Supplementary material:

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 Penny[2] and Barbara R. Holland[3]*

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

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

[3]School of Mathematics and Physics, University of Tasmania, Hobart, Australia.

* Corresponding author

Liat Shavit Grievink, The Edmond and Lily Safra Center for Brain Sciences

The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram,

Jerusalem 91904

Email: liat.shavitgrie@mail.huji.ac.il

Phone: ++972 (054)954 1102

Fax: ++972 (0)2 6226 2410

AU test, Table S1.

Table S1. AU test (Shimodaira ,2002) results for the original 13-taxon data and the 9 different bootstrap trees (using WAG+F+I+G). For simplicity, only the ingroup phylogenies are shown (the outgroup phylogeny was identical in all 9 trees), also we have replaced the subtree (Chaetosphaeridium,(Chara,(Marchantia,(Arabidopsis,Oryza)))) with Streptophyta as this exact subtree occurs in all 9 trees. Obs - the test statistic, au - the *p*-value of the approximately unbiased test, np - the bootstrap probability of the selection. Please see Shimodaira and Hasegawa 2001 for details regarding these values.

trees	obs	au	np
(((Mesostigma,Streptophyta),(Prototheca,Nephroselm))	-4.9	0.798	0.487
((Nephroselm,(Prototheca,(Mesostigma,Streptophyta)))	4.9	0.524	0.212
((Prototheca,(Nephroselm,(Mesostigma,Streptophyta)))	8.1	0.344	0.125
((Nephroselm,((Mesostigma,Prototheca),Streptophyta))	17.7	0.275	0.085
(((Mesostigma,Prototheca),(Nephroselm,Streptophyta))	23.5	0.232	0.059
(((Nephroselm,(Mesostigma,Prototheca)),Streptophyta)	28.5	0.081	0.014
((Mesostigma,(Streptophyta,(Prototheca,Nephroselm)))	23.3	0.073	0.018
((Mesostigma,(Prototheca,(Nephroselm,Streptophyta)))	34.5	0.036	0.005
((Mesostigma,(Nephroselm,(Prototheca,Streptophyta)))	52.0	9e-07	2e-06

Chi square test, Table S2.

Table S2. Data used for the Chi square test showing that sites are not missing at random (13-taxon dataset).

	No data missing	Missing data only in the 13-taxon dataset	Missing data in the 8-taxon (& 13) dataset	Total
Support S	590	750	1166	2506
Support B	1358	1212	1546	4116
Total	1948	1962	2712	6622

Testing removing sites at random

We explored two ways of producing an 8-taxon dataset without missing data - either starting by removing all sites with missing data from the 13-taxon alignment and then reducing to 8 taxa (row e of Table 2), or first reducing to 8 taxa and then removing sites with any missing data (row d of Table 2). The first option gave an alignment 1948 sites and, in most cases, a resulting tree with Mesostigma in the S position, whereas the second option gave an alignment of 3910 sites with, in most cases, Mesostigma in the B position (see Table 1). Tree reconstruction for the 8-taxon dataset (row c of Table 2) with missing data included resulted, in most cases, in a tree where Mesostigma is in the B position.

To explore if there was anything special about the 4674 sites with missing data and particularly about the 1962 sites that are in alignment e but not in alignment d we formed 100 alignments by deleting 4674 and 1962 sites at random from alignment c and d, respectively. Trees were then reconstructed using PhyML v3.0 under the models for which the use of different alignments altered the placement of Mesostigma. In most cases the random removal of sites did not alter the basal positioning (B) of Mesostigma (see Table S3). This suggests

that there may be something special about the sites with missing data, and particularly about the 1962 sites that are in alignment d but not in alignment e.

Table S3. Tree reconstruction results of datasets where sites were randomly removed from either (A) the 8-taxon data or (B) the 8-taxon data where sites with missing data were removed after taxon sampling. In both cases sites were randomly removed until the total length of the alignment was equal to that of the 8-taxon data where sites with missing data were removed prior to taxon sampling (1948 sites).

Model	8-taxon (6622 sites)	8-taxon reduced then cleaned (3910 sites)	8-taxon cleaned then reduced (1948 sites)	Random removal (A) (4674 sites removed)			Random removal (B) (1962 sites removed)		
				B	S	P	B	S	P
JTT	B	B	S	**80**	16	4	**91**	7	2
JTT+F	S	B	S	**77**	18	5	**88**	7	5
JTT+I	B	B	S	**73**	23	4	**70**	25	5
JTT+I+F	B	B	S	**69**	24	7	**67**	25	8
JTT+G	B	S	S	**62**	35	3	**-**	-	-
JTT+I+G	S	B	S	**54**	40	6	47	**52**	1
WAG	B	B	S	**83**	15	2	**93**	6	1
WAG+F	B	B	S	**82**	15	3	**91**	6	3
WAG+I	B	B	S	**73**	22	5	**71**	28	1
WAG+I+F	B	B	S	**66**	24	10	**69**	26	5
WAG+G	B	S	S	**60**	36	4	**-**	-	-
WAG+I+G	S	B	S	**56**	39	5	47	**51**	2
cpREV	B	B	S	**81**	15	4	**89**	8	3
cpREV+F	B	B	S	**76**	18	6	**81**	11	8
cpREV+I	B	B	S	**69**	22	9	**65**	31	4
cpREV+I+F	B	B	S	**61**	27	12	**61**	25	14
cpREV+I+G	S	B	S	**51**	40	9	43	**52**	5

Analysis of Goremykin et al. (2005) data

We measured the preference that each site in the alignment had for the tree shown in Fig 2A of Goremykin et al. (2005) versus the tree shown in Fig 2B (Table S4). Preference was the difference in site likelihoods under the GTR+I+G model. Specific parameters for this model were chosen by optimising the parameters on the NJ tree (fig 2C of Goremykin et al. 2005). We tested to check that these parameter values didn't differ too greatly depending on the tree they were optimised on and found that the differences in parameter values using any of tree 2A, 2B or 2C from Goremykin et al. (2005) where typically in the 3rd decimal place and never greater than 0.05.

Table S4

	Prefer grasses basal	Prefer *Amborella* + *Nymphaea* basal	Total
Sites with no missing data	9023 (17%)	45154 (83%)	**54177**
Sites with missing data	14648 (42%)	20611 (58%)	**35259**
Total	**23671**	**65765**	**89436**

We also repeated the analysis that lead to figure 3 in the main text for the Goremykin et al. (2005) dataset (Figure S1). Sites are ranked by preference for the tree in Fig 2A of Goremykin et al. over the tree in Fig 2B. The distribution of sites with missing data is highly non-random with respect to this ordering of sites.

Figure S1