Transcriptome Annotation

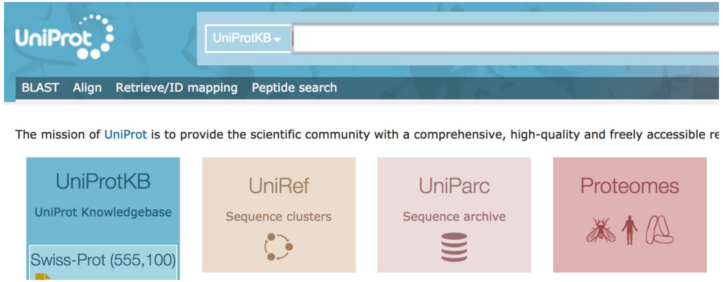
- **Why annotate?**
- Assign some biological identity/meaning/function to assembled transcripts

- **How do we get this information?**
- Look for sequence similarity of our transcripts with known genes in other orgs.

- **Can we just use a closely related organism’s genome, if it is available?**
- Not necessarily....

Databases: UniProtKB, UniRef

- **Databases for annotation**
  - well-curated & maintained
  - info from many species



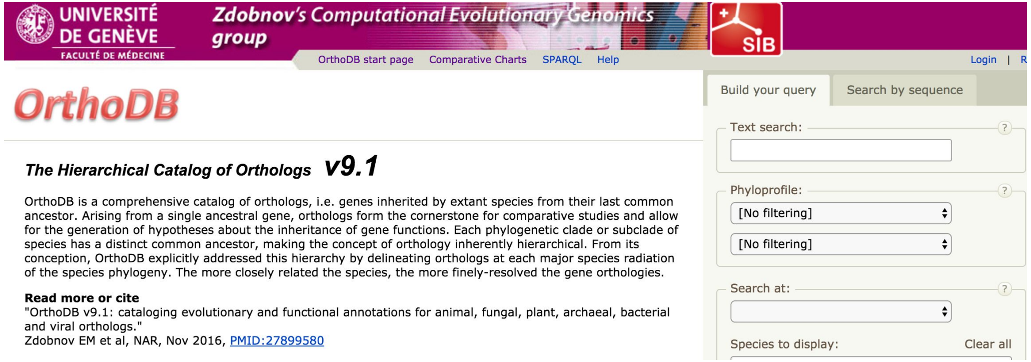
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible re

- **UniprotKB/UniRef**
  - Manually curated protein sequences + automated translations of genomes
  - **UniRef90**: sequences sharing 90% sequence similarity are clustered into a single entry (contains isoforms, homologs, etc)
  - Other data: biological/molecular function, domains, expression, PPIs

Magrane et al 2011

Databases: OrthoDB

- **OrthoDB**
  - catalog of protein-coding **orthologs** = genes in extant species arising from a single gene in a last common ancestor
  - delineates orthologs at each major radiation along species phylogeny
  - Other info: gene universality, duplicability, evolutionary rate, gene architecture



The Hierarchical Catalog of Orthologs **v9.1**

OrthoDB is a comprehensive catalog of orthologs, i.e. genes inherited by extant species from their last common ancestor. Arising from a single ancestral gene, orthologs form the cornerstone for comparative studies and allow for the generation of hypotheses about the inheritance of gene functions. Each phylogenetic clade or subclade of species has a distinct common ancestor, making the concept of orthology inherently hierarchical. From its conception, OrthoDB explicitly addressed this hierarchy by delineating orthologs at each major species radiation of the species phylogeny. The more closely related the species, the more finely-resolved the gene orthologies.

Read more or cite
"OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs."
Zdobnov EM et al, NAR, Nov 2016, PMID:27899580

Zdobnov et al 2017

Databases: PFAM and RFAM

- **PFAM**
  - catalog of protein families and **domains** (functional regions)
  - use HMMER to search against PFAM-A databases
  - HMMER uses Hidden Markov Model (HMM) approach to make more accurate predictions of remote homology than BLAST
- **RFAM**
  - Catalog of RNA families, mostly non-coding RNA genes
  - Uses covariance models to infer homology based on both sequence and secondary structure

Eddy 2004 “What is a hidden Markov Model?”
Finn et al 2016
Nawrocki et al 2016