ZOL855 F11 Data Analysis Project
December 15, 2011

HIGH EVOLUTIONARY DIVERGENCE OF GRASSES EXPLAINING THEIR HIGH DEGREE OF DIVERSITY TOWARDS PALM TREES

Sebastian Raschka

Introduction

A short while ago, researches initiated a whole genome *de novo* sequencing approach of *Phoenix dactylifera*. Since the date palm is an important crop for agriculture, the major goal was to lay the foundation for further experiments engaging the improvement of fruit quality, ripening time and early gender determination based on DNA markers (Al-Dous *et al.* 2011).

Furthermore, the authors exhibited a cladogram in their research article, which shows the taxonomy of all fully sequenced crops, based on taxonomy numbers from NCBI, including the latest sequenced *Phoenix dactylifera* (Fig. 1). To their surprise, the authors observed during the genome annotation process in a BLAST search that *Phoenix dactylifera* shares a higher similarity with the dicotyledonous *Vitis vinifera* (9,022 matches) in terms of predicted proteins, than it does with the *Oryza sativa* (5,094 matches) (Al-Dous *et al.*, 2011).

At a first glance, this observation is indeed astonishing, because one would expect that the monocotyledonous date palm is more similar to the monocotyledonous rice, than to the dicotyledonous grapevine, because of their relatedness with respect to the taxonomic tree.

However, as far as it is known, the authors didn't pursue this matter in scientific manner any further, yet (Al-Dous *et al.* 2011).

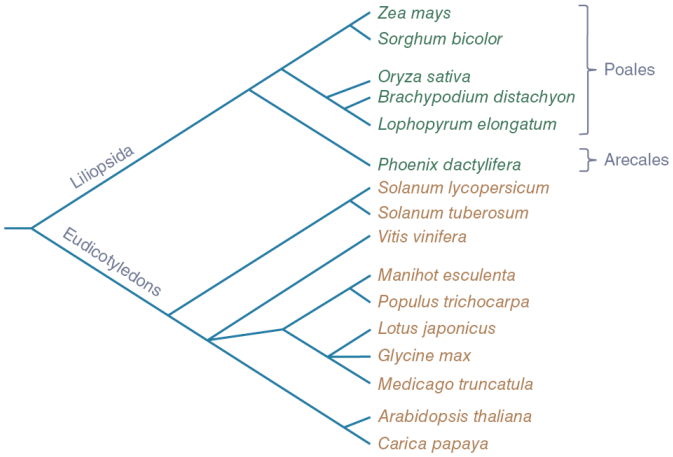

Fig. 1. Taxonomic tree of all crop taxa with publicly available genomes presented in the research paper of the whole genome sequencing of data palm (Al-Dous *et al.*, 2011).

As described by Douglas J. Futuyama (Futuyama, 2004) it is a common misconception to assume that most similar species are necessarily also most closely related to each other. In fact, relationships of taxa are described by common ancestry, which is inferred by shared derived character states, and not by overall similarity of all character states. In terms of the principle of "relatedness through common ancestry" a suitable hypothesis comes up claiming the high

evolutionary divergence of the *Poales*, which might explain the unexpected high degree of dissimilarity towards the date palm.

In order to investigate the divergence of the grasses, phylogenetic analyses with focus on phylogenetic distance measures were conducted, together with character tracings in order to identify the synapomorphies that give support for monophyletic groups of monocotyledonous plants and dicotyledons, respectively.

Therefore, conserved marker genes were chosen, which proved to be reliable in terms of reconstructing the angiosperm phylogeny in prior studies (Soltis *et. al.* 2011). A secondary selection criterion was the availability of these marker genes for the taxa of interest. On that account the choice was narrowed down to the *atpB* gene, a plastidial gene on the large single-copy region, which is coding for the beta subunit of the ATP synthase (Bovenberg et. al., 1984). Additionally, a nuclear gene - the 18S rDNA - coding for the 18S ribosomal RNA was used for the phylogenetic analyses. Although, it is a conserved region and a commonly used gene for plant phylogenetic analyses, a few limitations and shortcomings were previously described for this region. Within the 18S rRNA significant differences between the substitution rates of the more conserved stem regions of the RNA structure, and the less conserved loop regions were menbtioned. Thus, the 18S rDNA might contain less phylogenetic information than the *atpB* coding sequence (CDS), and was shown to have problems to resolve certain angiosperm phylogenies (Soltis *et al.*, 1999).

Material and Methods

Computer Software

Multiple sequence alignments, maximum parsimony analyses, neighbor-joining analyses, maximum likelihood analyses, and the referring phylogeny tests via bootstrap methods were executed in MEGA 5.05 (Tamura *et al.* 2011). The Bayesian Inference analyses were accomplished in MrBayes 3.20 (Ronquist and Huelsenbeck, 2011).

MacClade 4.08a was used for analyses of character evolution (Maddison and Maddison, 2005), and for Fu and Li's test DnaSP 5 was conducted (Librado and Rozas, 2009). Phylogenetic trees were visualized via FIGTree v. 1.3.1 (Morariu *et al.* 2008), where branch support values, order and group labels, and phylogenetic distances were added manually.

Taxa and Marker Genes

In sum, 19 different angiosperm species were used (Tab. 1) for these analyses, with three gymnosperm species as outgroup: *Welwitschia*, *Ginkgo*, and *Pinus*. Outgroups species were chosen on the basis of previous angiosperm studies (Soltis *et al.* 2011). The data set consisted of 1752 aligned characters per taxon for the 18S rDNA gene, and 1505 aligned character for the coding sequence of plastidial *atpB* gene; both genes were analyzed individually. For some species that were included in the analyses with the *atpB* CDS, 18S rDNA sequences were not available. On that account they were replaced by other family representatives. However, in many species the available 18S rDNA

sequences from NCBI were only partially complete. The sequence data were obtained via the Entrez search system from the NCBI database.
(http://www.ncbi.nlm.nih.gov/sites/gquery).

Tab 1. Shown are all taxa used in the phylogenetic analysis. Due to the limited availability of the 18S rDNA sequences in the ENTREZ database, some alternative taxa for these analyses had to be chosen. Superscripts behind the taxa names indicate in which analyses the referring taxa were included.

Ingroup	**Division**	**Group**	**Order**	**Species**
	Angiospermae	Monocotyledons	*Poales*	*Zea mays* $^{atpB\ 18S}$
				Oryza sativa $^{atpB\ 18S}$
				Sorghum bicolor $^{atpB\ 18S}$
				Hordeum vulgare ^{18S}
				Brachypodium distachyon ^{atpB}
			Arecales	*Phoenix dactylifera* $^{atpB\ 18S}$
				Phoenix canariensis $^{atpB\ 18S}$
				Arenga pinnata $^{atpB\ 18S}$
				Elaeis guineensis ^{atpB}
				Elaeis oleifera $^{atpB\ 18S}$
		Eudicotyledons	*Vitales*	*Vitis vinifera* ^{atpB}
				Vitis sp. ^{18S}
			Malpighiales	*Manihot esculenta* $^{atpB\ 18S}$
				Populus trichocarpa $^{atpB\ 18S}$
			Fabales	*Lotus japonicus* ^{atpB}
				Glycine max $^{atpB\ 18S}$
				Medicago truncatula $^{atpB\ 18S}$
			Brassicales	*Arabidopsis thaliana* $^{atpB\ 18S}$
				Carica papaya $^{atpB\ 18S}$
Outgroup	*Ginkgophyta*		*Ginkgoales*	*Ginkgo biloba* $^{atpB\ 18S}$
	Pinophyta		*Pinales*	*Pinus sylvestris* ^{atpB}
				Pinus Wallichinia ^{18S}
	Gnetophyta		*Welwitschiales*	*Welwitschia mirabilis* $^{atpB\ 18S}$

Multiple sequence alignments

For all alignments the algorithm MUSCLE within the software MEGA5.05 was used, with the suggested default parameters as described by Barry G. Hall (Hall, 2011). In terms of the *atpB* CDS codons were aligned, and the protein sequences were checked for correctness with reference to the coding frame. For the genomic 18S sequence an ordinary nucleotide alignment was conducted. Due to comparisons to alternative ClustalW alignments via the publicly available free online alignment comparison server AltAVisT (http://bibiserv.techfak.uni-bielefeld.de/altavist/), the MUSCLE alignment seemed to be more accurate. After the alignment was accomplished excessing sequences were manually trimmed to common start and end points. For the *atpB* CDS the the aligned sequences began with the start codon.

Additionally, the reliabilities of the *atpB* and 18S alignments were checked by using the free online server GUIDANCE (http://guidance.tau.ac.il/).

Maximum Parsimony Analysis

For the maximum parsimony analysis all sites were used (including all gaps and missing data from the alignment). The treespace has been searched searched via Close-Neighbor-Interchange method with 100 random initial starting trees at search level 3 for the most parsimonious trees. The phylogeny was tested using the bootstrap method with 2000 pseudoreplicates. Steps were not weighted (default weight = 1).

Neighbor-Joining Analysis

Before conducting a neighbor-joining analysis, the aligned data sets were tested for suitability in terms of calculating the overall mean distances as suggested by the authors of MEGA 5.05 (Nei and Kumar, 2000). As evolutionary model the maximum composite likelihood model was chosen, which was also suggested by Sudhir Kumar (Hall, 2011). For the test of phylogeny a bootstrapping with 2000 pseudoreplicates was performed. The gaps and missing data from the alignment were treated with pairwise deletion during these analyses.

Maximum Likelihood Analysis

Appropriate evolutionary models for the maximum likelihood analyses were inferred by the model test implemented in MEGA 5.05. On the basis of these results the General Time Reversible evolutionary model with Gamma-distributed rates was chosen to analyze the *atpB* CDS. For 18S rDNA sequences the Kimura-2-Parameter model with uniform rates was calculated as most appropriate for this data set.

For the heuristic tree search method Nearest-Neighbor-Interchange with an automatically created initial tree was used. and in order to test the phylogeny a bootstrap run with 2000 pseudoreplicates was accomplished.

Bayesian Inference Analysis

For the Bayesian analyses the General Time Reversible model with invariant sites and Gamma-distributed rates was used. The analyses were conducted in two parallel runs using four chains and one million generations, where every tenth generation was sampled. The results were summarized with a burnin value of 25%. The Bayesian analysis of 18S rDNA was stopped at a split-frequency value

of 0.000581, and for the *atpB* at a value of 0.001420, respectively.

Fu and Li's test

Within DnaSP Fu and Li's test with an outgroup was conducted (Fu and Li, 1993) to estimate the number of mutations between selected taxa from the 18S rDNA and *atpB* CDS data sets. As ingroup one representative from the order of *Poales (Oryza sativa)* and *Arecales (Phoenix dactylifera),* and two representatives from the group of the dicotyledons *(Arabidopsis thaliana* and *Carica papaya)* were chosen, based on their sequence completeness. The outgroup was *Ginkgo biloba* in this analysis.

Results

Alignments

Via a two nucleotide sequences BLAST alignment (Altschul *et al.*, 1998) the similarity of the atpB CDSs between *Vitis vinifera*, *Phoenix dactylifera*, and *Oryza sativa* was investigated. The result was that the grapevine shares a sequence similarity of 93% with date palm, where the similarity between the more closely related species, date palm and rice, is only 91%. This finding is consistent with the findings of the previous study of Al-Dous *et al.*, where the high protein similarity between the more distantly related species *Phoenix dactylifera* and *Vitis vinifera* was described (Al-Dous *et al.*, 2011). The comparison of the 18S rDNA sequences in this manner was waived, because of the incompleteness of the 18S sequence from *Vitis sp.*

When the MUSCLE alignment was done, a manual investigation revealed that most of the sequences aligned well. For the *atpB* sequences only one gap (six residues in length) was observed. The 18S sequence alignment showed more gaps, due to the incompleteness of the sequences for some taxa.

From the GUIDANCE analysis the resulting overall alignment scores for *atpB* and 18S were both 0.99, where 1.0 represents highest confidence in the alignment (Penn *et al.*, 2010). It was remarkable the fact, that the aligned 18S sequence of Arenga pinnata showed preternaturally many singletons, which might be due to sequencing errors.

atpB Analyses

For the analyses of the *atpB* CDS it was possible to construct the phylogenetic trees (Fig. 2-5) with high statistical support for the major monophyletic groups: *Dicotyledons, Monocotyledons, Arecales,* and *Poales.* However, within the clade *Arecales* the statistical support for the resulting phylogeny is low, and basically, only given for grouping the two *Phoenix* species as sister taxa. Taken together all used analyses resulted in the construction of the same phylogeny. Consistent phylograms from the neighbor-joining, maximum likelihood, and Bayesian analysis show a higher evolutionary distance for the *Poales* clade as for the Arecales clade, outbound from their assumed most common ancestor.

The maximum parsimony analysis resulted in one single most parsimonious tree (Fig. 3) with a treelength of 1185 steps (CI index = 0.61, RI index = 0.71). When the *Arecales*

clade was manually moved onto the dicotyledon branch, the treelength increased by 11 steps up to 1196 (CI index = 0.61, RI index = 0.70). The analysis of character phylogeny revealed that in sum 28 synapomorphic characters give support for the monocot clade, and 14 synapomorphies are unique for the dicot clade (Fig. 3). Character 66 was traced on the most parsimonious tree (Fig. 6 A), and the tree with the branches of the Arecales clades swapped (Fig. 6 B). Upon the branch swapping from the most parsimonious tree, the character CI decreased from 0.5 to 0.33, and the RI decreased from 0.8 to 0.78. The summary of the character evolution of the two trees (Fig. 7) shows that the most parsimonious tree is for 23 character shorter than the swapped tree, where the swapped tree is just for twelve characters shorter.

Analyses of 18S rDNA

In all used phylogenetic methods the phylogeny of the main phylogenetic groups could be reconstructed as described for the *atpB* analyses. However, the statistical support at deeper resolutions is mostly weak or even not given. Only the Bayesian analysis constitutes an exception having strong statistical support throughout, even at deeper resolutions. Consistently, all phylograms (Fig. 8, Fig. 10-11) show a high evolutionary distance for *Arenga pinnata*. The maximum parsimony analysis resulted in two most parsimonious trees (Fig. 9 shows the consensus tree), with treelengths of 592 steps (CI = 0.73, RI = 0.68). The difference between the both most parsimonious trees was either the grouping of *Arenga pinnata* as sister taxa to *Elaeis oleifera* one the one MP tree, or as sister taxa to the *Phoenix* species on the other MP tree. The tracing of character 652 in those MP trees is shown in Fig. 12. On one tree the character evolved with three steps with CI = 0.5 and RI = 0.8 (Fig 12 A), where it evolved with 4 steps on the second MP tree with CI = 0.33 and RI = 0.78 (Fig. 12 B). The summary of all characters that evolved different in the two MP, i.e. two, are shown in Fig. 13. Six synapomorphic characters for the monocot group, and seven synapomorphic characters for the dicot group were identified (Fig. 9). Five out of the six synapomorphic characters from the monocot group evolved from adenosine or thymine to guanosine or cytosine. On the other side five of seven synapomorphic characters of the dicot clade evolved from cytosine to thymine. The diagram in Fig. 14 summarizes the GC contents of the 18S and *atpB* sequences for the monocots and dicots. Where the average GC contents between the analyzed monocots and dicots are almost equal, a noticeable difference between the GC contents of 18S (monocots: 50.9%, dicots: 49.6%) sequence and *atpB* (monocots: 42.6%, dicots: 42.0%) was observed.

Fu and Li's test

Fu and Li's test with an outgroup revealed a total number of mutations of 261 of 1455 compared sequences (17.9%) for the *atpB* coding sequence. For 18S sequences the total number of mutations was 120 out of 1639 (7.3%).

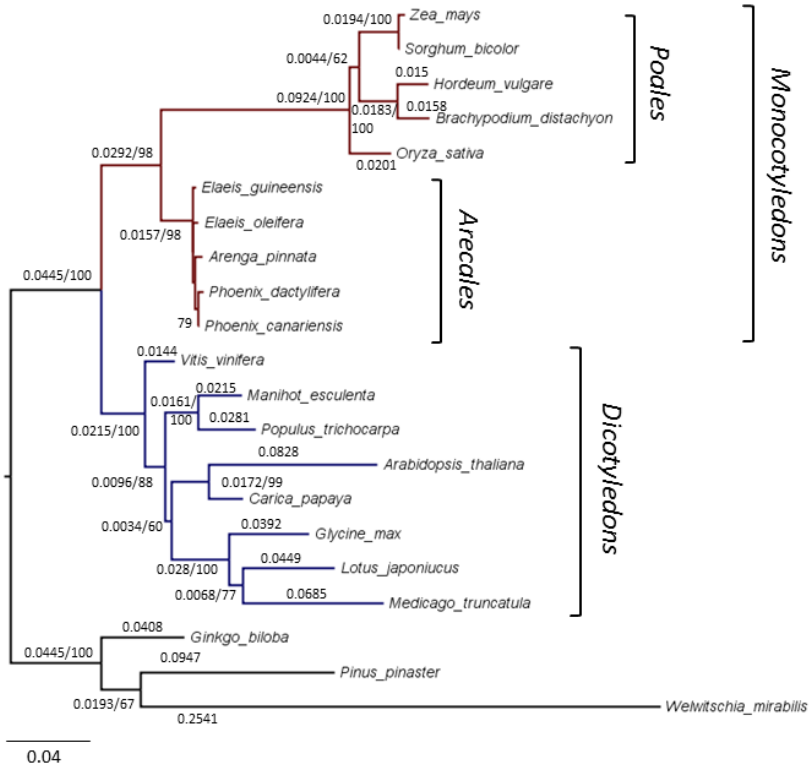

Fig. 2. Phylogram from the Neighbor-Joining analysis with *atpB* CDS. Decimal numbers on the branches represent branch lengths as phylogenetic distance measure (first number), and integers show bootstrap percentages (second number). Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. The branches of the monocotyledon clade are colored in blue and the dicotyledon clade in red, respectively.

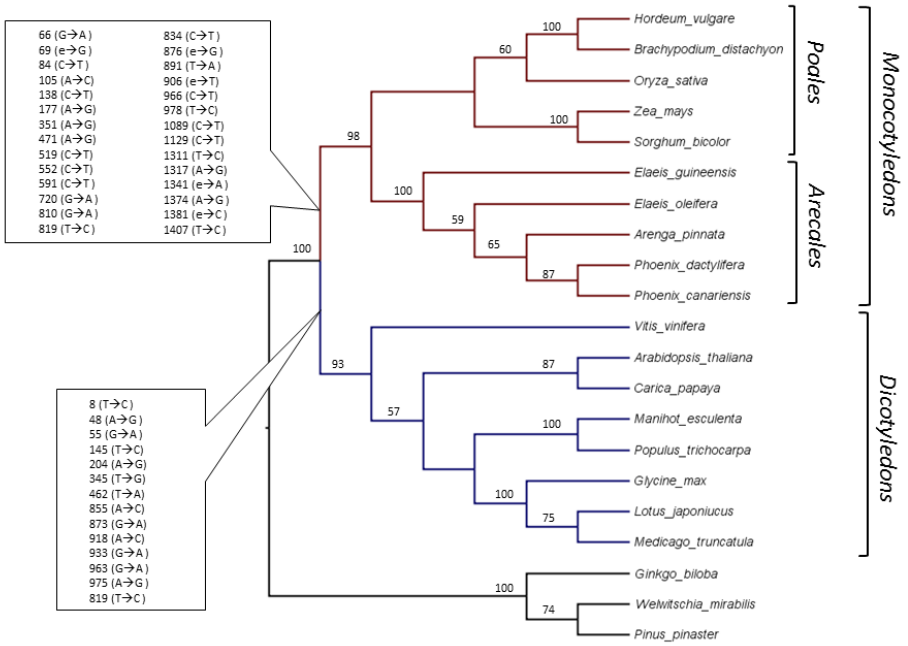

Fig. 3. Most parsimonious tree from the Maximum Parsimony analysis. with *atpB* CDS. Numbers on the branches are bootstrap percentages. Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. The branches of the monocotyledon clade are colored in blue and the dicotyledon clade in red, respectively. Callboxes on the left show the character positions that are synapomorphies for the monocots and dicots with their state change in referred to assumed ancestor state.

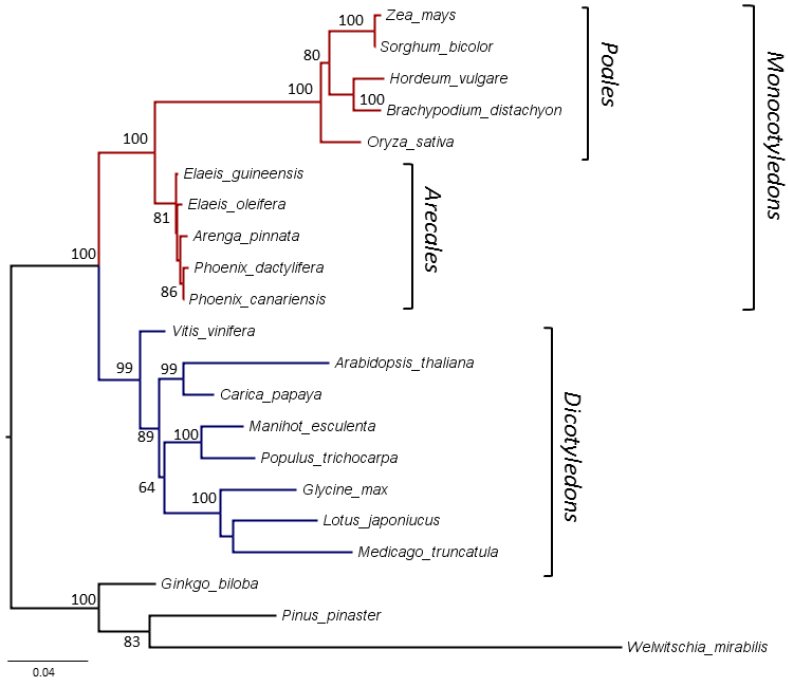

Fig. 4. Phylogram from the Maximum Likelihood analysis. with *atpB* CDS. Numbers on the branches are bootstrap percentages. Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. Callboxes on the left show the character positions that are synapomorphies for the monocots and dicots with their state change in referred to assumed ancestor state.

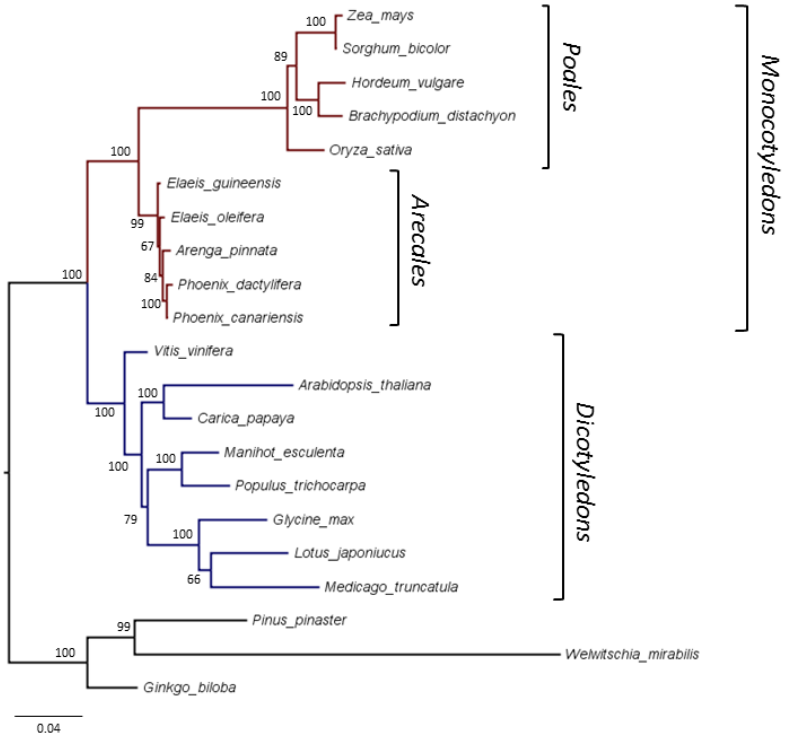

Fig. 5. Phylogram from the Bayesian inference analysis with atpB *CDS*. Numbers on the branches represent Bayesian support values in percent. Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. The branches of the monocotyledon clade are colored in blue and the dicotyledon clade in red, respectively.

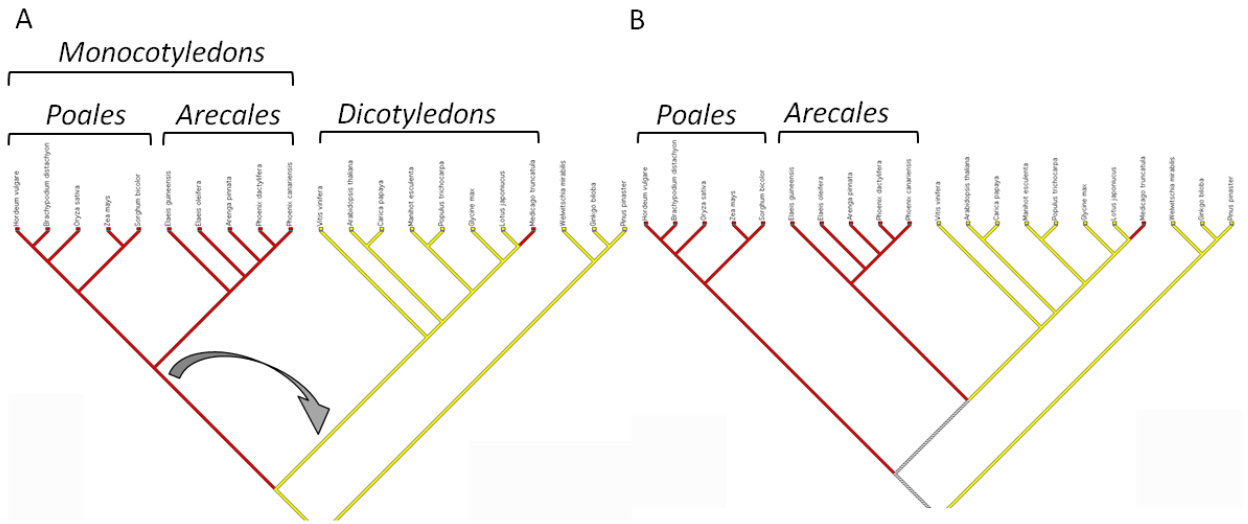

Fig. 6. Cladogram of the most parsimonious tree (A) from the *atpB* maximum parsimony analysis. The tree on the right is the result of a manual branch swapping, where the *Arecales* clade was moved into the dicotyledon group (B). The colors show the phylogeny of character 66. Character state guanine is shown in yellow, adenosine in red, respectively.

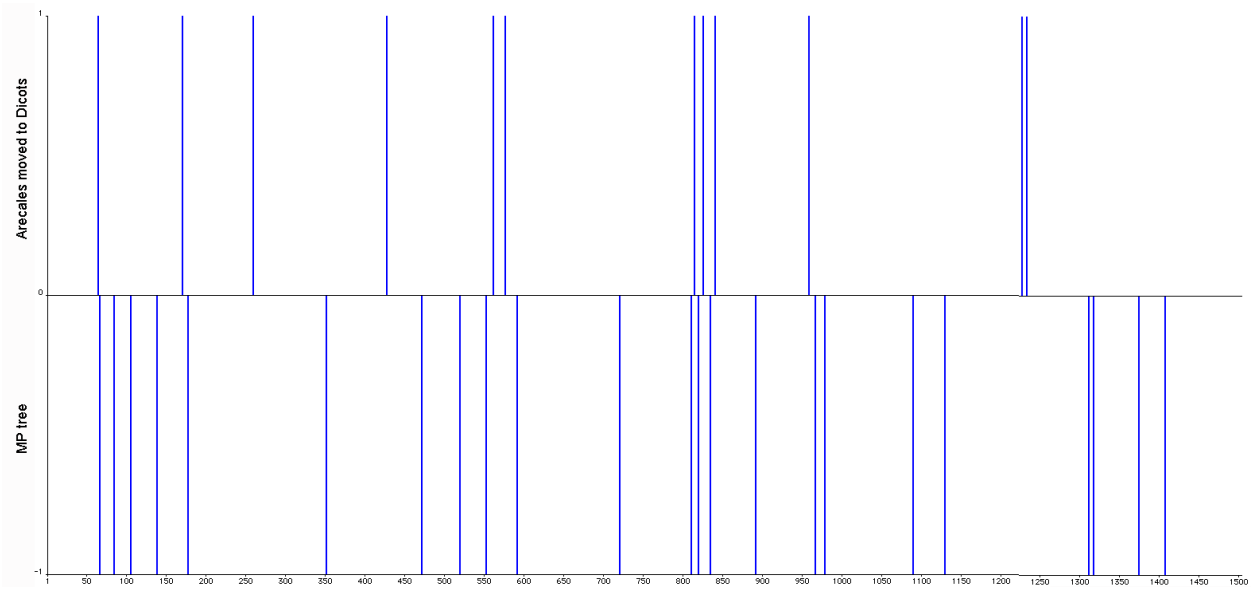

Fig. 7. Overview of all characters that evolve with fewer steps either on the most parsimonious tree (lower panel) or in the tree after the branch swapping (top panel) from Fig. 6. The referring character position is shown on the abscissa.

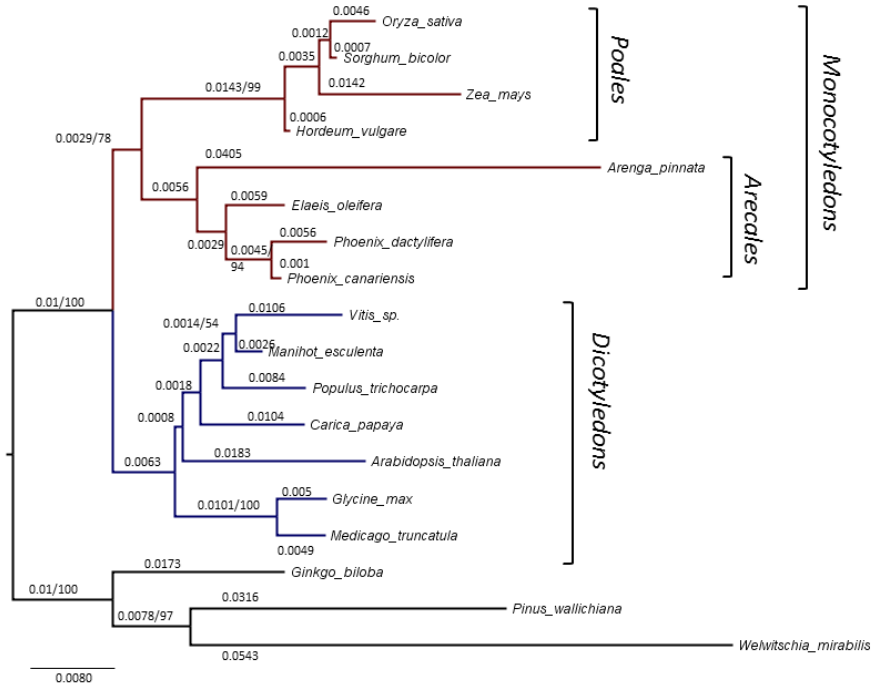

Fig. 8. Phylogram from the Neighbor-Joining analysis with 18S rDNA. Decimal numbers on the branches represent branch lengths as phylogenetic distance measure (first number), and integers show bootstrap percentages (second number). Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. The branches of the monocotyledon clade are colored in blue and the dicotyledon clade in red, respectively.

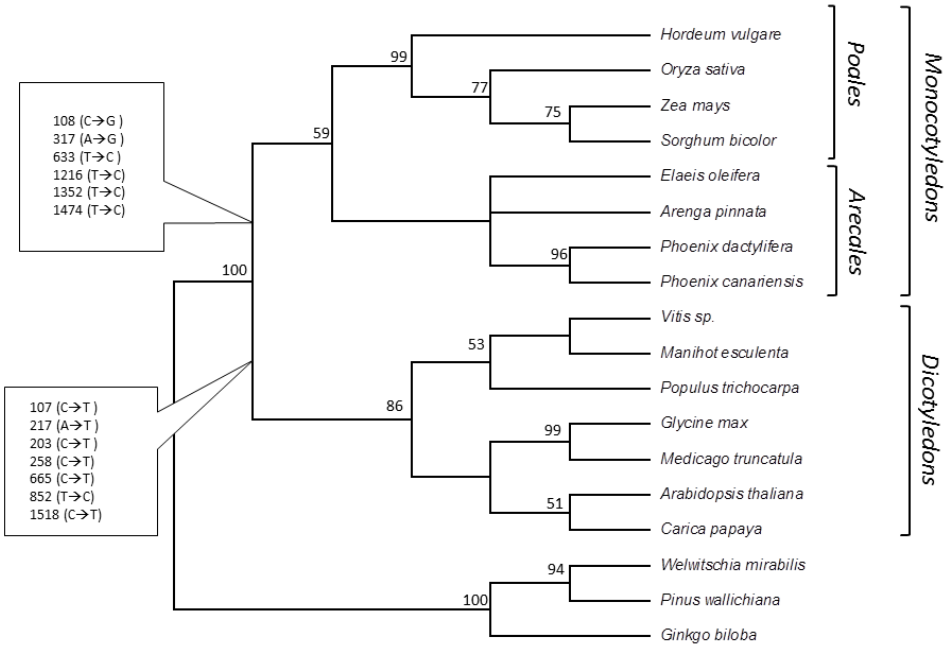

Fig. 9. Consensus tree from the Maximum Parsimony analysis. with 18S rDNA. Numbers on the branches are bootstrap percentages. Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. Callboxes on the left show the character positions that are synapomorphies for the monocots and dicots with their state change in referred to assumed ancestor state.

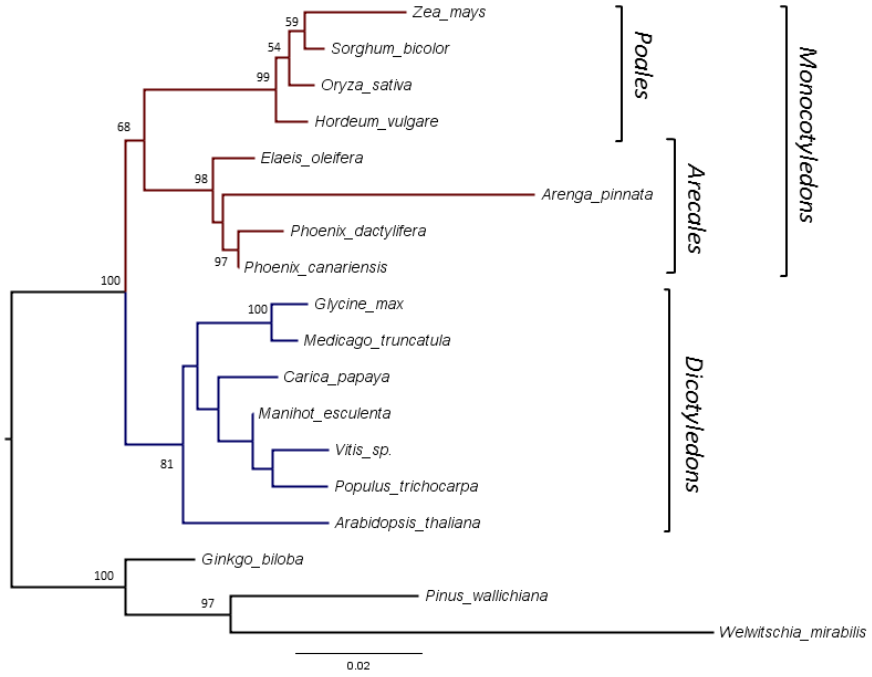

Fig. 10. Phylogram from the Maximum Likelihood analysis. with 18S rDNA. Numbers on the branches are bootstrap percentages. Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. Callboxes on the left show the character positions that are synapomorphies for the monocots and dicots with their state change in referred to assumed ancestor state.

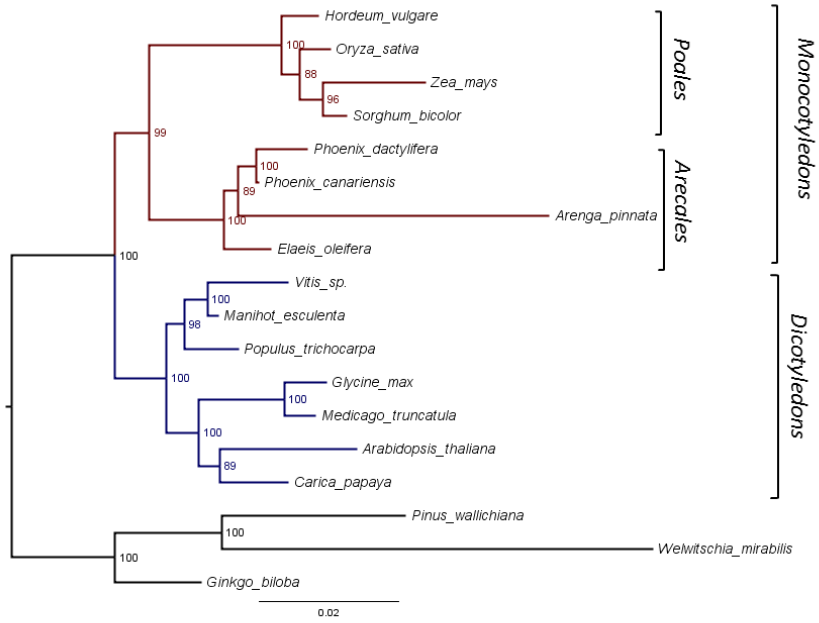

Fig. 11. Phylogram from the Bayesian inference analysis with 18S rDNA. Numbers on the branches represent Bayesian support values in percent. Brackets on the right condense the analyzed taxa to their associated orders and groups, respectively. Order and group names are based on taxonomies from NCBI. The branches of the monocotyledon clade are colored in blue and the dicotyledon clade in red, respectively.

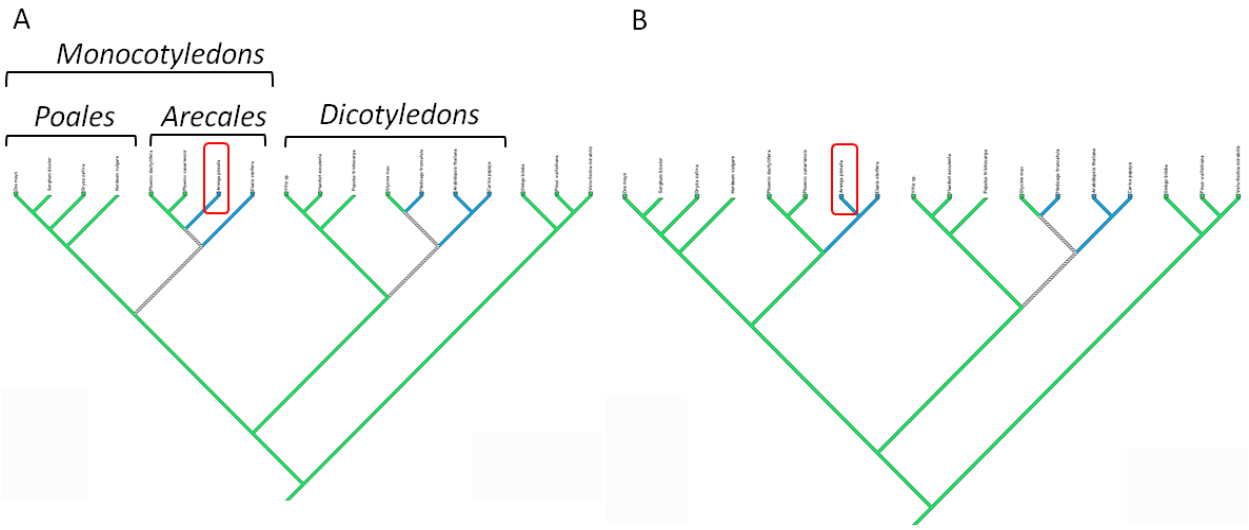

Fig. 12. Cladograms of the most parsimonious trees from the 18S maximum parsimony analysis. Shown is the phylogeny of character 652. The Character state cytosine is shown in green, thymine in blue, respectively. The difference between the two trees is the grouping of *Arenga pinnata* (red box) either as sister taxon of *Phoenix* (A) or as sister taxon of *Elaeis olifeira* (B).

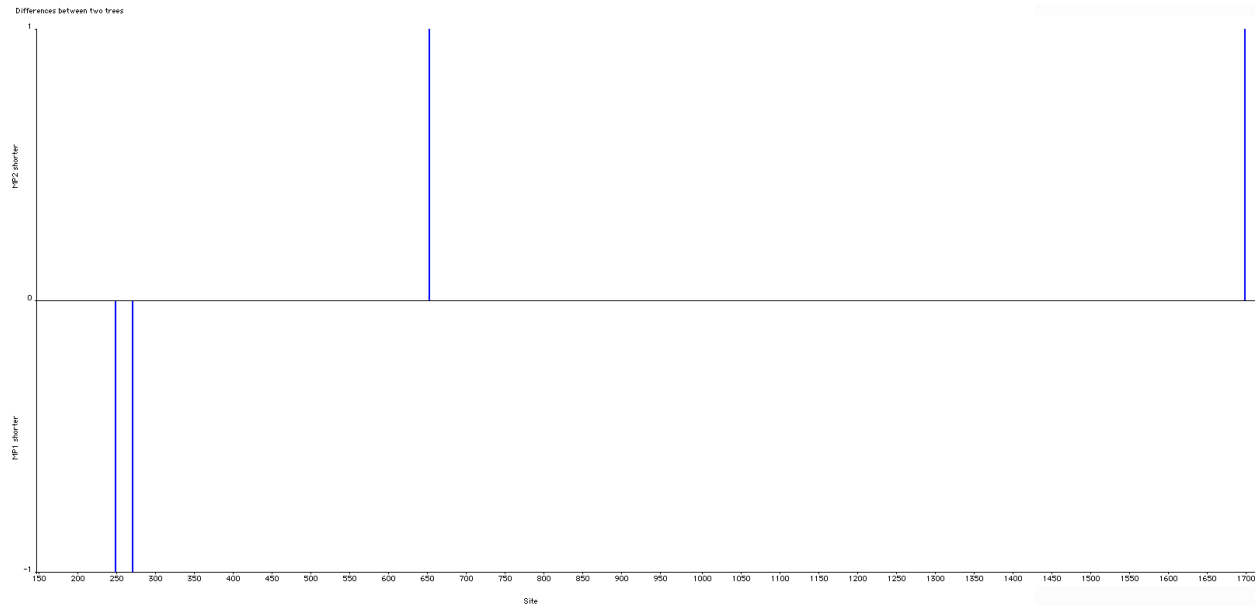

Fig. 13. Overview of all characters that evolve with fewer steps either one of the two most parsimonious trees. Upper panel shows these characters for the tree from Fig. 12 B, the lower panel for the tree from Fig. 12 A, respectively.

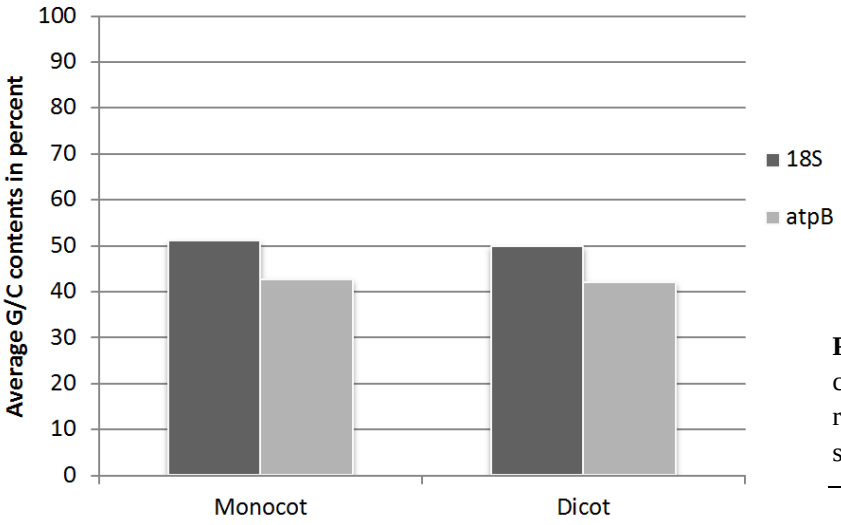

Fig. 14. Diagram shows the average GC contents for monocots and dicots with respect to the 18S and the *atpB* sequences.

Discussion

High bootstrap values for most of the monophyletic groups from the *atpB* analyses provide strong evidence that the phylogeny of the analyzed taxa was inferred correctly. The posterior probabilities from the Bayesian analysis are consistent with this observation. Only within the Arecales clade the statistical support values are relatively low or are even non-existent. This might be due to the fact that the evolutionary information from the *atpB* gene was not sufficient to reconstruct clear taxa relationships. The phylograms from the Bayesian, maximum likelihood, and neighbor-joining analyses are showing very short branches within the *Arecales* clade, which represent the low evolutionary distance between the palms with respect to the *atpB* CDS. The maximum parsimony analysis resulted in just one single MP tree, which supports the assumption that the tree topology is likely to be inferred correctly. The branch swapping, where the *Arecales* order was moved to the dicotyledons, shows a significant cost of character steps, as well as a decrease of the tree RI and CI, thus an alternative topology like this is highly unlikely.

Interestingly the *Poales* order shows a high divergence (Fig. 2, Fig. 4-5), in contrast to the *Arecales* order. This high evolutionary distance is consistent with the observation that *Phoenix dactylifera* shares a lower amino acid identity with *Oryza sativa* than with the more distantly related *Vitis vinifera*.

However, the high evolutionary distance of *Poales* in comparison to the *Arecales* order is less pronounced in the phylograms of the 18S analyses. Surprisingly, *Arenga pinnata,* a member of the palm family, shows by far the highest evolutionary distance within the angiosperms. This might be explained by the fact that the corresponding sequence has a high amount of singletons in comparison to the other sequences, which is supposed to be due to sequencing errors. The maximum parsimony analysis resulted in two MP trees, which only differed with respect of the

grouping of *Arenga pinnata*, supporting the assumption about the issue with this sequence.

When the overall mutation rates within the 18S and atpB sequences were inferred, for 18S the rate was unequivocally lower (7.3% to 17.9%), suggesting that 18S contains less evolutionary information for the phylogenetic analysis of these taxa than *atpB* does. A previous study about the limited utility of the 18S rDNA signal for phylogenetic analysis of land plants is consistent with this hypothesis (Soltis *et al.*, 1999)

After comparing the synapomorphic characters that support the dicot clade with respect to the 18S sequence, a change towards a higher GC content was initially suspected. After comparing the average GC contents between monocots and dicots then, this suspicion could be averted. But interestingly the 18S sequence is supposed to have a higher GC content than the *atpB* region. Because 18S is coding for a RNA, which is a structural part of the ribosome, this might be explained by a favorable gain in structural stability in paired RNA regions.

Conclusion

The *atpB* gene on the chloroplast seems to be a more useful marker for the inference of angiosperm phylogeny than the nuclear 18S rDNA. The signal from the *atpB* CDS revealed a high evolutionary distance of the *Poales* order, which explains the higher similarity of *Phoenix* to the more distant dicotyledon *Vitis*, than to the more closely related *Oryza*. This research shows that high similarity between species does not necessarily imply that they are also most closely related to each other. But in actual fact, as it was first described by the German biologist Willi Hennig in 1966, shared derived character states are what define the relationship through common ancestry (Futuyama, 2005).

Literature Cited

Al-Dous E. K., George B., Al-Mahmoud M. E., Al-Jaber M. Y., Wang H., Salameh Y. M., Al-Azwani E. K., Chaluvadi S., Pontaroli A. C., DeBarry J., Arondel V., Ohlrogge J., Saie I. J., Suliman-Elmeer K. M., Bennetzen J. L., Kruegger R. R., Malek J. A. (2011) De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nature Biotechnology **29:** 521-527

Altschul S., Madden T., Schaffer A., Zhang J. H., Zhang Z., Miller W., Lipman D. (1998) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Faseb Journal **12:** A1326-A1326

Bovenberg W. A., Howe C. J., Kool A. J., J. N. H. J. (1984) Physical mapping of genes for chloroplast DNA encoded subunit polypeptides of the ATPsynthase complex from Petunia hybrida. Current Genetics**:** 283-290

Chang C.-C., Chen H.-L., Li W.-H., Chaw S.-M. (2004) Dating the Monocot Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes. Journal of Molecular Evolution **58:** 424-441

Futuyama D. Y. (2005) Evolution, 1st Edition. Sinauer Associates, Incorporated, Sunderland, Massachusetts U.S.A

Fu Y. X., Li W. H. (1993) Statistical Tests of Neutrality of Mutations. Genetics **133:** 693-709

Hall B.G. (2011) Phylogenetic Trees Made Easy: A How-To Manual, 4th Edition. Sinauer Associates, Incorporated, Sunderland, Massachusetts U.S.A.

Librado, P. and Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25:** 1451-1452 | doi: 10.1093/bioinformatics/btp187.

Maddison D. R., Maddison. W. P. (2005) MacClade 4: Analysis of phylogeny and character evolution. Version 4.08a.

Morariu V. I., Srinivasan B. V., Raykar V. C., Duraiswami R., Davis L. S. (2008) **Automatic online tuning for fast Gaussian summation.** Advances in Neural Information Processing Systems (NIPS)

Nei M., Kumar S. (2000) Molecular Evolution and Phylogenetics, Oxford University Press, New York U.S.A.

Penn O., Privman E., Ashkenazy H., Landan G., Graur D., Pupko T. (2010) GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Research **38:** W23-W28

Ronquist F., Huelsenbeck J. P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572-1574

Soltis P. S. S. D. E., Wolf P.G., Nickrent D .L., Chaw S.-M., L. C. R. (1999) The Phylogeny of Land Plants Inferred from 18S rDNA Sequences: Pushing the Limits of rDNA Signal? Mol. Biol. Evol. **16:** 1774–1784

Soltis D. E., Smith S. A., Cellinese N., Wurdack K. J., Tank D. C., Brockington S. F., Refulio-Rodriguez N. F., Walker J. B., Moore M. J., Carlsward B. S., Bell C. D., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth C. A., Gitzendanner M. A., Sytsma K. J., Qiu Y. L., Hilu K. W., Davis C. C., Sanderson M. J., Beaman R. S., Olmstead R. G., Judd W. S., Donoghue M. J., Soltis P. S. (2011) Angiosperm phylogeny: 17 genes, 640 taxa. American Journal of Botany **98:** 704-730

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution **28:** 2731-2739

List of Abbreviations

atpB	-	ATPase subunit beta
CDS	-	Coding Sequence
CI	-	Consistency Index
GC	-	guanosine/cytosine
MP tree	-	most parsimonious tree
rDNA	-	ribosomal DNA
RI	-	Retention Index