Syst. Biol. 66(4):517–530, 2017

DOI:10.1093/sysbio/syw101
Advance Access publication November 16, 2016

Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses

JEREMY M. BROWN[1,*] AND ROBERT C. THOMSON[2]

[1]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Science Building, Baton Rouge, LA 70803, USA and;* [2]*Department of Biology, University of Hawai'i at Mānoa, 2538 McCarthy Mall, Edmondson Hall Rm 216, Honolulu, HI 96822, USA*

**Correspondence to be sent to: Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA; E-mail: jembrown@lsu.edu.*

Received 19 January 2016; reviews returned 20 October 2016; accepted 21 October 2016
Associate Editor: Mark Holder

Abstract.—As the application of genomic data in phylogenetics has become routine, a number of cases have arisen where alternative data sets strongly support conflicting conclusions. This sensitivity to analytical decisions has prevented firm resolution of some of the most recalcitrant nodes in the tree of life. To better understand the causes and nature of this sensitivity, we analyzed several phylogenomic data sets using an alternative measure of topological support (the Bayes factor) that both demonstrates and averts several limitations of more frequently employed support measures (such as Markov chain Monte Carlo estimates of posterior probabilities). Bayes factors reveal important, previously hidden, differences across six "phylogenomic" data sets collected to resolve the phylogenetic placement of turtles within Amniota. These data sets vary substantially in their support for well-established amniote relationships, particularly in the proportion of genes that contain extreme amounts of information as well as the proportion that strongly reject these uncontroversial relationships. All six data sets contain little information to resolve the phylogenetic placement of turtles relative to other amniotes. Bayes factors also reveal that a very small number of extremely influential genes (less than 1% of genes in a data set) can fundamentally change significant phylogenetic conclusions. In one example, these genes are shown to contain previously unrecognized paralogs. This study demonstrates both that the resolution of difficult phylogenomic problems remains sensitive to seemingly minor analysis details and that Bayes factors are a valuable tool for identifying and solving these challenges. [Expressed sequence tags; negative constraints; ortholog; posterior probability; ultraconserved elements.]

Phylogenetic studies now make routine use of genome-scale strategies to collect data sets containing tens, hundreds, or even thousands of sequenced loci. More data should reduce stochastic error and lead to topological resolution. As predicted, many published studies now provide highly resolved phylogenetic estimates with maximal support values for all bipartitions (posterior probabilities approximated to be = 1.0, bootstrap proportions = 100%), despite the fact that different data sets and studies often support fundamentally different conclusions about evolutionary relationships. For instance, different studies concerning the root of Metazoa, and the phylogenetic position of ctenophores and poriferans, have arrived at strongly contrasting phylogenetic inferences (Dunn et al. 2008; Philippe et al. 2009; Pick et al. 2010; Chang et al. 2015; Pisani et al. 2015). Recent attempts to reconstruct the phylogenetic relationships among major groups of birds are another high-profile example of resolution within and incongruence between data sets (Jarvis et al. 2014; Prum et al. 2015). The unsettling observation of strong conflict necessarily means that data sets and/or the methods used to analyze them differ in fundamental respects. To better understand these differences, researchers need a way to investigate how much phylogenetic information each data set contains and how this information is distributed across genes and sites. Unfortunately, typical measures of support offer little ability to meaningfully explore these patterns, because their estimated values often equal one or zero and they are therefore indistinguishable from one another.

Bayes factors (BFs; Kass and Raftery 1995) offer an alternative perspective on support for topological relationships in a Bayesian context. Defined as the ratio of marginal likelihoods comparing two hypotheses (H_1 and H_2),

$$Bayes\ Factor = \frac{P(D|H_1)}{P(D|H_2)} = \frac{\frac{P(H_1|D)}{P(H_2|D)}}{\frac{P(H_1)}{P(H_2)}},$$

a BF is closely linked to the posterior probabilities currently favored in most Bayesian phylogenetic studies. One convenient interpretation of the BF is the degree to which the support for two hypotheses, in this case the presence or absence of a particular bipartition, changes after observing the data (D). Stated another way, a BF quantifies the change in the odds favoring a hypothesis when comparing the prior to the posterior. When the prior odds are even $\left(\frac{P(H_1)}{P(H_2)} = 1\right)$, the BF simply reflects the posterior odds ratio, $\frac{P(H_1|D)}{P(H_2|D)}$.

There are several reasons why one might wish to estimate BFs instead of, or in addition to, posterior probabilities for phylogenetics. The first reason is perhaps primarily psychological. BFs, and particularly log(BF)s, offer a larger numerical range to measure support than posterior probabilities. The advantage of an expanded range becomes apparent if we think about rescaling posterior probabilities into posterior odds ratios when two different data sets strongly support H_1 with posterior probabilities of 0.99 and 0.999999. While both of these values are very close to 1, they actually have a >10,000-fold difference in their posterior odds ratios $\left(\frac{0.99}{0.01}\right.$ and $\left.\frac{0.999999}{10^{-6}}\right)$. A data set that produces a

posterior probability of 0.999999 in favor of a bipartition supports that relationship much more strongly than one that produces a posterior of 0.99. Similar considerations apply when both data sets seem to reject a hypothesis with posterior estimates near 0.

The second advantage to calculating BFs concerns numerical precision and explains why posterior odds ratios are rarely used in phylogenetics. When posterior probabilities are extreme (near 0 or 1), Markov chain Monte Carlo (MCMC) does not estimate these values with sufficient precision to distinguish between 0.99 and 0.999999, despite their very different interpretations. By calculating marginal likelihoods individually with well-behaved estimators (Lartillot and Philippe 2006; Fan et al. 2010; Xie et al. 2011) and taking their ratio, extreme BFs can be estimated accurately. If the prior probabilities of each hypothesis are also available, the prior odds ratio can be combined with the BF to estimate the posterior odds ratio.

The last reason BFs can be useful is the one most often invoked based on theoretical grounds—they do not depend on the prior probabilities of the two hypotheses, as opposed to posterior odds ratios. If the hypotheses concern the monophyly of a set of taxa (e.g., H_1 requires monophyly and H_2 requires non-monophyly), the standard discrete, uniform prior on topology will tend to favor H_2, sometimes strongly. These unbalanced prior odds are often an unintended consequence of the topology prior, and one might wish to know how strongly the data support each hypothesis irrespective of this prior effect. However, Bergsten et al. (2013) also note challenges to interpreting topological BFs when the data strongly support clades compatible with a focal clade, even if the focal clade itself does not have much direct support. They recommend constraining non-focal relationships that are well supported (i.e., giving them a prior probability of 1) to avoid these problems. Such constraints have the effect of reducing imbalanced prior odds on monophyly. Here, we follow the advice of Bergsten et al. (2013) both to avoid these issues and because the effect of prior odds is not our primary interest.

Despite the consensus that large phylogenetic data sets are desirable, there has been little agreement about the best way to assemble such data sets. All phylogenomic data collection strategies target certain regions of the genome, and use long chains of bioinformatic decisions involved in data cleanup and processing before producing alignments that are suitable for analysis. The logic that more data yield more confidence is intuitively appealing and has led the field to pay comparatively little attention to these long decision chains and how they affect the information content and quality of resulting data sets. Given the growing diversity of data collection strategies available to phylogenetic researchers, understanding the performance and tradeoffs among alternative approaches has become increasingly important. While methodological simplicity, amount of data produced, and costs of alternative approaches are all reasonably well understood, there are few quantitative comparisons of the phylogenetic information content of data sets that result from these alternative approaches.

The phylogenetic position of turtles within amniotes has been a subject of continuing discussion. Alternative analyses support several different phylogenetic hypotheses (Gaffney 1980; Rieppel and DeBraga 1996; DeBraga and Rieppel 1997; Rieppel and Reisz 1999; Lyson et al. 2010, 2012; Chiari et al. 2012; Crawford et al. 2012; Fong et al. 2012; Shaffer et al. 2013; Wang et al. 2013; Lu et al. 2013; Thomson et al. 2014; Schoch and Sues 2015), all with strong support. The question of turtle placement presents an ideal opportunity to examine issues relating to the characteristics of phylogenetic data sets generated under alternative approaches. The phylogenetic position of turtles has been examined with at least six distinct phylogenomic data sets in recent years. Two of these are based on *de novo* sequencing and assembly of full turtle genomes combined with existing genomes of other amniotes (Shaffer et al. 2013; Wang et al. 2013), a third combines data from these earlier genome sequencing efforts (Lu et al. 2013), a fourth makes use of *de novo* sequences generated from an EST library (Chiari et al. 2012), a fifth contains sequences from Ultra Conserved Elements (UCEs; Crawford et al. 2012), and a sixth makes use of Sanger sequencing of polymerase chain reaction (PCR) amplicons deployed at a large scale (Fong et al. 2012). These data sets differ widely both in how sequences were collected and in the number (and length) of sequenced genes. One data set (Fong et al. 2012) also differs sharply from the others in the number of taxa sampled.

The use of these six phylogenomic data sets is well suited for comparing the quantity and the quality of phylogenetic information produced by different approaches. While the phylogenetic position of turtles remains uncertain, the relationships among the remaining major lineages of amniotes are no longer debated: all six data sets recover the same relationships among non-testudine amniote lineages, as do most previous phylogenetic analyses that examine this clade (Fig. 1). The availability of multiple, large data sets with similar patterns of taxonomic sampling, coupled with the high degree of certainty through most of the amniote tree, presents the rare opportunity to quantitatively compare among a wide variety of alternative data collection strategies in the context of a phylogeny that is essentially known.

In this study, we use BFs to characterize the extent and quality of phylogenetic information across genes comprising these six data sets. Our results reveal that the majority of variation in phylogenetic information across genes resides in outliers with extreme BFs, which are entirely hidden by MCMC estimates of posterior probabilities. Genes with extreme support can be exceptionally important, because they wield an outsized influence on inferences when data are analyzed jointly. For one data set, the influence of fewer than 1% of genes (2 of 248) led to strong support for turtles as sister to crocodilians. Upon further investigation, we

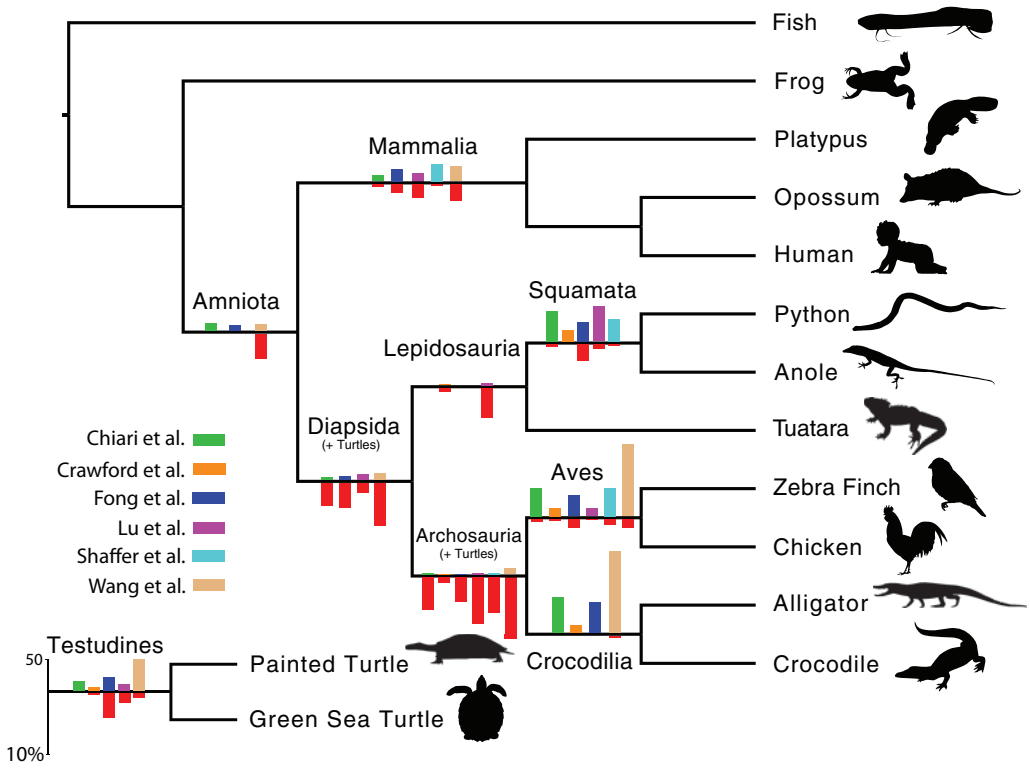

FIGURE 1. Overview of amniote phylogeny, with major groups labeled on internal branches. Representative species are samples of those sequenced in the six data sets used in this study. Colored, upward bars on each branch give the median 2ln(BF) value across genes supporting that relationship for each data set (see values in Table 1). Red, downward bars show the percentage of genes in each data set that strongly reject (2ln(BF) < −10) each clade (see values in Table 2). For Archosauria and Diapsida, we provide values for the monophyly of these groups along with turtles, since most studies suggest that turtles are a member of these clades. Silhouette images were obtained from PhyloPic (http://phylopic.org).

found these genes to contain previously unrecognized paralogs and their removal led to strong support for turtles as sister to archosaurs. BFs also demonstrate marked differences in overall information content and quality across data sets, with important implications for how each data set might be best analyzed. Some data sets contain a sizable proportion of genes that strongly reject known relationships. Despite these differences, all data sets consistently exhibited very little information to resolve the phylogenetic placement of turtles. The relative lack of information available to place turtles offers an explanation for the sensitivity of this placement to data and model selection.

METHODS

Data

Six phylogenomic data sets were taken from recent studies (Chiari et al. 2012; Crawford et al. 2012; Fong et al. 2012; Lu et al. 2013; Shaffer et al. 2013; Wang et al. 2013) that focused on resolving the phylogenetic placement of turtles among amniotes. The size of these data sets ranged from 75 to 1955 genes. All data sets were taken "as-is" from previous authors in order to maintain the influences of researcher decision-making on the extent and quality of phylogenetic information. Processing of these original data files involved only conversion of the file types to a common NEXUS format and, if necessary, standardizing the taxonomic names. These NEXUS files are available on Dryad at http://dx.doi.org/10.5061/dryad.8gm85.

Gene Characterization

The six studies used very different methods. Techniques ranged from Sanger sequencing of PCR amplicons (Fong et al. 2012) to collection of whole-genome assemblies using a variety of sequencing technologies (Shaffer et al. 2013; Wang et al. 2013). After data collection, each study also developed and used an independent data processing pipeline to perform sequence quality filtering on the raw data, assign homology, and carry out alignments. Due to the diversity of approaches employed, we expected differences in the amount and nature of data included in each of the data sets. We characterized these differences using five summary statistics described below.

(i) We calculated the percentage of missing data for each gene as the total number of missing and ambiguous bases ("N"s, "-"s, and "?"s) divided by the total number of bases included in the alignment (sequence length multiplied the number of taxa). (ii) We summarized alignment quality for each gene by employing the widely used Heads-or-Tails

(HoT) metric (implemented in COS v2.03 perl script at http://nsmn1.uh.edu/dgraur/scripts/HoT/), which measures the difference between a given alignment and an alignment generated from a reversed set of the same sequences (Landan and Graur 2007). Alternative measures of alignment consistency that rely on assessing consistency of alignments generated from alternative guide trees (e.g., GUIDANCE; Penn et al. 2010a, 2010b) appear to be sensitive to the number of taxa in an alignment, because sets of bootstrap trees are expected to vary more for alignments that have many taxa than for those that have few. Because the number of taxa in each alignment varied widely and frequently was small (<8), we chose to focus on the HoT method.

We selected a best-fit substitution model for each gene using MrModelTest v2.3 (Nylander 2004) and PAUP* v4.0b10 (Swofford 2003) under the Akaike's Information Criterion (AIC) criterion (Akaike 1974). These models were used both to characterize genes and in downstream phylogenetic analyses (detailed below). (iii) We measured each gene's "clockness" by estimating a maximum-likelihood (ML) gene tree using the AIC-selected substitution model in GARLI v2.0 (Zwickl 2006). We performed five replicated searches, terminating each after 5000 generations without an improvement in the log-likelihood score of at least 0.01 units. We then calculated the likelihood score of this tree in PAUP* v4.0b10 (Swofford 2003) both without a clock and under a clock constraint, rooting the tree at the midpoint in both cases and using two times the difference in log-likelihoods as the summary statistic for departure from the strict clock. (iv) We also characterized the rate of evolution for each gene as the sum of branch lengths in its ML gene tree. (v) Finally, we measured base frequency heterogeneity across taxa using a χ^2 test as implemented in PAUP* v4.0b10 (Swofford 2003) and recorded the χ^2 statistic as a summary statistic.

Phylogenetic Inference and BF Calculation

We performed Bayesian phylogenetic inference for each gene with MrBayes v3.2 (Ronquist et al. 2012), assuming the AIC-selected substitution models. We estimated the joint posterior probability distribution of tree topologies and other model parameters using two independent MCMC analyses that each employed four Metropolis-coupled chains per run. We allowed each chain to run for 10 million generations, sampled the parameter states every 10,000th generation, and discarded the first 2000 of these samples as burn-in. We quantified posterior support for each of the nine uncontroversial clades in the amniote backbone tree, which includes the monophyly of Aves, Crocodilia, Rhyncocephalia, Squamata, Mammalia, Archosauria, Lepidosauria, Diapsida, and Amniota, exclusive of Testudines (Fig. 1). Bipartition posterior probabilities were calculated only for those genes with a combination of taxa that could support or reject that bipartition. For instance, the posterior probability for diapsids was calculated only for genes that contained at least one archosaur, one lepidosaur, and two outgroups to the diapsids. All genes contained sequences for at least four taxa. All bipartition posterior probability calculations were carried out by filtering the full posterior distribution of tree topologies according to the relevant constraints using Dendropy v3.12.0 (Sukumaran and Holder 2010) with Python v2.7.6. We estimated the phylogeny for each of the six concatenated data sets using a similar strategy as above. For each data set, we used a general time reversible substitution model with gamma-distributed among site rate variation and a proportion of invariable sites, carrying out two independent MCMC analyses under identical settings as above.

Unlike all of the six original studies whose data we use here, we also quantified phylogenetic information using BFs (Kass and Raftery 1995). In our case, the alternative hypotheses are represented by the presence or absence of individual bipartitions (e.g., $H_1 =$ bipartition A is present and $H_2 =$ bipartition A is absent). To calculate the 2ln(BF) favoring the monophyly of each uncontroversial amniote clade, we estimated two marginal likelihoods for every gene across all six data sets. The first likelihood marginalized across topologies consistent with all nine uncontroversial, backbone amniote clades (or whatever subset of those nine could be examined with the taxa available for each gene). Hereafter, we will refer to constraints that enforce the presence of a bipartition as "positive constraints" and those that enforce the absence of a bipartition as "negative constraints". Therefore, the first likelihood employs positive constraints on all uncontroversial relationships and will be called the positive marginal likelihood for a bipartition. The second likelihood involved positive constraints on all non-focal, uncontroversial relationships, but a negative constraint for the focal bipartition. We refer to this value as the negative marginal likelihood for a bipartition. So, estimation of the negative marginal likelihood for the bipartition uniting birds would require the monophyly of all other uncontroversial groups (crocodilians, archosaurs, rhyncocephalians, squamates, mammals, lepidosaurs, diapsids, testudines, and amniotes), but would not allow the monophyly of birds. While turtles were constrained to be monophyletic, their position in the amniote tree was not constrained. We enforced positive constraints on non-focal clades to provide a more meaningful measure of support (Bergsten et al. 2013).

A different strategy was necessary to calculate 2ln(BF) values favoring particular placements for turtles, since there is no way to set up a single set of constraints in MrBayes v3.2.2 (Ronquist et al. (2012), or any other Bayesian phylogenetic software of which we are aware) that would allow an MCMC analysis to sample only a set of pre-specified topologies. As above, we calculated two marginal likelihoods for each turtle placement hypothesis. The first marginal likelihood constrained turtle placement to a single position in the tree (e.g., sister to archosaurs). The second marginal likelihood considered all other hypothesized positions for turtles

(e.g., sister to birds, crocodilians, diapsids, lepidosaurs, mammals, or all non-turtle amniotes). In practice, the ln(marginal likelihood) for each sister placement was initially estimated individually in MrBayes (see full details below). To avoid floating-point underflow, we calculated the ln(marginal likelihood) for a composite hypothesis that included several possible placements for turtles as

$$a + \ln\left(\sum_{i=1}^{n} e^{\ln(L_i)-a}\right) - \ln(n)$$

where n is the number of placements being considered, L_i is the marginal likelihood of the i-th placement, and a is the logarithm of the maximum L_i value. This expression gives the log of the average marginal likelihood across all placements, which is equivalent to the log marginal likelihood of the composite hypothesis because each of the relevant topologies has equal prior probability.

All marginal likelihoods were initially estimated in MrBayes v3.2.2 (Ronquist et al. 2012) using steppingstone sampling (Xie et al. 2011) with the appropriate topological constraints. Each steppingstone run employed two independent analyses, with four Metropolis-coupled chains apiece. These analyses were run for 1,000,000 generations with 50 steps used to move from the posterior to the prior. Steps in the power posteriors followed a Beta(0.4,1.0) distribution. An initial burn-in step was used before power posterior sampling began, and each step in the steppingstone employed a 25% burn-in. Analyses were spot-checked to ensure consistency in the marginal likelihood estimates across replicates. The mean of the two independent estimates was used to calculate 2ln(BF) values.

When checking the consistency of marginal likelihood estimates from MrBayes, we noticed that independent steppingstone runs related to the monophyly of uncontroversial relationships occasionally returned strongly divergent values, particularly when they involved negative constraints. Looking into this further, we discovered a bug in MrBayes that incorrectly shuts off all topology proposals for some combinations of constraints, even when the topology is not fully specified. Different runs were assigned different starting topologies stochastically and were unable to sample other topologies. To fix this problem, we modified the latest MrBayes source code (v3.2.5) so that topology moves are never shut off. During our consistency checks, we also found that for some runs employing negative constraints, mostly involving the analysis of large concatenated data sets, runs had trouble mixing properly. To ensure more accurate marginal likelihood estimates, we re-ran all analyses involving positive or negative constraints on the monophyly of uncontroversial clades using our modified version of MrBayes. To improve mixing and more rigorously check convergence, we increased the number of independent runs to 4, increased the number of chains per run to 16, and raised the temperature for Metropolis coupling to 0.2. With these two changes, steppingstone analyses exhibited much greater consistency in marginal likelihood estimates across runs, especially for single genes. In some of the concatenated analyses, we were never able to find settings for the Metropolis-coupled Markov chain that entirely eliminated mixing problems. However, the variation in estimated marginal likelihoods across runs was generally several orders of magnitude smaller than the estimated BFs. Nonetheless, we suggest that readers interpret 2ln(BF) values for concatenated data sets, particularly those of Lu et al. (2013) and Wang et al. (2013), with some caution. Since the steppingstone runs related to turtle placement did not seem to be affected by either the bug or the mixing problems, we did not re-run those analyses.

For BFs relating to the monophyly of uncontroversial relationships, the positive marginal likelihood was in the numerator and the negative marginal likelihood in the denominator. For BFs relating to turtle placement, the focal placement marginal likelihood was in the numerator and the log of the average marginal likelihood across other placements was in the denominator. Values of 0 for 2ln(BF) indicate complete ambiguity when comparing hypotheses, while values of greater than 10 are generally considered to indicate very strong evidence in support of the hypothesis in the numerator and values less than −10 are considered to indicate very strong evidence against that hypothesis (Kass and Raftery 1995).

Correlations Between Gene Characteristics and Phylogenetic Information

We examined correlations between the support that each gene provided for each uncontroversial clade (2ln(BF) values) and various characteristics of the genes, detailed above. We calculated Spearman's rank correlation coefficient (r_s) using functions from the "stats" library in R (R Core Team 2016). To determine which correlations were significant, we employed a resampling procedure. Both sets of values were randomly permuted 1000 times and r_s was calculated for each permutation. These values formed a null distribution, which we used to compute a two-tailed p-value.

To investigate whether certain genes were universally superior for resolving relationships in the amniote tree or whether individual genes provided information about different relationships, we also calculated r_s between the 2ln(BF) values for each pair of backbone bipartitions across genes. If the same genes tend to strongly support all backbone bipartitions [2ln(BF) ≫ 10] or be ambivalent about backbone bipartitions [−10 < 2ln(BF) < 10], we expect to see strong correlations. However, if genes vary in the support they provide for different backbone bipartitions, we expect low correlations.

Testing Paralogy in the Transcriptome Data of Chiari et al. 2012

While exploring the distributions of 2ln(BF) values across genes, we noticed that two genes in the data set

of Chiari et al. (2012) seemed to be strong outliers in the strength of their support for a sister relationship between crocodilians and turtles (alignments ENSGALG00000008916 and ENSGALG00000011434). To better understand what was driving this pattern, we explored the phylogenetic signal coming from these two genes in more detail. First, we calculated a site-likelihood profile for each gene using PAUP*4b10 (Swofford 2003) by specifying two trees that constrained turtles as sister to either crocodilians or archosaurs. Using AIC-chosen models of sequence evolution, we estimated ML model parameter values and branch lengths on each topology independently for each gene and then recorded log site likelihoods after optimization. To quantify each site's preference for the placement of turtles, we calculated differences in the log site likelihoods (equivalent to the log of the site likelihood ratio) between the two constraint trees. For comparative purposes, we also calculated site-likelihood profiles for the gene with the next highest 2ln(BF) value supporting crocodilian sister placement (ENSGALG00000001362) and the gene with the lowest 2ln(BF) value for crocodilian sister placement (ENSGALG00000005758). Comparisons of site-likelihood profiles (Supplementary Fig. S2 available on Dryad) across these four genes suggested to us that the orthology of the sequences in the outlier genes may be questionable.

To investigate the potential for paralogy in the outlier genes, we used each sequence in these alignments to search the most closely related reference genome in GenBank for similar regions using BLASTn 2.2.32+ (Altschul et al. 1997). For 7 out of 12 sequences, a reference genome for the same species was available. *Caiman* was searched against the *Alligator mississippiensis* reference, *Caretta* was searched against the *Chelonia mydas* reference, while the remaining turtles (*Chelonoidis*, *Emys*, and *Phrynops*) were searched against the *Chrysemys picta* reference. In each BLASTn search, we found 2–3 regions from each reference genome with strong similarity (>70%) that largely overlapped the query sequence. For all searches, the hits returned by BLASTn were not contiguous stretches in the reference genomes, suggesting that each alignment is made up of multiple exons. For each of the top hits with both high similarity and large overlap, the reference genome fragments returned by BLASTn were assembled into a single contig with GENEious (Kearse et al. 2012) against the original query sequence. We then performed Bayesian phylogenetic inference (as detailed above for the original alignments) using the expanded alignments.

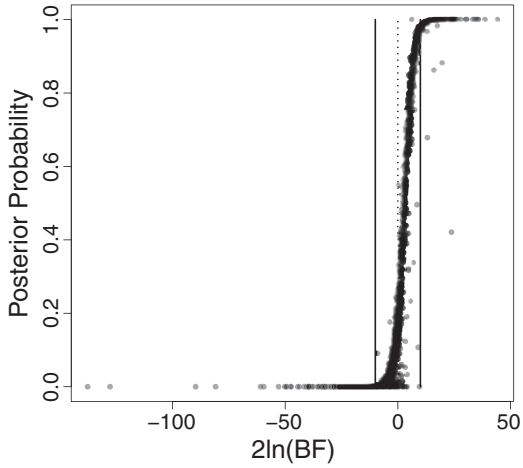

FIGURE 2. A comparison of MCMC posterior probability estimates to 2ln(BF) values for archosaur monophyly across all genes in the Shaffer data set. Solid vertical lines indicate the typical cutoffs of 10 or −10 for 2ln(BF) values interpreted as very strong support or rejection, respectively. The dashed vertical line indicates a 2ln(BF) value of 0.

RESULTS AND DISCUSSION

BFs Uncover Hidden Variation

BFs are able to differentiate between genes and data sets with varying amounts of information, even when they result in indistinguishable posterior probability estimates. For instance, Figure 2 shows the relationship between MCMC posterior probability estimates of archosaur monophyly and the corresponding 2ln(BF) values for 1943 genes from the Shaffer et al. (2013) data set. While posterior probabilities and 2ln(BF) values are tightly correlated for genes with intermediate support ($-10 < 2\ln(BF) < 10$), nearly a third of these genes (29%) fall in the tails stretching to the left ($2\ln(BF) < -10$) and right ($2\ln(BF) > 10$). For many other clades of interest, the proportion of genes in the tails is much higher. In these tails, MCMC posterior probability estimates are close to either 0 or 1, and genes appear to reject or support this relationship with equal strength. However, BFs reveal that variation in the strength of rejection or support is many-fold greater *within* tails than in the intermediate range, suggesting that coarse estimates of posterior probabilities provided by MCMC may obscure more variation among genes than it clarifies. Based on MCMC estimates of posterior probabilities alone, the genes that most strongly reject archosaur monophyly (in this case with $2\ln(BF) < -100$) would be impossible to identify. However, these outliers can have an outsized influence on inferences from joint analysis of all genes. For these amniote data sets, genes strongly rejecting the monophyly of well-established groups should certainly be considered suspect. We demonstrate below how investigation of genes with extreme support for turtle placement revealed unappreciated paralogy in a small proportion of alignments (<1%) that had an extraordinary influence on the inferred placement of turtles.

Sequencing Strategies Vary in Per-Gene Information Content

Gene-specific BFs for the six data sets analyzed here reveal striking patterns in the distribution of support across genes, data sets, and clades (Figs. 3 and 4). First, despite being supported with indistinguishable

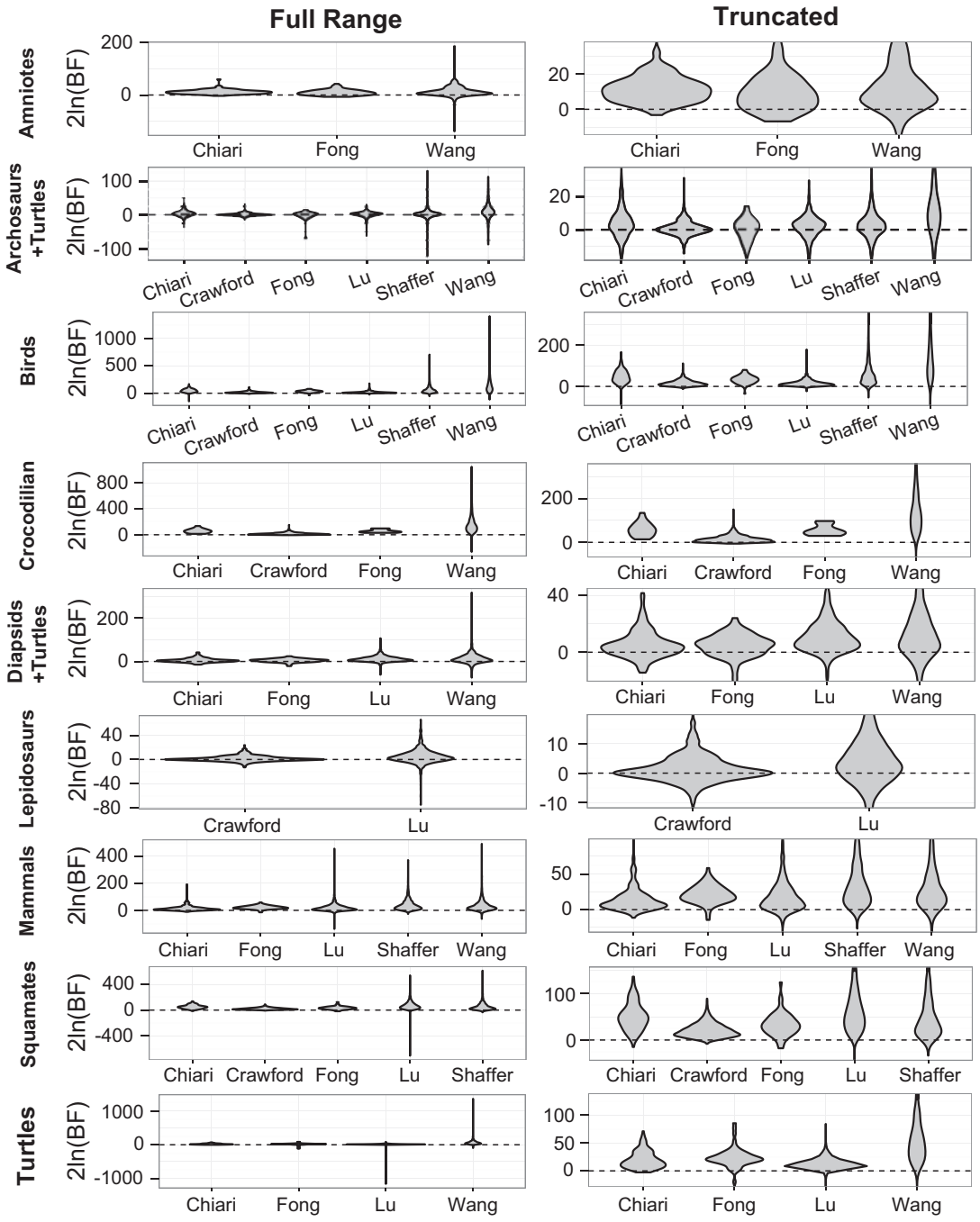

FIGURE 3. Summary of BF support for major clades in the amniote phylogeny. Each row includes violin plots of 2ln(BF) values for a different clade, showing the distribution of gene-specific values for all data sets with taxon sampling appropriate for a test of monophyly. Plots in the left column show the full range of 2ln(BF) values. Plots in the right column use a truncated range to facilitate comparison of central tendencies and minimize the influence of extreme values.

posterior probability estimates (e.g., 1.0), major clades varied extensively in the strength of support provided by individual genes. The monophyly of some groups, such as birds, crocodilians, squamates, and turtles, was very strongly supported ($2\ln(BF) \gg 10$) across nearly all genes. Others, such as amniotes, archosaurs+turtles, lepidosaurs, and diapsids+turtles, had much weaker and more conflicting support. Weaker support for some clades may be the result of short internal branches uniting that group, potentially caused by rapid diversification. Variation in support could also be driven by choices in taxon sampling. For instance, if two representatives of a group do not span its earliest divergence, then inferred support values actually pertain to a much more closely related clade with a longer subtending branch.

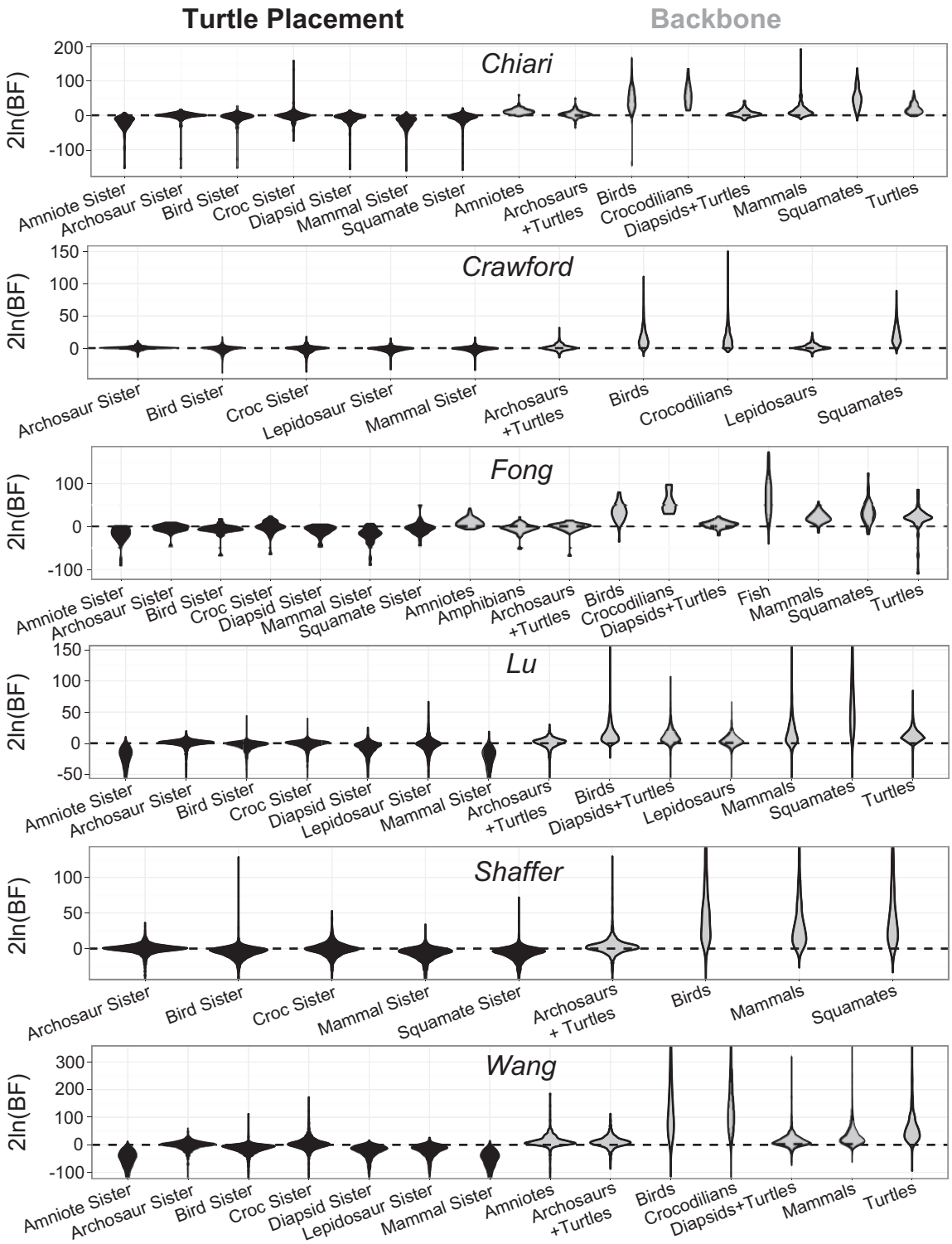

FIGURE 4. Summary of BF support for both the placement of turtles and the monophyly of major amniote clades, grouped by data set. Each row shows violin plots of gene-specific 2ln(BF) values. Support for different hypothesized placements of turtles is shown on the left in black. Support for the monophyly of major amniote clades is shown on the right in gray. Note that the *y*-axis is truncated for the data sets of Lu et al. (2013), Shaffer et al. (2013), and Wang et al. (2013) to highlight differences in central tendencies of the distributions and minimize the influence of outliers. The full scale is shown in Supplementary Figure S1 available on Dryad.

TABLE 1. BFs [2ln(BF)] in favor of major amniote clades for concatenated data sets or median values across genes

Data set	Amniota	Archosauria + Testudines	Aves	Crocodilia	Diapsida + Testudines	Lepidosauria	Mammalia	Squamata	Testudines
Chiari (concat)	814.8	853.3	13,040.2	2263.8	1124.0	—	3249	5492.1	3791.4
Chiari (median)	11.0	2.7	44.9	56.3	4.0	—	9.2	48.3	15.8
Crawford (concat)	—	596.6	16,738.2	18,068.0	—	952.1	—	19,706.2	6851.5
Crawford (median)	—	0.3	12.8	11.8	—	0.6	—	18.0	6.1
Fong (concat)	365.6	127.4	−11.1	881.7	364.8	—	752.1	51.9	17.2
Fong (median)	8.3	0.9	33.0	47.8	5.3	—	20.1	30.5	20.9
Lu (concat)	—	2994.9	20,965.9	—	15,864.2	6324.1	31,264.2	116,411.0	7998.4
Lu (median)	—	2.1	12.1	—	9.1	3.4	13.4	55.6	10.6
Shaffer (concat)	—	5654.7	92,465.7	—	—	—	48,200.0	36,071.7	—
Shaffer (median)	—	2.2	44.9	—	—	—	28.02	36.36	—
Wang (concat)	9580.4	11,594.9	171,707.3	219,512.1	14,619.3	—	39,114.8	—	90,671.7
Wang (median)	9.6	9.4	115.3	129.2	9.9	—	23.2	—	50.6

Second, data sets generated through different sequencing strategies exhibit consistent differences in their phylogenetic information as measured with BFs. In general, the UCE data set of Crawford et al. (2012) carried the least information on a per-gene basis (Table 1, Figs. 3 and 4). However, the large number of UCE loci produces a concatenated data set with a total information content on par with many others (Table 1). At the other end of the spectrum, those data sets based on orthologs selected from fully sequenced genomes (Lu et al. 2013; Shaffer et al. 2013; Wang et al. 2013) tended to have higher per-gene information content (Table 1, Figs. 3 and 4). While this difference was sometimes apparent in central tendencies across genes (e.g., higher median BFs), the most striking distinction was the presence of outlier genes with BFs of large magnitude. Unlike the UCE data set, which was fairly homogeneous in BFs across genes, the data sets derived from whole genomes frequently contained a small number of genes with BFs that were orders of magnitude more extreme than the median. Interestingly, while outliers often supported uncontroversial relationships, some also strongly *rejected* these relationships with surprising strength (Table 2).

Third, we found some interactions between clade and data set, particularly related to the number and extremity of outlier genes. For instance, the whole-genome data set of Wang et al. (2013) contained some outliers showing extremely strong support for nearly all uncontroversial clades. However, this data set also includes outliers that strongly reject monophyly of these same groups, notably amniotes and archosaurs. The data set of Shaffer et al. (2013) was similar in this regard, although taxon sampling was less extensive than Wang et al. (2013) so we could not calculate BFs for many clades of interest. Nonetheless, the Shaffer et al. (2013) data set contains genes with outlying positive support for several uncontroversial clades, but negative outliers primarily reject the monophyly of archosaurs. Differences between the Shaffer and Wang data sets are particularly noteworthy, since these studies started with a similar set of reference genomes from which they selected and aligned putative orthologs (although they included different de novo turtle genomes). The Lu et al. (2013) data set is notable in the number of strong negative outliers it contains, which reject the monophyly of lepidosaurs, squamates, turtles, and, to a lesser extent, mammals. Turtle monophyly is particularly interesting, because the Lu et al. (2013) data set contains extreme negative outliers with no positive outliers, while the Wang et al. (2013) data set contains extreme positive outliers with no negative outliers. The reasons why some genes exhibit such extreme support or rejection for uncontroversial clades are not immediately obvious, but could be driven by paralogy, poor alignment, or strong convergent evolution, all of which could be compounded or caused by high rates of evolution.

Two points regarding the distribution of information content across genes and data sets are important to keep in mind. First, the vast majority of this variation would go completely unrecognized if genes were only compared based on MCMC estimates of posterior probabilities. All genes with a $2\ln(BF) > 10$ for a particular bipartition would appear to "resolve" that part of the tree. However, such a coarse categorization results in enormous information loss that obscures fundamental differences among phylogenomic studies. Second, these distributions reveal that different phylogenomic data sets may behave very differently when conducting multi-gene phylogenetic inference, and the method of choice for analyzing individual phylogenomic data sets should perhaps depend as much on the distribution of information across genes as it does on the biological question that is being studied. When performing concatenated inference, genes with BFs of large magnitude can have an immense influence on the overall phylogenetic estimate. For instance, in our analyses (Figs. 3 and 4), outlier genes with 2ln(BF) values of >200 were discovered in several phylogenomic data sets. A gene with a 2ln(BF) of 200 corresponds to a BF of approximately 2.7×10^{43}. Such an overwhelming value implies a level of certainty that is difficult to put into words. If this gene proved unreliable for some reason (e.g., questionable homology, model misspecification, etc.), it would take 20 genes with $2\ln(BF) = -10$ to

TABLE 2. Tallies of genes rejecting well-established backbone relationships across data sets, with the final row giving the number of genes that reject at least one of the relationships listed in the other rows

	Chiari et al. 2012 248 genes 16 taxa	Crawford et al. 2012 1145 genes 10 taxa	Fong et al. 2012 75 genes 110 taxa	Lu et al. 2013 1638 genes 11 taxa	Shaffer et al. 2013 1955 genes 8 taxa	Wang et al. 2013 1113 genes 12 taxa
Amniota	0	—	0	—	—	45
Archosauria (+ Testudines)	13	10	3	123	112	111
Aves	1	2	1	2	16	16
Crocodilia	0	0	0	—	—	7
Diapsida (+ Testudines)	9	—	3	25	—	77
Lepidosauria	—	6	—	78	—	—
Mammalia	1	—	1	36	5	29
Squamata	1	0	2	11	4	—
Testudines	0	3	3	24	—	8
Total	24 (9.7%)	21 (1.8%)	10 (13.3%)	262 (16.0%)	132 (6.8%)	246 (22.1%)

counterbalance the outlier's effect (and we investigate one such case below). In essence, if an outlier gene is unreliable, the data set must contain many more reliable genes of typical information content to counterbalance its effects, which provides one explanation for why phylogenomic conclusions remain sensitive despite the large amount of data. This thought experiment suggests that the identification and careful scrutiny of outliers based on BFs might be a valuable tool for ensuring the accuracy of concatenated analyses. While we have not explicitly investigated the effect of outliers on hierarchical analyses that employ a probability distribution on gene tree topologies (e.g., multispecies coalescent analyses), gene tree estimates with artificially inflated certainty may also be problematic in that context. These results argue for careful consideration of the absolute fit between assumed models of sequence evolution and the sequence data in each gene.

Focusing on genes that strongly reject well-established backbone relationships (Table 2), we have estimated a minimum bound on the level of spurious phylogenetic signal that must be present in these data sets. The number of genes supporting clearly incorrect relationships conservatively ranged from 1.8% to 22.1%. These estimates are conservative for several reasons. First, we focused on a small set of well-supported relationships that could be compared across studies. If taxon sampling was more extensive and amniote relationships better understood, these values would almost certainly rise. Second, we only tallied those genes that strongly reject ($2\ln(BF) < -10$) backbone relationships. Inference or data-quality problems have likely affected other genes to more moderate degrees. Third, we can focus only on the effects of these problems in depressing support for known relationships. Systematic error in models of sequence evolution, in particular, may also artificially inflate support. Lastly, we have ignored any pre-filtering that took place before these data were published. Given that amniote relationships are so well known, genes that produced completely incoherent phylogenetic estimates might have been recognized and removed. This type of pre-filtering is less likely to occur for poorly understood groups. Despite these *caveats*, the percentage of genes with strongly spurious signal should still be concerning.

Comparatively Little Information to Place Turtles

All six phylogenomic data sets contained substantially less information about turtle placement than they did for most backbone relationships (Fig. 4). Median BFs for even the most strongly supported turtle placements (sister to archosaurs or crocodilians) were very small. Of the backbone relationships with BF distributions similar to turtle placements, they tended to correspond to old divergences (e.g., the monophyly of amniotes and amphibians). This difference in support, despite being consistent and strong, was not apparent from most analyses presented in the six original papers. Only the study of Fong et al. (2012) suggested that their data did not convincingly resolve the placement of turtles with the same degree of certainty as the other major clades in the tree.

Despite greater ambivalence about turtle placement, two possibilities were uniformly rejected by nearly all genes in all data sets: turtles sister to all other amniotes or sister to mammals. The only data set for which the distribution of gene-specific BFs is not noticeably lower for the mammalian sister placement is the UCE data from Crawford et al. (2015; taxon sampling in this data set did not allow testing of the placement sister to all other amniotes). The placement of turtles sister to diapsids or squamates (or lepidosaurs, depending on the taxon sampling) tended to have a distribution of 2ln(BF) values that was also shifted well below 0 for most data sets, although these positions were not as strongly rejected as the amniote and mammalian sister placements (Fig. 4). Interestingly, a small number of outlier genes did place turtles sister to squamates/lepidosaurs with strong support across three data sets (Fong et al. 2012; Lu et al. 2013; Shaffer et al. 2013). Future work will explore whether a predictable cause can explain this rare, yet strong, support for a turtle-lepidosaur sister relationship.

The distribution of BFs for bird, crocodilian, and archosaur sister placements were all similar, with marginally less enthusiasm for the bird sister placement in most data sets (Fig. 4). Each of these three placements was both supported and rejected by many genes in all data sets. When we analyzed each concatenated data set, we recovered strong posterior support for an archosaur sister placement in three cases (Crawford et al. 2012; Lu et al. 2013; Shaffer et al. 2013) and a crocodilian sister placement in the other three (Chiari et al. 2012; Fong et al. 2012; Wang et al. 2013; Supplementary Figs. S5, S7–S11 available on Dryad). Interestingly, the analyses preferred by the authors in all of the six original studies favored an archosaur sister placement for turtles, usually strongly, although with several important *caveats*. The preferred analyses of Fong et al. (2012) could not reject a crocodilian sister placement for turtles, and concatenated analyses with different gene and taxon sampling led to a variety of different placements. Wang et al. (2013) removed all third codon positions in their concatenated analyses (but did not discuss the justification for this decision). When included, turtles are recovered as sister to crocodilians. Chiari et al. (2012) originally inferred a strongly supported crocodilian sister placement for turtles using standard models of sequence evolution, but argued that these analyses were unduly influenced by "substitution saturation" at third codon positions. When third positions were removed or more sophisticated models of sequence evolution were employed, turtles were inferred to be sister to archosaurs with strong support. The distribution of BFs across genes in all six data sets highlights two explanations for this sensitivity. There is relatively little information to place turtles, so small differences in analytical choices can have a large effect on the outcome. In addition, outlier genes may play an oversized role in turtle placement, given the average ambivalence across genes in each data set as a whole.

Two aspects of the BF distributions for turtle placement are particularly interesting. First, the uniformly greater ambivalence about turtle placement relative to most backbone relationships across all data sets suggests that the lineage divergence event leading to turtles took place in close temporal proximity to another divergence event, likely that which separated birds from crocodilians. If so, less time would have elapsed and fewer changes would have accrued in the ancestor that unites the two sister lineages among turtles, birds, and crocodilians (likely birds and crocodilians), meaning that each gene tree contains less information about these relationships. Further, closely spaced speciation events would have increased the probability of stochastic variation in gene tree topologies due to incomplete lineage sorting. Second, genes seem to have a consistent preference for placing turtles sister to either archosaurs or crocodilians, but less so sister to birds. If turtles really are sister to archosaurs, as previous studies suggest, and there is coalescent variation in gene tree topologies, we should expect to see roughly as many genes placing turtles sister to birds as sister to crocodilians. Why, then, do we seem to see a consistent preference for the crocodilian sister placement? One possibility is that some aspect of molecular evolution not captured by typical phylogenetic models is convergent between turtles and crocodilians, leading to erroneous support for their monophyly. Another possibility is that the processes of selecting genes or aligning sequences within genes are biased toward finding similarities between crocodilians and turtles to the exclusion of birds.

The Outsized Influence of Paralogous Genes on Turtle Placement

While examining the distribution of BF values for turtle placement (Fig. 4), we were struck by two outlying genes in the Chiari et al. (2012) data set that strongly supported the placement of turtles as sister to crocodilians. These outlying values were notable both because of their large magnitude ($2\ln(BF) > 130$) and because the full data set of Chiari et al. (2012) prefers a placement of turtles sister to crocodilians under standard models of sequence evolution. To better understand the cause of this pattern and the effects it might have on inferences derived from the full data set, we performed three analyses. First, we calculated site-specific likelihood profiles to examine whether the strong opinions of these genes were driven by a small number of sites or were widespread across sites within each gene. We found that strong support for the crocodilian sister placement was widespread across sites, but that both genes had an unusual pattern of site-specific likelihood ratios (Supplementary Fig. S2 available on Dryad). Strong support was not confined to a particular region of the gene, to a particular codon position, nor was it specific to the crocodilian sister placement. A sizeable proportion of sites supported an archosaur sister placement. Based on these results, we suspected that the orthology of sequences in these alignments was questionable.

To examine this possibility, we next searched closely related reference genomes to identify all regions similar to the sequences contained in the original alignments. For both genes, we found multiple regions in each reference genome that had high similarity and large overlap with the query sequences from the original alignments. Adding all relevant sequences to these alignments, we conducted Bayesian phylogenetic inference and found that the original sequences do not form monophyletic groups (Supplementary Figs. S3 available on Dryad; Fig. 4). Instead, they seem to be drawn from multiple clades, each of which contains no more than one sequence from each reference genome. The precise history of duplications and losses for these genes is difficult to determine, but based on the inferred topologies the most plausible scenario in both cases is that the sequences originally selected for turtles and crocodilians are paralogous to those selected for the remaining taxa. If true, the strong support they seem

to confer on a crocodilian sister placement for turtles is spurious.

To understand how much influence these genes may have had on concatenated analyses, we inferred trees for the concatenated data set both including and excluding these paralogous genes. When these two genes are included, a crocodilian sister placement for turtles is supported with an estimated posterior probability of 1.0 (Supplementary Fig. S5 available on Dryad). When these genes are removed the concatenated data set still contains 246 genes, but turtles are now estimated to be sister to archosaurs with a posterior probability of 1.0 (Supplementary Fig. S6 available on Dryad). These results highlight the potential for a small number of genes with extreme preferences for particular relationships to exert undue influence on concatenated analyses of large numbers of genes.

Phylogenetic Information Content is Difficult to Predict

By calculating BFs for each combination of bipartition and gene across all data sets, we were able to ask if some genes are universally superior to others. In other words, do some genes have sufficient information to strongly resolve all branches, while others have little information to resolve any? If this were the case and we rank genes by how much support they provide to different bipartitions, we should expect to see strong correlations in these ranks. This pattern was not generally observed (Supplementary Table S1 available on Dryad). Rank correlations tended to be weak, suggesting that different genes carry information about different branches. The data set of Shaffer et al. (2013) seems closest to having a set of universally informative genes, as all rank correlations are positive and some are of large magnitude (Supplementary Table S1 available on Dryad).

We also summarized properties of these genes that have been suggested previously to influence the reliability of phylogenetic estimates, including rate of evolution, clockness, heterogeneity of base composition, amount of missing data, and alignment certainty. We were particularly interested in whether these properties could predict how strongly a gene would support (or reject) well-established relationships. To investigate these patterns, we calculated correlations between BF ranks and summary statistics across genes (Supplementary Table S2 available on Dryad). Only one property exhibited consistently negative correlations with BFs: alignment certainty. However, even for alignment certainty, the correlation was often non-significant or only of moderate magnitude when significant.

Overall, these results suggest that the most predictive feature of a gene's (un)reliability is its alignment certainty. Other characteristics may also be related to reliability, but these effects may be specific to certain bipartitions. We caution that these results do *not* suggest that any of these characteristics are necessarily unrelated to the quality of phylogenetic inference. Rather, the effect of some genic properties may be easier to predict and more consistent than others.

Future Prospects for the Use of Topological BFs

One reason BFs have only rarely been used to quantify topological support is the computational cost required to estimate them accurately. Current methods for estimating marginal likelihoods that are both reasonably accurate and widely implemented involve specially constructed MCMC analyses (Lartillot and Philippe 2006; Fan et al. 2010; Xie et al. 2011) specific to each topological hypothesis (e.g., $H_1 =$ bipartition A is present and $H_2 =$ bipartition A is absent). Estimating BFs for all bipartitions in a single unrooted tree topology with n tips then requires $2(n-3)$ independent MCMC analyses, which can be daunting for many studies. However, n is often not that large for many current phylogenomic data sets and each of these analyses can be conducted independently. Further time-savings are realized if researchers focus on specific bipartitions of greatest interest, perhaps those where disagreement is strongest between data sets.

New methods of marginal likelihood estimation that are both reasonably accurate and fast (e.g., the inflated density ratio, IDR, method; Arima and Tardella 2014) may become more widely available in the near future. IDR involves only the recycling of likelihood scores generated during posterior estimation, but would still require separate analyses to be run for each topological hypothesis. Recent work has also developed topological reference distributions that would allow increasingly accurate marginal likelihood estimation for hypotheses without a fixed topology (Holder et al. 2014; Wu et al. 2014). These developments should increase the accuracy of BF estimates for the types of topological hypotheses used in this study, although the most interesting patterns involve BFs that are orders of magnitude greater than the expected error in marginal likelihood estimation using the best current methods. We observed comparatively small variation in estimated marginal likelihoods across independent analyses.

Bergsten et al. (2013) recently published an insightful analysis of different approaches to calculating topological BFs and their potential pitfalls. In particular, they note that hypotheses involving constrained tree spaces, such as when testing monophyly, will tend to be strongly favored if prior distributions on topologies are very diffuse (e.g., uniform). The reason for this is the extraordinarily small prior probability of monophyly, when considering all topologies. As long as the posterior is more concentrated than the prior in a way consistent with monophyly, although perhaps not explicitly supporting it, the BF will be positive and often very large. Here, we have avoided this pitfall by leveraging extensive prior information about the backbone relationships among major amniote clades to specify a strongly informed topological

prior distribution. Similar prior information about the monophyly of major groups may be available in many circumstances where the application of BFs would be useful for investigating the relationships among those groups (e.g., the relationships among orders of placental mammals or amphibians). In such cases, the pitfalls of the "conventional" BF tests may be avoided. In cases where less prior information is available, we urge those interested in applying these tests to carefully consider how prior probability is distributed across their hypotheses of interest.

Conclusion

While MCMC estimates of posterior probabilities have dominated Bayesian phylogenetics as the preferred measure of support, they have important limitations. These limitations have become increasingly apparent and troublesome as the size of phylogenetic data sets has increased. The field needs alternate measures that can be used to more meaningfully compare support. Here, we have shown how BFs calculated with estimated marginal likelihoods can reveal marked and heretofore unrecognized variation in phylogenetic information. Understanding this variation can have a variety of downstream benefits, including more appropriate pairing of data sets and analytical methods, the ability to predict relationships most likely to be sensitive to methodological choices, and identification of outlier genes with surprising patterns of phylogenetic information. All of these should result in a more accurate and nuanced understanding of phylogenetic relationships, as well as patterns of molecular evolution. While we are not prepared to make any definitive statement about the phylogenetic position of turtles that was not already made in previous studies, we believe that we have laid the groundwork to resolve remaining questions (e.g., If turtles are sister to archosaurs, why do so many genes support a crocodilian sister placement?). We also hope that more studies will report BFs for different phylogenomic sequencing strategies to allow a general picture to emerge regarding their relative merits.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.8gm85.

Acknowledgements

Much of this research was conducted with high-performance computational resources provided by Louisiana State University (http://www.hpc.lsu.edu). V. Doyle, M. Hellberg, M. Holder, N. Lartillot, G. Mount, G. Naylor, and an anonymous reviewer provided helpful comments that greatly improved this manuscript. J. Thorne and C. Nasrallah generously provided the derivation of the expression for calculating the log of the average marginal likelihood across topological hypotheses while avoiding underflow.

Funding

J.M.B. and R.C.T. gratefully acknowledge financial support from startup funds provided by Louisiana State University and the University of Hawaii, as well as National Science Foundation awards DEB-1355071, DEB-1354506, and DBI-1356796.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Cont. 19:716–723.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Aci. Res. 25:3389–3402.

Arima S., Tardella L. 2014. Inflated density ratio (IDR) method for estimating marginal likelihoods in Bayesian phylogenetics. In: Chen M.-H., Kuo L., Lewis P.O., editors. Bayesian Phylogenetics (Methods, Algorithms, and Applications). Boca Raton, FL: Chapman & Hall/CRC. p. 25–57

Bergsten J., Nilsson A.N., Ronquist F. 2013. Bayesian tests of topology hypotheses with an example from diving beetles. Syst. Biol. 62:660–673.

Chang E.S., Neuhof M., Rubinstein N.D., Diamant A., Philippe H., Huchon D. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. USA 112:14912–14917.

Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10:65.

Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783–786.

Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J., Henderson J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of turtles. Mol. Phylogenet. Evol. 83:250–257.

DeBraga M., Rieppel O. 1997. Reptile phylogeny and the interrelationships of turtles. Zool. J. Linn. Soc. 120:281–354.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M. V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

Fan Y., Wu R., Chen M.-H., Kuo L., Lewis P.O. 2010. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28: 523–532.

Fong J.J., Brown J.M., Boussau B., Fujita M.K. 2012. A phylogenomic approach to vertebrate phylogeny supports a turtle-crocodilian affinity and a paraphyletic Lissamphibia. PLoS One 7:e48990.

Gaffney E. 1980. Phylogenetic relationships of the major groups of amniotes. In: Panchen A., editor. The terrestrial environment and the origin of land vertebrates. London: Academic Press. p. 593–610.

Holder M.T., Lewis P.O., Swofford D.L., Bryant D. 2014. Variable tree topology stepping-stone marginal likelihood estimation. In: Chen M.-H., Kuo L., Lewis P.O., editors. Bayesian Phylogenetics (Methods, Algorithms, and Applications). Boca Raton, FL: Chapman & Hall/CRC. p. 95–111.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray

D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. V Lovell P. V Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M. V., Alfaro-Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrom P., Edwards S. V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346:1320–1331.

Kass R.E., Raftery A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90: 773–795.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649.

Landan G., Graur D. 2007. Heads or tails: A simple reliability check for multiple sequence alignments. Mol. Biol. Evol. 24:1380–1383.

Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

Lu B., Yang W., Dai Q., Fu J. 2013. Using genes as characters and a parsimony analysis to explore the phylogenetic position of turtles. PLoS One 8:e79348.

Lyson T.R., Bever G.S., Bhullar B.-A.S., Joyce W.G., Gauthier J.A. 2010. Transitional fossils and the origin of turtles. Biol. Lett. 6:830–833.

Lyson T.R., Sperling E.A., Heimberg A.M., Gauthier J.A., King B.L., Peterson K.J. 2012. MicroRNAs support a turtle + lizard clade. Biol. Lett. 8:104–107.

Nylander J.A.A. 2004. MrModeltest v2. Program distributed by the author.

Penn O., Privman E., Ashkenazy H., Landan G., Graur D., Pupko T. 2010a. GUIDANCE: a web server for assessing alignment confidence scores. Nucl. Aci. Res. 38:W23–W28.

Penn O., Privman E., Landan G., Graur D., Pupko T. 2010b. An alignment confidence score capturing robustness to guide-tree uncertainty. Mol. Biol. Evol. 27:1759–1767.

Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E., Da Silva C., Wincker P., Le Guyader H., Leys S., Jackson D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19:706–712.

Pick K.S., Philippe H., Schreiber F., Erpenbeck D., Jackson D.J., Wrede P., Wiens M., Alié A., Morgenstern B., Manuel M., Wörheide G. 2010. Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships. Mol. Biol. Evol. 27:1983–1987.

Pisani D., Pett W., Dohrmann M., Feuda R., Rota-stabelli O., Philippe H. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA 112:15402–15407.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rieppel O., DeBraga M. 1996. Turtles as diapsid reptiles. Nature 384:453–455.

Rieppel O., Reisz R.R. 1999. The origin and early evolution of turtles. Annu. Rev. Ecol. Syst. 30:1–22.

Ronquist F., Teslenko M., Van Der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Schoch R.R., Sues H.-D. 2015. A middle Triassic stem-turtle and the evolution of the turtle body plan. Nature 523:584–587.

Shaffer H., Minx P., Warren D., Shedlock A.M., Thomson R.C., Valenzuela N., Abramyan J., Badenhorst D., Biggar K.K., Borchert G.M., Botka C.W., Bowden R.M., Braun E.L., Bronikowski A.M., Bruneau B.G., Buck L.T., Capel B., Castoe T.A., Czerwinski M., Delehaunty K.D., Edwards S. V, Fronick C.C., Fujita M.K., Fulton L., Graves T.A., Green R.E., Haerty W., Hariharan R., Hillier L.H., Holloway A.K., Janes D., Janzen F.J., Kandoth C., Kong L., de Koning J., Li Y., Literman R., Mardis E.R., McGaugh S.E., Minx P., Mork L., O'Laughlin M., Paitz R.T., Pollock D.D., Ponting C.P., Radhakrishnan S., Raney B.J., Richman J.M., St John J., Schwartz T., Sethuraman A., Shaffer B., Spinks P.Q., Storey K.B., Thane N., Vinar T., Warren D.E., Warren W.C., Wilson R.K., Zimmerman L.M., Hernandez O., Amemiya C.T. 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14:R28.

Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Swofford D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). version 4b10. Sunderland (MA): Sinauer Associates.

Thomson R.C., Plachetzki D.C., Mahler D.L., Moore B.R. 2014. A critical appraisal of the use of microRNA data in phylogenetics. Proc. Natl. Acad. Sci. 111:E3659–E3668.

Wang Z., Pascual-Anaya J., Zadissa A., Li W., Niimura Y., Huang Z., Li C., White S., Xiong Z., Fang D., Wang B., Ming Y., Chen Y., Zheng Y., Kuraku S., Pignatelli M., Herrero J., Beal K., Nozawa M., Li Q., Wang J., Zhang H., Yu L., Shigenobu S., Wang J., Liu J., Flicek P., Searle S., Wang J., Kuratani S., Yin Y., Aken B., Zhang G., Irie N. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat. Genet. 45:701–706.

Wu R., Chen M.-H., Kuo L., Lewis P.O. 2014. Consistency of marginal likelihood estimation when topology varies. In: Chen M.-H., Kuo L., Lewis P.O., editors. Bayesian Phylogenetics (Methods, Algorithms, and Applications). Boca Raton, FL: Chapman & Hall/CRC. p. 113–127

Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

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