RNA-Seq Data Analysis

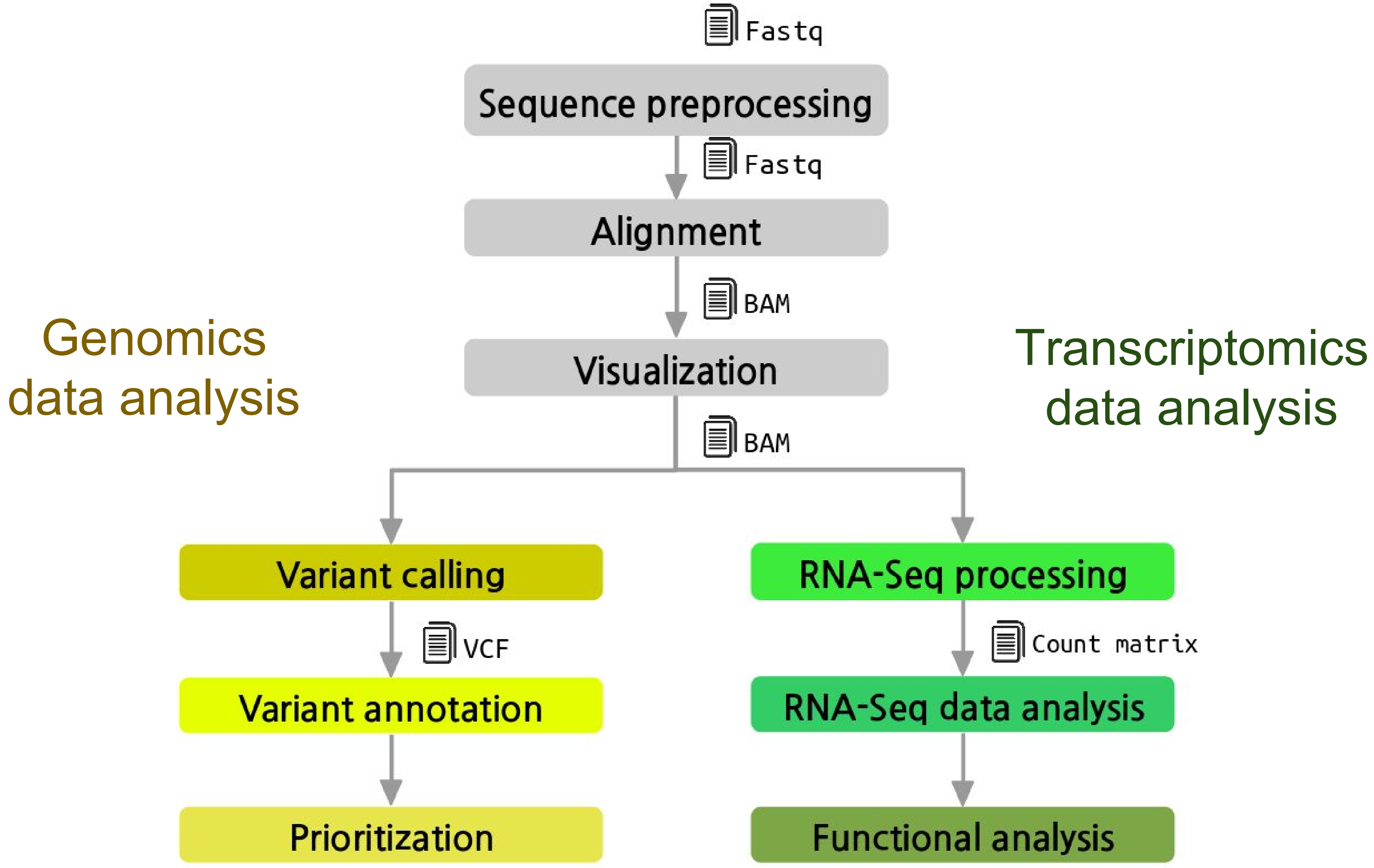
Marta R. Hidalgo

Máster en Biotecnología Biomédica
Universidad Politécnica de Valencia

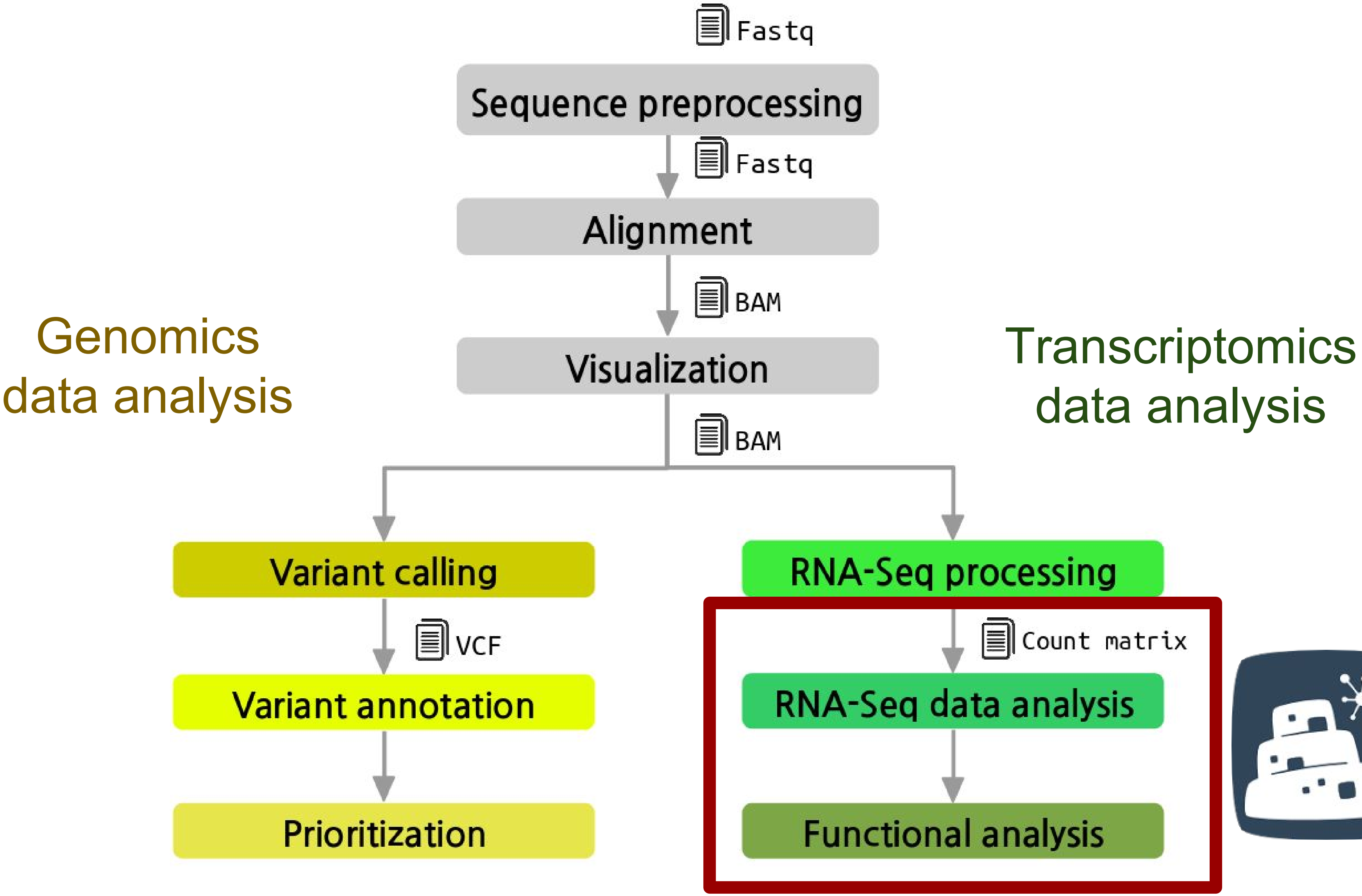
Unidad de
Bioinformática y
Bioestadística

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

NGS data analysis pipeline



NGS data analysis pipeline



Fastq

Sequence preprocessing

Fastq

Alignment

BAM

Visualization

BAM

Genomics data analysis

Transcriptomics data analysis

Variant calling

VCF

Variant annotation

Prioritization

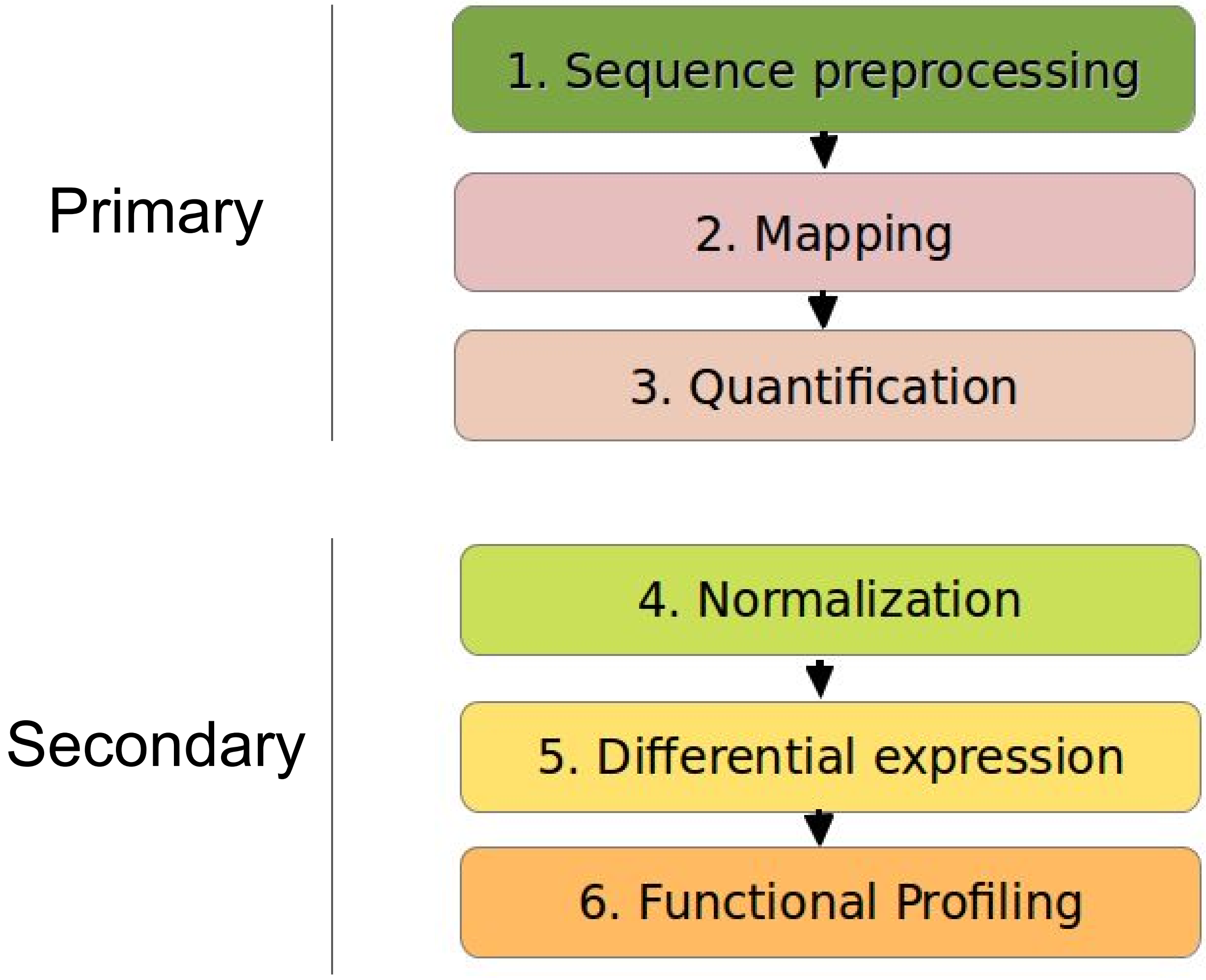
RNA-Seq processing

Count matrix

RNA-Seq data analysis

Functional analysis

RNA-seq data analysis pipeline



Primary

1. Sequence preprocessing
2. Mapping
3. Quantification

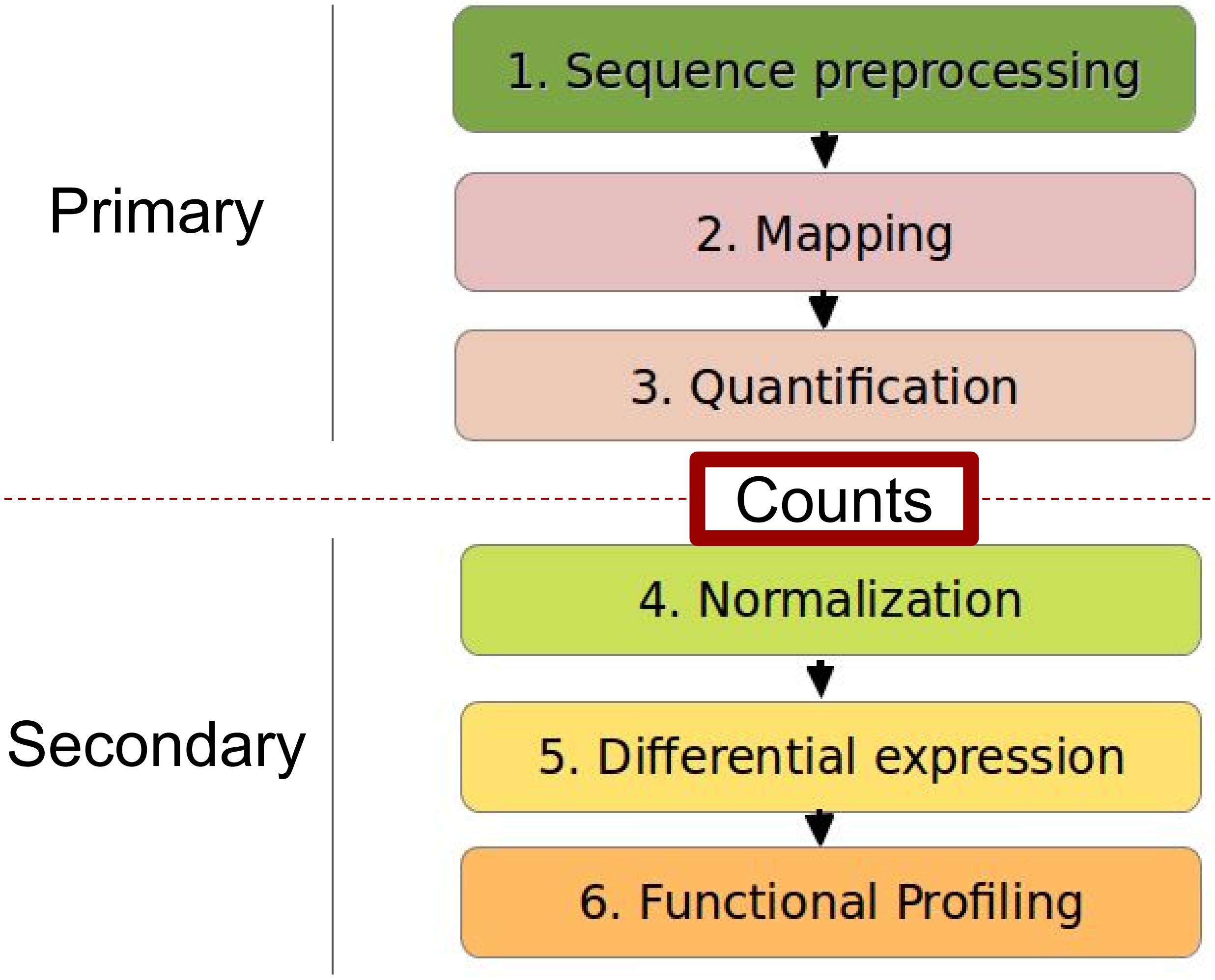
Secondary

4. Normalization
5. Differential expression
6. Functional Profiling

UNIX

```
me@linuxbox:~$
```

RNA-seq data analysis pipeline



Primary

1. Sequence preprocessing

2. Mapping

3. Quantification

Counts

Secondary

4. Normalization

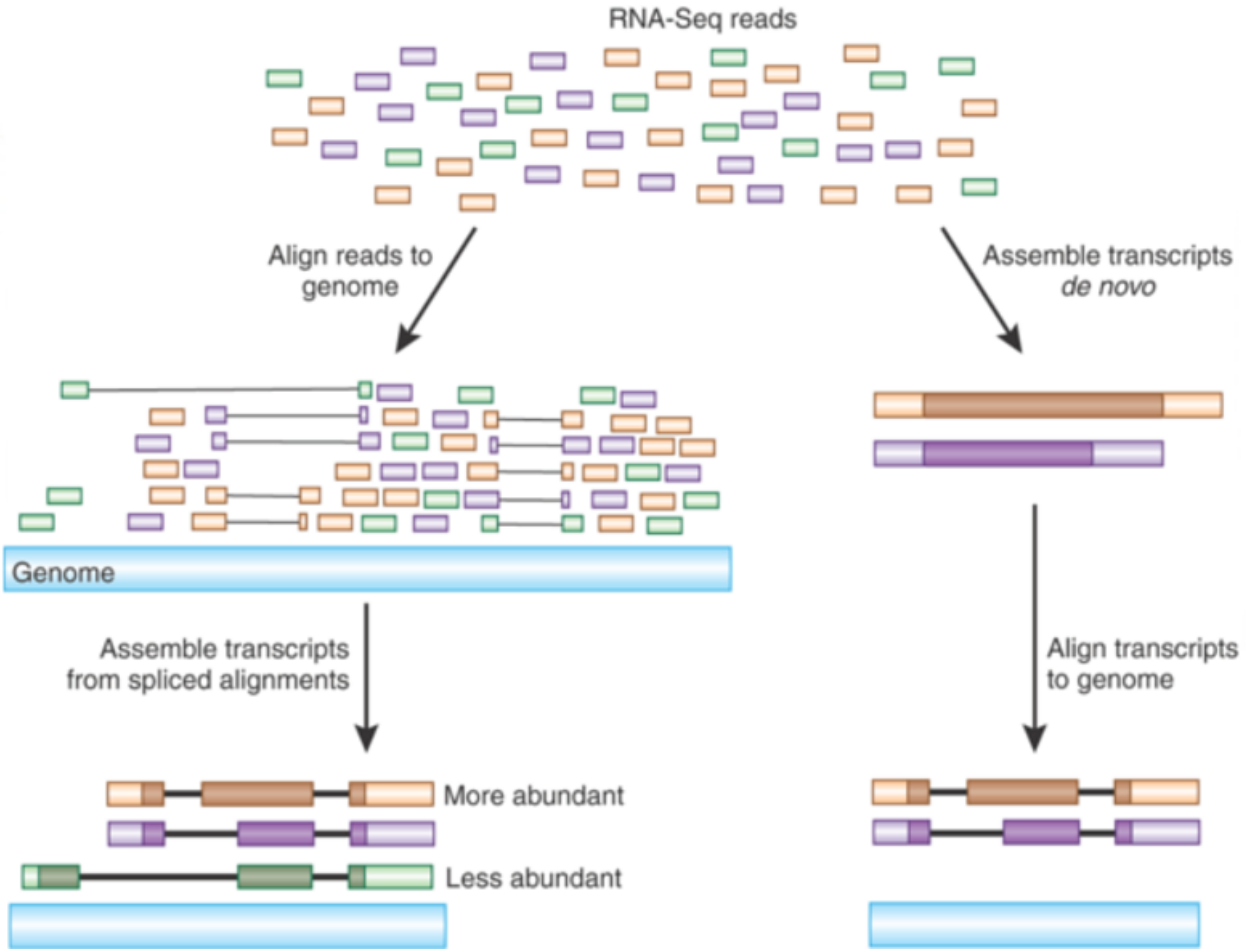
5. Differential expression

6. Functional Profiling

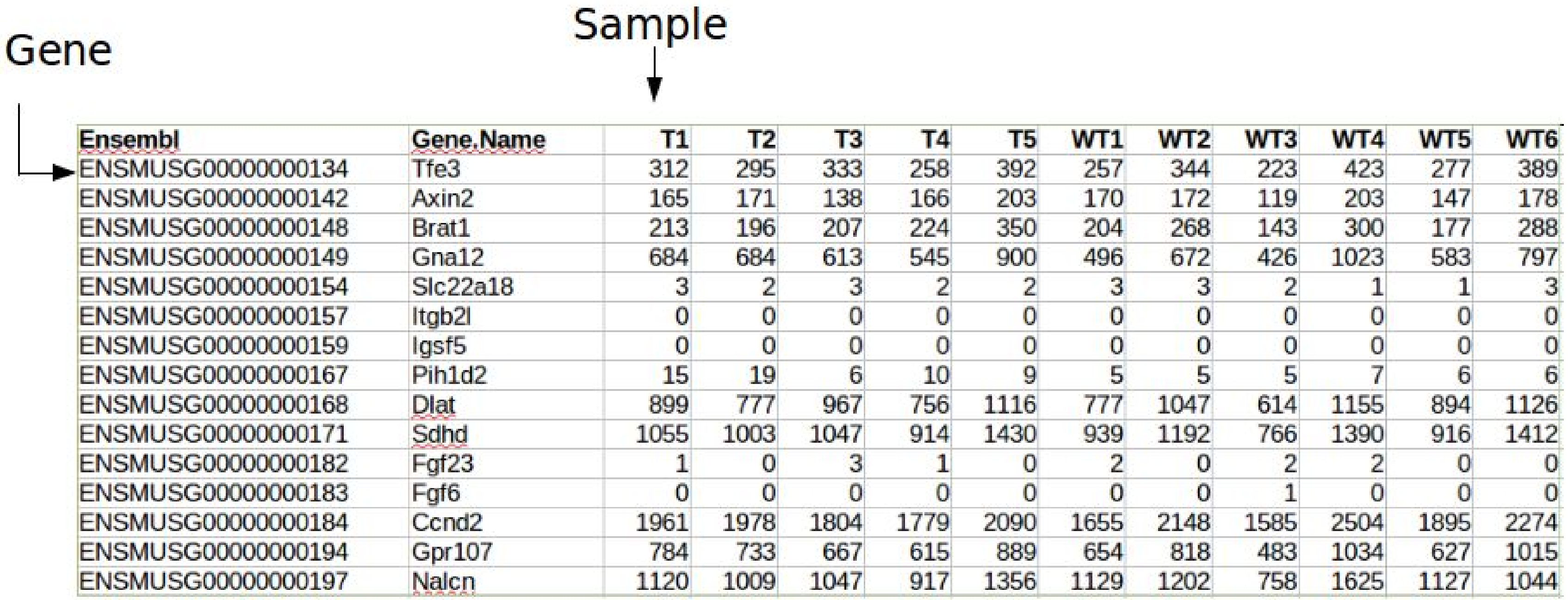
UNIX

```
me@linuxbox:~$
```

Counts


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Counts

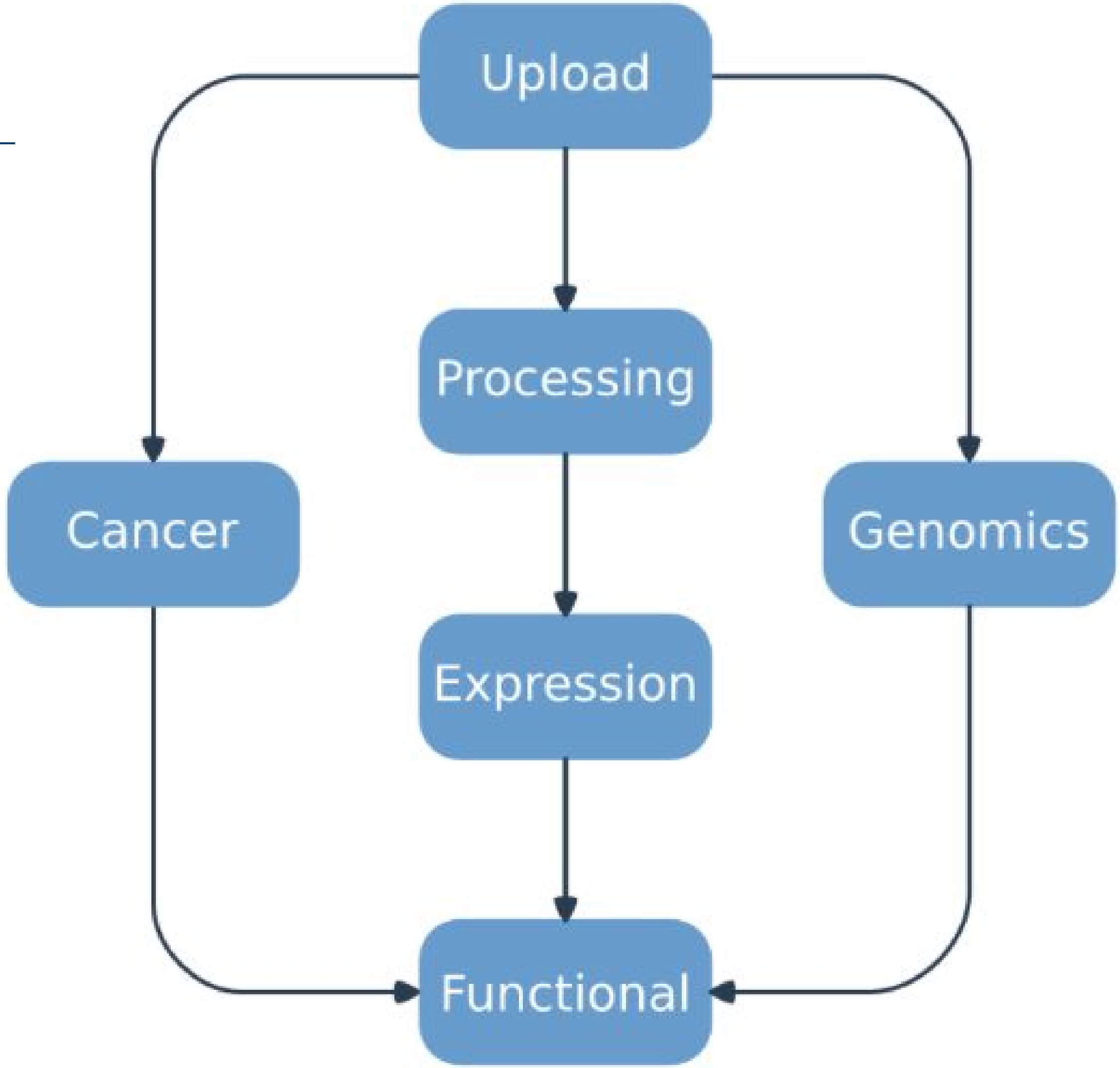


Ensembl	Gene.Name	T1	T2	T3	T4	T5	WT1	WT2	WT3	WT4	WT5	WT6
ENSMUSG00000000134	Tfe3	312	295	333	258	392	257	344	223	423	277	389
ENSMUSG00000000142	Axin2	165	171	138	166	203	170	172	119	203	147	178
ENSMUSG00000000148	Brat1	213	196	207	224	350	204	268	143	300	177	288
ENSMUSG00000000149	Gna12	684	684	613	545	900	496	672	426	1023	583	797
ENSMUSG00000000154	Slc22a18	3	2	3	2	2	3	3	2	1	1	3
ENSMUSG00000000157	Itgb2l	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000159	Igsf5	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000167	Pih1d2	15	19	6	10	9	5	5	5	7	6	6
ENSMUSG00000000168	Dlat	899	777	967	756	1116	777	1047	614	1155	894	1126
ENSMUSG00000000171	Sdhd	1055	1003	1047	914	1430	939	1192	766	1390	916	1412
ENSMUSG00000000182	Fgf23	1	0	3	1	0	2	0	2	2	0	0
ENSMUSG00000000183	Fgf6	0	0	0	0	0	0	0	1	0	0	0
ENSMUSG00000000184	Ccnd2	1961	1978	1804	1779	2090	1655	2148	1585	2504	1895	2274
ENSMUSG00000000194	Gpr107	784	733	667	615	889	654	818	483	1034	627	1015
ENSMUSG00000000197	Nalcn	1120	1009	1047	917	1356	1129	1202	758	1625	1127	1044

Babelomics 5

Babelomics 5

GENE EXPRESSION, GENOME
VARIATION AND FUNCTIONAL
PROFILING ANALYSIS SUITE



Upload

Processing

Cancer

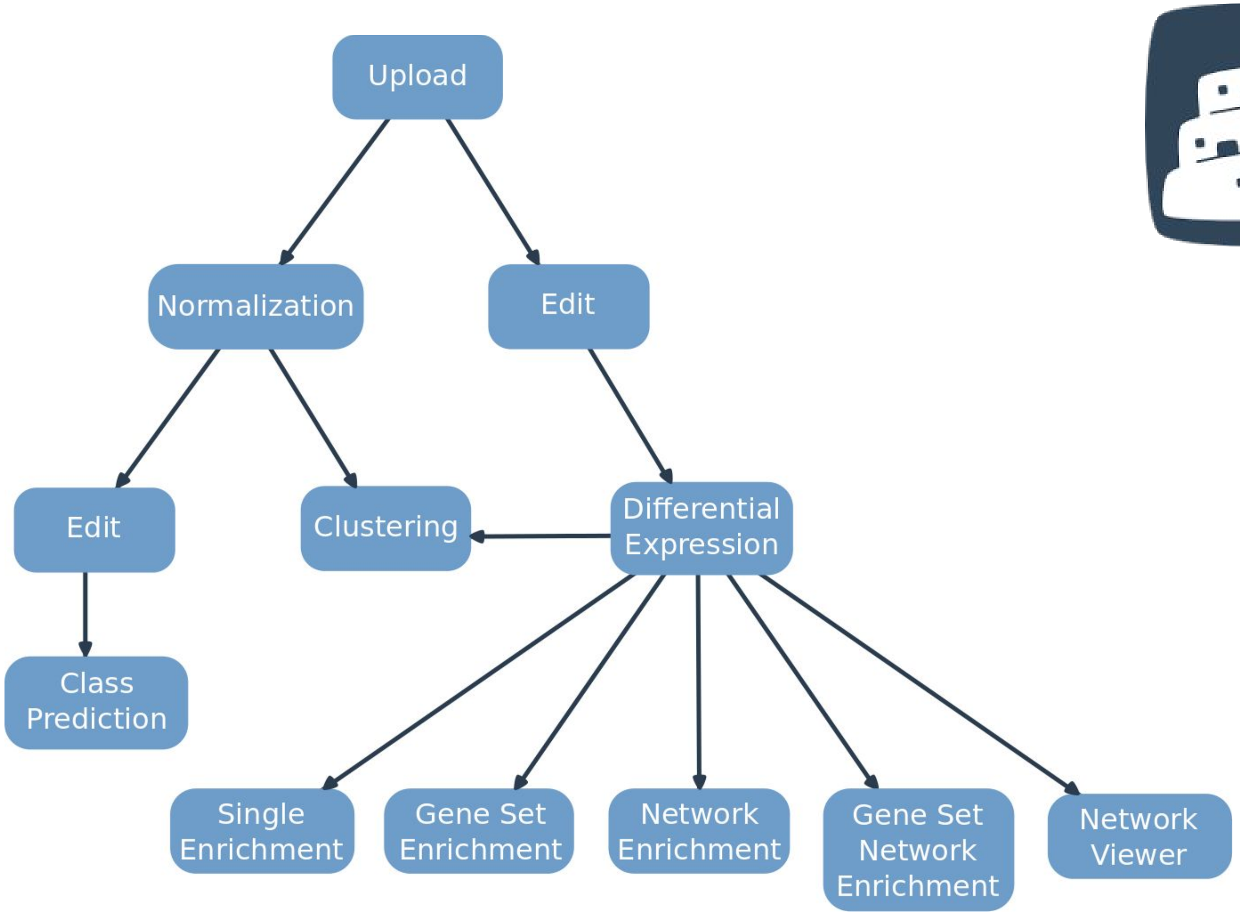
Genomics

Expression

Functional

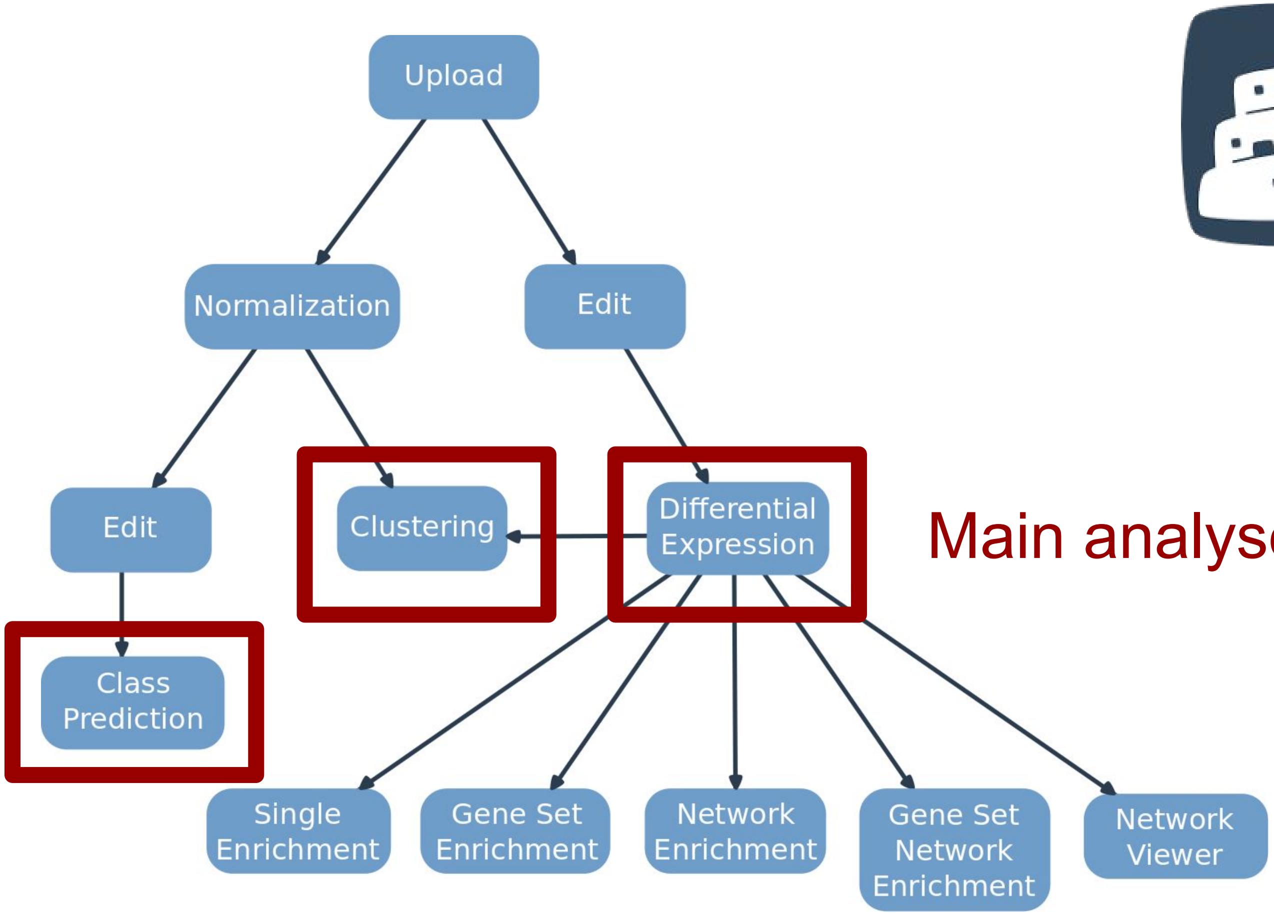
http://courses.babelomics.org

RNA-seq pipeline in Babelomics 5



http://courses.babelomics.org

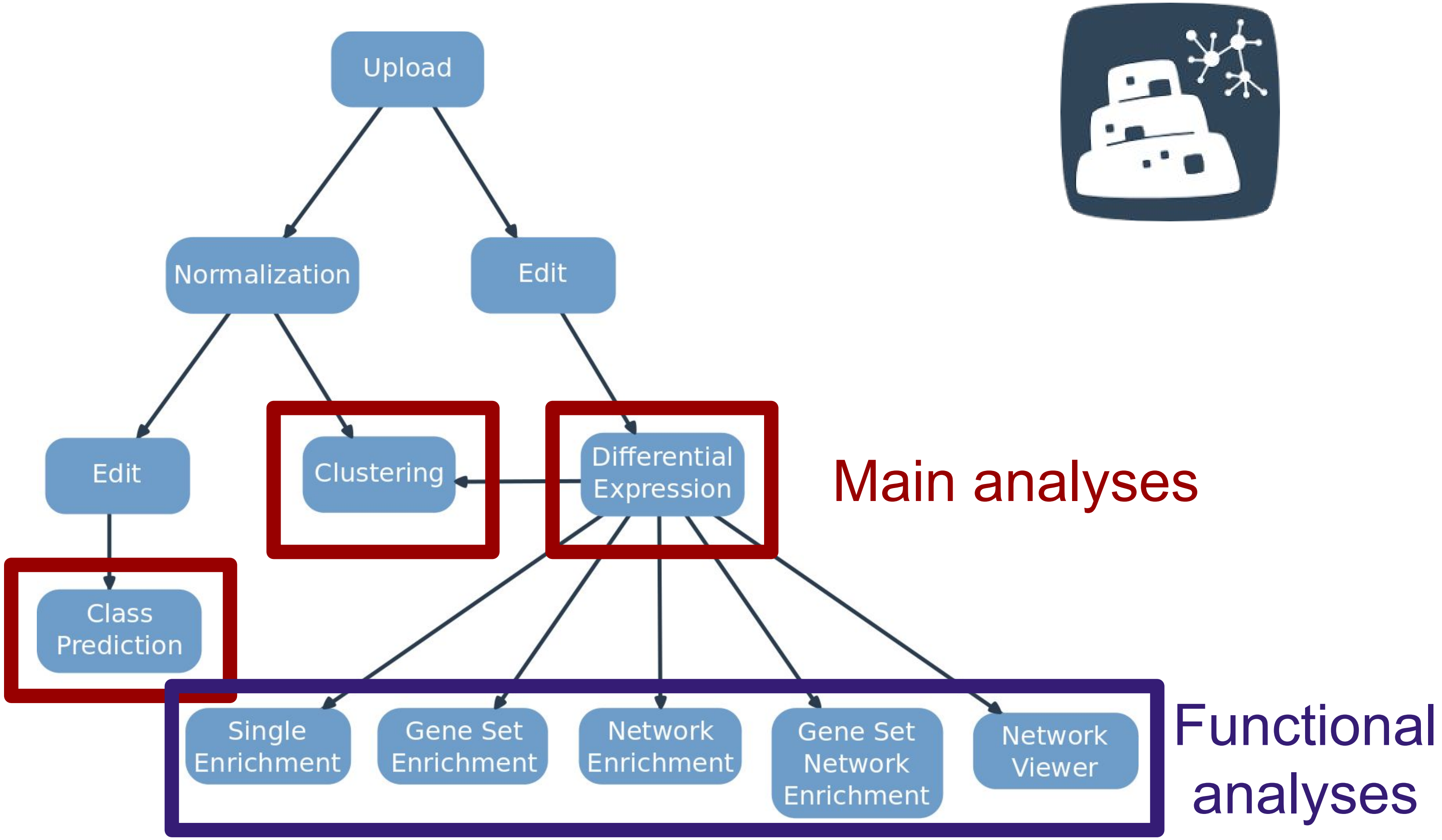
RNA-seq pipeline in Babelomics 5



Main analyses

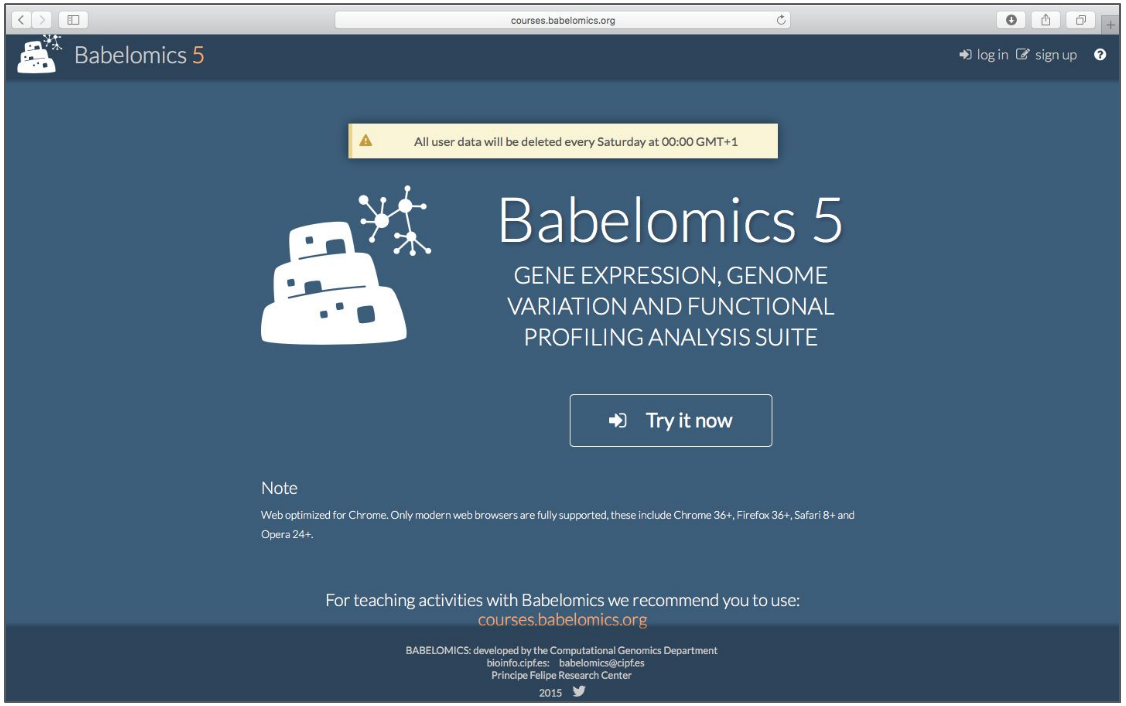
http://courses.babelomics.org

RNA-seq pipeline in Babelomics 5



http://courses.babelomics.org

Babelomics 5: web structure



All user data will be deleted every Saturday at 00:00 GMT+1

Babelomics 5

GENE EXPRESSION, GENOME VARIATION AND FUNCTIONAL PROFILING ANALYSIS SUITE

Note

Web optimized for Chrome. Only modern web browsers are fully supported, these include Chrome 36+, Firefox 36+, Safari 8+ and Opera 24+.

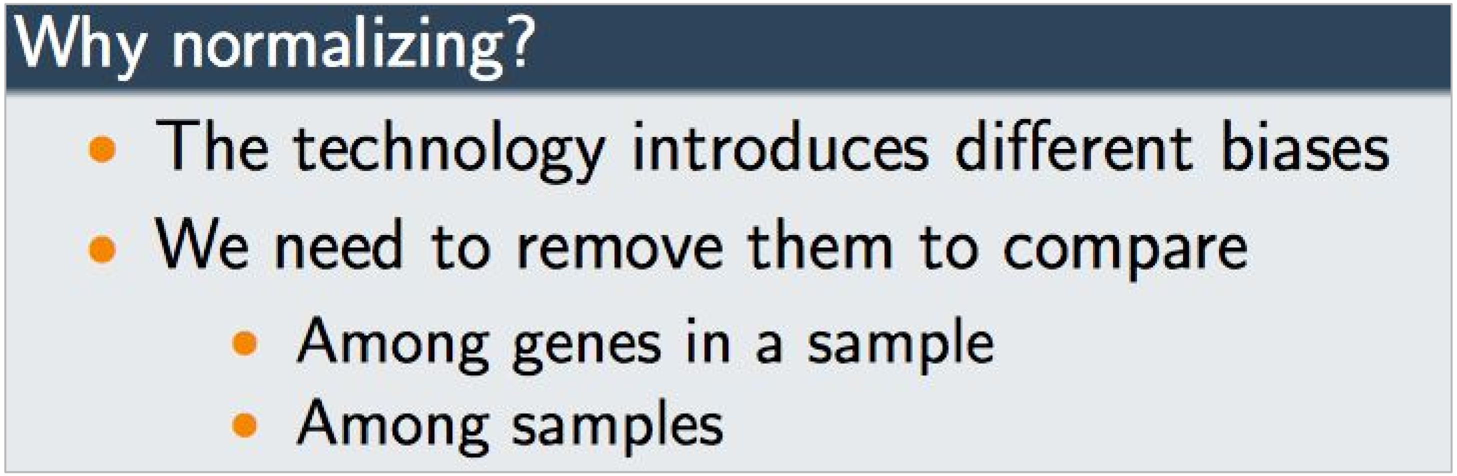
For teaching activities with Babelomics we recommend you to use:
courses.babelomics.org

BABELOMICS: developed by the Computational Genomics Department
bioinfo.cipf.es: babelomics@cipf.es
Principe Felipe Research Center
2015

http://courses.babelomics.org

Practical exercises: http://bioinfo.cipf.es/mbb

Normalization



Why normalizing?

- The technology introduces different biases
- We need to remove them to compare
  - Among genes in a sample
  - Among samples

Biases

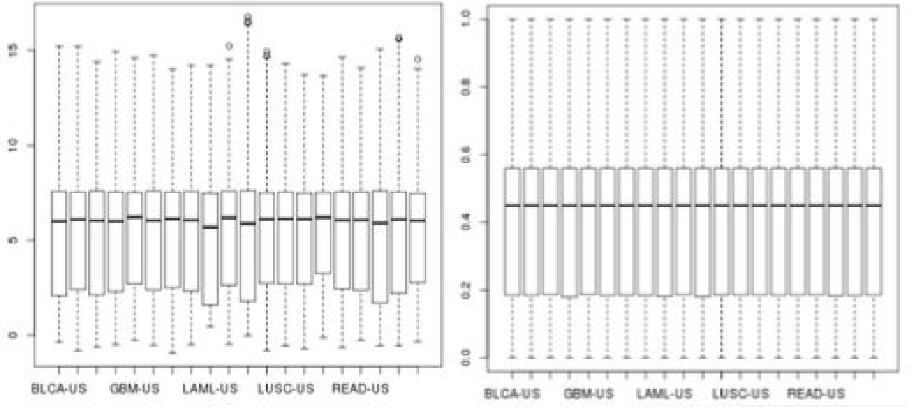
- Gene length: larger genes get more reads
- Library depth: deeper libraries get more reads
- RNA composition: some genes steal reads from other genes
- Others

Normalization methods

- Reads per kilobase per million (**RPKM**): Removes gene length and library depth biases.

$$\mathrm{RPKM} = \frac{\textit{total exon reads}}{\textit{mapped reads (millions)} * \textit{exon length (KB)}}$$

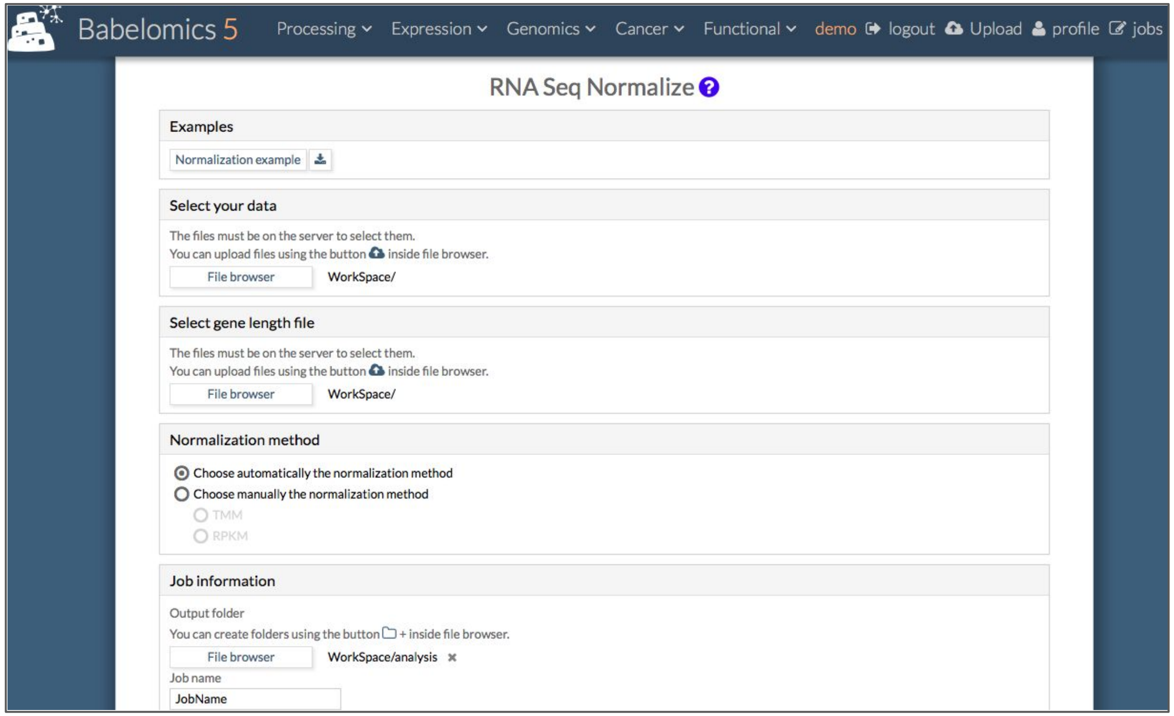
- Trimmed means of M-values (**TMM**): assumes only a few genes are differentially expressed and changes library depth.
- **Quantiles**: makes all sample distributions the same.



Normalization methods in Babelomics 5

- RPKM (gene length required)
- TMM
- TMM with gene length correction (gene length required)
- Automatic selection of the method based on the diagnostic test for differences in RNA composition from NOISeq.

Babelomics 5: normalization



RNA Seq Normalize

Examples

Normalization example

Select your data

The files must be on the server to select them.
You can upload files using the button inside file browser.

File browser WorkSpace/

Select gene length file

The files must be on the server to select them.
You can upload files using the button inside file browser.

File browser WorkSpace/

Normalization method

● Choose automatically the normalization method
○ Choose manually the normalization method
○ TMM
○ RPKM

Job information

Output folder
You can create folders using the button + inside file browser.

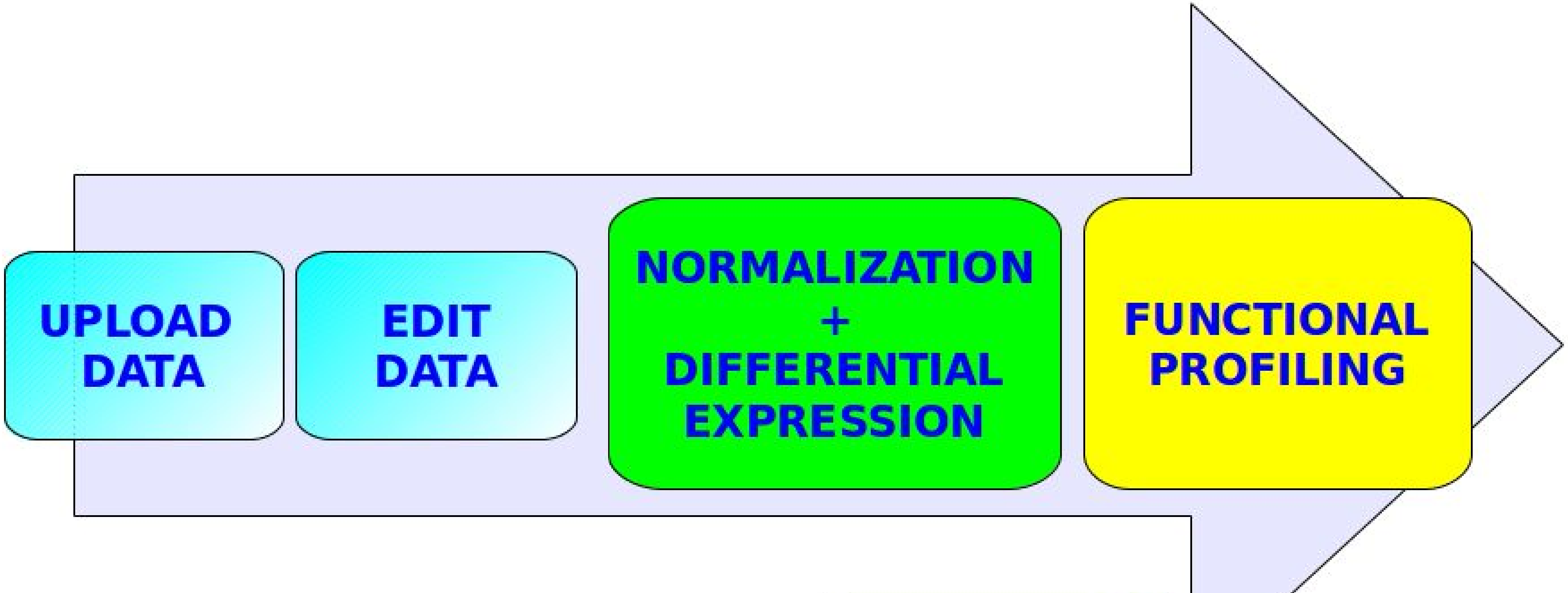
File browser WorkSpace/analysis ✖

Job name
JobName

http://courses.babelomics.org

Normalization exercise: http://bioinfo.cipf.es/mbb

Differential expression

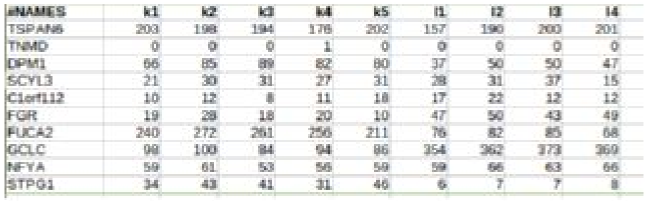


UPLOAD DATA

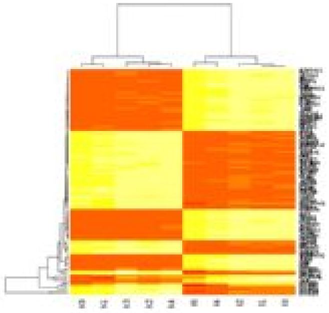
EDIT DATA

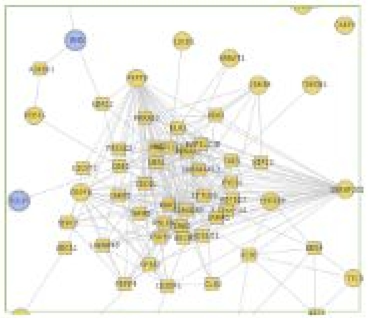
NORMALIZATION + DIFFERENTIAL EXPRESSION

FUNCTIONAL PROFILING



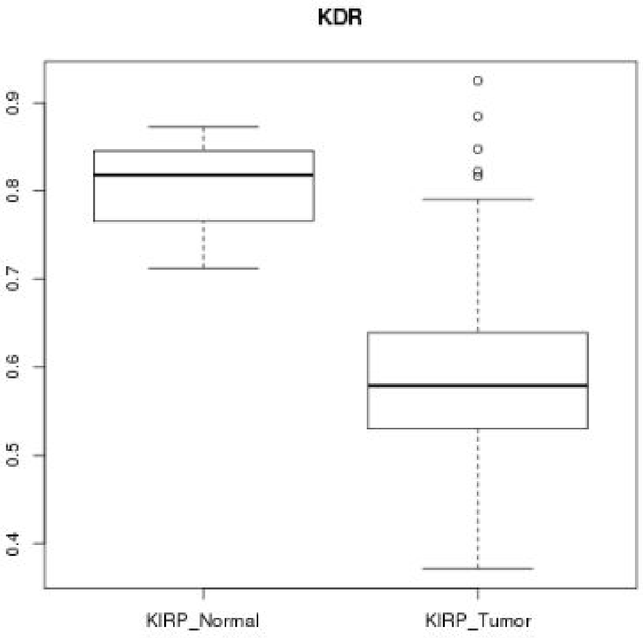
#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4
TSPAN6	203	198	194	176	202	157	190	200	201
TNMD	0	0	0	1	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47
SCYL3	21	30	31	27	31	28	31	37	15
C1orf112	10	12	8	11	18	17	22	12	12
FGR	19	28	18	20	10	47	50	43	49
FUCA2	240	272	261	256	211	76	82	85	68
GCLC	98	100	84	94	86	354	362	373	369
NFYA	59	61	53	56	59	59	66	63	66
STPG1	34	43	41	31	46	6	7	7	8

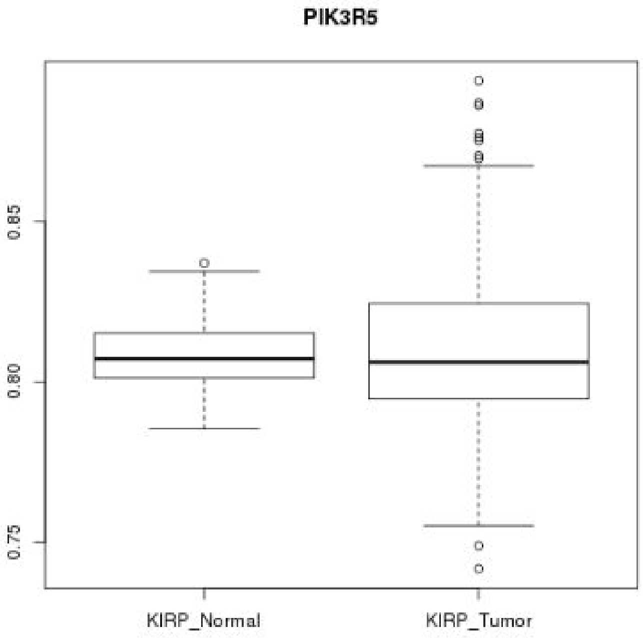




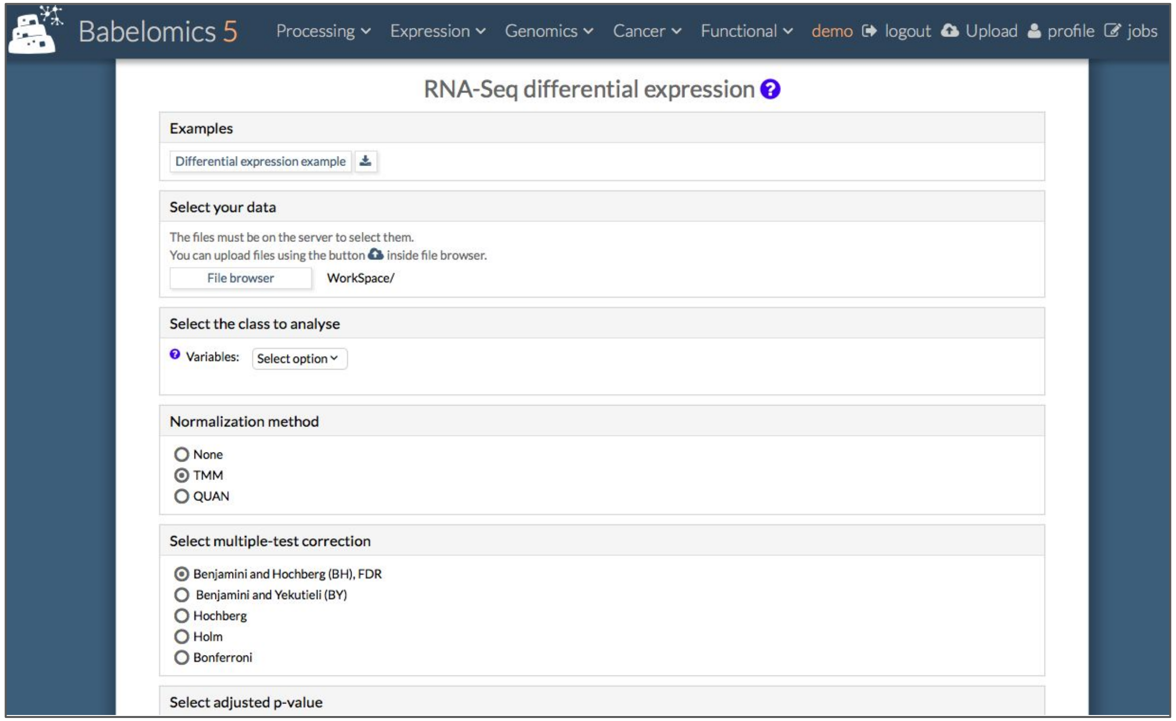
Differential expression

Genes that show statistically significant differences in expression level between conditions





Babelomics 5: differential expression



RNA-Seq differential expression

Examples

Differential expression example

Select your data

The files must be on the server to select them.
You can upload files using the button inside file browser.

File browser WorkSpace/

Select the class to analyse

Variables: Select option

Normalization method

○ None
● TMM
○ QUAN

Select multiple-test correction

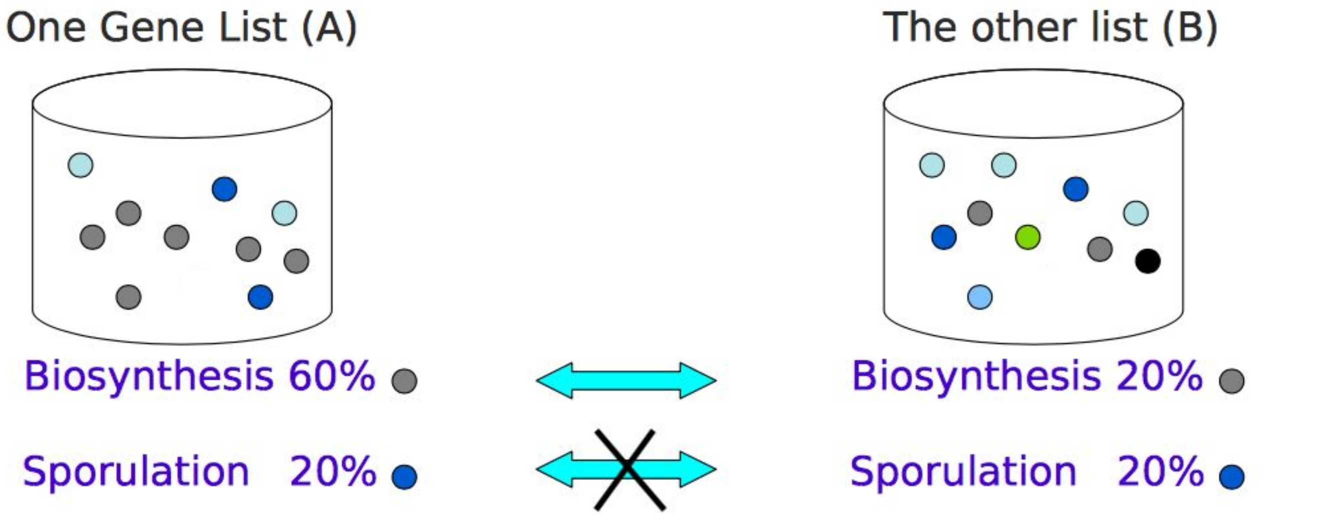
● Benjamini and Hochberg (BH), FDR
○ Benjamini and Yekutieli (BY)
○ Hochberg
○ Holm
○ Bonferroni

Select adjusted p-value

http://courses.babelomics.org

Normalization exercise: http://bioinfo.cipf.es/mbb

Functional profiling: single enrichment



One Gene List (A)

The other list (B)

Biosynthesis 60% ●

Biosynthesis 20% ●

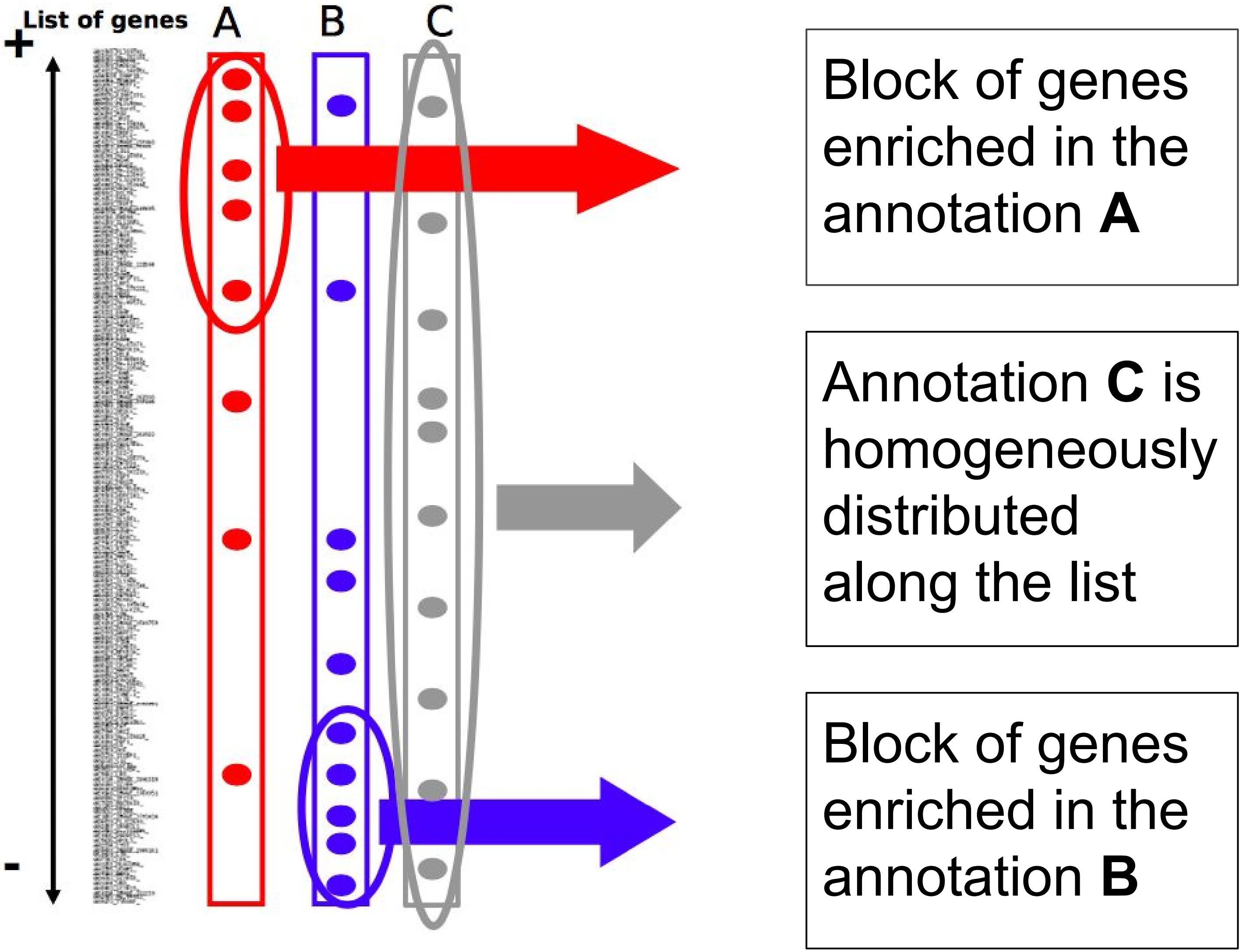
Sporulation 20% ●

Sporulation 20% ●

	A	B
Biosynthesis	**6**	**2**
No biosynthesis	**4**	**8**

Association test

Functional profiling: gene set enrichment

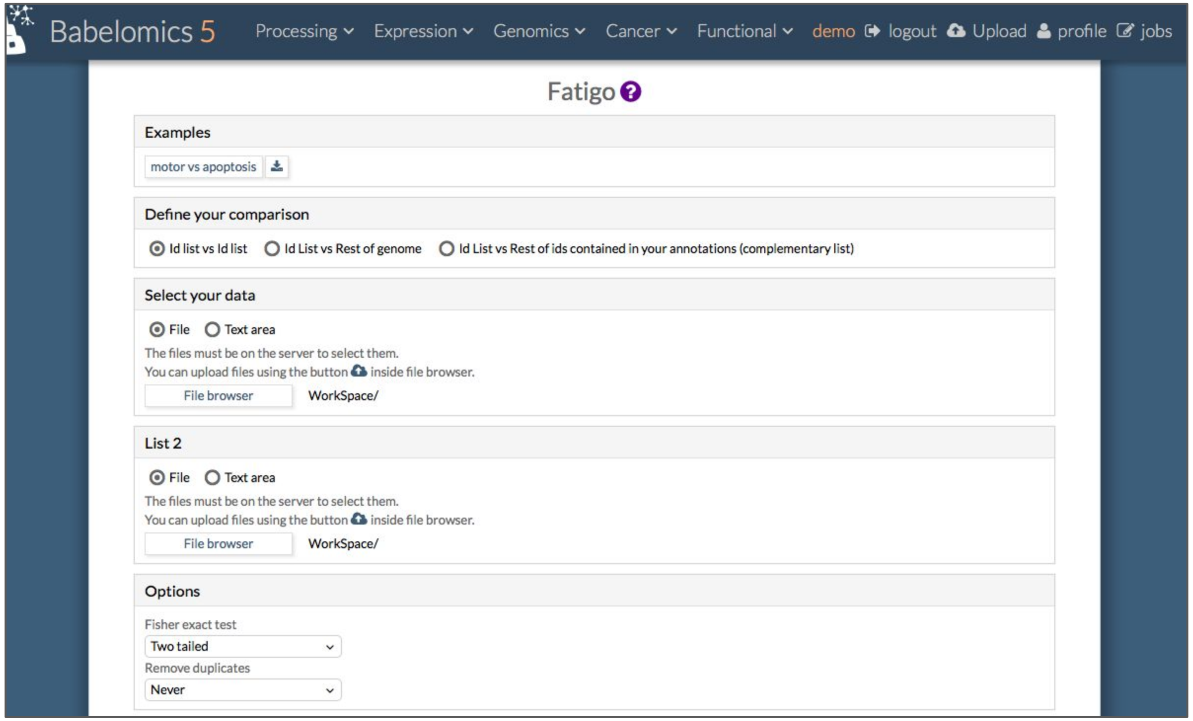


Block of genes enriched in the annotation **A**

Annotation **C** is homogeneously distributed along the list

Block of genes enriched in the annotation **B**

Babelomics 5: functional enrichment



Fatigo

Examples

motor vs apoptosis

Define your comparison

● Id list vs Id list ○ Id List vs Rest of genome ○ Id List vs Rest of ids contained in your annotations (complementary list)

Select your data

● File ○ Text area

The files must be on the server to select them.
You can upload files using the button inside file browser.

File browser WorkSpace/

List 2

● File ○ Text area

The files must be on the server to select them.
You can upload files using the button inside file browser.

File browser WorkSpace/

Options

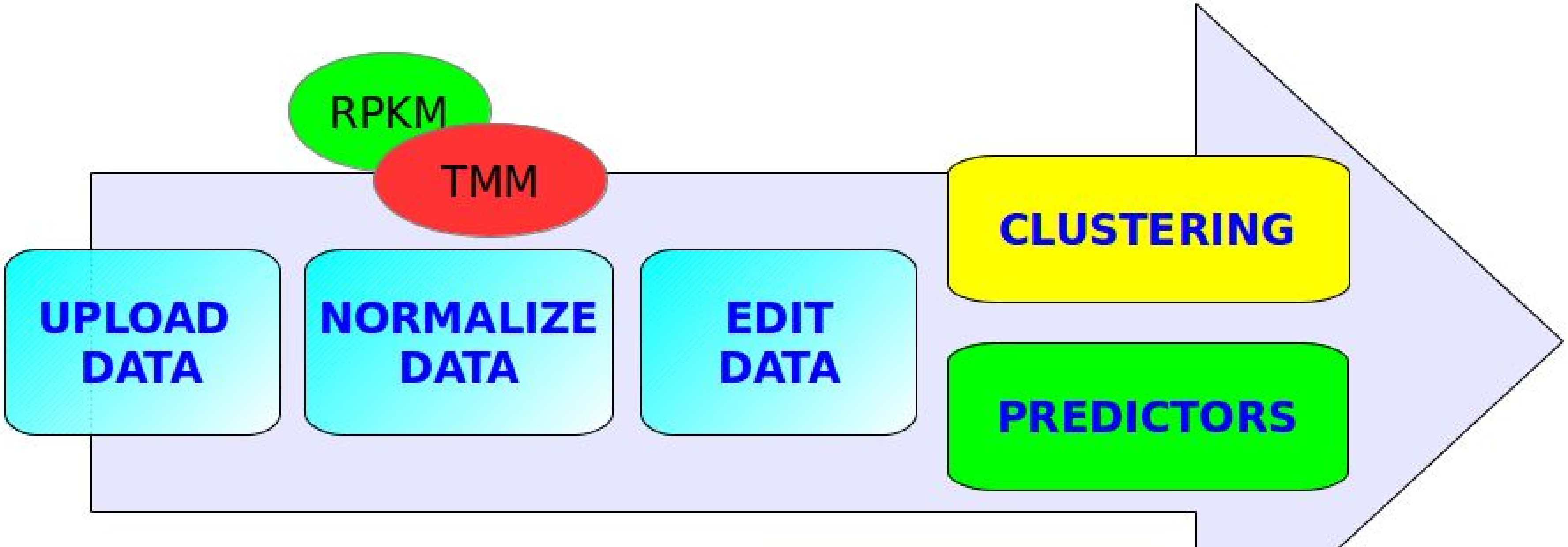
Fisher exact test: Two tailed

Remove duplicates: Never

http://courses.babelomics.org

Single enrichment exercise: http://bioinfo.cipf.es/mbb

Classification methods: unsupervised and supervised



RPKM

TMM

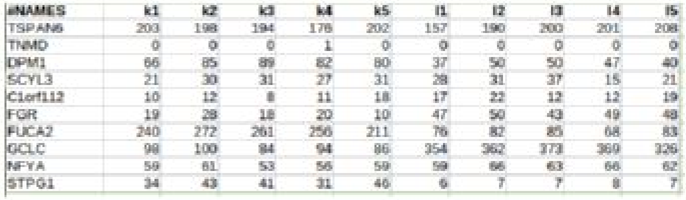
UPLOAD DATA

NORMALIZE DATA

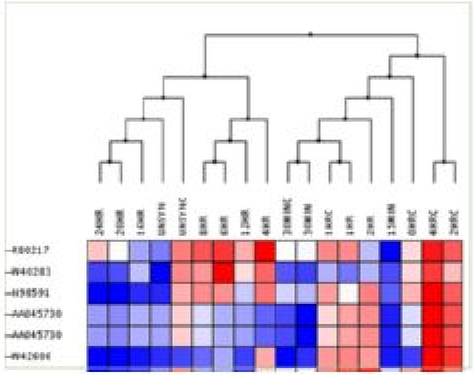
EDIT DATA

CLUSTERING

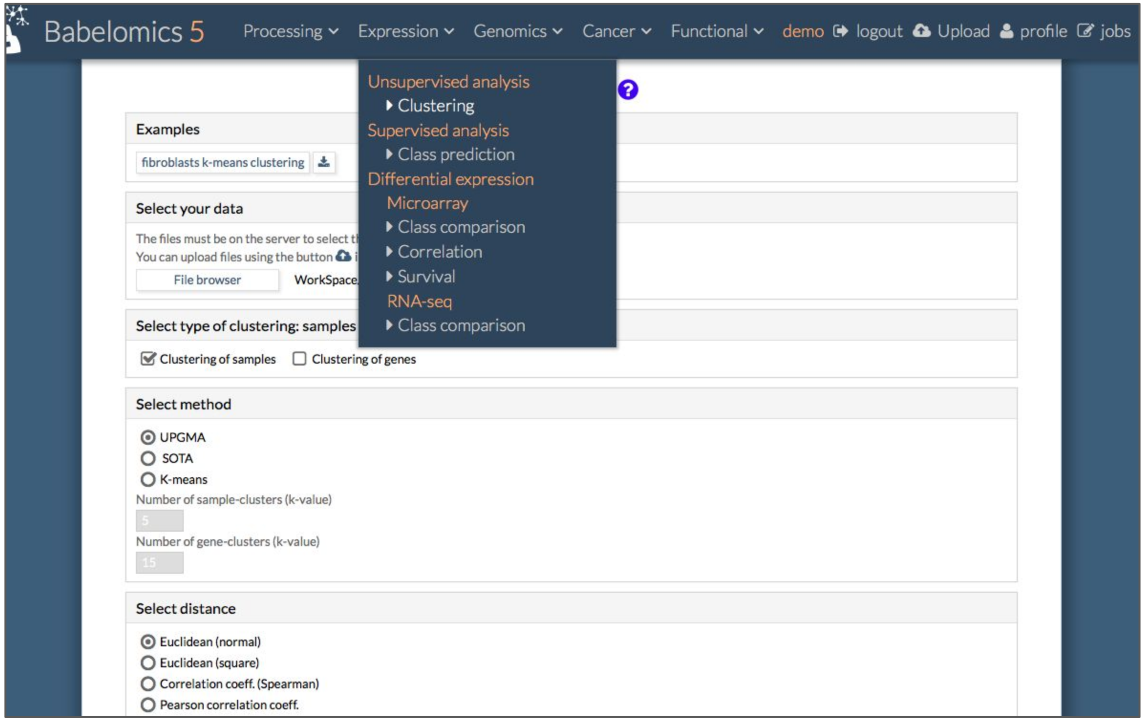
PREDICTORS



#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	100	84	94	86	354	362	373	369	326
NFYA	59	61	53	56	59	59	66	63	66	62
STPG1	34	43	41	31	46	6	7	7	8	7



Babelomics 5: classification



Examples

fibroblasts k-means clustering

Select your data

The files must be on the server to select t

You can upload files using the button i

File browser

WorkSpace.

Select type of clustering: samples

☑ Clustering of samples ☐ Clustering of genes

Select method

● UPGMA

○ SOTA

○ K-means

Number of sample-clusters (k-value)

5

Number of gene-clusters (k-value)

15

Select distance

● Euclidean (normal)

○ Euclidean (square)

○ Correlation coeff. (Spearman)

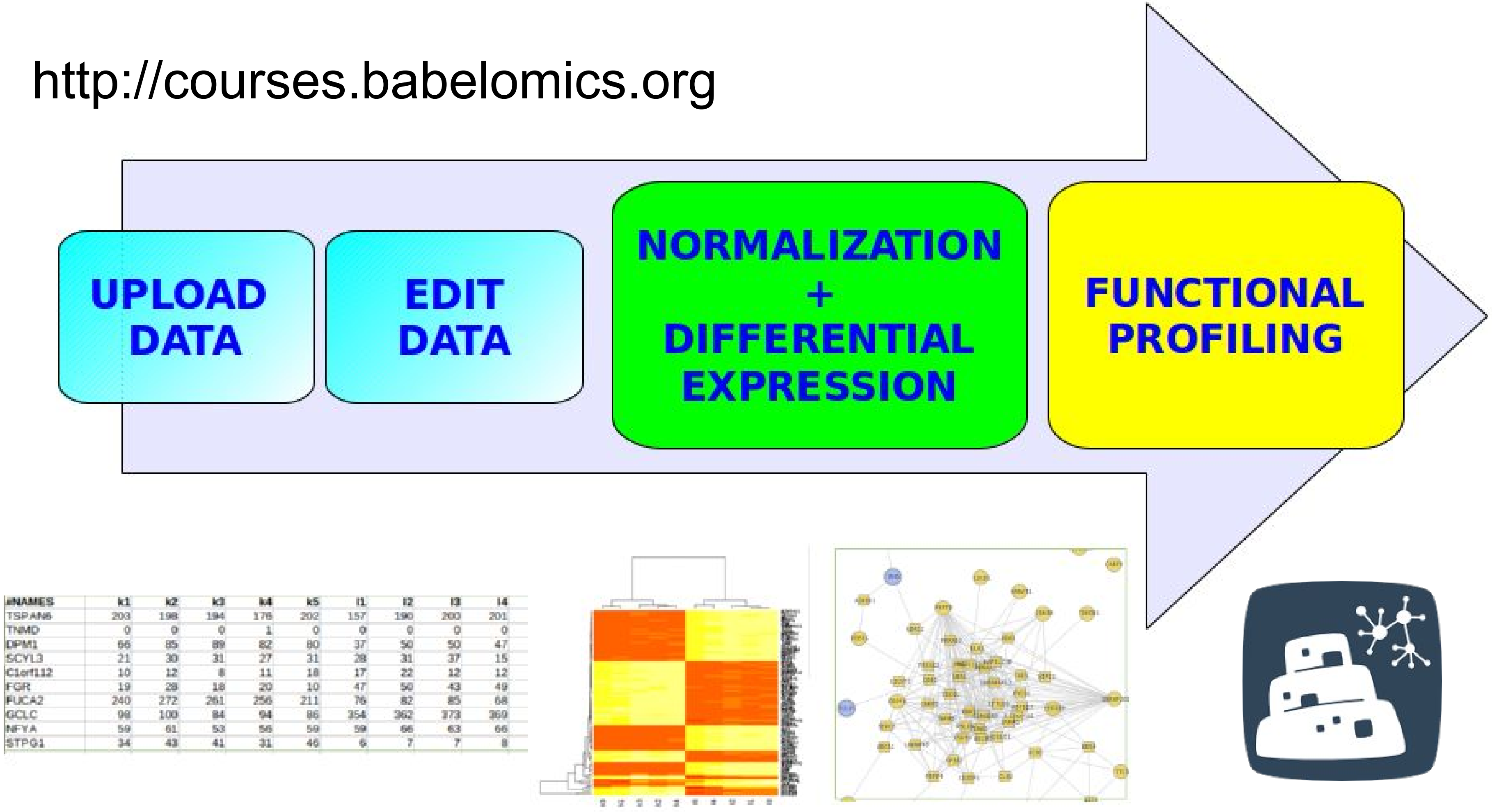
○ Pearson correlation coeff.

http://courses.babelomics.org

Clustering exercise: http://bioinfo.cipf.es/mbb

Babelomics 5: RNA-seq pipeline

http://courses.babelomics.org



UPLOAD DATA

EDIT DATA

NORMALIZATION + DIFFERENTIAL EXPRESSION

FUNCTIONAL PROFILING

#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4
TSPAN6	203	198	194	178	202	157	190	200	201
TNMD	0	0	0	1	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47
SCYL3	21	30	31	27	31	28	31	37	15
C1orf112	10	12	8	11	18	17	22	12	12
FGR	19	28	18	20	10	47	50	43	49
FUCA2	240	272	261	256	211	76	82	85	68
GCLC	98	100	84	94	86	354	362	373	369
NFYA	59	61	53	56	59	59	66	63	66
STPG1	34	43	41	31	46	6	7	7	8

Pipeline exercise: http://bioinfo.cipf.es/mbb