Chapter 3

Structural and Functional Annotation of Eukaryotic Genomes with GenSAS

Jodi L. Humann, Taein Lee, Stephen Ficklin, and Dorrie Main

Abstract

The Genome Sequence Annotation Server (GenSAS, https://www.gensas.org) is a secure, web-based genome annotation platform for structural and functional annotation, as well as manual curation. Requiring no installation by users, GenSAS integrates popular command line-based, annotation tools under a single, easy-to-use, online interface. GenSAS integrates JBrowse and Apollo, so users can view annotation data and manually curate gene models. Users are guided step by step through the annotation process by embedded instructions and a more in-depth GenSAS User's Guide. In addition to a genome assembly file, users can also upload organism-specific transcript, protein, and RNA-seq read evidence for use in the annotation process. The latest versions of the NCBI RefSeq transcript and protein databases and the SwissProt and TrEMBL protein databases are provided for all users. GenSAS projects can be shared with other GenSAS users enabling collaborative annotation. Once annotation is complete, GenSAS generates the final files of the annotated gene models in common file formats for use with other annotation tools, submission to a repository, and use in publications.

Key words Structural annotation, Functional annotation, Annotation pipeline, Manual curation

1 Introduction

While advances in sequencing and computational technologies, coupled with more affordable costs, are enabling researchers to routinely sequence genomes of interest, predicting genes and assigning biological relevance to the putative proteins that those genes encode remain challenging tasks for non-computational scientists. Eukaryotic genome annotation involves three major steps: identification and masking of repetitive DNA sequences, structural annotation, and functional annotation (see excellent review of process by Yandell and Ence [1]). Compared to prokaryotic organisms, eukaryotic genome sequences contain repetitive sequences that complicate the annotation process. Repeat identification and masking simply change the bases in repetitive regions to an "N" or "X" nucleotide, allowing downstream tools to ignore the

Martin Kollmar (ed.), *Gene Prediction: Methods and Protocols*, Methods in Molecular Biology, vol. 1962,
https://doi.org/10.1007/978-1-4939-9173-0_3, © The Author(s) 2019

repeat. During the structural annotation portion of the process, DNA landmarks such as protein-coding genes, tRNAs, and rRNAs, which can be determined based on the DNA sequence, are identified. After structural annotation, functional annotation is performed in silico to infer biological function to the proteins of the gene models. After the initial annotation of the genome, some genome sequencing projects (e.g., TAIR, https://www.arabidopsis.org/) also perform manual curation of the gene models to improve the quality of the annotation.

There are many different programs that annotate specific features, such as gene models, repeats, tRNAs, and rRNAs (examples in Table 1). Some tools are accessible online, while others require a local installation. There are also annotation pipelines that combine multiple annotation tools. MAKER2 [2] is an example of a gene prediction pipeline. MAKER2 runs three different gene prediction programs (SNAP, GeneMark-ES, AUGUSTUS) within the pipeline and also will align user-provided transcript, RNA-seq, and protein evidence to the genome. MAKER2 then uses the results from the gene prediction tools and alignments of provided evidence to generate a consensus gene model set. Another annotation pipeline is the NCBI Eukaryotic Genome Annotation Pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/). The NCBI pipeline first identifies repetitive DNA sequences for masking; then aligns transcripts, RNA-seq reads, and proteins to the genome sequence; and uses the aligned evidence to predict gene models. The NCBI pipeline also identifies miRNAs, tRNAs, rRNAs, snoRNAs, and snRNAs, in addition to predicting gene models.

The vision of GenSAS (https://www.gensas.org) is to provide a web-based, modular tool that requires no software installation and management and is easy for scientists of all skill levels to use via a graphical user interface. The major annotation steps of the GenSAS workflow include repeat identification and masking, evidence alignment to the genome, structural annotation (genes, rRNA, tRNA), functional annotation of gene models, optional manual editing of the gene models, and creation of final annotation files. The design of GenSAS allows for the addition of new annotation tools; thus the list of available tools in GenSAS can change. This chapter will discuss tools in GenSAS v6.0 (Table 1). Even if new tools are added to GenSAS, the GenSAS interface, and how users interact with GenSAS, remains the same. For the most current list of GenSAS tools, please see https://www.gensas.org/tools. The goal of this chapter is to describe how to use GenSAS and provide pointers on how to get the best annotation possible from using this annotation platform.

Table 1
Tools used in GenSAS

Tool	References	Website
PRINSEQ-lite	[3]	http://prinseq.sourceforge.net/
BUSCO	[4, 5]	https://busco.ezlab.org/
RepeatMasker		http://www.repeatmasker.org/
RepeatModeler		http://www.repeatmasker.org/
AUGUSTUS	[11, 12]	http://bioinf.uni-greifswald.de/augustus/
BRAKER2	[13]	http://exon.gatech.edu/GeneMark/braker1.html
GeneMark-ES	[14, 15]	http://exon.gatech.edu/GeneMark/
Genscan	[16]	http://genes.mit.edu/GENSCAN.html
GlimmerM	[17]	http://www.cbcb.umd.edu/software/glimmerm/
SNAP	[18]	https://github.com/KorfLab/SNAP
BLAST+	[19]	http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
BLAT	[20]	https://genome.ucsc.edu/FAQ/FAQblat.html; https://github.com/icebert/pblat
Diamond	[21]	https://github.com/bbuchfink/diamond
PASA	[22]	http://pasapipeline.github.io/
HISAT2	[23]	https://ccb.jhu.edu/software/hisat2/index.shtml
TopHat2	[24]	https://ccb.jhu.edu/software/tophat/index.shtml
getorf		http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/getorf.html
RNAmmer	[25]	http://www.cbs.dtu.dk/services/RNAmmer/
SSR Finder		GenSAS custom tool, MainLab Bioinformatics
tRNAScan-SE	[26, 27]	http://lowelab.ucsc.edu/tRNAscan-SE/
EVidenceModeler	[28]	http://evidencemodeler.github.io/
InterProScan	[29]	http://www.ebi.ac.uk/Tools/pfa/iprscan5/
Pfam	[30]	http://pfam.xfam.org/
SignalP	[31]	http://www.cbs.dtu.dk/services/SignalP/
TargetP	[32]	http://www.cbs.dtu.dk/services/TargetP/
One code to find them all	[33]	http://doua.prabi.fr/software/one-code-to-find-them-all
Apollo	[34]	http://apollo.berkeleybop.org/
JBrowse	[35]	https://jbrowse.org/

2 Preparing to Use GenSAS

2.1 GenSAS User Account

GenSAS is available at https://www.gensas.org. Users must register for an account using the "Create new account" link on the right side of the home page. User accounts keep data private, allow for sharing of projects with other GenSAS users, and enable users to log out of GenSAS while jobs are running. In order to ensure that GenSAS is available to as many users as possible, there are some limitations to user accounts and projects (Table 2).

2.2 Is Your Genome Ready for Annotation?

The quality of a genome annotation depends on several factors, but the most important factor is the quality of the genome assembly. The saying "garbage in, garbage out" applies to annotation. If the input genome is split in hundreds of thousands of contigs or scaffolds, smaller than the average gene size, the gene prediction programs will not be very effective. The method used to assemble the genome should have produced a file reporting the number of contigs, the minimum and maximum lengths, and other metrics like the N50 (a weighted average length of contigs where more weight is given to longer contigs). If a majority of the assembled contigs are not over the average gene length for your organism, the genome will not annotate well. If you do not have a metrics report

Table 2
GenSAS user account and project limitations

Limitation	Details
Projects expire after 60 days unless expiration reset by user	Users receive email reminders that a project will expire. If expiration is not reset before 60 days, the project will be deleted.
User accounts will remain active as long as users have an active GenSAS project	When all projects have expired, the user is not part of a shared project, and the user has not logged in for 6 months, the GenSAS account is deleted.
User accounts are limited to 250 GB of storage space on GenSAS server	This size limitation includes all user-uploaded files as well as results generated by GenSAS. This is for all projects combined. If the user reaches 250 GB, new jobs will not run until old projects/data are deleted to free up space.
Assembly files must be high quality	All genome assemblies uploaded to GenSAS are evaluated and only assemblies under 25,000 total sequences and with over 50% of those sequences larger than 2500 bp will be accepted for use in a GenSAS project.
Only seven jobs per user can run at one time	While GenSAS does submit jobs to a computational cluster, the cluster resources are not endless. Users can only have seven jobs running at one time. However, more than seven jobs can be submitted and as running jobs complete, the waiting jobs in the queue will run.

from the assembly program, PRINSEQ [3] is an easy tool to gather that data. It is available to use on the web (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi?home=1). GenSAS does run PRINSEQ as part of the sequence upload process, but it is highly recommended that you check your genome before loading it to GenSAS. If your genome is in good shape in regard to sequence number and length, another tool to determine the completeness of the assembly is BUSCO [4, 5]. BUSCO determines the percentage of the core orthologous genes that are present in the assembly of unannotated genomes. BUSCO can either be run using a virtual machine (https://busco.ezlab.org/) or downloaded and installed locally (https://gitlab.com/ezlab/busco). If your genome assembly is missing a significant number of conserved, core genes according to BUSCO, then it might be best to wait to annotate the genome until a better-quality assembly is available. GenSAS does allow users to run BUSCO on uploaded genome assemblies.

2.3 User-Provided Files

It is best to gather all the needed files before starting a GenSAS project. In addition to the genome assembly in the FASTA format, GenSAS also accepts many different types of evidence files. Table 3 lists the types of evidence files that can be provided for GenSAS projects. While having a high-quality assembly is important for a

Table 3
User-provided files for GenSAS project

File contents	File extension	Used for
Genome assembly	.fasta, .fa, .fas	Sequences to be annotated with GenSAS
Aligned repeats	.gff3	Displayed as track in JBrowse, can be used in consensus masked sequence job
Aligned transcripts	.gff3	Displayed as track in JBrowse, can be used in gene consensus job
Aligned Proteins	.gff3	Displayed as track in JBrowse, can be used in gene consensus job
Predicted genes, previous annotation	.gff3	Displayed as track in JBrowse, can be used in gene consensus job
Other features	.gff3	Displayed as track in JBrowse
Repeat libraries	.fasta, .fa, .fas	Can be used with RepeatMasker
Transcripts/ESTs	.fasta, .fa, .fas	Can be used with alignment tools and to train AUGUSTUS
Proteins	.fasta, .fa, .fas	Can be used with alignment tools and to train AUGUSTUS
Gene structures	.gb	Gene models from GenBank, can be used to train AUGUSTUS
Illumina RNA-seq reads	.fastq	Can be used with TopHat2 and HISAT2; Resulting alignments can be used to train AUGUSTUS or BRAKER2

good annotation, having a good collection of evidence that is specific to and originates from the organism genome being annotated is equally important. Species-specific evidence files, used with some of the annotation tools, are especially helpful for non-model organisms. Users only need to provide species-specific data, as GenSAS provides up-to-date common databases such as repeat sequence collections from Repbase [6] for use with RepeatMasker (Table 1) and transcript and peptide sequence collections from NCBI RefSeq [7], SwissProt [8], and TrEMBL [8] for use with alignment programs. GenSAS accepts data that have already been aligned to the genome assembly in the form of GFF3 files or sets of unaligned sequences in the FASTA or FASTQ format for use with alignment tools within GenSAS. Ideally, the evidence files should be from the same organism that the genome sequence originated from, but this is not always the case. If you do not have evidence files from your organism, try to find some data in public repositories, such as GenBank (https://www.ncbi.nlm.nih.gov/).

The GFF3 file format (http://gmod.org/wiki/GFF3) is a standard nine-column format that defines annotated features (e.g., gene, mRNA, exon, intron, etc.), their name, their type, and location in the genome sequence and is a common output of annotation tools. Examples of GFF3 files to use as evidence in GenSAS are outputs from other previously run annotation tools, previous versions of the genome annotation, and aligned repeats, transcripts, and proteins. The sequence names from the assembly file must match the sequence names in the first column of the GFF3 file. The GFF3 importers in GenSAS use the feature types in column three of the GFF3 file. Table 4 lists the feature types recognized by each GenSAS GFF3 importer. If the GFF3 file has sequence names that do not match the names in the assembly file or has feature types that GenSAS does not recognizes, then no features will be imported into GenSAS.

Table 4
Feature types recognized in column 3 by the GenSAS GFF3 importers

GenSAS GFF3 Importer	Recognized feature types
Repeats	repeat, repeat_region
Transcript Alignments	match, match_part
Protein Alignments	match, match_part
Gene Predictions	gene (required), transcript, mRNA, CDS, exon, five_prime_UTR, three_prime_UTR
Other Features	Any term in column 3

FASTA files of repeat, transcript, EST, or protein sequences that are specific to the genome being annotated can also be uploaded to GenSAS and are used with various alignment tools within GenSAS. FASTA files have the sequence name on one line that begins with a ">" and the nucleotide or protein sequence on the following line(s). In addition to FASTA files, two other file types can be used as evidence in GenSAS. Gene models from NCBI (https://www.ncbi.nlm.nih.gov/) can be uploaded and used as evidence to train the AUGUSTUS (Table 1) gene prediction program. The gene models must be in the GenBank (.gb) file format and need to have at least 100 sequences that are similar enough at the protein level to align to the genome being annotated. GenSAS also accepts Illumina RNA-seq reads, either as paired or non-paired reads in FASTQ format. It is highly recommended that the RNA-seq reads are filtered by quality prior to upload to GenSAS. RNA-seq reads are aligned to the genome using TopHat2 (Table 1), and the resulting alignment can be used to train AUGUSTUS.

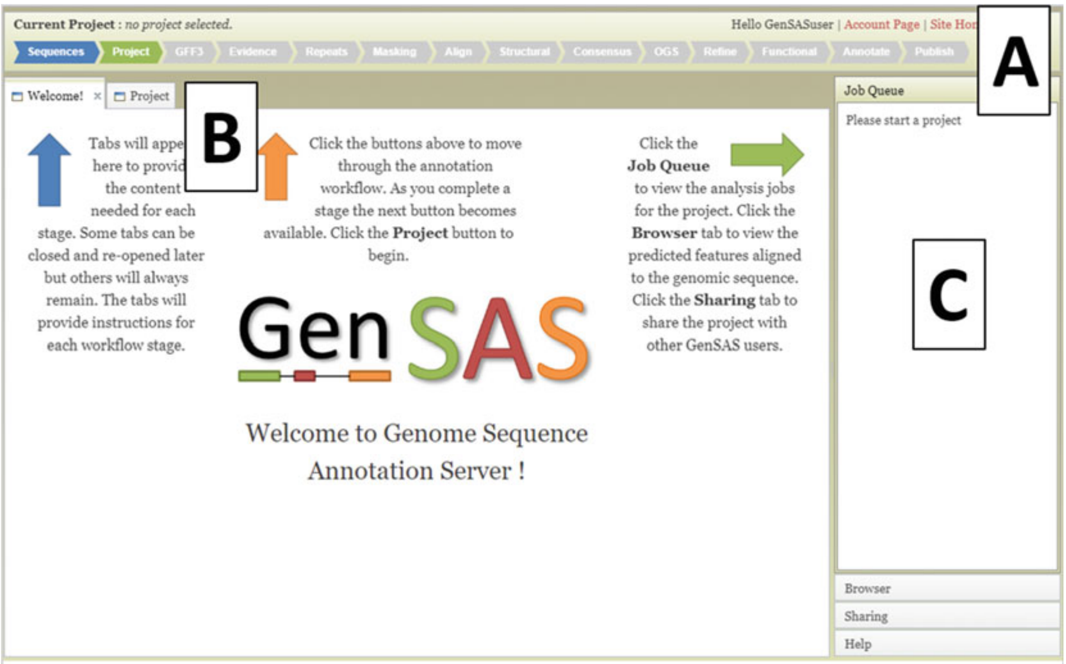
3 Generating a Genome Annotation with GenSAS

3.1 Getting Started with GenSAS

3.1.1 The GenSAS Interface

To access GenSAS, users log in on the homepage (https://www.gensas.org/) and then click on the "Use GenSAS" tab in the header menu. This opens the GenSAS interface which has three main sections. The header area (Fig. 1A) has a flowchart of the annotation process. The flowchart arrows can be clicked to navigate to each stage of the annotation process for the project. When the arrow is gray, it means that the step is not yet available. If the arrow is blue, the step is available for use. A green arrow indicates the current step of the project that is being viewed. The header also displays the name of the open GenSAS project in the upper left corner and links to the user's account details, the GenSAS homepage, and a logout link in the upper right corner. The center section of the GenSAS interface is the tab area (Fig. 1B) and is the main area of the interface. Different tabs open as the workflow progresses. See the last paragraph of this section for more information about the tabs. On the right side of the interface is an accordion menu (Fig. 1C). The accordion menu allows users to access the job queue, open JBrowse/Apollo, share projects with other GenSAS users, and access a "Help" section.

Users can see the status of submitted jobs and view job details in the Job Queue (Fig. 2). In the Job Queue, the "Status" of each job updates periodically or any time the user clicks "Update status." Clicking "View full report" opens a tab to see a more complete status report of each job in the overall job queue. As annotation jobs finish, the annotations are viewable in JBrowse, but users can also click the job name in the Job Queue and open the job results



Tabs will appe here to provid 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.

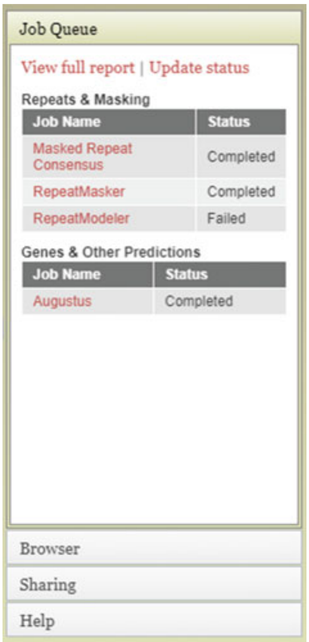
Click the buttons above to move through the annotation workflow. As you complete a stage the next button becomes available. Click the **Project** button to begin.

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.

Please start a project

Welcome to Genome Sequence Annotation Server !

Fig. 1 The GenSAS interface has three main sections. The header region (A) which has a flowchart of the annotation process, the tab area (B) where most user interactions with GenSAS occur, and the accordion menu (C) which has links to the Job Queue, to open JBrowse/Apollo, and to share projects between users



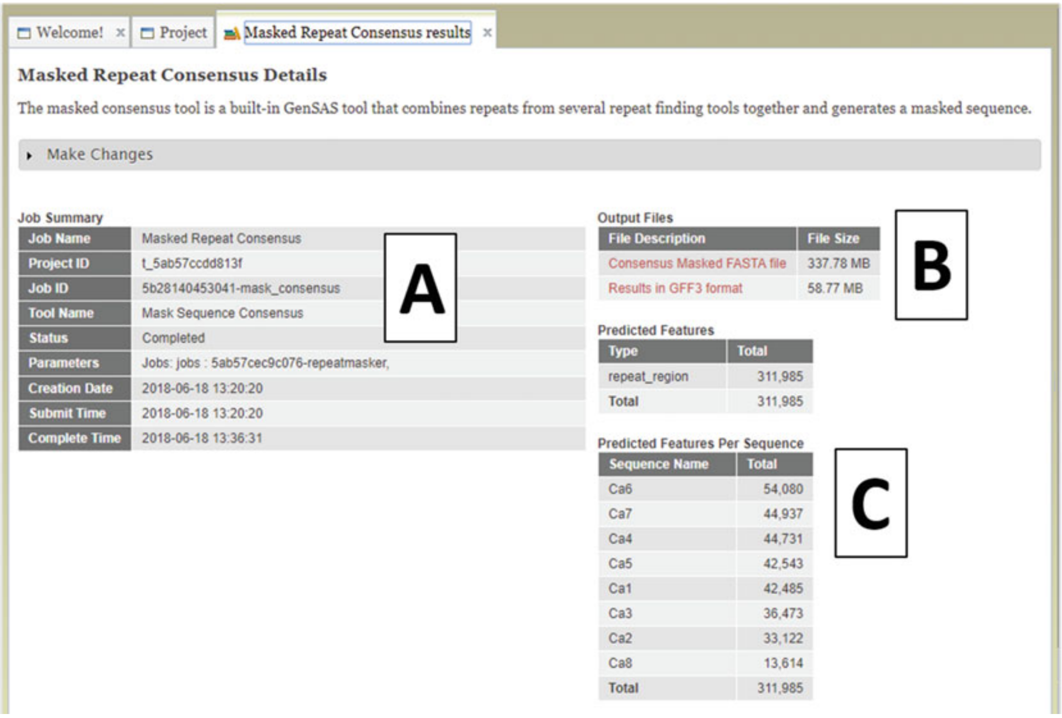
Repeats & Masking

Job Name	Status
Masked Repeat Consensus	Completed
RepeatMasker	Completed
RepeatModeler	Failed

Genes & Other Predictions

Job Name	Status
Augustus	Completed

Fig. 2 The GenSAS Job Queue lists the jobs associated with the project and their status. Clicking on the job name opens a results tab for that job



Masked Repeat Consensus Details

The masked consensus tool is a built-in GenSAS tool that combines repeats from several repeat finding tools together and generates a masked sequence.

Make Changes

Job Summary

Job Name	Masked Repeat Consensus
Project ID	t_5ab57ccdd813f
Job ID	5b28140453041-mask_consensus
Tool Name	Mask Sequence Consensus
Status	Completed
Parameters	Jobs: jobs : 5ab57cec9c076-repeatmasker,
Creation Date	2018-06-18 13:20:20
Submit Time	2018-06-18 13:20:20
Complete Time	2018-06-18 13:36:31

Output Files

File Description	File Size
Consensus Masked FASTA file	337.78 MB
Results in GFF3 format	58.77 MB

Predicted Features

Type	Total
repeat_region	311,985
Total	311,985

Predicted Features Per Sequence

Sequence Name	Total
Ca6	54,080
Ca7	44,937
Ca4	44,731
Ca5	42,543
Ca1	42,485
Ca3	36,473
Ca2	33,122
Ca8	13,614
Total	311,985

Fig. 3 The results tab for each job has three sections. The Job Summary section (A), the Output Files section (B), and some summary Tables (C)

tab (Fig. 3). In the job results tab, there is a Job Summary section (Fig. 3A) which lists the job name, the settings used, and the day and time of submission and completion. There is also an "Output Files" section which contains links to error and run log files, the raw output from the tool, and the GFF3 file that was loaded into JBrowse. Most tools also have some summary tables (Fig. 3C) that just provide a quick overview of the results. For most tools, multiple jobs of the same tool, with different parameters, can be submitted simultaneously, but the job names need to be unique. This allows users to experiment with different settings or evidence files in the same tool.

GenSAS uses an integrated instance of JBrowse to view data and the JBrowse plug-in, Apollo, for manual curation of annotations. To open JBrowse/Apollo, click on the "Browser" section of the accordion menu on the right of the GenSAS interface, and then click the "Open Apollo" button. The "Apollo" tab will open and has two sections. On the left is the JBrowse display (Fig. 4A), and on the right is the Apollo interface (Fig. 4B). The "Tracks" tab (Fig. 4C) is used to control which tracks, or the results from each tool, are visible. During the "Annotate" step of GenSAS, manual gene model editing can be performed using the "User-created Annotations" track (Fig. 4D). More details on how to use these

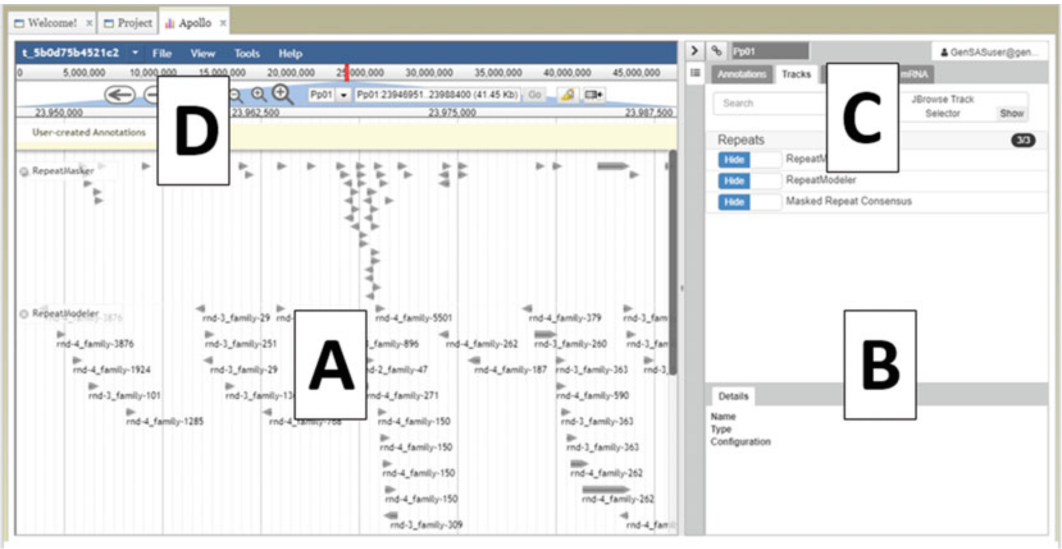
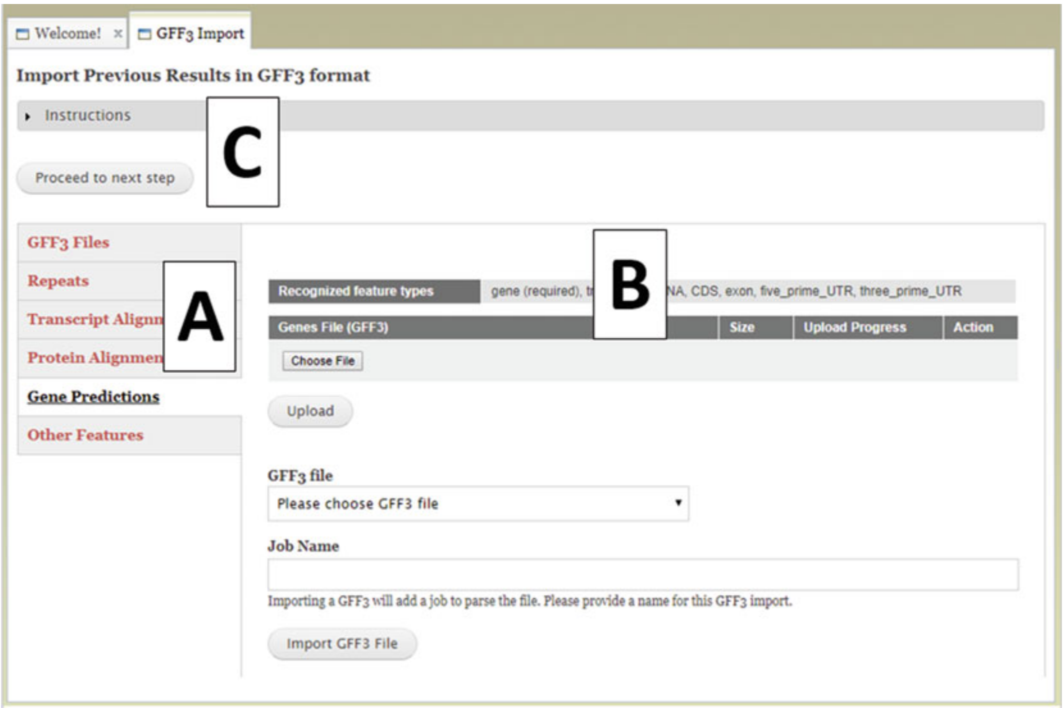


Fig. 4 The "Apollo" tab in GenSAS has the JBrowse interface on the left (A) and the Apollo interface on the right (B). Which tracks are visible is controlled through the "Tracks" Table (C) on the Apollo interface. During the "Annotate" step of GenSAS, user can drag gene models to the "User-created Annotations" track (D) and edit them

tools within GenSAS are in the GenSAS User's Guide (https://www.gensas.org/apolloJbrowse) as well as on the JBrowse and Apollo websites (Table 1).

GenSAS allows users to share their projects with other GenSAS users once the first job in the Job Queue completes. To share a project, click on the "Sharing" section of the accordion menu on the right, and then click "Share this project." A "Project Sharing" tab opens, and under the "Share this project" section, the name of the other user is entered. The owner of the project can grant the other user read only or full access to the project. With full access to the project, the other user can run annotation jobs and edit gene models with Apollo.

Most of the tabs in GenSAS have a similar layout. In general, if there are different job types or tools to select, there are clickable options on the left side of the tab (Fig. 5A) (*see* **Note 1**). The content in the center of the tab (Fig. 5B) will change depending on the option selected, and this is where job names are edited, tool settings are adjusted, or files are selected for upload. All the tabs have an expandable instructions section (Fig. 5C) that provides a brief overview of the annotation step. There are more detailed instructions in the GenSAS User's Guide (https://www.gensas.org/users_guide). There is also a "Proceed to next step" button under the "Instructions" section which moves the annotation process to the next stage.



Importing a GFF3 will add a job to parse the file. Please provide a name for this GFF3 import.

Fig. 5 An example of a GenSAS tab layout. Different options/tools are available to click on the left (A) and the content on the right (B) will change with each option. At the top of the tab is an "Instructions" section and a "Proceed to next step" button (C)

3.1.2 Genome Sequence Upload and Project Creation

The first steps of a GenSAS project involve loading files and creating a project (Fig. 6). Before starting a GenSAS project, the genome assembly file needs to be loaded by clicking on the "Sequences" arrow in the flowchart (Fig. 1A). On the Sequences tab, there are three options on the left side: Available Sequences, Upload Sequences, and Subset Sequences. "Available Sequences" just displays a table of sequences that the user has already loaded into GenSAS. Sequences from shared projects are not visible on this table to users who are not the owner of the sequence. Under "Upload Sequences," there is an interface to select a sequence file to upload and fields to select the sequence type (e.g., contig, scaffold, or pseudomolecule) and to enter the assembly version number (*see* **Note 2**). To help ensure that GenSAS users get quality results, the uploaded genome assembly metrics are determined using PRINSEQ (Table 1). GenSAS will only use assemblies that are 25,000 sequences or less and have more than 50% of the sequences over 2500 bases in length. Sequence files that do not meet these requirements are flagged and a file of just the sequences above 2500 bases is made available for use in the project, if desired. To use the file of sequences above 2500 bp, click on the "violated" label in the "Status" column of the Available Sequences table for

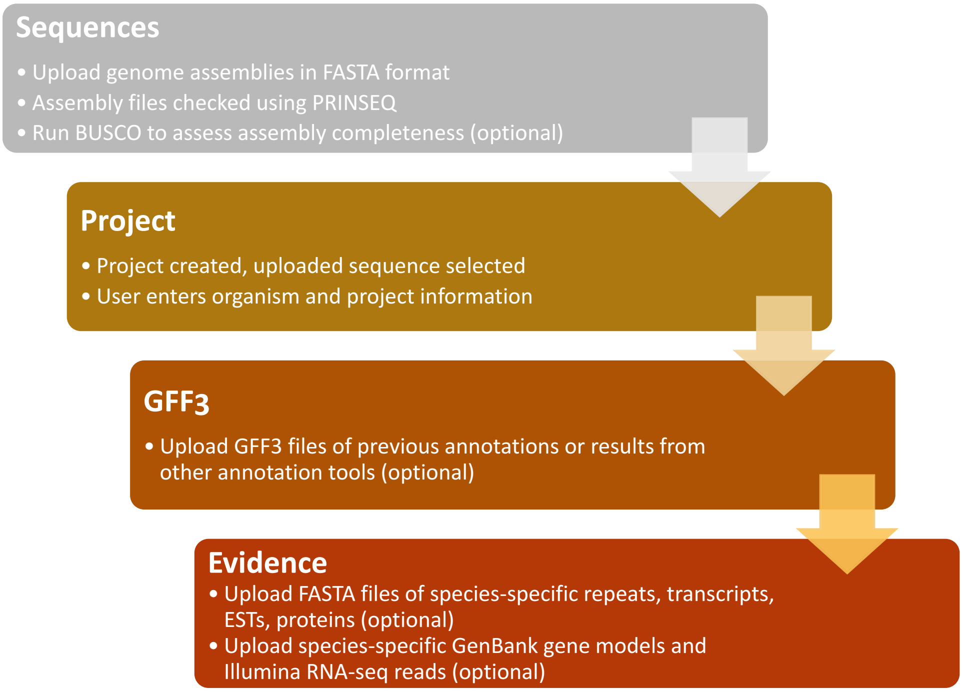

Sequences
• Upload genome assemblies in FASTA format
• Assembly files checked using PRINSEQ
• Run BUSCO to assess assembly completeness (optional)
Project
• Project created, uploaded sequence selected
• User enters organism and project information
GFF3
• Upload GFF3 files of previous annotations or results from other annotation tools (optional)
Evidence
• Upload FASTA files of species-specific repeats, transcripts, ESTs, proteins (optional)
• Upload species-specific GenBank gene models and Illumina RNA-seq reads (optional)

Fig. 6 An overview of the first steps of the GenSAS annotation process which include uploading files and project creation

that sequence set. A new tab opens and the option to use the filtered file is available to click. The "Subset Sequences" option allows users to select sequences from a previously loaded file by sequence name, or filter sequences by minimum size, and create a subset of sequences for use in a project. "Subset Sequences" is a good option to use to test or optimize the GenSAS workflow on a handful of contigs from their genome assembly before annotating the entire genome. Users can run BUSCO on uploaded genome assemblies by clicking on the "processed" label in the "Status" column for that sequence set in the Available Sequences Table. A new tab opens that has the stats from the PRINSEQ analysis and the option to run BUSCO is at the bottom. Users just need to select the appropriate BUSCO dataset and click "Run BUSCO" to start the job. When the job completes, the results will also be available on the tab where the BUSCO job was created.

Once the genome assembly is uploaded (*see* **Note 3**), a new project can be created by clicking on the "Project" arrow in the flowchart. On the "Project" tab, there are two options on the left side: "Load an Existing Project" and "Begin a New Project." Existing projects include those previously created by the user and projects shared by other GenSAS users. When the "Begin a New Project" option is selected, a web form with required and optional information fields appears. The required information includes project name, project type, genus and species, and selection of a sequence group. After the project is created, the "Project" tab displays summary details about the open project and allows users to reset the 60-day project expiration or delete the project. All GenSAS projects expire and are deleted after 60 days (Table 2), unless the user resets the expiration of the project from the Project tab. Users can reset the expiration on a project as often as needed to keep the project active. There is also a "Close this Project" button on the Project tab. If a current project is open, the project first must be closed in order to switch projects or create a new project.

3.1.3 Uploading Supporting Evidence to GenSAS

After project creation, the "GFF3" arrow of the flowchart becomes available. This is an optional step in the pipeline where supporting GFF3 files (discussed in Subheading 2.3) are uploaded to GenSAS. There are six options on the left of the "GFF3" tab: GFF3 Files, Repeats, Transcript Alignments, Protein Alignments, Gene Predictions, and Other Features. The "GFF3 Files" section is a list of all files previously uploaded by the user and can be selected for repeated use. The remaining five options are loaders for specific data types. All imported GFF3 files are visible in JBrowse as new tracks, and certain data types are also available for use in other steps of GenSAS. To load a GFF3 file, select the data type from the options on the left and then either upload a file, or select a previously uploaded GFF3 file, enter a name for the job, and then click "Import GFF3 File." A job will then appear in the Job Queue. After all GFF3 import jobs are started and in the job queue, or to skip this step, click the "Proceed to next step" button, and the "Evidence" step in the flowchart will be available to use. Users can return to the GFF3 step later if they would like to load more files and the import jobs do not have to be completed before moving to the next step.

FASTA files of species-specific repeats, transcripts, ESTs, and proteins as well as the GenBank files and RNA-seq reads discussed in Subheading 2.3 can optionally be loaded into GenSAS under the "Evidence" step. To load the FASTA files and GenBank genes file, first select the appropriate data type on the left side of the Evidence tab: Upload Repeat Libraries, Upload Transcripts & ESTs, Upload Proteins, or Upload Gene Structures. Then select the file(s) and click "Upload Files." To upload RNA-seq reads, select the "Upload Illumina RNA-seq." For RNA-seq files, the option is to load a paired set of read files or a single non-paired reads file (*see* **Note 2**). Once all the evidence files are loaded, click the "Proceed to next step" button.

3.2 Structural Annotation

3.2.1 Repeat Identification and Masking

The steps of structural annotation include repeat masking, aligning evidence to the genome, predicting gene structures, identifying tRNAs and rRNAs, and creating a consensus gene model set (Fig. 7). Under the "Repeats" step (optional), two repeat finding tools are available: RepeatMasker and RepeatModeler (Table 1). RepeatMasker relies on libraries of previously identified repeats to find repeats in the genome sequence. GenSAS provides the repeat collections from Repbase [6] de novo repeat finder and does not rely on previous evidence, which makes it especially good for non-model organisms where no repeat information is available. After the repeat jobs have completed running, the "Masking" step becomes available. Under the masking step, one or more masking jobs can be selected to create the masked consensus. If a GFF3 file of aligned repeats was imported, it will also be an option

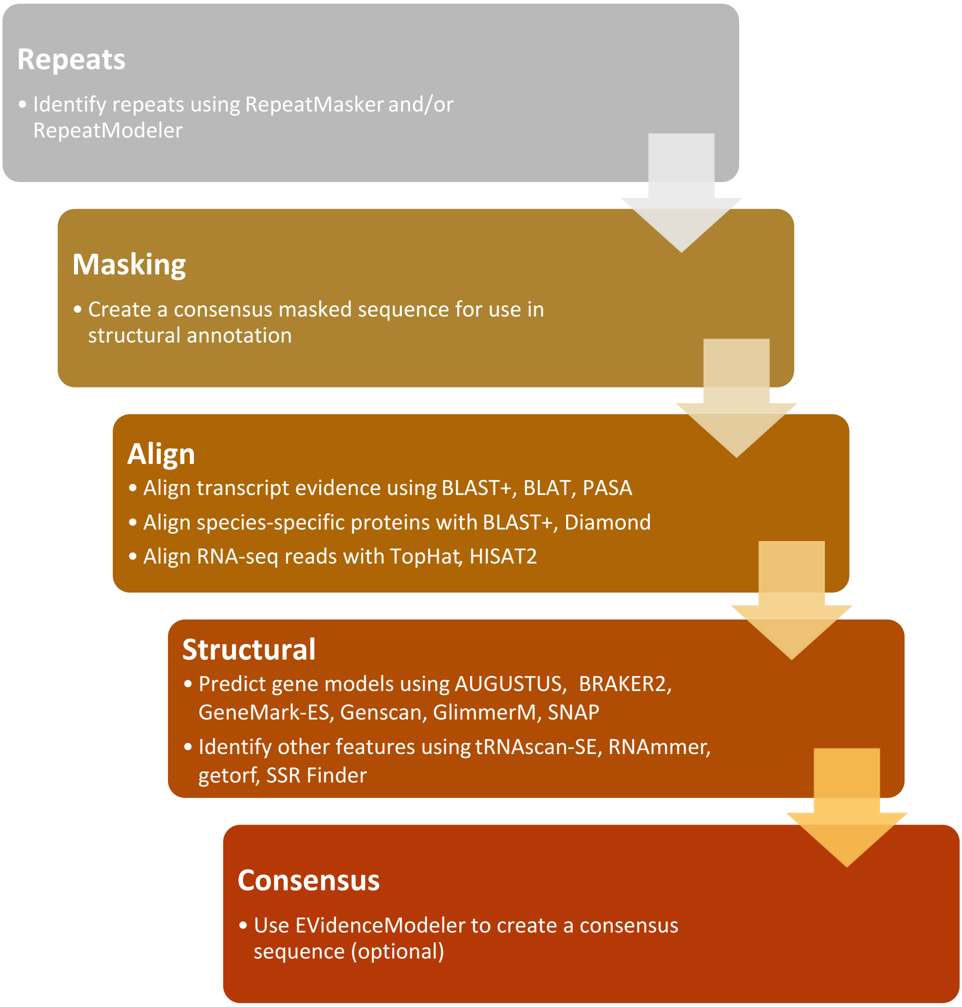

Repeats
• Identify repeats using RepeatMasker and/or RepeatModeler
Masking
• Create a consensus masked sequence for use in structural annotation
Align
• Align transcript evidence using BLAST+, BLAT, PASA
• Align species-specific proteins with BLAST+, Diamond
• Align RNA-seq reads with TopHat, HISAT2
Structural
• Predict gene models using AUGUSTUS, BRAKER2, GeneMark-ES, Genscan, GlimmerM, SNAP
• Identify other features using tRNAscan-SE, RNAmmer, getorf, SSR Finder
Consensus
• Use EVidenceModeler to create a consensus sequence (optional)

Fig. 7 An overview of the structural annotation steps of GenSAS

to use in the masked consensus. When the masked consensus is generated, GenSAS produces two versions. One version is hard-masked with the repeat region nucleotides converted to an "X." The other version is a soft-masked, and the repeat region nucleotides have been converted to a lower-case letter. The hard-masked sequence is used with the transcript and protein alignment tools in the "Align" step. For the compatible gene prediction tools under the "Structural" step, the user has the option of using the soft-masked sequence as the input or using the default hard-masked sequence. Once the masked consensus job is complete, the "Align" step will be available. If no repeat masking is desired, the unmasked sequence can be used in subsequent steps by not setting up a masked consensus job and just proceeding to the next step.

3.2.2 Alignment of Transcript and RNA-seq Evidence

The optional "Align" step has tools for aligning evidence to the genome sequence. Alignments of transcript and protein evidence can be useful when generating a genes consensus later in the annotation process and can be helpful during manual curation of the annotation. User-provided transcript evidence can be aligned using BLAST+, BLAT, or PASA (Table 1). For transcript alignments, GenSAS also has the NCBI RefSeq [7] transcript sets available for use. User-provided protein evidence can be aligned using BLAST+ and Diamond (Table 1). For the BLAST+ tool, users can also adjust the settings to change the specificity of the alignment. TopHat2 and HISAT2 (Table 1) are used to align the user-provided RNA-seq reads. Once all alignment jobs have been set up, the "Proceed to next step" button can be clicked to move on to the "Structural" step. Alignment jobs do not have to be completed to move on to the next step; however some tools under the "Structural" step do use results from the alignment tools. The alignment jobs need to complete before the results are available for use in the downstream annotation steps.

3.2.3 Gene Prediction and Other Structural Features

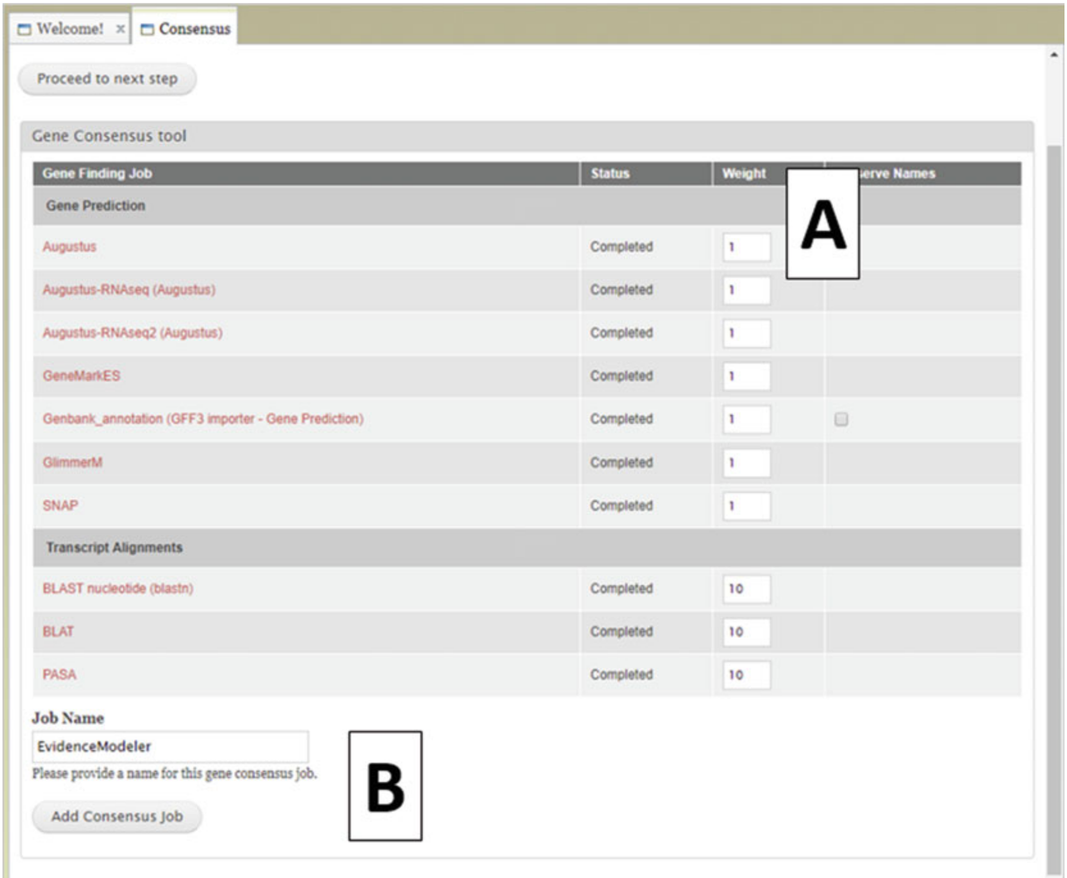
The "Structural" step has tools for gene prediction and for identifying other genetic elements. On the "Structural" tab, there are two options on the left side: "Gene Prediction" and "Other Features." Clicking on these options changes the visible list of available tools. Under "Gene Prediction," there are several tools to choose from. Some can be trained, while others rely on pre-set organism profiles. AUGUSTUS (Table 1) can either be used with the provided pre-trained datasets from model organisms or can be trained using user-provided evidence. To train AUGUSTUS, open the "Options for training AUGUSTUS" section under the AUGUSTUS setting page. There is the option to select four different file types, and in some cases, AUGUSTUS requires specific combinations of these options to work properly (Table 5). If the proper file combination is not selected, an error message will be displayed when the job is submitted. The BRAKER2 (Table 1) tool

Table 5
Data type combinations needed to train AUGUSTUS

Training option	Required data type to select in GenSAS (user-provided file type)
Genes and transcripts	"Gene Structures" (GenBank file) *and* "cDNA sequences" (FASTA file)
Proteins only	"Protein Sequences" (FASTA file)
Proteins and transcripts	"Protein Sequences" (FASTA file) *and* "cDNA sequences" (FASTA file)
RNA-seq reads	"BAM File," select results of TopHat2 job

can be trained with aligned RNA-seq evidence. For non-model organisms without any supporting evidence, a good tool might be GeneMark-ES (Table 1) which performs self-training. The three other gene prediction tools Genscan, GlimmerM, and SNAP (Table 1) have pre-installed profiles for model organisms. GenSAS will also parse user-uploaded results from FGENESH [9] for use in the annotation process, but FGENESH cannot be run on GenSAS due to license restrictions. Under the "Other Features" section of the "Structural" tab, there are four tools: getorf, RNAammer, SSR Finder, and tRNAscan-SE (Table 1). RNAammer identifies rRNAs, tRNAscan-SE finds tRNAs, SSR Finder identifies simple sequence repeats, and getorf finds open reading frames.

For all the structural annotation tools, once a job is submitted, the job name appears in the Job Queue. And as with other GenSAS jobs, each tool can be run multiple times, with different settings, provided each job has a unique name. It is very important once all these jobs complete that the results are critically looked at by the user. Some of these tools may perform better on certain genomes than others, and if the results are poor, then omitting those results from downstream steps is highly recommended. This is very important for the last part of the structural annotation process (Fig. 7), the "Consensus" step. During the consensus step, EVidenceModeler (EVM, Table 1) can be used to create a merged consensus gene set. EVM allows the user to assign weights to each data track that is used to generate the consensus. Higher weights (i.e., 10) indicate that the data are more experimentally based and should be trusted more. Lower weights (i.e., 1) indicate that the data are more theoretical or from mathematical predictions and might not be as accurate. All available tracks are present on a table on the "Consensus" tab (Fig. 8). GenSAS pre-populates the weights (Fig. 8A) and gives transcript alignments a weight of 10, protein alignments a weight of 5, and gene prediction tool results a weight of 1. The dataset weights can be edited by the user and if the user wants a track to be omitted just remove the weight from the box and leave it blank. EVM can be run multiple times with different weight settings if each job is assigned a unique name (Fig. 8B).



Gene Finding Job	Status	Weight	ierve Names
Gene Prediction			
Augustus	Completed	1	
Augustus-RNAseq (Augustus)	Completed	1	
Augustus-RNAseq2 (Augustus)	Completed	1	
GeneMarkES	Completed	1	
Genbank_annotation (GFF3 importer - Gene Prediction)	Completed	1	☐
GlimmerM	Completed	1	
SNAP	Completed	1	
Transcript Alignments			
BLAST nucleotide (blastn)	Completed	10	
BLAT	Completed	10	
PASA	Completed	10	

Job Name

EvidenceModeler

Please provide a name for this gene consensus job.

Fig. 8 The "Consensus" tab in GenSAS. Dataset weights (A) can be adjusted by the user before submitting the EVidenceModeler job. Users can submit multiple EVM jobs with different settings by editing the Job Name (B)

3.3 Functional Annotation

The functional annotation portion of GenSAS (Fig. 9) begins by selecting the Official Gene Set (OGS). The OGS is the gene model set on which the functional annotation tools will be run, manual curation can be performed on, and that the final annotation files will be generated from. The OGS is selected by the user from a list of available gene sets on GenSAS. This list includes gene predictions uploaded by the user at the "GFF3" step, results from the tools under the "Gene Predictions" section of the "Structural" step, and any jobs created at the "Consensus" step. It is up to the user to evaluate the results and to select the gene set that makes the most sense for the organism being annotated. Once an OGS is selected, the "Refine" step becomes available for use. Under the refine step, there is an option to run the OGS through PASA with transcript evidence to help further refine the gene structure junctions and start and stop positions. This step is optional, and it has

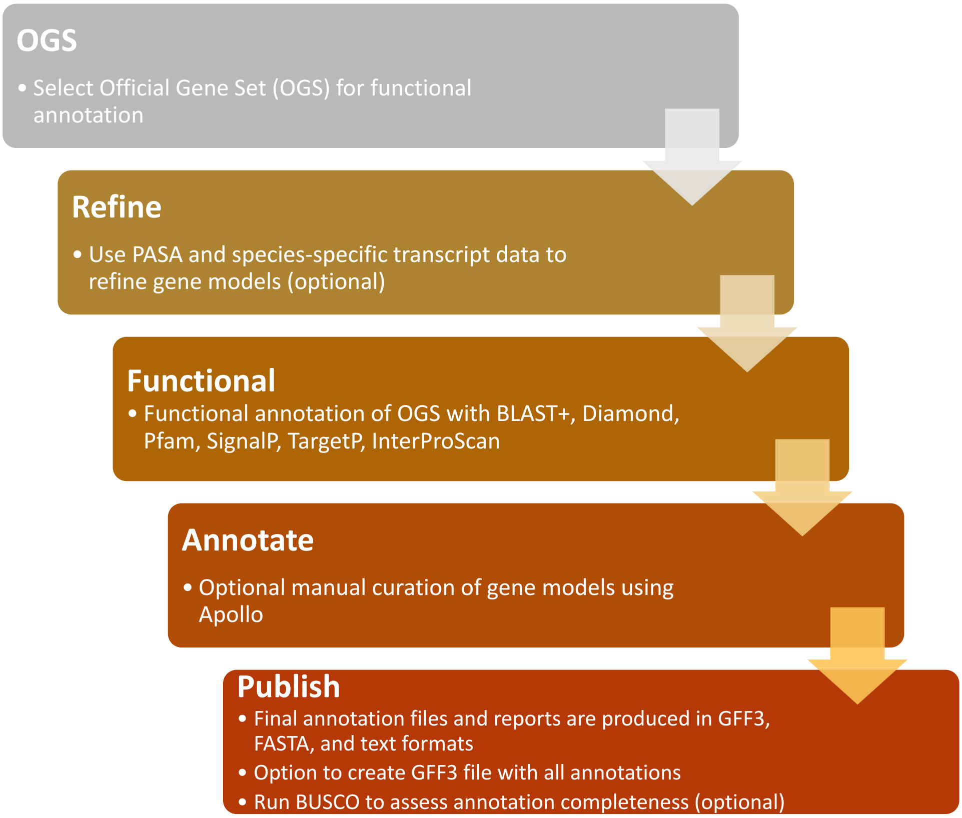
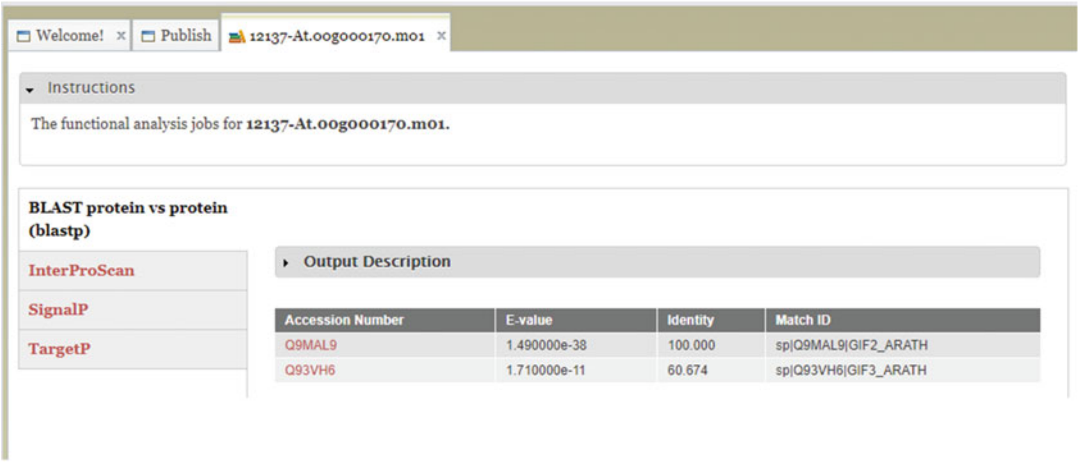

OGS
• Select Official Gene Set (OGS) for functional annotation
Refine
• Use PASA and species-specific transcript data to refine gene models (optional)
Functional
• Functional annotation of OGS with BLAST+, Diamond, Pfam, SignalP, TargetP, InterProScan
Annotate
• Optional manual curation of gene models using Apollo
Publish
• Final annotation files and reports are produced in GFF3, FASTA, and text formats
• Option to create GFF3 file with all annotations
• Run BUSCO to assess annotation completeness (optional)

Fig. 9 Overview of the functional annotation, manual curation, and final steps of a GenSAS project

been observed that this step only works well with transcript evidence from the same organism as the genome being annotated.

The “Functional” step is where functional annotation tools are run on the predicted proteins of the OGS. Jobs can be created by clicking on the five tool names on the left: BLAST+, Diamond, InterProScan, Pfam, SignalP, and TargetP (Table 1). For protein alignments with BLAST+ and Diamond, the available databases include SwissProt [8], TrEMBL [8], NCBI RefSeq proteins [7], and any user-provided protein files. Please note that when using a large protein database (i.e., TrEMBL), in conjunction with a large genome, the BLAST job will take quite a while to complete. The remaining four tools identify functional domains within the predicted proteins. Functional annotation jobs also appear in the Job Queue once submitted, but the results do not appear in JBrowse/Apollo as individual tracks like the structural annotation tools since the tools are only run on the predicted proteins, and not the entire genome sequence. To view the functional annotation results for the OGS genes, either open the results tab for each tool by clicking on the job name in the Job Queue and click on the mRNA name on



The functional analysis jobs for 12137-At.00g000170.m01.

Accession Number	E-value	Identity	Match ID
Q9MAL9	1.490000e-38	100.000	sp|Q9MAL9|GIF2_ARATH
Q93VH6	1.710000e-11	60.674	sp|Q93VH6|GIF3_ARATH

Fig. 10 Example mRNA details tab, which has links to the functional annotation job results on the left side. Clicking on each link displays the results for that mRNA

the summary table or right-click on the gene model in JBrowse and select "View putative annotation." Either of these methods will open the functional results tab for that mRNA (Fig. 10). As with the other GenSAS tabs, the results from each functional annotation tool can be selected on the left side, and when clicked, the content in the tab will change to display those results.

3.4 Manual Curation

After the "Functional" step, the "Annotate" step is available. At this step, the "User-created annotations" track, that is part of Apollo, is available to edit in JBrowse. Apollo allows users to manually curate the OGS prior to producing the final annotation files. With Apollo, users can edit intron-exon junctions, start and stop locations, and UTR lengths and add functional annotation notes. While manual curation is an optional step, it is highly recommended. Apollo was designed to allow for collaborative manual annotation efforts between many users and keeps track of which users have made edits. The sharing function of GenSAS allows for users to share their GenSAS project with other GenSAS users allowing those users to also do manual curation in Apollo. For more detailed directions on the manual curation functions of Apollo, please see http://genomearchitect.github.io/users-guide/. There is also a brief example of how to perform manual curation in the GenSAS User's Guide (https://www.gensas.org/annotate).

3.5 Final Annotation Files

When the annotation process is complete, the final files are produced under the "Publish" step. During the Publish step, GenSAS will merge any manually curated genes from Apollo into the OGS and run the functional annotation tools on the manually edited gene models. GenSAS will then rename all the gene models with a consistent naming scheme and add the assembly and annotation versions to the file names. GenSAS automatically selects the minimum files needed, such as all the FASTA and GFF3 files associated with the OGS and masked consensus. Users can also select specific tools and have GenSAS prepare the output files from those results as well. Functional annotation results are output as tab-delimited files. Please note that if any changes to the project are made in the previous steps of GenSAS, after the Publish step has been run, the Publish step needs to be run again to produce the newest version of the annotation files. Users also have the option to run BUSCO on the predicted proteins to assess the completeness of the annotation. GenSAS also produces summary reports related to the final annotation features and the tools that were used to produce the annotation. A summary table of genome annotation metrics (e.g., number of genes, CDS, mRNA, tRNA, rRNA, etc.) is produced with custom scripts, and a summary of the type of repeats present in the repeat consensus is generated using a script called "One code to find them all" (Table 1). GenSAS also produces a summary report of which tools were used to create the OGS and functional annotations and the tool versions and settings.

4 Future Development

The GenSAS development team is constantly looking for ways to make GenSAS better for the user, and feedback from the users drives the improvement of GenSAS. A couple more tools that will be added include the alignment tool GMAP [10] to provide more options for transcript alignments and the gene prediction tool MAKER2 [2].

5 Notes

1. We recommend using Chrome, Firefox, and Edge Internet browsers with GenSAS. Some users of the Safari Internet browser have had issues with the GenSAS interface displaying properly. The GenSAS interface may not display properly if the "zoom" function of the browser program is being used but will appear normal at 100% magnification.
2. Files over 2 GB in size might not load through the web interface. If you encounter problems loading large files to GenSAS, please contact us (https://www.gensas.org/contact), and we will provide a secure FTP location for file transfer. The FTP can also be used to transfer large files out of GenSAS if needed.
3. Once assembly files are uploaded to GenSAS, the file needs to be processed before it can be used in a GenSAS project. For

larger genomes, this takes a bit of time, and it may take several minutes before the genome assembly is available to select for project creation.

Acknowledgments

We acknowledge with thanks funding from the US Department of Agriculture—National Institute of Food and Agriculture (NIFA) award number 2011-67009-30030, NIFA NRSP10 award, and Washington State University (WSU). We also acknowledge the WSU Hort 550 (Bioinformatics for Research) graduate classes of 2015, 2016, and 2017 for testing and providing feedback about GenSAS.

References

1. Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13(5):329–342. https://doi.org/10.1038/nrg3174
2. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491. https://doi.org/10.1186/1471-2105-12-491
3. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6):863–864. https://doi.org/10.1093/bioinformatics/btr026
4. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351
5. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35(3):543–548. https://doi.org/10.1093/molbev/msx319
6. Bao WD, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11. https://doi.org/10.1186/s13100-015-0041-9
7. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189
8. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169. https://doi.org/10.1093/nar/gkw1099
9. Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7:S10.1–S1012. https://doi.org/10.1186/Gb-2006-7-S1-S10
10. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21(9):1859–1875. https://doi.org/10.1093/bioinformatics/bti310
11. Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465–W467. https://doi.org/10.1093/Nar/Gki458
12. Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32:

W309–W312. https://doi.org/10.1093/Nar/Gkh379

13. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32(5):767–769. https://doi.org/10.1093/bioinformatics/btv661
14. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18(12):1979–1990. https://doi.org/10.1101/gr.081612.108
15. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res 33 (20):6494–6506. https://doi.org/10.1093/nar/gki937
16. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268(1):78–94. https://doi.org/10.1006/jmbi.1997.0951
17. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. Genomics 59(1):24–31. https://doi.org/10.1006/geno.1999.5854
18. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59. https://doi.org/10.1186/1471-2105-5-59
19. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST plus : architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
20. Kent WJ (2002) BLAT - the BLAST-like alignment tool. Genome Res 12(4):656–664. https://doi.org/10.1101/Gr.229202
21. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1):59–60. https://doi.org/10.1038/nmeth.3176
22. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31 (19):5654–5666. https://doi.org/10.1093/Nar/Gkg770
23. Kim D, Landmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–U121. https://doi.org/10.1038/Nmeth.3317
24. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36
25. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35 (9):3100–3108. https://doi.org/10.1093/nar/gkm160
26. Lowe TM, Chan PP (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44(W1):W54–W57. https://doi.org/10.1093/nar/gkw413
27. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5):955–964. https://doi.org/10.1093/Nar/25.5.955
28. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 9(1):R7. https://doi.org/10.1186/Gb-2008-9-1-R7
29. Jones P, Binns D, Chang HY, Fraser M, Li WZ, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240. https://doi.org/10.1093/bioinformatics/btu031
30. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1): D279–D285. https://doi.org/10.1093/nar/gkv1344
31. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785–786. https://doi.org/10.1038/nmeth.1701
32. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953–971. https://doi.org/10.1038/nprot.2007.131
33. Bailly-Bechet M, Haudry A, Lerat E (2014) "One code to find them all": a perl tool to conveniently parse RepeatMasker output files.

Mob DNA 5:13. https://doi.org/10.1186/1759-8753-5-13

34. Unni D, Dunn N, Yao E, Buels R, Li Y, Holmes I, Elsik C, Lewis S (2017) GMOD/Apollo: Apollo2.1.0(JB#d3827c) (Version 2.1.0). Zenodo. https://doi.org/10.5281/zenodo.1295754

35. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. Genome Res 19 (9):1630–1638. https://doi.org/10.1101/gr.094607.109