Structural and functional annotation using a web-based platform for model and non-model organisms

Jodi L. Humann[1], Stephen P. Ficklin[1], Taein Lee[1], Chun-Huai Cheng[1], Heidi Hough[1], Sook Jung[1], Jill Wegrzyn[2], David Neale[3], and Dorrie Main[1]

[1]Washington State University, Pullman, WA; [2]University of Connecticut, Storrs, CT; [3]University of California, Davis, CA

Contact email: jhumann@wsu.edu, dorrie@wsu.edu

www.gensas.org

Abstract

The Genome Sequence Annotation Server v5.0 (GenSAS) merges several annotation tools into one, easy-to-use, web-based interface. GenSAS guides researchers through the annotation process from start to finish and integrates JBrowse for viewing of annotation tool results and Apollo for manual annotation, which can be a collaborative project between multiple users. GenSAS now supports the use of RNA-Seq reads from the Illumina platform for training gene prediction programs in addition to genome-specific supporting data files in the GFF3 and FASTA file formats. After adding the genome sequence and supporting evidence files, users proceed through repeat identification and masking, alignments of evidence to the genome sequence, structural annotation and gene prediction, designation or creation of the official gene set for gene model refinement, functional annotation, optional manual curation, and generation of files for publication.

Key features of GenSAS

- Completely web-based with no software to install and integrated JBrowse and Apollo.
- User accounts ensure data is kept private, progress is saved automatically and allows projects to be shared with other GenSAS users for collaborative annotation.
- Tools settings are customizable and can be used with user uploaded evidence files or with GenSAS provided global library files for repeat, transcript and protein alignments.
- Results are exported in GFF3 and FASTA file formats which can be used with many downstream analysis programs and for publication.

Tools integrated into GenSAS

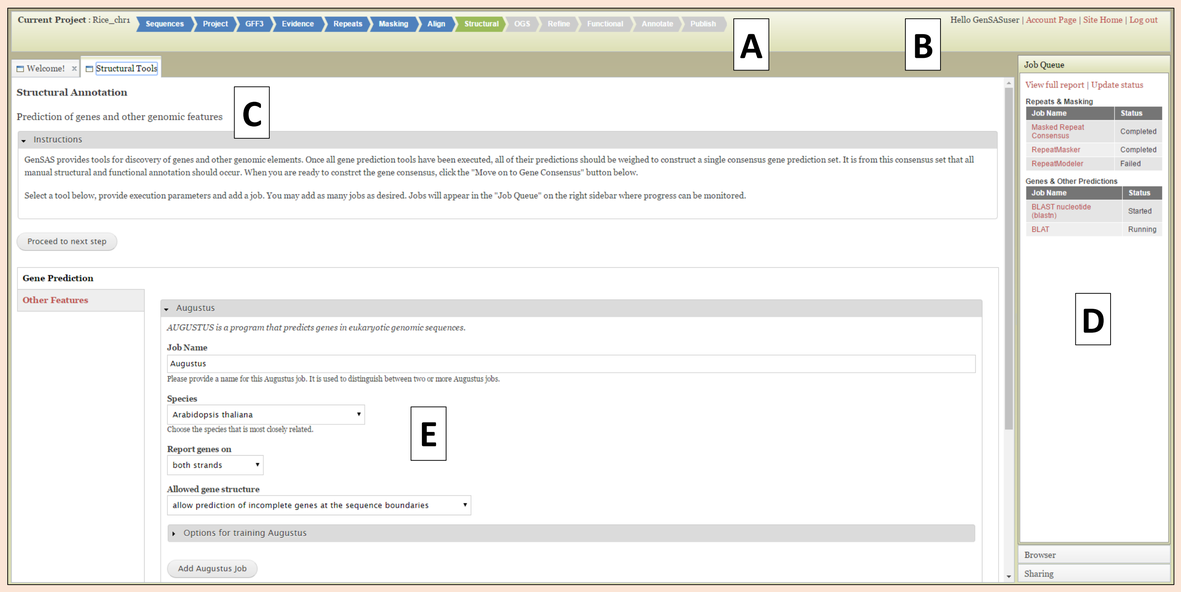
Supported DNA evidence	**Alignment tools**
ESTs and transcripts	BLAT, nucleotide BLAST, PASA
RNA-Seq reads	TopHat

Eukaryote structural annotation	**Available tools**
Repeat identification	RepeatMasker, RepeatModeler
Gene model prediction	Augustus, GeneMarkES, Genscan, GlimmerM, SNAP
Gene model consensus	EVidenceModeler

Prokaryote structural annotation	**Available tools**
Gene model prediction	GeneMarkS, Glimmer3

Other structural features	**Available tools**
Open reading frames (ORFs)	Getorf
Simple sequence repeats (SSRs)	SSR Finder
tRNAs	tRNAScan-SE
rRNAs	RNAmmer

Functional annotation	**Available tools**
Protein homology	Protein BLAST
Conserved protein signatures	InterProScan
Protein family homology	Pfam
Signal peptides	SignalP
Protein localization	TargetP



GenSAS user interface

A.) Project workflow: Shows current stage of the annotation process and enables navigation to different steps by clicking on step name.

B.) User and project info: Displays name of current project and provides links to logout or user account information.

C.) Primary work area: Different tabs appear as the annotation progresses. This is the main work area in GenSAS and there are instructions at the top of each tab.

D.) Accordion menu: The different sections of this menu allow jobs to be tracked, JBrowse/Apollo to be opened and projects shared with other GenSAS users.

E.) Tool settings: If adjustable parameters are available for the tool, they can easily be changed through the GenSAS interface.

New features of GenSAS v5.0

- ***Sequences Step:*** Before starting a project, users are allowed to upload sequences for annotation and create subsets of sequences from multiple sequences FASTA files. Sequences can be filtered by minimum size or sequence names.
- ***RNA-seq Reads:*** Users can now upload pre-processed Illumina RNA-seq reads to train Augustus. This is especially useful for non-model organisms. Either one non-paired read file or a set of paired read files can be added. Maximum RNA-seq data size is 100 Gb.
- ***Alignment Step:*** DNA alignments are now a separate step before structural annotation. This allows for RNA-seq data to be aligned to the genome with TopHat prior to running Augustus.
- ***New tools:*** GeneMarkES and GeneMarkS are now part of the Structural Annotation step and are excellent gene model predictors for eukaryotes and prokaryotes, respectively. RNAmmer has also been added for rRNA identification.
- ***Official Gene Set (OGS):*** Users designate or create an OGS before functional annotation. Any manual curation performed with Apollo, will be automatically integrated into the OGS at the Publish Step.

Work in progress

- Option to create single merged GFF3 file of annotation at Publish Step
- Integration of newest versions of Apollo and JBrowse
- Improvements in job submission to cluster to allow for BLAST jobs to run more efficiently

Funding provided by