Using Transcriptome and Genome Sequences To Find Coding Sequences For A Specific Protein

Benjamin King, MDIBL
Overall Process

Align a protein from one organism (human) to all skate proteins and find best hit from the skate

Align the best hit from the skate against all human proteins
Claim putative orthology if the best human protein hit is the sequence we started with.
Workflow

1. Retrieve a human protein sequence for a gene of interest

2. Align this human protein sequence to all skate transcriptome contigs

3. Retrieve the transcriptome contig sequence for best hit

4. Align best hit transcriptome contig against all human proteins
Non-Unique Alignments

Adapted from Nardone et al., 2004
Step 1: Direct Homology-based Assignments

**BLAST vs. MGI/RefSeq/Locus Link**
- >98% ID, >100 bp to mouse
- Match full CDS
- Assign Gene Name (complete)
  - Evidence: BLASTN
- Assign Gene Name (partial, etc.)
  - Evidence: BLASTN

**BLASTX vs. SP/PIR/SP_TREMBL**
- >85% ID, >90% length
- Match full length
- Assign protein name homolog/species (complete)
  - Evidence: BLASTX
- Assign protein name homolog/species (partial, etc.)
  - Evidence: BLASTX

- >70% ID, >70% length
  - Match full length
  - Assign "similar to protein name" (complete)
    - Evidence: BLASTX

- >50% ID, >50% length
  - Match full length
  - "weakly similar to protein name" (partial, etc.)
    - Evidence: BLASTX

**No Direct Homology-based Assignment – step 2**

**NOTE:** Use BLAST to Note frameshifts

Carol Bult, Jackson Lab