Review article

Statistics and Truth in Phylogenomics

Sudhir Kumar,*[1,2] Alan J. Filipski,[1] Fabia U. Battistuzzi,[1] Sergei L. Kosakovsky Pond,[3] and Koichiro Tamura[4]

[1]Center for Evolutionary Medicine and Informatics, Biodesign Institute, Arizona State University

[2]School of Life Sciences, Arizona State University

[3]Department of Medicine, University of California San Diego

[4]Department of Biological Sciences, Tokyo Metropolitan University, Tokyo, Japan

***Corresponding author:** E-mail: s.kumar@asu.edu.

Associate editor: Jeffrey Thorne

Abstract

Phylogenomics refers to the inference of historical relationships among species using genome-scale sequence data and to the use of phylogenetic analysis to infer protein function in multigene families. With rapidly decreasing sequencing costs, phylogenomics is becoming synonymous with evolutionary analysis of genome-scale and taxonomically densely sampled data sets. In phylogenetic inference applications, this translates into very large data sets that yield evolutionary and functional inferences with extremely small variances and high statistical confidence (*P* value). However, reports of highly significant *P* values are increasing even for contrasting phylogenetic hypotheses depending on the evolutionary model and inference method used, making it difficult to establish true relationships. We argue that the assessment of the robustness of results to biological factors, that may systematically mislead (bias) the outcomes of statistical estimation, will be a key to avoiding incorrect phylogenomic inferences. In fact, there is a need for increased emphasis on the magnitude of differences (effect sizes) in addition to the *P* values of the statistical test of the null hypothesis. On the other hand, the amount of sequence data available will likely always remain inadequate for some phylogenomic applications, for example, those involving episodic positive selection at individual codon positions and in specific lineages. Again, a focus on effect size and biological relevance, rather than the *P* value, may be warranted. Here, we present a theoretical overview and discuss practical aspects of the interplay between effect sizes, bias, and *P* values as it relates to the statistical inference of evolutionary truth in phylogenomics.

Key words: molecular evolution, statistical inference, phylogenetics, evolutionary tree, statistical bias, variance.

Introduction

Phylogenomics has been specifically defined as the use of phylogenetic methods to predict protein functions via evolutionary analysis of a gene and its homologs (e.g., Eisen et al. 1997; Eisen 1998), an application that traces its roots to the early 1960s when proteins arising from gene duplications were first identified (e.g., Zuckerkandl and Pauling 1965). In molecular systematics, phylogenomics is usually taken to denote the inference of historical relationships among species using genome-scale sequence data (e.g., Delsuc et al. 2005). Uniting these two disparate definitions, phylogenomics is now the molecular phylogenetic analysis of genome-scale data sets for predicting gene function, identifying traces of molecular adaptation, inferring evolutionary patterns of macromolecules, and establishing relationships and divergence times of genes and species.

All phylogenomic analyses are statistical in nature and involve inferring truth about the evolutionary history of sequences either as an intermediate tool or as an end point. Statistical phylogenetics provides a framework for estimating historical patterns, inferring intrinsic parameters of evolutionary processes, and testing hypotheses under the auspices of the neutral theory of molecular evolution (Kimura 1983; Nielsen 2005; Yang 2006; Nei et al. 2010). Statistics affords quantification of the uncertainty in an estimate that is most frequently expressed as the standard error of the estimate and related measures and tests whether an observed effect is expected under a particular model of reality. These tests are necessary because a small data set sampled from a large population (usually assumed to be effectively infinite) may not be representative, a fact that was originally realized in the early 20th century (Fisher 1925; Rao 1989). These methods have been important because the acquisition of data in most biological disciplines is still slow and expensive, which results in rather small data sets. Consequently, the ability to quantify confidence in a measurement or hypothesis based on limited data is fundamental to valid scientific analysis, and the use of *P* values, quite rightly, is omnipresent in life sciences literature today.

In contrast to these limited sample scenarios, phylogenomics is a field rich in data for drawing historical inferences. Today, it is possible to envision a time when an extensive set of homologous sequences for each gene and genomic segment from a large sampling of living species will become available for many groups. That is, we will have effectively complete, or at least densely sampled, taxonomic sequence data to make functional and evolutionary inferences. So, phylogenomic analysis will frequently involve effectively complete or nearly complete data from genomes and species that can be brought to bear on a specific problem. Although "completeness" is not a statistical property per se, it is becoming important to appreciate it in

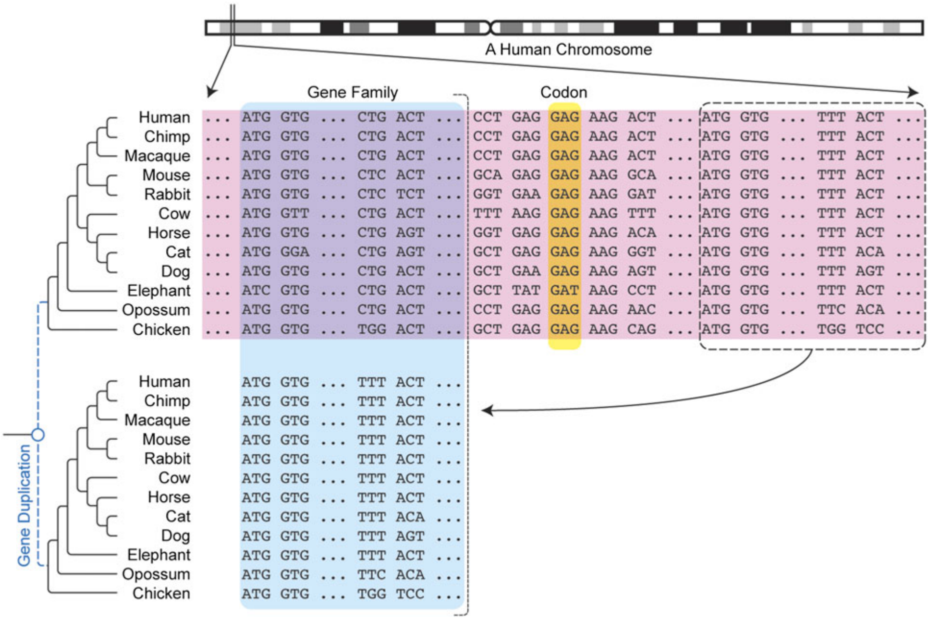
A Human Chromosome
Gene Family
Codon
Gene Duplication
Human ... ATG GTG ... CTG ACT ... CCT GAG GAG AAG ACT ... ATG GTG ... TTT ACT ...
Chimp ... ATG GTG ... CTG ACT ... CCT GAG GAG AAG ACT ... ATG GTG ... TTT ACT ...
Macaque ... ATG GTG ... CTG ACT ... CCT GAG GAG AAG ACT ... ATG GTG ... TTT ACT ...
Mouse ... ATG GTG ... CTC ACT ... GCA GAG GAG AAG GCA ... ATG GTG ... TTT ACT ...
Rabbit ... ATG GTG ... CTC TCT ... GGT GAA GAG AAG GAT ... ATG GTG ... TTT ACT ...
Cow ... ATG GTT ... CTG ACT ... TTT AAG GAG AAG TTT ... ATG GTG ... TTT ACT ...
Horse ... ATG GTG ... CTG AGT ... GGT GAG GAG AAG ACA ... ATG GTG ... TTT ACT ...
Cat ... ATG GGA ... CTG AGT ... GCT GAG GAG AAG GGT ... ATG GTG ... TTT ACA ...
Dog ... ATG GTG ... CTG ACT ... GCT GAA GAG AAG AGT ... ATG GTG ... TTT AGT ...
Elephant ... ATC GTG ... CTG ACT ... GCT TAT GAT AAG CCT ... ATG GTG ... TTT ACT ...
Opossum ... ATG GTG ... CTG ACT ... CCT GAG GAG AAG AAC ... ATG GTG ... TTC ACA ...
Chicken ... ATG GTG ... TGG ACT ... GCT GAG GAG AAG CAG ... ATG GTG ... TGG TCC ...
Human ATG GTG ... TTT ACT ...
Chimp ATG GTG ... TTT ACT ...
Macaque ATG GTG ... TTT ACT ...
Mouse ATG GTG ... TTT ACT ...
Rabbit ATG GTG ... TTT ACT ...
Cow ATG GTG ... TTT ACT ...
Horse ATG GTG ... TTT ACT ...
Cat ATG GTG ... TTT ACA ...
Dog ATG GTG ... TTT AGT ...
Elephant ATG GTG ... TTT ACT ...
Opossum ATG GTG ... TTC ACA ...
Chicken ATG GTG ... TGG TCC ...

FIG. 1. Anatomies of three types of phylogenomic data are discussed. (*A*) Genome-scale sequences for inferring evolutionary history of species, (*B*) a data set for tracing adaptive evolution for an individual codon, and (*C*) a multigene family sequence alignment for molecular phylogenetic analysis of gene duplications.

practice because the largest sample sizes possible to test specific hypotheses will ultimately be a reality in phylogenomics.

There are two different common scenarios. On one end, we have the complete genomes of many species for inferring their evolutionary relationships, whereas formerly, we had only a few genes. This quantity of data naturally gives us much reduced errors due to site sampling and estimation, which leads to very high power in the rejection of null hypotheses and high confidence values in establishing phylogenetic patterns (e.g., interior branch length in a tree is equal to zero). On the other end, we may also have "complete" available knowledge of the sequence of a codon across a large number of species within a relevant group when seeking tracks of adaptive evolution at individual codons in ancestral evolutionary lineages (fig. 1). In this kind of functional phylogenomic analysis, we need to estimate the numbers of nonsynonymous (amino acid altering) and synonymous (silent) substitutions over a long-term evolutionary history for a given codon. Because there are only three positions in a codon and the same positions often experience multiple substitutions, the estimates of nonsynonymous and synonymous substitutions per site will have large variances. This will afford us only low power in rejecting the null hypothesis of neutral evolution. This testing cannot be made more powerful if no more living species remain to be sampled or if they are uninformative on this point, as even the data from all species extant today constitute only an extremely sparse sample of all possible evolutionary trajectories. This limitation may not exist in detecting signatures of recent adaptive events within contemporary populations because they often consist of a large number of individuals and the population is continuously changing. This will lead to a very large data sets and powerful tests of selection (see reviews in Nielsen et al. 2007; Pool et al. 2010).

These realities of contemporary and emerging data sets are prompting biologists to carefully consider the biological importance of *P* values, as we move from analysis of a few genes to genome-scale data and also as we focus on tracking adaptive history of individual positions, codons, and segments in genomes. Needless to say, it is always important to establish statistical significance of any reported effect. However, statistical significance by itself need not be the only guide in phylogenomics, as there is a growing need to pay special attention to model assumptions and effect sizes. Effect size, in essence, is simply the relative (percentage) or absolute magnitude of difference observed between measurements. Effect sizes naturally relate directly to physical reality (truth), whereas *P* values are about our confidence in rejecting a null hypothesis. Because extremely significant *P* values can be obtained for very small effect sizes from very large data sets, the consideration and interpretation of effect sizes are becoming increasingly important for biology. For example, consider two large population samples of organisms that differ in mean weight by 0.01%. This difference is not likely to be biologically significant, though if the sample sizes are large enough, it well may be statistically so. Here, we emphasize the distinction between these two completely different notions (See also, Good 1992; Benjamini et al. 2001; Fidler et al. 2006; Gelman and Stern 2006; Hubbard and Armstrong 2006; Armstrong 2007; Nakagawa and Cuthill 2007; Porter 2008; Strimmer 2008; Ziliak and McCloskey 2008; Läärä 2009).

In the following sections, we explore the implications of the completeness of phylogenomic data sets, small and

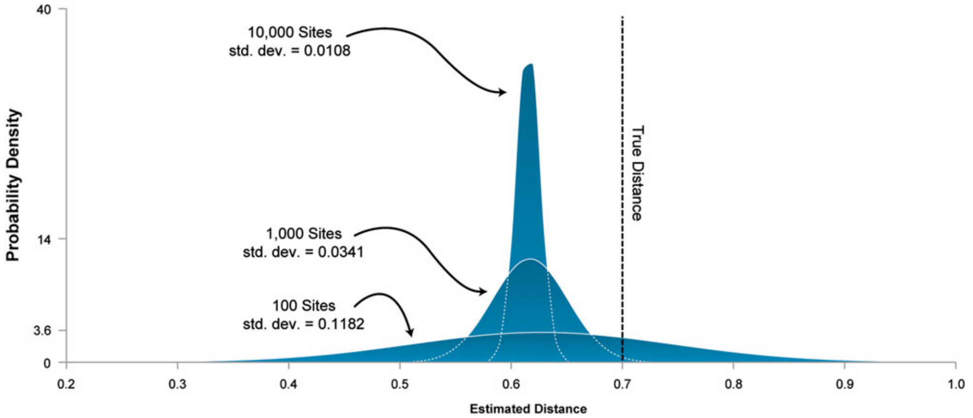

FIG. 2. An illustration of how large data sets allow an arbitrarily great reduction in the variance of an estimate without making it any more accurate. Pairs of DNA sequences with an evolutionary distance of 0.7 substitutions per site were generated according to a GTR (Lanave et al. 1984; Tavare 1986) of evolution using SeqGen (Rambaut and Grassly 1997). The evolutionary distance between simulated sequences was then estimated under the JC model (Jukes and Cantor 1969). The JC model is a special case of GTR; it does not model transition/transversion bias or base frequency biases, both of which are present in the simulated data. Therefore, the distance estimates will be biased. The figure shows how the distribution of estimates derived from 1,000 replicates narrows with increasing number of sites used (100 to 10,000 bp for the sequence length, in steps of a factor of 10). Each distribution was approximately normal, so normal curves are shown for simplicity. The mean estimate of distance under the JC model is close to 0.62 in each case since an overly simple model tends to underestimate distances. At the same time, the distribution of estimated distances narrows with increasing sequence length as described by the central limit theorem. As a result, the apparent precision of the estimate improves with increasing sequence length, but this improvement is spurious, as the mean estimate remains incorrect because of violations of model assumptions. Indeed, as the sequence length increases, the distances become, in a sense, less truthful, as they converge to a biased value and away from the true one. Thus, our confidence in an incorrect estimate can become arbitrarily high when bias is involved.

large, for navigating between the twin hazards of bias on one hand and lack of statistical power on the other and how these relate to accurate inference. Our discussions and examples here focus primarily on inferences drawn using a frequentist approach to statistics because of its widespread usage in phylogenomics studies, but we also discuss newer techniques that are being developed under modern statistical frameworks, including the Bayesian methods. Although our examples draw from phylogenomics research in animals, the discussion presented applies to other eukaryotes as well as to prokaryotes.

Statistical *P* values from the Analysis of Complete Data

Availability of genome sequences from all species in a group means that we have all of the available data on all observable differences to infer patterns of speciation and adaptation, if we assume negligible effect of within-species population variation (and horizontal gene transfer [HGT] events) on long-term historical patterns. This is a great boon to biologists because we are interested in understanding patterns and processes that gave rise to the contemporary diversity. However, when testing null hypotheses using genomic-scale data, *P* values from many commonly used statistical tests can be extremely low (highly significant). This is because *P* values from statistical tests provide guidance on whether the observed deviation from the expectation may be explained by chance alone. As we sample more and more of the data, the test becomes more powerful in rejecting the null hypothesis, even if only small deviations persist. Although the null hypothesis may be rejected with a highly significant *P* value for one set of assumptions for a given data set, both the effect size and the conclusion may change substantially when changes are made to the assumed model.

For example, an exome-scale squared correlation coefficient (r^2) of 0.1 between the ω (nonsynonymous to synonymous substitution rate ratio) and the number of synonymous substitutions per site was a matter of extensive analysis and interest because it was incredibly significant statistically ($P < 10^{-68}$) (Wyckoff et al. 2005). The correlation (effect size) is very small in this case, and it may be caused by unknown confounding factors or biases when estimating evolutionary distances (fig. 2) because a substitution model is needed to convert observed differences among sequences into the actual number of substitutions during the evolutionary history to account for multiple hits at the same site (reviewed in Yang 2006).

Indeed, a reanalysis of Wyckoff's data revealed that the observed correlation is not robust to the choice of model and the method of analysis (Li et al. 2009). Therefore, our enhanced ability to detect even the smallest correlations when using genome-scale data is hampered by a greater role of bias caused by the use of imperfect models. This is demonstrated by a simple example in figure 2, where estimates of evolutionary distances are obtained under a Jukes–Cantor (JC) model (Jukes and Cantor 1969). Pairs of DNA sequences with varying length (100 to 10,000 bp) with an evolutionary distance of 0.7 substitutions per site were generated according to a general time reversible (GTR) model (Lanave et al. 1984; Tavare 1986). For each sequence length, 1,000 sequence pairs were generated and estimates plotted. JC distances for different sequence pairs show a wide distribution because of random effects for short sequence lengths (100 bp). Even though the JC distance underestimates the true distance by 11% (0.62 vs. 0.70 substitutions per site), the distribution of JC distances contains the true distance (fig. 2). Furthermore, *P* value of the difference between the true and the estimated distance is not significant (two-sided *Z*-test at a 5% significance level) for 82% pairs of sequences of 100 bp each. Therefore, the null hypothesis will not be rejected for these. However, with an increase in sequence length, the distribution of JC distances narrows tremendously. For the case of length-10,000 sequences, the *P* value of the difference between the true and the estimated distance is significant at a 5% level for all sequence pairs, and all differences are significant at a 0.1% level for 100,000 bp sequences. The expected bias of 0.08 in the JC distance estimates is now much larger than the estimation variance, which reduces by a factor of 100 when using the 10,000 bp long sequences. This means that the abundance of data is allowing incorrect inferences to be assigned high significance when the nucleotide substitution model is incorrect. Therefore, in our opinion, it is better to place greater emphasis on the effect size and the robustness of the observed effect for a variety of models rather than on the *P* value associated with any particular genome-scale data analysis outcome. In general, high statistical significance in the presence of bias is not indicative of large effect size or biological importance of the result.

Multigene Family Phylogenetics

In contrast to the above situation with genome-wide estimates of biological quantities, limits on reported significance are experienced when we infer the evolutionary history of sequences in multigene families containing duplicated genes. The essential problem is that phylogenetic data available within a single locus is quite limited for most genes. A widely used practice in the field is to use a bootstrap test to determine the robustness of the phylogenetic tree inferred from an alignment of homologous sequences (Felsenstein 1985). Although the bootstrap support value has no simple frequentist interpretation as a probability, a bootstrap support of $\geq$95% is often considered to be roughly equivalent to the classical hypothesis-testing *P* value $\leq$ 0.05 (Efron et al. 1996).

In multigene family trees, rather low bootstrap support values are invariably encountered for many critical parts of the inferred phylogenetic tree, particularly those corresponding to gene duplication events. For example, the bootstrap support in the phylogenetic analysis of developmentally important genes is much less than 95% for a vast majority of gene duplications (e.g., Zhang and Nei 1996; Hughes et al. 2001; Nam et al. 2003). This places biologists in a quandary about whether to use or not to use branching patterns with low bootstrap support values when making biological interpretations. Because the lengths of the protein and DNA sequences are constrained in multigene family alignments, it is impossible to reduce the variance of the estimate and make the (bootstrap) statistical test more powerful.

Therefore, biologists will need to use the final estimates in drawing biological inference as long as the inferences are robust to the assumptions of substitution models and the phylogenetic method used, once all living species or species groups have been sequenced. Because many gene duplication events have produced orthologs in many species, it is possible in multigene phylogenetics to avoid including sequences (or species) that have evolved with unusual patterns of substitution or faster rates in order to generate consensus inferences. Such methods have been studied in the context of phylogenetic inference of species histories (e.g., Felsenstein 1988; Aguinaldo et al. 1997; Hassanin 2006), but they have remained relatively unexplored for the special case of multigene family phylogenetics.

Increasingly, Bayesian methods for phylogenetic inference and related estimation have emerged as alternatives to the bootstrap resampling approaches, especially when the bootstrap probabilities are low (See Yang 2005; Ronquist and Deans 2009 for reviews). And, it has become natural to use posterior probabilities (PPs) to assess confidence in the inferred phylogenies. PPs are estimates of the probability that a hypothesis is true given a data set and are explicitly based on assumptions about the prior distribution for that hypothesis. To distinguish from traditional *P* values, they are referred to as credibility values. Thus, they are not strictly comparable to bootstrap support but are often used for similar purposes. PP, for example, tends to be higher in value than bootstrap support. They are known to be sensitive to model violation, which may lead to inflated and high credibility values for contradictory inferences, a problem for which priors that reduce this effect have begun to emerge (Yang and Rannala 2005; Yang 2008). Bootstrap values, on the other hand, are not classical test probabilities in phylogenetics and are often considered conservative (Buckley 2002; Suzuki et al. 2002; Waddell et al. 2002; Alfaro et al. 2003; Douady et al. 2003; Erixon et al. 2003; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004).

Empirical analysis and computer simulations have now shown that PPs are close to the bootstrap support value when the latter is very high (Hillis and Bull 1993; Cummings et al. 2003; Erixon et al. 2003; Misawa and Nei 2003). Otherwise, PPs are thought to produce liberal estimates of confidence in results (Suzuki et al. 2002; Mar et al. 2005; Wrobel 2008). Bootstrapped PPs have been suggested as a hybrid alternative in which both ordinary Bayesian analysis and resampling are combined (e.g., Waddell et al. 2002; Douady et al. 2003). Recently, more advanced approximate maximum likelihood (ML) measures of branch support have been proposed that work efficiently on large data sets and combine frequentist and Bayesian frameworks (Anisimova et al. 2011). At a genomic scale, some promising Bayesian methodologies are becoming available to improve phylogenetic accuracy by inferring species and gene family trees simultaneously in a unified framework that models gene duplications and losses, gene- and species-specific rate variation, and sequence substitution simultaneously (e.g., Rasmussen and Kellis 2010).

Codon-Based Models for Diversifying and Directional Selection

Selection is another key property describing the evolutionary process and a major focus of phylogenomics investigations. Using statistical methods in evolutionary genetics, researchers frequently evaluate the strength of selection operating on individual codons over particular branches or regions of a phylogenetic tree (Anisimova and Kosiol 2009; Delport et al. 2009). The common approach is to estimate the rates of synonymous (d_S or K_s) and nonsynonymous (d_N or K_a) substitutions per site and compare them using a statistical test, typically using counting, ML, or Bayesian approaches (e.g., Nielsen and Yang 1998; Suzuki and Gojobori 1999; Nielsen and Huelsenbeck 2002). Making a further assumption that synonymous substitutions are selectively neutral, one statistically tests the ratio (ω) of these two rates at a codon or along an evolutionary lineage to infer adaptive or purifying selection (Hughes and Friedman 2008; Fletcher and Yang 2010). This approach has been applied in the analyses of individual genes, members of gene families, and in genome-wide scans (Zhang et al. 1998; Enard et al. 2002; Wong et al. 2004; Nielsen et al. 2005). Biological considerations suggest that ω should be a function of both the particular codon in the alignment and the branch in the phylogenetic tree—there is no a priori reason to believe that selective forces will affect any two sites or any two branches in exactly same way. However, a model that assigns a separate ω (ratio of nonsynonymous to synonymous rates) for each branch-site pair contains approximately twice as many parameters than the sample size (Posada 2008; Kosakovsky Pond et al. 2011) and is statistically untenable.

The simplest and early published approach is to estimate a single ω from all sites and branches jointly (e.g., Goldman and Yang 1994; Muse and Gaut 1994) and obtain the average value for the gene. Unless only a few sequences (≤ 5) are available (Nielsen 2005), this model is too simplistic because selective pressures vary both across sites and lineages. For larger samples, there are many approaches that have been designed to estimate ω for every site (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Kosakovsky Pond and Frost 2005; Massingham and Goldman 2005), with similar power and accuracy for alignments of sufficient divergence with 30 or more sequences (Kosakovsky Pond and Frost 2005). This can be either done directly by using a fixed effects type of statistical model (Kosakovsky Pond and Muse 2005; Massingham and Goldman 2005) or indirectly by fitting a discrete distribution of ω rates (random effects models) and then assigning each site to a particular class using a full or approximate Bayesian techniques (Yang et al. 2005; Huelsenbeck et al. 2006).

Such models have been used quite successfully in identifying sites subject to strong diversifying selective pressures (Sawyer et al. 2004; Kosakovsky Pond et al. 2006), but the effective sample size from which site-wise ω is derived is limited by the number of sequences in the alignment. More critically, if selective pressures fluctuate along the phylogenetic trees (as would be the case for selective sweeps followed by fixation), the assumption of temporally constant ω that site-wise models must make in order to gain power is violated. This has been partially remedied by developing models that seek evidence of such episodic selection, either directional (Seoighe et al. 2007; Delport et al. 2008; Kosakovsky Pond et al. 2008) or diversifying (Pupko and Galtier 2002; Guindon et al. 2004; Anisimova and Yang 2007; Kosakovsky Pond et al. 2011).

However, it may be impossible to identify rapid bursts of selection that influence only a few sites because the statistical basis of such inference will be limited to a few realizations of the evolutionary process (one or a few codons along a few lineages), and it is not clear how to increase the size of the sample by species-level sequencing. Two recent studies demonstrate that episodic diversifying selection can be identified but only if a sufficient number of sites (e.g., 10–15% for a typical gene) are affected by it and the size of the effect is large (e.g., $\omega = 4$) (Kosakovsky Pond et al. 2011; Yang and dos Reis 2011). This is indeed a very high bar to reach for individual positions.

Many of the standard modeling assumptions are likely to be false and will need to be relaxed in order to improve the power and accuracy of codon-based tests of selection. The assumptions that synonymous substitutions rates do not vary across a gene and are selectively neutral appear to be false, due to CpG hypermutability (e.g., mammals and plants), codon usage bias, and purifying selection against certain synonymous substitutions in humans, other mammals, and microbes (e.g., Suzuki and Gojobori 1999; Subramanian and Kumar 2003, 2006, ; Tamura, Subramanian, et al. 2004; Chamary and Hurst 2005; Kondrashov et al. 2006; Filipski et al. 2007; Mayrose et al. 2007; Resch et al. 2007; Gaffney and Keightley 2008; Ke et al. 2008; Lind et al. 2010; Ratnakumar et al. 2010). These effects can even appear in unexpected places, such as selection against CpG dinucleotides that has been observed in avian influenza viruses as they adapt to mammalian hosts (Jimenez-Baranda et al. 2011). Such effects can be partially overcome by incorporating synonymous rate variation or selection into the models (Kosakovsky Pond and Frost 2005; Yang and Nielsen 2008; Suzuki et al. 2009; Zhou et al. 2010). Also, ω-based methods can only measure substitution rates averaged over all possible amino acid residues; such models are gross under simplifications of the biological process, but work is underway to remedy this shortcoming (Doron-Faigenboim and Pupko 2007; Conant and Stadler 2009; Delport et al. 2010; Rodrigue et al. 2010).

Some of these issues can be circumvented by using codon-based models to define evolutionary metrics that extend standard genetic distances and compare the evolutionary process among nonhomologous genes (Kosakovsky Pond et al. 2010). Such approaches do not rely on P values of a specific test to decide if two genes are similar or dissimilar, which is fraught with issues of effect and sample size, but rather compare entire distributions of substitution rates between genes to measure how far apart their "selective" histories are. With such developments, statistical analysis of synonymous and nonsynonymous rates of change in individual codons and genes is becoming increasingly more useful for identifying codons within genes and genes in the genome that are ''likely candidates'' for adaptive change and prioritizing them for experimental validation, as the latter is necessary to test the statistical predictions. In most practical settings, scientists are using such comparative analysis of genomic data to generate biological hypotheses and narrowing down possibilities for test in the laboratory, without which the number of genes and positions are too large to tackle (Yang et al. 2009). Of course, the use of such procedures does not guarantee that all (or any) adaptively important genes or positions will get a high priority, and many challenges remain about the best practices in the use of available data and techniques and the vaildity of the associated conclusions (e.g., Yokoyama et al. 2008; Nozawa et al. 2009a, 2009b; Yang et al. 2009; Nei et al. 2010; Yang and dos Reis 2011). As mentioned earlier, such limitation would likely be overcome for detecting adaptive events within living populations with the fast accumulation of resequencing data that is increasing the number of synonymous and nonsynonymous variants observed (see reviews in Nielsen et al. 2007; Pool et al. 2010).

Genome-Scale Inference of Species Trees

Contrary to the plight of multigene family phylogenetics and codon-based tests of selection wherein the numbers of sites in the alignment are strictly limited, the evolutionary relationships of species can be inferred with increasingly longer sequences. Even though the process of HGT and other types of genome mixing may have obscured phylogenetic signals from the earliest diverging lineages in the tree of life (e.g., Doolittle 1999; Brown 2003; Philippe and Douady 2003; Lake and Rivera 2004; Ge et al. 2005; Gribaldo et al. 2010; Philippe et al. 2011), the possibility of definitely resolving the evolutionary relationships of major groups of species has greatly excited biologists (e.g., Rokas, Williams, et al. 2003; Zardoya and Suarez 2008; Genome 2009; Koonin et al. 2011). The reason for biologists' enthusiasm regarding phylogenomics is easy to illustrate using the available data from major groups of mammals (fig. 3). Two ML trees derived from 1,000 bp long alignments lack statistical resolution, despite the fact that each alignment contains a minimal number of insertions and deletions for mammalian noncoding sequence data (fig. 3*A*). The two trees differ topologically from each other in 22 places and from the accepted tree of mammals in 11 places.

Even though the consensus tree derived from individual trees of 992 noncoding sequence alignments is consistent with our current understanding of mammalian phylogeny (fig. 3*B*), the percentage of individual genomic-segment trees in which each clade was inferred is very low; only five partitions were recovered in more than 70% of the gene trees. Such results have also been observed in computer simulations modeled after mammalian gene evolution (e.g., Gadagkar et al. 2005). For many decades, such deficits of resolution were major sources of controversy; authors obtained different trees from varying data sets over time due to gene sampling errors (e.g., Castresana 2007; Huerta-Cepas et al. 2007). This problem is resolved by a genome-scale alignment of these data, which yields a phylogenetic tree with extremely high bootstrap support (fig. 4, base tree).

Does this mean that we now have a philosopher's stone in genome-wide alignments that will eventually reveal the true tree? In the current case, answering this question requires the assessment of whether the statistically significant inferences from one type of phylogenomic data are supported when using the other types of phylogenomic data (robustness of inference to data type sampled). Sequence alignments of thousands of protein-coding genes (from analyses of 250,221 coding nucleotide sites) are available for such a cross-validation experiment for figure 3 because the evolutionary history of coding and noncoding DNA is the same, assuming that orthology relationships are correctly inferred. In this case, one would expect the phylogenomic analysis of protein sequences and of different codon positions (first, second, and third) to yield the same evolutionary relationships. Indeed, a vast majority of phylogenetic groupings fall into this category. However, we found a few important differences, such as the position of the Afrotheria + Xenarthra clade (tenrec, elephant, hyrax, armadillo, and sloth) and the relative positioning of each of the Chiroptera, Artiodactyla, Perissodactyla, and Carnivora clades (fig. 4). Fortunately, these alternative resolutions of the tree are usually associated with relatively low bootstrap support (Nishihara et al. 2007; Wildman et al. 2007; Prasad et al. 2008). Still, bootstrap support is not always an accurate guide: Nishihara et al. (2007) obtained conflicting well-supported results for alternative eutherian trees depending on whether a concatenation- or separate-gene analysis strategy was used. Gadagkar et al. (2005) found such a problem with concatenation data analysis even for simulated data sets where there were no confounding factors or alignment biases and the correct substitution model was used.

It is now well appreciated that phylogenetic analysis of a concatenated alignment of genes can sometimes produce anomalously high levels of bootstrap support (e.g., Gadagkar et al. 2005; Seo 2008). Gadagkar et al. (2005) have advocated the reporting of consensus trees from individual genes (or genomic segments) to convey the presence of significant alternative phylogenetic signal. Conflicting high levels of support for alternative teleost phylogenies have been attributed to difficulties in alignment of noncoding regions (Negrisolo et al. 2010). In Bayesian analysis, methods are available to apply different models to different partitions, but identifying or confirming optimal partitions is sometimes problematic, and individual genes may still yield conflicting phylogenetic signal (Rokas, King, et al. 2003; Nylander et al. 2004; Brandley et al. 2005; Brown and Lemmon 2007). Substitution saturation and rate variation over time can also obfuscate phylogenetic signal and render data sets of little use (Penny et al. 2001; Mossel 2003). For example,

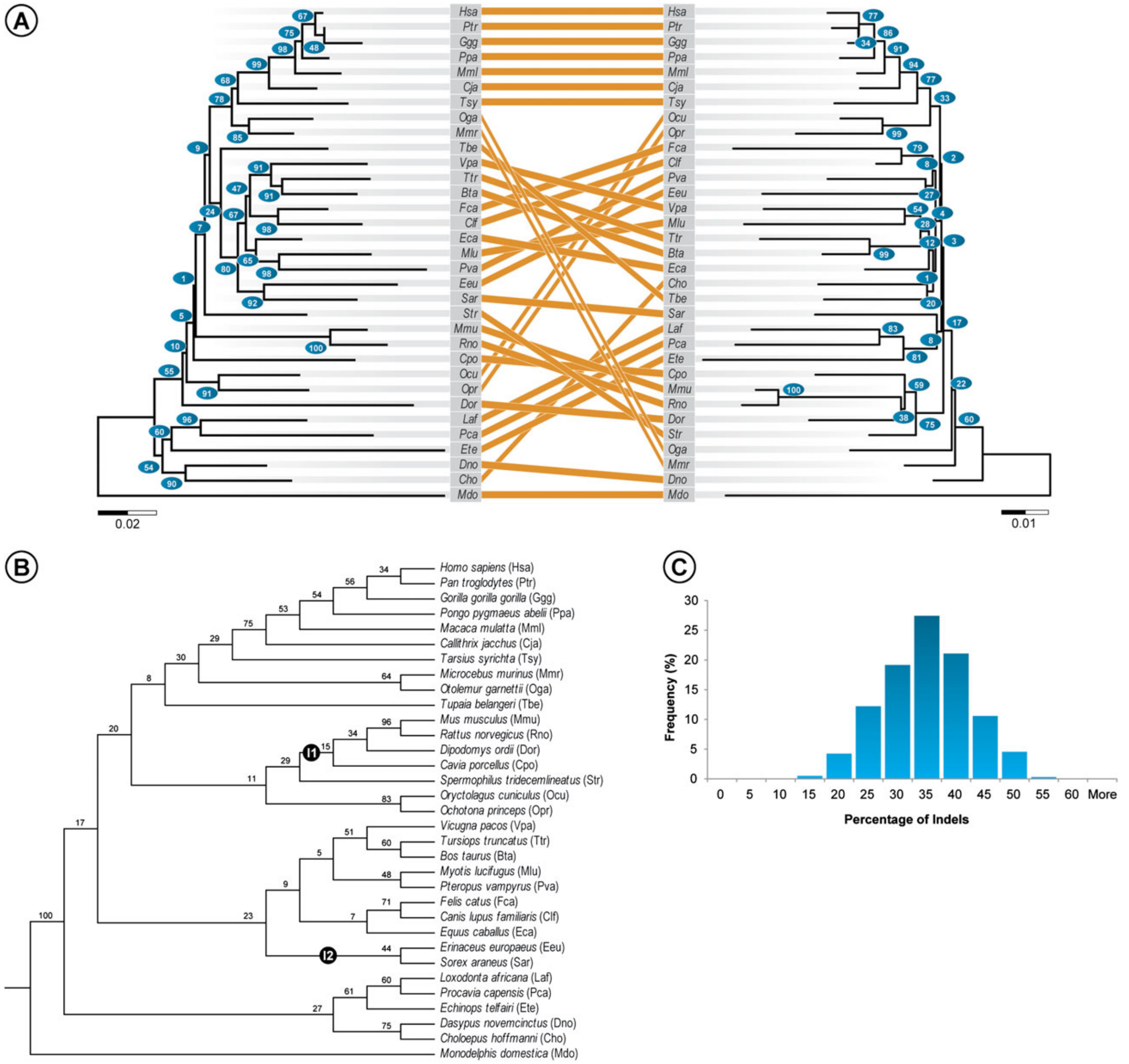

FIG. 3. Examples based on evolutionary relationships of 33 mammals inferred using a set of 992 noncoding DNA sequence alignments of 1,000 bp each. (*A*) Comparison of trees inferred from two 1,000 bp genomic segments containing the fewest insertions and deletions (11.4% and 12.6%, respectively). Bootstrap support obtained from 5,000 replicates is shown for both segments. Phylogenetic trees were inferred using maximum composite likelihood distances under a Tamura–Nei model (Tamura, Nei, et al. 2004) for neighbor joining analysis (Saitou and Nei 1987) with MEGA software (Kumar et al. 2008). The two trees differ in many places, showing that a sequence length of 1,000 bp is insufficient to reliably estimate many mammalian evolutionary relationships. (*B*) An extended majority rule consensus tree based on the 960 ML phylogenies inferred under a GTR Model of nucleotide substitution with gamma distribution of rates and invariant sites (GTR+Γ+I); ML tree inference failed to converge/complete for 32 data sets. Numbers on branches refer to the percentage of data sets (trees) in which the indicated cluster in the consensus tree was observed. Although the consensus tree topology is quite similar to the nominal University of California at Santa Cruz (UCSC) mammalian tree, differing only in the position of the bats, the low consensus numbers show that individual segment trees differ extensively. (*C*) A histogram is depicted showing the distribution of the percent bases involved in insertions or deletions in the 992 UCSC alignments. The alignments were extracted from the hg18 human genome alignment available from the UCSC Genome Browser at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz44way/ (Kuhn et al. 2009). Only the 32 placental and 1 marsupial species were used. We first divided each human chromosome into 1,000-bp segments and then all segments containing more than 600 sites with insertions and deletions for any placental or marsupial mammalian species in the alignment were discarded. This resulted in a total of 992 alignments of 1,000 bp each. I1 and I2 are two selected interior branches for which results are shown in figure 5.

Song et al. (2010) analyzed the inconsistent, but well-supported, phylogenomic trees from beetle mitochondrial data using several methods of inference and concluded that the problem was caused by failure to correctly model compositional heterogeneity and among-site rate variation.

Misleading Species History Inferences from Phylogenomics

Certainly, many phylogenomic investigations have yielded contrasting results. For example, the relative position of nematodes, arthropods, and chordates in the animal tree

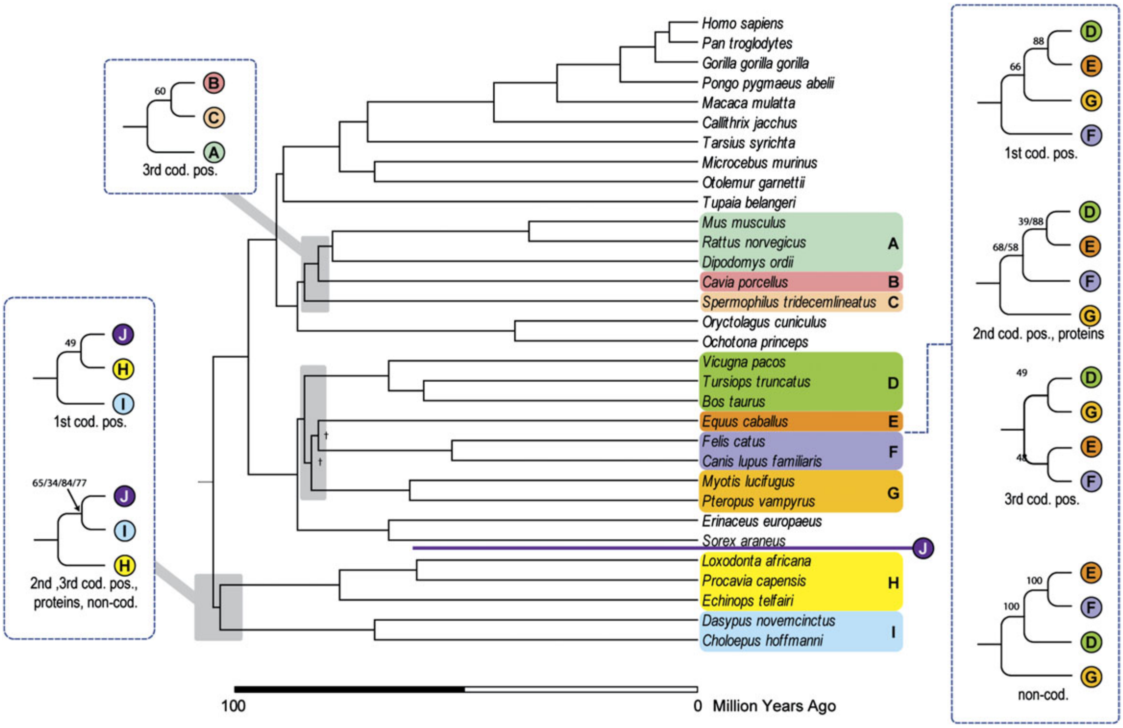

Fig. 4. Differences between the University of California at Santa Cruz (UCSC) tree of 32 placental mammals and neighbor joining (NJ) trees generated using five different sequence data partitions of similar size (first codon position, second codon position, third codon position, protein, and noncoding). The first three sets contain 83,407 bp, the protein set contains 83,407 amino acids, and the fifth data set consists of 100,000 noncoding DNA sites. The latter is a head-to-tail concatenation of 100 alignments of 1,000 bp homologous segments that have remained largely intact from insertions–deletions for the last 100 My (fewer than 20% of sites with an insertion or deletion). NJ trees were inferred using the maximum composite likelihood distance for DNA and the Jones–Taylor–Thornton substitution model for amino acids, respectively, in MEGA (Jones et al. 1992; Tamura et al. 2004; Kumar et al. 2008). Specific differences between the UCSC tree and the trees generated from different partitions are shown in the dotted boxes, along with bootstrap support values. Bootstrap support values for the main phylogeny (UCSC tree) were also calculated using 992,000 bp of noncoding DNA and were found to be 100% except for the two nodes flagged with a dagger (†), one of which had a bootstrap value of 82% (Perissodactyla as nearest neighbor to Carnivora) and the other of which (placement of bats as shown) was not present in the 992,000 bp tree.

has been highly controversial. Traditionally, chordates have been joined with arthropods into a Coelomata clade that excludes nematodes. Although protein sequence alignments supported this view, 18S ribosomal RNA studies supported a clade of molting animals (Ecdysozoa) in which nematodes and arthropods were more closely related to each other, to the exclusion of Chordates (Aguinaldo et al. 1997). More extensive genomic studies during the last decade seemed to favor the classical Coelomata hypothesis. However, these results have been questioned because the inclusion of the fast-evolving representative nematode (*Caenorhabditis elegans*) is thought to bias the phylogenetic inference (see an overview in Telford et al. 2008). Subsequent genomic-scale analyses reject the Coelomata hypothesis by attributing previous findings to inadequate choice of outgroups (Holton and Pisani 2010). Other examples of opposing statistical support include the phylogenetic position of myzostomids (parasitic organisms typically found on echinoderms), whose placement in the metazoan tree by use of 77 ribosomal proteins is strongly rejected by mitochondrial sequence data (Bleidorn et al. 2009).

In general, organellar genomes (e.g., mitochondrial, chloroplast, and plastid) have been widely used in phylogenetics for several reasons, including faster evolutionary rates of mitochondrial genes than nuclear genes (e.g., in mammals) and historical ease of sequencing their complete genomes due to relatively smaller sequence lengths. However, major discordances between nuclear and mitochondrial trees have been seen in the phylogenetic analysis of vertebrate proteins. The head-to-tail concatenation of all the protein-coding sequences in the mitochondrial genomes of 11 vertebrates rooted using an outgroup (lampreys or sea urchins) suggests, with high statistical confidence, that all mammals form a sister group to other vertebrates, including chicken, frog, and bony fish (Takezaki and Gojobori 1999). This inference is not generated by an analysis of nuclear DNA (e.g., Venkatesh et al. 2001). Many other examples have been cited, including cases where complete mitochondrial genomes lead to inference of incorrect trees with high confidence (Naylor and Brown 1998; Springer et al. 2001; Ballard and Whitlock 2004; Wahlberg et al. 2009). The lesson here is that genome-scale data do not always lead to a more correct (or consistent) picture of the evolutionary history of species. Sometimes, there is a need for the sampling of alternative species representatives or the use of alternative molecular characters, such as SINES (Shedlock et al. 2004) and location of introns (Krauss et al. 2008).

Although the deep branching in the tree of mammals and the Ecdysozoa/Coelomata controversies are two of many phylogenetic controversies in phylogenomics today, it is important to emphasize that most questions of molecular phylogeny do not present such problems. Such inquiries either correspond fairly well with earlier estimates from morphological data or they have been satisfactorily resolved by taking a phylogenomic approach (e.g., Springer et al. 2004). Nevertheless, persistent unresolved questions remain, and many of these are turning out to be difficult problems, a circumstance that is not clearly due to lack of data so much as the lack of adequate phylogenetic signal in the data (e.g., Soltis et al. 2004; Rokas and Chatzimanolis 2008).

Phylogenetic signal is a direct function of the length of the branch (in units of the expected number of substitutions per site) that defines the evolutionary relationship. Whenever two nodes are separated by a short branch, they become prime candidates for producing resolution inconsistencies when using different large-scale data sets, yet (in many cases) with very high bootstrap confidence. A high bootstrap confidence for multiple contrasting resolutions indicates that statistical bias is influencing the outcome because it is never possible to completely model and incorporate all parameters needed to describe an evolutionary process. Phylogeneticists refer to this bias as "systematic error" (Swofford et al. 1996), which has existed in all phylogenetic analyses to some extent, though it has been rather small compared with the sampling variance that is due to the finite length of sequence alignments in the past. For genome-scale data sets, the site-sampling variance reduces to virtually zero. In contrast, the systematic error does not appear to be reduced (and may even increase) with the increasing size of data sets and the antiquity of the species included, so it becomes the central determinant of statistical support. Although the ordinary phylogenetic bootstrap support estimation is expected to be unbiased (Efron et al. 1996), all statistical estimates of confidence (including the bootstrap test) are based on the assumption that the data are well described by the assumed models and that the sites are identically distributed and independent. Bias will occur when the models used for analyses are underparameterized (e.g., Erixon et al. 2003).

Bias in Phylogenomic Analyses

Bias is introduced in phylogenomics because statistical analyses require a model of sequence evolution, which can only approximate the true, but unknown, evolutionary process. In modeling evolutionary processes, we implicitly assume that positions and lineages are homogeneous in terms of their evolutionary patterns (mentioned below). However, each site and lineage may have evolved under unique evolutionary forces, some having more or less random and independent effects, others systematic and correlated. Natural selection varies over time depending upon the organism's environment, and it may involve correlated changes at different sites. Similarly, the pattern of mutation is not the same in all lineages and in all sites nor may we assume that its effects are linear and additive so that averaging accurately reflects the total effect. Having large data sets allows us to reduce sampling error but at the same time increases the probability that the basic assumption of homogeneity across sites and evolutionary time will be violated. Consequently, the traditional and highly desirable property of statistical estimations, where deviations cancel with increasing data set size, is overwhelmed by the bias introduced by violations of the underlying assumptions in phylogenomic inference. This problem persists despite the fact that statistical phylogenetics has advanced greatly in sophistication and enables the scientists to utilize complex evolutionary models for substitution probabilities, incorporate across-site rate variation, look for the presence of heterotachy due to differences in relative rate among lineages of different genes, and incorporate compositional heterogeneity among lineages (e.g., Shoemaker and Fitch 1989; Galtier and Gouy 1995; Yang 1996; Tamura and Kumar 2002; Kolaczkowski and Thornton 2004; Kumar and Filipski 2007; Rannala and Yang 2008). In fact, all methods of phylogenetic inference have their own inherent biases that can produce highly supported but erroneous results, even when the data have evolved with identical patterns (Gadagkar et al. 2005; Xia 2006; Susko 2008).

Because the pattern of nucleotide substitution is expected to vary among sites, partitioned analyses are carried out where different genomic segments and/or codon positions are allowed to have their own pattern (reviewed in Rannala and Yang 2008). The genome-scale data set is partitioned in different categories, and the ML trees are optimized accordingly, making it possible to relax the homogeneity across sites assumption. However, not all sites within each category have evolved with the same pattern because, for example, not all third codon positions are equivalent—they differ in their levels of codon degeneracy (0-, 2-, and 4-fold) as well as the type of degeneracy (purine vs. pyrimidine degeneracy). This means that a large number of site categories would need to be introduced upon a priori in order to avoid bias. Of course, it is not always possible to determine these categories, except those that are easily presented by molecular biological considerations (e.g., codon positions, genes, and degeneracy of the genetic code). Furthermore, once conflicting subsets of sites, or mosaic genomes, are determined, it is not necessarily clear how to account for the differences (Evans et al. 2010; Hallström and Janke 2010).

Even when subdividing the data set into homogeneous site collections, scientists frequently assume that the (expected) relative branch lengths in the true tree are the same across data partitions. Violation of this assumption can lead to the heterotachy problem (Lopez et al. 2002), which is known to mislead phylogenetic inference, especially when the interior branch length is short (Kolaczkowski and Thornton 2004; Gadagkar and Kumar 2005; Philippe et al. 2005). Such a detrimental effect of heterotachy may be decreased by optimizing branch lengths for each site partition independently (Kolaczkowski and Thornton 2008).

Sophisticated tests of how well a model fits a data set sometimes reveal a poor fit of real or simulated sequence data to an inferred tree (Bollback 2002; White et al. 2007). We know that, in many cases, underparameterization leads to underestimates of distances (Nei and Kumar 2000), but it is unclear how that affects phylogenetic reconstruction. In general, the science of understanding how model violations affect the accuracy of estimates of genome-scale trees is in its infancy (Ripplinger and Sullivan 2010). We do not yet have reliable and comprehensive ways of detecting or measuring different kinds of model violations in empirical data. When the interior branch lengths are not very short, the bias introduced by the use of approximate (albeit sophisticated) models to describe biological complexity does not severely affect the result enough to mislead us. And, the availability of large numbers of sequences and sites decreases the sampling error and provides a way to infer evolutionary relationships with high statistical support and accuracy.

However, high statistical significance for very short branches due to low site-sampling errors is a cause for suspicion of undetected sources of bias. One way to avoid bias caused by concatenation of heterogeneous data subsets is to use methods that enable us to group sites with similar substitution patterns together and then conduct analyses that model each partition individually (e.g., Shapiro et al. 2006; Bofkin and Goldman 2007; Rannala and Yang 2008). Of course, it is possible that we would get contradictory high confidence resolutions for different data sets, even with the most sophisticated models and conforming data. In such instances, because of closely spaced speciation events, there simply may not be enough of a phylogenetic signal in the genomes to resolve the evolutionary relationships! If we already are using a complete data set of a certain type, then barring development of better models, we need to look for alternative types of data to resolve the branching patterns.

Accurate determination of orthology is another important issue in phylogenomics for animals, plants, and microbes. This is particularly so in the latter, as microbial genomes undergo HGT and recombination extensively. In this case, a single genomic alignment is actually a mixture of multiple alternative phylogenies, and the genome phylogeny is not classical tree-like. This feature constrains the number of vertically inherited genes available to be small when large number of species are included, with increasing the number of genes requiring decreases in the taxon sampled (e.g., Ciccarelli et al. 2006; Pisani et al. 2007). Of the thousands of genes present in most bacterial species, for example, only a very small subset (numbering less than 50) is considered at the core of species phylogeny reconstruction (Charlebois and Doolittle 2004; Koonin and Wolf 2009). On the other hand, consensus tree methods (e.g., supertrees) enable the use of large numbers of genes for few species, often combining partially overlapping trees obtained with as few as four taxa. The problem then becomes how to tell whether alternative phylogenies that are detected from these reduced data sets reflect actual gene histories or simply artifact and bias (Bapteste et al. 2005; Comas et al. 2007; Ané 2011; Blair and Murphy 2011). With horizontally transferred genes estimated to be up to 60% of a genome depending on the species included in the analyses and the methods used to identify the horizontal transfer (Dagan and Martin 2006), the very concept of species is being questioned with no simple answer in sight for species that show high frequency of HGT (reviewed in Doolittle and Zhaxybayeva 2009; Vos 2011). In the presence of a fluid genome, the question is then which percentage of genes should be native to a genome to be able to define a species and how can we identify them.

An additional potential source of bias in the phylogenomics of microbes is due to the heterogeneity of evolutionary rates and differences in base composition biases among species in large data sets. Although present in any large group of organisms, this issue is particularly severe in microbes due to the large difference in G+C content. The constraints in evolutionary mechanisms imposed by compositional biases have been considered the culprit of some controversial phylogenetic placements, such as that of hyperthermophilic lineages at the base of the bacterial tree (Gribaldo and Philippe 2002). A similar situation is present in eukaryotic microbes, namely in the Plasmodium group (Apicomplexa) in which variable base compositions (G+C ~25% to ~45%) and amino acid usages are responsible for phylogenomic conflicts (Davalos and Perkins 2008). Overall, HGT, recombination, base composition bias, and other factors are now recognized to be major influences on microbial phylogenomics over both (evolutionary) time and (genome) space dimensions (Ciccarelli et al. 2006; Philippe et al. 2011).

Sequence Alignment as a Source of Bias

In the above discussion, we assumed that all the sequences were homologous, with site homology across sequences known without error. Except for closely related sequences, this assumption is rarely met. The assembly of phylogenomic data sets requires a multiple sequence alignment, and all commonly used multiple sequence alignment algorithms use a guide tree and yield data sets containing gaps that are needed to maximize site homologies (Kumar and Filipski 2007; Notredame 2007; Lunter et al. 2008; Rosenberg 2009). The choice of a guide tree and the handling of gaps during analysis are known to introduce problems in phylogenomics. Many studies have already found the effect of the presence of large amounts of indels in alignments to be substantial in inferring phylogenies accurately (e.g., Wiens 2006; Hartmann and Vision 2008; Löytynoja and Goldman 2008).

Highly supported but contradictory phylogenies can be produced when using implicit guide trees in the alignment of very long raw sequences. In a four-species example, the guide tree phylogeny becomes the inferred phylogeny when the number of sites is large. In data sets such as these, the variance of estimates of evolutionary parameters becomes vanishingly small with increasing sequence length. However, the alignment bias caused by errors in the guide tree will affect each position aligned, and it will not diminish with increasing sequence length (Kumar and Filipski 2007). Joint alignment and phylogeny estimation methods

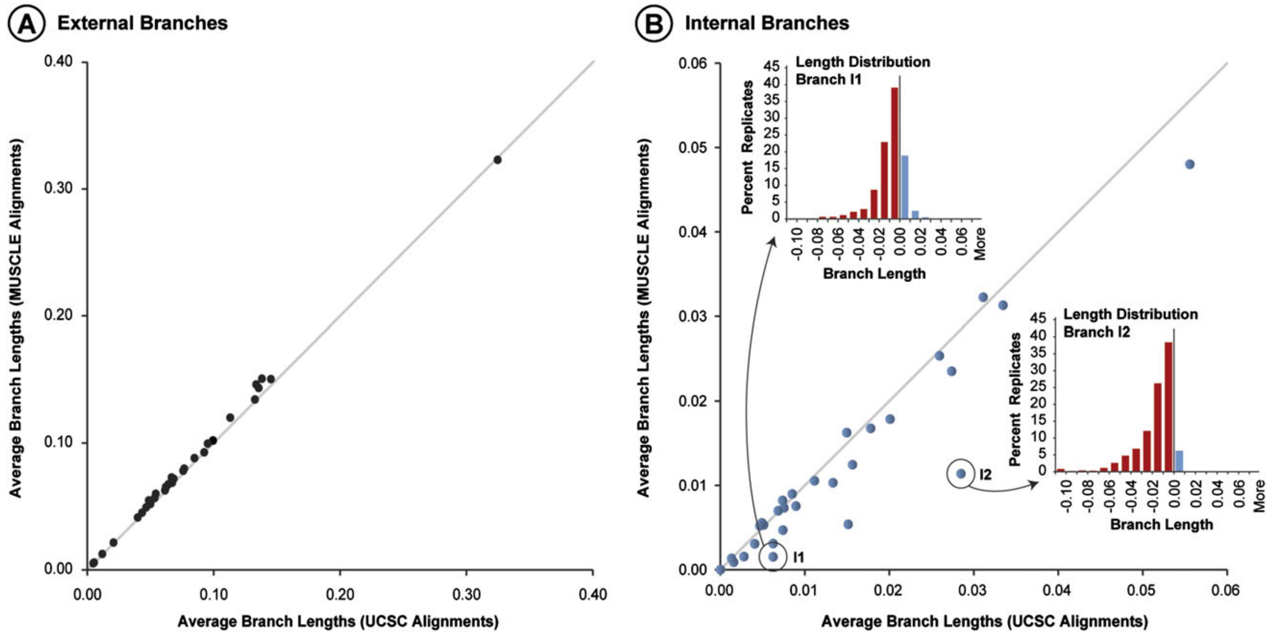

Fig. 5. Relationship of the interior branch lengths inferred from University of California at Santa Cruz (UCSC) database (Miller et al. 2007; Kuhn et al. 2009) and MUSCLE alignments for (*A*) external and (*B*) internal branches in the UCSC database phylogeny. Each point represents an average from the analysis of 992 data sets (1,000 bp each). For each data set, branch lengths were inferred by fitting the maximum composite likelihood distances onto the UCSC tree topology by employing the ordinary least squares (OLS) approach (Rzhetsky and Nei 1993) in MEGA5 (Tamura et al. 2011). The UCSC tree topology was created by Miller et al. (2007) as the one that seemed in best agreement with the published literature. OLS was chosen because it naturally allows for negative branch lengths that may occur when the topology used is not the optimal tree. MUSCLE alignments were conducted using the default options (Edgar 2004); UCSC generated alignments using the MULTIZ program (Blanchette et al. 2004) and the UCSC tree topology as described in detail in Miller et al. (2007). Histograms of differences between MUSCLE and UCSC branch lengths for individual 1,000 bp segment alignments are inset for two internal branches (I1 and I2), which show substantial difference in branch lengths between the UCSC and MUSCLE alignments. These two branches are marked by open circles in panel B scatter plot, and their positions in the UCSC phylogenetic tree are shown in figure 3*B*. Red bars in the histogram show the frequency of with which the use of segment-specific alignments produces smaller branch lengths than the UCSC alignment for the same segment.

are another promising approach, but it is not yet clear what their general properties may be in terms of bias and extensibility (Fleissner et al. 2005; Lunter et al. 2005; Redelings and Suchard 2005; Novák et al. 2008).

Genome-scale alignments for phylogenetic inference are usually generated by aligning individual segments of the genome and assembling them into a full alignment. This process is expected to adversely impact the total phylogenetic information, especially for short branches in the true tree. This is because the process of carrying out an independent multiple sequence alignment for each segment implicitly leads to the use of segment-specific guide trees in contemporary alignment programs. These guide tree topologies will differ significantly from the true tree and from each other, which would introduce conflicting signals in different segment alignments as the site homologies are optimized using the guide tree. To expose this effect, we simply realigned 992 homologous segments from 33 mammalian species separately using the MUSCLE software such that segment-specific guide trees are used (Miller et al. 2007; Kuhn et al. 2009). Then, the branch lengths of the known topology relating these species were calculated using the original alignments (which were generated using the known topology, *x* axis in fig. 5) and compared with the branch lengths obtained using the segment-specific guide trees (*y* axis in fig. 5). Although the external branches are estimated well for the original tree in the segment-specific alignments (fig. 5*A*), many internal branches are underestimated considerably due to guide-tree effect (fig. 5*B*). For example, a vast majority (>90%) of segment-specific alignments produce negative branch length for the original tree (fig. 5*B*, examples I1 and I2). These realities of practical data analyses are another source of bias in deciphering the branching order of hard-to-resolve parts of the tree of life.

Conclusions

Statistics has brought rigor to scientific investigation, and the concept of null hypothesis testing has been important in differentiating chance patterns from meaningful events in biology. However, the interpretation of and reliance on formal statistical significance in the field of phylogenomics comes with many challenges. We have discussed some of these challenges when phylogenetics is used in the identification of signatures of adaptive change, deciphering evolutionary dynamics of multigene families, and inferring species history. These discussions are not meant to distract from the mathematical correctness of the statistical methods nor their successful application in molecular evolution. Rather, our discussions highlight how the inferences can easily be misleading when the assumptions made by the methods are violated. The presence of any bias, however small, can lead to high statistical significance due to the extremely large genomic samples available to infer phylogenetic trees and estimate substitution and evolutionary parameters. Such biases arise because the fundamental assumption of independently and identically distributed data is not met in empirical genome-scale data sets. Furthermore, the impossibility of perfectly describing any biological process using mathematical models inevitably introduces biases. All models are wrong in some sense but that does not preclude some from being useful (Box and Draper 1987). The hard work is distinguishing the useful models from the misleading ones.

In our view, robustness of observed patterns across methods and data will become increasingly important in phylogenomics. In addition to the inference of species history, it will also hold true for the case of multigene phylogenetics and codon-based tests of adaptive evolution, where the universe of data (sites) is constrained by biological reality. Even though sources of many biases are known, their identification and measurement in particular cases are difficult due to the frequent requirement of knowledge of the evolutionary history of the sequences in question. For this reason, a greater emphasis needs to be placed on effect size when drawing biological conclusions. Furthermore, in order to proceed in understanding genomic biology, we must find ways to deal with the fact that we are now beginning to exhaust certain kinds of available data. To continue, we must, therefore, develop improved methods, devise more sophisticated models, and identify further sources of data that can supply different perspectives on the basic evolutionary questions of phylogeny and selection. Although we have focused on potential bias and modeling pitfalls and cautioned about statistical overconfidence, we do not see these things as intrinsic weaknesses, and we believe that the field of phylogenomics is poised to have great impact on the progress of both evolutionary science and modern medicine.

Acknowledgments

We thank Revak Raj Tyagi for his assistance with data analysis and Dr Jesse Taylor and anonymous reviewers for many insightful comments. We thank Carol Williams for her editing support. This research was supported by research grants from the National Institutes of Health to S.K. and A.J.F.

References

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489–493.

Alfaro ME, Zoller S, Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol Biol Evol.* 20:255–266.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biol Evol.* 3:246–258.

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol.* 60:685–699.

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol.* 26:255–271.

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol.* 24:1219–1228.

Armstrong JS. 2007. Significance tests harm progress in forecasting. *Int J Forecast.* 23:321–327.

Ballard JW, Whitlock MC. 2004. The incomplete natural history of mitochondria. *Mol Ecol.* 13:729–744.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol.* 5:33.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. 2001. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res.* 125:279–284.

Blair C, Murphy RW. 2011. Recent trends in molecular phylogenetic analysis: where to next? *J Hered.* 102:130–138.

Blanchette M, Kent WJ, Riemer C, et al. (12 co-authors). 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14:708–715.

Bleidorn C, Podsiadlowski L, Zhong M, Eeckhaut I, Hartmann S, Halanych KM, Tiedemann R. 2009. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol Biol.* 9:150.

Bofkin L, Goldman N. 2007. Variation in evolutionary processes at different codon positions. *Mol Biol Evol.* 24:513–521.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol.* 19:1171–1180.

Box GEP, Draper NR. 1987. Empirical model-building and response surfaces. New York: Wiley.

Brandley MC, Schmitz A, Reeder TW. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Syst Biol.* 54:373–390.

Brown JM, Lemmon AR. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Syst Biol.* 56:643–655.

Brown JR. 2003. Ancient horizontal gene transfer. *Nat Rev Genet.* 4:121–132.

Buckley TR. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst Biol.* 51:509–523.

Castresana J. 2007. Topological variation in single-gene phylogenetic trees. *Genome Biol.* 8:216.

Chamary J, Hurst L. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:R75.

Charlebois RL, Doolittle WF. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* 14:2469–2477.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.

Comas I, Moya A, Gonzalez-Candelas F. 2007. From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case. *Syst Biol.* 56:1–16.

Conant GC, Stadler PF. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol Biol Evol.* 26:1155–1161.

Cummings MP, Handley SA, Myers DS, Reed DL, Rokas A, Winka K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. *Syst Biol.* 52:477–487.

Dagan T, Martin W. 2006. The tree of one percent. *Genome Biol.* 7:118.

Davalos LM, Perkins SL. 2008. Saturation and base composition bias explain phylogenomic conflict in Plasmodium. *Genomics* 91:433–442.

Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, Kosakovsky Pond SL. 2010. CodonTest: modeling amino acid substitution preferences in coding sequences. *PLoS Comput Biol.* 6:e1000885.

Delport W, Scheffler K, Seoighe C. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog.* 4:e1000242.

Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform.* 10:97–109.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6:361–375.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–2129.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. *Genome Res.* 19:744–756.

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. *Mol Biol Evol.* 24:388–397.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol.* 20:248–254.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 5:113.

Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A.* 93:13429–13434.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8:163–167.

Eisen JA, Kaiser D, Myers RM. 1997. Gastrogenomic delights: a movable feast. *Nat Med.* 3:1076–1078.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418:869–872.

Erixon P, Svennblad B, Britton T, Oxelman B. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst Biol.* 52:665–673.

Evans NM, Holder MT, Barbeitos MS, Okamura B, Cartwright P. 2010. The phylogenetic position of Myxozoa: exploring conflicting signals in phylogenomic and ribosomal data sets. *Mol Biol Evol.* 27:2733–2746.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution.* 39:783–791.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet.* 22:521–565.

Fidler F, Burgman MA, Cumming G, Buttrose R, Thomason N. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv Biol.* 20:1539–1544.

Filipski A, Prohaska S, Kumar S. 2007. Molecular signatures of adaptive evolution. In: Pagel M, Pomiankowski A, editors. Evolutionary genomics and proteomics. Sunderland (MA): Sinauer Associates. p. 241–254.

Fisher RA. 1925. Statistical methods for research workers. Edinburgh, (UK): Oliver and Boyd.

Fleissner R, Metzler D, von Haeseler A. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst Biol.* 54:548–561.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol.* 27:2257–2267.

Gadagkar SR, Kumar S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Mol Biol Evol.* 22:2139–2141.

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol.* 304: 64–74.

Gaffney DJ, Keightley PD. 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. *BMC Evol Biol.* 8:265.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. *Proc Natl Acad Sci U S A.* 92:11317–11321.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol.* 3:e316.

Gelman A, Stern H. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *Am Stat.* 60:328–331.

Genome KCoS. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J Hered.* 100:659–674.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725–736.

Good IJ. 1992. The Bayes/non-Bayes compromise: a brief review. *J Am Stat Assoc.* 87:597–606.

Gribaldo S, Philippe H. 2002. Ancient phylogenetic relationships. *Theor Popul Biol.* 61:391–408.

Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. 2010. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat Rev Microbiol.* 8:743–752.

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A.* 101:12957–12962.

Hallström BM, Janke A. 2010. Mammalian evolution may not be strictly bifurcating. *Mol Biol Evol.* 27:2804–2816.

Hartmann S, Vision TJ. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol Biol.* 8:95.

Hassanin A. 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. *Mol Phylogenet Evol.* 38:100–116.

Hillis D, Bull J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol.* 42:182–192.

Holton TA, Pisani D. 2010. Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol Evol.* 2:310–324.

Hubbard R, Armstrong JS. 2006. Why we don't really know what "statistical significance" means: a major educational failure. *J Mark Educ.* 28:114–120.

Huelsenbeck J, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst Biol.* 53:904–913.

Huelsenbeck JP, Jain S, Frost SW, Kosakovsky Pond SL. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A.* 103:6263–6268.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. 2007. The human phylome. *Genome Biol.* 8:R109.

Hughes AL, da Silva J, Friedman R. 2001. Ancient genome duplications did not structure the human hox-bearing chromosomes. *Genome Res.* 11:771–780.

Hughes AL, Friedman R. 2008. Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. *Immunogenetics* 60:495–506.

Jimenez-Baranda S, Greenbaum B, Manches O, Handler J, Rabadan R, Levine A, Bhardwaj N. 2011. Oligonucleotide motifs that disappear during the evolution of influenza virus in humans increase alpha interferon secretion by plasmacytoid dendritic cells. *J Virol.* 85:3893–3904.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 8:275–282.

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

Ke S, Zhang XHF, Chasin LA. 2008. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res*. 18:533–543.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980–984.

Kolaczkowski B, Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol Biol Evol*. 25:1054–1066.

Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2006. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J Theor Biol*. 240:616–626.

Koonin EV, Puigbò P, Wolf YI. 2011. Comparison of phylogenetic trees and search for a central trend in the "forest of life". *J Comput Biol*. 18:917–924.

Koonin EV, Wolf YI. 2009. The fundamental units, processes and patterns of evolution, and the tree of life conundrum. *Biol Direct*. 4:33.

Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*. 22:1208–1222.

Kosakovsky Pond SL, Frost SD, Grossman Z, Gravenor MB, Richman DD, Brown AJ. 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput Biol*. 2:e62.

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*. 28: 3033–3043.

Kosakovsky Pond SL, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*. 22:2375–2385.

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol*. 25:1809–1824.

Kosakovsky Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD. 2010. Evolutionary fingerprinting of genes. *Mol Biol Evol*. 27:520–536.

Krauss V, Thummler C, Georgi F, Lehmann J, Stadler PF, Eisenhardt C. 2008. Near intron positions are reliable phylogenetic markers: an application to holometabolous insects. *Mol Biol Evol*. 25:821–830.

Kuhn RM, Karolchik D, Zweig AS, et al. (22 co-authors). 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res*. 37:D755–D761.

Kumar S, Filipski A. 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res*. 17:127–135.

Kumar S, Nei M, Dudley J, Tamura K. 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform*. 9:299–306.

Läärä E. 2009. Statistics: reasoning on uncertainty, and the insignificance of testing null. *Ann Zool Fennici*. 46:138–157.

Lake JA, Rivera MC. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol Biol Evol*. 21:681–690.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. *J Mol Evol*. 20:86–93.

Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in bayesian phylogenetics. *Syst Biol*. 53:265–277.

Li J, Zhang Z, Vang S, Yu J, Wong GK, Wang J. 2009. Correlation between Ka/Ks and Ks is related to substitution model and evolutionary lineage. *J Mol Evol*. 68:414–423.

Lind PA, Berg OG, Andersson DI. 2010. Mutational robustness of ribosomal protein genes. *Science* 330:825–827.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol*. 19:1–7.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320:1632–1635.

Lunter G, Miklos I, Drummond A, Jensen J, Hein J. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* 6:83.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res*. 18:298–309.

Mar JC, Harlow TJ, Ragan MA. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evol Biol*. 5:8.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169:1753–1762.

Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T. 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics* 23:i319–i327.

Miller W, Rosenbloom K, Hardison RC, et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*. 17:1797–1808.

Misawa K, Nei M. 2003. Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. *J Mol Evol*. 57(Suppl 1):S290–S296.

Mossel E. 2003. On the impossibility of reconstructing ancestral data and phylogenies. *J Comput Biol*. 10:669–676.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11:715–724.

Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc*. 82:591–605.

Nam J, dePamphilis CW, Ma H, Nei M. 2003. Antiquity and evolution of the MADS-box gene family controlling flower development in plants. *Mol Biol Evol*. 20:1435–1447.

Naylor GJ, Brown WM. 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst Biol*. 47:61–76.

Negrisolo E, Kuhl H, Forcato C, Vitulo N, Reinhardt R, Patarnello T, Bargelloni L. 2010. Different phylogenomic approaches to resolve the evolutionary relationships among model fish species. *Mol Biol Evol*. 27:2757–2774.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet*. 11:265–289.

Nielsen R. 2005. Statistical methods in molecular evolution. New York: Springer.

Nielsen R, Bustamante C, Clark AG, et al. (13 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. 3:e170.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet*. 8:857–868.

Nielsen R, Huelsenbeck JP. 2002. Detecting positively selected amino acid sites using posterior predictive P-values. *Pac Symp Biocomput*. 7:576–588.

Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.

Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol*. 8:R199.

Notredame C. 2007. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*. 3:e123.

Novák Á, Miklós I, Lyngsø R, Hein J. 2008. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*. 24:2403–2404.

Nozawa M, Suzuki Y, Nei M. 2009a. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A*. 106:6700–6705.

Nozawa M, Suzuki Y, Nei M. 2009b. Response to Yang et al: problems with Bayesian methods of detecting positive selection at the DNA sequence level. *Proc Natl Acad Sci U S A*. 106.:E96.

Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. *Syst Biol*. 53:47–67.

Penny D, McComish BJ, Charleston MA, Hendy MD. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J Mol Evol*. 53:711–723.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9:e1000602.

Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol*. 6:498–505.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol*. 5:50.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol*. 24:1752–1760.

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Res*. 20:291–300.

Porter TM. 2008. Signifying little. *Science* 320:1292.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol*. 25:1253–1256.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol*. 25:1795–1808.

Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc R Soc Lond B Biol Sci*. 269:1313–1316.

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

Rannala B, Yang Z. 2008. Phylogenetic inference using whole genomes. *Annu Rev Genomics Hum Genet*. 9:217–231.

Rao CR. 1989. Statistics and truth: putting chance to work. New Delhi (India): Council of Scientific & Industrial Research.

Rasmussen MD, Kellis M. 2010. A Bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol*. 28:273–290.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc B Biol Sci*. 365:2571–2580.

Redelings BD, Suchard MA. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst Biol*. 54:401–418.

Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol*. 24:1821–1831.

Ripplinger J, Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. *Mol Biol Evol*. 27:2790–2803.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*. 107:4629–4634.

Rokas A, Chatzimanolis S. 2008. From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. *Methods Mol Biol*. 422:1–12.

Rokas A, King N, Finnerty J, Carroll SB. 2003. Conflicting phylogenetic signals at the base of the metazoan tree. *Evol Dev*. 5:346–359.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.

Ronquist F, Deans AR. 2009. Bayesian phylogenetics and its influence on insect systematics. *Annu Rev Entomol*. 55:189–206.

Rosenberg MS. 2009. Sequence alignment: methods, models, concepts, and strategies. Berkeley (CA): University of California Press.

Rzhetsky A, Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol*. 10:1073–1095.

Saitou N, Nei M. 1987. The neighbor-joining method—a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406–425.

Sawyer SL, Emerman M, Malik HS. 2004. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol*. 2:E275.

Seo TK. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol*. 25:960–971.

Seoighe C, Ketwaroo F, Pillay V, et al. (11 co-authors). 2007. A model of directional selection applied to the evolution of drug resistance in HIV-1. *Mol Biol Evol*. 24:1025–1031.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol*. 23:7–9.

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

Shoemaker JS, Fitch WM. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol Biol Evol*. 6:270–289.

Soltis DE, Albert VA, Savolainen V, et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. *Trends Plant Sci*. 9:477–483.

Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Syst Entomol*. 35:429–448.

Springer MS, DeBry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. *Mol Biol Evol*. 18:132–143.

Springer MS, Stanhope MJ, Madsen O, de Jong WW. 2004. Molecules consolidate the placental mammal tree. *Trends Ecol Evol*. 19:430–438.

Strimmer K. 2008. A unified approach to false discovery rate estimation. *BMC Bioinformatics* 9:303.

Subramanian S, Kumar S. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res*. 13:838–844.

Subramanian S, Kumar S. 2006. Higher intensity of purifying selection on $>$90% of the human genes revealed by the intrinsic replacement mutation rates. *Mol Biol Evol*. 23:2283–2287.

Susko E. 2008. On the distributions of bootstrap support and posterior distributions for a star tree. *Syst Biol*. 57:602–612.

Suzuki Y, Glazko GV, Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci U S A*. 99:16138–16143.

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol*. 16:1315–1328.

Suzuki Y, Gojobori T, Kumar S. 2009. Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. *Mol Biol Evol*. 26:2275–2284.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. Sunderland (MA): Sinauer Associates. p. 407–514.

Takezaki N, Gojobori T. 1999. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. *Mol Biol Evol*. 16:590–601.

Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol*. 19:1727–1736.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A*. 101:11030–11035.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 28:2731–2739.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol*. 21:36–44.

Tavare S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. In: Robert M. Miura, editor. Lectures in mathematics in the life sciences. Providence (RI): American Mathematical Society. p. 57–86.

Telford MJ, Bourlat SJ, Economou A, Papillon D, Rota-Stabelli O. 2008. The evolution of the Ecdysozoa. *Philos Trans R Soc Lond B Biol Sci*. 363:1529–1537.

Venkatesh B, Erdmann MV, Brenner S. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc Natl Acad Sci U S A*. 98:11382–11387.

Vos M. 2011. A species concept for bacteria based on adaptive divergence. *Trends Microbiol*. 19:1–7.

Waddell PJ, Kishino H, Ota R. 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. *Genome Inform*. 13:82–92.

Wahlberg N, Weingartner E, Warren A, Nylin S. 2009. Timing major conflict between mitochondrial and nuclear genes in species relationships of Polygonia butterflies (Nymphalidae: Nymphalini). *BMC Evol Biol*. 9:92.

White WT, Hills SF, Gaddam R, Holland BR, Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. *Mol Biol Evol*. 24:2029–2039.

Wiens JJ. 2006. Missing data and the design of phylogenetic analyses. *J Biomed Inform*. 39:34–42.

Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, Guindon S, Gascuel O, Grossman LI, Romero R, Goodman M. 2007. Genomics, biogeography, and the diversification of placental mammals. *Proc Natl Acad Sci U S A*. 104:14395–14400.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051.

Wrobel B. 2008. Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. *J Appl Genet*. 49:49–67.

Wyckoff GJ, Malcom CM, Vallender EJ, Lahn BT. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet*. 21:381–385.

Xia X. 2006. Topological bias in distance-based phylogenetic methods: problems with over- and underestimated genetic distances. *Evol Bioinform Online*. 2:377–389.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol*. 11:367–372.

Yang Z. 2005. Bayesian inference in molecular phylogenetics. In: Gascuel O, editor. Mathematics of evolution and phylogeny. Oxford: Oxford University Press. p. 63–90.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 2008. Empirical evaluation of a prior for Bayesian phylogenetic inference. *Philos Trans R Soc B Biol Sci*. 363: 4031–4039.

Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol*. 28:1217–1228.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*. 25:568–579.

Yang Z, Nielsen R, Goldman N. 2009. In defense of statistical methods for detecting positive selection. *Proc Natl Acad Sci U S A*. 106:E95; author reply E96.

Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst Biol*. 54:455–470.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22:1107–1118.

Yokoyama S, Tada T, Zhang H, Britt L. 2008. Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. *Proc Natl Acad Sci U S A*. 105:13480–13485.

Zardoya R, Suarez M. 2008. Sequencing and phylogenomic analysis of whole mitochondrial genomes of animals. *Methods Mol Biol*. 422:185–200.

Zhang J, Nei M. 1996. Evolution of Antennapedia-class homeobox genes. *Genetics* 142:295–303.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A*. 95:3708–3713.

Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol Biol Evol*. 27:1912–1922.

Ziliak ST, McCloskey DN. 2008. The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor (MI): University of Michigan Press.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p. 97–166.