Supplementary Information

For

Widespread impact of horizontal gene transfer on plant colonization of land

Jipei Yue[1,2] Xiangyang Hu[1,3] Hang Sun[1] Yongping Yang[1,3] Jinling Huang[2]

[1]Key Laboratory of Biodiversity and Biogeography, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China

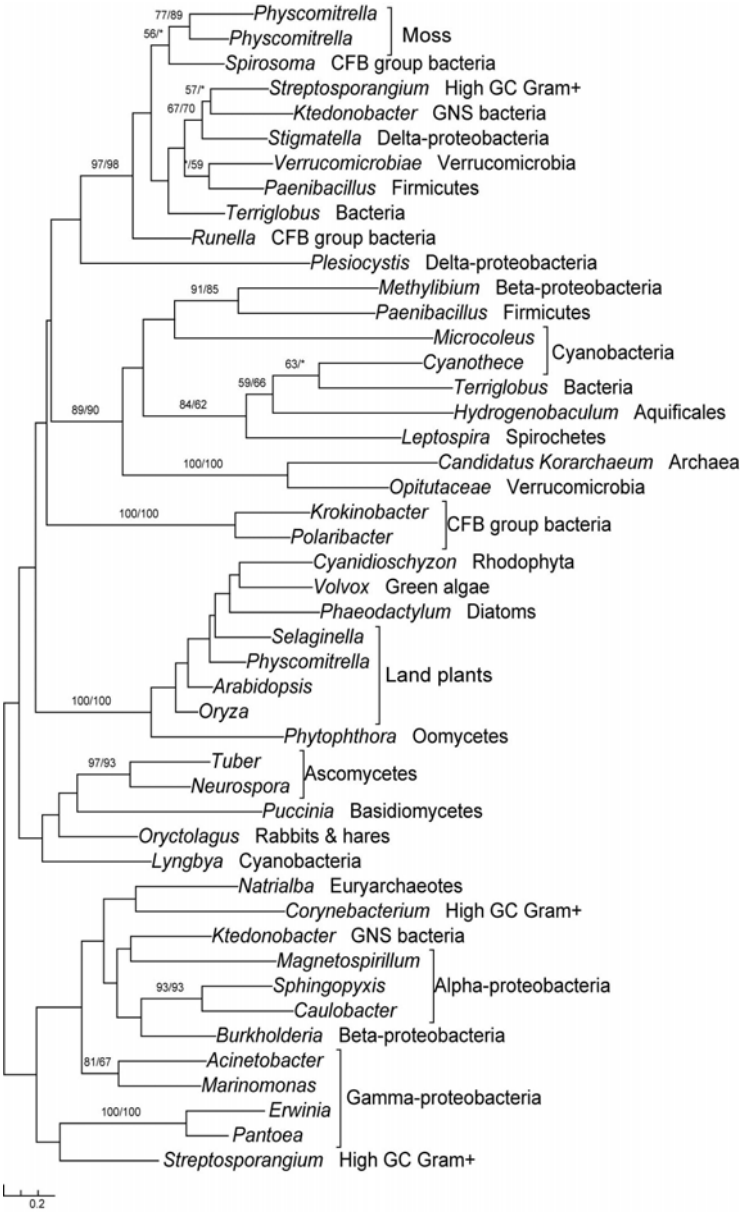
[2]Department of Biology, East Carolina University, Greenville, NC 27858, USA

[3]Institute of Tibet Plateau Research, Chinese Academy of Sciences, Kunming 650201, China

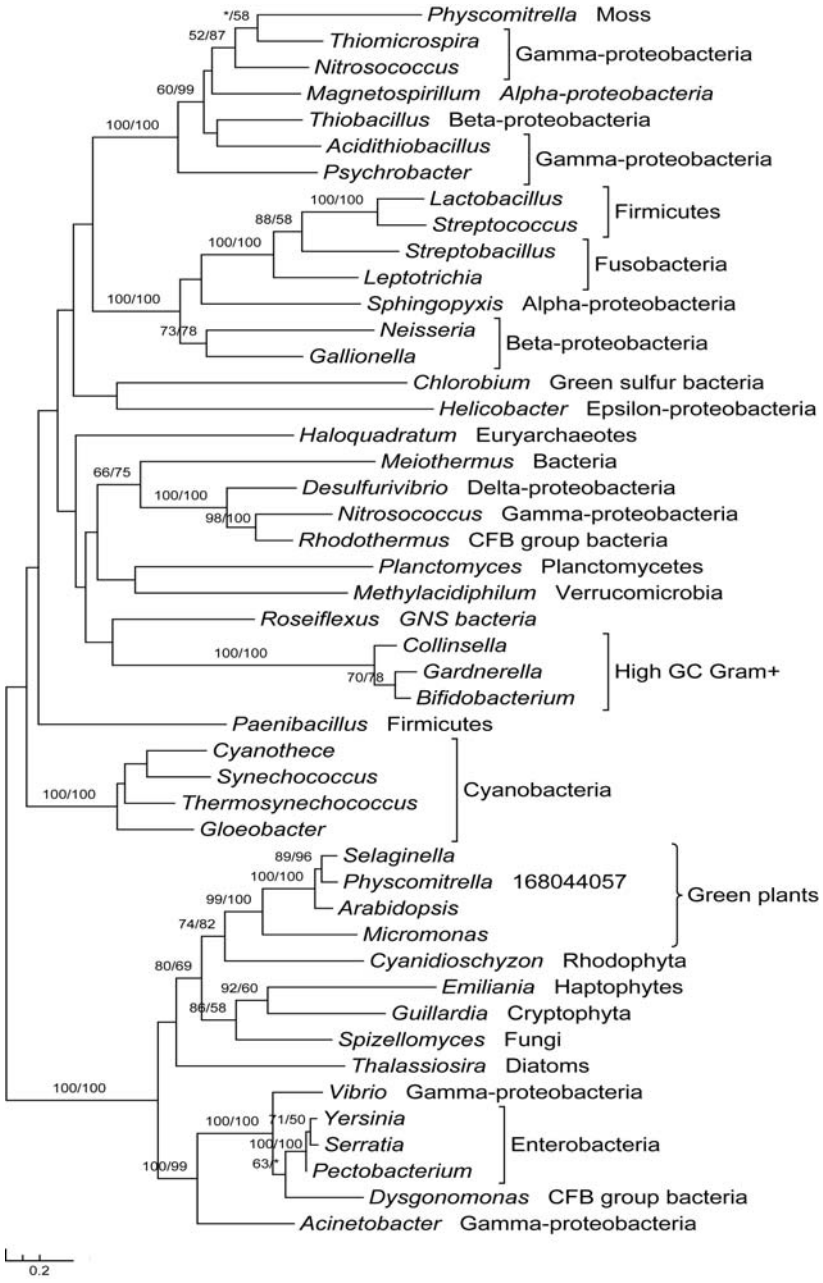
Correspondence: Jinling Huang

Email: huangj@ecu.edu

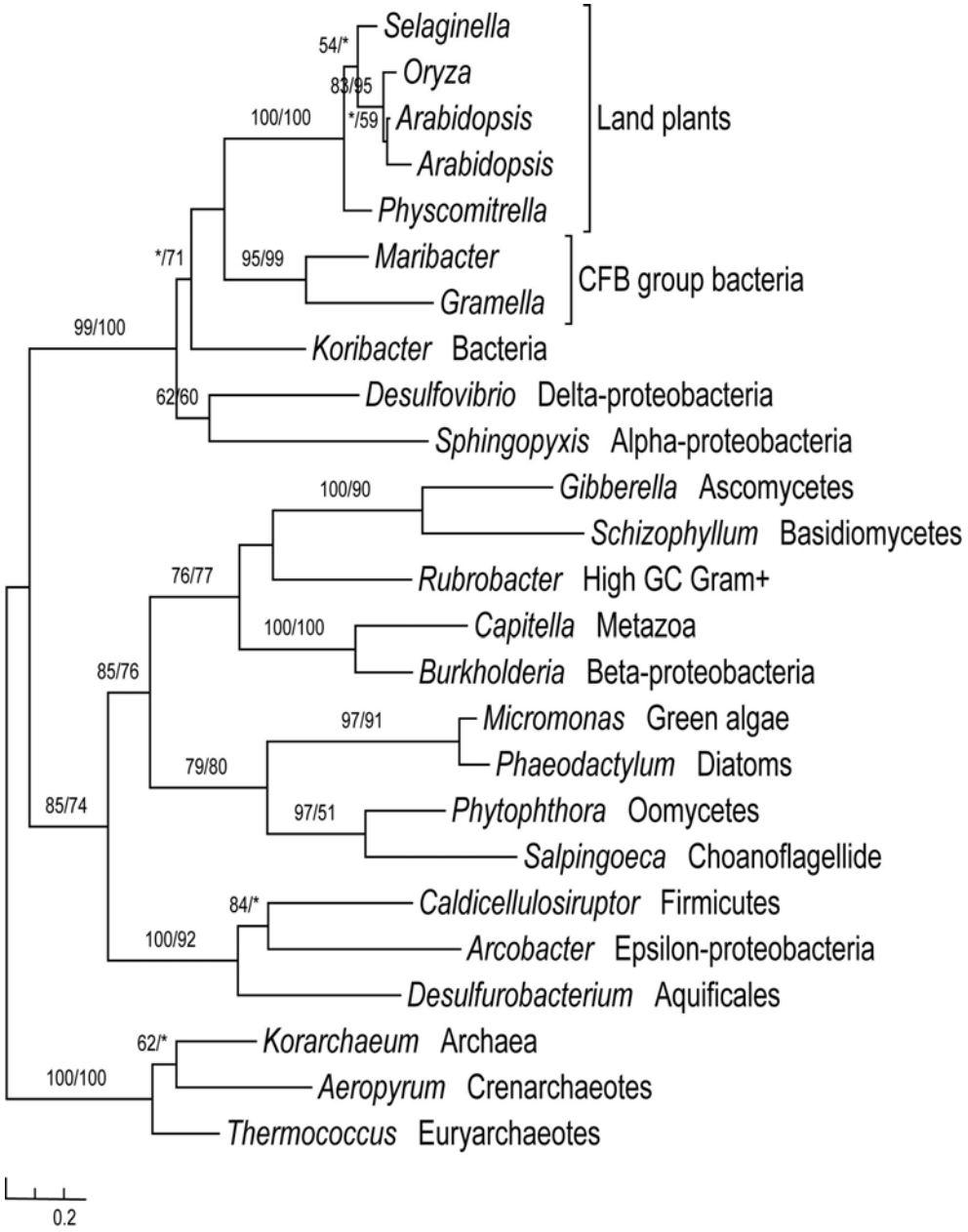
Fax: 252-328-4178



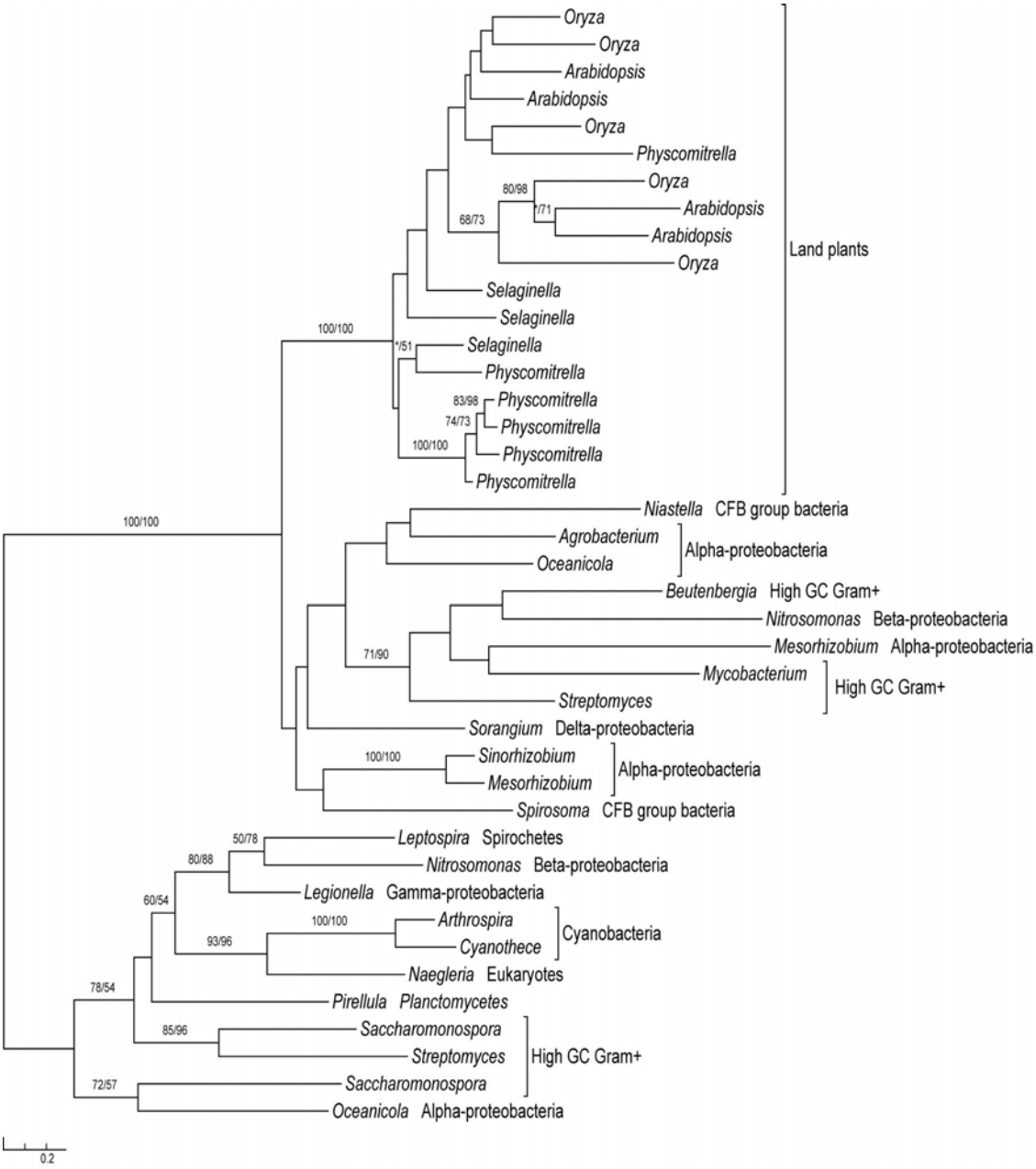
Supplementary Figure S1. Molecular phylogeny of FAD linked oxidase. Three copies of this gene exist in *Physcomitrella*, two of which (Genbank GI numbers 168012414, 168045341) group within bacterial sequences (upper part of the tree); the other copy (Genbank GI number 168012432) of this gene is a mitochondrial precursor in several other eukaryotes.



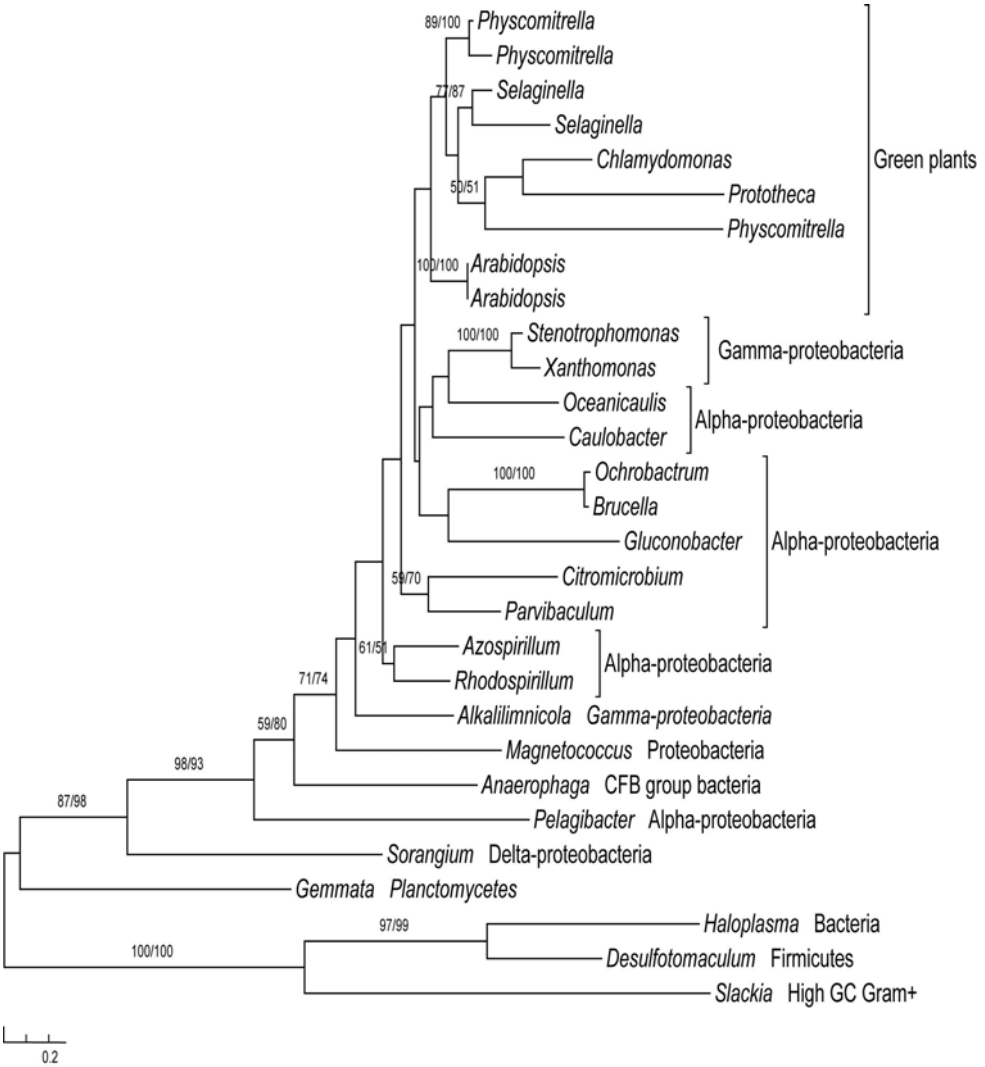
Supplementary Figure S2. Molecular phylogeny of phosphoenolpyruvate carboxylase (PEPCase). *Physcomitrella* sequence (Genbank GI number 168029489) forms a highly supported clade with proteobacterial homologs. Another gene copy in *Physcomitrella* (Genbank GI number 168044057) groups with homologs from green plants, red algae and other eukaryotes. In *Arabidopsis*, this copy is annotated as phosphoenolpyruvate carboxylase 2 (ATPPC2) and is targeted to chloroplasts. Some of other eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating likely mitochondrial origin.



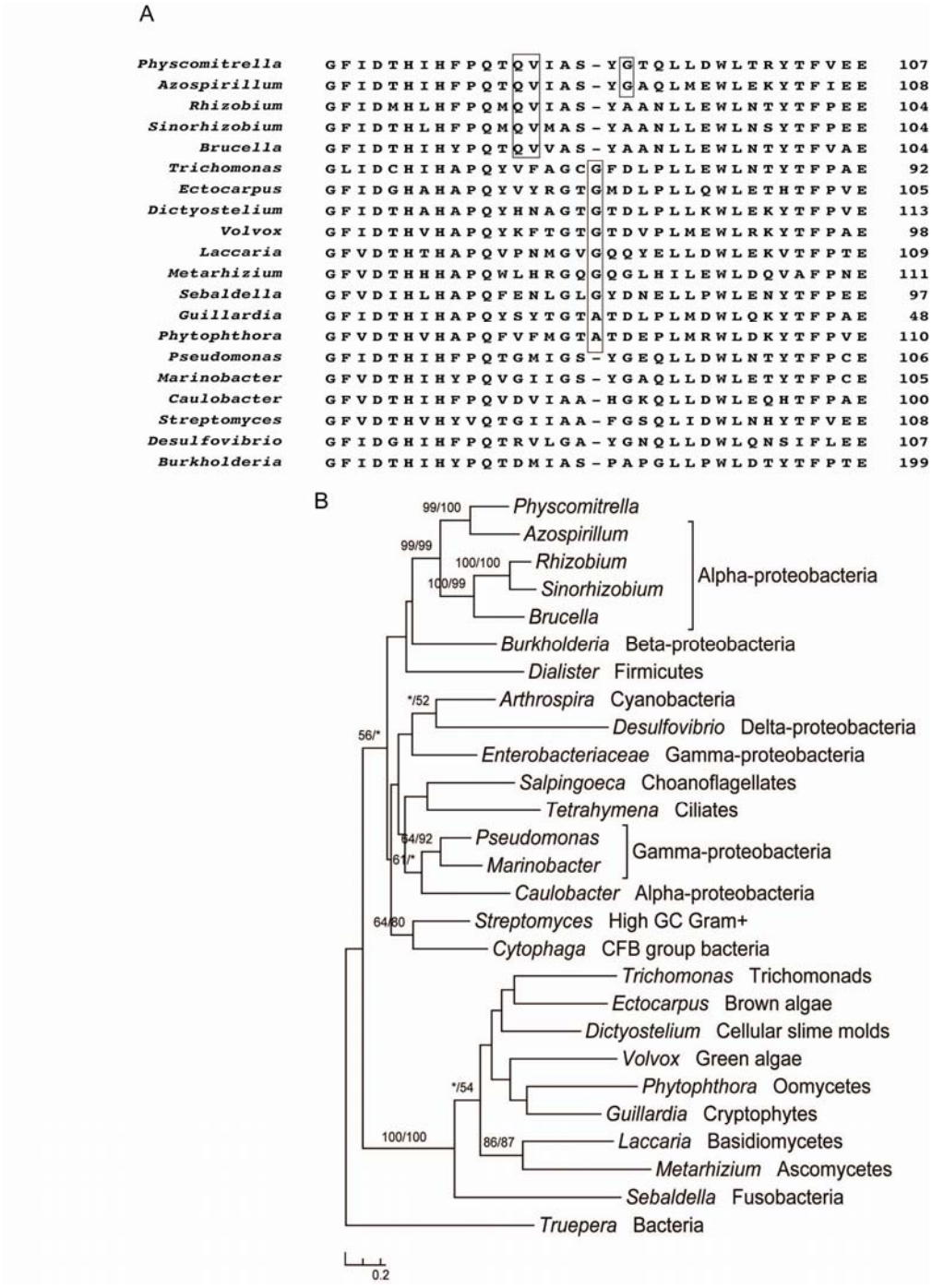
Supplementary Figure S3. Molecular phylogeny of arginase. *Physcomitrella* sequence (Genbank GI number 168024860) forms a highly supported clade with homologs from other land plants and bacteria. Their relationship is supported by several conserved amino acid residues and shared indels. Several other eukaryotic sequences form another clade with bacterial homologs. Some of these eukaryotic sequences were predicted by TargetP to be mitochondrial precursors, indicating that they are likely derived from mitochondria.



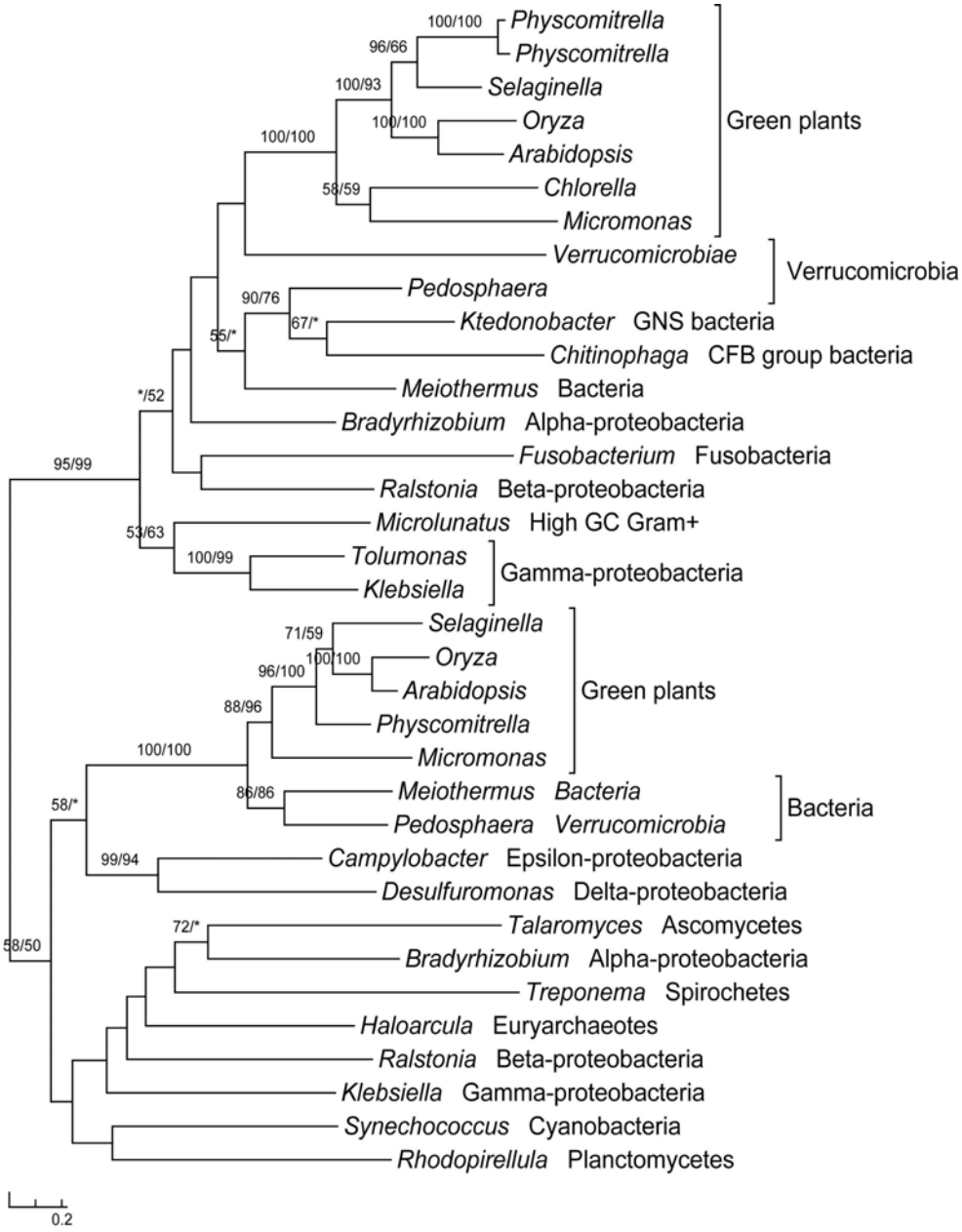
Supplementary Figure S4. Molecular phylogeny of YUCCA flavin monooxygenase (YUCCA3). *Physcomitrella* sequences (Genbank GI numbers 168013839， 168007310， 168038243，168047840，168059684) form a highly supported clade with other land plant and bacterial homologs. *Naegleria* sequence forms a clade with cyanobacterial homologs, which in turn is related to other bacterial homologs. The relationship of the two major sequence clades depicted in this gene tree is also supported by multiple conserved amino acid residues and shared indels.



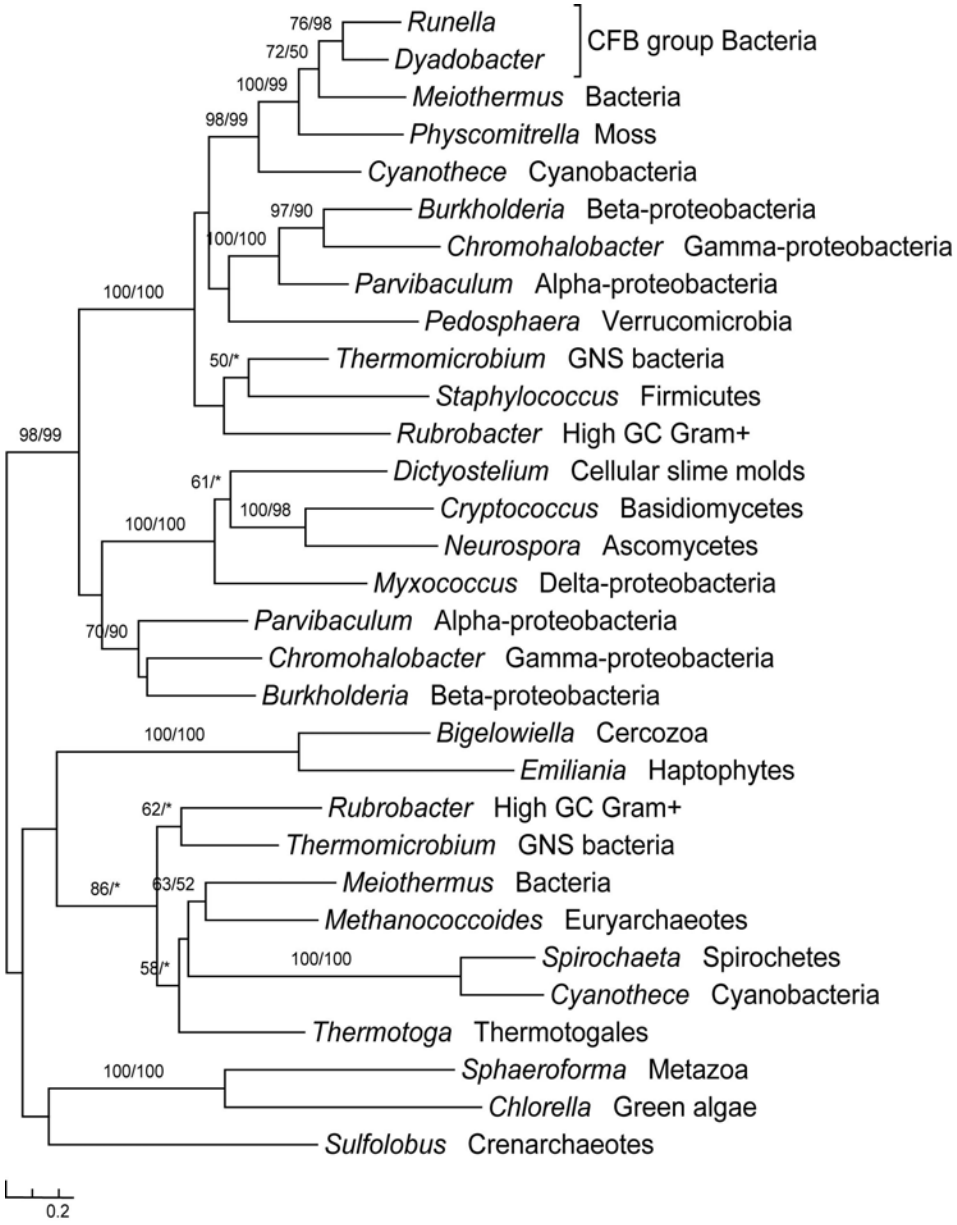
Supplementary Figure S5. Molecular phylogeny of glutamate-cysteine ligase (GCL). Identifiable homologs were only found in green plants and bacteria. *Physcomitrella* sequences (Genbank GI numbers 168009654, 168067242, 168052608) form a monophyletic group with green plant and proteobacterial homologs. No cyanobacterial homologs were found, indicating that this gene family in green plants is unlikely of plastid (cyanobacterial) origin.



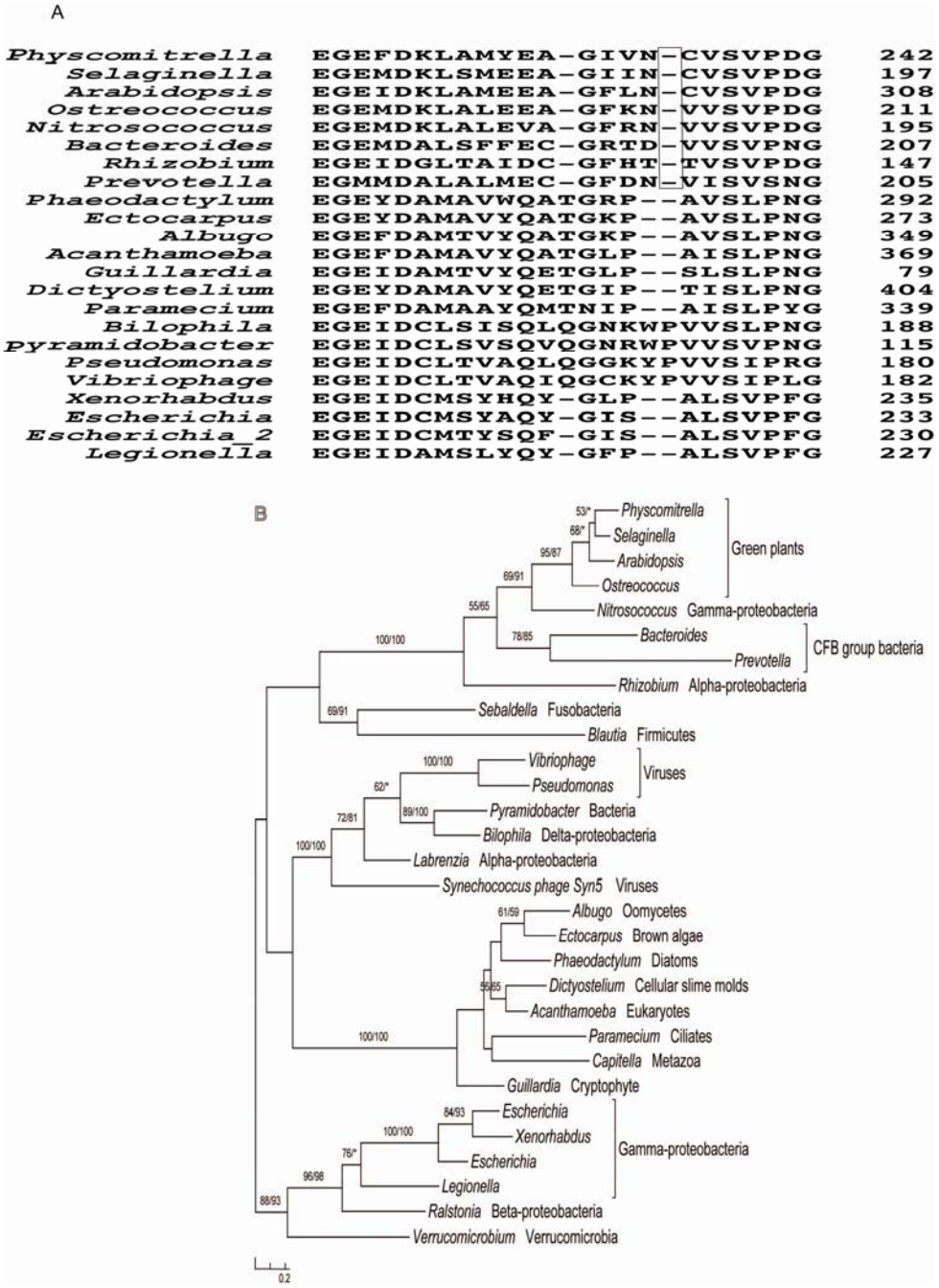
Supplementary Figure S6. Multiple protein sequence alignment (A) and molecular phylogeny of guanine deaminase (B). *Physcomitrella* sequence (Genbank GI number 168025229) has 55-68% identity with alpha-proteobacterial homologs. Several other eukaryotic sequences group with bacterial homologs. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating likely mitochondrial origin. The highly supported clades on the phylogenetic tree are also supported by several conserved amino acid residues and shared indels.



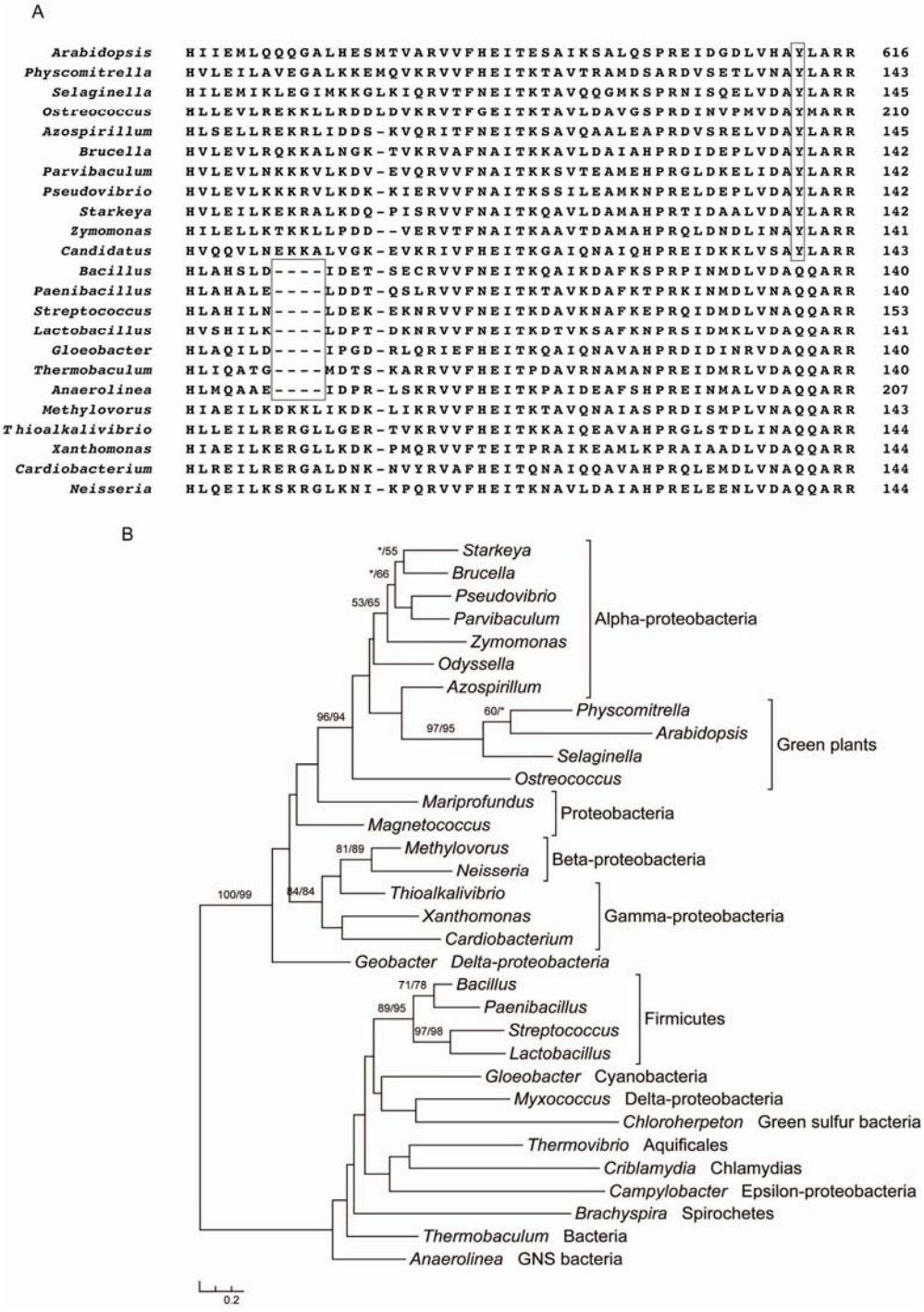
Supplementary Figure S7. Phylogenetic analyses of allantoate amidohydrolase (AAH) and ureidoglycolate amidohydrolase. Identifiable homologs of *Physcomitrella* AAH (upper part of the tree; Genbank GI numbers 167997139，168064079) are only found in green plants and bacteria. *Physcomitrella* ureidoglycolate amidohydrolase sequence (lower part of the tree; Genbank GI number 168010247) forms a highly supported clade with homologs from green plants and bacteria. The relationship of these two gene families is supported by multiple conserved amino acid residues.



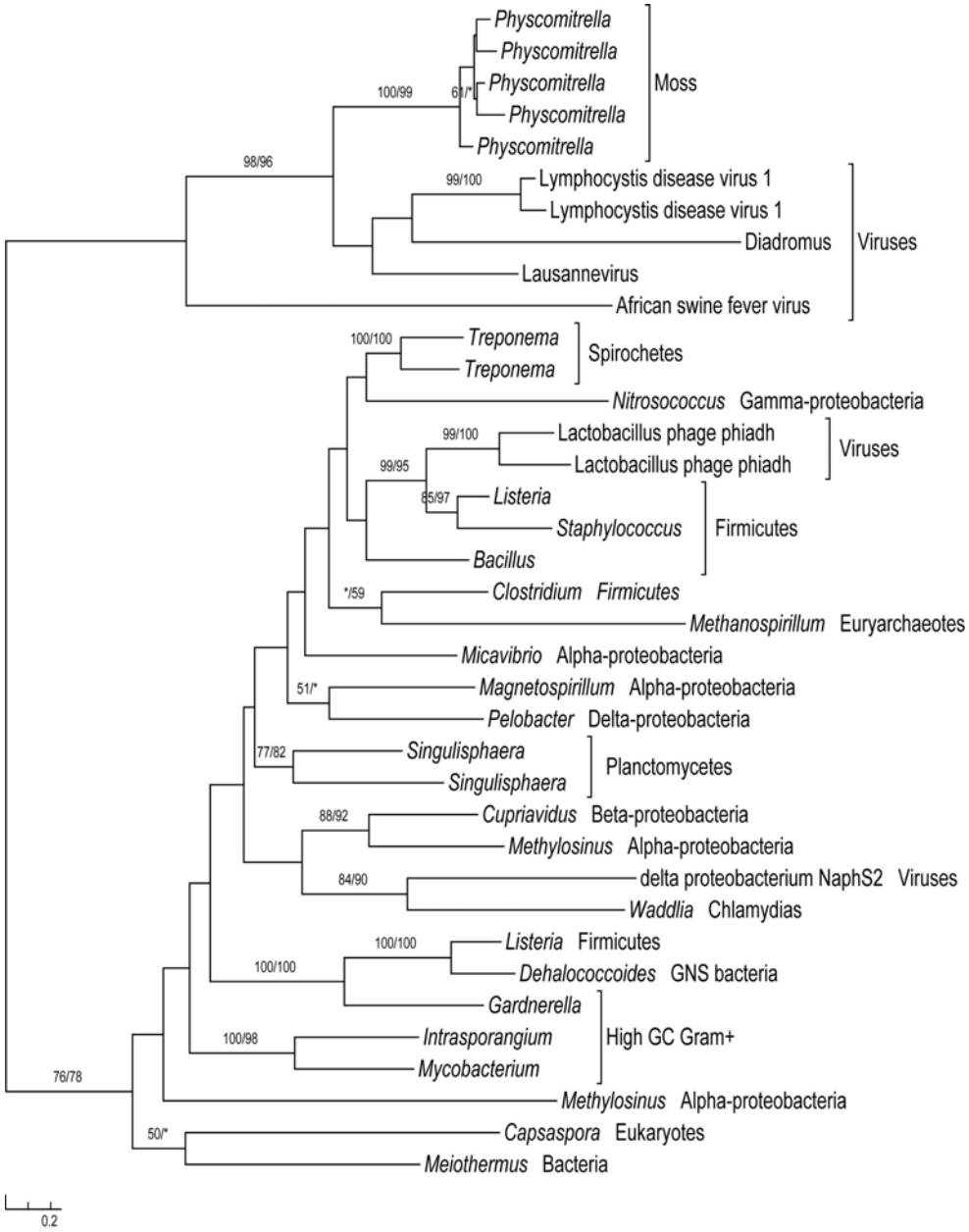
Supplementary Figure S8. Molecular phylogeny of glutamine synthetase. *Physcomitrella* sequence (Genbank GI numbers 168040136) has 59-61% identity with homologs of bacterial *Runella*, *Dyadobacter*, and *Meiothermus*. These sequences form a highly supported clade. Several other eukaryotic sequences group with miscellaneous bacterial homologs. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating a likely mitochondrial or alpha-preoteobacteria origin.



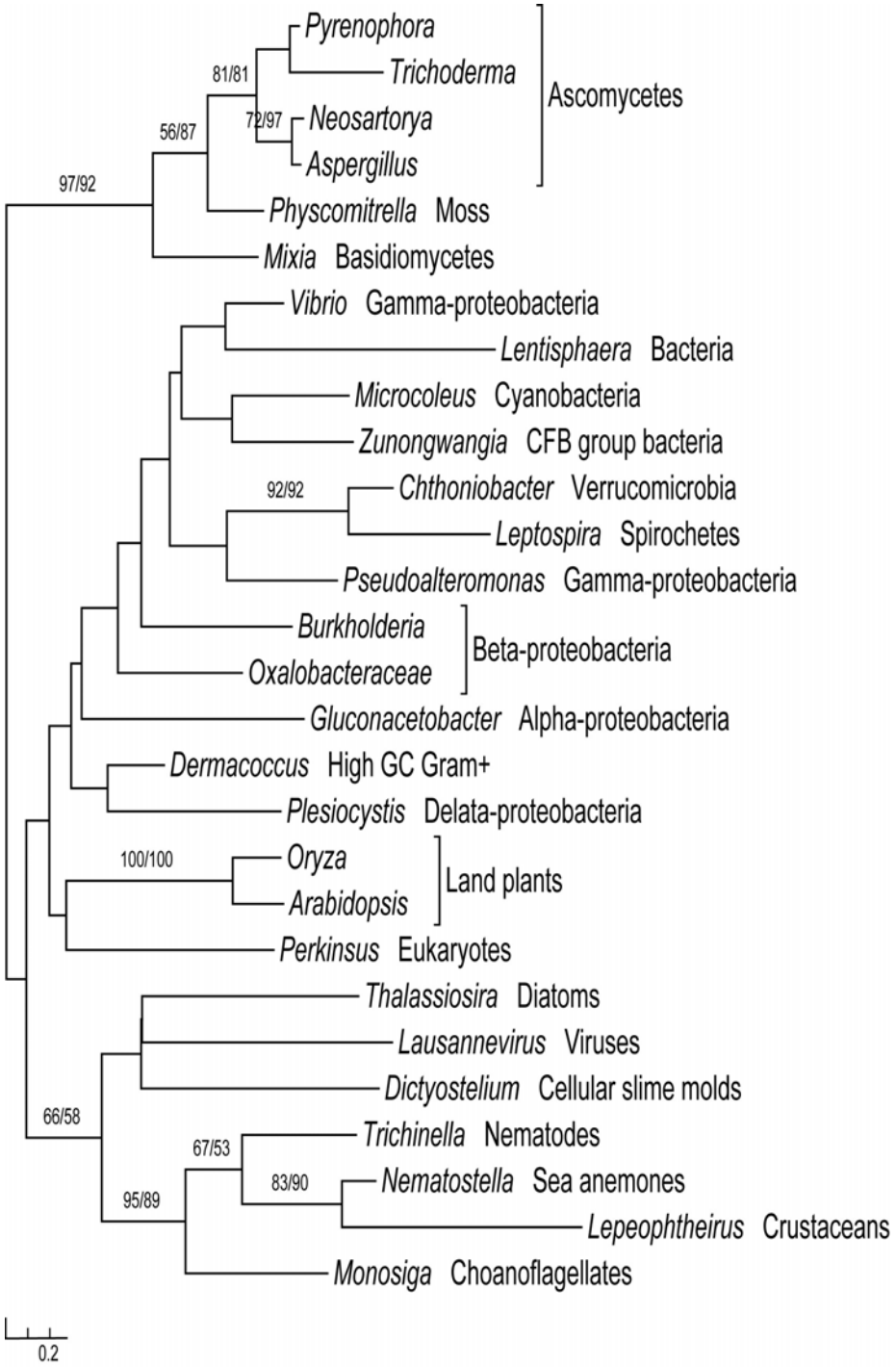
Supplementary Figure S9. Multiple protein sequence alignment (A) and molecular phylogeny of toprim domain-containing protein. *Physcomitrella* sequence (Genbank GI number 168040643) forms a clade with green plant and bacterial homologs. Other eukaryotic sequences form another clade. Some of these other eukaryotic sequences are predicted to mitochondrial precursors by TargetP, indicating likely mitochondrial origin. The closer relationship between green plant sequence and bacterial homologs is also supported by multiple shared indels.



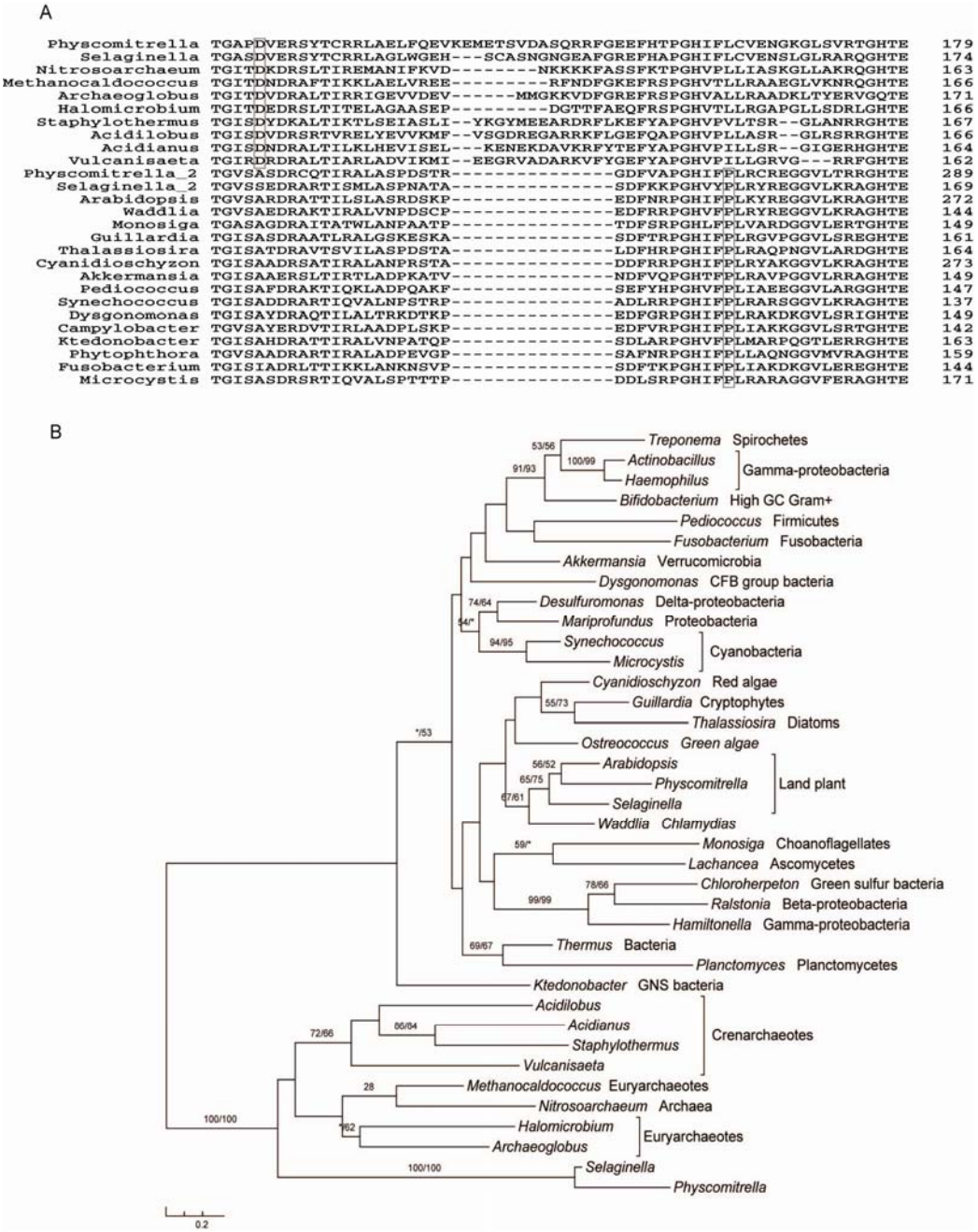
Supplementary Figure S10. Multiple protein sequence alignment (A) and molecular phylogeny of DNA topoisomerase I (B). *Physcomitrella* sequence (Genbank GI number 168037859) forms a clade with homologs from green plants and alpha-proteobacteria. Identifiable homologs were only found in green plants and bacteria. Highly supported clades on the phylogenetic tree are also supported by several conserved amino acid residues and shared indels.



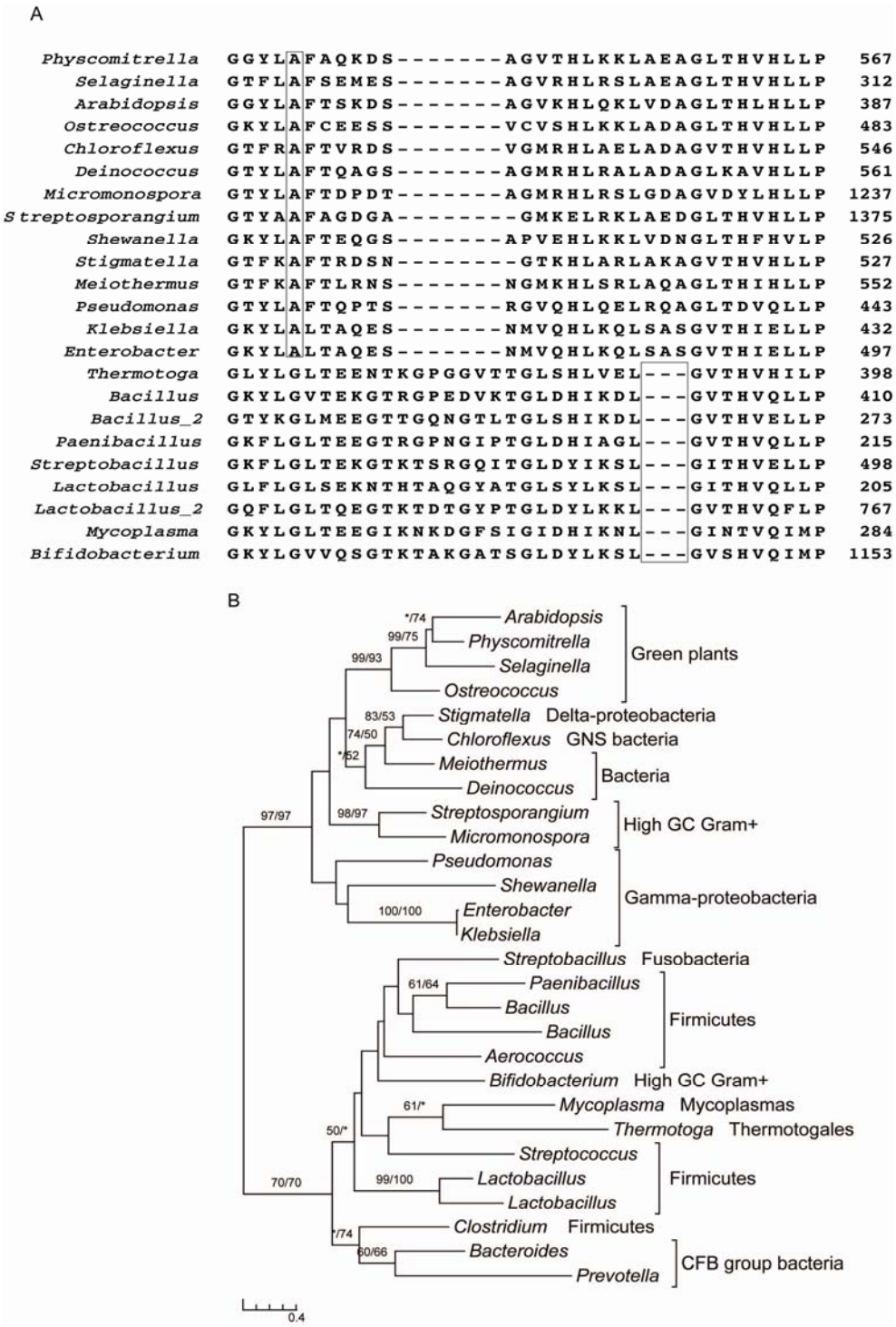
Supplementary Figure S11. Molecular phylogeny of phage/plasmid primase, P4 family. Identifiable homologs of moss sequences are only found in viruses and bacteria. *Physcomitrella* sequences (Genbank GI numbers 168026035, 168057313, 168032336, 168009191, 168041210) form a highly supported clade with viral homologs.



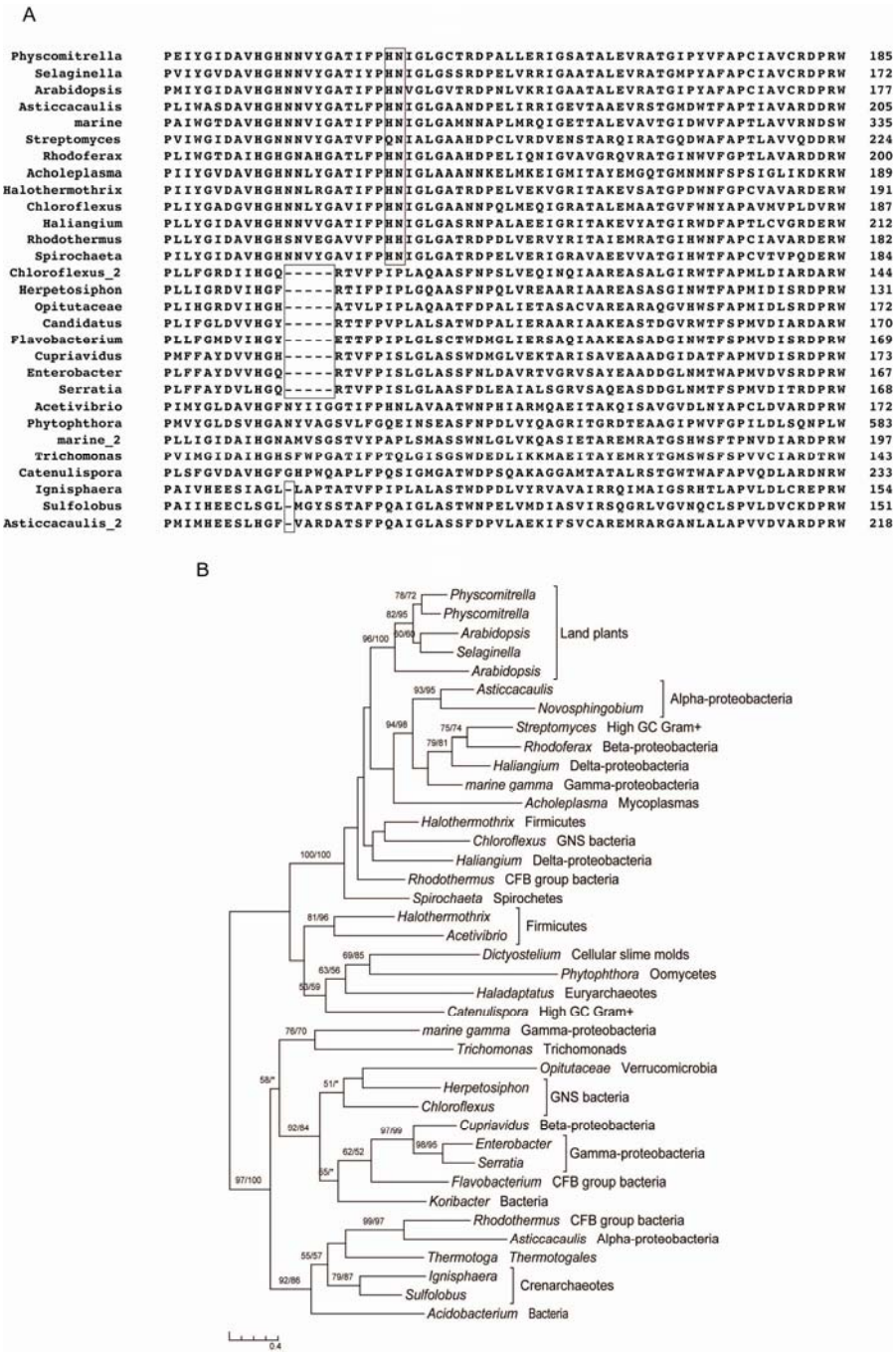
Supplementary Figure S12. Molecular phylogeny of DNA repair family protein. *Physcomitrella* sequence (Genbank GI numbers 168044851) has 50% identity with fungal homologs, and they form a highly supported clade. Several other eukaryotic sequences group with bacterial homologs. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating a likely mitochondrial origin.



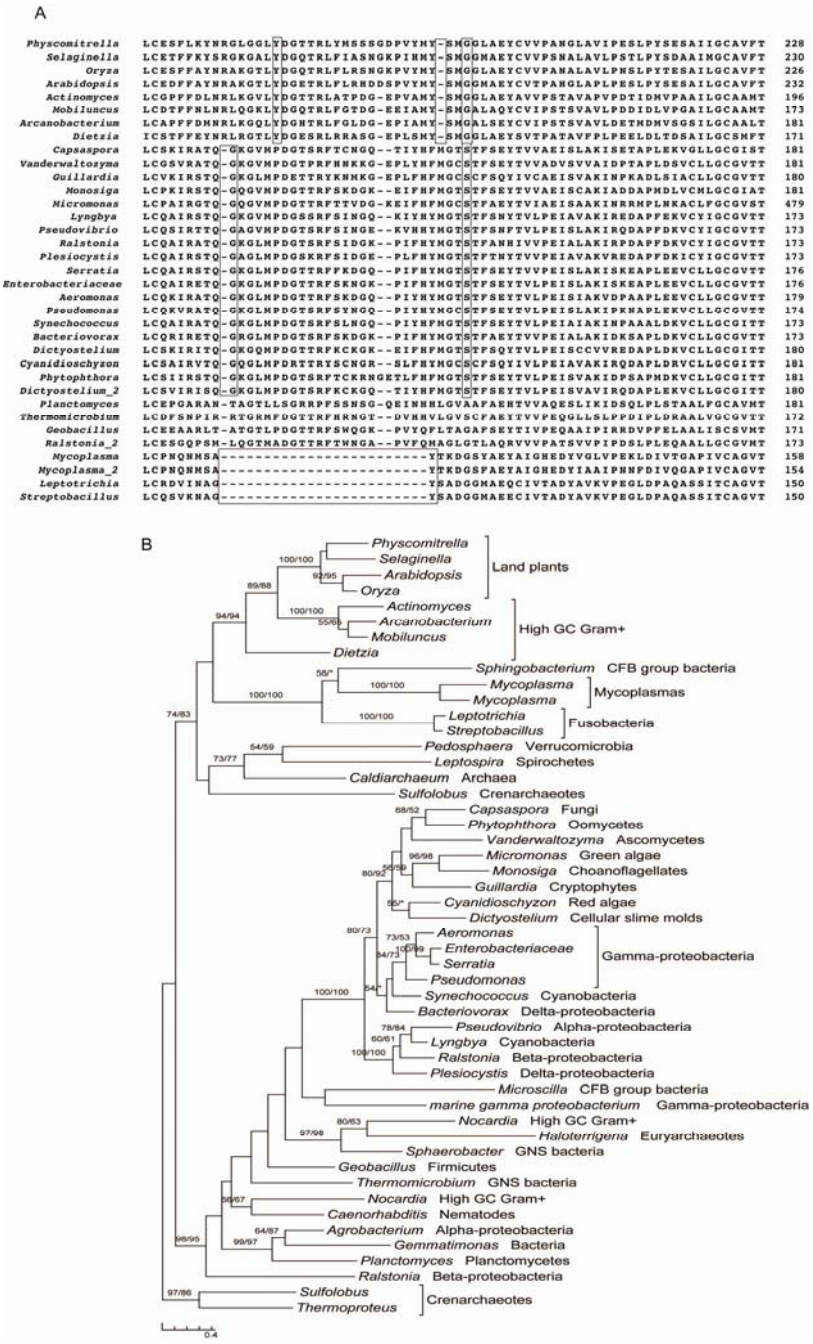
Supplementary Supplementary Figure S13. Multiple protein sequence alignment (A) and molecular phylogeny of 3, 4-dihydroxy-2-butanone 4-phosphate synthase (ribB) (B). BLAST result indicated that *Physcomitrella* sequence (Genbank GI number 168028296) has 31%-37% identity with homologs from archaea, whereas only 28% sequence identity with the other endogenous copy (GI number 168035901). This second copy is 332 aa longer and annotated as GTP cyclohydrolase-2 in *Arabidopsis*. It forms a clade with other eukaryotic homologs within bacterial sequences. This second copy was predicted by TargetP to be a chloroplast precursor, suggestive of possible plastid (or cyanobacterial) origin. The relationship of these two copies was also supported by several conserved amino acid residues and shared indels.



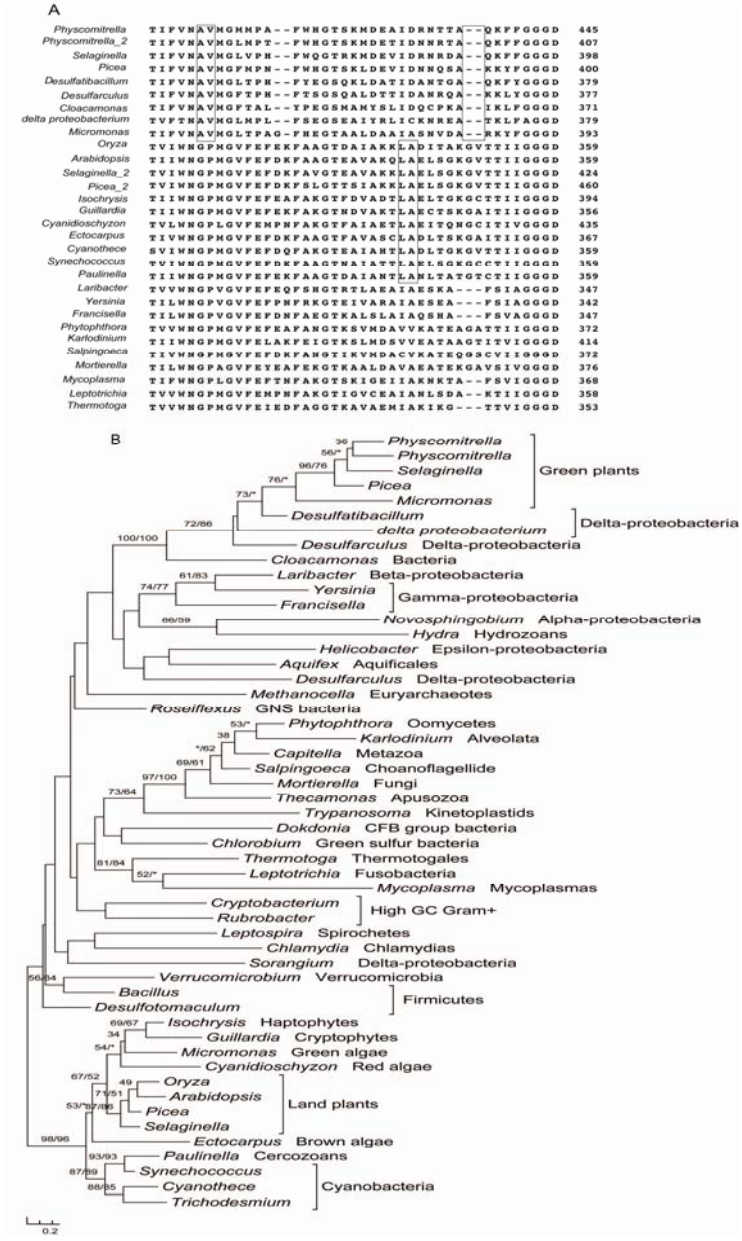
Supplementary Figure S14. Multiple protein sequence alignment (A) and molecular phylogeny of limit dextrinase (LDA) (B). *Physcomitrella* sequence (Genbank GI number 168038552) forms a highly supported clade with homologs of green plants and bacteria. No cyanobacterial homologs were found, indicating that this gene family in green plants most likely is not derived from plastids (cyanobacteria).



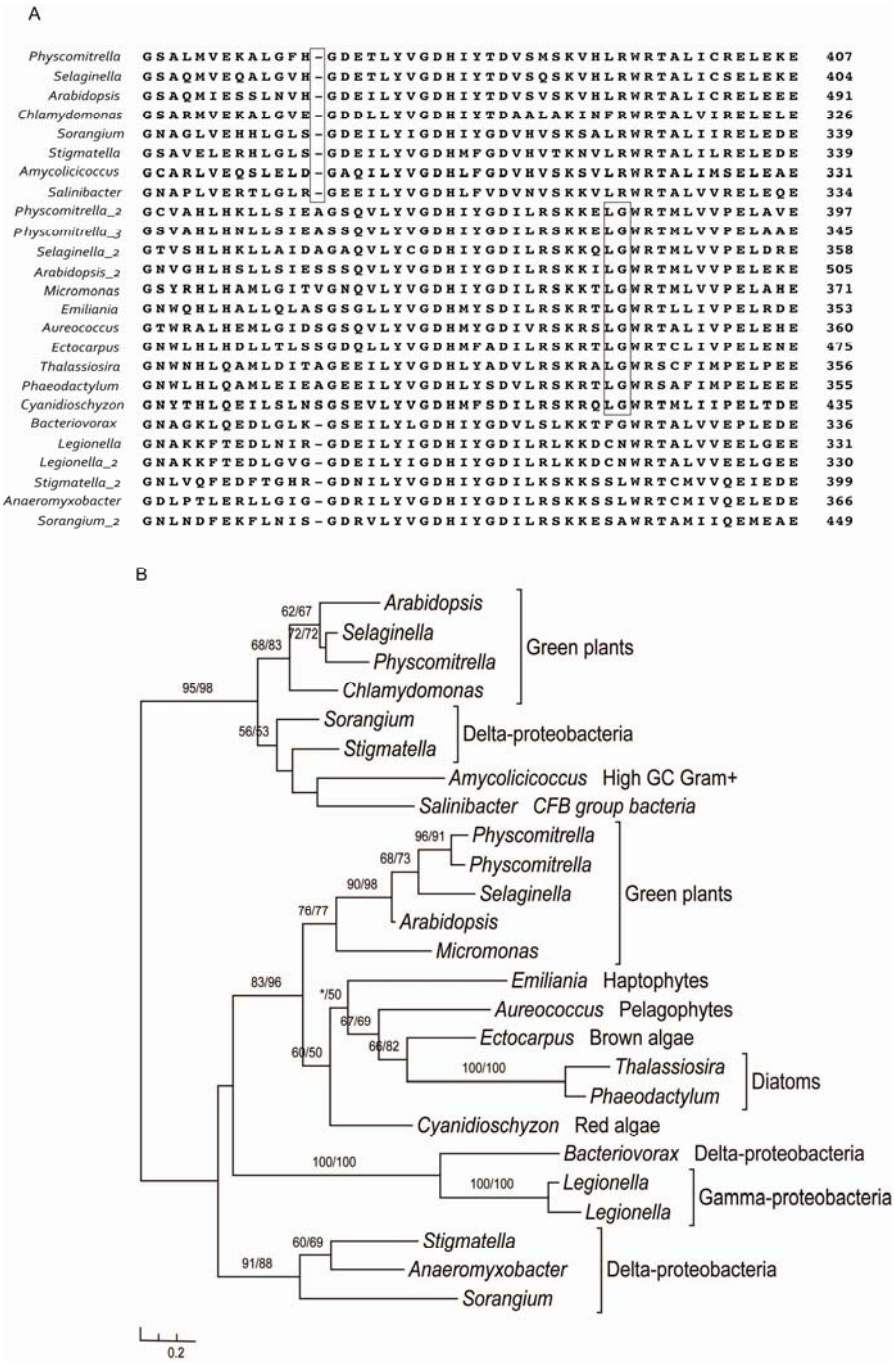
Supplementary Figure S15. Multiple protein sequence alignment (A) and molecular phylogeny of β-glucosidase (B). Identifiable homologs of *Physcomitrella* sequences are predominantly found in bacteria. *Physcomitrella* sequences (Genbank GI numbers 168069539, 168059435) form a monophyletic group with other land plant and bacterial homologs. This relationship is also supported by several conserved amino acid residues and shared indels. No cyanobacterial homologs were found, indicating that this gene family in land plants is unlikely of plastid (cyanobacterial) origin.



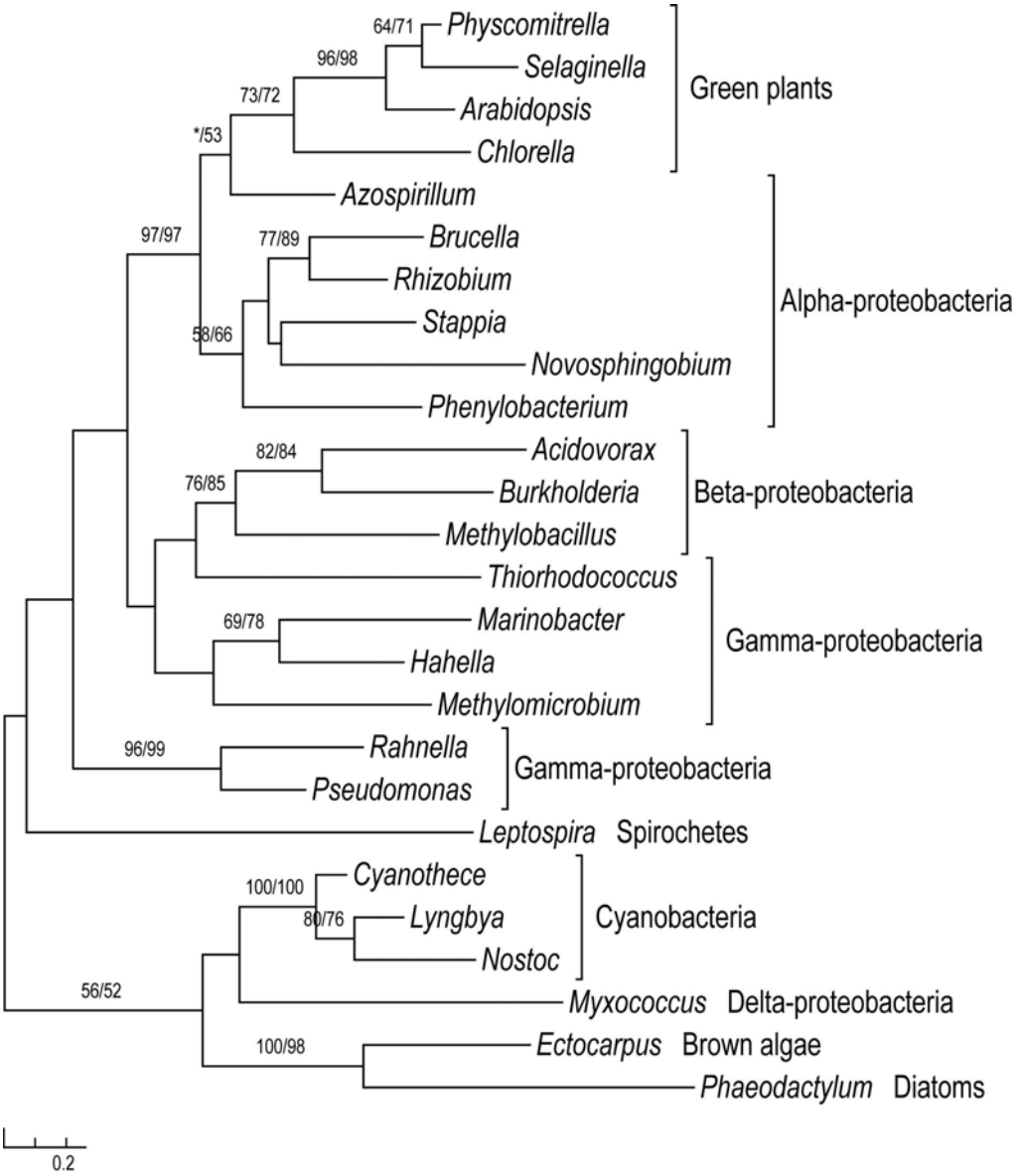
Supplementary Figure S16. Multiple protein sequence alignment (A) and molecular phylogeny of GroES-like zinc-binding alcohol dehydrogenase (B). Identifiable homologs of *Physcomitrella* sequences are predominantly found in bacteria. *Physcomitrella* sequence (Genbank GI number 168030245) groups with homologs from land plants and high GC gram-positive bacteria. Several other eukaryotic sequences form another monophyletic group with bacterial homologs, possibly derived from either ancient HGT or mitochondria. The highly supported clades on the phylogenetic tree are also supported by several conserved amino acid residues and shared indels.



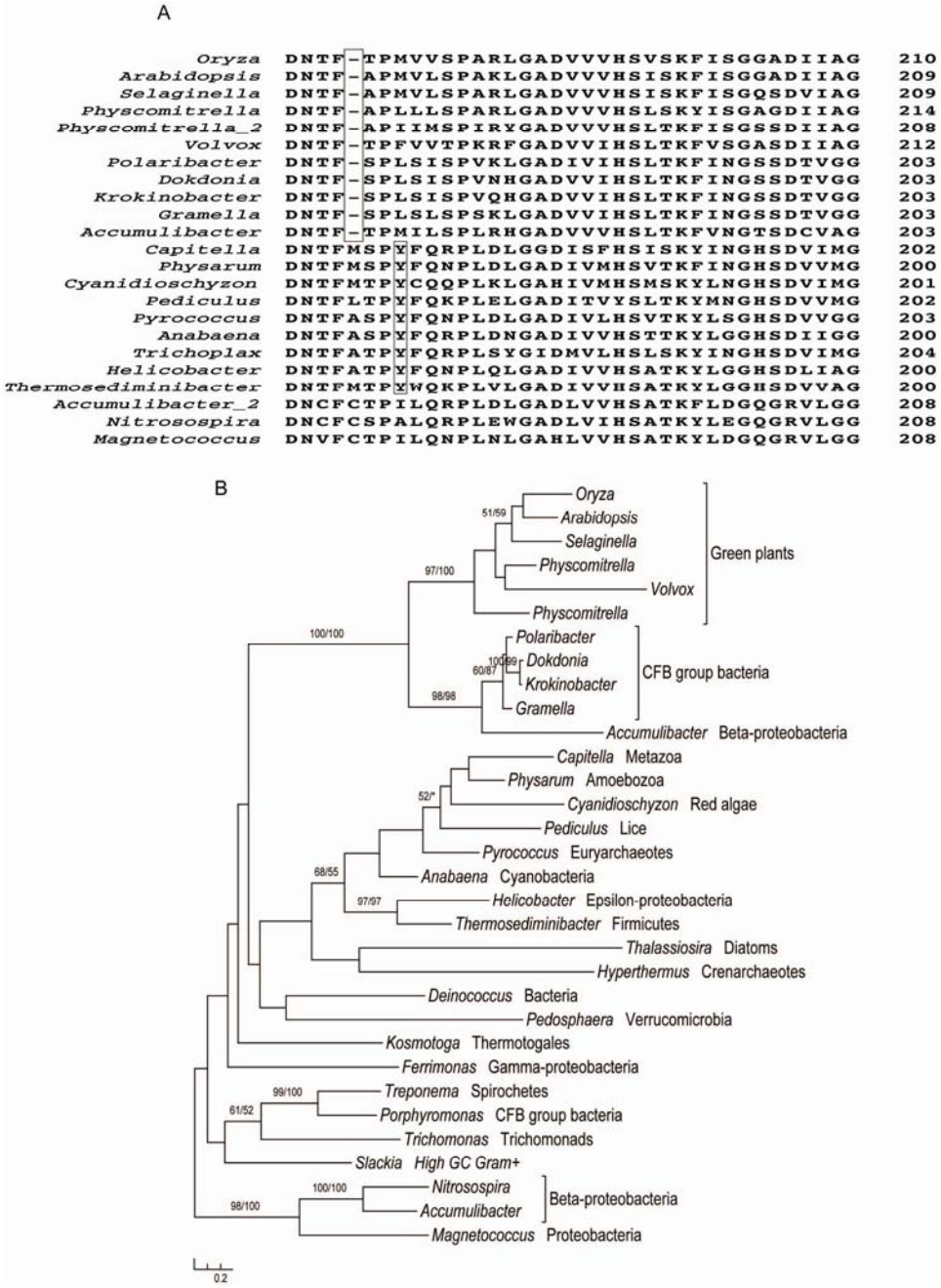
Supplementary Figure S17. Multiple protein sequence alignment (A) and molecular phylogeny of phosphoglycerate kinase (PGK) (B). *Physcomitrella* sequences (Genbank GI numbers 168058081, 168034630) form a clade with green plant and delta-proteobacterial homologs. Several other eukaryotic sequences form a moderately supported clade, which in turn groups with bacterial homologs. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating likely mitochondrial origin. Another gene copy of land plants forms a highly supported monophyletic group with cyanobacterial and other photosynthetic eukaryotic homologs. Some of these photosynthetic eukaryotic sequences are predicted by TargetP to be chloroplast precursors, indicating likely plastid (or cyanobacterial) origin. *Physcomitrella* sequence is absent from this clade, suggesting that either replacement of plastid homolog or gene acquisition following a loss of plastid copy in *Physcomitrella*.



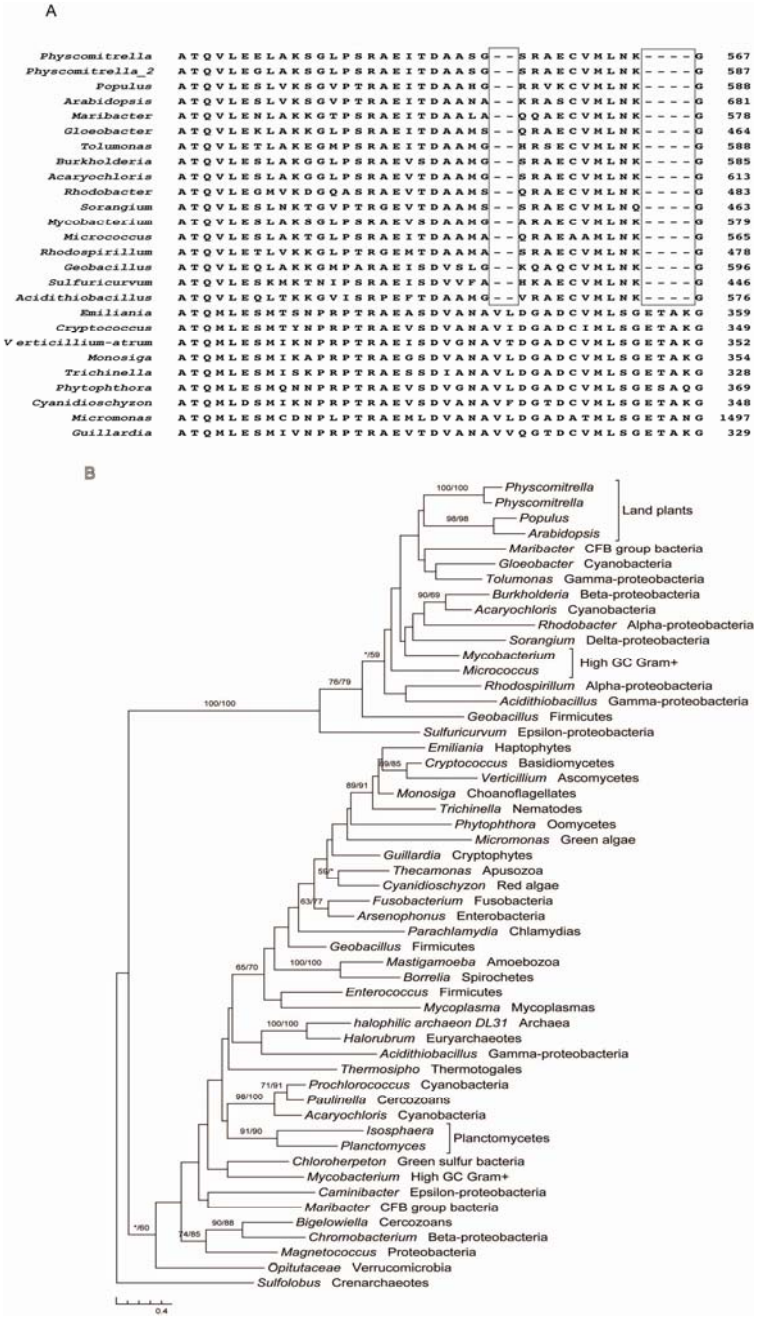
Supplementary Figure S18. Multiple protein sequence alignment (A) and molecular phylogeny of HAD-superfamily hydrolase (B). *Physcomitrella* sequence (Genbank GI number 168001220) forms a clade with green plant and bacterial homologs. Other two copies group with green plant and other photosynthetic eukaryotic homologs. In *Arabidopsis*, it is targeted to chloroplasts. Most of other eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating likely mitochondrial origin. The relationship of these gene copies is also supported by several conserved amino acid residues and shared indels.



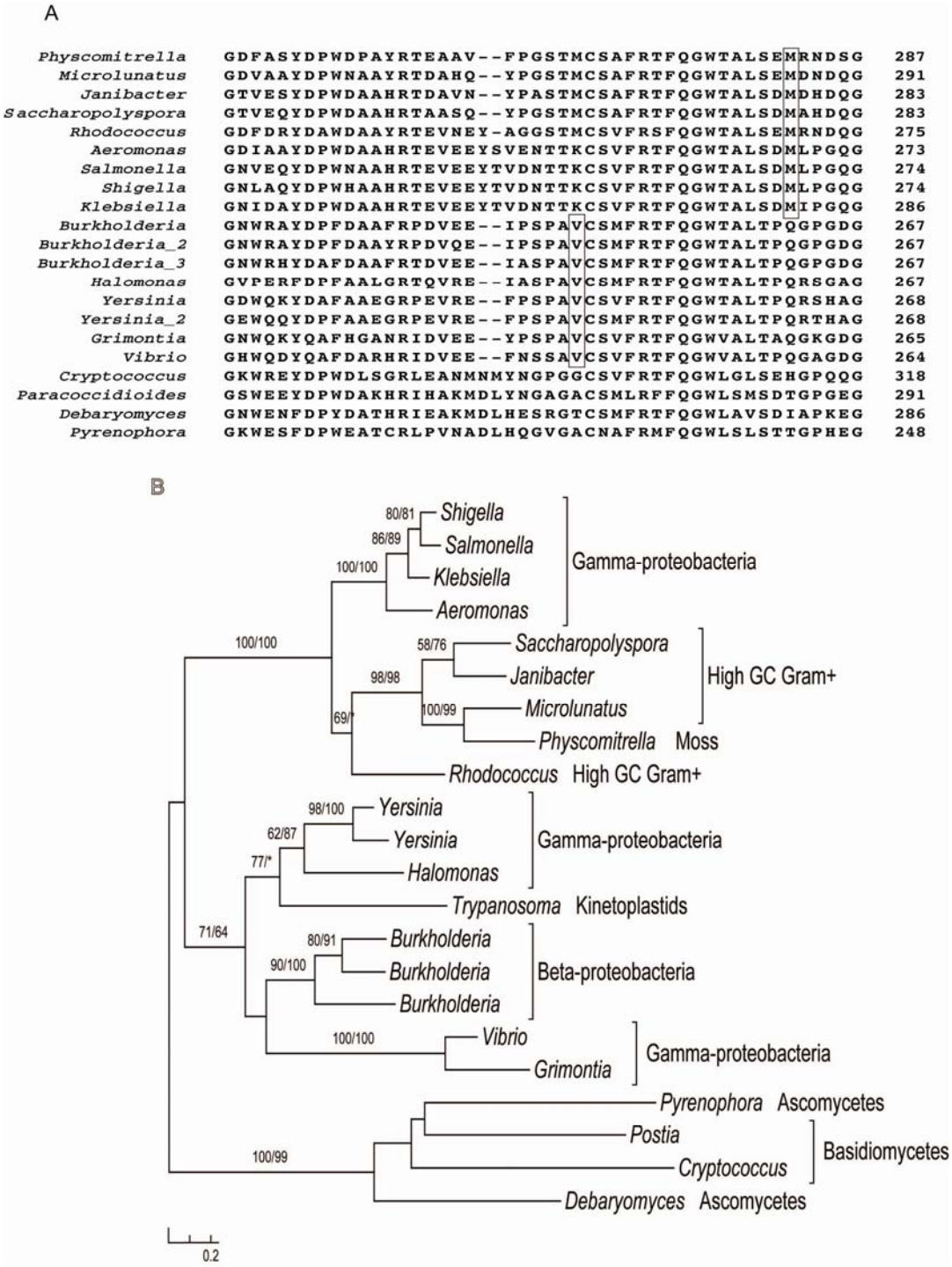
Supplementary Figure S19. Molecular phylogeny of Acyl-CoA N-acyltransferase. *Physcomitrella* sequence (Genbank GI number 168042611) forms a highly supported clade with homologs from green plants and alpha-proteobacteria. Two sequences from brown algae and diatoms group with cyanobacterial and delta-proteobacterial homologs with modest support. It is unclear whether these two sequences are of plastid origin.



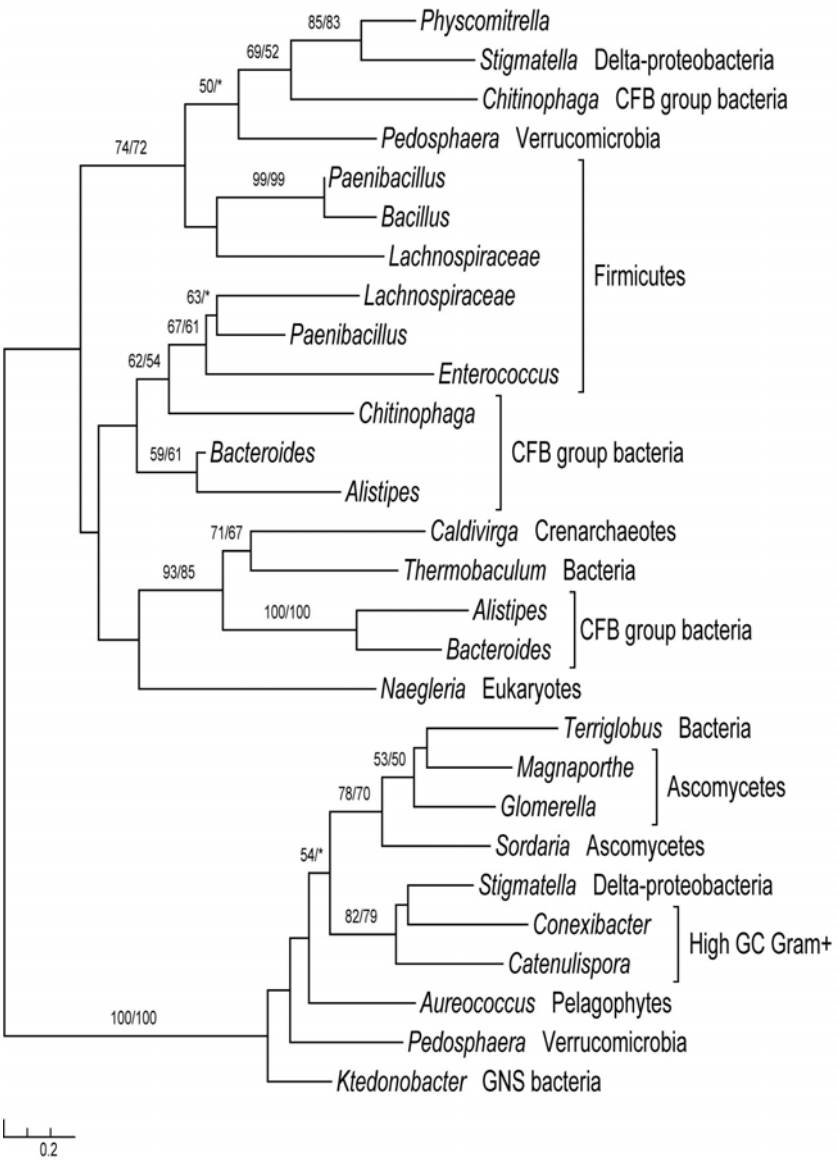
Supplementary Figure S20. Multiple protein sequence alignment (A) and molecular phylogeny of methionine gamma-lyase (MGL) (B). *Physcomitrella* sequences (Genbank GI numbers 168013924，168008405) form a clade with homologs from other green plants, CFB bacteria, and beta-proteobacterial *Accumulibacter*. Several other eukaryotic sequences group with bacterial homologs. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating likely mitochondrial origin. The close relationship between green plant sequences and bacterial homologs is also supported by some conserved amino acid residues and shared indels.



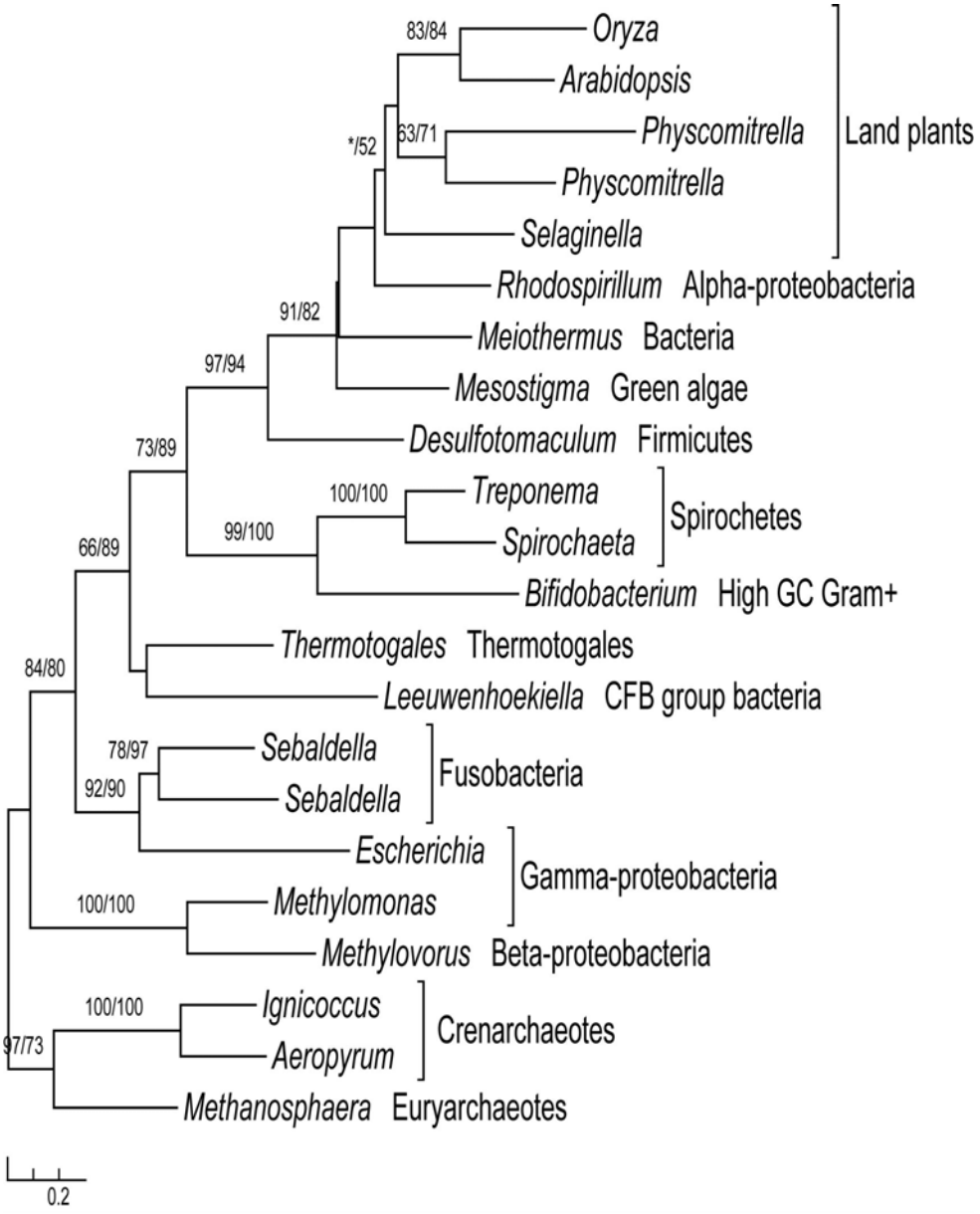
Supplementary Figure S21. Multiple protein sequence alignment (A) and molecular phylogeny of pyruvate kinase (B). *Physcomitrella* sequences (Genbank GI numbers 168053775, 168053903) form a clade with homologs from other land plants and miscellaneous bacteria. Green algal sequence forms another clade with homologs from other eukaryotes and bacteria. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating likely mitochondrial origin. This relationship is also supported by several conserved amino acid residues and shared indels. Land plants sequences form a group with bacterial sequences, including those from cyanobacteria. However, these land plant sequences are unlikely of plastid (cyanobacterial) origin because cyanobacterial orthologs are rarely found in this clade.



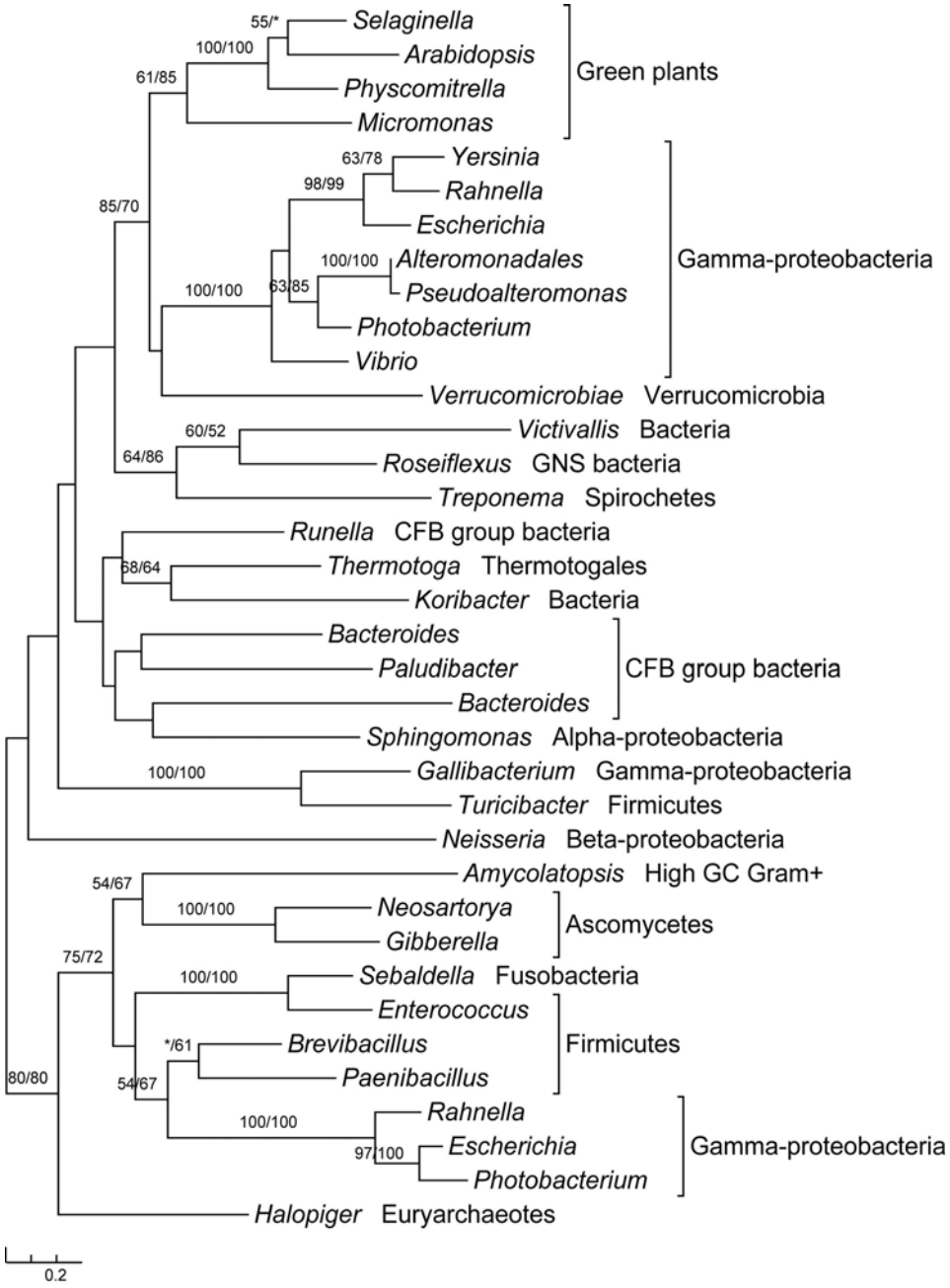
Supplementary Figure S22. Multiple protein sequence alignment (A) and molecular phylogeny of ybiU protein (B). *Physcomitrella* sequence (Genbank GI number 168021919) has 62-72% identity with homologs from high GC gram-positive bacteria. They form a highly supported clade in the phylogenetic tree. Their relationship is also supported by several conserved amino acid residues and shared indels.



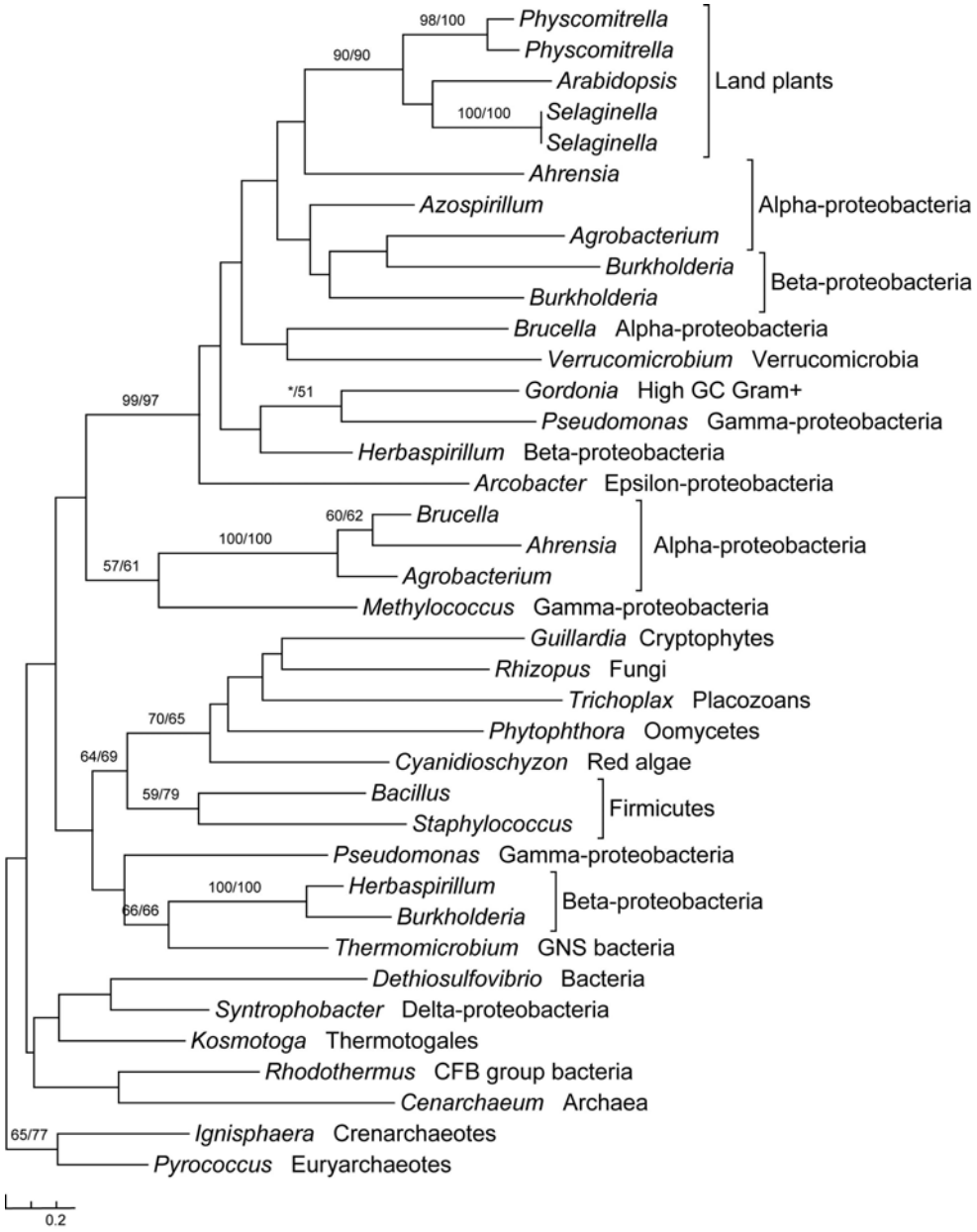
Supplementary Figure S23. Molecular phylogeny of glycoside hydrolase. BLAST result indicated that *Physcomitrella* sequence (Genbank GI number 168052263) has 50% identity with homologs from *Stigmatella aurantiaca*, and they form a monophyletic group in the phylogenetic tree. Several other eukaryotic sequences group with miscellaneous bacterial homologs. Some of these eukaryotic sequences were predicted by TargetP to be located in mitochondria, suggesting that they might be of mitochondrial (or alpha-proteobacterial) origin.



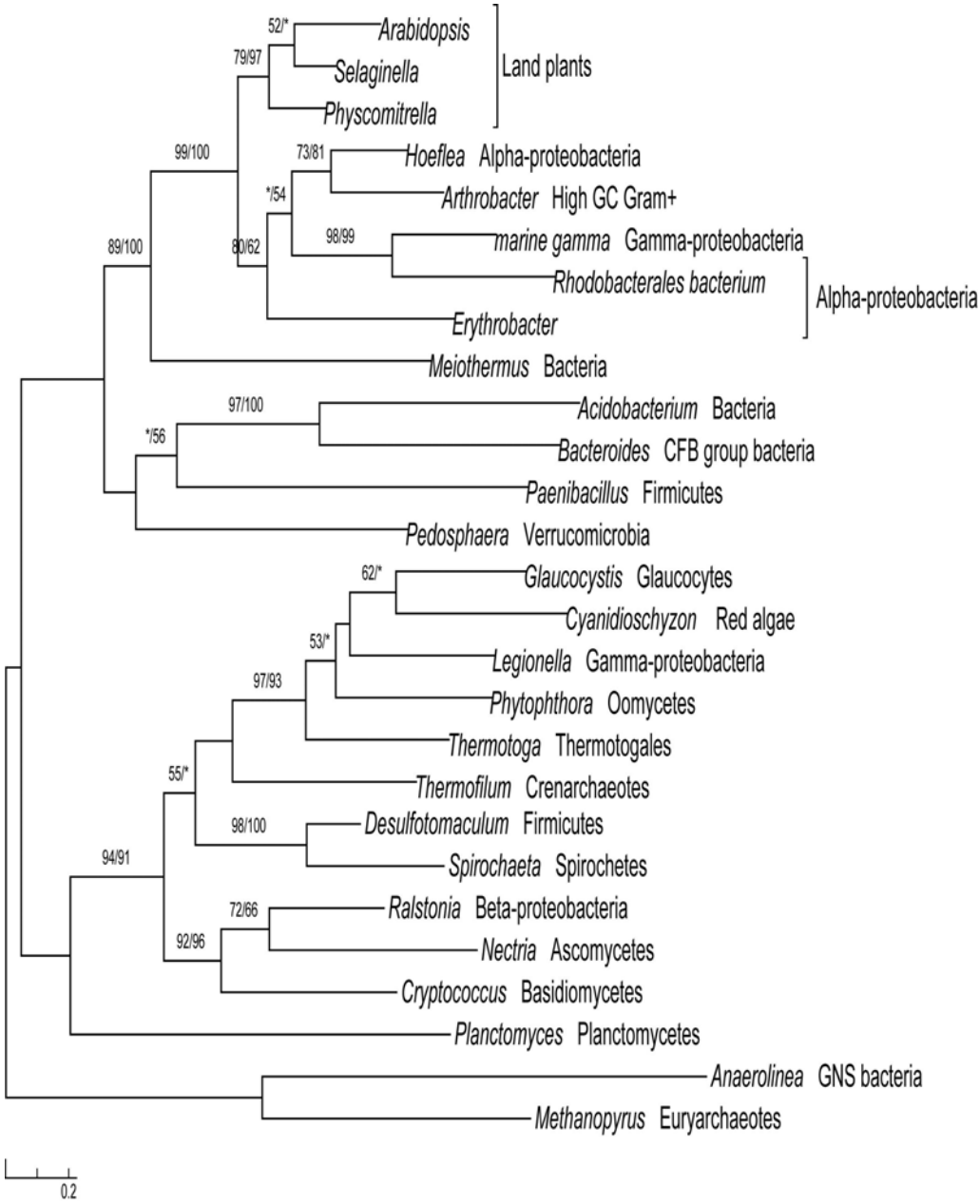
Supplementary Figure S24. Molecular phylogeny of sugar isomerase (SIS) family. Identifiable homologs of *Physcomitrella* sequences are only found in green plants and bacteria. No cyanobacterial homologs were found, indicating that this gene family in green plants is unlikely of plastid (cyanobacterial) origin.



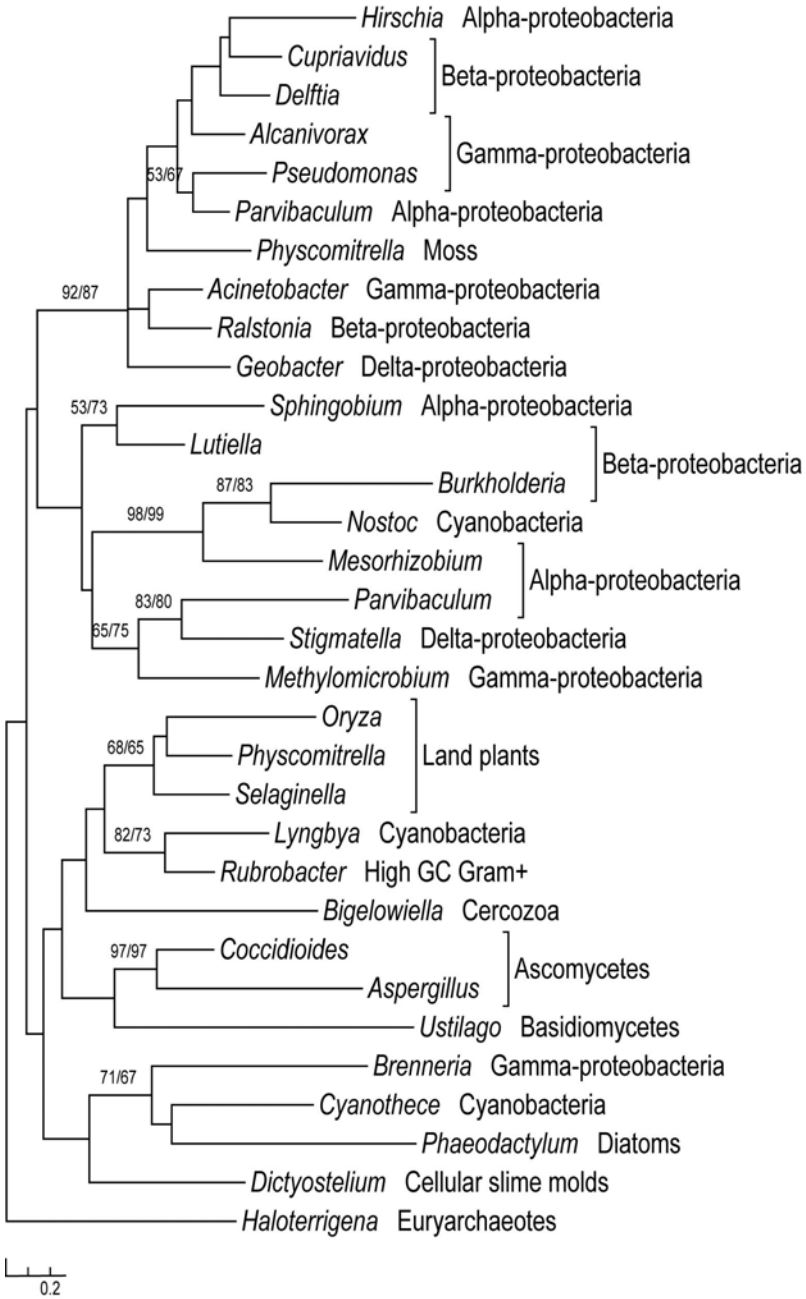
Supplementary Figure S25. Molecular phylogeny of glycoside hydrolase family 2. Identifiable homologs of moss sequences are predominantly found in bacteria. *Physcomitrella* sequence (Genbank GI number 168036598) forms a clade with homologs of green plants, gamma-proteobacteria and verrucomicrobia. No cyanobacterial homologs were identified, suggesting that green plant sequences are unlikely of plastid (cyanobacterial) origin. Two fungal sequences form another moderately supported clade with homologs from miscellaneous bacteria.



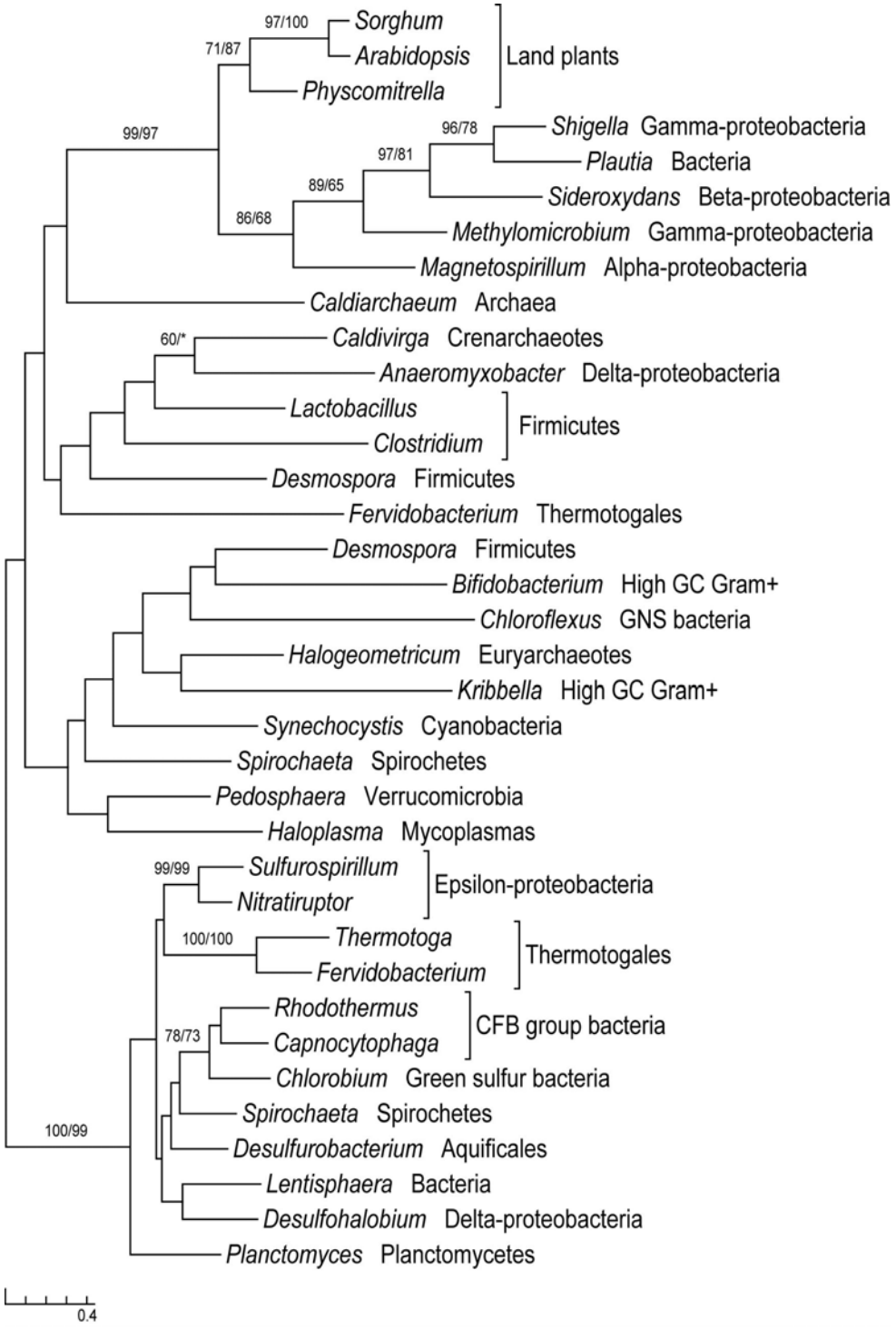
Supplementary Figure S26. Molecular phylogeny of hydroxypyruvate reductase 2 (HPR2). *Physcomitrella* sequences (Genbank GI numbers 167997717，168037243) form a strongly supported monophyletic group with land plant and bacterial homologs. This relationship is supported by several conserved amino acid residues and shared indels. Several other eukaryotic sequences form a moderately supported clade with bacterial homologs. Some of these eukaryotic sequences were predicted by TargetP to be mitochondrial precursors, suggesting that they are likely of mitochondrial (alpha-proteobacterial) origin.



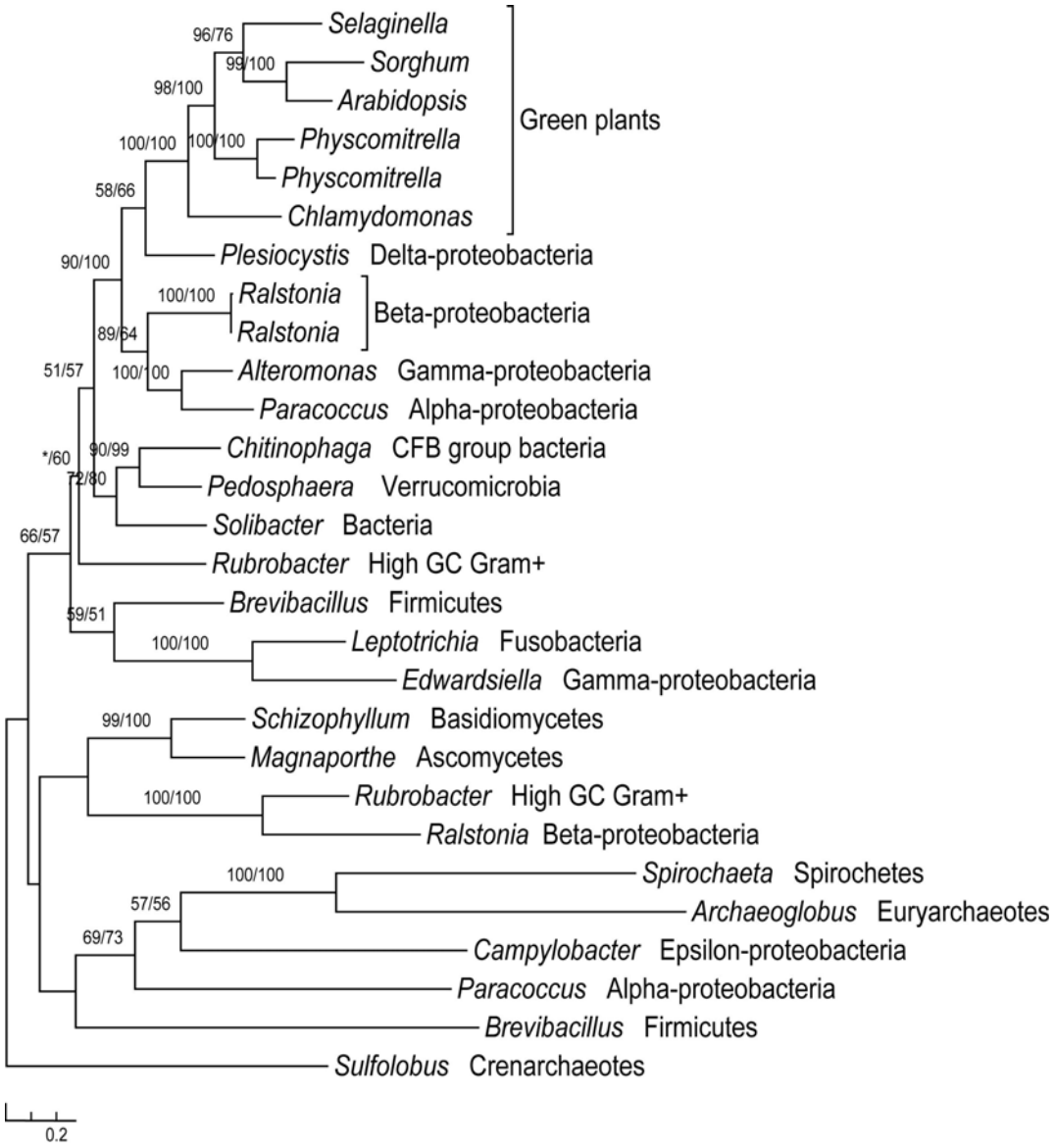
Supplementary Figure S27. Molecular phylogeny of inositol 2-dehydrogenase like protein. *Physcomitrella* sequence (Genbank GI number 168003329) forms a clade with homologs from land plants and bacteria (mostly alpha-proteobacteria). Their relationship is also supported by several conserved amino acid residues and shared indels. Several other eukaryotic sequences group with bacterial and archaeal homologs. Some of these eukaryotic sequences are predicted by TargetP as mitochondrial precursors, indicating likely mitochondrial origin.



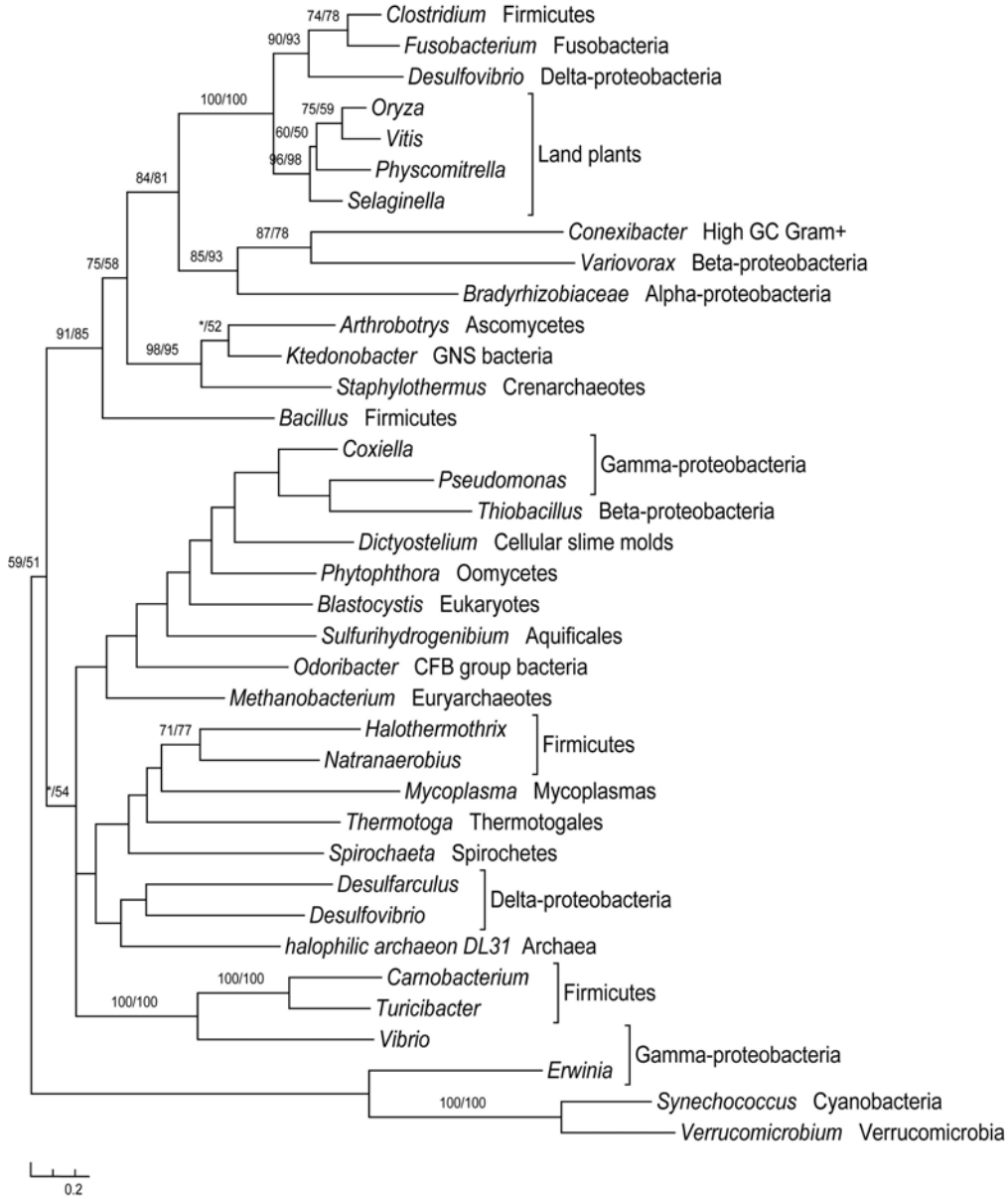
Supplementary Figure S28. Molecular phylogeny of short-chain dehydrogenase/reductase SDR. Identifiable homologs of *Physcomitrella* sequences are predominantly found in bacteria. *Physcomitrella* sequence (Genbank GI number 168031790) groups within a large bacterial clade. Another *Physcomitrella* copy (GI number 168036475) groups with other eukaryotic and bacterial homologs. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, suggesting that they are likely of mitochondrial (alpha-proteobacterial) origin.



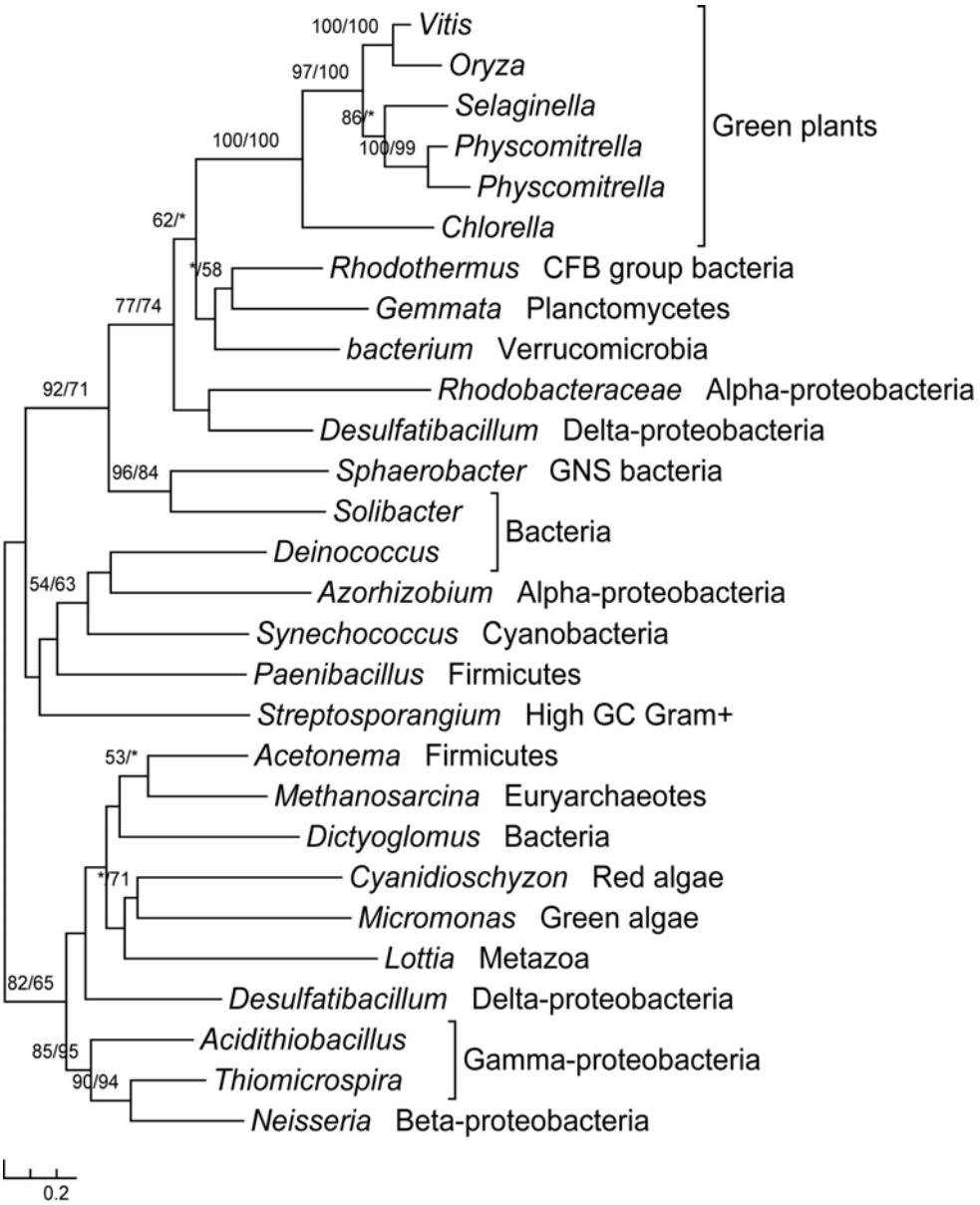
Supplementary Figure S29. Molecular phylogeny of ATP-binding cassette I1 (ABCI1) transporter. Identifiable homologs were only found in land plants and bacteria. *Physcomitrella* sequence (Genbank GI number 168004297) forms a highly supported monophyletic group with land plant and proteobacterial homologs, and their relationship is supported by several conserved amino acid residues and shared indels.



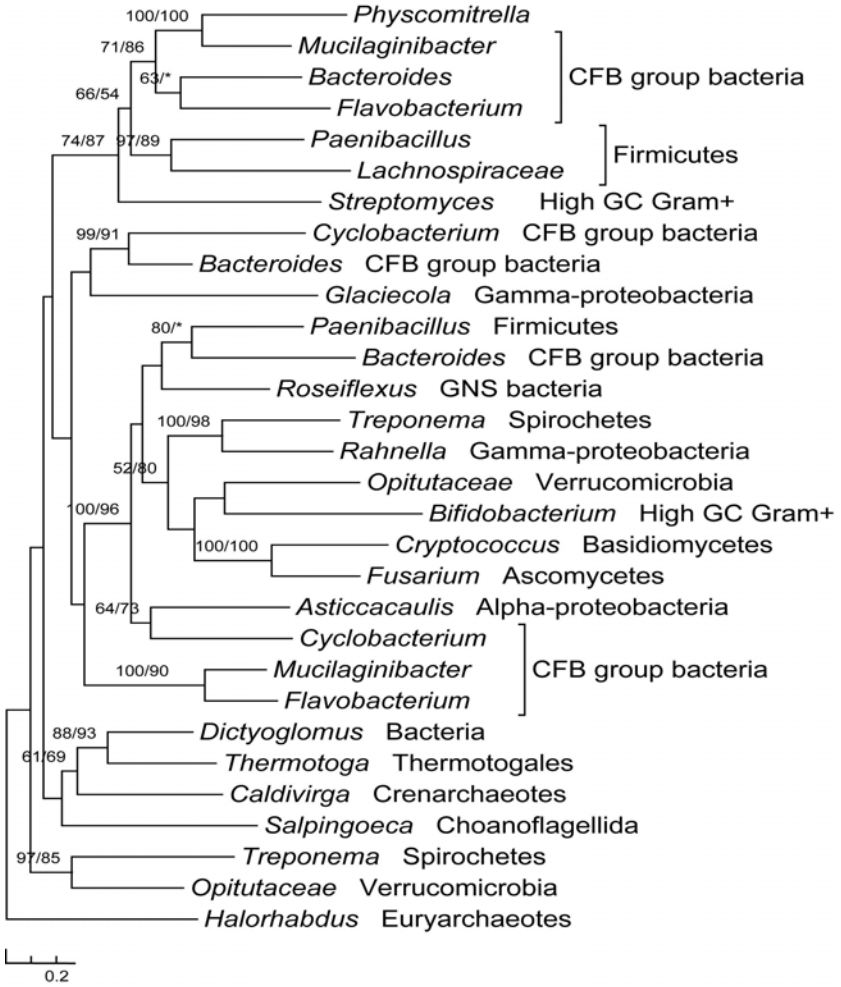
Supplementary Figure S30. Molecular phylogeny of uracil permease. *Physcomitrella* sequences (Genbank GI numbers 168012184, 168043133) form a highly supported clade with homologs from green plants and proteobacteria. This relationship is supported by several conserved amino acid residues and shared indels.



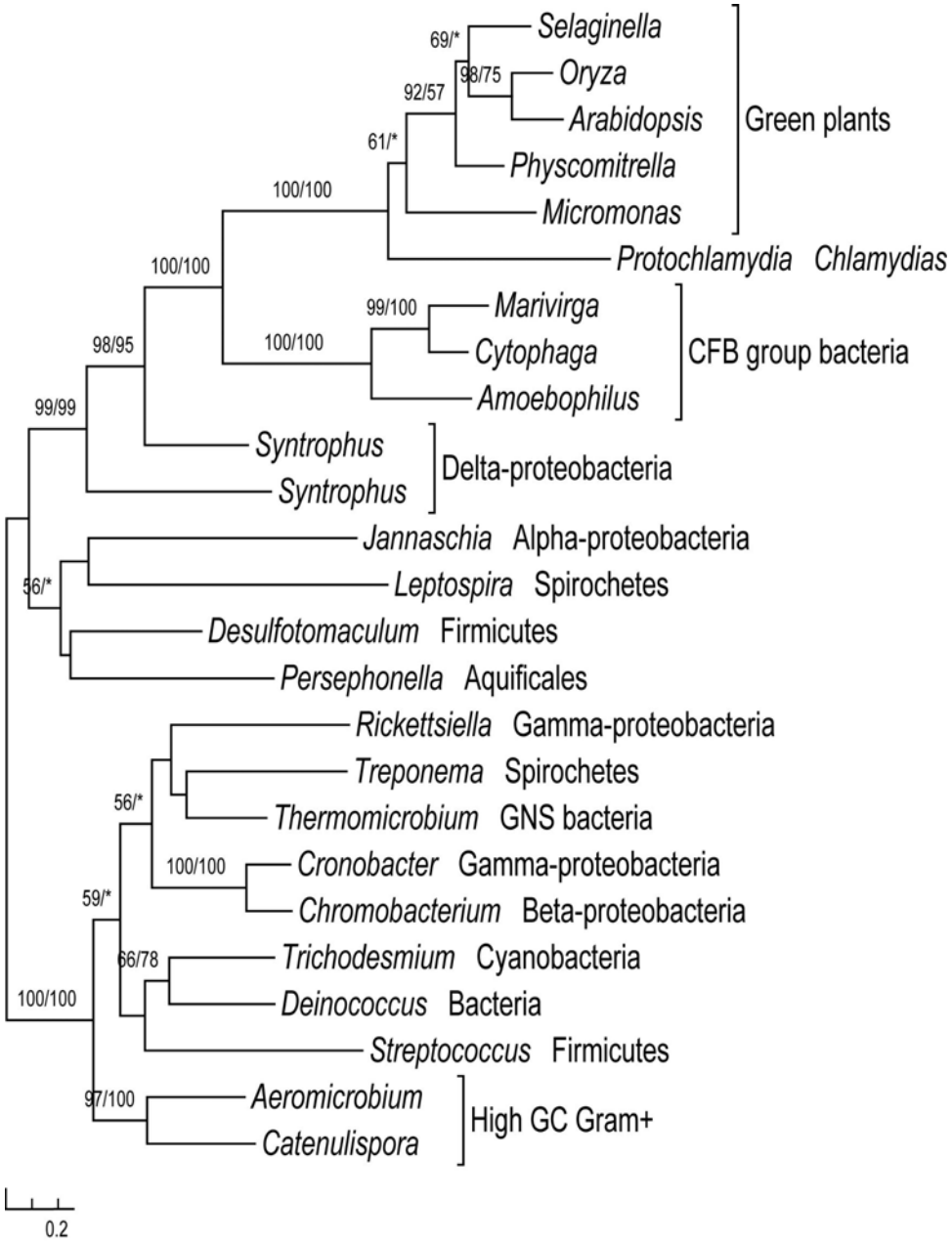
Supplementary Figure S31. Molecular phylogeny of amidohydrolase family. *Physcomitrella* sequence (Genbank GI number 168021897) forms a highly supported clade with homologs from other land plants and miscellaneous bacteria. This relationship is supported by multiple conserved amino acid residues and shared indels. Several other eukaryotic sequences group within bacterial homologs. Some of these eukaryotic sequences are predicted by TargetP to be mitochondrial precursors, indicating a likely mitochondrial or alpha-proteobacterial origin.



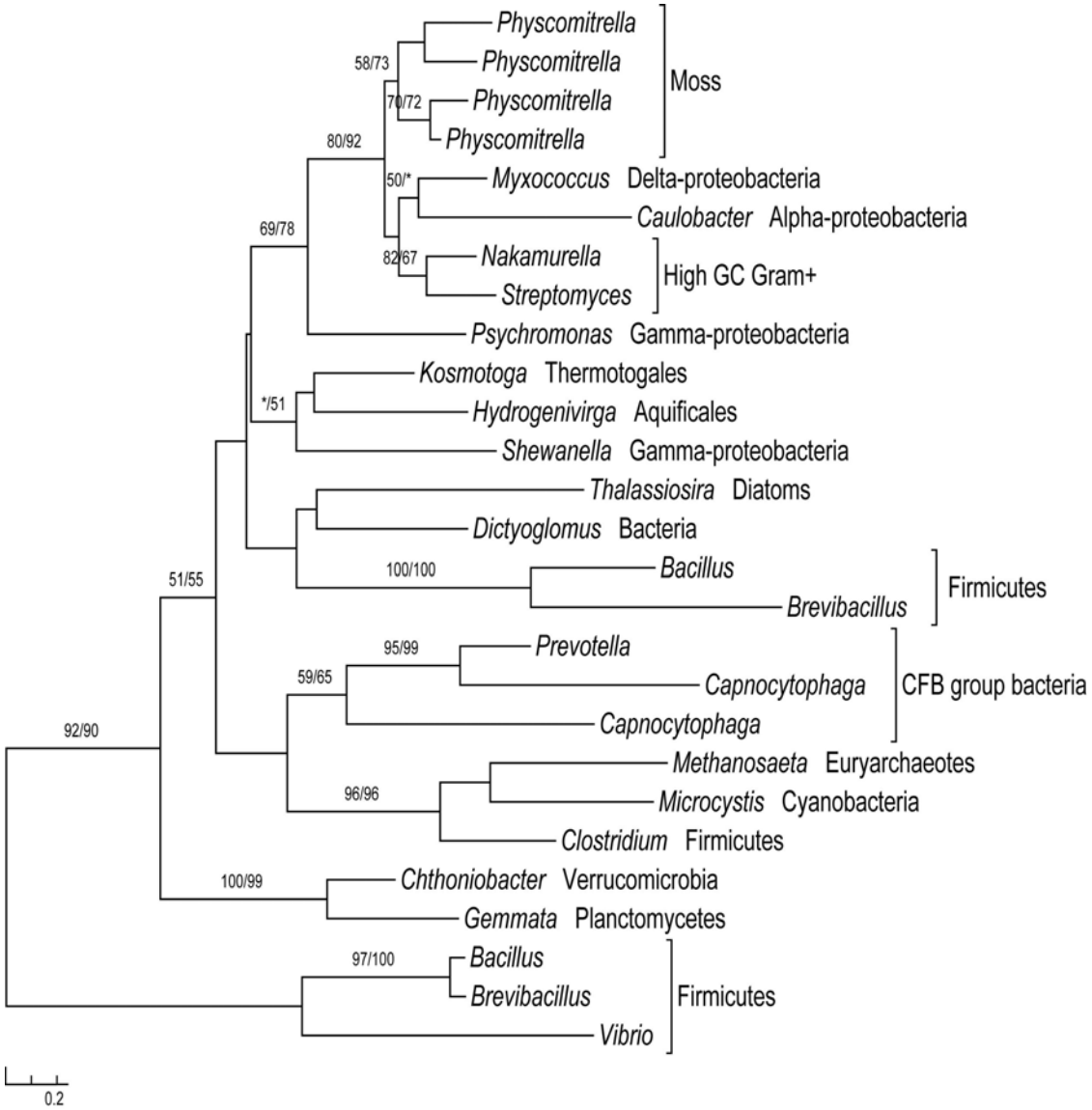
Supplementary Figure S32. Molecular phylogeny of amidase family. *Physcomitrella* sequences (Genbank GI numbers 168042262，168003211) form a highly supported clade with homologs from green plants and bacteria. Their relationship is also supported by multiple conserved amino acid residues and shared indels. Several other eukaryotic sequences form a clade with bacterial homologs. Some of them are predicted by TargetP to be mitochondrial precursors, indicating a likely mitochondrial origin.



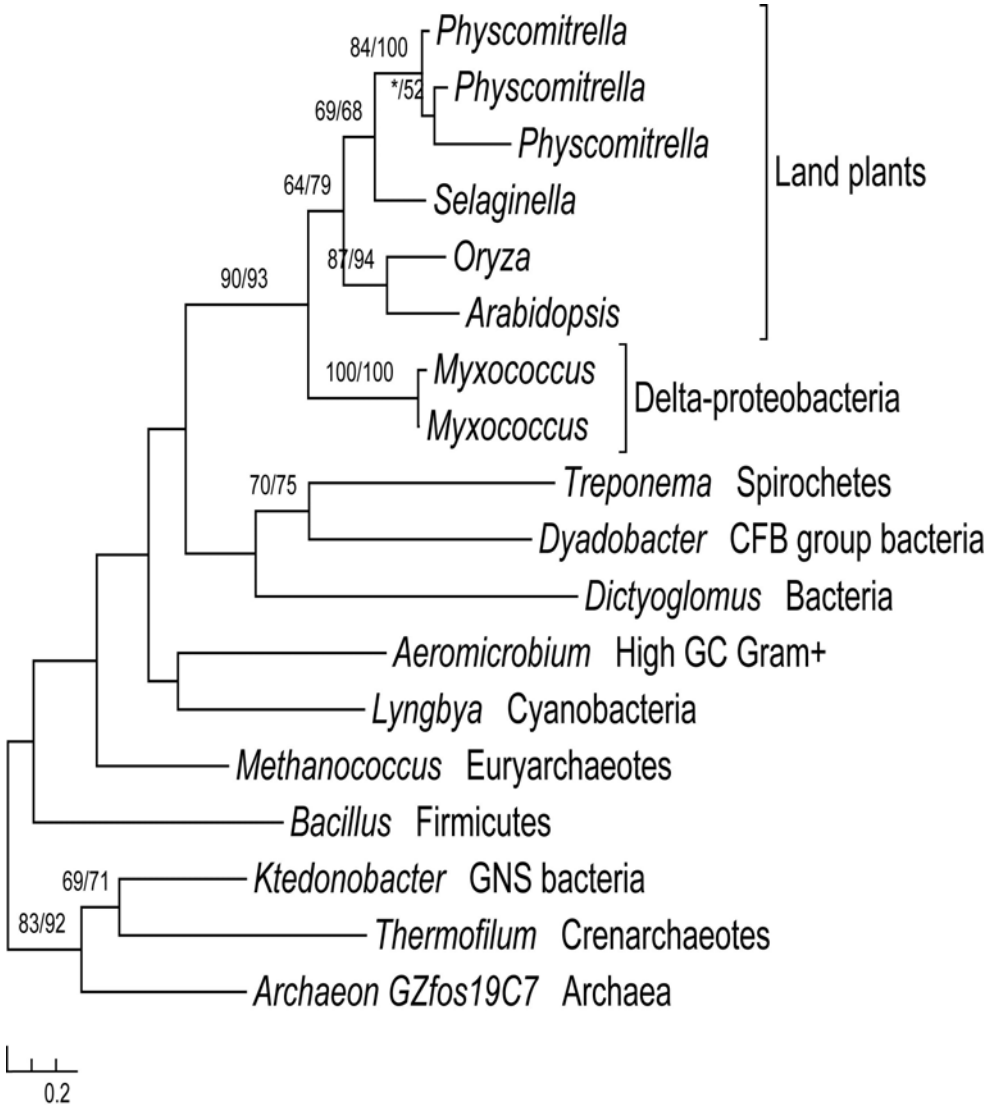
Supplementary Figure S33. Molecular phylogeny of alpha-L-rhamnosidase. Identifiable homologs of the *Physcomitrella* sequence are predominantly found in bacteria. *Physcomitrella* sequence (Genbank GI number 168031461) has 49% identity with homolog of *Mucilaginibacter paludis*, these two sequences form a highly supported monophyletic group.



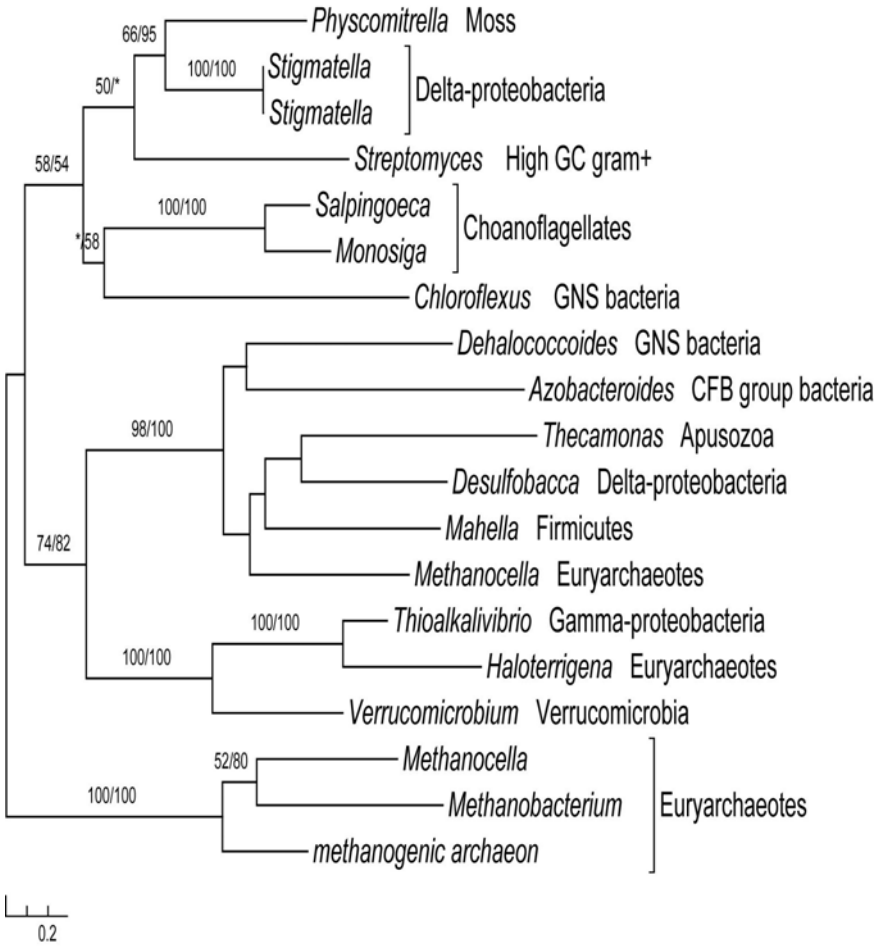
Supplementary Figure S34. Molecular phylogeny of D-alanine-D-alanine ligase. *Physcomitrella* sequence (Genbank GI number 168012025) forms a highly supported clade with homologs from other green plants, *Protochlamydia* and CFB bacteria. Identifiable homologs of the *Physcomitrella* sequence were only found in green plants and bacteria.



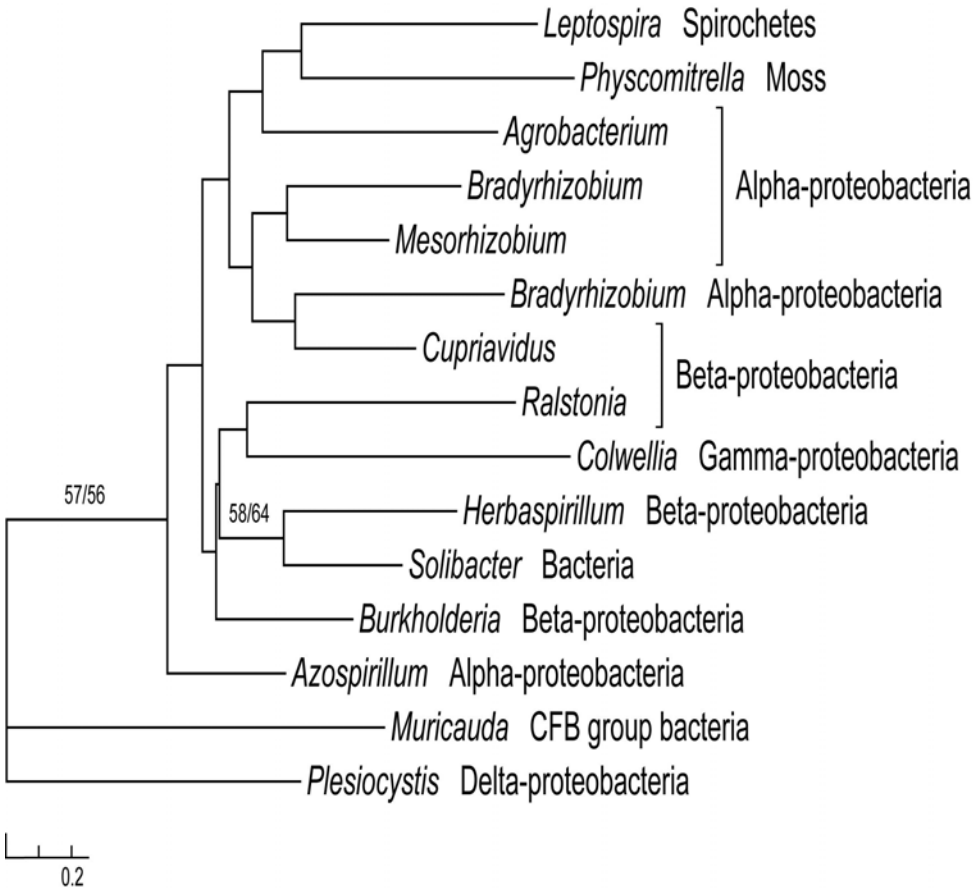
Supplementary Figure S35. Molecular phylogeny of M6 family peptidase. Identifiable homologs of *Physcomitrella* sequences are predominantly found in bacteria. *Physcomitrella* sequences (Genbank GI numbers 168013514，168019010，168032091, 168048759) form a clade with homologs from proteobacteria and high GC gram positive bacteria. Their relationship is supported by multiple conserved amino acid residues and shared indels. Other than two diatom sequences, no other eukaryotic homologs were found.



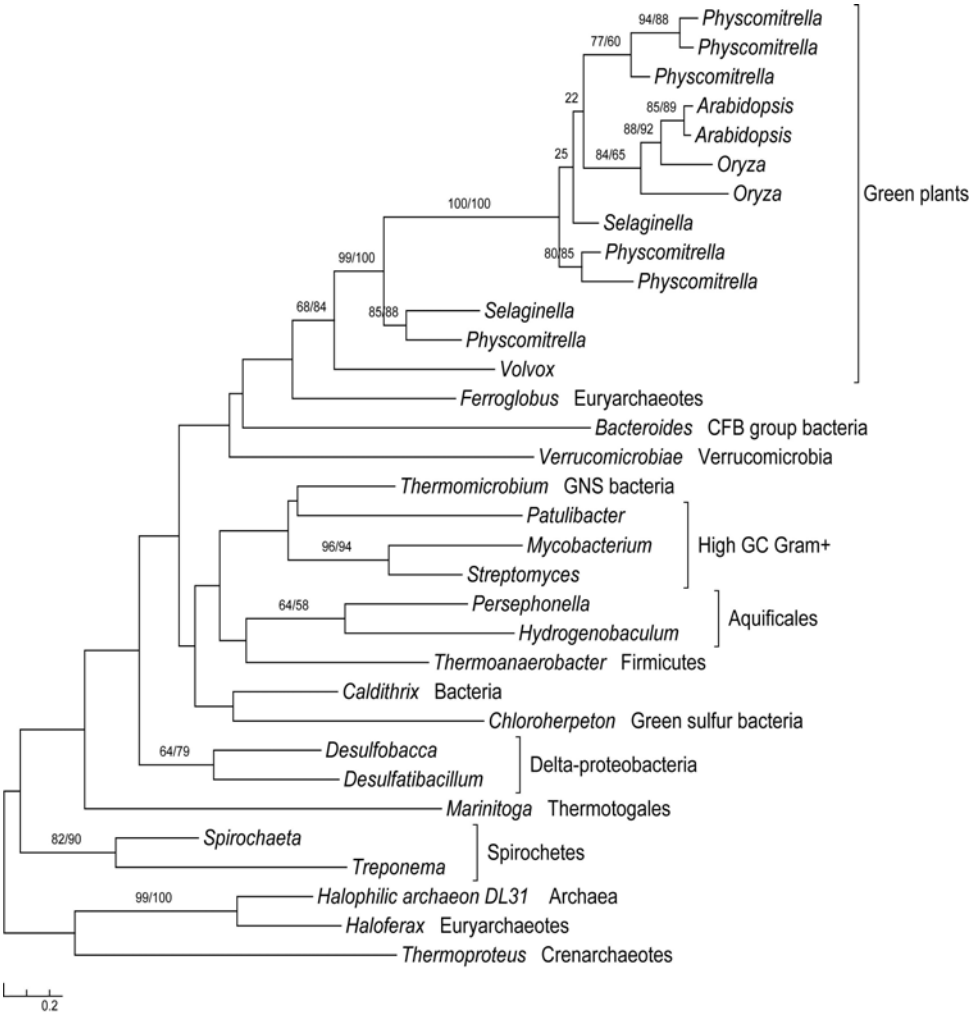
Supplementary Figure S36. **Molecular phylogeny of PfkB family kinase.** *Physcomitrella* sequences (Genbank GI numbers 167998254, 168016595, 168021833) form a highly supported clade with homologs from other land plants and delta-proteobacteria. No other eukaryotic homologs were identified.



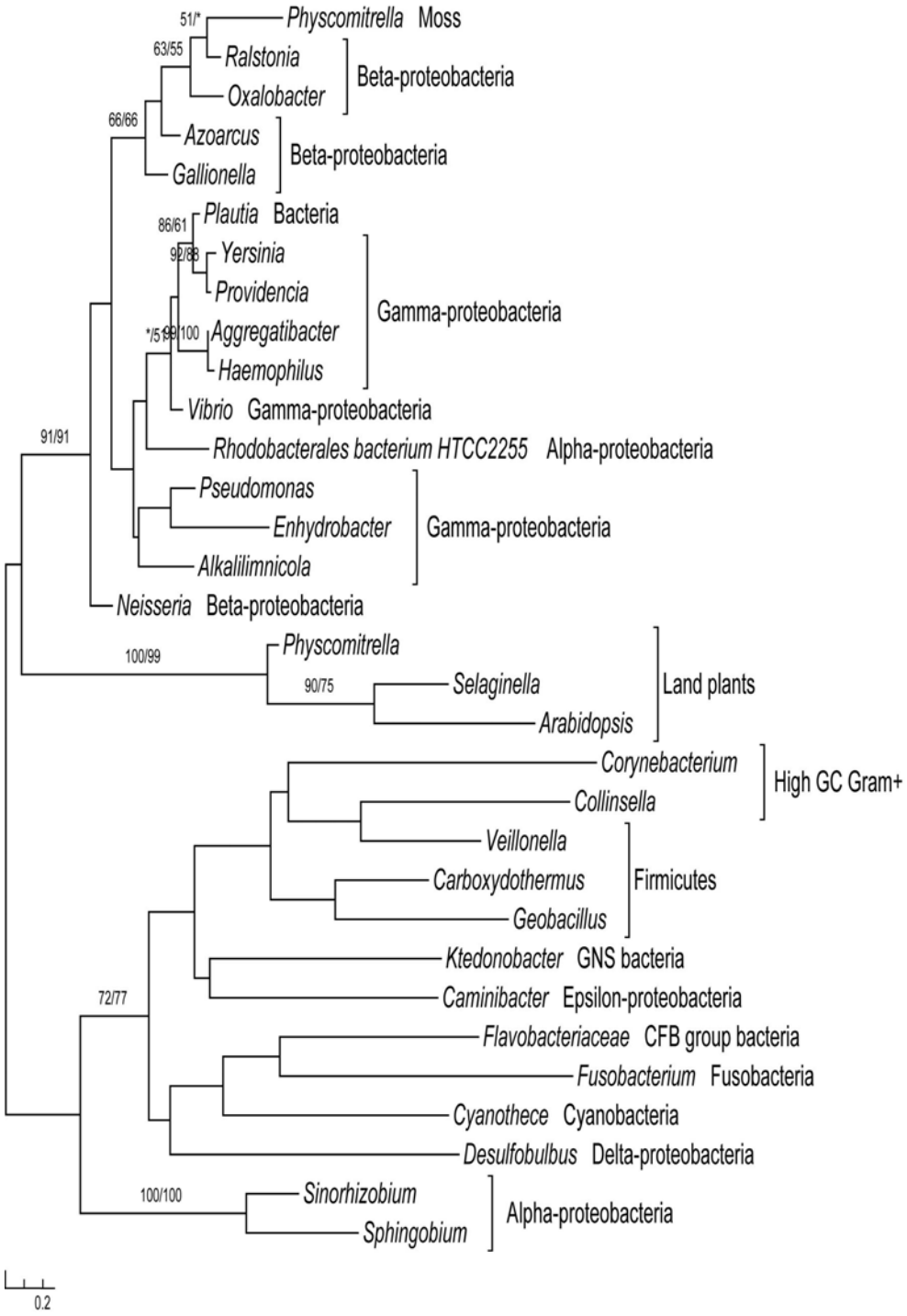
Supplementary Figure S37. Molecular phylogeny of 1, 4-dihydroxy-2-naphtoate octaprenyl transferase. *Physcomitrella* sequence (Genbank GI number 168009868) has 42-46% identity with homologs from the delta-proteobacterial *Stigmatella*. These sequences form a clade.



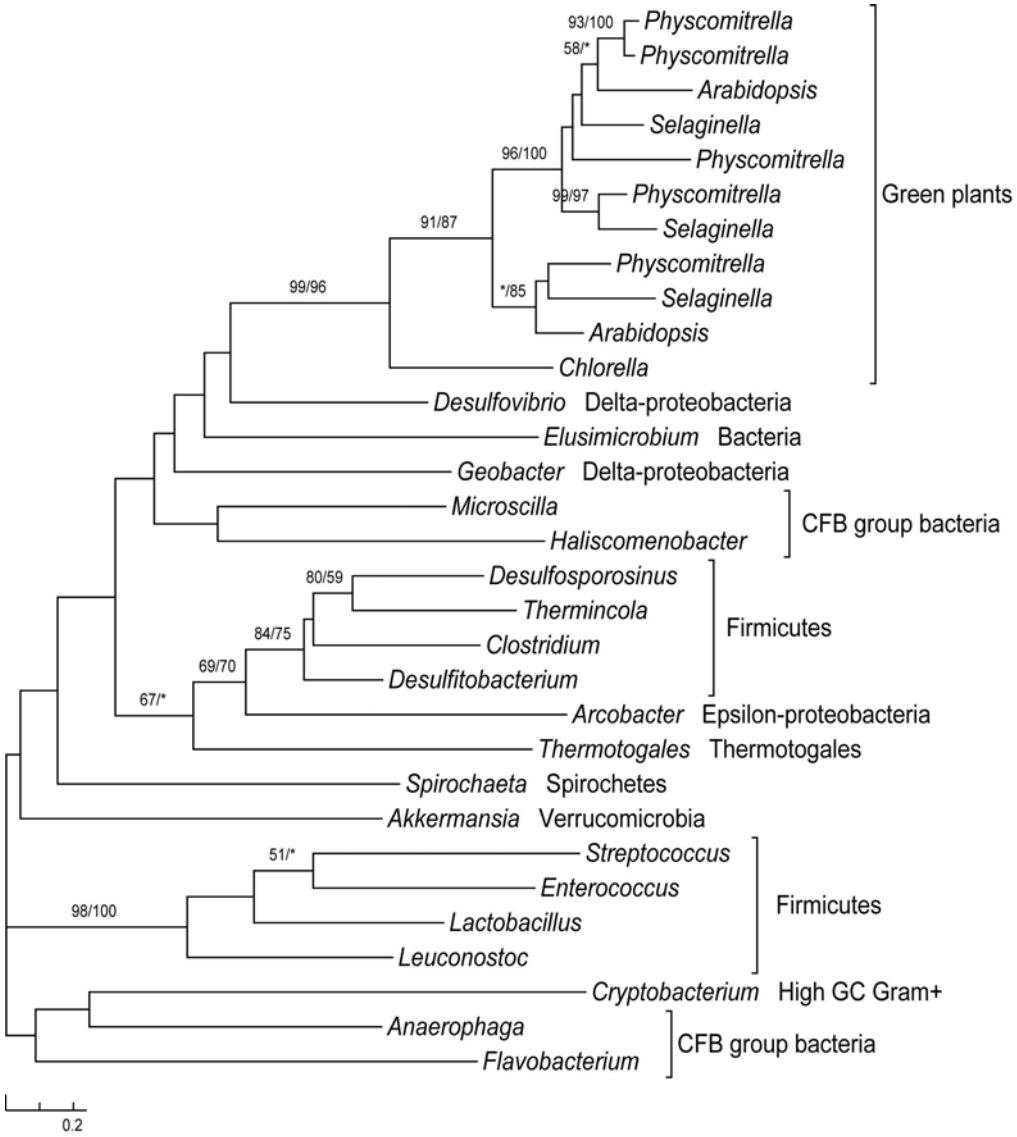
Supplementary Figure S38. Molecular phylogeny of dienelactone hydrolase family. Identifiable homologs were only found in *Physcomitrella* and bacteria. Because only 103 reliably aligned amino acid residues were used to construct the phylogeny, the bootstrap values for most clades identified in the phylogenetic tree are low. The *Physcomitrella* sequence is located on a long scaffold and therefore should not result from sequence contamination.



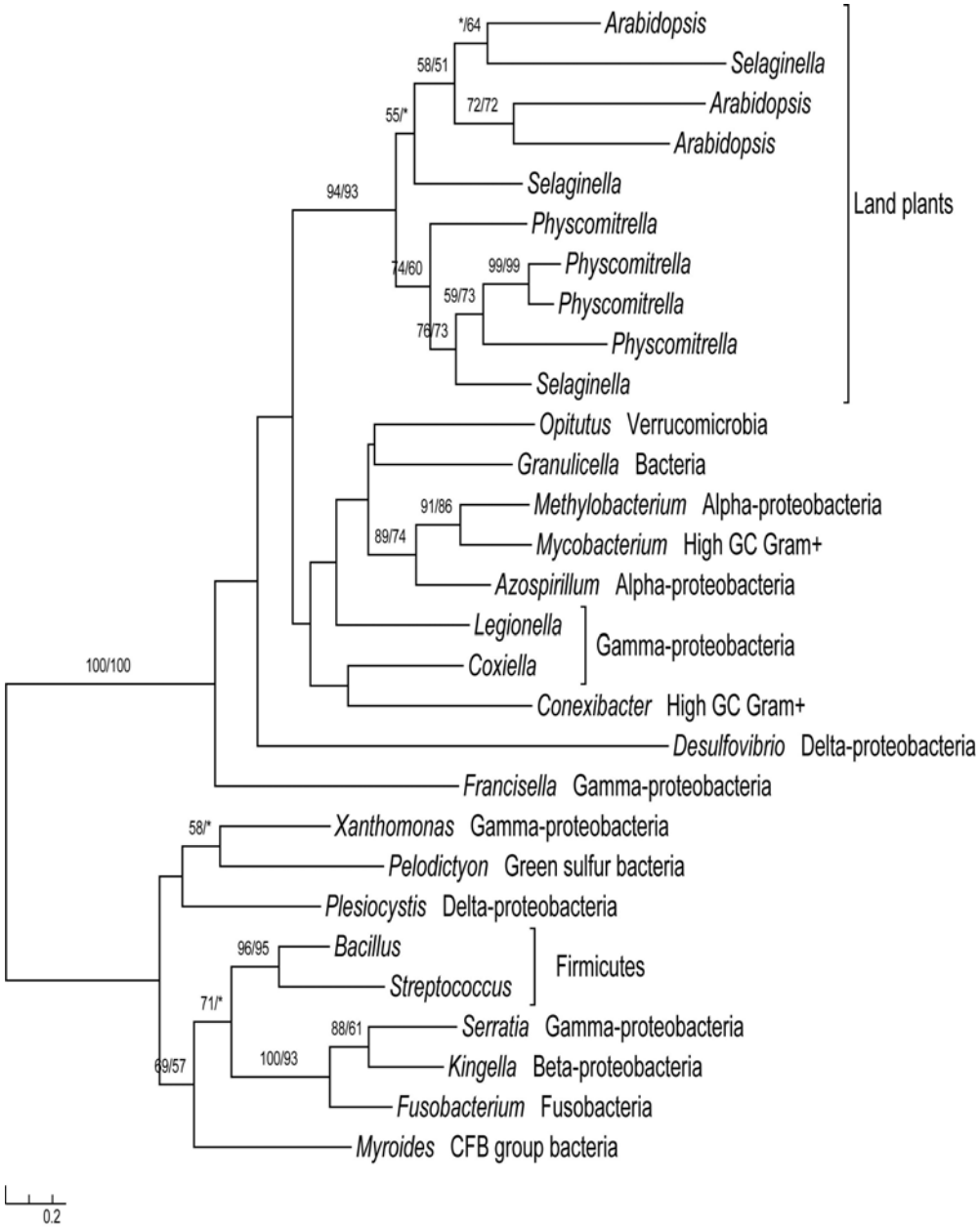
Supplementary Figure S39. Molecular phylogeny of wound-responsive family protein. Identifiable homologs were only found in green plants and prokaryotes. *Physcomitrella* sequences (Genbank GI numbers 168031639, 168046102, 168006875, 168012338, 168023049, 168014136) form a clade with homologs from other green plants. Only 95 reliably aligned amino acid residues was used to construct the phylogeny, therefore the support values for most clades are generally low. No cyanobacterial homologs were found, indicating that this gene family in green plants is unlikely of plastid (cyanobacterial) origin.



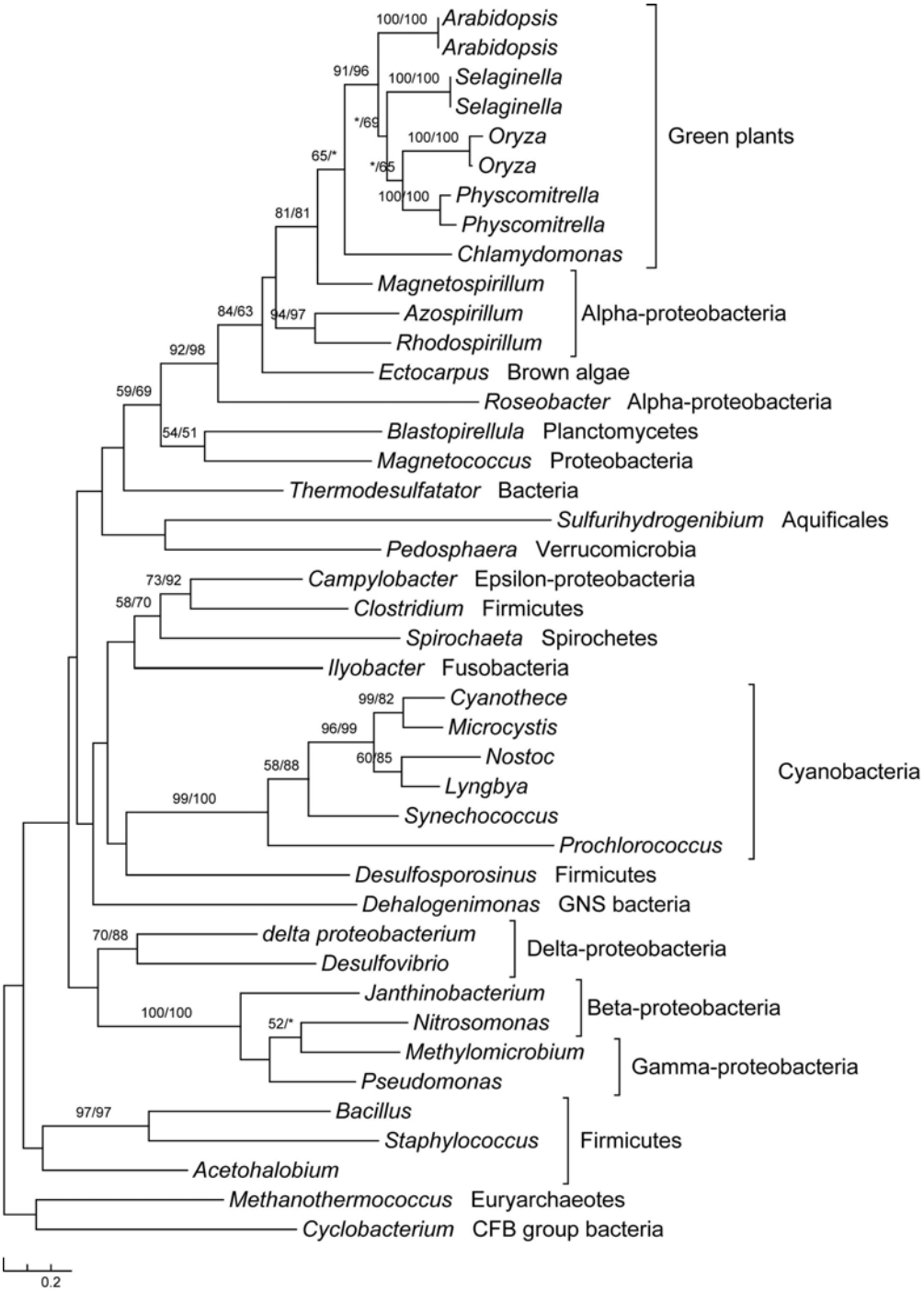
Supplementary Figure S40. Molecular phylogeny of ribosomal protein S6. One *Physcomitrella* sequence (Genbank GI number 162662263) forms a clade with homologs of proteobacteria and was likely acquired through HGT. Another *Physcomitrella* gene copy (Genbank GI number 168013078) form a clade with other land plant homologs. In *Arabidopsis*, this copy is annotated as regulator of fatty-acid composition 3 (RFC3) and is targeted to chloroplasts, suggesting the RFC3 in land plant is likely of plastid (cyanobacterial) origin.



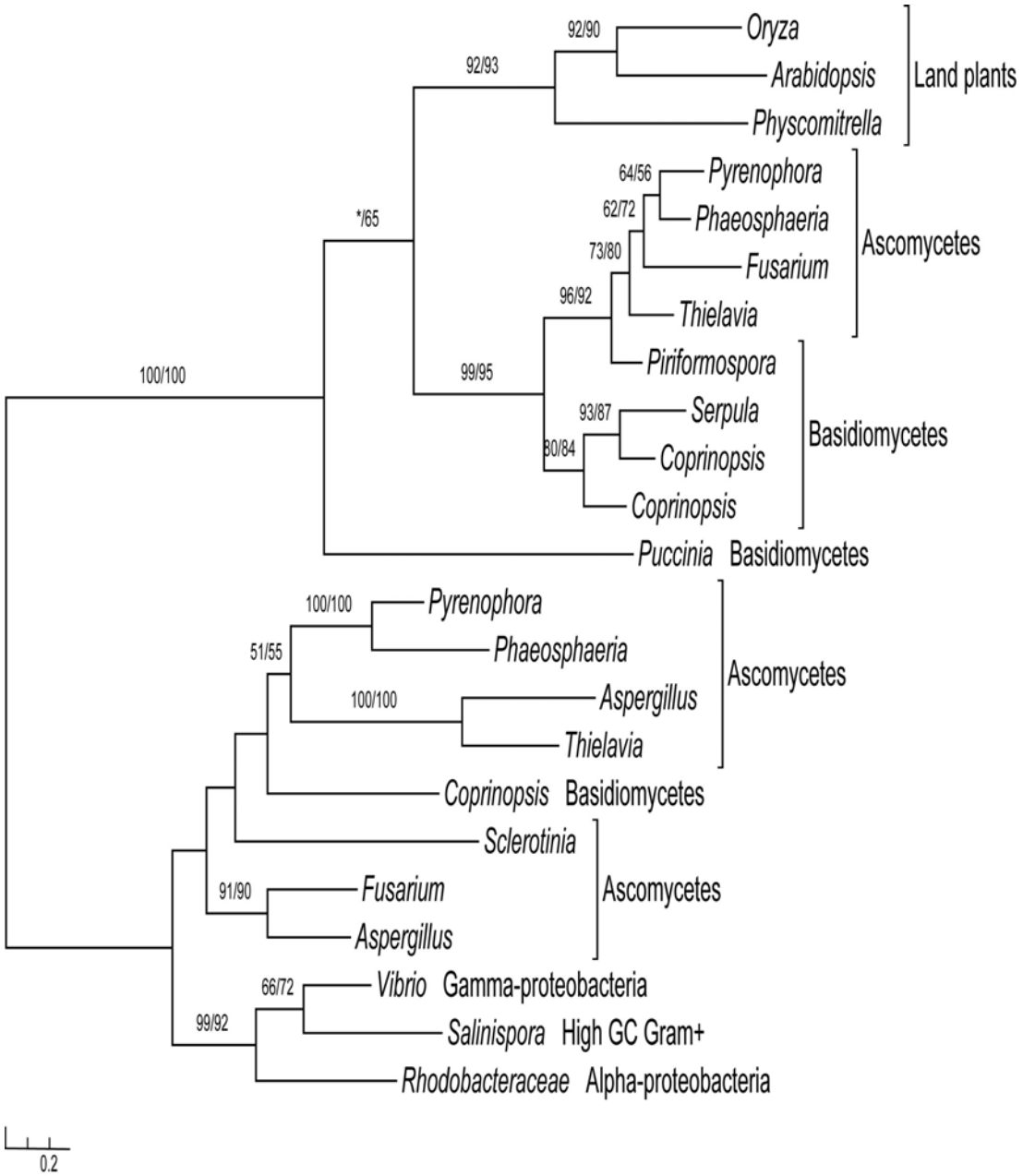
Supplementary Figure S41. Molecular phylogeny of fatty acyl-ACP thioesterases B (FATB). Identifiable homologs of *Physcomitrella* sequences (Genbank GI numbers 167998911, 168035219, 168024004, 168044508, 168036485) were only found in green plants and bacteria. No cyanobacterial homologs were found, indicating that this gene family in green plants is unlikely of plastid (cyanobacterial) origin.



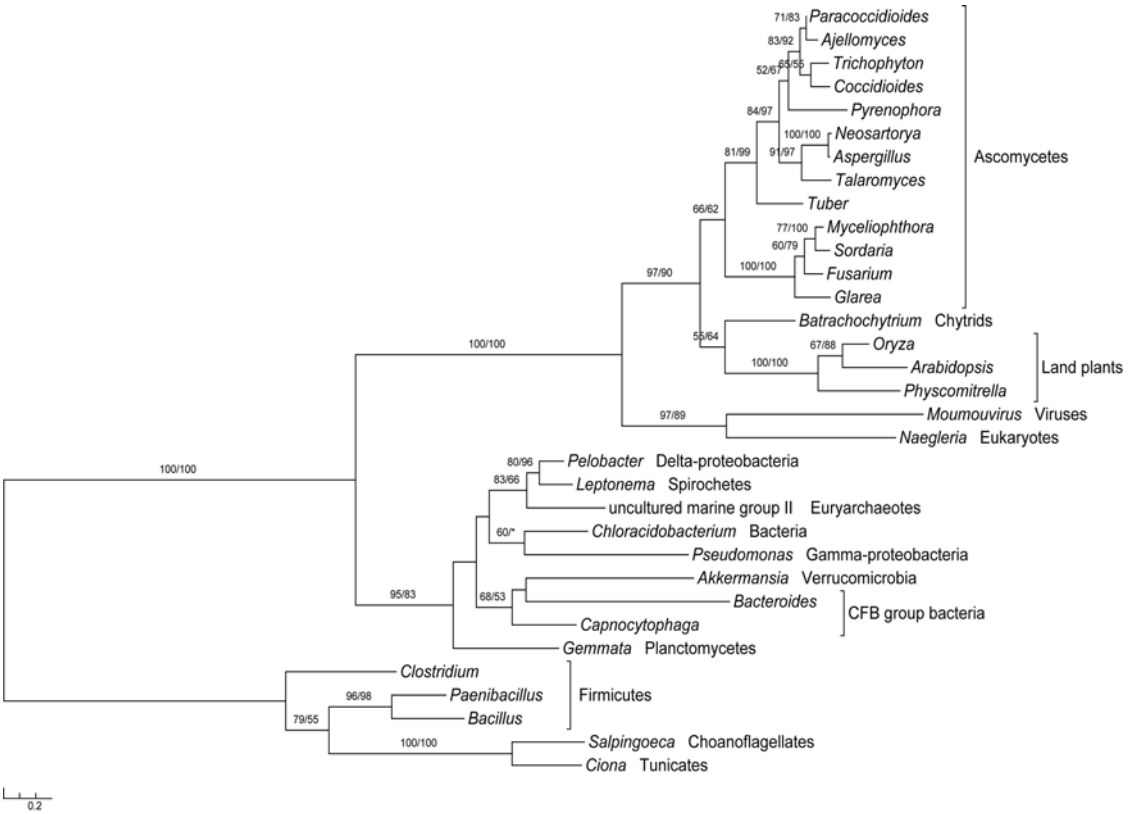
Supplementary Figure S42. Molecular phylogeny of HAD superfamily, subfamily IIIB acid phosphatase. *Physcomitrella* sequences (Genbank GI numbers 168062119, 168062518, 168033997, 168032668) form a highly supported clade with homologs of other land plants and bacteria. Their relationship is supported by several conserved amino acid residues and shared indels.



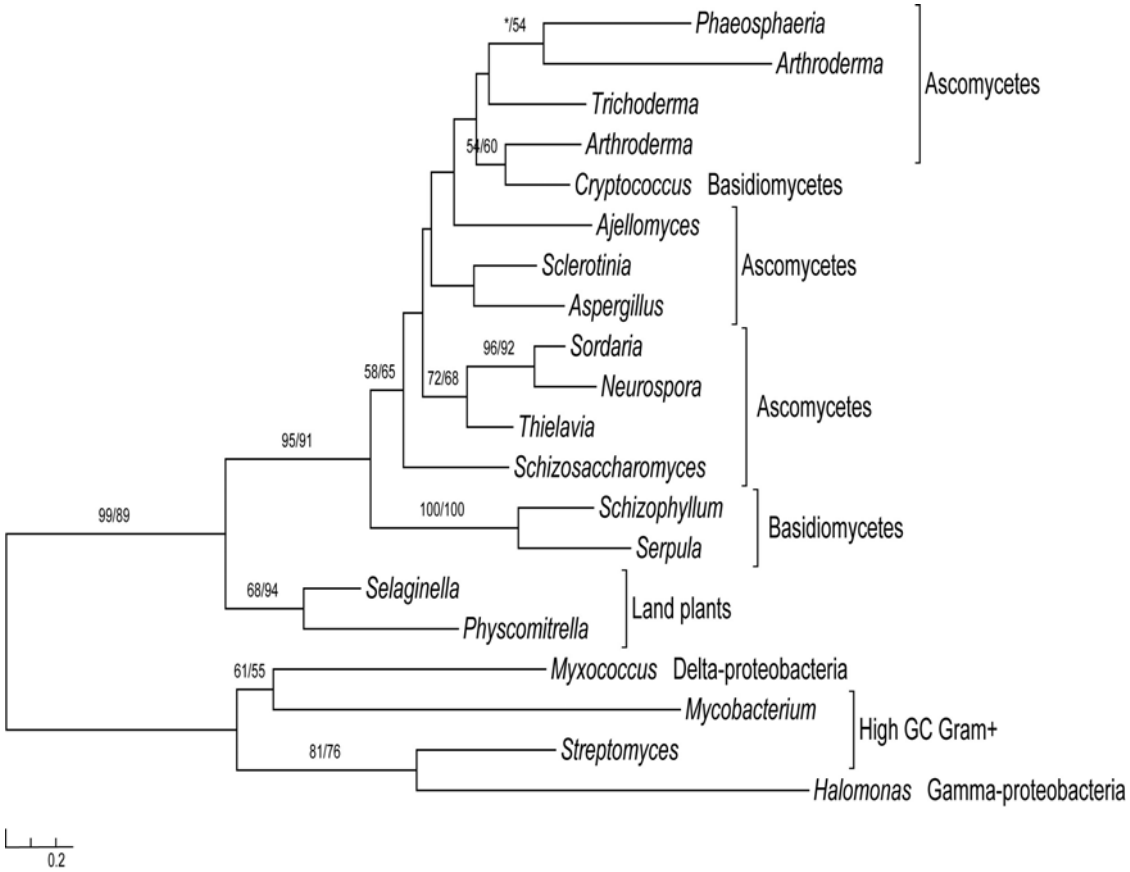
Supplementary Figure S43. Molecular phylogeny of N-acetyl-gamma-glutamyl-phosphate reductase（argC). *Physcomitrella* sequences (Genbank GI numbers 168037402，168031643) form a clade with other green plant homologs, which is in turn related to alpha-proteobacterial homologs. It is likely that brown alga *Ectocarpus* acquired this gene through an independent HGT event. However, the scenario of gene duplication in the common ancestor of brown algae and green plants and subsequent differential gene losses cannot be confidently excluded.



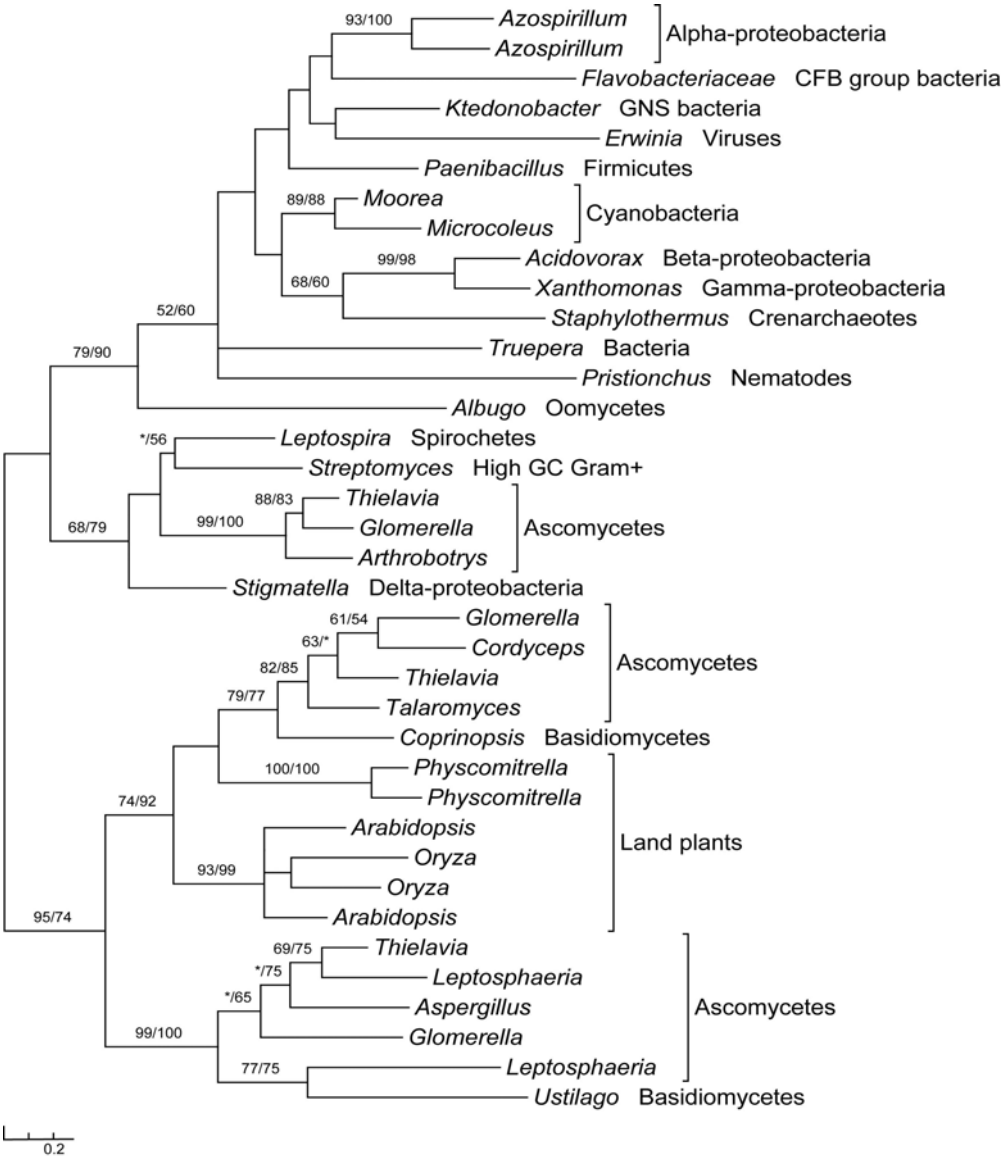
Supplementary Figure S44. Molecular phylogeny of NRPS-like enzyme. Identifiable homologs were only found in land pants, fungi and bacteria. *Physcomitrella* sequence (Genbank GI number 168054351) forms a highly supported clade with other land plant homologs, which in turn groups within the fungal clade.



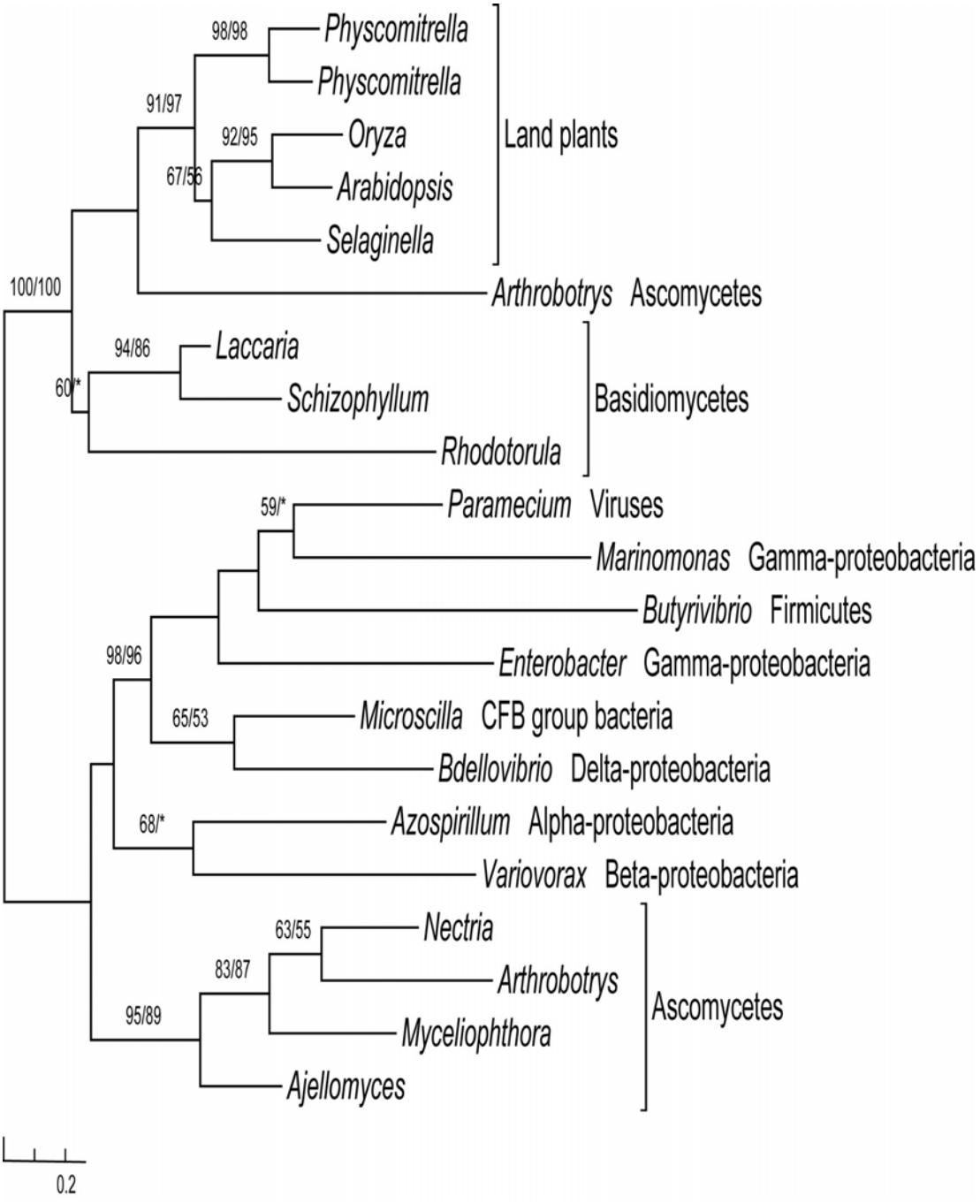
Supplementary Figure S45. Molecular phylogeny of flotillin-like protein. Identifiable homologs of *Physcomitrella* sequence are predominantly found in fungi and bacteria. *Physcomitrella* sequence (Genbank GI number 168017323) and other land plant sequences group within fungal homologs. These land plant and fungal sequences form a highly supported clade. *Naegleria* sequence is closely related to *Moumouvirus* homolog and they also share multiple conserved amino acid residues.



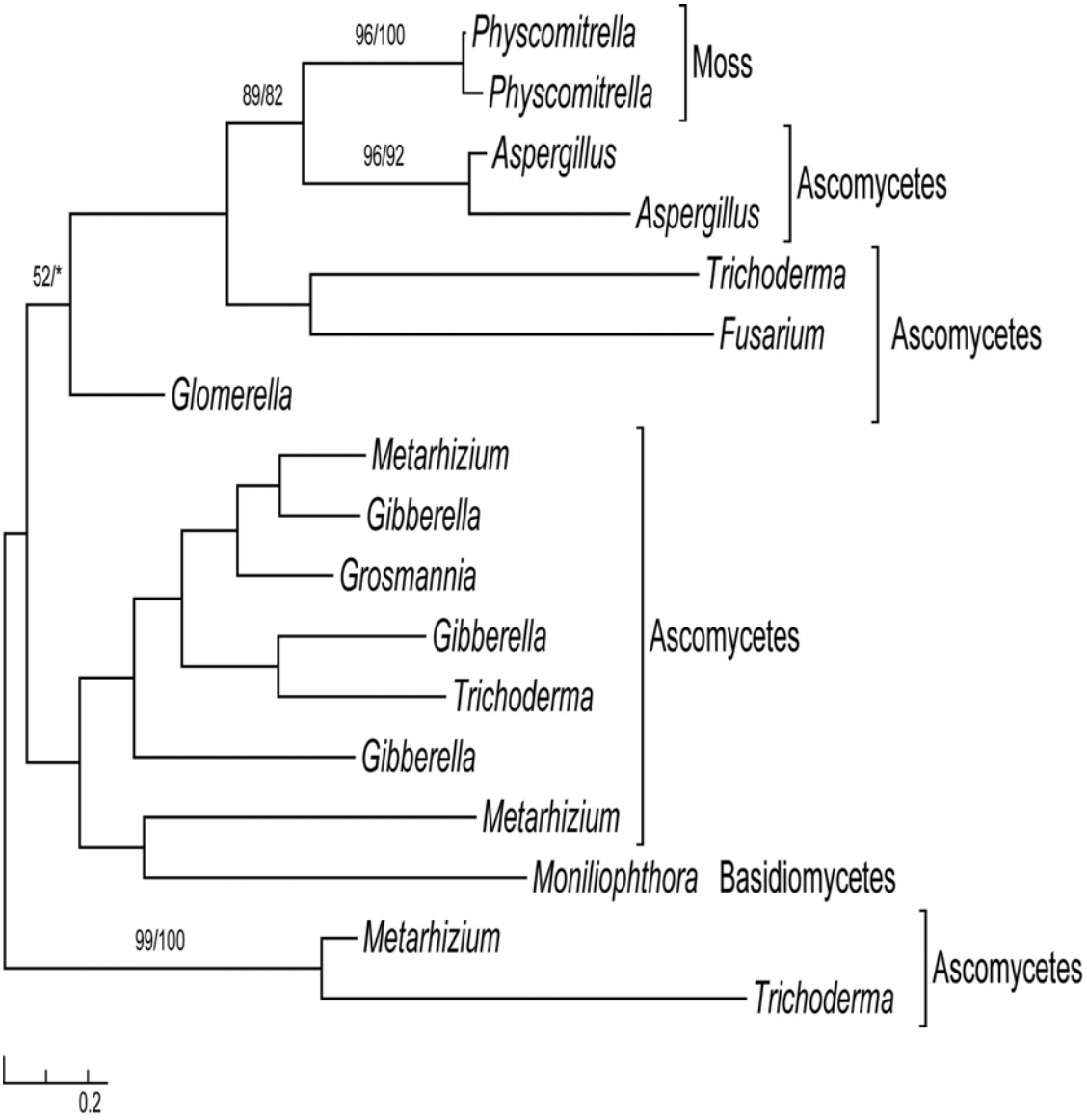
Supplementary Figure S46. Molecular phylogeny of hemerythrin HHE domain protein. Identifiable homologs of *Physcomitrella* sequence were only found in club moss, fungi and bacteria. *Physcomitrella* sequence (Genbank GI number 168061715) forms a highly supported clade with club moss sequences, which in turn is related to a fungal sequence clade. It is unclear whether this gene in land plants was acquired from fungi or bacteria.



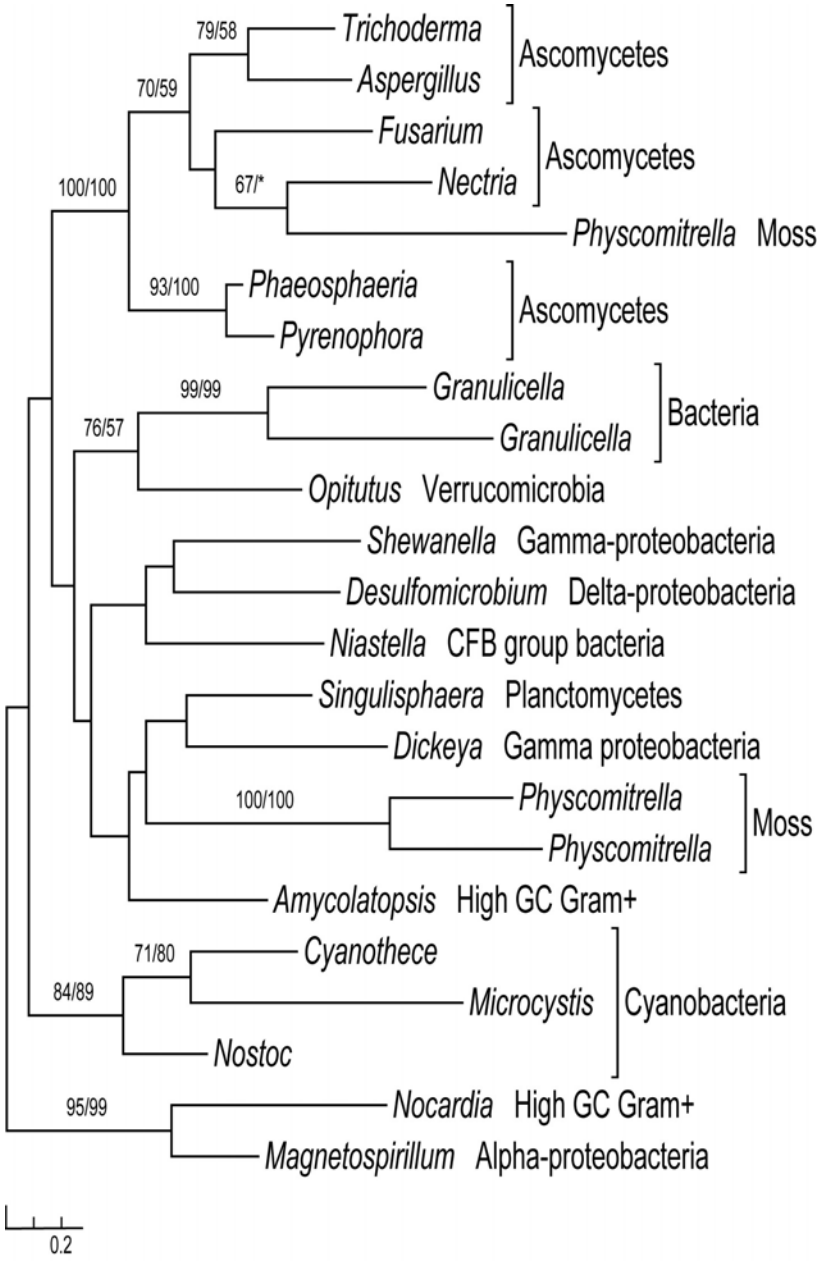
Supplementary Figure S47. Molecular phylogeny of gycosyl hydrolase family. *Physcomitrella* sequences (Genbank GI numbers 168029996, 168033792) form a monophyletic group with homologs from other land plants and fungi. Several other eukaryotic sequences group with bacterial homologs. It is unclear whether these sequences are of mitochondrial origin.



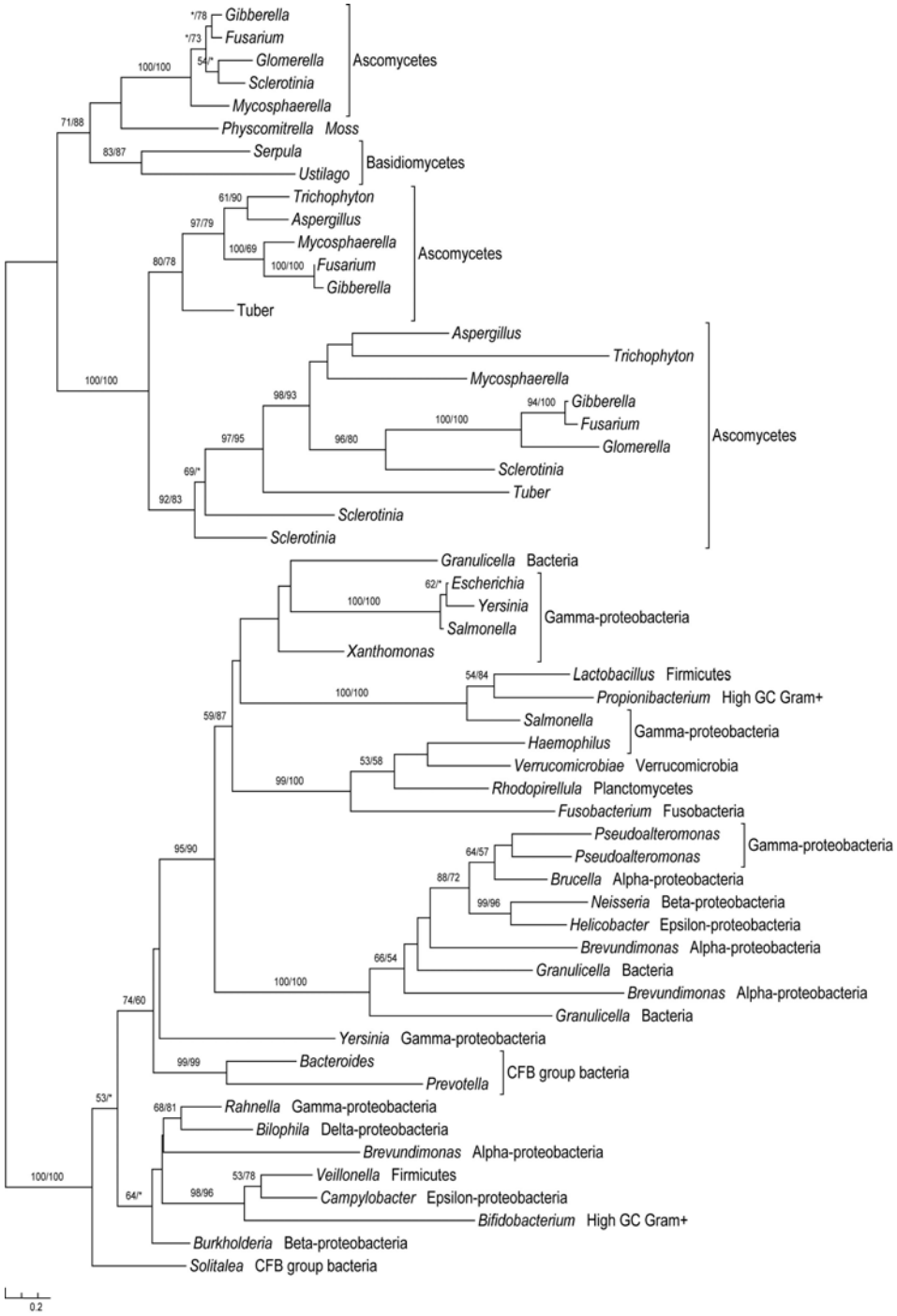
Supplementary Figure S48. Molecular phylogeny of β-1,4-mannosyl-glycoprotein. *Physcomitrella* sequences (Genbank GI numbers 168005754, 168016522) form a monophyletic group with other land plant homologs, which in turn groups within the fungal clade.



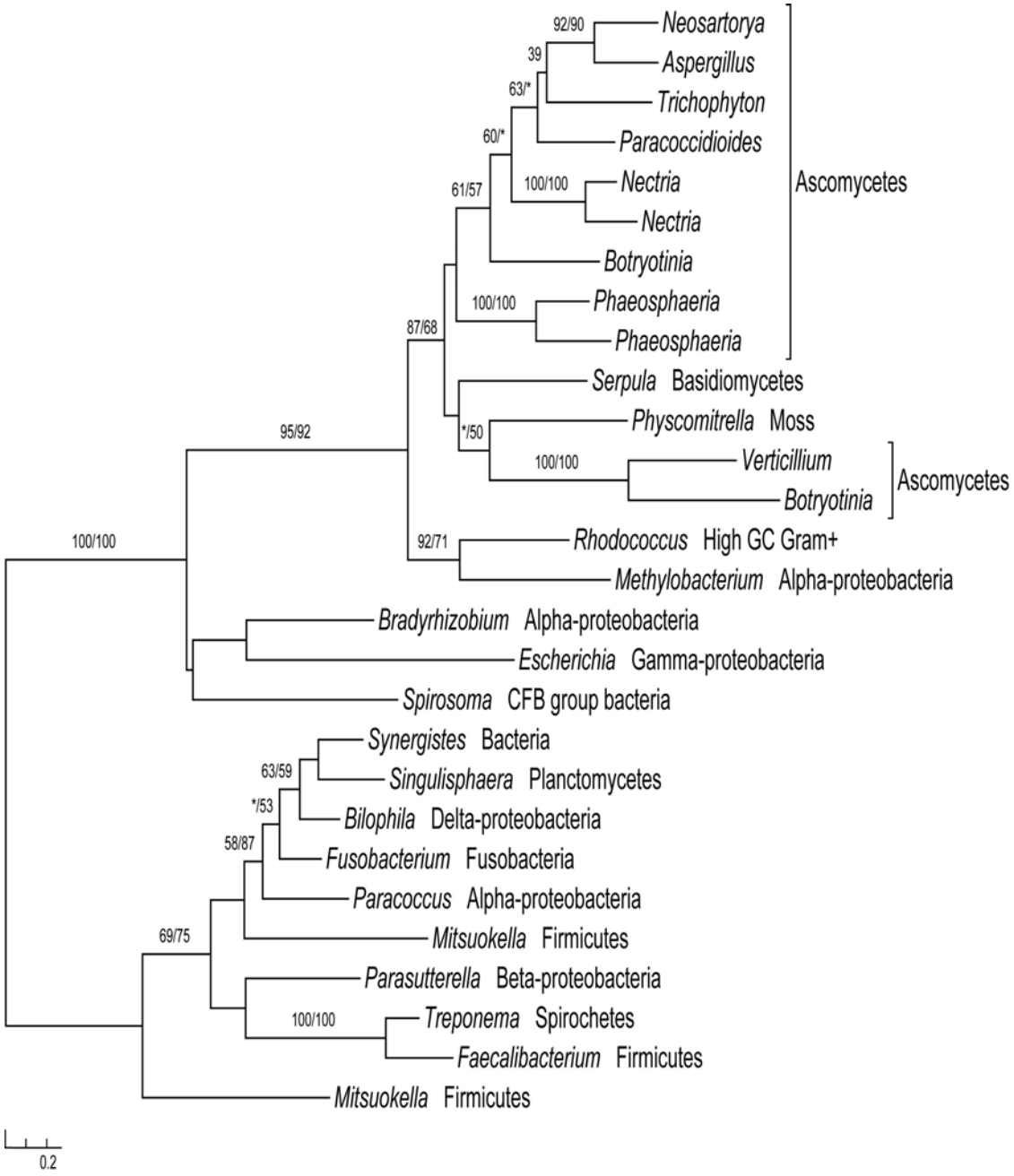
Supplementary Figure S49. Molecular phylogeny of killer toxin protein (KP4). Identifiable homologs of *Physcomitrella* sequences (Genbank GI numbers 168050142, 168050096) are only presented in fungi. *Physcomitrella* sequences form a highly support clade with *Aspergillus* homologs. It is likely that *Physcomitrella* acquired this gene from *Aspergillus* or other fungi.



Supplementary Figure S50. Molecular phylogeny of peptidoglycan binding domain containing protein. Identifiable homologs of *Physcomitrella* sequences are predominantly found in fungi and bacteria. One *Physcomitrella* sequence (Genbank GI number 168050434) appears to be derived from ascomycetes. Other two moss copies (Genbank GI numbers 168036135, 168019630) group with bacterial homologs but without strong support.



Supplementary Figure S51. Molecular phylogeny of l-fucose permease. Identifiable homologs of *Physcomitrella* sequences were only found in fungi and bacteria. *Physcomitrella* sequence (Genbank GI number 167997115) groups within the fungal clade, suggesting that this gene in *Physcomitrella* is likely of fungal origin, which confirms a previous study by Richards et al. (2009).



Supplementary Figure S52. Molecular phylogeny of a hypothetical protein. Moss sequences (Genbank GI number 168028121) groups within a large fungal clade. Identifiable homologs of *Physcomitrella* were only found in fungi and bacteria. The fungal origin of this moss sequence confirms a previous study by Richards et al. (2009).

Supplementary Table S1. Acquired genes identified in *Physcomitrella patens*. Homologs of heterokaryon incompatibility (HET) superfamily were only founded in *P. patens* and fungi. The identification of this gene family was exclusively based on taxonomic distribution and no phylogeneitc tree was constructed for this family. Asterisks indicate genes that were also reported by earlier studies. "y" indicates existence of indels shared by potential donors and recipient organisms in multiple sequence alignments.

Putative gene product	**GI number**	**Gene function**	**Putative donor**	**Taxonomic distribution**	**indels**	**Figure**	**Homologous locus in *Arabidopsis***
Subtilase family	168024416	Proteolysis	Bacteria	Land plants		Fig.1a	AT2G04160
	168049684						AT2G04160
	168009784						AT3G14240
	168006037						AT3G14240
	168051252						AT4G34980
	168034558						AT2G19170
	168033556						AT2G19170
	168017764						AT2G19170
	168000855						AT2G19170
	168003990						AT4G30020
Arginase	168024860	Polyamine biosynthesis	Bacteria	Land plants	y	Fig. S3	AT4G08900
Acyl-activating enzyme 18 (AAE18)	168064848	Auxin biosynthesis	Bacteria	Green plants	y	Fig. 2b	AT1G55320
	168042921						
	168064660						
YUCCA flavin monooxygenase (YUC3)	168013839	Auxin biosynthesis	Bacteria	Land plants	y	Fig. S4	AT4G28720
	168007310						AT1G04610
	168038243						AT5G25620
	168047840						AT1G04610
	168059684						AT4G13260

Glutamate-cysteine ligase (GCL)	168009654	Glutathione synthesis	Proteobacteria	Green plants		Fig. S5	AT4G23100
	168067242						
	168052608						
Wound-responsive family protein	168031639	Defense response	Bacteria	Green plants		Fig. S39	AT1G19660
	168046102						AT1G19660
	168006875						AT1G75380
	168012338						AT1G19660
	168023049						AT1G19660
	168014136						AT1G19660
HAD superfamily, subfamily IIIB acid phosphatase	168062119	Herbivorous insect resistance	Bacteria	Land plants	y	Fig. S42	AT4G29260
	168062518						
	168033997						
	168032668						
NRPS-like enzyme	168054351	Oxidative stress resistance	Fungi	Land plants		Fig. S44	AT4G18540
N-acetyl-gamma-glutamyl-phosphate reductase （argC)	168037402	Cadmium stress response	α-proteobacteria	Green plants		Fig. S43	AT2G19940
	168031643						
HAD-superfamily hydrolase	168001220	Cold stress response	Bacteria	Green plants	y	Fig. S18	AT5G48960
Killer toxin Protein (KP4)	168050142	Pathogen resistance	Ascomycetes	Moss		Fig. S49	no homology
	168050096						
Flotillin-like protein	168017323	Endocytosis	Ascomycetes	Land plants		Fig. S45	AT5G25260
Allantoate amidohydrolase (AAH)	167997139	Purine degradation	Bacteria	Green plants	y	Fig. S7	AT4G20070
	168064079						
Ureidoglycolate amidohydrolase (UAH)	168010247	Purine degradation	Bacteria	Green plants		Fig. S7	AT5G43600
Guanine deaminase (GDA)	168025229	Purine degradation	α-proteobacteria	Moss	y	Fig. S6	no homology

PfkB family kinase	167998254 168016595 168021833	Vitamin B6 salvaging	δ-proteobacteria	Land plants		Fig. S36	AT5G58730
Methionine gamma-lyase (MGL)	168013924 168008405	L-methionine degradation	CFB bacteria	Green plants	y	Fig. S20	AT1G64660
Glutamine synthetase (GS)	168040136	Glutamine biosynthesis	CFB bacteria	Moss		Fig. S8	no homology
3,4-dihydroxy-2-butanone 4-phosphate synthase (ribB)	168028296	Riboflavin biosynthesis	Euryarchaeotes	Land plants	y	Fig. S13	no homology
Hemerythrin HHE domain protein	168061715	Iron homeostasis	Ascomycetes	Land plants		Fig. S46	no homology
Hydroxypyruvate reductase 2 (HPR2)	167997717 168037243	Photorespiratory	Bacteria	Land plants	y	Fig. S26	AT1G79870
Inositol 2-dehydrogenase like protein	168003329	Pollen germination and tube growth	α-proteobacteria	Land plants	y	Fig. S27	AT4G17370
Peptidoglycan binding domain containing protein	168050434	Peptidoglycan binding	Ascomycetes	Moss		Fig. S50	no homology
Sugar isomerase (SIS) family	168059735 168019301	Sugar binding	α-proteobacteria	Green plants		Fig. S14	AT5G52190
Limit dextrinase (LDA)	168038552	Starch biosynthesis	Bacteria	Green plants	y	Fig. S14	AT5G04360
β-glucosidase	168069539 168059435	Cellulose degradation	Bacteria	Land plants	y	Fig. S15	AT5G04885 AT5G20950
Gycosyl hydrolase family	168029996 168033792	Carbohydrate methabolism	Ascomycetes	Land plants		Fig. S47	AT1G13130 AT3G26140
Glycoside hydrolase	168052263	Carbohydrate methabolism	δ-proteobacteria	Moss		Fig. S23	no homology

Glycoside hydrolase family 2	168036598	Carbohydrate methabolism	γ-proteobacteria	Green plants		Fig. S25	AT3G54440
α-L-rhamnosidase	168031461	Carbohydrate methabolism	CFB bacteria	Moss		Fig. S33	no homology
FAD linked oxidase	168012414	Oxygen-dependent oxidoreductases	CFB bacteria	Moss		Fig. S1	no homology
	168045341						
Short-chain dehydrogenase/reductase SDR	168031790	Oxidation-reduction	Proteobacteria	Moss		Fig. S28	no homology
Fatty acyl-ACP thioesterases B (FATB)	167998911	Fatty acid biosynthesis	Bacteria	Green plants		Fig. S41	AT1G08510
	168035219						AT1G08510
	168024004						AT1G08510
	168044508						AT1G08510
	168036485						AT4G13050
1,4-dihydroxy-2-naphthoate octaprenyltransferase	168009868	Menaquinone biosynthesis	δ-proteobacteria	Moss		Fig. S37	no homology
Phosphoenolpyruvate carboxylase (PEPCase)	168029489	Carbon fixation	γ-proteobacteria	Moss		Fig. S2	no homology
GroES-like zinc-binding alcohol dehydrogenase family	168030245	Glycolysis	High GC gram+	Land plants	y	Fig. S16	AT5G63620
Pyruvate kinase	168053775	Glycolysis	Bacteria	Land plants	y	Fig. S21	AT3G49160
	168053903						
Phosphoglycerate kinase (PGK)	168058081	Glycolysis	δ-proteobacteria	Land plants	y	Fig.S17	no homology
	168034630						
ATP-binding cassette I1 (ABCI1) transporter	168004297	Molecular transport	Bacteria	Land plants	y	Fig. S29	AT1G63270
Uracil permease	168012184	Nucleobase transport	Bacteria	Green plants	y	Fig. S30	AT5G03555

	168043133						
L-fucose permease*	167997115	Sugar transport	Ascomycetes	Moss		Fig. S51	no homology
β-1,4-mannosyl-glycoprotein	168005754	Glycosyl transfering	Basidiomycetes	Land plants		Fig. S48	AT5G14480
	168016522						
DNA repair family protein	168044851	DNA replication	Ascomycetes	Moss		Fig. S12	no homology
Toprim domain-containing protein	168040643	DNA replication	Bacteria	Green plants	y	Fig. S9	AT1G30680
DNA topoisomerase I	168037859	DNA replication	Proteobacteria	Green plants	y	Fig. S10	AT4G31210
phage/plasmid primase, P4 family	168026035	DNA replication	Viruses	Moss		Fig. S11	no homology
	168057313						
	168032336						
	168009191						
	168041210						
Ribosomal protein S6	162662263	RNA binding	β-proteobacteria	Moss		Fig. S40	no homology
M6 family peptidase	168013514	Peptidase activity	Bacteria	Moss		Fig. S35	no homology
	168019010						
	168032091						
	168048759						
Amidohydrolase family	168021897	Hydrolase activity	Bacteria	Land plants	y	Fig. S31	no homology
Amidase family protein	168042262	Acrylonitrile metabolism	Bacteria	Green plants	y	Fig. S32	AT5G07360
	168003211						
D-alanine-D-alanine ligase family	168012025	Peptidoglycan biosynthesis	Chlamydiae/CFB bacteria	Green plants		Fig. S34	AT3G08840
Dienelactone hydrolase family	168063002	Hydrolase activity	Bacteria	Moss		Fig. S38	no homology
Vein Patterning 1 (VEP1)	168003008	Vascular development	Bacteria	Land plants	y	Fig. 1b	AT4G24220
Heterokaryon incompatibility (HET) family	168037338	Heterokaryon formation	Fungi	Moss		No fig	no homology
	168042194						

	168042266						
	168042178						
	168043854						
	168066323						
	168042180						
	168041122						
	168043902						
	168043828						
	168042274						
	168041824						
	168007634						
	168043856						
	168042184						
	168042182						
	168015126						
	168049214						
	168062785						
	168057911						
ybiU protein	168021919	Unknown	High GC gram+	Moss	y	Fig. S22	no homology
Acyl-CoA N-acyltransferase	168042611	Unknown	α-proteobacteria	Green plants		Fig. S19	At2g23390
Hypothetical protein*	168028121	Unknown	Ascomycetes	Moss		Fig. S52	no homology

Supplementary Table S2. Sources of additional sequences used for database construction.

Species	Taxonomic group	Data type	Taxid
Acanthamoeba castellanii	Amoebozoa	EST	5755
Acetabularia acetabulum	Viridiplantae	EST	35845
Alexandrium tamarense	Alveolata	Genome	2925
Allomyces macrogynus	Fungi	EST	28583
Amphidinium carterae	Alveolata	Genome	2961
Antonospora locustae	Fungi	EST	278021
Astasia longa	Fungi	EST	3037
Barachionus plicatilis	Metazoa	Genome	10195
Bigelowiella natans	Rhizaria	Genome	227086
Capitella teleta	Metazoa	Genome	283909
Chlorella vulgaris	Viridiplantae	Genome	3077
Cyanidioschyzon merolae	Eukaryota	EST	45157
Cyanophora paradoxa	Glaucocystophyceae	EST	2762
Diplonema papillatum	Diplonemida	Genome	91374
Emiliania huxleyi	Eukaryota	Genome	2903
Euglena gracilis	Euglenozoa	EST	3039
Glaucocystis nostochinearum	Glaucocystophyceae	Genome	38271
Guillardia theta	Cryptophyta	EST	55529
Hartmannella vermiformis	Amoebozoa	EST	5778
Heterocapsa triquetra	Alveolata	EST	66468
Histiona aroides	Jakobida	EST	392300
Hyperamoeba dachnya	Amoebozoa	EST	181200
Isochrysis galbana_CCMP_1323	Haptophyceae	EST	37099
Jakoba bahamiensis	Excavata	EST	221721
Jakoba libera	Excavata	EST	143017
Karenia brevis	Alveolata	EST	156230
Karlodinium micrum	Alveolata	Genome	342587
Lottia gigantea	Metazoa	EST	225164
Malawimonas californiana	Excavata	EST	221722
Mastigamoeba balamuthi	Amoebozoa	EST	108607
Mesostigma viride	Viridiplantae	Genome	41882
Mortierella verticillata	Fungi	EST	78898
Neocallimastix patriciarum	Fungi	EST	4758
Nephroselmis olivacea	Viridiplantae	EST	31312
Nuclearia simplex_strain_2	Opisthokonta	EST	154970
Oxytricha trifallax	Alveolata	EST	94289

Paracercomonas marina	Rhizaria	EST	372086
Pavlova lutheri	Haptophyceae	EST	2832
Physarum polycephalum	Amoebozoa	Genome	5791
Phytophthora ramorum	Stramenopiles	Genome	164328
Polysphondylium pallidum	Amoebozoa	EST	13642
Polytomella parva	Viridiplantae	EST	51329
Porphyra yezoensis	Rhodophyta	EST	2788
Proterospongia choanojuncta	Choanoflagellida	EST	218848
Prototheca wickerhamii	Viridiplantae	EST	3111
Reclinomonas americana	Jakobida	Genome	48483
Rhizopus oryzae	Fungi	EST	64495
Saitoella complicata	Fungi	Genome	5606
Sawyeria marylandensis	Heterolobosea	EST	194530
Scenedesmus obliquus	Viridiplantae	EST	3088
Seculamonas ecuadoriensis	Jakobida	Genome	221724
Sphaeroforma arctica	Opisthokonta	EST	72019
Spironucleus vortens	Fornicata	Genome	58336
Spizellomyces punctatus	Fungi	EST	109760
Stachyamoeba lipophora	Heterolobosea	EST	463046
Streblomastix strix	Oxymonadida	EST	222440
Taphrina deformans	Fungi	EST	5011
Tetrahymena thermophila	Alveolata	Genome	5911
Thecamonas trahens	Eukaryota	EST	529818