Eukaryotic Genome Annotation

Igor Grigoriev, Genome Annotation Group - PGF

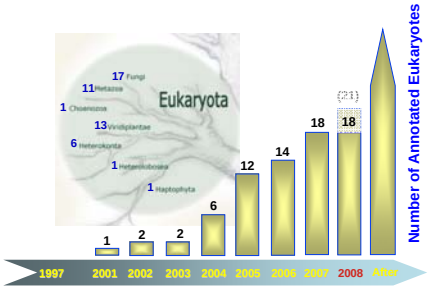
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

U.S. DEPARTMENT OF ENERGY
Office of Science

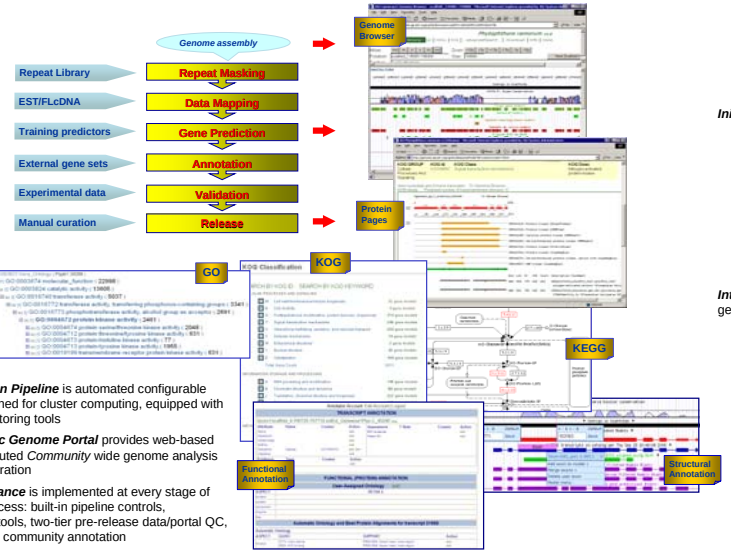
Abstract

We annotated over 50 eukaryotic genomes including 41 (80%) in 2006-2008. Annotation throughput has increased ~10 fold in the last 5 years. We improved annotation process, automated the annotation pipeline, and further developed genome analysis tools integrated in the Eukaryotic Genome Portal.

Post-annotation analysis of these genomes resulted in 3-fold increase of the number of genome publications in Nature and Science. These were enabled by JGI Community Annotation program focused on achieving higher quality of annotations, teaching and training new users, and building active user communities around the genomes sequenced at JGI. This unique across genome sequencing centers program was well received by user communities.



Annotation Pipeline and Eukaryotic Genome Portal



JGI Annotation Pipeline is automated configurable pipeline, designed for cluster computing, equipped with graphical monitoring tools

JGI Eukaryotic Genome Portal provides web-based tools for distributed *Community* wide genome analysis and manual curation

Quality assurance is implemented at every stage of annotation process: built-in pipeline controls, benchmarked tools, two-tier pre-release data/portal QC, user feedback, community annotation

Staged Portal Release

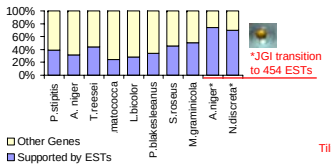
Initial release: blastable and downloadable assembly and ESTs

Intermediate release: genome browser with preliminary set of gene models and lines of evidence

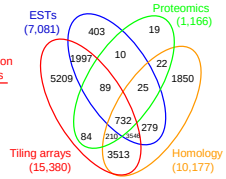
Final release: final gene set with annotations, analyses, and tools for manual curation

Validation of Annotations and Genome Analysis

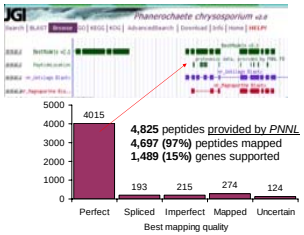
ESTs Support 25-80*% of Predicted Genes



Combined Lines of Evidence

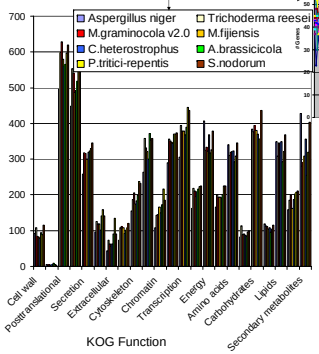


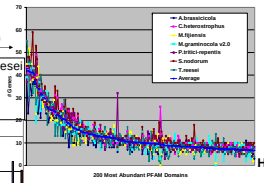
Proteomics Support



4,825 peptides provided by *PNNL*
4,697 (97%) peptides mapped
1,489 (15%) genes supported

Comparative Analysis of Functional Annotations (KOG and PFAM)

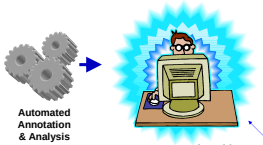




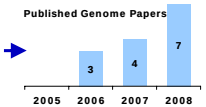
Homology and EST support for *H.annosum* gene models predicted by different annotation pipelines

	Eugene	Genmark	Fgenesh	JGI Pipe
Number of gene models	11,547	9,609	8,409	12,270
Models with partial EST support	5544	3829	4567	5248
with full length EST support	2538	1182	2896	3073
EST coverage per gene	77.7%	68.2%	80.8%	79.1%
Supported splice sites	41,581	40,808	45,498	47,671
Models with homology support	6758	6043	5750	7214
Model coverage	64%	60%	68%	69%
Models with homology and EST support	2894	2172	2720	2953

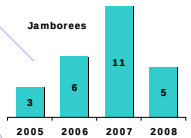
Community Annotation



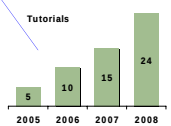
Published Genome Papers



Jamborees



Tutorials

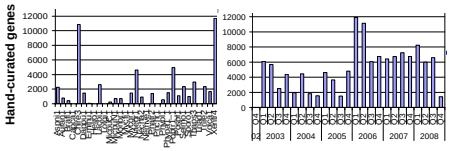


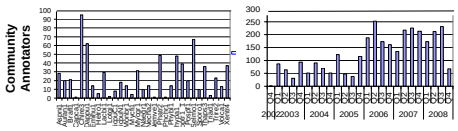
Communication with Users:

- Initial hand-shake meetings (project scope, community inputs)
- Genome Steering Committee calls to organize community annotation
- User notification about releases and feedback collection
- In-person, phone, video training and web tutorial
- Genome workshops (jamborees) and User Community Meetings
- User requests management and tracking

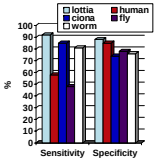
2006-2008: ~800 active annotators built ~8,000 models and hand curated ~70,000 genes in 50 eukaryotic genomes

Per Organism **Per Quarter**

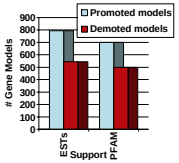




Value of Manual Inputs to Annotation



Trained gene predictors



Manually curated genes

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.