Gramene GeneTrees:

A comprehensive phylogenomics database in plants and other model Eukaryotes

William Spooner[1], Joshua C. Stein[1], Sharon Wei[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

Since the completion of the Arabidopsis thaliana genome in 2000, more than 20 plant genomes have been sequenced, with the number set to increase rapidly in the coming years. It is now becoming possible to apply phylogenomics-based techniques from metazoan biology to the plant arena. We have applied an automated methodology, EnsemblCompara GeneTrees, to proteins predicted from a wide range of genomes to develop the first comprehensive plant phylogenomics resource. This consists of protein-level phylogenetic trees between twelve whole genomes; four dicotyledon plants (Arabidopsis lyrata, Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera), three monocotyledon plants (Oyrza sativa Japonica Group, Oryza sativa Indica Group, Sorghum bicolor), and five model metazoa/fungi (Caenorhabditis elegans, Ciona intestinalis, Drosophila melanogaster, Homo sapiens, Saccharomyces cerevisiae). Validation of our data through comparison data with similar resources and well-studied gene families suggests accurate and consistent results both for summaries of the database as a whole and for individual example trees. The GeneTrees form a component of the Gramene database (http://www.gramene.org), an established resource for comparative plant genomics, and form a useful platform for the study of plant molecular evolution as well as functional annotation of newly sequenced genomes.

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.

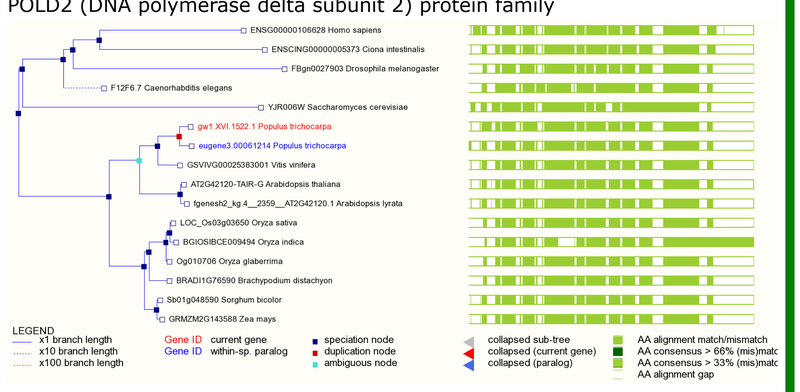
References

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..

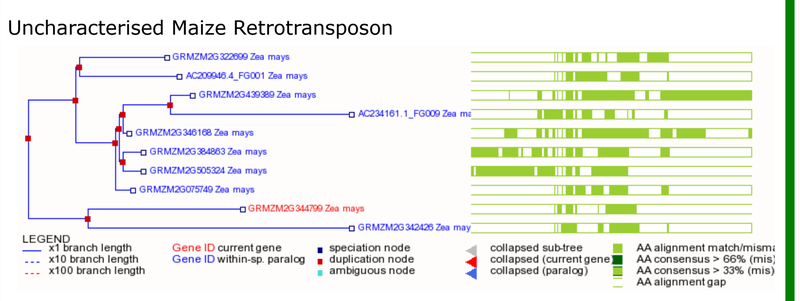
Visualisation

Example GeneTrees

POLD2 (DNA polymerase delta subunit 2) protein family

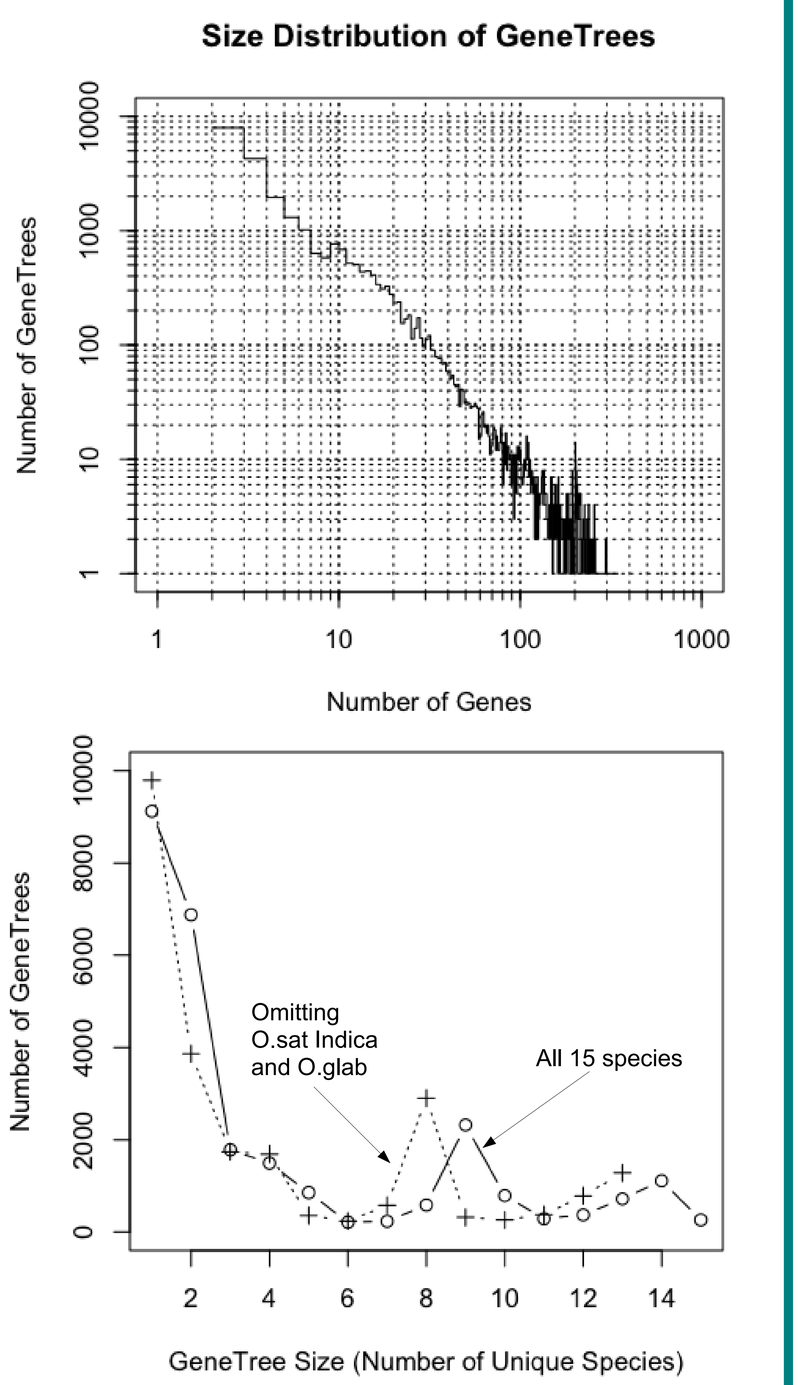


Uncharacterised Maize Retrotransposon



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

Tree Sizes



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) and an incomplete assembly from Oryza glaberrima added to a total species count of fifteen, 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%)
	Oryza glaberrima (African rice)	2,467	2,120 (86%)	81 (3%)	347 (14%)
	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
,-+	Oryza	1,302	1,997	1.53	17.68	0.55
| `--->	*Oryza glaberrima*	1,395	2,448	1.75	13.40	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