



Genome **S**equence **A**nnotation **S**erver

An easy to use, web-based platform for individual and collaborative structural and functional genome annotation

Jodi Humann Stephen Ficklin, Taein Lee, Chun-Huai Cheng, Sook Jung, Jill Wegrzyn, David Neale, Dorrie Main*

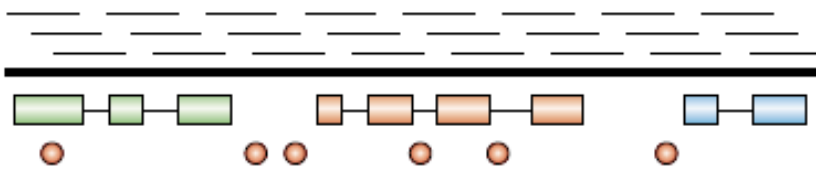
dorrie@wsu.edu

What is DNA annotation and why do it?

- **Getting the DNA sequence is only the first step**
- **Need to know the biological relevance of the DNA sequence**
- **Annotated sequence can be used to find putative genes of interest for study**

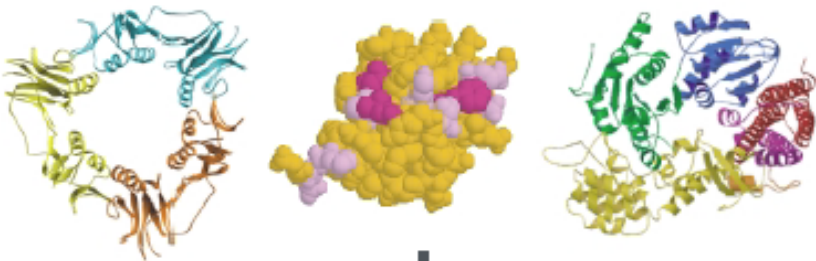
Where?

Nucleotide-level annotation



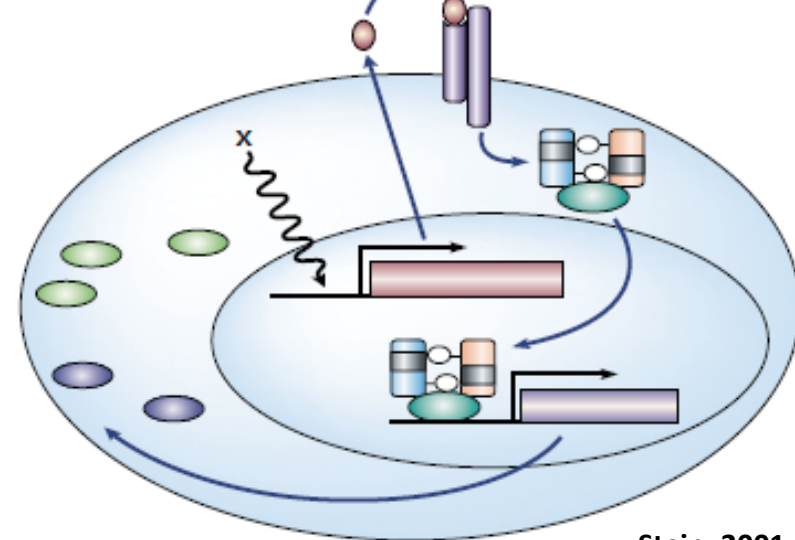
What?

Protein-level annotation



How?

Process-level annotation



Step 1: Nucleotide level

Genes, ORFs, genetic markers, tRNA, rRNA, ncRNA, repeats, regulatory elements

Step 2: Protein level

Translate genes and ORFs into proteins, search for homologs, assign putative function

Step 3: Process level

Assign GO terms, do lab experiments (mutagenesis, transcriptomics, RNA silencing, etc.)

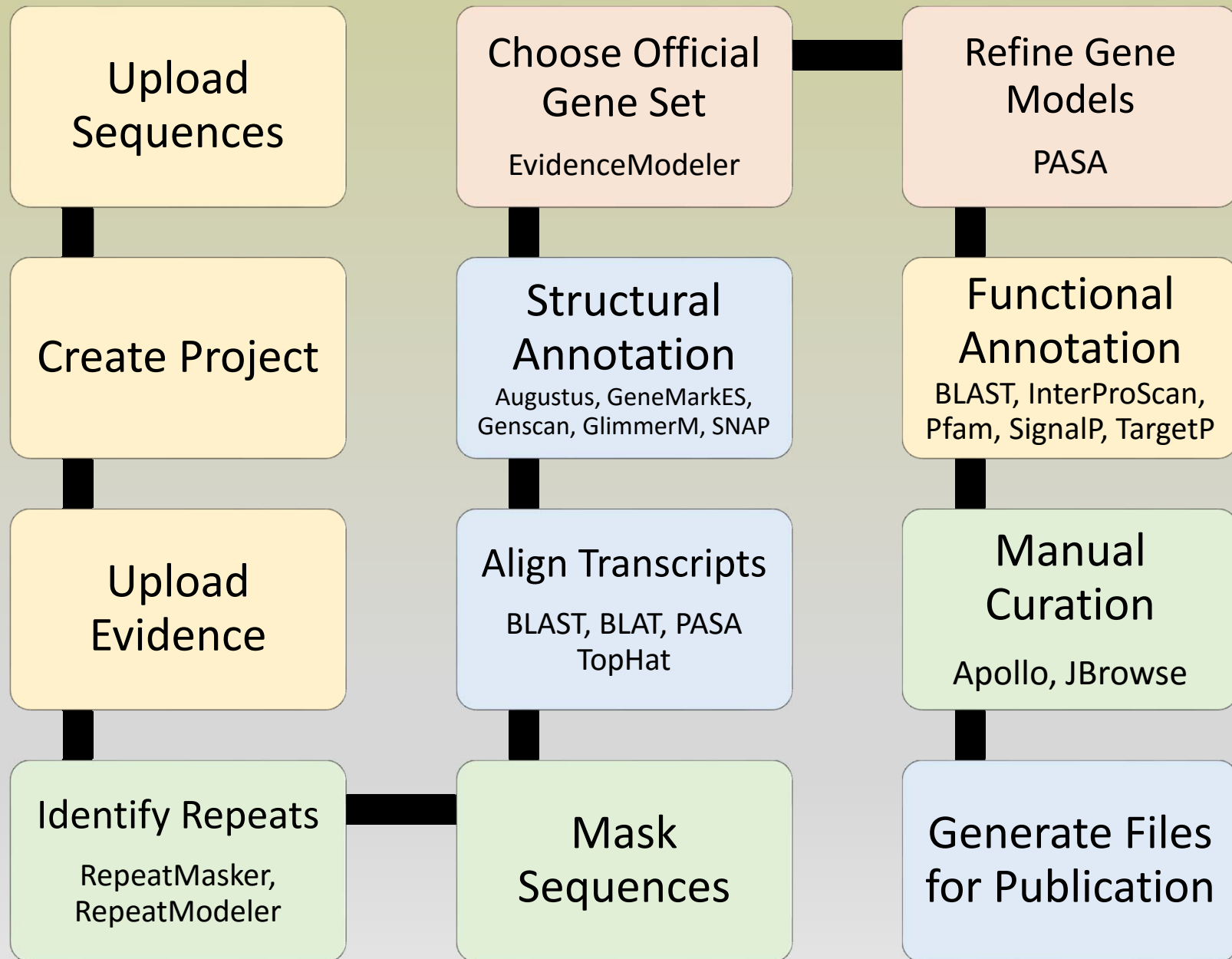
What scientists want

- **Current annotation tools:**
 - **Many tools available, but run independently of each other**
 - **Most of the tools are run via the command line and require server access**
- **Scientists want a platform that:**
 - **Is a single location for DNA annotation**
 - **Does not require management of computing equipment and software tools**
 - **Is easy to use and can be adapted to a variety of DNA sequences**

What is GenSAS?

- **A single website that combines numerous annotation tools into one interface**
- **User accounts keep data private and secure as well as allow users to share data for collaborative annotation**
- **Easy-to-use interfaces, with integrated instructions allow researchers at all skill levels to annotate DNA**

GenSAS annotation process



Welcome to GenSAS

The Genome Sequence Annotation Server (GenSAS) is an online tool that provides a pipeline for whole genome structural and functional annotation. Users can upload genome sequences and select from a variety of tools for repeat masking, prediction of gene models and other structural features as well as functional annotation tools. GenSAS integrates with **JBrowse** and **Apollo** to provide visualization and editing. Please see our video tutorial under the "Help" tab.

What's Coming!

Look for these improvements soon!

The screenshot displays the GenSAS v4.0 web interface. The top navigation bar includes 'Project', 'Sequences', 'Files', 'Reports', 'Masking', 'Genes', 'Consensus', 'Apollo', 'JBrowse', 'Annotations', and 'Status'. The 'Available Tools' section is highlighted, showing options for 'Gene Predictions' (Transcript Alignments, Protein Alignments, Other Features) and 'Other Features' (Repeat Masking, Repeat Identification, Repeat Classification, Repeat Annotation, Repeat Masking, Repeat Identification, Repeat Classification, Repeat Annotation). The 'Apollo' genome browser is shown below, displaying a genomic track with gene models and annotations. The 'Job Queue' table on the right shows the status of various jobs, including 'Repeat Masking', 'Repeat Identification', 'Repeat Classification', 'Repeat Annotation', 'Repeat Masking', 'Repeat Identification', 'Repeat Classification', and 'Repeat Annotation'.

User login

Username *

Password *

- [Create new account](#)
- [Request new password](#)

Log in

Beta-testing GenSAS v5.0,
to be released August



Tabs will appear here to provide the content needed for each stage. Some tabs can be closed and re-opened later but others will always remain. The tabs will provide instructions for each workflow stage.



Click the buttons above to move through the annotation workflow. As you complete a stage the next button becomes available. Click the **Project** button to begin.



Welcome to Genome Sequence Annotation Server !

Click the **Job Queue**



to view the analysis jobs for the project. Click the **Browser** tab to view the predicted features aligned to the genomic sequence. Click the **Sharing** tab to share the project with other GenSAS users.

Job Queue

Please start a project

Browser

Sharing

GenSAS welcome tab provides users with a quick overview of what each of the three screen sections do.

- **Sequence Tab:**
 - Single sequence or multi-sequence fasta file
 - Please make sure your assembly is good quality
 - New to v5.0, users can upload a multi-sequence FASTA file and create a subset based on sequence names or minimum size
 - If no subset is created all sequences in multi-sequence FASTA file are analyzed with the same parameters
- **Project Tab:**
 - Open existing project or shared project
 - Create new project

- **GFF3 Tab (optional):**
 - Previous annotations
 - Output from other tools
- **Evidence Tab (optional):**
 - EST, mRNA sequences
 - Repeat motifs
 - Protein sequences
 - NCBI gene structures for organism
 - Pre-processed Illumina RNA-Seq reads

**The more organism specific data you have,
the better the annotation will be**

- **Repeats Tab:**
 - RepeatMasker – evidence based repeat finder
 - RepeatModeler – *de novo* repeat finder
 - Can run each tool multiple times with different parameters by changing job name
- **Masking Tab:**
 - Look at the results in JBrowse and choose which set(s) to use in the consensus
 - Masked consensus is then used as input for the annotation tools unless the user elects to skip repeat masking

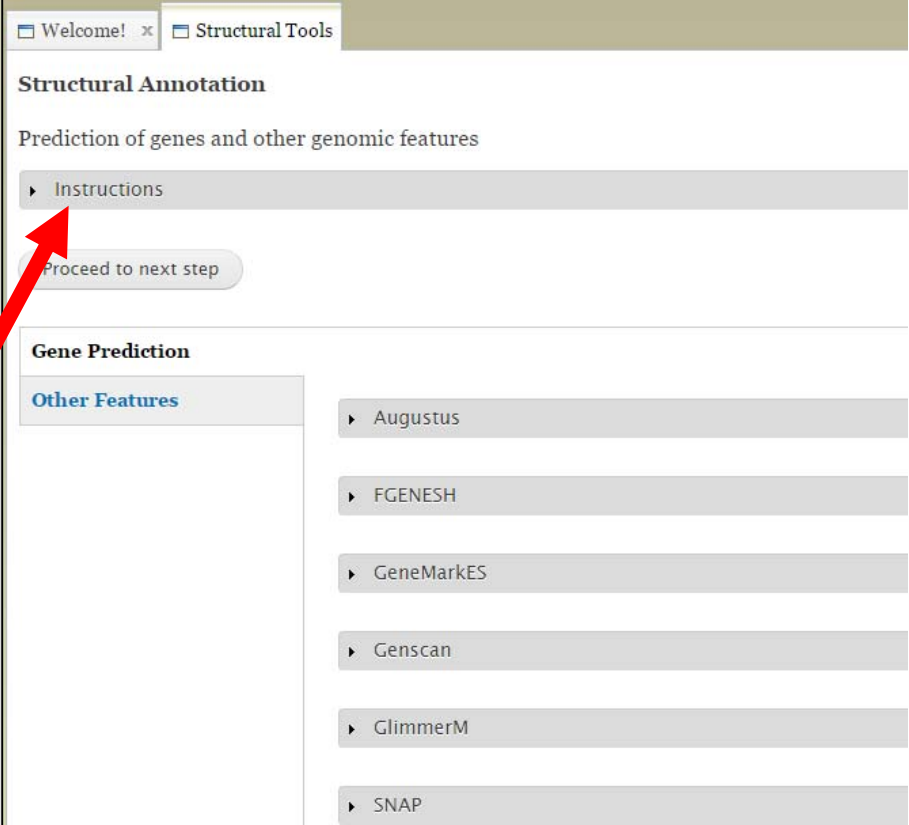
- **Align Tab:**

- New in v5.0, Align step has been added to allow users to align RNA-Seq data for training the gene prediction programs
- Also can align full-length transcripts and proteins

- **Structural Tab:**

- Gene prediction programs
- SSR Finder
- tRNAScanSE

All tabs have an Instructions section that can be opened and collapsed



Welcome! x Structural Tools

Structural Annotation

Prediction of genes and other genomic features

▸ Instructions

Proceed to next step

Gene Prediction

Other Features

- Augustus
- FGENESH
- GeneMarkES
- Genscan
- GlimmerM
- SNAP

- **OGS Tab:**
 - New to v5.0, Official Gene Set tab allows user to designate gene model set for manual annotation process and final publication
 - Use previous annotation, output from single gene predictor or generate consensus using EvidenceModeler
- **Refine Tab:**
 - Use PASA and RNA evidence to refine OGS gene models

The Official Gene Set (OGS)

▶ Instructions

[Create Consensus](#)

[Use Existing](#)

Please select only one of the following jobs to serve as the official gene set.

	Job
<input type="radio"/>	Augustus
<input type="radio"/>	Augustus-no-training (Augustus)
<input type="radio"/>	BLAST nucleotide (blastn)
<input type="radio"/>	PASA
<input type="radio"/>	Tophat

Job Queue

[View full report](#) | [Update status](#)

Repeats & Masking

Job Name	Status
Masked Repeat Consensus	Completed
RepeatMasker	Completed
RepeatModeler	No results

Genes & Other Predictions

Job Name	Status
Augustus	Completed
Augustus-no-training	Completed
BLAST nucleotide (blastn)	Completed
PASA	Completed
PASA Consensus Refinement	Started
Tophat	Completed

Browser

Sharing

- Job status can be monitored through Job Queue
- Progress through GenSAS is automatically saved
- Users can log off GenSAS and jobs will continue running
- While jobs are running, users can look at the completed results in Apollo/JBrowse
- Once the project has results, users can share the project with other GenSAS users for collaborative annotation

- **Functional Tab:**

- **OGS gene models and other user selected gene models are functionally annotated**

Functional Analysis

▶ [Instructions](#)

Proceed to next step

Job Selection

Please choose a previously run job for functional analysis, then add one more jobs rom the the tool selection below.

- PASA Consensus Refinement
- Genes Consensus
- Augustus
- GeneMarkES
- SNAP

BLAST protein vs protein (blastp)

[InterProScan](#)

[Pfam](#)

[SignalP](#)

[TargetP](#)

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Job Name

BLAST protein vs protein (blastp)

>

Show 10 entries Filter:

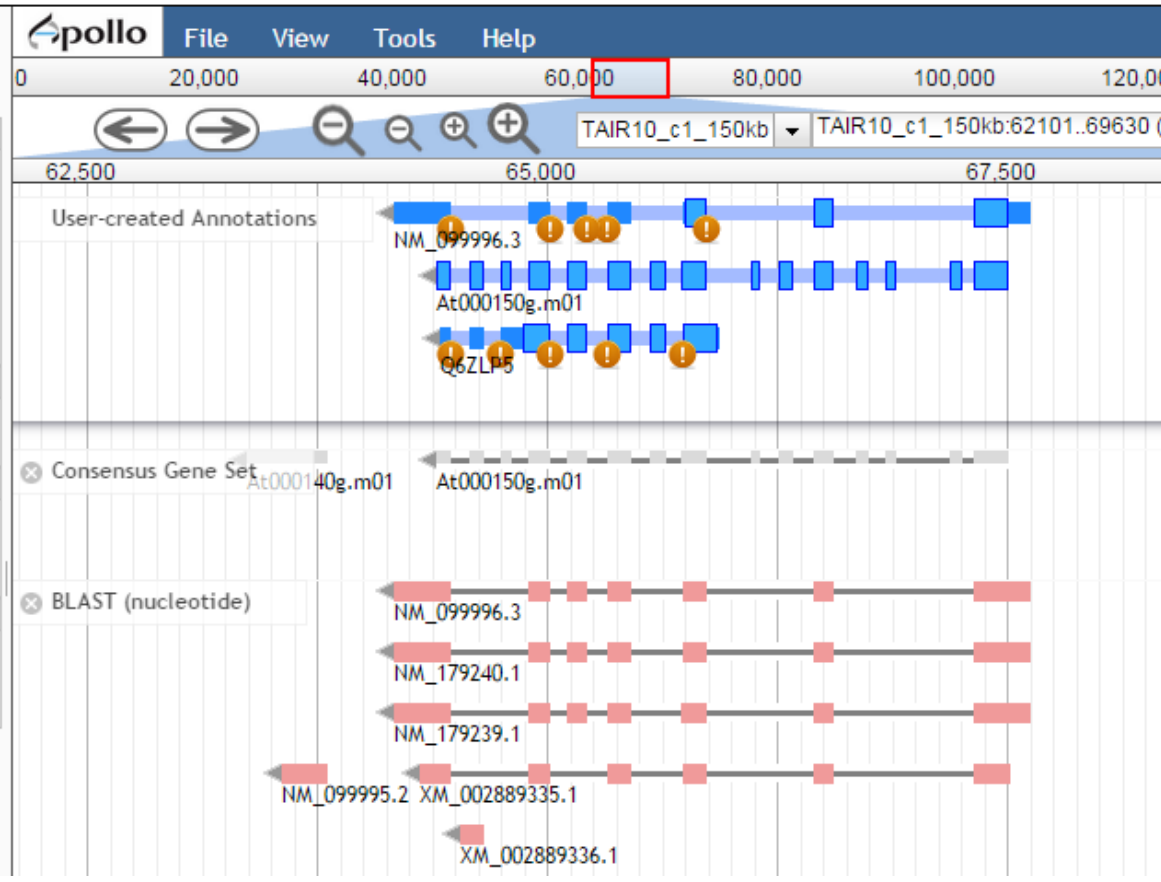
Track	Feature name	Feature type	Last modified	Editor	Owner
<input type="checkbox"/> TAIR10_c1_150kb	Q6ZLP5	mRNA	Tue Jan 06 14:56:47 PST 2015	GenSASuser	GenSASuser
<input type="checkbox"/> TAIR10_c1_150kb	null	gene	Tue Jan 06 14:56:47 PST 2015	null	GenSASuser
<input type="checkbox"/> TAIR10_c1_150kb					

[Reload JBrowse](#) | [View JBrowse](#)

Showing 1 to 3 of 3 entries

Available Tracks

- Gene Predictions 6
 - Augustus
 - Augustus-complete genes only
 - Consensus Gene Set
 - Genscan
 - GlimmerM
 - SNAP
- Protein Alignments 1
 - BLAST (proteins)
- Repeats 2
 - RepeatMasker
 - RepeatMasker-slow
- Transcript Alignments 2
 - BLAST (nucleotide)
 - BLAT



- **Manual annotation from Apollo are automatically merged into OGS at Publish Step**

Publish

▶ Instructions



Notice: Despite best efforts to ensure publishable files contain accurate information, there is always the chance that an unknown bug or issue may affect results. Please review all files before public release.

▼ Available Results for Publishing

Please choose the jobs to be included in the published release for this project.

Consensus Masking

- Masked Repeat Consensus

The repeat masked consensus job created the FASTA sequence on which all other predictions were made. This job should be included in any published release.

Repeats & Masking

- RepeatMasker (RepeatMasker)

Your masked consensus sequence should be all you need to publish. But you can include individual repeat finding jobs if desired.

Gene Refinement

- PASA Consensus Refinement (PASA Consensus Refinement)

The set of jobs used for automated refinement of gene models.

Gene Consensus

- Genes Consensus (EvidenceModeler)

The job used for creation of a consensus of gene model.

Gene Predictions

- Augustus (Augustus)

**GenSAS exports data in GFF3
and FASTA formats**

Supported by

