Babelomics

Microarray Data Analysis
Differential Gene Expression

Valencia, March 2011

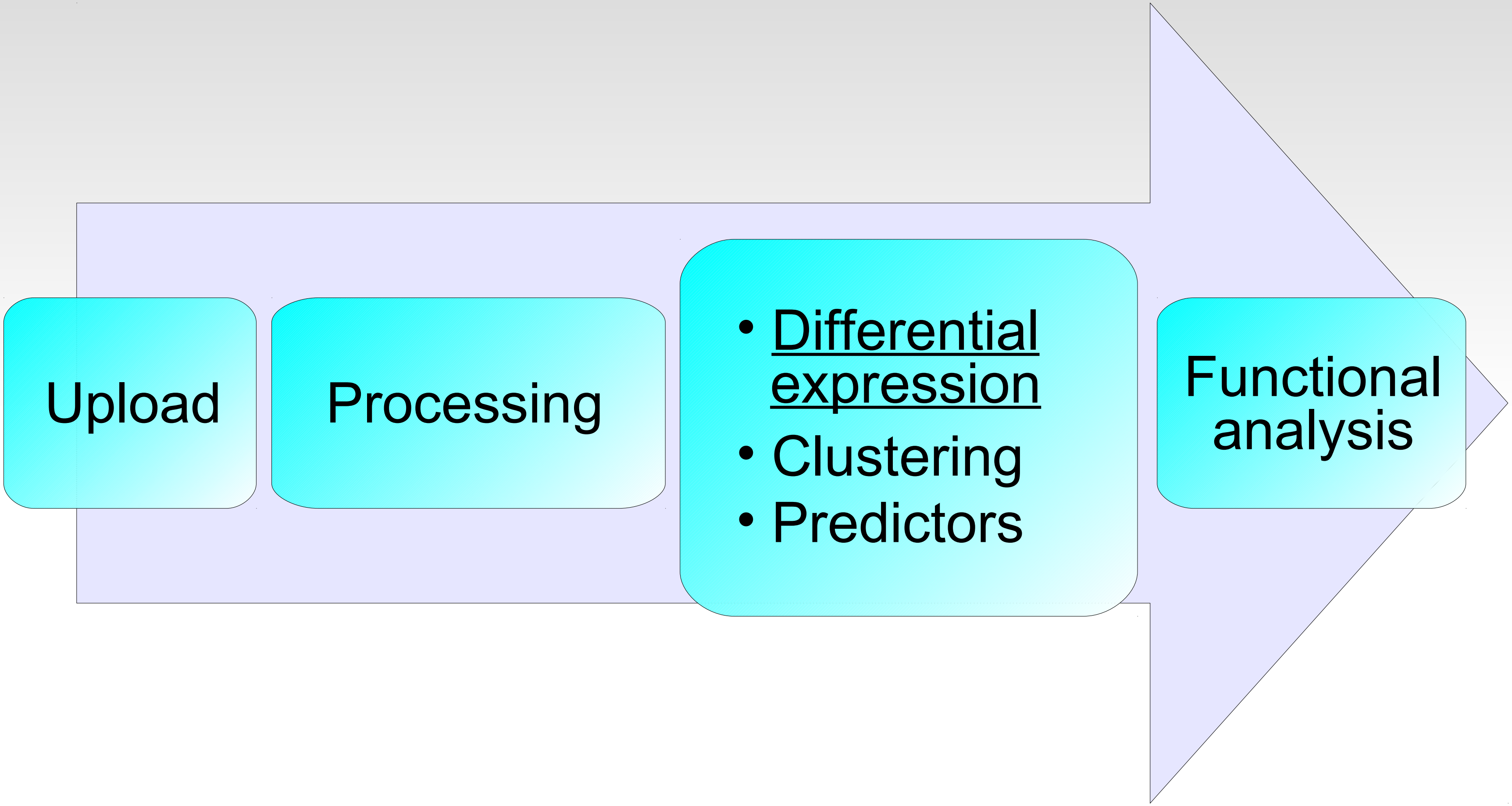
Francisco García, fgarcia@cipf.es
Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)

INB

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

er ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

Data analysis workflow



Upload

Processing

- Differential expression
- Clustering
- Predictors

Functional analysis

Differential Expression

Class comparison

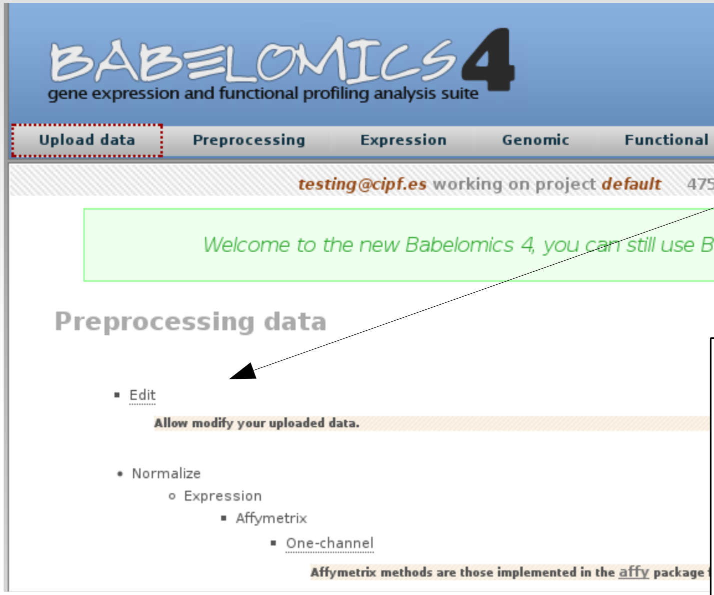
Correlation

Survival

Time / dosage series

Input

1. After normalization, we will add information relative to experimental design



testing@cipf.es working on project default 475.71 Mb of 1.00 Gb (46.46%)

Welcome to the new Babelomics 4, you can still use Babelomics 3 at: http://babe

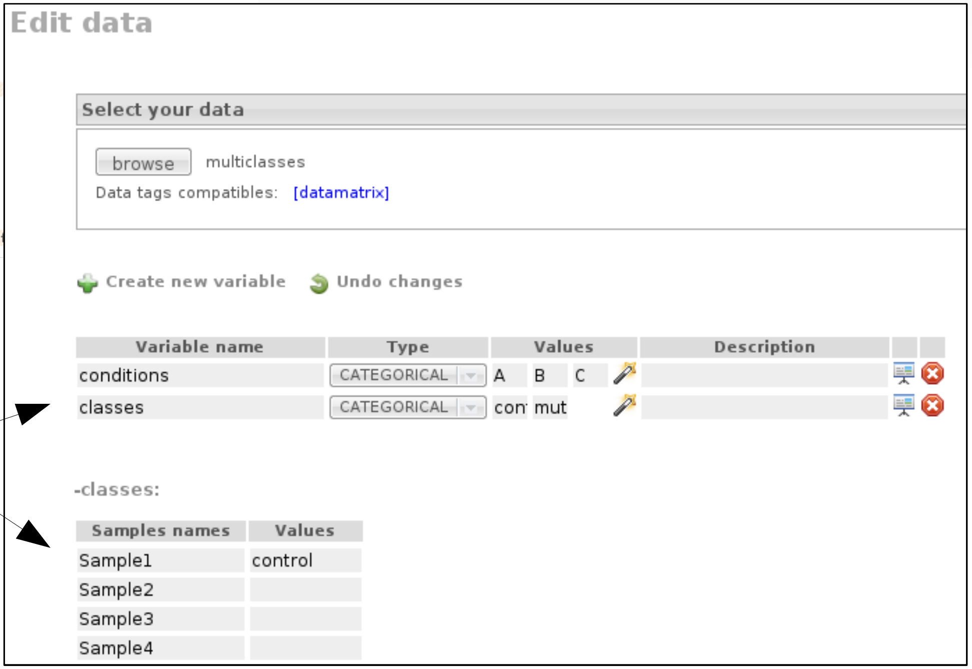
Preprocessing data

- Edit

  Allow modify your uploaded data.

- Normalize
  - Expression
    - Affymetrix
      - One-channel

        Affymetrix methods are those implemented in the affy package



Edit data

Select your data

multiclasses

Data tags compatibles: [datamatrix]

Variable name	Type	Values	Description
conditions	CATEGORICAL	A B C	
classes	CATEGORICAL	con mut	

-classes:

Samples names	Values
Sample1	control
Sample2	
Sample3	
Sample4	

2. Assigning values of variables to each array

Input

Array names

Arrays

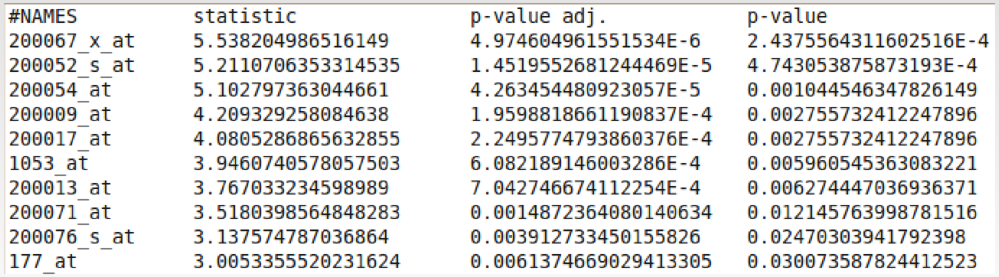
Tab separated file

genes

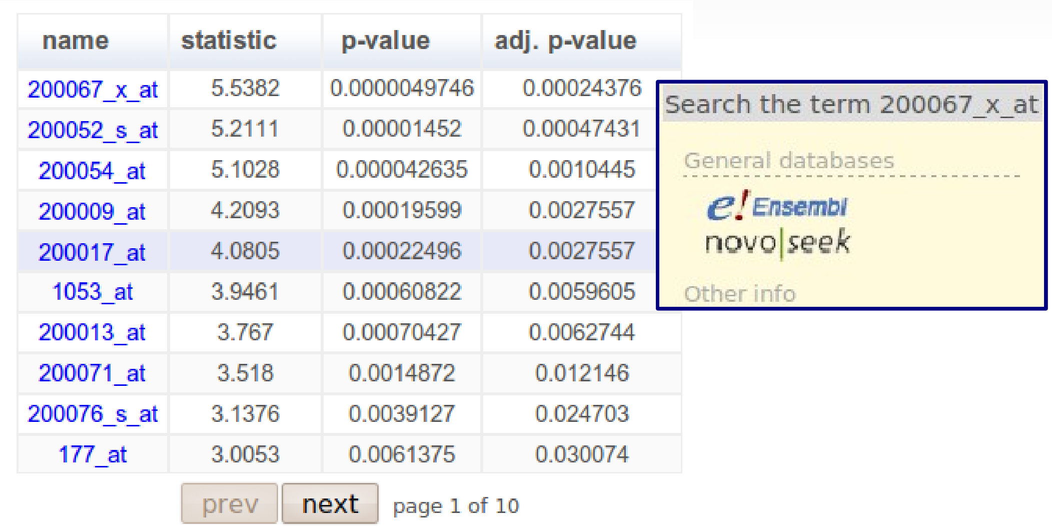
#NAMES	**col1**	**col2**	**col3**	**col4**	**col5**	**col6**	**col7**
YGR138C	-1.23	-0.81	1.79	0.78	-0.42	-0.69	0.58
YPR156C	-1.76	-0.94	1.16	0.36	0.41	-0.35	1.12
YOR230W	-2.19	0.13	0.65	-0.51	0.52	1.04	0.36
YAL018C	-1.22	-0.98	0.79	-0.76	-0.29	1.54	0.93
YBR287W	-1.47	-0.83	0.85	0.07	-0.81	1.53	0.65
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

......

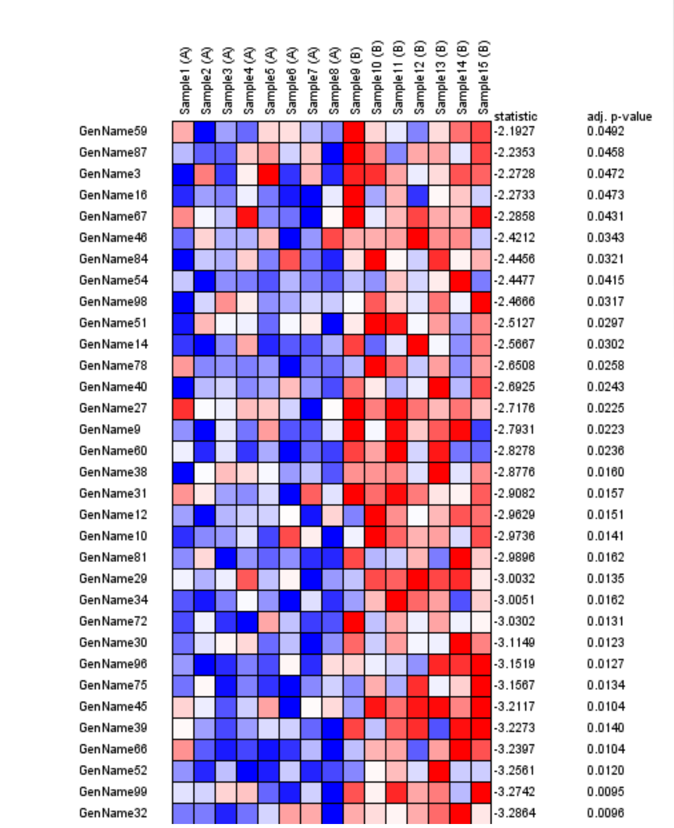
Results



```
#NAMES          statistic                p-value adj.             p-value
200067_x_at     5.538204986516149        4.974604961551534E-6     2.4375564311602516E-4
200052_s_at     5.2110706353314535       1.4519552681244469E-5    4.743053875873193E-4
200054_at       5.102797363044661        4.263454480923057E-5     0.001044546347826149
200009_at       4.209329258084638        1.9598818661190837E-4    0.002755732412247896
200017_at       4.0805286865632855       2.2495774793860376E-4    0.002755732412247896
1053_at         3.9460740578057503       6.082189146003286E-4     0.005960545363083221
200013_at       3.767033234598989        7.042746674112254E-4     0.006274447036936371
200071_at       3.5180398564848283       0.0014872364080140634    0.012145763998781516
200076_s_at     3.137574787036864        0.003912733450155826     0.02470303941792398
177_at          3.0053355520231624       0.0061374669029413305   0.030073587824412523
```

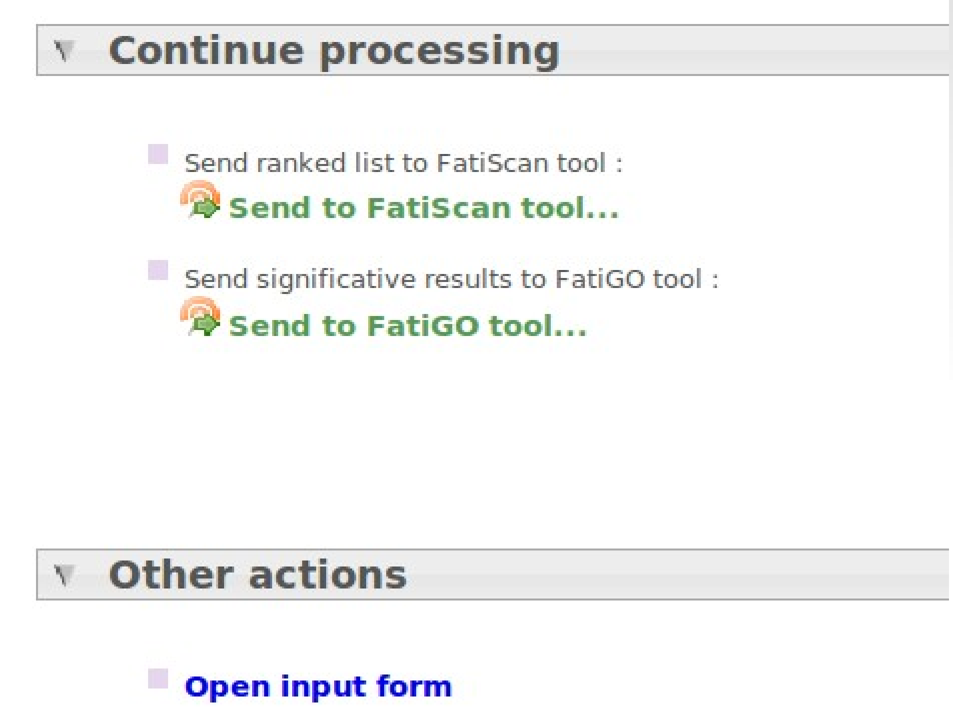


name	statistic	p-value	adj. p-value
200067_x_at	5.5382	0.0000049746	0.00024376
200052_s_at	5.2111	0.00001452	0.00047431
200054_at	5.1028	0.000042635	0.0010445
200009_at	4.2093	0.00019599	0.0027557
200017_at	4.0805	0.00022496	0.0027557
1053_at	3.9461	0.00060822	0.0059605
200013_at	3.767	0.00070427	0.0062744
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



	statistic	adj. p-value
GenName59	-2.1927	0.0492
GenName87	-2.2353	0.0458
GenName3	-2.2728	0.0472
GenName16	-2.2733	0.0473
GenName67	-2.2858	0.0431
GenName46	-2.4212	0.0343
GenName84	-2.4456	0.0321
GenName54	-2.4477	0.0415
GenName98	-2.4666	0.0317
GenName51	-2.5127	0.0297
GenName14	-2.5667	0.0302
GenName78	-2.6508	0.0258
GenName40	-2.6925	0.0243
GenName27	-2.7176	0.0225
GenName9	-2.7931	0.0223
GenName60	-2.8278	0.0236
GenName38	-2.8776	0.0160
GenName31	-2.9082	0.0157
GenName12	-2.9629	0.0151
GenName10	-2.9736	0.0141
GenName81	-2.9896	0.0162
GenName29	-3.0032	0.0135
GenName34	-3.0051	0.0162
GenName72	-3.0302	0.0131
GenName30	-3.1149	0.0123
GenName96	-3.1519	0.0127
GenName75	-3.1567	0.0134
GenName45	-3.2117	0.0104
GenName39	-3.2273	0.0140
GenName66	-3.2397	0.0104
GenName52	-3.2561	0.0120
GenName99	-3.2742	0.0095
GenName32	-3.2864	0.0096

Results



Continue processing

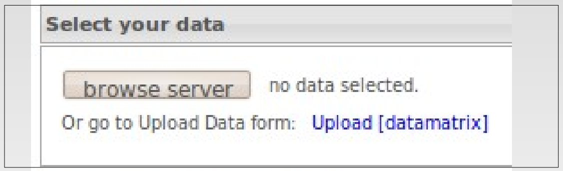
- Send ranked list to FatiScan tool :
  Send to FatiScan tool...
- Send significative results to FatiGO tool :
  Send to FatiGO tool...

Other actions

- **Open input form**

Different experimental designs

Class comparison

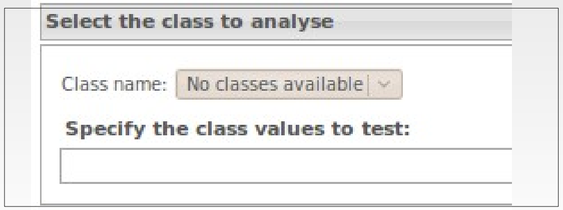


Select your data

browse server   no data selected.

Or go to Upload Data form: Upload [datamatrix]

1. data

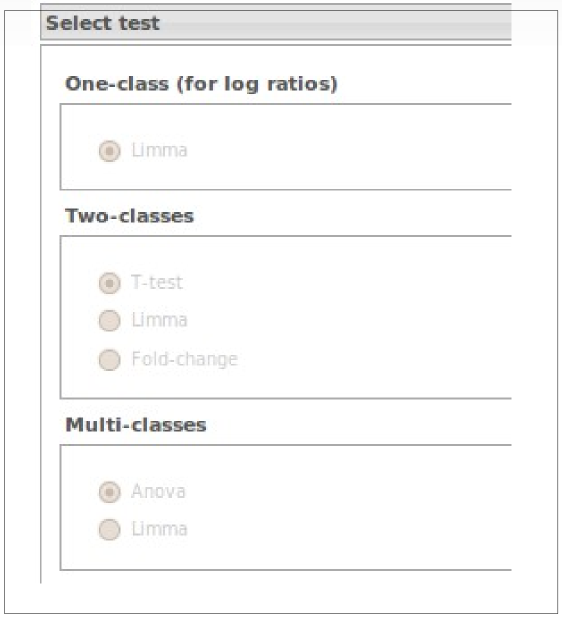


Select the class to analyse

Class name: No classes available

Specify the class values to test:

2. classes



Select test

One-class (for log ratios)

● Limma

Two-classes

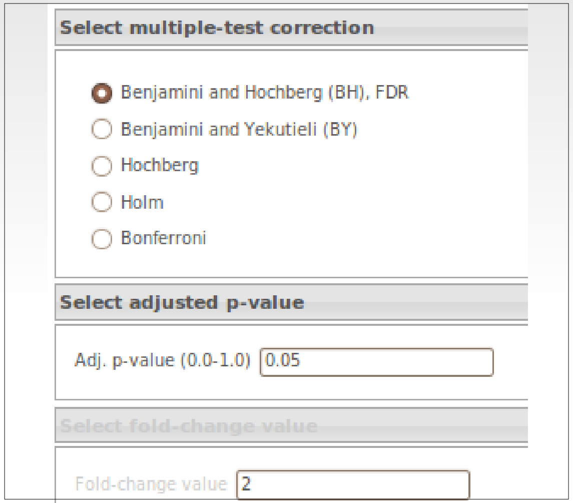
● T-test
○ Limma
○ Fold-change

Multi-classes

● Anova
○ Limma

3. methods

4. adj. p-value



Select multiple-test correction

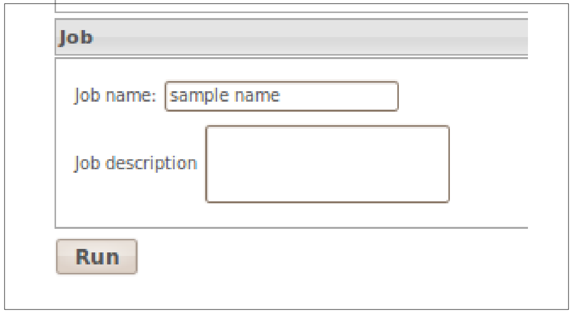
● Benjamini and Hochberg (BH), FDR
○ Benjamini and Yekutieli (BY)
○ Hochberg
○ Holm
○ Bonferroni

Select adjusted p-value

Adj. p-value (0.0-1.0) 0.05

Select fold-change value

Fold-change value 2



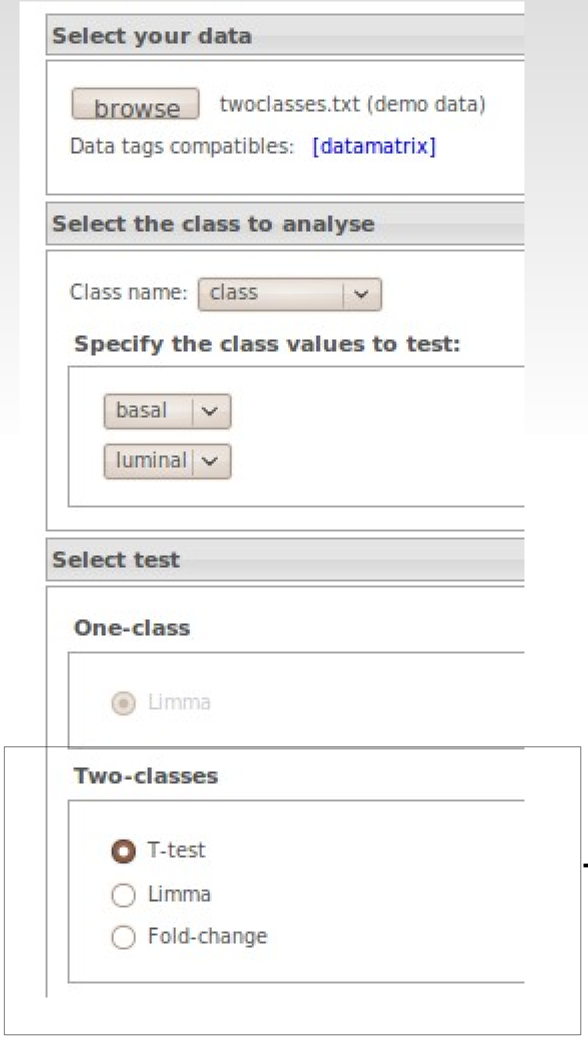
Job

Job name: sample name

Job description

5. job name

Two-classes *form*



Select your data

browse twoclasses.txt (demo data)

Data tags compatibles: [datamatrix]

Select the class to analyse

Class name: class

Specify the class values to test:

basal

luminal

Select test

One-class

● Limma

Two-classes

● T-test

○ Limma

○ Fold-change

Methods:

Limma, t-test:

$$H_o: \ \mu_1 = \mu_2$$
$$H_a: \ \mu_1 \neq \mu_2$$

Fold-change:

$$\mathrm{Log}_2 \left(\overline{y}_1 \ / \ \overline{y}_2 \right)$$

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

t – test for a gene expression

For each gene, we check if its mean expression is equal or different across the **two** classes

H_0: $\mu_1 = \mu_2$

Null hypothesis: the mean expression is **equal** n both groups.

H_a: $\mu_1 \neq \mu_2$

Alternative hypothesis: the mean expression is **different** between the groups.

Mean in group 1

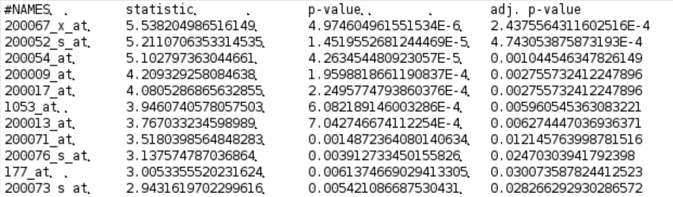
Mean in group 2

Test Statistic: $T = \dfrac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$

Estimation of the variability of the differences

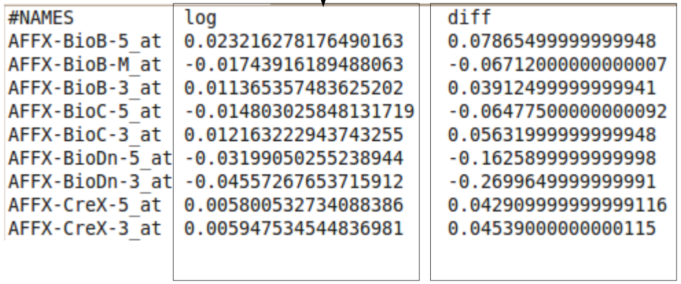
Two-classes *results*

Limma, t-test

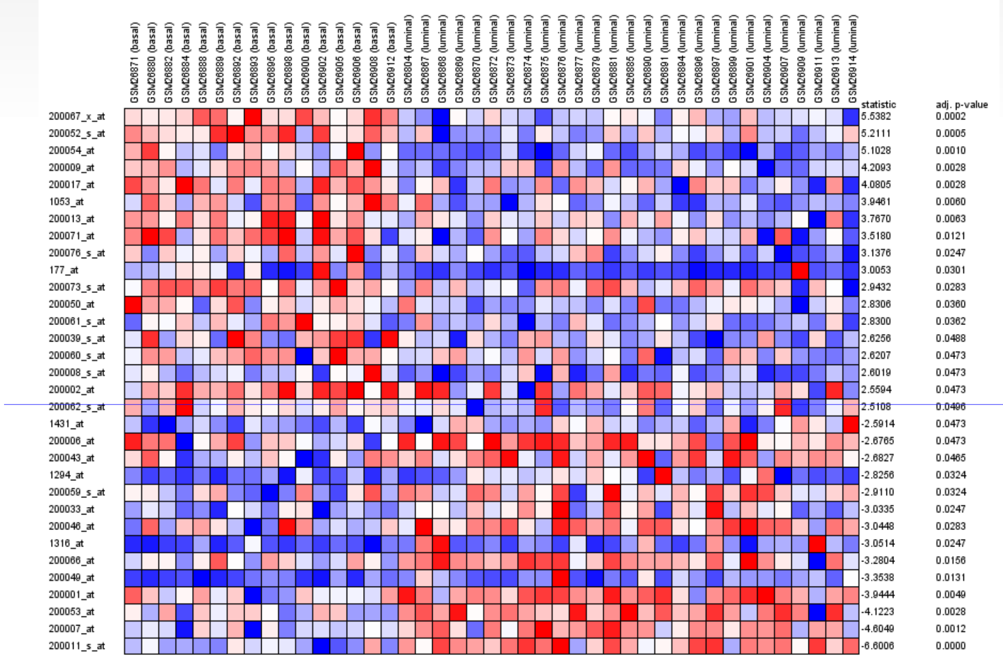


```
#NAMES          statistic             p-value                 adj. p-value
200067_x_at     5.538204986516149     4.974604961551534E-6    2.4375564311602516E-4
200052_s_at     5.2110706353314535    1.4519552681244469E-5   4.743053875873193E-4
200054_at       5.10279736304466l     4.263454480923057E-5    0.001044546347826149
200009_at       4.209329258084638     1.9598818661190837E-4   0.002755732412247896
200017_at       4.0805286865632855    2.2495774793860376E-4   0.002755732412247896
1053_at         3.9460740578057503    6.082189146003286E-4    0.005960545363083221
200013_at       3.767033234598989     7.042746674112254E-4    0.006274447036936371
200071_at       3.5180398564848283    0.0014872364080140634   0.012145763998781516
200076_s_at     3.137574787036864     0.003912733450155826    0.02470303941792398
177_at          3.0053355520231624    0.0061374669029413305   0.030073587824412523
200073 s at     2.9431619702299616    0.00542108668753043l    0.028266292930286572
```

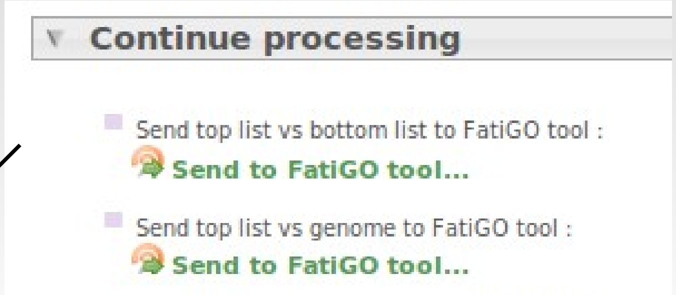
Fold-change



```
#NAMES          log                    diff
AFFX-BioB-5_at  0.023216278176490163   0.07865499999999948
AFFX-BioB-M_at  -0.01743916189488063   -0.06712000000000007
AFFX-BioB-3_at  0.011365357483625202   0.03912499999999941
AFFX-BioC-5_at  -0.014803025848131719  -0.06477500000000092
AFFX-BioC-3_at  0.012163222943743255   0.05631999999999948
AFFX-BioDn-5_at -0.03199050255238944   -0.16258999999999998
AFFX-BioDn-3_at -0.04557267653715912   -0.26996499999999991
AFFX-CreX-5_at  0.005800532734088386   0.042909999999999116
AFFX-CreX-3_at  0.005947534544836981   0.04539000000000115
```

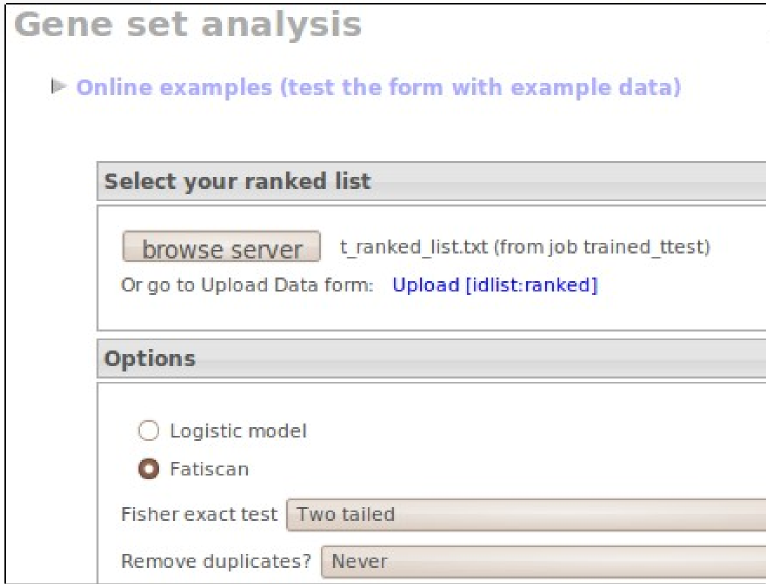


Two-classes *results*



Continue processing

- Send top list vs bottom list to FatiGO tool :
  Send to FatiGO tool...
- Send top list vs genome to FatiGO tool :
  Send to FatiGO tool...



Gene set analysis

▸ Online examples (test the form with example data)

Select your ranked list

browse server t_ranked_list.txt (from job trained_ttest)

Or go to Upload Data form: Upload [idlist:ranked]

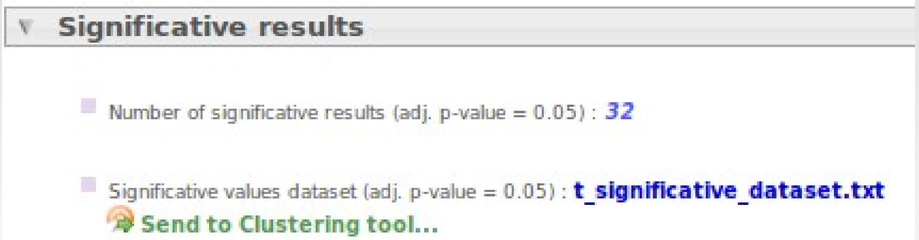
Options

○ Logistic model

● Fatiscan

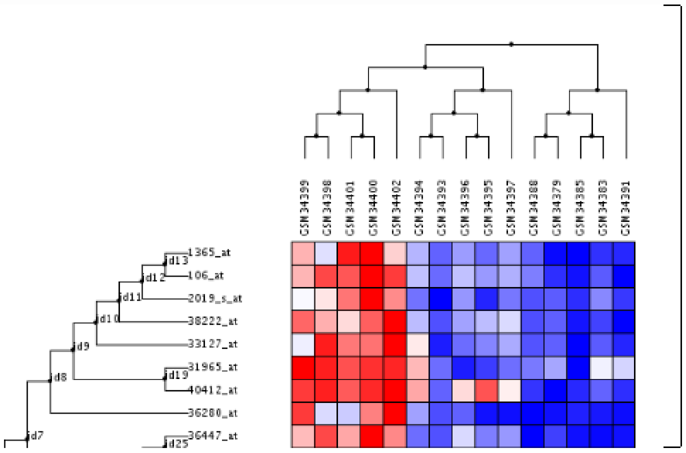
Fisher exact test: Two tailed

Remove duplicates?: Never

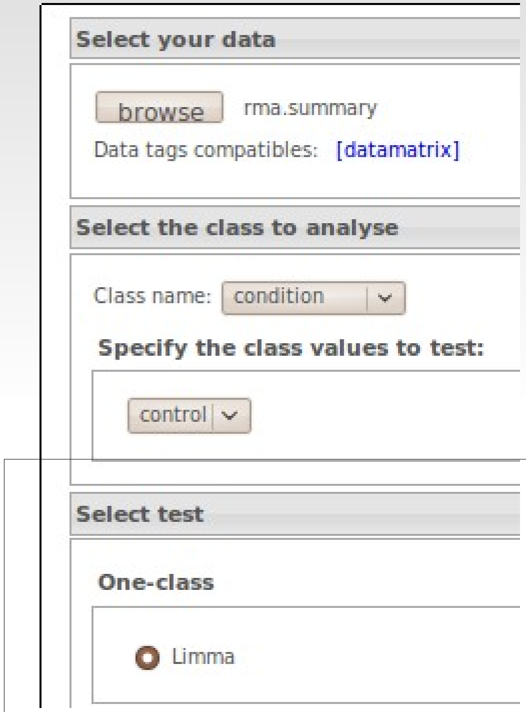


Significative results

- Number of significative results (adj. p-value = 0.05) : *32*
- Significative values dataset (adj. p-value = 0.05) : **t_significative_dataset.txt**
  Send to Clustering tool...



One-class



Select your data

rma.summary

Data tags compatibles: [datamatrix]

Select the class to analyse

Class name: condition

Specify the class values to test:

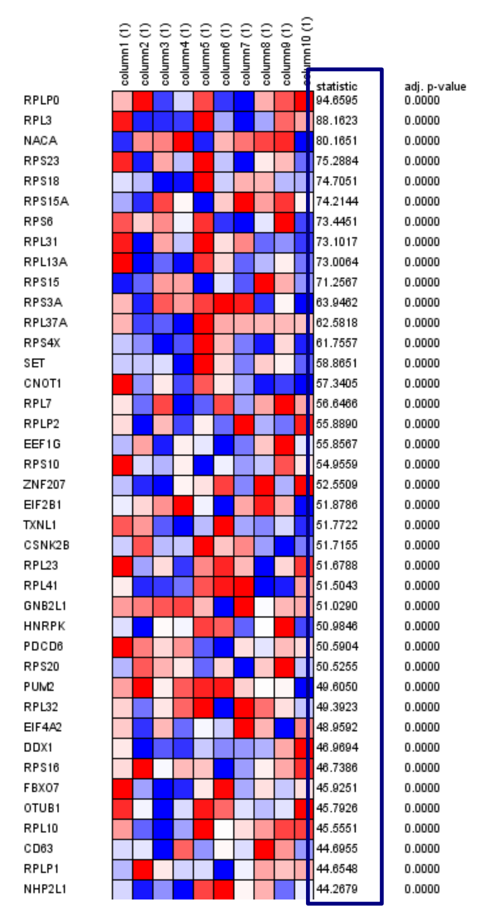
control

Select test

One-class

● Limma

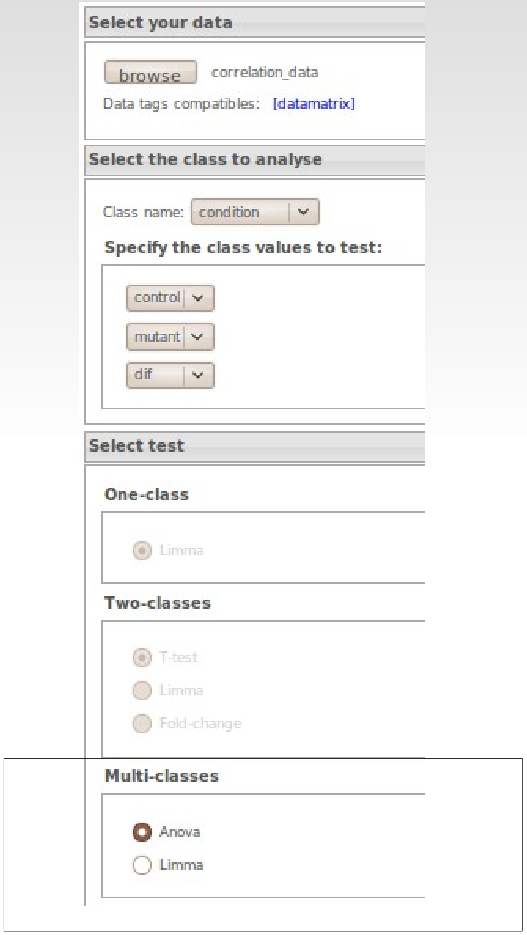
H_o: $\mu = 0$
H_a: $\mu \neq 0$



	statistic	adj. p-value
RPLP0	94.6595	0.0000
RPL3	88.1623	0.0000
NACA	80.1651	0.0000
RPS23	75.2884	0.0000
RPS18	74.7051	0.0000
RPS15A	74.2144	0.0000
RPS6	73.4451	0.0000
RPL31	73.1017	0.0000
RPL13A	73.0064	0.0000
RPS15	71.2567	0.0000
RPS3A	63.9462	0.0000
RPL37A	62.5818	0.0000
RPS4X	61.7557	0.0000
SET	58.8651	0.0000
CNOT1	57.3405	0.0000
RPL7	56.6466	0.0000
RPLP2	55.8890	0.0000
EEF1G	55.8567	0.0000
RPS10	54.9559	0.0000
ZNF207	52.5509	0.0000
EIF2B1	51.8786	0.0000
TXNL1	51.7722	0.0000
CSNK2B	51.7155	0.0000
RPL23	51.6788	0.0000
RPL41	51.5043	0.0000
GNB2L1	51.0290	0.0000
HNRPK	50.9846	0.0000
PDCD6	50.5904	0.0000
RPS20	50.5255	0.0000
PUM2	49.6050	0.0000
RPL32	49.3923	0.0000
EIF4A2	48.9592	0.0000
DDX1	46.9694	0.0000
RPS16	46.7386	0.0000
FBXO7	45.9251	0.0000
OTUB1	45.7926	0.0000
RPL10	45.5551	0.0000
CD63	44.6955	0.0000
RPLP1	44.6548	0.0000
NHP2L1	44.2679	0.0000

Genes ordered by statistic

Multi-classes



Select your data

browse correlation_data

Data tags compatibles: [datamatrix]

Select the class to analyse

Class name: condition

Specify the class values to test:

control

mutant

dif

Select test

One-class

○ Limma

Two-classes

○ T-test

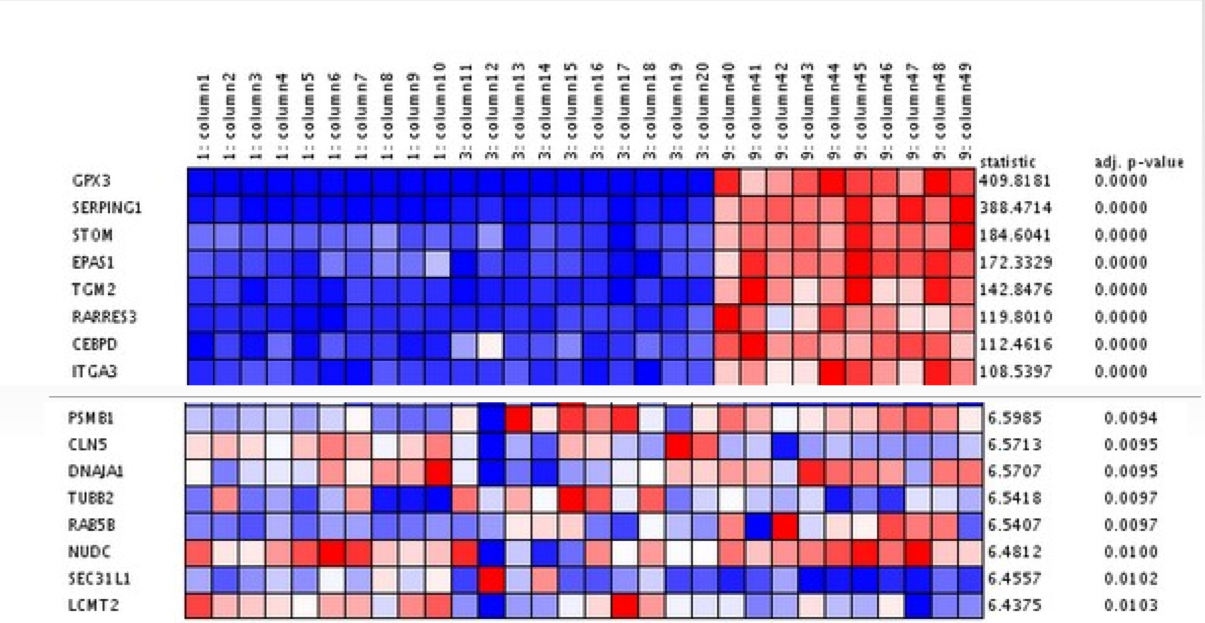
○ Limma

○ Fold-change

Multi-classes

● Anova

○ Limma



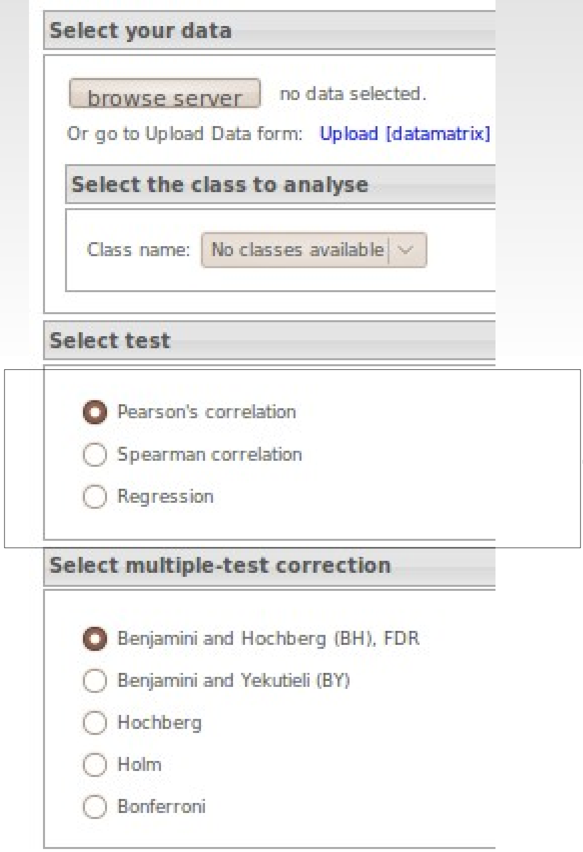
	statistic	adj. p-value
GPX3	409.8181	0.0000
SERPING1	388.4714	0.0000
STOM	184.6041	0.0000
EPAS1	172.3329	0.0000
TGM2	142.8476	0.0000
RARRES3	119.8010	0.0000
CEBPD	112.4616	0.0000
ITGA3	108.5397	0.0000
PSMB1	6.5985	0.0094
CLN5	6.5713	0.0095
DNAJA1	6.5707	0.0095
TUBB2	6.5418	0.0097
RAB5B	6.5407	0.0097
NUDC	6.4812	0.0100
SEC31L1	6.4557	0.0102
LCMT2	6.4375	0.0103

```
#NAMES       statistic           p-value                adj. p-value
GPX3         409.8180802593008   1.9528823003156504E-18 1.945070771143877E-16
SERPING1     388.47140956564357  5.341282971471628E-17  4.092244491989032E-15
STOM         184.60411994407977  2.220446049250313E-16  7.37188088351104E-14
EPAS1        172.33291938517297  4.440892098500626E-16  1.1057821325266559E-13
TGM2         142.84763740196715  4.3298697960381105E-15 8.625100633707915E-13
RARRES3      119.80095119621501  3.752553823233029E-14  6.229239346566828E-12
```

H_o: $\mu_1 = \mu_2 = \ldots = \mu_n$

H_a: not H_o

Gene expression related to a continuous variable, *form*



Select your data

browse server no data selected.

Or go to Upload Data form: Upload [datamatrix]

Select the class to analyse

Class name: No classes available

Select test

● Pearson's correlation

○ Spearman correlation

○ Regression

Select multiple-test correction

● Benjamini and Hochberg (BH), FDR

○ Benjamini and Yekutieli (BY)

○ Hochberg

○ Holm

○ Bonferroni

→

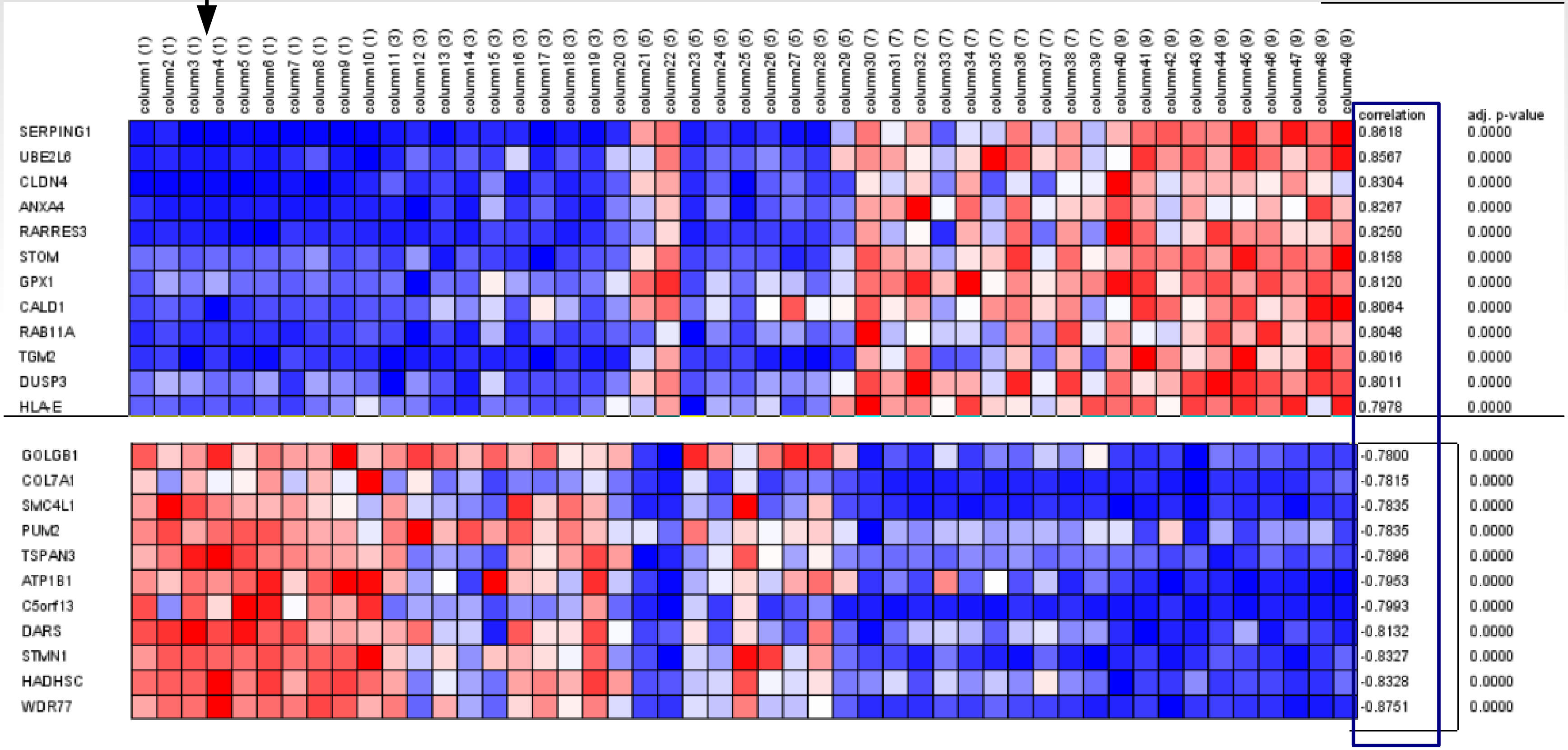
methods

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

Correlation *results*

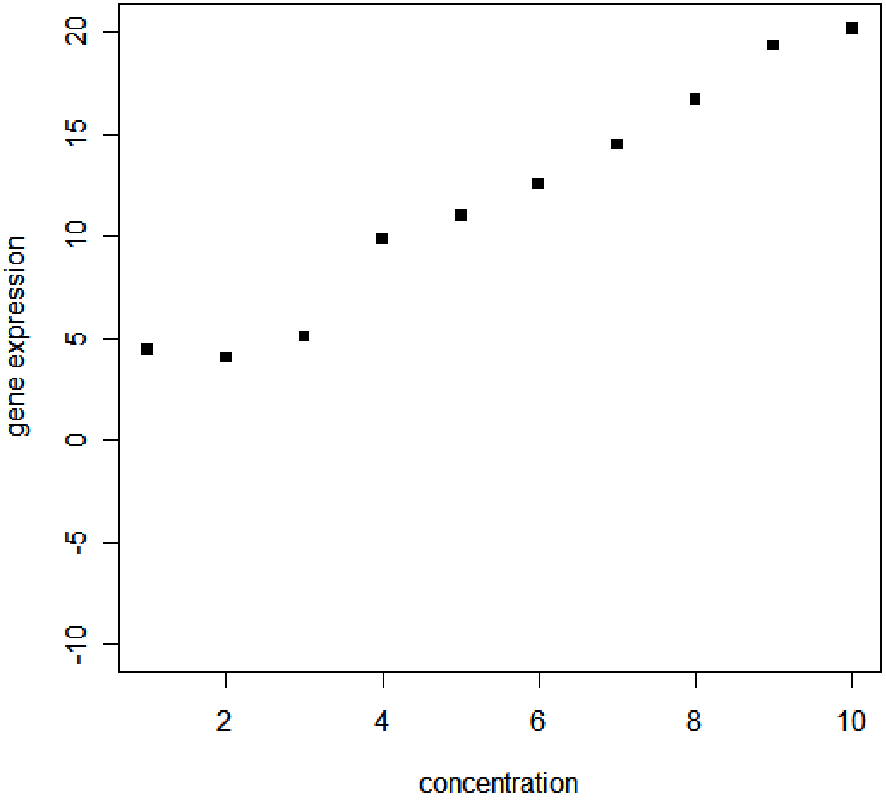
Arrays ranked according to the independent variable



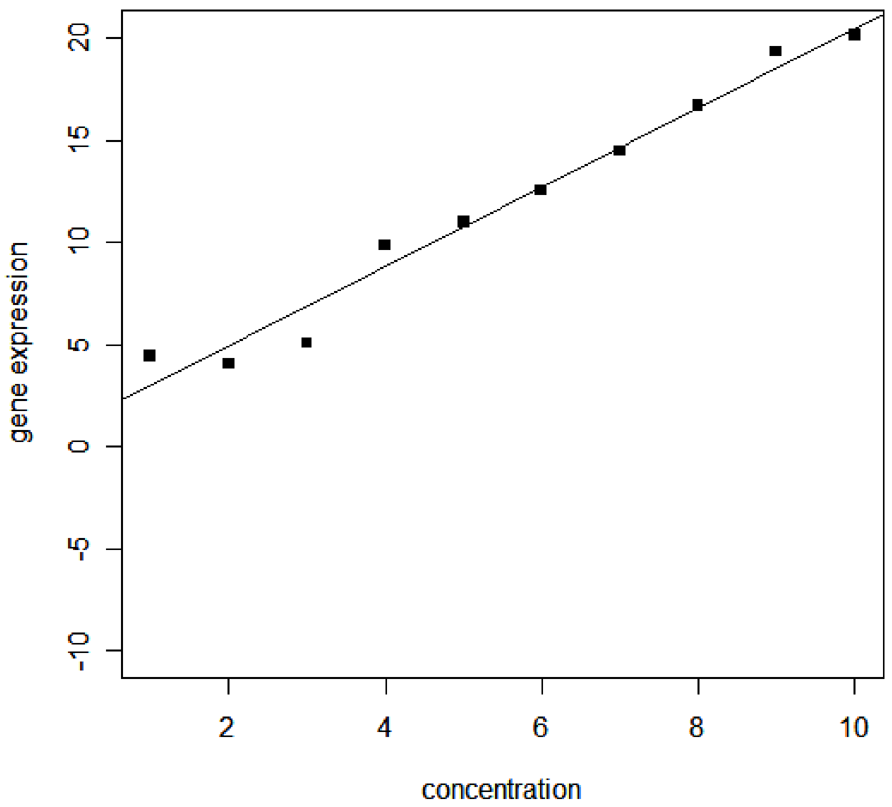
	correlation	adj. p-value
SERPING1	0.8618	0.0000
UBE2L6	0.8567	0.0000
CLDN4	0.8304	0.0000
ANXA4	0.8267	0.0000
RARRES3	0.8250	0.0000
STOM	0.8158	0.0000
GPX1	0.8120	0.0000
CALD1	0.8064	0.0000
RAB11A	0.8048	0.0000
TGM2	0.8016	0.0000
DUSP3	0.8011	0.0000
HLA-E	0.7978	0.0000
GOLGB1	-0.7800	0.0000
COL7A1	-0.7815	0.0000
SMC4L1	-0.7835	0.0000
PUM2	-0.7835	0.0000
TSPAN3	-0.7896	0.0000
ATP1B1	-0.7953	0.0000
C5orf13	-0.7993	0.0000
DARS	-0.8132	0.0000
STMN1	-0.8327	0.0000
HADHSC	-0.8328	0.0000
WDR77	-0.8751	0.0000

Genes ranked by correlation to the continuous variable

Regression

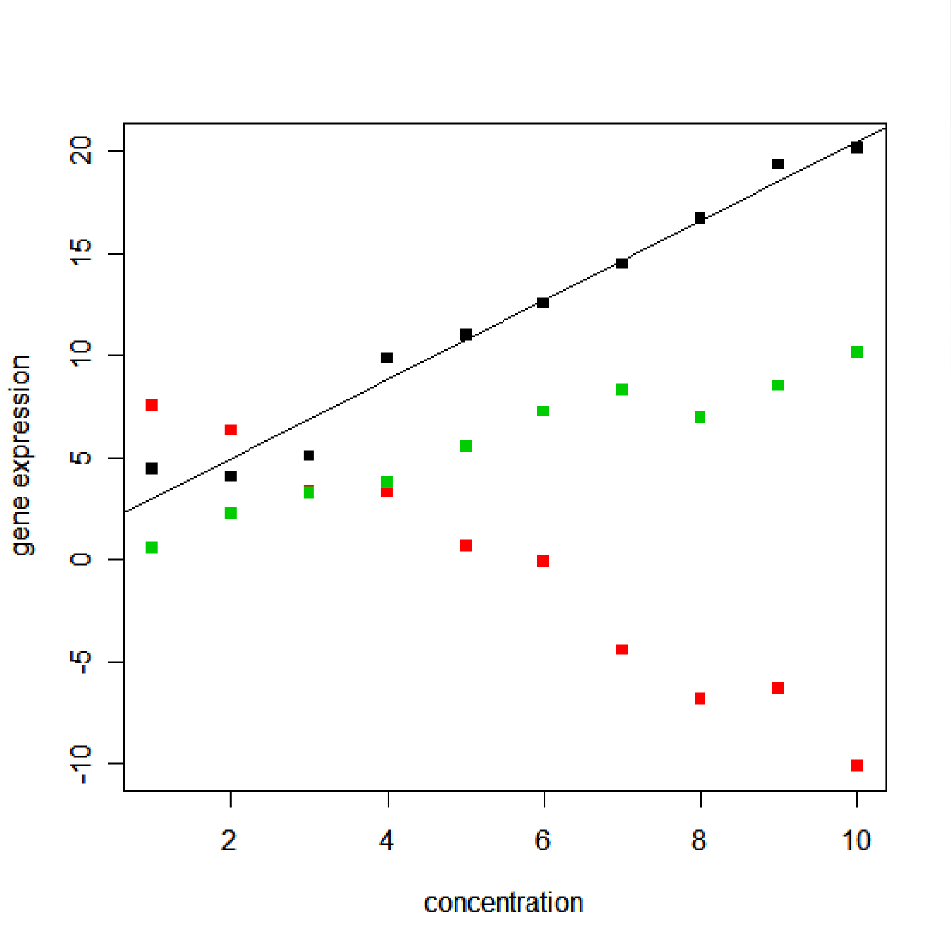


Regression



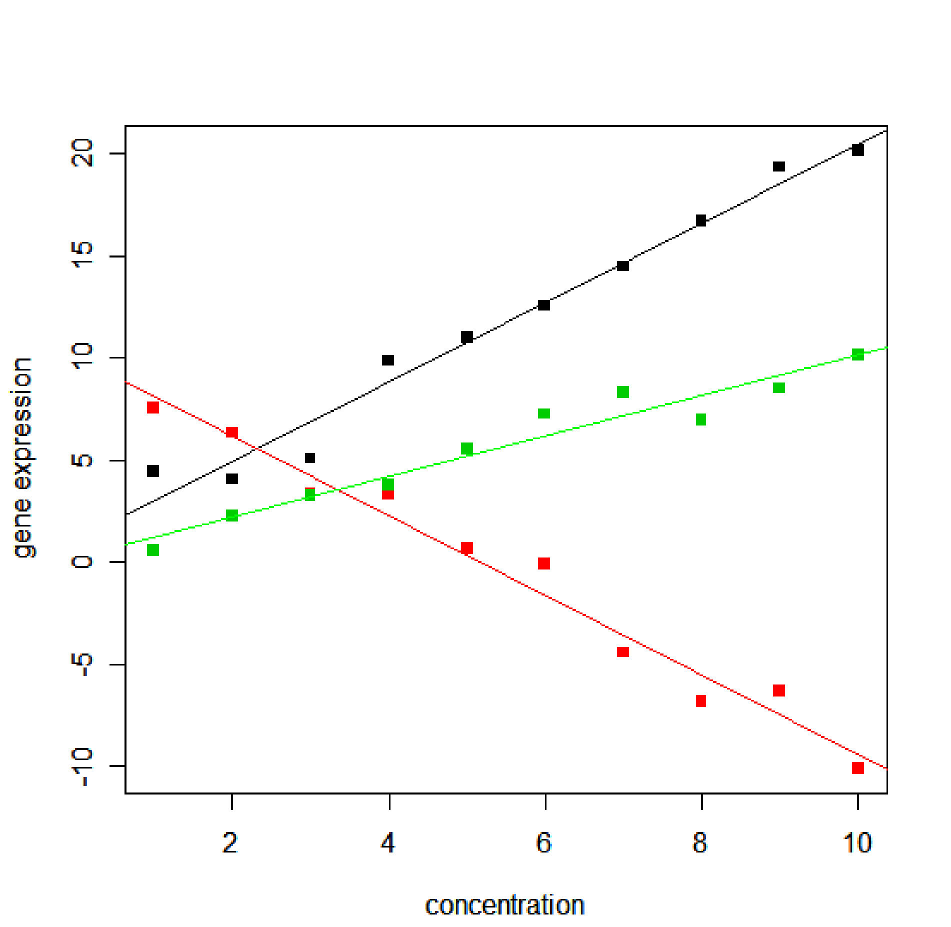
gene1 slope

Regression



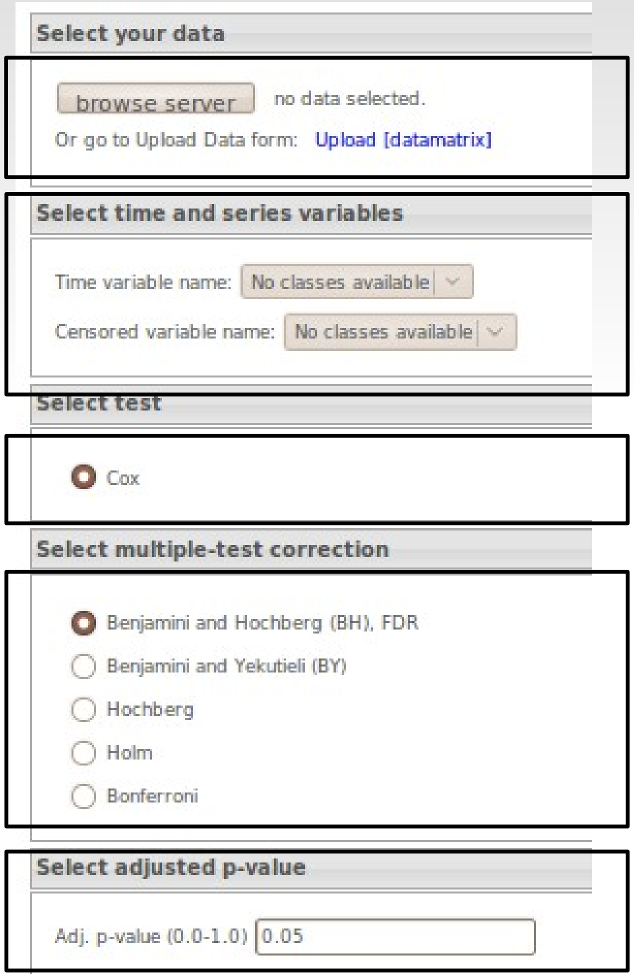
gene1 slope

Regression



gene1 slope
gene2 slope
gene3 slope
…

Survival *form*



Select your data

no data selected.

Or go to Upload Data form: Upload [datamatrix]

Select time and series variables

Time variable name: No classes available

Censored variable name: No classes available

Select test

● Cox

Select multiple-test correction

● Benjamini and Hochberg (BH), FDR

○ Benjamini and Yekutieli (BY)

○ Hochberg

○ Holm

○ Bonferroni

Select adjusted p-value

Adj. p-value (0.0-1.0) 0.05

← 1. Expression

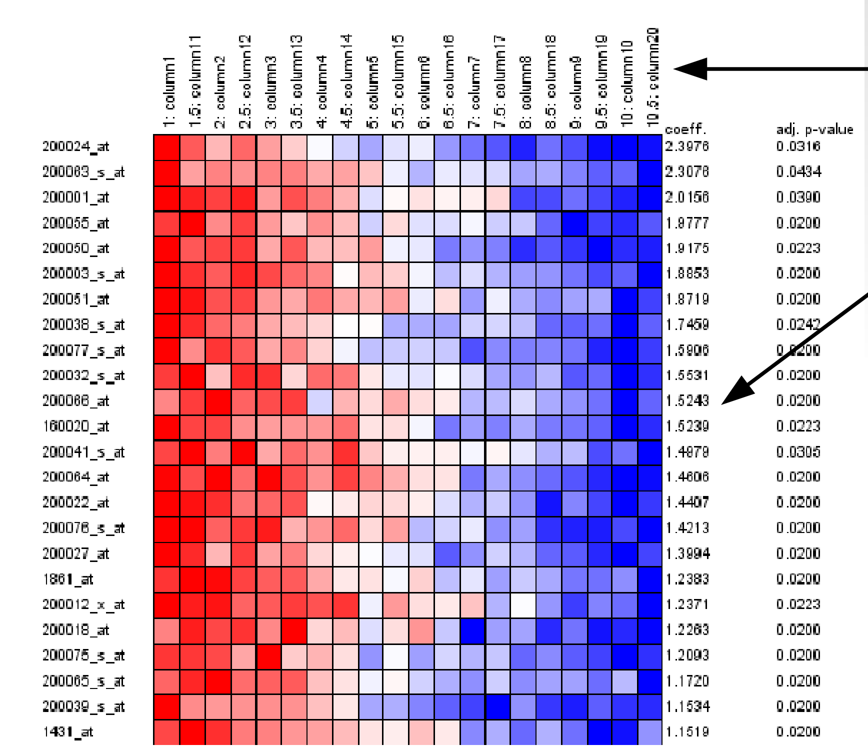
← 2. Survival times & censoring variable

← 3. Cox proportional hazards regression model

← 4. Multiple-test correction

← 5. adjusted p-value

Survival *results*



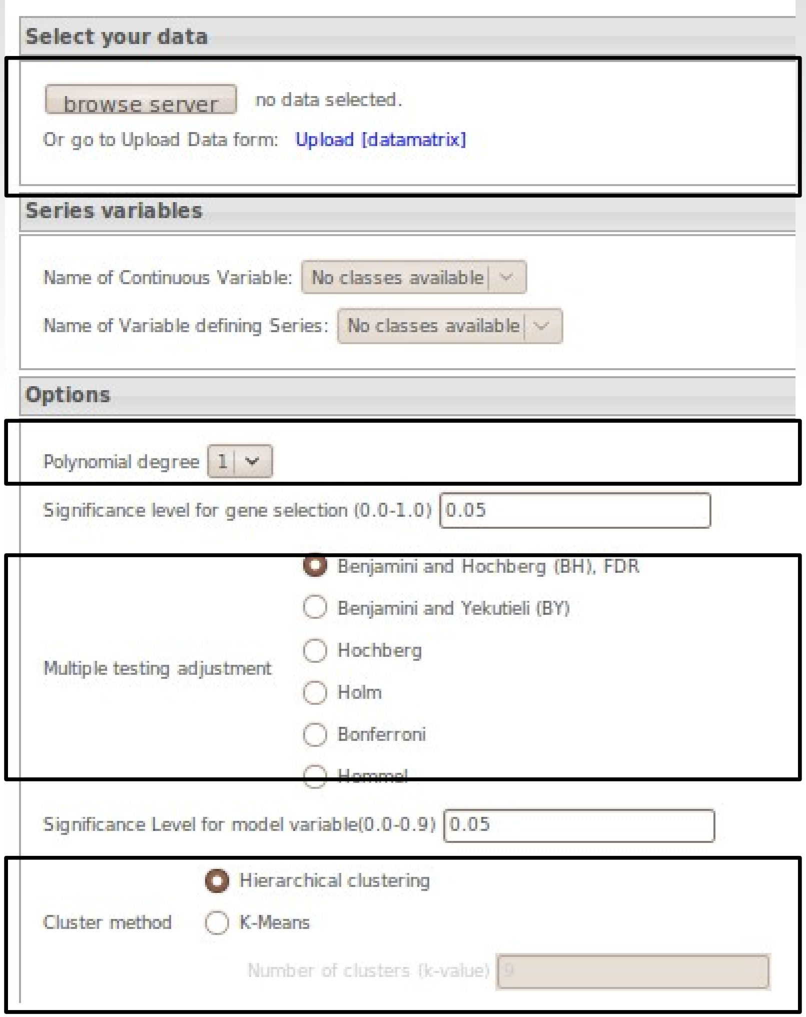
	coeff.	adj. p-value
200024_at	2.3976	0.0316
200083_s_at	2.3076	0.0434
200001_at	2.0156	0.0390
200055_at	1.9777	0.0200
200050_at	1.9175	0.0223
200003_s_at	1.8853	0.0200
200051_at	1.8719	0.0200
200038_s_at	1.7459	0.0242
200077_s_at	1.5906	0.0200
200032_s_at	1.5531	0.0200
200066_at	1.5243	0.0200
160020_at	1.5239	0.0223
200041_s_at	1.4978	0.0305
200064_at	1.4606	0.0200
200022_at	1.4407	0.0200
200076_s_at	1.4213	0.0200
200027_at	1.3994	0.0200
1861_at	1.2383	0.0200
200012_x_at	1.2371	0.0223
200018_at	1.2263	0.0200
200075_s_at	1.2093	0.0200
200065_s_at	1.1720	0.0200
200039_s_at	1.1534	0.0200
1431_at	1.1519	0.0200

Arrays ranked according to the survival time

Genes ranked by their relationship with survival time

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

Time course analysis / Dose analysis, *form*



Select your data

browse server no data selected.

Or go to Upload Data form: Upload [datamatrix]

Series variables

Name of Continuous Variable: No classes available

Name of Variable defining Series: No classes available

Options

Polynomial degree 1

Significance level for gene selection (0.0-1.0) 0.05

Multiple testing adjustment
- ● Benjamini and Hochberg (BH), FDR
- ○ Benjamini and Yekutieli (BY)
- ○ Hochberg
- ○ Holm
- ○ Bonferroni
- ○ Hommel

Significance Level for model variable(0.0-0.9) 0.05

Cluster method
- ● Hierarchical clustering
- ○ K-Means

Number of clusters (k-value) 9

← Expression data

← Time variable and series classification

← Complexity of model

← Significance level

← Multiple testing

← Clustering

Time course analysis / Dose analysis, *example*

✓ Arabidopsis.

✓ 4 series: **control and 3 treatments** (cold, salt, heat).

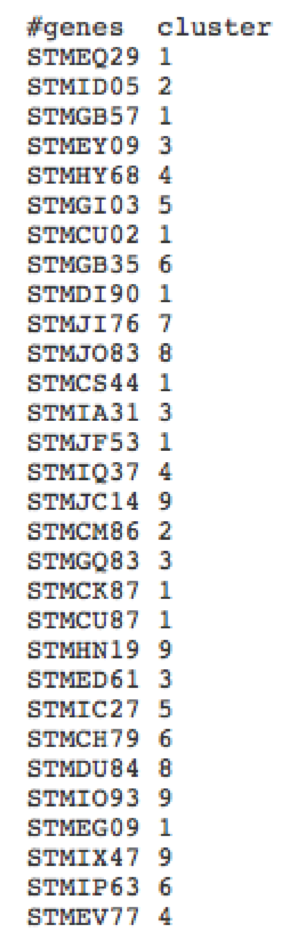
✓ 3 time points.

✓ 3 replicates.

What do we want?

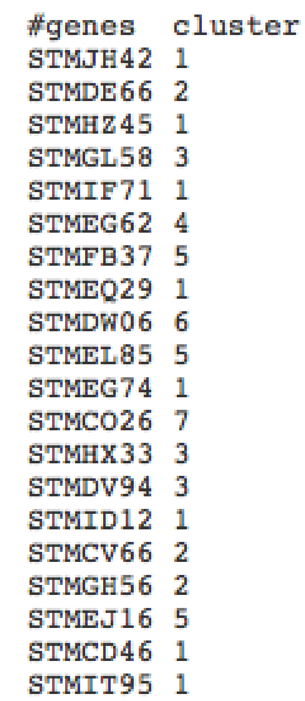
Time course analysis / Dose analysis, *results*

Control



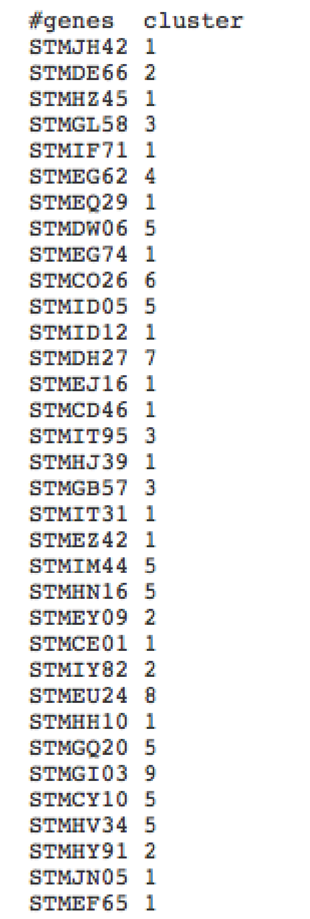
#genes	cluster
STMEQ29	1
STMID05	2
STMGB57	1
STMEY09	3
STMHY68	4
STMGI03	5
STMCU02	1
STMGB35	6
STMDI90	1
STMJI76	7
STMJO83	8
STMCS44	1
STMIA31	3
STMJF53	1
STMIQ37	4
STMJC14	9
STMCM86	2
STMGQ83	3
STMCK87	1
STMCU87	1
STMHN19	9
STMED61	3
STMIC27	5
STMCH79	6
STMDU84	8
STMIO93	9
STMEG09	1
STMIX47	9
STMIP63	6
STMEV77	4

Cold vs Control



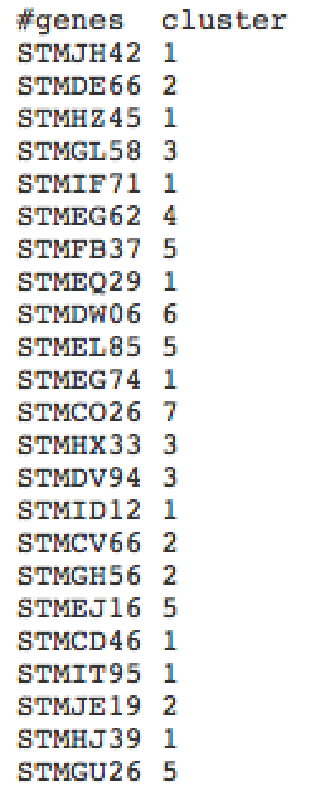
#genes	cluster
STMJH42	1
STMDE66	2
STMHZ45	1
STMGL58	3
STMIF71	1
STMEG62	4
STMFB37	5
STMEQ29	1
STMDW06	6
STMEL85	5
STMEG74	1
STMCO26	7
STMHX33	3
STMDV94	3
STMID12	1
STMCV66	2
STMGH56	2
STMEJ16	5
STMCD46	1
STMIT95	1

Salt vs Control



#genes	cluster
STMJH42	1
STMDE66	2
STMHZ45	1
STMGL58	3
STMIF71	1
STMEG62	4
STMEQ29	1
STMDW06	5
STMEG74	1
STMCO26	6
STMID05	5
STMID12	1
STMDH27	7
STMEJ16	1
STMCD46	1
STMIT95	3
STMHJ39	1
STMGB57	3
STMIT31	1
STMEZ42	1
STMIM44	5
STMHN16	5
STMEY09	2
STMCE01	1
STMIY82	2
STMEU24	8
STMHH10	1
STMGQ20	5
STMGI03	9
STMCY10	5
STMHV34	5
STMHY91	2
STMJN05	1
STMEF65	1

Heat vs Control



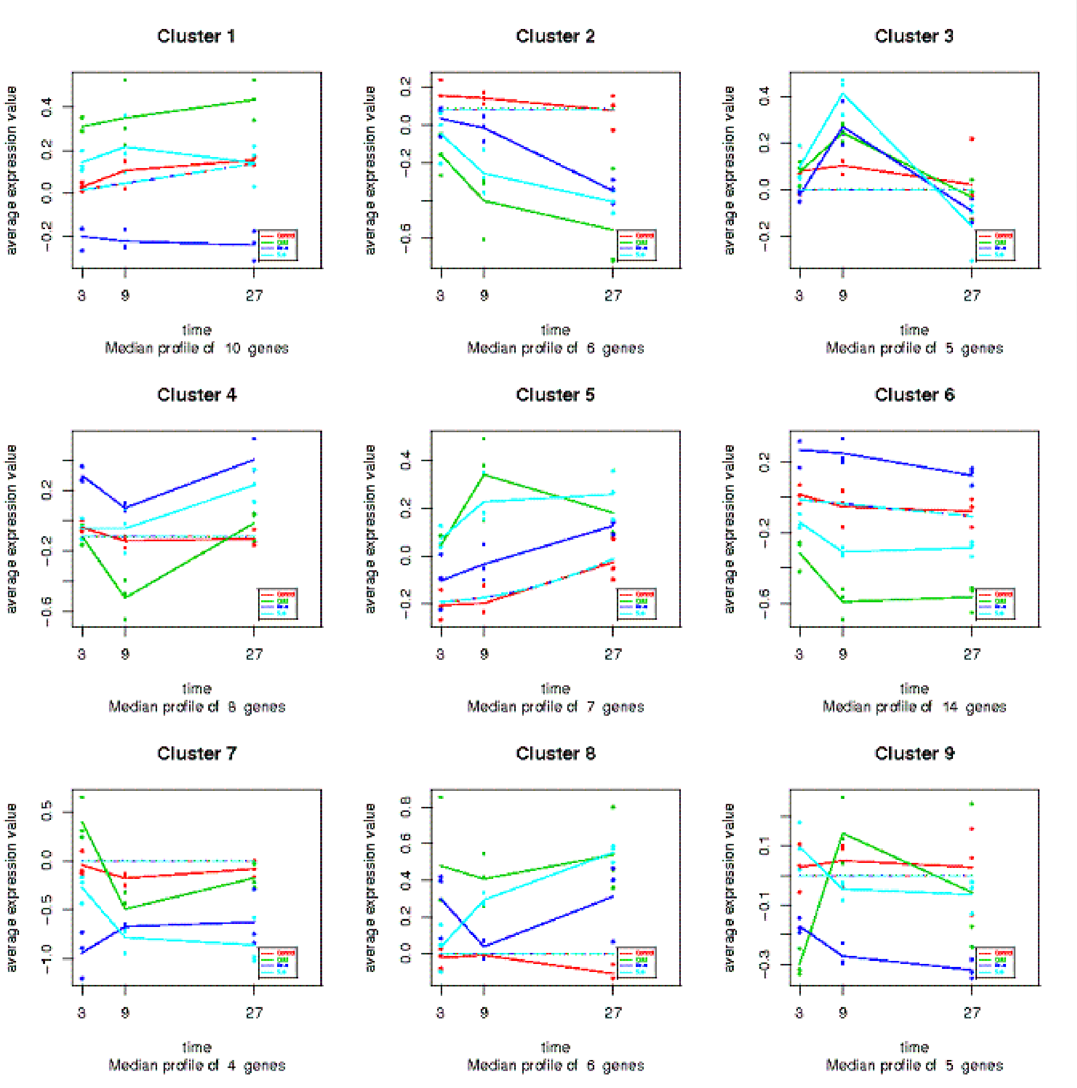
#genes	cluster
STMJH42	1
STMDE66	2
STMHZ45	1
STMGL58	3
STMIF71	1
STMEG62	4
STMFB37	5
STMEQ29	1
STMDW06	6
STMEL85	5
STMEG74	1
STMCO26	7
STMHX33	3
STMDV94	3
STMID12	1
STMCV66	2
STMGH56	2
STMEJ16	5
STMCD46	1
STMIT95	1
STMJE19	2
STMHJ39	1
STMGU26	5

Time course analysis / Dose analysis, *results*

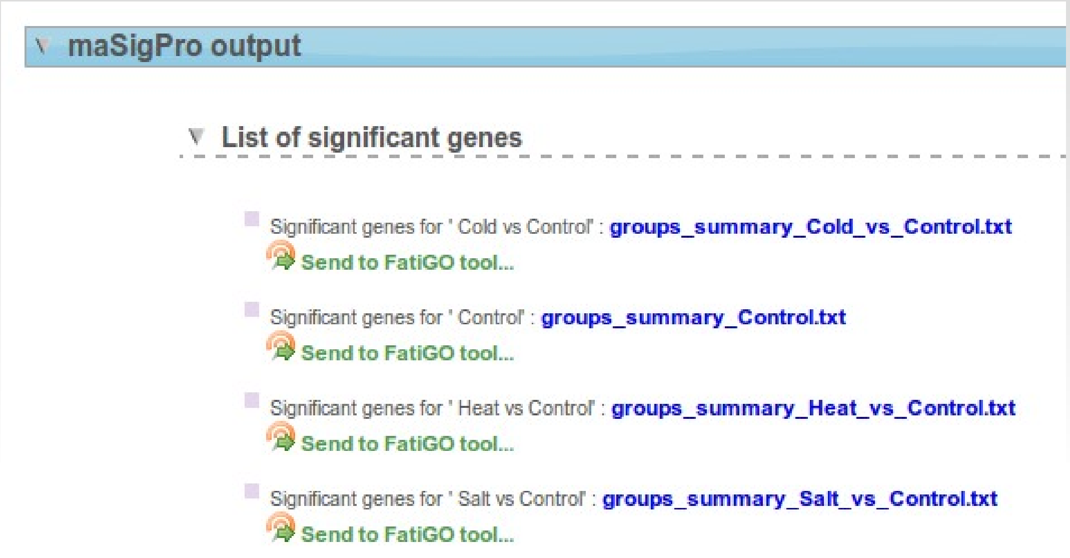
Cold vs Control

#genes	cluster
STMFB37	1
STMDW06	2
STMEL85	1
STMCO26	3
STMHX33	1
STMHQ28	1
STMCV66	4
STMDH27	4
STMIT95	5
STMJE19	6
STMHJ39	5
STMGU26	1
STMEZ42	5
STMCV36	4
STMIM44	6
STMEM39	5
STMHY68	4
STMGH85	1
STMGQ20	2
STMGI03	7
STMJN05	8
STMDB75	7
STMFB38	4
STMDJ03	9
STMGB35	5
STMDU19	9
STMDE59	4
STMCF08	6
STMHK44	7
STMJI76	3
STMEM80	6
STMIA39	6
STMIO60	8

210 significant genes



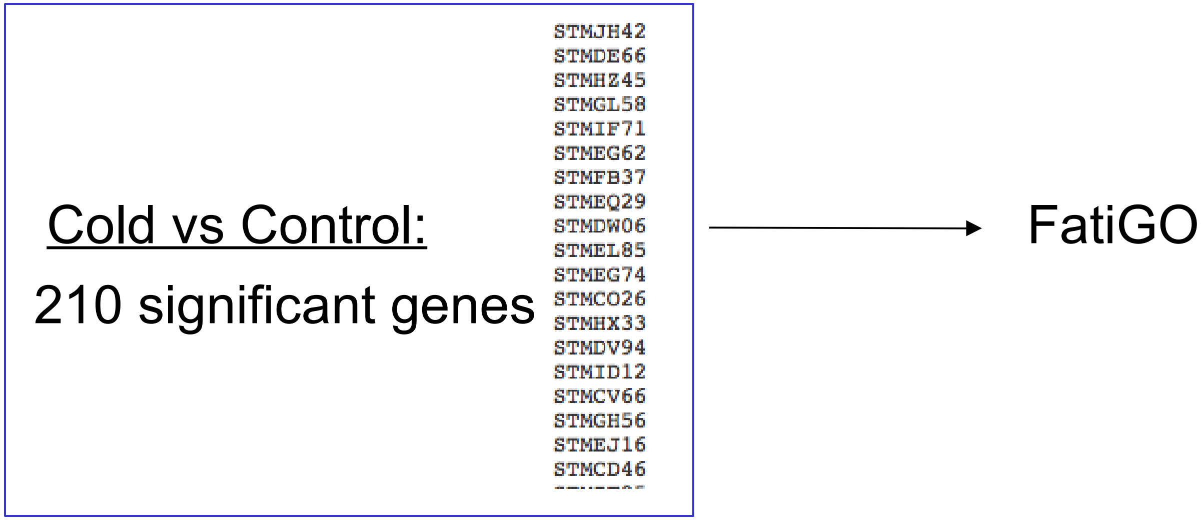
Time course analysis / Dose analysis. Redirecting to functional tools



maSigPro output

List of significant genes

- Significant genes for ' Cold vs Control' : groups_summary_Cold_vs_Control.txt
  Send to FatiGO tool...
- Significant genes for ' Control' : groups_summary_Control.txt
  Send to FatiGO tool...
- Significant genes for ' Heat vs Control' : groups_summary_Heat_vs_Control.txt
  Send to FatiGO tool...
- Significant genes for ' Salt vs Control' : groups_summary_Salt_vs_Control.txt
  Send to FatiGO tool...



Cold vs Control:

210 significant genes

```
STMJH42
STMDE66
STMHZ45
STMGL58
STMIF71
STMEG62
STMFB37
STMEQ29
STMDW06
STMEL85
STMEG74
STMCO26
STMHX33
STMDV94
STMID12
STMCV66
STMGH56
STMEJ16
STMCD46
```

Remember...

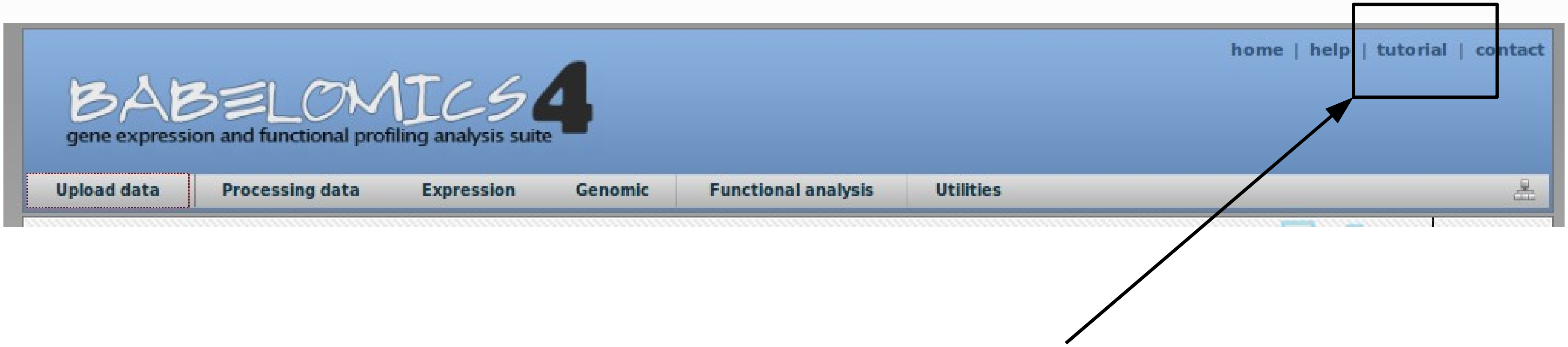
- Babelomics allows us to analyze Differential Expression in **different experimental scenarios.**

- Differential Expression needs **normalized data** from Normalization Babelomics or other tool.

- These results can be **functional interpretated** using several tools in Babelomics: FatiGo, Logistic Models, Snow,...

Let's practise!

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


home | help | tutorial | contact
BABELOMICS 4
gene expression and functional profiling analysis suite
Upload data
Processing data
Expression
Genomic
Functional analysis
Utilities

Go to the tutorial:

Expression Data Analysis / Differential Expression