Protocol

The *Saccharomyces* Genome Database: Gene Product Annotation of Function, Process, and Component

J. Michael Cherry[1]

Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120

An ontology is a highly structured form of controlled vocabulary. Each entry in the ontology is commonly called a term. These terms are used when talking about an annotation. However, each term has a definition that, like the definition of a word found within a dictionary, provides the complete usage and detailed explanation of the term. It is critical to consult a term's definition because the distinction between terms can be subtle. The use of ontologies in biology started as a way of unifying communication between scientific communities and to provide a standard dictionary for different topics, including molecular functions, biological processes, mutant phenotypes, chemical properties and structures. The creation of ontology terms and their definitions often requires debate to reach agreement but the result has been a unified descriptive language used to communicate knowledge. In addition to terms and definitions, ontologies require a relationship used to define the type of connection between terms. In an ontology, a term can have more than one parent term, the term above it in an ontology, as well as more than one child, the term below it in the ontology. Many ontologies are used to construct annotations in the *Saccharomyces* Genome Database (SGD), as in all modern biological databases; however, Gene Ontology (GO), a descriptive system used to categorize gene function, is the most extensively used ontology in SGD annotations. Examples included in this protocol illustrate the structure and features of this ontology.

MATERIALS

Equipment

Internet-connected computer with web browser

METHOD

The GO (Gene Ontology Consortium 2013) is the major ontology used to communicate functional characteristics of gene products. This protocol introduces tools available at SGD to use GO annotations. The statistical treatment of a collection of genes with GO annotations uses techniques typically called term enrichment. There are many different algorithms and tools that implement these methods of determining significantly shared annotations between a set of genes. However, there is little agreement on which metric determines the best enrichment tool. This is partly because of the number of components used to determine enrichment, the sets of annotations, version of the ontology, amount of filtering by evidence, and the background frequency used. Huang et al. (2009) provide a discussion on term enrichment methods that are used to determine which annotations are significant. Shah et al. (2013) provide background on the use of gene enrichment with specific case examples. SGD has chosen to integrate the GO TermFinder tool (Boyle et al. 2004) and it is available in the Analysis menu. GO Term Enrichment requires an input list of genes that have been

[1]Correspondence: cherry@stanford.edu

Cite this protocol as *Cold Spring Harb Protoc*; doi:10.1101/pdb.prot088914

GO annotated and considers all the genes' annotations to determine which GO terms are represented more than expected by chance.

At any step in this protocol support is available via the comprehensive Help documents maintained by SGD. These can be accessed via the help button at the top of each page or, for specific features, a small red button with a question mark in its center is provided and will link to the help pages specific for that feature.

The SGD is continually updated; therefore, specific items presented in this protocol may not appear on the SGD website exactly as described.

1. Open the URL, www.yeastgenome.org, in any modern web browser (e.g., Chrome, Firefox or Safari).
2. Enter "RTT103" in the Search box and press return to go directly to the *RTT103* Locus Summary page. Click on the Gene Ontology tab.

   Rtt103p is a protein involved in RNA polymerase II transcription termination and is also involved in the regulation of Ty1 transposition. Manual curation includes four annotations to GO biological process terms. Of the four annotations two are to the same term, "mRNA 3′-end processing." The difference between these two annotations is the Evidence code that was defined. For each manually created GO annotation the biocurator reviews the published work and determines the appropriate GO term that describes the experimental result. The creation of the annotation also includes identifying the type of evidence reported for the result (http://www.geneontology.org/page/guide-go-evidence-codes). The code IMP stands for Inferred from Mutant Phenotype and IGI stands for Inferred from Genetic Interaction. In Kim et al. (2004) two forms of results were available and thus both were used to create individual annotations. The IMP annotation is based on an investigation of a RTT103 knockout and that for IGI is based on a synthetic genetic array (SGA) assay where the rtt103 deletion strain was crossed with other deletion strains. The SGA assay showed that RTT103 makes a synthetic lethal with REF2, CTK1 and CTK2. Manual annotations are also presented for Molecular Function and Cellular Component with IDA and IPI (Inferred from Physical Interaction) evidence.
3. Scroll down the *RTT103* Gene Ontology page to view the High-throughput annotations section.

   This section includes results reported in published large-scale assays. In this example there are no HTP annotations.
4. Scroll down the *RTT103* Gene Ontology page to view the Computational predictions annotations section.

   These annotations are not reviewed by curators and are the result of computational analyses, such as protein motif searches, BLAST analysis, and rule systems from several database projects such as UniProt (http://www.uniprot.org).
5. Scroll to the bottom of the *RTT103* Gene Ontology page and view the Shared Biological Processes section.

   This type of visualization is on many of the SGD pages and is a useful alternative to tables of information. In this example you can see the Biological Process annotations (green boxes) annotated to RTT103 (yellow circle). In addition, other genes are included (dark gray circles) that share annotations with RTT103.
6. While not available on the RTT103 visualization, you can typically explore the network by moving the slider at the bottom to increase or decrease the number of shared annotation terms.
7. Manipulate the network by dragging the circles and squares.

   Changing the placement of the nodes can enhance the view.
8. Double click on a node. The circles go to genes and the squares go to GO Term pages. For example, find and click on "mRNA 3″-end processing" to read complete information about this term.

   The term's details include a network display of the parent and child terms. Further down all annotations to this GO term are provided as a table.
9. Select the "GO Term Finder" option from the Analyze pull-down in the purple bar at the top of the page to explore the annotations associated with a list of genes using the GO Term Finder. Enter the following four gene names in "Step 1: Query Set" box: PCF11, TFA1, TFA2, and RTT103. Do not use commas to separate the gene names, rather use space or return. In the "Step 2: Choose Ontology and Set Cutoff" box, select function and then click on Search to use the default settings.

Cite this protocol as *Cold Spring Harb Protoc*; doi:10.1101/pdb.prot088914

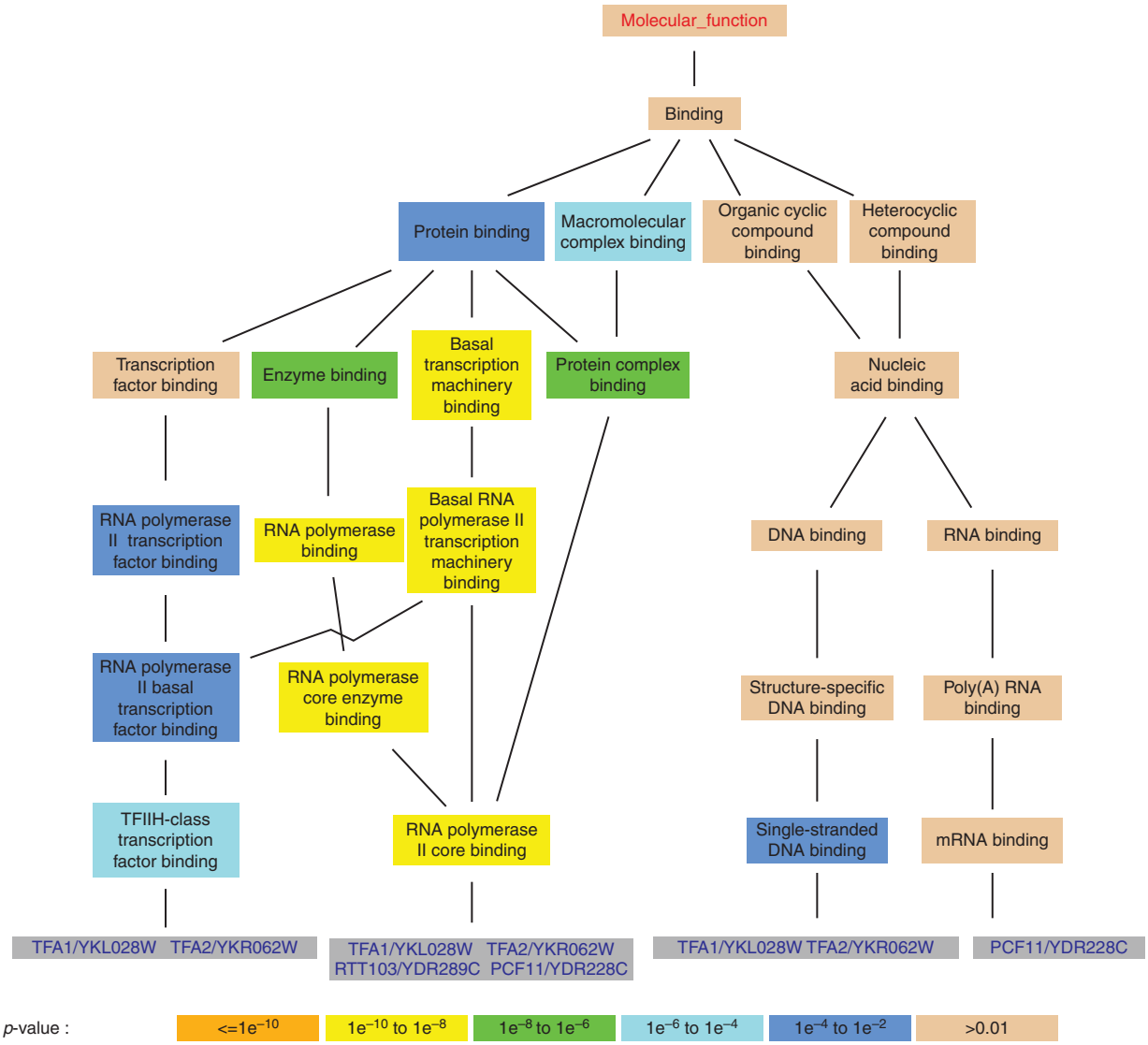

FIGURE 1. Screenshot (redrawn for clarity) of the Gene Ontology Term Finder output graphic. This is the result from performing term enrichment on the annotations of the four genes: PCF11, TFA1, TFA2, and RTT103. The graphic shows the organization of a portion of the Molecular Function graph. Each manually curated annotation for these four genes is shown. The term enrichment results are indicated by color-coded terms. In this graphic the most significant terms are shown in yellow, followed by green, cyan, blue and beige. The significance indicates the likelihood that this set of genes would be annotated to the term by chance.

The results may take a few minutes to return. The graphic (Fig. 1) shows the GO terms that describe the annotations for this set of genes, color-coded by significance. The significance indicates the likelihood that this set of genes would be annotated with this set of GO Terms by chance. The graphic also shows which GO terms have been used to annotate the genes. In this case, using the Molecular Function annotations, the shared terms with greater significance are in yellow such as "RNA polymerase II core binding"; however, parent terms are also highlighted. The table at the bottom of the page includes the p*-values for the results.*

ACKNOWLEDGMENTS

I am grateful to all the present and past staff of the *Saccharomyces* Genome Database project for their dedication to accuracy and service to life science educators and researchers. I also want to thank the yeast research community for their support and suggestions. This work was supported by the National

Human Genome Research Institute (grant number U41 HG001315), and Funding for open access charge was provided by the National Institutes of Health. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health.

REFERENCES

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20:** 3710–3715.

Gene Ontology Consortium. 2013. Gene Ontology annotations and resources. *Nucleic Acids Res* **41:** D530–D535.

Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37:** 1–13.

Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, Buratowski S. 2004. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432:** 517–522.

Shah NH, Cole T, Musen MA. 2013. Chapter 9: Analyses using disease ontologies. *PLoS Comput Biol* **8:** e1002827. doi: 10.1371/journal.pcbi.1002827.

Cite this protocol as *Cold Spring Harb Protoc*; doi:10.1101/pdb.prot088914

The *Saccharomyces* Genome Database: Gene Product Annotation of Function, Process, and Component

J. Michael Cherry

Cold Spring Harb Protoc; doi: 10.1101/pdb.prot088914