OrthoDB v10: samplong the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs

Kriventseva et al. 2019 *Nucleic Acids Research*

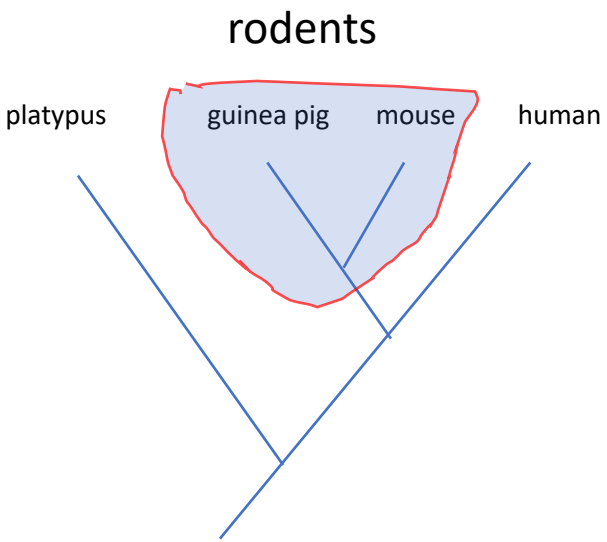
OrthoDB

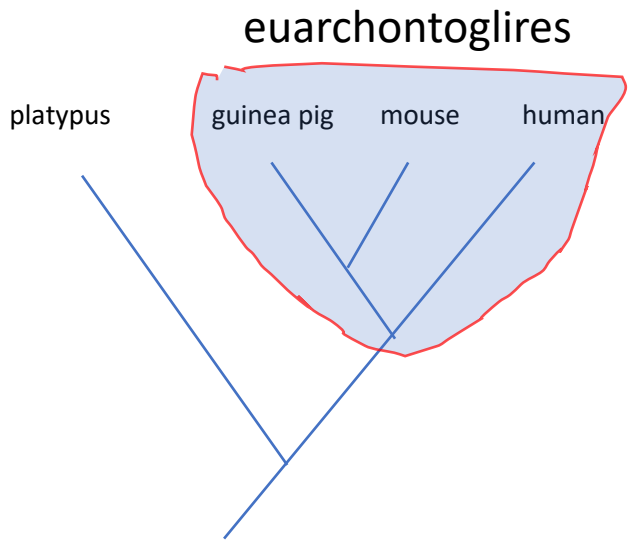
- Genomes can theoretically give us ***all the genes*** for an organism
- Homology can be used to hypothesize about gene functions
- Orthology – homology of genes due to speciation
  - orthologs
  - orthogroups (OGs) – groups of homologs with respect to their species radiation
  - the **cornerstone** of comparative genomics.

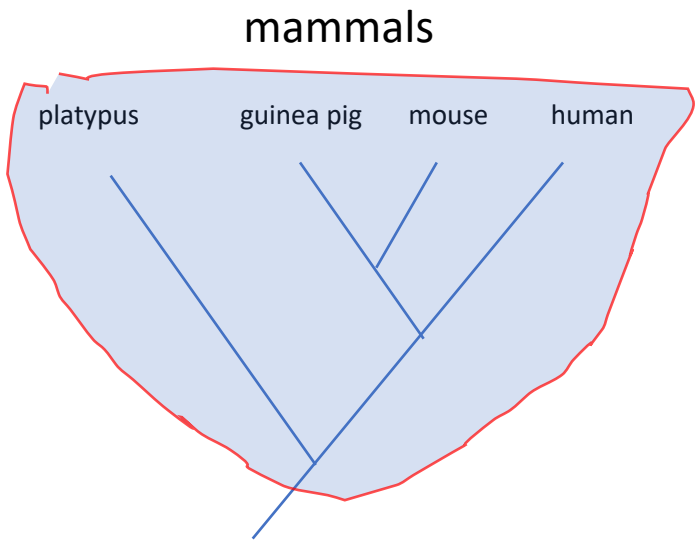
OrthoDB

- Based on the idea that genomics is inherently hierarchical
- Each group of species has a common ancestor
- Ortholog delineation is applied at each radiation

i.e., what are the orthologs for:







OrthoDB contains

- Functional annotations
  - Genes:
    - collects functional data attached to genes from other databases
      - (NCBI, UniProt, Gene Ontology etc.)
      - KEGG pathways
      - Online Mendelian Inheritance in Man (OMIM)
  - Orthogroups
    - aggregations corresponding to gene-level annotations
    - essentially uses text matching to determine the best scoring single phrase in common with all the genes in an orthogroup
  - Functional categories:
    - Clusters of Orthologous Genes (COG)
    - Gene Ontology (GO)
    - Kyoto Encyclopedia of Genes and Genomes

KEGG Pathways

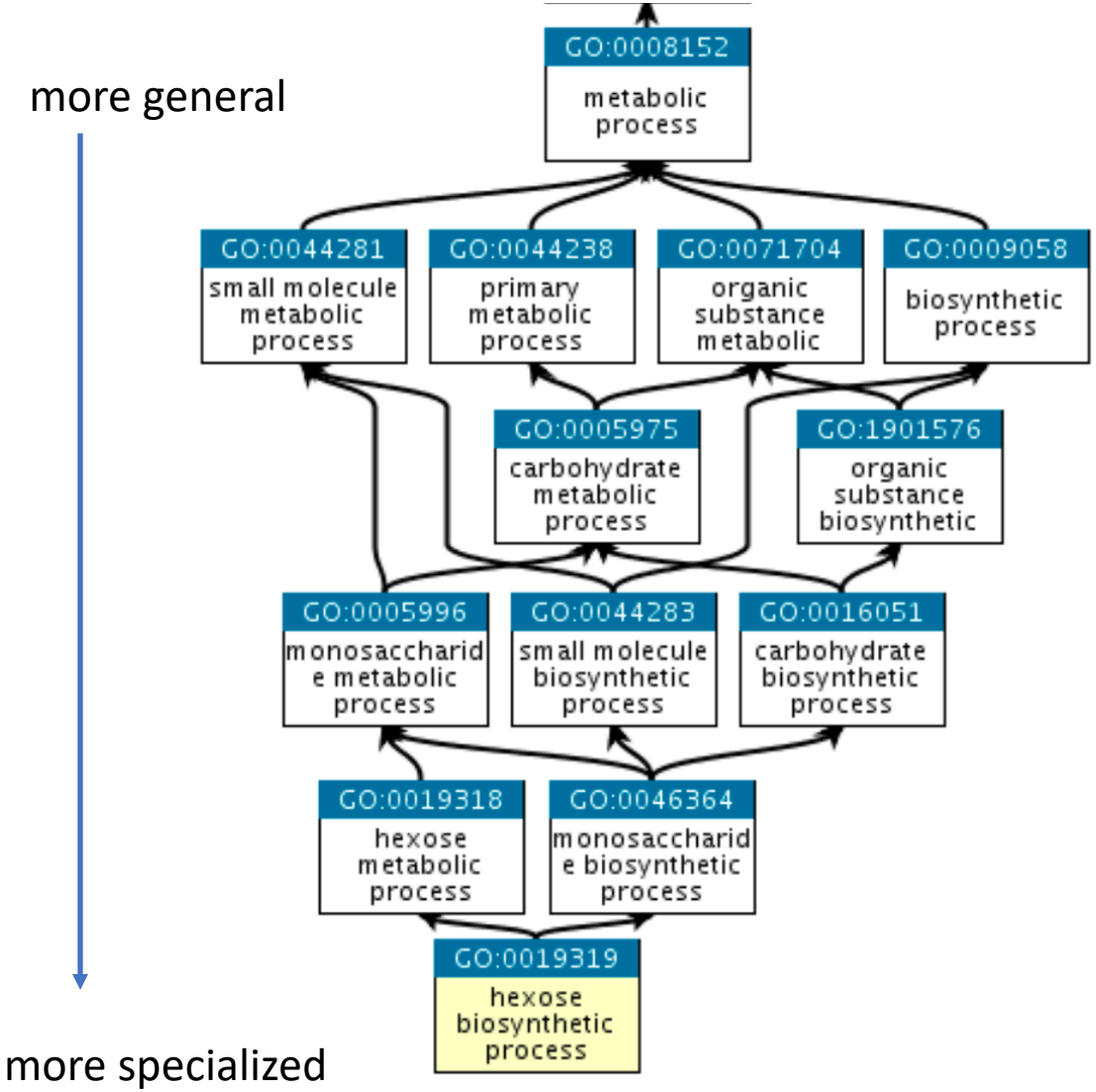
- A pathway is a series of interactions among molecules in a cell that leads to a certain product or change in a cell (Wikipedia)

- https://www.genome.jp/brite/br08901
- Metabolic pathways
- Genetic pathways
- Signal transduction pathways

Gene Ontology (geneontology.org)

- Ontologies consist of a set of classes (or terms or concepts) with relations that operate between them.
- For the biological domain, the Gene Ontology (GO) describes three aspects:
  - Molecular Function – what gene products do
  - Cellular Component – where in the cell a gene product performs its function
  - Biological Process – broader programs accomplished by molecular activities
- GO is like a graph, where each term is a node and relationships between terms are edges between the nodes
- Hierarchical, meaning it contains “parent” and “child” terms

Gene Ontology (geneontology.org)



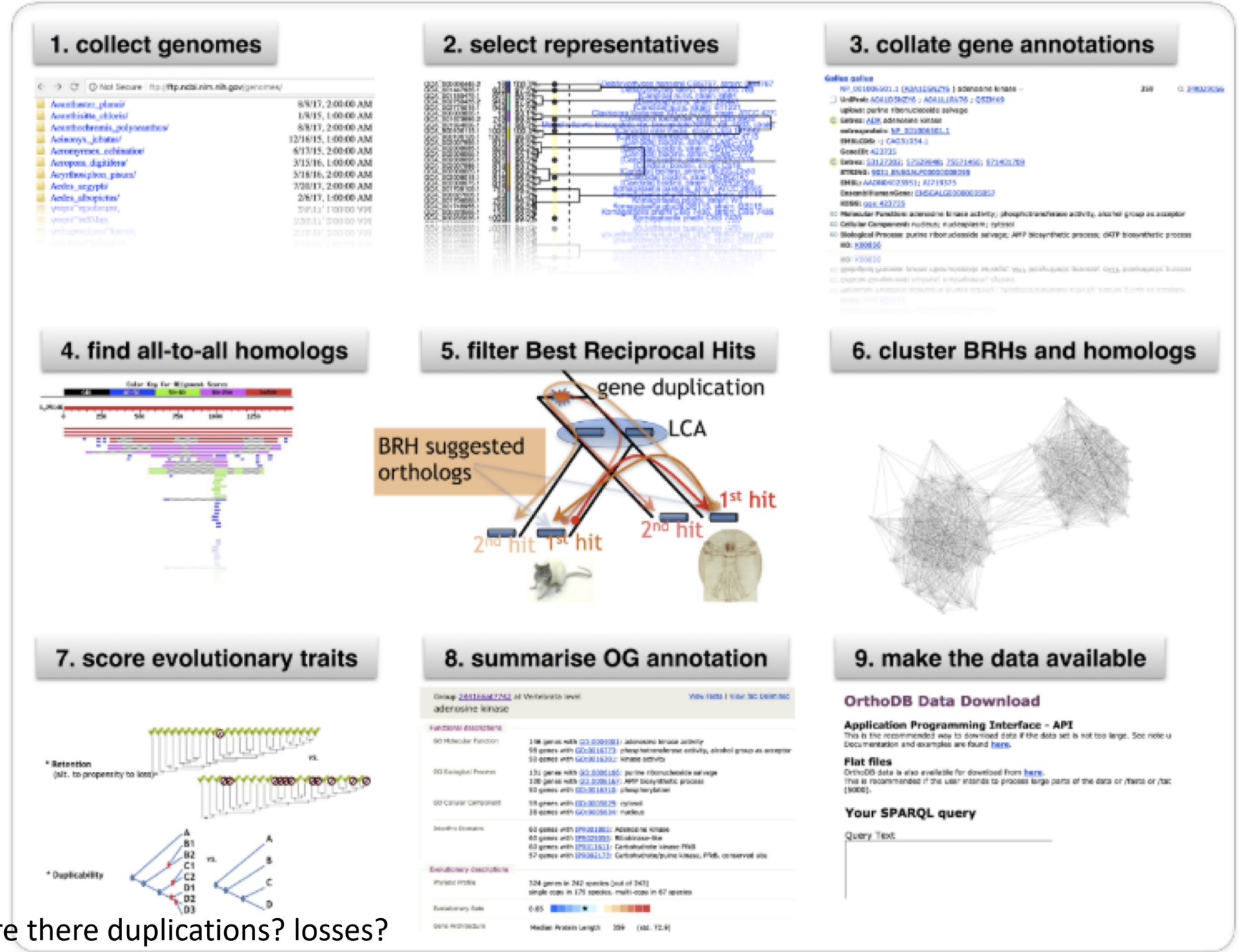
- "hexose biosynthetic process" has two parents
  - "hexose metabolic process"
  - "monosaccharide biosynthetic process"
    - "biosynthetic process" is a subtype of "metabolic process" AND hexose is a subtype of monosaccharide

OrthoDB contains

- Evolutionary annotations
  - phyletic profile: the proportion of species in an orthogroup with orthologs
    - Does a gene family have a function that is necessary for life (“basal”)?
      - Orco – co-receptor for insect odorant-sensing
      - or is the gene lineage-specific, reflecting unique adaptations (26Ab)
  - evolutionary rate: sequence divergence at the protein level
    - can indicate strength of selection acting on gene over deep time
    - slower=purifying selection (Orco); faster=positive selection (26Ab)
  - sibling groups: sequence uniqueness of the orthologs
    - can allow the discovery of closely related gene families with similar functions

graph-based phase:

tree-based phase:



are there duplications? losses?