

Annotation Guidelines

Details Highlighted in Yellow vary according to server and Organism specifics

Genome portal address: <http://genome.jgi-psf.org/GenspV>

NOTE: You must be granted access to the annotation tools in order to annotate!

Annotators

If you have been granted access to a password-protected genome portal then you will automatically have annotation privileges. If the genome portal is publicly available, you will need to be granted access to the tools. All annotators must have a registered account. You may register here: <https://signon.jgi-psf.org/register> Once you've registered an account you will need to request access for a particular genome portal.

An updated list of annotators is generally maintained at a **Community maintained site**. This list will include the name of the annotator, their affiliation and contact information, the specific area of **Organism** biology that they represent, and the gene families or the pathways that the annotator intends to cover. Please contact the **Principal Collaborator (contact information is maintained on the Organism portal)** to obtain the list.

Annotation process

The Joint Genome Institute generated gene models for protein-coding genes using several algorithms. In an attempt to cut down on the redundancy, we filter the models based on homology (percent id and percent cover, completeness of model) and EST support to create a virtual track, called **FilteredModels**, which represents those choices. Models initially chosen for the FilteredModels track are automatically entered into the GeneCatalog, which is represented on the browser as the **GeneCatalog** track. This GeneCatalog track will eventually constitute our reference list of gene models, which will be submitted to GenBank. The annotation process consists of deciding whether the chosen model in the GeneCatalog track is correct, and if not, removing it from the track and promoting an alternative gene model or creating a new one. Once manual curation starts, the curators go through and search for gene models of interest. You are in charge of curating various annotation fields, as described below, and recording interesting findings that describe the genome biology of **Organism**.

Manual curation involves the following activities:

Curate gene model structure

1. Promote a model from the pool to the catalog (if missing or incorrect)
2. Demote a model from the catalog (if incorrect or another represents the locus better)
3. Report incomplete models, gene duplicates in the genome and pseudogenes
4. Create/fix a new model using TrackEditor (if missing or incorrect)

Curate functional annotation

1. Assign gene name (synonym)
2. Assign/correct function/description/define
3. Correct/add gene ontology

Describe the gene in relation to its membership/phylogeny within gene families

1. Report gene family membership
2. Produce amino-acid alignment with the other family members
3. Produce phylogenetic tree of the genes to better understand their homologies

Manual Annotation and Curation

Where should I start?

- * [Visit the JGI Help! page](#) to get familiar with Genome Portal tools
- * Read this document to get familiar with annotation process
- * Register as an annotator
- * Find your genes of interest
- * Assess gene structure of the model in GeneCatalog and other models at the same locus, pick the best one or create a new model
- * Go to annotation page to promote/demote gene model and assign name and function

How do I find a gene of interest?

There are several different entry points into the portal to find your gene of interest.

- * KEGG browser – you can search or browse through KEGG metabolic and regulatory pathways to retrieve information about enzymes, pathways and proteins related to JGI-predicted genes.

* KOG browser – KOG is a eukaryote-specific version of the Clusters of Orthologous Groups (COG) tool for identifying ortholog and paralog proteins. The KOG browser provides a way to find JGI-predicted genes by KOG classification or ID.

* GO browser – you can browse through the Gene Ontology (GO) and find genes related to GO terms. GO arranges biological terms under three organizing principles: molecular function, biological process and cellular component.

* Advanced search – this tool allows you to locate genes, transcripts and proteins in one of two ways: by directly searching for keywords contained in user- and automatically-created annotations, or by indirectly searching for keywords characterizing various protein sequences that have high-scoring alignments with the transcripts of interest. Detailed information on how to perform these searches can be found on the advanced search page.

* BLAST/BLAT search – if you have a known sequence, you may use this tool to align your sequence of interest to the genomic sequence, the FilteredModel set or to all of the gene predictions generated for **Organism**.

More detailed information for these tools can be found on the [JGI portal Help! page](#).

How do I know the selected gene model is good or not?

Compare the selected model with:

1. available supporting evidence you see at a particular locus on either the [genome browser](#) (ESTs, homology, genome conservation, proteomics), or on the [protein page](#) (alignments, domains, completeness) or through any custom analysis you would like to do (eg, Blast).

2. other models in different tracks generated by gene predictors used in automated annotation (you may find up to 10 different models per locus).

Pick the best available model. If it is different from the one in the GeneCatalog, replace it by changing the disposition on the [annotation page](#). If no available models are good enough, you can create your own in the user track using the [Track Editor tool](#).

To begin, prioritize your annotation effort. Gene models should be attacked in declining interest based on their possible functions in biological processes and based on the amount of supporting data. For example, protein features on the genome scaffolds can have (1) good support – including some mix of computational predictions, tBlastn

matches to selected proteins, gapped (Blat) EST matches, or Blastx matches, (2) unmatched protein features that have protein support but lack sequence similarity to an annotated protein, (3) weak features that lack protein support but can have numerous good EST matches.

Next, determining the best available track in the genome browser. For a protein feature with sequence similarity matches to other proteomes, this process consists of identifying the best blastx match to the scaffold sequence or the best Blastp match to a predicted protein sequence.

Using the best tracks as guides, modify the boundaries of the FilteredModel.

- (1) [Should the gene model be merged with, or split from, another locus?]
- (2) [Is the gene model missing untranslated regions?] (Is there evidence that more scaffold sequence is transcribed upstream or downstream from the edges of the gene model?)
- (3) [Are there distant exons at either ends of the gene model?]
- (3) [Are the intron and exon boundaries correct?]
- (4) [Have all alternative splice variants been identified?]
- (5) [Are there degenerate genes annotated as protein coding loci?]

A note on pseudogenes

It is far easier to identify a pseudogene than to demonstrate that a locus is a functional gene. Real genes are often outnumbered by their pseudogenes. However, a very recent pseudogene may be hard to distinguish from a working gene; there are cases of long ORFs with ATG start and a single stop codon, frequent transcription, yet have defective translation initiation.

Pseudogenes of moderate age are usually easily recognizable. They have accumulated many defects across the coding sequence but not so many that homology to the parent gene is blurred. A single internal stop codon is not completely persuasive as there are cases of mRNA editing. Plus, draft genome sequences do contain errors.

Older pseudogenes may have drifted off to the point of marginal recognizability. Determining a good parental sequence and event boundaries become problematic. Nonetheless, ancient pseudogenes can still establish whether a given protein domain was present at the time of formation – these are still identifiable long after point mutations have largely obliterated alignments. (Absence of a domain might also mean retro-insertion of an alternatively spliced mRNA.)

How do I replace the model in GeneCatalog?

Use the "disposition" field on the annotation page to promote or demote models to and from the GeneCatalog. You can enter:

- * "Catalog" if the model is correct
- * "Demote" if it is not a gene or if you want to replace it with a different model

Please do not forget to "Demote" the incorrect models, otherwise they will appear concurrently with the correct ones.

Note: The "Demote" option does not delete the model or its annotation from the database. It simply removes it from the Catalog track.

How do I create a new/better model?

First, expand all model tracks and evaluate all models at the same locus. Likely, a better model was already generated by one of gene predictors but, for some reason, was not promoted to the GeneCatalog. If none of the models is good, create your own in the user track using a user-friendly interface described at <http://Genome.jgi-psf.org/Tutorial/tutorial/editor.html>

The Track Editor tool allows you to:

- * Create a new model by copying an existing model
- * Edit a new model
- * Add existing exons to a new model
- * Create an *ab initio* model

Once editing is finished, protein analysis will automatically be run on the newly edited model only if the model has been released. Releasing a model does not automatically add it to the GeneCatalog. You must go to the annotation page (or model web page) and set the disposition to "Catalog". The automatically catalogued gene model will remain in the GeneCatalog at that locus unless it is manually deleted.

How do I access the annotation page?

There are multiple ways of accessing annotation pages

- 1) Through the gene model's Protein page

Just click on the protein page link where it says "view/modify manual annotation". This takes you to the Transcript Annotation tool, which displays annotation information for the gene. If you are a registered annotator, you can annotate the genome with information about the gene you are viewing.

2) Via Advanced Searching directly against annotations

The returned results include a relevance score and several links. For a gene or transcript, the following links appear:

- * "P" Link to the Protein page for the gene or transcript
- * "T" Link to the transcript annotation page
- * "G" Link to a Genome Browser-based view of the gene or transcript

3) Through the GO/KEGG/KOG functional tools

You can see a list of the predicted gene models' products (proteins) in the JGI genome that are assigned to a particular functional category – GO+children / metabolic pathway / orthologous gene cluster, etc.

For a gene or transcript, the following links appear:

- * "P" to see the Protein page for a gene product
- * "A" Link to the transcript annotation page

How do I annotate a gene?

You must be granted access to the annotation tools in order to enter and edit manual curations. If you do not have access please contact the **Principal Collaborator** for permission.

The "transcript" page offers several fields that can be filled individually by clicking on the add link (or edit if you have already entered annotation for this gene). When completing these fields, bear in mind that your annotation will eventually be translated into a Gene entry in NCBI. [Click here for an example of a NCBI Gene entry](#)

****IMPORTANT NOTE:**** If another annotator has already entered a curation into one of the fields please DO NOT add another entry. If you disagree with their entry please contact them to discuss this. Their user id, which appears under the "Creator" column, is linked to their email address to facilitate contact. If there is an automatic entry that you disagree with you should then edit that so that there is a single entry only.

TRANSCRIPT ANNOTATION							
Tutorial/scaffold 7:1126936-1128699 e_gw1.7.242.1 Hide							
Attribute	Value	Creator	Action	Assessment	? Note	Creator	Action
Name			add	EST evidence			add
Description			add	Needs GO			add
Model Notes			add				
Define			add				
Disposition	Catalog	AUTOMATIC	add edit				
Literature			add				
Evidence	Type	Creator	Action				
			add				

Name – (known as the “gene” in GenBank entries) The gene name nomenclature is organism specific, so for internal consistency refer to your community’s standards. Please do not deviate from the gene naming rules. If there are concerns, please send your comments to the **Principal Collaborator** by email. Names must be unique and you will not be allowed to enter a gene name if that name has already been assigned to another gene model. Do not assign a name if the gene function is unknown. If no names are assigned to a gene, a numerical identifier (locus_tag) will be automatically assigned to every gene at the time of GenBank submission (e.g. GENSP_12300 where “12300” is the JGI protein id for that gene).

Define – (known as the “product” in GenBank entries) Unlike the gene name, this field is mandatory. It should be a short (<85 characters) precise description of the gene and gene product, and if possible, it should include the gene's main function(s). It should include the full standard name of the protein, but no acronyms, EC numbers, or species names, all of which can be listed in description field. Very often, the define of a related entry in Swissprot can be used.

Define rules and examples

There is no formatting (*italics*, **bold face**, underline) nor punctuation possible for Define entries. The species and gene name (e.g. *Genus species* GENSP_010204) will be added automatically to create the final define, so please do not enter these designations.

- If genes encode proteins for which experimental evidence is present the Define should then list the function, e.g. pyruvate carboxylase
- For most genes in the Organism genome, the function is deduced by sequence similarity to functionally characterized genes in other genomes. In such cases the Define should specify similarity, e.g.

pyruvate carboxylase-like protein *****Please note: the words “similar to” and “putative” are no longer accepted by NCBI. If you enter “similar to pyruvate carboxylase” in the defline it will be changed to “pyruvate carboxylase-like protein” when submitted to GenBank.**

- If genes have unreliable similarity deduction but contain a conserved domain, the Defline should read "pyruvate carboxylase domain-containing protein".
- For genes encoding proteins with unknown function, Deflines should read "hypothetical protein".
- For pseudogenes, the Defline should read "xyz expressed pseudogene" if there is EST evidence for its transcription or "xyz pseudogene" if there are no transcripts.

****A note about alternative splice variants:** If you promote two or more models to the GeneCatalog at a given locus then the defline **MUST** specify that these are variants. The defline should end with “variant 1” , “isoform A” or “homolog 1” for the first splice variant and “variant 2”, “isoform B” or “homolog 2” for the second splice variant.

Gene Descriptions, Model notes, Dispositions and examples

Description (known as the “note” in GenBank entries) can be as detailed as needed, provided that the information is accurate and useful to researchers not familiar with the type of protein. Include information about the functions of the protein, its domains, splicing variants, interactions or subcellular location, comments about its phylogenetic origin, relationship to paralogs and orthologs (though NCBI will not accept gi ids or accession numbers so don’t include those), clustering with genes of related function, or overlap with neighboring genes, etc. With proper caution, you can input your gene function and evolutionary hypotheses, suggesting further study that should be pursued.

Example of a gene description:

[Function] Metallothionein regulates the intra-cellular concentrations of essential metals and offers protection from metal poisoning by binding free metal ions or undergoing exchange reactions with metals bound to other ligand. [Regulation] MTN is regulated by metals, with MTN

transcript increasing in response to elevated metals in tissues. [Domain] The 18 cysteine residues are arrayed in characteristic Cys-xaa-yaa-Cys, Cys-x-Cys, and Cys-Cys motifs, establishing it as a class 1 metallothionein. [Splicing] There are two identified MTN4 gene transcripts (MTN4-1, MTN4-2). These transcripts share CDS and 5' UTR. The 3'UTR of MTN4-2 is expanded, which contains 2 introns that are not shared with MTN4-1. [Phylogeny] Four MTN loci are identified. MTN1 and MTN4 are duplicates. Organism MTN genes are monophyletic within the Metallothionein phylogeny that includes homologues from 2 insect species and from 8 crustacean species (total of 21 aligned amino acid sequences).

Literature as bibliographic references documenting the gene can also be placed here. Rather than whole reference, use the Pubmed identifiers (PMID) directly copied from the bottom of the Entrez page. For example: PMID: 12185496 . This will appear as a clickable link on the JGI transcript page, and later in the Bibliography section of the Gene entry.

Model notes are not indexed for searching nor are they submitted to GenBank, but are useful for any detailed analysis. If the model appears correct, choose "no issue" from the pull-down menu. Otherwise, you can either change the model (see above), or simply place your comments here, for future use by yourself or others. Indicate whether you suspect misassembly, e.g. "C-terminus probably represented by xxx-xxx on scaffold zzz". State how the model needs to be modified, e.g. "6th intron not supported by EST data", "probably misses an N-terminal extension of 120-150 residues", or "should be fused with upstream gene model", etc. You can also place here a FASTA (<80 character lines) of your version of the transcript or predicted protein. In case of splicing variants / alternative transcription starts, you should generate a new model AND describe variants in the model notes (if biologically significant, enter also in the description). As usual, be concise and precise. Any notes about the structure of the model should be followed up by using the Track Editor tool to edit the gene model structure as noted.

Disposition field is used to decide whether the gene model should appear in the Catalog. You can enter "Catalog" if the model is correct and should be included in the GeneCatalog, or "Demote" if you do not want to include it in the GeneCatalog and you want to replace it with a model from another track. Do not forget to "Demote" the erroneous models, otherwise they will appear concurrently with the correct ones in the GeneCatalog. In the case of alternative splicing, you may promote two or more gene models to the GeneCatalog at the same locus. If you do so, the name and/or define MUST indicate that the models are splice variants.

Note: "Demote" does not delete the model or its annotation; it just removes it from the Catalog track.

EST evidence If in your opinion the gene model is supported by available ESTs please choose "Yes" in this field.

How do I keep track of curated models?

Search page if you already worked with a gene in the Portal and know its model name, protein or transcript identifiers, you can enter them here.

Make a note of your most interesting/bizarre findings, for it may be useful when it comes to writing your manuscript and may be selected for inclusion in the main genome paper. For example, a new gene family, an unusual domain structure, evidence for horizontal gene transfer, an unsuspected metabolic pathway, functional gene clustering, shortest exon, alternatively spliced variants with biological significance, overlapping genes, gene in an intron, etc... whatever deserves attention!!

Keep a list of interesting gene model names, protein page URLs or links created for the browser. Using the following link, you can access all models you have curated so far:

<http://genome.jgi-psf.org/cgi-bin/showMyAnnotation?db=GenspV>

Functional Annotation

The functional annotations of genes are based on the Gene Ontology and KOG classifications. Please refer to <http://www.geneontology.org> and http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12969510&dopt=Citation for details. At the time that we transfer the annotated genome sequence to NCBI, information will be fed automatically from the Automatic Ontology field, unless you fill in this field, so enter data here only if the automatic assignment has failed (false positives are not infrequent). Although this categorization may seem crude or even pointless at times, it is our only tool to automatically group genes by function, pathway etc. So please consider the functional annotation an important part of your effort to make the genome a useful tool in the long run.

Note: No online tool is available for annotating RNA genes at the moment.

A large body of automatic annotation is already available on the protein and transcript pages. Carefully read the **JGI Help!** page. It gives an excellent overview of the many tools available to identify the function of a

gene. The Search page allows you to search all models (searches the annotation of the protein hits). It also allows searching the items in the various alignment tracks and the gene models by name. The Advanced Search page searches only the Catalog and Filtered Model tracks. Fields searched include the automatic annotation and model name, plus the user-entered description, defline and gene name. Use the GO, KEGG and KOG pages for lists of genes that have been automatically assigned to a pathway or function.