Gramene Compara GeneTrees:
A Phylogenomics Resource for Plants

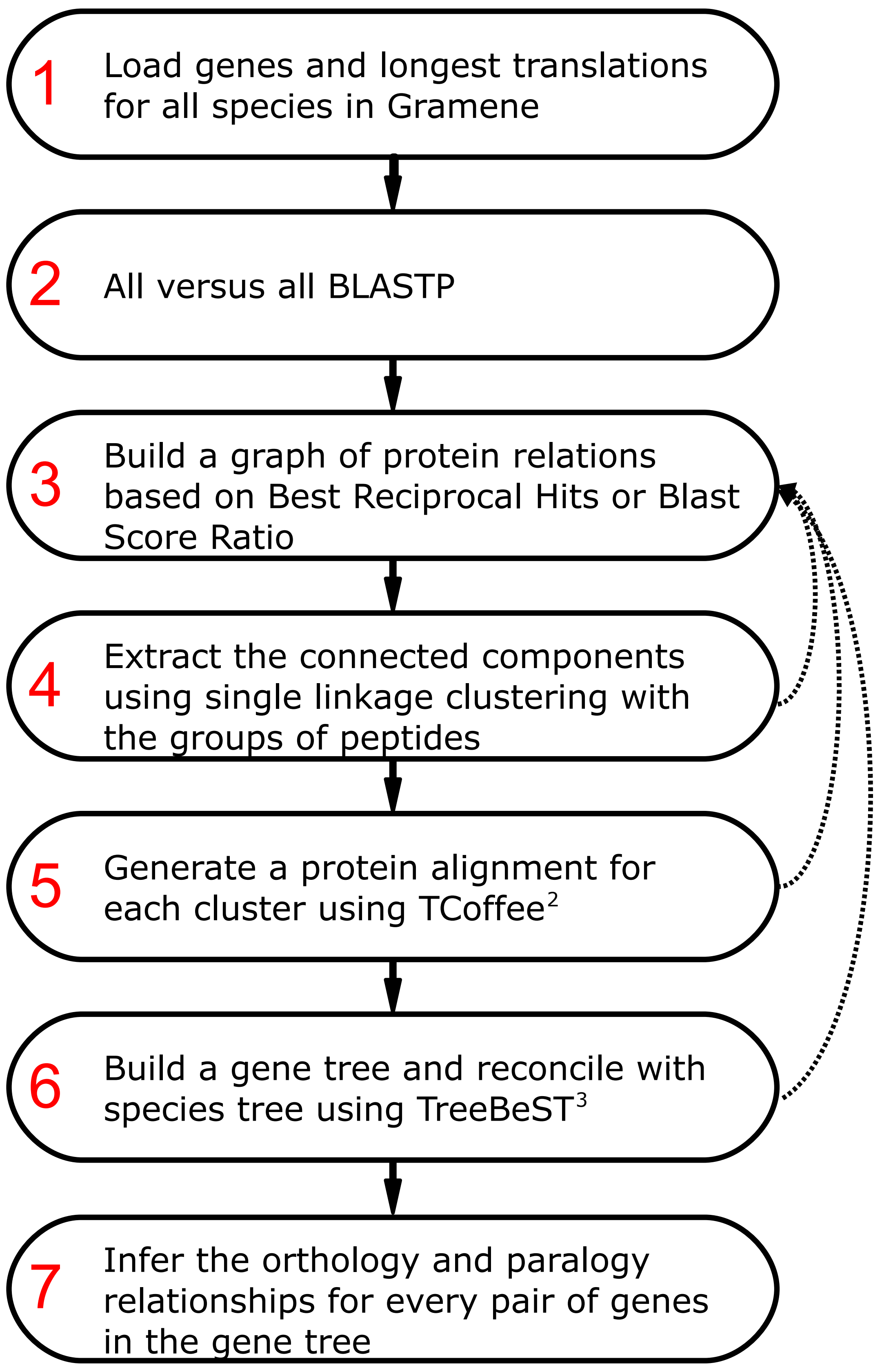
William Spooner[1], Joshua C. Stein[1], Sharon Wei[1], Liya Ren[1], Doreen Ware[1][2]
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
[2]USDA-ARS NAA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY 14853

Abstract

Comparative functional genomics allows researchers to trace evolutionary histories of genes and traits. We present the database and web-site of Gramene Compara GeneTrees. The previously published Ensembl method uses clustering to define gene families and phylogenetic reconstruction to define orthologs and paralogs. It was applied to the complete genomes of six grass-lineage and four eudicot species, generating 27,031 families. As expected many species-specific genes and families were identified and most were attributed to differentially expanded transposable elements. Approximately 77% of all genes was assigned an orthologous relationship and 76% paralogous. Secondary phylogenetic analyses of called orthologs were in agreement with the expected species tree, demonstrating internal consistency of the method. Concordance of InterPro annotation was evaluated when compared between rice and Arabidopsis orthologs. Synteny between rice and sorghum, which was preserved from both speciation and an ancient whole genome duplication event, was used to demonstrate good sensitivity and specificity of ortholog and paralog calls. The Gramene website environment is integrated with genome browsers, comparative maps and functional annotations, including gene ontology and InterPro. The Compara database adds a new level of tools to aid researchers in the making inferences of function and strategies for gene annotation.

Data Generation

Ensembl Compara Gene Tree Pipeline[1]



1 Load genes and longest translations for all species in Gramene

2 All versus all BLASTP

3 Build a graph of protein relations based on Best Reciprocal Hits or Blast Score Ratio

4 Extract the connected components using single linkage clustering with the groups of peptides

5 Generate a protein alignment for each cluster using TCoffee[2]

6 Build a gene tree and reconcile with species tree using TreeBeST[3]

7 Infer the orthology and paralogy relationships for every pair of genes in the gene tree

Funding

This work was initially supported (2001-2004) by the USDA Initiative for Future Agriculture and Food Systems (IFAFS) (grant no. 00-52100-9622) and a Cooperative State Research and Education Service (CSREES) agreement through the USDA Agricultural Research Service (grant no. 58-1907-0-041). For the years 2004-2007 this work was supported by the National Science Foundation (NSF) PGI grant award #0321685. Current work is being supported by the NSF Plant Genome Research Resource grant award #0703908.

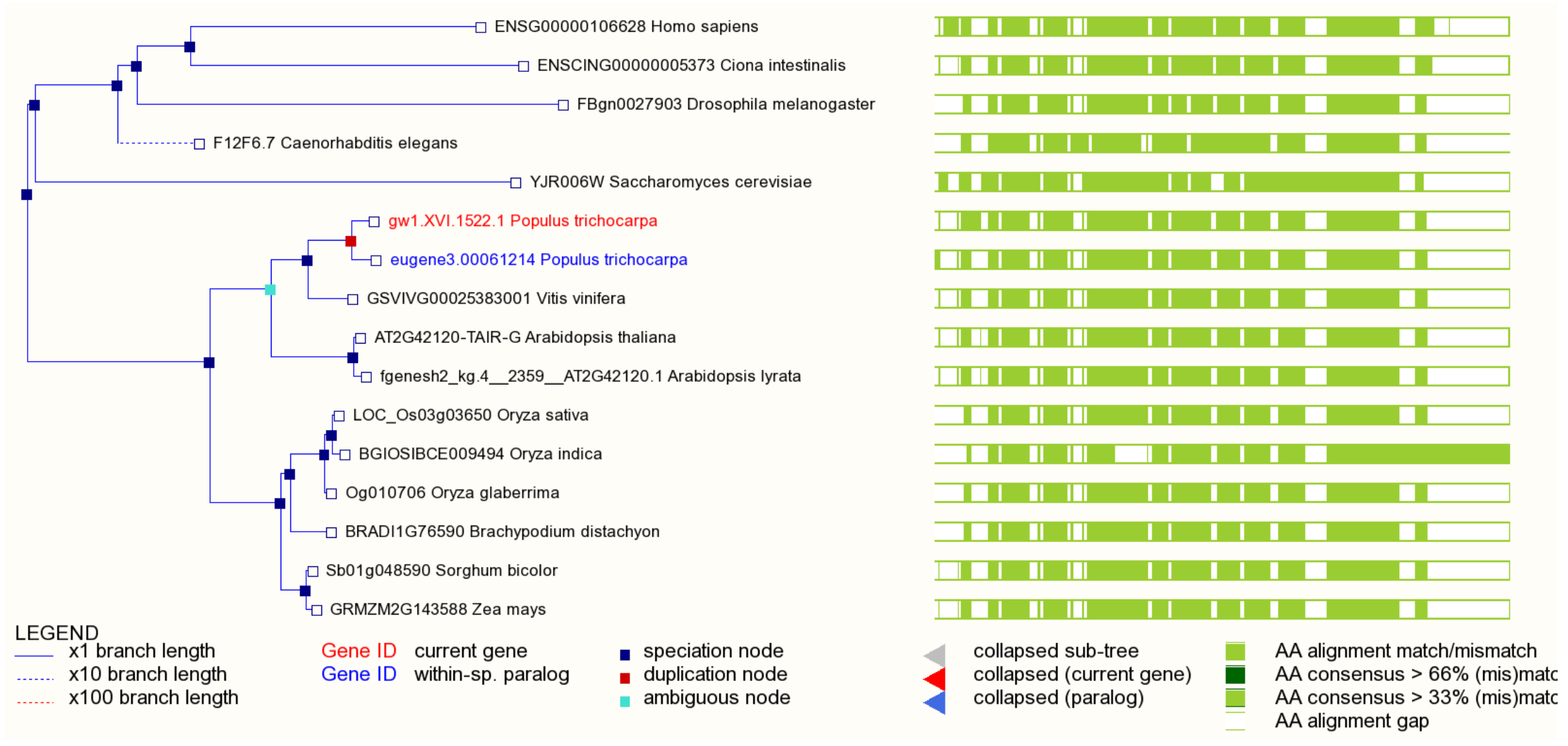
References

1. Vilella A.J., et al. (2008). *Genome Res.* Pre-print: doi:10.1101/gr.073585.107
2. Edgar, R.C. (2004). *Nucleic Acids Res* **32:** 1792-1797.
3. Li, H. (2008). http://treesoft.sourceforge.net/treebest.shtml
4. Kent, W.J.,et al. (2003). *Proc Natl Acad Sci U S A* **100:** 11484-11489..
5. Kent, W.J., (2002). *Genome Res.* 12: 656-664..

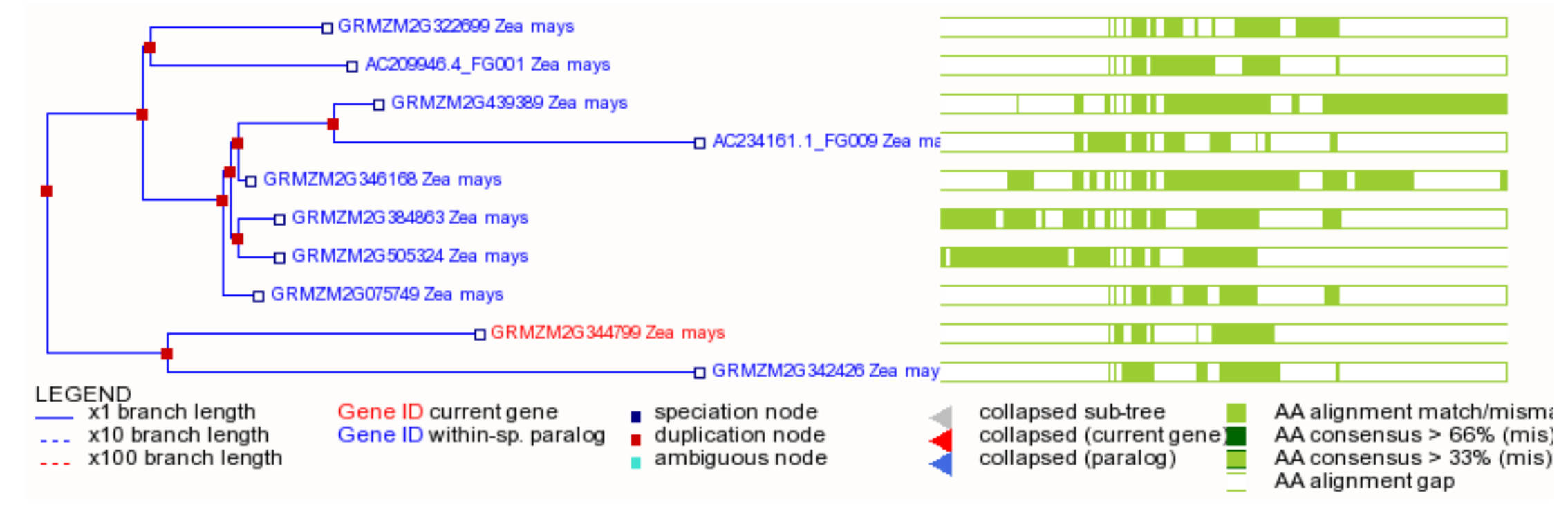
Visualisation

Example GeneTrees

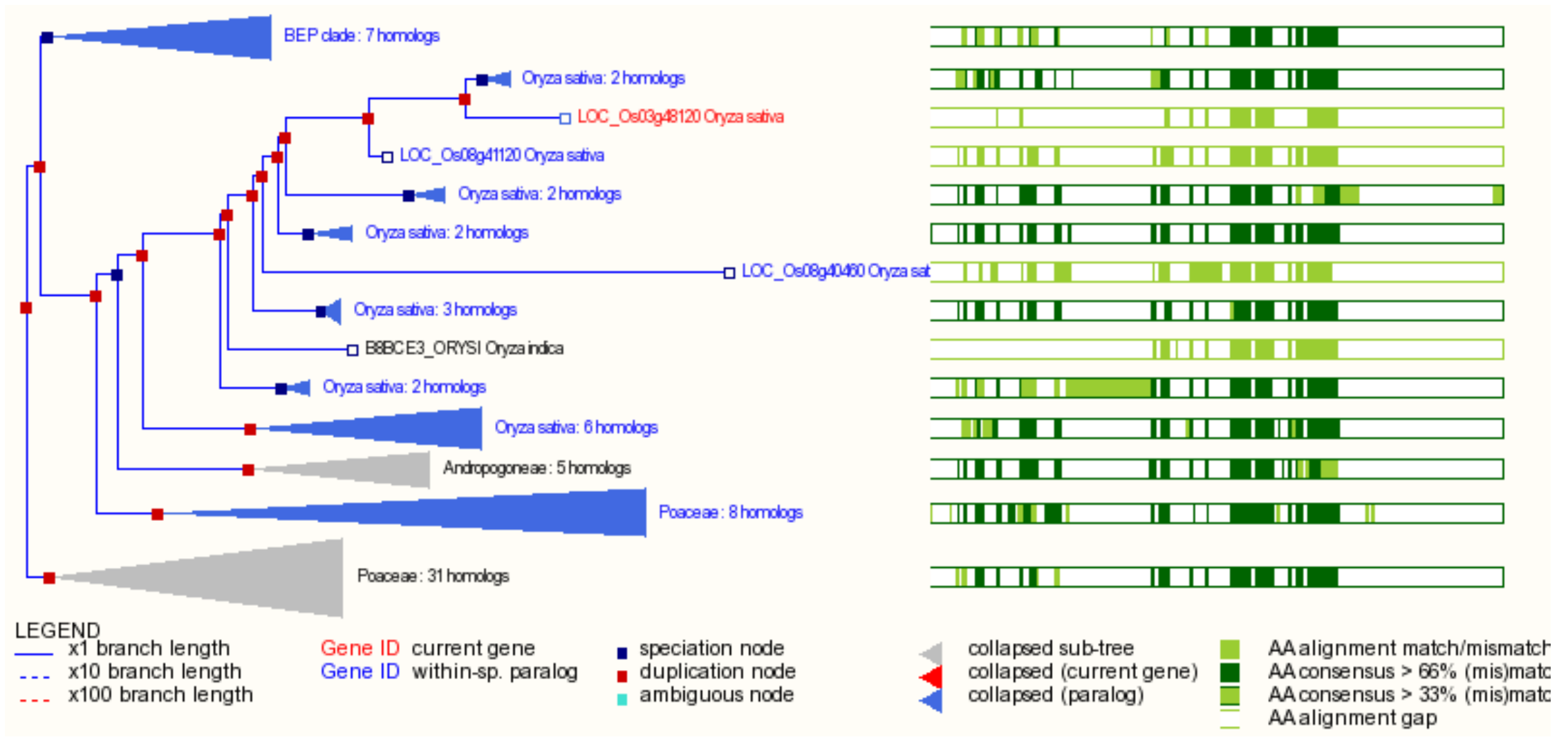
POLD2 (DNA polymerase delta subunit 2) protein family


ENSG00000106628 Homo sapiens
ENSCING00000005373 Ciona intestinalis
FBgn0027903 Drosophila melanogaster
F12F6.7 Caenorhabditis elegans
YJR006W Saccharomyces cerevisiae
gw1.XVI.1522.1 Populus trichocarpa
eugene3.00061214 Populus trichocarpa
GSVIVG00025383001 Vitis vinifera
AT2G42120-TAIR-G Arabidopsis thaliana
fgenesh2_kg.4__2359__AT2G42120.1 Arabidopsis lyrata
LOC_Os03g03650 Oryza sativa
BGIOSIBCE009494 Oryza indica
Og010706 Oryza glaberrima
BRADI1G76590 Brachypodium distachyon
Sb01g048590 Sorghum bicolor
GRMZM2G143588 Zea mays
LEGEND
x1 branch length
x10 branch length
x100 branch length
Gene ID current gene
Gene ID within-sp. paralog
speciation node
duplication node
ambiguous node
collapsed sub-tree
collapsed (current gene)
collapsed (paralog)
AA alignment match/mismatch
AA consensus > 66% (mis)matc
AA consensus > 33% (mis)matc
AA alignment gap

Uncharacterised Maize Retrotransposon

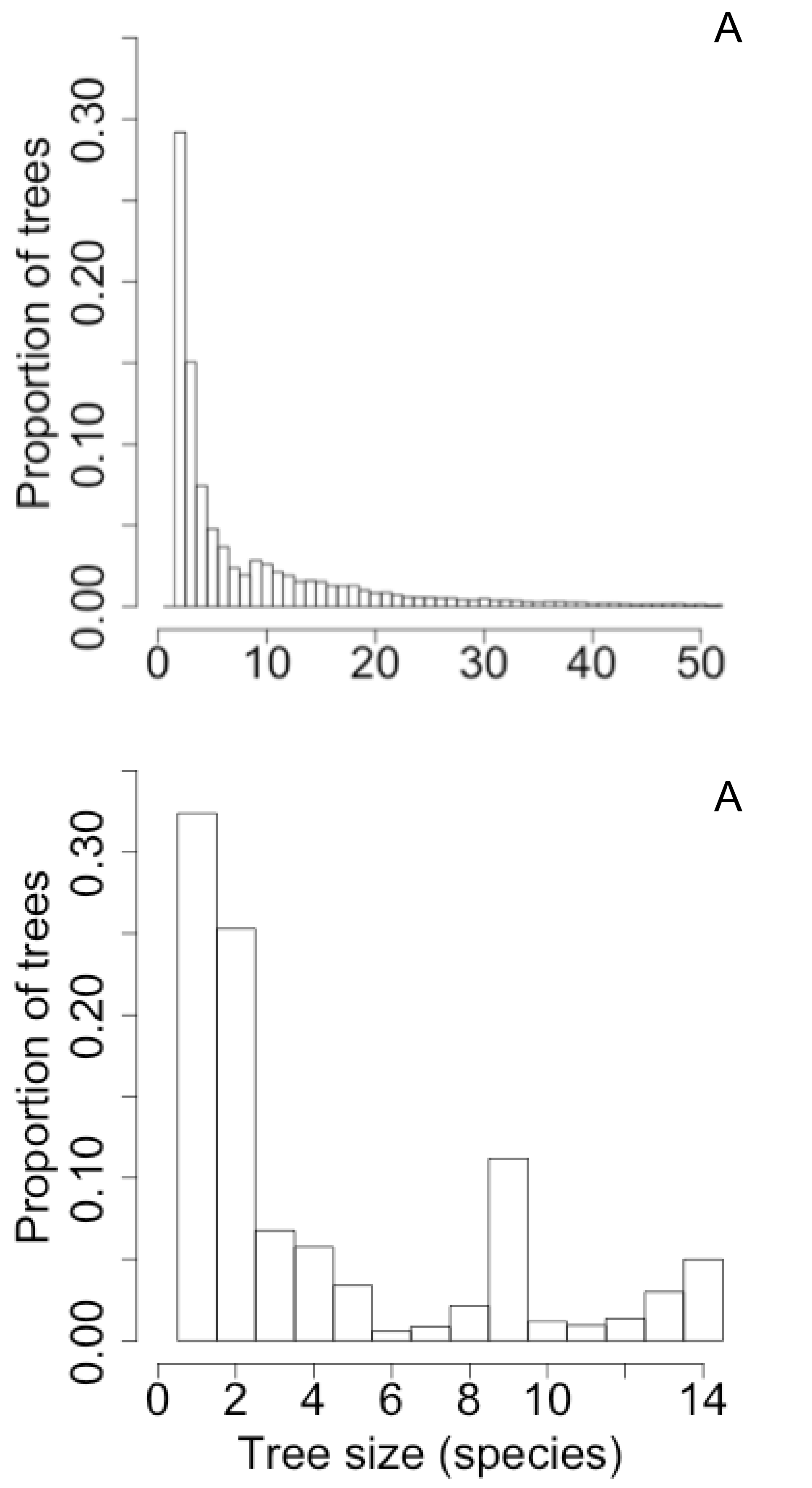


69-member BTB/POZ family showing extensive rice-specific expansion



Tree Sizes

Characteristics of gene family clusters. A. Distribution of cluster sizes with respect to number of genes. B. Distribution of cluster sizes with respect to number of species per tree.



Data

Protein-coding gene predictions from eight fully-sequenced plant species were supplemented with similar data from five fungi/metazoan species. Predictions from a second Oryza sativa cultivar (Indica Group) added to a total species count of fourteen, and a total gene count of 402,192. The EnsemblCompara GeneTree method produced a total of 27,031 individual GeneTrees comprising of a total of 687,302 nodes; 357,182 leaves representing the same number of genes (88.8% of input genes), 180,383 nodes representing gene speciation events, and 149,741 nodes representing gene duplication events.

Species

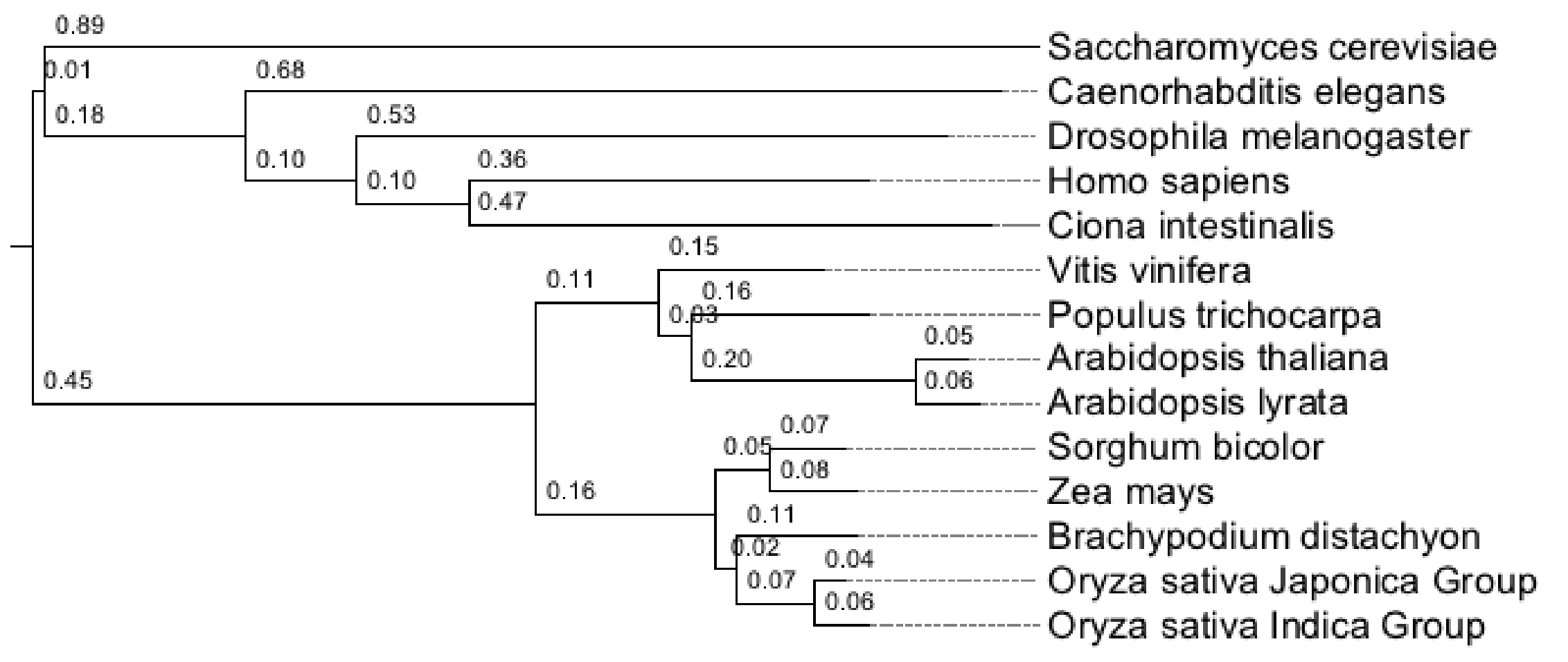
Clade	Species	Total Genes	Genes in trees	Genes in single-species tree	Singleton Genes
Mono-cotyledon Plants	*Oryza sativa* (Japonica rice)	57,995	51,359 (90%)	11,254 (19%)	6,636 (11%)
	Non TE-genes	*41,775*	*35,793 (86%)*	*1,969 (4%)*	*5,982 (14%)*
	Annotated TE-genes	*16,220*	*15,566 (96%)*	*9,285 (57%)*	*654 (4%)*
	Oryza sativa (Indica rice)	38,861	35,071 (89%)	511 (1%)	3,790 (10%)
	Brachypodium distachyon (false brome)	25,532	24,564 (96%)	384 (2%)	968 (4%)
	Sorghum bicolor (sorghum)	34,496	32,730 (95%)	1,584 (5%)	1,766 (5%)
	Zea mays (maize)	32,540	30,258 (93%)	1,025 (3%)	2,282 (7%)
Eudi-cotyledon Plants	*Arabidopsis thaliana* (thale cress)	31,280	29,550 (94%)	2,468 (8%)	1,730 (6%
	Non TE-genes	*27,379*	*26,030 (95%)*	*244 (1%)*	*1,349 (5%)*
	Annotated TE-genes	*3,901*	*3,520 (90%)*	*2,244 (63%)*	*381 (10%)*
	Arabidopsis lyrata (lyrate rockcress)	32,667	30,136 (92%)	1,839 (6%)	2,531 (8%)
	Populus trichocarpa (poplar)	38,449	33,903 (88%)	3,483 (9%)	4,546 (12%)
	Vitis vinifera (grape)	30,434	26,794 (88%)	1,979 (7%)	3,640 (12%)
Fungi/ Metazoa	*Homo sapiens* (human)	22,294	18,774 (84%)	3,610 (16%)	3,520 (16%)
	Ciona intestinalis (sea squirt)	14,180	11,182 (79%)	1,930 (14%)	2,998 (21%)
	Drosophila melanogaster (fruit fly)	14,141	11,223 (79%)	2,084 (15%)	2,918 (21%)
	Caenorhabditis elegans (nematode)	20,158	15,415 (76%)	6,775 (34%)	4,743 (24%)
	Saccharomyces cerevisiae (yeast)	6,698	4,103 (61%)	989 (15%)	2,595 (39%)

Phylogenetic Context

Phylogeny	Taxon	Trees with taxon	Nodes at taxon	Average nodes per Tree	% Dupl-ication Nodes	Species Intersect-ion Score
,--->	*Oryza sativa* Japonica	14,310	66,666	4.66	22.96	1.00
,-+	*Oryza sativa*	11,100	36,273	3.27	18.56	0.84
| `--->	*Oryza sativa* Indica	11,735	37,572	3.20	6.66	1.00
,-+	BEP clade	7,548	19,217	2.55	3.44	0.62
| `---->	*Brachypodium distachyon*	7,953	28,195	3.55	12.88	1.00
,-+	Poaceae	8,689	35,043	4.03	33.42	0.67
| | ,----->	*Sorghum bicolor*	9,026	39,526	4.38	17.19	1.00
| `-+	Andropogoneae	7,532	22,355	2.97	12.91	0.66
| `----->	*Zea mays*	8,281	38,127	4.60	20.64	1.00
,-+	Magnoliophyta	7,505	23,719	3.16	39.68	0.63
| | ,--->	*Arabidopsis thaliana*	8,793	33,971	3.86	13.01	1.00
| | ,-+	Arabidopsis	8,151	32,470	3.98	29.31	0.89
| | | `--->	*Arabidopsis lyrata*	8,849	35,074	3.96	14.08	1.00
| | ,-+	Rosids	6,177	11,765	1.90	4.56	0.59
| | | `----->	*Populus trichocarpa*	8,552	49,174	5.75	31.06	1.00
| `-+	Core eudicotyledons	7,056	22,066	3.13	28.23	0.50
| `------->	*Vitis vinifera*	7,749	35,496	4.58	24.52	1.00
+	Eukaryota	3,358	5,765	1.72	25.92	0.49
| ,--->	*Homo sapiens*	6,393	27,981	4.38	32.90	1.00
| ,-+	Chordata	4,142	6,349	1.53	10.11	0.71
| | `--->	*Ciona intestinalis*	5,068	15,040	2.97	25.65	1.00
| ,-+	Coleomata	4,371	6,815	1.56	13.21	0.57
| | `----->	*Drosophila melanogaster*	5,189	15,063	2.90	25.49	1.00
| ,-+	Bilateria	3,882	7,725	1.99	29.45	0.56
| | `------->	*Caenorhabditis elegans*	5,236	23,668	4.52	34.87	1.00
`-+	Fungi/Metazoa	2,158	2,398	1.11	9.55	0.41
`--------->	*Saccharomyces cerevisiae*	2,838	5,344	1.88	23.22	1.00

Consensus Tree

Scaled consensus species tree from BASEML on concatenated proteins from 2845 individual trees touching all species of interest


0.89
0.01
0.18
0.68
0.10
0.53
0.10
0.36
0.47
0.45
0.11
0.15
0.16
0.03
0.20
0.05
0.06
0.16
0.05
0.07
0.08
0.11
0.02
0.07
0.04
0.06
Saccharomyces cerevisiae
Caenorhabditis elegans
Drosophila melanogaster
Homo sapiens
Ciona intestinalis
Vitis vinifera
Populus trichocarpa
Arabidopsis thaliana
Arabidopsis lyrata
Sorghum bicolor
Zea mays
Brachypodium distachyon
Oryza sativa Japonica Group
Oryza sativa Indica Group