Genic evidence that gnetophytes are sister to all other seed plants

Yinzhi Zhang[1*], Zhiming Liu[2]

[1]Jiangsu bioinformatics professional committee. [2]Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Science.

*e-mail: raulee@outlook.com.

Abstract

Gnetophytes, comprising three relict genera, *Gnetum*, *Welwitchia* and *Ephedra*, are a morphologically diverse and enigmatic assemblage among seed plants. Despite recent progress on phylogenomic analyses or the insights from the recently decoded *Gnetum* genome, the relationship between gnetophytes and other seed plant lineages is still one of the outstanding, unresolved questions in plant sciences. Here, we showed that phylogenetic studies from nuclear genes support the hypothesis that places gnetophytes as sister to all other extant seed plants and so this hypothesis should not be ruled out according to phylogenetic inference based on nuclear genes. However, this extraordinarily difficult phylogenetic problem might never be solved by phylogenetic inference based gene tree under various artificial selection. Hence, we adopted a novel approach, comparing gene divergence among different lineages, to solve the conflicts by showing that gnetophytes actually did not gained a set of genes like the most recent common ancestor (MRCA) of other seed plants. This distinct gene evolution pattern could not be explained by random gene lost as in other seed plants but should be interpreted by the early divergence of gnetophytes from rest of seed plants. With such a placement, the gymnosperms are paraphyletic and there should be three distinct groups of living seed plants: gnetophytes, non-gnetophytes gymnosperms and angiosperms.

Key words: gnetophytes, seed plants evolution, phylogenetic inferences, gene proliferation

Introduction

Although the living seed plants exhibit a high degree of species richness and extensive morphological variation, extant seed plant taxa represent only a relic of their former diversity. The extensive extinction and long independent evolution and convergence in both morphological characters and genome sequences, has made it extremely difficult to correctly trace phylogenetic relationships among seed plant lineages. This problem has been triggering debates for over several decades and the central issue is the placement of an enigmatic group of gymnosperms called the genophtes (Fig. 1): The 'anthophyte hypothesis' was first proposed because of shared morphological similarities with angiosperms. This hypothesis placed gnetophytes as most closely related to angiosperms (Crane, 1985; Doyle & Donoghue, 1986). However, this hypothesis had been ruled out due to the rise and development of molecular research (Doyle et al., 1994; Wickett et al., 2014). Most molecular studies based on plastid and nuclear sequence have show strong support to the hypothesis that gnetophytes should be placed closely related to conifers. It mainly includes three hypotheses: either as sister to Pinaceae ('Gnepine'), to cupressophytes ('Gnecup'), or to all conifers ('Gnetifer'), depending on the different data sets, sites selection, and inferring methods involved (Ruhfel et al., 2014; Wickett et al., 2014). Moreover, sophisticated phylogenomic analyses based on large-scale data matrices are also pointing towards the hypothesis that gnetophytes are most closely related to conifers (Wickett et al., 2014; Ran et al., 2018). Interestingly, the other two remaining hypotheses, that have received little attention, but have also been constantly supported using different sets of gene sequence and various phylogenomic reconstruction methods, either place gnetophytes as sister to all other gymnosperms (Sanderson et al., 2000; Rai et al., 2008; Li et al., 2017) or even all seed plants (Frohlich & Parker, 2000; Schmidt & Schneider-Poetsch, 2002; Lee et al., 2011; Chen et al., 2016). Notably, the 'Seed plants sister' hypothesis has not been supported with phylogenetic tree inferred from nuclear genes to our current knowledge and tends to be ignored in phylogenetic testing (Wickett et al., 2014).

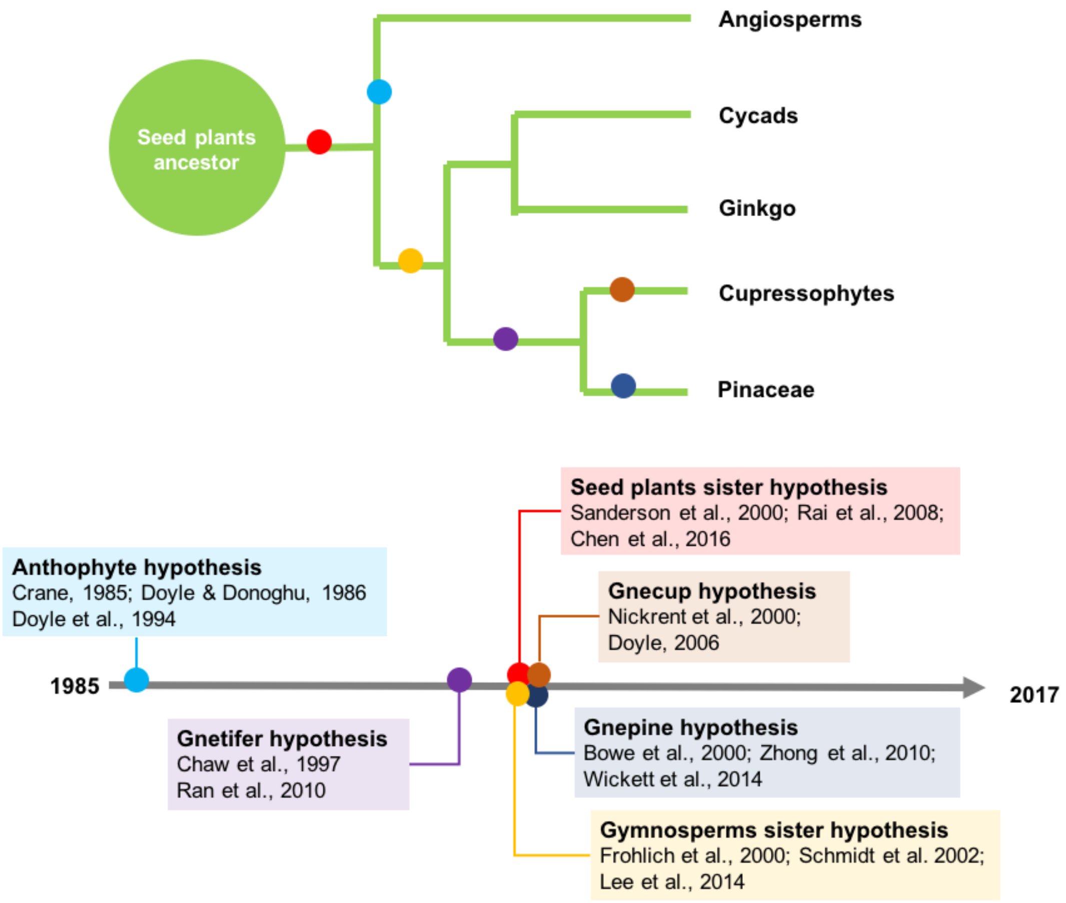

Angiosperms
Cycads
Seed plants ancestor
Ginkgo
Cupressophytes
Pinaceae
Seed plants sister hypothesis
Sanderson et al., 2000; Rai et al., 2008; Chen et al., 2016
Anthophyte hypothesis
Crane, 1985; Doyle & Donoghu, 1986 Doyle et al., 1994
Gnecup hypothesis
Nickrent et al., 2000; Doyle, 2006
1985
2017
Gnetifer hypothesis
Chaw et al., 1997 Ran et al., 2010
Gnepine hypothesis
Bowe et al., 2000; Zhong et al., 2010; Wickett et al., 2014
Gymnosperms sister hypothesis
Frohlich et al., 2000; Schmidt et al. 2002; Lee et al., 2014

Figure 1. Six conflicting hypotheses of the phylogenetic position of gnetophytes inferred by diverse sets of molecular markers or morphological character clustering. The anthophytes hypothesis was first proposed due to morphological character cladistic and once triggered great debates on history. Couple of genes or alignments from large matrices of nuclear or plastid genome compartments were conducted and provide conflicting topologies on phylogenetic positions of gnetophytes.

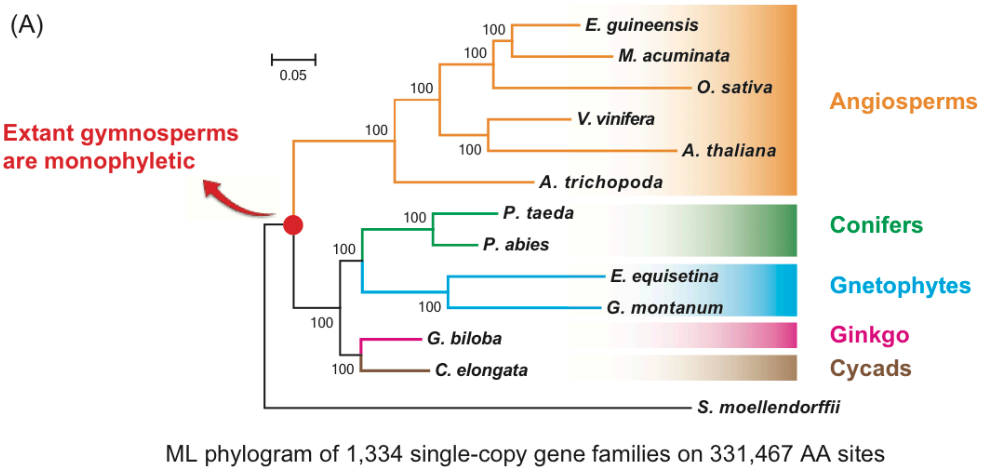
Phylogenetic inferences on nuclear genes

To test the strength of the emerging consensus hypothesis that gnetophytes are closed to conifers, we measured the phylogenetic signals from a data matrix of 13 land plants species (using nine whole genome assemblies and four whole transcriptome data sets, Table 1). We used the whole genome assembly of *G. montanum*, and transcriptomic data from *Ephedra equisetina* (Wan et al., 2018, see '*Gnetophytes Genome Project*', which also contains information on data production, taxon sampling, assembly and annotation). We also used genomic data from nine published plant genomes and transcriptome data of Cycads and *Ginkgo* (Table 1; six gymnosperms - *G. montanum*, *E. equisetina*, *Picea abies*, *P. taeda*, *G. biloba*, and *Cycas elongata*; six angiosperms - *A. trichopoda*, *Vitis vinifera*, *Arabidopsis thaliana*, *Musa acuminata*, *Elaeis guineensis*, *and Oryza sativa*, and; one non-seed plant - *S. moellendorffii*). Protein genes of species with genomes were from database (Table 1). For species with transcriptomic data, we used Genewise 2.4.1 (Birney et al., 2004) to generate gene structures based on proteins to the assembled *G. montanum* sequence. Protein gene from all these species were used for phylogenetic analyses.

Table 1. Species and genome assemblies used in this study

Species	Sources of data
Selaginella moellendorffii	http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10
Picea abies	http://congenie.org
Pinus taeda	http://pinegenome.org/pinerefseq/
Ginkgo biloba	ftp://ftp.plantbiology.msu.edu/ pub/data/MPGR/Ginkgo biloba/ (EST data)
Cycas elongata	http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154384
Gnetum montanum	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA339497
Ephedra equisetina	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA339497
Amborella trichopoda	ftp://ftp.ensemblgenomes.org/pub/plants/release-25/plants/
Musa acuminata	ftp://ftp.ensemblgenomes.org/pub/release-25/plants/
Oryza sativa	ftp://ftp.ensemblgenomes.org/pub/release-25/plants/
Vitis vinifera	ftp://ftp.ensemblgenomes.org/pub/release-25/plants/
Arabidopsis thaliana	ftp://ftp.ensemblgenomes.org/pub/release-25/plants/
Elaeis guineensis	ftp://ftp.ensemblgenomes.org/pub/release-25/plants/

Similarities between these proteins were detected using an all-against-all BLASTp version 2.2.26 (Altschul et al., 1997) with an E-value of 1e^{-10}. Only alignments between gene pairs that had >0.5 coverage of each sequence were retained for analysis. The software OrthoMCL (Li et al., 2003), based on a Markov cluster algorithm was applied to find orthogroups (with an inflation value of 1.5, to extract orthologous and paralagous proteins). Then we retrieved a single copy sequence in per species, to generate a 1,334 gene dataset. Then the alignments of 331,467 AA sites were concatenated to a super alignment matrix. ProtTestwas used to select the best fit model (LG+Γ4 model) for amino acid replacement and RA×ML (v8.0.19) was used to reconstruct a maximum likelihood tree. Robustness of the maximum likelihood tree was assessed using the bootstrap method (100 pseudo-replicates). A ML tree supporting "gnetophytes around the conifers" were shown in Fig. 2A.



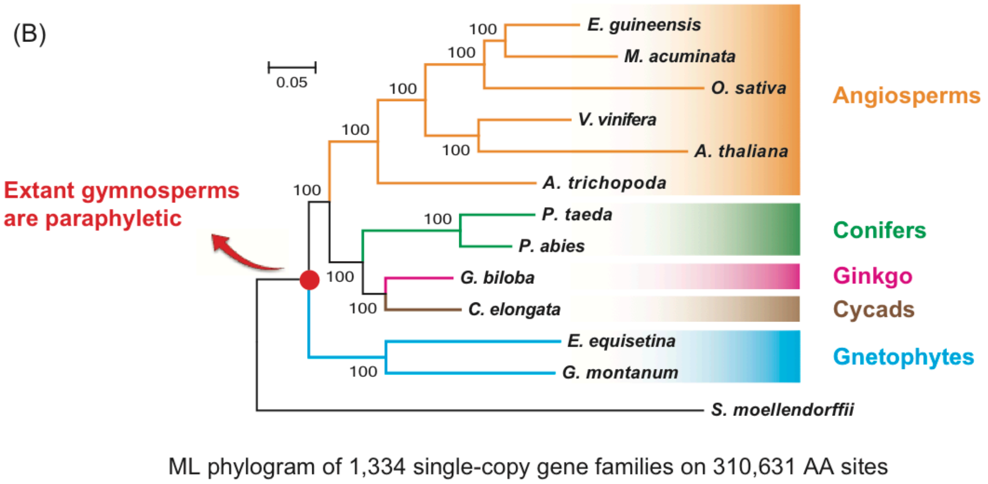


Figure 2. Phylogenetic inference based on nuclear genes with different sites selection. The alignments of 331,467 AA sites in Figure 2A (or 310,631 AA sites in Figure 2B) were concatenated to a super alignment matrix. ProtTest was used to select the best fit model (LG+Γ4 model) for amino acid replacement and RA×ML (v8.0.19) was used to reconstruct a maximum likelihood tree. Robustness of the maximum likelihood tree was assessed using the bootstrap method (100 pseudo-replicates).

Considering the following three factors: (a) The extinctions of the ancestor of modern angiosperms; (b) angiosperms have greater average rate of substitution than gymnosperms; (c) the slow evolutionary rate of the gymnosperms, the distinct evolution rates among extent angiosperms and gymnosperms as well as the lack of ancient angiosperm ancestor sequence, simultaneously leads to that gnetophytes will contain much more same amino-acids with the non-gnetophyte gymnosperms than extant angiosperms, which will outweigh the real evolutionary signals during phylogenetic inference. Hence, to reduce the effect of (a) and (b), we retained sites where at least one of the six gymnosperms has same amino-acid with at least one of the six angiosperms. Finally, 310,631 (94%) of the total 331,467 sites were retained and used to do the phylogenic tree construction following the above approach. Notably, a ML tree (with 100% bootstrap support) indicating "gnetophytes as sisters to other seed plants" based on these sites were generated and showed in Fig. 2B.

Though the 'gnetophytes as sister to other seed plants' hypothesis has been reported previously, either by using whole plastid sequence data, plastid proteins, or mitochondrial genes (data were not shown) analyzed with a range of methods including maximum parsimony and maximum likelihood (Frohlich & Parker, 2000; Schmidt & Schneider-Poetsch, 2002; Lee et al., 2011; Chen et al., 2016), this hypothesis tends to be ruled out based on nuclear loci (Wickett et al., 2014). Here, for the first time, we showed that this hypothesis could also be supported from phylogenetic inference from nuclear genes. Given this, a hypothesis of gnetophytes being sister to all other seed plants cannot be ruled out on the basis of phylogenetic trees inferred from nuclear loci. However, resolving phylogenetic relationships among extant seed plants has been shown as an extraordinarily difficult problem only based on gene sequences (Wickett et al., 2014). Hence, we think there is little hope that the mainstream approach of gene tree construction resulting from different treatments of the data and methods of analysis will solve this problem fundamentally. We should adopt other novel approach to provide better resolution of relationships among major seed plant clades.

Gene evolution patterns among representative seed plants

Because of the inconsistencies in signals using typical phylogenetic approaches, we developed an indirect approach to examine patterns of gene proliferation among major seed plant clades. In total, five different protein coding gene sets in *Amborella trichopoda* were compared against the relevant protein coding genes in *Sellaginella moellendorffii* (Fig. 3, supplementary table 1-5). Only genes of *A. trichopoda* having at least a score > 50 with *S. moellendorffii* were used in the boxplot. The determination and selection methods are as follows:

(a) We defined an “n1” set of single copy orthoMCL genes (Li et al., 2003), which are those groups shared by *G. montanum*, *G. biloba*, *P. taeda* and *A. trichopoda* (supplementary table S2).

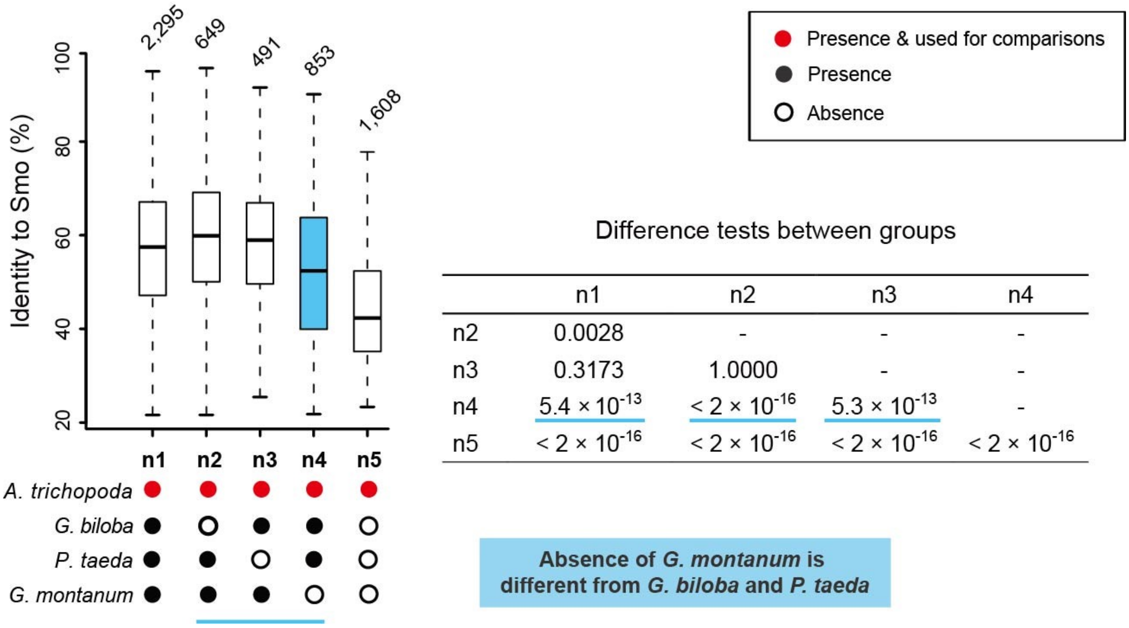
(b) The “n2” gene set comprised protein-coding genes considered to be missing in *G. biloba* but present in the other seed plants, *G. montanum*, *P. taeda* and *A. trichopoda.* We used BLASTp, with an E-value threshold of 1e^{-5}, to query protein sets from *G. montanum* and *P. taeda* against protein sets from *G. biloba* and *A. trichopoda*. Genes shared by *G. montanum* and *P. taeda*, that also had ≥ 10% higher scores (score=identity*coverage) when queried against *A. trichopoda* compared with *G. biloba*, were considered to be genes that are absent in *G. biloba*.

(c) The “n3” set were protein genes considered to be missing in *P. taeda* but present in the other seed plants, *G. montanum*, *G. biloba*, *A. trichopoda.* We aligned by BLASTp with an E-value of 1e^{-5} protein sets from *G. biloba* and *G. montanum* against protein sets from *P. taeda* and *A. trichopoda*. Genes shared by *G. biloba* and *G. montanum*, that also had ≥ 10% higher scores (score=identity * coverage) when aligned against *A. trichopoda* compared with *P. taeda*, were considered to be genes that are absent in *P. taeda*.

(d) The “n4” set consisted of protein genes considered to be missing in *G. montanum* but present in the other seed plants, *G. biloba*, *P. taeda*, *A. trichopoda.* We aligned by BLASTp with an E-value of 1e^{-5} protein sets from

G. biloba and *P. taeda* against protein sets from *G. montanum* and *A. trichopoda*. Genes shared by *G. biloba* and *P. taeda*, that also had ≥ 10% higher scores (score=identity*coverage) when aligned against *A. trichopoda* compared with *G. montanum*, were considered to be genes that are absent in *G. montanum*.

(e) The "n5" set were protein genes considered to occur only in *A. trichopoda* specific families.



Difference tests between groups

	n1	n2	n3	n4
n2	0.0028	-	-	-
n3	0.3173	1.0000	-	-
n4	5.4×10^{-13}	$< 2 \times 10^{-16}$	5.3×10^{-13}	-
n5	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$

Figure 3. Genic signatures for gnetophytes being distinct from other seed plants. Comparisons of *A. trichopoda* gene sets that carry orthologues present in different combinations of gymnosperms (BLASTp gene identities against *S. moellendorffii* - Smo). The "n1" group contains genes found in all seed plants investigated and were likely found in the seed plant ancestor, while "n5" represents genes present only in *A. trichopoda*, genes that are likely to have diverged substantially only in the angiosperm lineage). When *G. montanum* orthologous genes are absent ("n4"), the dataset has significantly lower overall BLASTp identities than in datasets where orthologues are absent in other gymnosperms (e.g. "n2" and "n3").

We mainly focused on genes that were present in *Amborella trichopoda* but were absent in one of the gymnosperms (i.e. absent in either *G. montanum*, or *Ginkgo biloba*, or *Pinus taeda*). The "n1" set, with a speculative orthologue in all the seed plant lineages, were considered to be genes that have been retained from the seed plant ancestor (Supplementary Table 1). In contrast the "n2", "n3" and "n4" sets have an orthologue missing in *G. biloba*, *P. taeda* or *G. montanum* (Supplementary Table 2-4). The "n5" set of genes (Supplementary Table 5) is considered to be specific to *A. trichopoda* or gained after angiosperms separated from gymnosperms. Our analyses of the sequence identities of the "n1"-"n5" gene sets showed that the sequence identity distribution of the "n5" set is lower than the "n1" set ($p < 2 \times 10^{-16}$, Fig. 3). Remarkably, the "n4" set, representing genes thought to be absent in *G. montanum*, also has a lower sequence identity distribution than that of the "n2" set (orthologous genes considered to be absent in *G. biloba*, $p < 2 \times 10^{-16}$, Fig. 3) and the "n3" set (orthologous genes considered to be absent in *P. taeda*, $p < 5.3 \times 10^{-13}$, Fig. 3). When *G. montanum* orthologous genes are absent ("n4"), the dataset has significantly lower overall BLASTp identities than in datasets where orthologues are absent in other gymnosperms (e.g. "n2" and "n3"). What becomes apparent is that orthologous genes in *G. montanum* (i.e. n1, n2, and n3 sets; Fig. 3) play a notable role in comparison among all gene sets identity distributions (Fig. 3).

How to interpret this gene evolution patterns? It's important to note that these analyses are not affected by the different substitution rate among *G. montanum*, *G. biloba*, and *P. taeda* since all the genes of "n1"-"n5" sets are from *A. trichopoda*. Firstly, "n5" set is deservedly considered later obtained by *A. trichopoda* after its split with other gymnosperms and the genes from "n5" set are more young than other genes from other sets (Fig. 4). Hence, the "n5" set genes has less sequence similarities (or identities) with the ancient *S. moellendorffii* genes. It is expected that because comparison between orthologs of ancient, conserved ("old") genes would be more similar to each other, and have higher sequence identities than orthologs of less conserved, or more rapidly diverging ("yound") genes. Secondly, "n2" genes and "n3" genes show no obvious difference, which indicated that “n2” genes and “n3” genes were obtained before the split of *G. montanum*, *G. biloba*, *P. taeda* and *A. trichopoda* (in other words, these genes were obtained in the common ancestor of *G. montanum*, *G. biloba*, *P. taeda* and *A. trichopoda*). The absence of these genes in *G. biloba* or *P. taeda* is due to random lost during their independent evolution (Fig. 4). And this speculation is confirmed by that there are also no obvious difference between “n1” genes and “n2”/“n3” genes since “n1” represents genes shared by *G. montanum*, *G. biloba*, *P. taeda* and *A. trichopoda* (in other words, these genes were also obtained in the common ancestor of *G. montanum*, *G. biloba*, *P. taeda* and *A. trichopoda*). In other words, “n2”/“n3” genes were as “old” as “n1” genes. Notably, the ‘gene absence’ patterns could not be explained by random lost in *G. montanum* but should be interpreted by the early divergence of gnetophytes from rest of seed plants because “n4” has much lower identities than “n1”, “n2”, and “n3”. In other words, part (red G2 in Fig. 4) of “n4” genes were actually obtained only by the common ancestor of *G. biloba*, *P. taeda* and *A. trichopoda* with early split of *G. montanum* and these “n4” genes are “younger” than “n1”, “n2”, and “n4” but “older” than “n5” genes. The observed gene evolution pattern is only consistent with the hypothesis that gnetophytes are sister to all other extant seed plants (Fig. 4). Notably, that there is little overlap (Wayne comparison in Fig. 4) between these three sets (‘n2’, ‘n3’ and ‘n4’) also confirms than our approach can detect specific gene absence in each species, which are not affected by the gene loss in other nodes (such as L4 or L2 in Fig. 4).

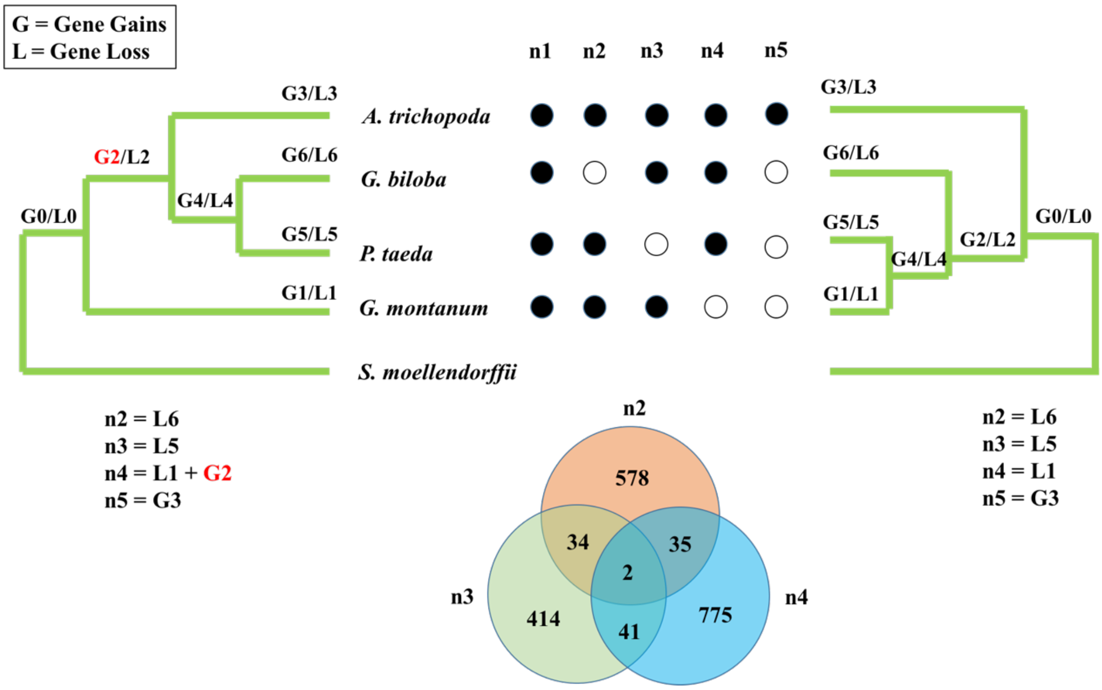


Figure 4. Interpretation of distinct Genic signatures for gnetophytes based on gene gains/loss mechanism with 'Seed plants sister'(right) hypothesis and 'closed to conifers' hypothesis. More gene loss detected in *G. montanum* and 'n2' genes harboring less identity is due the G2 part of genes (genes gained by the MRCA of other seed plants). The right topology tree can not explain the distinct patterns found in *G. montanum.*

For additional, we examined in detail some exemplar multigene families of the "n4" set using phylogenetic approaches and observed two families that are consistent with the 'Seed plants sister' hypothesis. For example, from the germin-like protein family the orthologue GLP7 is missing from *G. montanum* and is broadly shared by non-gnetophytes gymnosperms and angiosperms (see Supplementary Materials online and Supplementary Fig. 1). In the same way, sub-clades of the Phenylalanine Ammonia Lyase gene family are found to be only shared by non-gnetophytes gymnosperms and angiosperms (Supplementary Fig. 2, Supplementary Table 4; Bagal et al., 2012).

Discussion

We argue here that the hypotheses 'gnetophytes are sister to all other seed plants' can not be ruled out based only on the phylogenetic trees inferred from nuclear loci. Assuming so, there remains the potential for profound shifts in our understanding of the origin and divergence of many genetic, genomic, biochemical, metabolic and morphological characters in seed plant evolution. For example, the two paralogues GgWOXX and GgWOXY should be consider to have been lost in the their MRCA of other seed plants after the split with gnetophytes (Wan et al., 2018). And if gymnosperms may not be monophyletic, then many characters used to be considered to be derived in gnetophytes (Wan et al., 2018) may in fact be ancestral characters (Fig. 5, e.g. intron structures, lack of WGD, pfam domains).

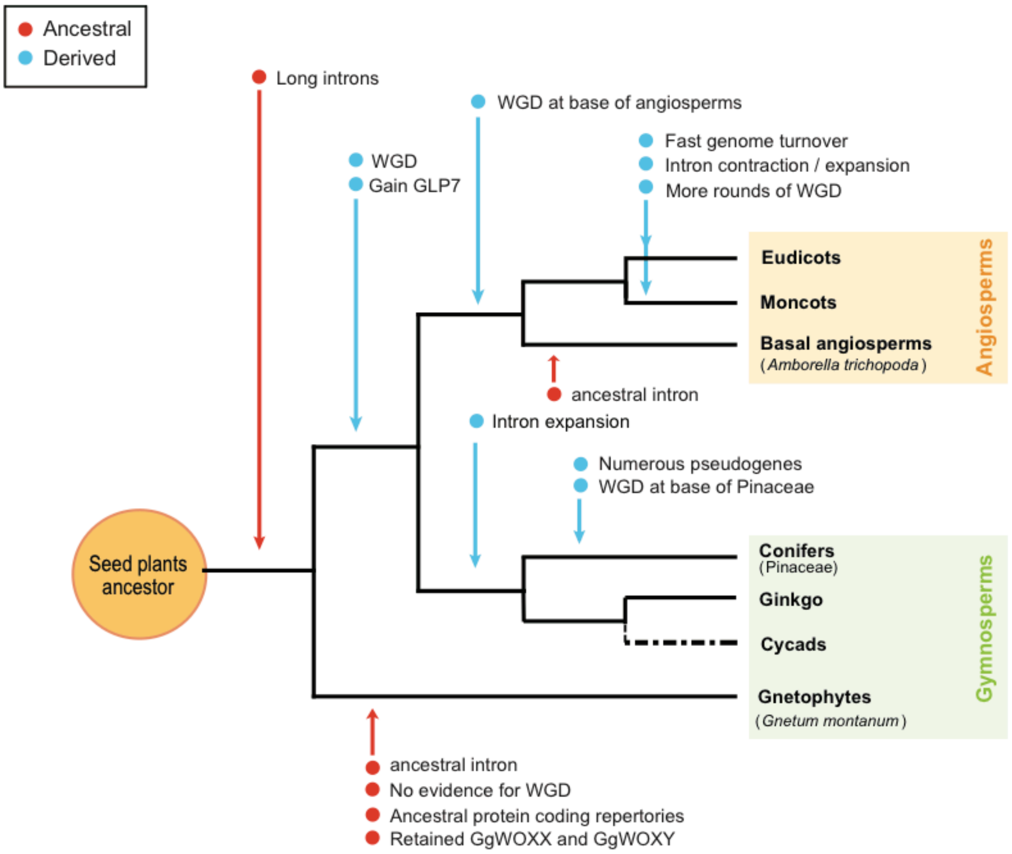


Figure 5. Reinterpretation of genome evolution patterns across seed plants based on the hypotheses gnetophytes are sister to all other seed plants.

ACKNOWLEDGEMENTS

Thanks to Prof. Antonis Rokas, who provide guidance and formula on gene-wise log-likelihood scores (GLS), and also to Dr. Yanting Hu for her kindly help on Cycads transcriptome data transfer. We also thank Prof. Andrew R. Leitch, Ilia J. Leitch, Yves Van de Peer, Laura Kelly, Honglei Li, Yang Liu, Min Yang, Qingfeng Wang, Tao Wan, Jinbo Zhang, and Ji Li for the useful comments on the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict interest.

AUTHOR CONTRIBUTIONS

Yinzhi Zhang conceived the study and led the manuscript preparation. Yinzhi Zhang worked on the data matrix construction; Zhiming Liu mostly contributed to the phylogenetic analyses of PAL, GLP families. Yinzhi Zhang mostly contributed to the analyses of gene family divergent pattern comparison.

ADDITIONAL INFORMATON

Supplementary information including materials, methods, figures and tables are available at *BioRxiv* Online.

REFERENCES

Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, *21*(9), 2104–2105. https://doi.org/10.1093/bioinformatics/bti263

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.https://doi.org/10.1093/nar/25.17.3389

Bagal, U. R., Leebens-Mack, J. H., Lorenz, W. W., & Dean, J. F. (2012). The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage. *BMC Genomics 13*, 1–9. https://doi.org/10.1186/1471-2164-13-S3-S1

Bar-Hen, A., Mariadassou, M., Poursat, M.-A., & Vandenkoornhuyse, P. (2008). Influence function for robust phylogenetic reconstructions. *Molecular Biology and Evolution*, *25*(5), 869–873. https://doi.org/10.1093/molbev/msn030

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. Genome research, 14(5), 988-995. http://www.genome.org/cgi/doi/10.1101/gr.1865504.

Brown, J. M., & Thomson, R. C. (2016). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, *66*(4), 517–530. https://doi.org/10.1093/sysbio/syw101

Burleigh, J. G., & Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American Journal of Botany*, *91*(10), 1599–1613. https://doi.org/10.3732/ajb.91.10.1599

Carlquist, S. (1996). Wood, bark and stem anatomy of New World species of *Gnetum*. *Botanical Journal of the Linnean Society*, *120*(1), 1–19. https://doi.org/10.1111/j.1095-8339.1996.tb00476.x

Castoe, T. A., de Koning, A. J., Kim, H. M., Gu, W., Noonan, B. P., Naylor, G., ... Pollock, D. D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, *106*(22), 8986–8991. https://doi.org/10.1073/pnas.0900233106

Chen, Z.-D., Yang, T., Lin, L., Lu, L.-M., Li, H.-L., … Sun, M. (2016). Tree of life for the genera of Chinese vascular plants. *Journal of Systematics and Evolution*, *54*(4), 277–306. https://doi.org/10.1111/jse.12219

Crane, P. R. (1985). Phylogenetic analysis of seed plants and the origin of angiosperms. *Annals of the Missouri Botanical Garden. 72*(4), 716–793. https://doi.org/10.2307/2399221

Donoghue, M. J., & Scheiner, S. M. (1992). The evolution of the endosperm: a phylogenetic account. In: Wyatt, R. (Eds.), Ecology and evolution of plant reproduction (pp. 356–389). New York: Chapman & Hall.

Doyle, J. A. (2012). Molecular and fossil evidence on the origin of angiosperms. *Annual Review of Earth and Planetary Sciences*, *40*(1), 301–326. https://doi.org/10.1146/annurev-earth-042711-105313

Doyle, J. A., & Donoghue, M. J. (1986). Seed plant phylogeny and the origin of angiosperms: An experimental cladistic approach. *The Botanical Review*, *52*(4), 321–431. https://doi.org/10.1007/BF02861082

Doyle, J. A., Donoghue, M. J., & Zimmer, E. A. (1994). Integration of morphological and ribosomal RNA data on the origin of angiosperms. *Annals of the Missouri Botanical Garden*, *81*(3), 419–450. https://doi.org/10.2307/2399899

Drew, B. T., Ruhfel, B. R., Smith, S. A., Moore, M. J., Briggs, B. G., Gitzendanner, M. A., … Soltis, D. E. (2014). Another look at the root of the Angiosperms reveals a familiar tale. *Systematic Biology*, *63*(3), 368–382. https://doi.org/10.1093/sysbio/syt108

Drouin, G., Daoud, H., & Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular Phylogenetics and Evolution*, *49*(3), 827–831.

https://doi.org/10.1016/j.ympev.2008.09.009

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Finster, S., Legen, J., Qu, Y., & Schmitz-Linneweber, C. (2012). Land plant RNA editing or: don't be fooled by plant organellar DNA sequences. In: Bock R., Knoop V. (Eds.), *Genomics of Chloroplasts and Mitochondria* (pp. 293–321). Dordrecht: Springer.

Friedman, W. E. (1990). Double fertilization in *Ephedra*, a nonflowering seed plant: its bearing on the origin of angiosperms. *Science*, *247*(4945), 951–954. https://doi.org/10.1126/science.247.4945.951

Frohlich, M. W., & Parker, D. S. (2000). The mostly male theory of flower evolutionary origins: From genes to fossils. *Systematic Botany*, *25*(2), 155–170. https://doi.org/10.2307/2666635

Gatesy, J., Meredith, R. W., Janecka, J. E., Simmons, M. P., Murphy, W. J., & Springer, M. S. (2016). Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. *Cladistics*, *33*(3), 295–332. https://doi.org/10.1111/cla.12170

Guo, W., Zhu, A., Fan, W., & Mower, J. P. (2017). Complete mitochondrial genomes from the ferns *Ophioglossum californicum* and *Psilotum nudum* are highly repetitive with the largest organellar introns. *New Phytologist*, *213*(1), 391–403. https://doi.org/10.1111/nph.14135

Hall, T. A. (1999). BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Ickert-Bond, S. M., & Renner, S. S. (2016). The Gnetales: Recent insights on their morphology, reproductive biology, chromosome numbers, biogeography, and divergence times. *Journal of Systematics and Evolution*, *54*(1), 1–16. https://doi.org/10.1111/jse.12190

Katoh, K., Kuma, K. I., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, *33*(2), 511–518. https://doi.org/10.1093/nar/gki198

Kimball, R. T., Wang, N., Heimer-McGinn, V., Ferguson, C., & Braun, E. L. (2013). Identifying localized biases in large datasets: A case study using the avian tree of life. *Molecular Phylogenetics and Evolution*, *69*(3), 1021–1032. https://doi.org/10.1016/j.ympev.2013.05.029

Knoop, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Current Genetics*, *46*(3), 123–139. https://doi.org/10.1007/s00294-004-0522-8

Lee, E. K., Cibrian-Jaramillo, A., Kolokotronis, S.-O., Katari, M. S., Stamatakis, A., Ott, M., … DeSalle, R. (2011). A functional phylogenomic view of the seed plants. *PLoS Genetics*, *7*(12), e1002411. https://doi.org/10.1371/journal.pgen.1002411

Li, L. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes.

Genome Research, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Li, Z., De La Torre, A. R., Sterck, L., Cánovas, F. M., Avila, C., Merino, I., … Van de Peer, Y. (2017). Single-copy genes as molecular markers for phylogenomic studies in seed plants. *Genome Biology and Evolution*, *9*(5), 1130–1147. https://doi.org/10.1093/gbe/evx070

Liu, Y., Cox, C. J., Wang, W., & Goffinet, B. (2014). Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Systematic Biology*, *63*(6), 862–878. https://doi.org/10.1093/sysbio/syu049

Mathews, S. (2009). Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. *American Journal of Botany*, *96*(1), 228–236. https://doi.org/10.3732/ajb.0800178

McCoy, S. R., Kuehl, J. V., Boore, J. L., & Raubeson, L. A. (2008). The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. *BMC Evolutionary Biology*, *8*(1), 130. https://doi.org/10.1186/1471-2148-8-130

Nickrent, D. L., Parkinson, C. L., Palmer, J. D., & Duff, R. J. (2000). Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Molecular Biology and Evolution*, *17*(12), 1885–1895. https://doi.org/10.1093/oxfordjournals.molbev.a026290

Qiu, Y.-L., Li, L., Wang, B., Xue, J.-Y., Hendry, T. A., Li, R.-Q., … Chen, Z.-D. (2010). Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *Journal of Systematics and Evolution*, *48*(6), 391–425. https://doi.org/10.1111/j.1759-6831.2010.00097.x

Rai, H. S., Reeves, P. A., Peakall, R., Olmstead, R. G., & Graham, S. W. (2008). Inference of higher-order conifer relationships from a multi-locus plastid data set. *Botany*, *86*(7), 658–669. https://doi.org/0.1139/B08-062

Ran, J.-H., Gao, H., & Wang, X.-Q. (2010). Fast evolution of the retroprocessed mitochondrial rps3 gene in Conifer II and further evidence for the phylogeny of gymnosperms. *Molecular Phylogenetics and Evolution*, *54*(1), 136–149. https://doi.org/10.1016/j.ympev.2009.09.011

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., & Burleigh, J. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*, *14*(1), 23. https://doi.org/10.1186/1471-2148-14-23

Rydin, C., Källersjö, M., & Friis, E. M. (2002). Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: Conflicting data, rooting problems, and the monophyly of conifers. *International Journal of Plant Sciences*, *163*(2), 197–214. https://doi.org/10.1086/338321

Sanderson, M. J., Wojciechowski, M. F., Hu, J.-M., Khan, T. S., & Brady, S. G. (2000). Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Molecular Biology and Evolution*, *17*(5), 782–797. https://doi.org/10.1093/oxfordjournals.molbev.a026357

Schliep, K. P. (2010). Phangorn: Phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593. https://doi.org/10.1093/bioinformatics/btq706

Schmidt, M., & Schneider-Poetsch, H. A. W. (2002). The evolution of gymnosperms redrawn by phytochrome genes: The Gnetatae appear at the base of the gymnosperms. *Journal of Molecular Evolution*, *54*(6), 715–724. https://doi.org/10.1007/s00239-001-0042-9

Shavit Grievink, L., Penny, D., & Holland, B. R. (2013). Missing data and influential sites: Choice of sites for phylogenetic analysis can be as important as taxon sampling and model choice. *Genome Biology and Evolution*, *5*(4), 681–687. https://doi.org/10.1093/gbe/evt032

Shen, X.-X., Hittinger, C. T., & Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, *1*(5), 0126. https://doi.org/10.1038/s41559-017-0126

Simmons, M. P. (2017). Relative benefits of amino-acid, codon, degeneracy, DNA, and purine-pyrimidine character coding for phylogenetic analyses of exons. *Journal of Systematics and Evolution*, *55*(2), 85–109. https://doi.org/10.1111/jse.12233

Stamatakis, A., Ludwig, T., & Meier, H. (2004). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, *21*(4), 456–463. https://doi.org/10.1093/bioinformatics/bti191

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577. https://doi.org/10.1080/10635150701472164

Wan, T., Liu, Z.-M., Li, L.-F., Leitch, A. R., Leitch, I. J., Lohaus, R., … Wang, X.-M. (2018). A genome for gnetophytes and early evolution of seed plants. *Nature Plants*, *4*(2), 82–89. https://doi.org/10.1038/s41477-017-0097-2

Wang, X. Q., & Ran, J. H. (2014). Evolution and biogeography of gymnosperms. *Molecular Phylogenetics and Evolution*, *75*, 24–40. https://doi.org/10.1016/j.ympev.2014.02.005

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., … Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, *111*(45), E4859–E4868. https://doi.org/10.1073/pnas.1323926111

Zhong, B., Yonezawa, T., Zhong, Y., & Hasegawa, M. (2010). The position of Gnetales among seed plants: Overcoming pitfalls of chloroplast phylogenomics. *Molecular Biology and Evolution*, *27*(12), 2855–2863. https://doi.org/10.1093/molbev/msq170