *G. texensis II*; Figure 6.2). Despite the co-distribution of haplotypes between the two *G. texensis* clusters, and their geographic overlap with haplotypes from *G. rubens*, the gene tree shows that all but one of the *G. rubens* haplotypes (including four shared haplotypes) are limited to just one lineage of *G. texensis* (what we are calling the *G. texensis I* cluster; Figure 6.2).

This complicated genealogical structure, with haplotypes of *G. rubens* mixed with haplotypes from *G. texensis I* but not *G. texensis II*, could arise from gene flow or reflect the incomplete sorting of gene lineages from a common ancestor. However, the series of tests used to evaluate these possibilities suggest that the polyphyletic structure of the gene tree reflects the geographic history of divergence and not gene flow (see details below). The complexity of these species histories no doubt also contributes to a serious violation of the assumption the divergence models used by typical coalescent-based likelihood or Bayesian approaches for estimating population genetic parameters (reviewed in Excoffier & Heckel 2006), thus precluding their use in this study. This unusual structure, particularly within *G. texensis*, motivated the use of the novel alternative analyses presented here (see also Knowles & Maddison 2002; Fagundes et al. 2007). However, only a small subset of historical scenarios was evaluated, as with any statistical phylogeographic study because of the enormous space of potential histories (e.g., varying population sizes and migration rates that might differ between species and change among populations and over the species history) and simplifying assumptions might affect the results (Knowles & Maddison 2002). Without additional data (i.e., multiple loci), it also is not possible to evaluate the extent to which the patterns of differentiation observed in the mitochondrial sequences are an accurate reflection of the species' histories. Despite these caveats, the approach and combination of test that were devised identify a biologically interesting model of species divergence – namely, peripatric speciation (e.g., Harrison 1991; Knowles et al. 1999). Under this model, past geographic substructure may have contributed to the origin of the cricket species, even though such regional division is not apparent today (e.g., the two *G. texensis* genetic clusters are co-distributed across the species range). The implications of this model, including how these findings might motivate future potential studies, as well as supporting evidence from comparison with other taxa in the region, are discussed below.

The role of historical regional substructure in species divergence

There is no behavioral, morphological or other evidence whatsoever suggesting the existence of two cryptic species within the currently recognized *G. texensis*, or significant differentiation between populations of *G. texensis* that are distributed sympatrically

versus allopatrically with respect to *G. rubens* (Table 6.3). This indicates that the general lack of *G. rubens* haplotypes within the *G. texensis II* cluster does not reflect some differential gene flow owing to reproductive isolation. Despite a number of undescribed cryptic species within North American *Gryllus* (unpublished data, D. Weissman and D. Gray), the unimodal distribution of intra-specific variation in reproductive characters (e.g., song and female preference traits; Gray & Cade 2000, 1999; Izzo & Gray 2004)) and lack or correspondence with the two mitochondrial genetic clusters indicates a lack of cryptic species in *G. texensis*. Moreover, the statistical phylogeographic test, which revealed that the degree of genetic differentiation between sympatric *G. texensis* and *G. rubens* was greater (not smaller as expected) than the genetic distance between allopatric *G. texensis* and *G. rubens*, also indicates that recent gene flow is not a sufficient explanation for the lack of reciprocal monophyly between the species, although low levels of past or present hybridization are certainly not precluded by these analyses.

The genealogical structure is consistent with two population lineages of *G. texensis* with incomplete sorting of ancestral polymorphisms between this subdivided ancestor and the more recently derived species *G. rubens* (Figure 6.4b). The apparently large population sizes of these crickets are indeed consistent with this observation; tens or even hundreds of thousands of these crickets engage in eruptive flights every summer and fall (Cade 1979) and the total numbers of *G. texensis* could easily be several million or more (W. H. Cade, personal communication), greatly reducing the rate of lineage sorting. Furthermore, our simulations also show support for the origination of *G. rubens* from one of the two ancestral lineages of *G. texensis* (Figure 6.6). Such a scenario would be consistent with a geographic scenario in which proto-*G. rubens* became geographically isolated, perhaps in peninsular Florida during the climate-induced distributional shifts caused by the Pleistocene glacial cycles. Our data indicating recent population expansion of *G. rubens* combined with significant isolation-by-distance further supports this model. Such regional substructure has been documented in other taxa from the area (e.g., Avise & Smith 1974), including both plants and animal species (see comparative phylogeographic review of 148 taxa, Soltis et al. 2006). In the absence of additional loci in the crickets for genetic analysis, these comparative studies identify the plausibility of the model of species divergence proposed by our study (i.e., the observed patterns of

differentiation in COI are not simply a reflection of the historical dynamics of the mitochondrial genome owing to a discordance between the observed gene tree and the actual species history, Maddison 1997). Nonetheless, multi-locus data are not only important for testing the proposed peripatric model of divergence, but it would also be very helpful for deriving accurate population genetic parameter estimates from coalescent-based likelihood or Bayesian approaches (reviewed in Excoffier & Heckel 2006) and provide a more biologically realistic model of divergence.

In contrast to the distinct biogeographic patterns that mirror these past discontinuities in many of the taxa from the general study area (e.g., regions divided by the Apalachicola River, Appalachian Mountains, Tombigbee River, and the Mississippi River; reviewed in Soltis et al. 2006), no such correspondence between geography and genetic divergence is apparent in *G. texensis*. While any regional substructure that might have existed in the historical past of *G. texensis* (Fig. 4b) apparently has been eroded by migration, a lack of evidence for recent gene flow between sympatric *G. texensis* and *G. rubens* suggests that divergence in behavioral traits are an effective reproductive barrier (Gray 2005; Gray & Cade 2000). However, this will need to be confirmed with additional genetic data. Interestingly, genetic evidence suggest that species boundaries are quite porous in other related and recently derived *Gryllus* taxa (Broughton & Harrison 2003), suggesting that gene flow homogenizes species differences except for those characters for which divergence is maintained by selection. Further investigations into the demography of speciation of *G. texensis* and *G. rubens* will provide an important context for identifying whether the mode of speciation actually differs among these recently derived North American gryllines (i.e., divergence with or without gene flow) and what it is about the species-specific traits that confer a more (or less) effective reproductive barrier to gene flow.

REFERENCES

Avise, J. C. 2000. *Phylogeography: the history and formation of species*. Cambridge, Massachusetts: Harvard University Press.

Avise, J. C. & Smith, M. H. 1974. Biochemical genetics of sunfish. I. geographic variation and subspecies intergradation in the bluegill, *Lepomis macrochirus*. *Evolution*, **28**, 42-56.

Broughton, R. E. & Harrison, R. G. 2003. Nuclear gene genealogies reveal historical, demographic and selective factors associated with speciation in field crickets. *Genetics*,**163**, 1389-1401.

Buckley, T., M. Cordeiro, D. Marshall, and C. Simon. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (Maoricicada Dugdale). Systematic Biology, **55**, 411-425.

Cade, W. H. 1979. Field cricket dispersal flights measured by crickets landing at lights. *Texas Journal of Science*, **XXXI,** 125-130.

Calendini, F. & Martin, J.-F. 2005. PaupUP v1.0.2032.22590 Beta. A free graphical frontend for Paup* Dos software.

Carstens, B.C. & Knowles, L.L. 2007. Shifting distributions and speciation: species divergence during rapid climate change. *Molecular Ecology*, **16**, 619-627

Duffy, S., Burch, C. L. & Turner, P. E. 2007. Evolution of host specificty drives reproductive isolation among RNA viruses. *Evolution*, **61**, 2614-2622.

Doebeli, M., Dieckman, U., Metz, J. A. J. & Tautz, D. 2005. What we have also learned: adaptive speciation in theoretically plausible. *Evolution*, **59**, 691-695.

Excoffier L. & G. Heckel. 2006. Computer programs for population genetics data analysis: a survival guide. *Nature Review of Genetics* **7**, 745-758.

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., & Excoffier, L. 2007. Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences, USA*, **104,** 17614-17619.

Gavrilets, S., Vose, A., Barluenga, M., Salzburger, W. & Meyer, A. 2007. Case studies and mathematical models of ecological speciation. 1. cichlids in a crater lake. *Molecular Ecology*, **16**, 2893-2909.

Gray, D. A. 2005. Does courtship behavior contribute to species-level reproductive isolation in field crickets? *Behavioral Ecology*, **16,** 201-206.

Gray, D. A., Barnfield, P., Seifried, M. & Ritchards, M. R. 2006. Molecular divergence between*Gryllus rubens* and *Gryllus texensis*, sister species of field crickets (Orthoptera: Gryllidae). *Canadian Entomologist*, **138,** 305-313.

Gray, D. A. & Cade, W. H. 2000. Sexual selection and speciation in field crickets. *Proceedings of the National Academy of Sciences, USA*, **97,** 14449-14454.

Gray, D. A. & Cade, W. H. 1999. Quantitative genetics of sexual selection in the field cricket, *Gryllus integer*. *Evolution*, **53**, 848-854.

Gray, D. A., Walker, T. J., Conley, B. E. & Cade, W. H. 2001. A morphological means of distinguishing females of the cryptic field cricket species, *Gryllus rubens* and *G. texensis* (Orthoptera: Gryllidae). *Florida Entomologist*, **84,** 314-315.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.*, **41,** 95-98.

Harrison, R. G. 1991. Molecular-Changes at Speciation. A*nnual Review of Ecology and Systematics*, **22**, 281-308.

Harrison, R. G. 1979. Speciation in North American field crickets: evidence from electrophoretic comparsons. *Evolution*, **33,** 1009-1023.

Hey, J. 2005. The number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biology, **36**, e193.

Hickerson, M., Dolman, G., and C. Mortiz. 2006. Comparative phylogeographic summary statistics for testing simultaneous vicariance. *Molecular Ecology*, **15**, 209-223

Higashi, M., Takimoto, G. & Yamamura, N. 1999. Sympatric speciation by sexual selection. *Nature (London)*, **402,** 523-526.

Higgie, M., Chenoweth, S. & Blows, M. W. 2000. Natural selection and the reinforcement of mate recognition. *Science, Wash.*, **290,** 519-521.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337-338.

Izzo, A. S. & Gray, D. A. 2004. Cricket song in sympatry: species specificity of song without reproductive character displacement in *Gryllus rubens*. *Annals of the Entomological Society of America*, **97,** 831-837.

Jensen, J. L., Bohonak, A. J. & Kelley, S. T. 2005. Isolation by distance, web service. *BMC Genetics*, **6,** 13.

Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., McCarter, J., Wakeley, J. & Hey, J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics*, **156**, 1913-1931.

Knowles, L.L. 2001. Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshoppers. *Molecular Ecology*, **10,** 691-701.

Knowles, L. L. 2004. The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17,** 1-10.

Knowles, L. L. & W. P. Maddison. 2002. Statistical phylogeography. *Molecular Ecology*, **11**, 2623-2635.

Knowles, L. L., Futuyma, D. J., Eanes, W.F. & Rannala, B. 1999. Insights into speciation mode from historical demography in the phytophagous beetle Ophraella. *Evolution*, **53**, 1846-1856.

Knowles, L.L. & Carstens, B.C. 2007. Estimating a geographically explicit model of population divergence. *Evolution*, **61**, 477-493.

Lande, R. 1981. Models of speciation by sexual selection on polygenic traits. *Proceedings of the National Academy of Sciences, USA*, **78,** 3721-3725.

Leaché, A. D., Crews, S. C. & Hickerson, M. J. 2007. Two waves of diversification in mammals and reptiles of Baja California revealed by hierarchical Bayesian anlaysis. Biology Letters, 3, 646-650.

Linnen, C. R. & Farrell, B. D. 2007. Mitonuclear discordance is caused by rampant mitochondrial introgression in Neodiprion (Hymenoptera: Diprionidae) sawflies. *Evolution*, **61**, 1417-1438.

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

Masta, S. E. & Maddison, W. P. 2002. Sexual selection driving diversification in jumping spiders. *Proceedings of the National Academy of Sciences, USA*, **99,** 4442-4447.

Nei, M. 1987. *Molecular Evoluitonary Genetics*. New York, NY: Columbia University Press.

Payne, R. J. H. & Krakauer, D. C. 1997. Sexual selection, space and speciation. *Evolution*, **51,** 1-9.

Perret, M., Chautems, A., Spichiger, R., Barraclough, T. G. & Savolainen, V. 2007. The geographical pattern of speciation and floral diversification in the neotropics: the tribe Sinningieae (Gesneriaceae) as a case study. *Evolution*, **61**,1641-4660.

Peters, J. L., Zhuravlev, Y., Fefelov, I., Logie., A. & Omland, K. E. 2007. Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (*Anas* spp.). *Evolution*, **61**, 1992-2006.

Richards, C.L. & Knowles, L.L. 2007. Tests of phenotypic and genetic concordance and their application to the conservation of Panamanian golden frogs (Anura, Bufonidae). *Molecular Ecology*, **16**, 3119-3133.

Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19,** 2496-2497.

Schneider, S., Roessli, D. & Excoffier, L. 2000. *Arlequin: A software for population genetics data analysis. Ver 2.000*.: Genetics and Biometry Lab, Department of Anthropology, University of Geneva. http://lgb.unige.ch/arlequin/software/.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America*, **87,** 651-701.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, 15, 4261-4293.

Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland, Massachusetts: Sinaeur Associates.

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

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123,** 585-595.

Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, **110**, 325–344

Takimoto, G., Higashi, M. & Yamamura, N. 2000. A deterministic genetic model for sympatric speciation by sexual selection. *Evolution*, **54,** 1870-1881.

Turner, G. F. & Burrows, M. T. 1995. A model of sympatric speciation by sexual selection. *Proceedings of the Royal Society of London Series B*, **260,** 287-292.

Vines, T. H. & Schluter, D. 2006. Strong assortative mating between allopatric sticklebacks as a by-product of adaptation to different environments. *Proceedings of the Royal Society of London Series B*, **273,** 911-916.

Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R., and A. Di Rienzo. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences, USA*, **102,** 18508-18513.

Wakeley, J. 2003. Inferences about the structure and history of populations: coalescents and intraspecific phylogeography. Pp. 193-215 in R. Singh and M. Uyenoyama, eds. *The Evolution of Population Biology*. Cambridge University Press, Cambridge.

Walker, T. J. 1998. Trilling field crickets in a zone of overlap (Orthoptera: Gryllidae: *Gryllus*). *Annals of the Entomological Society of America*, **91,** 175-184.

Walker, T. J. 2000. Pulse rates in the songs of trilling field crickets (Orthoptera: Gryllidae: *Gryllus*). *Annals of the Entomological Society of America*, **93,** 565-572.

West-Eberhard, M. J. 1983. Sexual selection, social competition, and speciation. *Quarterly Review of Biology*, **58,** 155-182.