eLIFE elifesciences.org

RESEARCH ARTICLE

A multispecies coalescent model for quantitative traits

Fábio K Mendes[1]*, Jesualdo A Fuentes-González[1,2], Joshua G Schraiber[3,4,5], Matthew W Hahn[1,6]

[1]Department of Biology, Indiana University, Bloomington, United States; [2]School of Life Sciences, Arizona State University, Tempe, United States; [3]Department of Biology, Temple University, Philadelphia, United States; [4]Center for Computational Genetics and Genomics, Temple University, Philadelphia, United States; [5]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, United States; [6]Department of Computer Science, Indiana University, Bloomington, United States

Abstract We present a multispecies coalescent model for quantitative traits that allows for evolutionary inferences at micro- and macroevolutionary scales. A major advantage of this model is its ability to incorporate genealogical discordance underlying a quantitative trait. We show that discordance causes a decrease in the expected trait covariance between more closely related species relative to more distantly related species. If unaccounted for, this outcome can lead to an overestimation of a trait's evolutionary rate, to a decrease in its phylogenetic signal, and to errors when examining shifts in mean trait values. The number of loci controlling a quantitative trait appears to be irrelevant to all trends reported, and discordance also affected discrete, threshold traits. Our model and analyses point to the conditions under which different methods should fare better or worse, in addition to indicating current and future approaches that can mitigate the effects of discordance.

DOI: https://doi.org/10.7554/eLife.36482.001

***For correspondence:** fkmendes@indiana.edu

Competing interests: The authors declare that no competing interests exist.

Funding: See page 15

Received: 07 March 2018
Accepted: 02 July 2018
Published: 14 August 2018

Reviewing editor: Antonis Rokas, Vanderbilt University, United States

Introduction

Understanding how traits evolve through time is one of the major goals of phylogenetics. Phylogenetic inferences made about traits can include estimating a trait's evolutionary rate and ancestral states, determining whether the evolution of a trait is influenced by natural selection, and establishing whether certain character states make speciation and extinction more or less likely (***Harvey and Pagel, 1991***; ***O'Meara, 2012***; ***Garamszegi, 2014***). Despite the variety of questions one can ask, and the plethora of different discrete and continuous traits that can be studied, it has long been recognized that in order to make inferences about trait evolution it is imperative to consider how the species carrying these traits are related (***Felsenstein, 1985***). Phylogenetic comparative methods model traits as evolving along a phylogeny, and therefore often require one, or sometimes multiple, species trees as input (***Pagel, 1999***; ***O'Meara, 2012***; ***Hahn and Nakhleh, 2016***).

The unprecedented increase in the availability of molecular data has been a boon to the construction of well-supported species trees — that is, those with high levels of statistical support. Thanks to advances in sequencing technology, species trees are now denser, taller, and better resolved. In contrast to the high levels of support provided by genome-scale data, phylogenomic studies have also revealed topological discordance between gene trees to be pervasive across the tree of life (***Pollard et al., 2006***; ***White et al., 2009***; ***Hobolth et al., 2011***; ***Brawand et al., 2014***; ***Zhang et al., 2014***; ***Suh et al., 2015***; ***Pease et al., 2016***). Gene trees can disagree with one another and with the species tree because of technical reasons — for example, model misspecification, low phylogenetic

signal, or the mis-identification of paralogs as orthologs — but also as a result of biological phenomena such as incomplete lineage sorting (ILS), introgression, and horizontal gene transfer (**Maddison, 1997**). Among the latter, ILS is well-studied due to its conduciveness to mathematical characterization (**Hudson, 1983**; **Tajima, 1983**; **Pamilo and Nei, 1988**), in addition to being an inevitable result of population processes (**Edwards, 2009**). Going backwards in time, ILS is said to occur when lineages from the same population do not coalesce in that population, but instead coalesce in a more ancestral population. If these lineages then happen to coalesce first with others from more distantly related populations, the gene tree will be discordant with the species tree.

The realization that genealogical discordance is the rule rather than the exception (in species trees with short internal branches) has been followed closely by a growing awareness that failing to consider microevolutionary-scale processes in phylogenetic inferences can be problematic (**Kubatko and Degnan, 2007**; **Edwards, 2009**; **Mendes and Hahn, 2016**; **Mendes and Hahn, 2018**). In the context of species tree estimation, this has led to the development of popular methods that account for processes such as ILS and introgression (e.g. **Liu, 2008**; **Than et al., 2008**; **Liu et al., 2009**; **Heled and Drummond, 2010**; **Larget et al., 2010**; **Bryant et al., 2012**; **Mirarab and Warnow, 2015**; **Solís-Lemus and Ané, 2016**). However, the development of models incorporating discordance in order to deal with trait evolution have lagged behind those estimating species trees.

One way that gene tree discordance can affect inferences about trait evolution is by increasing the risk of hemiplasy. Hemiplasy is the production of a homoplasy-like pattern by a non-homoplastic event (**Avise and Robinson, 2008**), generally because a character-state transition has occurred on a discordant gene tree. Consider the example shown in **Figure 1**: trait one is underlain by a gene whose topology is discordant with the species tree; a single state transition occurs only once along the branch leading to the ancestor of species *A* and *C*. However, if one attempts to infer the history of transitions on the species tree, two spurious transitions (for instance, on branches leading to *A* and *C*) must be invoked. The same occurs with trait two (**Figure 1**), but on the other discordant gene tree. Unless the gene tree underlying a discrete trait is concordant with the species tree (such as trait three in **Figure 1**), ignoring its topology can lead one to believe that homoplasy has happened, when in fact it has not — this is due to hemiplasy.

Recent work on the relevance of gene tree discordance to phylogenetic inferences has demonstrated that hemiplasy is widespread and problematic. At the molecular level, hemiplasy can cause apparent substitution rate variation, can spuriously increase the detection of positive selection in coding sequences, and can lead to artefactual signals of convergence (**Mendes and Hahn, 2016**; **Mendes et al., 2016**). In datasets with high levels of gene tree discordance, the fraction of all substitutions that are hemiplastic can be quite high (**Copetti et al., 2017**). In contrast, the manner in which genealogical discordance might affect studies of complex trait evolution is still not well understood. While past work has investigated how the genetic architecture of complex traits interacts with genetic drift to influence patterns of variation between populations and species (**Lynch, 1988**, **Lynch, 1989**, **Lynch, 1994**; **Whitlock, 1999**; **Ovaskainen et al., 2011**; **Zhang et al., 2014**), an interesting and still unanswered question is whether genealogical discordance can have an effect on these traits.

As continuous traits are often underlain by a large number of loci, a significant fraction of them could have discordant gene trees in the presence of ILS or introgression. Trait-affecting substitutions on discordant internal branches (those that are absent from the species tree; **Mendes and Hahn, 2018**) of such trees would then increase the similarity in traits between more distantly related species, while decreasing that of more closely related species. Such an effect could consequently affect the inferences from phylogenetic comparative methods about these quantitative traits. On the other hand, the most frequent gene tree in a data set is generally expected to be concordant with the species tree (except in cases of anomalous gene trees; **Degnan and Rosenberg, 2006**). As a consequence, we might expect that the contribution to traits from loci with concordant gene trees would outweigh the signal introduced by loci with discordant gene trees, possibly making phylogenetic inferences about continuous traits more robust to hemiplasy relative to discrete traits. In other words, a reasonable hypothesis is that gene tree discordance should only be problematic for traits controlled by a small number of loci, but not for those controlled by many loci (**Hahn and Nakhleh, 2016**).

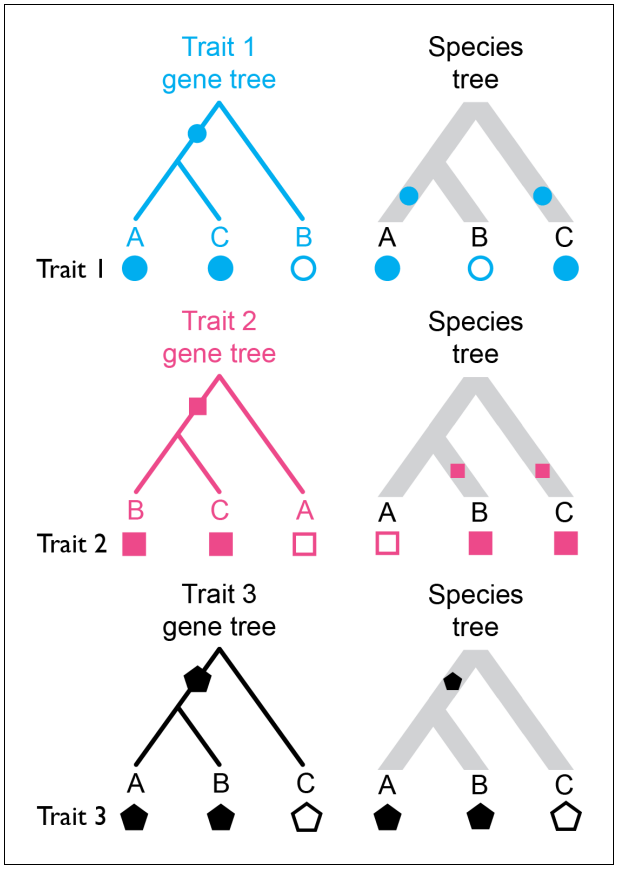


Figure 1. Three distinct discrete traits with their states mapped to the gene trees they evolved on, and to the species tree. Hollow and filled shapes represent the ancestral and derived states, respectively, with character-state transitions being indicated by filled shapes along internal branches. Traits 1 and 2 undergo a single character-state transition in their evolutionary history, but when the states are resolved on the species tree, a homoplasy-like (yet not truly homoplastic) pattern emerges (i.e. hemiplasy). Trait three has evolved along a gene tree that matches the species tree in topology, and so no hemiplasy occurs.

DOI: https://doi.org/10.7554/eLife.36482.002

Here, we present a model of quantitative traits evolving under the multispecies coalescent that accounts for gene tree discordance, deriving the expected variances and covariances in quantitative traits under this model. We then apply traditional phylogenetic comparative methods to data simulated under the coalescent framework in order to better understand the impact of discordance on phylogenetic inference. Our framework makes it possible to vary levels of ILS and the number of loci controlling a quantitative trait (cf. ***Schraiber and Landis, 2015***), and so we also address whether variation in genetic architecture has an effect on phylogenetic inference. Finally, we use the threshold model (***Wright, 1934***; ***Felsenstein, 2005***) to investigate whether discretizing quantitative traits makes inferences about them more or less robust to the potential effects of gene tree discordance and hemiplasy.

Characterizing trait distributions in the three-species case under the coalescent and Brownian motion models

Before deriving results for the multispecies coalescent model using a three-species phylogeny, we present expectations for quantitative traits under Brownian motion (BM), a diffusion model commonly used in phylogenetic comparative methods. Although there are multiple possible interpretations of the underlying microevolutionary dynamics that lead to BM, we compare our model to it because it is tractable and popular, thus providing a clear baseline that is likely to be informative about the behavior of our model. Under BM, trait values from multiple species will exhibit a

multivariate normal distribution with the covariance structure given by the phylogeny (**Felsenstein, 1973**). More specifically, in the case of n species, the variances within species and covariances between species are given by $\boldsymbol{V} = \sigma^2\boldsymbol{T}$, the variance-covariance matrix. Here, σ^2 is the evolutionary rate parameter, which measures how much change is expected in an infinitesimal time interval. $\boldsymbol{T}$ is an n x n matrix whose off-diagonal entries, t_{ij}, are lengths of the internal branches subtending the ancestor of species i and j, and whose diagonal entries correspond to the lengths of the paths between each species and the root. For the phylogeny in **Figure 2a**, and $\sigma^2 = 1$:

$$\boldsymbol{V} = \sigma^2\boldsymbol{T} = \sigma^2\begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix} = \begin{bmatrix} 5 & 4 & 0 \\ 4 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix} \quad (1)$$

For example, the BM expected variance in species A, $Var_{BM}(A)$, corresponds to the rate parameter multiplied by the length of the path extending from the root to the tip, and therefore evaluates to five. Note that $Var_{BM}(A)$ is not the population trait variance observed among individuals of A, but the expected variance in species A's mean trait value, resulting from evolution along the lineage leading to A. The covariance between species A and B, $Cov_{BM}(A, B)$, corresponds to the rate parameter multiplied by the length of the branch shared by these two lineages, which evaluates to four in this example.

Given the species tree topology in **Figure 2a**, the expected variance in trait value within any species, A, B, or C, is also readily derived under a neutral coalescent model (for a complete derivation, see section 1.1 in Appendix 1). Although many quantitative traits may be under selection, the selection coefficient on an individual allele is proportional its effect size (**Keightley and Hill, 1988**). This implies that loci with small effects will experience very small selection coefficients, and that neutral expectations for genealogical quantities should still be justified. The expected trait value variance within species A or B is given by:

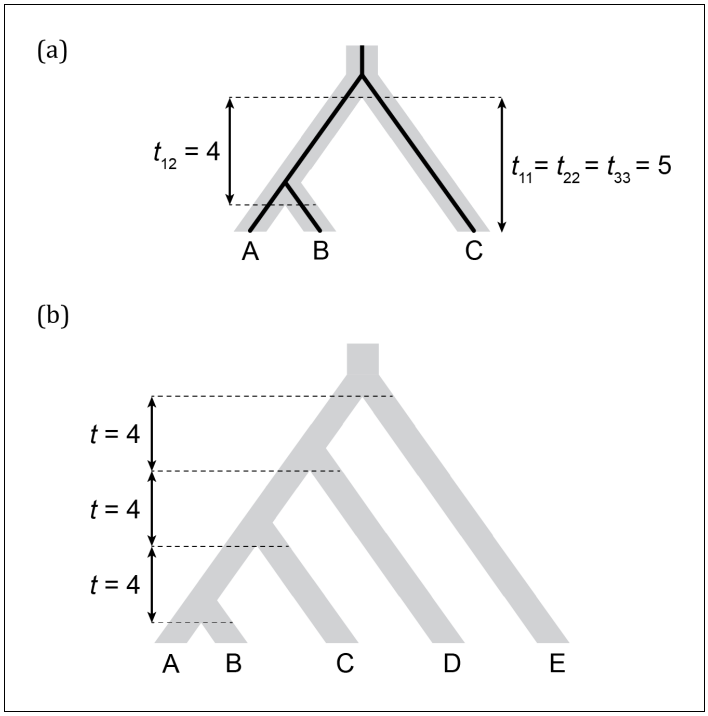


Figure 2. Three- and five-taxon trees with branch lengths indicated. (**a**) Three-species phylogeny (and a concordant gene tree within it) and its corresponding T matrix entries. (**b**) Five-species phylogeny used in coalescent simulations for PCM analyses. Branch lengths are indicated in units of $2N$ generations.
DOI: https://doi.org/10.7554/eLife.36482.003

$$Var_{Coal} = 2N\mu\sigma_M^2\left[t_e + \left(1 - e^{-t/2N}\right)\left(\frac{t}{2N} + 1\right) + \left(e^{-t/2N}\right)\left(\frac{t}{2N} + 1 + \frac{1}{3}\right)\right]$$

t t_e A B A B N μ σ_M^2

Equation 2

t_e A B e^{-tN} A B e^{-tN} t N t N C **Equation 2** t N t_e C,

.

$$Cov_{Coal}(A,B) = 2N\mu\sigma_M^2\left[\left(1 - e^{-t/2N}\right)\left(1 + \left(\frac{t}{2N} - \left(1 - \frac{t/2N}{e^{t/2N} - 1}\right)\right) + \left(\frac{1}{3}e^{-t/2N}\right)\right]$$

$$Cov_{Coal}(A,C) = Cov_{Coal}(B,C) = 2N\mu\sigma_M^2\left(\frac{1}{3}e^{-t/2N}\right)$$

A C B C A B C t_e

Equations 3 and 4

Figure 2a $\approx$ N $2N$ / σ_M^2 μ / Var-$_{Coal}$ / Var_{BM} / .

Gillespie and Langley, 1979, **Edwards and Beerli, 2000**

A C B C .

$Cov_{Coal}(A, C)$ $Cov_{BM}(A, C)$ **Equation 4** **Equation 1** A B A B **Figure 2a** $Cov_{Coal}(A, B)$ $Cov_{BM}(A, B)$

determine the value of such a trait. However, expected trait values in the coalescent are not exactly the same as those expected under the classical BM model, even in the absence of genealogical discordance and given a fixed phylogeny. While expected covariances will be identical between models if no genealogical discordance is present, expected variances will still differ; this difference will be accentuated with larger ancestral population sizes. This result will therefore affect any parameters being estimated — such as the evolutionary rate, σ^2 — that depend on expected species variances. Below, we explore how the expectations under the coalescent and BM models can further differ in the presence of ILS and discordance.

Consequences of genealogical discordance to quantitative traits: the 'deep coalescence' effect and hemiplasy

We can predict from the expectations laid out above that the variances and covariances under the coalescent model will change in the presence of discordance. In contrast, the BM model will have the same expectations because it does not consider genealogical discordance — the species tree is a fixed parameter. In order to characterize the effects of discordance on variances within species and covariances between species, we considered five different scenarios with increasing percentages of gene tree discordance (0, 15, 30, 45 and 60% discordance, respectively). We used the three-species phylogeny (***Figure 2a***) for its mathematical tractability, and in addition to computing the expectations of these measures (using ***Equations 2–4***), we simulated 1000 data sets under each of the five scenarios. This simulation procedure is illustrated in ***Figure 3***, where for each locus underlying a quantitative trait, mutations are thrown down at random along the genealogy and mutational effects of each mutation are drawn from a distribution determined by σ^2_M. Simulations were repeated for different numbers of loci affecting the trait: 5, 15, 25, 50 and 100. In keeping with the usual practice in

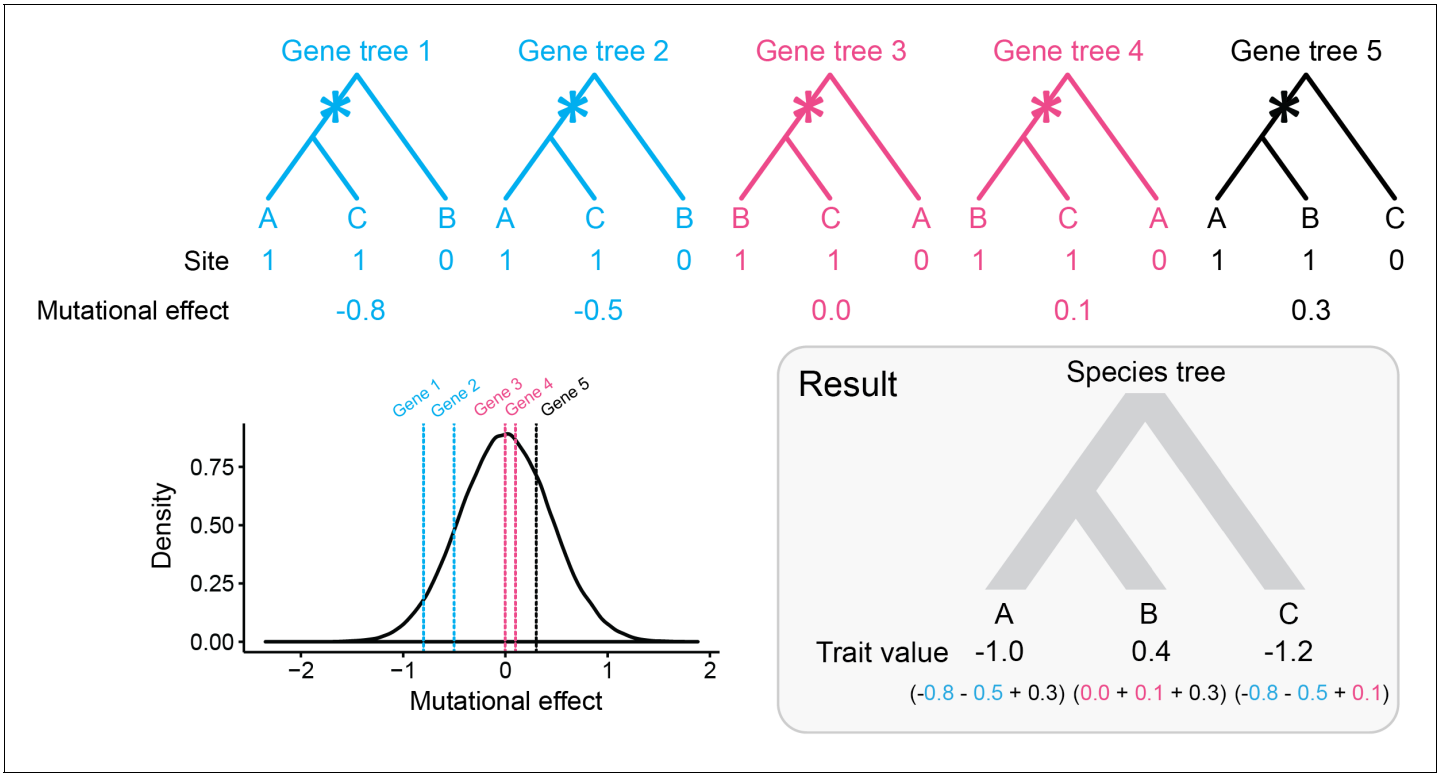


Figure 3. A single continuous traits controlled by five loci, four of which have discordant gene trees. (**a**) Genealogies of the loci controlling the trait. Asterisks represent mutations at a given site in each of the five loci. Ancestral alleles (0) have no effect on the trait value. Derived alleles (1) have their random effects on the trait value drawn from a mutational effect distribution (see (**b**)). (**b**) Mutational effect distribution of derived alleles. The distribution has a mean of zero and unit variance. (**c**) The outcome of a simulation consists of one trait value per species, which correspond to the sum of all derived allele mutational effects coming from all loci controlling the trait.

DOI: https://doi.org/10.7554/eLife.36482.004

comparative analyses of employing a single, static species tree, levels of ILS were increased by multiplying ancestral population sizes by incrementally larger factors — the topology and branch lengths of the species tree were kept constant (see Materials and methods for details).

We observed an overall good match between the observed and expected variances and covariances (***Figure 4*** and ***Figure 5a***). Under the coalescent model, larger ancestral population sizes make coalescent waiting times longer, and result not only in more ILS and more gene tree discordance, but also in taller trees on average. As expected (***Equations 2–4)***, data sets simulated with larger N therefore had higher variances and covariances (***Figure 4a–b***). We refer to this phenomenon as the 'deep coalescence' (DC) effect. The DC effect occurs due to the increase in average gene tree height, relative to the species tree height, under the coalescent model with large population sizes (cf. ***Gillespie and Langley, 1979***). We stress that (i) this effect is *not* due to discordance, and (ii) not only variances, but covariances among lineages that share a history in the species tree, are affected. The latter happens because, as mentioned above, it will take longer for any two lineages to coalesce given a larger population size (the parameter N controls this time in ***Equations 3 and 4***). Consequently, the waiting time for the last coalescent event (which determines the length of the internal branch) will also be longer, leading to higher covariances between pairs of descendant species.

The number of loci did not influence variances and covariances (results not shown), which is expected. This is because the standard deviations of the mutational effect distributions used in our simulations (σ_M^2) are scaled to keep trait-value variances constant with changing numbers of loci, thus ensuring a fair comparison between models with different numbers of loci. This follows the standard logic of the infinitesimal model, that is, the larger the number of loci controlling a trait, the smaller the effect each mutation should have on the trait value (***Fisher, 1919***; for more details in the context of the coalescent model, see ***Schraiber and Landis, 2015***).

Finally, under the coalescent model, gene tree discordance *does* have an effect: the covariance between species A and C (and between B and C) increases with more ILS relative to the covariance between species A and B (***Figure 5a***). Recall that when there is no discordance there is no covariance between non-sister species, because they do not share an evolutionary history. Discordant gene trees offer the opportunity for non-sister species to have a shared history, and covariance increases. As a result, there is an increased similarity between non-sister species in quantitative traits due to hemiplasy in the underlying gene trees. Ultimately, the effect of hemiplasy on continuous traits is to make covariances between different pairs of species converge on the same value (***Figure 5b***). This makes intuitive sense, as in the limit all three topologies will be equally frequent, resulting in equal covariances between all pairs of species.

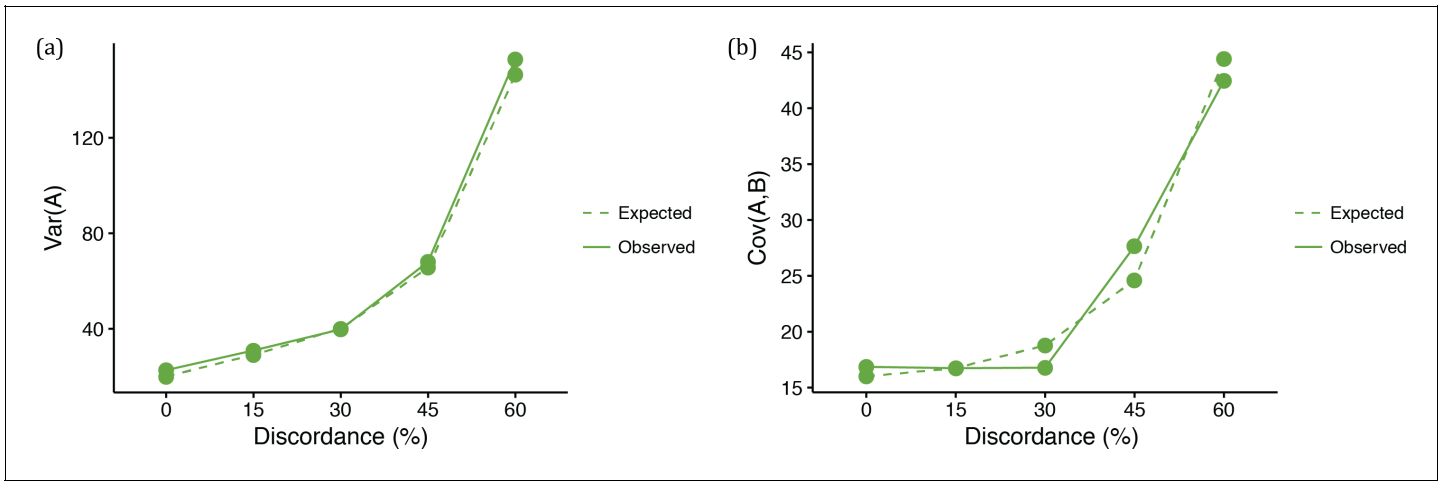


Figure 4. Quantitative trait summaries under varying levels of genealogical discordance, for a three-taxon species tree. (**a**) Expected and observed variances in trait values of species A in each of the five ILS conditions. Expected values come from ***Equation 2***. (**b**) Expected and observed covariances between species A and B in each of the five ILS conditions. Expected values come from ***Equation 3***.

DOI: https://doi.org/10.7554/eLife.36482.005

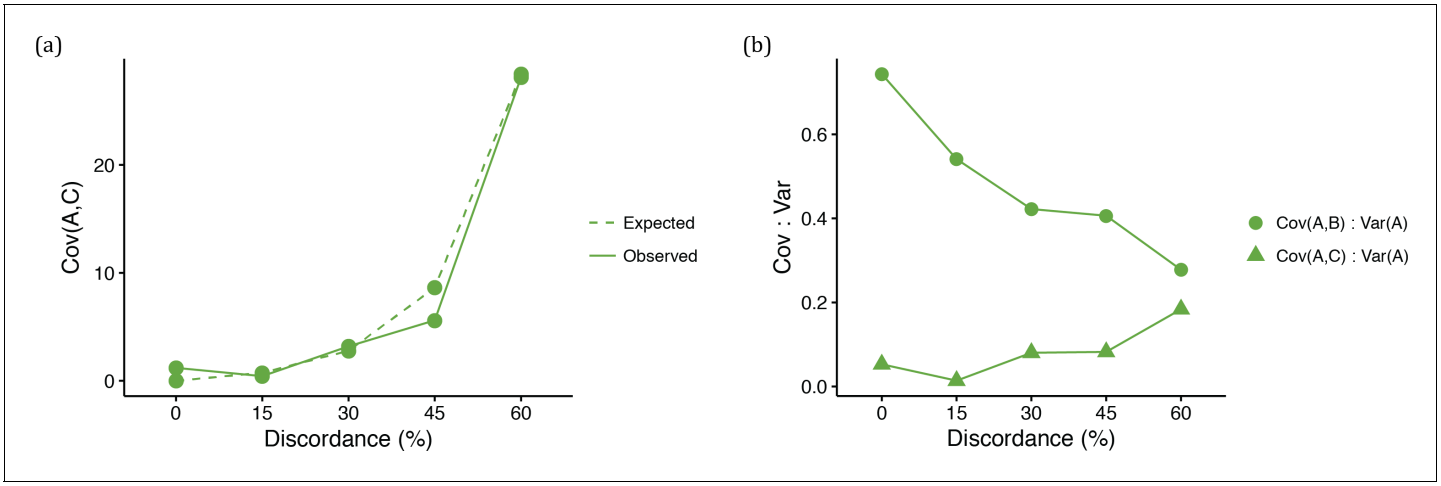


Figure 5. Quantitative trait summaries under varying levels of genealogical discordance, indicating the effect of hemiplasy for a three-taxon species tree. (a) Expected and observed covariances between species *A* and *C* in each of the five ILS conditions. Expected values come from ***Equation 4***. (b) Observed covariances between a pair of species normalized by the variance in species *A*, for all five ILS conditions.
DOI: https://doi.org/10.7554/eLife.36482.006

We emphasize that the aforementioned effects were observed despite the fact that the concordant topology was always the most common, and that substitutions on discordant trees and discordant branches occurred in only a fraction of the loci underlying the continuous trait. Furthermore, the number of loci does not seem to strongly affect our results (see below), as the difference between covariances were similar regardless of the number of loci controlling the trait. This suggests that our model applies to all quantitative traits, not just those controlled by a large number of loci.

Hemiplasy increases inferred evolutionary rates and decreases phylogenetic signal

The results from our model demonstrate that in the presence of ILS, quantitative traits will have both larger variances in each species as well as positive covariances between non-sister species. These outcomes suggest that hemiplasy could affect inferences involving quantitative traits, as traditional models (BM or otherwise) do not allow for shared trait variation along branches that do not exist in the species tree.

We first investigated the impact of discordance and hemiplasy on estimates of a commonly inferred parameter, the BM evolutionary rate, σ^2. We estimated σ^2 from data simulated along a five-species asymmetric phylogeny (***Figure 2b***). Simulating data for five species allows for more ILS (and a greater effect of hemiplasy; ***Mendes and Hahn, 2016***) relative to the three-species case, due to the larger number of possible gene tree topologies (105 in the former case versus the three possible topologies in the latter). Again, we simulated data under five ILS conditions with different percentages of gene tree discordance (0, 20, 40, 60 and 80% discordant trees, respectively) by keeping the phylogeny constant and increasing population sizes. As in the three-species case, we simulated continuous traits controlled by 5, 15, 25, 50 and 100 loci.

As mentioned above, increasing ancestral population sizes increases both ILS and the average height of gene trees with two main resulting patterns: (i) expected covariances between non-sister species will increase (due to hemiplasy), and (ii) expected variances within species will increase (due to deep coalescence). Because BM does not model the number, topology, or lengths of the gene trees underlying a continuous trait, we predicted that both outcomes would be accounted for when inferring rates under the BM model as spuriously higher evolutionary rates. Indeed, we observed a positive correlation between the estimated σ^2 under BM and ILS (***Figure 6a***). This pattern was the same for all data sets, irrespective of the number of loci controlling the trait.

We also reasoned that another major consequence of hemiplasy — resulting from the changes in expected covariances in trait values between pair of species — would be the reduction of the average phylogenetic signal with increasing ILS. This is because the effect of hemiplasy on quantitative

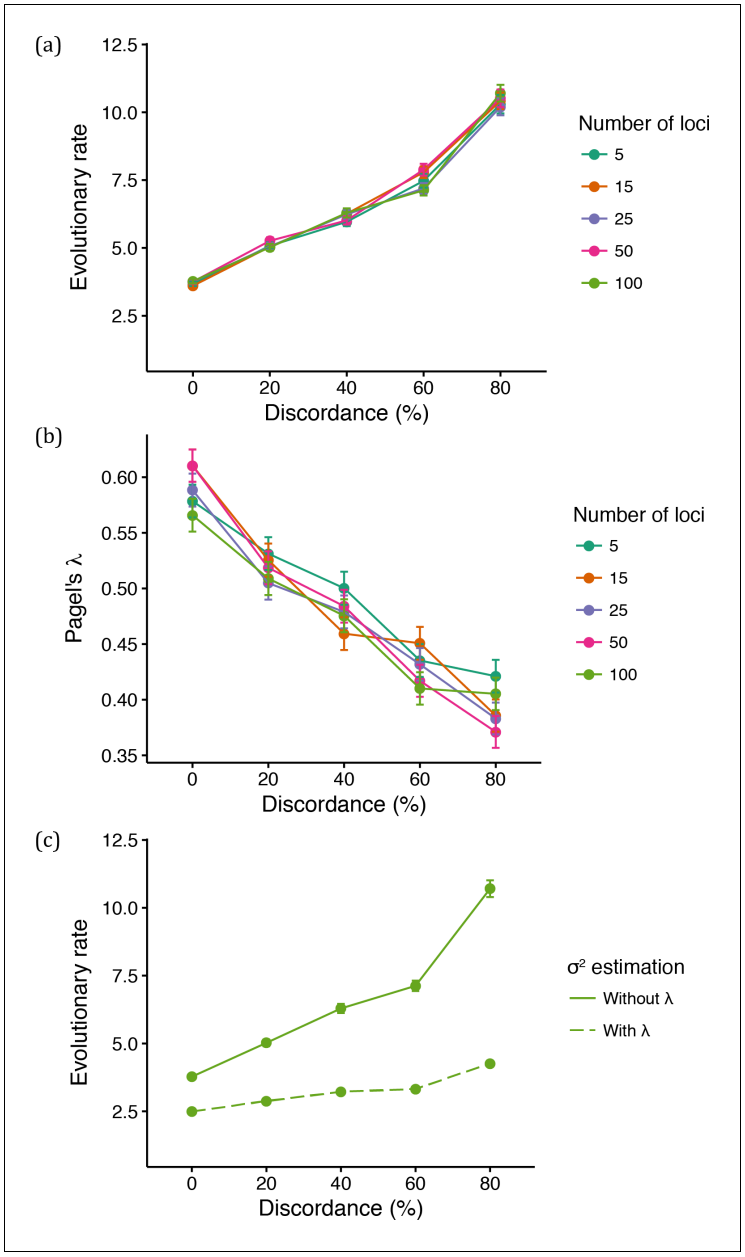


Figure 6. Mean parameter estimates under varying levels of genealogical discordance and number of loci controlling quantitative trait. (**a**) Mean evolutionary rate for different number of loci controlling the simulated continuous trait and different levels of discordance. (**b**) Mean value of Pagel's λ for different number of loci controlling the simulated continuous trait and different levels of discordance. (**c**) Mean evolutionary rate when 100 loci control the trait ('Without λ' is the same as shown in (**a**); in 'With λ', the rate was estimated with Pagel's λ).
DOI: https://doi.org/10.7554/eLife.36482.007

traits is to make covariances between more closely related species become smaller relative to covariances between more distantly related species. The more hemiplasy, the less should the covariances resemble values that would be observed for a trait evolving along the species tree, and thus the phylogenetic signal should be lower. We measured the phylogenetic signal in each replicated simulation by estimating a commonly used parameter, Pagel's λ (where λ = 1 indicates a trait evolving according to BM along a species tree, and λ <1 indicates lower phylogenetic signal; ***Pagel, 1999***; ***Freckleton et al., 2002***). As expected, estimates of λ decreased on average with increasing ILS

(*Figure 6b*), reflecting the lower phylogenetic signal of traits partly determined by discordant gene trees.

Given the results from Pagel's λ, we attempted to distinguish the contribution of the DC effect (i.e. overall increase in variances and covariances) from that of hemiplasy (i.e. relative change in covariances) to the spurious increase in σ^2. The parameter λ can be thought of as a species tree branch-stretching parameter (*O'Meara, 2012*): we predicted that when estimating σ^2 in the presence of λ, the latter would act as a 'buffer' parameter absorbing the effect of hemiplasy by becoming reduced itself (as shown in *Figure 6b*). Indeed, evolutionary rates were much lower when estimated in the presence of λ (*Figure 6c*, 'With λ'), but still remain higher in data sets simulated with increasing levels of ILS. This is because while λ can absorb the effect of hemiplasy by shrinking internal branches, it cannot account for the DC effect resulting from the increased average gene tree heights in higher ILS conditions.

These results suggest that both the DC effect and hemiplasy contribute to the increase in estimates of σ^2. In BM model terms, understanding the impact of the DC effect on higher estimates of σ^2 is straightforward: if the tree (reflected in matrix ***T***, *Equation 1*) is held constant and all variances and covariances (the entries of ***V***, *Equation 1*) become larger, then σ^2 must become larger. But our results also suggest that the effect of hemiplasy is comparable to the DC effect, and possibly of even greater magnitude in the presence of more ILS. This observation is perhaps less intuitive, but indicates that σ^2 must become much higher to account for the difference between the observed covariances (i.e. off-diagonal entries of ***V***) and expected covariances, given the observed variances and ***T***. Assuming that quantitative traits evolve according to the coalescent model, larger ancestral population sizes and genealogical discordance can thus lead to an overestimation of σ^2 and to lower values of λ, and will likely affect all comparative methods that can make use of such parameters, not just the BM model. We point the curious reader to Appendix 1 (section 2.3) for a thorough theoretical treatment on how expected trait variances and covariances under the two models should differ, and why these differences can lead to the reported estimates of σ^2 and λ.

Hemiplasy can increase the false positive rate in phylogenetic hypothesis testing

Many studies test the hypothesis that groups of species differ in measured traits due to factors other than phylogenetic relatedness. We addressed whether hemiplasy could also interfere with this type of phylogenetic hypothesis testing. The comparative method of choice we used was the phylogenetic ANOVA (*Garland et al., 1993*). As in traditional ANOVA, this method allows the comparison of mean trait values across groups of species. Within a linear model framework, the phylogenetic ANOVA also corrects for the inflation of degrees of freedom caused by the non-independence of trait value errors around the regression line (which can be estimated by looking at the residuals) — which results from the hierarchical nature of the phylogenetic relationships among species (*Felsenstein, 1985*; *Garland et al., 1993*; *Uyeda et al., 2018*). This correction allows the approximation of the true number of degrees of freedom through simulations of trait values along the phylogeny (given some model of trait evolution — BM in our case). The simulations collectively comprise an empirical *F* distribution that is then used in hypothesis testing (*Garland et al., 1993*).

Our prediction was that increasing levels of ILS and of hemiplasy would increase the false positive rate of phylogenetic ANOVAs. We tested this prediction by conducting phylogenetic ANOVAs on the five-species simulations. Hypothesis testing consisted of comparing the null hypothesis that a pair of species had the same mean trait value as the remaining three species, against the alternative hypothesis of different means. This procedure was repeated on each of the 1000 replicates, for all possible groupings of two species versus three species; we then recorded the average number of times per replicate the p value was significant ($p<0.05$).

As predicted, we observed a positive correlation between ILS levels and the mean number of times the null hypothesis was rejected in the phylogenetic ANOVA; this trend was unaffected by the number of loci underlying the trait (*Figure 7*). This result suggests that an arbitrary group of species is, on average, more likely to have a spuriously (and significantly) smaller or larger mean trait value than the remaining species in the presence of gene tree discordance.

The sharing of similar trait values by non-sister species is an expected byproduct of higher expected covariances among those species when there is gene tree discordance. Such changes in

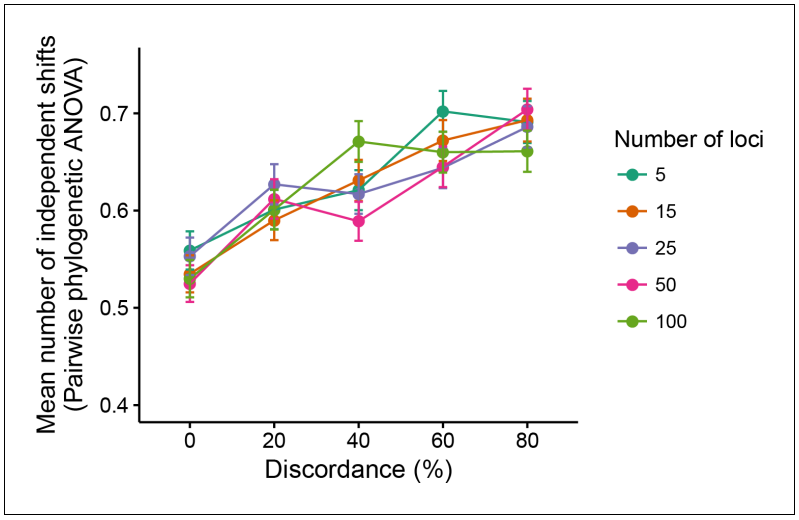


Figure 7. Mean number of independent trait-value shifts (i.e. significant phylogenetic ANOVA tests) among all possible groupings of two versus three species.
DOI: https://doi.org/10.7554/eLife.36482.008

expected covariances between pairs of species trait values are a symptom of hemiplasy (***Figure 5***), but not of the DC effect. We thus believe that hemiplasy not only contributes to the incorrect estimation of parameters such as the evolutionary rate, but can also play a role in increasing the false positive rate of phylogenetic comparative methods.

Threshold traits are strongly affected by hemiplasy

As demonstrated above, the magnitude of the effect of hemiplasy on phylogenetic inferences is consistently proportional to the observed levels of gene tree discordance in a data set. Our results also suggest that the number loci underlying a quantitative trait does not matter to the expected trends from such inferences. One remaining question, however, is whether hemiplasy can have an effect on a threshold trait — that is, a discrete trait that has a continuous character underlying it (***Wright, 1934***; ***Felsenstein, 2005***), and if the genetic architecture of such trait is relevant to this effect.

Addressing this question is straightforward, as we only need to treat our simulated continuous traits as the underlying 'liability' of a threshold character. By choosing an arbitrary threshold of one standard deviation above the mean continuous trait value (over all replicates and all species), we coded all simulated trait values as either '0' (if below the threshold), or '1' (if above). Defining a threshold using a dispersion measure such as the standard deviation, instead of a fixed value, allows us to account for the higher variances expected in replicates under higher ILS conditions.

Before laying out our predictions for how the effect of hemiplasy on threshold characters should be manifested, we first define a few terms used in the discussion that follows. A 'trait pattern' consists of the threshold character states (from a single replicate) observed at the tips of the tree. Given tree ((((*A*,*B*),*C*),*D*),*E*) (the tree we used in the simulations; ***Figure 2b***), trait pattern '11000' signifies species *A* and *B* sharing state '1' (both had liabilities above the threshold) and species *C*, *D* and *E* sharing state '0' (the three species had liabilities below the threshold). A congruent (informative) trait pattern can be produced by character-state transitions occurring on internal branches that are present in the species tree; thus trait patterns '11000' and '11100' are congruent. Conversely, an incongruent trait pattern is the result of either homoplastic or hemiplastic evolution: multiple true character-state transitions, or transitions on internal branches of discordant gene trees that are absent from the species tree, respectively. Trait patterns such as '01100' and '11010', for example, are incongruent.

If hemiplasy affects threshold traits as it does continuous traits, we predicted that higher ILS levels would lead to a larger number of incongruent trait patterns, and to a lower number of congruent trait patterns. As expected, counts of incongruent informative trait patterns increased with

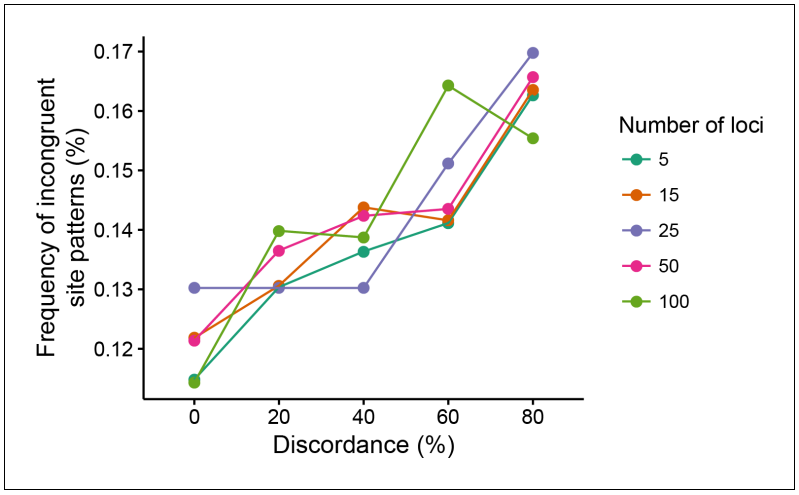


Figure 8. Frequency of incongruent trait patterns (out of all informative trait patterns) for threshold traits. Each combination of level of discordance and number of loci was simulated 1000 times.
DOI: https://doi.org/10.7554/eLife.36482.009

increasing ILS levels (***Figure 8***); congruent trait patterns likewise decreased. Furthermore, the same trend was observed from simulations where the liability character was underlain by few or many loci (***Figure 8***). This suggests that even in the case of threshold traits, larger numbers of loci comprising the genetic architecture will not mitigate the effect of hemiplasy.

Incongruent trait patterns are interesting because they can be suggestive of convergent evolution, or of correlated evolution when more than one trait exhibits similar patterns. While we do not further investigate the behavior of phylogenetic comparative methods applicable to discrete characters here, it is clear nonetheless that in the presence of gene tree discordance inferences from incongruent patterns can be misleading about the number of times a trait has evolved. This is because misidentification of the branches along which character states are inferred to change will be more likely in the presence of gene tree discordance. Overall, the more genealogical discordance is present in a data set, the more likely it is that a discrete trait will exhibit an incongruent pattern by chance, simply as a result of the stochasticity of the coalescent process.

Discussion

In the present study, we propose a multispecies coalescent model for quantitative trait evolution that incorporates the genealogical discordance that underlies such traits. We use this model to investigate whether discordance can affect phylogenetic inferences, mainly through the phenomenon of hemiplasy. We considered ILS to be the sole cause of gene tree discordance. By employing coalescent theory, we demonstrate that in the absence of ILS the coalescent and traditional BM model are equivalent with respect to the expected covariances between species trait values, but differ in terms of the species expected trait variances. In the presence of ILS, hemiplasy causes the expected covariance in trait values between pairs of more distantly related species to increase.

The increased covariance due to hemiplasy leads to error in estimates of two parameters commonly studied under the BM model, namely, the evolutionary rate, σ^2, and Pagel's λ. Hemiplasy consistently led to an overestimation of σ^2, and to lower λ estimates. Moreover, errors were also observed when conducting comparative analyses such as the phylogenetic ANOVA, whose false positive rate was increased by greater levels of genealogical discordance and hemiplasy. Finally, by treating quantitative traits as a liability character underlying a threshold trait, we found that hemiplasy affects the number of times such traits appeared incongruent with the species tree. All of the aforementioned results held irrespective of the number of loci controlling the quantitative trait.

We have focused the comparisons of trait-value expectations from our model (which assumes no selection) to those under BM. The latter can be seen as equivalent to a quantitative genetics model in which many genes have small effects on a selectively unconstrained character (***Lande, 1976***;

Harmon, 2017). However, evidence suggests that even when BM fits the data well, the evolutionary rate parameter does not seem to be meaningful as a quantitative genetic measure of drift (*Estes and Arnold, 2007*; *Harmon et al., 2010*; *Uyeda et al., 2011*). We note again that the motivation for investigating BM (and methods that employ BM, e.g. the phylogenetic ANOVA) as a model for learning about quantitative traits was not the similarity of its assumptions to the theoretical and simulation conditions considered here, but instead: (i) its popularity, tractability, and centrality in these types of studies, and (ii) the fact that BM-based models do not explicitly incorporate the number of loci, their gene trees, nor the effects of each locus on a quantitative trait of interest. These qualities of BM are helpful in elucidating the properties of our model, in addition to being suggestive of how more realistic BM-based models would fare in the face of discordance.

There are extensions to BM that can explicitly include certain forms of selection (e.g. the Ornstein-Uhlenbeck model; *Butler and King, 2004*) but whose behavior we have not investigated. While these models incorporate selection, it is still unlikely that the parameters of such models can be interpreted as population-genetic quantities. For the purposes of the analyses presented here, whether or not BM can accommodate traits evolving under a mix of drift and selection is relevant only insofar as this model can escape the effects of hemiplasy. Based on our results, we expect that models that do not explicitly consider genealogical discordance can at least in principle be vulnerable to hemiplasy, regardless of whether they include selection.

Although we have only presented theoretical results for trees with a small number of tips, our results are likely to hold even when the species tree is large. This is because the total number of taxa is not especially relevant to the multispecies coalescent process; instead, it is the number of lineages in each phylogenetic 'knot' (i.e. the number of lineages involved in ILS; *Ané et al., 2007*). Extensions of our results to knots with four or five taxa are straightforward, though with larger numbers than this the total number of topologies that need to be considered grows unwieldy. Similarly, the assumptions used by our model about constant population sizes and mutation rates are only required to be true across phylogenetic knots, where the relevant coalescent events occur. This may be only a very small fraction of the total amount of time represented by large trees. We do not expect violations of these assumptions elsewhere in the tree — outside of knots — to have a qualitative effect any different than they would in non-coalescent models. Of course, the larger the periods of time covered by any such analysis, the harder it can be to infer parameters of interest, as in those circumstances even estimating phylogenies can be a daunting task. In such cases, ILS may be the least of a phylogeneticist's concerns. Finally, when analyzing our simulated data here we employed the true, known species tree used in the simulations, but in typical phylogenetic analyses of real data an estimate of the phylogeny is used. These facts will certainly cause results from real-world studies to deviate from expectations.

While we found that discordance can indeed affect traditional phylogenetic methods for studying traits with complex genetic architectures, it is likely that the results from our simple simulations are conservative. This is because we assumed only additivity of mutations and a Gaussian mutational effect distribution. The presence of dominance, epistasis, and broader or skewed mutational effect distributions are likely to compound the effects of hemiplasy. Moreover, while our assessment of phylogenetic methods is by no means exhaustive, it is unlikely that the trends we report here are exclusive to the approaches we investigated. Methods that compare models with one versus multiple evolutionary rates across a tree (*O'Meara et al., 2006*), or that estimate branch lengths from quantitative traits (*Felsenstein, 1973*), for example, could be affected by hemiplasy in the same way that nucleotide substitution models are (*Mendes and Hahn, 2016*). Similarly, methods addressing the correlation between discrete traits (e.g. *Pagel, 1994*) could also have increased false positive rates if hemiplasy acts on multiple traits in similar ways. Hemiplasy is also expected to broaden the confidence intervals around ancestral state reconstructions of quantitative traits (*Martins and Hansen, 1997*), making it harder to infer significant shifts in trait means and to place such shifts on internal branches of the species tree. While recently proposed methods that study BM models over species networks do represent a step forward in the presence of discordance due to hybridization and introgression (e.g. *Dwueng-Chwuan and O'Meara, 2015*; *Bastide et al., 2018*), these methods do not account for either deep coalescence or the full spectrum of genealogical discordance.

Given our results, it is reasonable to ask whether and when traditional phylogenetic comparative methods for quantitative traits are appropriate. ILS is expected to act when there are short internode distances in a species tree, regardless of how far in the past the rapid succession of speciation

Schluter et al., 1997, *Freckleton and Harvey, 2006*

Mendes and Hahn, 2016

λ

λ

Materials and methods

Simulations under the multi-species coalescent model for quantitative traits

Schraiber and Landis, 2015 *Figure 3*

Harmon, 2017

Figure 3b
ms *Hudson, 2002*

A. *B..* *C.* *A.* *B..* *C..* *D..* *E.*

,

N

θ /

Parameter estimation and hypothesis testing under Brownian motion

σ λ

geiger *Harmon et al., 2008*

σ λ λ *Figure 6c* *geiger*

geiger

Discretization of quantitative trait values using the threshold model

Wright, 1934, ***Felsenstein, 2005***

.

Acknowledgements

Additional information

Funding

Funder	Grant reference number	Author
National Science Foundation	DBI-1564611	Matthew W Hahn Fábio K Mendes
National Institutes of Health	R35 GM124745	Joshua G Schraiber

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

, ,

Author ORCIDs

.

.

Decision letter and Author response

https://doi.org/10.7554/eLife.36482.017

https://doi.org/10.7554/eLife.36482.018

Additional files

Supplementary files

- DOI: https://doi.org/10.7554/eLife.36482.010
- DOI: https://doi.org/10.7554/eLife.36482.011

Data availability

.

.

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Mendes FK, Fuentes-González JA, Schraiber JG, Hahn MW	2018	Data from: A multispecies coalescent model for quantitative traits	https://doi.org/10.5061/dryad.m2s1735	Available at Dryad Digital Repository under a CC0 Public Domain Dedication

References

Ané C
Molecular Biology and Evolution **24**.

Avise JC . *Systematic Biology* **57**.

Bastide P
Systematic Biology.

Brawand D
Nature **513**.

Bryant D
. *Molecular Biology and Evolution* **29**.

Butler MA . *The American Naturalist* **164**.

Copetti D
PNAS **114**.

Degnan JH *PLoS Genetics* **2**.

Dwueng-Chwuan J *bioRxiv* .

Edwards S .
Evolution **54**.

Edwards SV *Evolution* **63**.

Estes S .
The American Naturalist **169**.

Felsenstein J *American Journal of Human Genetics* **25**.

Felsenstein J *The American Naturalist* **125**.

Felsenstein J
Philosophical Transactions of the Royal Society B: Biological Sciences **360**.

Fisher RA
Transactions of the Royal Society of Edinburgh **52**.

Freckleton RP .
The American Naturalist **160**.

Freckleton RP *PLoS Biology* **4**.

Garamszegi LZ *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*

Garland T
Systematic Biology **42**.

Gillespie JH *Journal of Molecular Evolution* **13**.

Hahn MW *Evolution* **70**.

Harmon LJ .
Bioinformatics **24**.
Harmon LJ

Evolution **64**.
Harmon LJ

Harvey PH *The Comparative Method in Evolutionary Biology*
Heled J *Molecular Biology and*
Evolution **27**.
Hobolth A

Genome
Research **21**.
Hudson RR *Evolution* **37**.

Hudson RR *Bioinformatics* **18**.

Keightley PD
Genetical Research **52**.

Kubatko LS
Systematic Biology **56**.

Lande R *Evolution* **30**.

Larget BR .
Bioinformatics **26**.

Liu L . *Bioinformatics* **24**.

Liu L
Systematic Biology **58**.
Lynch M *Evolution*
42.
Lynch M
Evolution; International Journal of Organic Evolution **43**.

Lynch M . *Ecological Genetics* .

Maddison WP *Systematic Biology* **46**.

Martins EP .
The American Naturalist **149**.

Mendes FK *Systematic*
Biology **65**.
Mendes FK
Molecular Biology and Evolution **33**.

Mendes FK *Systematic Biology* **67**.

Mirarab S .
Bioinformatics **31**.

O'Meara BC
Evolution **60**.

O'Meara BC . *Annual Review of Ecology,*
Evolution, and Systematics **43**.
Ovaskainen O
Genetics **189**.

Pagel M .
Proceedings of the Royal Society B: Biological Sciences **255**.

Pagel M *Nature* **401**.

Pamilo P *Molecular Biology and Evolution* **5**.

Pease JB
PLOS Biology **14**.

Pollard DA
Drosophila. *PLoS Genetics* **2**.

Schluter D *Evolution* **51**.

Schraiber JG
Theoretical Population Biology **102**.

Simons YB
PLoS Biology **16**.

Solís-Lemus C
PLoS Genetics **12**.

Suh A
PLoS Biology **13**.

Tajima F *Genetics* **105**.

Than C .
BMC Bioinformatics **9**.

Uyeda JC *PNAS* **108**.

Uyeda JC *Systematic Biology* **106**.

White MA
PLoS Genetics **5**.

Whitlock MC *Genetical Research* **74**.

Wright S *Genetics* **19**.

Zhang G
Science **346**.

Appendix 1

DOI: https://doi.org/10.7554/eLife.36482.012

1. Deriving expected variances and covariances in quantitative trait values under the coalescent model

1.1 Variance in the three species case

S X i

.

$$\mathrm{Var}(X_i) = 2N\mu\sigma_M^2 \quad B_{\mathrm{root},i},$$

N μ σ_M^2

σ_M^2

σ^2 $\mu\sigma_M^2$ $B_{\mathrm{root},i}$

i *gene trees* S

$B_{\mathrm{root},i}$.

$$B_{\mathrm{root},i} = t_e + \left(1 - e^{-t/2N}\right)\left(\frac{t}{2N} + 1\right) + \left(e^{-t/2N}\right)\left(\frac{t}{2N} + 1 + \frac{1}{3}\right),$$

t t_e S $1 - e^{-t/2N}$

$e^{-t/2N}$

$t_e + \frac{t}{2N} + 1$ $t_e + \frac{t}{2N} + 1 + \frac{1}{3}$

Equation 2 .

$$E(\mathrm{Var}(X_i)) = 2N\mu\sigma_M^2\left[t_e + \left(1 - e^{-t/2N}\right)\left(\frac{t}{2N} + 1\right) + \left(e^{-t/2N}\right)\left(\frac{t}{2N} + 1 + \frac{1}{3}\right)\right].$$

2.Covariances in the three species case

i j .

$$\mathrm{Cov}\left(X_i, X_j\right) = 2N\mu\sigma_M^2 \quad B_{\mathrm{root,MRCA}_{(i,j)}},$$

$B_{\mathrm{root,MRCA}_{(i,j)}}$

i j

$S = ((\mathrm{A}, \mathrm{B}), \mathrm{C})$ $i = \mathrm{A}$ $j = \mathrm{B}$ $B_{\mathrm{root,MRCA}_{(i,j)}}$.

$$B_{\mathrm{root,MRCA}_{(\mathrm{A,B})}} = \left(1 - e^{-t/2N}\right)\left(1 + \left(\frac{t}{2N} - \left(1 - \frac{t/2N}{e^{t/2N} - 1}\right) + \left(\frac{1}{3}e^{-t/2N} \quad 1\right),\right.\right.$$

$e^{-t/2N}$ $\frac{1}{3}$

$1 - e^{-t/2N}$

$B_{\mathrm{root,MRCA}_{(\mathrm{A,B})}}$ $1 - \frac{t/2N}{e^{t/2N} - 1}$

Mendes and Hahn, 2018

$B_{\mathrm{root,MRCA}_{(\mathrm{A,B})}}$ $2N$

$\frac{1}{3}e^{-t/2N}$

$2N \quad \frac{t}{2N} = 1$

Equation 3 .

$$E(\mathrm{Cov}(X_A, X_B)) = 2N\mu\sigma_M^2\left[\left(1-e^{-t/2N}\right)\left(\left(1+\left(\frac{t}{2N}-\left(1-\frac{t/2N}{e^{t/2N}-1}\right.\right.\right.\right. \quad +\left(\frac{1}{3}e^{-t/2N}\right.\right]\left.\right($$

$i = \mathrm{A} \quad i = \mathrm{B} \quad j = \mathrm{C} \quad B_{\mathrm{root,MRCA}_{(i,j)}}$.

$$B_{\mathrm{root,MRCA}_{(\mathrm{A,C})}} = B_{\mathrm{root,MRCA}_{(\mathrm{B,C})}} = \frac{1}{3}e^{-t/2N} \quad 1.$$

Equation S.5 $2N$

$\frac{1}{3}e^{-t/2N} \quad 2N \quad \frac{1}{2N} = 1$

Equation 4 .

$$E(\mathrm{Cov}(X_A, X_C)) = E(\mathrm{Cov}(X_B, X_C)) = 2N\mu\sigma_M^2\left(\frac{1}{3}e^{-t/2N}\right. \quad .$$

2. An alternative derivation for variances and covariances in quantitative traits, with further considerations

n

$2N$, $2N$

L

$X_{i,l} \quad l \quad X_i \quad X$

i

$$X_i = \sum_{l=1}^{L} X_{i,l}.$$

$$\begin{aligned}\mathrm{Var}(X_i) &= \sum_{l=1}^{L}\mathrm{Var}\ X_{i,l} \\ &= L\mathrm{Var}\ X_{i,l}\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}\ X_i, X_j &= \sum_{l=1}^{L}\sum_{k=1}^{L}\mathrm{Cov}\ X_{i,l}, X_{j,k} \\ &= \sum_{l=1}^{L}\mathrm{Cov}\ X_{i,l}, X_{j,l} \\ &= L\mathrm{Cov}\ X_{i,l}, X_{j,l}\end{aligned}$$

$$\mathrm{Cov}\ X_{i,l}, X_{j,k} = 0 \quad k \neq l$$

2.1. The variance of a single sample from a species

$X \quad i$

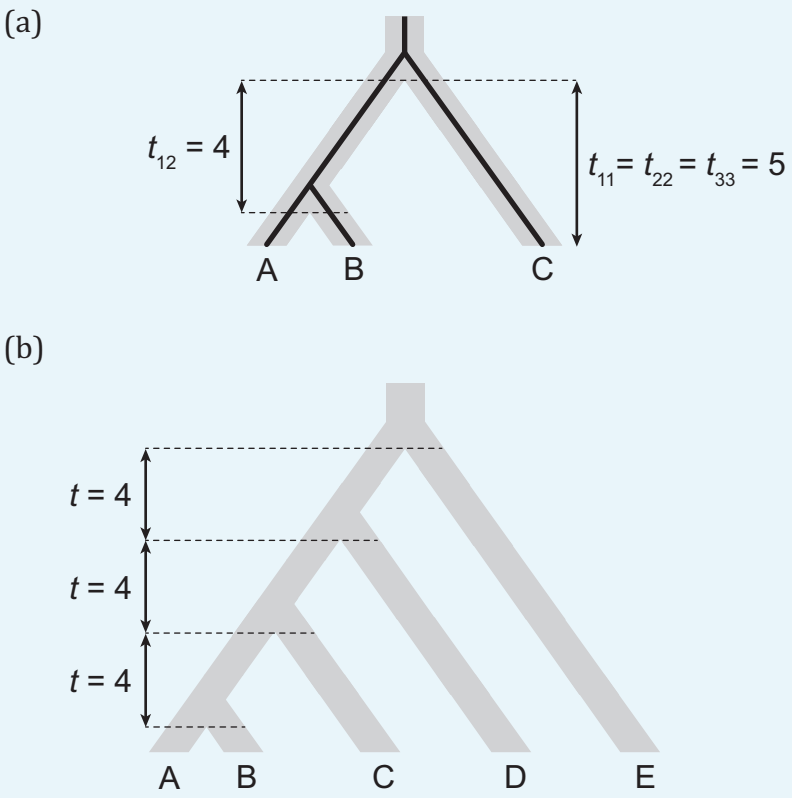
G_l .

$$\mathrm{Var}\ X_{i,l}|G_l = \mu\sigma_M^2 T_{\mathrm{MRCA},l}$$

$T_{\mathrm{MRCA,l}}$ *coalescence*

$T_{\mathrm{MRCA},l}$

Appendix 1—figure 1



Appendix 1—figure 1.

t_{ij} t_{AB} $t_{(AB)C}$ t_{MRCA}
$T_{ij,l}$ $T_{AB,l}$ $T_{(AB)C,l}$ $T_{\mathrm{MRCA},l}$ l

DOI: https://doi.org/10.7554/eLife.36482.013

p_k k *all species*

$$p_2 = 1 - e^{-t/2N} \qquad p_3 = e^{-t/2N}$$

Equations S.3 and S.6

.

$$\begin{aligned}
\mathrm{Var}\ X_{i,l} &= E\ \mathrm{Var}\ X_{i,l}|G_l + \mathrm{Var}\ E\ X_{i,l}|G_l \\
&= E\ \mathrm{Var}\ X_{i,l}|G_l \\
&= E\ \mu\sigma_M^2 T_{\mathrm{MRCA},l} \\
&= \mu\sigma_M^2 \left(t_{\mathrm{MRCA}} + 4N \sum_{k=2}^{n} \left(1 - \tfrac{1}{k}\ p_k \right. \right. \\
&= \mu\sigma_M^2 \left(t_{\mathrm{MRCA}} + 4N - 2N \sum_{k=2}^{n} \tfrac{2}{k} p_k \right. \left. \right)
\end{aligned}$$

t_{MRCA}

Appendix 1—figure 1

$T_{\mathrm{MRCA},l}$

t_{MRCA}

k $4N \sum_{k=2}^{n} 1 - \frac{1}{k}\ p_k$

$4N(1 - 1/k)$ k

$2N$ p_k k

p_k

X .

$$\mathrm{Var}(X_i) = L\mu\sigma_M^2 \quad t_{\mathrm{MRCA}} + 4N - 2N\sum_{k=2}^{n}\left(\frac{2}{k}p_k\right).\left($$

Equations S.14 and S.3

2.2. The covariance between samples from two species

l G_l

.

$$\mathrm{Cov}\ X_{i,l}, X_{j,l}|G_l \ = \mu\sigma_M^2\ T_{\mathrm{MRCA},l} - T_{ij,l}\ ,$$

$T_{ij,l}$ coalescence i j $T_{AB,l}$ **Appendix 1—figure 1** p_k k all species .

$$\begin{aligned}\mathrm{Cov}\ X_{i,l}, X_{j,l} &= E\ \mathrm{Cov}\ X_{i,l}, X_{j,l}|G_l \quad + \mathrm{Cov}\ E\ X_{i,l} \quad , E\ X_{j,l}\\ &= E\ \mathrm{Cov}\ X_{i,l}, X_{j,l}|G_l\\ &= \mu\sigma_M^2 E\ T_{\mathrm{MRCA},l} - T_{ij,l}\\ &= \mu\sigma_M^2\left[\left(t_{\mathrm{MRCA}} + \sum_{k=2}^{n}\left(4N\ 1 - \tfrac{1}{k}\ p_k \quad - \quad t_{ij} + 2N\right.\right.\right.\left(\right.\\ &= \mu\sigma_M^2\left(t_{\mathrm{MRCA}} - t_{ij} + 2N - 2N\sum_{k=2}^{n}\left(\frac{2}{k}p_k\right.\right.\right.\end{aligned}$$

t_{ij} i j t_{AB} **Appendix 1—figure 1** $T_{\mathrm{MRCA},l}$ $T_{ij,l}$.

$$\mathrm{Cov}\ X_i, X_j \ = L\mu\sigma_M^2 \quad t_{\mathrm{MRCA}} - t_{ij} + 2N - 2N\sum_{k=2}^{n}\left(\frac{2}{k}p_k\right).\left($$

Equation S.16 and (S.17)

t_{MRCA} p_k $i = \mathrm{A}$ **Equation S.17** **Equations S.6 and S.7** $j = \mathrm{B}$ $j = \mathrm{C}$

2.3. A comparison of the covariance structure under the Brownian motion and coalescent models

$L\mu\sigma_M^2 = \sigma^2$ σ^2 $L\mu\sigma_M^2 t_{\mathrm{MRCA}} = \sigma^2 t_{\mathrm{MRCA}}$ σ^2 $L\mu\sigma_M^2(t_{\mathrm{MRCA}} - t_{ij}) =$ $\sigma^2(t_{\mathrm{MRCA}} - t_{ij})$ σ^2 ij .

$$\Sigma_{ij} = \Sigma_{ij}^{(BM)} + \quad 2N\ 1+\delta_{ij}\ L\mu\sigma_M^2 - 2NL\mu\sigma_M^2 \sum_{k=2}^{n} \left(\frac{2}{k} p_k\right) \cdot \left($$

$\Sigma_{ij}^{(BM)}$ δ_{ij} $2N\ 1+\delta_{ij}\ L\mu\sigma_M^2 - 2NL\mu\sigma_M^2 \sum_k \left(\frac{2}{k} p_k\right.$ $\Sigma_{ij}^{(BM)}$ Σ_{ij} equally Σ_{ij}

2.3.1. The covariance structure of the BM and coalescent models in the limiting cases of no ILS vs. maximum ILS

$\sum_{k=2}^{n} \frac{2}{k} p_k = E\ \frac{2}{K}$ K $E(K)$.

$E\ \frac{2}{K}$ $\frac{2}{E(K)}$ $\frac{2}{n}$

$E(K) \le n$ $\sum_{k=2}^{n} \left(\frac{2}{k} p_k\right.$ $p_n = 1$ $p_k = 0$ $2 \le k \le n-1$.

$$\sum_{k=2}^{n} \left(\frac{2}{k} p_k\right. = \frac{2}{n}.$$

.

$$\sum_{k=2}^{n} \left(\frac{2}{k} p_k\right. \le \sum_{k=2}^{n} \left(\frac{2}{2} p_k\right.$$
$$= \sum_{k=2}^{n} \left(p_k\right.$$
$$= 1.$$

$p_2 = 1$ $p_k = 0$ $3 \le k \le n$.

$$\sum_{k=2}^{n} \left(\frac{2}{k} p_k\right. = 1.$$

.

$$(\mathrm{Maximum\,ILS}) \frac{2}{n} \le \sum_{k=2}^{n} \left(\frac{1}{k} p_k\right. \le 1 (\mathrm{no\,ILS}).$$

.

$$(\text{No ILS})\Sigma_{ij}^{(BM)} + 2N\delta_{ij}L\mu\sigma_M^2 \leq \Sigma_{ij} \leq \Sigma_{ij}^{(BM)} + 2NL\mu\sigma_M^2\left(1-\frac{2}{n}+\delta_{ij}\right) \quad (\text{Maximum ILS}).$$

$2N\ 1+\delta_{ij}\ L\mu\sigma_M^2$ $2NL\mu\sigma_M^2\sum_k\left(\frac{2}{k}p_k$ Σ_{ij}

$\Sigma_{ij}^{(BM)}$ $i \neq j$

Σ_{ij}

$\Sigma_{ij}^{(BM)}$ *even in the absence of ILS*

$\Sigma_{ij}^{(BM)}$ i j

2.3.2 Asymptotic behavior when internal coalescence is rare

Figure 5b

$$p_n = 1 \qquad p_k = 0 \qquad 2 \leq k \leq n-1\ .$$

$$\sum_{k=2}^{n}\left(\frac{2}{k}p_k = \frac{2}{n}.\right.$$

$$\begin{aligned}\Sigma_{ij} &= \Sigma_{ij}^{(BM)} + 2N\ 1+\delta_{ij}\ L\mu\sigma_M^2 - 2NL\mu\sigma_M^2\frac{2}{n}\\ &= \Sigma_{ij}^{(BM)} + 2NL\mu\sigma_M^2\ 1-\tfrac{2}{n}+\delta_{ij}\\ &\sim 2NL\mu\sigma_M^2\ 1-\tfrac{2}{n}+\delta_{ij}\ ,\end{aligned}$$

N