

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
 - Database 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 $\log_2(\text{Fold-Change}) > 1$
- 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 : $\log_2(\text{Fold-Change}) > 1$
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 $\log_2(\text{Fold-Change}) > 1$

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 string, use `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
ENSG000000081189	1716	1806	8410	97
ENSG000000106772	3063	3316	12095	13
ENSG000000124942	309	415	5096	61

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
c29

Delimited by
Tab

From
29: Filter on data 28

Email notification
 Send an email notification when the job completes.

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

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

■ 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_upgenes_lfc1_padj005.txt

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 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

Annotation Summary Results

Current Gene List:
siMitfvssiLuc_upgenes_lfc1_padj005

Current Background: Homo sapiens

- Disease (2 selected)
- Functional_Annotations (5 selected)
- Gene_Ontology (3 selected)
- General_Annotations (0 selected)
- Interactions (1 selected)
- Literature (0 selected)
- Pathways (3 selected)
- Protein_Domains (4 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

DAVID: Database for Annotation, Visualization, and Integrated Discovery (Laboratory of Human Retrovirology and I...)

https://david.ncicrf.gov/chartReport.jsp?annot=55,88,85,86,91,92,78,27,35,43,90,1,3,52,5 80 %

Aucun traqueur connu par Firefox n'a été détecté sur cette page.

Annotation	Count	Score	P-value
GOTERM_BP_DIRECT	9	1,2	1,9E-5 8,5E-3
KEGG_PATHWAY	17	2,2	2,0E-5 8,5E-3
INTERPRO	14	1,8	2,2E-5 6,3E-3
GOTERM_BP_DIRECT	13	1,7	2,9E-5 7,8E-3
GOTERM_BP_DIRECT	19	2,4	3,0E-5 1,1E-2
INTERPRO	23	2,9	4,3E-5 9,4E-3
GOTERM_BP_DIRECT	23	2,9	4,3E-5 1,5E-2
SMART	11	1,4	4,4E-5 1,2E-2
GOTERM_MF_DIRECT	13	1,7	4,9E-5 8,2E-3
GOTERM_CC_DIRECT	21	2,7	5,0E-5 3,1E-3
UP_KW_DOMAIN	14	1,8	5,3E-5 5,1E-4
KEGG_PATHWAY	30	3,8	5,9E-5 6,3E-3
KEGG_PATHWAY	21	2,7	6,2E-5 6,3E-3
GOTERM_BP_DIRECT	8	1,0	6,8E-5 2,1E-2
UP_SEQ_FEATURE	7	0,9	7,8E-5 2,9E-2
INTERPRO	23	2,9	8,5E-5 1,5E-2
UP_SEQ_FEATURE	19	2,4	9,0E-5 2,5E-2

pi3k-akt

Tout surligner Respecter la casse Respecter les accents et diacritiques

Gene Report

Current Gene List: siMitfvssiLuc_upgenes_lfc1_padj005

Current Background: Homo sapiens

780 DAVID IDs

30 record(s)

ENSEMBL_GENE_ID	GENE NAME	Related Genes	Species
ENSG00000186469	G protein subunit gamma 2(GNG2)	RG	Homo sapiens
ENSG00000049130	KIT ligand(KITLG)	RG	Homo sapiens
ENSG00000181072	cholinergic receptor muscarinic 2(CHRM2)	RG	Homo sapiens

Download File

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

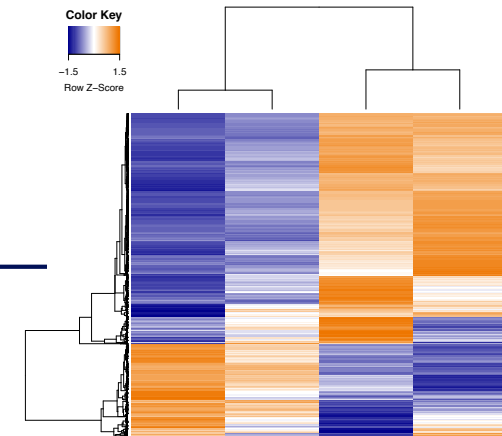
Data represented :

■ Expression

- Normalized and divided by gene length
→ to compare the expression level of several genes

■ Expression variation

- $\log_2(\text{Fold-Change})$
 \log_2 → over- and under-expression are on symmetric scales
- Z-score
→ row z-score = $[\text{Value} - \text{mean}(\text{row})] / \text{standard deviation}(\text{row})$

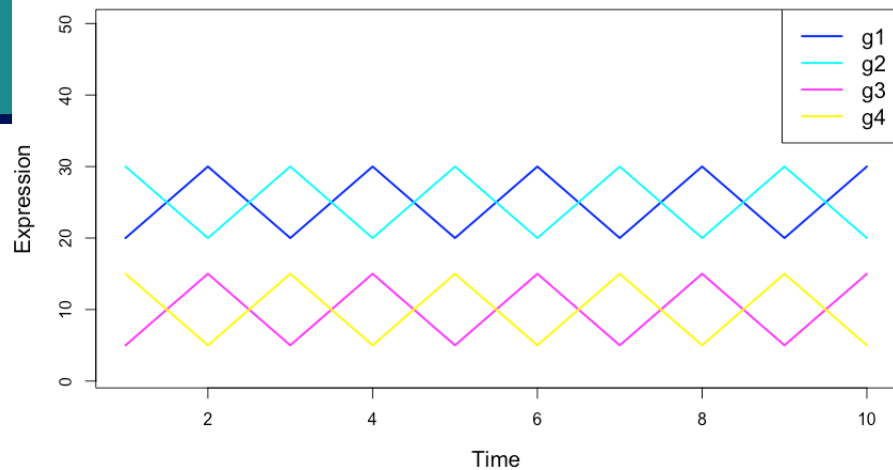
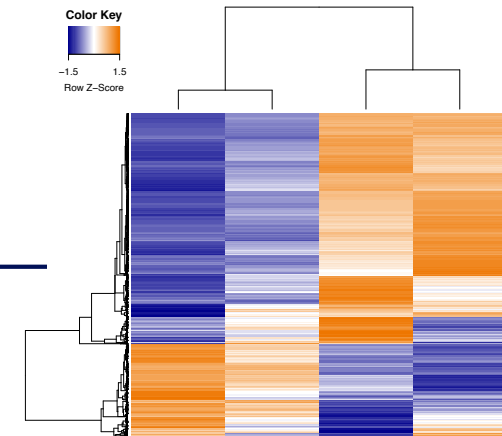


Heatmap and clustering

■ Hierarchical clustering

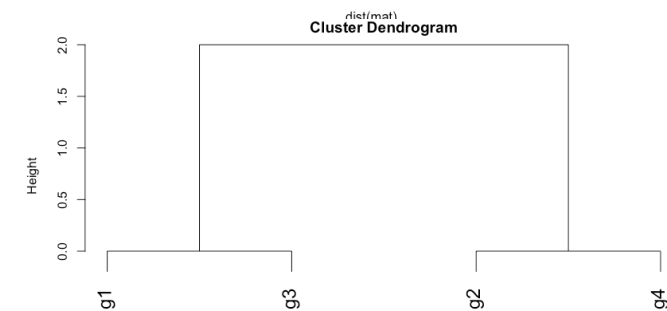
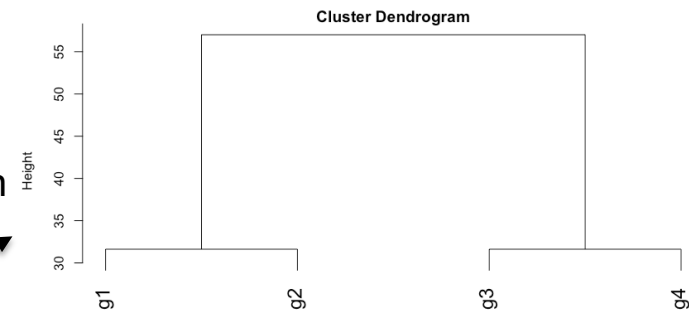
■ Distance measure

- 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

