JGI Plant Genomics Gene Annotation Pipeline

Shengqiang Shu[1], David M. Goodstein[1], David Hayes[1], Therese Mitros[2] & Dan Rokhsar[1]

[1]Lawrence Berkeley National Laboratory/DOE Joint Genome Institute, Walnut Creek, California 94598, USA.

[2]University of California – Berkeley, Berkeley, California

**To whom correspondence should be addressed*: S. Shu (sqshu@lbl.gov @lbl.gov)

July 2014

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

JGI Plant Genomics Gene Annotation Pipeline

Shengqiang Shu

Richard Hayes, Therese Mitros, David Goodstein and Daniel Rohksar

Abstract

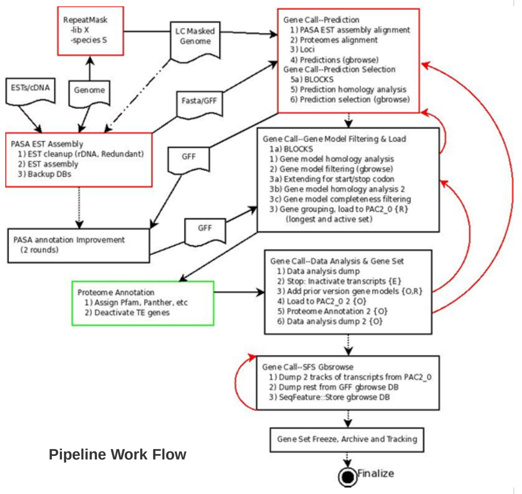
JGI plant genomics gene annotation pipeline is a streamlined pipeline integrating multiple components. It provides one or more optional redo operations in each component to allow parameter tweaking to suit each genome characteristics so a naïve or experienced user can make high quality gene annotation in a timely manner. It has been used for gene annotation/re-annotation of many JGI flagship plant genomes and other genomes, producing stable and high quality gene models.

Methods

Prerequisites: genome fasta, transcriptome fasta, related species proteomes, repeat library

Steps (detailed work flow shown below):

1) Run RepeatModeler to get repeat library if not available
2) RepeatMask genome with repeat library
3) Run PASA to assembly transcripts
4) Run BLASTX and EXONERATE on homolog proteomes
5) Define locus based on EXONERATE hits and transcript assembly alignments
6) Make initial gene predictions for each locus by GenomeScan, FGENESH+ and FGENESH_EST
7) Select one best prediction for each locus
8) Run PASA on the selected predictions for annotation improvement like adding UTRs and alternative transcripts
9) Filter gene models
10) Run protein domain analysis on the filtered gene models and deactivate TE genes
11) Build browser database/datastore for visualization


RepeatMask
-lib X
-species S
LC Masked Genome
ESTs/cDNA
Genome
Fasta/GFF
PASA EST Assembly
1) EST cleanup (rDNA, Redundant)
2) EST assembly
3) Backup DBs
GFF
GFF
PASA annotation Improvement
(2 rounds)
Proteome Annotation
1) Assign Pfam, Panther, etc
2) Deactivate TE genes
Gene Call--Prediction
1) PASA EST assembly alignment
2) Proteomes alignment
3) Loci
4) Predictions (gbrowse)
Gene Call--Prediction Selection
5a) BLOCKS
5) Prediction homology analysis
6) Prediction selection (gbrowse)
Gene Call--Gene Model Filtering & Load
1a) BLOCKS
1) Gene model homology analysis
2) Gene model filtering (gbrowse)
3a) Extending for start/stop codon
3b) Gene model homology analysis 2
3c) Gene model completeness filtering
3) Gene grouping, load to PAC2_0 {R}
(longest and active set)
Gene Call--Data Analysis & Gene Set
1) Data analysis dump
2) Stop: Inactivate transcripts {E}
3) Add prior version gene models {O,R}
4) Load to PAC2_0 2 {O}
5) Proteome Annotation 2 {O}
6) Data analysis dump 2 {O}
Gene Call--SFS Gbsrowse
1) Dump 2 tracks of transcripts from PAC2_0
2) Dump rest from GFF gbrowse DB
3) SeqFeature::Store gbrowse DB
Gene Set Freeze, Archive and Tracking
Finalize

Pipeline Work Flow

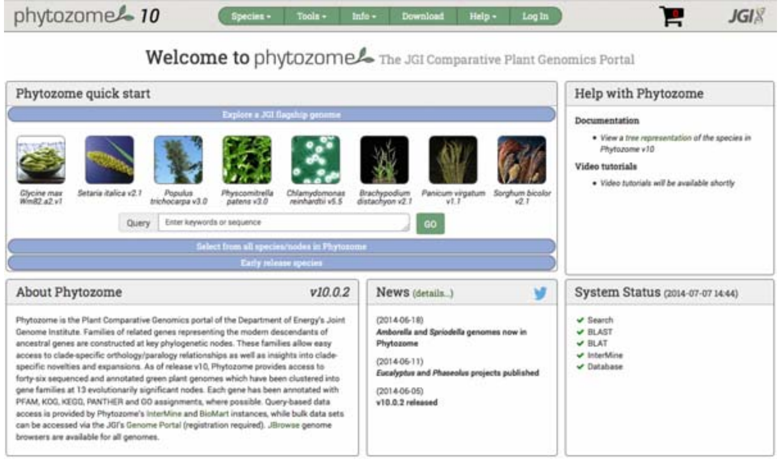
Contact: S. Shu (sqshu@lbl.gov)

Introduction

Plant genomes vary in size and are highly complex with a high amount of repeats, genome duplication and tandem duplication. Gene encodes a wealth of information useful in studying organism and it is critical to have high quality and stable gene annotation. Thanks to advancement of sequencing technology, many plant species genomes have been sequenced and transcriptomes are also sequenced. To use these vastly large amounts of sequence data to make gene annotation or re-annotation in a timely fashion, an automatic pipeline is needed. JGI plant genomics gene annotation pipeline, called integrated gene call (IGC), is our effort toward this aim with aid of a RNA-seq transcriptome assembly pipeline. It utilizes several gene predictors based on homolog peptides and transcript ORFs. See Methods for detail.

Here we present genome annotation of JGI flagship green plants produced by this pipeline plus Arabidopsis and rice except for chlamy which is done by a third party. The genome annotations of these species and others are used in our gene family build pipeline and accessible via JGI Phytozome portal whose URL and front page snapshot are shown below.

http://phytozome.jgi.doe.gov/pz/portal.html


phytozome 10
Species
Tools
Info
Download
Help
Log In
JGI
Welcome to phytozome The JGI Comparative Plant Genomics Portal
Phytozome quick start
Explore a JGI flagship genome
Glycine max Wm82.a2.v1
Setaria italica v2.1
Populus trichocarpa v3.0
Physcomitrella patens v3.0
Chlamydomonas reinhardtii v5.5
Brachypodium distachyon v2.1
Panicum virgatum v1.1
Sorghum bicolor v2.1
Query
Enter keywords or sequence
GO
Select from all species/nodes in Phytozome
Early release species
Help with Phytozome

Documentation

- View a tree representation of the species in Phytozome v10

Video tutorials

- Video tutorials will be available shortly

About Phytozome v10.0.2

Phytozome is the Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute. Families of related genes representing the modern descendants of ancestral genes are constructed at key phylogenetic nodes. These families allow easy access to clade-specific orthology/paralogy relationships as well as insights into clade-specific novelties and expansions. As of release v10, Phytozome provides access to forty-six sequenced and annotated green plant genomes which have been clustered into gene families at 13 evolutionarily significant nodes. Each gene has been annotated with PFAM, KOG, KEGG, PANTHER and GO assignments, where possible. Query-based data access is provided by Phytozome's InterMine and BioMart instances, while bulk data sets can be accessed via the JGI's Genome Portal (registration required). JBrowse genome browsers are available for all genomes.

News (details...)

(2014-06-18)
Amborella and Spriodella genomes now in Phytozome

(2014-06-11)
Eucalyptus and Phaseolus projects published

(2014-06-05)
v10.0.2 released

System Status (2014-07-07 14:44)

✓ Search
✓ BLAST
✓ BLAT
✓ InterMine
✓ Database

Results

Land plant protein coding gene structure is shown to be remarkably stable in term of exon number and exon length, and to some extent intron length as well (shown in far right tables).

Using percentage of gene models with PFAM domain as a yardstick, our automatic pipeline produced results approaching human curated Arabidopsis and many genomes are better than rice, also a well-known annotated genome. Except for chlamy (a non land plant), genome annotation with lower assigned PFAM domain reflects largely on quality of inputs.

	Gene	Trans	%PFAM	G. Size	%Exonic
G. max	56,044	88,647	75.1	979M	9.66
P. trichocarpa	41.335	73,013	74.9	434M	15.33
B. distachyon	31,694	42,868	70.7	272M	19.38
S. italica	35,471	40,599	66.4	406M	12.01
S. bicolor	33,032	39,441	67.3	727M	8.48
P. virgatum	98,007	125,439	55.6	1,698M	7.69
P. patens	26,610	42,392	68.4	473M	10.44
C. reinhardtii	17,741	19,526	51.8	111M	52.03
A. thaliana	27,416	35,386	77.4	120M	34.19
O. sativa	39,049	49,061	62.0	375M	15.21

Results

	Number of Exons				
	25%	Median	Mean	75%	Max
G. max	3	5	6.5	9	78
P. trichocarpa	2	5	6.3	9	76
B. distachyon	2	4	5.8	8	72
S. italica	2	3	5.0	7	60
S. bicolor	2	4	5.3	7	65
P. virgatum	2	3	4.5	6	60
P. patens	2	5	6.3	8	70
C. reinhardtii	4	7	8.9	11	173
A. thaliana	2	4	5.9	8	79
O. sativa	2	3	4.9	7	78

	Exon Length				
	25%	Median	Mean	75%	Max
G. max	90	151	284	330	9,509
P. trichocarpa	89	149	277	322	7,911
B. distachyon	92	158	306	366	9,439
S. italica	91	157	288	345	7,851
S. bicolor	94	173	361	451	14,531
P. virgatum	94	171	316	385	18,512
P. patens	93	162	313	378	10,268
C. reinhardtii	95	156	369	351	12,274
A. thaliana	89	147	262	300	7,761
O. sativa	90	159	318	369	15,363

	Intron Length				
	25%	Median	Mean	75%	Max
G. max	105	225	519	599	18,215
P. trichocarpa	101	180	380	483	10,053
B. distachyon	94	149	419	500	16,886
S. italica	92	136	334	418	6,930
S. bicolor	96	150	480	491	18,859
P. virgatum	92	137	391	422	18,637
P. patens	153	215	274	312	7,589
C. reinhardtii	164	227	173	314	82,837
A. thaliana	86	100	165	168	11,602
O. sativa	96	168	416	501	18,327

References

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res, 31, 5654-5666.

Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2011 <http://www.repeatmasker.org>.

Yeh, R.-F., Lim, L. P., and Burge, C. B. (2001) Computational inference of homologous gene structures in the human genome. Genome Res. 11: 803-816.

Salamov, A. A. and Solovyev, V. V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res 10, 516-22.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.