Genome Annotation Using ASAP

https://asap.genetics.wisc.edu/asap/home.php
ASAP (A Systematic Annotation Package) houses biological information in the form of sequences, annotations, experimental data sets, and more. As a member of the community, we hope that you will use this resource as a part of your research. We also invite you to contribute to ASAP directly using its community annotation features. This manual describes what types of information can be contributed and how to enter that information into the database.

**How do I benefit from contributing annotations and having my work represented in ASAP? My genome of interest is already annotated and in GenBank.**

If you use any database, you want an up-to-date snapshot of the current state of knowledge in your field of enquiry. While GenBank entries may only be updated by authors, ASAP annotations can be made by any registered annotator and are publicly accessible moments after submission. The author of each annotation is identified, and each annotation is supported by evidence, enabling everyone to assess their accuracy prior to curation. If the whole community participates, the value of the database for each user increases enormously. This way you and your colleagues have access through ASAP to the information you need now, including:

- Information about a gene or genome from scientists other than the GenBank depositors.
- Correction of errors.
- New analysis or experimental data which changes the interpretation of prior observations.
- Links to your own publications provide the evidence for your annotations.

Jeffrey M. Perkel (no connection with ASAP) wrote in support of community annotators in the June 2006 issue of *The Scientist* (volume 20, issue 6, page 71) “Why You Should Be Annotating: scientists who rely on accurate gene predictions should share in the burden of creating them”.

You are the expert: whether you work on a single protein, a gene cluster, a pathway, or on DNA sequences underlying these entities, you likely have specific knowledge about genes or proteins that would improve the accuracy and completeness of the database. Community annotation is an important part of the ASAP mission and we invite you to participate.

**Table of Contents**

---

Page 2
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers Contribute to ASAP</td>
<td>4</td>
</tr>
<tr>
<td>What Types of Information Can I Add to ASAP?</td>
<td>4</td>
</tr>
<tr>
<td>Introduction to ASAP</td>
<td>5</td>
</tr>
<tr>
<td>Getting Started</td>
<td>6</td>
</tr>
<tr>
<td>Finding Features</td>
<td>7</td>
</tr>
<tr>
<td>How to Add an Annotation</td>
<td>10</td>
</tr>
<tr>
<td>Writing Annotations</td>
<td>11</td>
</tr>
<tr>
<td>Annotation Types</td>
<td>11</td>
</tr>
<tr>
<td>Name/Synonym</td>
<td>11</td>
</tr>
<tr>
<td>Product/Alternative Product Name</td>
<td>11</td>
</tr>
<tr>
<td>Function</td>
<td>12</td>
</tr>
<tr>
<td>Mutant Phenotype</td>
<td>12</td>
</tr>
<tr>
<td>Overexpression Phenotype</td>
<td>13</td>
</tr>
<tr>
<td>Genetic Interaction</td>
<td>13</td>
</tr>
<tr>
<td>Molecular Interaction</td>
<td>14</td>
</tr>
<tr>
<td>Regulation</td>
<td>14</td>
</tr>
<tr>
<td>GO Terms</td>
<td>15</td>
</tr>
<tr>
<td>Comment</td>
<td>15</td>
</tr>
<tr>
<td>Evidence Codes</td>
<td>16</td>
</tr>
<tr>
<td>Experimental</td>
<td>16</td>
</tr>
<tr>
<td>Sequence Analysis</td>
<td>16</td>
</tr>
<tr>
<td>Published Annotation</td>
<td>17</td>
</tr>
<tr>
<td>Nucleotide/Protein Sequence Similarity</td>
<td>17</td>
</tr>
</tbody>
</table>
Researchers Contribute to ASAP

Whether you work on a single protein, a gene cluster, a pathway, you likely know specific things about genes or proteins that the community could use in the database. Community annotation is an important part of ASAP and we would like to invite you to contribute.

What Types of Information Can I Add to ASAP?

Researcher contributed annotations generally fall into three categories.

Names and Synonyms

Usually 3-4 letters long, it is important for the database to have all of the names that a gene was ever called so that searching the database with any of the names will find the gene.

Functional Classification

What does the protein do? Usually the result of a number of experiments, the function annotation describes the “big picture” role of the protein in the cell. The product annotation describes what the protein is, such as the enzymatic activity that a protein has (i.e. hexokinase) or the common name that the protein has been given (i.e. sigma 70).

Phenotypes, Interactions, and regulation

In the course of experimentation researchers often generate mutants, overexpress proteins, observe physical and genetic interactions, and determine the regulation of gene expression. Usually the results of single experiments, these annotations describe concisely what defects a mutant has, what molecules does the protein bind, or how the gene is regulated.
Introduction to ASAP

ASAP was developed as a resource for the annotation of completely sequenced bacterial genomes. Initial stages of annotation involves identification of genes, RNAs, and other regions of the genome that have significance. Such subsets of the genome are termed “Features”. There are many different types of Features, and ASAP uses the same names for these Features as GenBank. There are over 50 Feature types, but many of them are rarely used. Let’s look at the most commonly used Feature types.

**CDS**
A CDS is a coding sequence, and is synonymous with ORF. This is the Feature type that is most commonly annotated by researchers, as proteins are usually the focus of most research. While the CDS technically refers to the DNA sequence that can encode for a protein, protein based annotations are associated with these Feature types.

**Gene**
Gene refers to a region of the genome that contains a CDS and its mRNA and promoter and regulatory elements. Poorly defined by GenBank as a “region of biological interest...”, a Gene’s boundaries are somewhat subjective. Gene as a word is sometimes casually synonymous with ORF or CDS, and the technical definition is given here to distinguish the two Feature types. Genes should only be annotated when the annotation is specific to the gene and would not be appropriate for the CDS.

**-10 signal/-35 signal/RBS**
These are the key binding sites for the transcriptional machinery and the ribosome. They are given the name of the CDS that they are associated with. They are usually identified by sequence analysis, but are sometimes experimentally verified.

**tRNA/rRNA/misc RNA**
These Feature types are transcribed into RNA gene products. Small regulatory RNAs that are common in bacteria are misc RNAs.
Sometimes there is a region of the genome that isn’t accurately described by any of the available Feature types. Misc features can be just about anything.

There are many other types of Features, this is just a brief overview of some of the more well-known Feature types. Let’s go through the process of logging in and how to put annotations into ASAP.

Getting Started

In order to contribute annotations, you must be registered and logged in. If you are not registered, email us to request an account.

**NOTE:** Registration usually takes less than 24 hours, but is not instant.

**LOG IN!**

Point your browser to: https://asap.genetics.wisc.edu/asap/home.php

**NOTE:** If you are not prompted to log in, you may be logged in as a Guest, and will need to click “Log In” on the menu bar

Welcome to ASAP

The ASAP database contains genome sequence capabilities. Sign up (umliss@wisc.edu) as:

Or log on as a registered user:

- User name: 
- Password: 
- Session Type: Secure

Log on
Finding Features

As we discussed before, most of the time you will be annotating a CDS. We’ll go through the process of finding a CDS using ASAP’s search tools.

You can search by gene name, featureID, or a number of other criterion using the advanced search.

Most of the time you will be searching by name and you will know which species and strain of the CDS you are interested in. In this case, it is often easier to search only one genome. Click on the link in the “Anno” column for the genome of interest to reach the individual genome’s search page. Here you can search by myriad criterion, but it is most common to find Features by searching for Feature name. Your results will be displayed in a separate page.
There are many different types of annotations that appear on a CDS’s main page. The top of the page contains basic information about the feature such as type (in this case CDS), name, genome, and size. Links to the DNA and Protein sequence can also be found below in the Sequence section, and you can specify if you would like any upstream or downstream sequence when retrieving DNA sequence. Links to other databases with additional information.
about the feature are also present. The Context section contains a link (“Browse sequence in GaPP”) that opens a new window with a browser that allows you to see the genes in the same region of the genome and their relative size (shown below). Double clicking on any of the genes (pink boxes) will open that gene’s page in a new window.

In sections further down the feature page, you will find more detailed information such as annotations (See previous page). Annotations are generally single line pieces of information about the gene (or gene product). Annotations will be discussed in detail later in this manual. Also further down the feature page, you will find sections of overlapping features (Context Table), a list of orthologs from other organisms (Comparative), links to BLAST results querying this feature against every genome in the ASAP database (Pre-Run Search Results), and links to experimental data from functional genomics experiments when available (Experimental).

In order to add an annotation, select the type of annotation you want to add and click on the “Add” button (you must be logged in). There are over 40 different types of annotations that can be added. In the next few pages, we will look at the types of annotations most commonly added by researchers.
How to Add an Annotation

Let’s go through the process of actually adding an annotation. In our fictitious example, we will add a new name to a CDS. After selecting “name” from the pull-down menu and clicking the “Add”, as described on the previous page, you will be directed to the Add Annotation page. You will be presented with a form where you will need to supply the name, evidence, and originating organism.

Data Field
The is the text box where you would type, in this case, the name you wish to enter.

Evidence
Every annotation is associated with evidence. In our example, a published paper with a PubMed ID number uses the name, so we have selected Published Annotation. Evidence will be discussed in more detail later.

Organism
In this example the paper dealt with Escherichia coli strain MG1655, which is a K-12 strain. This information was entered, but is optional.
Writing Annotations

Here we describe some of the more common types of annotation. This describes how the entry in the Data Field text box should be written. For simplicity, CDS and protein will be used almost interchangeably.

Annotation Types

Name/Synonym
Definition: Name generally refers to the main or primary name that a CDS has. All other names that have been used to describe the CDS are Synonyms. Importantly, names of proteins are not annotated as names, but rather as Products (see below). In bacteria, CDS names are typically three lower case letters, usually followed by a capital letter, or sometimes a number.

How to Annotate: Simply enter the 3-4 letters/numbers of the Name or Synonym, with no additional text or punctuation.

Product/Alternate Product Name
Definition: Analogously to a CDS’s Name and Synonym, proteins are described by Product, the primary name of the protein or gene product, and Alternate Product Names, which are other names that describe the protein or gene product.

How to Annotate: Enter the Product name in the field. Product names that are simply capitalized versions of the gene Name (or Synonym) should not be annotated.

Example:

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Beta-Galactosidase</th>
</tr>
</thead>
</table>

Function
Definition: The biological role that a protein performs. Basically, what the protein does for the cell. Usually the result of multiple experiments.
How to Annotate: Usually beginning with a verb (regulates, transports, induces, degrades, involved in, required for...), **Function** lines describe what the protein does, or what process it is involved in.

Example:

<table>
<thead>
<tr>
<th>Annotation</th>
</tr>
</thead>
</table>
| *Annotates have a maximum length of 4000 characters and may not include quotes*
| regulates transcription of oxidative stress-response genes |

**Note:** Do not confuse **Product** with **Function**. “Hexokinase” is a **Product** annotation, “catalyzes the first step of glycolysis” is a **Function** annotation.

**Mutant Phenotype**

Definition: An observable difference between a mutant strain and its wild-type parent.

How to Annotate: Lines should begin with a description of the mutant followed by a semicolon, then a description of the phenotype.

Examples:

<table>
<thead>
<tr>
<th>Annotation</th>
</tr>
</thead>
</table>
| *Annotations have a maximum length of 4000 characters and may not include quotes*
| abcD::kan mutant; null mutant shows increased sensitivity to UV irradiation |
| OR |
| defG-3238A mutant; mutant with altered active site is unable to utilize raffinose as a carbon source |

**Overexpression Phenotype**

Definition: An observable difference between a strain overproducing a protein and its wild-type parent.

How to Annotate: Like **Mutant Phenotype**, **Overexpression Phenotype** requires a description of the method of overexpression followed by a semicolon. Include plasmid name and promoter information if available/relevant. After the semicolon describe the phenotype.
Genetic Interaction
Definition: A phenotype of a strain harboring more than one mutation or overexpressed protein that provides additional information than the individual mutant or overexpression phenotypes. One common situation is two individual single mutants that show no phenotypes, but the double mutant does show a phenotype. Another common Genetic Interaction is the suppression of a phenotype of a mutant by overexpressing or mutating a different protein.

How to Annotate: A description of the mutations or proteins being overexpressed should precede a semicolon, with the description of the interaction following.

Examples:
**Molecular Interaction**

Definition: The binding of a protein to any molecule. Usually the protein is binding another protein or DNA, but can also be metabolites or other small molecules.

How to Annotate: State the molecule that the protein binds and the technique used to determine the interaction. It is important to include the technique for molecular interactions because of the range of reliability of various methods.

**Examples:**

<table>
<thead>
<tr>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>binds the tppB promoter in DNAse I footprinting assays</td>
</tr>
<tr>
<td>interacts with YopN by MALDI-TOF</td>
</tr>
</tbody>
</table>

**NOTE:** For protein-protein interactions, make sure that you also add the reciprocal Molecular Interaction annotation to the other CDS’s page.

**Regulation**

Definition: Anything that affects the expression patterns of a CDS. This could be environmental factors such as nutrients, temperature, or stressors. Sometimes a Mutant Phenotype or Overexpression Phenotype will be altered expression of a CDS. Add that information in a Regulation annotation on the affected gene’s page.

How to Annotate: Start your annotation with “Expression regulated by” followed by the factor and then a semicolon. Finish the annotation by explaining the conditions and effect.

**Examples:**

<table>
<thead>
<tr>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>expression regulated by dsrA; expression is reduced in a strain overexpressing dsrA</td>
</tr>
<tr>
<td>expression regulated by acetate; expression is induced by acetate</td>
</tr>
</tbody>
</table>

OR
GO terms
ASAP is configured to accept GO (Gene Ontology) terms. To add a GO term select the type of GO term you wish to add and select the GO term ID from the pull down bar:

The GO term will appear on the CDS’s page with the term that is hyperlinked to the definition of the term and its place in the hierarchy.

ASAP has all GO evidence codes in the evidence pull down menu, and all GO annotations require GO evidence. For more information on GO and GO evidence codes, see www.geneontology.org.

Comment
Sometimes you may have something interesting to say about a CDS or protein, but it doesn’t fit nicely into any category of annotation. Comment is used for these purposes.

Examples:

Annotations have a maximum length of 4000 characters and may not include quotes

transcription start site has been determined to be a nucleotide at position -58

OR

multiple specific monoclonal antibodies have been raised

Evidence Codes
Every annotation requires evidence. This allows anyone viewing the annotation to evaluate the reliability of an annotation and have access to the source reference. In ASAP you are required to select a type of evidence and supply a reference. The reference can be a database or a website, but is usually going to be a published journal article. These articles have PubMed ID numbers which will be the reference for the paper. We will review the most common evidence codes.
NOTE: At first, the number of evidence codes that there are to choose from is a bit intimidating, but it’s actually not that bad.

Experimental
Most annotations will probably have experimental evidence, which simple means that the annotation was based on the results of an experiment. Supply the PubMed ID.

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Experimental (supply PubMed ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence Reference</td>
<td>Supply the value indicated by the selected evidence type above</td>
</tr>
<tr>
<td></td>
<td>12345678</td>
</tr>
</tbody>
</table>

Sequence Analysis
As a researcher you may look at the sequence of a protein and predict its function or something else about the protein. This is a sequence analysis and if it is published you may annotate with this code and supply a PubMed ID.

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Published Sequence Analysis (supply PubMed ID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence Reference</td>
<td>Supply the value indicated by the selected evidence type above</td>
</tr>
<tr>
<td></td>
<td>12345678</td>
</tr>
</tbody>
</table>

If a sequence analysis is not published, use Unpublished Sequence Analysis and supply your name and email address separated by a comma.

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Unpublished Experimental Evidence (supply Author Name and Email)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence Reference</td>
<td>Supply the value indicated by the selected evidence type above</td>
</tr>
<tr>
<td></td>
<td>Brad Anderson, <a href="mailto:bdanderson@wisc.edu">bdanderson@wisc.edu</a></td>
</tr>
</tbody>
</table>

Published Annotation
This evidence code is often used for names and synonyms. If an author publishes a report on a previously uncharacterized CDS and gives it a name, that is a published annotation. A second use is in the following case. In the introduction of many papers, there is a lot of information about proteins that were not discovered or characterized in the paper itself, but by previous work. This information is referenced in the bibliography. While it is desirable to have the original paper and refer to it and use an Experimental evidence code, sometimes it is not possible or practical. Using this evidence code allows you to add the annotation without digging endlessly through the literature.
Nucleotide/Protein Sequence Similarity

If a gene or protein sequence resembles that of another gene or protein and a similar function, role, or other property is inferred, the evidence code is probably Protein Sequence Similarity. Sometimes the nucleotide sequence is used, so Nucleotide Sequence Similarity should be used in those cases. The reference here is usually another protein, one with experimentally verified properties. If the reference protein is in ASAP, you can use the ASAP FeatureID. Otherwise a GenPept or Swiss-Prot accession number will suffice.

Sequence Analysis vs. Sequence Similarity

Sometimes it is difficult to decide whether to use Sequence Analysis or Similarity. If you analyze a protein sequence and observe hydrophobic regions and hypothesize that the protein may be embedded in the membrane, that is a Sequence Analysis. If you BLAST your sequence and notice that it has high similarity to membrane spanning proteins, that is Sequence Similarity. Often it is not this clear, and it would seem that both evidence codes would be appropriate. In order to resolve this, think of what would be more useful to the end user. If you have a paper that discusses the protein sequence and explains its similarity to other proteins, use Sequence Analysis and reference the paper. If it is unpublished, or the paper does not go into any depth about the protein, you may be better off referencing the related protein and using Sequence Similarity.
Organism
One final field in the Add Annotation form is the organism field. Many times experiments will be done in a strain that is closely related to a genome in ASAP. In these cases, the corresponding gene in the related strain should be annotated in ASAP, but we can keep track of what organism the experiments were originally done using this field. The entry should be made in this order:

Genus species strain (E. coli pathotype)

Pathotype is optional and is UPEC, EHEC, EPEC, etc.

<table>
<thead>
<tr>
<th>Reference Organism</th>
<th>Optionally enter an organism to associate with this annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli EDL933 (EHEC)</td>
<td></td>
</tr>
</tbody>
</table>

All of the organisms whose genome is in ASAP are listed in the pull-down menu, or you can type in your own.

Other Attributes of ASAP
We have covered how to add annotations, but there is much more that can be done in ASAP. You can add orthologs and paralogs, change the coordinates of a feature, upload functional genomics data, and more. If you would like to do any of these things, or anything else, please contact us and someone will assist you by either adding the information themselves, or showing you how to do it. We are here to serve the community and welcome any comments, criticism, or suggestion that you may have.

Contact Information
Visit ASAP online to login or view as a guest:

https://asap.genetics.wisc.edu/asap/home.php