Dear Ciliate Researchers,

Tetrahymena Genome Database needs your help to improve the functional annotation of genes in anticipation of our grant renewal submission in Spring 2013. Our evaluation will be based in part on the success we have had collecting information submitted by the user community. We are one of the only genome databases to open annotation directly to the research community, and proving the viability of this model is vital to our renewed funding.

One simple – but significant – contribution you can make is to name the genes you are working on and provide short descriptions about their functions. We welcome these types of data for any gene whose function can be readily determined, whether they have been published or not. In fact, some of the easiest genes to identify and name may never be mentioned in a Tetrahymena publication because they are so well studied in other systems. Below are some simple guidelines for naming genes that we think are easy to follow, yet rigorous enough to provide accurate names in most cases. (If not, we can always change them later – that's the beauty of a Wiki.)

Bioinformatics steps to assist in naming Tetrahymena genes.

Gene Names must conform to the published and community-accepted guidelines: three capital letters followed by one or two digits (e.g. PDD1). This makes it possible for text-based search engines to easily identify Gene Names on this site and in publications. A Gene Name is often an abbreviation of a phrase describing the gene's function. For example: PDD = Programmed DNA Degradation. The numbers following the abbreviation allow it to be used to describe more than one gene. This is useful when describing genes that belong to the same family (e.g. the myosin genes MYO1, MYO2....MYO13) or that have related functions (e.g. PDD1, PDD2, PDD3). Do not assign different meanings to the same abbreviation; choose a different abbreviation for the unrelated gene instead.

Note: If the name you would like to give a gene does not follow these published standards (for example, a Tetrahymena homolog of PCNA cannot be given a four letter name), you can enter this name in the Alias field. Please take the time to assign a standardized name to the gene as well, for our own records in the community. The standardized name may omit some aspect of the name found in other systems, or overlap with the name of some unrelated gene. While this is unfortunate, the benefits of naming genes in a consistent way largely outweigh any confusion that may arise.

I. Find the *Tetrahymena* homolog of a known gene from another organism

A. Starting with a protein from another organism

This approach may be particularly useful for looking at Tetrahymena genes annotated as 'predicted protein' in TGD, which have a high co-expression correlation coefficient with a gene of interest that you are already studying. These directions will show you how to find genes that are best reciprocal BLAST hits of homologs in another genome. Genes that are one another's best hits between two genomes are typically orthologous, allowing us to confidently transfer the name and annotations given for the other gene to our Tetrahymena gene. Comparing with the Saccharomyces genome is particularly useful, since they use the same naming conventions we do, and because yeast has 1/3 of the genes Tetrahymena has, eliminating much of the confusion when deciding on a name. Use other eukaryotic databases (such as human, mouse, Drosophila, Arabidopsis) for protein sequences of genes that are not encoded in the *S. cerevisiae* genome.

Steps

1. BLAST search a protein sequence from another organism against the *Tetrahymena* genome. Do this by copying the amino acid sequence of the gene into the BLAST server at Tetrahymena Genome Database (TGD): http://www.ciliate.org/blast/blast_link.cgi

Choose 'blastp' from the Program pulldown menu and 'T. thermophila Amino Acid (protein)' from the Database menu.

2. If you get one or more strong matches (e value = 1e-05 or lower), click on the top match to get to the related gene page on TGD. Scroll to the bottom of the gene page, and copy the Tetrahymena protein sequence.

3. BLAST this protein back against the other (original) organism's genome. If it is a reciprocal best match (meaning that the original gene of interest is at the top of the "match" list when searched with the *Tetrahymena* protein sequence) it is likely an ortholog.

4. To be sure this name is appropriate, BLAST the gene at NCBI. Go to BLAST on NCBI (choose from right-hand menu on NCBI homepage). Choose "protein-BLAST", and then "non-redundant database" on the protein BLAST page (this is normally the default). The results returned should show you if there are alternative (sometimes better) names for the gene you are looking at.

5. If all of these conditions are satisfied, feel free to give the Tetrahymena gene the same name as this homolog, *provided it follows the Tetrahymena naming conventions.* If it has a nonstandard name, consider naming it with an acronym of three letters and one or two numbers that are found in the homolog's name.

6. Since Tetrahymena has almost 25,000 genes, there are often many homologs (paralogs) of any gene you find in yeast, or other model system. Returning to the BLAST search at TGD, check to see if the next few genes in the list hit the same gene in the model organism. If so, these are probably expanded members of the gene family. Consider giving these names the same acronym, but different numbers. If you have any questions about gene names or other annotations, feel free to email: ciliate-curator@bradley.edu

7. If it happens that there is no reciprocal best BLAST hit for your gene of interest, it is likely that the gene family in either Tetrahymena or the source organism for your gene has many paralogs. It may also be that a failure to find a reciprocal best hit for a given gene could be the result of mis-prediction of the correct gene structure. If this is the case, email us and we may be able to guide you through the process.

B. Starting with a *Tetrahymena* protein sequence

Over the years I've found that BLASTing Tetrahymena genes against a smaller, well-defined set of genes is more effective than searching nr at NCBI. At NCBI, the top hits are likely to be poorly studied genes from an obscure plant or Alveolate. These were usually named based only on their similarity to a model organism in the first place, so using them to name our genes puts a layer of uncertainty in our annotations. A better approach is to compare against the budding yeast genome. *S. cerevisiae* has a fairly minimal, but mostly complete, set of genes shared by all eukaryotes. It's also the best annotated genome available, and many of the names were given based on wet lab studies rather than sequence similarity. This should be the default we use. However, the yeast genome is very small and has lost a fair number of genes that were retained in Tetrahymena and other eukaryotes. If you wish to name and annotate these genes, you should BLAST your gene at the genome databases for mouse, fly, *Arabidopsis*, *C. elegans*, or *Dictyostelium*.

Saccharomyces Genome Database	www.yeastgenome.org
Mouse Genome Informatics	www.informatics.jax.org
FlyBase	www.flybase.org
The Arabidopsis Information Resource	www.arabidopsis.org
WormBase	www.wormbase.org
dictyBase	www.dictybase.org

Steps

1. Go to the BLAST server at Saccharomyces Genome Database:
http://www.yeastgenome.org/cgi-bin/blast-sgd.pl
Choose "BLASTP" from the first pulldown menu, and select "Open Reading Frames" from the list of databases.

2. Paste the *Tetrahymena* protein sequence into the search box. Run the search.

3. If the top hit is lower than 1e-05, check to see if it is a reciprocal best BLAST hit by following the steps in Section A.

If you are studying a pathway that has a more complex set of genes in that budding yeast, try BLASTING the human genome database, by selecting this database at the NCBI site. http://blast.ncbi.nlm.nih.gov/

II. Naming a gene based on the presence of conserved domains

Not all proteins in *Tetrahymena* will have a clear ortholog, but many have conserved domains that provide enough information to justify giving a putative gene a name. Examples where this approach has been used effectively in gene naming are the ABC transporter genes, Rab genes, cyclin genes, and myosin genes.

Take the *Tetrahymena* protein sequence of interest and look for conserved domains using the BLASTP server at NCBI (see Section A, step 4). BLASTP at NCBI automatically searches the sequence for domains at CDD, NCBI's Conserved Domain Database, which includes domains found in Pfam, SMART, and other sources. If domains are present, they will pop up before your BLAST results page. You can always return to the domain search by following the link at the top of your Results page. If the domain covers the majority of your sequence, you can choose a name based on the letters found in the domain. Also, if the domain is distinctive to a gene family, you may feel confident giving it a name based on this similarity.

Updating TGD Wiki

Steps

1. After you have determined the identity of your gene, you should first make sure that the desired name is not already being used for a different gene. You can either enter the name into the Quick Search and see if it already exists, or you can look at the list of genes that have been named at TGD:

http://ciliate.org/index.php/show/namedgenes

2. Log in to TGD Wiki. If you need a username and password, contact us at: ciliate-curator@bradley.edu

3. On the gene's page, click Edit next to Identifiers and Description. Enter the gene's name in the Gene Name box. (Ex. PDD1) Note that the name must conform to the naming guidelines.

> Note: Try to avoid naming genes as the Tetrahymena homolog of (Important Gene) = TXX1. If all the genes were to start with "T", we could only generate 26x26=676 gene prefixes.

4. In the Name Description box, enter the phrase containing the letters that make up the prefix of the gene. (Ex. Programmed DNA Degradation)

5. Update the Headline section if you find it to be incorrect, or if you have discovered new information that is not found there. This text will show up as the Description seen on the Gene Page. Possible functions or homologs are good pieces of information to include in the Headline. Separate phrases with semicolons rather than periods in this section, and avoid complete sentences. There is an upper limit of 240 characters that can be listed in the Headline, so try to be succinct.

6. Click Update to save your information, then click Go Back to see the changes on the gene page.

7. The new gene name will now show up in the list of named genes: http://ciliate.org/index.php/show/namedgenes

8. If you have entered a report on the data used to name the gene at SUPRDB, enter the SUPRDB ID in the Associated Literature section. See the SUPRDB Edit Guide below for more information.

Welcome to SUPRDB!

SUPRDB collects unpublished scholarly reports and makes them available on the web. You can then use the SUPRDB ID and link provided by our site to support annotations at other databases.

> **At the end of this protocol you will find a guide for entering reports that

describe BLAST data used for naming genes.

Guide to using SUPRDB

1. Write your report in your favorite word processing program.

    a. Keep formatting to a minimum. Some types of formatting are incompatible with SUPRDB. If you receive an error while trying to upload your text, try selecting your text and hitting the "Erase Formatting" icon in the toolbar. You can reformat your text using the tools at SUPRDB. Alternatively, using the "Paste and match formatting" option can work.

2. Log in to SUPRDB. If you don't have an account, contact any person who does and ask them to sign you up. Any user with an account has the ability to register other users.

3. Add a New Project. The button for this is at the top of the page.

4. Type in the name of the first author. Standard format is (First Name) (Middle Initial) (Last Name). To add more authors, click the Add Another Author link and a new box will appear.

5. Enter the Project Title. Also select a Project Type. You must select at least one Project Type. If you do not see a phrase that describes the work you have done, email us and we will add an appropriate description for you to select.

6. Before hitting Next, which will save your progress and allow you to enter the text and figures of your report, double check the Authors and Title. You will not be able to alter these once these are entered. You will be able to add additional Project Types on the next page. When you are happy with your entries, hit Next to proceed.

7. The SUPRDB ID of your project will now be displayed in the upper left corner of the page. This ID is unique to your report, and can be used to link from other databases.

8. To add text to the Abstract, Introduction, Methods, Results, Discussion, and References sections, click "Edit this section" to the right of each portion of the report. A box will appear. Copy and paste text for these sections from your original document. You may need to format some of the text using the toolbar options.

9. Upload figures in .jpg format only. There is a maximum size SUPRDB will allow for figures.

10. Add more Project Types, Constructs, Genes, and GO Annotations as appropriate.

    a. When adding GO annotations, add only the numerical portion of the GO ID.

11. Hit Save this Section and View Project when you are done proofreading your entry.

Adding BLAST data to support Gene Names at TGD:

Please take the time to fill out a SUPRDB report, even if a BLAST result with a similar score may be found under the Homologs section of the TGD Gene Page. This will let other people know that someone has looked to see if better names exist, and if other paralogs exist.

1. Consider as Title such as "BLAST Analysis of TET1 (TTHERM_#######)", or "BLAST Analysis of the TET gene family", depending on how many genes are analyzed in the study.

2. Provide a link to the TGD page for each gene mentioned in the report by entering their TTHERM IDs with the Add Feature option.

3. The text in each section does not need to be lengthy, but the report should provide putative functions of each gene based on sequence similarity. It should also describe how the search was performed: names of sequences, programs, and databases that were used.

4. In the Results section, provide E-values and Scores for the comparisons used to characterize each gene. You may wish to include screenshots of the BLAST results to show the extent of the similarity.

5. Once you are done writing your report, visit the TGD page for each of the genes you characterize. Log in to TGD and load the SUPR ID (found at the top of the SUPRDB page) of your report under Associated Literature. TGD will automatically provide a link to the SUPRDB page of your report. You can also enter GO Annotations from your BLAST data using the SUPR ID. Annotations using BLAST data should receive the evidence code ISA (Inferred from Sequence Alignment).

6. Don't forget to add the gene's new name at TGD when you're done!

Thank you very much for your contributions to these community resources.

- The Staff at Ciliate.org