Molecular Phylogenetics and Evolution 64 (2012) 12–20

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Random roots and lineage sorting

Jeffrey A. Rosenfeld [a,c], Ansel Payne [b], Rob DeSalle [c,*]

[a] IST/High Performance and Research Computing, University of Medicine and Dentistry of New Jersey, Newark, NJ 07103, USA
[b] American Museum of Natural History, Richard Gilder Graduate School, New York, NY 10024, USA
[c] American Museum of Natural History, Sackler Institute for Comparative Genomics, New York, NY 10024, USA

ARTICLE INFO

*Article history:*
Received 2 August 2011
Revised 11 February 2012
Accepted 27 February 2012
Available online 14 March 2012

*Keywords:*
Lineage sorting
Random rooting
Phylogenies
Gene tree species tree

ABSTRACT

Lineage sorting has been suggested as a major force in generating incongruent phylogenetic signal when multiple gene partitions are examined. The degree of lineage sorting can be estimated using the coalescent process and simulation studies have also pointed to a major role for incomplete lineage sorting as a factor in phylogenetic inference. Some recent empirical studies point to an extreme role for this phenomenon with up to 50–60% of all informative genes showing incongruence as a result of lineage sorting. Here, we examine seven large multi-partition genome level data sets over a large range of taxonomic representation. We took the approach of examining outgroup choice and its impact on tree topology, by swapping outgroups into analyses with successively larger genetics distances to the ingroup. Our results indicate a linear relationship of outgroup distance with incongruence in the data sets we examined suggesting a strong random rooting effect. In addition, we attempted to estimate the degree of lineage sorting in several large genome level data sets by examining triads of very closely related taxa. This exercise resulted in much lower estimates of incongruent genes that could be the result of lineage sorting, with an overall estimate of around 10% of the total number of genes in a genome showing incongruence as a result of true lineage sorting. Finally we examined the behavior of likelihood and parsimony approaches on the random rooting phenomenon. Likelihood tends to stabilize incongruence as outgroups get further and further away from the ingroup. In one extreme case, likelihood overcompensates for sequence divergence but increases random rooting causing long branch repulsion.



## 1. Introduction

The gene tree species tree phenomenon (Pamilo and Nei, 1988; Maddison, 1997) has been inserted as a major problem in population genetics, speciation studies, and systematics. In systematics, the gene tree problem can result in single gene trees showing incongruence with each other. Two approaches at different extremes (Miyamoto and Fitch, 1995) have been posited as solutions to problems of incongruence in systematics – taxonomic congruence (consensus) and character congruence (concatenation). More recently the taxonomic congruence solution has involved utilization of coalescent theory (Maddison and Knowles, 2006; Edwards et al., 2007; Ané et al., 2007; Carstens and Knowles, 2007; Than and Nakhleh, 2009; Edwards, 2009; Knowles, 2009; Ané, 2010; Knowles and Kubatko, 2010; Bansal et al., 2010; Yu et al., 2011) and other approaches (Rosenberg, 2002; Meng and Kubatko, 2009; McCormack et al., 2009).

Empirical studies using multiple individuals from closely related species and usually less than ten gene partitions have revealed considerable incongruence of individual gene tree/species tree with each other and with the concatenated hypothesis. A classic example of this kind of approach is the study from Machado and Hey (2003) and Machado et al. (2002) on the closely related *Drosophila* species cluster – *D. pseudoobscura pseudoobscura*, *D. pseudoobscura bogatana*, and *D. persimilis*, where no two genes in their data set result in the same topology. Similar empirical studies using genome level information with dozens (and sometimes thousands) of genes have also claimed large amounts of incongruence of individual gene trees with each other and with an overall concatenated hypothesis. One of the more extreme examples of these kinds of studies is the Drosophila 12 Genomes analysis where close to 50% of the greater than 12,000 genes examined support a hypothesis that is not in line with either accepted taxonomy or with the concatenated hypothesis (Pollard et al., 2006).

Fig. 1 summarizes some of these genome level studies and gives the upper limit of the estimation of the number of phylogenetically "inaccurate" genes in each. Several simulation studies have also demonstrated a high degree of incongruence of single gene partitions with each other and with the overall concatenated hypothesis generated from the simulated data. Degnan and colleagues (Degnan and Rosenberg, 2006; Degnan and Salter, 2005, 2009;

* Corresponding author.
*E-mail addresses:* rosenfj1@umdnj.edu (J.A. Rosenfeld), anselpayne@gmail.com (A. Payne), desalle@amnh.org (R. DeSalle).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.02.029

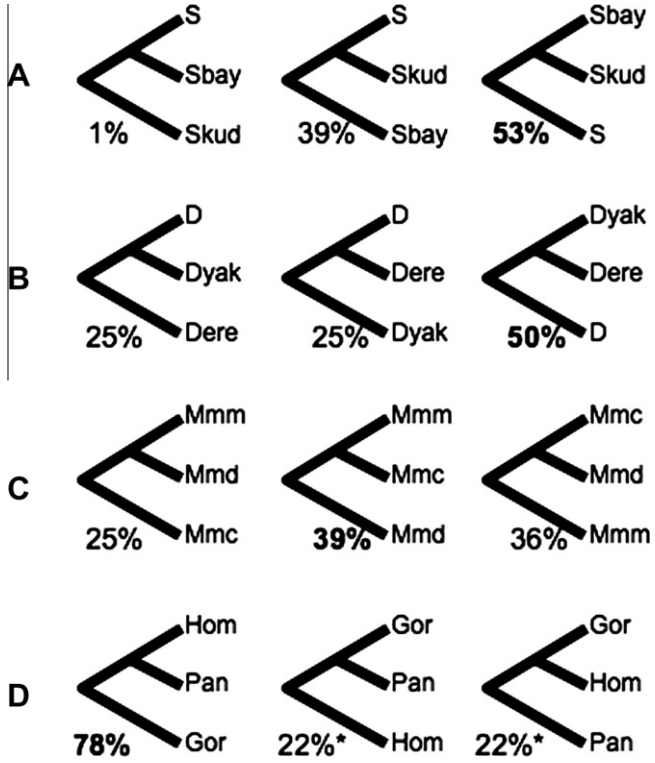



**Fig. 1.** Estimates of incongruence in four published genome-level studies. (A) Possible trees from the Rokas et al. (2003) study showing the three topologies for the yeast species considered in this paper. The percentages here do not add up to 100% because there are some gene topologies that are not relevant to the three shown for this example. S stands for the three ingroup species – *S. cerevisae*, *S. paradoxus* and *S. mikatae*. Sbay stands for *S. bayanus* and Skud stands for *S. kudriavzeii*. (B) Possible trees from the Pollard et al. (2006) study showing the three topologies considered in this paper. D stands for the three ingroup species – *D. melanogaster*, *D. sechelia* and *D. simulans*. (C) Possible trees from the White et al. (2009) study showing the three topologies considered in this paper for the Mus musculus subspecies. Mmm stands for *Mus musculus musculus*, Mmd stands for *Mus musculus domesticus* and Mmc stands for *Mus musculus castaneus*. (D) Possible trees from the Hobolth et al. (2007) study showing the three major topologies for the human (Hom), chimp (Pan), gorilla (Gor) trichotomy considered in this paper. Percentages of gene partitions (A and B) or chromosomal regions (C and D) are shown below each topology. The percentages in bold are for the concatenated tree of each data set.

Kubatko and Degnan, 2007) have pioneered these kinds of studies and have made some interesting observations about tree topology and tree symmetry and the incongruence of single gene partitions. These three kinds of studies have led researchers to the conclusion that lineage sorting is a pervasive problem in the systematics of closely related species. Coalescent theory predicts such behavior of gene trees in relation to the species tree, and so it is important to examine the phenomenon in detail.

One of the issues not examined in great detail in most of these studies is the role of outgroup choice on the incongruence of single gene partitions with each other and with a concatenated hypothesis. Wheeler (1990) pointed out that if an outgroup is chosen that is overly distant from the ingroup being studied, such an outgroup would root "randomly" to the ingroup. This conclusion was a logical extension of examination of long branch attraction, but differed in that, in the case of random rooting, only one taxon (the outgroup) has an inordinate long branch. Several other researchers have addressed the problem of outgroup choice from a theoretical vantage (Watrous and Wheeler, 1981; Farris, 1982; Maddison et al., 1984; Lyons-Weiler et al., 1998; Milinkovitch and Lyons-Weiler, 1998; Smith, 1992), but these studies were done before the advent of genome level information as a source of characters for phylogenetic analysis. Some empirical studies have also examined this problem but again not at the genome level (Qiu et al., 2001; Lartillot et al., 2007; Graham et al., 2002).

Gatesy et al. (2007) examined this problem in the Rokas et al. (2003) genome level yeast data set. In the original analysis of the yeast data set, Rokas et al. (2003) observed that nearly 40% of the gene partitions favored a tree topology at odds with the strongly supported concatenated hypothesis. Subsequent analyses (Taylor and Piel, 2004; Phillips et al., 2004; Hedtke et al., 2006; Gatesy and Baker, 2005; Gatesy et al., 2007) of this dataset have explored the problem of incongruence. Gatesy et al. (2007) showed that the taxa used in the original analysis were randomly rooting onto a stable network. In fact, all 106 gene partitions in this data set give the same ingroup network. By examining the distance of the outgroup to the ingroup taxa in this study, they also showed that as the outgroup gets further away from the ingroup, the number of incongruent rooted topologies gets larger.

This communication examines the impact of rooting on incongruence in large multipartition data sets. We have compiled seven large matrices from a broad array of taxa for this purpose. In this study we first examine the effect of increasing the distance of outgroups from the ingroup. We next examine the claim of rampant lineage sorting in these data sets. Finally we examine the limits of parsimony and likelihood analysis in accommodating this problem of random rooting.

## 2. Materials and methods

### 2.1. Data sets

We used seven large data sets ranging in number of partitions from 13 (Collubines) to over 19,000 (*Drosophila* 12 genomes) and from 30,000 characters (*D. pseudoobscura* data set) to 35,000,000 characters (*Drosophila* 12 genomes). All matrices were augmented with NEXUS character set statements for partitioned analysis. The names in parentheses are how we refer to each data set in this paper. All partitioned data sets are available as Nexus formatted Supplemental Tables (1 through 7).

*D. pseudoobscura* (*pseud*): For this data set we created a matrix with three ingroup taxa (*D. persimilis, D. p. pseudoobscura* and *D. p. bogotana*), and four outgroups (*D. miranda*, *D. melanogaster*, *D willistoni* and *D virilis*), by starting with the Machado and Hey (2003) data set and then searching the data base for genes present in all seven taxa. Statistics for the matrix can be found in Table 1). Fig. 2A shows the tree topology of the accepted relationships of these flies.

*Drosophila 12 genomes* (*dro12*): Incongruence in this group of flies has been shown by Wong et al. (2007) and Pollard et al., (2006). This data set (Table 1) was created from the raw Drosophila 12 Genome data base (http://rana.lbl.gov/drosophila/). In this data set, Pollard et al. (2006) originally suggested that when examining the relationships of *D. melanogaster*, *D. yakuba* and *D. erecta*, rampant lineage sorting was the source of incogruence. We focus on

**Table 1**
Parsimony metrics for the seven matrices used in this study. Data matrix names are from the text. ConcPI stands for number of Concatenated phylogenetically informative characters. ConcTL stands for total length of the Concatenated tree and ConcCI stands for the consistency index of the concatenated tree.

| Matrix | Characters | Partitions | Conc PI | Conc TL | Conc CI |
|---|---|---|---|---|---|
| pseudo | 48,277 | 32 | 6577 | 25,478 | 0.80 |
| dro12 | 35,646,040 | 12,270 | 7,511,306 | 54,864,641 | 0.60 |
| yeast | 127,337 | 106 | 63,248 | 268,991 | 0.53 |
| mouse | 1,145,155 | 4108 | 62,705 | 78,622 | 0.84 |
| coli | 33,738 | 13 | 7914 | 26,738 | 0.44 |
| colII/ape | 51,848 | 55 | 17,971 | 93,122 | 0.33 |

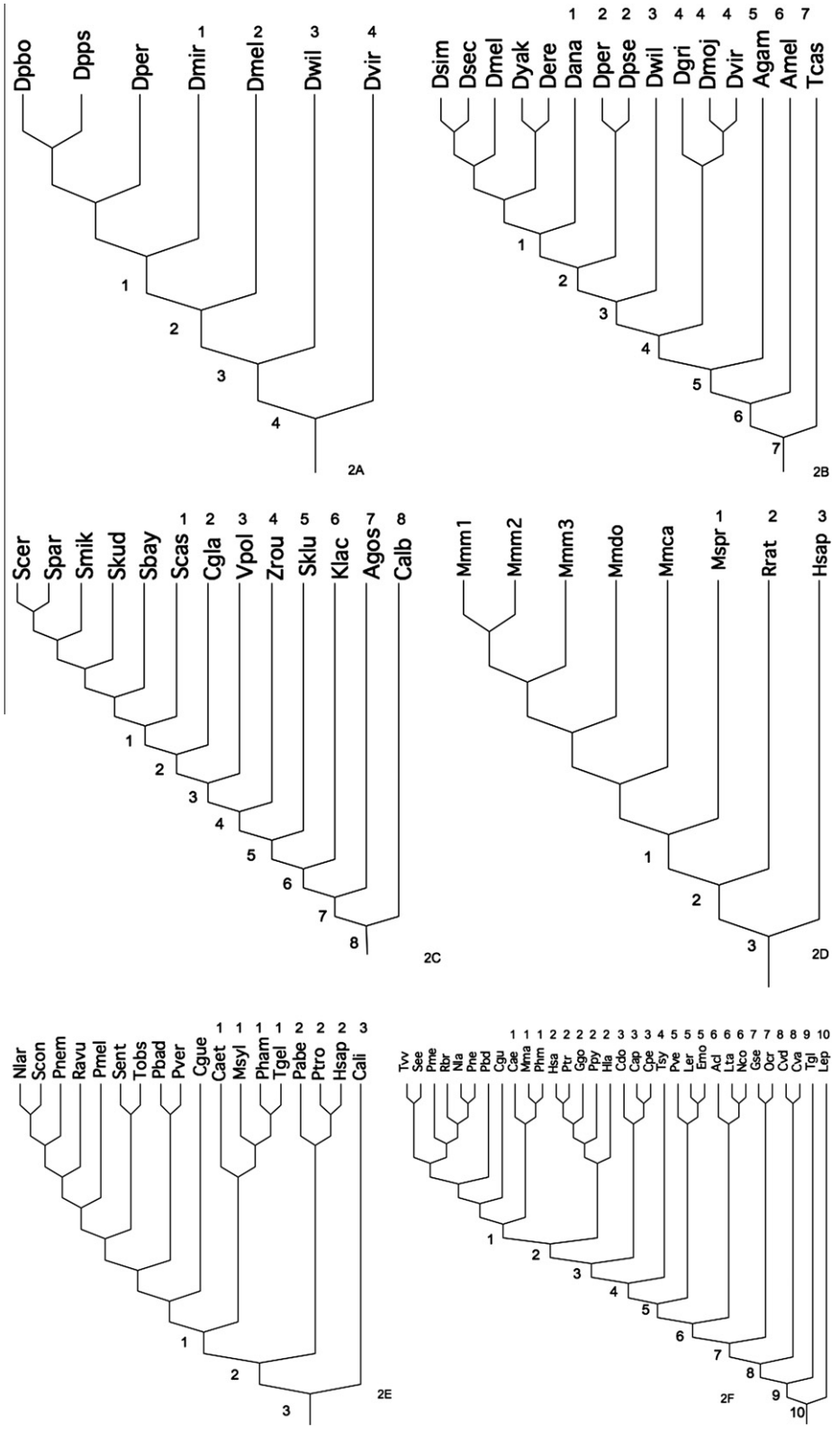



**Fig. 2.** Phylogenetic trees for each of the data sets examined in this study. Numbers above the species designation refer to outgroups and are numbered from closest possible outgroup to most distant outgroup. The nodes where the numbered outgroups are relevant are similarly numbered. (A) Tree for the pseud dataset. (B) Tree for the dro12 data set. (C) Tree for the yeast data set. (D) Tree for the mouse data set. (E) Tree for the coll data set. (F) Tree for the coll data set. Abbreviations are in Supplemental Table 9. These abbreviations are also the same for Figs. 3 and 4.

that taxonomic problem as well as the relationships of the *D. sechelia, D. simulans* and *D. melanogaster* species triad. We have used an amino acid matrix for this data set. The accepted topology for this data set is shown in Fig. 2B. Since this data set is so large (>35,000,000 characters) we subdivided it into seven manageable subdatasets.

*Yeast* (*yeast*): We used the Rokas et al. (2003) data set as a base for this data set (obtained from the authors). We added five new outgroup taxa to the Rokas et al. (2003) data set representing novel phylogenetic distances over what was used in the original study (*Candida glabrata, Kluyveromyces polysporus* [*V. polyspora*], *Zygosaccharomyces*, *Kluyveromyces lactis, Zygosaccharomyces rouxii*, and *Ashbya gossypii*) using the phylogenetic hypothesis for fungi in Dujon (2010). The original phylogenetic problem in the Rokas et al. (2003) data set involved the relationships of two species *S. bayanus* and *S. kudriavzevii*. We also examine the relationships of the *Saccharomyces mikatae, Saccharomyces cerevisiae* and *Saccharomyces paradoxus* triad. Summary of this data set can be found in Table 1. The accepted tree for this data set is shown in Fig. 2C (Dujon, 2010).

*Mouse* (*mouse*): We based this matrix on the study of White et al. (2009). This study examined the relationships of several inbred strains of lab mouse. The triad *Mus mus musculus, M.m. domesticus* and *M.m. castaneus* was the focus of the White et al. (2009) study. Since over ten inbred strains of *M.m. musculus* were also examined we used the three most highly diverged *M.m. musculus* strains for this task. Because the data matrix in White et al., was not available, we constructed a matrix (Table 1) from the raw data obtained from http://www.sanger.ac.uk/Projects/M_musculus/. This data set is part of a project to use Illumina sequencing to re-sequence important laboratory mouse strains. The gene sequences that comprise the matrix for each strain were extracted and then compared to each other, and the human and rat genomes using BLAT.

*Primate* (*ape*): We examined the relationships of the great apes using the data set of Perelman et al. (2011). From this study we expanded the matrix to include the whole mitochondrial genomes of as many primates in the original matrix as possible. We then trimmed the number of taxa from the original 166 down to 50 and included only those taxa with full mitochondrial genomes. The resolution of the human, chimp and gorilla relationships has been a continuing question. Hobolth et al. (2007) estimated that nearly 20% of a sequenced gene region on the X chromosome supported trees that were at odds with the now accepted (gorilla, (human, chimp)) topology. Therefore we examined the relationships of these taxa as well as how orangutan and gibbon are related to the aforementioned trio of species. Perelman et al. (2011) data set allowed us to use outgroups with over 10 different phylogenetic distances from the ingroup to test hypothesis about outgroup randomness (Table 1; Fig. 2E).

*Colubine I* (*colI*): Incongruence of trees generated from mitochondrial versus nuclear markers for colubines has been suggested by Ting et al. (2008) and Roos et al. (2011). The latter study constructed a matrix to examine the possible conflict between mitochondrial and nuclear DNA gene trees for Colubine primates. Their data set contained 17 taxa for 13 gene partitions (one of which is the mitochondrial genome). The phylogenetic problem examined by Roos et al. (2011) involved the topology of the Asian versus African Colubines using two potential outgroups – Apes and Old World monkeys. Specifically, the relationships of the African Colubines, *Pilocolobus* and *Procolobus* to *Colobus* show incongruence amongst gene partitions for the 13 genes used by Roos et al., 2011. In addition, within the Asian Colubines, the Odd-nosed monkeys and langurs show topological incongruence amongst gene partitions. A New World Monkey (*Callithrix*) was added to the original Roos et al. (2011) matrix to increase the number of outgroups for this data set to three. The accepted tree for these data is shown in Fig. 1E and Table 1 summarizes the characteristics of the data set.

*Colubine II* (*colII*): Perelman et al. (2011) data set was also used to address the same topology questions addressed in the Colubine I data set. The accepted tree for these data is shown in Fig. 2F and Table 1 summarizes the characteristics of the data set.

### 2.2. Phylogenetic analysis

All tree building analyses were accomplished in PAUP (Swofford, 2002). For all data sets, trees were generated using parsimony and likelihood with Branch and Bound (bandb) searches. In most analyses the likelihood settings were lset nst = 2 rates = gamma shape = estimate. When we examined the yeast data set we also used a more parameter rich model (lset nst = 6 rates = gamma shape = estimate). Searches were trivial in all cases as there were either 3, 4 or 5 ingroup taxa in analyses plus a single outgroup. Phylogenetic analyses were scripted with command files that kept ingroup taxa (either 3, 4 or 5 of them) stable with the swapping of a single outgroup taxa into the analysis separately. Each tree generated in this study was saved as a Nexus file Newick tree. A file for each single outgroup analysis for each matrix was generated and these were then parsed to tabulate the number of trees for different hypotheses generated for the different matrices and different outgroups. Raw data and summary excel files are available from the authors on request.

### 2.3. Analysis of changing tree topologies by varying outgroup

We used two measures to quantify the impact of outgroup manipulation. First, the ratio of the number of "bad" (B) phylogenetic hypotheses to "good" (G) phylogenetic hypotheses (the "B/G ratio") was computed by dividing the number of trees with inaccurate topologies by the number of trees with the accepted phylogeny. So for instance if three taxa (A, B and C) were being examined and the tree topology with taxon A as basal was the accepted topology and it occurs in 100 genes in an analysis and there are 50 trees generated by genes with B and C as basal, then the B/G ration would be 50/100 gene partitions or 0.5. Our second measure was to take the total number of genes that produce an accurate topology and divide it by the total number of genes that gave a resolved topology (called the G percentage). These measures were correlated in most cases so we present here only the B/G ratio. The genetic distance between each outgroup to the ingroup taxa in each data matrix was then calculated using PAUP (with a K2P model) and plotted versus the B/G ratio using linear regression.

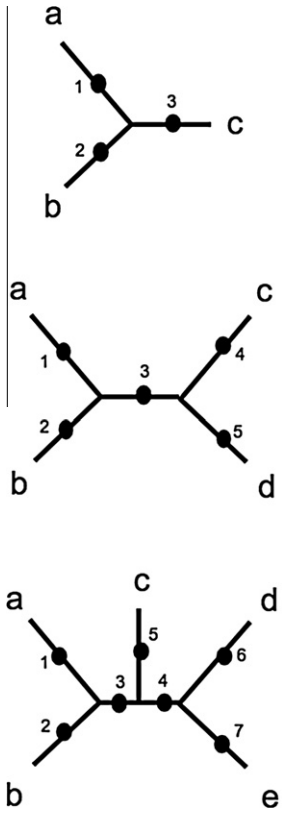
### 2.4. Rooting strategies

Fig. 3 shows the possible points of rooting for networks with 3, 4 and 5 taxa. The three taxa network has three root points, a stable four taxa network has five root points and a stable five taxa network has seven root points. The rest of the figure details the strategy for swapping outgroups in and out of analyses and which taxa were used to keep the ingroup stable. In this way we were able to examine the effect of using further and further distanced roots when there are three, four and five ingroup taxa. We also examined three taxon statements within the ingroups that allowed us to add two "new" but extremely close outgroup taxa. We were able to do this in six cases (the pseudo matrix only used three ingroup taxa to begin with). Fig. 4 shows the strategy for manipulating the outgroup taxa with the three "closer" ingroup taxa (we call these analyses alternative or "alt").

## 3. Results and discussion

### 3.1. Impact of increasing root distance on tree topology

The general trend observed when increasing outgroup distance is that the B/G ratio (see methods) increased in a linear fashion. Fig. 5 shows this trend for both MP and ML analysis. In all cases, the slopes of the curves in this figure are positive and in all but

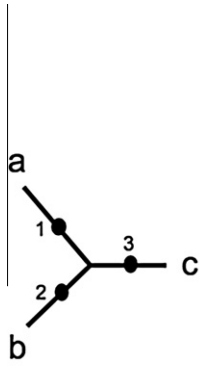



| Three Taxa | a | b | c |
|---|---|---|---|
| dro12 | Dyak | Dere | Dmel (Dsec,Dsim) |
| pseud | Dpbo | Dpps | Dper |
| yeast | Sbay | Skud | Smik (Spar,Scer) |
| mouse | Mmd | Mmc | Mmm3 (Mmm2,Mmm1) |
| coll | Cgue | Pver (Pbad) | Sent (Nlar,Scon,Ravu,Pnel,Tobs) |
| colll | Cgue | Pbad | Tvv (See,Pme,Rbr,Nlar,Pnel) |
| ape | Hsa | Ptr | Ggo (Pab) |

| Four Taxa | a | b | c | d |
|---|---|---|---|---|
| dro12 | Dyak | Dere | Dmel | Dsim (Dsec) |
| pseud | -- | -- | -- | -- |
| yeast | Sbay | Skud | Smik | Spar (Scer) |
| mouse | Mmdo | Mmca | Mmm3 | Mmm2 (Mmm1) |
| coll | Cgue | Pver (Pbad) | Sen | Nlar (Scon,Ravu,Pnel,Tobs) |
| colll | Cgue | Pbad | Tvv | See (Pme,Rbr,Nlar,Pnel) |
| ape | Hsa | Ptr | Ggl | Pab |

| Five Taxa | a | b | c | d | e |
|---|---|---|---|---|---|
| dro12 | Dyak | Dere | Dmel | Dsim | Dsec |
| pseud | -- | -- | -- | -- | -- |
| yeast | Sbay | Skud | Smik | Spar | Scer |
| mouse | Mmdo | Mmca | Mmm3 | Mmm2 | Mmm1 |
| coll | Cgue | Pver (Pbad) | Sent | Nlar | Scon (Ravu,Pnel,Tobs) |
| colll | Cgue | Pbad | Tvv | See | Pme (Rbr,Nlar,Pnel) |
| ape | Hsa | Ptr | Ggo | Pab | Hla |

**Fig. 3.** Rooting points for networks with three (A), four (B) and five (C) ingroup taxa. The root points are numbered 1 through 5 and the taxa are numbered a through e. The table to the right explains the combinations of ingroup taxa we used to manipulate outgroup addition for each of the seven data sets we used. For instance, for the dro12 data set and three ingroup taxa, taxon "a" would be Dyak, taxon "b" would be Dere and taxon "c" would be either Dmel, Dsec or Dsim. The abbreviations in the table are the same as in Fig. 2 and can be found in Supplemental Table 9.




| | dr12 | yeast | coll | colll | mouse | ape |
|---|---|---|---|---|---|---|
| a | Dmel | Scer | Nas | Nla | Mmm1 | Hsp |
| b | Dsim | Spar | Prs | Pme | Mmm2 | Ptr |
| c | Dsec | Smik | Sem | See | Mmm3 | Ggo |
| OG1 | Dyak,Dere | Sbay | Pil | Pbd | Mmdo | Ppy |
| OG2 | Dana | Skud | Col | Cgu | Mmca | Hla |
| OG3 | Dpse, Dper | Scas | Mac etc | Mma,Hsp | Mspre | Cgu etc |
| OG4 | Dwili | Sklu | Hom etc | Cpe etc | Rrat | Mma etc |
| OG5 | Dv Dg, Dm | Calb | Cal | Tsy | Hsap | Tsy |
| OG6 | Agam | Cgla | ---- | Lct | ---- | Lct |
| OG7 | Amel | Vpol | ---- | Lta | ---- | Lta |
| OG8 | Tcas | Zrou | ---- | Gse | ---- | Gse |
| OG9 | ---- | Klac | ---- | Cva | ---- | Cva |
| OG10 | ---- | Agos | ---- | Tgl | ---- | Tgl |
| OG11 | ---- | ---- | ---- | Lep | ---- | Lep |

**Fig. 4.** Strategy for examining six of the seven data sets for "closer" ingroup taxa (by closer we mean phylogenetically closer). The tree shows the three ingroup taxa labeled as a, b and c. The dots refer to the possible rooting points of the network and these are numbered 1 through 3. Outgroups are listed in order of increasing distance from the ingroups (OG1 through OG11). Abbreviations are the same as in Fig. 2. We omitted the pseud data set from this analysis because there were only three ingroup taxa to begin with.

one case the correlation coefficient is significant at the $P < 0.1$ level and better (Table 2). This result indicates that there is a positive correlation of increasing genetic distance of an outgroup with increasing phylogenetic inaccuracy. Another way to summarize this observation is to note that the closer the ingroup is to the outgroup, (by closer we mean phylogenetically closer) the fewer incongruent gene partitions are found.

Because there are seven potential root points on a stable network with five ingroup taxa, this means that there are more opportunities for random rooting for this number of taxa than for three

or four ingroup taxa. We therefore tested whether the number of potential root points was a factor in the correlation. To do this, we examined each data set by including just three ingroup and then just four ingroup taxa. Supplemental Table 8 shows the results of this analysis and indicates that while the correlation starts to erode when the number of root points in a network is reduced, there are still strong correlations for most analyses (Supplemental Table 3).

Another striking result of these analyses is that the B/G ratio intercept when distance is 0.0 is either negative or very close to zero. This result suggests that the ratio of inaccurate to accurate gene trees goes to zero as the outgroup gets closer and closer to the ingroup. The only departure from this general trend for five taxa, is the yeast data set where the B/G intercept for MP is 0.14 and the same intercept for ML is 0.20. For four and three taxa ingroup analyses, the B/G ratio intercept shows greater departure from zero than for five taxa (Tables 2 and 3). These results together with the assumption that the relationship of distance to B/G ratio is linear suggest that the amount of lineage sorting at the point of divergence of closely related species is closer to zero than previously inferred. The reason for the discrepancy lies in the extreme distance of the chosen outgroups to the ingroups in nearly all of the matrices we examined. The major departure from this trend is the mouse matrix, where the slopes of the regressions for this data set are either 0 or slightly significantly different from 0.

### 3.2. Examining very closely related taxa for the random outgroup effect

In at least three of the data sets (Rokas et al., 2003; Pollard et al., 2006; White et al., 2009) we examined, the previous analyses suggested large numbers of genes experiencing lineage sorting in taxa that were not the most closely related taxa in the data set. For the yeast data set, the major problems attributed to incongruence

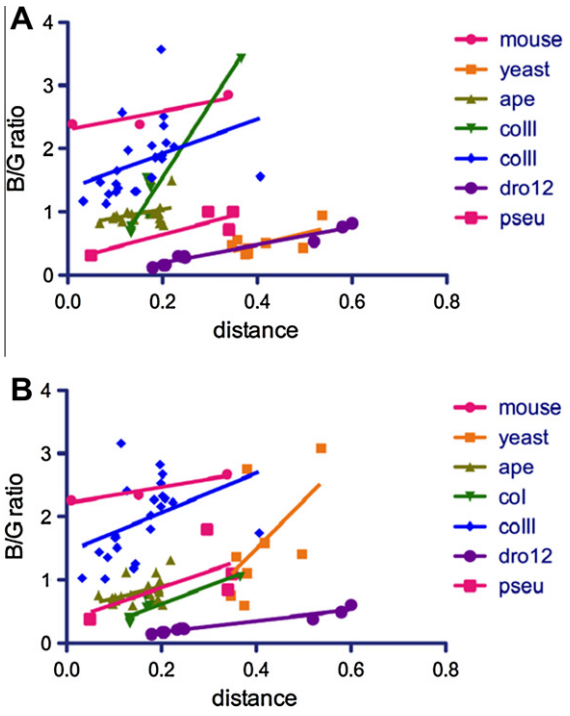



**Fig. 5.** Linear regressions for the seven data sets in the analysis. A shows the regressions for the parsimony analyses and B shows the regressions for likelihood analyses. Legends for the different data sets are shown as insets in each graph. Sequence distance (calculated using the K2P model) is graphed on the *X* axis and B/G ratio ("bad" to "good" ratio; see text) is graphed on the *Y* axis.

**Table 2**
Linear regression statistics for the seven data sets analyzed with five ingroup taxa.

| Dataset | Slope | $r^2$ | Intercept | Significance |
|---|---|---|---|---|
| Ape_5T_MP | 1.397 ± 0.8441 | 0.1205 | −0.5452 | ** |
| Ape_5T_ML | 1.830 ± 0.8007 | 0.207 | −0.289 | ** |
| ColI_5T_MP | 11.59 ± 0.9370 | 0.9623 | 0.06703 | **** |
| ColI_5T_ML | 2.883 ± 0.5556 | 0.8178 | −0.01384 | *** |
| ColII_5T_MP | 2.746 ± 1.516 | 0.1351 | −0.498 | * |
| ColII_5T_ML | 3.141 ± 1.539 | 4.165 | 0.3184 | ** |
| Yeast_5T_MP | 1.870 ± 0.8686 | 0.4358 | 0.1453 | ** |
| Yeast_5T_ML | 7.621 ± 4.272 | 0.3466 | 0.2045 | * |
| pseud_5T_MP | 2.006 ± 0.7904 | 0.7631 | −0.1186 | * |
| pseud_5T_ML | 2.605 ± 2.326 | 0.3854 | −0.1361 | ns |
| Dro12_5T_MP | 8.896 ± 0.7933 | 0.9402 | 0.006766 | **** |
| Dro12_5T_ML | 14.36 ± 1.186 | 0.9483 | 0.06703 | **** |
| Mouse_5T_MP | 1.467 ± 0.7264 | 0.8031 | −1.566 | * |
| Mouse_5T_ML | 1.286 ± 0.3329 | 0.9372 | −1.719 | * |

*$p < 0.10$, **$P < 0.05$, ***$p < 0.01$, ****$p < 0.001$. Data matrix names are from the text. 5T stands for 5 taxa and MP and ML stand for Maximum Parsimony and Maximum Likelihood respectively.

involved Sbay and Sklu, when three taxa – Smik, Scer and Sper – are actually more closely related to each other than the examined triad. In the fly data set, the triad of species *D. melanogaster, D. yakuba* and *D. erecta* were examined for lineage sorting, when three taxa – *D. melanogaster, D. simulans* and *D. sechelia* are actually more closely related to each other than the examined triad. Finally, in the mouse data set, the three taxa examined were a strain of *Mus musculus musculus, M.m. domesticus* and *M.m. castaneus*, while there are at least ten other *M.m. musculus* strains that are more closely related to each other than the three taxa examined. In addition, by shifting the three ingroup taxa to examine the potential for lineage sorting to even more closely related entities, this also allows for the use of closer outgroup taxa. For example, in the yeast data set, where for the five ingroup taxa – Sklu, Sbay, Smik, Scas and Sper are used as ingroups, if we examine the relationships of the three most closely related taxa in the group, then Sklu and Sbay become outgroups. These two taxa are more closely related to the three ingroup taxa than the originally designated outgroups (see Fig. 3). Table 3 shows the results of altering the ingroup taxa to include the most closely related taxa in each data set (see Fig. 6). This table demonstrates that in four out of five cases we examined for maximum parsimony, adding closer and closer outgroups results in lower B/G ratios. The intercept values for significant regressions are all either slightly negative or very close to zero, indicating that when the outgroups chosen are extremely close to the ingroup species, the number of incongruent genes is expected to go to zero or nearly so.

We also used these analyses to give us an estimation of the degree of lineage sorting amongst closely related taxa. Table 4 shows the best estimate of percentage of genes that are incongruent with the concatenated hypothesis and hence candidates for lineage sorting. Not surprisingly the data set with the smallest number of partitions shows a relatively large percentage of gene partitions that are incongruent with the concatenated tree (44% for MP and 50% for ML). This result may be due to the small number of gene partitions that are included in this data set (13 partitions). In fact, when the same phylogenetic question is examined using the larger Perelman et al. (2011) data set (ColII) with 55 partitions the percentage of incongruent genes drops to 20% for MP and 41% for ML. The Mouse data set is also highly incongruent and hence shows the largest percentage of potential lineage sorting. This observation could also be the result of the large number of extremely closely related strains that are in this data set. The divergence of these mouse lineages is almost certainly very recent (Yang et al., 2007; White et al., 2009), and in fact, the taxa examined in this data set are all considered part of the same species (*M.m. musculus,*

**Table 3**
Regression analysis for "closer" three taxa as ingroup analysis.

| Data set | Slope | $r^2$ | Intercept | Significance |
|---|---|---|---|---|
| Ape_alt_MP | 4.211 ± 0.8661 | 0.4962 | −0.01699 | **** |
| Ape_alt_ML | 4.781 ± 1.137 | 0.4241 | −0.03605 | **** |
| ColI_alt_MP | 1.645 ± 1.401 | 0.04258 | −0.1259 | * |
| ColI_alt_ML | 0.8885 ± 1.910 | 0.02635 | −0.8228 | ns |
| ColII_alt_MP | 1.117 ± 0.4580 | 0.4596 | −0.2322 | ** |
| ColII_alt_ML | −0.3905 ± 1.077 | 0.00385 | 1.065 | ns |
| Yeast_alt_MP | 0.7189 ± 0.567 | 0.1673 | 0.04428 | * |
| Yeast_alt_ML | 2.507 ± 0.2869 | 0.9052 | 0.001863 | **** |
| Dro_alt_MP | 1.410 ± 0.05667 | 0.9841 | 0.03688 | **** |
| Dro_alt_ML | 1.902 ± 0.05792 | 0.9908 | 0.02081 | **** |
| Mouse_alt_MP | 0.004891 ± 0.1 | 0.0002145 | −190.6 | ns |
| Mouse_alt_ML | 2.605 ± 2.326 | 0.3854 | −0.1361 | ns |

*$p < 0.10$, **$P < 0.05$, ***$p < 0.01$, ****$p < 0.001$. Data matrix names are from the text. alt stands for the alternative taxon set using closer three taxa sets and MP and ML stand for Maximum Parsimony and Maximum Likelihood respectively.

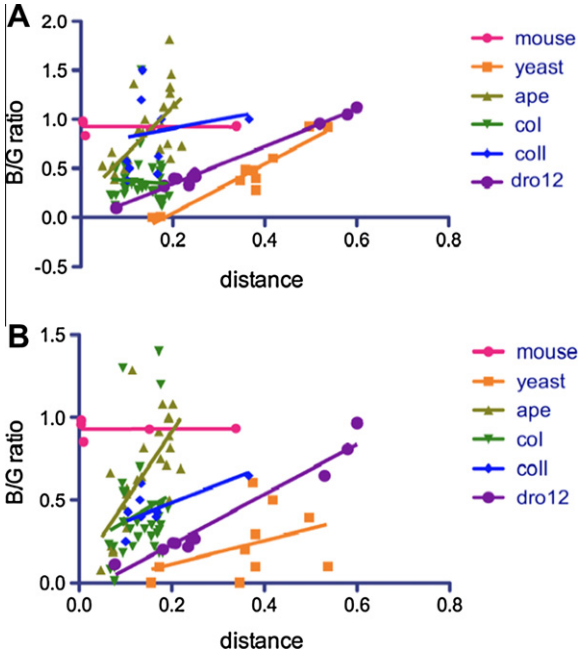



**Fig. 6.** Linear regressions for the data sets where the "closer" outgroup analyses were accomplished. A shows the regressions for the parsimony analyses and B shows the regressions for likelihood analyses. Legends for the different data sets are shown as insets in each. Sequence distance (calculated using the K2P model) is graphed on the *X* axis and B/G ratio (see text) is graphed on the *Y* axis. We omitted the pseud data set from this analysis because there were only three ingroup taxa to begin with.

**Table 4**
Best estimate of percentage of genes that are incongruent with the concatenated hypothesis at the time of divergence of species and hence candidates for "true" lineage sorting. We used the intercept values for each of the analyses shown in Figs. 5 and 6 to compile these estimates. Percentages are calculated for just those genes that gave trees that were resolved (informative) and for all genes (total) in the data sets. Maximum Parsimony is represented by MP and Maximum Likelihood is represented by ML.

| | Percent informative | | Percent total | |
|---|---|---|---|---|
| | MP | ML | MP | ML |
| dr12 | 0.036 | 0.020 | 0.016 | 0.010 |
| mouse | 0.495 | 0.331 | 0.082 | 0.027 |
| ColI | 0.444 | 0.500 | 0.153 | 0.090 |
| ape | 0.142 | 0.320 | 0.054 | 0.145 |
| ColII | 0.200 | 0.416 | 0.072 | 0.384 |
| yeast | 0.000 | 0.000 | 0.000 | 0.000 |

*M.m. castaneus* and *M.m. domesticus* and several inbred strains of *M.m. musculus*). It is not surprising therefore that incongruence of gene trees is pervasive in this data set because of the recentness of reticulation.

The three data sets with large numbers of genes and strongly differentiated taxa – dro12, ape and yeast – all show a similar pattern of having very few genes that are incongruent with the concatenated topology. The yeast data set in particular, shows no incongruence when the most closely related three ingroups are examined. This result is entirely consistent with the Gatesy et al. (2007) result where all 106 yeast genes give the same 5 taxa network topology when the outgroup problem is removed. The *Drosophila* 12 genomes data set shows at most 8% of the genes in the data set as being incongruent with the accepted (which is also the concatenated) topology. This percentage can be contrasted with the original suggestion that in Drosophila up to 50% of the genes are incongruent and thus candidates for lineage sorting. The ape data set for parsimony is incongruent at 14% of the informative partitions with the accepted (which is also the concatenated) topology. This estimate is for the degree of incongruence for the chimp human gorilla trichotomy. Recently Hobolth et al. (2011) have estimated the degree of lineage sorting with orangutan-human-chimp to be 1.5%, which is more in line with our analysis here. When the percentage of incongruent gene partitions relative to the total number of partitions in each data set is estimated, the level of incongruence over the entire data set is reduced even more (for MP this ranges from 0% to 15% and for ML from 0% to 38%). This variance is most likely caused by the range of taxonomic systems examined in this study.

### 3.3. Assessing the impact of likelihood analysis versus parsimony analysis with random outgroups

Another observation that can be made from Tables 2–4 is that for the likelihood approach, only about half of the analyses give significant correlations, suggesting, not surprisingly, that ML and MP are behaving differently in the way they process information from the distant outgroups. Changing the likelihood model to include more parameters, has little impact on the overall results (data not shown). This result suggests that in some cases ML tends to stabilize the number of accurate gene partitions, whereas MP tends to allow for the observation of the incremental decrease of accuracy with increasing distance.

The yeast dataset requires a detailed examination with respect to comparing MP and ML. The following analysis involves considering the three closest Yeast ingroups only (scer, spar, smik). The number of genes giving the three different topologies these three ingroups can generate during MP analysis is given in Table 5. Note

**Table 5**
Number of gene partitions (out of 106) that agree with the accepted topology (also the concatenated topology) for the yeast three taxon data set that includes Scas, Smik and Spar. ML1 model parameters were nst = 2, rates = gamma, shape = estimate, ML2 model parameters were nst = 6, rates = gamma, shape = estimate. MP is the maximum parsimony analysis.

| Outgp | ML1 | ML2 | MP |
|---|---|---|---|
| Agos | 55 | 59 | 101 |
| Calb | 53 | 65 | 102 |
| Cgla | 71 | 77 | 99 |
| Klac | 66 | 70 | 100 |
| Sbay | 105 | 102 | 103 |
| Scas | 77 | 78 | 104 |
| Sklu | 82 | 86 | 102 |
| Skud | 106 | 106 | 106 |
| Vpol | 70 | 80 | 101 |
| Zrou | 75 | 77 | 103 |

that for both ML analyses (ML1 and ML2) the number of inaccurate gene partitions is much larger than for MP. We attribute the poor performance of the ML approach in this case to a previously recognized phenomenon called long branch repulsion (Siddall, 1998; Siddall and Whiting, 1999). This phenomenon results when two taxa have extremely long branches and are actually each others' closest relative. Siddall (1998) showed that maximum likelihood overcompensates for the two long branches and repulses them to accrue a position in a topology that is "inaccurate". This is exactly the problem we observe for the yeast data set. In this case, any outgroup we choose outside of scud and sbay is at least three times more distant to the ingroups (scer, smik, sper) and the longest branch of the ingroups is the smik branch which is 60% longer than either of the scer or sper branches. Hence the "correct" topology is avoided as a result of repulsion of smik and the outgroup.

## 4. Conclusions

In this study we have attempted to characterize the random outgroup problem and how it could be misinterpreted as lineage sorting. We come to four conclusions: (1) Increasing the genetic and phylogenetic distance of the outgroup to ingroups increases the frequency of gene partitions that produce incongruent topologies; (2) While some studies have claimed as much as 2/3 of gene partitions are sorting (mice), and most settle somewhere near 50% (yeast, drosophila and primates), the actual percentage of gene partitions that are sorting may be much less, perhaps even less than 10% in most cases where large numbers of gene partitions have been examined (Table 2); (3) ML and MP behave differently when dealing with the random outgroup or long branch outgroup. Specifically, ML tends to decrease the effect of using outgroups further and further away from the ingroup, while MP is more susceptible to the showing the effect. In addition we show in at least one case where the ML approach over compensated for the random outgroup effect and showed a strong case of "long branch repulsion". (4) It is obvious from the results of this study and several other theoretical studies that outgroup choice is extremely important in phylogenetic analysis. Our results suggest that outgroups at extreme distances from the ingroup will mimic lineage sorting and produce roots randomly. As with other theoretical treatments of outgroup choice we recommend choosing the closest outgroup taxon possible to minimize the impact of this rooting problem.

An examination of the simulation studies mentioned earlier is illuminating in the context of the random rooting affect. Kubatko and Degnan (2007), used simulation studies to assess the occurrence of incongruent tree topologies under the coalescent. In their study they used four taxa to examine the dynamics of the coalescent. They divided the trees that four taxa can produce into the following categories: matching tree (MT; the tree that matches the concatenated analysis), symmetrical trees (St) and swapped trees (ST). Symmetrical trees are those trees that maintain the network topology of the MT. In other words, they are very much like the trees we have discussed at length in this paper where roots can attach randomly and give an "altered" topology. Kubatko and Degnan (2007) call these "anomalous". The STs are trees that change the MT network topology and are only discoverable in analyses with greater than four taxa. Kubatko and Degnan (2007) showed that under the coalescent the St's could become quite predominate in their simulated data sets, but under no coalescent model did the ST trees rise to any significant percentage of the total. We point out that the alternative trees that we examined for the analyses in this paper are anomalous trees with respect to the MT.

The frequency of lineage sorting has been focused upon as a problem in modern systematics (Edwards, 2009; Knowles, 2009; Degnan and Rosenberg, 2009). Several authors using genome level data sets have suggested that the amount of lineage sorting is alarming. We do not deny the existence of lineage sorting, as coalescent theory predicts that gene trees and the species tree don't necessarily have to be congruent in all cases. With studies that utilize several closely related taxa within species, the problem of lineage sorting and incongruence could indeed be extreme as a result of reticulation. However, we suggest that much of the lineage sorting observed so far in genome level data sets can be attributed to something much simpler – random rooting.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2012.02.029.

## References


Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426.

Ané, C., 2010. Reconstructing concordance trees and testing the coalescent model from genome-wide data sets. In: Knowles, L.L., Kubatko, L.S. (Eds.), Estimating Species Trees: Practical and Theoretical Aspects. Wiley-Blackwell, Hoboken, NJ, pp. 35–52.

Bansal, M.S., Burleigh, J.G., Eulenstein, O., 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. BMC Bioinform. 11 (Suppl. 1), S42.

Carstens, B.C., Knowles, L.L., 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. Syst. Biol. 56, 400–411.

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, 762–768.

Degnan, J.H., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37.

Degnan, J.H., Rosenberg, N.A., 2009. Trends Ecol. Evol. 24 (6), 332–340, Epub 2009 March 21, Review.

Dujon, B., 2010. Yeast evolutionary genomics. Nat. Rev. Genet. 11 (7), 512–524, Review.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104, 5936–5941, Gene tree discordance, phylogenetic inference and the multispecies coalescent.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

Farris, J.S., 1982. Outgroups and parsimony. Syst. Zool. 31, 328–334.

Graham, S.W., Olmstead, R.G., Barrett, S.C., 2002. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. Mol. Biol. Evol. 19 (10), 1769–1781.

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

Hobolth, A., Christensen, O.F., Mailund, T., Schierup, M.H., 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3 (2), e7, Epub 2006 November 30.

Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H., Mailund, T., 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21 (3), 349–356, Epub 2011 January 26.

Gatesy, J., Baker, R.H., 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst. Biol. 54 (3), 483–492.

Gatesy, J., DeSalle, R., Wahlberg, N., 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56 (2), 355–363.
Knowles, L.L., Kubatko, L.S., 2010. Estimating species trees: practical and theoretical aspects. Wiley-Blackwell, Hoboken, NJ.
Knowles, L.L., 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58, 463–467.
Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56 (1), 17–24.
Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7 (Suppl. 4), S4.
Lyons-Weiler, J., Hoelzer, G.A., Tausch, R.J., 1998. Optimal outgroup analysis. Biol. J. Linn. Soc. 64, 493–511.
Machado, C.A., Hey, J., 2003. The causes of phylogenetic conflict in a classic Drosophila species group. Proc. Biol. Sci. 270 (1520), 1193–1202.
Machado, C.A., Kliman, R.M., Markert, J.A., Hey, J., 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. Mol. Biol. Evol. 19 (4), 472–488.
Maddison, W.P., Donoghue, M.J., Maddison, D.R., 1984. Outgroup analysis and parsimony. Syst. Zool. 33, 83–103.
Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.
Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.
McCormack, J.E., Huang, H., Knowles, L.L., 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58 (5), 501–508, Epub 2009 August 20.
Meng, C., Kubatko, L.S., 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor. Popul. Biol. 75 (1), 35–45, Epub 2008 November 5.
Milinkovitch, M.C., Lyons-Weiler, J., 1998. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. Mol. Phylogenet. Evol. 9, 348–357.
Miyamoto, M.M., Fitch, W.M., 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44, 64–76.
Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5, 568–583.
Perelman, P., Johnson, W.E., Roos, C., Seuánez, H.N., Horvath, J.E., Moreira, M.A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M.P., Silva, A., O'Brien, S.J., Pecon-Slattery, J., 2011. A molecular phylogeny of living primates. PLoS Genet. 7 (3), e1001342, Epub 2011 March 17.
Phillips, M.J., Delsuc, F., Penny, D., 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21, 1455–1458.
Pollard, D., Iyer, V.N., Moses, A.M., Eisen, M.B., 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2, e173.
Qiu, Y.L., Lee, J., Whitlock, B.A., Bernasconi-Quadroni, F., Dombrovska, O., 2001. Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? Amborella, Nymphaeales, Illiciales, Trimeniaceae, and Austrobaileya. Mol. Biol. Evol. 18 (9), 1745–1753.
Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425 (6960), 798–804.
Roos, C., Zinner, D., Kubatko, L.S., Schwarz, C., Yang, M., Meyer, D., Nash, S.D., Xing, J., Batzer, M.A., Brameier, M., Leendertz, F.H., Ziegler, T., Perwitasari-Farajallah, D., Nadler, T., Walter, L., Osterholz, M., 2011. Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. BMC Evol. Biol. 24 (11), 77.
Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.
Siddall, M.E., 1998. Success of parsimony in the four-taxon case. Mol. Biol. Evol. 13, 1187–1191Siddall, M.E., . Long-branch repulsion by likelihood in the Farris Zone. Cladistics 14, 209–220.
Siddall, M.E., Whiting, M., 1999. Long-branch abstractions. Cladistics 15, 9–24.
Smith, A.B., 1992. Rooting molecular trees: problems and strategies. Biol. J. Linn. Soc. 51, 279–292.
Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.
Taylor, D.J., Piel, W.H., 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21.
Than, C., Nakhleh, L., 2009. Species tree inference by minimizing deep coalescences. PLoS Comput. Biol. 5 (9), e1000501, Epub 2009 September 11.
Ting, N., Tosi, A.J., Li, Y., Zhang, Y.P., Disotell, T.R., 2008. Phylogenetic incongruence between nuclear and mitochondrial markers in the Asian colobines and the evolution of the langurs and leaf monkeys. Mol. Phylogenet. Evol. 46 (2), 466–474, Epub 2007 November 28.
Watrous, L.E., Wheeler, Q.E., 1981. The out-group comparison method of character analysis. Syst. Zool. 30, 1–11.
Wheeler, W.C., 1990. Nucleic acid sequence phylogeny and random outgroups. Cladistics 6, 363–367.
White, M.A., Ané, C., Dewey, C.N., Larget, B.R., Payseur, B.A., 2009. Fine scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5, e1000729.
Wong, A., Jensen, J.D., Pool, J.E., Aquadro, C.F., 2007. Phylogenetic incongruence in the *Drosophila melanogaster* species group. Mol. Phylogenet. Evol. 43 (3), 1138–1150, Epub 2006 September 9.
Yang, H., Bell, T.A., Churchill, G.A., 2007. Pardo-Manuel de Villena F. On the subspecific origin of the laboratory mouse. Nat. Genet. 39, 1100–1107.
Yu, Y., Than, C., Degnan, J.H., Nakhleh, L., 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst. Biol. 60 (2), 138–149, Epub 2011 January 19.