Jill Keeney
NSF Yeast ORFan network

Defining and Searching for ORFans

Clarification of definitions used in describing ORFs and ORF function is important in defining ORFans. Three relevant terms used by SGD to describe ORFs are verified, uncharacterized and dubious. The definitions below are taken from the SGD glossary page (https://sites.google.com/view/yeastgenome-help/sgd-general-help/glossary).

> Verified ORFs: ORFs for which experimental evidence exists that a gene product is produced in *S. cerevisiae*. Generally, these have obvious orthologs in one or more other *Saccharomyces* species. Most named genes are in this class. Evidence from large-scale analyses that indicates an ORF may be biologically relevant is sometimes but not always enough to upgrade an ORF from "Uncharacterized" to "Verified", depending on the individual case.
>
> Uncharacterized open reading frame (ORF): ORF likely to encode an expressed protein, as suggested by the existence of orthologs in one or more other species, but for which there are no specific experimental data demonstrating that a gene product is produced in *S. cerevisiae*. While most Uncharacterized ORFs have systematic names only (e.g., YKL100C), a few have been given genetic names (e.g., PAU8). Evidence from large-scale analyses that indicates an ORF may be biologically relevant is sometimes but not always enough to upgrade an ORF from "Uncharacterized" to "Verified", depending on the individual case.
>
> Dubious open reading frame (ORF): ORF is unlikely to encode an expressed protein. Dubious ORFs may meet some or all of the following criteria: 1) the ORF is not conserved in other *Saccharomyces* species; 2) there is no well-controlled, small-scale, published experimental evidence that a gene product is produced; 3) a phenotype caused by disruption of the ORF can be ascribed to mutation of an overlapping gene; and 4) the ORF does not contain an intron. Many ORFs classified as "Dubious" are small and overlap a larger ORF of the class "Verified" or "Uncharacterized"; however, overlap with another ORF does not mandate that an ORF be classified as "Dubious."

Note that genes in any of these categories can be described as "unknown" based on Gene Ontology. We are using Gene Ontology to define ORFans.

GO: Gene Ontology

The following description is from https://sites.google.com/view/yeastgenome-help/sgd-general-help/glossary. The Gene Ontology project page has general information and links.

> "The Gene Ontology (GO) project was established to provide a common language to describe aspects of a gene product's biology. The use of a consistent vocabulary allows genes from different species to be compared based on their GO annotations. For each of three categories of biological information--molecular function, biological process, and cellular component--a set of terms has been selected and organized. Each set of terms uses a controlled vocabulary, and parent-child relationships between terms are defined. This combination of a controlled vocabulary with defined relationships between items is referred to as an ontology. Within an ontology, a child may be a "part of" or an example ("instance") of its parent. There are three independently organized controlled vocabularies, or gene ontologies, one for molecular function, one for biological process, and one for cellular component. Many-to-many parent-child relationships allowed in the ontologies. A gene may be annotated to any level in an ontology, and to more than one item within an ontology."

Here are definitions of GO terms (taken from http://www.geneontology.org/).

Cellular Component
These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

Biological Process
A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions. Examples of broad biological process terms are "cellular physiological process" or "signal transduction". Examples of more specific terms are "pyrimidine metabolic process" or "alpha-glucoside transport". The general rule to assist in distinguishing between a biological process and a molecular function is that a process must have more than one distinct steps.

A biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Molecular Function
Molecular function terms describe activities that occur at the molecular level, such as "catalytic activity" or "binding activity". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are "catalytic activity" and "transporter activity"; examples of narrower functional terms are "adenylate cyclase activity" or "Toll receptor binding".

It is easy to confuse a gene product name with its molecular function; for that reason, GO molecular functions are often appended with the word "activity".

Defining an ORFan

The definition of GO terms is central to our definition of an ORFan. Strictly defined, an ORFan would be an ORF for which the GO term in all three gene ontologies (molecular function, biological process, cellular component) are unknown. However, we are being much more flexible. An ORF described as "unknown" under any of these GO terms is considered an ORFan for the purposes of this grant project. Many ORFans are annotated as unknown for one or two of these categories that have been pulled out in screens or have enough data to relate them to a specific area of interest.

This definition includes a large number of *S. cerevisiae* ORFs. The number of ORFs currently defined as unknown under each GO term can be found on SGD as follows:

On SGD, from the top menu: Sequence ---> reference genome ---> genome snapshot.

Scroll down to the summary of GO annotations. Click on the buttons to see the name and number of each top ranking term for each GO category. Mouse over the bar to get the actual number of ORFs.

As of 6/14/2017, the number of ORFs "annotated to unknown" were:

Biological Process-1811

Molecular Function-2634

Cellular Component-1322

(Note: the links to feature type on this page do not work. A method to generate a list of unknown genes in each category is described below.)

Finding specific ORFans

There are many, many ORFans. You may wish to search (or have your students search) for ORFans with some data to link them to a particular process, function or component.

One way to search for ORFans is to go to SGD and put the term "unknown" in the search bar in the upper right of the page. On the left of the search results page, click on "genes". There are approximately 1000 ORFs that result from this search. (Note that this method searches the gene description, not the GO terms for each ORF entry, so the number of resulting ORFs is less than the number listed above under each GO term). Many of these ORFs are uncharacterized or dubious (see definitions above). Due to variation in gene descriptions used, even the verified ORFs in this list range from all three GO terms unknown (LEE1/YPL054W) to all three GO terms with some annotation, and thus not an ORFan (SLX8/YER116C).

On the left side of the search page, the ORF list is broken down by specific GO terms. These links are where to find ORFans with some data linking them to specific GO terms. You will need to do some scrolling and reading, as some non-ORFans end up on this search list because of their gene names, such as FUN-Function Unknown Now. Apparently the function is now known, but the gene name was not changed.

You can easily download a list for a specific GO term, for example "nucleotide binding (direct)". Clicking on this GO terms shows 37 results for this search (unknown, gene, nucleotide binding direct). In the upper right, change the display from "list" to "wrapped". You can download that list, but even better is to choose "analyze". There are four options for tools. Choose yeast mine (far right), and it will load your list into yeast mine. Name your list and save it. You can then manage columns to include various descriptors and other information. I include:

-systematic name
-standard name
-gene name
-gene qualifier (verified, uncharacterized, dubious)
-brief description
-gene description

That table can then be downloaded as a .txt file and opened in excel. You will need to go through the list and confirm which are ORFs of unknown function. Alternatively, you can work with the filters options on yeast mine to sort out genes with unknown in the description.

Another way to search for ORFans is to search on SGD for a phenotype or process. For example, put the term "transposition" in the search bar in the upper right of the page. On the left of the search results page, click on "genes". There are ~475 ORFs that result from this search. In the upper right, change the display from "list" to "wrapped". You can download that list, but even better is to choose "analyze" from the menu above the list. There are four options for tools. Choose yeast mine (far right), and it will load your list into yeast mine. Name your list and save it. You can then manage columns to include various descriptors and other information. I include:

-systematic name
-standard name
-gene name
-gene qualifier (verified, uncharacterized, dubious)
-brief description
-gene description

That table can then be downloaded as a .txt file and opened in excel. You will need to go through the list and confirm which are ORFs of unknown function. Alternatively, you can work with the filters options on yeast mine to sort out genes with unknown in the description.

Note that instead of genes, one could choose "phenotype", and get a gene list under each phenotype. Unfortunately, there is no analyze option. To move the gene list to yeast mine, select and copy the list and paste into yeast mine in a new window (on a PC, the copy to clipboard and then pasting into yeast mine does not seem to work).

FYI: SGD has a YouTube channel with videos on various topics.

https://www.youtube.com/channel/UCnTiLvqP2aYeHEaJI7m9DUg/videos

Using YeastMine to generate a list of ORFs defined as "unknown" within each GO term

GO terms have ID numbers to used to define each GO category. This can be used to generate a list of ORFs annotated to unknown within each GO term. It is worth creating an account with YeastMine so that you can log in and save your searches.

Navigate to the YeastMine homepage:
http://yeastmine.yeastgenome.org/yeastmine/begin.do

In the top menu bar, click on "templates"

Click on the template "GO ID ---> Genes"

Enter one of the following in the lookup search box:

"GO:0003674" to search for molecular function unknown

"GO:0008150" to search for biological process unknown

"GO:0005575" to search for cellular component unknown

Results of searches will be saved if you are logged in. You can customize columns, sort according to columns and also download lists.

Other notes:

1. Cautionary note on terminology...this instruction is given on SGD within YeastMine help:

> "Uncharacterized genes (i.e. genes for which a function hasn't been identified) can be identified using the Gene Ontology annotation to the root node: Molecular_Function (GO:0003674) unknown."

Note in this description the confusion of terminology between "unknown" as a GO term descriptor and "uncharacterized gene". Be sure to understand the difference.

2. More on GO term annotations. This information is from the GO site, concerning the use of the ID numbers for "unknown" (given above) for compiling gene annotations:

> "Used for annotations when information about the molecular function, biological process, or cellular component of the gene or gene product being annotated is not available.
>
> Use of the ND evidence code indicates that the annotator at the contributing database found no information that allowed making an annotation to any term indicating specific knowledge from the ontology in question (molecular function, biological process, or cellular component) as of the date indicated. This code should be used only for annotations to the root terms, molecular function ; GO:0003674, biological process ; GO:0008150, or cellular component ; GO:0005575, which, when used in annotations, indicate that no knowledge is available about a gene product in that aspect of GO."

(http://www.geneontology.org/page/nd-no-biological-data-available)