A Genome-Scale Investigation of Incongruence in Culicidae Mosquitoes

Yuyu Wang[1,2], Xiaofan Zhou[2], Ding Yang[1,*], and Antonis Rokas[2,*]

[1]Department of Entomology, China Agricultural University, Beijing, China

[2]Department of Biological Sciences, Vanderbilt University

*Corresponding author: E-mail: dyangcau@126.com; antonis.rokas@vanderbilt.edu.

Accepted: November 19, 2015

Abstract

Comparison of individual gene trees in several recent phylogenomic studies from diverse lineages has revealed a surprising amount of topological conflict or incongruence, but we still know relatively little about its distribution across the tree of life. To further our understanding of incongruence, the factors that contribute to it and how it can be ameliorated, we examined its distribution in a clade of 20 Culicidae mosquito species through the reconstruction and analysis of the phylogenetic histories of 2,007 groups of orthologous genes. Levels of incongruence were generally low, the three exceptions being the internodes concerned with the branching of *Anopheles christyi*, with the branching of the subgenus *Anopheles* as well as the already reported incongruence within the *Anopheles gambiae* species complex. Two of these incongruence events (*A. gambiae* species complex and *A. christyi*) are likely due to biological factors, whereas the third (subgenus *Anopheles*) is likely due to analytical factors. Similar to previous studies, the use of genes or internodes with high bootstrap support or internode certainty values, both of which were positively correlated with gene alignment length, substantially reduced the observed incongruence. However, the clade support values of the internodes concerned with the branching of the subgenus *Anopheles* as well as within the *A. gambiae* species complex remained very low. Based on these results, we infer that the prevalence of incongruence in Culicidae mosquitoes is generally low, that it likely stems from both analytical and biological factors, and that it can be ameliorated through the selection of genes with strong phylogenetic signal. More generally, selection of genes with strong phylogenetic signal may be a general empirical solution for reducing incongruence and increasing the robustness of inference in phylogenomic studies.

Key words: maximum likelihood, gene tree, bootstrap support (BS), bipartition, internode certainty (IC).

Recent advances in DNA sequencing technologies provide great opportunities for using genome-scale data to reconstruct phylogenetic history (Rokas and Abbot 2009; Hittinger et al. 2010; Faircloth et al. 2012; Lemmon et al. 2012). However, recent phylogenomic studies in diverse taxonomic groups, including plants (Zhong et al. 2013; Wickett et al. 2014), fungi (Hess and Goldman 2011; Salichos and Rokas 2013), and animals (Song et al. 2012; Jarvis et al. 2014), have shown that a large number of individual gene trees are topologically incongruent with each other. For example, a recent analysis of 1,070 orthologs from 23 yeast genomes identified 1,070 distinct gene trees, which were all incongruent with the phylogeny inferred from concatenation analysis (Salichos and Rokas 2013). Surprisingly, nearly half the internodes of the yeast phylogeny exhibited very low internode certainty (IC) values (Salichos and Rokas 2013), a measure of topological conflict (Salichos and Rokas 2013; Salichos et al. 2014). Similarly, the analysis of 32 plant taxa found 182 distinct topologies in a set of 184 gene trees (Zhong et al. 2013), and the analysis of 447 nuclear genes from 37 mammal species revealed 440 distinct topologies (Song et al. 2012).

Incongruence between gene trees can stem from analytical or biological factors. A wide variety of analytical factors can lead to failure to accurately infer a gene tree; these can be either due to stochastic error (e.g., insufficient sequence length or taxon samples) or due to systematic error (in case of departure from model assumptions, Jeffroy et al. 2006). In contrast, a number of biological factors can lead to gene trees that are actually distinct from each other and from the species tree. Examples of biological factors include incomplete lineage sorting (ILS), hidden paralogy, horizontal gene transfer, as well as gene duplication and loss, recombination and natural selection (Galtier and Daubin 2008; Degnan and Rosenberg 2009; Fontaine et al. 2015). Although gene tree incongruence

caused by analytical factors can be potentially reduced by some data filtering approaches, such as using genes with high phylogenetic information content (Dell'Ampio et al. 2014), slowly evolving genes (Betancur-R et al. 2014), genes with stationary base composition (Romiguier et al. 2013), genes with strong phylogenetic signals (Salichos and Rokas 2013), as well as internode-specific genes (Chen et al. 2015), incongruence stemming from biological factors cannot (Fontaine et al. 2015; Nater et al. 2015; Suh et al. 2015).

Although conflict between gene trees has been reported in analyses of phylogenomic data matrices from diverse plant, fungal, and animal taxa, we still know relatively little about the distribution of incongruence across the tree of life. Mosquitoes in the genus *Anopheles* represent an excellent lineage for investigating incongruence for two reasons. First, the draft genomes of 16 anophelines from Africa, Asia, Europe, and South America, representing a variety of geographic locations and ecological conditions and a range of evolutionary distances from each other, were recently sequenced (Neafsey et al. 2015). In addition to these 16 newly sequenced *Anopheles* genomes, the genomes of two additional *Anopheles* species, namely *A. gambiae* (Holt et al. 2002) and *Anopheles darlingi* (Marinotti et al. 2013), as well as the genomes of two other species belonging to the subfamily Culicinae, namely *Aedes aegypti* (Nene et al. 2007) and *Culex quinquefasciatus* (Arensburger et al. 2010), are also available.

The second reason is the demonstrated presence of incongruence in the *Anopheles* phylogeny, particularly within the *Anopheles gambiae* species complex (Besansky et al. 1994; Hittinger et al. 2010; Fontaine et al. 2015). Specifically, a genome-wide investigation of the relationships between the five species belonging to the *A. gambiae* complex, namely *A. gambiae*, *Anopheles arabiensis*, *Anopheles quadriannulatus*, *Anopheles melas*, and *Anopheles merus*, reported extensive introgression (Fontaine et al. 2015), prompting the authors of an associated commentary to ponder whether the notion of a bifurcating species phylogeny is a meaningful way to describe the evolutionary relationships among species in the complex (Clark and Messer 2015). Remarkably, it appears that the topology inferred from concatenation analysis, albeit strongly supported, is likely incorrect (Fontaine et al. 2015). This very high degree of incongruence raises the question on whether it is localized between species in the *A. gambiae* complex or whether it is also present in other parts of the *Anopheles* phylogeny.

Low Levels of Incongruence in Culicidae Phylogeny

In this study, we assembled a data set of 2,007 groups of orthologous genes (henceforth referred to simply as genes) from 20 Culicidae mosquito genomes (table 1). Maximum likelihood (ML) concatenation analysis of the 2,007-gene data matrix produced a species phylogeny in which all internodes exhibited 100% bootstrap support (BS) (fig. 1). Summarizing the 2,007 gene trees into an extended Majority Rule Consensus (eMRC) phylogeny or using them as input to construct a coalescent-based species phylogeny resulted in topologies that were identical to the concatenation phylogeny (fig. 1). Interestingly, 12 out of 17 internodes in the eMRC phylogeny had a gene-support frequency (GSF) of greater than 80%. Two of the remaining five internodes are associated with the branchings of *Anopheleschristyi* (GSF = 53%) and subgenus *Anopheles* (GSF = 62%), respectively, whereas the other three internodes show very low GSF values (33–43%) and all reside within the *A. gambiae* complex (fig. 1).

One thousand one hundred twenty-six of the 2,007 gene trees are unique, which means that about half of the gene trees do not agree (by at least one internode) with each other, or with species phylogeny supported by concatenation, eMRC, and coalescent-based approaches. The average normalized Robinson–Foulds (Robinson and Foulds 1981) tree distance between the 2,007 gene trees and the species phylogeny (0.21) was lower than that generated by an all-pairs comparison between the 2,007 gene trees (0.29) (fig. 2).

To quantify incongruence, we used IC which evaluates support for a given internode according its frequency in a given set of trees jointly with that the most prevalent conflicting bipartitions in the same set of trees (Salichos and Rokas 2013; Salichos et al. 2014). Examination of the eMRC phylogeny showed that 10 out of 17 internodes had IC values equal or greater than 0.70 and another two values greater than 0.55. The remaining five internodes, namely the branchings of *A. christyi* and subgenus *Anopheles* as well as the internodes within the *A. gambiae species complex*, had IC values less than 0.25. The branching of *A. christyi* depicts the bipartition (*A. gambiae* complex, *A. christyi*) (GSF = 53, IC = 0.08; fig. 1), which conflicts with the bipartition ([*A. christyi*, *A. epiroticus*],[remaining 18 species]), whose GSF is 27, yielding an IC value of 0.08. The branching of subgenus *Anopheles* depicts the bipartition (subgenus *Cellia*, subgenus *Anopheles*) (GSF = 62, IC = 0.19; fig. 1), which conflicts with the bipartition ([*Anopheles albimanus*, *A. darlingi*, *Anopheles atroparvus*, *Anopheles sinensis*],[remaining 16 species]), whose GSF is 21, yielding an IC value of 0.19.

The incongruence observed in internodes within the *A. gambiae* complex is much higher than the rest of the Culicidae phylogeny. The average GSF in the three internodes within the complex is 37.33, a value much smaller than the average GSF of 87.86 that is observed in the rest of the Culicidae phylogeny (fig. 1 and table 1). Similarly, the IC values of internodes in the *A. gambiae* complex rank first, third, and fifth lowest among the 17 internodes in the Culicidae phylogeny (fig. 1).

Table 1

The Effect of Using Genes and Bipartitions with Strong Phylogenetic Signal on the Culicidae Phylogeny

Treatment	Treatment Details	Average GSF	TC	RTC	Number of Internodes with Increased GSF	Number of Internodes with Decreased GSF	Number of Internodes with Increased IC	Number of Internodes with Decreased IC
Default analysis	Default analysis	87.86	10.80	0.64	NA	NA	NA	NA
Selection of genes whose ML trees have high average BS	Genes with average BS $\geq$ 70% (1,818 genes)	90.07	11.11	0.65	5	0	5	0
	Genes with average BS $\geq$ 80% (1,379 genes)	92.14	11.70	0.69	11	0	10	0
	Genes with average BS $\geq$ 90% (378 genes)	95.29	12.77	0.75	13	0	14	0
	Genes with average BS $\geq$ 95% (66 genes)	96.43	13.02	0.77	15	0	14	1
Selection of genes whose ML trees have high TC	Using only the 1,818 genes with the highest TC	89.93	11.14	0.66	4	0	5	0
	Using only the 1,379 genes with the highest TC	92.07	11.68	0.69	11	0	9	0
	Using only the 378 genes with the highest TC	95.64	12.81	0.75	14	0	13	1
	Using only the 66 genes with the highest TC	96.21	12.91	0.76	15	0	14	1
Selection of bipartitions with high BS in the ML trees of genes	Using only bipartitions that have $\geq$ 70% BS	NA	12.40	0.73	NA	NA	14	0
	Using only bipartitions that have $\geq$ 80% BS	NA	12.88	0.76	NA	NA	14	0
	Using only bipartitions that have $\geq$ 90% BS	NA	13.24	0.78	NA	NA	13	0
	Using only bipartitions that have $\geq$ 95% BS	NA	13.34	0.78	NA	NA	13	0

NOTE.—The columns correspond to: the specific filtering of genes or bipartitions with strong phylogenetic signal tested (treatment and treatment details), the average GSF of the internodes of the Culicidae eMRC phylogeny (average GSF), the TC of the Culicidae eMRC phylogeny, the RTC of the Culicidae eMRC phylogeny, the numbers of internodes of the Culicidae eMRC phylogeny in which GSF increases or decreases by more than 3%, and the numbers of internodes of the Culicidae eMRC phylogeny in which IC increases or decreases by more than 0.03. As the maximum value of IC for a given internode is 1, the maximum value of TC for a given phylogeny is the number of internodes, which in this case is 17. In the analyses concerned with the use of bipartitions, only those bipartitions that displayed BS greater or equal to 70%, 80%, 90%, or 95% in the ML trees of the 2,007 genes were used to construct eMRC phylogenies, which were then compared with the default analysis. NA, not applicable.

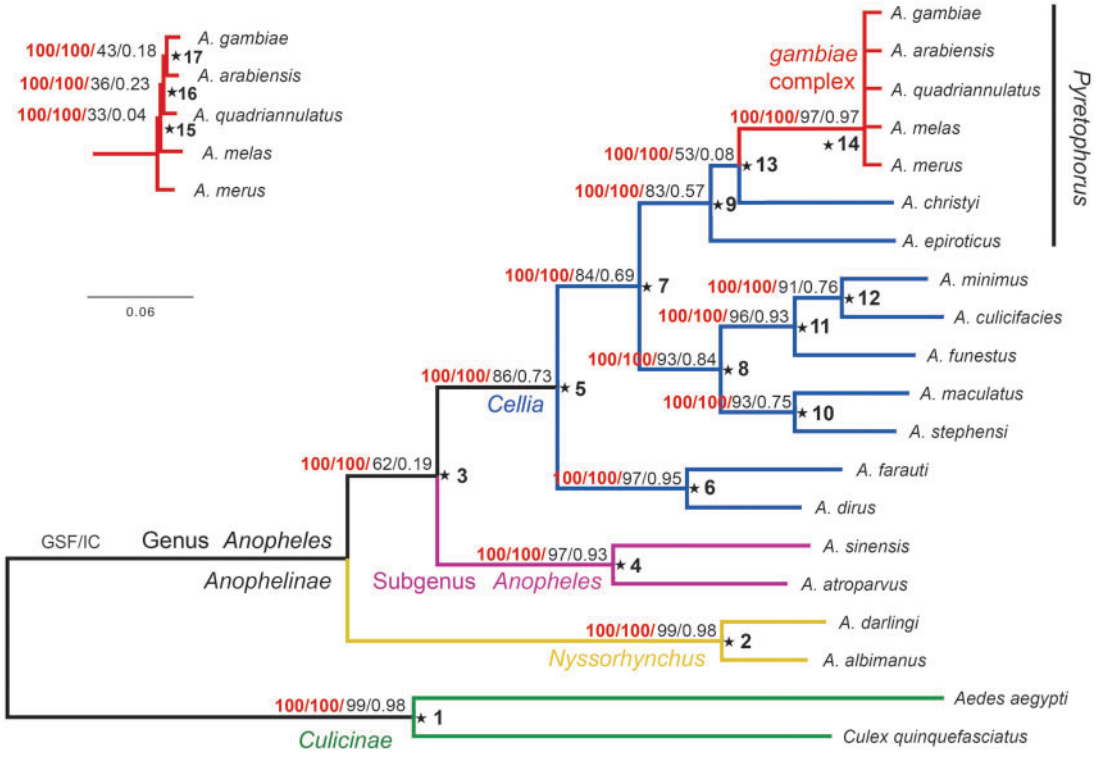

Fig. 1.—The Culicidae species phylogeny recovered from the concatenation analysis of 2,007 genes using ML. Asterisks denote internodes that received 100% BS by the concatenation analysis. The same topology is also recovered by the eMRC phylogeny as well as by the coalescent phylogeny (ASTRAL and STAR) of the 2,007 individual gene trees. Black values near internodes correspond to GSF and IC, respectively. Bold red values correspond to support values of the coalescent phylogeny based on ASTRAL and STAR, respectively. The scale bar is in units of nucleotide substitutions per site.

To measure the degrees of conflict for every internode, IC can be more informative than GSF (Salichos and Rokas 2013; Salichos et al. 2014). For example, the placement of *Anopheles stephensi* and the placement of *Anopheles funestus* received 93% and 96% GSF, whereas their ICs were 0.75 and 0.93, respectively. This difference in the IC values of the two internodes despite similar GSF values is a result of the secondary conflicting signal difference. Specifically, whereas the most prevalent conflicting bipartition to the placement of *A. stephensi* has a GSF of 4%, the most prevalent conflict to the placement of *A. funestus* has a GSF of only 1%.

Using Genes with Strong Phylogenetic Signal Reduces Incongruence

To test whether using genes with stronger phylogenetic signal can reduce incongruence, we analyzed four data sets comprising genes whose ML trees had average BS values across all internodes greater than or equal to 70% (1,818 genes), 80% (1,379 genes), 90% (378 genes), or 95% (66 genes), and four data sets comprising the 1,818, 1,379, 378, or 66 genes whose ML trees had the highest tree certainty (TC) values. Note that gene selection was solely based on the strength of phylogenetic signal exhibited in their gene trees (measured by BS or TC) without any consideration to the topology supported. The concatenation analysis as well as the eMRC analysis was redone each time when the new data set was selected. We found that the GSF and IC values of the vast majority of internodes increased as the stringency of the BS and TC filters increased (supplementary table S2, Supplementary Material online), suggesting that selecting genes with high average BS or high TC significantly reduced incongruence in the Culicidae phylogeny (table 1).

We also tested whether using internodes with high BS can reduce the incongruence by extracting bipartitions with BS values greater than or equal to 70%, 80%, 90% or 95% from every ML tree of the 2,007 genes and then used them to construct the eMRC phylogenies. Importantly, the use of highly supported bipartitions allows one to quantify a given internode's IC from only the subset of bipartitions that highly support or conflict with that internode (Salichos and Rokas 2013; Salichos et al. 2014). Compared to the phylogeny of figure 1, this practice significantly increased IC values for ≥ 13 internodes (table 1 and supplementary table S2, Supplementary Material online).

Even though the GSF and IC values of the internodes concerned with the branching of the subgenus *Anopheles*, with

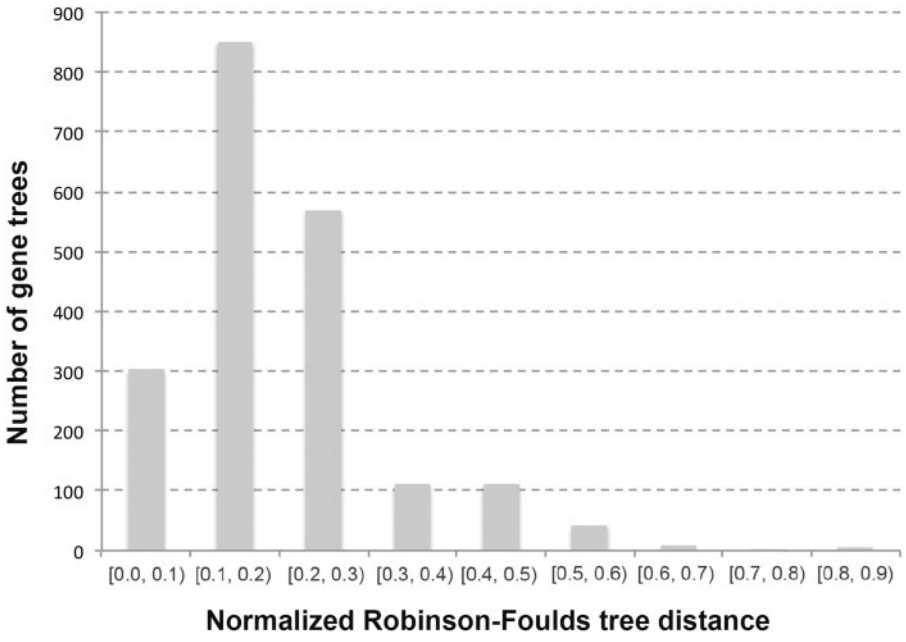

Fig. 2.—The distribution of the agreement between the bipartitions present in the 2,007 individual gene trees and the concatenation phylogeny, measured using the normalized Robinson–Foulds tree distance. The phylogeny of the 20 Culicidae species analyzed in this study is unrooted and contains 17 nontrivial bipartitions.

the branchings within the *A. gambiae* species complex, and with the branching of *A. christyi*, increased as the stringency of the BS and TC filters increased, the values themselves were still lower even for the strictest filters (supplementary table S2, Supplementary Material online). Thus, incongruence in these internodes may be the result of biological factors such as ILS, gene duplication and loss, or introgression (Fontaine et al. 2015).

The Relationship between Incongruence and Gene Alignment Length

Both average BS as well as TC values are positively correlated with genes' alignment lengths (alignment length vs. BS, $r = 0.50$, P value $< 2.2e\text{-}16$; alignment length vs. TC, $r = 0.56$, P value $< 2.2e\text{-}16$; supplementary fig. S1, Supplementary Material online). How does incongruence behave if we use genes with the same lengths? To resolve this question, we created a new data matrix that contained only the first 999 bp of the sequence alignment of the 1,340 genes that were 999 bp or longer (genes with shorter alignment lengths were excluded) and re-analyzed levels of incongruence in the Culicidae phylogeny. The results are quite similar to the results from the 2,007-gene data matrix; 12 of the 17 internodes exhibit high GSF and IC values, whereas internodes within the *A. gambiae* species complex as well as internodes associated with the placement of *A. christyi* and the subgenus *Anopheles* show low GSF and very low IC values (fig. 3). Thus, although the average BS and TC values were positively correlated with gene alignment length, using loci that have the same alignment lengths does not appear to substantially decrease or increase the incongruence present in this phylogenomic data matrix.

We also examined whether the selection of genes or bipartitions with strong phylogenetic signal in this set of 1,340 alignment length-standardized genes reduced incongruence. We tested three data sets comprising genes whose ML trees showed average BS values across all internodes that were greater than or equal to 70% (1,138 genes), 80% (603 genes) or 90% (45 genes) (no gene had average BS greater or equal to 95%), and three data sets comprising the 1,138, 603 or 45 genes whose trees had the highest TC. Almost all the GSF and IC values of every internode increased as the value of the BS or TC filter increased (table 2 and supplementary table S3, Supplementary Material online). Using genes or internodes with high BS or IC values also significantly reduced the observed incongruence (table 2 and supplementary table S3, Supplementary Material online). Similarly, selecting internodes with high BS decreased incongruence by extracting only those bipartitions that display BS values greater than or equal to 70%, 80%, 90% or 95% from every one of the 1,340 genes' ML trees. This practice significantly increased IC values for $\geq$14 internodes relative to the phylogeny of figure 3 (table 2 and supplementary table S3, Supplementary Material online). However, the GSF and IC values of the internodes concerned with the branching of the subgenus *Anopheles*, the branching of *A. christyi*, as

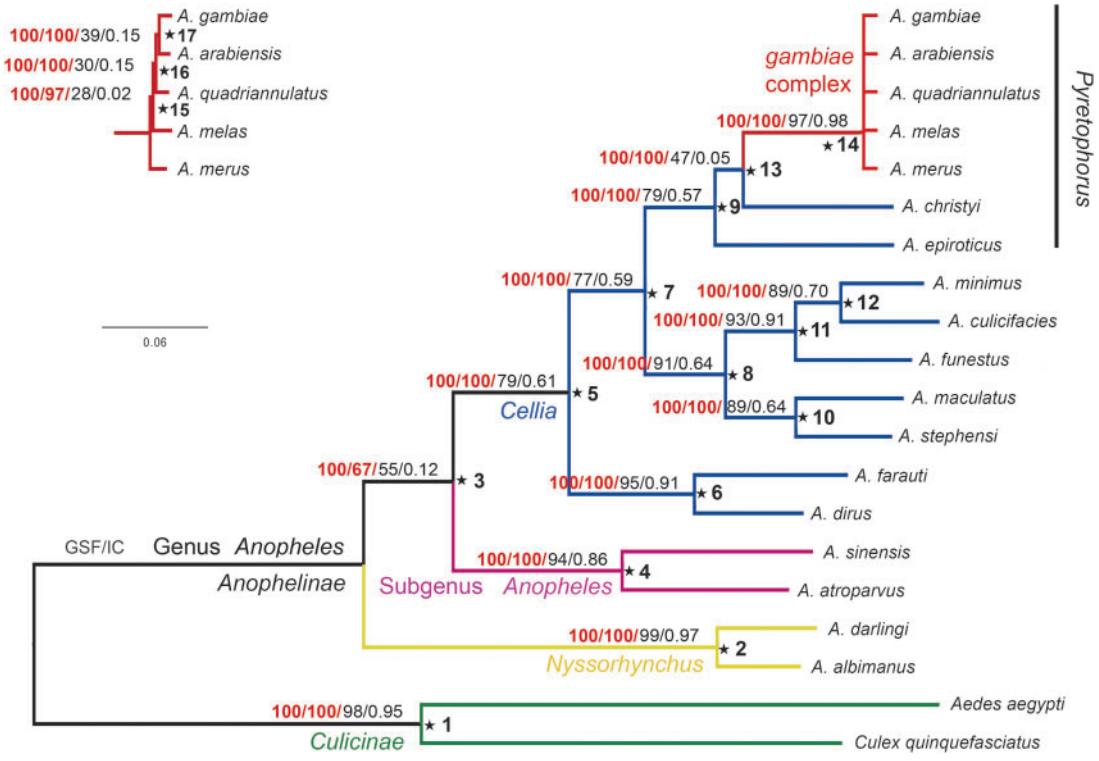

Fig. 3.—The Culicidae species phylogeny recovered from the concatenation analysis of first 999 bp sequence of 1,340 genes using ML. Asterisks denote internodes that received 100% BS by the concatenation analysis. The same topology is also recovered by the eMRC phylogeny as well as the coalescent phylogeny (ASTRAL and STAR) of the 1,340 individual gene trees. Black values near internodes correspond to GSF and IC, respectively. Bold red values correspond to support values of the coalescent phylogeny based on ASTRAL and STAR, respectively. The scale bar is in units of nucleotide substitutions per site.

well as with branchings within the *A. gambiae* species complex, were still lower even for the strictest filters.

Conclusion

In summary, analyses of a 2,007-gene phylogenomic data matrix from 20 Culicidae mosquito genomes showed that incongruence was low and localized to specific branches. Similar to previous studies, the use of genes or bipartitions with strong phylogenetic signal (quantified either through the use of BS or IC values) substantially reduced the observed incongruence. However, the GSF and IC values of the internodes concerned with the branchings of *A. christyi*, the subgenus *Anopheles*, as well as with branchings within the *A. gambiae* species complex remained very low. Combined with the observation that many of the *A. gambiae* species complex internodes are very short, the observed incongruence is consistent with previous inferences of extensive introgression within the *A. gambiae* species complex (Fontaine et al. 2015). Short internode length makes biological factors, such as ILS or introgression, the most likely explanation for the incongruence observed in the branching of *A. christyi*. In contrast, the internode associated with the branching of the subgenus *Anopheles* and the subgenus *Celia* is much longer suggesting that this incongruence is more likely to be due to analytical factors. Very similar results were obtained with a 1,340-gene phylogenomic data matrix in which all genes had the same length, arguing that the well-known correlation between alignment length and phylogenetic signal did not have a major influence on phylogenetic reconstruction in this lineage. What's more, they add to the body of evidence (Salichos and Rokas 2013) showing that the selection of genes with strong phylogenetic signal can reduce incongruence and increase the robustness of phylogenetic inference. Thus, this strategy may be a general empirical solution for ameliorating incongruence in phylogenomic studies.

Materials and Methods

Data Matrix Construction

We used the complete sets of annotated orthology data of 20 Culicidae mosquito genomes (supplementary table S1, Supplementary Material online) from http://cegg.unige.ch/orthodbmoz2 (Neafsey et al. 2015). We selected 2,008 single-copy genes which that contained sequences from all 20 species as our initial data set. The nucleotide sequences

Table 2
The Effect of Using Genes and Bipartitions with Strong Phylogenetic Signal on the Culicidae Phylogeny Based on the First 999 bp of Every Gene's Alignment

Treatment	Treatment Details	Average GSF	TC	RTC	Number of Internodes with Increased GSF	Number of Internodes with Decreased GSF	Number of Internodes with Increased IC	Number of Internodes with Decreased IC
1,340 genes	First 999 bp sequence of every gene (1,340 genes)	84.43	9.98	0.59	NA	NA	NA	NA
Selection of genes whose	Genes with average BS ≥ 70% (1,138 genes)	87.36	10.34	0.61	8	0	5	0
ML trees have	Genes with average BS ≥ 80% (603 genes)	91.00	11.17	0.66	14	0	10	1
high average BS	Genes with average BS ≥ 90% (45 genes)	95.43	12.25	0.72	14	0	13	2
Selection of genes whose ML	Using only the 1,138 genes with the highest TC	87.29	10.34	0.61	10	0	7	0
trees have high TC	Using only the 603 genes with the highest TC	90.71	11.16	0.66	13	0	10	0
	Using only the 45 genes with the highest TC	94.86	12.23	0.72	14	0	13	1
Selection of bipartitions with	Using only bipartitions that have ≥ 70% BS	NA	11.82	0.70	NA	NA	14	0
high BS in the ML trees	Using only bipartitions that have ≥ 80% BS	NA	12.53	0.74	NA	NA	15	0
of genes	Using only bipartitions that have ≥ 90% BS	NA	13.10	0.77	NA	NA	15	0
	Using only bipartitions that have ≥ 95% BS	NA	13.33	0.78	NA	NA	15	0

NOTE.—The columns correspond to: the specific filtering of genes or bipartitions with strong phylogenetic signal tested (treatment and treatment details), the average GSF of the internodes of the Culicidae eMRC phylogeny (average GSF), the TC of the Culicidae eMRC phylogeny, the RTC of the Culicidae eMRC phylogeny, the numbers of internodes of the Culicidae eMRC phylogeny in which GSF increases or decreases by more than 3%, and the numbers of internodes of the Culicidae eMRC phylogeny in which IC increases or decreases by more than 0.03. As the maximum value of IC for a given internode is 1, the maximum value of TC for a given phylogeny is the number of internodes, which in this case is 17. In the analyses concerned with the use of bipartitions, only those bipartitions that displayed BS greater or equal to 70%, 80%, 90%, or 95% in the ML trees of the 1,340 genes were used to construct eMRC phylogenies, which were then compared with the default analysis. NA, not applicable.

of all genes were translated to amino acids. A series of different data sets was constructed using custom Perl scripts.

Gene Alignment

We aligned all genes using the MAFFT software, version 7.182 (Katoh and Toh 2008) based on their amino acid sequence, using the default settings (automatic selection of the appropriate strategy, from L-INS-i, FFT-NS-i, and FFT-NS-2, according to data size). Then, we used PAL2NAL (Suyama et al. 2006) to translate amino acid sequence alignments to codon sequence alignments, and the "automated" option of trimAl (Capella-Gutierrez et al. 2009) to trim the amino acid sequence alignments. Trimmed segments of the amino acid sequence alignments were deleted from their corresponding codon sequence alignments using custom Perl scripts. Following trimming, our data matrix consisted of 2,007 genes from 20 species and contained no missing data.

To test how incongruence varied independent of gene alignment length, we also generated a data matrix that was comprised of only the first 999 bp of sequence from each gene. Since there were 667 genes shorter than 999 bp, this data matrix contains 1,340 genes from 20 species, every one of which has a 999 bp long alignment and does not contain any missing data.

Gene Tree Inference

For the codon sequence alignment of each gene, the un-rooted phylogenetic tree under the optimality criterion of ML was inferred using RAxML, version 8.0.20 (Stamatakis 2014), under the GTRGAMMA model and with the values of the nucleotide base frequencies fixed to "observed" and those of the substitution rate parameters estimated from the data (raxmlHPC-PTHREADS-SSE3 -T 8 -f a -x 12345 -p 12345 -N 100 -m GTRGAMMA -s ALIGNMENT -n NAME). For the concatenation analysis, codon sequence alignments from all genes were analyzed as a single super-matrix. The un-rooted concatenation species phylogeny was inferred through a single ML search under the GTRGAMMA model in RAxML, version 8.0.20 (Stamatakis 2014), with the values of the nucleotide base frequencies fixed to "observed" and those of the substitution rate parameters estimated from the data (raxmlHPC-PTHREADS-SSE3 -T 8 -f a -x 12345 -p 12345 -N 100 -m GTRGAMMA -s ALIGNMENT -n NAME). In all cases, robustness in inference was assessed via bootstrap resampling (100 replicates). Note that the RAxML software first infers the topologies for each of the bootstrap replicates and then searches for the best-scoring ML tree using every fifth bootstrap replicate tree as a starting tree.

The un-rooted eMRC phylogeny that consisted of those bipartitions that appear in more than half of the ML gene trees, as well as of additional compatible bipartitions that appear in less than half of the gene trees, was inferred from the CONSENSE program in the Phylogeny Inference Package,

version 3.696 (PHYLIP; J. Felsenstein, University of Washington, Seattle; http://evolution.genetics.washington.edu/phylip.html). The eMRC phylogeny of bipartitions with high BS was constructed using custom Perl scripts. As the divergence of Culicinae and Anophelinae lineages is well established, all phylogenies shown in figures have been midpoint rooted at the internode that separates these two lineages for easier visualization.

The coalescent species phylogeny was estimated using 100 replicates of multi-locus bootstrapping in ASTRAL (Mirarab and Warnow 2015) (java -Xmx36000M -jar astral.4.7.8.jar -i TREECOLLECTION -o OUTPUT -b BS_PATH -r 100), and using the online version of the STAR software with 100 rooted bootstrap replicates trees of every gene (Liu et al. 2009; http://bioinformatics.publichealth.uga.edu/SpeciesTreeAnalysis/STAR/STAR.php).

Tree Distance Estimation

Distances between trees were estimated using the normalized Robinson–Foulds tree distance (Robinson and Foulds 1981) as calculated by RAxML, version 8.0.20 (Stamatakis 2014) (raxmlHPC-PTHREADS-SSE3 -T 8 -f r -z TREECOLLECTION -m GTRGAMMA -n NAME).

Evaluation of Incongruence

IC, TC, and relative TC (RTC) (Salichos and Rokas 2013; Salichos et al. 2014) were calculated using RAxML, version 8.0.20 (Stamatakis 2014) (raxmlHPC-PTHREADS-SSE3 -T 8 -f i -t REFERENCETREE -z TREECOLLECTION -m PROTGAMM AAUTO -n NAME).

We used these average BS and TC values to construct eight subsets of 2,007 orthogroups: four with genes that have average BS values greater than or equal to 70% (1,818 genes), 80% (1,379 genes), 90% (378 genes), or 95% (66 genes), respectively, as well as four data sets comprising the 1,818, 1,379, 378, or 66 genes whose ML trees have the highest TC values, respectively.

For every gene from the 2,007-gene data matrix, we also extracted all bipartitions from its ML tree that have BS values greater than or equal to 70%, 80%, 90%, and 95%, respectively. We then used each one of these four sets of highly supported bipartitions to construct eMRC species phylogenies with custom Perl scripts.

For the 1,340-gene data matrix, in which each gene's alignment was 999 bp long, we constructed six subsets: three with genes that have average BS values that are greater than or equal to 70% (1,138 genes), 80% (603 genes), or 90% (45 genes), respectively, (no gene had average BS values greater than or equal to 95%); as well as three data sets comprising the 1,138, 603, or 45 genes whose ML trees have the highest TC values, respectively.

For every gene from the 1,340-gene data matrix, we also extracted all bipartitions from its ML tree that have BS values greater than or equal to 70%, 80%, 90%, and 95%. We then used each one of these four sets of highly supported bipartitions to construct eMRC species phylogenies with custom Perl scripts.

Data Availability

All data and analyses described in this study are deposited at Figshare under the accession 10.6084/m9.figshare.1566851.

Supplementary Material

Supplementary figure S1 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This work was partially supported by the China Scholarship Council fellowship program (No. 201406350100 to Y.W.), the National Natural Science Foundation of China (No. 31320103902 to D.Y.), the National Institutes of Health (NIAID, AI105619 to A.R.), and the National Science Foundation (DEB-0844968 and DEB-1442113 to A.R.).

Literature Cited

Arensburger P, et al. 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science 330:86–88.

Besansky NJ, et al. 1994. Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors. Proc Natl Acad Sci U S A. 91:6885–6888.

Betancur-R R, Naylor G, Ortí G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst Biol. 63:257–262.

Capella-Gutierrez S, Silla-Martinez J, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Chen M, Liang D, Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst Biol. 64:1104–1120.

Clark AG, Messer PW. 2015. Conundrum of jumbled mosquito genomes. Science 347:27–28.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 24:332–340.

Dell'Ampio E, et al. 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol Biol Evol. 31:239–249.

Faircloth BC, et al. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol. 61:717–726.

Fontaine MC, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc B Biol Sci. 363:4023–4029.

Hess J, Goldman N. 2011. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. PLoS ONE 6:e22783.
Hittinger CT, Johnston M, Tossberg JT, Rokas A. 2010. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc Natl Acad Sci U S A. 107:1476–1481.
Holt RA, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298:129–149.
Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231.
Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 9:286–298.
Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol. 61:727–744.
Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. Syst Biol. 58:468–477.
Marinotti O, et al. 2013. The Genome of *Anopheles darling*, the main neotropical malaria vector. Nucleic Acids Res. 41:7387–7400.
Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.
Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst Biol. 64:1000–1017.
Neafsey DE, et al. 2015. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. Science 347:1258522.
Nene V, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316:1718–1723.
Robinson DR, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci 53:131–147.
Rokas A, Abbot P. 2009. Harnessing genomics for evolutionary insights. Trends Ecol Evol. 24:192–200.
Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol Biol Evol. 30:2134–2144.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.
Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol Biol Evol. 31:1261–1271.
Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A. 109:14942–14947.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13:e1002224.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:609–612.
Wickett NJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 111:E4859–E4868.
Zhong B, Liu L, Yan Z, Penny D. 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18:492–495.

Associate editor: Daniel Sloan