Syst. Biol. 60(3):358–365, 2011

DOI:10.1093/sysbio/syq097
Advance Access publication on February 8, 2011

Taxon Sampling and the Optimal Rates of Evolution for Phylogenetic Inference

Jeffrey P. Townsend[1,*] and Christoph Leuenberger[2]

[1]*Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06520-8106, USA; and* [2]*Département de mathématiques, Université de Fribourg, Chemin du Musée 3, 1705 Fribourg, Switzerland;*
[*]*Correspondence to be sent to: Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06520-8106, USA; E-mail: jeffrey.townsend@yale.edu.*

Received 13 April 2010; reviews returned 21 July 2010; accepted 20 December 2010
Associate Editor: Richard Glor

Optimal rates of evolution and the informativeness of characters for phylogenetic inference have received increasing attention as the effort to distinguish phylogenetic signal and noise from large data sets intensifies. Ascertaining optimal rates for character evolution and predicting sequences featuring high phylogenetic utility are challenging tasks for which little theory has been developed. We argue for the usage of predictive theoretical tools that identify phylogenetic signal for quartets of taxa. We demonstrate analytically that, under an infinite states model, phylogenetically optimal rates of character evolution increase with greater taxon sampling. Finally, we argue for the development of increasingly sophisticated tools for the prediction of phylogenetic informativeness that incorporate higher taxon sampling and that directly account for phylogenetic noise.

In a recent paper, Klopfstein et al. (2010) evaluate and critique profiles of phylogenetic informativeness (Townsend 2007), a method designed to inform the choice of markers for phylogenetic inference, and to interpret of the power of data sets to resolve short deep internodes in the history of life. To motivate their critique, Klopfstein et al. (2010) provide two analyses. First, they profile the phylogenetic informativeness of CO1 and 28S rRNA and reconstruct phylogeny for diplazontine parasitoid wasps. Second, they perform simulations to test the optimal rate of character change against a prediction based on four-taxon theory in Townsend (2007), addressing data sets of increasing numbers of taxa. With regard to their empirical data set, Klopfstein et al. (2010) find fault with the relation between the quartet-based (Townsend 2007) phylogenetic informativeness profile and the results of phylogenetic inference. Furthermore, they argue that their simulations indicate a dramatic trend in the optimal rate of change of a character for phylogenetic inference, such that the greater the taxon sampling, the slower the optimal rate of change of the characters.

Here, we dispute these claims. Their discussion of the empirical data set does not establish reasonable expectations of the predictor and does not account for an important caveat to the Townsend (2007) phylogenetic informativeness approach. Furthermore, their simulated scenario can be more directly and comprehensively addressed with mathematical analysis. Regardless of how it is addressed, however, their scenario tests for an outcome that is not synonymous with informative data under increased taxon sampling. Consequently, Klopfstein et al. (2010) in fact predict the opposite relation between optimal rate of evolution and degree of taxon sampling to that which should be expected.

The Informativeness and Performance of CO1 and 28S

Klopfstein et al. (2010) remark that, in their analysis of the Townsend (2007) phylogenetic informativeness profiles based on their diplazontine wasp data set, CO1 exhibits an informativeness that is higher across the entire history of their phylogeny, by a factor ranging from four at the root to much more at the tips. In contrast, they show, using the concatenated-alignment tree as the tree for comparison, that CO1 outperforms 28S rRNA by only a few additional nodes “correctly” resolved (33 were resolved correctly by CO1, whereas 26 were resolved correctly by 28S rRNA). However, we would point out that the higher phylogenetic informativeness of CO1 provides no clear expectation for how much better it will perform in terms of specific nodes; that performance depends completely on the distribution of how hard the unresolved nodes are to resolve. For instance, it could be the case that half of the internodes in a given phylogeny are long and recent and the corresponding clades are easy to resolve, and half are extremely short and deep and thus recalcitrant to any resolution. In that case, two genes featuring truly dramatically different informativeness could perform identically in terms of number of nodes resolved. Moreover, even if the distribution of difficulty of resolution across nodes is highly uniform, there is no reason to expect that an x-fold difference in informativeness would result in an x-fold difference in number of nodes resolved. It may require an x-fold difference in amount of informative data to resolve one additional node, an x^2-fold difference to solve two additional nodes, an x^3-fold difference to solve three additional nodes, etc. Although which gene performs better should generally correlate with the quantitative metric of informativeness, the degree and pattern of resolution of internodes as a consequence of that informativeness depends fundamentally on the true length of the relevant internodes.

Klopfstein et al. (2010) more saliently take issue with the fact that the CO1 phylogenetic informativeness profile, while declining at deeper time scales, remains higher than the informativeness profile for 28S rRNA over the most deep nodes (Fig. 1). There are far fewer deep nodes than recent nodes, but if attention is restricted just to the deeper subset of nodes, 28S yielded higher node support than CO1. This acute and valid observation illustrates an issue that informativeness-based studies such as Townsend et al. (2008) have neglected: the importance of the primary caveat expressed in Townsend (2007) to use of the phylogenetic informativeness profile for prediction of the utility of sequences for phylogenetic inference. In particular, the CO1 informativeness profile is on the decline during the deeper epoch of concern, whereas the 28S rRNA informativeness profile is on the rise. As discussed in Townsend (2007), when profiling phylogenetic informativeness to select character sets, the informativeness profile conveys the historical epochs during which a character or set of characters are most likely to provide informative phylogenetic signal but does not discount for the misleading effects of noise (homoplasy) caused by convergence to the same character state in divergent lineages. Such convergence will occur more in faster evolving sites than in slower evolving sites. Thus, for instance, when the height of the informativeness curves are equal at a time in history, character sets are best selected that exhibit phylogenetic informativeness profiles that peak deeper than, rather than more recently than, the epoch of interest. This choice should minimize selection of characters that may have too frequently evolved to convergent states, reducing support for correct nodes and possibly supporting incorrect nodes as well. The issue is that characters that are highly informative early in history rapidly become sources of phylogenetic noise due to multiple hits for deeper divergences. Figure 1 depicts this effect in a purely diagrammatic fashion as a "rain shadow of noise" behind the early peak of informativeness of CO1 from the data set of Klopfstein et al. (2010). Although CO1 shows considerable continued potential for signal based on the height of its profile deep in the phylogeny, it also shows extensive potential for noise. In contrast, the informativeness profile of 28S rRNA is still on the rise, so that its signal is very unlikely to be swamped by noise. The same pattern occurs in the three other studies cited by Klopfstein et al. (2010). The balance of signal and noise based on the rate of evolution of characters is complicated and depends on the lengths of the relevant internodes in a way that cannot be addressed by the asymptotic theory of Townsend (2007), which addresses signal alone.

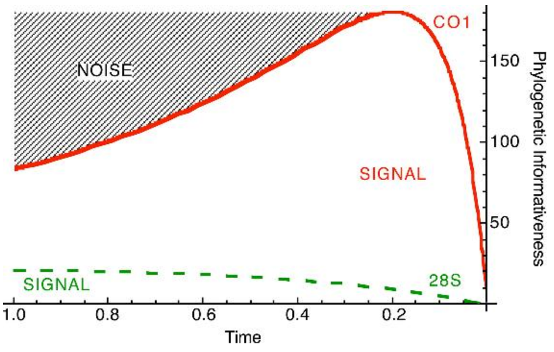

FIGURE 1. The Townsend (2007) phylogenetic informativeness profiles for 28S and CO1 markers from the diplazontine parasitoid wasp data set of Klopfstein et al. (2010). The area under the curves for each gene represents predicted phylogenetic signal. Phylogenetic noise is not characterized by the method of Townsend (2007), however, for illustration, a "rain shadow" of phylogenetic noise has been depicted deeper than the recent peak of informativeness of CO1.

INTERPRETING PROFILES OF PHYLOGENETIC INFORMATIVENESS

Given these considerations, how should profiles of informativeness be interpreted? Viewing a profile yields more than just a quantification of signal that will generally correlate with utility. It also gives a qualitative impression of the potential for phylogenetic noise. Consider two genes whose profile is at the same height for a particular epoch. Whichever gene has an informativeness profile that peaks more deeply and that declines backward in time more slowly would be preferable, as it is predicted to yield less phylogenetic noise. In the most recent portion of a phylogenetic informativeness profile, when the profile is rising as it goes back in time, noise is likely to manifest to a lesser degree compared with signal, and there is less need to consider the impact of phylogenetic noise. Once the informativeness profile has crested, however, there are fewer and fewer sites evolving at the optimal rate for phylogenetic informativeness, and there are increasing numbers of sites that are evolving more rapidly than optimal that are therefore at risk of delivering phylogenetic noise. Sites that provide high signal in recent history deliver noise for deep history; thus, the difference between the recent peak of signal and the level of signal deeper in history is roughly proportional to the degree to which noise will be an issue with a locus at that depth in the phylogeny. However, this proportionality across time cannot be directly compared with the level of signal across time to create a quantitative metric that incorporates noise because the balance of signal and noise depends critically on the length of the internode in question. If the internode is short, noise has a much greater effect. If the internode is long, signal may easily outweigh noise. Thus, there is no way to quantify signal versus noise with a phylogenetic informativeness profile across time alone, even though they both may be depicted as in Figure 1. The relative scaling of their importance depends in a critically important way upon the length of the short deep internode.

A more complete theory that directly integrates noise by estimating a length and thus the relative scaling for this internode is certainly desirable. Until then, signal alone as in the phylogenetic informativeness profile correlates with resolution achievable across loci. Profiles may productively be used as a quantitative metric. Generally such usage will improve inference. (But not always! It would not be a "prediction" if it always predicted correctly.) Thus, experimental design should not only incorporate evaluation of the relative quantification inherent to profiles of phylogenetic informativeness, but when feasible should incorporate judicious discounting of the expected value of sets of characters whose profiles of informativeness peak more recently than the epoch of interest. Moreover, additional theory that accounts for the predicted length of the internodes in question and therefore the scaling relationship between signal and noise should be a high priority.

Informativeness and the Optimal Rate of Evolution with Increased Taxon Sampling

Klopfstein et al. (2010) are also critical of the explicit use of quartet-based taxon sampling in the phylogenetic informativeness derivation by Townsend (2007). They note that calculation of probabilities associated with synapomorphies on short deep internodes in other tree topologies are possible. However, no tree topology with fewer than four taxa features a short deep internode; the three-taxon derivation of Klopfstein et al. (2010) fails to account for an unmodeled root taxon. Only with a root taxon would the strong phylogenetic signal of a deep synapomorphy uniting two pairs of sister taxa manifest. There are both theoretical (Bandelt and Dress 1986) and empirical (Townsend 2007; Mahon and Neigel 2008; Townsend et al. 2008; Schoch et al. 2009) reasons to believe that the quartet results of Townsend (2007) may be productively applied to trees with greater than four taxa. Exact predictions of informativeness for more complex evolutionary histories are challenging to derive. However, some restricted solutions for higher taxon sampling are possible. In general, empirical and simulation results have justified a claim that greater in-group taxon sampling permits the effective usage of faster evolving characters (Graybeal 1998; Hillis 1998; Poe 2003; Hedtke et al. 2006), leading to a prediction of a slightly higher optimal rate. Correspondingly, Townsend and López-Giráldez (2010) derive a faster optimal rate for when an additional taxon is added to character data already collected for a complete quartet.

In contrast, Klopfstein et al. (2010) perform simulations that are interpreted to "show that the optimum evolutionary rate decreases with increasing number of taxa." They simulated data sets for the four-taxon case and for data sets of 8, 16, and 32 taxa (see fig. 1 from Klopfstein et al. 2010), counting the number of times that a nucleotide pattern in accord with only one single change on the short interior branch manifested. They observed that the pattern occurred more frequently with lower and lower rates of character change for the data sets with greater and greater taxon sampling. They thus show that the probability of a never-reversed synapomorphy decreases rapidly with more taxa, as each taxon has a chance of reversing the synapomorphy. The same result may be derived analytically. Given $2n$ species ($n \geq 2$) and their states, we can assign them to the $2n$ leaves of a phylogenetic tree with n exterior branches emanating from each of the two internal nodes (Fig. 3). There are $(2n!)/(2(n!)^2)$ ways to do this. The probability that mutations occur exactly on the short branch only is

$$P(\lambda) = \mathrm{e}^{-2nT\lambda}(1 - \mathrm{e}^{-\lambda\epsilon}).$$

The optimal rate to maximize this probability can be determined analytically:

$$\lambda_{opt} = -\frac{1}{\epsilon}\log\left(1 - \frac{\epsilon}{2nT}\right) \approx \frac{1}{2nT}.$$

(This result can be shown to hold true, asymptotically for $n \rightarrow \infty$, under the assumption of a Jukes–Cantor model as well.) Consistent with the simulations of Klopfstein et al. (2010), this optimal rate tends to 0 as $n \rightarrow \infty$. Consistent with Townsend (2007), $n = 2$ yields $1/(4T)$. However, concluding from these results that "optimal" rates of change for actual phylogenetic inference decrease with increasing taxon sampling is not justified. This pattern of perfect bipartition is necessary for informativeness for a quartet but becomes exceedingly unlikely as the number of taxa increases. For large analyses, such perfect bipartitions are surely not the source of information providing resolution of phylogenetic trees. That information likely lies with the larger numbers of perfectly bipartitioned quartet subtrees that are key to the theory in Townsend (2007). As the number of taxa are increased, it becomes easier to find individual bipartitioned quartets among all the taxa. Although utterly unreversed synapomorphies decrease in probability with increased taxon sampling, observation of a single 2 + 2 outcome, where two taxa share an unreversed synapomorphy and two taxa on the other side of an internode share an ancestral state or parallel synapomorphy, should increase in probability with greater taxon sampling. Below, we analytically demonstrate this increase.

Klopfstein et al. (2010) claim that the optimal rate for a character decreases with increased taxon sampling. If so, the optimal rate should be lower for five taxa than for four taxa. However, calculation reveals that such a 2 + 2 outcome occurs more often with higher rates in a five-taxon tree for which all taxa are added de novo to the study. Consider the goal of resolving the AB/CD quartet in the tree depicted in Figure 2, where

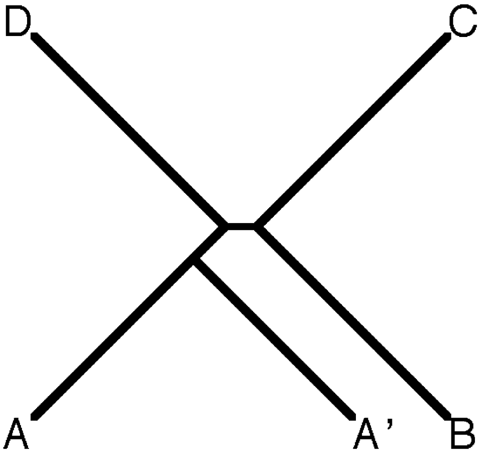

FIGURE 2. A five-taxon phylogenetic tree in which the quartet of interest are composed of taxa labeled A, B, C, and D. The additional taxon that will also be sampled is labeled A'.

five species (A, B, C, D, and A′) are to be sampled. Taxon sampling is assumed to be deep because recent sampling will have minimal impact on inference of the deep internode (Graybeal 1998; Poe 2003; Townsend and López-Giráldez 2010). For the leaves to exhibit informative states, at least one substitution must occur along the short internal branch (ϵ), no mutations may occur along the exterior branches of B, C, and D, and mutations along at most one of the exterior branches A and A′ may occur. The probability for this event to occur is

$$P(\lambda) = e^{-3T\lambda}(1-e^{-\epsilon\lambda})(1-(1-e^{-T\lambda})^2)$$
$$\approx \epsilon\lambda(2e^{-4T\lambda} - e^{-5T\lambda})$$

for small ϵ. This probability is maximal for $\lambda_{\text{opt}} = 0.294/T$, which is, as predicted, slightly larger than the $1/(4T)$ revealed for a simple quartet in Townsend (2007). Consistent with this finding, an increase in the optimal rate has also been demonstrated for the addition of character data for a fifth taxon to a known unresolved quartet with previously identified quartet character data (Townsend and López-Giráldez 2010).

Addressing trees larger than five taxa becomes increasingly technically difficult. However, two divergent simplified models, a "many-sister-pairs" model and a "two-hard-polytomy" model are tractable for arbitrarily high levels of taxon sampling. Any tree with exclusively deep taxon sampling should lie somewhere between these two models. Consider a many-sister-pairs tree (Fig. 3) for which we require at least one mutation to occur on all n or on all but one of the short interior branches from the root (each of length ϵ), but we conservatively require none on to occur on any external branches. This assumption is conservative because the direction of our model error due to this assumption would favor slower rates in general and thus would err on the side of favoring the interpretation of Klopfstein et al. (2010). The probability of this event under these assumptions is

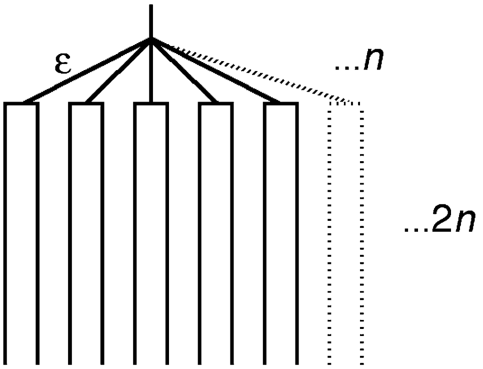

FIGURE 3. Phylogenetic tree with $2n$ paired taxa with leaf length T, with all pairs connecting to a star node via n internodes of length ϵ.

$$P(\lambda) = ((1-e^{-\epsilon\lambda})^n + n\, e^{-\epsilon\lambda}(1-e^{-\epsilon\lambda})^{n-1})e^{-2nT\lambda}$$
$$\approx n\epsilon^{n-1}\lambda^{n-1}e^{-2nT\lambda}.$$

Taking the derivative yields an optimal rate of

$$\lambda_{\text{opt}} \approx \frac{1}{2T} - \frac{1}{2nT},$$

which agrees with Townsend (2007) for $n=2$, and monotonically increases with n, in stark contrast to the conclusion of Klopfstein et al. (2010).

It could be argued that requiring so many sister clades to be resolved (i.e., requiring mutations to have occurred on so many short branches) places an unrealistic optimality on higher rates. This argument may be countered by considering the opposite extreme to it, a two-hard-polytomies model (Fig. 4), again with arbitrarily large taxon sampling. The tree in Figure 4 features a single short deep internode of length ϵ that separates two-hard polytomies, one with n_A subtending lineages and the other with n_B subtending lineages. To produce truly synapomorphic states, one or more mutations must occur along the short interior branch, with probability $1 - e^{-\lambda\epsilon} \approx \lambda\epsilon$, where the approximation holds for small values of epsilon. Also, to produce truly synapomorphic states, two or more branches on either side of the short deep internode must remain unchanged over time T, each with probability $e^{-\lambda T}$. The number of subtending branches exhibiting a state that traces its ancestry without interruption to the internal branch then follows a binomial distribution: n_A tries each with probability $e^{-\lambda T}$ on the left side of the short internal branch and n_B tries each with probability $e^{-\lambda T}$ on the right side of the short internal branch. By defining the optimal rate as that which maximizes the number of taxa that are separable at the internal branch and taking into account the need for two or more taxa to exhibit common ancestry on each side, the optimal rate can be derived (see Appendix).

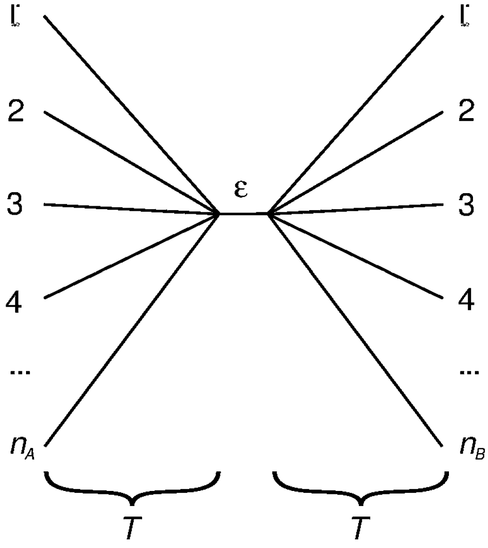

FIGURE 4. Phylogenetic tree composed of a deep short internode of length ϵ adjoining two star nodes with n_A and n_B subtending taxa, each with leaf length T.

Within the two-hard-polytomy model, two extreme cases accounting for arbitrarily high taxon sampling exemplify the results. For a fully balanced tree in which $n_A = n_B$, the only case for which this optimum can be derived analytically is $n = 2$, and the result is $\lambda_{\text{opt}} = 1/(4T)$, which was already derived in Townsend (2007). For larger n, the optimal value can still be determined numerically (Fig. 5a). The values of λ_{opt} increase with greater n from $1/(4T)$ toward a limit of $1/T$. Thus, for this balanced tree, the optimal rate of character change for revealing states that reflect common ancestry increases with increasing taxon sampling. The other extremal case is that of a completely unbalanced tree in which $n_B = 2$. The optimal rate λ_{opt} for this case is also derived in the Appendix, and numerical analyses of that result are charted across a range of levels of taxon sampling in Figure 5b. The rate first grows from $\lambda_{\text{opt}} = 1/(4T)$ (cf. Townsend 2007) to a maximal value $\lambda_{\text{opt}} = 0.366/T$ for $n_A = 7$, then gently falls to the limit value $1/(3T)$ as $n_A \to \infty$.

Intermediately balanced trees for general group sizes n_A and n_B would presumably fall between these two limits. Further numerical analysis and asymptotic results on general group sizes (see Appendix) are fully consistent with an optimal rate that always increases above $1/(4T)$ and that in most cases, monotonically rises with increasing taxon sampling.

Optimality of Rate in Finite-State Models

Infinite states models like those above accurately characterize the probability of a true signal corresponding to to an unreversed synapomorphy, but do not additionally discount for positively misleading data that can arise as a consequence of convergence of character state (homoplasy). The theory of optimal rates in the case of Markov substitution models (like the Jukes–Cantor model, see Felsenstein 2004) is more involved than in the case of an infinite states model. In the second part of the Appendix, we demonstrate that analysis of a finite-state model continues to yield an increase in the optimal rate with increased taxon sampling. The corresponding profiles for two simple tree topologies (Fig. 6) demonstrate that under the assumption of a Jukes–Cantor model of base substitution, the optimal rate increases with additional taxon sampling.

Conclusions

We have demonstrated that the claim that optimal rates of character change decrease with increased taxon sampling is unfounded. Using an infinite states model that accurately characterizes the probability of observing states that are identical by descent, we have presented analytical results supporting our point of view. A finite-state model, as well as an extensive empirical analysis and simulation-based literature applying finite-states models (Graybeal 1998; Hillis 1998; Poe 2003; Hedtke et al. 2006), agrees with our infinite-states models, arguing that increased taxon sampling permits the productive usage of faster evolving characters for phylogenetics. On the other hand, we agree with (Klopfstein et al. 2010) that, as is the case for all products of limited

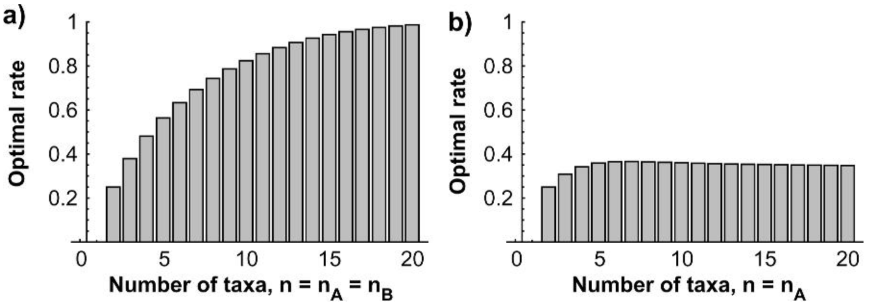

FIGURE 5. a) Optimal rates for balanced trees (with number of taxa $n = n_A = n_B$) for an infinite states model. b) Optimal rates of unbalanced trees (with number of taxa $n = n_A, n_B = 2$).

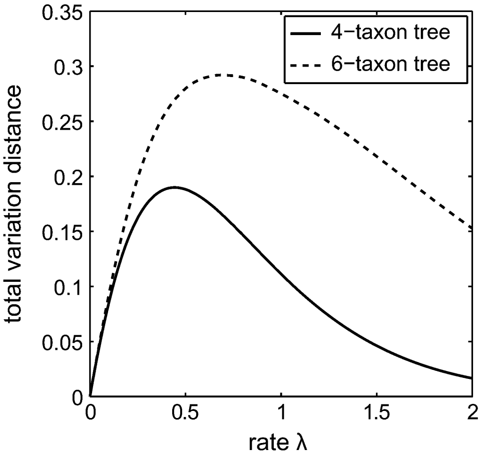

FIGURE 6. Variation distances $\mathbb{E}(\Delta^+)/\epsilon$ for a four-taxon tree and for a six-taxon tree.

models, profiles of phylogenetic informativeness should be used with careful consideration of the expressed caveats for their performance. In particular, profiles of phylogenetic informativeness only indicate the epochs in which phylogenetic signal is expected to be maximized for quartets and do not discount for likely phylogenetic noise from the convergence of fast-evolving sites. This lack of discounting is likely to be especially egregious for genes like cytochrome b (Townsend et al. 2008) or cytochrome oxidase (Klopfstein et al. 2010) that exhibit extreme variance in rates across sites and that therefore are likely to mingle signal and noise on similar time scales. In our point of view, however, discontinuing use of phylogenetic informativeness profiles without a suitable alternative is deeply unwise. They are a valuable tool that need only be conscientiously applied to improve phylogenetic experimental design. Future theoretical advances may chip away at the assumptions underlying phylogenetic informativeness profiles, leading to methods for exhaustively or partially quantifying the effects on predictions of utility of increased taxon sampling or phylogenetic noise induced by convergence to a common state. Until more comprehensive methodologies become available that predict the utility of phylogenetic markers, for the purposes of experimental design, research studies should generally trend toward use of characters that are evolving slower than the optimum to avoid phylogenetic noise. At the same time, a research study involving large numbers of deeply branching taxa should trend from slower than the quartet-based optimum toward the selection of faster evolving character sets. Thus, experimental design employing a quantitative profile of the phylogenetic signal should continue to be accompanied by a thoughtful understanding of focused issues relating to taxon sampling, tree structure, and phylogenetic noise.

REFERENCES

Bandelt H., Dress A. 1986. Reconstructing the shape of a tree from observed dissimilarity data. Adv. Appl. Math. 7(3):309–343.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47(1):9–17.

Hedtke S., Townsend T., Hillis D. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55(3):522–529.

Hillis D. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47(1):3–8.

Klopfstein S., Kropf C., Quicke D. 2010. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). Syst. Biol. 59(2):226–241.

Mahon B., Neigel J. 2008. Utility of arginine kinase for resolution of phylogenetic relationships among brachyuran genera and families. Mol. Phylogenet. Evol. 48(2):718–727.

Poe S. (2003). Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52(3):423–428.

Schoch C., Sung G., López-Giráldez F., Townsend J., Miadlikowska J., Hofstetter V., Robbertse B., Matheny P., Kauff F., Wang Z., Gueidan C., Andrie R., Trippe K., Ciufetti L., Wynns A., Fraker, E., Hodkinson B., Bonito G., Groenewald J., Arzanlou M., de Hoog G., Crous P., Hewitt D., Pfister D., Peterson K., Gryzenhout M., Wingfield M., Aptroot A., Suh S., Blackwell M., Hillis D., Griffith G., Castlebury L., Rossman A., Lumbsch H., Lücking R., Bdel B., Rauhut A., Diederich P., Ertz D., Geiser D., Hosaka K., Inderbitzin P., Kohlmeyer J., Volkmann-Kohlmeyer B., Mostert L., O'Donnell K., Sipman H., Rogers J., Shoemaker R., Sugiyama J., Summerbell R., Untereiner W., Johnston P., Stenroos S., Zuccaro A., Dyer P., Crittenden P., Cole M., Hansen K., Trappe J., Yahr R., Lutzoni F., Spatafora, J. 2009. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. Syst. Biol. 58(2):224.

Townsend J. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56(2):222.

Townsend J., López-Giráldez F. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst. Biol. 59(4):446-457.

Townsend J., López-Giráldez F., Friedman R. 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol. 67(5):437–447.

APPENDIX

The Optimal Substitution Rate for Two-Group Trees

Infinite alleles model.—We consider an ultrametric tree with two internal nodes a and b at height T, connected, via the root, by a short internal branch of length $\epsilon \ll T$. Node a is connected to $n_A \geq 2$ external branches of lengths T and node b to $n_B \geq 2$ external branches of lengths T, totaling $n_A + n_B \geq 4$ leaves. The short internal branch thus separates the leaves into the two groups A and B. Each character evolves along the tree according to an infinite states model with substitution rate λ. For any two separable taxa, we would be able to tell from the data whether or not they belong to the same group. Our goal is to derive the optimal substitution rate λ_{opt} for which the expected number of separable taxa is maximal.

To produce truly synapomorphic states, one or more substitutions must occur along the short interior branch. Let Z be an indicator variable taking the value 1 in this case, such that $\mathbb{P}(Z=1) = 1 - e^{-\lambda\epsilon} \approx \lambda\epsilon$, where the approximation holds for small values of epsilon. Also, to produce truly synapomorphic states, two or more branches on either side of the short deep internode must remain unchanged over time T, each with probability $\mathrm{e}^{-\lambda T}$.

To unburden the notation, we will from now on assume the normalization $T = 1$. For the case of general T, the optimal rates that we derived must simply be divided by T, for example, instead of $\lambda_{\text{opt}} = 1/4$ we have $\lambda_{\text{opt}} = 1/4T$.

Let X_A be a random variable that counts the number of branches from group A on which there have been *no* substitutions. X_A follows the binomial distribution

$$X_A \sim \mathcal{B}(n_A, \mathrm{e}^{-\lambda}).$$

The random variable X_B is similarly defined for group B and is binomially distributed as $\mathcal{B}(n_B, \mathrm{e}^{-\lambda})$.

For the given tree topology, there are separable taxa if and only if the data contain two states α and β, which both occur more than once. In that case, there have been one or more substitutions along the interior branch and no substitutions exactly on the branches ending in states α (in group A) and β (in group B). On all other branches, there have been substitutions. The number of separable taxa is then $x_A + x_B$. The random variable S counting the separable taxa is thus defined by

$$S = X_\alpha + X_\beta = \begin{cases} X_A + X_B, & \text{if } X_A, X_B \geq 2, \text{ and } Z = 1; \\ 0, & \text{else,} \end{cases}$$

where X_α and X_β denote the number of taxa in state α and β, respectively. Here,

$$X_\alpha = X_A \cdot \operatorname{Ind}(X_A \geq 2, X_B \geq 2, Z = 1)$$

and

$$X_\beta = X_B \cdot \operatorname{Ind}(X_A \geq 2, X_B \geq 2, Z = 1).$$

Here, Ind(event) denotes an indicator variable that takes value 1 if event takes place and value 0 otherwise. The distribution of X_α (for $k = 2, \ldots, n_A$) is then

$$\begin{aligned} \mathbb{P}(X_\alpha = k) &= \mathbb{P}(Z = 1) \cdot \mathbb{P}(X_B \geq 2) \cdot \mathbb{P}(X_A = k) \\ &= c(\lambda, n_B) \cdot \binom{n_A}{k} \mathrm{e}^{-k\lambda}(1 - \mathrm{e}^{-\lambda})^{n_A - k}, \end{aligned}$$

where

$$\begin{aligned} c(\lambda, n_B) &= \mathbb{P}(Z = 1) \cdot \mathbb{P}(X_B \geq 2) \\ &= \lambda\epsilon(1 - (1 - \mathrm{e}^{-\lambda})^{n_B} - n_B\, \mathrm{e}^{-\lambda}(1 - \mathrm{e}^{-\lambda})^{n_B - 1}). \end{aligned} \tag{A.1}$$

The distribution of X_β is defined similarly. The optimal rate, by definition, maximizes the expectation of S:

$$\lambda_{\text{opt}} = \underset{\lambda}{\operatorname{argmax}}\, \mathbb{E}(S).$$

Clearly, $\mathbb{E}(S) = \mathbb{E}(X_\alpha) + \mathbb{E}(X_\beta)$. Moreover, we have

$$\begin{aligned} \mathbb{E}(X_\alpha) &= c(\lambda, n_B) \sum_{k=2}^{n_A} k \binom{n_A}{k} \mathrm{e}^{-k\lambda}(1 - \mathrm{e}^{-\lambda})^{n_A - k} \\ &= c(\lambda, n_B) \left(\sum_{k=0}^{n_A} k \binom{n_A}{k} \mathrm{e}^{-k\lambda}(1 - \mathrm{e}^{-\lambda})^{n_A - k} \right. \\ &\qquad \left. - n_A\, \mathrm{e}^{-\lambda}(1 - \mathrm{e}^{-\lambda})^{n_A - 1} \right) \\ &= c(\lambda, n_B)(\mathbb{E}(X_A) - n_A \mathrm{e}^{-\lambda}(1 - \mathrm{e}^{-\lambda})^{n_A - 1}). \end{aligned}$$

Because X_A is a binomial variable, $\mathbb{E}(X_A) = n_A\, \mathrm{e}^{-\lambda}$. Thus,

$$\mathbb{E}(X_\alpha) = c(\lambda, n_B) n_A\, \mathrm{e}^{-\lambda}(1 - (1 - \mathrm{e}^{-\lambda})^{n_A - 1}). \tag{A.2}$$

We will first consider in more detail two extreme cases: Totally balanced trees for which $n_A = n_B$ and totally unbalanced trees for which $n_B = 2$.

For a totally balanced tree, we have by symmetry $\mathbb{E}(S) = 2\mathbb{E}(X_\alpha)$. Thus, λ_{opt} maximizes expression (A.2) and we define $n = n_A = n_B$. The only case for which this optimum can be found analytically is $n = 2$ and the result is $\lambda_{\text{opt}} = 1/4$. This result has already been derived in Townsend (2007). For larger n, the optimal value can be determined numerically. As can be seen from Figure 5a, the values of λ_{opt} increase—with increasing n—from $\lambda_{\text{opt}} = 1/4$ toward the limit of $\lambda_{\text{opt}} = 1$.

In the case of an extremely unbalanced tree, $n_B = 2$,

$$c(\lambda, 2) = \lambda\epsilon\, \mathrm{e}^{-2\lambda}$$

and thus

$$\mathbb{E}(X_\alpha) = \lambda\epsilon n_A\, \mathrm{e}^{-3\lambda}(1 - (1 - \mathrm{e}^{-\lambda})^{n_A - 1}).$$

Switching the roles of n_A and n_B in equation (A.2),

$$\begin{aligned} \mathbb{E}(X_\beta) &= c(\lambda, n_A) 2\, \mathrm{e}^{-\lambda}(1 - (1 - \mathrm{e}^{-\lambda})^{2-1}) \\ &= 2\lambda\epsilon\, \mathrm{e}^{-2\lambda}(1 - (1 - \mathrm{e}^{-\lambda})^{n_A} \\ &\qquad - n_A\, \mathrm{e}^{-\lambda}(1 - \mathrm{e}^{-\lambda})^{n_A - 1}). \end{aligned}$$

Putting these together,

$$\begin{aligned} \mathbb{E}(S) &= \mathbb{E}(X_\alpha + X_\beta) \\ &= \lambda\epsilon\, \mathrm{e}^{-2\lambda}[n_A\, \mathrm{e}^{-\lambda}(1 - 3(1 - \mathrm{e}^{-\lambda})^{n_A - 1}). \\ &\qquad + 2(1 - (1 - \mathrm{e}^{-\lambda})^{n_A})]. \end{aligned}$$

Figure 5b displays the dependence of λ_{opt} on n_A for fixed $n_B = 2$. The rate first grows from $\lambda_{\text{opt}} = 1/4$ to a maximal value $\lambda_{\text{opt}} = 0.366$ for $n_A = 7$ and then gently falls to the limit value $1/3$ as $n_A \to \infty$.

Let us finally consider the case of general group sizes n_A and n_B. Define $\lambda_{\text{opt}}^{n_A, n_B}$ as the value which maximizes the expectation

$$\begin{aligned}\mathbb{E}(S) &= \mathbb{E}(X_\alpha + X_\beta) \qquad \text{(A.3)}\\ &= \lambda\epsilon\, \mathrm{e}^{-\lambda} n_A[(1-(1-\mathrm{e}^{-\lambda})^{n_A-1})\cdot\ldots\\ &\quad \ldots\cdot(1-(1-\mathrm{e}^{-\lambda})^{n_B} - n_B\,\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})^{n_B-1})]\\ &\quad + \lambda\epsilon\, \mathrm{e}^{-\lambda} n_B[(1-(1-\mathrm{e}^{-\lambda})^{n_B-1})\cdot\ldots\\ &\quad \ldots\cdot(1-(1-\mathrm{e}^{-\lambda})^{n_A} - n_A\,\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})^{n_A-1})].\end{aligned}$$

The maximum of this expression can be determined numerically but not analytically. We can state the following asymptotic results. If n_A is very large compared with n_B, then we need only consider the dominant terms in the above expression and we get that $\lambda_{\text{opt}}^{n_A,n_B}$ approximately maximizes the term

$$\begin{aligned}\mathbb{E}(S) \approx \mathbb{E}(X_\alpha) \approx \lambda\epsilon\, \mathrm{e}^{-\lambda} n_A(1-(1-\mathrm{e}^{-\lambda})^{n_B}\\ - n_B\,\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})^{n_B-1}).\end{aligned}$$

We define

$$\lambda_{\text{opt}}^{\infty,n_B} := \lim_{n_A\to\infty} \lambda_{\text{opt}}^{n_A,n_B}.$$

Some values that have been obtained numerically follow:

$$\lambda_{\text{opt}}^{\infty,2} = 0.333,\ \lambda_{\text{opt}}^{\infty,3} = 0.444,$$

$$\lambda_{\text{opt}}^{\infty,4} = 0.533,\ \lambda_{\text{opt}}^{\infty,5} = 0.607,$$

$$\lambda_{\text{opt}}^{\infty,10} = 0.841,\ \lambda_{\text{opt}}^{\infty,15} = 0.949.$$

These values are strictly increasing with increasing n_B. In particular, one can show that they tend to the limit $\lambda_{\text{opt}}^{\infty,\infty} = 1$. For a very large number of taxa, the optimal rate is thus close to 1 as long as both groups are relatively large. If n_A and n_B are large, we also can approximate from equation (A.3) that if the rate is optimal (i.e., ≈ 1),

$$\mathbb{E}(S) \approx \frac{\epsilon}{e}(n_A + n_B) = 0.368\epsilon(n_A + n_B).$$

This expectation means that for two-group trees with many long external branches and for which neither of the two groups is composed of few taxa, the optimal rate is roughly 1, and the expected fraction of separable taxa is 0.37 times the fraction of the short interior branch with respect to the long external branches.

Markov models of base substitution.—Let $\mathcal{T}$ be a phylogenetic tree with n leaves and any topology. The data at the leaves of $\mathcal{T}$ are created by a Markov model of base substitution with Poisson rate λ and rate matrix $\mathbf{Q}$. We denote by D a vector of states at the leaves of the tree (e.g., $D = \{AATG\}$ for $n = 4$).

We concentrate on an interior branch of length ϵ whose nodes we call a and b. Let $\mathcal{E}$ be the event ("signal") that the states at a and b differ. As we are interested in short branches only, we write

$$\mathbb{P}(\mathcal{E}) = \epsilon\lambda + O(\epsilon^2).$$

One can think of $\mathbb{P}(\mathcal{E})$ as the "prior" probability of the signal. When data D are observed, the corresponding "posterior" probability of the signal is given by Bayes' formula:

$$\mathbb{P}(\mathcal{E}|D) = \frac{\mathbb{P}(D|\mathcal{E})\mathbb{P}(\mathcal{E})}{\mathbb{P}(D)}.$$

We call a state vector D *informative* for the signal $\mathcal{E}$ if the posterior exceeds the prior, that is,

$$\mathbb{P}(\mathcal{E}|D) > \mathbb{P}(\mathcal{E})$$

or equivalently $\mathbb{P}(D|\mathcal{E}) > \mathbb{P}(D)$. Let us define the "excess of posterior over prior" by

$$\Delta^+(D) := \max\{0, \mathbb{P}(\mathcal{E}|D) - \mathbb{P}(\mathcal{E})\}. \qquad \text{(A.4)}$$

One can view $\Delta^+(D)$ as a measure of how much additional knowledge the observation of D adds as far as the detection of the signal is concerned. Noninformative state vectors add nothing, only informative state vectors contain a surplus of information over the prior probability. Loosely speaking, an informative state with, say, $\Delta^+ = 4\%$ is doubly as precious as an informative state with $\Delta^+ = 2\%$ because it gives us double us much "extra knowledge" over the prior probability of $\mathcal{E}$. The expectation $\mathbb{E}(\Delta^+)$ of the random variable Δ^+ represents the mean informativeness per site. A mutation rate λ which maximizes $\mathbb{E}(\Delta^+)$ will be most informative as far as detection of $\mathcal{E}$ is concerned because it contains the greatest mean informativeness per site. One can prove that

$$\begin{aligned}\mathbb{E}(\Delta^+) &= \frac{\epsilon\lambda}{2}\sum_{D\in\mathcal{D}} |\mathbb{P}(D|\mathcal{E}) - \mathbb{P}(D|\overline{\mathcal{E}})| + O(\epsilon^2)\\ &= \epsilon\lambda\cdot \mathrm{TV}(\mathbb{P}(\cdot|\mathcal{E}), \mathbb{P}(\cdot|\overline{\mathcal{E}})) + O(\epsilon^2), \qquad \text{(A.5)}\end{aligned}$$

where $\mathrm{TV}(\cdot,\cdot)$ denotes the total variation distance between the two probability measures.

In Figure 6, we display the function $\mathbb{E}(\Delta^+)/\epsilon$ in dependence of the rate λ for two toy examples of phylogenetic trees with an underlying Jukes–Cantor model of base substitution (for which the model specifications are $\pi_i = 1/4$, $q_{ij} = 1/3$): a four-taxon tree with $n_A = n_B = 2$ branches of length $T = 1$ each and a six-taxon tree with $n_A = n_B = 3$ branches of length $T = 1$. Observe that the optimal rate increases with additional taxon sampling. For the four-taxon tree, 84 of the 256 state vectors qualify as informative. For small rates λ only, the *AACC*-pattern states have an appreciable posterior excess Δ^+, although for larger rates, other state vectors gain in relative importance. Table 1 gives a few numerical values.

TABLE 1. Informativeness of states in four-taxon tree

λ	$\Delta^+(AACC)$	$\Delta^+(AACG)$	$\Delta^+(ACGT)$
0.1	56.1×10^{-3}	3.7×10^{-3}	0.2×10^{-3}
0.3	26.1×10^{-3}	5.2×10^{-3}	1.0×10^{-3}
0.5	12.6×10^{-3}	4.0×10^{-3}	1.3×10^{-3}
1.0	2.3×10^{-3}	1.3×10^{-3}	0.8×10^{-3}

Syst. Biol. 60(3):366–375, 2011

DOI:10.1093/sysbio/syq099
Advance Access publication on February 4, 2011

Bayesian Estimation of Substitution Rates from Ancient DNA Sequences with Low Information Content

Simon Y. W. Ho[1,2,*], Robert Lanfear[1], Matthew J. Phillips[1], Ian Barnes[3], Jessica A. Thomas[1,3], Sergios-Orestis Kolokotronis[4], and Beth Shapiro[5]

[1]*Centre for Macroevolution and Macroecology, Research School of Biology, Australian National University, Canberra, ACT 0200, Australia;*
[2]*School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia;*
[3]*School of Biological Sciences, Royal Holloway University of London, Egham, Surrey TW20 0EX, UK;*
[4]*Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA; and*
[5]*Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA;*
Correspondence to be sent to: School of Biological Sciences, Macleay Building A12, University of Sydney, Sydney, NSW 2006, Australia; E-mail: simon.ho@sydney.edu.au.

Received 17 September 2009; reviews returned 28 February 2010; accepted 3 November 2010
Associate Editor: Kelly Zamudio

Rates of molecular evolution have been shown to vary significantly among nucleotide sites, loci, and taxa. In addition to these forms of rate heterogeneity, there is evidence that molecular rates vary with the timescale over which they are estimated. One of the most striking observations has been that of elevated mutation rates over very short timescales, such as those presented in studies of pedigrees (e.g., Howell et al. 2003; Millar et al. 2008) and mutation accumulation lines (e.g., Denver et al. 2000; Haag-Liautard et al. 2008). In contrast, much lower rates are observed over evolutionary timescales, as estimated in phylogenetic analyses calibrated with reference to paleontological or geological data.

The disparity between rates of spontaneous mutation and evolutionary substitution can exceed an order of magnitude. Intermediate rates are expected between these two ends of the spectrum, but there has been disagreement over the exact form of the decline from the mutation rate to the substitution rate. Some authors have suggested that elevated mutation rates are very short-lived, perhaps persisting for only a small number of generations (Macaulay et al. 1997; Gibbons 1998). More recently, it was proposed that the estimated rate decays exponentially over tens to hundreds of thousands of years, producing a "time dependence" of rates, whereby the magnitude of the inferred rate depends on the age of the calibration used in the analysis (Ho et al. 2005, 2007c; Penny 2005; Ho and Larson 2006). Although some of the original evidence for this hypothesis has been challenged (Emerson 2007; Bandelt 2008), there has been a steady accumulation of empirical and theoretical support for a prolonged elevation of short-term rates (e.g., Genner et al. 2007; Burridge et al. 2008; Henn et al. 2009; Peterson and Masel 2009; Soares et al. 2009). This has included compelling evidence from analyses of ancient DNA (aDNA) in which the sampling times of the heterochronous sequences are able to provide calibrating information for estimating rates (e.g., Lambert et al. 2002; Barnes et al. 2007; Ho et al. 2007b; Hay et al. 2008; Subramanian et al. 2009a).

In a recent critique, Debruyne and Poinar (2009) have claimed that the high rate estimates obtained in Bayesian analyses of aDNA data are an unintended consequence of analyzing short sequences. According to their "signal-dependent artifact" hypothesis, aDNA-based rate estimates depend almost entirely on the information content in the sequence alignment. The essence of the criticism is that the posterior distribution of the rate becomes so wide that the posterior mean becomes an upwardly biased estimator. This behavior has been noted in previous studies of heterochronous data with low information content (e.g., Ho et al. 2007c; Firth et al. 2010). However, Debruyne and Poinar go on to state that "the [rate] acceleration phenomenon is certainly of much lower magnitude than has been previously reported by Ho et al. (2005)" (p. 358). This is a misleading comparison because our study was based almost exclusively on analyses of modern DNA (isochronous sequences) using internal-node calibrations which, as argued by Debruyne and Poinar, are able to overcome the signal-dependent artifact. In fact, much of the evidence for time-dependent rates has come from analyses of isochronous data (e.g., Genner et al. 2007; Burridge et al. 2008; Henn et al. 2009; Soares et al. 2009; Papadopoulou et al. 2010).

Resolving the concerns over Bayesian rate estimates from aDNA is important for several reasons. First, aDNA sequences typically range in age from 10^2 to 10^5 years, thereby filling a crucial calibration gap between the time periods covered by pedigrees (usually $< 10^2$ years) and fossil-calibrated species phylogenies (usually $> 10^6$ years). Second, because the ages of aDNA sequences can provide sufficient calibrating information for estimating rates (Drummond et al. 2003), these data make it possible to circumvent problems associated with choosing and implementing calibrations at internal nodes (Emerson 2007; Ho and Phillips 2009; Firth et al. 2010). In particular, terminal-node calibrations remove the need for assumptions about genetic divergence being correlated with the divergence of species or populations, which can be dubious because of the uncertainty posed by ancestral polymorphism (Charlesworth et al. 2005; Peterson and Masel 2009). Consequently, if rates can be accurately estimated from aDNA data, some insight can be gained into the underlying causes of time-dependent rates (Ho et al. 2007c).

There is still uncertainty regarding the factors driving the time dependence of rates. Previous studies have considered the possibility of contributions from incomplete purifying selection, calibration error, sequencing error, aDNA damage, ancestral polymorphism, saturation, and model misspecification, among others (Ho et al. 2005, 2007c; Woodhams 2006; Henn et al. 2009; Loogväli et al. 2009; Peterson and Masel 2009; Soares et al. 2009; Subramanian et al. 2009a). Debruyne and Poinar have added to this list with their suggestion that mean posterior rate estimates are upwardly biased for data sets with low information content. Distinguishing among these various factors is crucial to future studies of recent divergence times, evolutionary rates, and the molecular evolutionary process in general.

Below, we investigate the two major aspects of the critique by Debruyne and Poinar. The first of these is that the posterior mean provides a biased measure of the rate in Bayesian analyses of data sets with low information content. To examine this issue, we perform new analyses of sequence data simulated using known evolutionary parameters. We assess the relationship between sequence variation and estimated rate under a range of simulation conditions, including various rates and sequence lengths.

The second major aspect of the critique by Debruyne and Poinar is that the artifactual rate estimates from aDNA data are governed by the information content of the alignments, as measured by the number of variable sites. Indeed, Debruyne and Poinar base their entire signal-dependence model on analyses of alignments of varying length, which they regard as a suitable proxy for information content. Although this might be appropriate for isochronous data, we argue that it does not provide the full picture for heterochronous data because the ages of the tips represent a crucial part of the phylogenetic and temporal signal. We propose that the amount of information contained in these ages depends on their structure and spread, including the length of the sampling interval in relation to the period spanned by the genealogy of the sequences (Drummond et al. 2003; Firth et al. 2010). To investigate this, we perform new analyses of 18 published aDNA alignments to assess whether the ages of the sequences in these data sets provide sufficient calibrating information for estimating rates. The results of these analyses show that most real data sets appear to have satisfactory temporal structure and signal.

The results of our new analyses indicate that the "signal dependence" hypothesis has limited relevance to the majority of real aDNA data sets. Our results also suggest that the signal dependence cannot be regarded as an analogue to time dependence, unless one is willing to accept the validity of equating alignment length with temporal depth in aDNA data. Moreover, our results highlight the importance of other factors, including the distribution of sampling times, choice of population size prior, and the use of appropriate summary statistics in analyses of heterochronous sequence data that exhibit low variation.

New Analyses in Response to Debruyne and Poinar

Here, we build upon a simulation study that was presented in one of our previous evaluations of Bayesian rate estimation using aDNA data (Ho et al. 2007b). Debruyne and Poinar have challenged the results of this study, criticising two aspects of our analyses. First, they argue that the rates estimated from the simulated data are more precise than those obtained from real aDNA data. Although this observation is correct, these results are an expected consequence of simulation-based analysis: the evolutionary models for nucleotide substitution and demographic history used in the analysis of the simulated data are chosen to match the conditions under which the data were generated. This is adopted as standard practice to make it easier to isolate the effects of the factor(s) of interest.

The second criticism of the simulation study of Ho et al. (2007b) is that the substitution rate used in the simulations is too high, with Debruyne and Poinar stating that the rate is "25-fold the estimate of the substitution rate for the mt genome of vertebrates" (p. 350). However, this simulation rate was inspired by published estimates from the mitochondrial D-loop (Lambert et al. 2002; Shapiro et al. 2004), whereas Debruyne and Poinar compare this rate to that estimated from their elephantid data, which is based on whole mitochondrial genomes analyzed over a phylogenetic timeframe. Indeed, the vast majority of published aDNA data sets comprise sequences from the D-loop, which exhibits much higher mutation and substitution rates than does the rest of the mitochondrial genome in vertebrates. This also calls into question the design of the main analysis presented in their critique, in which subsamples from the complete mitochondrial genomes of woolly mammoths were taken to be representative of real aDNA data sets.

Nevertheless, the high rate used in our simulation could be viewed as a legitimate problem if short-term rates were not actually elevated. This led Debruyne and Poinar to pose the question: "what would the accuracy and precision of the posterior rate of change be if a slower rate of substitution, in the range of the interspecific mitochondrial substitution rates (between 1 and 2 × 10^{-8} substitutions/site/year) were applied to simulate the same sequence data?" (p. 350). In response to this question, and to address some of their other concerns, we present the results of a detailed simulation study below.

Simulation Study

We conducted analyses of simulated aDNA data to investigate the performance of Bayesian rate estimation. The amount of rate estimation bias is quantified under various combinations of simulation rate and sequence length, including conditions that might match those commonly encountered in real aDNA research. We investigate the impact of varying the population-size prior, and we compare the performance of different posterior measures of the rate.

Materials and methods.—Sequence evolution was simulated using Seq-Gen (Rambaut and Grassly 1997) on random trees generated according to a coalescent model with a constant population size of 10^5. Each simulated data set comprised 31 time-stamped, nonrecombining sequences, with ages of 0, 1000, 2000, ..., 30,000 years. All sequences were generated according to the Jukes–Cantor model of nucleotide substitution (Jukes and Cantor 1969), with rate homogeneity among sites and among branches. Simulations were performed with 3 different substitution rates (1×10^{-8} substitutions/site/year, 5×10^{-8} substitutions/site/year, and 1×10^{-7} substitutions/site/year) and 2 sequence lengths (100 and 1000 bp), representing the range of characteristics of typical aDNA data sets and encompassing conditions expected to generate sequence alignments with low information content. One thousand replicate data sets were generated for each combination of sequence length and rate. Apart from the substitution rate and sequence length, the simulations are identical to those described in the "uniform sampling regime" in our previous study (Ho et al. 2007b).

Substitution rates were estimated from the simulated data sets using the Bayesian phylogenetic software BEAST 1.4.8 (Drummond and Rambaut 2007). To match the simulation conditions, the Jukes–Cantor substitution model was assumed and a constant-size coalescent prior was chosen for the tree. A uniform prior of $[0, \infty)$ was chosen for the substitution rate. Posterior distributions of parameters were obtained by Markov chain Monte Carlo (MCMC) sampling, with samples drawn every 500 steps over a total of 2×10^7 steps, with the first 10% of samples discarded as burn-in. To compare different posterior measures of the substitution rate, the mean, median, and mode of the posterior rate distribution were calculated for each analysis. Effective sample sizes of parameters were examined to check for acceptable MCMC mixing and sufficient sampling from the posterior.

For any given data set, the estimates of rate and population size are closely tied. The population size prior can be influential in the estimation of rates, particularly when the data set is relatively uninformative. We investigated this issue by performing three sets of analyses, differing only in the population size prior: 1) population size fixed to its true (simulation) value of 10^5; 2) population size given a uniform prior of $[0, \infty)$; and 3) population size given a uniform prior of $[10^0, 10^9]$, representing a range of values that could be considered biologically plausible for vertebrates. Note that in all these analyses, "population size" is actually given as $N_e\tau$, the product of the effective population size (N_e) and generation time in years (τ).

Results.—The performance of rate estimation varied considerably among the three sets of simulations, providing a strong indication of the influence of the population size prior (Table 1). When the population size is fixed to its true (simulation) value of 10^5, estimates of rates are accurate and precise. The 95% highest posterior density (HPD) interval of the substitution rate included the simulation value at least 95% of the time. As noted by Debruyne and Poinar, the mean posterior rate estimates reveal that there is considerable overestimation of the rate when there is low information content or little sequence variability in the data set (low substitution rate and/or short sequence length). However, this bias disappears in the more informative data sets. As the posterior rate distributions are leptokurtic, the medians are less biased than the means. The posterior mode, which represents the maximum *a posteriori* estimate of the rate, appears to provide an unbiased measure across all combinations of substitution rate and sequence length.

A different pattern emerges when the population size is given an unbounded uniform prior distribution (Table 1). Many of the MCMC analyses failed to converge, yielding posterior samples with effective sample sizes not exceeding 100 and with the population size tending toward infinity and the rate tending toward zero. The percentage of analyses that failed to converge ranged from 10.2% to 99.2% across the 6 simulation settings (Fig. 1). If these problematic replicates are removed, the remaining replicates appear to yield reasonable estimates of the substitution rate (Table 1). The 95% HPD interval of the rate, although considerably wider than when the population size was fixed to its correct value, included the simulation value at least 96% of the time. Plausible, unimodal estimates of the population size were obtained in the MCMC analyses that showed signs of convergence. However, in almost all the simulation settings, the rate was overestimated by the mean, median, and mode. This could be a direct consequence of removing the replicates that produced unconverged MCMC analyses because those would have been the data sets with stochastically lower information content (i.e., driven by a smaller number of substitutions and thus producing lower rate estimates). Taking this into consideration, it is difficult to establish whether the estimation bias is genuine or whether it results from taking a biased sample of the simulation replicates.

When the population size is constrained to a range of biologically plausible values (10^0–10^9), yet another

TABLE 1. Summary of results from the simulation study, averaged across 1000 replicates. For the simulations with a population size prior of Uniform[0,∞), results were only summarized from the replicates that exhibited acceptable MCMC convergence. Further details are given in the text

Prior on population size	True rate (substitutions/site/year)	Length (bp)	Posterior rate estimate (substitutions/site/year) Mean	Median	Mode	Mean size of 95% HPD interval (substitutions/site/year)	95% HPD coverage[a]
Fixed to 10^5	1.00×10^{-8}	100	2.32×10^{-8}	1.87×10^{-8}	1.05×10^{-8}	5.63×10^{-8}	0.98
Fixed to 10^5	1.00×10^{-8}	1000	1.20×10^{-8}	1.14×10^{-8}	1.01×10^{-8}	1.67×10^{-8}	0.96
Fixed to 10^5	5.00×10^{-8}	100	6.74×10^{-8}	6.17×10^{-8}	5.10×10^{-8}	1.15×10^{-7}	0.96
Fixed to 10^5	5.00×10^{-8}	1000	5.31×10^{-8}	5.20×10^{-8}	4.97×10^{-8}	4.51×10^{-8}	0.96
Fixed to 10^5	1.00×10^{-7}	100	1.20×10^{-7}	1.13×10^{-7}	1.00×10^{-7}	1.67×10^{-7}	0.97
Fixed to 10^5	1.00×10^{-7}	1000	1.04×10^{-7}	1.03×10^{-7}	9.99×10^{-8}	7.18×10^{-8}	0.95
Uniform[0,∞)	1.00×10^{-8}	100	1.68×10^{-7}	1.21×10^{-7}	3.57×10^{-8}	4.80×10^{-7}	1.00
Uniform[0,∞)	1.00×10^{-8}	1000	2.82×10^{-8}	2.56×10^{-8}	2.08×10^{-8}	5.49×10^{-8}	0.97
Uniform[0,∞)	5.00×10^{-8}	100	1.92×10^{-7}	1.65×10^{-7}	1.07×10^{-7}	4.32×10^{-7}	0.96
Uniform[0,∞)	5.00×10^{-8}	1000	5.89×10^{-8}	5.71×10^{-8}	5.35×10^{-8}	8.10×10^{-8}	0.98
Uniform[0,∞)	1.00×10^{-7}	100	2.66×10^{-7}	2.40×10^{-7}	1.86×10^{-7}	5.28×10^{-7}	0.97
Uniform[0,∞)	1.00×10^{-7}	1000	1.03×10^{-7}	1.01×10^{-7}	9.78×10^{-8}	1.10×10^{-7}	0.97
Uniform[10^0,10^9]	1.00×10^{-8}	100	3.53×10^{-8}	7.34×10^{-9}	8.31×10^{-9}	1.68×10^{-7}	1.00
Uniform[10^0,10^9]	1.00×10^{-8}	1000	9.07×10^{-9}	6.71×10^{-9}	2.48×10^{-9}	2.52×10^{-8}	0.81
Uniform[10^0,10^9]	5.00×10^{-8}	100	5.20×10^{-8}	2.75×10^{-8}	8.48×10^{-9}	1.84×10^{-7}	0.83
Uniform[10^0,10^9]	5.00×10^{-8}	1000	4.66×10^{-8}	4.48×10^{-8}	3.53×10^{-8}	7.26×10^{-8}	0.87
Uniform[10^0,10^9]	1.00×10^{-7}	100	8.44×10^{-8}	5.80×10^{-8}	2.18×10^{-8}	2.49×10^{-7}	0.78
Uniform[10^0,10^9]	1.00×10^{-7}	1000	9.79×10^{-8}	9.63×10^{-8}	9.11×10^{-8}	1.09×10^{-7}	0.91

[a] Proportion of simulations in which the 95% HPD interval of the rate contained the true (simulation) value.

picture materialises. Coverage by the 95% HPD intervals was poorer, with the simulation value being excluded from the 95% HPD interval up to 22% of the time (Table 1). The mean size of the 95% HPD interval is smaller than in the analyses without any restrictions on the population size, although the disparity disappears as the number of variable sites in the alignment increases. The posterior mode is no longer the best summary of the rate, probably because the constraints on population size also impose restrictions on the values that can be taken by the substitution rate. In some cases, the posterior distribution of the rate is implicitly constrained, leading to a distorted mode. On the other hand, the posterior mean appears to provide a reasonably accurate estimate of the true substitution rate (Table 1), although it is possible that this is partly an unintended consequence of the population size constraints. That is, the mean posterior rate might only be accurate as a result of the population size priors constraining the substitution rate to reasonable values, even in the absence of real information on rates in the data. This effect could potentially explain some of the published rate estimates from uninformative aDNA sequence alignments, which have taken seemingly plausible values in spite of the low information content of the data.

aDNA Data Sets

Published aDNA data sets vary considerably in terms of their sequence lengths and underlying substitution rates as well as the temporal structure and spread of the samples. It would be useful to evaluate the information content in these data sets to determine whether they can produce reliable estimates of substitution rates and divergence times. One significant facet of heterochronous data that is overlooked by the use of diversity statistics (Depaulis et al. 2009), and in the analyses of information content performed by Debruyne and Poinar, is that the ages of the sequences form an important component of the information content (e.g., Firth et al. 2010). This stems from the fact that the sequence ages are used for calibrating estimates of substitution rates. A potential problem in analyses of heterochronous data is that rate estimates could be an artifact of the sampling ages.

Here, we use a date randomization test to investigate temporal structure in 18 published aDNA data sets. This test involves reanalyzing each data set after randomly shuffling the ages of the sequences and follows several previous studies of heterochronous data (de Bruyn et al. 2009; Miller et al. 2009; Subramanian et al. 2009b; Firth et al. 2010). The date randomization analysis is able to provide some insight into whether the structure and spread of the sequence ages are sufficient to provide reliable information on the rate underlying the evolution of the data set. If the original rate estimate is recovered in the date-randomized data sets, then there is insufficient temporal structure in the original data set and the rate estimate cannot be supported (Firth et al. 2010).

Materials and methods.—Using the Bayesian phylogenetic method implemented in BEAST v1.5.4 (Drummond and Rambaut 2007), we analyzed 18 published aDNA alignments: 16 of the 19 aDNA data sets analyzed by Ho et al. (2007b), the 11 mitogenome alignment of woolly mammoths examined by Debruyne and Poinar, and a muskox D-loop alignment (Campos et al. 2010). We excluded three data sets from the study by Ho et al. (2007b): the *Chlorobium* and nene alignments contained too few ancient sequences for the randomization test,

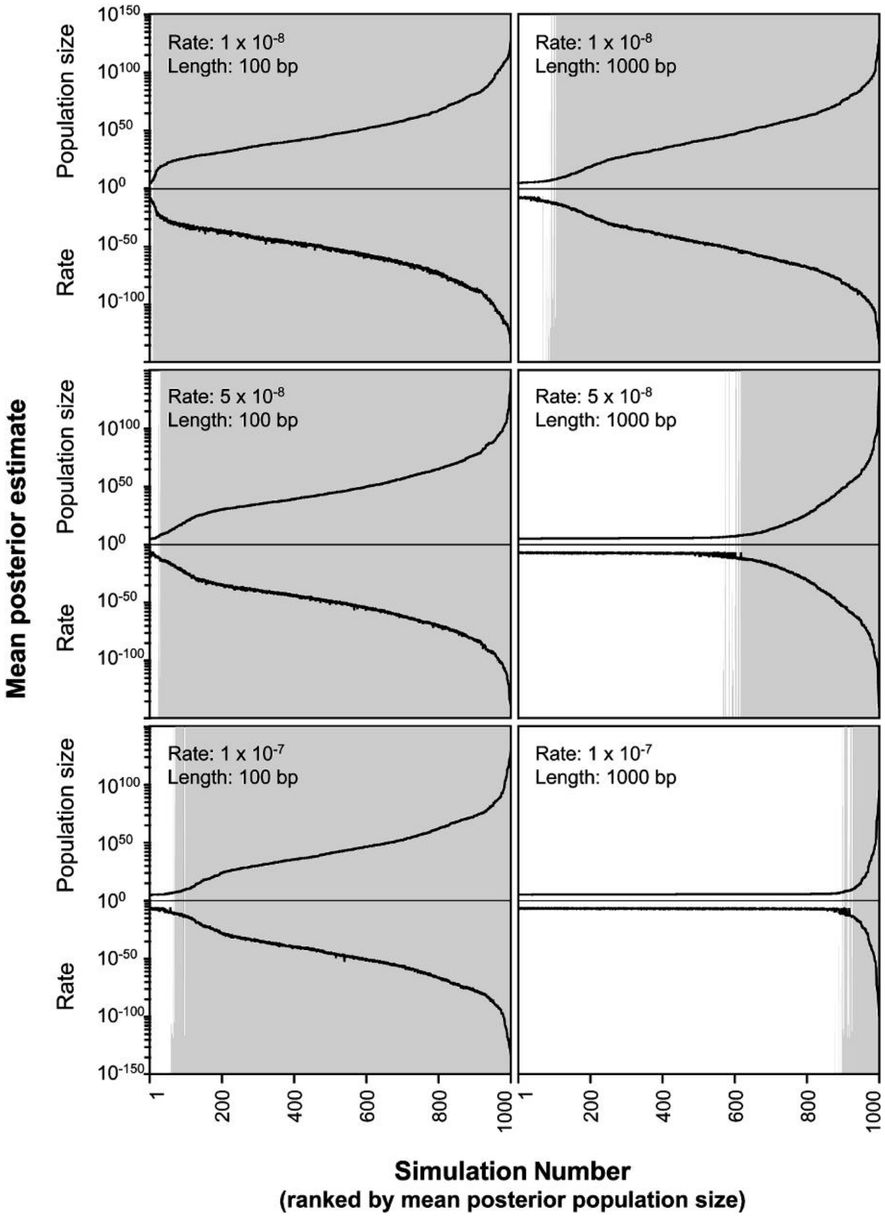

FIGURE 1. Graphs showing the correspondences between mean posterior population size, mean posterior rate, and MCMC convergence for Bayesian analyses of data generated under 6 different simulation conditions (3 different rates and 2 different sequence lengths). The results were obtained using an uninformative population size prior (uniform from 0 to ∞). Each panel shows the results from analyzing 1000 replicates, ranked from left to right by ascending mean posterior population size (top curve). The mean posterior rate estimate for the corresponding data set is also displayed on the same scale (lower curve), showing a close relationship with the estimated population size. Each simulation is given a gray vertical line in the background if the effective sample size for the posterior likelihood is below 100, which suggests a lack of convergence to the stationary distribution. For each MCMC analysis, samples were drawn from the posterior every 500 steps over a total of 2×10^7 steps, with the first 10% of samples discarded as burn-in.

whereas the muskox alignment is superseded by the larger data set published by Campos et al. (2010). The basic characteristics of the 18 data sets are outlined in Table 2, with further details available in the original publications.

Substitution models were selected by comparison of Bayesian information criterion scores, with the number of aligned sites taken as the sample size for the penalty term. Owing to the intraspecific nature of the data sets, models that allowed a proportion of invariable sites were excluded. All data sets were treated as unpartitioned, and a constant-size coalescent prior was specified for the topology and divergence times. All analyses were repeated using a Bayesian skyride demographic model (Minin et al. 2008). The better demographic model (constant size or Bayesian skyride) was chosen on the basis of visual inspection of the results. In each analysis, samples from the posterior were drawn every 5×10^3 steps from a total of 5×10^7 steps, with the first 10% being discarded as burn-in. Where necessary,

TABLE 2. Details of aDNA alignments analyzed using the date randomization test described in the text

Species		Region	Sequences (ancient + modern)	Age range[a] (years)	Length (bp)	Variable sites	Result of date randomization test
Adélie penguin	*Pygoscelis adeliae*	D-loop	96 + 380	6424	347	159	√
Arctic fox	*Alopex lagopus*	D-loop	8 + 41	16,000	291	23	√
Aurochs	*Bos primigenius*	D-loop	41 + 0	10,300	360	34	√
Bison	*Bison priscus*	D-loop	150 + 32	60,400	615	170	√
Boar	*Sus scrofa*	D-loop	81 + 7	5400	572	47	√
Bowhead whale	*Balaena mysticetus*	D-loop	99 + 68	51,000	453	72	Fail
Brown bear	*Ursus arctos*	D-loop	36 + 57	59,000	193	69	√
Cave bear	*Ursus spelaeus*	D-loop	26 + 0	53,470	288	31	Fail
Cave hyaena	*Crocuta crocuta spelaea*	D-loop	10 + 0	13,140	366	27	Fail
Cave lion	*Panthera leo spelaea*	D-loop	23 + 0	46,275	213	12	√
Cow	*Bos taurus*	D-loop	36 + 91	8065	410	65	√
Horse	*Equus caballus*	D-loop	12 + 33	28,340	348	70	√
Maize	*Zea mays*	*adh2*	9 + 11	4500	190	26	Fail
Moa	*Pachyornis mappini*	D-loop	14 + 0	4912	241	20	Fail
Muskox	*Ovibos moschatus*	D-loop	114 + 16	45,740	682	203	√
Tuco-tuco	*Ctenomys sociabilis*	*cytb*	45 + 1	10,208	253	13	√
Woolly mammoth	*Mammuthus primigenius*	D-loop	32 + 0	35,970	741	42	Fail
Woolly mammoth	*Mammuthus primigenius*	Mitogenome	11 + 0	38,030	16,484	112	Fail

[a] Age of oldest sequence minus age of youngest sequence.

the number of MCMC steps was doubled or tripled in order to achieve an effective sample size >100 for the rate estimate.

The sequence ages in each of the 18 aDNA data sets were then randomly reassigned. This randomization was performed 20 times for each data set using the Java application SiteSampler v1.1 (Ho and Lanfear 2010). Bayesian phylogenetic analyses were performed using the same settings as described above for the original data. For each date-randomized data set, the demographic model was chosen to match that selected for the original data.

Results.—The posterior rate estimates from the 18 data sets are shown in Figure 2. It is interesting to note that among the 7 data sets that failed the date randomization test not all produced rate estimates with wide 95% HPD intervals. In these cases, the modal posterior rate was similar to the mean posterior rate (results not shown).

To investigate the potential presence of signal-dependent biases in these estimates, we considered the mean posterior rates in relation to the characteristics of the data sets from which they were estimated. Debruyne and Poinar hypothesize that the mean posterior rate estimate should be exponentially related to the amount of information in the data set, as reflected by the alignment length. We examined 4 measures of information content: the number of aligned sites, the number of variable sites, the number of sequences, and the product of the number of sites and sequences in the alignment. Excluding the mitogenome alignment of woolly mammoths, which represents an outlier and is nonindependent of the D-loop alignment from the same species, we find no evidence that any of these measures are related to the mean posterior rate estimate in the remaining 17 aDNA data sets ($r^2 < 0.1$ and $P > 0.2$ in all cases). However, more than 40% of the variation in rate estimates could be explained by an exponential relationship with the age range of the sequences in each data set ($r^2 = 0.431$ and $P = 0.004$).

Further insight into the temporal structure within the data sets was gained through the date randomization analyses. Eleven alignments passed the randomization test and seven failed (Fig. 2; Table 2). In addition to the results presented in this study, previous date randomization analyses of aDNA from tuatara (Subramanian et al. 2009b) and elephant seals (de Bruyn et al. 2009) have indicated that these two data sets contain sufficient temporal information to produce meaningful estimates of substitution rates. Among the data sets that failed the date randomization test, the bowhead whale alignment is noted for its low sequence diversity, with the observed variation dominated by singleton mutations (Borge et al. 2007). The maize alignment is a small data set comprising sequences sampled over a short time frame (Freitas et al. 2003). Notably, both of the mammoth alignments (D-loop and complete mitochondrial genome) failed the date randomization test.

DISCUSSION

Our analyses of simulated and real data show that the signal-dependent artifact highlighted by Debruyne and Poinar is unlikely to have contributed substantially to the published rate estimates from aDNA data sets. Our simulated data sets cover a range of sequence lengths and substitution rates, including those seen in real aDNA alignments. Regardless of the prior on population size, the posterior mean provides an unbiased estimate of the rate for the 1000 bp data sets simulated using a rate of 1×10^{-7} substitutions/site/year. These parameters are broadly similar to those of the mitochondrial D-loop in vertebrates. For the less informative alignments investigated here, including those simulated using lower rates, there is some degree of estimation bias unless the population size is fixed to its

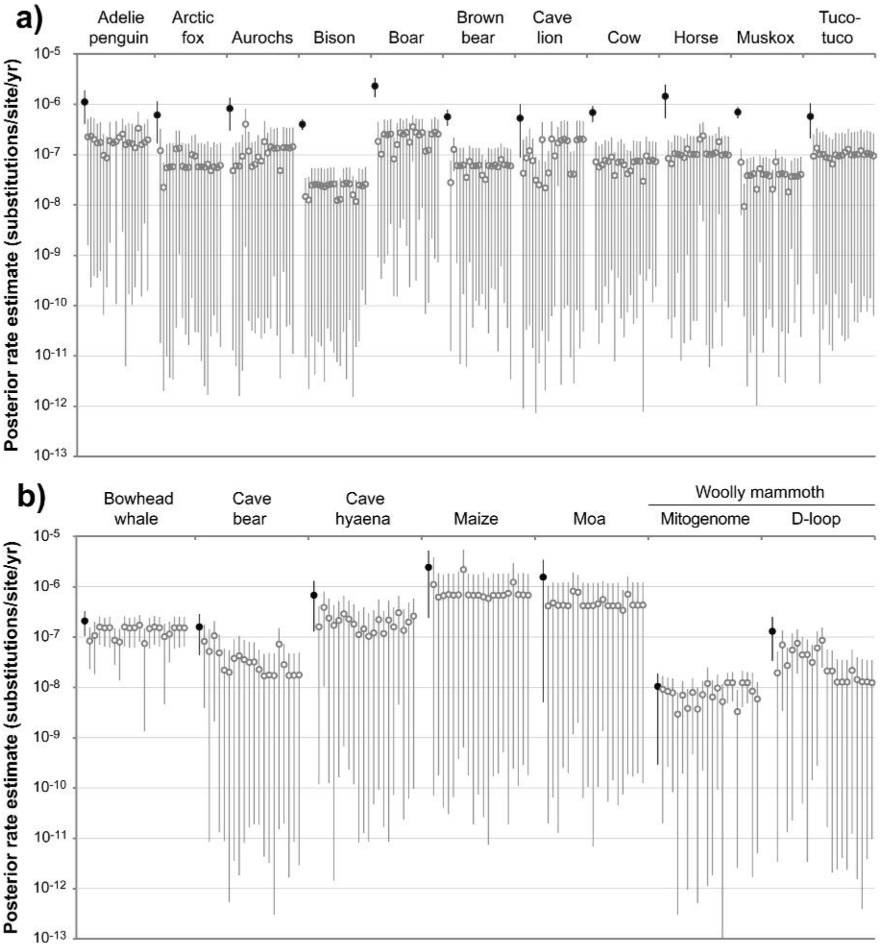

FIGURE 2. Estimates of substitution rates from a variety of aDNA alignments. For each data set, the first data point represents the rate estimated from the original data set (filled circles), whereas the remaining 20 data points (unfilled circles) represent the rates estimated from replicates in which the ages of the tips were randomly shuffled. Alignments were deemed to "pass" the date- randomization test if the mean posterior rate estimate from the original data set is not included in any of the 95% HPD intervals from the date-randomized replicates. a) Rate estimates from alignments that passed the date randomization test. b) Rate estimates from alignments that failed the date randomization test.

simulation value. However, the rates used for the simulations in this study are conservatively low because they are based on phylogenetic estimates. If short-term rates are actually elevated, as posited by the hypothesis of time-dependent rates, then the particular estimation biases observed in this study might be irrelevant to the majority of real aDNA alignments.

The results of our simulation analyses confirm that the posterior mean can be a biased measure of the substitution rate, as indicated by Debruyne and Poinar. However, the posterior estimates provide acceptable coverage because the 95% HPD intervals on the rates usually included the simulation value. The posterior mode, which is equivalent to the maximum *a posteriori* estimate of the rate, appears to be the best measure when the data set has low information content. Nevertheless, it can become distorted when a strongly bounded, informative prior is specified for the population size (or, presumably, for the substitution rate or age of the root). In view of these results, it might be most appropriate to report various summaries of the posterior distribution of rates and other parameters of interest. The three measures examined here, the mean, median, and mode, converge to the same value for the most informative data sets.

Contrary to the claim by Debruyne and Poinar, our analyses suggest that sequence variability is not the sole factor determining the performance of rate estimation. Other factors, such as the ages of the sequences, and probably the structure of the underlying genealogy, are also very important features of any aDNA data set. Evidently, the population size prior is highly influential in some of the analyses performed here. From a practical viewpoint, it is usually more feasible to use an informative prior for the age of the root rather than the population size. This is because effective population size and generation time are often difficult to estimate reliably, whereas the age of the root can sometimes be inferred from independent palaeontological or biogeographic sources.

Taking into account the results of all the date randomization analyses, the sampling ages of seven real aDNA data sets (bowhead whale, cave bear, cave hyaena, maize, moa, woolly mammoth D-loop, and woolly mammoth mitogenomes) were found to produce artifactual rate estimates using the date randomization test. This result suggests that some of the published aDNA alignments do not contain sufficient temporal information to support reliable estimation of rates and timescales. Randomization of sequence ages represents a potentially useful technique for investigating the validity of rate estimates from heterochronous data, including those from aDNA and serially sampled viruses (de Bruyn et al. 2009; Miller et al. 2009; Subramanian et al. 2009b; Firth et al. 2010).

It is interesting to note that the alignment of complete mitochondrial genomes from woolly mammoths failed the date randomization test. This suggests that the analyses performed by Debruyne and Poinar might be misleading, being based on a data set that is unable to yield plausible posterior estimates without strong prior information on the population size or root age. There are probably several reasons for the poor performance of the mammoth mitogenomic data set. First, the alignment comprises only a small number of sequences. Second, the mitochondrial tree of mammoths has a highly unusual structure, with a very deep split separating the two major clades (Gilbert et al. 2008). This is reflected in the imprecision of the date estimates that are obtained when only the ages of the tips are used for calibration (Debruyne et al. 2008; Gilbert et al. 2008; Debruyne and Poinar 2009). Third, mammoth mitochondrial DNA has evolved at an exceptionally low rate, a phenomenon that is mirrored in the mammoth nuclear genome (Hofreiter 2008; Miller et al. 2008). For example, the substitution rate in elephantids has been much lower than that in hominoid primates, which in turn have been evolving more slowly than other primates (Steiper et al. 2004). This also calls into question the analyses performed by Debruyne and Poinar in which subsamples of the mammoth mitogenomes were assumed to be representative of typical aDNA alignments. In practice, short aDNA alignments have almost exclusively come from the D-loop, which is the most variable portion of the vertebrate mitochondrial genome. The subsampling procedure used by Debruyne and Poinar will tend to include portions of the mitogenome that are evolving much more slowly, leading to exceptionally uninformative data sets.

Debruyne and Poinar recommend that the bias due to signal dependence can be overcome through the employment of “deep” calibrations, for example, at the root of the tree. Often, this is neither possible nor appropriate in analyses of heterochronous data. If rates were truly time-dependent, then such analyses would need to be performed in a relaxed-clock framework to allow the rate to vary between younger and older branches (Korsten et al. 2009). If a strict molecular clock is assumed, as in the analyses done by Debruyne and Poinar, then rate homogeneity across different timescales is invoked as an a priori assumption. Therefore, although this would seemingly address the problem posed by time-dependent rates, it only does so by assuming that the problem does not exist (Ho et al. 2007c). A suggested solution to this problem is to limit the analysis to third codon sites or synonymous sites, which are putatively subject to a much lesser degree of selective constraint (Briggs et al. 2009; Subramanian et al. 2009a; Endicott et al. 2010). In any case, reliable internal-node calibrations are rarely available in population-level analyses (Ho and Phillips 2009).

Although we have demonstrated that time-dependent rates are unlikely to be driven by a signal-dependent artifact, the findings obtained in the present study do not necessarily validate published estimates of rates from aDNA data. Such estimates can be detrimentally affected by a variety of other confounding factors, including misspecification of the demographic model (Emerson 2007; Ho et al. 2007c; Miller et al. 2009; Navascués and Emerson 2009; Subramanian et al. 2009b). Furthermore, postmortem damage can produce spurious polymorphisms in aDNA sequences, which can lead to biased estimates of rates (Ho et al. 2005, 2007a). Samples in several of the data sets have not been directly radiocarbon dated, but their ages have been inferred by stratigraphic correlation (layer dating). Rate estimates from these data sets, including the arctic fox, Adelie penguin, aurochs, boar, maize, and tuco-tuco, will be somewhat less reliable than those from data sets with directly dated samples. However, higher rate estimates have been obtained from a wide range of aDNA data sets, sourced from a variety of taxa with different demographic histories and biological characteristics, indicating that they should not be dismissed lightly. Combined with the exceptionally high rates estimated in studies of pedigrees and mutation accumulation lines, these results suggest that further empirical and theoretical investigations into the nature of time-dependent rates could be productive.

By their very nature, most aDNA data sets have low information content. Although the situation is changing, as high-throughput sequencing techniques allow complete mitochondrial genomes to be sequenced from conspecific individuals (Gilbert et al. 2008; Briggs et al. 2009; Stiller et al. 2009; Ho and Gilbert 2010), short alignments are likely to remain a common feature of aDNA studies in the near future. In these studies, the important question is not whether the information content is low, but whether it is sufficient for performing the analyses of interest.

Funding

This work was supported by the Australian Research Council and the Defense Advanced Research Projects Agency.

Acknowledgments

The authors thank K.Z., C.B., and R.D. for helpful comments and suggestions that considerably improved the paper.

References

Bandelt H.J. 2008. Clock debate: when times are a-changin': time dependency of molecular rate estimates: tempest in a teacup. Heredity. 100:1–2.

Barnes I., Shapiro B., Lister A., Kuznetsova T., Sher A., Guthrie D., Thomas M.G. 2007. Genetic structure and extinction of the woolly mammoth, *Mammuthus primigenius*. Curr. Biol. 17:1072–1075.

Borge T., Bachmann L., Björnstad G., Wiig Ø. 2007. Genetic variation in Holocene bowhead whales from Svalbard. Mol. Ecol. 16:2223–2235.

Briggs A.W., Good J.M., Green R.E., Krause J., Maricic T., Stenzel U., Lalueza-Fox C., Rudan P., Brajkovic D., Kucan Z., Gusic I., Schmitz R., Doronichev V.B., Golovanova L.V., de la Rasilla M., Fortea J., Rosas A., Pääbo S. 2009. Targeted retrieval and analaysis of five Neanderthal mtDNA genomes. Science. 325: 318–321.

Burridge C.P., Craw D., Fletcher D., Waters J.M. 2008. Geological dates and molecular rates: fish DNA sheds light on time dependency. Mol. Biol. Evol. 25:624–633.

Campos P.F., Willerslev E., Sher A., Orlando L., Axelsson E., Tikhonov A., Aaris-Sorensen K., Greenwood A.D., Kahlke R.D., Kosintsev P., Krakhmalnaya T., Kuznetsova T., Lemey P., MacPhee R., Norris C.A., Shepherd K., Suchard M.A., Zazula G.D., Shapiro B., Gilbert M.T. 2010. Ancient DNA analyses exclude humans as the driving force behind late Pleistocene musk ox (*Ovibos moschatus*) population dynamics. Proc. Natl. Acad. Sci. U.S.A. 107:5675–5680.

Charlesworth B., Bartolomé C., Noël V. 2005. The detection of shared and ancestral polymorphisms. Genet. Res. 86:149–157.

de Bruyn M., Hall B.L., Chauke L.F., Baroni C., Koch P.L., Hoelzel A.R. 2009. Rapid response of a marine mammal species to holocene climate and habitat change. PLoS Genet. 5:e1000554.

Debruyne R., Chu G., King C.E., Bos K., Kuch M., Schwarz C., Szpak P., Grocke D.R., Matheus P., Zazula G., Guthrie D., Froese D., Buigues B., de Marliave C., Flemming C., Poinar D., Fisher D., Southon J., Tikhonov A.N., MacPhee R.D., Poinar H.N. 2008. Out of America: ancient DNA evidence for a new world origin of late quaternary woolly mammoths. Curr. Biol. 18:1320–1326.

Debruyne R., Poinar H.N. 2009. Time dependency of molecular rates in ancient DNA data sets, a sampling artifact? Syst. Biol. 58:348–360.

Denver D.R., Morris K., Lynch M., Vassilieva L.L., Thomas W.K. 2000. High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. Science. 289:2342–2344.

Depaulis F., Orlando L., Hänni C. 2009. Using classical population genetics tools with heterochroneous data: time matters! PLoS ONE. 4:e5541.

Drummond A.J., Pybus O.G., Rambaut A., Forsberg R., Rodrigo A.G. 2003. Measurably evolving populations. Trends Ecol. Evol. 18:481–488.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Emerson B.C. 2007. Alarm bells for the molecular clock? No support for Ho et al.'s model of time-dependent molecular rate estimates. Syst. Biol. 56:337–345.

Endicott P., Ho S.Y.W., Stringer C. Forthcoming 2010. Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. J. Hum. Evol. 59:87–95.

Firth C., Kitchen A., Shapiro B., Suchard M.A., Holmes E.C., Rambaut A. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol. Biol. Evol. 27:2038–2051.

Freitas F.O., Bendel G., Allaby R.G., Brown T.A. 2003. DNA from primitive maize landraces and archaeological remains: implications for the domestication of maize and its expansion into South America. J. Archaeol. Sci. 30:901–908.

Genner M.J., Seehausen O., Lunt D.H., Joyce D.A., Shaw P.W., Carvalho G.R., Turner G.F. 2007. Age of cichlids: new dates for ancient lake fish radiations. Mol. Biol. Evol. 24:1269–1282.

Gibbons A. 1998. Calibrating the mitochondrial clock. Science. 279: 28–29.

Gilbert M.T.P., Drautz D.I., Lesk A.M., Ho S.Y.W., Qi J., Ratan A., Hsu C.-H., Sher A., Dalén L., Götherström A., Tomsho L.P., Rendulic S., Packard M., Campos P.F., Kuznetsova T., Shidlovskiy F., Tikhonov A., Willerslev E., Iacumin P., Buigues B., Ericson P.G., Germonpré M., Kosintsev P., Nikolaev V., Nowak-Kemp M., Knight J.R., Irzyk G.P., Perbost C.S., Fredrikson K.M., Harkins T.T., Sheridan S., Miller W., Schuster S.C. 2008. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. Proc. Natl. Acad. Sci. U.S.A. 105:8327–8332.

Haag-Liautard C., Coffey N., Houle D., Lynch M., Charlesworth B., Keightley P.D. 2008. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. PLoS Biol. 6:e204.

Hay J.M., Subramanian S., Millar C.D., Mohandesan E., Lambert D.M. 2008. Rapid molecular evolution in a living fossil. Trends Genet. 24:106–109.

Henn B.M., Gignoux C.R., Feldman M.W., Mountain J.L. 2009. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. Mol. Biol. Evol. 26:217–230.

Ho S.Y.W., Gilbert M.T.P. 2010. Ancient mitogenomics. Mitochondrion. 10:1–11.

Ho S.Y.W., Heupink T.H., Rambaut A., Shapiro B. 2007a. Bayesian estimation of sequence damage in ancient DNA. Mol. Biol. Evol. 24:1416–1422.

Ho S.Y.W., Kolokotronis S.-O., Allaby R.G. 2007b. Elevated substitution rates estimated from ancient DNA. Biol. Lett. 3:702–705.

Ho S.Y.W., Lanfear R. 2010. Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. Mitochondrial DNA. 21:138–146.

Ho S.Y.W., Larson G. 2006. Molecular clocks: when times are a-changin'. Trends Genet. 22:79–83.

Ho S.Y.W., Phillips M.J. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. 58:367–380.

Ho S.Y.W., Phillips M.J., Cooper A., Drummond A.J. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. 22:1561–1568.

Ho S.Y.W., Shapiro B., Phillips M., Cooper A., Drummond A.J. 2007c. Evidence for time dependency of molecular rate estimates. Syst. Biol. 56:515–522.

Hofreiter M. 2008. DNA sequencing: Mammoth genomics. Nature. 456:330–331.

Howell N., Smejkal C.B., Mackey D.A., Chinnery P.F., Turnbull D.M., Herrnstadt C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am. J. Hum. Genet. 72:659–670.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editors. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

Korsten M., Ho S.Y.W., Davison J., Pahn B., Vulla E., Roht M., Tumanov I.L., Kojola I., Andersone-Lilley Z., Ozolins J., Pilot M., Mertzanis Y., Giannakopoulos A., Vorobiev A.A., Markov N.I., Saveljev A.P., Lyapunova E.A., Abramov A.V., Mannil P., Valdmann H., Pazetnov S.V., Pazetnov V.S., Rokov A.M., Saarma U. 2009. Sudden expansion of a single brown bear maternal lineage across northern continental Eurasia after the last ice age: a general demographic model for mammals? Mol. Ecol. 18:1963–1979.

Lambert D.M., Ritchie P.A., Millar C.D., Holland B., Drummond A.J., Baroni C. 2002. Rates of evolution in ancient DNA from Adélie penguins. Science. 295:2270–2273.

Loogväli E.-L., Kivisild T., Margus T., Villems R. 2009. Explaining the imperfection of the molecular clock of hominid mitochondria. PLoS ONE. 4:e8260.

Macaulay V.A., Richards M.B., Forster P., Bendall K.E., Watson E., Sykes B., Bandelt H.-J. 1997. mtDNA mutation rates—no need to panic. Am. J. Hum. Genet. 61:983–985.

Millar C.D., Dodd A., Anderson J., Gibb G.C., Ritchie P.A., Baroni C., Woodhams M.D., Hendy M.D., Lambert D.M. 2008. Mutation and evolutionary rates in adelie penguins from the Antarctic. PLoS Genet. 4:e1000209.

Miller H.C., Moore J.A., Allendorf F.W., Daugherty C.H. 2009. The evolutionary rate of tuatara revisited. Trends Genet. 25:13–15.

Miller W., Drautz D.I., Ratan A., Pusey B., Qi J., Lesk A.M., Tomsho L.P., Packard M.D., Zhao F., Sher A., Tikhonov A., Raney B., Patterson N., Lindblad-Toh K., Lander E.S., Knight J.R., Irzyk G.P., Fredrikson K.M., Harkins T.T., Sheridan S., Pringle T., Schuster S.C. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 456:387–390.

Minin V.N., Bloomquist E.W., Suchard M.A. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459–1471.

Navascués M., Emerson B.C. 2009. Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. Mol. Ecol. 18:4390–4397.

Papadopoulou A., Anastasiou I., Vogler A.P. 2010. Revisiting the insect mitochondrial molecular clock: the mid-Aegean trench calibration. Mol. Biol. Evol. 27:1659–1672.

Penny D. 2005. Relativity for molecular clocks. Nature. 426:183–184.

Peterson G.I., Masel J. 2009. Quantitative prediction of molecular clock and K_a/K_s at short timescales. Mol. Biol. Evol. 26:2595–2603.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Shapiro B., Drummond A.J., Rambaut A., Wilson M.C., Matheus P.E., Sher A.V., Pybus O.G., Gilbert M.T., Barnes I., Binladen J., Willerslev E., Hansen A.J., Baryshnikov G.F., Burns J.A., Davydov S., Driver J.C., Froese D.G., Harington C.R., Keddie G., Kosintsev P., Kunz M.L., Martin L.D., Stephenson R.O., Storer J., Tedford R., Zimov S., Cooper A. 2004. Rise and fall of the Beringian steppe bison. Science. 306:1561–1565.

Soares P., Ermini L., Thomson N., Mormina M., Rito T., Röhl A., Salas A., Oppenheimer S., Macaulay V., Richards M.B. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. 84:1–20.

Steiper M.E., Young N.M., Sukarna T.Y. 2004. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence. Proc. Natl. Acad. Sci. U.S.A. 101:17021–17026.

Stiller M., Knapp M., Stenzel U., Hofreiter M., Meyer M. 2009. Direct multiplex sequencing (DMPS)—a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. Genome Res. 19:1843–1848.

Subramanian S., Denver D.R., Millar C.D., Heupink T., Aschrafi A., Emslie S.D., Baroni C., Lambert D.M. 2009a. High mitogenomic evolutionary rates and time dependency. Trends Genet. 25: 482–486.

Subramanian S., Hay J.M., Mohandesan E., Millar C.D., Lambert D.M. 2009b. Molecular and morphological evolution in tuatara are decoupled. Trends Genet. 25:16–18.

Woodhams M. 2006. Can deleterious mutations explain the time dependency of molecular rate estimates? Mol. Biol. Evol. 23:2271–2273.

Syst. Biol. 60(3):375–386, 2011

DOI:10.1093/sysbio/syq100
Advance Access publication on February 8, 2011

Are Transposable Element Insertions Homoplasy Free?: An Examination Using the Avian Tree of Life

Kin-Lan Han[1,2,3,*], Edward L. Braun[1], Rebecca T. Kimball[1], Sushma Reddy[4,5], Rauri C. K. Bowie[6,7], Michael J. Braun[2,3], Jena L. Chojnowski[1], Shannon J. Hackett[5], John Harshman[5,8], Christopher J. Huddleston[2], Ben D. Marks[9], Kathleen J. Miglia[10], William S. Moore[10], Frederick H. Sheldon[9], David W. Steadman[11], Christopher C. Witt[12,13], and Tamaki Yuri[1,2,14]

[1]*Department of Biology, University of Florida, Gainesville, FL 32611, USA;*
[2]*Department of Vertebrate Zoology, Smithsonian Institution, Suitland, MD 20746, USA;*
[3]*Behavior, Ecology, Evolution, and Systematics Program, University of Maryland, College Park, MD 20742, USA;*
[4]*Department of Biology, Loyola University Chicago, Chicago, IL 60626, USA;*
[5]*Zoology Department, Field Museum of Natural History, Chicago, IL 60605, USA;*
[6]*Museum of Vertebrate Zoology and* [7]*Department of Integrative Biology,*
University of California Berkeley, Berkeley, CA 94720, USA;
[8]*4869 Pepperwood Way, San Jose, CA 95124, USA;*
[9]*Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA;*
[10]*Department of Biological Sciences, Wayne State University, Detroit, MI 48202, USA;*
[11]*Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA;*
[12]*Department of Biology and* [13]*Museum of Southwestern Biology,*
University of New Mexico, Albuquerque, NM 87131, USA; and
[14]*Sam Noble Oklahoma Museum of Natural History, University of Oklahoma, Norman, OK 73072, USA;*
[*]*Correspondence to be sent to: Department of Biology, University of Florida, PO Box 118525, Gainesville, FL 32611, USA;*
E-mail: hankin@ufl.edu.

Received 14 December 2009; reviews returned 1 April 2010; accepted 3 November 2010
Associate Editor: Frank Anderson

[Rare genomic changes] provide an independent source of phylogenetic information largely immune from some of the problems that affect primary sequence data.—Rokas and Holland (2000)

In an attempt to find the true evolutionary tree of life, phylogeneticists have searched for "perfect" characters—those free of homoplasy. Rare genomic changes (RGCs) are infrequent mutations such as transposable element (TE) insertions, intron gains or losses, gene order changes, inversions, gene duplications, and even fusion/fissions of protein domains (Rokas and Holland 2000). RGCs are candidates for perfect characters, as they are believed to exhibit little or no homoplasy for two reasons. First, they typically accumulate slowly, with some types of RGCs accumulating so slowly that they are useful for defining the deepest branches in the tree of life (Keeling and Doolittle 1997; Stechmann and Cavalier-Smith 2002). Other RGC types, however, such as TE insertions, accumulate rapidly enough to be useful for defining more closely related groups (e.g., Watanabe et al. 2006; Kaiser et al. 2007). Second, regardless of their rate of accumulation, RGCs are thought to have a large state space (Steel and Penny 2000), which means that independent RGCs can be distinguished and are unlikely to be interpreted as homologous (Rokas and Holland 2000; Shedlock and Okada 2000; Ray et al. 2006). For example, TEs can insert into almost any position in the genome in two different orientations. Additionally, the existence of multiple TE types and subtypes (Jurka 1998; Wicker et al. 2005, 2007) makes it possible to identify independent insertion of different types based upon their sequence. Finally, most insertions include only part of the complete TE sequence, so independent insertions may be different segments of the original even if they are of the same subtype and in the same orientation.

Despite the reasons to expect RGCs to be perfect homoplasy-free characters, many different RGCs can exhibit homoplasy (Ray et al. 2006; Gibb et al. 2007). Although even very rare events like protein domain fusion/fissions can be reversed (Braun and Grotewold 2001; Braun 2003), the most commonly invoked explanation for RGCs that appear homoplastic are differences between individual gene trees associated with specific RGCs and the species tree (Fig. 1a) (Hillis 1999; Shedlock and Okada 2000; Shedlock et al. 2004; Ray et al. 2006; Sasaki et al. 2006; Nishihara et al. 2009; but see Murphy et al. 2007, for a possible exception). In fact, the only available statistical method for RGC analyses (Waddell et al. 2001) assumes that conflicts among RGCs reflect lineage sorting, thus it uses a coalescent model (Hudson 1992) to predict the distribution of character states. Consequently, this model assumes RGCs that appear to conflict with the species tree can be explained by hemiplasy, a situation where lineage sorting gives rise to the illusion of homoplasy with respect to the species tree (Avise and Robinson 2008). Hemiplasy is expected to be more likely to occur on short internodes in the species tree, whereas bona fide homoplasy is most likely to occur on long internodes because the probability that a specific gene tree conflicts with a species tree is typically related to the length of the relevant internal branches (e.g., Pamilo and Nei 1988; Degnan and Rosenberg 2009). Because coalescent models only account for conflict due to hemiplasy, the models proposed for analyses of RGC data will have to be expanded if RGCs also exhibit homoplasy.

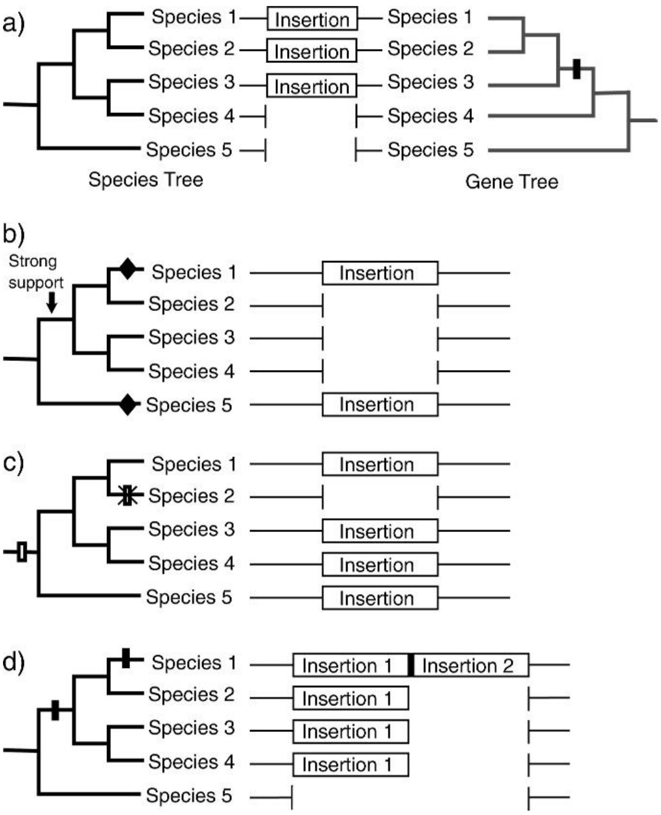

FIGURE 1. Potential complex TE insertion patterns. Solid bars indicate homoplasy-free insertions (those exhibiting a retention index of 1.0). Open bars represent the insertions that were subsequently deleted (with the deletion represented by a X over an open bar). Diamonds represent independent insertions in distinct lineages. a) A TE insertion associated with a gene tree (right) that is inconsistent with the species tree (left) will appear homoplastic due to lineage sorting. This situation was recently designated "hemiplasy" to distinguish it from true homoplasy due to multiple origins of a genomic feature (Avise and Robinson 2008). b) Multiple insertions at the same site in divergent taxa, shown is a case where there is an insertion at identical sites in two different taxa, but the strong phylogenetic support for placing these taxa in different clades suggests that these insertions are independent. c) Insertion and subsequent complete deletion of the TE in some taxa; shown is a case where an insertion appears in all but one taxon within a clade suggesting excision of the entire insertion from this taxon. d) Multiple insertions at the same site in some but not all taxa, shown here by a single insertion in the ancestor to Species 1–4, with a second insertion at the same site in Species 1. The insertions in Species 1 can be of the same type or of different types. Unlike the other scenarios shown here, this pattern of insertions does not have the potential to be misleading, although it does suggest the existence of hot spots for TE insertions and/or fixations. Duplications and other types of sequence changes (e.g., inversions) also have the potential to create complex insertion patterns similar to the examples presented here, so their interpretation can be difficult.

Insertions of TEs, specifically retrotransposons, are the RGCs most commonly used in vertebrate phylogenetics (Shedlock and Okada 2000; Kriegs et al. 2006; Nishihara et al. 2006a; Ray et al. 2006; Kaiser et al. 2007; Kriegs et al. 2007; Treplin and Tiedemann 2007). The presumption that RGCs do not exhibit homoplasy has even prompted conclusions based on single-TE insertions. However, some inferences supported by individual TEs, such as the phylogenetic position of the enigmatic rockfowl, *Picathartes* spp. (Treplin and Tiedemann 2007), and the phylogenetic position of the Japanese quail, *Coturnix japonica* (Kaiser et al. 2007; Kriegs et al. 2007), conflict with large-scale nucleotide and total evidence phylogenies (Barker et al. 2004; Crowe et al. 2006; Cox et al. 2007; Hackett et al. 2008; Kimball and Braun 2008). Even phylogenetic hypotheses based upon more than one TE insertion (e.g., Kriegs et al. 2006) can show conflict with large-scale studies of nucleotides or other RGCs (e.g., Kriegs et al. 2006 compared with Murphy et al. 2007; Wildman et al. 2007; Prasad et al. 2008). Several patterns of TE distribution are possible (Fig. 1). Although conflicts with the species tree are one potential pattern (Fig. 1a), other potential patterns of TE distribution can also lead to conflict (e.g., Fig. 1b,c). It is unclear how much of the conflict observed in published studies can be explained by conflicts among gene trees (hemiplasy) rather than homoplasy.

The argument that TE insertions exhibit little or no homoplasy is ultimately based upon assumptions about their biology. TEs are divided into two major classes that exhibit fundamental mechanistic differences: retrotransposons (Class I elements) use a “copy-and-paste” mechanism with an RNA intermediate, whereas DNA transposons (Class II elements) typically use a “cut-and-paste” mechanism with a DNA intermediate (Finnegan 1989; Wicker et al. 2007). Retrotransposons are generally more common than DNA transposons in eukaryotes and they are less likely to undergo precise (or nearly precise) excision (Labrador and Corces 1997; Wicker et al. 2007). Most TEs used in vertebrate phylogenetics such as L1 elements (e.g., Nishihara et al. 2006a, 2009) and chicken repeat 1 (CR1) elements (e.g., Kaiser et al. 2007; Kriegs et al. 2007), are retrotransposons that share an insertion mechanism called target-primed reverse transcription (Luan et al. 1993; Ichiyanagi and Okada 2008). Briefly, an endonuclease nicks the target DNA to generate a DNA strand with a free 3′-hydroxyl that is able to act as a primer for reverse transcription of the retrotransposon RNA. This mechanism has the potential to result in a bias toward specific insertion sites depending on the degree of endonuclease specificity, which appears to range from very strong (e.g., Xiong and Eickbush 1988; Feng et al. 1998) to relatively weak (e.g., Jurka 1997; Ichiyanagi and Okada 2008). There are likely to be a number of factors, in addition to endonuclease specificity, that can alter patterns of TE insertion accumulation over evolutionary time. Thus, predicting the probability that specific TE types will exhibit homplasy remains difficult, making it critical to evaluate this empirically.

Phylogenetic analyses using TEs have identified them using one of two methods. First, specific TE insertions can be targeted for polymerase chain reaction (PCR) amplification from all the taxa of interest (e.g., Sasaki et al. 2004; Kaiser et al. 2007). Second, TEs can be identified in silico by searching large-scale homologous sequences or even whole genomes (e.g., Kriegs et al. 2006). Although comparing large-scale genomic regions is less biased, the set of organisms with sufficient genomic data available remains limited. Some large-scale phylogenetic data sets have a sufficient amount of noncoding sequence to apply the second method with the added advantage of broader taxon sampling. Thus, searching phylogenetic data sets may improve our understanding of TE insertion patterns as well as help to identify TEs that are phylogenetically informative.

The large-scale avian phylogenetic data published by Hackett et al. (2008) is suitable for this last approach. This study included a large amount of noncoding data from 169 avian species (representing all orders, most nonpasserine families, and all major passerine clades), providing a much more extensive taxon sampling than is currently available for genome sequences, where only the chicken genome has been examined (Wicker et al. 2005). Using data from Hackett et al. (2008) and related papers (Chojnowski et al. 2008; Harshman et al. 2008; Yuri et al. 2008), we 1) establish the distribution of TE insertions and determine their potential to resolve phylogenetic questions in birds; 2) ask whether all TE insertions in this data set represent perfect, or homoplasy-free, characters on the Hackett et al. (2008) tree; and 3) assess the types of TEs found in a broad diversity of birds.

Materials and Methods

Sequencing and Alignment

Because most TEs in coding regions are selected against, we focused on screening noncoding DNA. We screened the data available from recent studies that examined avian phylogeny using noncoding sequences (Chojnowski et al. 2008; Hackett et al. 2008; Harshman et al. 2008). One locus, *HMGN2*, was poorly sampled in the previous studies, so we collected additional data from some of the same taxa used in those studies (deposited in GenBank with accession numbers HM439436-HM439451). Introns, coding exons, and untranslated regions (UTRs; noncoding exon regions) were identified using the annotation of the chicken genome (International Chicken Genome Sequencing Consortium 2004) and other vertebrate genomes (Hubbard et al. 2007). Sequences were aligned as described in previous publications (Chojnowski et al. 2008; Hackett et al. 2008; Harshman et al. 2008).

TE Identification

We used individual introns or UTRs as queries to search for homology to TEs from all organisms in Repbase (Jurka et al. 2005) using the CENSOR software tool (Kohany et al. 2006). To allow careful comparison of TE insertion boundaries, TE insertion positions identified using CENSOR were mapped onto the multiple sequence alignments using a C++ program written by E.L.B. In some cases, we modified alignments to better match the novel information about TE boundaries. Upon further examination of the alignments, some insertions were found in additional taxa not identified by CENSOR, though in all these cases, these taxa had short sequences (<40 bp) that appeared homologous based upon the alignment and so we considered them as representing TE insertions. For comparison, we also searched for TEs using RepeatMasker (Smit et al. 2004), another database of repetitive elements with a different search algorithm.

In addition, all insertions >40 bp in length in the original alignments were identified and examined to determine whether any could be identified as TEs. The cut off of 40 bp was used because all TE insertions identified by CENSOR were longer than 40 bp, suggesting that it is difficult to reliably identify TEs shorter than these sequences. All insertions that were not identified as TEs in the initial CENSOR or RepeatMasker searches were rerun through CENSOR using just the inserted region rather than the entire intron. Length differences between paleognaths and neognaths cannot be classified as insertions or deletions because the basal split in birds is between paleognaths and neognaths (e.g., Groth and Barrowclough 1999; Harshman et al. 2008) so they were excluded from consideration (none of these length differences appeared to be due to TEs).

Examining Homoplasy and Gene-Tree Topologies

Each TE insertion, including those that appeared to be homologous but were too short (e.g., those that were <40 bp) to be identified through CENSOR, was coded as a binary character (present/absent) for each taxon and mapped onto the nucleotide-derived tree of Hackett et al. (2008). In those cases where we were missing sequence data for specific taxa, we assumed that the taxon with missing data had the same character state as its sister taxon. When the distribution of TE insertions conflicted with the Hackett et al. (2008) tree, we determined whether the observed pattern could reflect the gene tree in the region surrounding the TE insertion rather than homoplasy. To examine gene trees, we used GARLI 0.96b8 (Zwickl 2006) to generate the maximum likelihood (ML) tree using the general time reversible model with Γ-distributed rates and invariant sites (GTR+Γ+inv) model for the locus containing the TE insertion (excluding the sites in the TE itself), and we examined support for that tree using 100 bootstrap replicates. We also examined the phylogenetic signal at individual sites by using PAUP* 4.0b10 (Swofford 2003) to calculate site likelihoods given the ML tree for the locus and the optimal tree rearranged to require only a single RGC and the GTR+Γ+inv model. This analysis allowed us to determine whether sites clustered near the TE insertion supported a different gene tree than the remainder of the locus, which is expected to be the case if recombination had occurred near the insertion site. All trees and data matrices were deposited in TreeBase (S10968). Additionally, the TE character matrix and sequence alignments are available on http://www.biology.ufl.edu/earlybird/.

Testing for Among-Locus Variation in the TE Insertion Rate

To test the hypothesis that TE insertion rates are equal across loci, we compared the simplest evolutionary model, a global Poisson model with equal rates (e.g., Braun and Kimball 2001), to the more general negative binomial (NB) model, which allows variable rates. In both models, the expected number of TE insertions at a locus is proportional to the length (Len) of the locus and the treelength (TL), which is the sum of the branch lengths for the relevant taxa. The rate of TE insertion (λ_{tr}) can be expressed as the expected number of insertions per base pair of noncoding DNA per myr. To estimate λ_{tr}, we used the average length of noncoding DNA at each locus and approximated TL by summing branch lengths of the Hackett et al. (2008) tree after making the tree ultrametric by nonparametric rate smoothing (Sanderson 1997). Divergence times were approximated by applying a calibration to the rate-smoothed tree that assumed the origin of Neoaves was 100 Ma (a consensus estimate based upon the studies retrieved from Hedges et al. 2006). To correct for taxa that were missing for specific loci in the Hackett et al. (2008) tree, we pruned the relevant taxa while retraining the time-calibrated branch length information to calculate TL. To accomplish this, we exported a matrix of patristic distances given the complete time-calibrated tree and used least squares to fit those distances to trees generated by pruning appropriate taxa. Thus, our measure of TL represents the total amount of time (in myr) available for TE insertions given all the sequence data available for any specific locus. The ML estimate of λ_{tr} given k observed TE insertions at a locus is proportional to the probability of observing that number of substitutions given Len and TL using Equation (1):

$$P(k|\lambda_{\mathrm{tr}}, \mathrm{Len}, \mathrm{TL}) = \frac{(\lambda_{\mathrm{tr}}[\mathrm{Len} \times \mathrm{TL}])^k e^{-\lambda_{\mathrm{tr}}[\mathrm{Len}\times\mathrm{TL}]}}{k!}. \quad (1)$$

The NB model is similar, but it adds a nonnegative variance inflation parameter (c) to the other variables used in Equation (1):

$$\begin{aligned} P(k|\lambda_{\mathrm{tr}}, \mathrm{Len}, \mathrm{TL}, c) = {} & \frac{(\lambda_{\mathrm{tr}}[\mathrm{Len} \times \mathrm{TL}])^k}{k!} \\ & \times \frac{\Gamma(1/c + k)}{\Gamma(1/c)(\lambda_{\mathrm{tr}}[\mathrm{Len} \times \mathrm{TL}] + 1/c)^k} \\ & \times \left(1 + \frac{\lambda_{\mathrm{tr}}[\mathrm{Len} \times \mathrm{TL}]}{1/c}\right)^{-1/c}. \quad (2) \end{aligned}$$

The likelihood ratio test is straightforward because the NB and Poisson models differ by a single parameter (Equation (2) reduces to Equation (1) when $c = 0$). This allows us to compare the null hypothesis of equal rates of TE insertion at different loci to the alternative hypothesis of variable rates of TE insertion across loci using a likelihood ratio test.

Results

We identified 66 distinct insertions of TEs by searching 34 introns and 2 UTRs (∼14 kb per species) from 17 loci (Table 1; see also online Table S1 for a complete list of the TE insertions that we identified, available from http://www.sysbio.oxfordjournals.org/). Neither of the UTRs had a TE insertion in any taxon, therefore we focus on introns hereafter. Two loci (comprising 4 introns) lacked TE insertions in any intron, and an additional 6 introns (distributed in 3 loci) lacked TE insertions despite the presence of TE insertions in other introns sequenced for those loci (Table 1). One insertion, a polinton (a DNA transposon; Kapitonov and Jurka 2006), was not identified in the initial CENSOR screen using the entire intron, but it was identified when CENSOR was used to examine the inserted sequence alone.

Table 1. Loci and introns searched for transposons, and the type of identified TE insertions

Locus	Chr.[a]	Mean	Mean	Number of TE insertions		Types
		length[b]	GC[b]	By locus	By intron	
ALDOB	Z			4		
Intron 3		493.8	0.44		2	CR1, SINE
Intron 4		153.3	0.49		0	
Intron 5		215.4	0.45		2	CRI, ERV
Intron 6		478.4	0.40		0	
Intron 7		150.8	0.48		0	
CLTC	19			4		
Intron 6		733.6	0.40		2	CR1, ERV
Intron 7		634.0	0.42		2	CR1
CLTCL1	15			2		
Intron 7		477.5	0.42		2	CR1, ERV
CRYAA	1			1		
Intron 1		940.3	0.55		1	CR1
EEF2	28			17		
Intron 5		346.9	0.49		5	CR1
Intron 6		298.2	0.49		1	CR1
Intron 7		178.6	0.55		3	CR1
Intron 8		376.1	0.51		8	CR1
FGB	4			6		
Intron 4		596.7	0.35		0	
Intron 5		542.6	0.37		2	CR1, Polinton
Intron 6		181.8	0.35		0	
Intron 7		826.9	0.35		4	CR1
GH1	27			9		
Intron 2		637.0	0.52		7	CR1
Intron 3		365.3	0.49		2	CR1
HMGN2	23			8		
Intron 2		353.8	0.37		1	CR1
Intron 3		314.4	0.39		1	CR1
Intron 4		347.1	0.4		5	CR1
Intron 5		421.5	0.42		1	CR1
IRF2	4			1		
Intron 2		607.4	0.40		1	CR1
MB	1			2		
Intron 2		694.1	0.46		2	CR1
MUSK	Z			2		
Intron 3		602.2	0.39		2	CR1, SINE
MYC	2			1		
Intron 2		317.5	0.46		1	CR1
PCBD1	6			7		
Intron 2		353.0	0.46		2	CR1
Intron 3		512.9	0.52		5	CR1
RHO	12			0		
Intron 1		910.8	0.52		0	
Intron 2		107.8	0.70		0	
Intron 3		219.1	0.66		0	
TGFB2	3			2		
Intron 5		570.1	0.44		2	CR1
TPM1	10			0		
Intron 6		459.6	0.39		0	

[a] Chr. = Chromosome.
[b] Excludes TEs.

As expected from previous studies (e.g., Wicker et al. 2005), CR1 retroelements were the most common TEs in our data set (60 out of 66; Table 2). CR1 insertions are also the TE type most commonly targeted for avian phylogenetics (Watanabe et al. 2006; Kaiser et al. 2007; Kriegs et al. 2007; Treplin and Tiedemann 2007). Three of the remaining TEs were endogenous retroviruses (ERVs), another group of elements that are relatively common in the chicken genome (Wicker et al. 2005; Weiss 2006; Huda et al. 2008). There were two short interspersed repetitive elements (SINEs; a DeuSINE [Nishihara et al. 2006b] and an RTE-related SINE [Jurka 2008]) and a polinton (Kapitonov and Jurka 2006), all of which are rare TEs in the chicken genome. RepeatMasker identified most of the CR1 insertions that were found by CENSOR, but it failed to identify the other TE types, so the remainder of the results focuses on the CENSOR output.

Full-length CR1s, ERVs, and polintons are longer than 4 kb (Haas et al. 1997; Huda et al. 2008), but all insertions in our data were partial insertions that ranged from approximately 40–900 bp for CR1 insertions and 60–600 bp for ERVs (though the PCR conditions used would have been unlikely to amplify introns with full-length insertions). Although SINEs are relatively short, the SINE insertions in our data set are also partial. The partial insertions of TEs evident in our data set are typical of the majority of TE insertions found in genomic surveys of birds and other organisms (Petrov et al. 2003; Wicker et al. 2005; Abrusán et al. 2008), and these partial insertions are typical of the TE insertions used for phylogenetics (Kriegs et al. 2006, 2007; Kaiser et al. 2007).

Of the 82 large (>40 bp) insertions in the alignments, 73% were identified as TEs. Not only were most insertions attributable to TEs but also the TE insertions were larger on average (∼300 bp) than other large insertions (∼125 bp). This suggests that TE insertions explain much of the large-scale size variation in the intron data sets (Chojnowski et al. 2008; Hackett et al. 2008; Harshman et al. 2008).

TE insertions were more common in some loci than in others, even after correcting for the amount of noncoding DNA sequenced (Table 1). In fact, we could reject the equal-rate Poisson model in favor of the NB model using the likelihood ratio test ($2\delta = 16.55$, $P < 0.0001$, df = 1), indicating that the rate of TE insertion/fixation varied across loci. The 2 loci with the largest number of TE insertions were *EEF2* (17 TEs; with ∼1140 bp of intronic sequence per species) and *GH1* (9 TEs with ∼740 bp of intronic sequence per species).

Some clades experienced more TE insertions than others (Fig. 2). For example, a superordinal clade

TABLE 2. TE types and insertion patterns

Type	Number of insertions	Number of autapomorphic insertions	TE insertion Pattern 1b[a] (Fig. 1b)	TE insertion Pattern 1c (Fig. 1)	TE insertion Pattern 1d[a] (Fig. 1)
CR1	60	34	4 (2 unique sites)	2	7 (4 taxa)
ERV	3	2	0	0	0
SINE	2	2	0	0	1 (with a CR1)
Polinton	1	0	0	0	0

[a] When 2 independent insertions are hypothesized to occur at the same site (e.g., Fig. 1b,d), we counted each independently.

comprising the paraphyletic Coraciiformes (kingfishers, rollers, bee-eaters, hoopoes, and hornbills) and the Piciformes (woodpeckers, barbets, jacamars, and puffbirds) had 19 TE insertions, almost 30% of the TE insertions identified. Most (11) of these 19 insertions were specific to the Piciformes. Other clades with numerous insertions include the Cuculiformes (cuckoos and anis) with 9, four of which were unique to the Yellow-billed cuckoo (*Coccyzus americanus*), and the Charadriiformes (shorebirds and their allies) with 6 insertion events in the Lari (gulls) and Scolopaci (sandpipers) (Fig. 2). In other orders, such as Anseriformes (ducks and geese), we found no TE insertions. Although the number of ERVs was quite small, the distribution was also skewed, with two (of three) ERV insertions found in a single order, the Galliformes (chickens, turkeys, pheasants, and their allies).

CR1 elements, like other types of TEs (e.g., Boissinot et al. 2000), are divided into subtypes that can be distinguished based upon their sequence (Vandergon and Reitman 1994). There is typically a small number of complete and actively transcribed retrotransposons in genomes. These intact TEs, often referred to as "master genes," give rise to many copies inserted throughout the genome and the subtype of all insertions will correspond to the subtype of the master gene. Because the master gene for a specific subtype can remain active for a relatively long time period, one or two subtypes may dominate in a clade (Kriegs et al. 2007). Consistent with this, we found that clades with large numbers of TE insertions generally had multiple insertions of a single subtype. For example, the most common CR1 subtype in the Coraciiformes and Piciformes clade and the Cuculiformes was F2, whereas the most common subtype in Charadriiformes was Y4.

Subtype identification, however, was problematic in some cases. For example, what appeared to be a homologous insertion was identified as a different CR1 subtype in closely related species (e.g., in *EEF2* intron 8, a TE insertion shared by all 22 passerines sampled was identified as five different subtypes). Furthermore, in some cases, RepeatMasker identified different subtypes than CENSOR for the same insertion (not shown). Given this, three important factors should be considered before using subtype identification: 1) the length of the insertion, because short insertions will have retained less information about subtypes than longer insertions; 2) the age of the insertion, because older insertions have undergone more mutation and may be harder to identify; and 3) the database used to identify subtypes.

Most TE insertions could be mapped onto the Hackett et al. (2008) tree (and even very divergent phylogenies such as those in Sibley and Ahlquist 1990 and Livezey and Zusi 2007) without homoplasy (Fig. 2). Indeed, the majority of TE insertions (38 insertions) were autapomorphic given our taxon sample (Table 2). The synapomorphic insertions occurred on relatively long branches on the phylogeny (see Fig. 3 in Hackett et al. 2008) and generally defined clades that were already well supported by analyses of nucleotide substitutions (Hackett et al. 2008), and thus provided no new phylogenetic information. Most of these united families or more derived groupings (Fig. 2), with only 11 insertions uniting orders or deeper-level clades. Ten of these 11 insertions united well accepted, monophyletic orders (Passeriformes [perching birds], Psittaciformes [parrots], Piciformes, Trogoniformes [trogons], Cuculiformes, and Columbiformes [doves]; some of these orders were united by two insertions). The remaining deeper-level insertion united Coraciiformes and Piciformes, which is a well-supported superordinal group in Hackett et al. (2008).

After careful examination of all the alignments, we identified a small number of sites that exhibited more complex patterns of TE insertion (Fig. 1; Table 2). We split these into two categories. The first category appeared to reflect insertion "hot spots" in the genome, whereas the second category appeared to reflect homoplasy in that the TEs were within a single clade, but the insertion did not map onto well-supported nodes without homoplasy (Fig. 1c).

TEs in the hotspot category appeared to be independent insertions at identical or nearly identical (within a few nucleotides) sites. We found six potential hot spots that were characterized by two patterns of insertion at these sites (Fig. 1b,d). At four of these sites, it appeared that two independent insertions had occurred at the same site in the same taxon (Fig. 1d, Species 1); for these, we scored each insertion event independently. Of these four cases, one was identified because the insertions were in different directions. The second was identified because the same region of the CR1 was included in each insertion event. In the third case, the two insertions represented different segments of a CR1 and did not align well to a single subtype. In the last

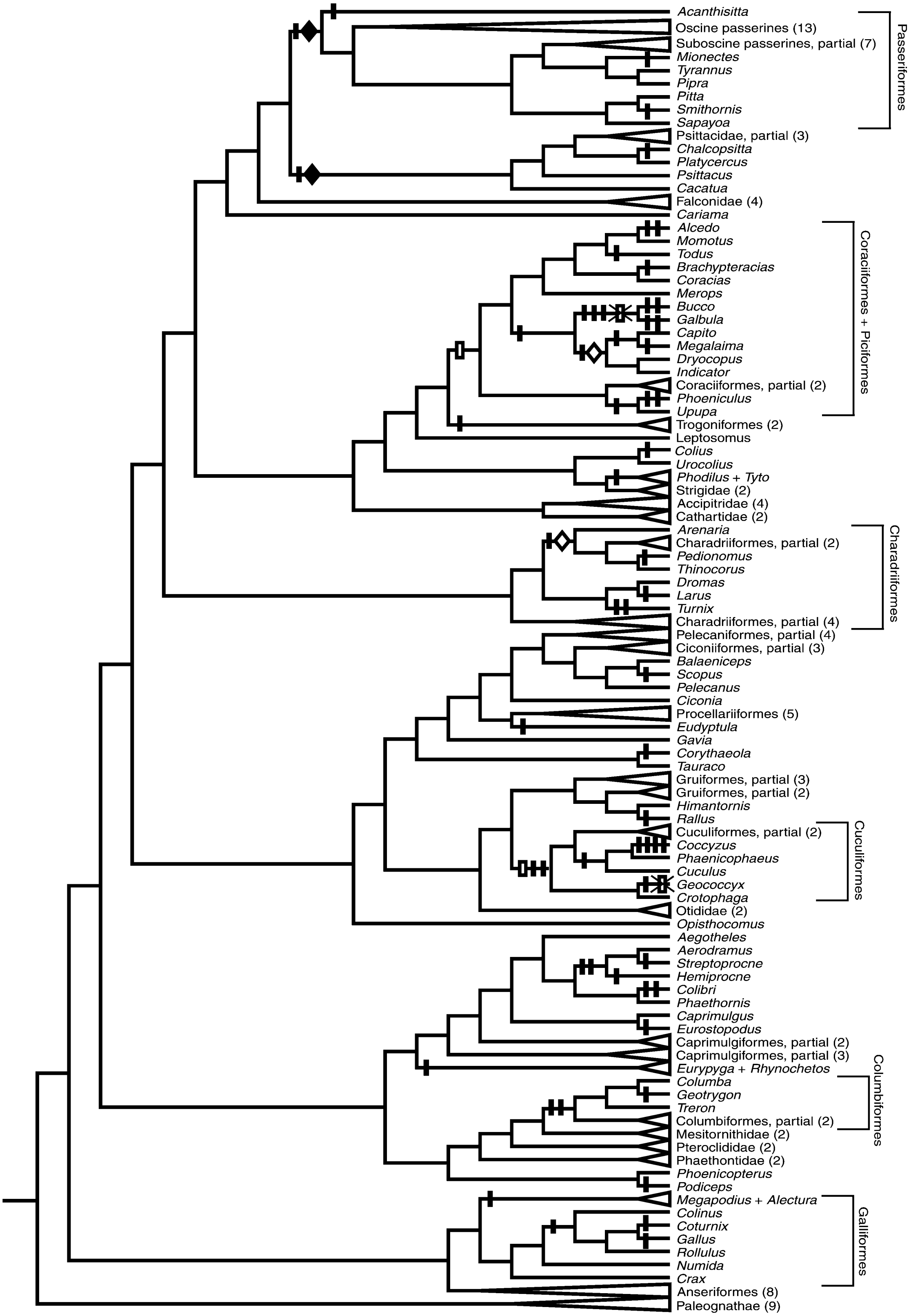

FIGURE 2. Phylogeny from Hackett et al. (2008) showing hypothesized TE insertions. Clades with no insertion events are collapsed for simplicity, and the number of taxa included in that clade is noted in parentheses. Symbols used are identical to Figure 1. There are 2 pairs of independent insertions with one pair represented by an open diamond and the other pair by a closed diamond.

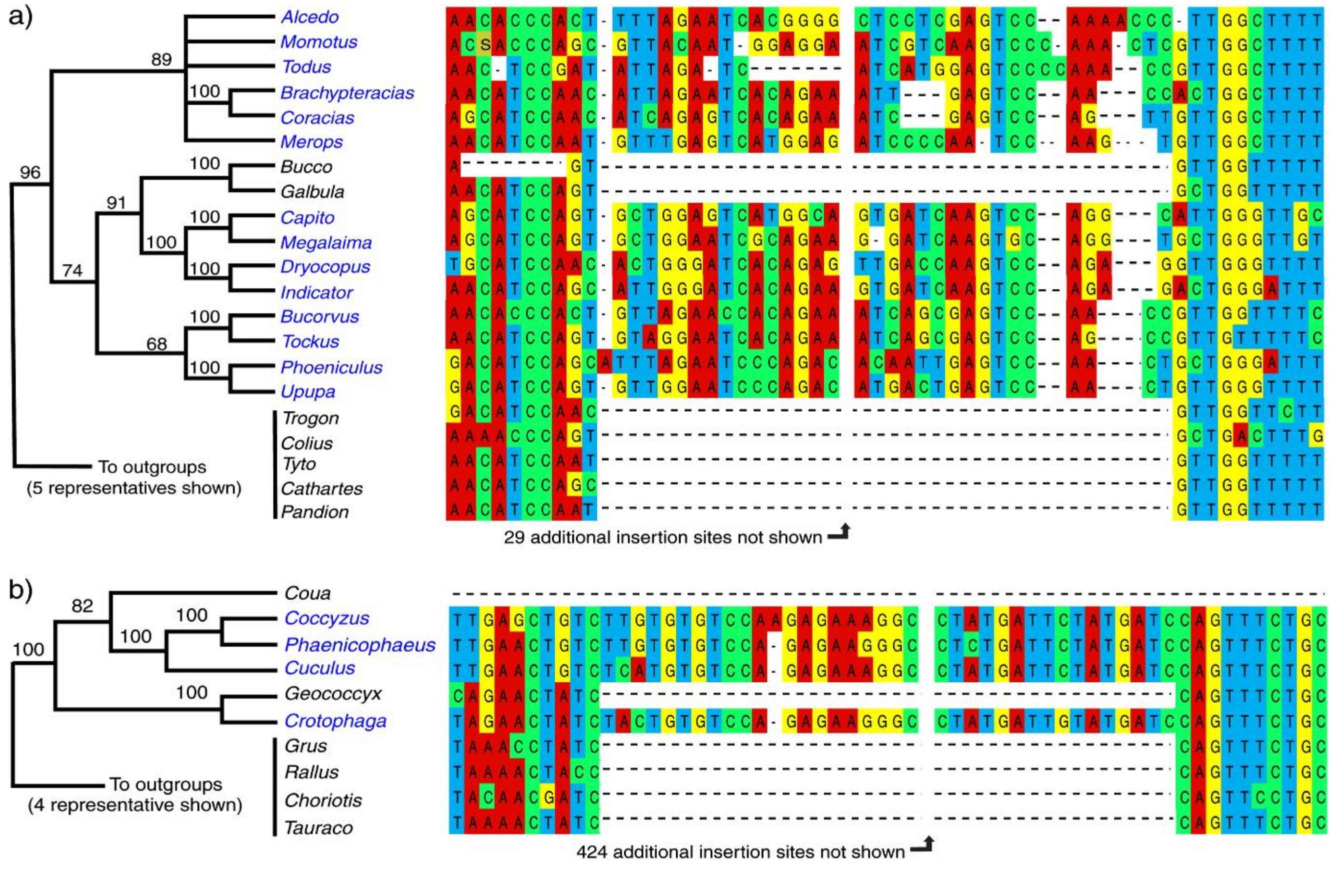

FIGURE 3. ML bootstrap analysis and alignments of loci that include insertions that appear homoplastic. ML bootstrap support >50% are shown. Analyses were run with all taxa for which we had data, although only the clades of interest are shown for each gene tree. Alignments are for the same taxa, with several related taxa shown for comparison. a) *HMGN2*, showing the absence of the insertion in *Bucco* and *Galbula* sequences. b) *GH1*, showing the absence of the insertion in *Geococcyx*. For this locus, *Coua* had a large deletion spanning the entire region (including much of the flanking intron). Sequences for one cuculiform, *Centropus*, could not be obtained for *GH1* and so it is not included in this figure.

case, there were different types (SINE and CR1). Some of these cases could be explained without invoking a hotspot model. The first two of these double-TE insertions (those that involve the same region of a specific element) could reflect duplication of the insertion (combined with an inversion for the TE insertion in different directions), whereas the third could involve a deletion event in the middle of an insertion combined with a high mutation rate (necessary to explain our observation that the two regions do not align to the same subtype with a high degree of identity). However, the insertion of distinct TE types (e.g., the SINE and CR1) at the same site must reflect independent insertions. Regardless of the specific mechanism(s) responsible for each of these insertions, it is clear that hot spots can be detected using a data set of the size we examined here and that determining whether specific TE insertions are homologous has the potential to be difficult.

Two sites with insertions at essentially the same site in different orders (i.e., similar to Species 1 and 5 in Fig. 1b) were also identified, suggesting the observed TE insertions had independent origins in each of the lineages, and providing additional evidence for the existence of hot spots in avian genomes. The first example of an independent insertion was identified because there was a large phylogenetic distance between clades with the TE insertions (Fig. 2 open diamonds). Uniting the clades with the insertions would require rearranging multiple strongly supported branches in the Hackett et al. (2008) phylogeny that are congruent with other estimates of avian phylogeny (e.g., Livezey and Zusi 2007). Furthermore, assuming that these insertions (TE insertions 24 and 25 in Table S1) have a single origin would render another TE (TE insertion 36 in Table S1) homoplastic and increase the number of changes necessary to explain the distribution of a second TE insertion (TE insertion 50 in Table S1; also see below for more information about this insertion). The second example was identified because the TE insertions were in different directions (Fig. 2 filled diamonds), although it remains possible that this pattern reflects a single insertion followed by a precise inversion of the TE region. Neither of these TE insertions were phylogenetically misleading (presuming independent insertion can be identified through the patterns we observed), but they do provide evidence for the existence of hot spots for TE insertion and/or fixation in the genome. In these situations, careful examination of the sequences (examining directionality, the segments of the TE present in the insertion, and whether the TE insertions are divergent types), as well as phylogenetic structure, helped identify insertions that were likely to be independent.

We identified two insertions in the homoplasy category (Fig. 1b). In *HMGN2* intron 4, a CR1 insertion uniting the Piciformes and Coraciiformes was absent in a single clade within the Piciformes (Fig. 3a). Specifically, this insertion (TE insertion 50 in Table S1) was absent in the suborder Galbulae, represented by *Bucco* and *Galbula* in Hackett et al. (2008) but present in other Piciformes and the outgroup (Coraciiformes). In *GH1* intron 2, an insertion (TE insertion 38 in Table S1) was found in all Cuculiformes for which we had sequence data in this region except *Geococcyx* (Fig. 3b). This TE insertion was present in *Crotophaga*, the sister taxon of *Geococcyx*. Thus, both of these TE inserts are homoplastic given the Hackett et al. (2008) tree.

The apparent homoplasy associated with the TE insertions in the *HMGN2* and *GH1* loci has several potential explanations. Errors in the Hackett et al. (2008) tree are an unlikely explanation because the relevant branches are well supported by many lines of evidence, including morphology (Livezey and Zusi 2007) and other molecular studies (e.g., Ericson et al. 2006). Additionally, both of the insertions that appear homoplastic conflict with other TEs; the *HMGN2* insertion conflicts with an insertion in another locus (*GH1*; TE insertion 36 in Table S1), whereas the *GH1* insertion conflicts with a second insertion in the same locus (TE insertion 39 in Table S1) and an insertion in another locus, *EEF2* (TE insertion 26 in Table S1). Therefore, the conclusion of TE insertion homoplasy is independent of the Hackett et al. (2008) topology.

Although lineage sorting is a possible explanation for the taxonomic distribution of the TE insertions in *HMGN2* and *GH1* because gene tree–species tree conflicts are known to occur (Degnan and Rosenberg 2009), several lines of evidence indicate that TE homoplasy (e.g., Fig. 1c) is more likely than hemiplasy (i.e., Fig. 1a) for the distribution of TE insertions in *HMGN2* and *GH1*. Examination of the gene trees for *HMGN2* and *GH1* (excluding the insertion) indicates that the insertions occur on relatively long branches (see online Fig. S1, available from http://www.sysbio.oxfordjournals.org/) and there is strong support for a gene-tree topology in conflict with the insertion. In principle, lineage sorting could be reconciled with both the distribution of TE insertions and the estimates of gene trees we obtained (Fig. 3) by invoking recombination or gene conversion. These phenomena predict that sites supporting an alternative topology (a topology consistent with the distribution of the TE insertion) would be found near the insertion; however, a pattern of sites supporting a topology congruent with the TE was not evident (data not shown), suggesting neither recombination nor gene conversion is likely. Instead, the distribution we observed was most consistent with either a precise deletion of the CR1 in the ancestor to these taxa (without leaving a molecular signature of the insertion as occurs with some TE types; cf. Shedlock et al. 2004) or independent insertions of the identical portion of a CR1 in multiple ancestors within the clade. Further research on the mechanisms of insertion and deletion for these TEs may reveal the most plausible pathway but either alternative could lead to incorrect phylogenetic conclusions.

Discussion

We found TE insertions in the intron partitions of most genes that we examined, consistent with the expectation that they are located throughout avian genomes. Almost all large insertions in our alignments were TEs, suggesting that TEs explain much of the observed variation in intron length. In agreement with recent studies that have used TE insertions for phylogenetic estimation, we found many insertions that defined widely accepted clades. However, we also found evidence of homoplasy. Although concern about homoplasy in TE data is not novel (e.g., Hillis 1999; Miyamoto 1999; Ray et al. 2006), many authors have suggested that apparent homoplasy of TEs with respect to the species tree can be explained by hemiplasy (e.g., Shedlock and Okada 2000; Kriegs et al. 2006; Nishihara et al. 2006a; Ray et al. 2006; Kaiser et al. 2007; Kriegs et al. 2007; Treplin and Tiedemann 2007). Nevertheless, lineage sorting is an unlikely explanation for the two cases of homoplasy we identified. Instead our data suggest that hot spots for TE insertions (and/or the fixation of TE insertions) reduce state space for this type of RGC, that precise deletion of these TEs can occur or that both phenomena contribute to homoplasy in avian TEs.

Insertion Sites, Hot spots, and Deletions

Regions of the genome can be TE-free (Simons et al. 2007), and when those observations are combined with our study, it seems clear that rates of TE insertion and/or fixation exhibit substantial variation across the avian genome. An exceptionally large number of TE insertions were found in specific introns, suggesting that they are hot spots for TE insertion or fixation. In fact, we identified four sites with multiple insertions (e.g. Fig. 1d) and two sites in which insertions occurred independently in divergent taxa (e.g., Fig. 1b). St. John and Quinn (2008) noted that recent CR1 insertions frequently had a TTCT sequence flanking the 3′ end of the insertion, suggesting a bias toward insertion at sites with this specific motif. This observation is consistent with the target-primed reverse transcription mechanism of retrotransposon insertion (see above), which involves endonuclease-mediated nicking of the target DNA followed by base pairing between conserved elements at the end of the TE (e.g., TTCT for CR1 elements) and the target sequence. We did not find the TTCT sequence flanking any of the CR1 insertions we identified, though St. John and Quinn (2008) reported that the motifs degraded and were mostly associated with very recent insertions. Thus, our failure to identify conserved TTCT motifs suggests that the insertions we identified are too ancient for preservation of the motif, although it is also possible that the elements we identified inserted through a variation of this mechanism.

Excision of TEs also has the potential to contribute to the observed phylogenetic distribution of insertions (for another possible example, see Murphy et al. 2007). In fact, the most parsimonious explanation for the homoplasy in *HMGN2*, assuming that insertions and deletions are weighted equally, would be insertion followed by a precise deletion of the entire insertion in some taxa within a clade. The alternative hypothesis, which is less parsimonious given equal weighting of insertions and deletions, would require three independent insertions given the Hackett et al. (2008) topology. Although the degree to which natural selection favors deletion of TEs is not known, selection may favor deletion for at least some TE classes (Petrov et al. 2003) and the potential for homoplasy due to TE deletions should not be ignored.

Phylogenetic Considerations

TE insertions retain a strong phylogenetic signal and have substantial potential for phylogenetic analyses. They exhibit very little homoplasy (Fig. 2); the retention index (RI) of TE insertions on the Hackett et al. (2008) tree is 0.97, much greater than that of Hackett et al. (2008) sequence data on the same tree ($\mathrm{RI}_{\mathrm{intron}} = 0.52$, $\mathrm{RI}_{\mathrm{coding\ exons}} = 0.54$, $\mathrm{RI}_{\mathrm{UTR}} = 0.58$). However, most of the TE insertions identified here were autapomorphic or united more recently diverged clades (e.g., they united families) that were already well supported by sequence data. A major reason for this may be the structure of the avian tree in which many clades arose during a short period (Chojnowski et al. 2008; Hackett et al. 2008). This means that many of the deep branches in the avian tree of life are very short, making the probability of accumulating a synapomorphic insertion on these internodes quite low (see Braun and Kimball 2001) and inflating the probability of hemiplasy. Consistent with the low probability of observing insertions that occurred along these short branches, all the synapomorphic insertions we observed occurred upon the longer internodes in the Hackett et al. (2008) tree that are well supported in nucleotide analyses. Another potential reason for the absence of TE insertions that unite groups defined by these short branches deep in the avian tree of life is that older insertions may be difficult to identify due to a bias toward deletion of these elements or the accumulation of other mutations over time that can obscure TE identification. Regardless of the basis for the pattern we observe, our results suggest that TE insertions may have the greatest potential to be phylogenetically informative within orders and families in birds where insertion events are easier to identify and characterize.

The observation that independent TE insertions can occur at the exact same site in the same or different taxa, or can be precisely deleted, suggests that care needs to be taken in assigning character states for phylogenetic analyses. Although subtype identification could help to clarify complex patterns of TE insertion, subtype identification is also complicated by the accumulation of both point mutations and indels after the insertions occur. Indeed, the bias toward deletion at the 5′ end (Abrusán et al. 2008) has the potential to result in short remnants of CR1s that cannot be reliably identified by subtype. In addition, the master gene model for retrotransposons predicts that many insertions within a specific lineage are likely to be the same subtype (Watanabe et al. 2006; Kriegs et al. 2007), further limiting the ability of subtype identification in teasing out more complex situations. In all cases, however, careful examination of the sequences and the alignment will help establish boundaries and aid in determining whether specific TE insertions are likely to be independent or shared.

Our results are consistent with analyses of the chicken genome (Wicker et al. 2005) and suggest that it might be most profitable to continue targeting CR1s for avian phylogenetics (e.g., Watanabe et al. 2006; Kaiser et al. 2007; Kriegs et al. 2007; Treplin and Tiedemann 2007) rather than the less common ERVs and SINEs. The ERVs we identified occurred at a lower frequency (∼5% of insertions) in our data than in the chicken genome (∼15% of insertions) (Huda et al. 2008). This may either reflect our more limited genomic sampling, or it may indicate that the chicken (or Galliformes as a whole) may have more ERV insertions than other birds. The latter hypothesis is consistent with the observation that two of the three ERVs we identified were in members of the Galliformes, although the small number of ERVs identified does not allow us to draw firm conclusions.

Conclusions

The TE insertions identified here provide support for a number of branches in the avian tree of life (Fig. 2). It is clear that TEs have the potential to provide additional evidence regarding relationships when nucleotides provide surprising or conflicting results. We found that having sequence data helped to clarify the independence of insertions, emphasizing the importance of sequencing TE insertions. Our results also suggest that TEs should not be viewed as perfect characters exempt from homoplasy. Instead, TE insertions present many of the same challenges for phylogenetic analyses as other types of data, such as nucleotide sequences. Available statistical methods for the analysis of TEs assume that any apparent homoplasy is due to differences between gene trees and species trees (Waddell et al. 2001). However, hemiplasy due to gene tree–species tree conflicts were not consistent with the homoplasy evident in our study. Ultimately, analytical methods for RGCs that can accommodate both hemiplasy and homoplasy are likely to prove more useful. An even more productive approach may be to develop methods that can integrate data from TE insertions into large-scale analyses of nucleotide sequences, potentially along with information about other types of RGCs. Integrated approaches of this type will ultimately allow analyses that can recover accurate phylogenomic estimates using all available information.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found at http://www.sysbio.oxfordjournals.org/.

Funding

This work was supported by the U.S. National Science Foundation Assembling the Tree of Life Program (DEB-0228682 to R.T.K., E.L.B., and D.W.S., DEB-0228675 to S.J.H., DEB-0228688 to F.H.S., and DEB-0228617 to W.S.M.).

Acknowledgments

This manuscript was improved by comments from J. Gordon Burleigh, Carol L. Chaffee, Julie M. Allen as well as two anonymous reviewers and our editors. Naomi R. Iahusz-Valez helped test insertion rate heterogeneity.

References

Abrusán G., Krambeck H.J., Junier T., Giordano J., Warburton P.E. 2008. Biased distributions and decay of long interspersed nuclear elements in the chicken genome. Genetics. 178:573–581.

Avise J.C., Robinson T.J. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. Syst. Biol. 57:503–507.

Barker F.K., Cibois A., Schikler P., Feinstein J., Cracraft J. 2004. Phylogeny and diversification of the largest avian radiation. Proc. Natl. Acad. Sci. U.S.A. 101:11040–11045.

Boissinot S., Chevret P., Furano A.V. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. 17:915–928.

Braun E.L. 2003. Innovation from reduction: gene loss, domain loss and sequence divergence in genome evolution. Appl. Bioinformatics. 2:13–34.

Braun E.L., Grotewold E. 2001. Fungal zuotin proteins evolved from MIDA1-like factors by lineage-specific loss of MYB domains. Mol. Biol. Evol. 18:1401–1412.

Braun E.L., Kimball R.T. 2001. Polytomies, the power of phylogenetic inference, and the stochastic nature of molecular evolution: a comment on Walsh et al. (1999). Evolution. 55:1261–1263.

Chojnowski J.L., Kimball R.T., Braun E.L. 2008. Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. Gene. 410:89–96.

Cox W.A., Kimball R.T., Braun E.L. 2007. Phylogenetic position of the New World quail (Odontophoridae: eight nuclear loci and three mitochondrial regions contradict morphology and the Sibley-Ahlquist tapestry. Auk. 124:71–84.

Crowe T.M., Bowie R.C.K., Bloomer P., Mandiwana T.G., Hedderson T.A.J., Randi E., Pereira S.L., Wakeling J. 2006. Phylogenetics, biogeography and classification of, and character evolution in, gamebirds (Aves: Galliformes): effects of character exclusion, data partitioning and missing data. Cladistics. 22:495–532.

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

Ericson P.G.P., Anderson C.L., Britton T., Elzanowski A., Johansson U.S., Källersjö M., Ohlson J.I., Parsons T.J., Zuccon D., Mayr G. 2006. Diversification of Neoaves: integration of molecular sequence data and fossils. Biol. Lett. 2:543–547.

Feng Q., Schumann G., Boeke J.D. 1998. Retrotransposons R1Bm endonuclease cleaves the target sequence. Proc. Natl. Acad. Sci. U.S.A. 95:2083–2088.

Finnegan D.J. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5:103–107.

Gibb G.C., Kardailsky O., Kimball R.T., Braun E.L., Penny D. 2007. Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. Mol. Biol. Evol. 24: 269–280.

Groth J.G., Barrowclough G.F. 1999. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. Mol. Phylogenet. Evol. 12:115–123.

Haas N.B., Grabowski J.M., Sivitz A.B., Burch J.B.E. 1997. Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. Gene. 197:305–309.

Hackett S.J., Kimball R.T., Reddy S., Bowie R.C.K., Braun E.L., Braun M.J., Chojnowski J.L., Cox W.A., Han K.-L., Harshman J., Huddleston C.J., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Steadman D.W., Witt C.C., Yuri T. 2008. A phylogenomic study of birds reveals their evolutionary history. Science. 320:1763–1768.

Harshman J., Braun E.L., Braun M.J., Huddleston C.J., Bowie R.C.K, Chojnowski J.L., Hackett S.J., Han K.-L., Kimball R.T., Marks B.D., Miglia K.J., Moore W.S., Reddy S., Sheldon F.H., Steadman D.W., Steppan S.J., Witt C.C., Yuri T. 2008. Phylogenomic evidence for multiple losses of flight in ratite birds. Proc. Natl. Acad. Sci. U.S.A. 105:13462–13467.

Hedges S.B., Dudley J., Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 22: 2971–2972.

Hillis D.M. 1999. SINEs of the perfect character. Proc. Natl. Acad. Sci. U.S.A. 96:9979–9981.

Hubbard T.J.P., Aken B.L., Beal K., Ballester B., Caccamo M., Chen Y., Clarke L., Coates G., Cunningham F., Cutts T., Down T., Dyer S.C., Fitzgerald S., Fernandez-Banet J., Graf S., Haider S., Hammond M., Herrero J., Holland R., Howe K., Howe K., Johnson N., Kahari A., Keefe D., Kokocinski F., Kulesha E., Lawson D., Longden I., Melsopp C., Megy K., Meidl P., Ouverdin B., Parker A., Prlic A., Rice S., Rios D., Schuster M., Sealy I., Severin J., Slater G., Smedley D., Spudich G., Trevanion S., Vilella A., Vogel J., White S., Wood M., Cox T., Curwen V., Durbin R., Fernandez-Suarez X.M., Flicek P., Kasprzyk A., Proctor G., Searle S., Smith J., Ureta-Vidal A., Birney E. 2007. Ensembl 2007. Nucleic Acids Res. 35:D610–D617.

Huda A., Polavarapu N., Jordan I.K., McDonald J.F. 2008. Endogenous retroviruses of the chicken genome. Biol. Direct. 3:9.

Hudson R.R. 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics. 131:509–512.

Ichiyanagi K., Okada N. 2008. Mobility pathways for vertebrate L1, L2, CR1, and RTE clade retrotransposons. Mol. Biol. Evol. 25: 1148–1157.

International Chicken Genome Sequencing Consortium 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695–716.

Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons Proc. Natl. Acad. Sci. U.S.A. 94:1872–1877.

Jurka J. 1998. Repeats in genomic DNA: mining and meaning. Curr. Opin. Struct. Biol. 8:333–337.

Jurka J. 2008. RTE-related SINE family from a horse. Repbase Rep. 8:378.

Jurka J., Kapitonov V.V., Pavlicek A., Klonowski P., Kohany O., Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467.

Kaiser V.B., van Tuinen M., Ellegren H. 2007. Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in galliform birds. Mol. Biol. Evol. 24:338–347.

Kapitonov V.V., Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 103:4540–4545.

Keeling P.J., Doolittle W.F. 1997. Widespread and ancient distribution of a noncanonical genetic code in diplomonads. Mol. Biol. Evol. 14:895–901.

Kimball R.T., Braun E.L. 2008. A multigene phylogeny of Galliformes supports a single origin of erectile ability in non-feathered facial traits. J. Avian Biol. 39:438–445.

Kohany O., Gentles A.J., Hankus L., Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 7:474.

Kriegs J.O., Churakov G., Kiefmann M., Jordan U., Brosius J., Schmitz J. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 4:e91.

Kriegs J.O., Matzke A., Churakov G., Kuritzin A., Mayr G., Brosius J., Schmitz J. 2007. Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC Evol. Biol. 7:190.

Labrador M., Corces V.G. 1997. Transposable element-host interactions: regulation of insertion and excision. Annu. Rev. Genet. 31: 381–404.

Livezey B.C., Zusi R.L. 2007. Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. II. Analysis and discussion. Zool. J. Linn. Soc. 149:1–95.

Luan D.D., Korman M.H., Jakubczak J.L., Eickbush T.H. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 72:595–605.

Miyamoto M.M. 1999. Molecular systematics: perfect SINEs of evolutionary history? Curr. Biol. 9:R816–R819.

Murphy W.J., Pringle T.H., Crider T.A., Springer M.S., Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res. 17:413–421.

Nishihara H., Hasegawa M., Okada N. 2006a. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc. Natl. Acad. Sci. U.S.A. 103:9929–9934.

Nishihara H., Maruyama S., Okada N. 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc. Natl. Acad. Sci. U.S.A. 106:5235–5240.

Nishihara H., Smit A.F.A., Okada N. 2006b. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 16:864–874.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Petrov D.A., Aminetzach Y.T., Davis J.C., Bensasson D., Hirsh A.E. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol. Biol. Evol. 20:880–892.

Prasad A.B., Allard M.W., NISC Comparative Sequencing Program, Green E.D. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol. Biol. Evol. 25: 1795–1808.

Ray D.A., Xing J., Salem A.H., Batzer M.A. 2006. SINEs of a nearly perfect character. Syst. Biol. 55:928–935.

Rokas A., Holland P.W.H. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. 15:454–459.

Sanderson M.J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218–1231.

Sasaki T., Takahashi K., Nikaido M., Miura S., Yasukawa Y., Okada N. 2004. First application of the SINE (short interspersed repetitive element) method to infer phylogenetic relationships in reptiles: an example from the turtle superfamily Testudinoidea. Mol. Biol. Evol. 21:705–715.

Sasaki T., Yasukawa Y., Takahashi K., Miura S., Shedlock A.M., Okada N. 2006. Extensive morphological convergence and rapid radiation in the evolutionary history of the Geoemydidae (Old World pond turtles) revealed by SINE insertion analysis. Syst. Biol. 55: 912–927.

Shedlock A.M., Okada N. 2000. SINE insertions: powerful tools for molecular systematics. Bioessays. 22:148–160.

Shedlock A.M., Takahashi K., Okada N. 2004. SINEs of speciation: tracking lineages with retroposons. Trends Ecol. Evol. 19: 545–553.

Sibley C.G., Ahlquist J.E. 1990. Phylogeny and classification of birds: a study in molecular evolution. New Haven (CT): Yale University Press.

Simons C., Makunin I.V., Pheasant M., Mattick J.S. 2007. Maintenance of transposon-free regions throughout vertebrate evolution. BMC Genomics. 8:470.

Smit A.F.A., Hubley R., Green P. 2004. RepeatMasker Open-3.0 [Internet]. Available from: http://www.repeatmasker.org. (Accessed July 15, 2009).

St. John J., Quinn T.W. 2008. Recent CR1 non-LTR retrotransposon activity in coscoroba reveals an insertion site preference. BMC Genomics. 9:567.

Stechmann A., Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science. 297:89–91.

Steel M., Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

Swofford D.L. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sunderland (MA): Sinauer Associates.

Treplin S., Tiedemann R. 2007. Specific chicken repeat 1 (CR1) retrotransposon insertion suggests phylogenetic affinity of rockfowls (genus *Picathartes*) to crows and ravens (Corvidae). Mol. Phylogenet. Evol. 43:328–337.

Vandergon T.L., Reitman M. 1994. Evolution of chicken repeat 1 (CR1) elements: evidence for ancient subfamilies and multiple progenitors. Mol. Biol. Evol. 11:886–898.

Waddell P.J., Kishino H., Ota R. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Inform. 12:141–154.

Watanabe M., Nikaido M., Tsuda T.T., Inoko H., Mindell D.P., Murata K., Okada N. 2006. The rise and fall of the CR1 subfamily in the lineage leading to penguins. Gene. 365:57–66.

Weiss R.A. 2006. The discovery of endogenous retroviruses. Retrovirology. 3:67.

Wicker T., Robertson J.S., Schulze S.R., Feltus F.A., Magrini V., Morrison J.A., Mardis E.R., Wilson R.K., Peterson D.G., Paterson A.H., Ivarie R. 2005. The repetitive landscape of the chicken genome. Genome Res. 15:126–136.

Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M., Panaud O., Paux E., SanMiguel P., Schulman A.H. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8:973–982.

Wildman D.E., Uddin M., Opazo J.C., Liu G., Lefort V., Guindon S., Gascuel O., Grossman L.I., Romero R., Goodman M. 2007. Genomics, biogeography, and the diversification of placental mammals. Proc. Natl. Acad. Sci. U.S.A. 104:14395–14400.

Xiong Y., Eickbush T.H. 1988. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. Mol. Cell. Biol. 8:114–123.

Yuri T., Kimball R.T., Braun E.L., Braun M.J. 2008. Duplication and accelerated evolution of growth hormone gene in passerine birds. Mol. Biol. Evol. 25:352–361.

Zwickl D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [Ph.D. dissertation]. Austin (TX): University of Texas at Austin.