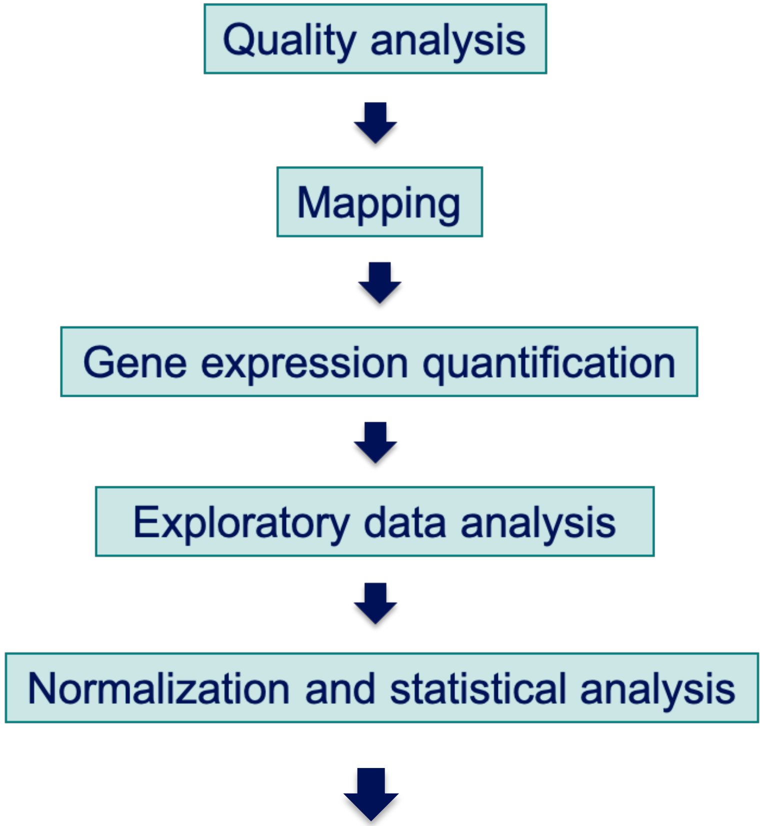
Functional analysis of RNA-seq data

Céline Keime
keime@igbmc.fr

Analysis of RNA-seq data



Quality analysis

Mapping

Gene expression quantification

Exploratory data analysis

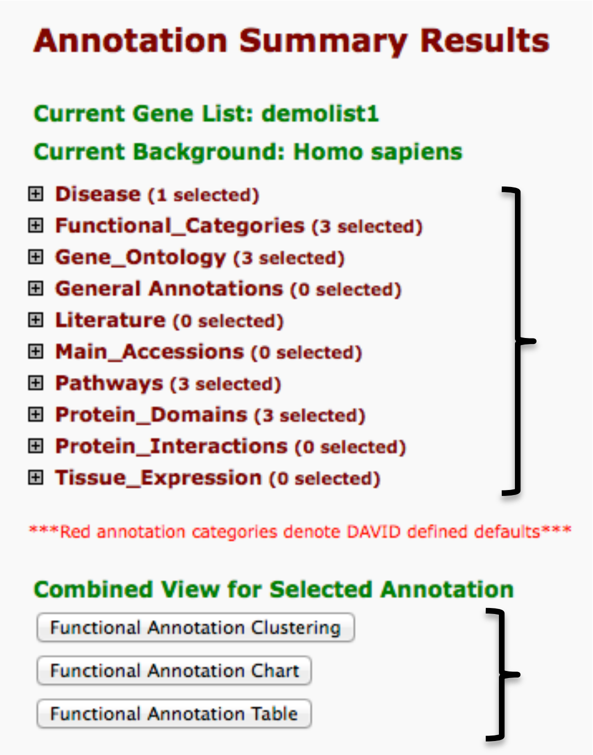
Normalization and statistical analysis

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
  - 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

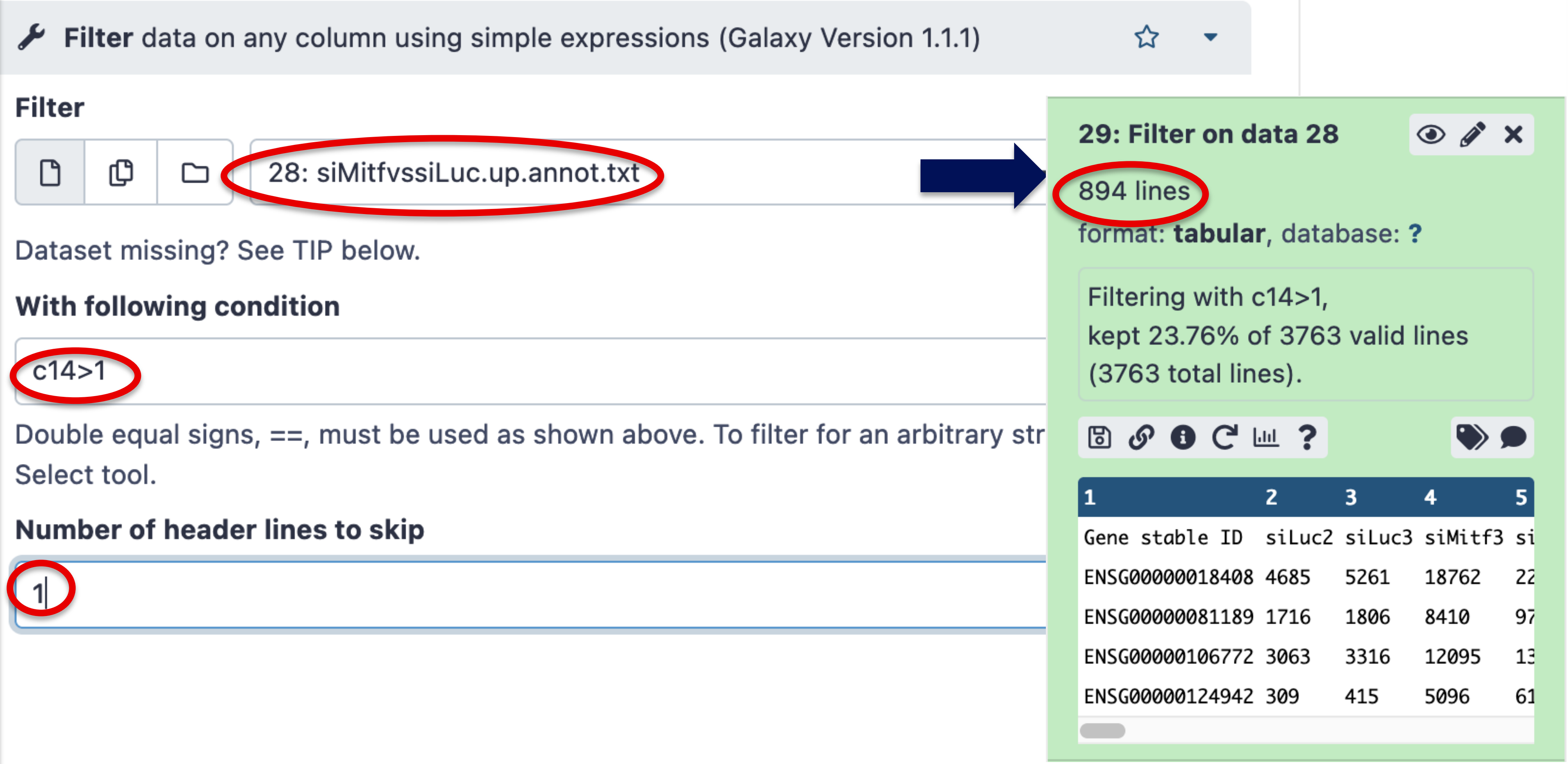
       In your history or dataset 21 in "NGS data analysis training Strasbourg" history
     - Threshold : $\log_2$(Fold-Change) > 1

       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. Select Ensembl gene ID corresponding to these genes using the **Cut** tool on Galaxy, and **save** the result in a file with **.txt** extension.
  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 1.1.1)

Filter

28: 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 str
Select tool.

Number of header lines to skip

1

29: Filter on data 28

894 lines

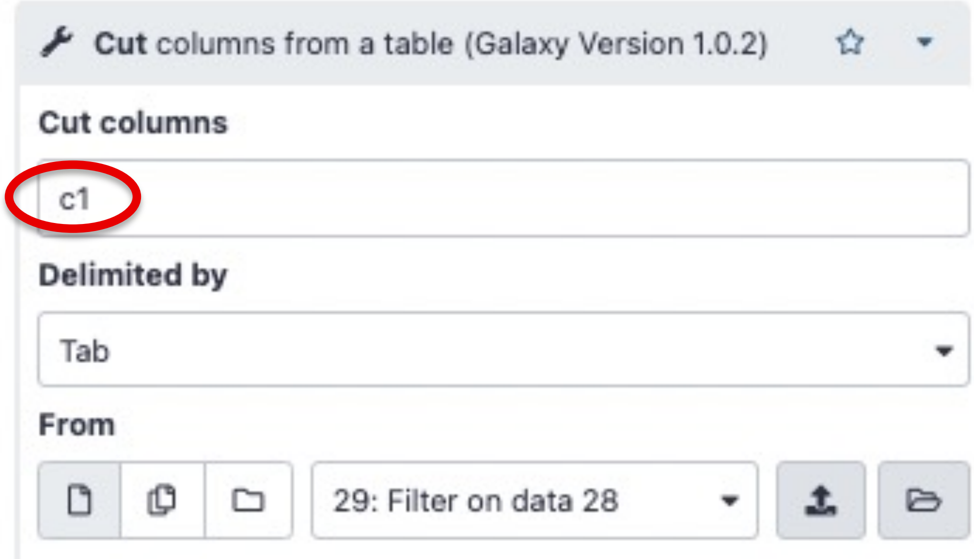
format: **tabular**, database: **?**

Filtering with c14>1,
kept 23.76% of 3763 valid lines
(3763 total lines).

1	2	3	4	5
Gene stable ID	siLuc2	siLuc3	siMitf3	si
ENSG00000018408	4685	5261	18762	22
ENSG00000081189	1716	1806	8410	97
ENSG00000106772	3063	3316	12095	13
ENSG00000124942	309	415	5096	61

2. Create a list of Ensembl gene IDs

- Select Ensembl gene IDs in the previous table



Cut columns from a table (Galaxy Version 1.0.2)

Cut columns

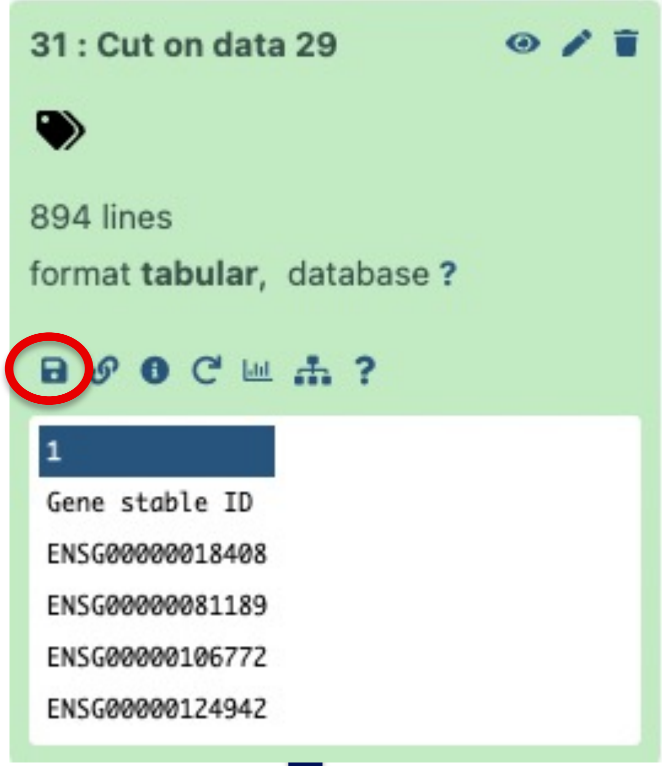
c1

Delimited by

Tab

From

29: Filter on data 28



31 : Cut on data 29

894 lines

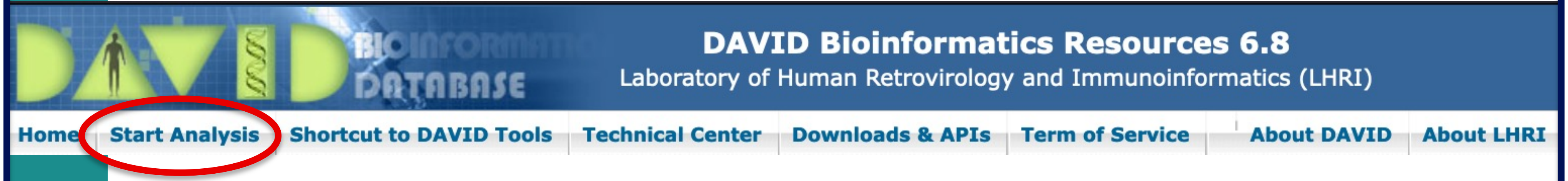
format **tabular**, database **?**

```
1
Gene stable ID
ENSG00000018408
ENSG00000081189
ENSG00000106772
ENSG00000124942
```

siMitfvssiLuc_upgenes_lfc1_padj005**.txt** file

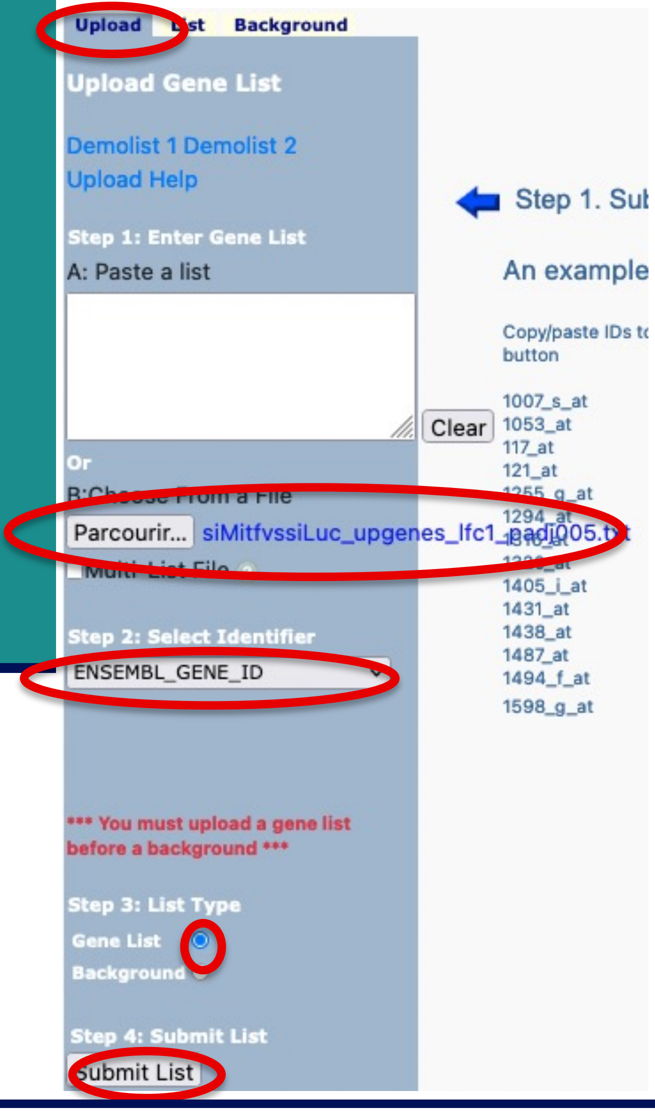
3. Analyse your gene list using DAVID

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



3. Analyze your gene list using DAVID

- Enter your gene list



Upload | List | Background

Upload Gene List

Demolist 1 Demolist 2

Upload Help

Step 1: Enter Gene List

A: Paste a list

Clear

Or

B:Choose From a File

Parcourir... siMitfvssiLuc_upgenes_lfc1_padj005.txt

☐ Multi-List File

Step 2: Select Identifier

ENSEMBL_GENE_ID

*** You must upload a gene list before a background ***

Step 3: List Type

Gene List ●

Background ○

Step 4: Submit List

Submit List

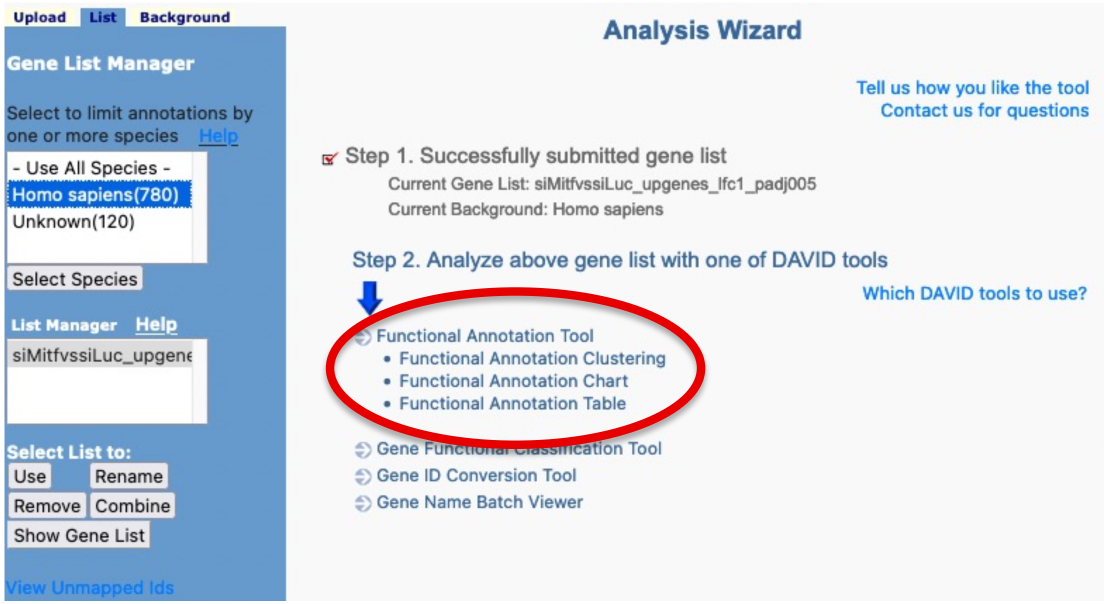
Step 1. Sub

An example

Copy/paste IDs t
button

1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at

- Analyze your list



Upload | List | Background

Gene List Manager

Select to limit annotations by one or more species Help

- Use All Species -
Homo sapiens(780)
Unknown(120)

Select Species

List Manager Help

siMitfvssiLuc_upgene

Select List to:

Use | Rename
Remove | Combine
Show Gene List

View Unmapped Ids

Analysis Wizard

Tell us how you like the tool
Contact us for questions

☑ Step 1. Successfully submitted gene list

Current Gene List: siMitfvssiLuc_upgenes_lfc1_padj005

Current Background: Homo sapiens

Step 2. Analyze above gene list with one of DAVID tools

Which DAVID tools to use?

Functional Annotation Tool
- Functional Annotation Clustering
- Functional Annotation Chart
- Functional Annotation Table

Gene Functional Classification Tool

Gene ID Conversion Tool

Gene Name Batch Viewer

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 ?

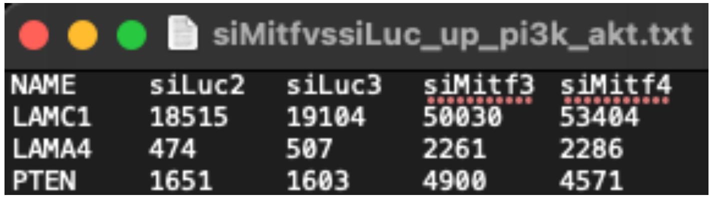
   Which genes are annotated with the most enriched GO biological process term ?

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



siMitfvssiLuc_up_pi3k_akt.txt

```
NAME     siLuc2   siLuc3   siMitf3  siMitf4
LAMC1    18515    19104    50030    53404
LAMA4    474      507      2261     2286
PTEN     1651     1603     4900     4571
```

   - → use Heatmapper (http://www.heatmapper.ca/expression/) to perform the heatmap

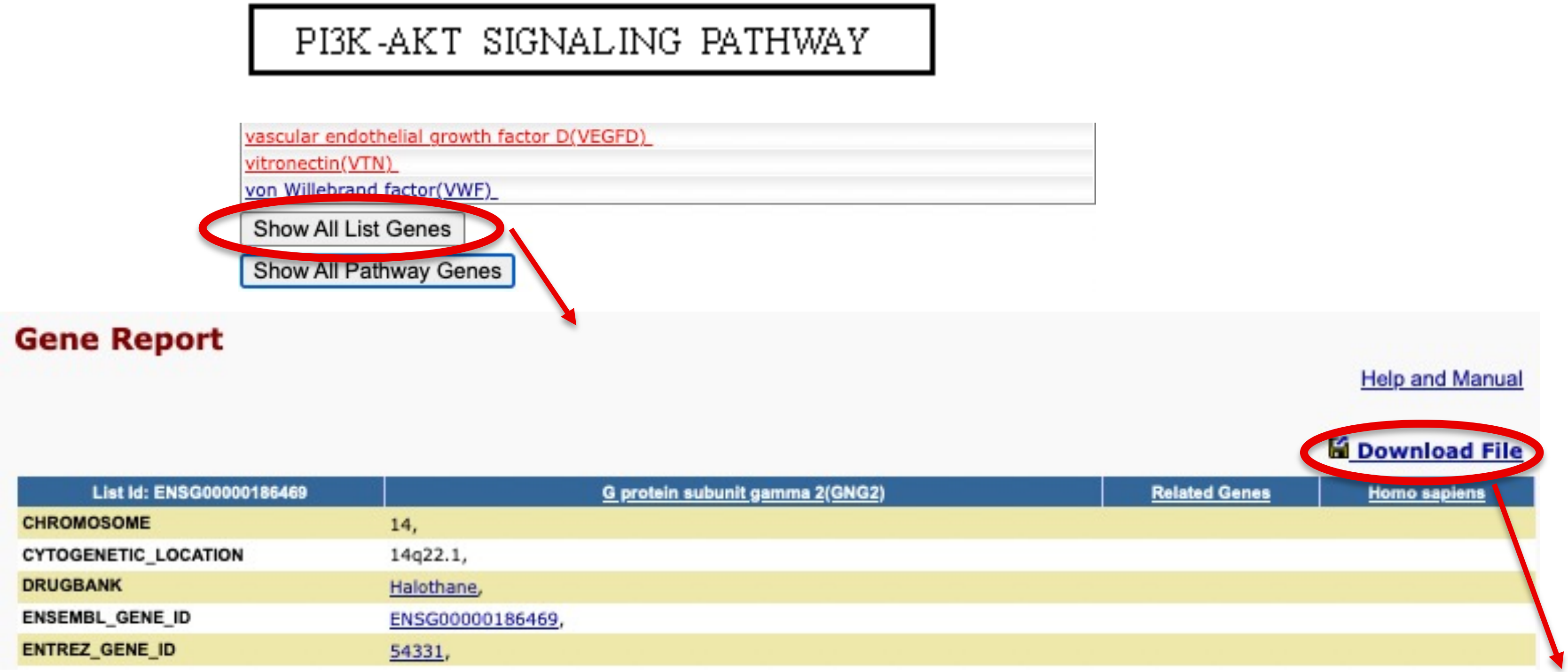
Exercise : functional analysis

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

At the bottom of PI3K-AKT signalling pathway page,

click on “Show All Pathway Genes”

then click on “Download File”



PI3K-AKT SIGNALING PATHWAY

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

Gene Report

List Id: ENSG00000186469	G protein subunit gamma 2(GNG2)	Related Genes	Homo sapiens
CHROMOSOME	14,		
CYTOGENETIC_LOCATION	14q22.1,		
DRUGBANK	Halothane,		
ENSEMBL_GENE_ID	ENSG00000186469,		
ENTREZ_GENE_ID	54331,		

pi3k_akt_signalling_genes.txt

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.

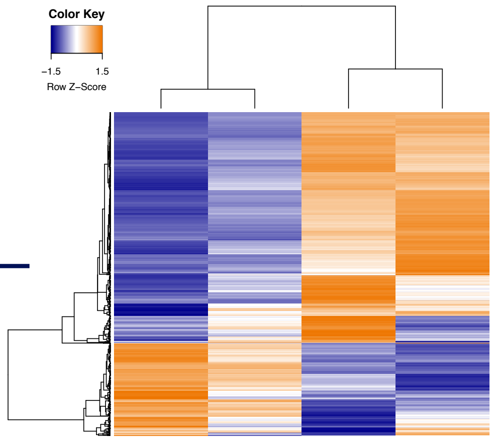
- Import pi3k_akt_signalling_genes.txt file on Galaxy
- On Galaxy, join siMitfvssiLuc.up.annot.txt with pi3k_akt_signalling_genes.txt on their common column (Ensembl gene ID)

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 :

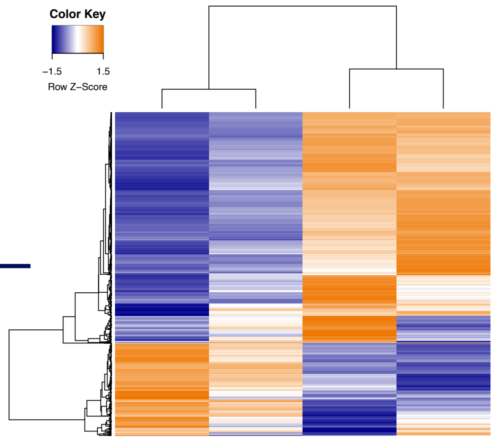
  - Expression
    - Normalized and divided by gene length

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

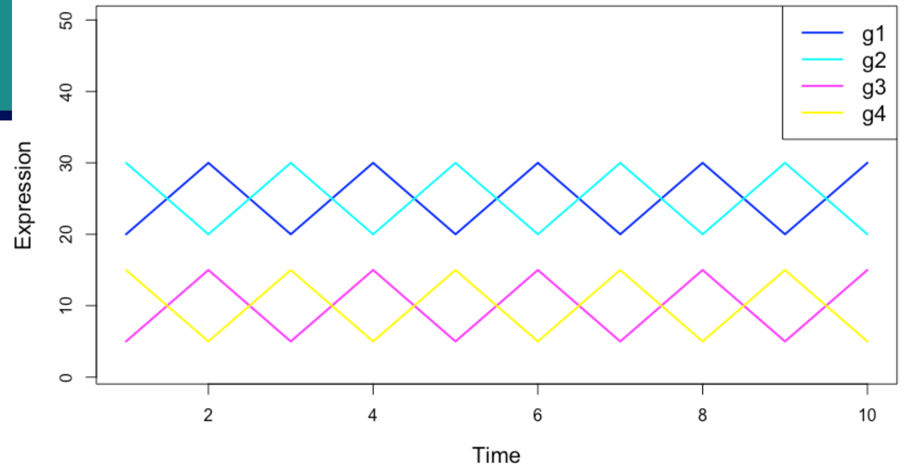
    log2 → over- and under-expression are on symmetric scales
    - Z-score

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

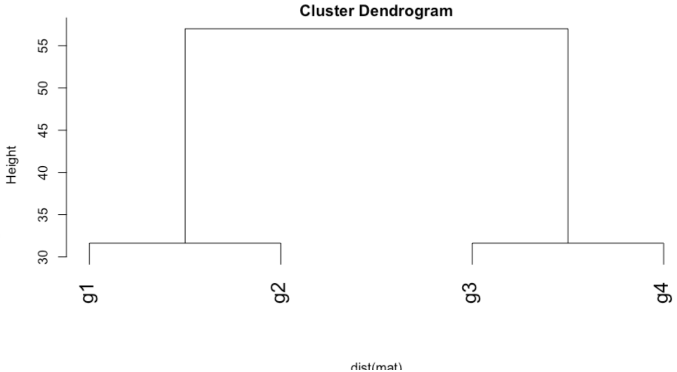
Heatmap and clustering



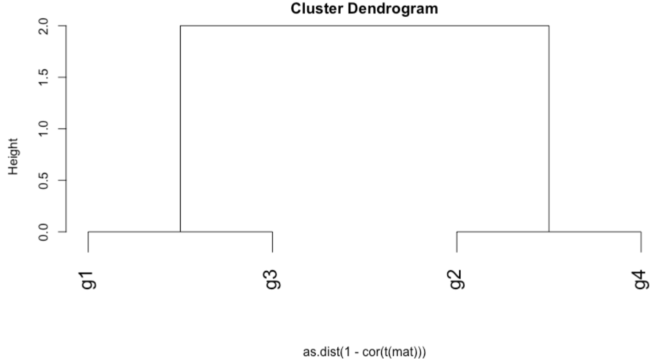
- Hierarchical clustering
  - Distance measure
    - Pairwise distance of all data points
    - Default in a lot of clustering software : Euclidean
    - If you want to group genes with similar expression patterns (i.e. on the shape of the expression profiles) : 1-correlation



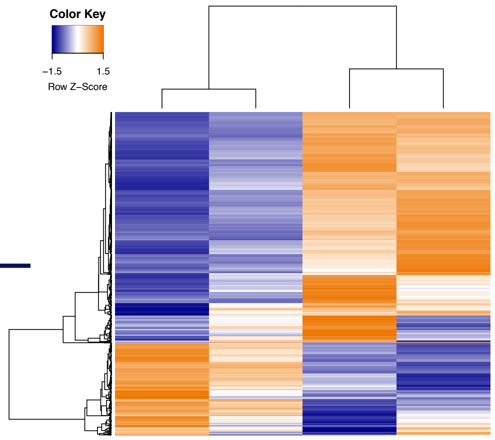
Euclidean distance



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