Methodology

The clustering was performed on the protein-coding gene using TribesMCL [1]. This software uses a Markov cluster (MCL) algorithm for grouping proteins into families based on a pre-computed sequence pairwise similarity matrix. We ran TribeMCL with different inflations from less to more stringent thresolds (ie. 1.2, 2, 3 and 5) that correspond to level 1, 2, 3 and 4 in the website [2]. The Pairwise similarity matrix was obtained by running Protein-Protein BLAST (v2.2.24+) with an e-value of 1e-5

Cluster Annotation

Clusters are checked by human curators using a dedicated interface. We add annotations, including family names defined via a consensus from existing gene and protein pattern annotations (e.g. UniProt-SwissProt, InterPro, Pirsf, Kegg, GO) for the sequences composing the clusters. The tool sums up high quality annotations available in external databases for protein sequences of a cluster. Some statistics have been made to spot clusters with specific InterPro family motifs. However, we do not check each member of the clusters at this stage. Annotation and analyses are a ongoing process. To check the curation status, please look at the signs.

Graphical signs in GreenPhylDB

Gene family lists

The clustering step and addtional annotation steps allowed us to define various lists of gene family.

- **Annotated gene family list:**
  Gene families or subfamilies manually annotated by an annotator and validated by the administrator. More information about our annotation

strategy is presented **here**. Each "validated" family is classified using confidence levels presented above (high-normal-unknown-suspicious-clustering error)

- **Species specific list**
  Gene families containing sequences from only one species after the clustering step.

- **Phylum specific list including species-specific families**
  Gene families containing sequences that belong to the same phylum including species-specific families underlying this specific phylum.

- **Phylum specific list including species-specific families**
  Gene families containing sequences that belong to the same phylum excluding species-specific gene families.

- **Transcription Factor list**
  Gene families identified as transcription factors
- **Plant specific family list**
  Gene families do not showing any similarity with the other major kingdom: Archea, Bacteria and Eukaryote (excluding plants).
  To define this list, 10 representative (using CD HIT software with ajusted parameters for each family) gene sequences from each family were submitted to BLAST (e-value: 1e-5) against the reference sequences from **NCBI RefSeq** (release 56: fungi, invertebrate, microbial, protozoa, vertebrate_mammalian, vertebrate_other, viral). Only families with no match were tagged as "plant specific".

- **GO Browser**

This browser allows to search gene families using a subset of the GO term classification (**v1.2**). GO terms are representative to gene families if one of the sequence matches an UniProt or a **InterPro** domain linked to a GO term. The rules are defined as follow:
If a gene family matches with at least one UniProt, or if 60 % of gene family members contains the same InterPro domain (arbitrary fixed thresold to define IPR specifc families), then this gene family is flagged with the corresponding GO annotation.

Phylogenetic analyses method

We developed a phylogenomics pipeline for ortholog inference. We validated the full procedure using test sets of orthologs and paralogs for *Oryza sativa* and *Arabidopsis thaliana* to demonstrate that this method outperforms pairwise methods for ortholog predictions.

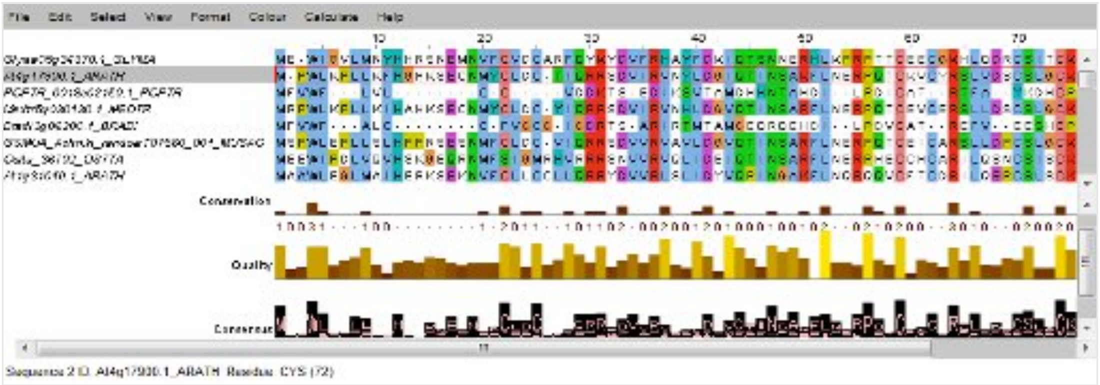
Filtering

Before processing annotated clusters, we filter alternative splices keeping the longest splice form.

Multiple Alignment

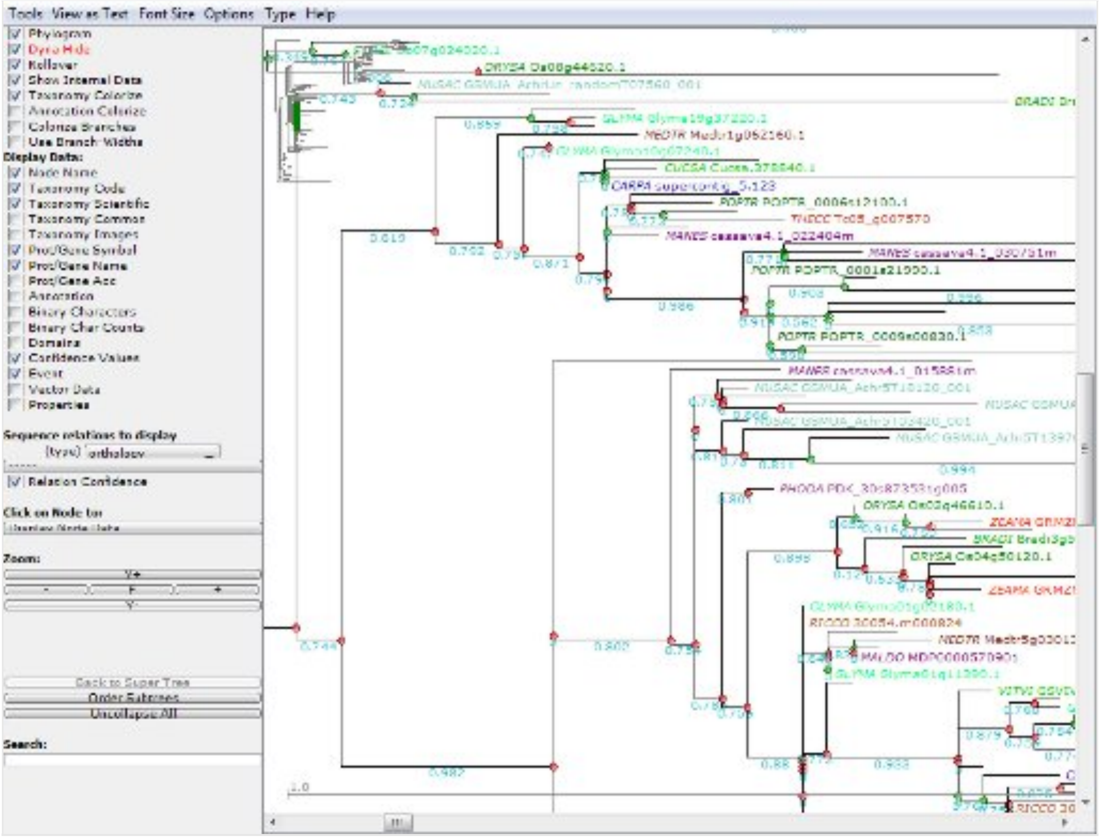
Multi-alignment is one of the major steps in phylogenomic construction. The objective is to identify and align the characteristic domains of the gene families.They were generated using the MAFFT software **[8]**. Different parameters are applied according to the size of the cluster because MAFFT offers a range of multiple alignment methods including alignment of a very large number of sequences, a feature needed either for very large multigene families or when a large number of species is employed.

Then, we applied a masking procedure **[9]** to the optimized alignment to detect and remove amino acid columns/ positions containing either no or a low phylogenetic signal. Cluster alignements may be visualized directly from the website with Jalview **[14]**.



Tree construction

We use PhyML software **[10]** (NNI + SH-Like) with aLRT as branch support. PHYML is one of the fastest maximum-likelihood tree reconstruction methods for generation of large trees with an acceptable CPU computing time. PHYML first constructs a BioNJ tree using the Neighbor- Joining tree algorithm and then optimizes this tree to improve the likelihood at each iteration. Trees can be visualized from the website using the Archeopteryx applet **[12]**.



☑ Phylogram
☑ Dyna Hide
☑ Rollover
☑ Show Internal Data
☑ Taxonomy Colorize
☐ Annotation Colorize
☐ Colorize Branches
☐ Use Branch Widths

Display Data:

☑ Node Name
☑ Taxonomy Code
☑ Taxonomy Scientific
☐ Taxonomy Common
☐ Taxonomy Images
☑ Prot/Gene Symbol
☑ Prot/Gene Name
☐ Prot/Gene Acc
☐ Annotation
☐ Binary Characters
☐ Binary Char Counts
☐ Domains
☑ Confidence Values
☑ Event
☐ Vector Data
☐ Properties

Sequence relations to display
(type) orthology

☑ Relation Confidence

Click on Node to:
Display Node Data

Tree rooting

Phylogenetic trees are rooted using RAPGreen v54[11] and a species tree generated from the NCBI Taxonomy.

Ortholog inference

We used a improved version of RAP (called Rap-Green) for the gene tree reconciliation and to ortholog/paralog predictions

References

1. Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-1584 (2002).
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215(3):403-410.

3. Zdobnov E.M. and Apweiler R. "InterProScan - an integration platform for the signature-recognition methods in InterPro" Bioinformatics, 2001, 17(9): p. 847-8.
4. Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
5. Timothy L. Bailey and Michael Gribskov, "Combining evidence using p-values: application to sequence homology searches", Bioinformatics, Vol. 14, pp. 48-54, 1998.
6. Schneider M, Bairoch A, Wu CH, Apweiler R. Plant protein annotation in the UniProt Knowledgebase.Plant Physiol. (2005) 138:59-66
7. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M.; From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354-357 (2006).
8. Katoh K, Kuma K, Toh H, Miyata T: MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic acids research 2005, 33(2):511-518
9. Salvador Capella-Gutierrez; Jose M. Silla-Martinez; Toni Gabaldon. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009 25: 1972-1973.
10. Guindon S, Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003, 52(5):696-704.
11. Dufayard J-F., Duret L., Penel S., Gouy M., Rechenmann F. and Perriere G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases, Bioinformatics, 21 (11): 2596-2603, 2005.
12. Zmasek CM, Eddy SR: ATV: display and manipulation of annotated

phylogenetic trees. Bioinformatics (Oxford, England) 2001, 17(4):383-384.

13. Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009) "Jalview Version 2 - a multiple sequence alignment editor and analysis workbench" Bioinformatics
14. Salse,J., Abrouk,M., Murat,F., Quraishi,U.M. et Feuillet,C. (2009) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. Brief Bioinform, 10, 619-630, 10.1093/bib/bbp037.