

Omics-based biomarkers detection

Francisco García García
Bioinformatics & Biostatistics Unit. CIPF



Unidad de
Bioinformática y
Bioestadística



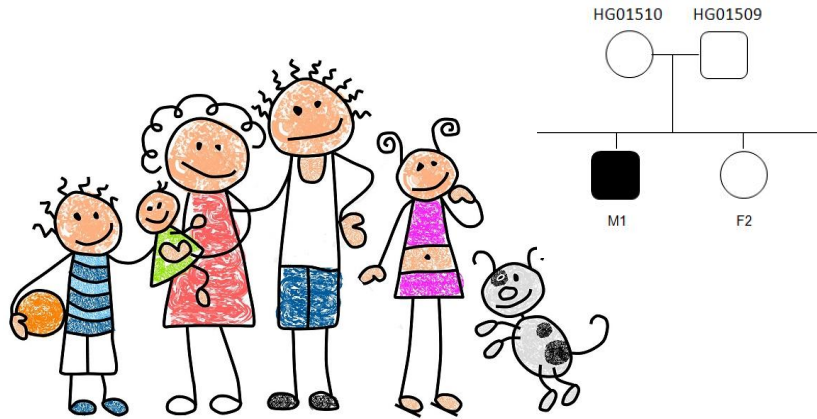
PRINCIPE FELIPE
CENTRO DE INVESTIGACION



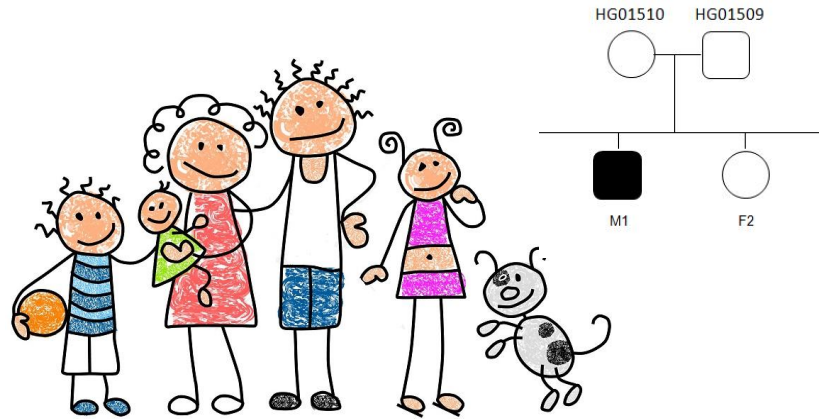
WODA

WEB-BASED OMICS DATA ANALYSIS

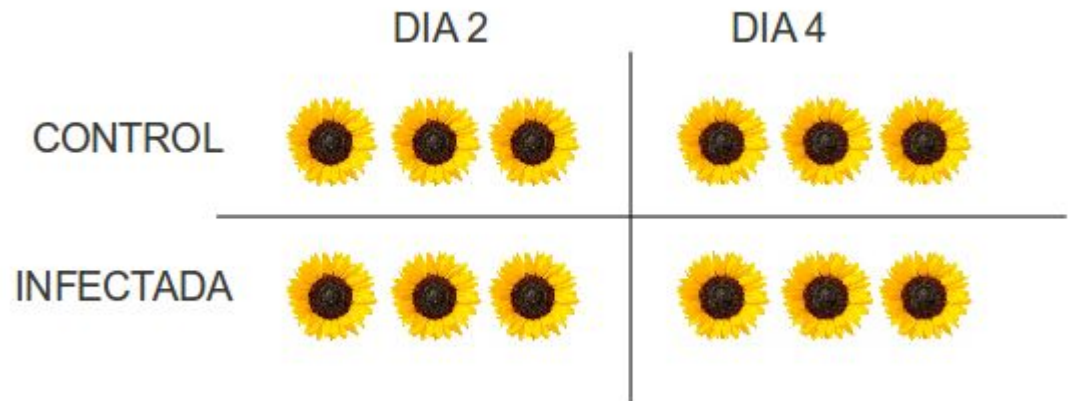
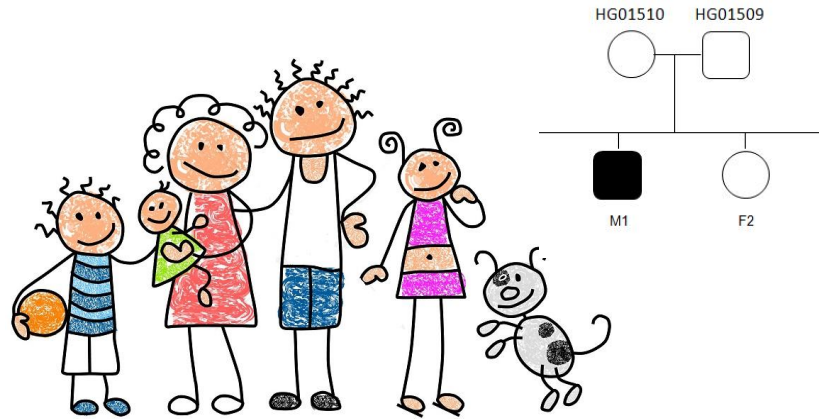
Detecting biomarkers



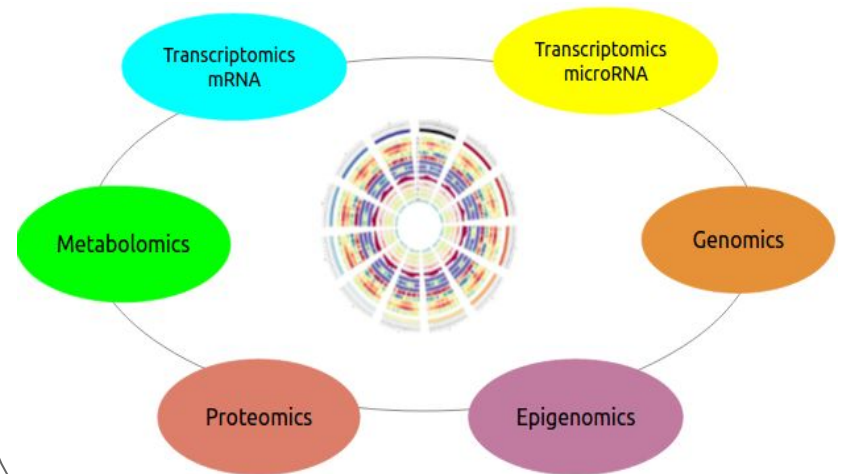
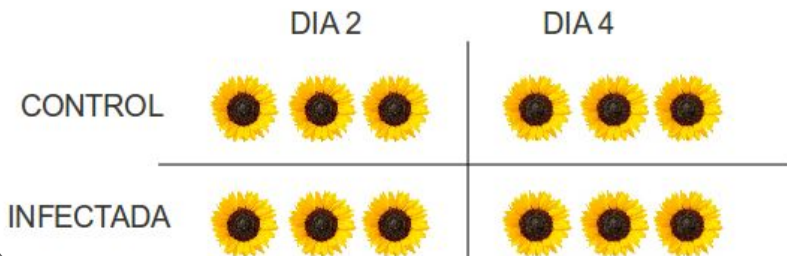
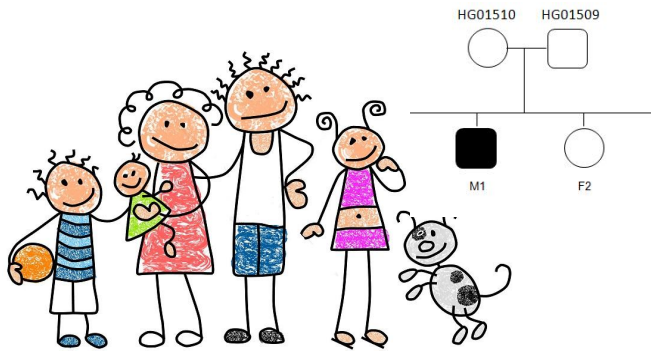
Detecting biomarkers



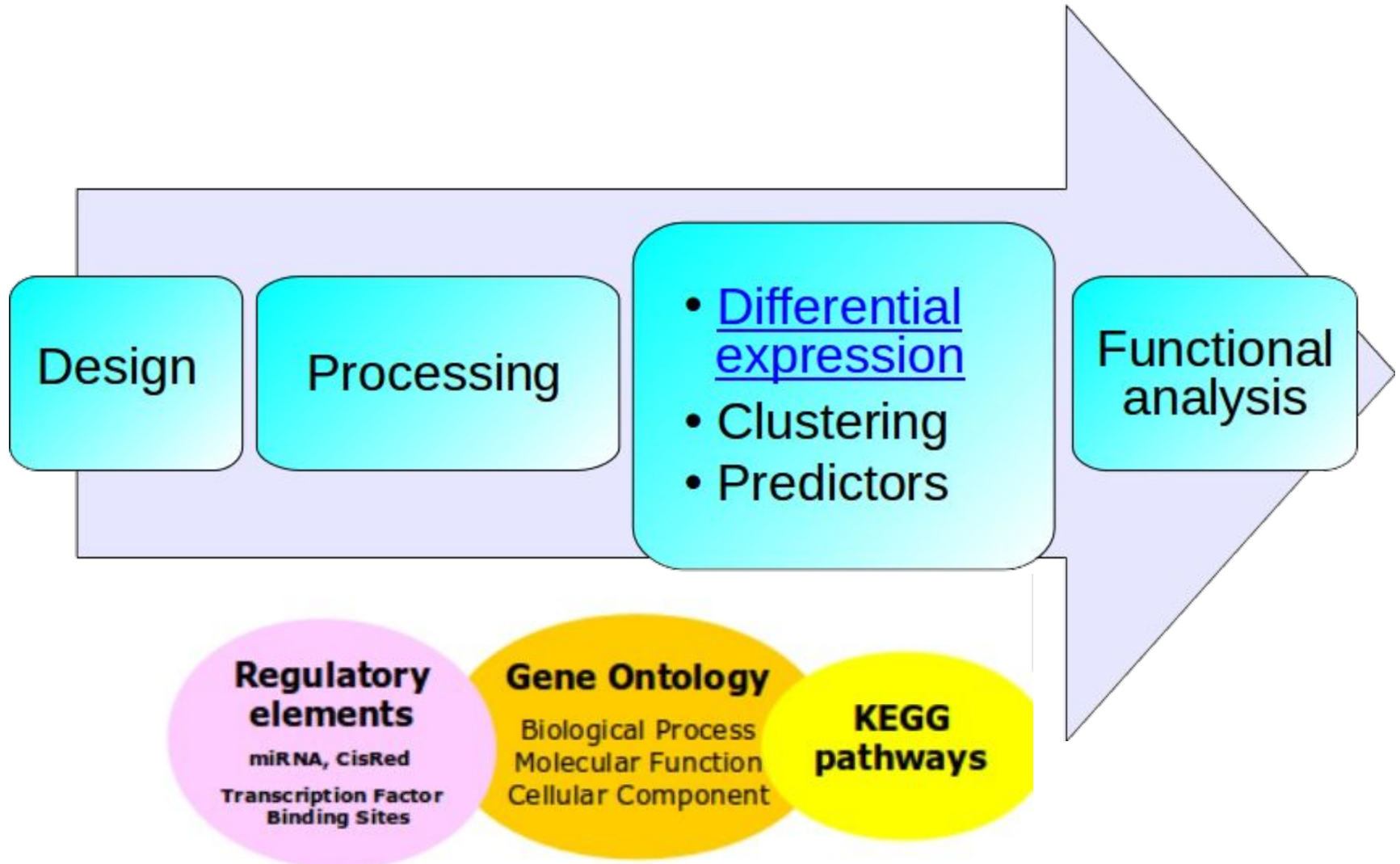
Detecting biomarkers



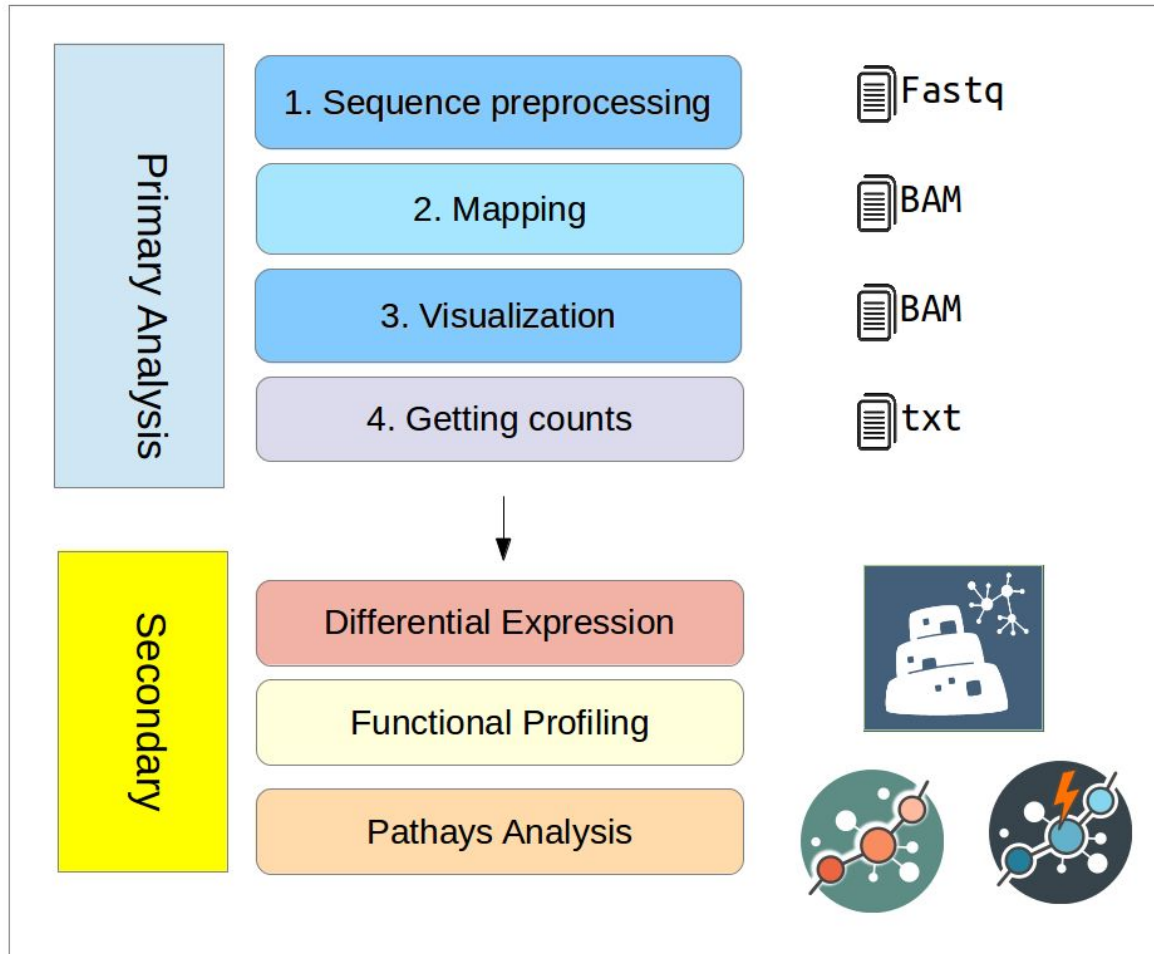
Detecting biomarkers



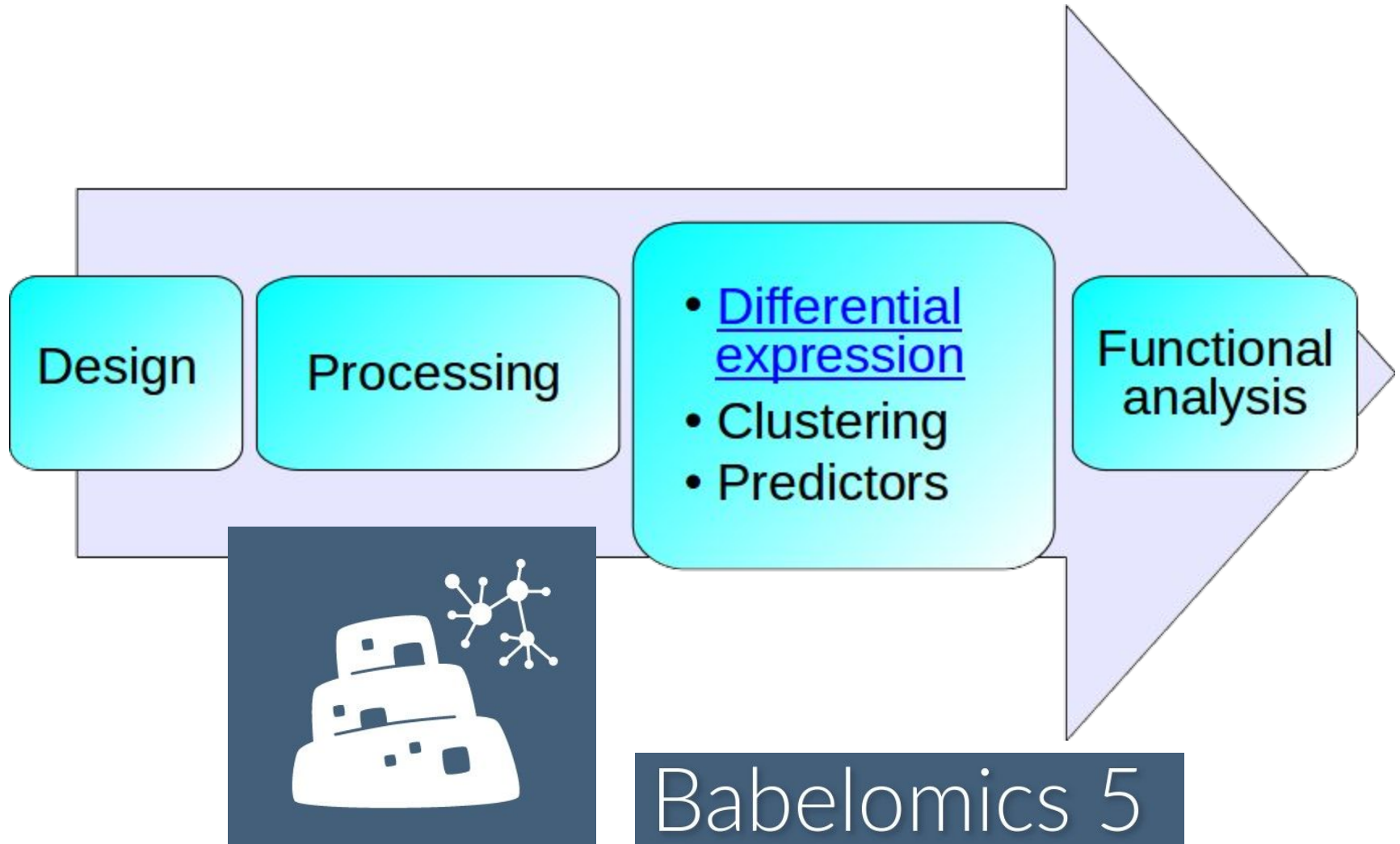
Data analysis workflow



Data analysis workflow



Data analysis workflow



Input

Samples names

Samples

Tab separated file

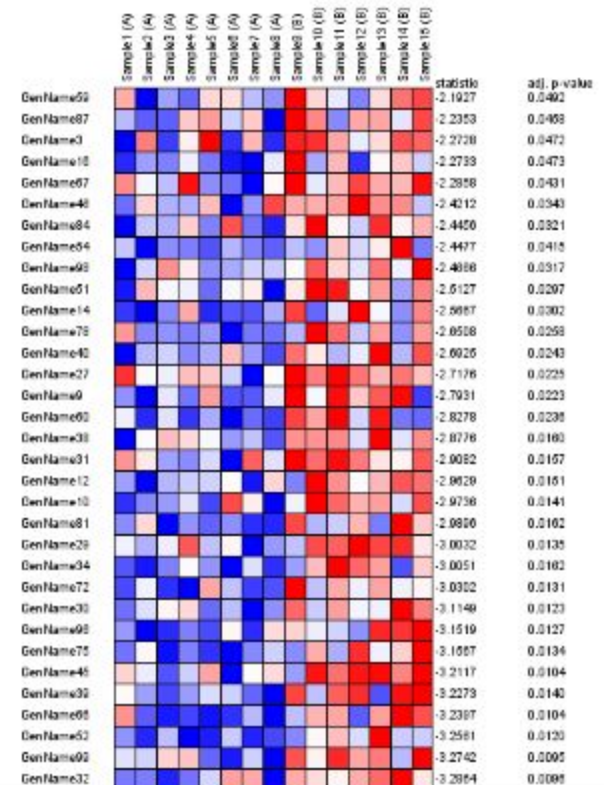
genes

#NAMES	col1	col2	col3	col4	col5	col6	col7
YGR138C	-1.23	-0.81	1.79	0.78	-0.42	-0.69	0.58
YPR156C	-1.76	-0.94	1.16	0.36	0.41	-0.35	1.12
YOR230W	-2.19	0.13	0.65	-0.51	0.52	1.04	0.36
YAL018C	-1.22	-0.98	0.79	-0.76	-0.29	1.54	0.93
YBR287W	-1.47	-0.83	0.85	0.07	-0.81	1.53	0.65
YCL075W	-1.04	-1.11	0.87	-0.14	-0.80	1.74	0.48
YDR055w	-1.57	-1.17	1.29	0.23	-0.20	1.17	0.26
YOR358W	-1.53	-1.25	0.59	-0.30	0.32	1.41	0.77
YBR006W	-1.76	-0.72	0.13	-0.01	-0.23	1.30	1.28
YBR241C	-1.39	-0.42	-0.08	-0.29	-0.65	1.85	0.98
YCR021c	-1.52	-0.99	0.26	0.04	-0.42	1.43	1.19
YCR061W	-1.57	-0.39	0.33	-0.54	-0.51	1.59	1.09
YDL024c	-1.27	-1.14	0.57	-0.30	-0.47	1.46	1.14
YDR298C	-1.49	-0.87	0.41	-0.47	-0.25	1.38	1.29
YER141w	-1.69	-0.60	0.00	0.41	-0.62	1.45	1.05

.....

Results

name	statistic	p-value	adj. p-value
200067_x_at	5.5382	0.0000049746	0.00024376
200052_s_at	5.2111	0.00001452	0.00047431
200054_at	5.1028	0.000042635	0.0010445
200009_at	4.2093	0.00019599	0.0027557
200017_at	4.0805	0.00022496	0.0027557
1053_at	3.9461	0.00060822	0.0059605
200013_at	3.767	0.00070427	0.0062744
200071_at	3.518	0.0014872	0.012146
200076_s_at	3.1376	0.0039127	0.024703
177_at	3.0053	0.0061375	0.030074



Detecting biomarkers



Expression ▾

Unsupervised analysis

- ▶ Clustering

Supervised analysis

- ▶ Class prediction

Differential expression

Microarray

- ▶ Class comparison
- ▶ Correlation
- ▶ Survival

RNA-seq

- ▶ Class comparison

A. **Continuous** variables:

- Metabolomics
- Proteomics
- Transcriptomics arrays
- Experimental data

Different experimental designs



Expression ▾

Unsupervised analysis

- ▶ Clustering

Supervised analysis

- ▶ Class prediction

Differential expression

Microarray

- ▶ Class comparison
 - ▶ Correlation
 - ▶ Survival
- RNA-seq
- ▶ Class comparison

Class comparison

Methods:

Limma, t-test:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

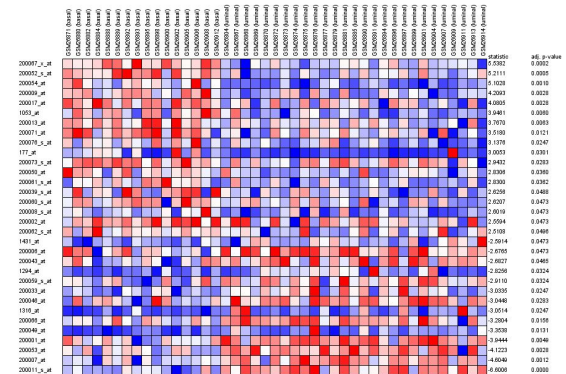
Fold-change:

$$\text{Log}_2 \left(\bar{y}_1 / \bar{y}_2 \right)$$

$$\bar{y}_1 - \bar{y}_2$$

$$H_0: \mu_1 = \mu_2 = \dots = \mu_n$$

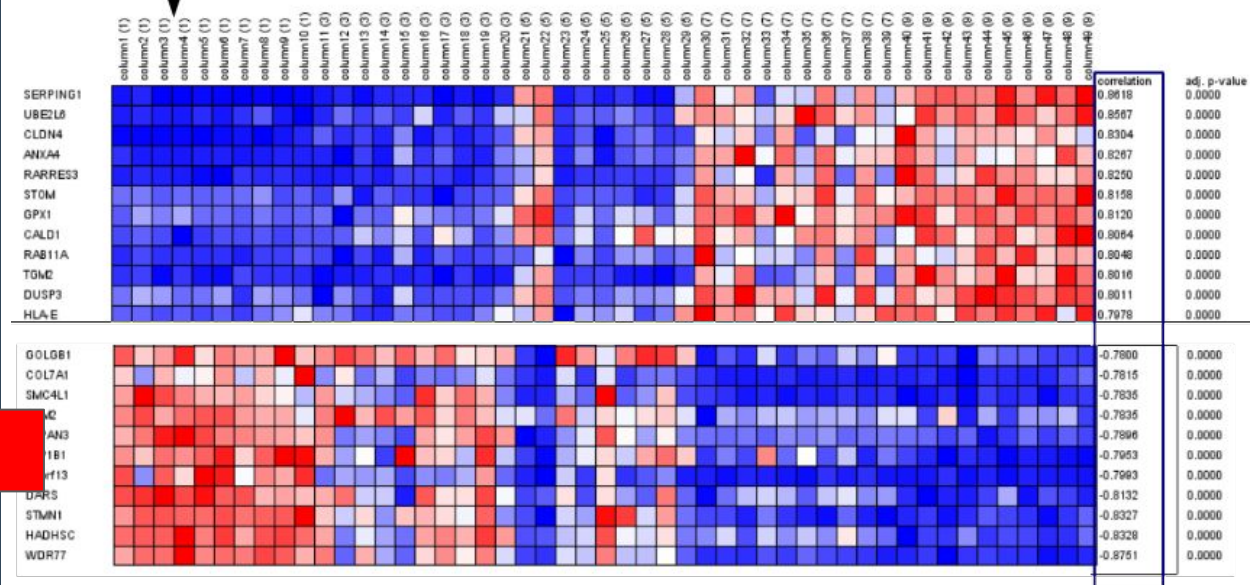
$$H_a: \text{not } H_0$$



Different experimental designs

Correlation

Samples ranked according to the independent variable



Genes ranked by correlation to the continuous variable

Expression ▾



Unsupervised analysis

▸ Clustering

Supervised analysis

▸ Class prediction

Differential expression

Microarray

▸ Class comparison

▸ Correlation

▸ Survival

RNA-seq

▸ Class comparison

Different experimental designs



Expression ▾

Unsupervised analysis

- ▶ Clustering

Supervised analysis

- ▶ Class prediction

Differential expression

Microarray

- ▶ Class comparison

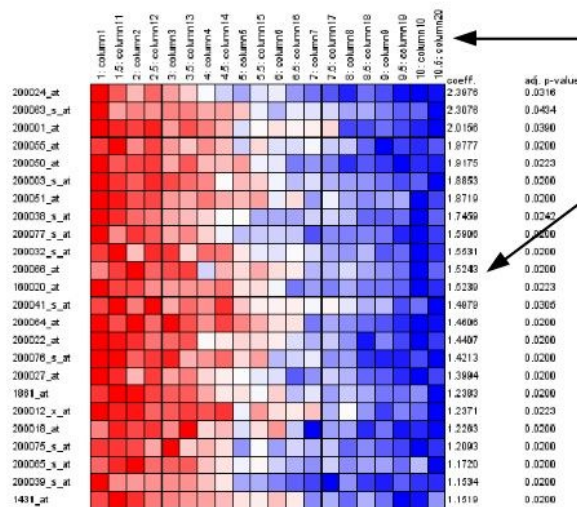
- ▶ Correlation

- ▶ Survival

RNA-seq

- ▶ Class comparison

Survival



Samples ranked according to the survival time

Genes ranked by their relationship with survival time

- ▶ Cox model coefficients
- ▶ Estimate for the statistics
- ▶ p-values

Detecting biomarkers



Expression ▾

Unsupervised analysis

- ▶ Clustering

Supervised analysis

- ▶ Class prediction

Differential expression

Microarray

- ▶ Class comparison
- ▶ Correlation
- ▶ Survival

RNA-seq

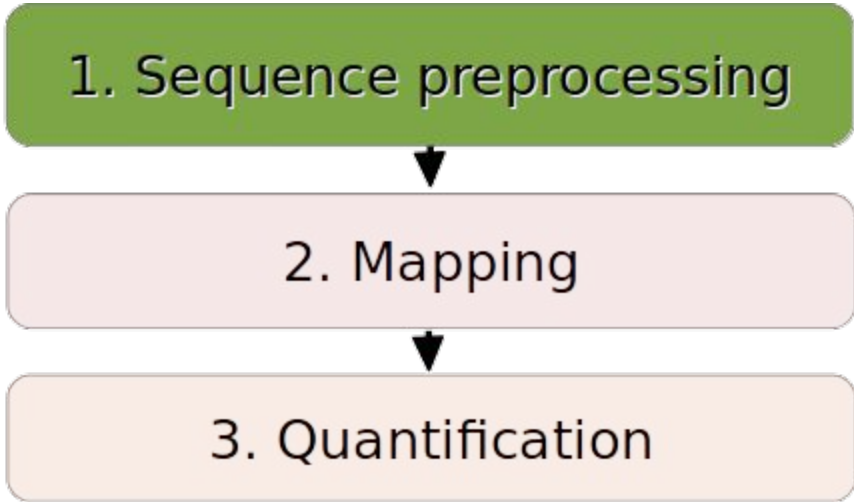
- ▶ Class comparison

B. Discrete variables:

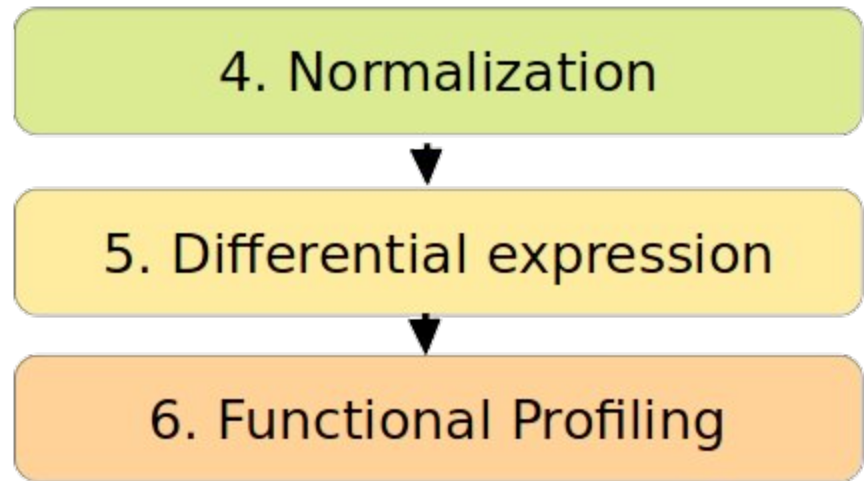
- RNA-Seq
- Experimental data

RNA-Seq Data Analysis Pipeline

Primary



Secondary



Fastq format

- We could say “it is a fasta with **qualities**”:
 - 1. Header (like the fasta but starting with “@”)
 - 2. Sequence (string of nt)
 - 3. “+” and sequence ID (optional)
 - 4. Encoded quality of the sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%++)(%%%),1***-+*'))**55CCF>>>>>CCCCCCC65
```

BAM/SAM format

@PG ID:HPG-Aligner VN:1.0
@SQ SN:20 LN:63025520

HWI-ST700660_138:2:2105:7292:79900#2@0/1 16 20 76703 254 76= * 0 0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
GJJGJLGGFLILGGIEIFEKEDELIGLJIHJFIKKFELFIKLFGLGHKKGJLFIIGKFFEFFEFGKCKFHHCFF AS:i:254 NH:i:1 NM:i:0

HWI-ST700660_138:2:2208:6911:12246#2@0/1 16 20 76703 254 76= * 0 0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
HHJFHLGFFLILEGIKIEEMGEDLIGLHIHJFIKKFELFIKLEFGKGHEKHJLFHIGKFFDFEFGKDKFHHCFF AS:i:254 NH:i:1 NM:i:0

HWI-ST700660_138:2:1201:2973:62218#2@0/1 0 20 76655 254 76M * 0 0
AACCCCAAAAATGTTGGAAGAATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGGACATACTTCT
FEFFGHHHGGHFKCCJKFHIGIFFIFLDEJKGJGGFKIHLFIJGIEGLDEDLFLGEIIMHHIKL\$BBGFFJIEHE AS:i:254 NH:i:1 NM:i:1

HWI-ST700660_138:2:1203:21395:164917#2@0/1 256 20 68253 254 4M1D72M * 0 0
NCACCCATGATAGACCAGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGTGTTCTGTATTCCTCAGGACACAGA
#4@ADEHFJFEJDHJGKEFIHGHGBGFHHFIICEIIFKKIFHEGJEHHGLELEGKJMFGGGLEIKHLFGKIKHDG AS:i:254 NH:i:3 NM:i:1

HWI-ST700660_138:2:1105:16101:50526#6@0/1 16 20 126103 246 53M4D23M * 0 0
AAGAAGTGCAAACCTGAAGAGATGCATGTAAAGAATGGTTGGGCAATGTGCGGCAAAGGGACTGCTGTGTTCCAGC
FEHIGGHIGIGJ6FCFHJIFFLJJCJGJHGFKKKKGJJKHFFKIFFFKHFLKHGKJLJGKILLEFFLIHJIEIB AS:i:368 NH:i:1 NM:i:4

SAM Specification:

<http://samtools.sourceforge.net/SAM1.pdf>

counts file

Gene

Sample



<u>Ensembl</u>	<u>Gene Name</u>	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

Fastq format

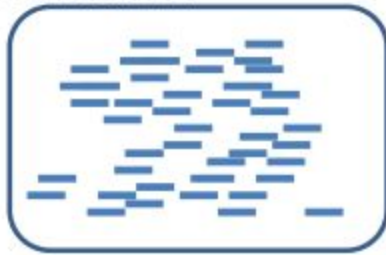
- We could say “it is a fasta with **qualities**”:
 - 1. Header (like the fasta but starting with “@”)
 - 2. Sequence (string of nt)
 - 3. “+” and sequence ID (optional)
 - 4. Encoded quality of the sequence

```
@SEQ_ID  
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT  
+  
!''*(((((***+))%%%++)(%%%),1***-+*'))**55CCF>>>>>CCCCCCC65
```


General context

Sequencing Reads

Individual A



Reference Genome



Sequencing depth

reads

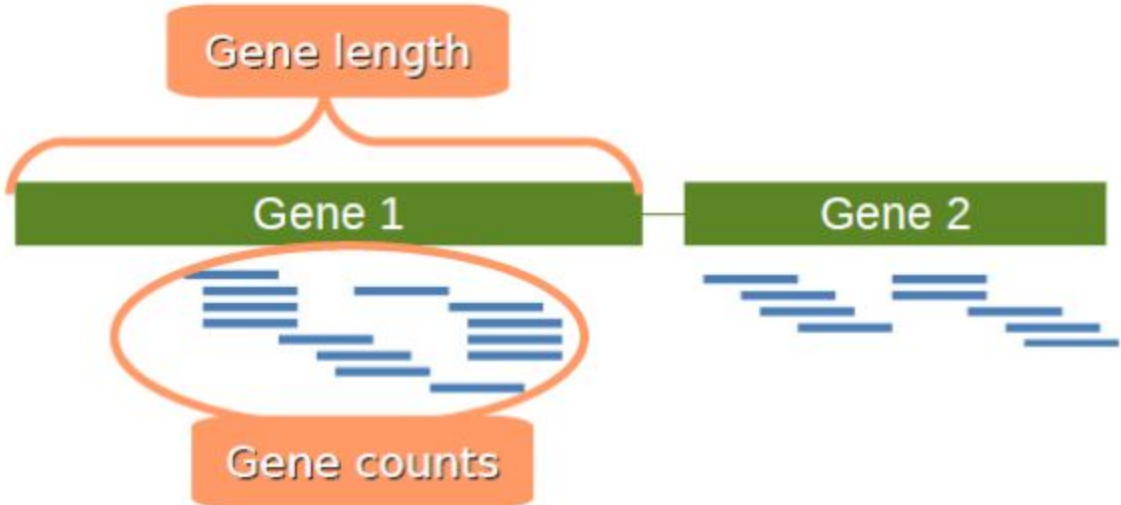


Gene length

Gene 1

Gene 2

Gene counts



Count Normalization

- **Transcript length:** *within* library
- **Library size:** *between* libraries
- Many **other biases** ...
 - Differences on the read count distribution among samples.
 - GC content of the gene affects the detection of that gene (Illumina)
 - sequence-specific bias is introduced during the library preparation

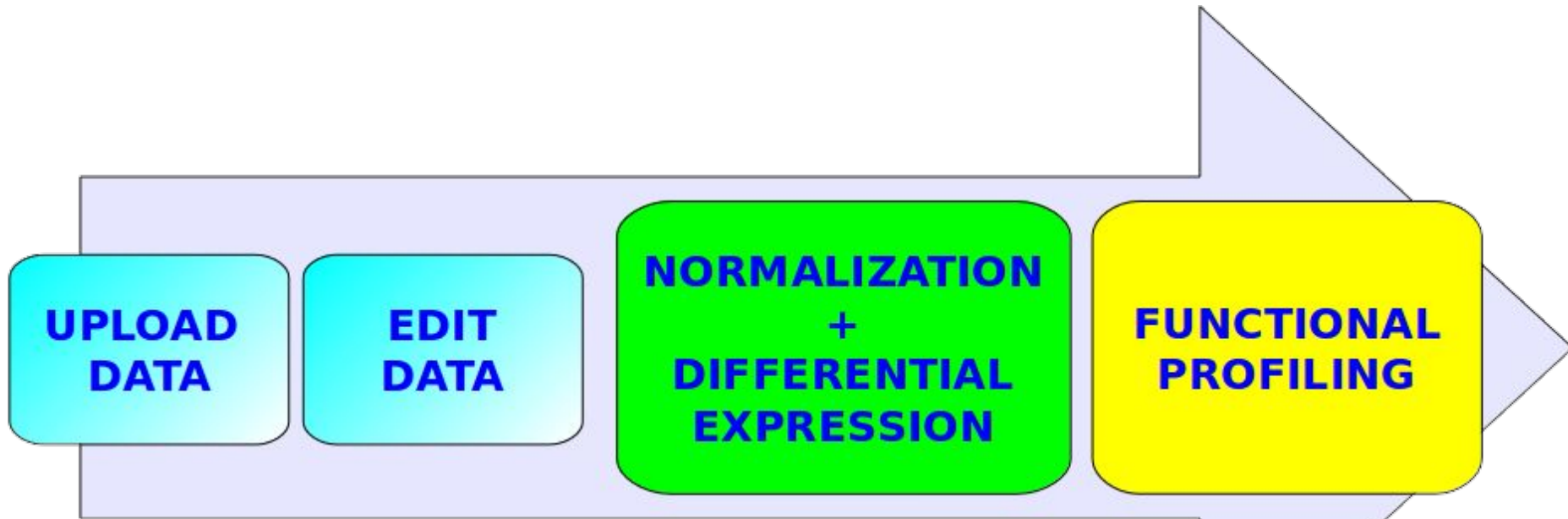
Count Normalization

- **RPKM**: Reads Per Kilobase of the transcript per Million mapped reads

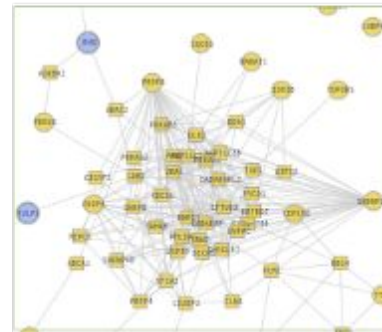
$$RPKM = 10^9 \times \frac{C}{N * L}$$

- **C** is the number of mappable reads mapped onto the gene's exons.
- **N** is the total number of mappable reads in the experiment.
- **L** is the total length of the exons in base pairs.
- Fragments Per Kilobase of exon per Million fragments mapped (FPKM),

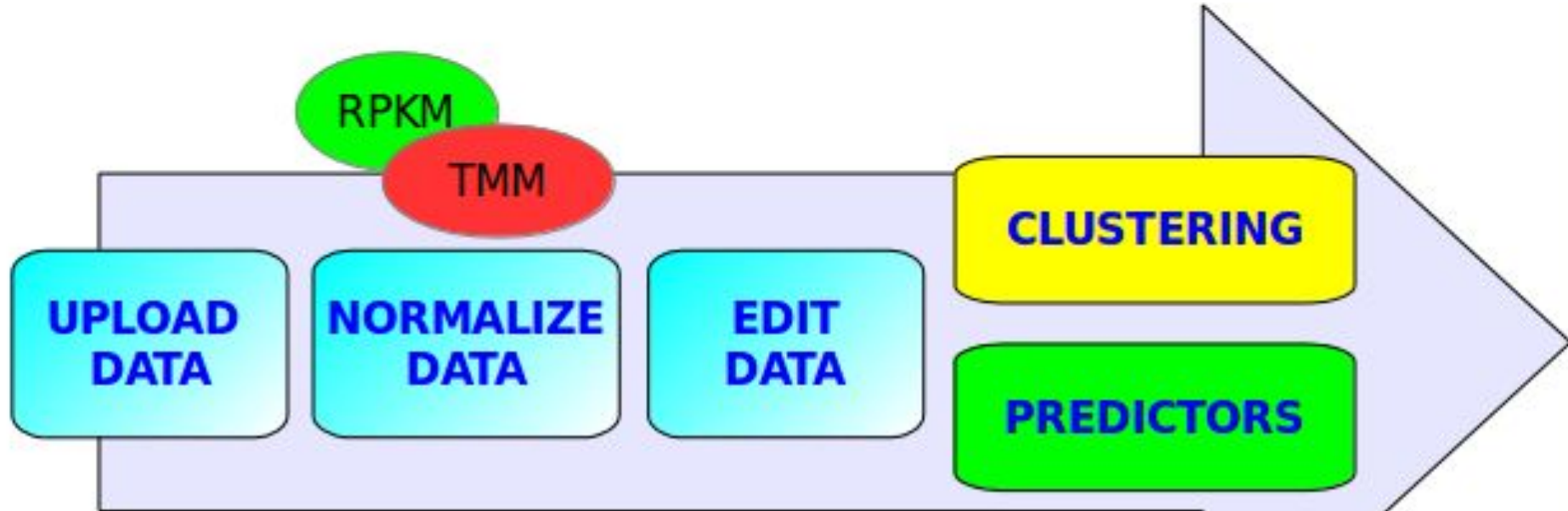
Working in Babelomics



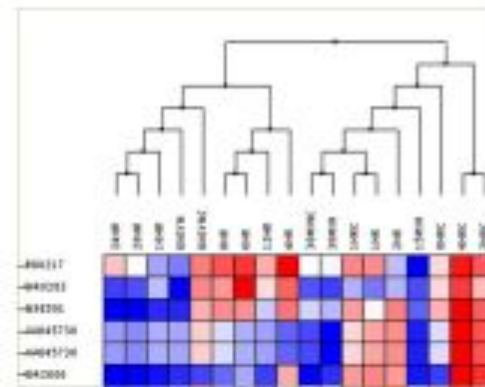
#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN5	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
OPM1	66	05	09	82	00	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C10orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	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	48	6	7	7	8	7



Working in Babelomics



HNAMES	K1	K2	K3	K4	K5	L1	L2	L3	L4	L5
TSPVW6	202	106	194	178	202	157	190	200	200	208
INMD	0	0	0	1	0	0	0	0	0	0
OPM1	66	66	66	62	60	37	30	30	47	40
SCV3	21	30	30	27	31	26	31	37	18	21
Kcom12	10	12	8	11	18	17	20	12	12	18
FGF	18	28	18	30	30	47	50	43	48	48
FUCA2	340	272	260	256	211	70	82	95	68	63
GCLC	90	100	94	84	80	254	262	272	309	326
HPYA	59	61	53	58	59	59	66	63	66	62
STPG5	34	43	45	31	40	0	7	7	8	7



Any question?

