
Postia placenta
Jamboree
Asilomar
March 18th and 19th
2007

Annotation guidelines for Postia placenta genome

Welcome to genome annotation. Three approaches to finding a biological story in a tangle of sequence data are to find protein family differences, to reconstruct biochemical pathways and identify whole genome differences. While we are searching for and collecting this data, we "annotate" the gene models so that you and others are able to gain meaningful information. The Genome Portal features a graphical interface created for the genome, and it provides a starting point for our annotation.

Annotation Fields

Name: This is a contentious issue, the *name*. Names usually have a 3 – 5 letter or number with number or letter version designation, like hsp70a. This gene name was assigned based on experimental evidence in *Drosophila*. However, we are not doing that here. The presence of a hsp70a-like gene model in our genome is not sufficient evidence for identical function to the *Drosophila* heat shock protein. Thus, we are going to use a place holder, until which time some brave researcher experimentally establishes function. This place will take the form of the proteinId (field on the protein page) preceded by a 3 letter code for the organism: Ppl. Thus the name will be, for example, Ppl134532. (Generally, this identifier is not be altered during annotation. However, we may decide to append an allele-designating suffix, ie. "A" or "B". This issue will be resolved in the next few days.)

Defline: This field is mandatory. It will be part of the field "Name" in the "General protein information" section, and will accompany the sequence in Fasta format, immediately following the name. It should be a short (<85 characters) and accurate description of the gene product and if possible its main function(s). It should include the full standard name of the protein, but no acronyms or abbreviations. Examples:

Heat Shock 70kDa Protein, cytosolic

Carbamoyl-phosphate synthase L chain

In addition, GO terms should be used whenever possible! As in the case of naming issues (above), there is often some uncertainty as to precise function. Therefore, we have in the past used qualifying terms to describe our confidence. This system is based on both the SwissProt rules in combination with Bernard Henrissat's qualifying terms for carbohydrate active enzymes. Generally, calls are guided by homology with a protein that has been experimentally validated. The rules are as follows:

No Qualifier: >=80%id and 80% coverage
Transfer name and defline altering format where needed

example: Heat Shock 70kDa Protein

Candidate >=70%id and 70%coverage to known gene (and less than the above)
example: Candidate Heat Shock 70Kd Protein

Related to >=50% id and 50% coverage to known gene
example: Related to Heat Shock 70Kd Protein

Below this threshold -hypotheticals (most likely automatically annotated as conserved hypotheticals)

example: Hypothetical Heat Shock 70Kd Protein

You will notice that there is also functional information through Interpro domains, KOG, GO, etc. We propose that an annotator should be allowed to increase confidence by one level if these other types of evidence corroborates the function, also, if an investigation into the critical residues is used (say by MSA and trimming) then the next highest level of confidence could be used. To ensure consistency, we will review and discuss this process during out initial meeting.

Description: The entry in this field will transfer to the NCBI gene page in the field "Comment". It can be as long as you want, as long as the information is accurate and useful to researchers not familiar with this type of protein. Include information about the protein's function(s), its domains, interactions or subcellular location, comments about its phylogenetic origin, relationship to paralogues and orthologues, clustering with genes of related function, or overlap with neighboring genes etc... With proper caution, you can input your wildest guesses here. (Don't get carried away. Remember, your name will be associated with this description!) This field will be searchable, together with the Name and Defline fields, through the Advanced Search tool. Examples:

Heat Shock Protein 70, cytosolic; dnaK-type molecular chaperone; involved in protein folding and disaggregation; *probable allelic variant of Ppl13452

Carbamoyl-phosphate synthase (glutamine-dependent) Large chain (EC 6.3.5.5) (CarB); first step in pyrimidine biosynthesis; probably chloroplast-localized; more similar to bacterial than to higher plant enzyme; downstream of EC 1.2.3.4, also in pathway

* Allelic variants will be present for many gene models. Typically, the alleles are easily recognized. We will review the criteria.

Bibliography: If possible, place here the most important pieces of literature on this gene (also from other organisms). Rather than full references, use the Pubmed identifier. Example: PMID: 16156644

Model notes: Will not be searched, but may be useful for any detailed analysis. If the model appears structurally incorrect, please find one that is for that locus, or correct it if possible. If not possible, describe the problem by using a selection from the drop-down menu.

Needs GO: This should be filled only if you want the Gene Ontology consortium to create a new category for this gene.

Functional annotation: This is based on Gene Ontology data. Please refer to http://www.geneontology.org/ for a full description. Information can be entered manually, or fed automatically from the Automatic Ontology field. This will be described more in the Jamboree or phone training sessions.

Groups to annotate.

Come publication time, it will be necessary to collect all text from participating members in order to collate into a single publication or consecutive publications. Due to page limits in most journals, sections are cut or transferred to supplemental material online. In the past we have included all persons contributing to the annotation as authors, regardless of whether their sections were used or not.

For this project, we would like to see thorough annotation on the gene categories listed below. We will however will gratefully accept anything you annotate and we will attempt to incorporate it into a publication.

Potential Categories	*Responsible Persons (tentative)
Carbohydrate active enzymes	Bernard Henrissat, Randy Berka
P450s	Hiro and Jagjit Yadav
Copper radical oxidases	Phil Kersten
FAD-dependent oxidoreductases	Phil Kersten and others
Multicopper oxidases:	Luis with help from Urusula
Proteins involved in iron homeostasis	
Peroxidases	David Hibbett
Quinone reductase	Ken Hammel
Lipases-Esterases	
Proteases	
Oxalate decarboxylases	
Hydrophobins	
Small molecular weight extracellular glycoproteins	
Secreted hypothetical proteins:	Cullen, Hibbett

Other possibilities:
transposons
mating type
Secondary metabolism

Secretion Pathway
Transcription Factors
Core Metabolism
Replication, Repair, Recombination
Translation
Signaling

*Assignments will be made and revised by mutual consent during the jamboree. Please let us know if there are categories of particular interest to you.