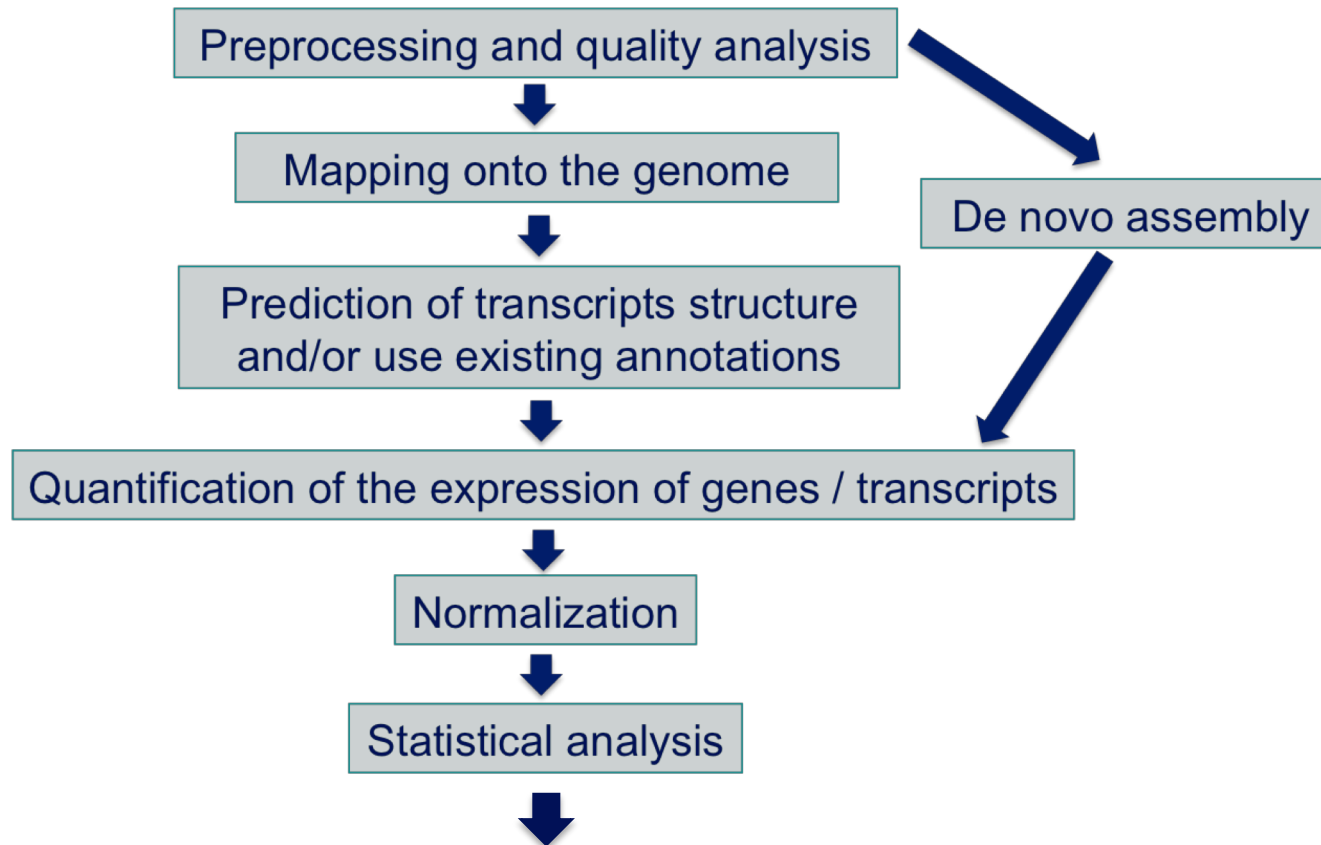




Functional analysis of RNA-seq data

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
 - 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

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

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

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

Execute

History

search datasets

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

The screenshot shows the Galaxy web interface. On the left, the 'Cut columns from a table (Galaxy Version 1.0.2)' tool is configured. The 'Cut columns' field contains 'c28', which is circled in red. The 'Delimited by' field is set to 'Tab'. The 'From' field is set to '44: Filter on data 43', with a blue arrow pointing to the right. Below the tool configuration is a blue 'Execute' button. On the right, the 'History' panel shows a search bar and a list of datasets. The top dataset is 'NGS data analysis training - RNAseq' (41 shown, 5 deleted, 7.48 GB). Below it is the current dataset, '46: Cut on data 44', which is highlighted in green. This dataset has 612 lines and is in tabular format using the hg38 database. A red circle highlights the 'Save' icon in the history panel. Below the history panel, a blue arrow points down to the filename 'siMitfvssiLuc_upgenes_lfc1_padj005.txt file'.

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.

i 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:

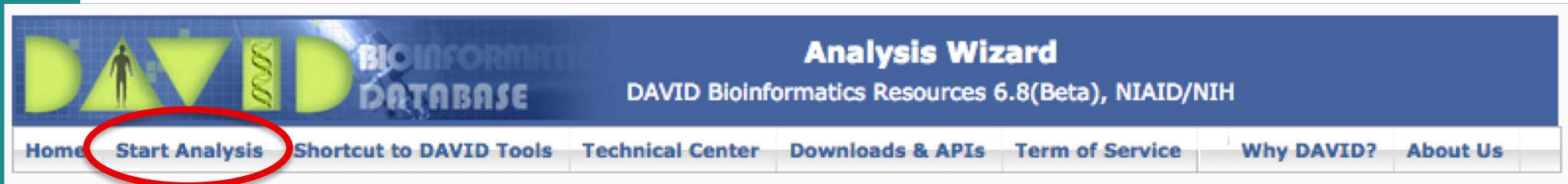
Cutting columns 1 and 3 from:

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 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

OFFICIAL_GENE_SYMBOL

Step 3: List Type

Gene List
Background

Step 4: Submit List

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(510)
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 “Ras signalling pathway”.
Are other genes from your list member of this pathway ?