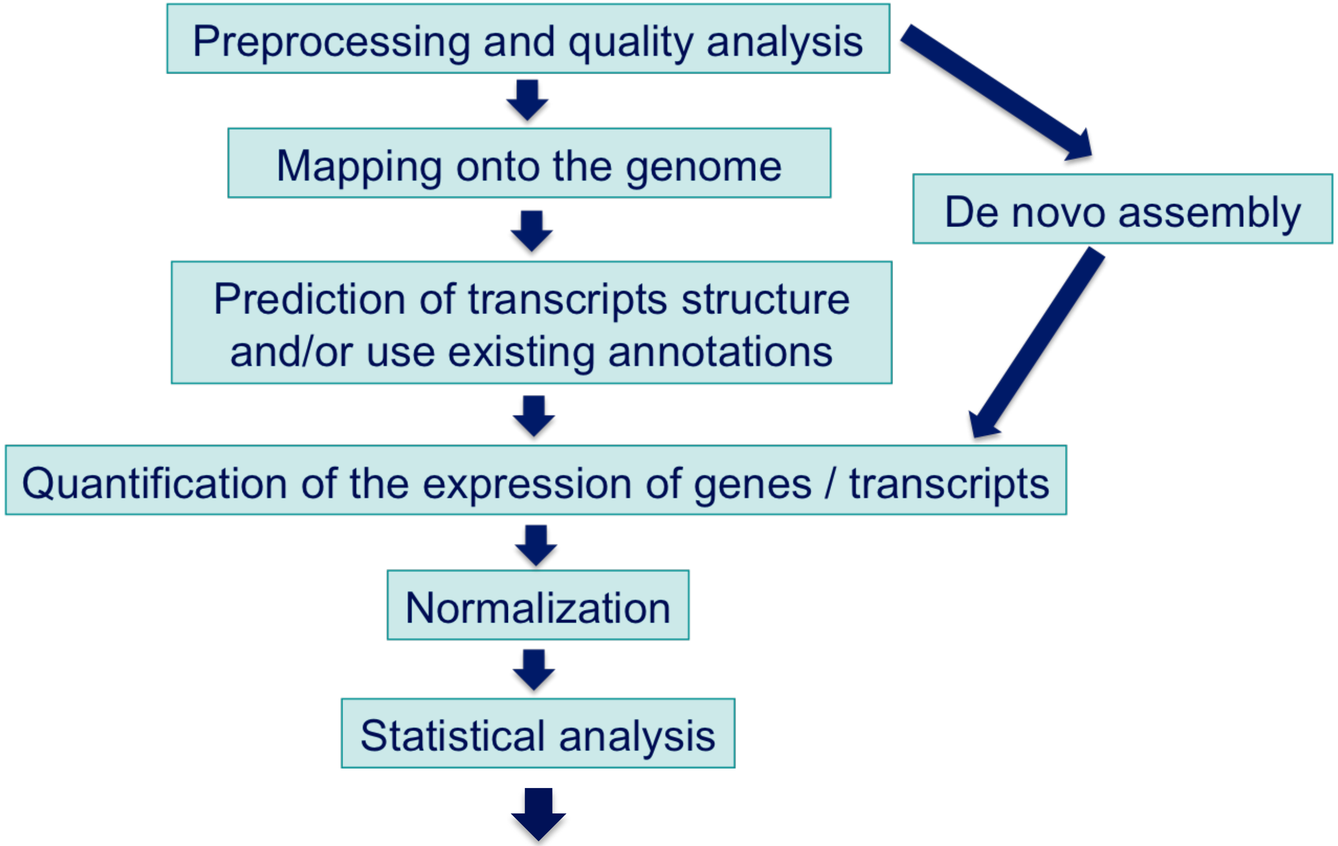
Functional analysis of RNA-seq data

Céline Keime
keime@igbmc.fr

Analysis of RNA-seq data



Preprocessing and quality analysis

Mapping onto the genome

De novo assembly

Prediction of transcripts structure and/or use existing annotations

Quantification of the expression of genes / transcripts

Normalization

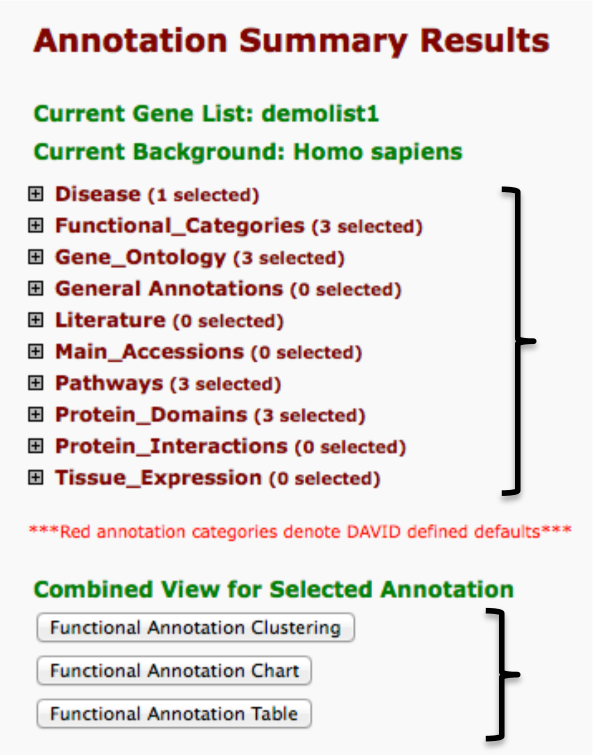
Statistical analysis

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



Annotation Summary Results

Current Gene List: demolist1

Current Background: Homo sapiens

- ⊞ **Disease** (1 selected)
- ⊞ **Functional_Categories** (3 selected)
- ⊞ **Gene_Ontology** (3 selected)
- ⊞ **General Annotations** (0 selected)
- ⊞ **Literature** (0 selected)
- ⊞ **Main_Accessions** (0 selected)
- ⊞ **Pathways** (3 selected)
- ⊞ **Protein_Domains** (3 selected)
- ⊞ **Protein_Interactions** (0 selected)
- ⊞ **Tissue_Expression** (0 selected)

Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table

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

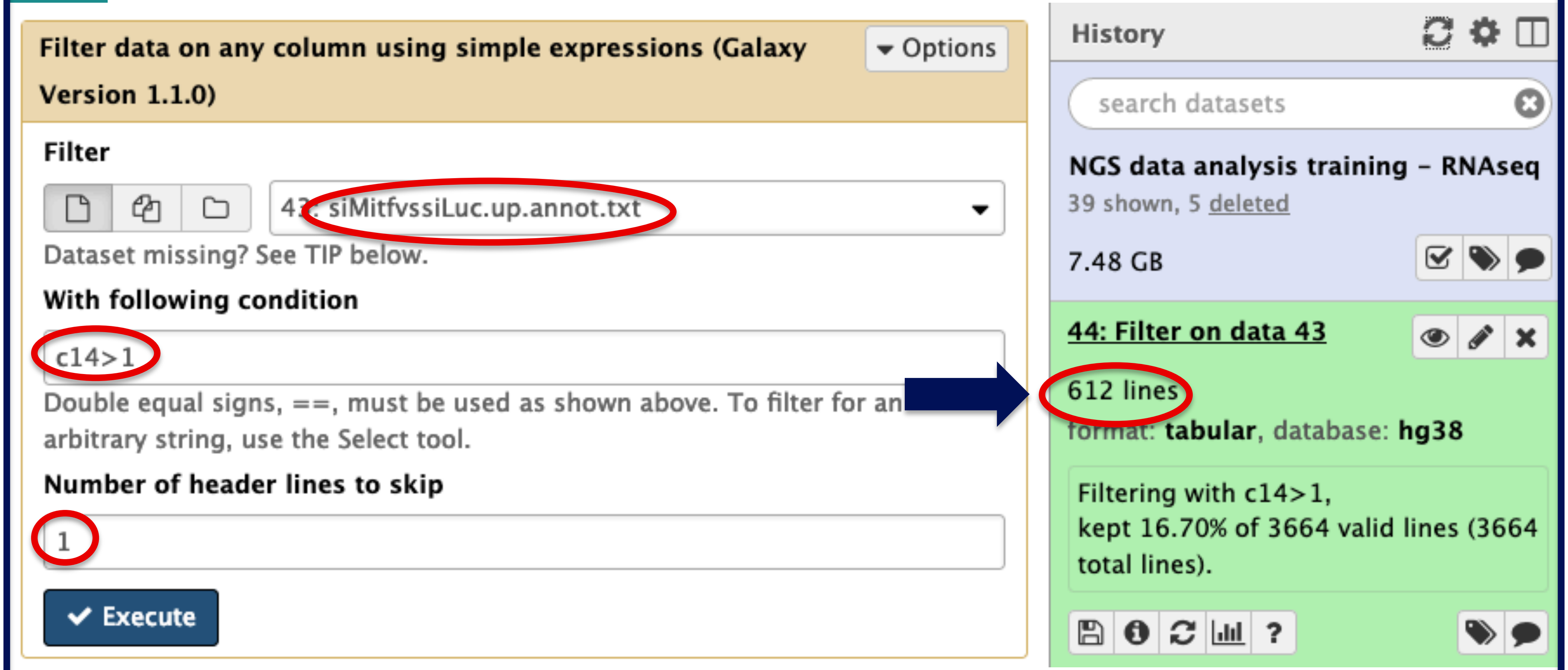
- Use DAVID to perform functional analysis of genes significantly over-expressed in siMitf vs siLuc samples

1. Select over-expressed genes using the filter tool on GalaxEast
   - Proposed thresholds :
     Adjusted p-value < 0.05 and $\log_2$(Fold-Change) > 1

   Genes in siMitfvssiLuc.up.annot.txt file have already been selected according to this adjusted p-value threshold
   (cf "Threshold of statistical significance" in SARTools advanced parameters)
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



Filter data on any column using simple expressions (Galaxy Version 1.1.0)

Filter

43: siMitfvssiLuc.up.annot.txt

Dataset missing? See TIP below.

With following condition

c14>1

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

1

NGS data analysis training – RNAseq

39 shown, 5 deleted

7.48 GB

44: Filter on data 43

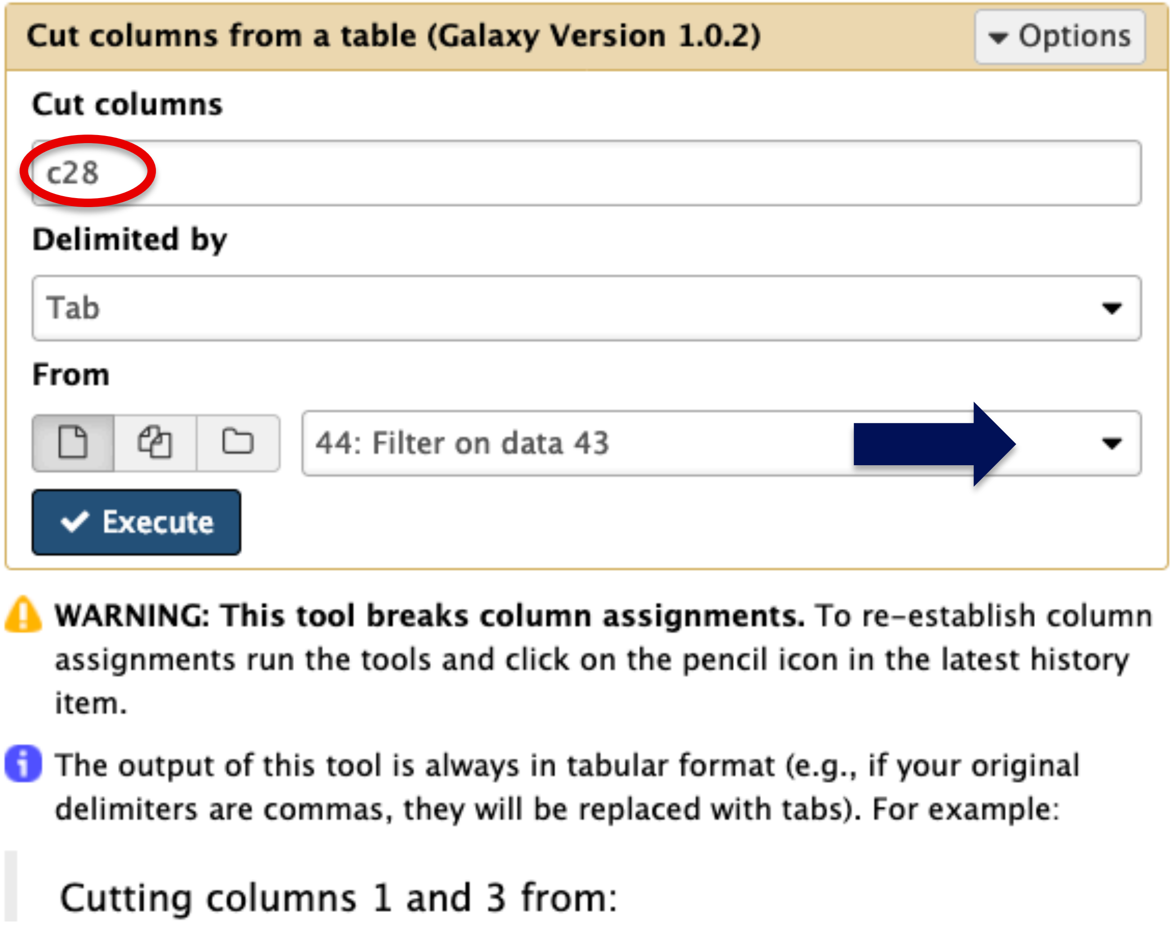
612 lines

format: tabular, database: hg38

Filtering with c14>1,
kept 16.70% of 3664 valid lines (3664 total lines).

2. Create a list of gene names

- Select associated gene names in the previous table



Cut columns from a table (Galaxy Version 1.0.2)

Cut columns

c28

Delimited by

Tab

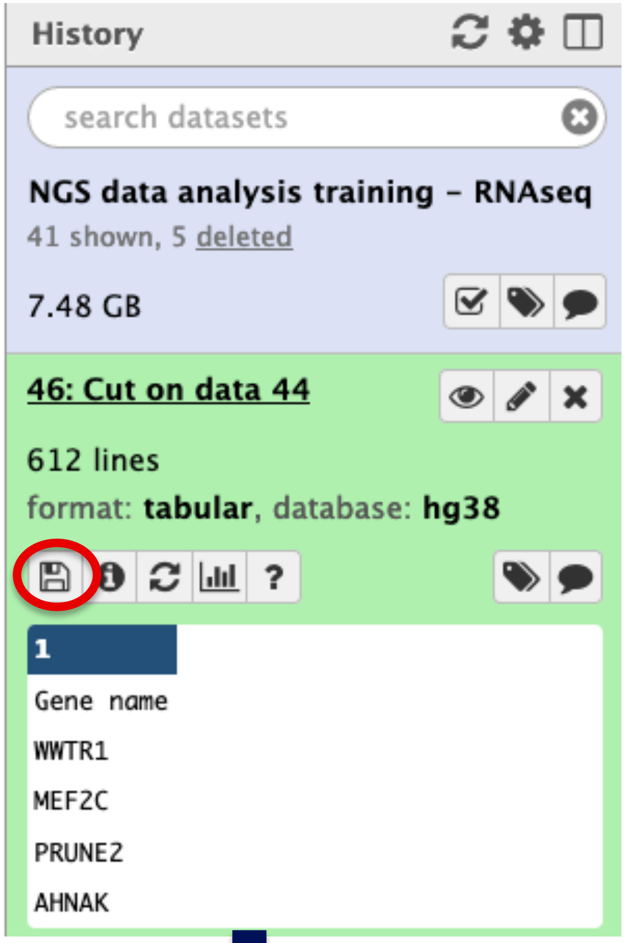
From

44: Filter on data 43

WARNING: This tool breaks column assignments. To re-establish column assignments run the tools and click on the pencil icon in the latest history item.

The output of this tool is always in tabular format (e.g., if your original delimiters are commas, they will be replaced with tabs). For example:

Cutting columns 1 and 3 from:



History

NGS data analysis training - RNAseq

41 shown, 5 deleted

7.48 GB

46: Cut on data 44

612 lines

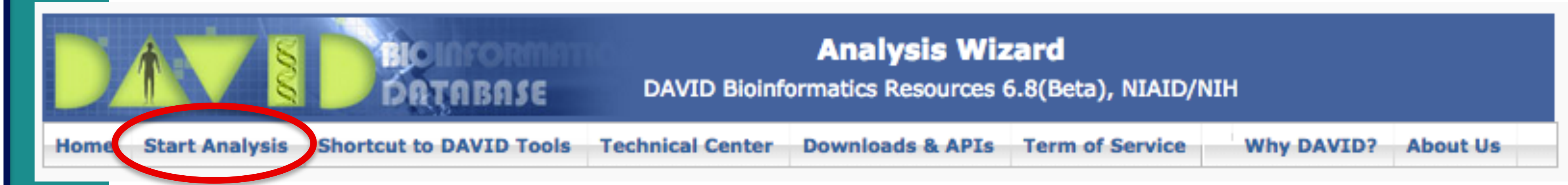
format: **tabular**, database: **hg38**

1
Gene name
WWTR1
MEF2C
PRUNE2
AHNAK

siMitfvssiLuc_upgenes_lfc1_padj005.txt file

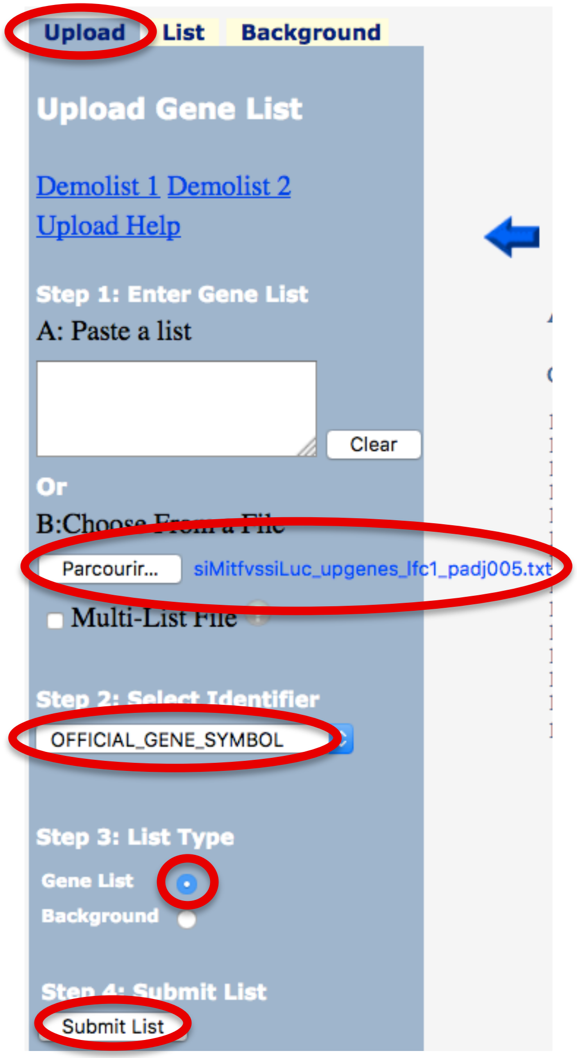
3. Analyse your gene list using DAVID

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



3. Start DAVID analysis

- Enter your gene list



Upload Gene List

Demolist 1 Demolist 2
Upload Help

Step 1: Enter Gene List

A: Paste a list

Or

B:Choose From a File

siMitfvssiLuc_upgenes_lfc1_padj005.txt

□ Multi-List File

Step 2: Select Identifier

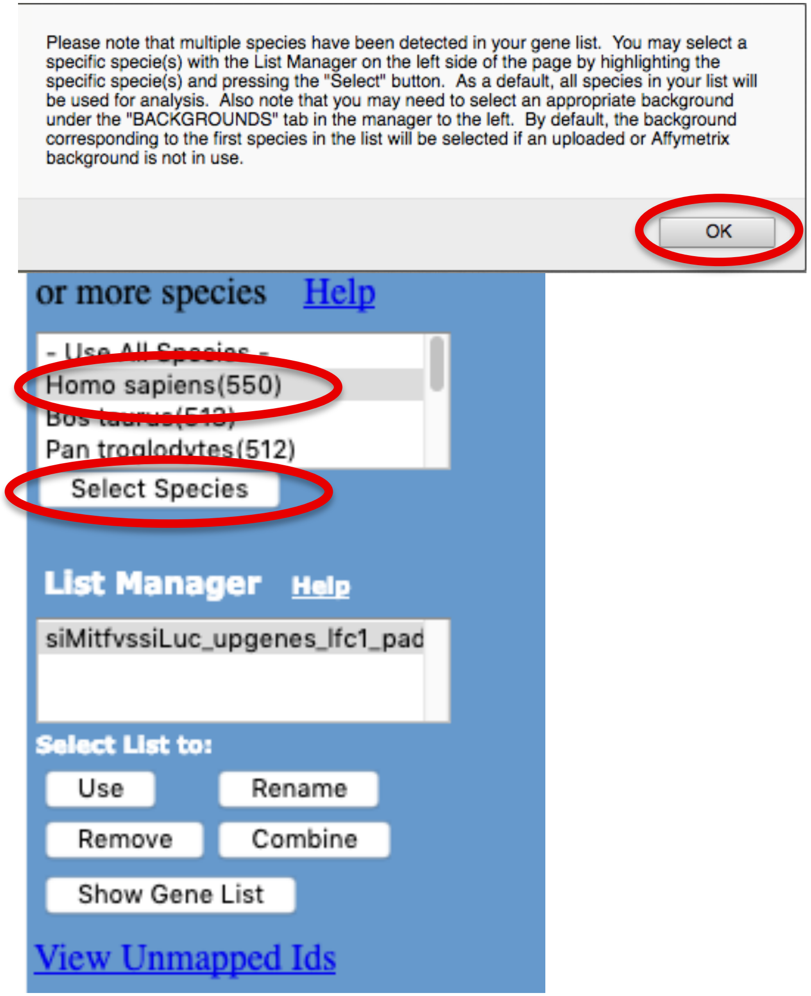
OFFICIAL_GENE_SYMBOL

Step 3: List Type

Gene List ●
Background ○

Step 4: Submit List

- Select species



Please note that multiple species have been detected in your gene list. You may select a specific specie(s) with the List Manager on the left side of the page by highlighting the specific specie(s) and pressing the "Select" button. As a default, all species in your list will be used for analysis. Also note that you may need to select an appropriate background under the "BACKGROUNDS" tab in the manager to the left. By default, the background corresponding to the first species in the list will be selected if an uploaded or Affymetrix background is not in use.

or more species Help

- Use All Species -
Homo sapiens(550)
Bos taurus(513)
Pan troglodytes(512)

List Manager Help

siMitfvssiLuc_upgenes_lfc1_pad

Select List to:

View Unmapped Ids

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
   “PI3K-Akt signalling pathway”.
   Are other genes from your list member of this pathway ?

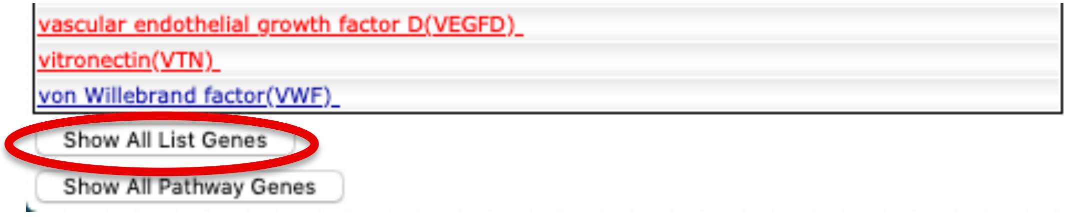
Exercise : functional analysis

4. 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 GalaxEast, and use Heatmapper (http://www.heatmapper.ca/expression/) to perform the heatmap

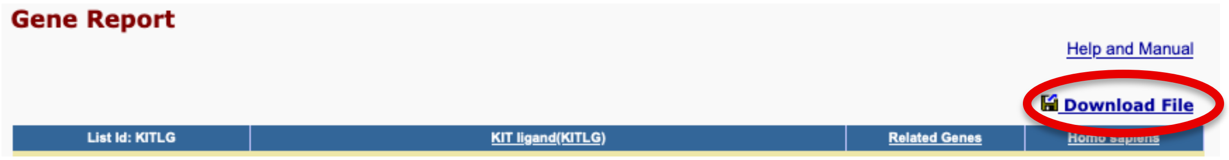
   1. Download list genes in PI3K-Akt signalling pathway from DAVID :

      Click on “Show all list genes” on the bottom of the page representing PI3K-AKT signalling pathway*



vascular endothelial growth factor D(VEGFD)
vitronectin(VTN)
von Willebrand factor(VWF)

then right click on Download File (top right) and save link target on disk



Gene Report

→ pi3k_akt_signalling_genes.txt

* You should be on this page at the end of question 3. Otherwise you will find this page in DAVID Functional Annotation Table by searching « PI3K » and clicking on the corresponding link (PI3K-Akt signalling pathway)

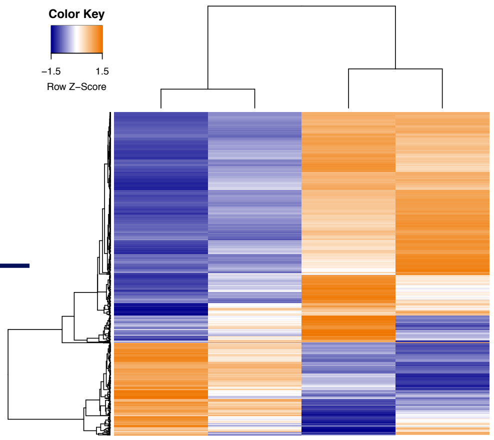
Exercise : functional analysis

We will join the file obtained at step 1 with siMitfvssiLuc.up.annot.txt using the common column (containing gene symbol) → We will thus retain only PI3K-Akt signalling genes from siMitfvssiLuc.up.annot.txt file.

2. In order to keep header during this join, rename the column “ID” from file pi3k_akt_signalling_genes.txt to “Gene name” (same name as the column containing gene symbol in siMitfvssiLuc.up.annot.txt file)
3. Import pi3k_akt_signalling_genes.txt file on GalaxEast
4. On Galaxeast, join siMitfvssiLuc.up.annot.txt with pi3k_akt_signalling_genes.txt on common column (Gene name)
5. On GalaxEast, prepare a file with 5 columns : Gene name and four columns containing normalized read counts in the four samples (use “cut” tool and the results obtained at step 4).
6. Download this file and change file extension to txt
7. Use this file to perform an heatmap representing the variation of expression of these genes in the four RNAseq samples using Heatmapper (http://www.heatmapper.ca/expression/)

   after changing the name of the first column to NAME

Heatmap and clustering



- Heatmap

  Colour-scaled representation of the data

  Data represented :

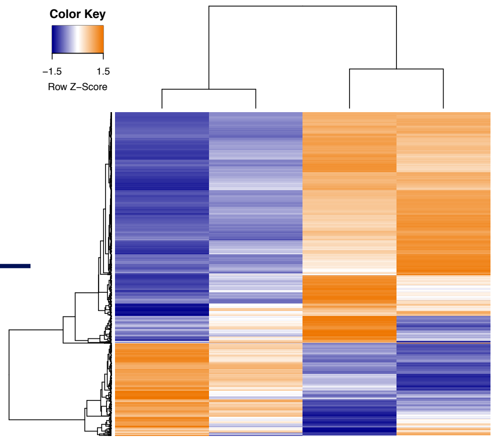
  - Expression
    - Normalized and divided by gene length

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

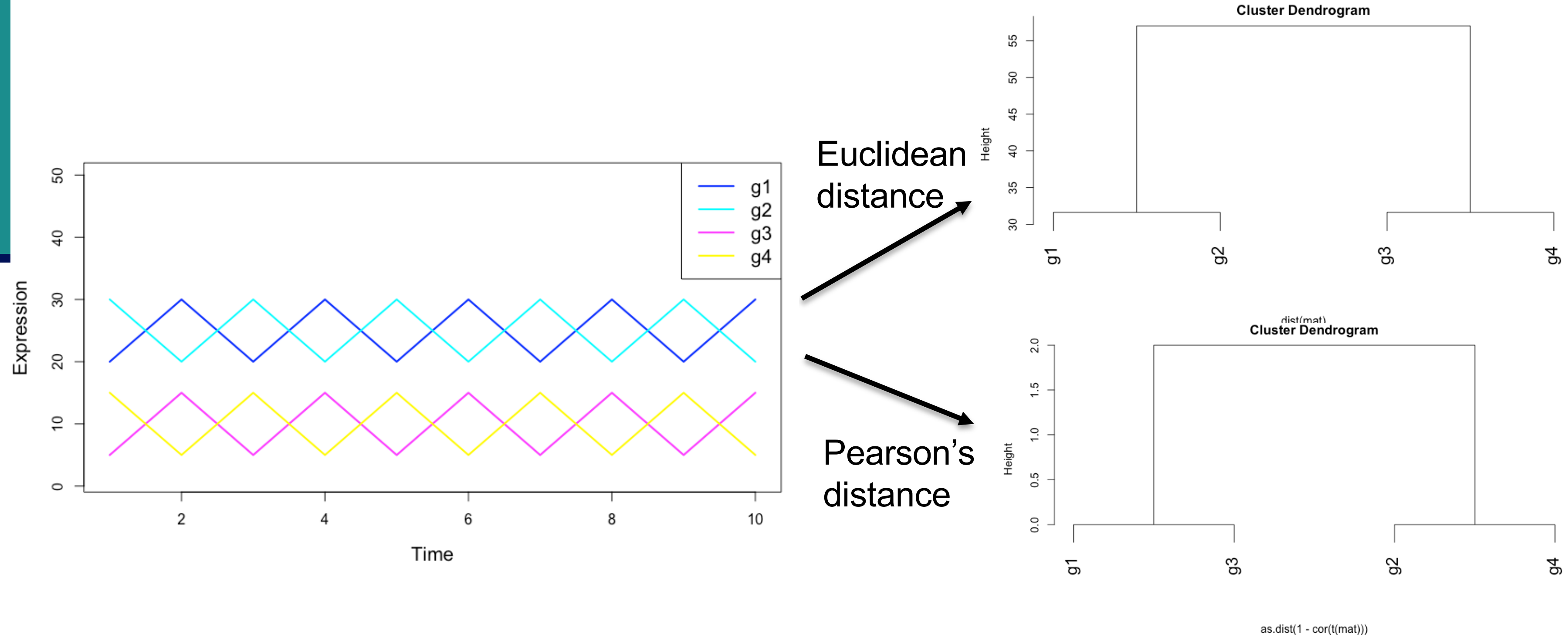
    log2 → over- and under-expression are on symmetric scales
    - Z-score

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

Heatmap and clustering



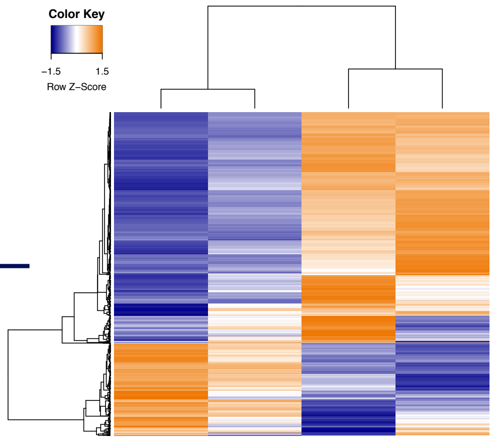
- Hierarchical clustering
  - Distance measure
    - Pairwise distance of all data points
    - Default in a lot of clustering software : Euclidean
    - If you want to group genes with similar expression patterns (i.e. on the shape of the expression profiles) : 1-correlation



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