# Representing Pseudogenes and related features at SGD

Rama Balakrishnan

Saccharomyces Genome Database

Stanford University

# SO Pseudogene definition

- SO-A sequence that closely resembles a known functional gene,at another locus within a genome, that is non-functional as a consequence of (usually several) mutations that prevent either its transcription or translation (or both). In general, pseudogenes result from either reverse transcription of a transcript of their "normal"paralog (SO:0000043) (in which case the pseudogene typically lacks introns and includes a poly(A) tail) or from recombination(SO:0000044) (in which case the pseudogene is typically a tandem duplication of its "normal" paralog). On occasion a pseudogene is functional as a consequence being "captured" by a non-paralogous gene,it is then known as a "captured_pseudogene"(SO:0100042).

# SGD pseudogene definition

- Two forms of pseudogenes-
  - Processed (mRNA transcript is reverse transcribed and re-integrated into the genome)
  - Unprocessed (arise from duplication of the gene in the genome and subsequent disablement)
- SGD-A DNA sequence similar to that of a functional gene within the yeast genome, but rendered non-functional through mutation. Pseudogenes typically result from gene duplication events within the same genome, followed by mutation, so that they are no longer transcribed or translated. Open Reading frames (ORFs) that are found to be non-functional in S288C are labelled as pseudogenes in SGD even though these ORFs may be functional in other strains of S.cerevisiae.Two forms of pseudogenes occur- processed and unprocessed. In yeast there are no processed pseudogenes and SGD has annotated the non-processed pseudogenes.

# At SGD….

- No known ‘processed pseudogenes’ in yeast
- Unprocessed pseudogenes
  - DNA sequence that has a start and a stop in the Systematic sequence (S288C strain)
  - Has an internal disablement (sequencing errors?)
  - Is duplicated within the genome (% identity>90%)
  - We look at other evidence like conservation in other Saccharomyces species, strains
  - We look at the literature to see if there is supporting evidence
- Example: FDH2 (YPL275W and YPL276W)

# Example of pseudogene in SGD

# What other types of mutations in S288C are we trying to classify?

- Sequences with a start and stop, and an internal stop or disablement in S288C, but are good ORFs in other Saccharomyces species, strains
- Had we not compared these sequences with other strains, species, we would have discarded them
- ORFs that are disabled but not duplicated within the S288C genome
- Example: CRS5/YOR031W
- What SO terms would apply?

# Current SO terms that were considered

- --Pseudogenic

  -----pseudogene

  -----decayed_exon
- Pseudogenic: A non-functional descendent of a functional entity
- decayed_exon: A non-functional descendent of an exon
- sequence_variant (SO:0000109)definition: A region of sequence where variation has been observed.
- mutation_affecting_coding_sequence (SO:1000054)definition: Any of the amino acid coding triplets of a gene are affected by the DNA mutation

# Suggestion

- Is there a term to indicate an internal mutation?
- Can we add a new term ‘blocked reading frame’ (synonym: closed reading frame)?