



# *Genomics analysis at FMI*

## *Motivation*



## *Summary*

- local galaxy
- FMI specific tools (increasing)
- reduce workload of standard analyses
- R objects can be used seamlessly between galaxy and R

# *Motivation*







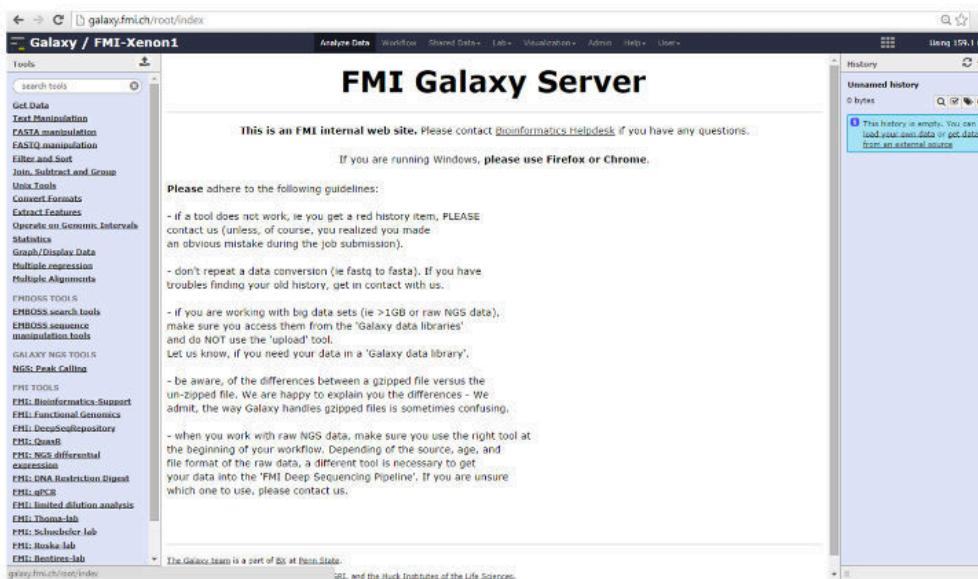
dreamstime.com



dreamstime.com

# Infrastructure and setup

- *internal FMI server*
- *224 users*
- *~ 100 FMI specific tools*
- *~ 400 jobs per month*



- ~ 400 jobs per month

The screenshot shows the FMI Galaxy Server homepage. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Lab, Visualization, Admin, Help, and User. A search bar at the top left contains the URL "galaxy.fmi.ch/root/index". On the left, a sidebar lists various tool categories: Get Data, Text Manipulation, FASTA manipulation, FASTQ manipulation, Filter and Sort, Join, Subtract and Group, Unix Tools, Convert Formats, Extract Features, Operate on Genomic Intervals, Statistics, Graph/Display Data, Multiple regression, and Multiple Alignments. Below these are sections for EMBOSSTOOLS, GALAXY NGS TOOLS, and FMI TOOLS, each listing specific tool names. The main content area features a large header "FMI Galaxy Server" and a message: "This is an FMI internal web site. Please contact [Bioinformatics Helpdesk](#) if you have any questions." It also advises users running Windows to "please use Firefox or Chrome". A section titled "Please adhere to the following guidelines:" provides several tips for using the server. At the bottom, it states "The Galaxy team is a part of BX at Penn State." The right side of the screen shows a "History" panel titled "Unnamed history" which is currently empty, with a note: "This history is empty. You can load your own data or get data from an external source".

Galaxy / FMI-Xenon1

Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Tools

search tools

Get Data

Text Manipulation

FASTA manipulation

FASTQ manipulation

Filter and Sort

Join, Subtract and Group

Unix Tools

Convert Formats

Extract Features

Operate on Genomic Intervals

Statistics

Graph/Display Data

Multiple regression

Multiple Alignments

EMBOSS TOOLS

EMBOSS search tools

EMBOSS sequence manipulation tools

GALAXY NGS TOOLS

NGS: Peak Calling

FMI TOOLS

FMI: Bioinformatics-Support

FMI: Functional Genomics

FMI: DeepSeqRepository

FMI: QuasR

FMI: NGS differential expression

FMI: DNA Restriction Digest

FMI: qPCR

FMI: limited dilution analysis

FMI: Thoma-lab

FMI: Schuebeler-lab

FMI: Roska-lab

FMI: Bentires-lab

History

Unnamed history

0 bytes

This history is empty. You can load your own data or get data from an external source

FMI Galaxy Server

This is an FMI internal web site. Please contact [Bioinformatics Helpdesk](#) if you have any questions.

If you are running Windows, **please use Firefox or Chrome**.

Please adhere to the following guidelines:

- if a tool does not work, ie you get a red history item, PLEASE contact us (unless, of course, you realized you made an obvious mistake during the job submission).
- don't repeat a data conversion (ie fastq to fasta). If you have troubles finding your old history, get in contact with us.
- if you are working with big data sets (ie >1GB or raw NGS data), make sure you access them from the 'Galaxy data libraries' and do NOT use the 'upload' tool.  
Let us know, if you need your data in a 'Galaxy data library'.
- be aware, of the differences between a gzipped file versus the un-zipped file. We are happy to explain you the differences - We admit, the way Galaxy handles gzipped files is sometimes confusing.
- when you work with raw NGS data, make sure you use the right tool at the beginning of your workflow. Depending of the source, age, and file format of the raw data, a different tool is necessary to get your data into the 'FMI Deep Sequencing Pipeline'. If you are unsure which one to use, please contact us.

The Galaxy team is a part of BX at Penn State.

RI, and the Huck Institutes of the Life Sciences.

# Infrastructure and setup

- *internal FMI server*
- *224 users*
- *~ 100 FMI specific tools*
- *~ 400 jobs per month*

The screenshot shows the FMI Galaxy Server interface. At the top, there's a navigation bar with links for Analyze Data, Workflow, Shared Data, Lab, Visualization, Admin, Help, and User. Below the navigation bar, the main content area has a title "FMI Galaxy Server". A message says, "This is an FMI internal web site. Please contact [Bioinformatics Helpdesk](#) if you have any questions. If you are running Windows, **please use Firefox or Chrome**." It also includes guidelines: "Please adhere to the following guidelines:" followed by two bullet points about tool usage and history items. On the left, a sidebar lists various tool categories: Tools, Get Data, Text Manipulation, FASTA manipulation, FASTQ manipulation, Filter and Sort, Join, Subtract and Group, Unix Tools, Convert Formats, Extract Features, Operate on Genomic Intervals, Statistics, Graph/Display Data, Multiple regression, and Multiple Alignments. On the right, a "History" panel titled "Unnamed history" shows a message: "This history is empty. You can load your own data or get data from an external source".

# Storage

*big data files accessible*

- *directly within in tools*
- *linked as data libraries*

Affymetrix array  
prepository

~ 20000 .CEL files



NFS mount

galaxy  
server

NGS repository

~ 7500 fastq files

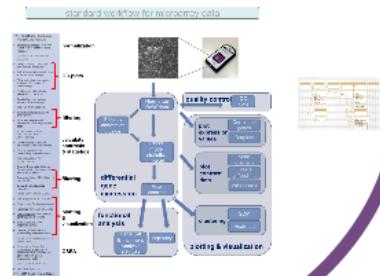


NFS mount

# *Standard analyses*

## *Microarray analysis with LIMMA*

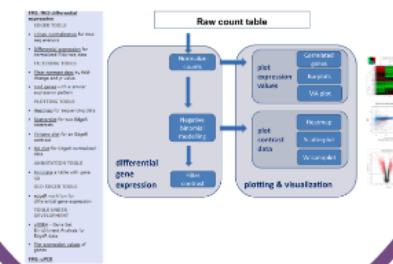
~ 850 Affymetrix array analyses



## *RNA-seq analysis with edgeR*

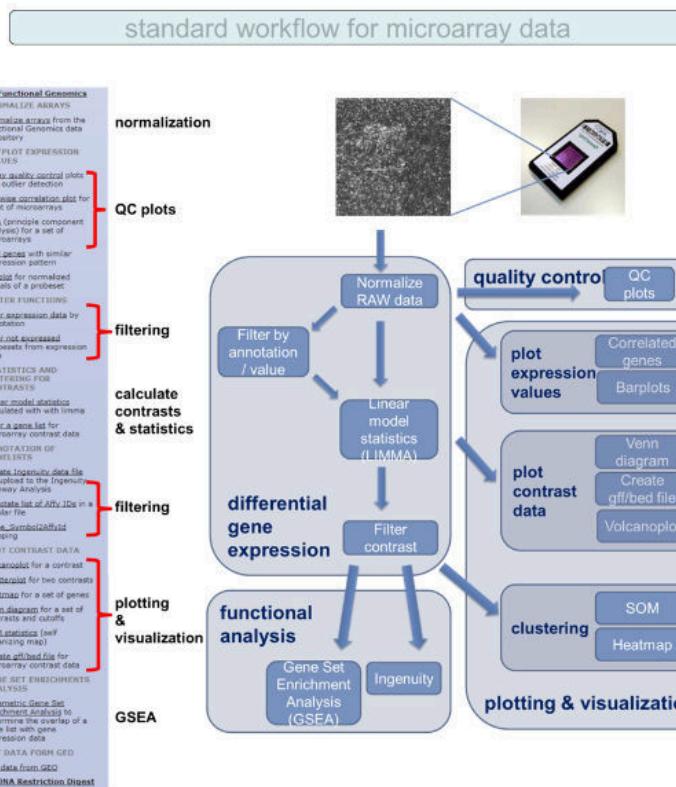
~ 150 NGS differential gene expression analyses

standard workflow for RNA-seq count data



# Microarray analysis with LIMMA

~ 850 Affymetrix array analyses



# standard workflow for microarray data

**FMI: Functional Genomics**

**NORMALIZE ARRAYS**

- Normalize arrays from the Functional Genomics data repository

**QC/PLOT EXPRESSION VALUES**

- Array quality control plots and outlier detection
- Pairwise correlation plot for a set of microarrays
- PCA (principle component analysis) for a set of microarrays
- Find genes with similar expression pattern
- Barplot for normalized signals of a probeset

**FILTER FUNCTIONS**

- Filter expression data by annotation
- Filter not expressed probesets from expression data

**STATISTICS AND FILTERING FOR CONTRASTS**

- Linear model statistics calculated with limma
- Filter a gene list for microarray contrast data

**ANNOTATION OF GENELISTS**

- Create Ingenuity data file for upload to the Ingenuity Pathway Analysis
- Annotate list of Affy IDs in a tabular file
- Gene\_Symbol2AffyId mapping

**PLOT CONTRAST DATA**

- Volcanoplot for a contrast
- Scatterplot for two contrasts
- Heatmap for a set of genes
- Venn diagram for a set of contrasts and cutoffs
- SOM statistics (self organizing map)
- Create gff/bed file for microarray contrast data

**GENE SET ENRICHMENTS ANALYSIS**

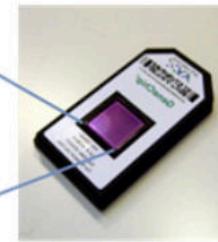
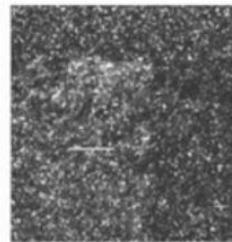
- Parametric Gene Set Enrichment Analysis to determine the overlap of a gene list with gene expression data

**GET DATA FORM GEO**

- Get data from GEO

**FMI: DNA Restriction Digest**

## normalization



## QC plots

## filtering calculate contrasts & statistics

## filtering

## plotting & visualization

## GSEA

## Normalize RAW data

## Filter by annotation / value

## Linear model statistics (LIMMA)

## differential gene expression

## Filter contrast

## functional analysis

## Gene Set Enrichment Analysis (GSEA)

## quality control

## QC plots

## plot expression values

## Correlated genes Barplots

## plot contrast data

## Venn diagram Create gff/bed file Volcanoplot

## clustering

## SOM Heatmap

## plotting & visualization





# RNA-seq analysis with edgeR

~ 150 NGS differential gene expression analyses

standard workflow for RNA-seq count data

**FMI: NGS differential expression**  
EDGER TOOLS

- EdgeR normalization for RNA-seq analysis
- Differential expression for normalized RNA-seq data

FILTERING TOOLS

- Filter contrast data by fold-change and p-value
- Find genes with a similar expression pattern

PLOTTING TOOLS

- Heatmap for sequencing data
- Scatterplot for two EdgeR contrasts
- Volcano plot for an EdgeR contrast
- MA plot for EdgeR normalized data

ANNOTATION TOOLS

- Annotate a table with gene ids

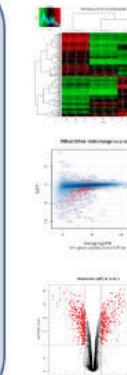
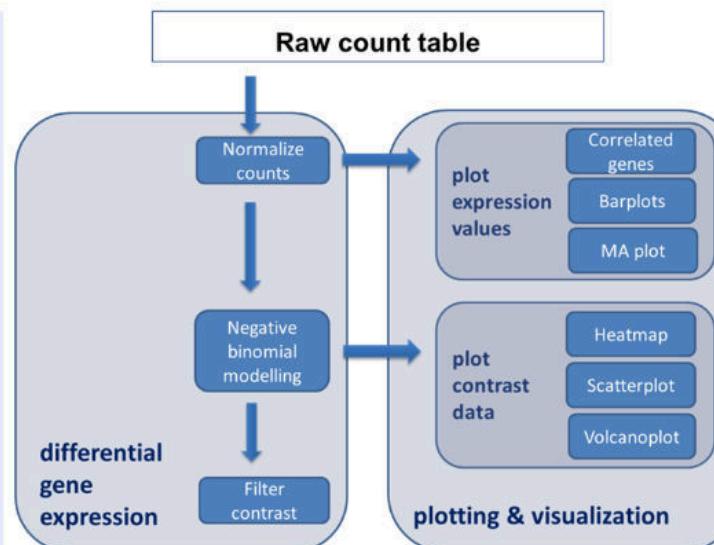
OLD EDGER TOOLS

- edgeR workflow for differential gene expression

TOOLS UNDER DEVELOPMENT

- pGSEA - Gene Set Enrichment Analysis for EdgeR data
- Plot expression values of genes

**FMI: qPCR**



# standard workflow for RNA-seq count data

## FMI: NGS differential expression

### EDGER TOOLS

- [EdgeR normalization](#) for RNA-seq analysis

- [Differential expression](#) for normalized RNA-seq data

### FILTERING TOOLS

- [Filter contrast data](#) by fold-change and p-value
- [Find genes](#) with a similar expression pattern

### PLOTTING TOOLS

- [Heatmap](#) for sequencing data
- [Scatterplot](#) for two EdgeR contrasts
- [Volcano plot](#) for an EdgeR contrast
- [MA plot](#) for EdgeR normalized data

### ANNOTATION TOOLS

- [Annotate](#) a table with gene ids

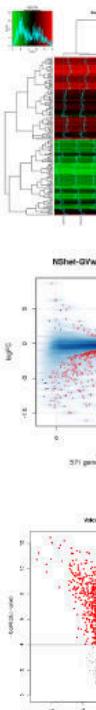
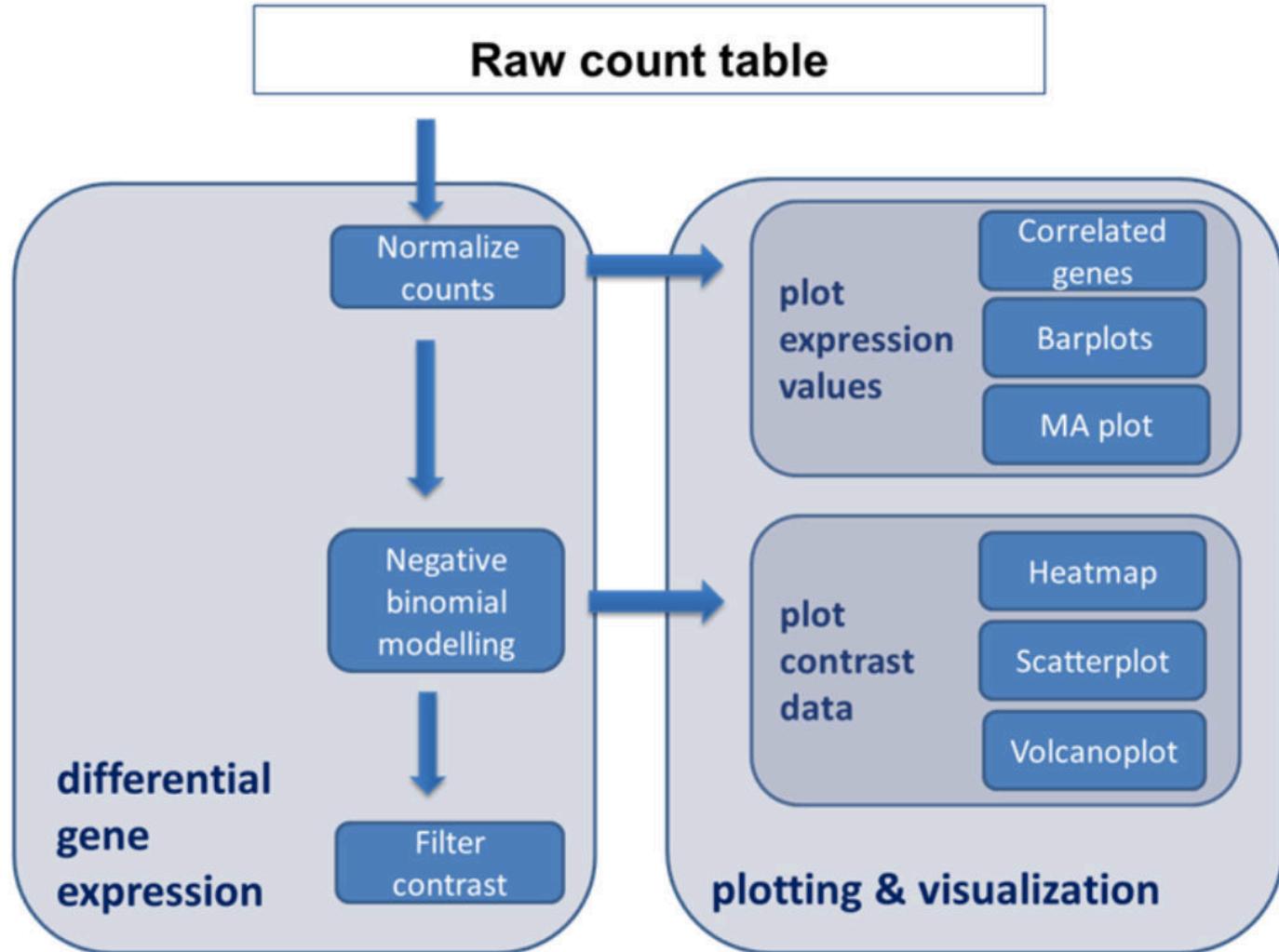
### OLD EDGER TOOLS

- [edgeR workflow](#) for differential gene expression

### TOOLS UNDER DEVELOPMENT

- [pGSEA](#) - Gene Set Enrichment Analysis for EdgeR data
- [Plot expression values of genes](#)

### FMI: qPCR



correlated

genes

Barplots

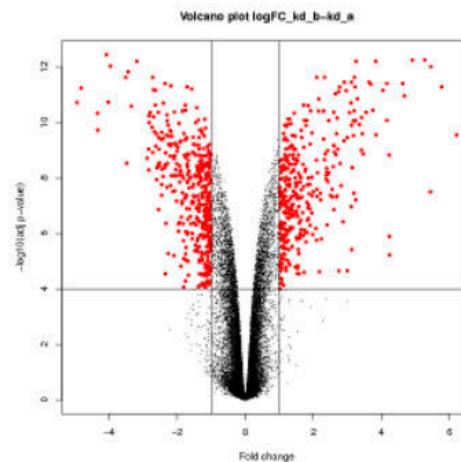
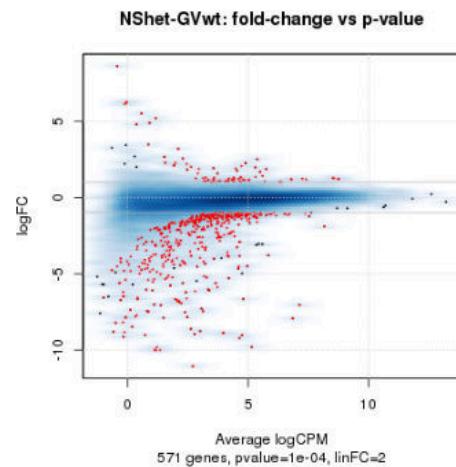
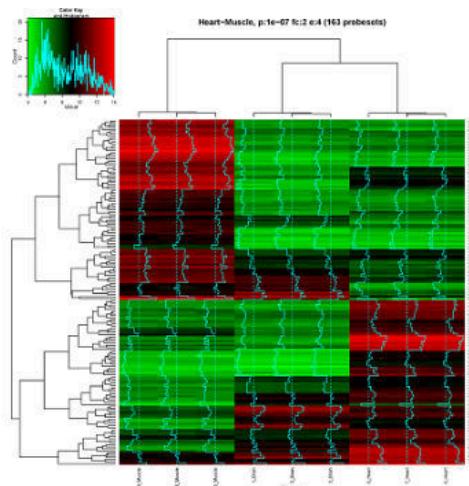
MA plot

Heatmap

Scatterplot

Volcanoplot

ualization



# Datatypes

*tabular and binary data types created for different R objects*

- *tabular datatype for user readable meta data*
- *binary datatype as easily accessible and readable data storage*

```
<datatype extension="exphe" type="galaxy.datatypes.tabular:exphe" display_in_upload="true" />
<datatype extension="exset" type="galaxy.datatypes.binary:exset" mimetype="application/octet-stream" display_in_upload="true"/>
<datatype extension="conphe" type="galaxy.datatypes.tabular:conphe" display_in_upload="true" />
<datatype extension="cont" type="galaxy.datatypes.binary:cont" mimetype="application/octet-stream" display_in_upload="true" />
<datatype extension="edgernorm" type="galaxy.datatypes.binary:edgernorm" mimetype="application/octet-stream" display_in_upload="true" />
<datatype extension="edgerpheno" type="galaxy.datatypes.tabular:edgerpheno" display_in_upload="true" />
<datatype extension="edgerdge" type="galaxy.datatypes.binary:edgerdge" mimetype="application/octet-stream" display_in_upload="true" />
<datatype extension="edgercont" type="galaxy.datatypes.tabular:edgercont" display_in_upload="true" />
```

*tabular and binary data types created for different R objects*

- *tabular datatype for user readable meta data*
- *binary datatype as easily accessible and readable data storage*

```
<datatype extension="exphe" type="galaxy.datatypes.tabular:exphe" display_in_upload="true" />
<datatype extension="exset" type="galaxy.datatypes.binary:exset" mimetype="application/octet-stream" display_in_upload="true"/>
<datatype extension="conphe" type="galaxy.datatypes.tabular:conphe" display_in_upload="true" />
<datatype extension="cont" type="galaxy.datatypes.binary:cont" mimetype="application/octet-stream" display_in_upload="true" />
<datatype extension="edgernorm" type="galaxy.datatypes.binary:edgernorm" mimetype="application/octet-stream" display_in_upload="true" />
<datatype extension="edgerpheno" type="galaxy.datatypes.tabular:edgerpheno" display_in_upload="true" />
<datatype extension="edgerdge" type="galaxy.datatypes.binary:edgerdge" mimetype="application/octet-stream" display_in_upload="true" />
<datatype extension="edgercont" type="galaxy.datatypes.tabular:edgercont" display_in_upload="true" />
```

```
Console C:/Users/tiroloff/Downloads/ 
> load("Galaxy120-[Normalized_DGE_object].edgernorm")
Loading required package: edgeR
Loading required package: limma
> ls()
[1] "dge"
Warning messages:
1: package 'edgeR' was built under R version 3.0.2
2: package 'limma' was built under R version 3.0.2
> dge
An object of class "DGEList"
$counts
      HR1.R2.1_DMSO MBA.PR_HR1.P.2_BYL MBA.PR_HR1.P.3_BYL719 MBA.PR_HR1.P.3_DMSO
NM_001001130        303           451           449           437
NM_001001144       1072          1321          1362           968
NM_001001152        122            319           269           146
NM_001001160         15             10            12            15
NM_001001176         9              13             9            45
      MBA.PR_HR1.R2.1.ON_BYL719 MBA.PR_HR1.R2.1.ON_DMSO
NM_001001130           430           396
NM_001001144          1313          1300
NM_001001152           183           201
NM_001001160            20             8
NM_001001176           32            25
28842 more rows ...

$samples
          group lib.size norm.factors
HR1.R2.1_DMSO    control 61531434      1
MBA.PR_HR1.P.2_BYL control 51500844      1
MBA.PR_HR1.P.3_BYL719 control 54682219      1
MBA.PR_HR1.P.3_DMSO     test  63461504      1
MBA.PR_HR1.R2.1.ON_BYL719 test  50446360      1
MBA.PR_HR1.R2.1.ON_DMSO     test  57703484      1
```

*load history items in R  
for in-depth analysis*





## *Summary*

- local galaxy
- FMI specific tools (increasing)
- reduce workload of standard analyses
- R objects can be used seamlessly between galaxy and R

# *Credits*



*Hans-Rudolf*



*FMI users*