JGI Plant Genomics Group Annotation Process

Shengqiang Shu, Richard D. Hayes, David M. Goodstein

Introduction

The JGI Plant Genomics annotation process employs an automated system for annotation of complex, highly repetitive plant genomes. Annotation includes both structural annotation of protein-coding genes and potential pseudo-genes, as well as domain identification and multiple classifications of gene translation peptides. The results of the annotation process are available through the Phytozome website (http://www.phytozome.net), which provides tools for genome browsing, comparative analysis and data capture.

Requirement

Annotation requires a set of fasta files of the genome assembly, libraries of appropriate repeat families if available, multi-fasta files of same-species and related-species ESTs/cDNAs, multi-fasta of proteomes from related, already annotated species,

Procedure

Annotation procedes through the following steps: Repeat-masking, EST assembly, alignment of EST assemblies and seed peptides against the genome, loci determination, gene prediction, gene filtering, annotation improvement (reconciliation of predictions with EST evidence, calling of splice variants), at each stage intermediate predictions are available for viewing in Gbrowse, to validate that parameter settings (e.g., maximum allowed intron, required EST coverage for calling alternate splices) are having the expected effects, and to examine outliers.

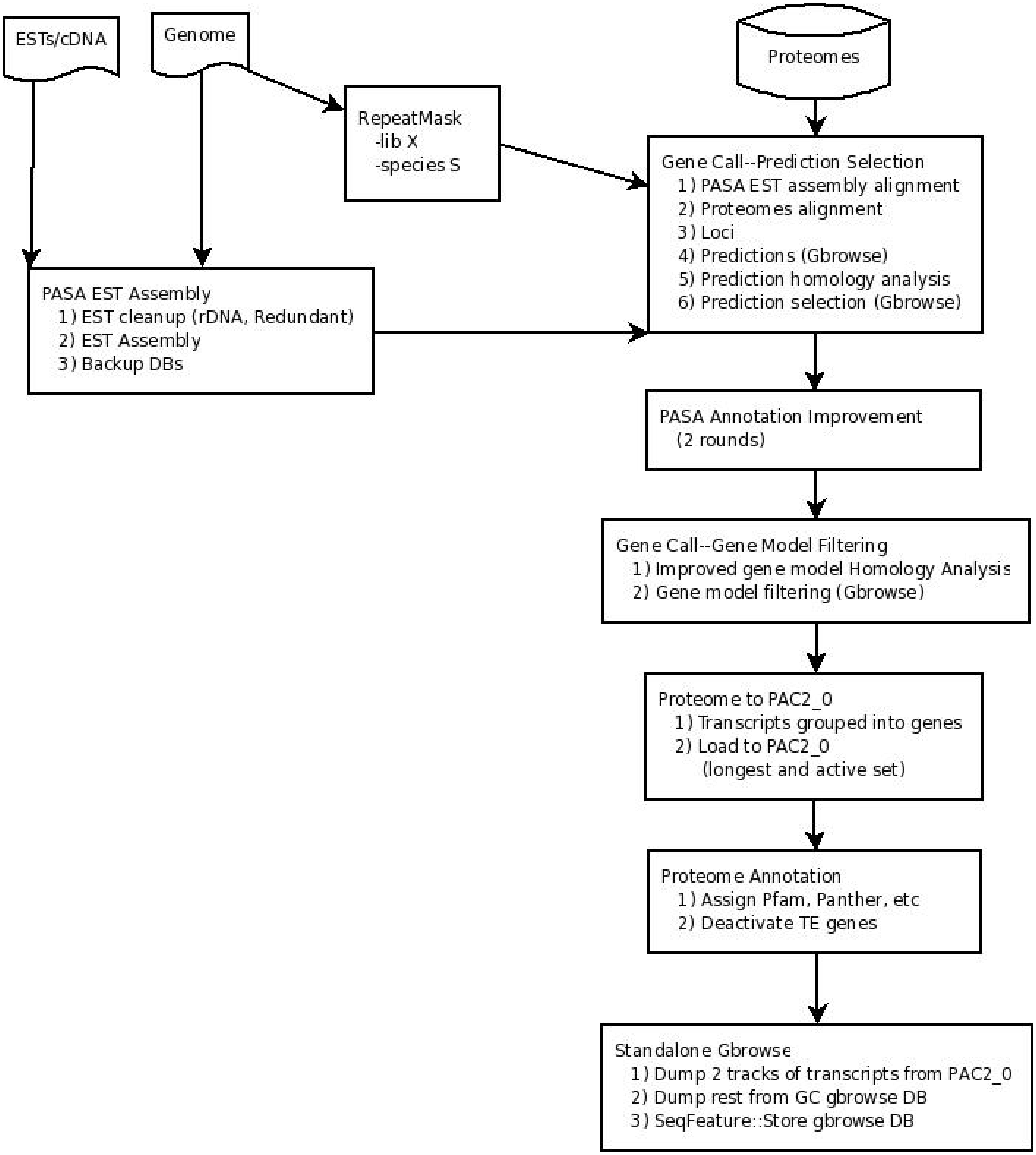
ESTs/cDNA
Genome
RepeatMask
-lib X
-species S
Proteomes
Gene Call--Prediction Selection
1) PASA EST assembly alignment
2) Proteomes alignment
3) Loci
4) Predictions (Gbrowse)
5) Prediction homology analysis
6) Prediction selection (Gbrowse)
PASA EST Assembly
1) EST cleanup (rDNA, Redundant)
2) EST Assembly
3) Backup DBs
PASA Annotation Improvement
(2 rounds)
Gene Call--Gene Model Filtering
1) Improved gene model Homology Analysis
2) Gene model filtering (Gbrowse)
Proteome to PAC2_0
1) Transcripts grouped into genes
2) Load to PAC2_0
(longest and active set)
Proteome Annotation
1) Assign Pfam, Panther, etc
2) Deactivate TE genes
Standalone Gbrowse
1) Dump 2 tracks of transcripts from PAC2_0
2) Dump rest from GC gbrowse DB
3) SeqFeature::Store gbrowse DB

Figure 1. JGI Plant Genomics annotation process

Repeat masking

de novo repeat identification and custom repeat library construction

It is advantageous to mask genomes for known transposable element (TE) sequence prior to an initial annotation, as this removes viral open reading frames and TE-associated enzymes such as integrase and transposase from the annotation pipeline. Yet, the genomes we annotate often lack pre-masked sequence either because they are newly sequenced, or repeat families in a particular species remain uncharacterized. In these cases, we use the RepeatModeler software suite (http://www.repeatmasker.org/RepeatModeler.html) to perform *de novo* repeat family prediction using only the assembled genome as input. RepeatModeler is a software wrapper that performs repeat family consensus sequence predictions using both the RECON (Bao and Eddy, 2002) and RepeatScout (Price *et al*., 2005) algorithms. First, 40 Mbp of the genomic assembly is sampled at random and RECON is run on this sample to produce the first set of predictions. Consensus sequences are built based on BLAST homology, and the resulting predicted repeat families then masked by RepeatMasker (Smit and Green, http://www.repeatmasker.org). A second round of prediction begins with an additional 3 Mbp sampling of the remaining unmasked genome and analysis by RepeatScout and predictions are again masked. RepeatScout rounds are repeated, increasing sampling by 3 times at the start of each round, until the entire genomic assembly sequence has been sampled and tested.

To correct for potential false positives (repetitive domains within, e.g., kinases), the final RepeatModeler predicted consensus sequence dataset is searched for functional PFAM and Panther domains. For a given putative repeat family, if no domains or only known TE domains it will be retained as a “true” repeat. Inversely, if only domains from known FP (false positives) sets are found, the family is rejected and removed from further analysis. Mixtures of TE and FP domain homology, or any other these domains in combination with previously unexamined domain classes require careful manual review. Often, large mosaic repeat consensus predictions will include TE domains clustered in one open reading frame, but have FP or otherwise enzymatic domains in another open reading frame. These mosaics are also removed from the final repeat family dataset to minimize the likelihood of erroneously masking sequence corresponding to enzyme classes of eukaryotic origin that are evolutionarily evolved from viral genes or that might represent degenerate TE found in introns or adjacent to valid eukaryotic coding sequence. Similarly, repeat families with homology only to ambiguous domains, e.g. PF00098 Zinc Knuckle, are also removed. Newly encountered FP or TE domains that can be conclusively included in one dataset or another, based on domain description and a literature search, are included in the curated test datasets in future runs of RepeatModeler.

The final set of predicted repeat family consensus sequence is combined with the *RepBase* set of typical short tandem repeats, e.g. (CAG)n, and low complexity sequence, e.g. AT-rich, to form a complete custom library for genome masking via RepeatMasker (options: -xsmall -gccalc [-lib libfile | -species “speciesName”]).

Gene Prediction

We use a homology-based approach to generate protein-coding gene models in plant genomes. Both EST ORFs and peptide alignments are used as seeds for GenomeScan (Yeh, 2001) and FGENESH (Salamov, 2000).

EST and peptide homolog data set selection

ESTS and cDNA within species and/or from related species are collected from various sources including the JGI sequence production group, NCBI GenBank repository, and collaborators on particular plant genome projects.

Proteomes of representative genomes on the plant tree of life that are well annotated are chosen as homolog seeds. For the annotation of rosid genomes, we use either *Populus trichocarpa* or *Glycine max*, or both (Eurosid 1), *Arabidopsis thaliana* (Eurosid 2), and *Vitis vinifera*. For grasses, we use *Oryza sativa* and *Sorghum bicolor.*

EST alignment and assembly with PASA

EST assemblies within species are made by PASA (Haas, 2003) using 95% identity, and 50% or more coverage depending how continuous genome assembly is. EST assemblies from related species are made by PASA with relaxed identity criteria, typically 80%, and 70% coverage. Regarding 70% coverage in related species PASA, we are more concerned with species transcriptome sequence divergence rather than with losing broken genes due to discontinuous genome assembly.

EST and peptide alignment to genome

EST assemblies are aligned to the soft-masked genome by BLAT (ref) and alignments with greater than 95% identity and 90% coverage are retained. Proteome peptides are aligned to the soft-masked genome by BLASTX (-e 1E-5). For both alignment types, a maximum allowed intron size is enforced. Initially chosen based on values from annotations of related genomes, this setting will often be refined as the annotation progresses.

Gene calling

Both EST assembly alignments and BLASTX peptide alignments are combined to define likely gene loci by joining overlapping alignments on the same strand. The locus region to be used for gene prediction will be extended by (typically) 1Kbp upstream and downstream if possible (i.e., without overlapping other loci), to allow for the possibility of additional terminal exons in a predicted model (compared to homologus peptides in other species) with poor EST support.

Genomic sequence from each locus and EST assembly ORFs (open reading frames) and peptides of the locus are input to GenomeScan, FGENESH+ and FGENESH_EST. We pick the best seed (the best BLASTX hit to locus genomic sequence) from EST ORF translations and peptides for FGENESH+, or the best seed of peptides and FGENESH's protein alignment of EST ORF translations as splicing information for FGENESH_EST.

Peptides derived via translation of gene models predicted by GenomeScan, FGENESH+ and FGENESH_EST are BLASTPed to the original seeding proteome sets. At each locus, the models are sorted by best predicted-peptide-to-seeding-peptide C-score (see below) and alignment coverage.

The best predictions for each locus are selected using multiple positive factors: C-score, peptide homology coverage, transcript splicing and coverage scores by EST assembly alignments, and one negative factor: overlap with repeats.

The selected predicted transcripts are fed into the PASA run where we had EST assemblies made to validate and improve gene models. If we have related species PASA, we do that PASA gene model improvement first and feed the improved gene models to within species PASA. PASA adds alternative transcripts based on EST/cDNA evidence.

Gene filtering

Peptides from PASA improved gene model transcripts are BLASTPed to proteomes used in loci generation to obtain best C-score and homology coverage. Genes will be filtered out if they overlap with repeats for > 20% of CDS, have weak homology, are too short or have internal stop codons (very rare).

Primary and alternate splice form identification

The filtered transcripts are grouped into genes if their CDS overlap on the same strand. The longest CDS transcript is chosen to be primary (ties are broken using overall transcript length). and rest of transcripts in the gene become alternative transcripts.

Gene naming

Transcript names take the form of *PREFIXnnnnnn* where *PREFIX* is a “meaningful” string (related to species name, though the actual abbreviation used will be driven by standards of each particular plant community), and *nnnnnn* is autogenerated and continuous from chromsome 5’ to 3’ and chromosome to chromosome. Gene name is the primary transcript name plus ‘.g’. For genomes where JGI controls the naming conventions, the gene name is in the form of *PREFIXnnnnnn* where *PREFIX* is 3 letters, first letter of genus in uppper case and first 2 letters of the species name, and the transcript name is its gene name plus '.N' where is N is ordinal number of the transcript (if there are splice variants) and "1" is always used for the primary transcript. A numerical gap of between 10 and 100 is included in the *nnnnnn* suffixes between 2 adjacent genes, to provide room for subsequent gene splits, new gene calls based on additional EST evidence, etc.

Proteome Annotation

Standard peptide analysis and classification

The peptides of all protein-coding genes are annotated as follows:

1. PFAM A/B (Sonnhammer, et. al.) domain calls and PANTHER (Guo, et. al.) family classification are assigned via hidden Markov model (HMM) profile scanning using TimeLogic's *Decypher* hardware-accelerated platform.
2. KOG (Koonin, et. al.) classification via NCBI rpsBLAST against the KOG PSSM (position specific scoring matrix) set.
3. E.C (Enzyme Classification) and KO (KEGG Orthology) numbers (http://www.genome.jp/) transferred by mutual-best-hit from a set of plant-specific curated gene sets from KEGG
4. GO (Gene Ontology) terms via pfam2go (ftp://ftp.geneontology.org/pub/go/external2go/pfam2go)

After domain assignment, genes with TE components (>= 30% assigned domains are known TE) are removed from final gene set

Phytozome PlantFAM classification

For peptides derived from genomes that will not be incorporated into core PlantFAM families (i.e., be used in *de novo* construction of gene families), one or more PlantFAM classifications are assigned via two methods:

1. at least one MBH (mutual best hit) between the peptide in question and a member of a given PlantFAM (with evalue < 1e-10 and overlap >= 0.65)
2. HMMscan (HMMER3, Sean Eddy et. al) hits between PlantFAM profiles and the peptide in question with evalue < 1e-25 AND evalue < (1e+20)*evalue(best PlantFAM hit to this peptide).

Implementation

The JGI plant genome annotation process consists of the following components/steps: ab initio repeat annotation if the repeat library does not exist, genome repeat masking, PASA assembly of ESTs/cDNAs and gene model improvement, final gene prediction and selection, and peptide annotation. Each component is implemented in perl scripts that call external publically available modules, as well as custom parsing and filtering modules. The entire process runs within the SAPS pipeline system (also in written in Perl), with parallelism implemented where appropriate.

Additional notes

Definition of C-score

We use the metric “C-score” as a measure of similarity between a protein from one predicted proteome and the proteins from a second predicted proteome. The C-score for protein X in one species and protein Y in a second species (CXY) is defined as the BLAST score of X against Y divided by the best BLAST score for protein X against all of the proteins in species Y. The C-score can be used to detect the presence of both orthologs (defined as mutual best BLAST hits) as well as potential paralogs. If X and Y are mutual best hits, then CXY and CYX will both equal 1. Recent paralogs of X will have a C-score of slightly less than 1 relative to Y; similarly, recent paralogs of Y will have a C-score of slightly less than 1 relative to X.

References

1. 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. http://nar.oupjournals.org/cgi/content/full/31/19/5654[Nucleic Acids Res, 31, 5654-5666].

2. Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0.*1996-2004 <http://www.repeatmasker.org>.

3. 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.

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

5. Smit AFA, and Hubley, R. RepeatModeler. http://www.repeatmasker.org/repeatmodeler.html

6. Bao Z. and Eddy S.R. (2002) Automated *de novo* Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Research*, 12:1269-1276.

7. Price AL, Jones NC, Pevzner PA. (2005) De novo identification of repeat families in large

genomes. *Bioinformatics*. 21 Suppl 1, i351-8.

8. Smit AFA, and Green P. RepeatMasker. http://www.repeatmasker.org

9. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA., A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* 5(2), 2004.

10. Sonnhammer EL, Eddy SR, Durbin R., Pfam: a comprehensive database of protein domain families based on seed alignments, Proteins 29(3), pp 405-20, 1997.

11. Mi H, Guo N, Kejariwal A, Thomas PD., PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways., Nucleic Acids Res 35, pp D247-52, 2007.