GRAMENE COMPARA: A PHYLOGENOMICS RESOURCE FOR PLANTS

Sharon Wei[1], Zhenyuan Lu[1], William Spooner[1], Joshua C. Stein[1], Andy Yates[3], Paul Kersey[3], Doreen Ware[1,2]
1 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
2 USDA ARS NAA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY, USA 148533
3 EMBL, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD UK

http://www.gramene.org
http://plants.ensembl.org

European Bioinformatics Institute
Cold Spring Harbor Laboratory
Oregon State University
Cornell University

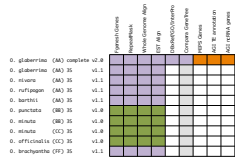
Abstract

The integration of genome annotation with evolutionary analysis, often referred to as phylogenomics, is a powerful strategy in the study of gene structure and function, and is a compelling motivation for acquiring complete genome sequences. The Gramene Project (www.gramene.org) in collaboration with Ensembl Plants (http://plants.ensembl.org) provides a comprehensive platform for comparative genomics in plants, utilizing the Ensembl Compara pipelines and database structure. The site offers data and visualizations of whole genome alignments, synteny analysis, phylogenetic trees, and ortholog/paralog designations. Release 34 includes the whole genomes of six monocots (rice japonica, rice indica, African rice, sorghum, Brachypodium, and maize), five dicots (Arabidopsis, A. lyrata, grape, soybean and poplar), two basal land plants (Physcomitrella, and Selaginella) and the green alga Chlamydomonas. Through collaboration with AGI (http://www2.genome.arizona.edu/), Gramene also hosts partial genomes of seven non-cultivated members of the Oryza genus. Comparative analyses include gene tree construction with ortholog/paralog inference, pairwise synteny maps based on phylogenetically-determined orthologs, 8-way multi-species whole genome alignments with ancestor reconstruction using the Enredo/Pecan/Ortheus pipeline and pairwise whole genome alignments using blastz-chain-net pipeline. These data are fully integrated with other Gramene resources, including gene and protein-level annotations, GO ontology, genome browsers, diversity data, and pathways. In addition, Rice japonica, Arabidopsis, Physcomitrella and Chlamydomonas are included as plant representatives in the Clade Compara to offer a broad view of homologous relationships from across a wide range of taxonomic groups from eukaryotes to prokaryotes. These data are accessible through genome browser, via the public MYSQL database, and ftp flat file downloads. We describe details of this resource and demonstrate its use in multiple applications, including the definition of duplication events, large and small-scale rearrangements, annotation inconsistencies, and comparison of gene-family diversity across species. The availability of this platform provides unique opportunities to elucidate the evolutionary history of flowering plants.

Genome-scale Datasets in Gramene

Clade	Species	Genome	Gene Anotation	Variome	Pathways
Monocots	Oruza sativa Japonica (Asian rice)	■	■		■
	Oryza sativa Indica (Asian rice)	■			
	Oryza glaberrima (African rice)	■			
	Brachypodium distachyon	■	■		■
	Sorghum bicolor	■			
	Zea mays (maize)	■	■		■
Eudicots	Arabidopsis thaliana	■	■		■
	Arabidopsis lyrata	■			
	Populus trichocarpa (poplar)	■	■		■
	Vitis vinifera (grape)	■	■		
	Glycine max (soybean)	■			
Bryophytes	Physcomitrella patens (moss)	■			
lycophytes	Selaginella moellendorffii (Spikemoss)	■			
Chlorophyta	Chlamydomonas.reinhardtii (green algae)	■			

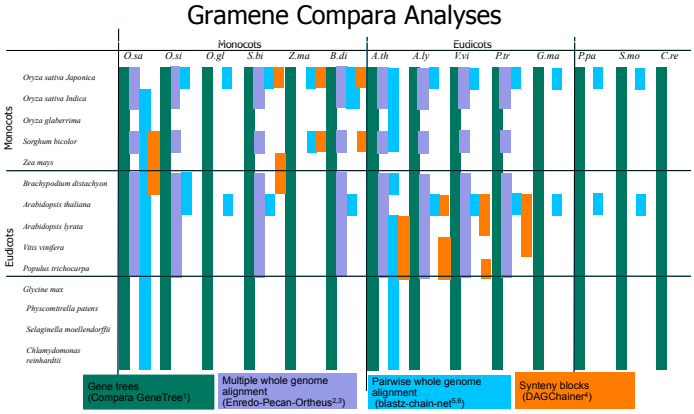
Partial Genomes with the rice genus (Oryza)



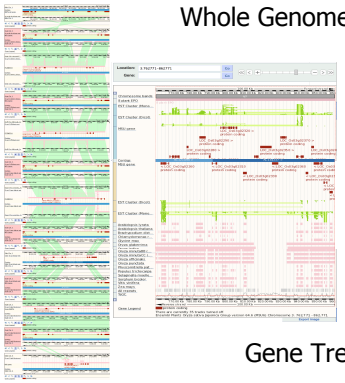
CSHL analysis complete (Gramene 33)
Repeat for updated assembly (Gramene 34)
Community annotation (AG//MIPS)
Oryza-centered Compara Trees (interim release)

Gramene 34 hosts 14 whole genomes and 8 partial genomes of chromosome 3 short arm from Oryza Genome Evolution project (previously known as OMAP http://www.omap.org/). All of the 14 whole genomes have community gene annotations, most of them have fgenesh prediction, protein domain annotation, cross references to other databases, GO annotations, some have variation data and pathways. The OGE partial genomes have fgenesh gene predictions. They have compara gene tree built within the oryza clade, and whole genome alignments against the same reference genome Oryza sativa Japonica.

Gramene Compara Analyses

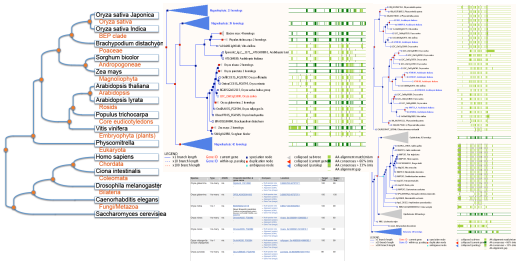


Whole Genome Alignment



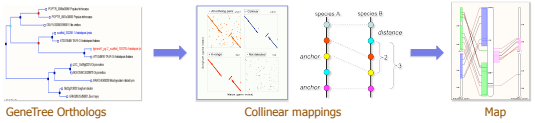
Blastz-net[4,5] alignments are provided for closely related pairs of species. The alignments are the results of post-processing the raw BlastZ results. In the first step, original blocks are chained according to their location in both genomes. The netting process chooses for the reference species the best sub-chain in each region. In gramene34, each full genome was aligned against Rice Japonica and Arabidopsis thaliana, the OGE chr3S partial genomes were aligned against Rice Japonica only. The alignments can be viewed on Multi-species page (left) and Location page (right). Genome rearrangements can be detected from the Multi-species view, in this example, there is an inversion between Oryza sativa Japonica and Brachypodium.
The EPO[2,3] multiple whole genome alignment analysis was performed using 8 genomes as a pilot run at Gramene32. The 8 genomes are *O.sativa Japonica, O.sativa Indica, S.bicolor, B. distachyon, A.thaliana, A.lyrata, V.vinifera, P. trichocarpa.*

Gene Tree



The middle image is the phylogenetic tree for DL (Drooping leaf) gene. The gramene34 gene trees were built using the twenty two core gramene databases, with the addition of five ensembl databases including Caenorhabditis elegans (nematode worm), Ciona intestinalis (sea squirt), Drosophila melanogaster (fruit fly), Homo sapiens (human) and Saccharomyces cerevisiae (yeast) to provide taxonomic depth. The left species tree gives a rough overview of the taxonomic coverage of Gramene gene trees. From a total of 771,369 input proteins, 50,440 GeneTree families were constructed, comprising 688,056 individual genes. In this example tree, number of genes is 129, number of duplication nodes is 56.
The right example is a Pan-taxonomic Gene tree build in collaboration with EnsemblGenome to expand the taxonomic coverage. In this example, the number of genes is 203, the number of duplication nodes is 49. The inferred orthologs and paralogs can be view in a table format.

Synteny Detection



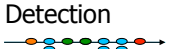
Syntenic genes and regions were calculated using gene orthologs predicted from Gramene Protein Compara analysis. DAGchainer[4] was used to identify colinear chains. To increase sensitivity, chaining was created on the basis of gene order, excluding positions of non-orthologous genes,rather than using gene coordinates. Chains were required to have at least five colinear genes with no more than ten intervening genes between neighbors. Resulting gene-pairs were classified as"syntenic:colinear". Next, we searched for additional syntenies among non-colinear genes to account for small-scale rearrangements and assembly errors. This also allowed members of tandemly duplicated clusters to be classified as syntenic, as they were often missed when strict colinearity was enforced. For a non-colinear gene in species1, nearest flanking colinear anchors were identified and mapped to the irrespective positions in species2. If the non-colinear ortholog in species2 was located within five genes of either anchor, then the non-colinear pair of genes was classified as "syntenic:in-range". Most of these occurred within already defined colinear chains. The coordinates of colinear chains was used to define syntenic regions.

Automated Detection of Split Genes

Species	Putative SplitGenes
Populus trichocarpa	1181
Sorghum bicolor	1087
Oryza sativa Japonica	916
Vitis vinifera	520
Oryza sativa Indica	365
Zea mays	280
Arabidopsis lyrata	202
Arabidopsis thaliana	137
Brachypodium distachyon	101

An application of gramene comparative genomics data to QC gene models; Non-overlapping protein alignments for a pair of paralogs in a gene tree is evidence for a gene split.
Whole genome alignments can be used as additional evidence for annotations that are inconsistent between species. If the split fragments are within 1Mbase, on the same genomic strand, then the paralogs are classified as contiguous. If they are on different regions, then the split is putative. Contiguous splits may indicate incorrect gene annotation, whereas putative splits are characteristic of genome mis-assembly. The table shows counts of putative split genes in various species; these range from 100 in brachypodium to 1000 in poplar.

Tandem Duplicate Detection



Species	Clusters	Genes	Largest	Function
Rice (J)	2519	7054	24	Phytosulfokine receptor-like
Sorghum	2182	5927	19	Chalcone-stilbene synthase like
Maize	1871	4564	22	Domain of unknown function
Arabidopsis	1738	4581	28	ECA1 gametogenesis related family

An application combining gramene gene tree with genomic locations of the member genes to detect Tandem Duplication genes. We can interrogate paralogs to infer tandem duplicates. A tandem is called if adjacent paralogs have no more than 2 intervening unrelated genes. The numbers of tandem clusters and genes for four species are shown. We typically see 2000 clusters containing 5000 genes for a given species. The largest clusters are often over 20 genes in length, and are from a variety of protein families. It is possible, therefore, to classify duplication nodes in the gene trees as those resulting from tandem duplication events and, through analysis of syntenic relationships, of those resulting from segmental duplication or ploidy.

References

1. Vilella AJ, Severin J, Ureta-Vidal A, Durbin R, Heng L, Birney E. *EnsemblCompara GeneTrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates. Genome Research 2009. 19: 327-335*
2. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. *Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18:1814-1828*
3. Paten B, et al. *Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 2008;18:1829-1843*
4. Haas BJ, et al. 2004." *DAGchainer: a tool for mining segmental genome duplications and synteny " Bioinformatics 20(18):3643-6*
5. Schwartz S et al., *"Human–Mouse Alignments with BLASTZ Genome Res.;13(1):103-7.*
6. Kent WJ et al., *"Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes" Proc Natl Acad Sci U S A., 2003;100(20):11484-9*

The Gramene-Ensembl Plants collaboration

Gramene and Ensembl Plants are collaborating to maintain a common set of reference databases for plant genomic sequence and annotation, integrating data for important plant species generated from across the globe.

weix@cshl.edu
Sharon Wei
516-367-8828

Cold Spring Harbor Laboratory