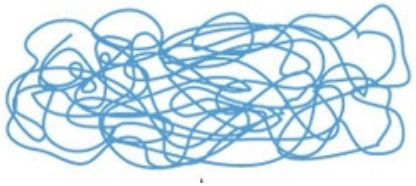
RNA-Seq Data Analysis from Babelomics 5

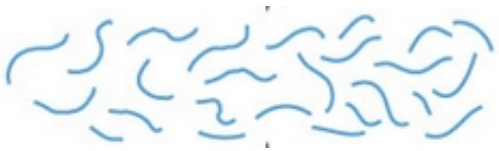
MDA course on NGS Data Analysis
Valencia, 29 Sep 2015

Francisco García
fgarcia@cipf.es

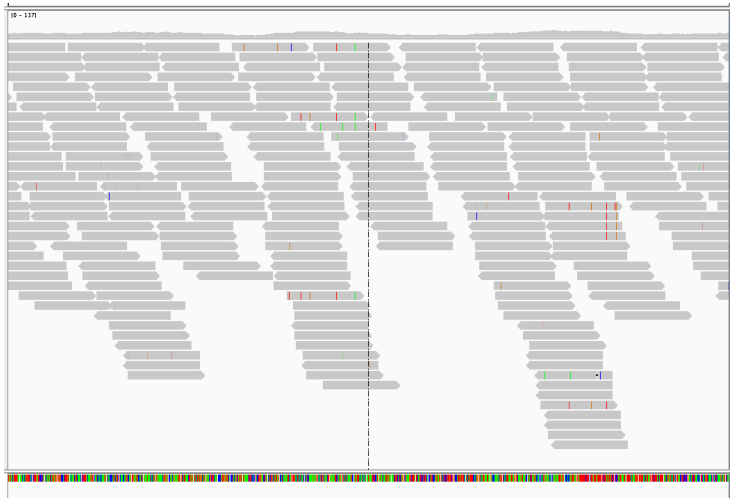
NGS technologies

How do these technologies work ?



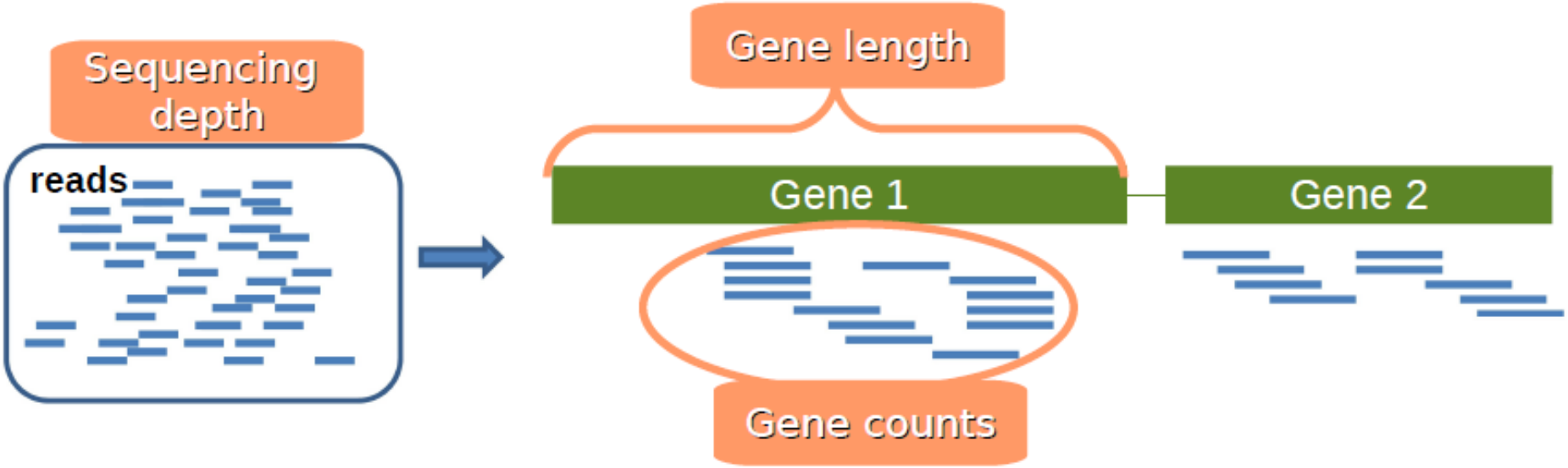


Reference genome



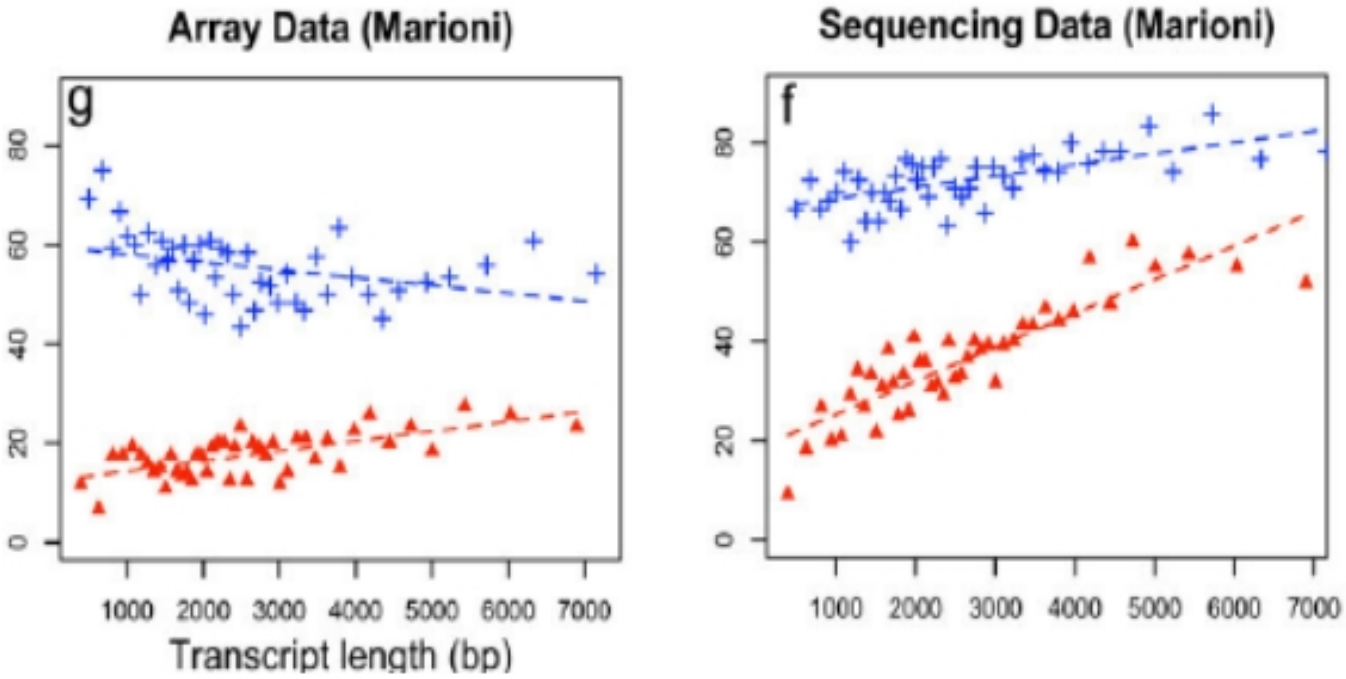
General context

Sequencing Reads



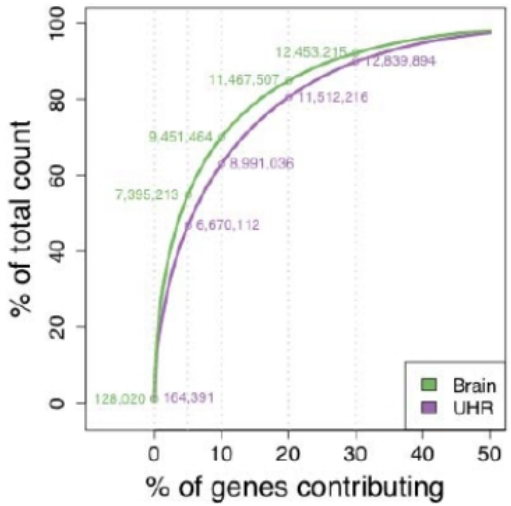
Gene/transcript length dependence

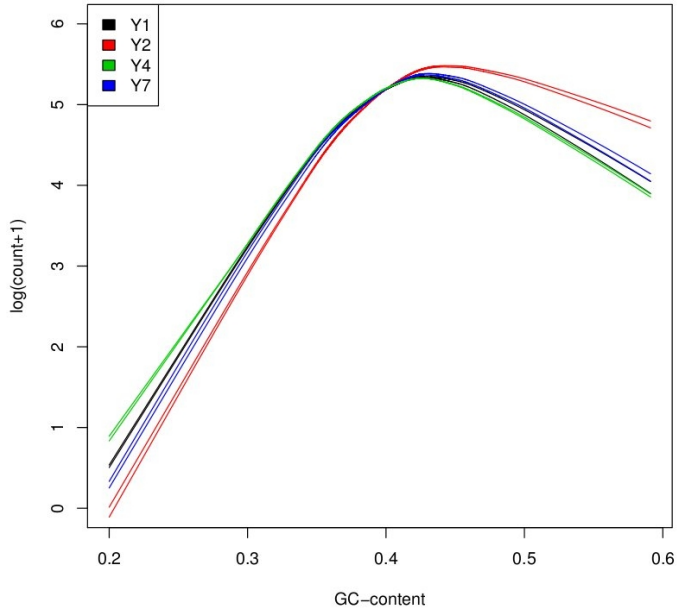
- Counts are proportional to...
  - the transcript length
  - the mRNA expression level.



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),

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
GIJGJLGGFLILGGIEIFEKEDELIGLJIHJFIKKFELFIKLFFGLGHKKGJLFIIGKFFEFFEFGKCKFHHCCCF    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
HHJFHLGFFLILEGIKIEEMGEDLIGLHIHJFIKKFELFIKLEFGKGHEKHJLFHIGKFFDFFEFGKDKFHHCCCF    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
FEFFGHHHGGHFKCCJKFHIGIFFIFLDEJKGJGGFKIHLFIJGIEGFLDEDFLFGEIIMHHIKL$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@ADEHFJFFEJDHJGKEFIHGHBGFHHFIICEIIFFKKIFHEGJEHHGLELEGKJMFGGGLEIKHLFGKIKHDG    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
   FEHIGGHIGIGJI6FCFHJIFFLJJCJGJHGFKKKKGIJKHFFKIFFFKHFLKHGKJLJGKILLEFFLIHJIEIIB    AS:i:368       NH:i:1  NM:i:4
```

SAM Specification:
http://samtools.sourceforge.net/SAM1.pdf

Counts

Gene

Sample

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

RNA-Seq Data Analysis Pipeline

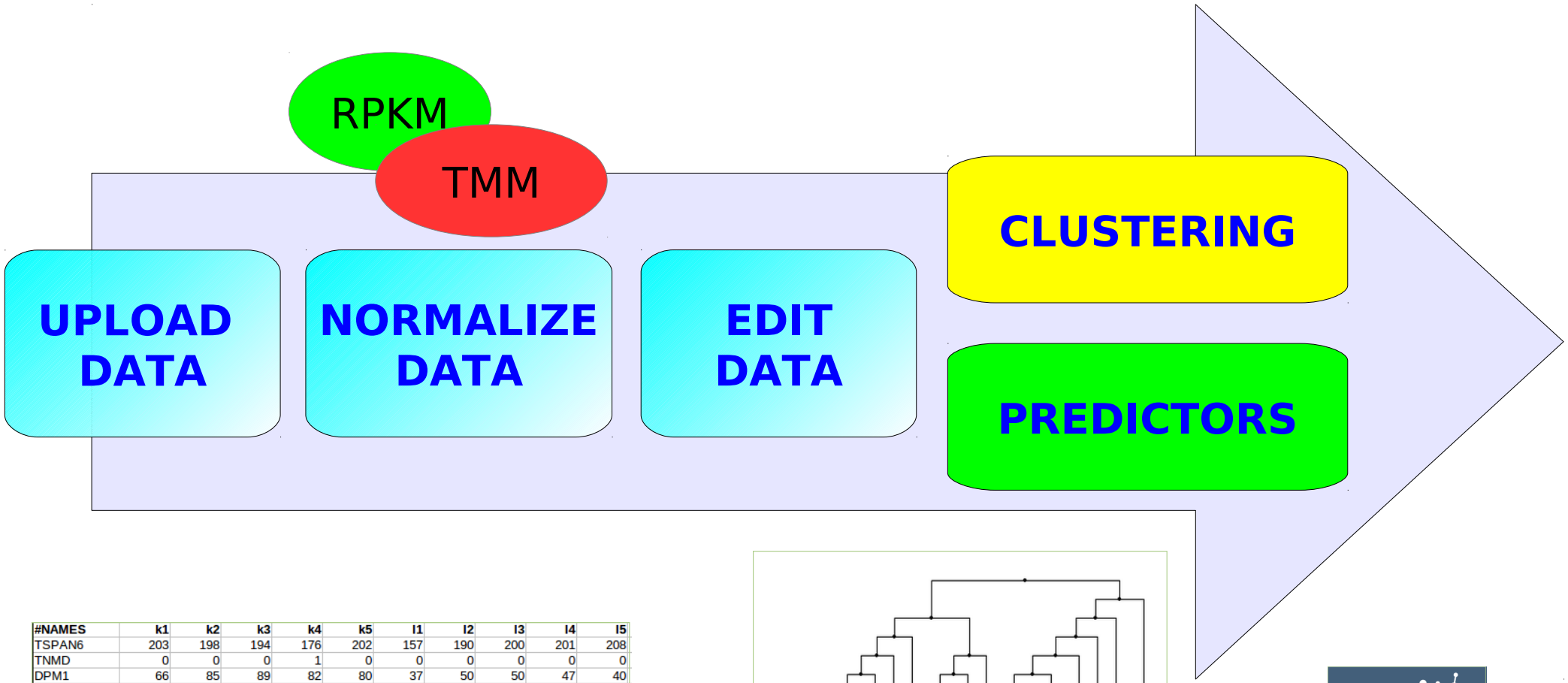
Primary

1. Sequence preprocessing → Fastq
2. Mapping → BAM
3. Quantification → Count matrix

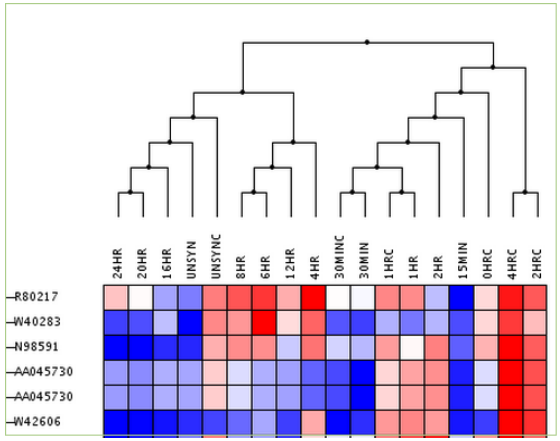
Secondary

4. Normalization
5. Differential expression
6. Functional Profiling

Supervised and Unsupervised Classification



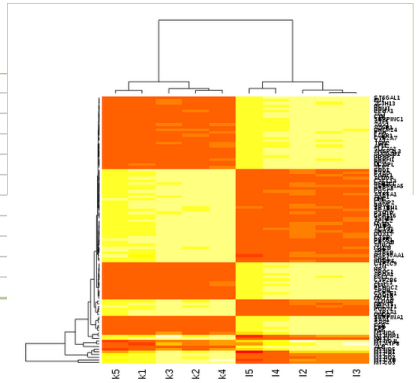
#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	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

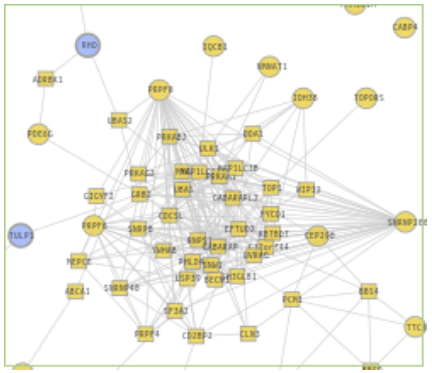


Differential Expression

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





More information

Nucleic Acids Research Advance Access published April 20, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv384

Babelomics 5.0: functional interpretation for new generations of genomic data

Roberto Alonso[1,2], Francisco Salavert[1,3], Francisco Garcia-Garcia[1], Jose Carbonell-Caballero[1], Marta Bleda[4], Luz Garcia-Alonso[1], Alba Sanchis-Juan[5], Daniel Perez-Gil[5], Pablo Marin-Garcia[5], Ruben Sanchez[1,6], Cankut Cubuk[1], Marta R. Hidalgo[1], Alicia Amadoz[1], Rosa D. Hernansaiz-Ballesteros[1], Alejandro Alemán[1,3], Joaquin Tarraga[1], David Montaner[1], Ignacio Medina[7] and Joaquin Dopazo[1,2,3,6,*]

Babelomics tutorial:

http://babelomics.bioinfo.cipf.es/