Analysis of ChIP-seq data (answers to questions)

Stéphanie Le Gras
(slegras@igbmc.fr)

Exercise 1: mapping statistics

- 2.
  - Click on the button ⚙ an select “create new”
  - Click on the history name “Unnamed history”, erase “Unnamed history”, enter “ChIP-seq data analysis” and press enter

- 3.
  - Click on Shared Data (top menu) and select “Data Libraries”
  - Click on “NGS data analysis training ” > “ChIPseq” > “mapping”
  - Select mitf.bam and ctrl.bam datasets (tick boxes beside dataset names)
  - Click on the button to History
  - Select history: ChIP-seq data analysis
  - Click on Import
  - Go back to the main page by clicking on “Analyzed data” (top menu)

Exercise 1: mapping statistics

- 4
  - Search for “flagstat” in the search field (tool panel)
  - Click on the name of the tool
  - Click on to select multiple datasets
  - Select all 2 datasets
  - Click on Execute

Sample name	No. of raw reads	No. of aligned reads
MITF	31,334,257	23,124,393
Ctrl	29,433,042	19,949,607

Exercise 2: duplicate reads estimate

- 1.
  - Search for “markdup” in the search field (tool panel)
  - Click on the name of the tool
  - Click on to select multiple datasets
  - Select the 2 bam files
  - Select validation stringency: Silent
  - Click on Execute
  - Open the datasets “MarkDuplicates on data * : MarkDuplicate metrics”

Sample name	No. of raw reads	No. of aligned reads	No. of duplicate reads
MITF	31,334,257	23,124,393	16,901,318
Ctrl	29,433,042	19,949,607	15,151,227

Exercise 3: Visualization of the data

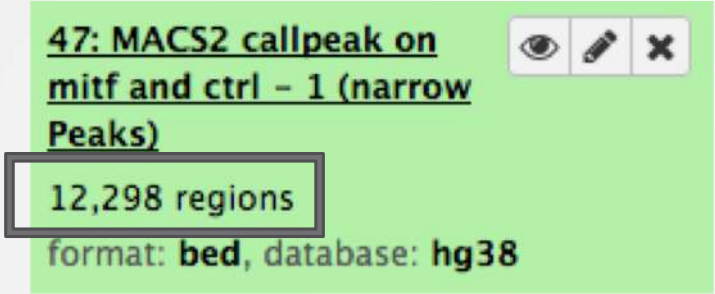
- 1.
  - Idh1 -> No peak
  - Eef2 -> No peak
  - AP1S2 -> Peak,
  - PABPC1l -> No peak
  - Park7 -> No peak
  - Pmel -> Peak
  - Cdk2 -> Peak
  - Actb -> No peak

Exercise 4: peak calling

- 1.
  - Search for “macs2 callpeak” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - ChIP-Seq Treatment File: mitf.bam
    - ChIP-Seq Control File: ctrl.bam
    - Effective genome size: Human
    - Outputs: select Peaks as tabular file, summits, Summary page (html), Plot in PDF
    - Click on 

Exercise 5: peak calling

- 2.
  - There is 12,298 peaks



- 3. Look at the HTML dataset

```
#2 finished!
#2 predicted fragment length is 75 bps
#2 alternative fragment length(s) may be 75 bps
#2.2 Generate R script for model : MACS2_model.r
```

  - The d value estimated by MACS seems a bit small. Let's try to re-run MACS with the expected fragment size : 200

Exercise 5: peak calling

- 4.
  - Click on the name of one of the datasets generated by Macs2.
  - Click on ↻ to display Macs2 form with the same parameters as for the previous run of Macs2
  - In Build Model, select Do not build the shifting model (--nomodel)
  - Enter 200 in the text box "The arbitrary extension size in bp"
  - Click on ✔ Execute
- 5.
  - 7,352 peaks are now found

  - NOTE: the graphs (showing the d values estimate) are no longer generated

Exercise 6: peak annotation

- 1.
  - Search for “homer annot” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Homer peaks OR BED format: MITF peaks - narrow peaks dataset (2nd run of Macs2)
    - Genome version: hg38
  - Click on ✔ Execute
- 2.
  - The Homer annotatePeaks tool generates two datasets: a log file and a tabular file containing annotated peaks.
  - Click on the ✎ of the dataset which contain annotated peaks.
  - Click on the Datatype tab
  - Select **tabular** in the drop down list “New Type:”
  - Click on Save

Exercise 6: peak annotation

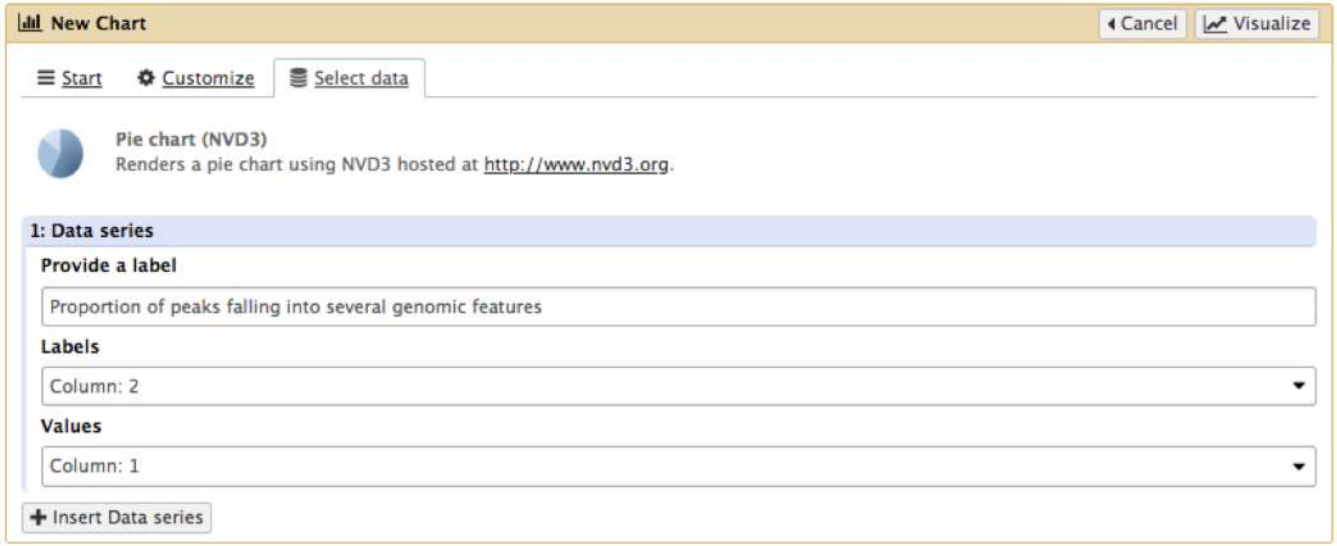
- 3.
  - Search for “histogra” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Dataset: tabular file which contains annotated peaks
    - Numerical column for x axis: column: 10
    - Plot title: Frequency of peaks relative to TSS
    - Label for x axis: Distance to TSS
  - Click on ✔ Execute
- 4.a.
  - Search for “Cut” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Cut columns: c8
    - Delimited by: Tab
    - From: tabular file which contains annotated peaks
  - Click on ✔ Execute

Exercise 6: peak annotation

- 4.b.
  - Search for “Remove” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Remove first: 1
    - From: resulting dataset after 4.b
  - Click on ✔ Execute
- 4.c.
  - Search for “Count” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - from dataset: resulting dataset after 4.c
    - Count occurrences of values in column(s): column: 1
    - Delimited by: Whitespaces
    - How should the results be sorted?: With the most common values first
  - Click on ✔ Execute

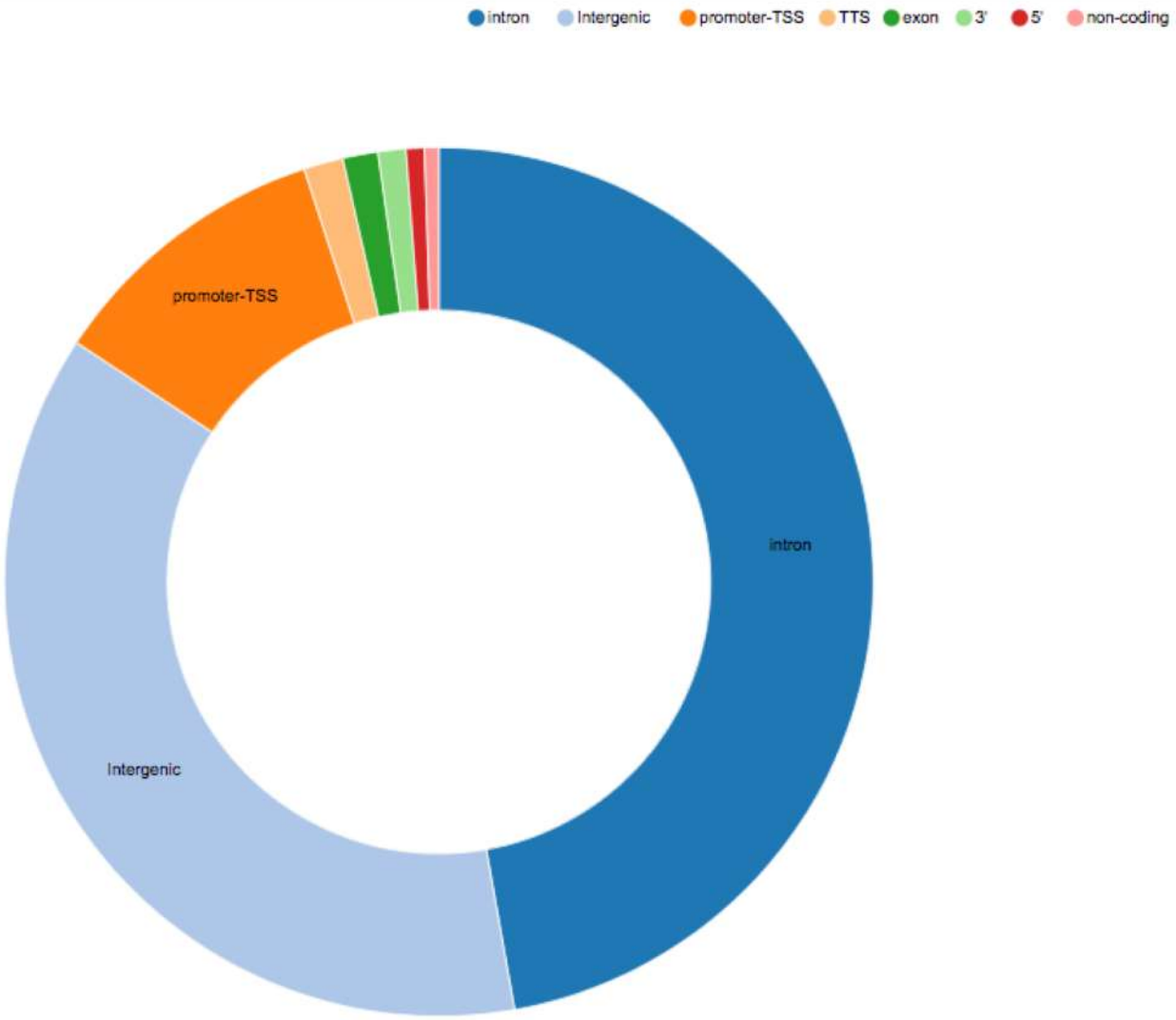
Exercise 6: peak annotation

- 4.d.
  - Expand the box of the dataset generated in 4.d, click on  and select Charts
  - Double click on Pie charts
  - Click on editor (top right)
  - Go to the Select data tab:
    - Provide a label: Proportion of peaks falling into several genomic features.
    - Labels: Column: 2
    - Values: Column: 1



  - Click on Visualize

Exercise 6: peak annotation



0:Proportion of peaks falling into several genomic features.

Exercise 7: *de novo* motif discovery

- 1.a
  - Search for “Sort” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Sort Dataset: dataset with peak summits
    - on column: Column: 5
    - with flavor: Numerical sort
    - everything in: Descending order
  - Click on ✔ Execute
- 1.b
  - Search for “select first” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Select first: 800
    - From: dataset generated in 1.a
  - Click on ✔ Execute

Exercise 7: *de novo* motif discovery

- 2.a
  - Import the file which contains chromosome lengths
  - Click on Shared Data (top menu) and select “Data Libraries”
  - Click on “Chromosome length”
  - Select the dataset named hg38.len (tick boxes beside dataset names)
  - Click on the button “To history”
  - Select history: ChIP-seq data analysis
  - Click on “Import”
  - Go back to the main page by clicking on “Analyzed data” (top menu)
- Run slopBed
  - BED/VCF/GFF file: MACS14_in_Galaxy_summits.bed
  - Genome file: hg38.len
  - Choose what you want to do: Increase the BED/VCF/GFF entry by the same number of base pairs in each direction. (default)
  - Number of base pairs: 50
  - Click on 

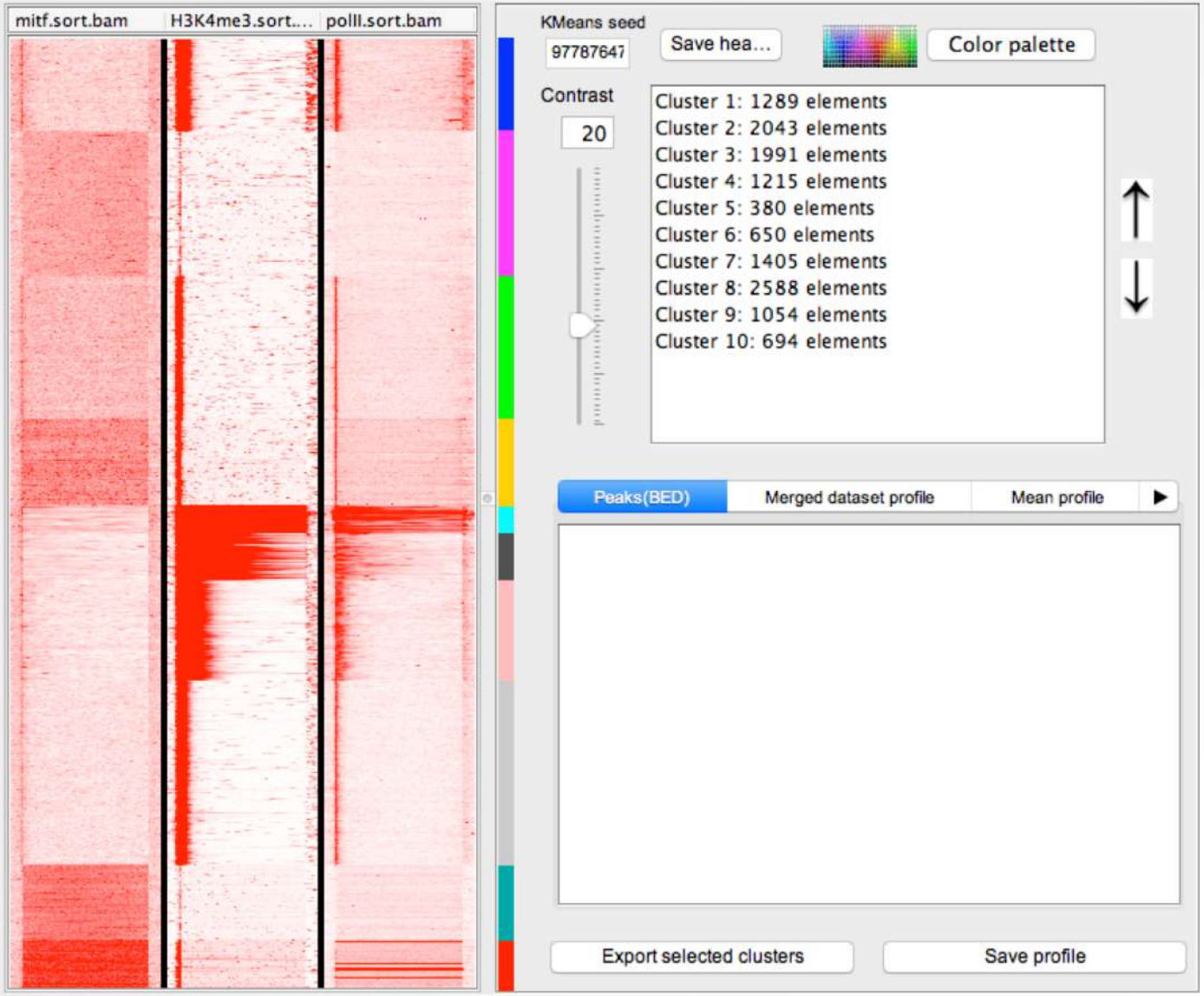
Exercise 7: *de novo* motif discovery

- 3.
  - Search for “extract” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Fetch sequences for intervals in: the dataset generated in 2.c
    - Interpret features when possible: No
    - Click on ✔ Execute
- 4.
  - Expand the box of the dataset generated in 3 and click on to download the file
- 5.
  - Go to MEME-chIP website and run the tool with the fasta file you’ve just downloaded and with default parameters.

Exercise 8: Clustering

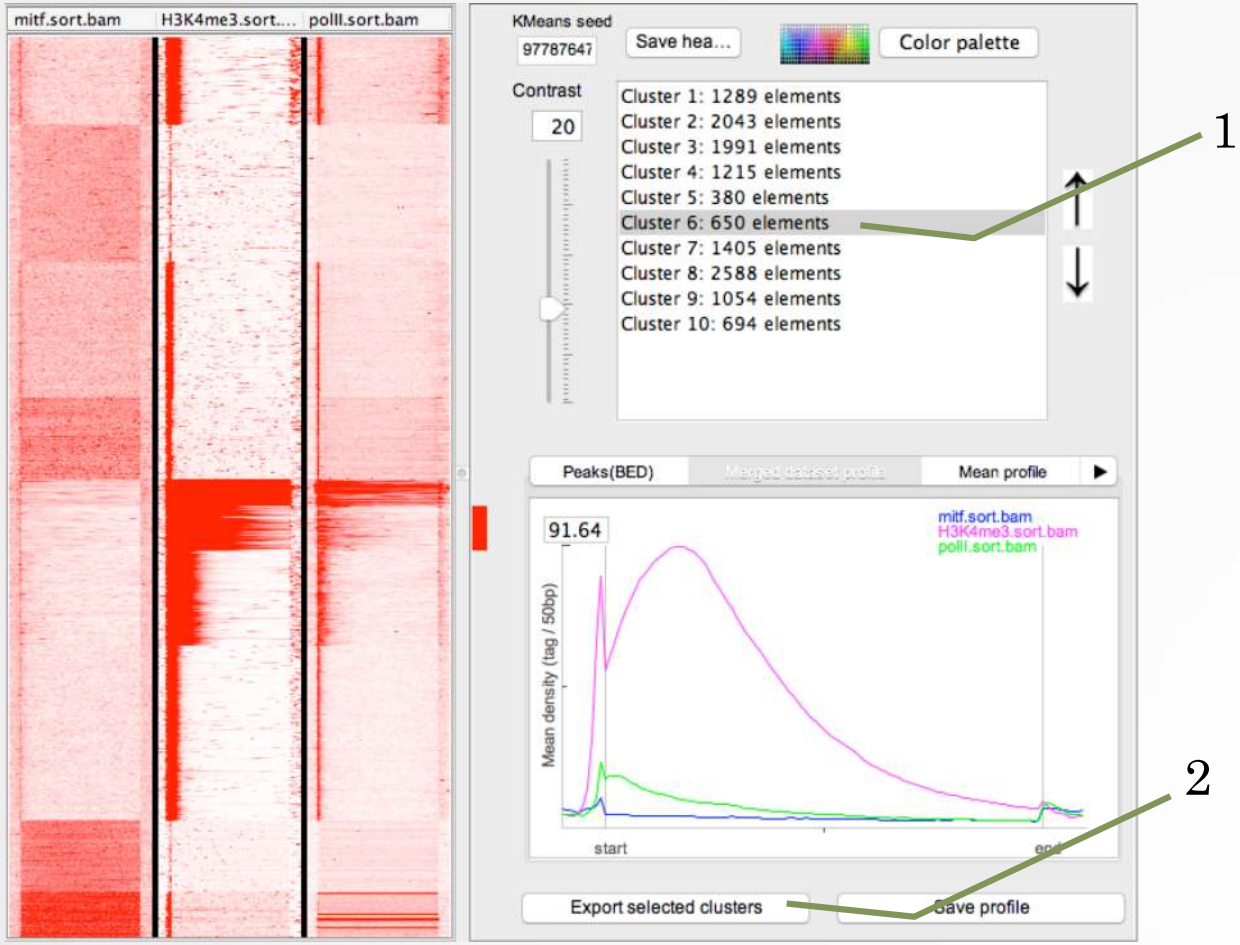
- 1.
  - Select clusters 1, 5, 7, 8 and click on Export Selected clusters
  - Import the file previously exported as reference coordinates. Click on browse, go to the directory which contains the file and click on open.
  - Click on Extract data
  - Click on Clustering

Exercise 8: Clustering



Exercise 8: Clustering

- 2.
  - Click on cluster 6 (1)
  - Click on Export selected clusters (2)



Exercise 8: Clustering

- Go to DAVID website https://david.ncifcrf.gov/
- Click on Function Annotation (left menu)
- Fill in the form:
  - Copy and paste Ensembl Gene IDs from the Cluster6.xls file in the Paste a list text field
  - Select Identifier (drop down list): ENSEMBL_GENE_ID
  - List Type: Gene List
  - Submit List
- Select: Continue to Submit IDs That DAVID Could Map
- Select to limit annotations by one or more species (left panel)
  - Select Homo sapiens (410)
  - Click on Select Species
- Click on Functional Annotation Clustering
- Keep all default
- Click on Functional Annotation Clustering

Exercise 8: Clustering

https://david.ncifcrf.gov/term2term.jsp?annot=59,12,87,88,30,38,46,3,5,55,53,70,79&currentList=0

DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
410 DAVID IDs

Options Classification Stringency Medium

Rerun using options Create Sublist

43 Cluster(s)

Download File

Annotation Cluster 1	Enrichment Score: 12.63			Count	P_Value	Benjamini
UP_KEYWORDS	Ribosomal protein	RT		29	1.1E-19	2.6E-17
KEGG_PATHWAY	Ribosome	RT		25	6.2E-18	8.6E-16
UP_KEYWORDS	Ribonucleoprotein	RT		33	7.0E-18	8.4E-16
GOTERM_BP_DIRECT	SRP-dependent cotranslational protein targeting to membrane	RT		21	3.9E-17	5.2E-14
GOTERM_MF_DIRECT	structural constituent of ribosome	RT		29	6.3E-17	3.8E-14
GOTERM_BP_DIRECT	translation	RT		30	1.2E-16	7.5E-14
GOTERM_BP_DIRECT	viral transcription	RT		21	1.5E-15	7.0E-13
GOTERM_BP_DIRECT	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	RT		21	5.1E-15	1.7E-12
GOTERM_BP_DIRECT	translational initiation	RT		22	7.2E-15	1.9E-12
GOTERM_BP_DIRECT	rRNA processing	RT		26	1.2E-14	2.7E-12
GOTERM_CC_DIRECT	ribosome	RT		22	2.2E-13	5.7E-11
GOTERM_MF_DIRECT	poly(A) RNA binding	RT		54	5.4E-12	9.3E-10
GOTERM_CC_DIRECT	cytosolic small ribosomal subunit	RT		11	3.3E-9	4.3E-7
GOTERM_CC_DIRECT	cytosolic large ribosomal subunit	RT		11	1.1E-7	4.9E-6
GOTERM_CC_DIRECT	focal adhesion	RT		16	1.5E-3	2.8E-2
GOTERM_MF_DIRECT	RNA binding	RT		14	1.3E-1	8.3E-1