Annotation of Eukaryotic Genomes

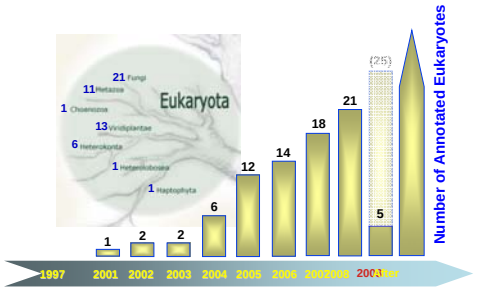
Igor Grigoriev*, Andrea Aerts, Frank Korzeniewski, Alan Kuo, Robert Otillar, Asaf Salamov, Xueling Zhao, Kemin Zhou
Genome Annotation Group, JGI; *ivgrigoriev@lbl.gov

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

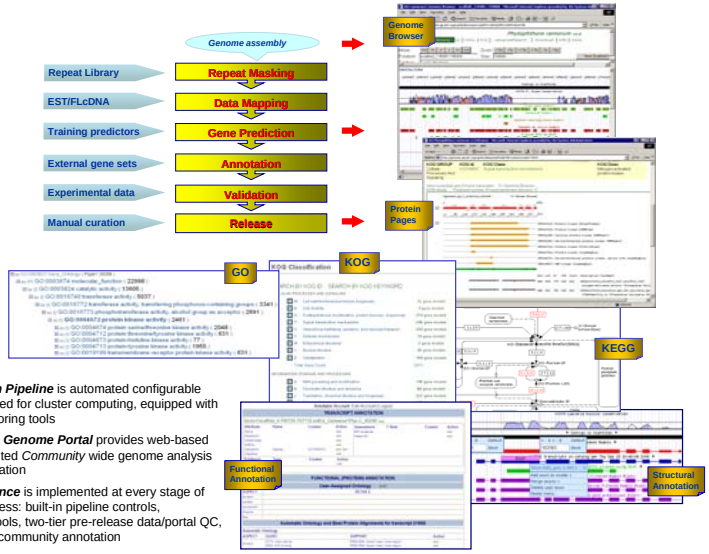
U.S. DEPARTMENT OF ENERGY
Office of Science

Abstract

We annotated over 60 eukaryotic genomes increasing annotation throughput ~10 fold in the last 5 years with improved annotation process, automated annotation pipeline, and further developed genome analysis tools integrated in the Eukaryotic Genome Portal. We developed a Community Annotation model, unique across sequencing centers, to achieve higher quality of annotations, teach and train new users, and build stronger genome user communities. These tools will serve as a platform for programmatic approach to scaling up sequencing and analysis of microbial eukaryotic, for example, JGI Fungal Program, along the directions of (i) sequencing phylogenetic breadth, (ii) deeper sampling DOE mission relevant organisms, and (iii) exploring ecological diversity. A diverse portfolio of microbial eukaryotes (68 fungal and 21 algal projects at different stages in JGI pipeline) and DOE mission positions the JGI to sequence a broad range of microbial eukaryotes.



Annotation Pipeline and Eukaryotic Genome Portal

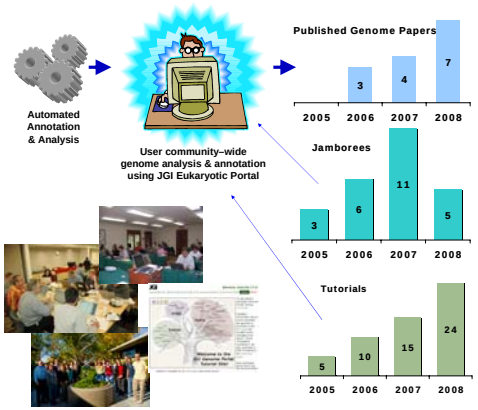


JGI Annotation Pipeline is automated configurable pipeline, designed for cluster computing, equipped with graphical monitoring tools

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

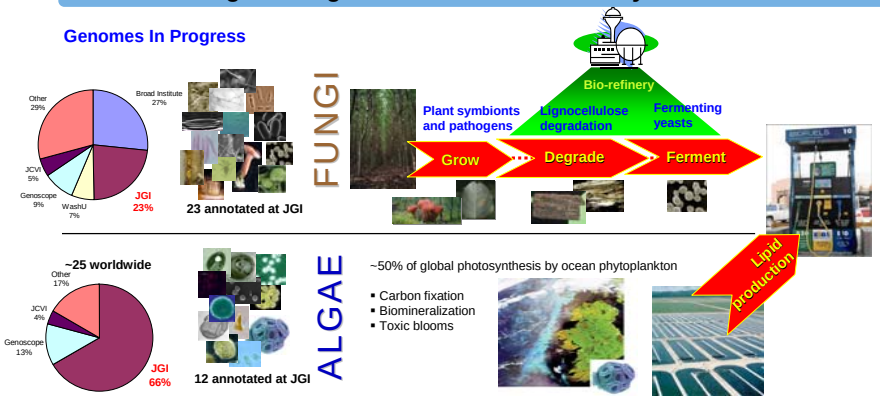
Community Annotation



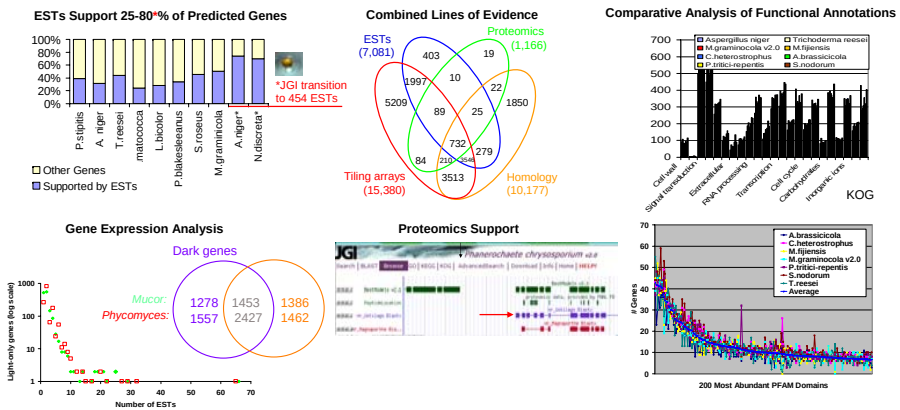
2006-present: 800+ active annotators hand curated ~80,000+ genes in 50 eukaryotic genomes

Fungi and Algae for DOE Mission: Today and Tomorrow

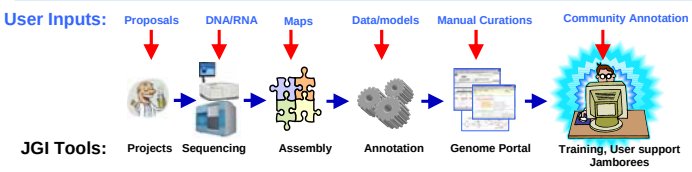
Genomes In Progress



Validation of Annotations and Genome Analysis



Programmatic Tools



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+EST support	2894	2172	2720	2953

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.