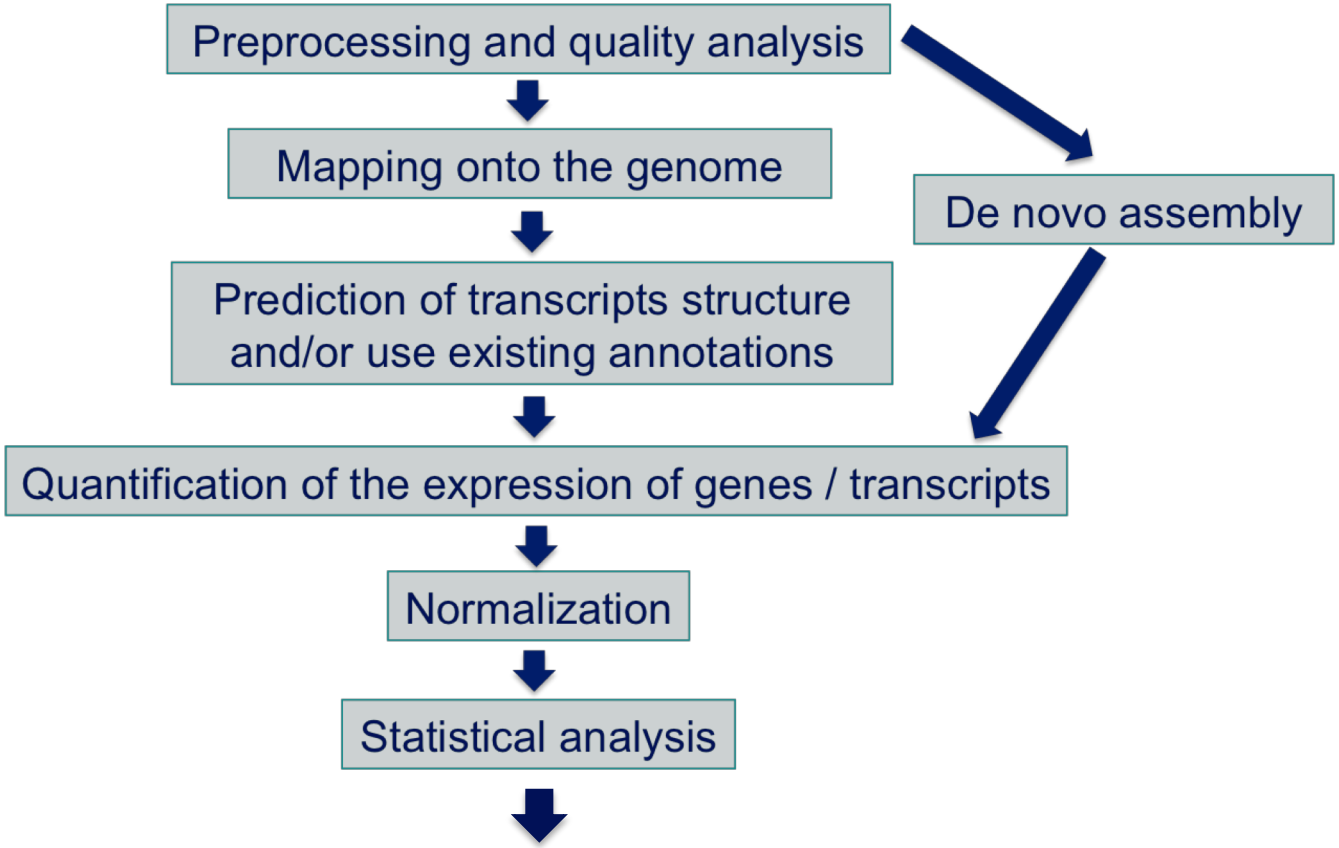
Functional analysis of RNA-seq data

Céline Keime
keime@igbmc.fr

Analysis of RNA-seq data

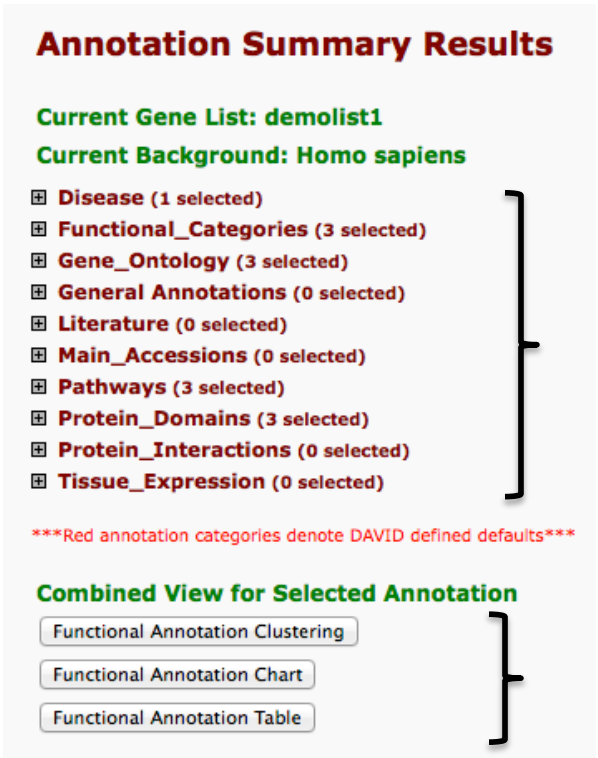


Functional enrichment analysis, pathway analysis, integration with other data, …

Functional analysis

- A lot of functional analysis tools available
  - Initially developed for microarray data
  - e.g. GO tools listed in
    http://geneontology.org/docs/go-enrichment-analysis/
  - Methods specific to RNA-seq data
    - Bioconductor packages
      - Goseq (Young et al., Genome Biology 2010;11:R14)
      - SeqGSEA (Wang et al. BMC Bioinformatics 2013, 14(Sup5):S16)
    - GSAASeqSP (Xiong et al Scientific Reports 2014; 4:6347)
- DAVID will be used for this practical session because
  - graphical interface & free software
- DAVID
  - **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery
  - https://david.ncifcrf.gov/
  - A very interested article describing how to use DAVID :
    Huang et al. Nature Protocols 2009;4(1):44-57.

DAVID



Different sources of annotation

- Disease (OMIM)
- Gene Ontology
- Pathways (KEGG, Biocarta)
- Protein Domains (InterPro, SMART)
- Protein Interaction (BIND)
- …

Different tools

- Functional Annotation Clustering
  - Cluster functionally similar terms associated with a gene list into groups
- Functional Annotation Chart
  - Identify enriched annotation terms associated with a gene list
- Functional Annotation Table
  - Query associated annotations for all genes from a list

Gene Ontology

- Defines concepts/classes used to describe gene function and relationships between these concepts
- Classifies functions along three aspects
  - Molecular function
    - Molecular activities of gene products
  - Cellular component
    - Where gene products are active
  - Biological process
    - Pathways and larger processes made up of the activities of multiple gene products

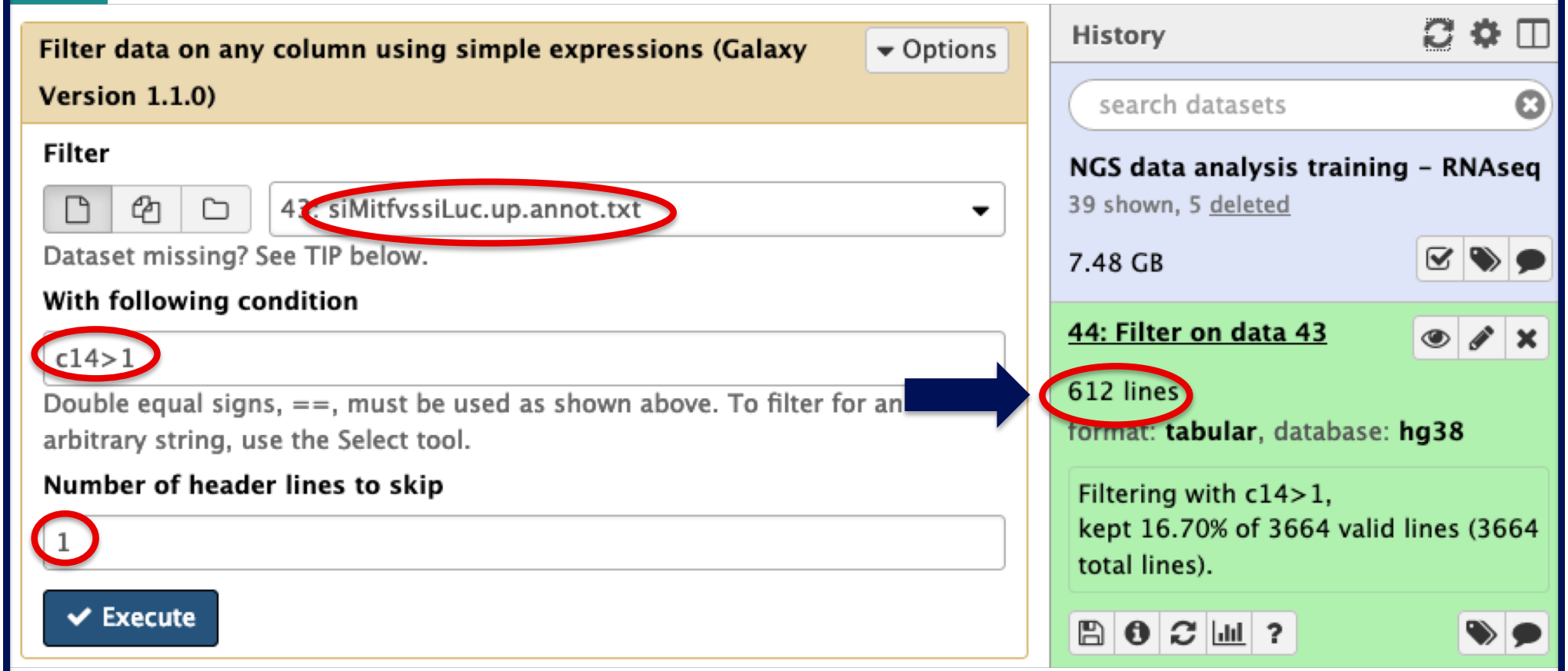
Exercise : functional analysis

- Use DAVID to perform functional analysis of genes significantly over-expressed in siMitf vs siLuc samples

1. Select over-expressed genes using the filter tool on GalaxEast
   - Proposed thresholds :
     Adjusted p-value < 0.05 and $\log_2$(Fold-Change) > 1
2. Create a file with gene name for all these genes using the cut tool on GalaxEast
3. Analyse this gene list using DAVID

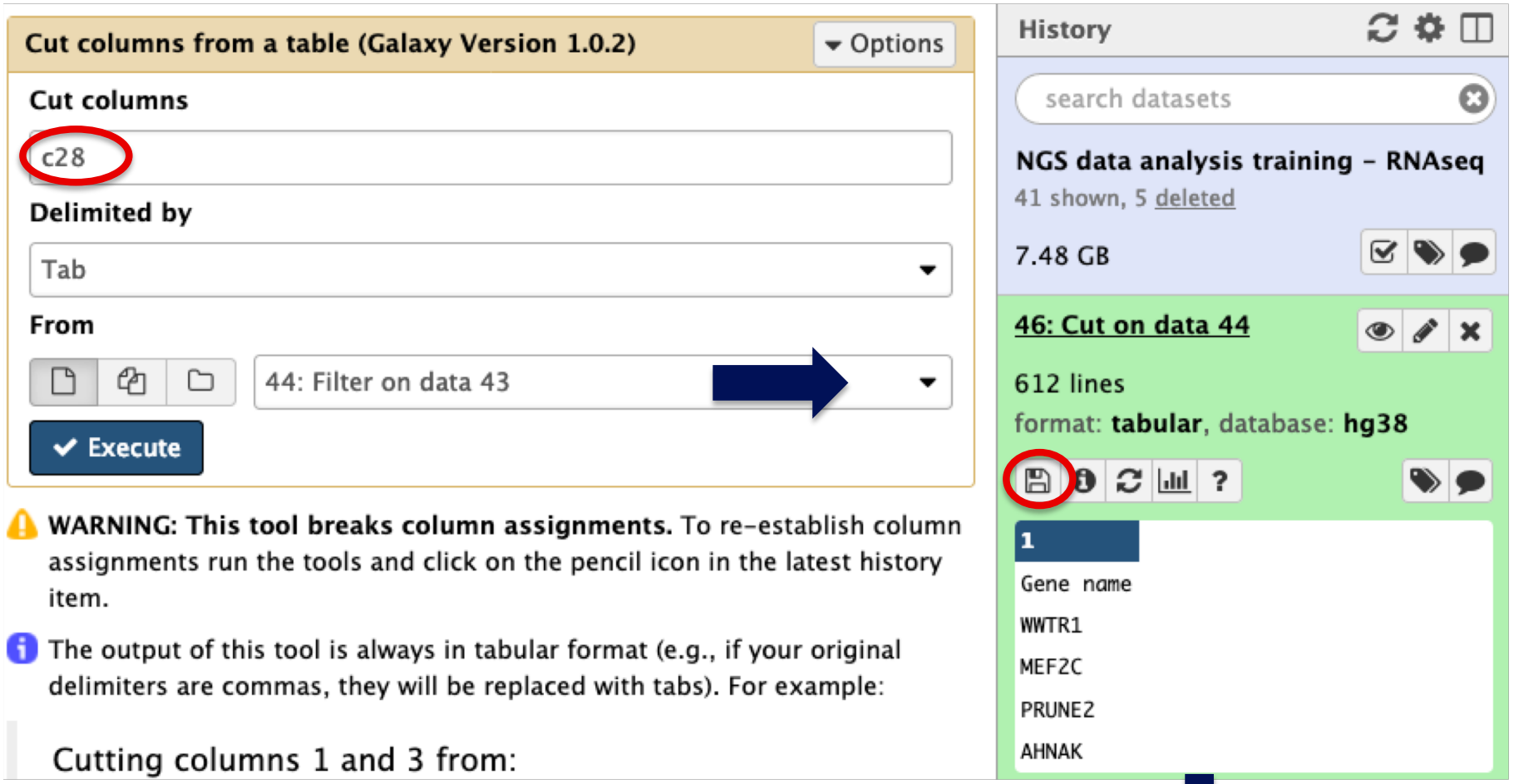
1. Select over-expressed genes

- Among significantly differentially expressed genes, select genes with $\log_2$(Fold-Change) > 1



2. Create a list of gene names

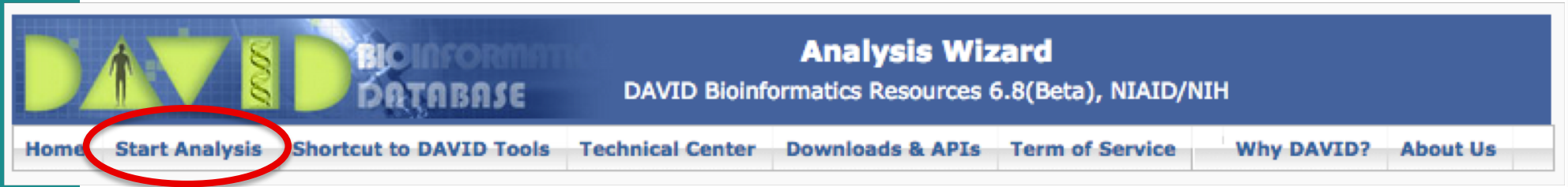
- Select associated gene names in the previous table



siMitfvssiLuc_upgenes_lfc1_padj005.txt file

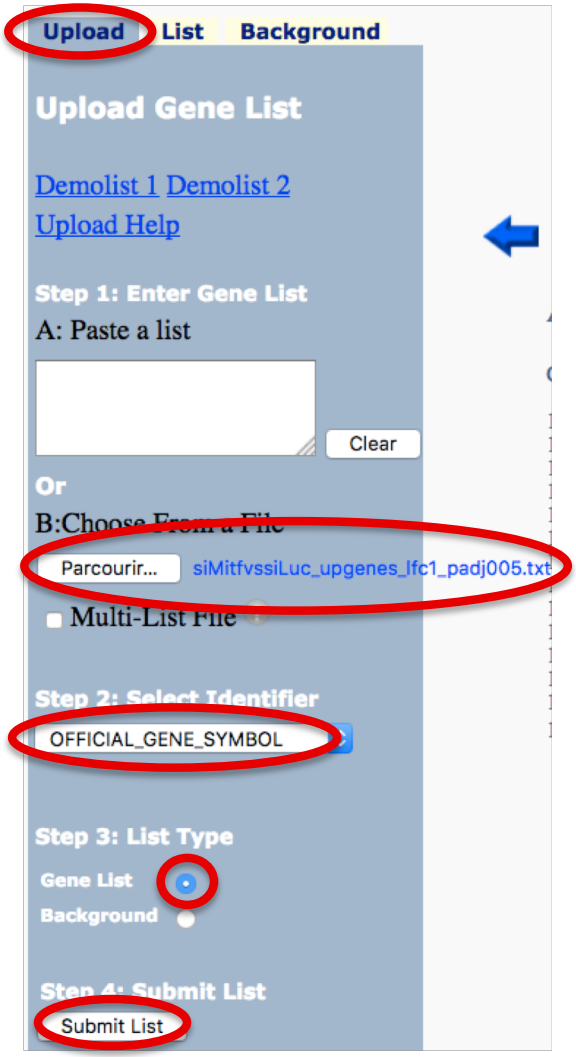
3. Analyse your gene list using DAVID

- Go to https://david.ncifcrf.gov
- Click on Start Analysis

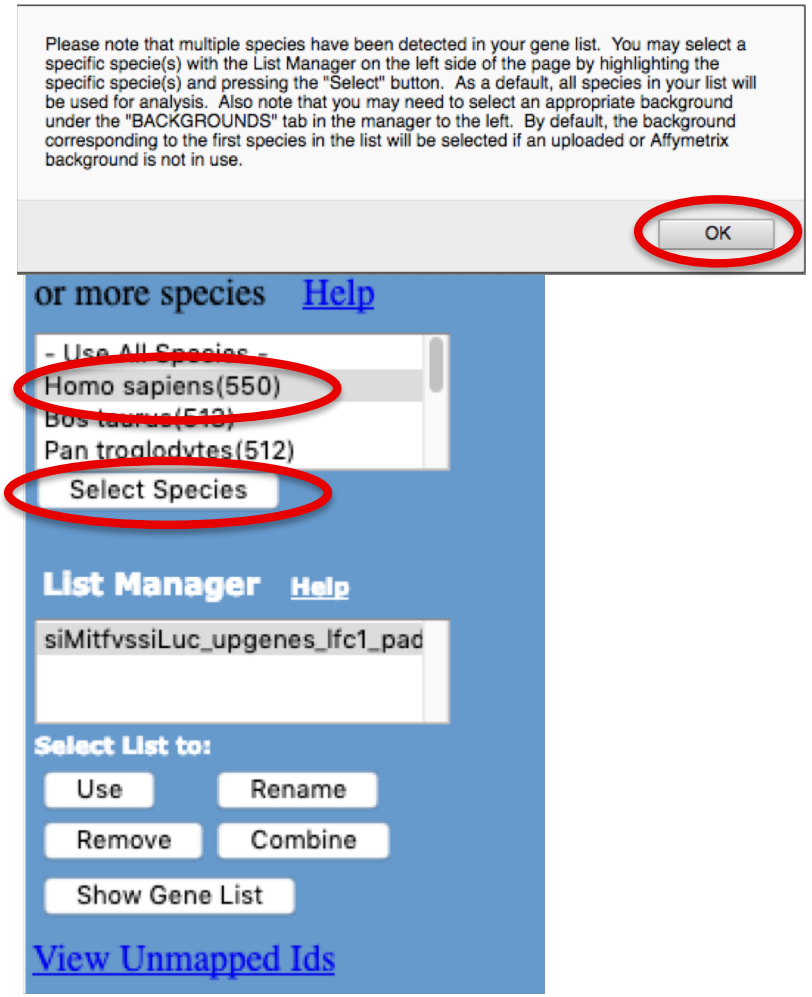


3. Start DAVID analysis

- Enter your gene list



- Select species



Exercise : functional analysis

1. What are the 10 most enriched functional annotation terms among annotations of the genes from your list ?

   How many genes are annotated with each of these terms ?

   Which genes are annotated with the most enriched term ?

2. As you see redundancy in previous results, it could be interesting to cluster functionally similar terms into groups.

   Look at the results of this clustering, for example for the first identified cluster.

   Click on  to visualize members of this cluster (genes and annotations).

3. *KIT ligand* (*KITLG*) gene is a member of this cluster.

   What are all associated annotations for this gene ?

   Among these annotations you will find the KEGG pathway "Ras signalling pathway".
   Are other genes from your list member of this pathway ?