

Missing Data and Influential Sites: Choice of Sites for Phylogenetic Analysis Can Be As Important As Taxon Sampling and Model Choice

Liat Shavit Grievink^{*1}, David Penny², and Barbara R. Holland³

¹The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Israel

²Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

³School of Mathematics and Physics, University of Tasmania, Hobart, Australia

*Corresponding author: E-mail: liat.sg@mail.huji.ac.il; liats80@hotmail.com.

Accepted: February 20, 2013

Abstract

Phylogenetic studies based on molecular sequence alignments are expected to become more accurate as the number of sites in the alignments increases. With the advent of genomic-scale data, where alignments have very large numbers of sites, bootstrap values close to 100% and posterior probabilities close to 1 are the norm, suggesting that the number of sites is now seldom a limiting factor on phylogenetic accuracy. This provokes the question, should we be fussy about the sites we choose to include in a genomic-scale phylogenetic analysis? If some sites contain missing data, ambiguous character states, or gaps, then why not just throw them away before conducting the phylogenetic analysis? Indeed, this is exactly the approach taken in many phylogenetic studies. Here, we present an example where the decision on how to treat sites with missing data is of equal importance to decisions on taxon sampling and model choice, and we introduce a graphical method for illustrating this.

Key words: maximum likelihood, site likelihood, *Mesostigma*, missing data, influential sites, taxon sampling.

The importance of both taxon sampling and model choice to the accuracy of phylogenetic studies has been well documented; however, the importance of the treatment of missing data has been examined less frequently. Huelsenbeck (1991) was one of the first to look at the issue. More recently, there has been a disagreement in the literature as to how problematic missing data can be. Lemmon et al. (2009) found that missing data could lead to biases for both maximum likelihood and Bayesian inference. Other authors have argued that incomplete taxa are typically beneficial for phylogenetic inference, that is, that more data are better even if they are not complete (Wiens and Morrill 2011; Wiens and Tiu 2012). We agree with the findings of Roure et al. (2013) who argue that the crux of this difference in opinion is that the simulation study of Lemmon et al. (2009) introduced sites with missing data where those sites were generated by a different process than the rest of the alignment, whereas in the simulation study of Wiens and Morrill (2011), sites with missing data were generated by the same process as at other sites. We think the debate about missing data can be more usefully

framed by the general statistical concept of whether or not data are missing at random (Allison 2002).

To illustrate this point of view, we report on a mitochondrial data set used to study deep divergences in the plant kingdom, including the deeply diverging green alga *Mesostigma*; we consider not only the overall likelihoods but also individual site likelihoods, we also examine the effect of selecting sites for analysis based on whether or not they contain missing data. Our example is a significant one for plant evolution, and the background is important for understanding why we use this example. Plants and green algae comprise two major phyla: Streptophyta (land plants and their green algal relatives) and Chlorophyta (most green algae). *Mesostigma viride* (common name *Mesostigma*) is the only known member of Mesostigmatales. Because it is such an isolated taxon, with a lineage which is likely to extend back a billion years, it is unsurprising that *Mesostigma* has been difficult to place accurately as it is expected to be prone to long-branch attraction (Felsenstein 1978; Hendy and Penny 1989). Some phylogenetic analyses have placed it as basal to all other greens

(before the split of Streptophyta and Chlorophyta, see Rogers et al. 2007 and references therein), whereas others (Cocquyt et al. 2009 and references therein) find it is sister to Streptophyta (fig. 1). The data sets used in these previous analyses included nuclear, plastid, and mitochondrial sequences.

Our basic data set for this example has mitochondrial sequences from 13 taxa (Rodríguez-Ezpeleta et al. 2007); it expands (in genes and taxa) the eight-taxon data set of Turmel et al. (2002). The WAG + F + G model used by Rodríguez-Ezpeleta et al. (2007) placed *Mesostigma* as sister to Streptophyta, although the tree was only weakly supported. This was in contradiction to the earlier findings of Turmel et al. (2002), which had placed *Mesostigma* basal to Streptophyta plus Chlorophyta. Nonetheless, because the Rodríguez-Ezpeleta et al. (2007) tree was congruent with their analysis of nuclear data and with previous single gene phylogenies, they concluded that the Turmel et al. (2002) placement of *Mesostigma* basal to Streptophyta plus Chlorophyta was an artifact. After adjusting their data set to the taxa used by Turmel et al. (2002), the authors suggested that the discrepancy was due to sparser taxon sampling combined with a failure to account for rate heterogeneity among sites but that the number of sites used was less important.

Reanalysis of *Mesostigma* Data

The models used in the two original studies (Turmel et al. 2002; Rodríguez-Ezpeleta et al. 2007) differ (JTT and WAG, respectively), and so the trees cannot be compared directly. For that reason, we estimated the phylogeny using PhyML v 3.0 (Guindon and Gascuel 2003) under each of the models, as well as the best-fit model selected by the program ProtTest (Abascal et al. 2005), with all combinations of +F, +I, and +G

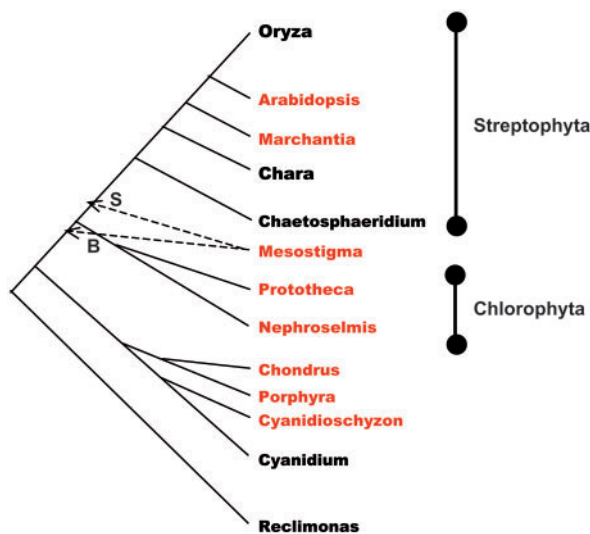


FIG. 1.—The two competing trees. *Mesostigma* is positioned either sister to the Streptophyta (S) or is basal to both Streptophyta and Chlorophyta (B). The taxa included in the eight-taxon data set are marked in red.

(these correspond to estimating the amino acid frequencies, the proportion of invariant sites, and the gamma distribution of rates across sites). The CpREV + F + I + G model was selected as the best-fit model; this is perhaps an unexpected result because the sequences are mitochondrial, but the CpREV model is based on chloroplasts. Other models (e.g., the mitochondrial MtREV) were included in the set of models that were tested but were not selected. The model used in Rodríguez-Ezpeleta et al. (2007) (WAG + G + F) was the fourth best model (with Δ AIC, the difference in Akaike Information Criterion score from the best-fit model, of 886.01). The JTT model used by Turmel et al. (2002) had a Δ AIC of 11,271.59 and was one of the worst-fit models, even when the data set was reduced to their taxon sampling (Δ AIC of 7,394.48). However, the sites in Turmel et al. (2002) are a subset of the sites in our data set.

The 8- and 13-taxon data sets differ in the number of sites containing missing data (defined throughout as either gaps or ambiguous character states). We were interested to determine the effect of different treatments of missing data on tree reconstruction, so we considered five different combinations of taxon and site sampling:

- The 13-taxon data set used by Rodríguez-Ezpeleta et al. (2007).
- The 13-taxon data set with sites containing missing data removed (13-taxon clean).
- The 13-taxon data set reduced to the eight taxa used by Turmel et al. (2002).
- The eight-taxon data set with sites containing missing data removed (eight-taxon reduced then cleaned).
- The 13-taxon clean data set reduced to the 8-taxon sample (eight-taxon cleaned then reduced).

The positions of *Mesostigma* in the resulting phylogenies and relevant bootstrap supports are listed in table 1.

In the 13-taxon data set with all 6,622 sites, only nine trees are ever seen in the bootstrap trees, regardless of the model used. Of these nine trees, only two can be rejected by the approximately unbiased (AU) test of Shimodaira (2002), performed using the program CONSEL (Shimodaira and Hasegawa 2001), under the WAG + F + I + G model (supplementary table S1, Supplementary Material online). Streptophyta is always monophyletic and the basal taxa always group together, but the positions of *Mesostigma*, and the two taxa from Chlorophyta (*Nephroselmis* and *Prototheca*) are uncertain. As our main purpose here is to analyze the effects of site sampling and taxon sampling, for most of our analyses, we make the simplifying assumption that *Nephroselmis* and *Prototheca* form a monophyletic group (Chlorophyta). This leaves only two trees to consider: the tree found by Turmel et al. (2002), which we refer to as the *Mesostigma* basal (the B tree), and the tree found by Rodríguez-Ezpeleta et al. (2007), which we refer to as the *Mesostigma* with Streptophyta (the S tree).

Table 1

The Positioning of Mesostigma in Trees Estimated Using Three Different Models (JTT, WAG, and CpREV) and Combination of +I, +G, and +F

Model	Original 13-Taxa (6,622 Positions)	13-Taxa Clean (1,948 Positions)	8-Taxa (6,622 Positions)	8-Taxa Reduced Then Cleaned (3,910 Positions)	8-Taxa Cleaned Then Reduced (1,948 Positions)
JTT	P (30,8,62)	S (82,1,15)	B (5,94,1)	B (7,91,2)	S (71,22,7)
JTT + F	P (36,12,52)	S (88,0,12)	B (7,89,4)	B (5,95,0)	S (74,11,12)
JTT + I	P (45,4,51)	S (96,0,4)	B (18,78,4)	B (20,78,2)	S (81,8,8)
JTT + I + F	P (57,8,35)	S (89,0,10)	B (26,73,1)	B (29,69,2)	S (76,13,10)
JTT + G	S (70,4,26)	S (83,0,16)	B (25,75,0)	S (47,52,1)	S (74,11,14)
JTT + G + F	S (68,1,31)	S (88,0,11)	S (33,65,2)	S (50,48,2)	S (84,7,7)
JTT + I + G	S (78,2,20)	S (90,0,9)	B (43,54,3)	S (50,49,1)	S (81,10,9)
JTT + I + G + F	S (68,0,32)	S (91,0,9)	S (45,55,0)	S (54,46,0)	S (78,11,11)
WAG	S (26,13,61)	S (83,2,12)	B (6,92,2)	B (6,93,1)	S (66,20,11)
WAG + F	P (28,13,59)	S (79,3,14)	B (8,92,0)	B (8,91,1)	S (67,19,9)
WAG + I	P (44,8,48)	S (93,1,5)	B (18,82,0)	B (26,74,0)	S (77,9,8)
WAG + I + F	P (45,2,53)	S (94,0,6)	B (15,82,3)	B (24,75,1)	S (82,7,10)
WAG + G	S (54,6,40)	S (87,0,13)	B (37,62,1)	S (46,54,0)	S (87,6,6)
WAG + G + + F	S (65,2,33)	S (78,0,21)	S (31,69,0)	S (43,56,1)	S (84,5,9)
WAG + I + G	S (66,2,32)	S (76,0,21)	B (28,71,1)	S (38,62,0)	S (84,7,9)
WAG + I + G + F	S (63,4,33)	S (86,0,10)	S (42,57,1)	S (50,50,0)	S (82,5,10)
CpREV	S (41,11,48)	S (79,3,15)	B (13,84,3)	B (5,93,2)	S (73,17,8)
CpREV + F	P (43,5,52)	S (83,0,15)	B (9,90,1)	B (11,85,4)	S (65,15,18)
CpREV + I	S (52,6,42)	S (90,0,10)	B (12,86,2)	B (22,72,6)	S (62,14,15)
CpREV + I + + F	P (45,5,50)	S (83,1,13)	B (16,81,3)	B (23,73,4)	S (78,3,16)
CpREV + G	S (69,2,29)	S (84,0,14)	S (35,64,1)	S (45,55,0)	S (86,4,10)
CpREV + G + F	S (72,3,25)	S (83,1,16)	S* (39,61,0)	S (41,55,4)	S (80,6,11)
CpREV + I + G	S (71,1,28)	S (90,0,9)	B (44,53,3)	S (42,54,4)	S (74,11,12)
CpREV + I + + G + F	S* (70,1,29)	S* (89,0,9)	S (34,63,3)	S* (38,59,3)	S* (84,3,11)

NOTE.—“S” indicates that Mesostigma is sister to Streptophyta, “B” indicates that Mesostigma is basal to green plants, and “P” indicates that Mesostigma is sister to Prototheca (fig. 1). The best-fit model, found using ProtTest, for each of the settings is marked with an *. Numbers in brackets show bootstrap support for the S split, B split, and P split in turn.

Table 1 presents that both site sampling (i.e., treatment of missing data) and taxon sampling influence tree reconstruction. Our results show that while site sampling only affects tree reconstruction for the 13-taxon data set when the distribution of rate heterogeneity, *G*, is not estimated, the eight-taxon data set is sensitive to site sampling even when *G* is estimated. Exclusion of sites with missing data typically increases support for the positioning of Mesostigma with Streptophyta for both the 8-taxon and 13-taxon data sets, whereas inclusion of sites with missing data results in its basal positioning. In contrast, exclusion of sites in a random manner most frequently does not affect the topology (see [supplementary material text and table S3, Supplementary Material](#) online). We also found that the eight-taxon cleaned then reduced data set places Mesostigma with Streptophyta, regardless of model choice (column 5 in table 1).

Site Likelihoods and Missing Data

The likelihood of a tree for a particular site in an alignment is the probability of that site pattern, given the tree (both topology and edge weights) and a substitution model. The overall likelihood is then the product of site likelihoods or the sum of

site log likelihoods. Normally, only overall likelihoods are considered when determining which tree provides a better explanation for the observed data.

The strong effect of site sampling led us to examine the likelihood of each site for the two competing trees and for each of the five data sets (table 2). For the original 13-taxon data set, there are many more sites supporting the basal positioning of Mesostigma. Nevertheless, Mesostigma on the ML tree is still with Streptophyta. Figure 2 shows histograms of the differences in site log likelihoods, as calculated using the phangorn package in R (Schliep 2010) using the WAG model for the two competing positions of Mesostigma (B and S) for each of the five data sets.

We decided to further investigate the relationship between a site’s preference for the S tree over the B tree and the presence of missing data. We measured preference for the S tree over the B tree by difference in log-likelihood score under the WAG model; this model was chosen as it was one of the models where site sampling caused a change in the preferred tree (table 1). For the eight-taxon data set, we ordered the sites in terms of preference for the S tree and plotted the cumulative proportion of sites with missing data (fig. 3). If data were missing at random with respect to the two

Table 2

Summary of Site Likelihoods Using the WAG Model for the 8- and 13-Taxon Data Sets, with and without the Removal of Missing Data

Data Set	No. of Taxa	Treatment of Sites with Missing Data	Mesostigma Position in ML Tree	No. of Sites Preferring Position S	No. of Sites Preferring Position B	Total Number of Sites	Average Difference in Likelihood between Trees
a	13	Included	S	2,506	4,116	6,622	0.0017
b	13	Excluded	S	<u>1,457</u>	491	1,948	0.0134
c	8	Included	B	2,768	3,854	6,622	0.0074
d	8	Excluded after taxon sampling	B	1,280	<u>2,630</u>	3,910	0.003
e	8	Excluded before taxon sampling	S	<u>1,510</u>	438	1,948	0.0138

NOTE.—“S,” within Streptophyta; “B,” basal to green plants. The Majority of sites are underlined and marked in bold.

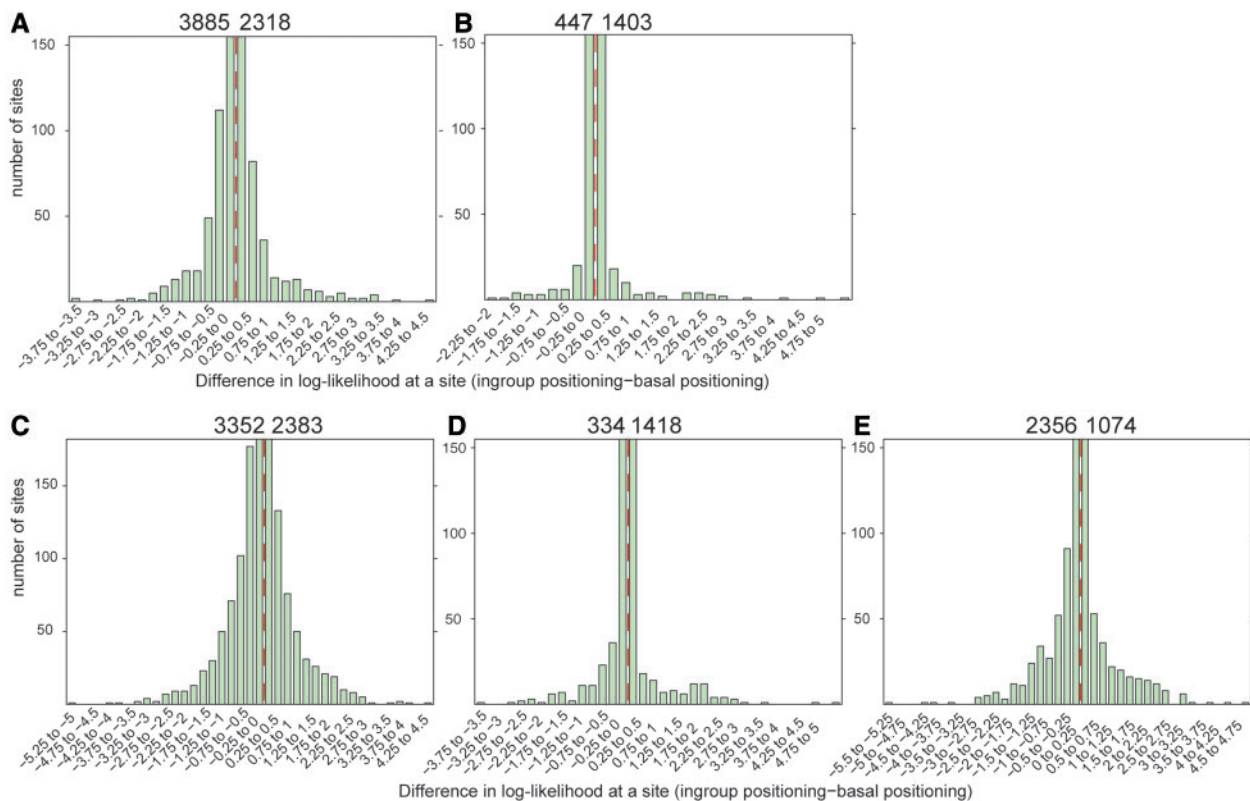


FIG. 2.—Truncated histograms of the differences in site likelihood for the two competing positions of Mesostigma for the five data sets of table 2. For each site, the log likelihoods are calculated for the two positions (S vs. B, see fig. 1) and then subtracted. For example, in (a) most sites (3,885) support position S, but the distribution is not symmetrical; a small number of sites (<1%) support B very strongly and dominate the larger number of sites supporting position S. (b–e) are the other data sets from table 2.

phylogenetic hypotheses (S tree vs. B tree), then we would expect the plot in figure 3 to be a straight line. However, in this case, we can see from the s-shaped curve that sites with the least preference for the S tree are more likely to contain missing data.

Interestingly, we found that for the original 13-taxon data set, approximately 62% of the sites supported the basal positioning of Mesostigma (B), whereas Mesostigma is placed with Streptophyta (S) in the ML tree. The percentage of sites

supporting position S increases when sites with missing data are removed. For the eight-taxon data set with missing data included, approximately 58% of the sites support position B (the ML tree for this data set). With sites with missing data excluded, approximately 78% of the remaining sites support position S, the ML tree in this case. These results, together with the low bootstrap support found by Rodríguez-Ezpeleta et al. (2007), suggest that site sampling is an important and problematic factor in this case.

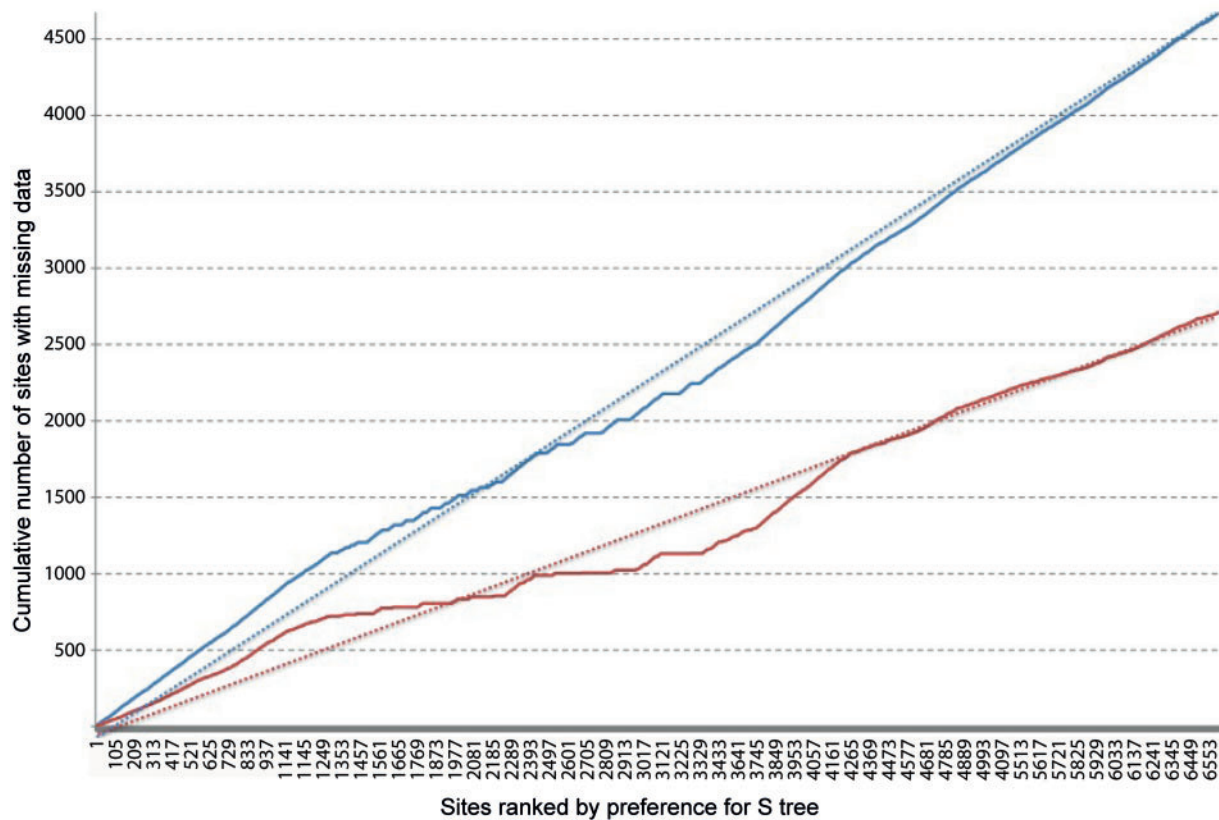


Fig. 3.—Nonrandomness of sites with missing data. Sites in the eight-taxon alignment have been ranked in order of increasing level of preference for the S tree over the B tree (x axis), and the y axis shows the cumulative total number of sites with missing data. The solid blue line records sites with missing data in the 13-taxon data (4,674 in total), and the solid red line records sites with missing data in the eight-taxon data (2,712 in total). Dashed straight lines show the expectation if sites with missing data were allocated randomly with respect to level of preference.

We used a χ^2 test (see [supplementary table S2, Supplementary Material](#) online) to evaluate whether the sites with missing data are randomly distributed with respect to support for the two trees. If the distribution of missing data is not random (an important concept in statistics in general, Allison 2002), the removal of sites with missing data could bias tree inference. Notably, the test rejected the hypothesis that the sites containing missing data are random with respect to their support for the two competing positions for *Mesostigma* ($\chi^2 = 77.99$; $df = 2$; P value = $2.2e^{-16}$). In general, for phylogenetic data, it seems unlikely that data will be missing at random, and therefore, the decision regarding the inclusion or removal of sites with missing data may be very influential. An important implication of this is that the relatively common practice in phylogenetics of removing sites that contain gaps before tree inference may cause a systematic bias.

The nonrandom distribution of missing data is not unique to this data set. We have, for example, considered a data set of Goremykin et al. (2005). For this data set, there is a disagreement regarding the position of grasses within the angiosperm group; figure 2 of the article by Goremykin et al. shows three alternatives. The ML tree using the GTR+I+G model

supports the ingroup positioning of the grasses (fig. 2A of Goremykin et al.), but parsimony and NJ analyses give topologies where the grasses are basal (fig. 2B and 2C of Goremykin et al.). We found that of sites without missing data, 83% preferred the ingroup positioning to the grasses basal position preferred by parsimony analyses (i.e., had a better likelihood), but of sites with missing data, only 58% preferred the ingroup positioning. (Further details on calculation of site likelihoods and the equivalent of figure 3 for the data set of Goremykin et al. are shown in the [supplementary material, table S4](#), and [fig. S1, Supplementary Material](#) online.)

Influential Sites

Bar-Hen et al. (2008) demonstrated that it may not be the majority of sites that determine which tree topology is preferred; sometimes a small number of highly influential sites may drive the results of phylogenetic inference. If these influential sites contain missing data, then the decision on whether to exclude or keep such sites may be the main determinant of the tree topology. In the case of the *Mesostigma* data, it does appear that a small number of highly influential sites seem to

be driving the choice of tree topology (fig. 2). Indeed, the removal of 1% of the sites, those with the highest absolute difference in log likelihood for the two competing positions, results in a phylogeny where *Mesostigma* groups with *Prototheca* (note that this position was not ruled out by an AU test of the original 13-taxon data set, see [supplementary table S1, Supplementary Material](#) online). In addition, using the WAG model, the basal positioning of *Mesostigma* has a higher log likelihood (−890,702.84) than with Streptophyta position (−890,711.16). The results using WAG + F + I + G are similar, although the difference in log likelihoods is smaller (−84,738.51 vs. −84,738.58 for the basal and ingroup positioning, respectively).

Conclusions

The fact that missing data are not random with respect to competing phylogenetic hypotheses means that decisions about how to treat sites with missing data can have a dramatic effect on the estimated phylogeny. In terms of resolving the position of *Mesostigma*, overall our results (particularly the AU test) suggest that this data set is not sufficient to infer the position of *Mesostigma* with confidence.

There are many factors that might explain, or at least correlate with, the presence or absence of missing data in both this and other data sets. Our definition of missing data here includes gap characters, and it seems likely that the process by which indels are generated could differ among sites, for instance, sites with higher overall rate of evolution may also be more likely to contain indels. Other authors have found that removal of fast evolving sites has an effect on phylogenetic inference (Philippe 2000; Goremykin et al. 2010). Given that branch lengths are a product of rate and time, it is possible that missing states are more likely to occur on long branches. This implies that in a heterotachous scenario where different rates apply for different combinations of sites and edges (Philippe et al. 2005; Lockhart et al. 2006), the chance of a site containing missing data might be correlated with the number of edges in the tree with fast rates for that site, and in turn, it might also correlate with any long-branch attraction effects.

This study of *Mesostigma* also highlighted an interesting enigma in that, for the 13-taxon data set there is a difference between the tree supported by the majority of sites and the tree selected by ML. The majority of sites support the basal position for *Mesostigma* as sister to all green plants. Nevertheless, because of a relatively small number of high influence sites, the sum of log likelihoods for the within-Streptophyta position was larger.

We hope that this letter will sharpen the debate on whether or not sites with missing data should be included in phylogenetic analyses. Simulation studies have demonstrated that if the sites that contain missing data are generated by the same process as complete sites, then their inclusion will not be

detrimental to phylogenetic inference (Wiens and Morrill 2011). For real data, the question remains as to whether or not missing data are missing “at random” or if sites with missing data follow a different process to complete sites. Results such as Roure et al. (2013) and Wiens and Tiu (2012) suggest that for many data sets, missing data may not be problematic, but the two examples presented here demonstrate that sometimes it will be. The approaches we put forward for assessing whether or not data are missing at random with respect to competing phylogenetic hypotheses should be a useful step in many phylogenetic analyses.

Supplementary Material

Supplementary figure S1 and tables S1–S4 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org>).

Acknowledgments

This work was supported by the New Zealand Marsden fund (05-MAU-033 to B.R.H.), the Australian Research Council (FT100100031), and the Alexander von Humboldt foundation (postdoctoral fellowship to L.S.G.).

Literature Cited

- Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104–2105.
- Allison PD. 2002. Missing data. Los Angeles (CA): Sage.
- Bar-Hen A, Mariadassou M, Poursat MA, Vandenkoornhuysen P. 2008. Influence function for robust phylogenetic reconstruction. *Mol Biol Evol.* 25:869–873.
- Cocquyt E, et al. 2009. Gain and loss of elongation factor genes in green algae. *BMC Evol Biol.* 9:39.
- Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool.* 27:401–410.
- Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Mol Biol Evol.* 22:1813–1822.
- Goremykin VV, Nikiforova SV, Bininda-Emonds OR. 2010. Automated removal of noisy data in phylogenomic analyses. *J Mol Evol.* 71: 319–331.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.
- Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Syst Zool.* 38:297–309.
- Huelsenbeck JP. 1991. When are fossils better than extant taxa in phylogenetic analysis? *Syst Zool.* 40:458–469.
- Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst Biol.* 58:130–145.
- Lockhart P, et al. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. *Mol Biol Evol.* 23:40–45.
- Philippe H. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc Biol Sci.* 267:1213–1221.
- Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol.* 5:50.
- Rodríguez-Ezpeleta N, Philippe H, Brinkmann H, Becker B, Melkonian M. 2007. Phylogenetic analyses of nuclear, mitochondrial, and plastid

- multigene data sets support the placement of *Mesostigma* in the Streptophyta. *Mol Biol Evol.* 24:723–731.
- Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*. *Mol Biol Evol.* 24:54–62.
- Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. *Mol Biol Evol.* 30:197–214.
- Schliep KP. 2010. phangorn: phylogenetic analysis in R. *Bioinformatics* 27: 592–593.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 51:492–508.
- Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.
- Turmel M, Otis C, Lemieux C. 2002. The complete mitochondrial DNA sequence of *Mesostigma viride*. *Mol Biol Evol.* 19:24–38.
- Wiens JJ, Morrill MC. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst Biol.* 60: 719–731.
- Wiens JJ, Tiu J. 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. *PLoS One* 7:e42925.

Associate editor: David Bryant