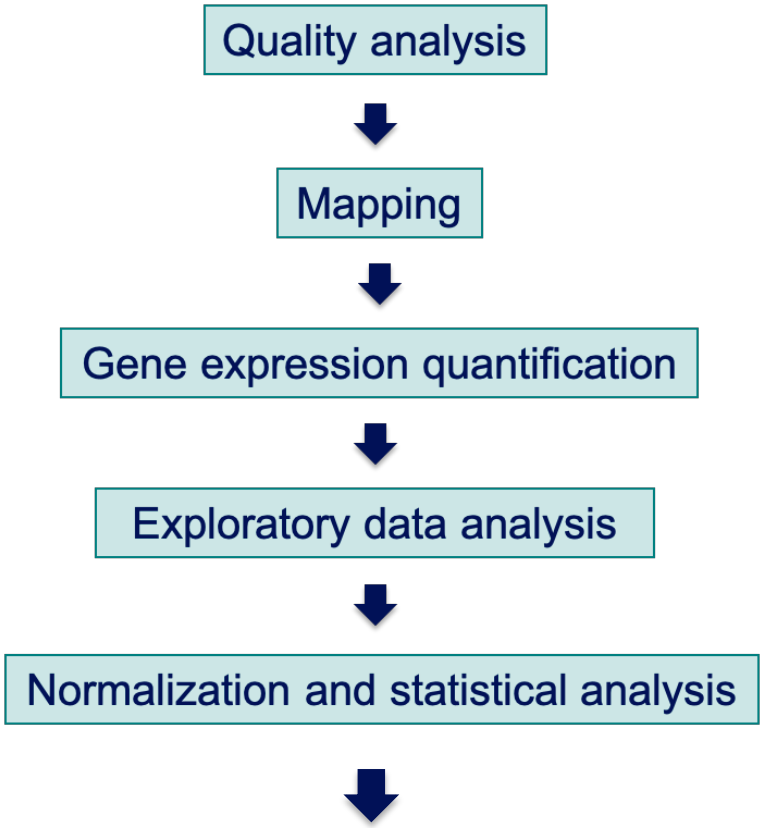
Functional analysis of RNA-seq data

Céline Keime
keime@igbmc.fr

Analysis of RNA-seq data

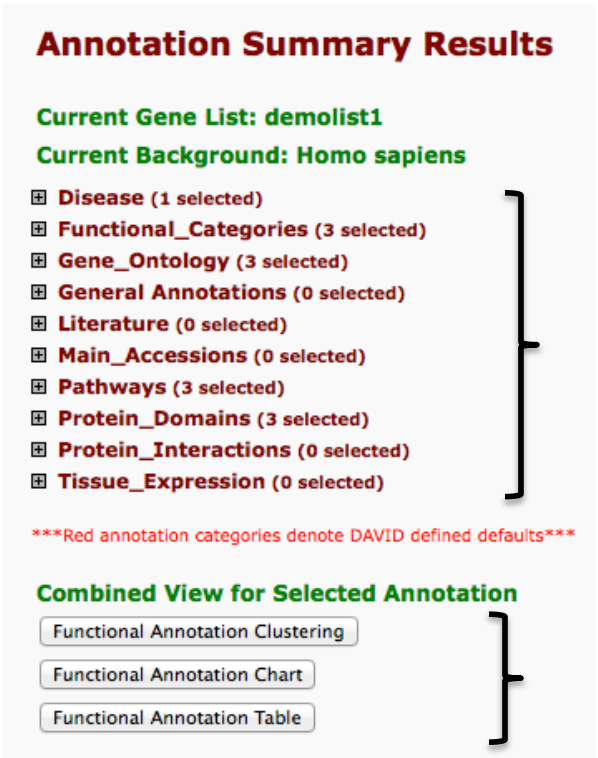


Functional enrichment analysis, pathway analysis, integration with other data, …

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

Exercise : functional analysis

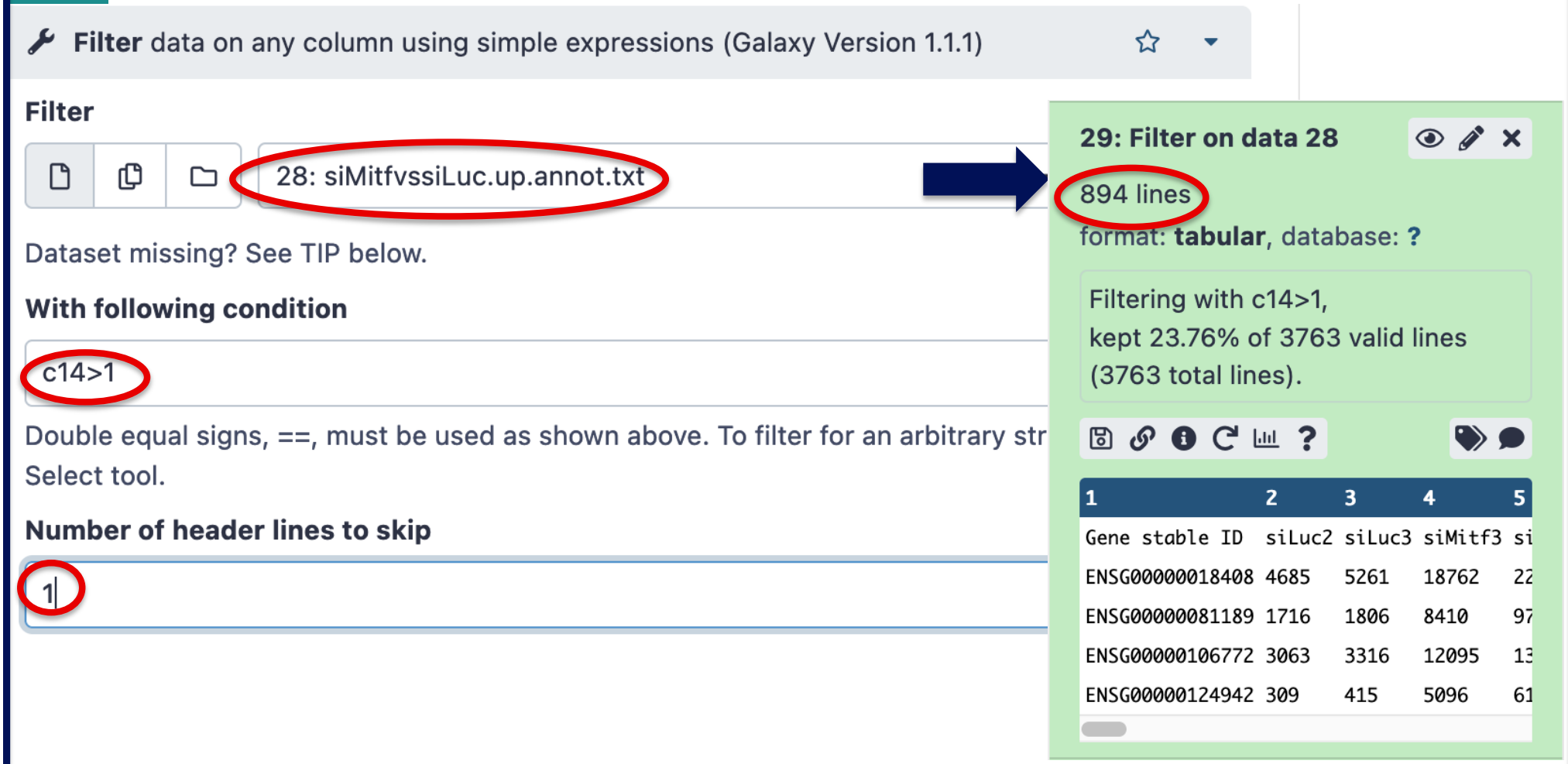
- Use DAVID to perform functional analysis of genes significantly over-expressed in siMitf vs siLuc samples
  - Using the thresholds : adjusted p-value < 0.05 and $\log_2$(Fold-Change) > 1
- For this purpose :
  1. Select over-expressed genes using the **Filter** tool on Galaxy
     - Input dataset : siMitfvssiLuc.up.annot.txt

     In your history or dataset 21 in "NGS data analysis training Strasbourg" history
     - Threshold : $\log_2$(Fold-Change) > 1

     Indeed, genes in siMitfvssiLuc.up.annot.txt file have already been selected with adjusted p-value < 0.05
     (cf "Threshold of statistical significance" in SARTools advanced parameters)
  2. Create a file with Ensembl gene ID for all these genes using the **Cut** tool on Galaxy
  3. Analyse this gene list using DAVID

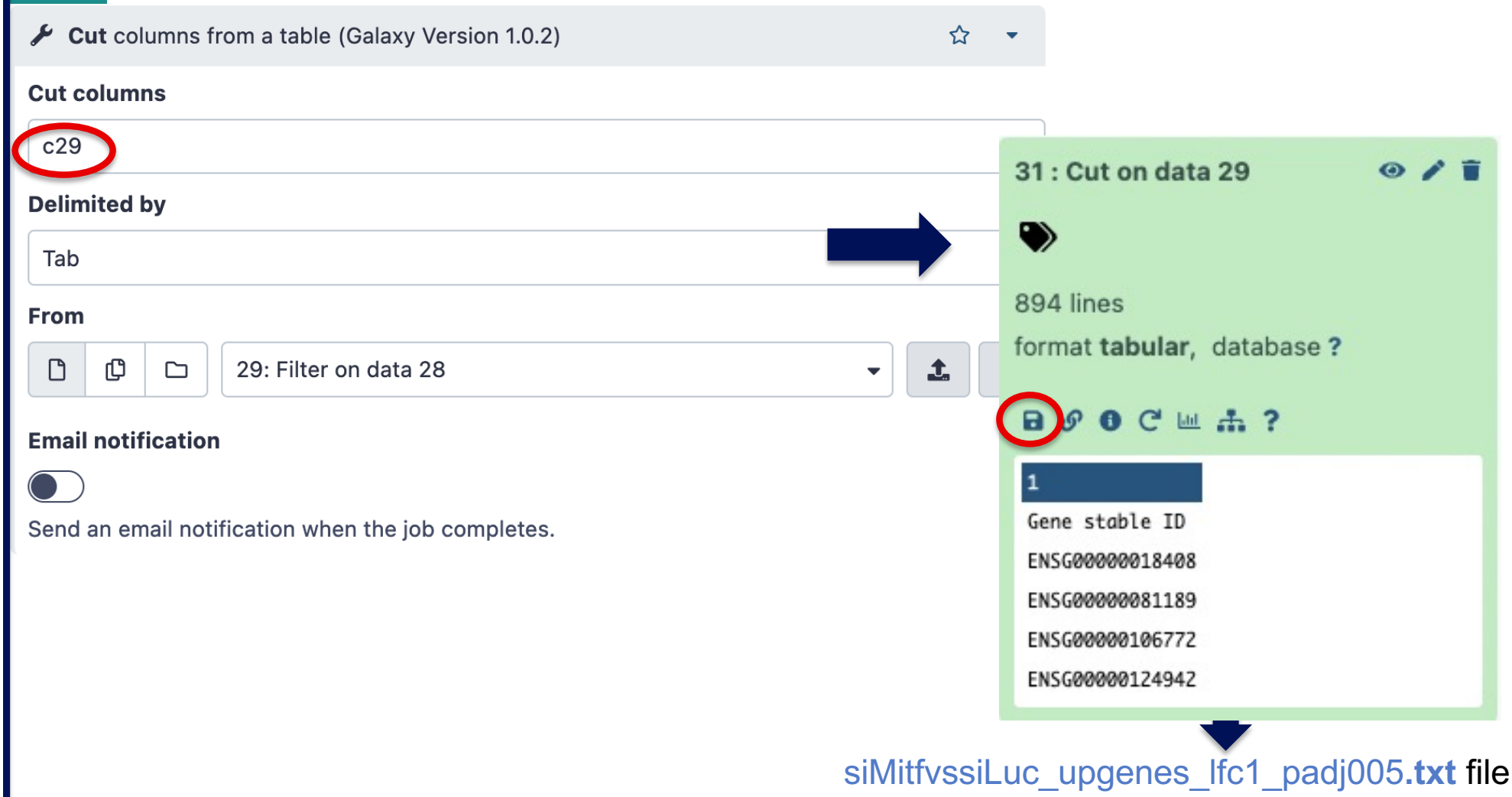
1. Select over-expressed genes

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



2. Create a list of gene names

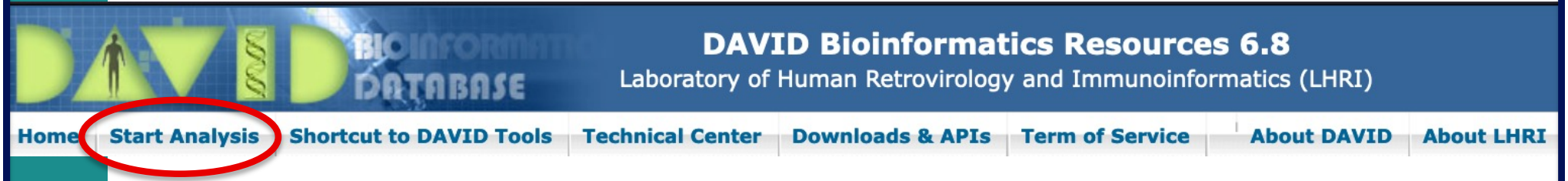
- Select associated gene names in the previous table



siMitfvssiLuc_upgenes_lfc1_padj005**.txt** file

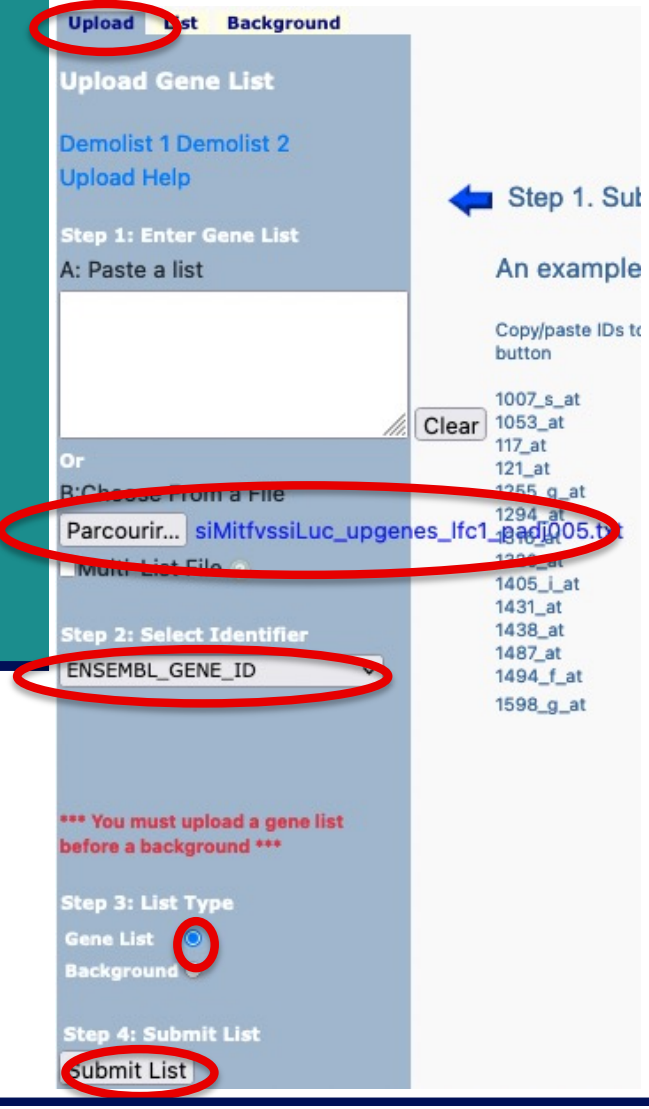
3. Analyse your gene list using DAVID

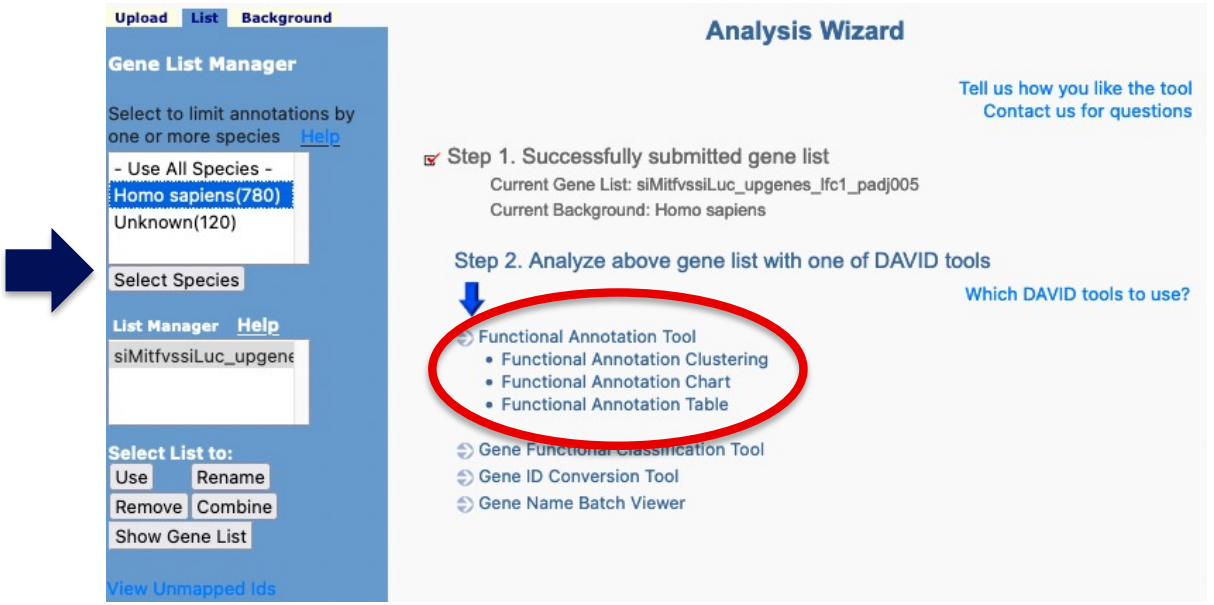
- Go to https://david.ncifcrf.gov
- Click on Start Analysis



3. Analyze your gene list using DAVID

- Enter your gene list
- Analyze your list





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 GO biological process term ?

2. *KIT ligand* (*KITLG*) gene is annotated with this GO term.

   What are all associated annotations for this gene ?

   Among these annotations you will find the KEGG pathway "PI3K-Akt signalling pathway".
   Are other genes from your list member of this pathway ?

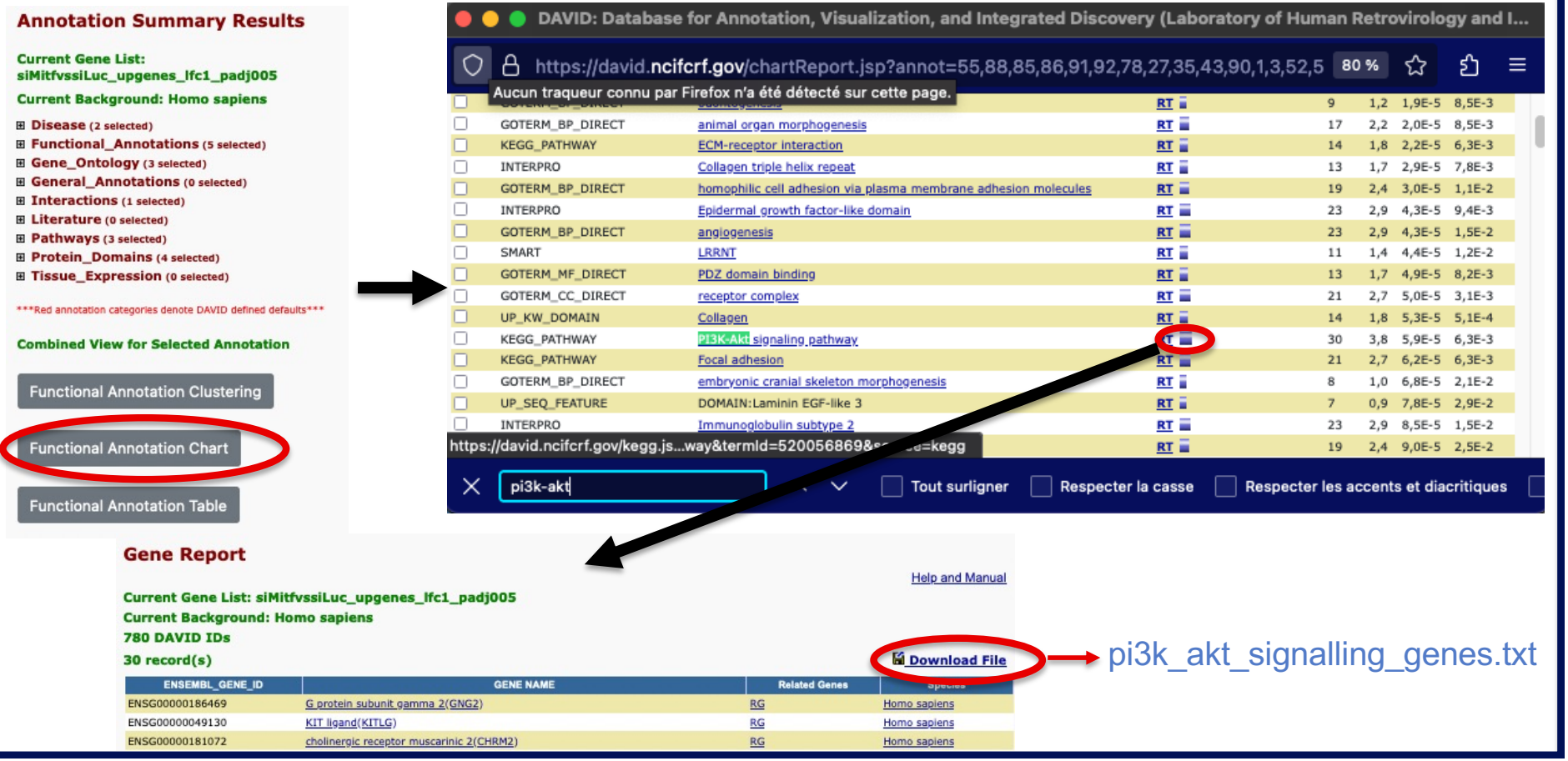
3. We would like to represent on an heatmap the variation of expression of all these genes (list genes in PI3K-Akt signalling pathway) in the four samples

   → Prepare a file with the normalized read counts for these genes in all samples using Galaxy and use Heatmapper (http://www.heatmapper.ca/expression/) to perform the heatmap

Exercise : functional analysis

3.1. Download list genes in PI3K-Akt signalling pathway from DAVID :

In “Functional Annotation Chart” results search for “PI3K-Akt”,
display the corresponding genes,
and then right click on Download File (top right) and save link target on disk



Exercise : functional analysis

3.2. On Galaxy, we will **join** the file obtained at step 3.1 with siMitfvssiLuc.up.annot.txt using the common column (containing Ensembl gene ID) → We will thus retain only PI3K-Akt signalling genes from siMitfvssiLuc.up.annot.txt file.

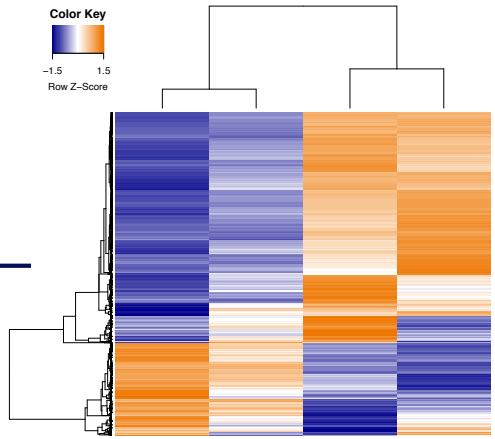
- Import pi3k_akt_signalling_genes.txt file on Galaxy
- On Galaxy, join siMitfvssiLuc.up.annot.txt with pi3k_akt_signalling_genes.txt on their common column (Ensembl gene ID)

3.3. On Galaxy, prepare a file with 5 columns : Gene name and four columns containing normalized read counts in the four samples (use the **Cut** tool and results obtained at step 3.2).

- Download this file
- Change file extension to txt and the name of the first column to NAME

3.4. Use this file to perform an **heatmap** representing the variation of expression of these genes in the four RNA-seq samples using Heatmapper (http://www.heatmapper.ca/expression/)

Heatmap and clustering



- Heatmap

  Colour-scaled representation of the data

  Data represented :

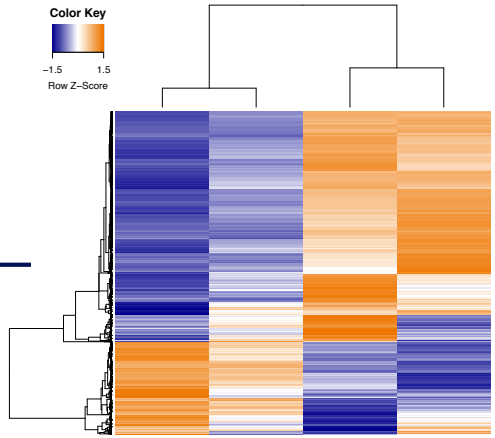
  - Expression
    - Normalized and divided by gene length

    → to compare the expression level of several genes
  - Expression variation
    - $\log_2$(Fold-Change)

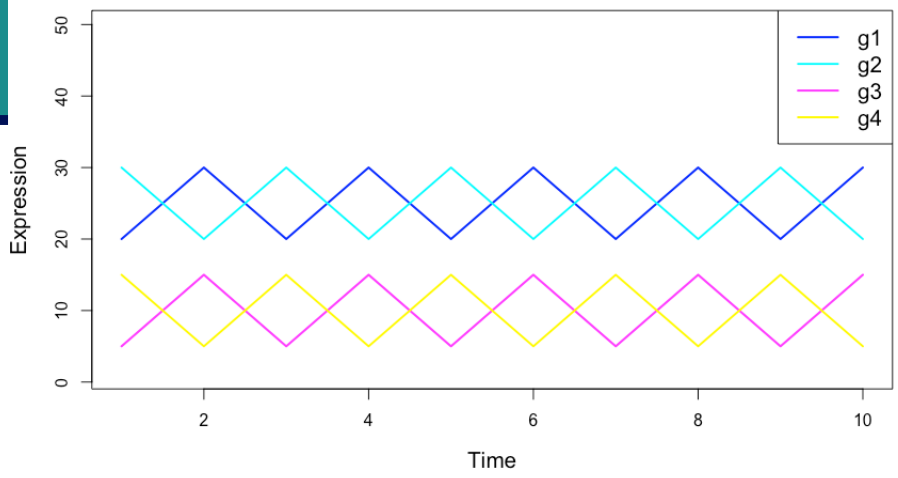
    log2 → over- and under-expression are on symmetric scales
    - Z-score

    → row z-score = [Value – mean(row)] / standard deviation(row)

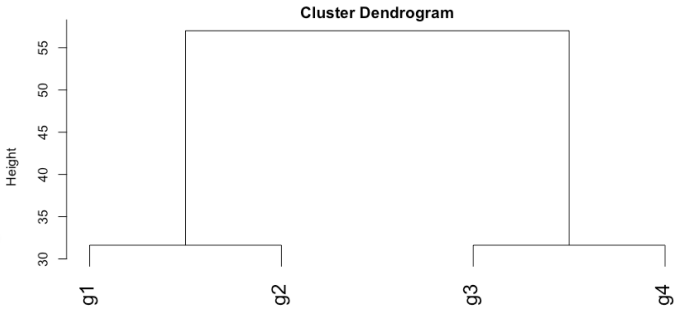
Heatmap and clustering



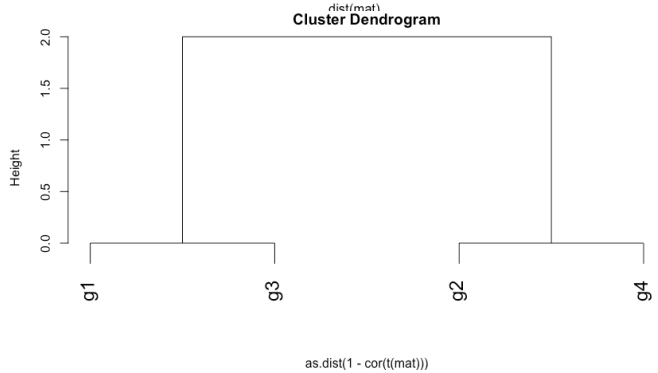
- Hierarchical clustering
  - Distance measure
    - Pairwise distance of all data points
    - Default in a lot of clustering software : Euclidean
    - If you want to group genes with similar expression patterns (i.e. on the shape of the expression profiles) : 1-correlation



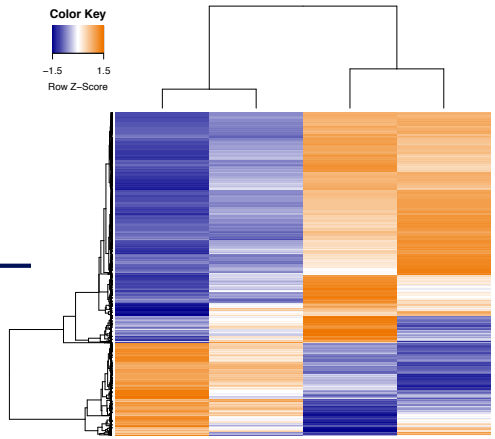
Euclidean distance



Pearson's distance



Heatmap and clustering



- Hierarchical clustering
  - Distance measure
    - Pairwise distance of all data points
    - Default in a lot of clustering software : Euclidean
    - If you want to group genes with similar expression patterns (i.e. on the shape of the expression profile) : 1-correlation
    - To group points
  - Clustering method
    - To join groups of points
    - Average : distance between two groups = average distance between all pairs of points from the two different groups