Single-copy genes as molecular markers for phylogenomic studies in seed plants

- Supplementary Materials -

Zhen Li[1,2,3], Amanda R. De La Torre[4,5], Lieven Sterck[1,2,3], Francisco M. Canovas[6], Concepcion Avila[6], Irene Merino Sierra[7], José Antonio Cabezas[8], María Teresa Cervera[8], Pär K. Ingvarsson[4,9], Yves Van de Peer[1,2,3,10,*]

[1] Department of Plant Biotechnology and Bioinformatics, Ghent university, Technologiepark 927, 9052 Ghent, Belgium

[2] Center for Plant Systems Biology, VIB, Technologiepark 927, 9052 Ghent, Belgium

[3] Bioinformatics Institute Ghent, Technologiepark 927, 9052 Ghent, Belgium

[4] Department of Ecology and Environmental Science, Umeå University, Linneaus väg 6, SE-90187 Umeå, Sweden

[5] Department of Plant Sciences, University of California Davis, Davis CA 95616, USA.

[6] Departamento de Biologia Molecular y Bioquimica, Facultad de Ciencias, Instituto Andaluz de Biotecnologia, Universidad de Malaga, Campus Universitario de Teatinos s/n, 29071 Malaga, Spain

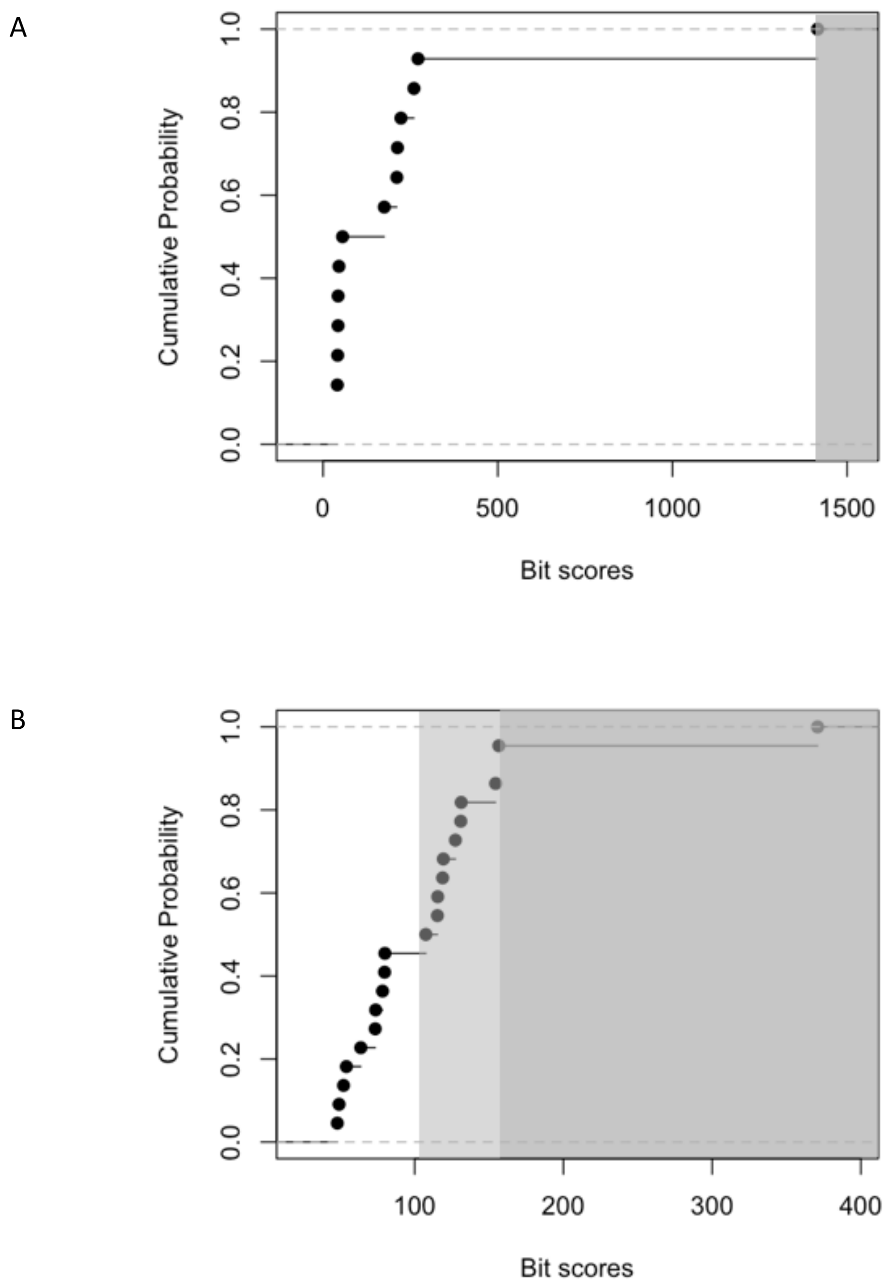
[7] Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences and the Linnean Center for Plant Biology, PO-Box 7080, SE-75007 Uppsala, Sweden

[8]INIA, Center Forest Research (CIFOR), E-28040 Madrid, Spain

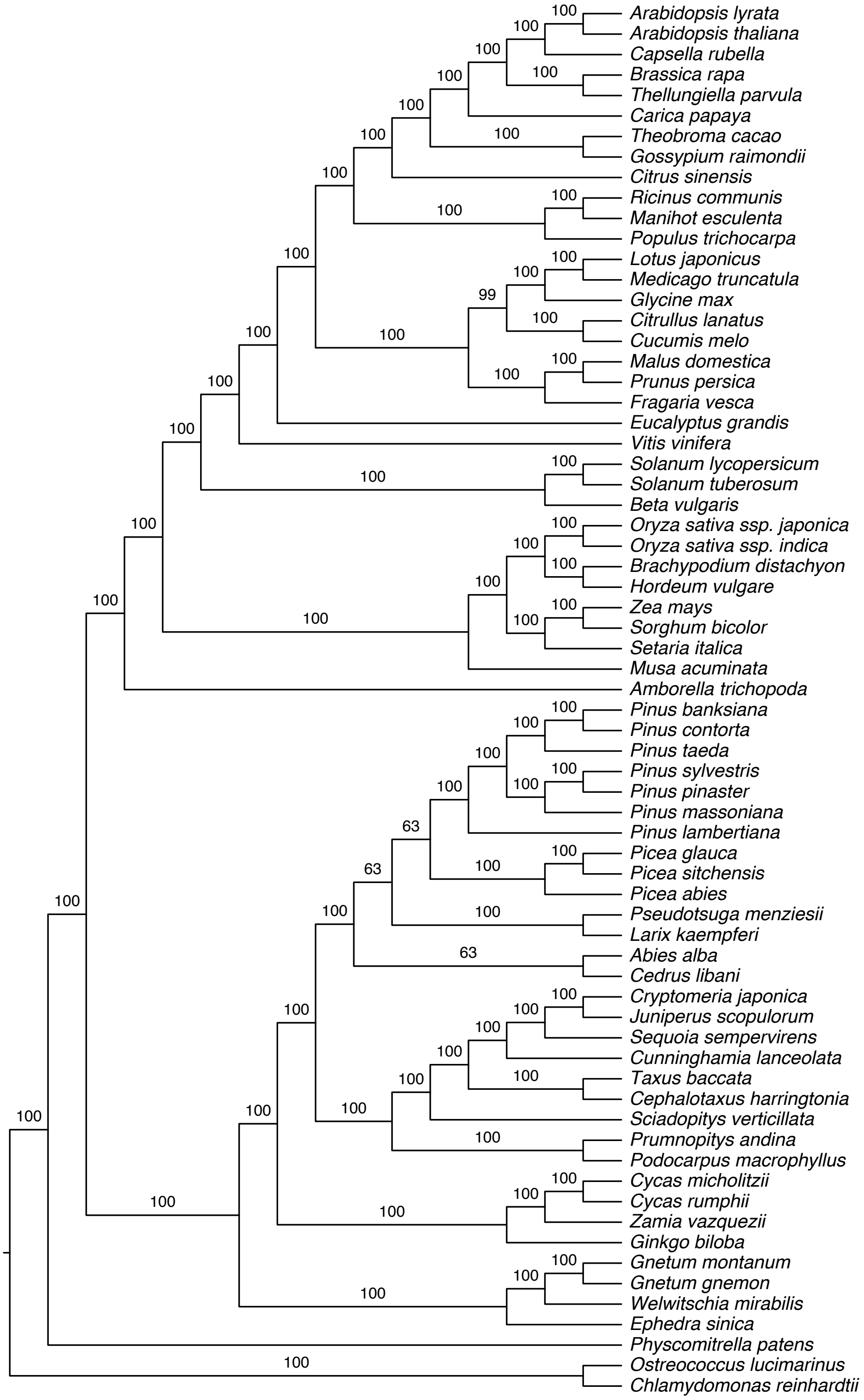
[9] Umeå Plant Centre, SE-90187, Umeå, Sweden

[10] Genomics Research Institute, University of Pretoria, Hatfield Campus, Pretoria 0028, South Africa.

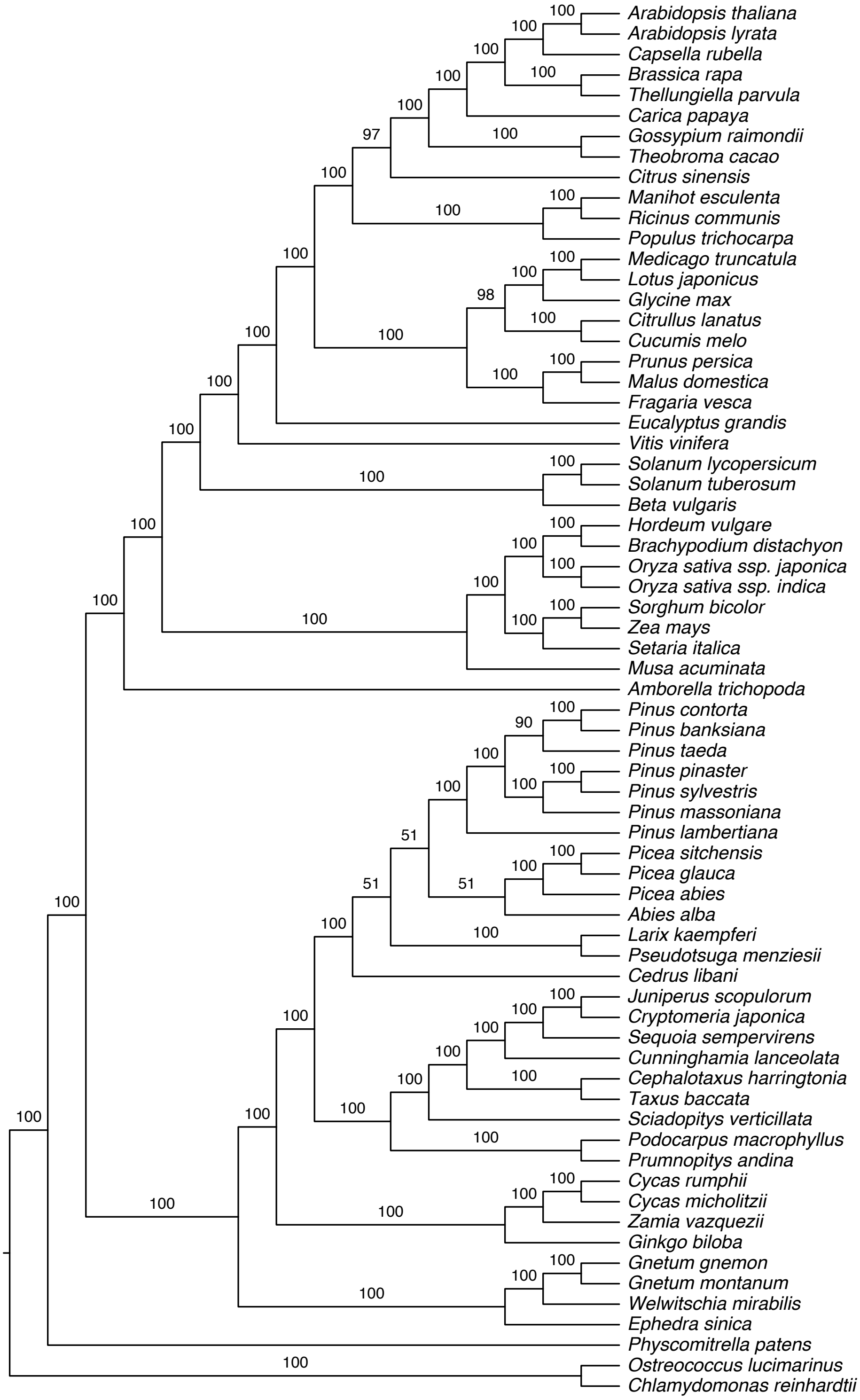
[*]Author for correspondence: Yves Van de Peer (yves.vandepeer@psb.vib-ugent.be)



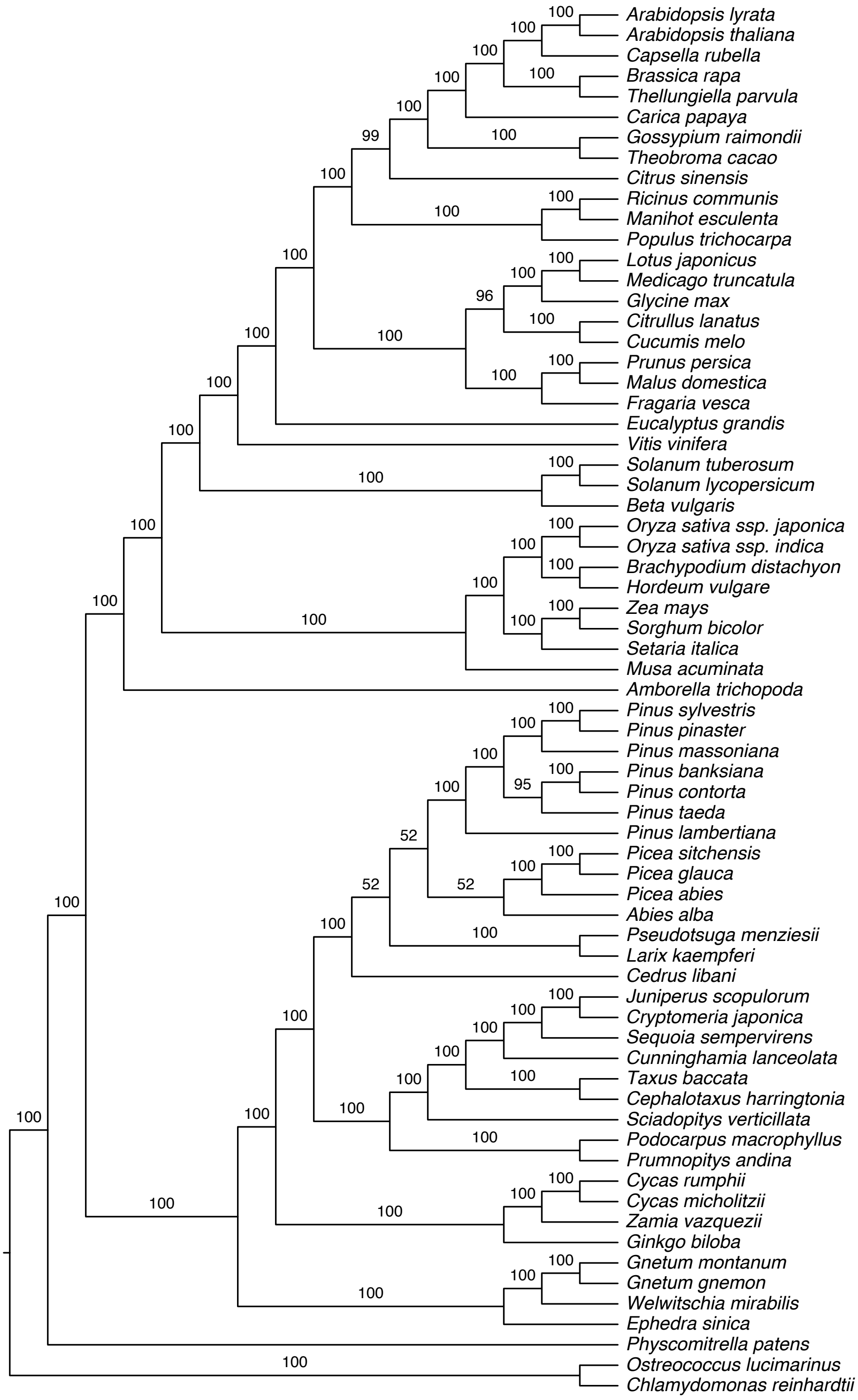
Supplementary Figure 1. Example showing the process to assign genes to a gene family accounting for 95% of the cumulative probability of bit-scores. The dark grey blocks in both A and B show regions with over 95% cumulative probability. Dots (i.e. hits) falling in the dark grey region are allocated to the gene family. In B, the light grey block denotes a region where hits have similar E-values (ΔE-value < 10×10^{20}) to the hit with the smallest Bit-score at the 95% border.



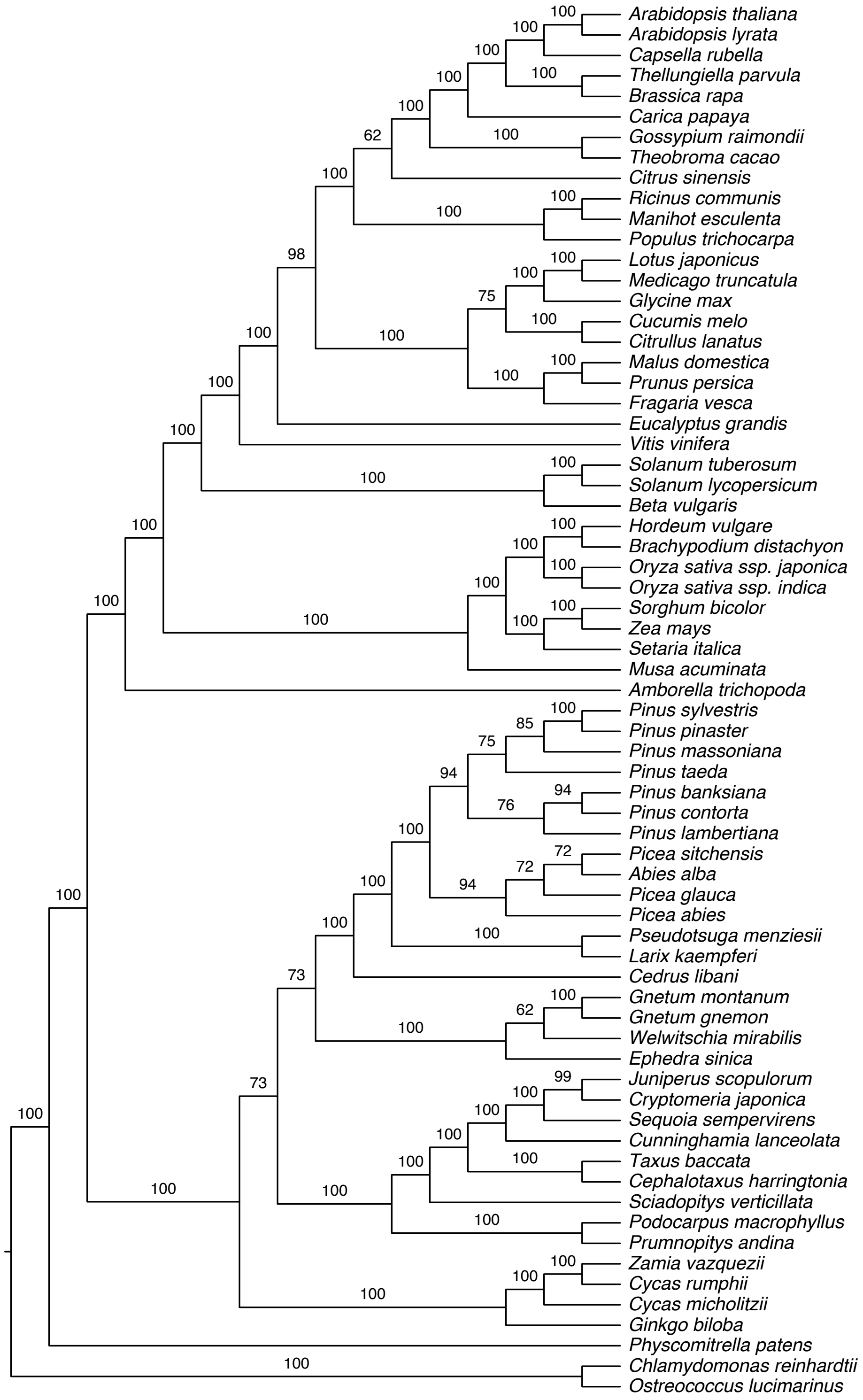
Supplementary Figure 2. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions with one partition. Numbers on branches represent bootstrap values.



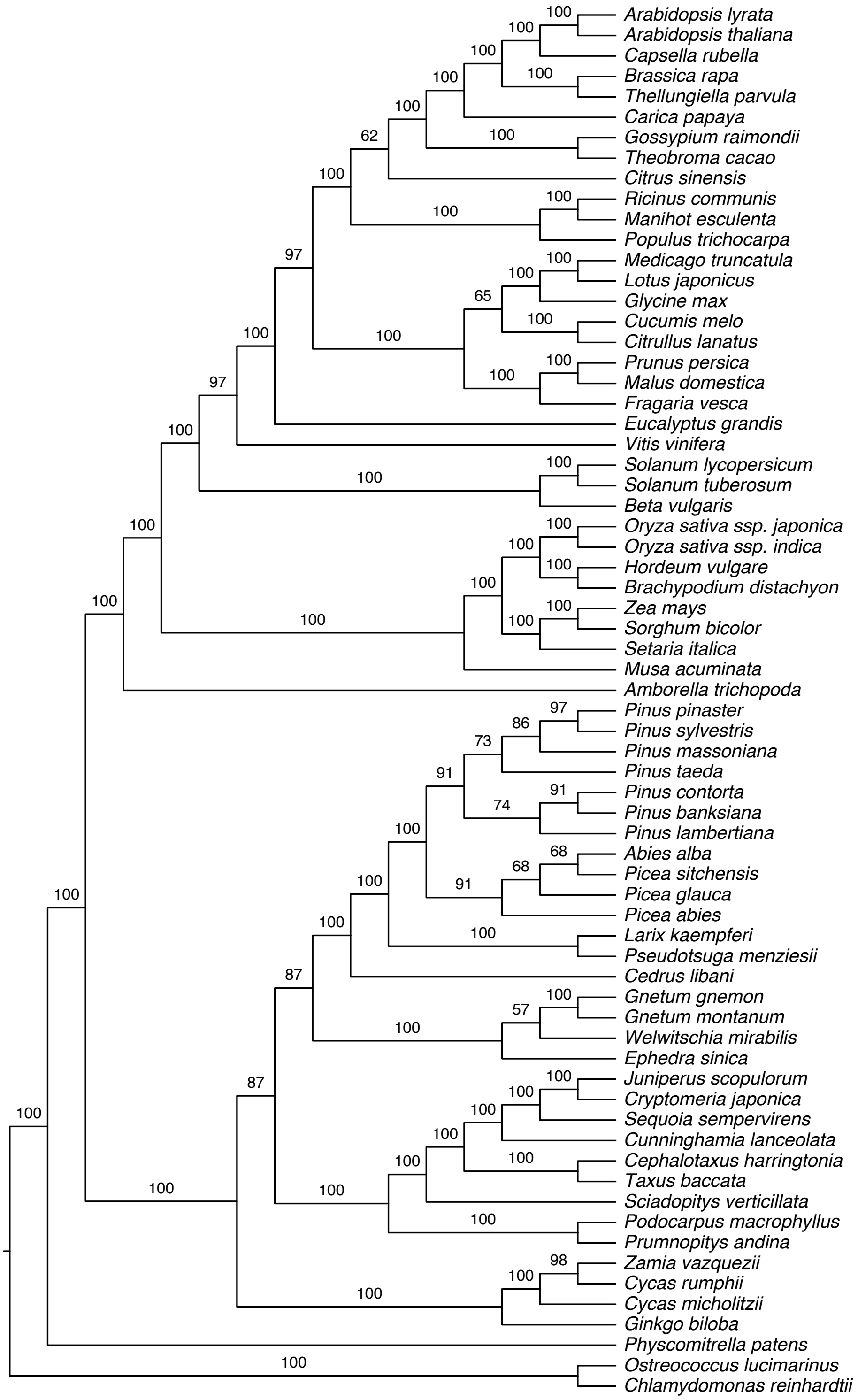
Supplementary Figure 3. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions, but with 1st and 2nd codon partitions as one partition and 3rd codon partition as another. Numbers on branches represent bootstrap values.



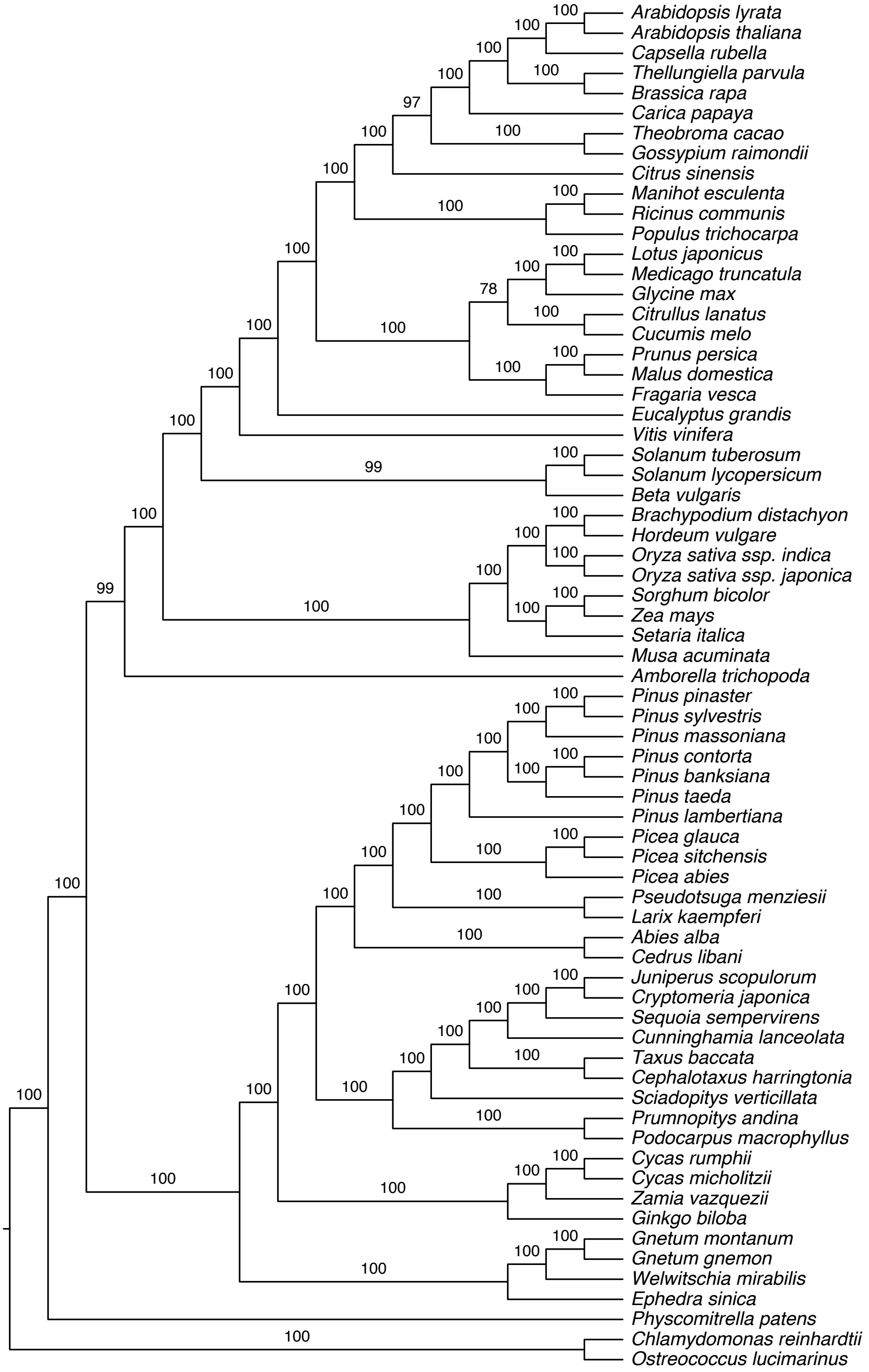
Supplementary Figure 4. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions but with three partitions, i.e. one for each codon position. Numbers on branches represent bootstrap values.



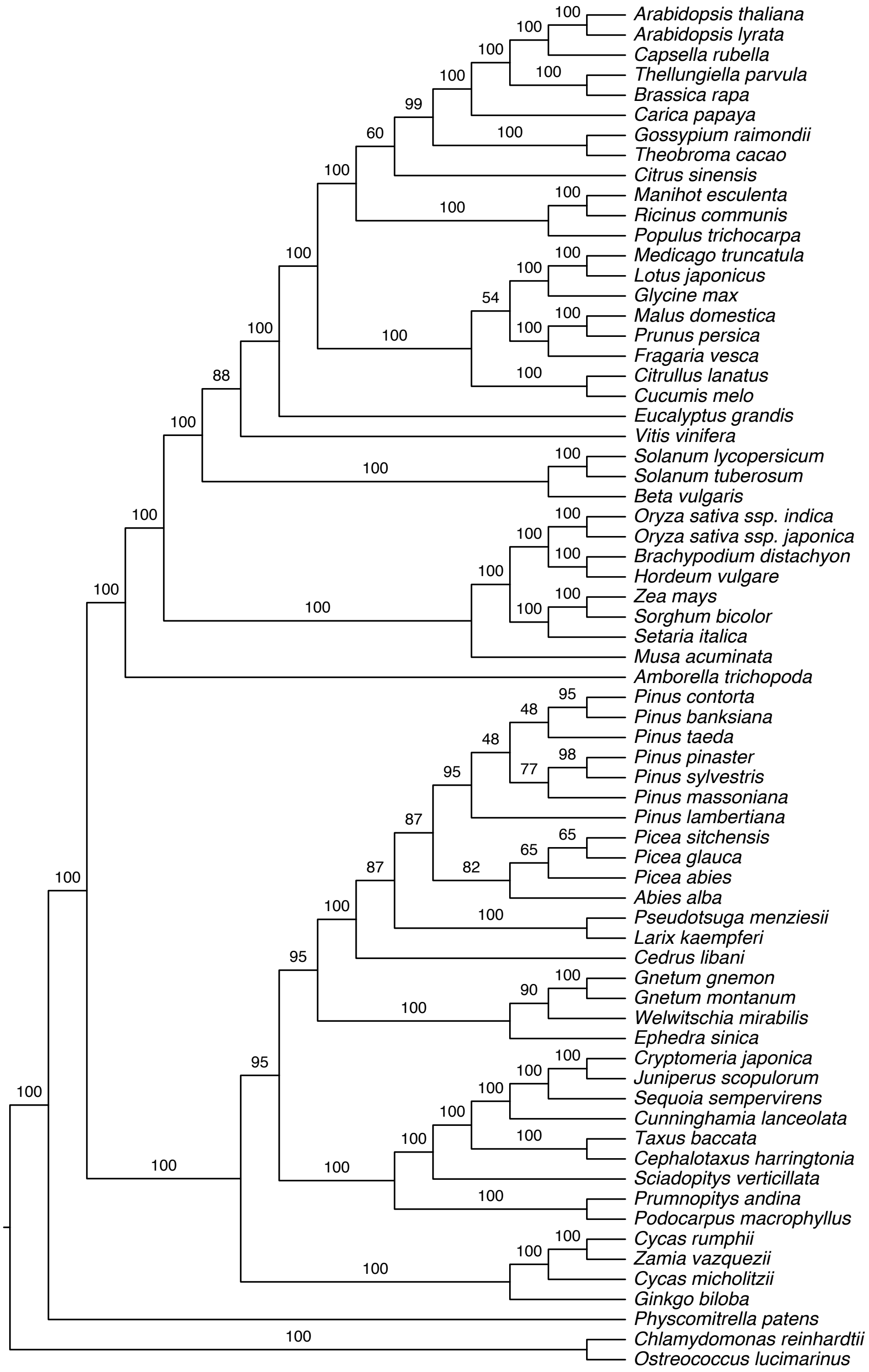
Supplementary Figure 5. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants with 3rd codon positions removed and with position 1 and 2 as one partition. Numbers on branches represent bootstrap values.



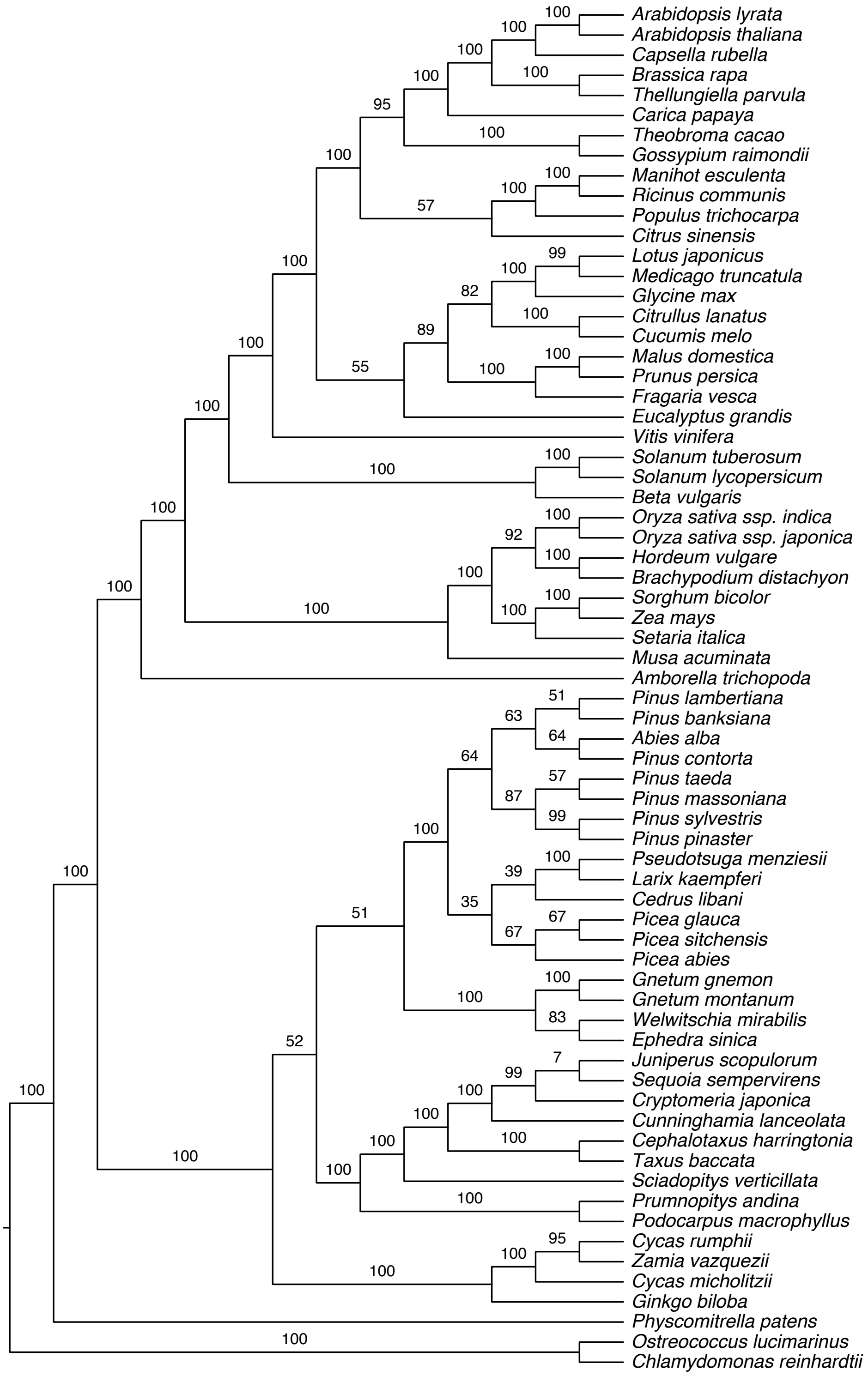
Supplementary Figure 6. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants with 3rd codon positions removed and with position 1 and 2 used as separate partitions. Numbers on branches represent bootstrap values.



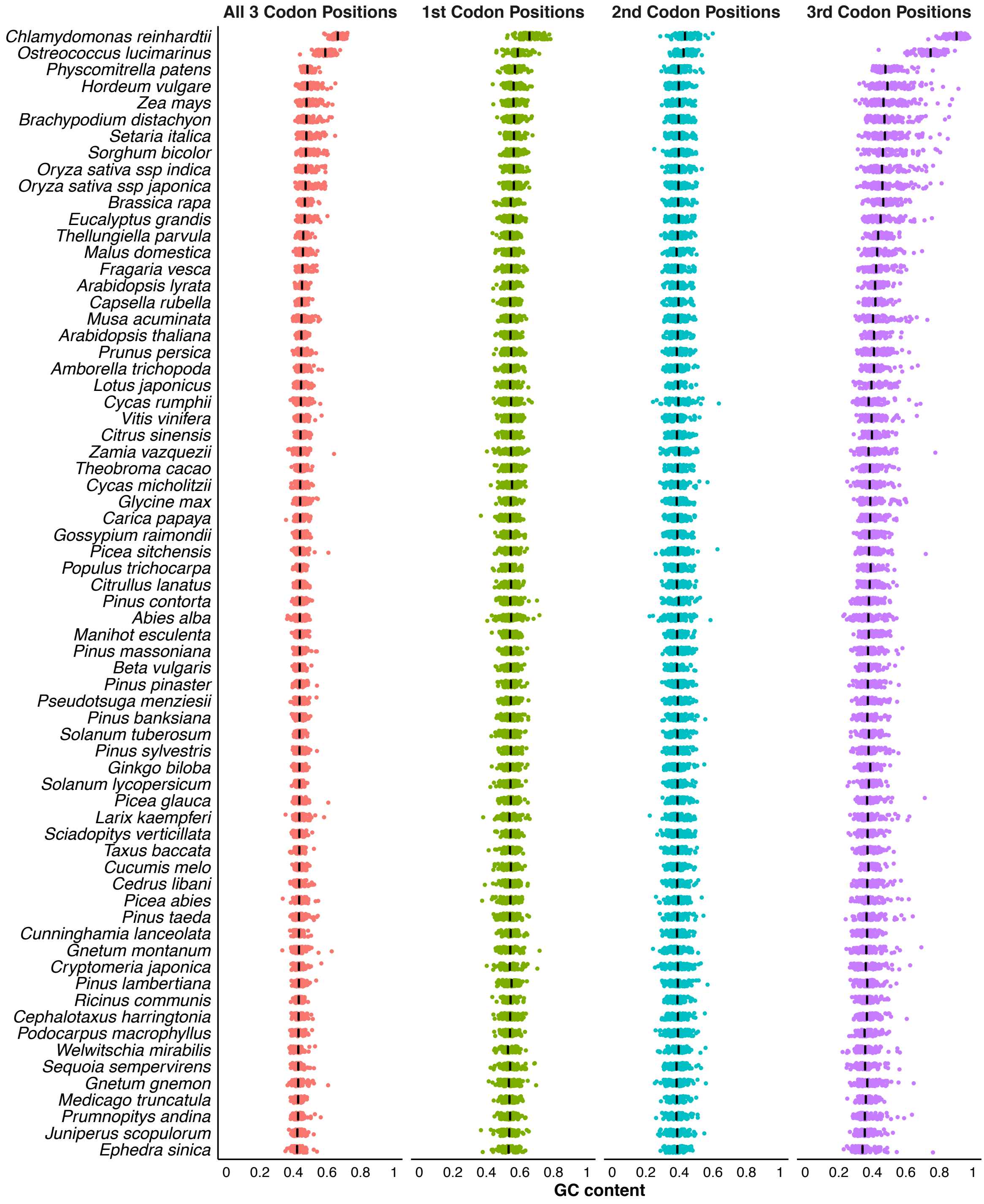
Supplementary Figure 7. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants only with 3rd codon positions. Numbers on branches represent bootstrap values.



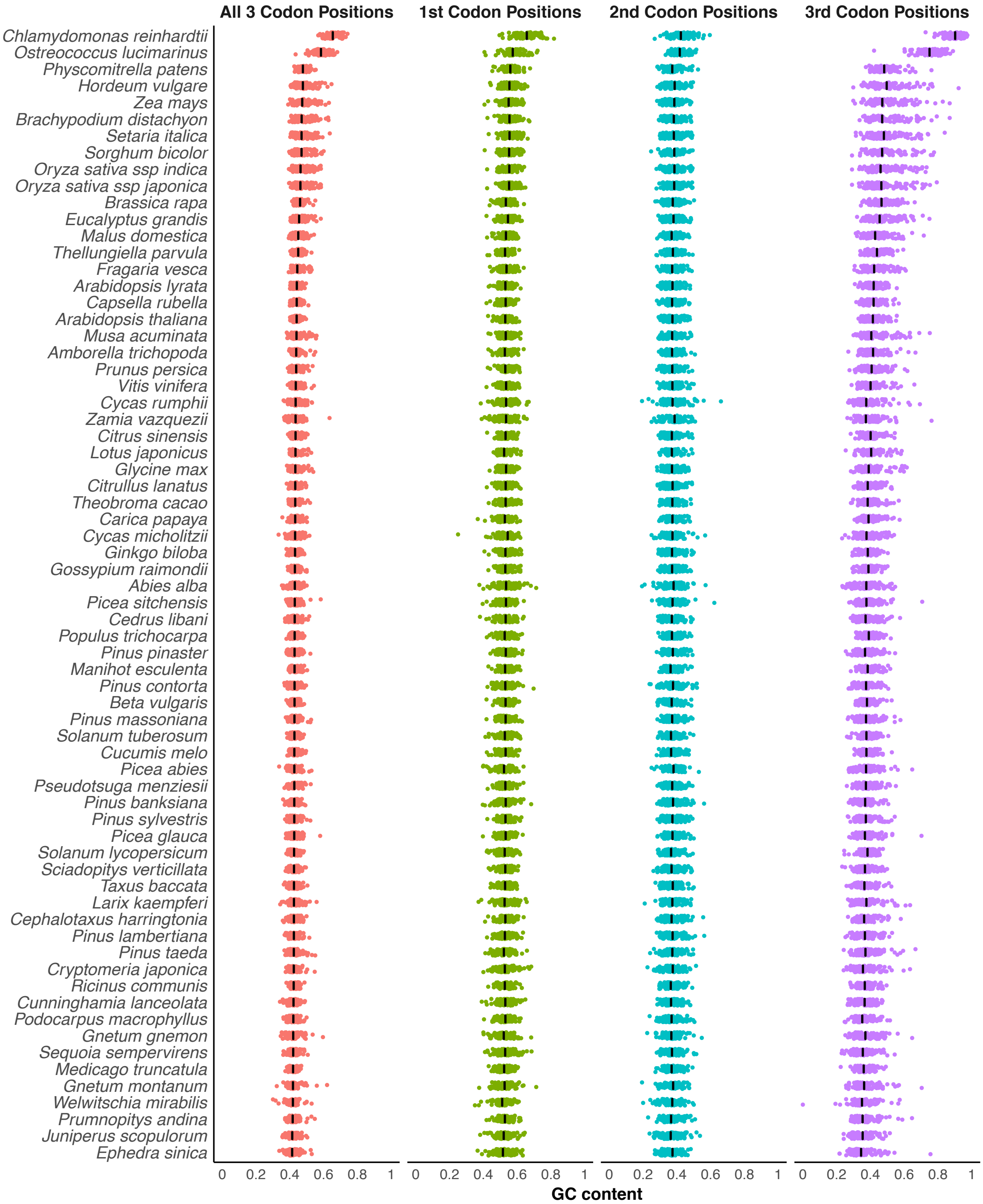
Supplementary Figure 8. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants only with 1st codon positions. Numbers on branches represent bootstrap values.



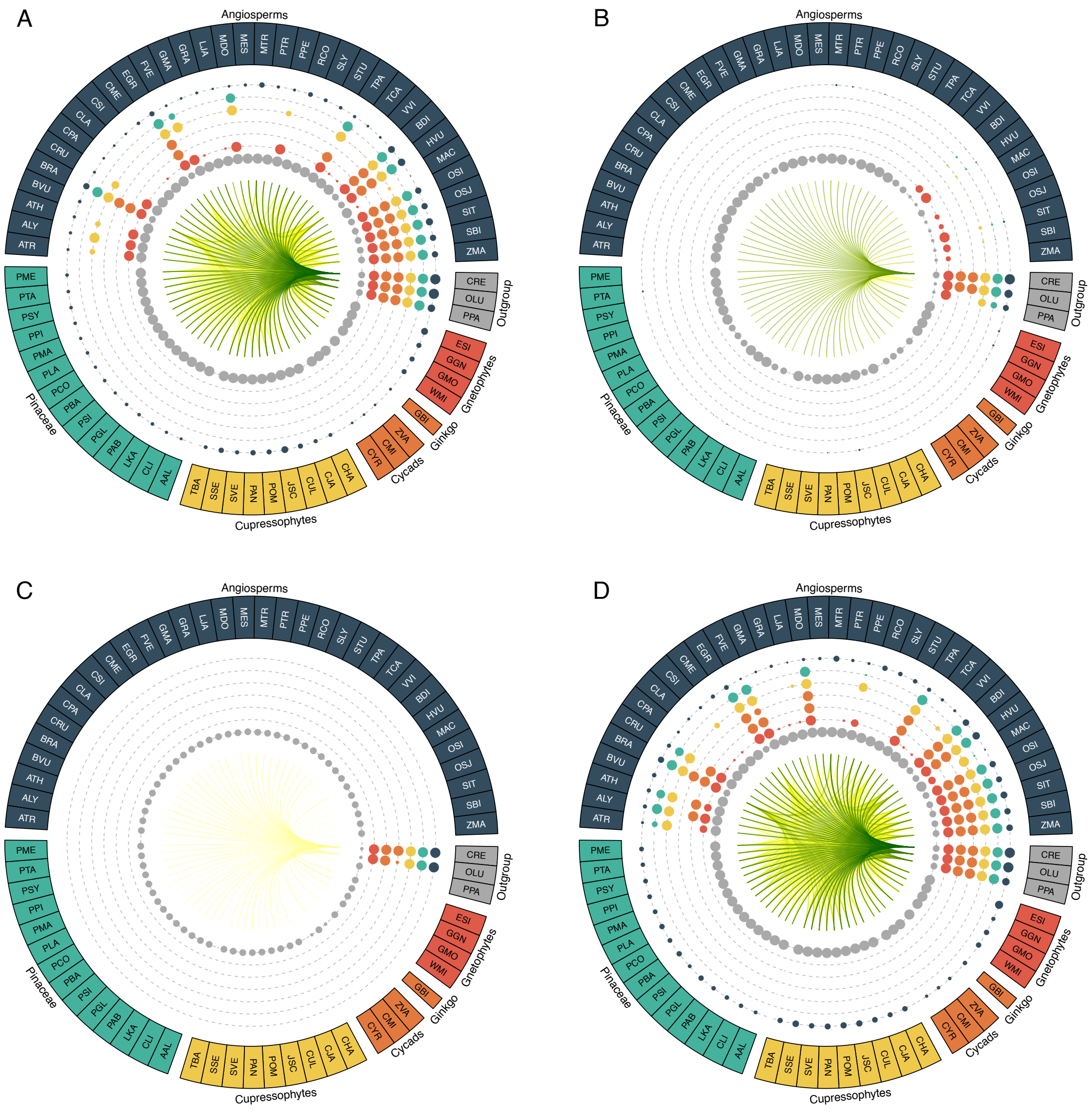
Supplementary Figure 9. Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants only with 2nd codon positions. Numbers on branches represent bootstrap values.



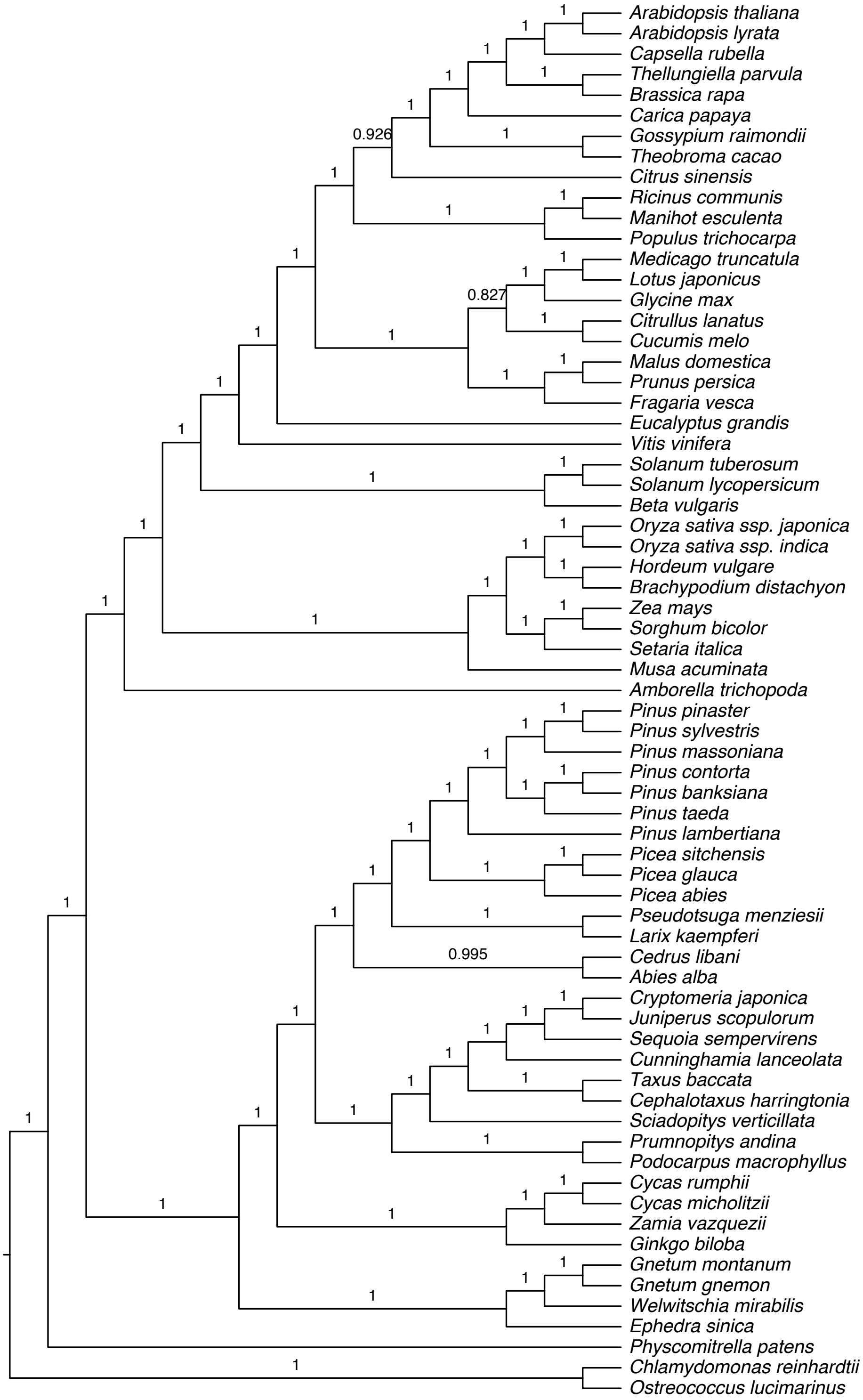
Supplementary Figure 10. GC content of the 106 phylogenetic markers at 1st, 2nd, and 3rd codon positions for the species studied.



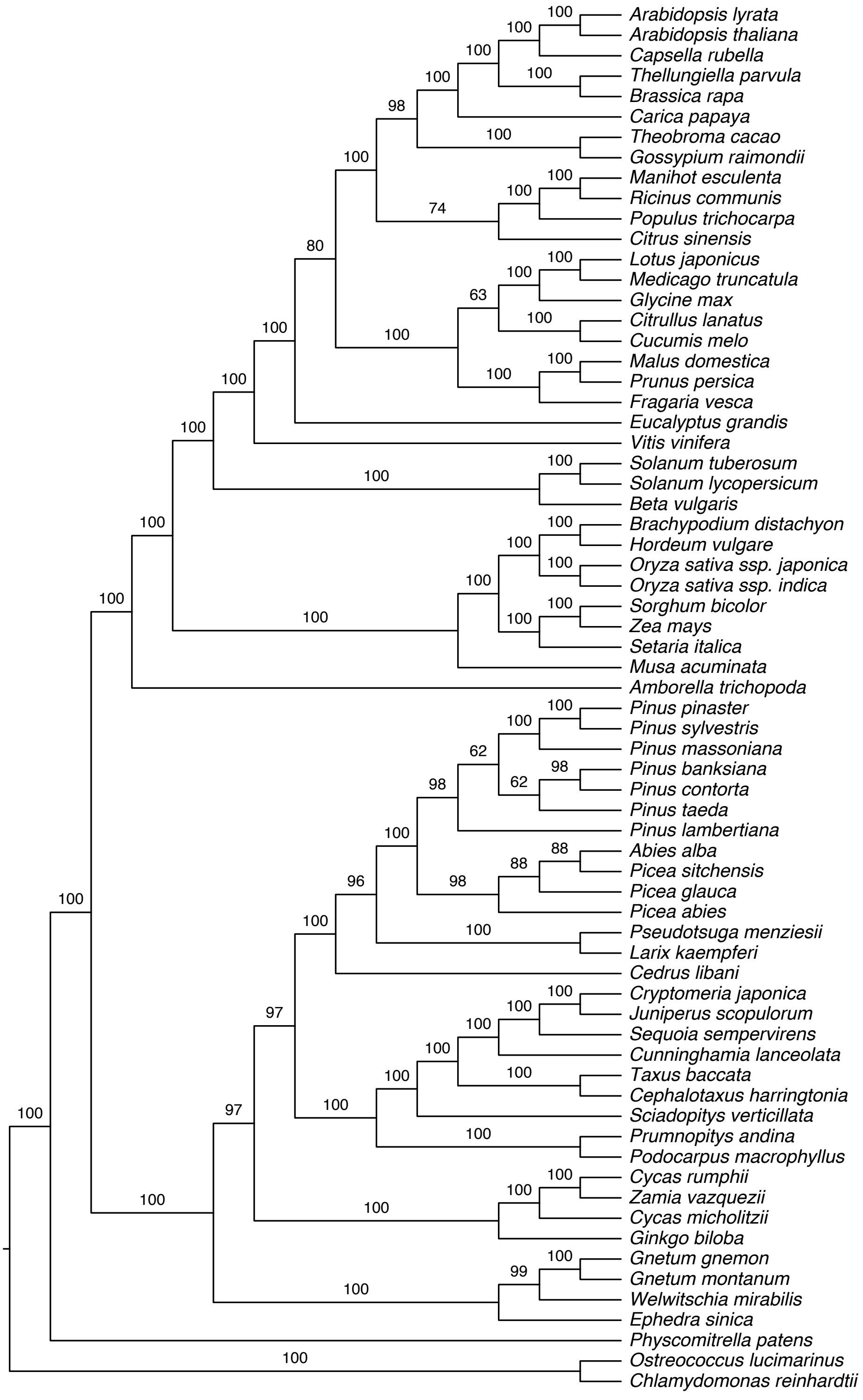
Supplementary Figure 11. GC content of the 106 phylogenetic markers at 1st, 2nd, and 3rd codon positions for the species studied after removing sites that encode the same amino acids.



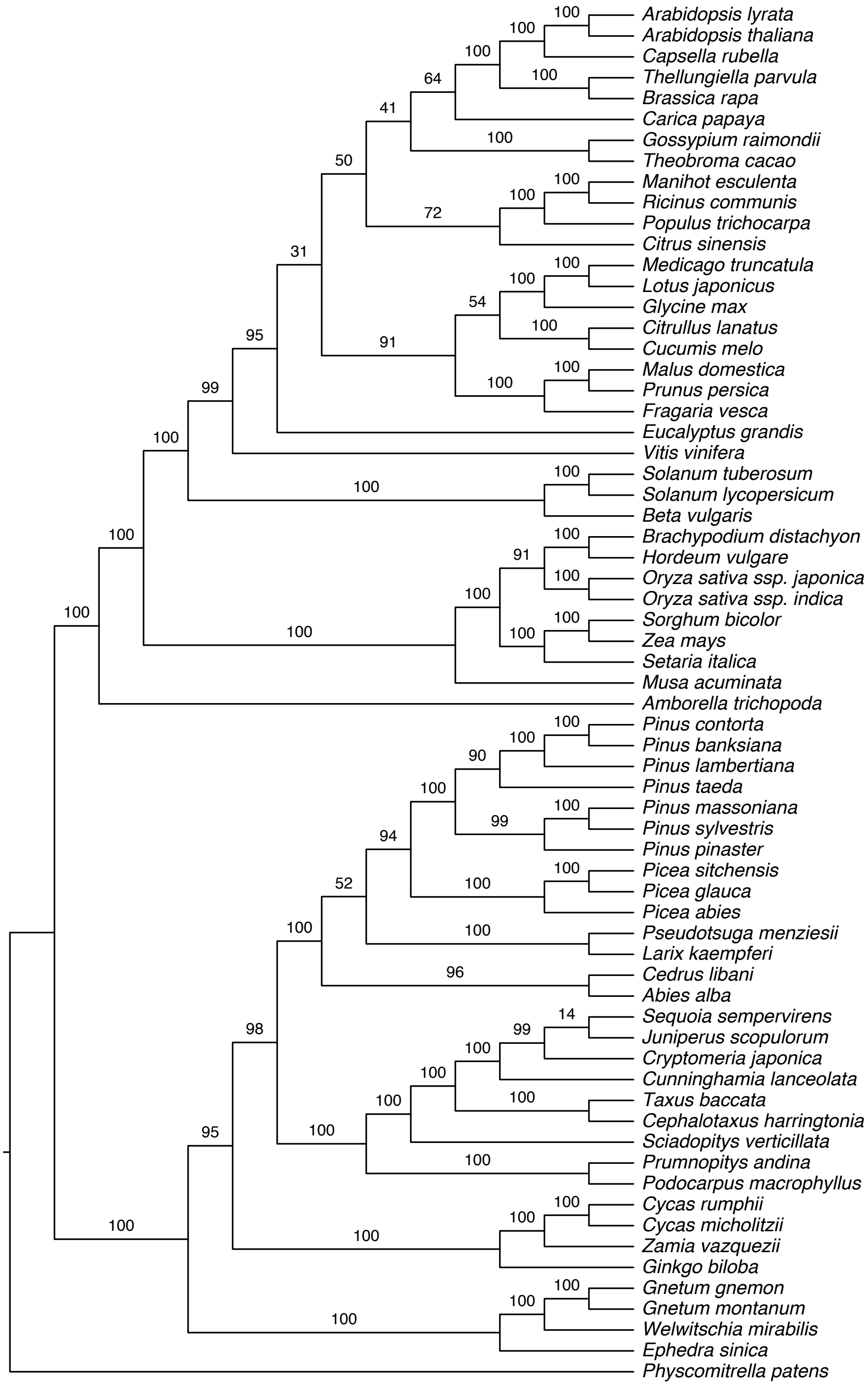
Supplementary Figure 12. Comparison of GC content in the concatenated alignment (A) and at each codon position (B, C, and D) from 106 genes in 68 species after removing sites that encode the same amino acids. Dot size correlates with the number of species in each lineage (group) that have a significantly different GC% (Wilcox test, $P < 1 \times 10^{-3}$) with the species compared with (colors of dots correspond to the compared lineages). Lines connecting any two species represent significant difference in GC content, with most significant in green and weakest in yellow (1×10^{-3}). The full names for the species can be found in Supplementary Table 3.



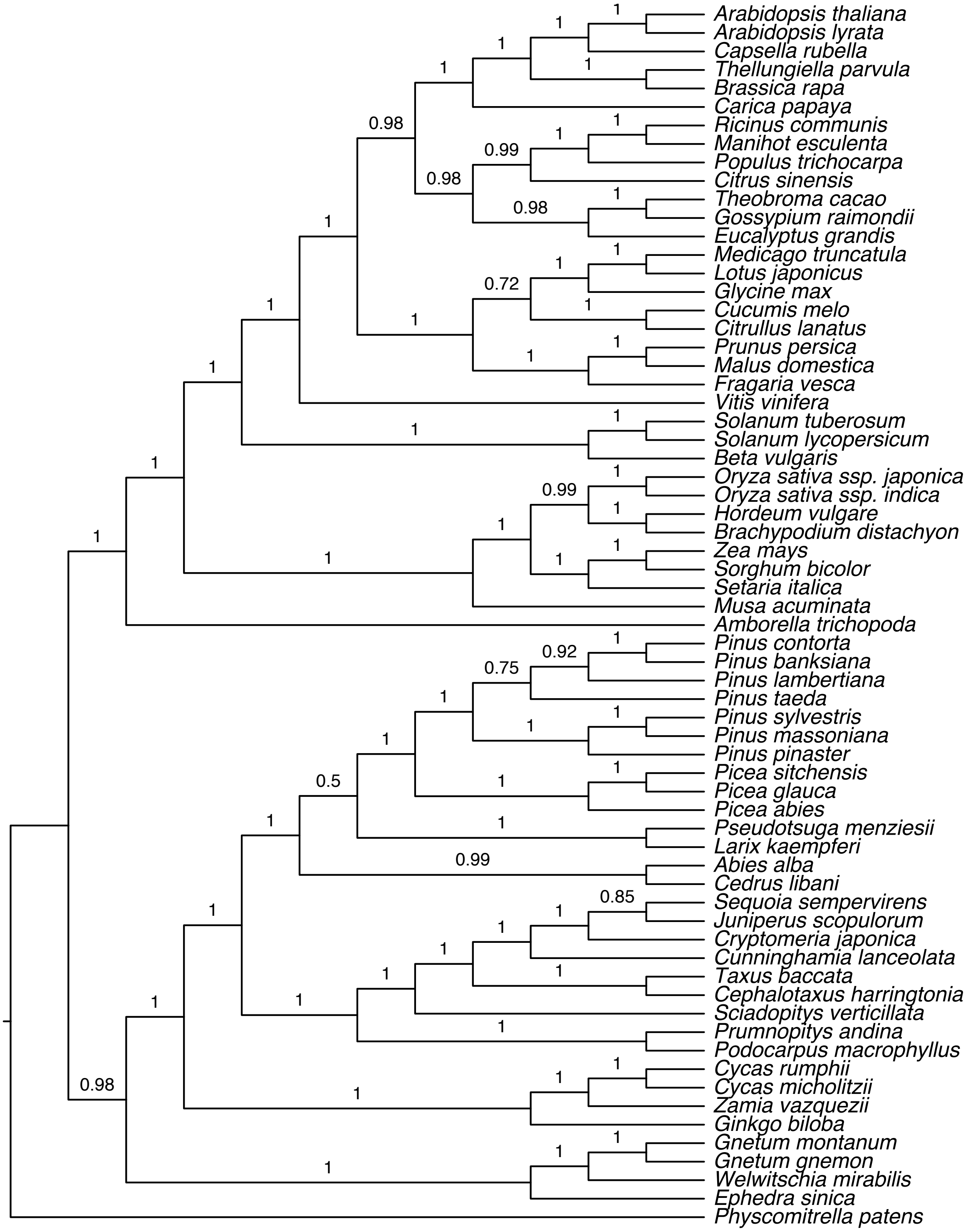
Supplementary Figure 13. Maximum likelihood tree inferred from a concatenated codon alignment of 106 single-copy genes using the Goldman and Yang (GY) model. Numbers on branches represent support values from the SH-like approximate likelihood-ratio test.



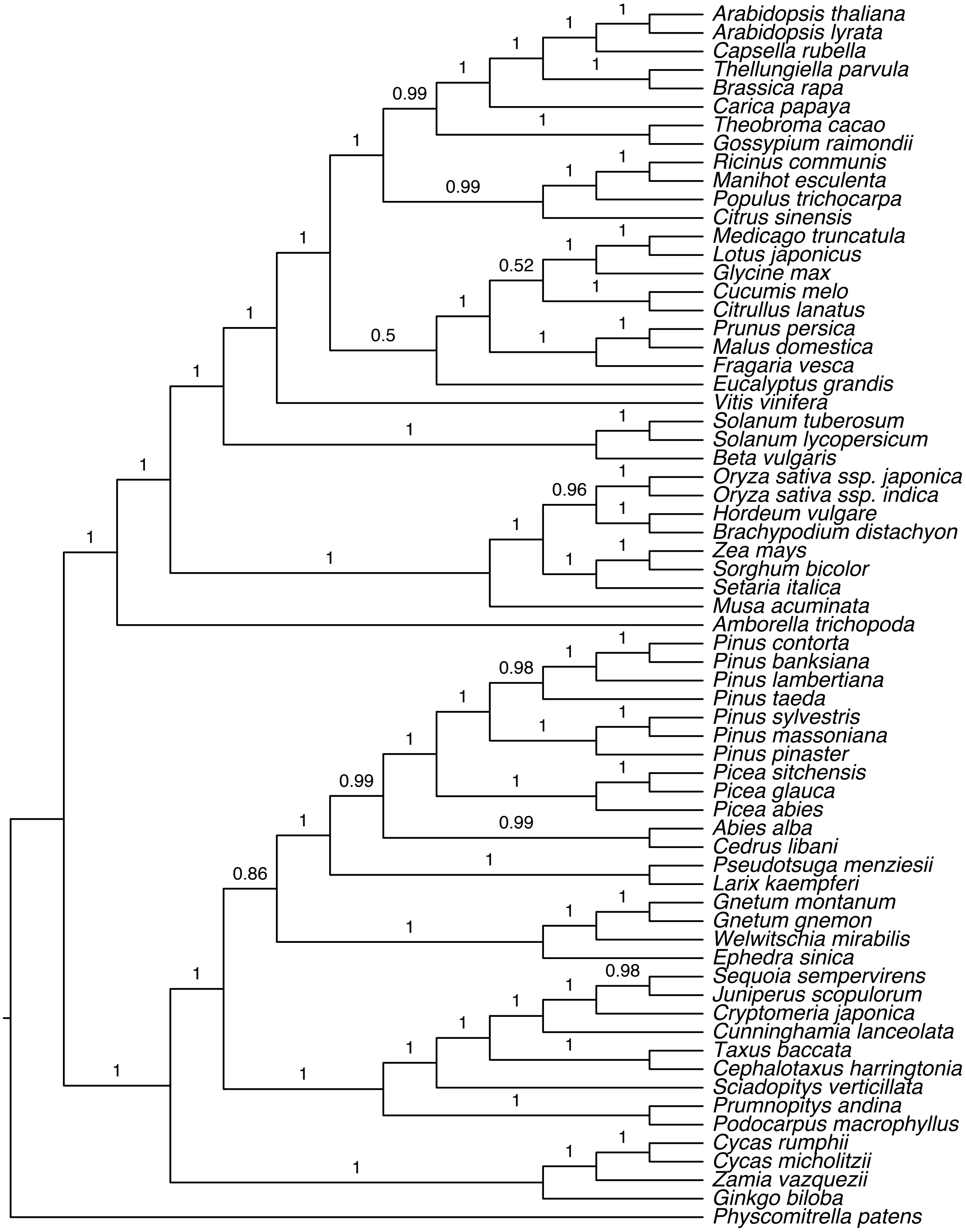
Supplementary Figure 14. Maximum likelihood tree inferred from a concatenated amino acid alignment of 106 single-copy genes using the JTT+I+GAMMA+F model. Numbers on branches represent bootstrap values.



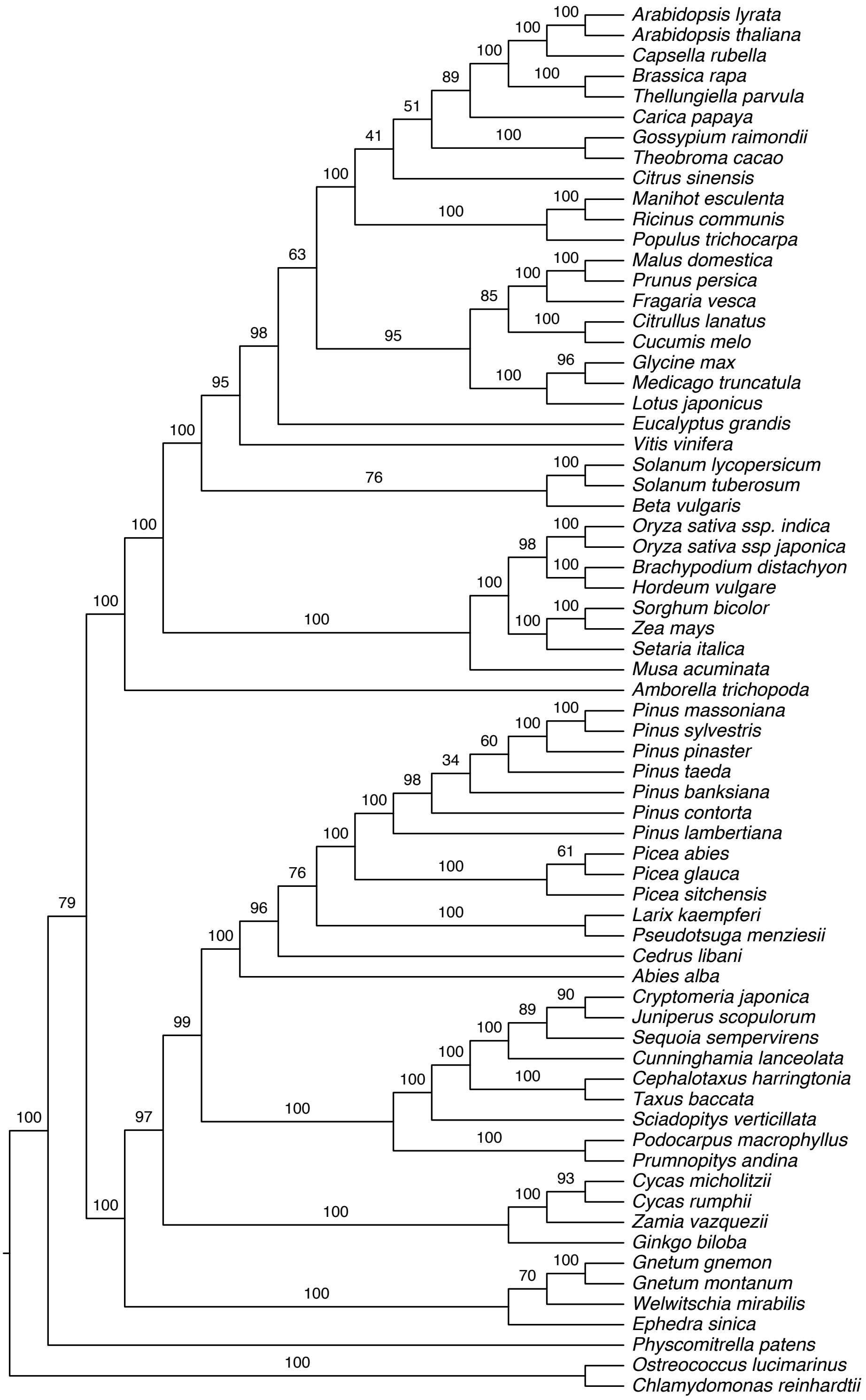
Supplementary Figure 15. Maximum likelihood tree inferred from a reduced concatenated amino acid alignment of 106 single-copy genes using the JTT+I+GAMMA+F model. Numbers on branches represent bootstrap values.



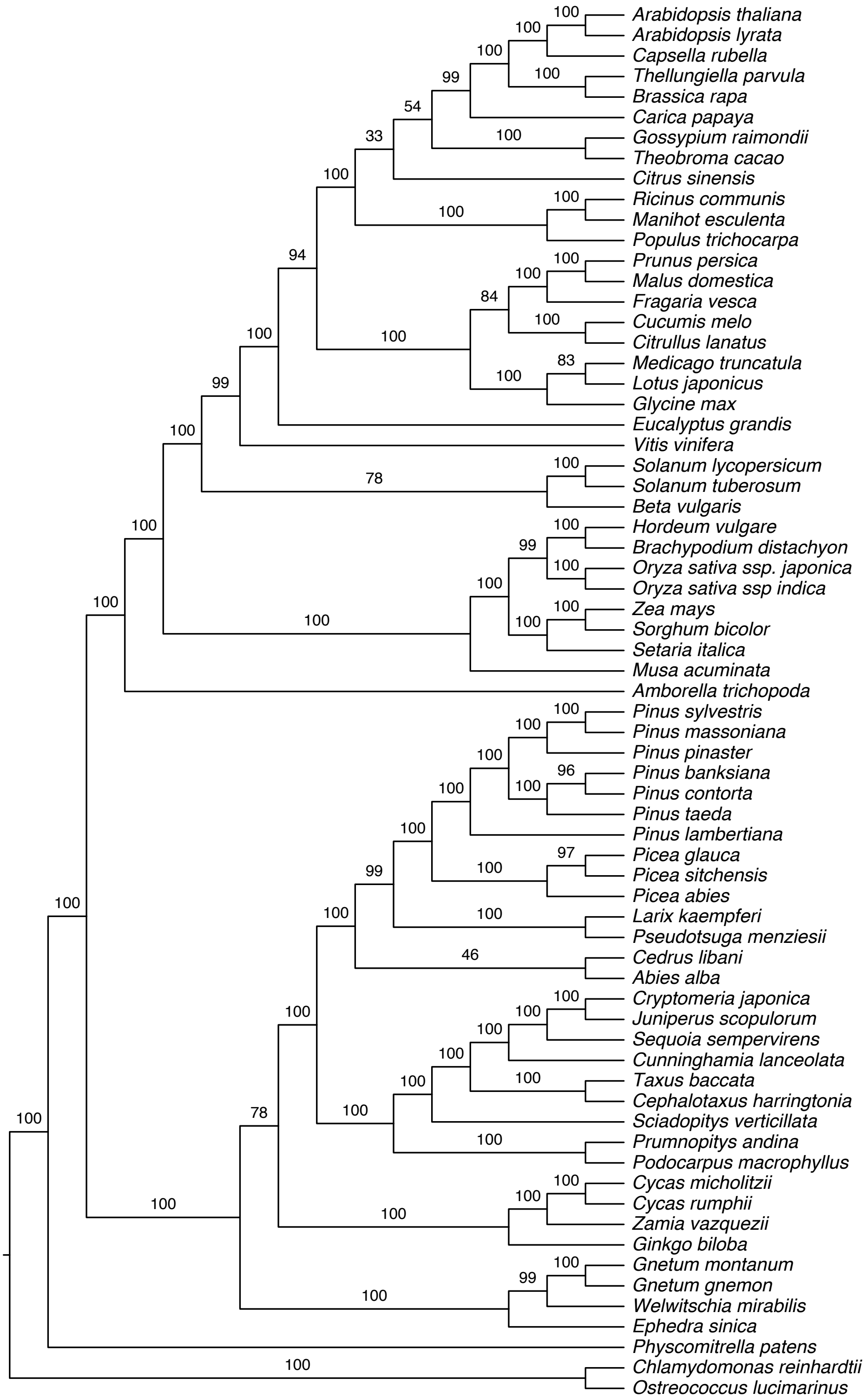
Supplementary Figure 16. Bayesian phylogenetic tree based on the reduced amino acid concatenation of 106 single-copy genes under CAT model. Numbers indicate posterior probabilities.



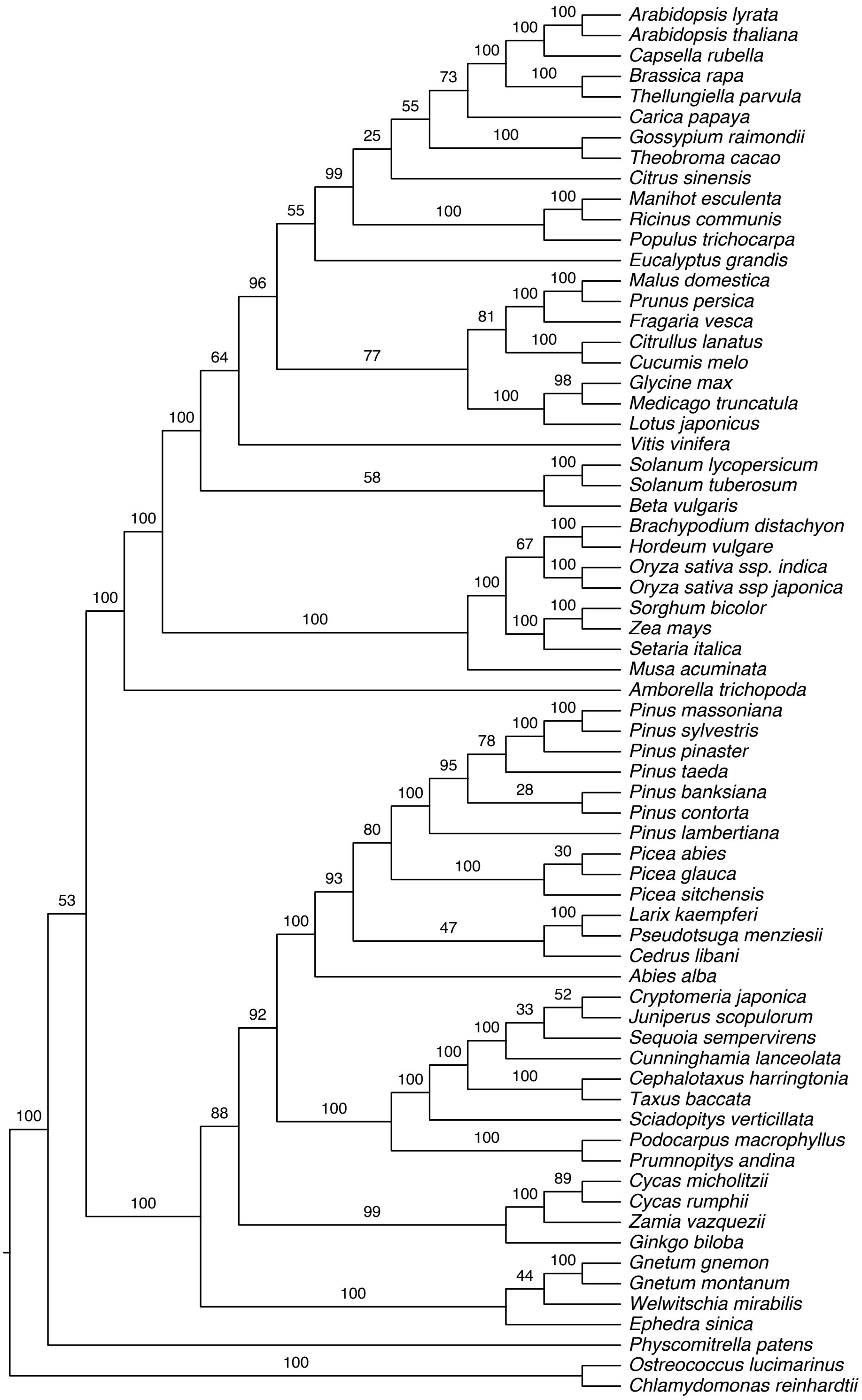
Supplementary Figure 17. Bayesian phylogenetic tree based on the reduced amino acid concatenation of 106 single-copy genes under CAT+GTR model. Numbers indicate posterior probabilities.



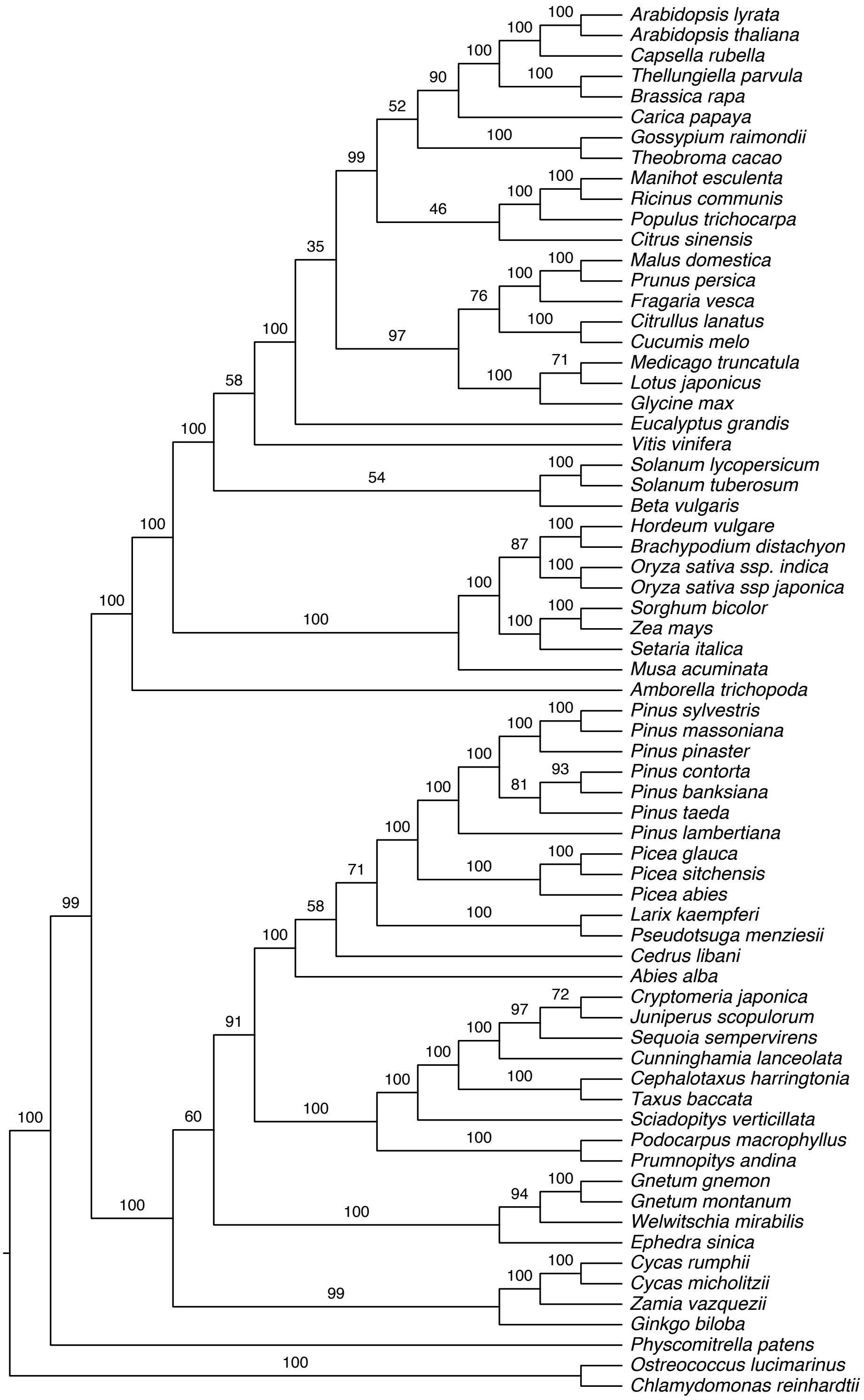
Supplementary Figure 18. Coalescent based tree inferred from gene trees of 106 single-copy genes by STAR. Numbers on branches represent bootstrap values.



Supplementary Figure 19. Coalescent based tree inferred from gene trees of 106 single-copy genes by ASTRA-II. Numbers on branches represent bootstrap values.



Supplementary Figure 20. Coalescent based tree inferred from gene trees of 106 single-copy genes by STAR. Gene trees were built without considering 3rd codon positions. Numbers on branches represent bootstrap values.



Supplementary Figure 21. Coalescent based tree inferred from gene trees of 106 single-copy genes by ASTRAL-II. Gene trees were built without considering 3rd codon positions. Numbers on branches represent bootstrap values.

Supplementary Table 1. Statistics of the sequencing reads in *Pinus pinaster* and *Pinus sylvestris*

	Tissue	Raw Data	# Reads for Assembly	# Bases for Assembly	% Reads for Assembly
Pinus pinaster	Shoot Apical Meristem	548,328	489,417	193,513,935	89.26%
	Shoot Apical Meristem	511,488	450,986	246,681,796	88.17%
	Cortex Root	595,575	388,906	144,683,636	65.30%
	Cortex Hypocotyl	552,098	445,605	238,646,186	80.71%
	Pith Hypocotyl	238,995	177,643	34,516,172	74.33%
	Pith Hypocotyl	141,112	110,838	31,040,865	78.55%
	Vascular Developing Root	596,858	511,873	351,624,842	85.76%
	Cortex Developing Root	466,708	380,969	251,267,660	81.63%
	Root Apical Meristem	475,279	422,455	153,747,617	88.89%
	Root Apical Meristem	605,083	535,391	338,062,907	88.48%
	Vascular Root	297,094	173,094	78,592,365	58.26%
	Vascular Root	448,120	262,767	129,947,819	58.64%
	Vascular Cotyledon	726,853	663,607	424,191,066	91.30%
	Mesophyll Cotyledon	776,974	720,010	460,528,808	92.67%
	Pith Hypocotyl	631,059	573,629	353,808,071	90.90%
	Vascular New Needle	747,050	691,860	472,097,944	92.61%
	Vascular Hypocotyl	678,337	607,026	363,679,917	89.49%
	Developing Needle	747,508	702,247	460,245,151	93.95%
	Mesophyll New Needle	682,094	601,778	328,943,737	88.23%
	Vascular Root	712,551	673,643	335,482,084	94.54%
	PPIN_454_Reads_Reg1.RL6.sff	493,631	394,947	151,121,677	80.01%
	PPIN_454_Reads_Reg2.RL6.sff	724,677	562,481	212,858,345	77.62%
	Total	**12,397,472**	**10,541,172**	**5,755,282,600**	**85.03%**
Pinus sylvestris	Early Embryo (E1)	603,508	539,840	250,762,894	89.45%
	Cleavage (E2)	625,195	573,611	279,768,670	91.75%
	Dominant Embryo (E3DO)	749,430	711,685	506,198,526	94.96%
	Megagametophyte (E3SU)	745,590	708,247	499,030,847	94.99%
	Dominant Embryo (E4)	817,722	780,954	552,917,241	95.50%
	Subordinate Embryos (M1)	758,018	725,345	512,694,678	95.69%
	Megagametophyte (M2)	712,754	671,451	465,102,991	94.21%
	Megagametophyte (M3)	800,707	768,722	521,997,230	96.01%
	Megagametophyte (M4)	789,571	754,048	516,959,353	95.50%
	Total	**6,602,495**	**6,233,903**	**4,105,432,430**	**94.42%**

Supplementary Table 2. Integrated transcriptomes for gymnosperms from public databases

Species	TreeGenes*	PlantGDB**	oneKP	# Transcripts	# Proteins
Abies alba	25,419	-	-	20,796	17,279
Cedrus libani	-	-	70,595	65,491	37,730
Cephalotaxus harringtonia	13,997	-	-	13,927	16,578
Cryptomeria japonica	347	24,299	-	21,922	19,688
Cunninghamia lanceolata	62,140	-	68,622	70,761	48,057
Cycas micholitzii	-	-	54,202	51,130	28,901
Cycas rumphii	-	10,901	22,908	21,844	12,619
Ephedra sinica	-	-	57,134	51,088	26,873
Ginkgo biloba	-	10,210	48,343	45,631	30,404
Gnetum gnemon	10,221	6,193	-	12,994	14,889
Gnetum montanum	-	-	70,316	65,123	32,549
Juniperus scopulorum	-	-	61,299	57,857	35,004
Larix kaempferi	57,484	-	-	53,274	40,692
Pinus banksiana	16	13,040	-	12,036	13,421
Pinus contorta	32	13,570	-	12,605	14,984
Pinus lambertiana	19,509	-	-	19,292	20,561
Pinus massoniana	69,738	-	-	55,525	46,423
Podocarpus macrophyllus	12,266	-	-	12,209	14,296
Prumnopitys andina	-	-	45,616	42,269	30,894
Pseudotsuga menziesii	393,638	9,857	-	246,282	149,717
Sciadopitys verticillata	11,955	-	51,723	40,310	30,035
Sequoia sempervirens	11,517	-	-	11,462	13,461
Taxus baccata	10,554	-	54,249	44,059	32,062
Welwitschia mirabilis	-	-	6,606	6,261	6,052
Zamia vazquezii	-	7,657	50,336	33,979	24,619

** TreeGenes includes ESTs, cDNAs, TSAs, and 454 assemblies*

***Cycas rumphii, Ginkgo biloba, Gnetum gnemon, and Zamia vazquezii are directly from PlantGDB*

Supplementary Table 3. The short name and clade of species used in the current study.

Species	Short name	Clade
Cephalotaxus harringtonia	CHA	Cupressophytes
Cryptomeria japonica	CJA	Cupressophytes
Cunninghamia lanceolata	CUL	Cupressophytes
Juniperus scopulorum	JSC	Cupressophytes
Podocarpus macrophyllus	POM	Cupressophytes
Prumnopitys andina	PAN	Cupressophytes
Sciadopitys verticillata	SVE	Cupressophytes
Sequoia sempervirens	SSE	Cupressophytes
Taxus baccata	TBA	Cupressophytes
Zamia vazquezii	ZVA	Cycads
Cycas micholitzii	CMI	Cycads
Cycas rumphii	CYR	Cycads
Amborella trichopoda	ATR	Angiosperms
Arabidopsis lyrata	ALY	Angiosperms
Arabidopsis thaliana	ATH	Angiosperms
Beta vulgaris	BVU	Angiosperms
Brassica rapa	BRA	Angiosperms
Capsella rubella	CRU	Angiosperms
Carica papaya	CPA	Angiosperms
Citrullus lanatus	CLA	Angiosperms
Citrus sinensis	CSI	Angiosperms
Cucumis melo	CME	Angiosperms
Eucalyptus grandis	EGR	Angiosperms
Fragaria vesca	FVE	Angiosperms
Glycine max	GMA	Angiosperms
Gossypium raimondii	GRA	Angiosperms
Lotus japonicus	LJA	Angiosperms
Malus domestica	MDO	Angiosperms
Manihot esculenta	MES	Angiosperms
Medicago truncatula	MTR	Angiosperms
Populus trichocarpa	PTR	Angiosperms
Prunus persica	PPE	Angiosperms
Ricinus communis	RCO	Angiosperms
Solanum lycopersicum	SLY	Angiosperms
Solanum tuberosum	STU	Angiosperms
Thellungiella parvula	TPA	Angiosperms
Theobroma cacao	TCA	Angiosperms
Vitis vinifera	VVI	Angiosperms
Ginkgo biloba	GBI	Ginkgo
Ephedra sinica	ESI	Gnetophytes

Species	**Short name**	**Clade**
Gnetum gnemon	GGN	Gnetophytes
Gnetum montanum	GMO	Gnetophytes
Welwitschia mirabilis	WMI	Gnetophytes
Brachypodium distachyon	BDI	Angiosperms
Hordeum vulgare	HVU	Angiosperms
Musa acuminata	MAC	Angiosperms
Oryza sativa ssp indica	OSI	Angiosperms
Oryza sativa ssp japonica	OSJ	Angiosperms
Setaria italica	SIT	Angiosperms
Sorghum bicolor	SBI	Angiosperms
Zea mays	ZMA	Angiosperms
Chlamydomonas reinhardtii	CRE	Outgroup
Ostreococcus lucimarinus	OLU	Outgroup
Physcomitrella patens	PPA	Outgroup
Abies alba	AAL	Pinaceae
Cedrus libani	CLI	Pinaceae
Larix kaempferi	LKA	Pinaceae
Picea abies	PAB	Pinaceae
Picea glauca	PGL	Pinaceae
Picea sitchensis	PSI	Pinaceae
Pinus banksiana	PBA	Pinaceae
Pinus contorta	PCO	Pinaceae
Pinus lambertiana	PLA	Pinaceae
Pinus massoniana	PMA	Pinaceae
Pinus pinaster	PPI	Pinaceae
Pinus sylvestris	PSY	Pinaceae
Pinus taeda	PTA	Pinaceae
Pseudotsuga menziesii	PME	Pinaceae