

I. Definitions of Gene Name Fields in Database:

A family name refers to the name of a group of genes that share some elements of function and sequence, for example PFAM motifs. The transcription factors that bind DNA through a basic helix-loop-helix structure belong to the bHLH family of transcription factors. In many cases there are already defined families and sub-families that one can identify by following the thread leading from the defined ortholog of the sea urchin gene in question. If you are not sure about the family leave the box blank.

For the tubulin case see:

<http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR000217>

Here "tubulin" is the family name and beta-tubulin is the "child" which I would use as the common name: Sp-beta-tubulin.

The common name is the name that identifies the gene according to the rules we drafted in the past weeks. For example, the sea urchin homolog of the brachyury protein will have the common name of Sp-Bra. Synonyms are Sp-Ta, T-box protein and T-protein. NB: In this case the gene family is the T-box gene family.

It is important not to confuse general terms connoting function with gene families, viz. "transcription factor" is not a gene family. Remember that the Interpro, PFAM and SwissProt pages on the ortholog to your gene are good sources of information on family structure and synonyms.

II. Gene Naming Conventions Proposed by Andy Cameron

After looking at the various schemes used in other sequenced genomes and talking to Paul Sternberg of Wormbase, I propose a scheme for gene names and gene symbols to be used in the annotation of the purple sea urchin genome. The gene symbol uses the specification of the mouse nomenclature with the addition of a sea urchin identifier, e.g., Sp-Otx. The specification is described on the mouse Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine. In particular the nomenclature page "Quick Guide to Nomenclature for Genes" (URL:http://www.informatics.jax.org/mgihome/nomen/short_gene.shtml[August, 2005]). The symbol description below is cited as coming from Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT, and the members of the Mouse Genome Database Group. 2003. MGD: The Mouse Genome Database. Nucleic Acids Res 31: 193-195.

SYMBOLS

- 1) Symbols are 3-5 characters, add additional characters as necessary up to 10.
- 2) Symbols begin with an uppercase letter followed by all lowercase letters.
- 3) Use punctuation only to separate two adjacent numbers (e.g., Lamb1-2) or for designating related (e.g., Es10-rs1). sequences and pseudogenes (e.g., Adh5-ps1).
- 4) Gene Family members: Use common stem (or root) symbol (e.g., see Cldn#).

- 5) If there is a homolog in vertebrates try to use the vertebrate symbol and any family convention used in vertebrates.
- 6) For single homologous genes from invertebrates not found in vertebrates use the original symbol from that species but include the word homolog at the end of the name followed by the name of the species in parentheses (e.g., symbol: Cdc20; name: cell division cycle 20 homolog (*S. cerevisiae*)).
- 7) If there is more than one sea urchin homolog for the invertebrate gene, assign the serial number after the word "homolog" (e.g., symbols: Atoh1 and Atoh2; names: atonal homolog 1 (*Drosophila*) and atonal homolog 2 (*Drosophila*) respectively.)
- 8) If the invertebrate/prokaryotic gene is similar to the sea urchin gene but is not determined to be a homolog, use the letter "l" to denote "-like" designations (e.g., symbol: Ash2l; name: ash2 (absent, small, or homeotic)-like (*Drosophila*)).

NAMES

- 1) Should be brief and specific.
- 2) Should convey the character or function of the gene.
- 3) In most cases there will be a usable name from the homolog.