Use of GO Terms to Understand the Biological Significance of Microarray Differential Gene Expression Data

Ramón Díaz-Uriarte and Joaquín Dopazo
rdiaz@cnio.es, jdopazo@cnio.es

http://bioinfo.cnio.es/∼rdiaz

Bioinformatics Unit
Centro Nacional de Investigaciones Oncológicas (CNIO)
(Spanish National Cancer Center)

Outline of talk

- Scenario and objectives
- Outline of methodology
  - Gene Ontology
  - Fisher's test and adjustments for multiple testing
- Analysis of kidney and testis data
- Alternatives to some of the steps
- Discussion (virtues and limitations) and conclusion

Scenario

- DNA array technology yields information on the expression of thousands of genes.
- We might be interested in understanding, e.g., which genes differ between organs (data from Project Normal).
- We can examine each gene selecting only those that show significant differences using an appropriate statistical model, and correcting for multiple testing.

- The threshold, thus, is based on conventional levels (e.g., Type I error rate of 0.05) attending exclusively to statistical criteria.
- This approach, however, does not necessarily provide any cues of the biological significance of these differences.

Objective

Illustrate the use of a simple method that can help us understand the biological relevance of statistical differences in gene expression data (e.g., among organs of mouse).

- Examine significant differences in the distribution of terms related to biological processes or molecular functions.

Objective

Illustrate the use of a simple method that can help us understand the biological relevance of statistical differences in gene expression data (e.g., among organs of mouse).

- Examine significant differences in the distribution of terms related to biological processes or molecular functions.
- Information on biological process or molecular functions can be obtained from Gene Ontology (GO).

Outline of methodology (I)

1. Analyze microarray expression data to sort genes based on how much they differ between organs:

$$y_{ijkl} = \mu + dye_i + mouse_j + organ_k + error_{ijkl}.$$

Sort genes based on *t* statistic for organ effect.

E.g., comparing kidney and testis:

- Very large *t*: much more expressed in kidney;
- Very small *t* —large abs. value—: much more expressed in testis.

Outline of methodology (II)

2. Use the ordering information to group genes. E.g.,

- $t > threshold$: more expressed in kidney;
- $t < -threshold$: more expressed in testis;
- $-threshold \leq t \leq threshold$: none of the above;

Similar approach if using ANOVA *F*-ratios or *p*-values (except no information on directionality).

Outline of methodology (III)

3. Examine which GO terms differ between the two groups we formed in the previous step.

Interlude: Gene Ontology (GO)

- Gene Ontology (GO) provides a structured vocabulary for annotation of eukaryotic genes. Three categories:
  - biological process;
  - molecular function;
  - cellular component;
- Terms hierarchically structured in a network (DAG); more general to more specific.
- Data can be annotated at varying levels (depending on available information).

Outline of methodology (III)

3. Examine which GO terms differ between the two groups we formed in the previous step.
   - For a level (e.g. 3) of GO, obtain GO terms associated with each gene.

Outline of methodology (III)

3. Examine which GO terms differ between the two groups we formed in the previous step.
   - For a level (e.g. 3) of GO, obtain GO terms associated with each gene.
   - For each GO term, test for difference in its frequency between the two groups.

Outline of methodology (III)

3. Examine which GO terms differ between the two groups we formed in the previous step.
   - For a level (e.g. 3) of GO, obtain GO terms associated with each gene.
   - For each GO term, test for difference in its frequency between the two groups.
   - Adjust for multiple testing: adjusted *p*-values for each GO term.

3.b. Difference in frequency of terms

For each GO term prepare 2x2 table:

	Genes $t > 10$	Genes $t < -10$
reproduction term present	a	b
reproduction term absent	c	d

Fisher's exact test for 2x2 contingency table: test association of rows and columns (e.g., class $t > 10$, compared to class $t < -10$, increases or decreases the chances of having the term "reproduction"?).

3.c. Adjustments for multiple testing

- Testing as many hypotheses as GO terms: use of *p*-value $\Rightarrow$ excessive number of false rejections $\equiv$ conclude that too many GO terms have a different frequency between the two groups.

3.c. Adjustments for multiple testing

- Testing as many hypotheses as GO terms: use of *p*-value $\Rightarrow$ excessive number of false rejections $\equiv$ conclude that too many GO terms have a different frequency between the two groups.
- Used permutation-based step-down minP method of Westfall & Young: obtained adjusted *p*-values.

Summary of methodology (I)

- Analyze microarray expression data to sort genes based on how much they differ between organs, using an appropriate statistical model.

Summary of methodology (I)

- Analyze microarray expression data to sort genes based on how much they differ between organs, using an appropriate statistical model.
- Form pairs of groups based on the sorted differences:

Summary of methodology (I)

- Analyze microarray expression data to sort genes based on how much they differ between organs, using an appropriate statistical model.
- Form pairs of groups based on the sorted differences:
  - $t > threshold$

Summary of methodology (I)

- Analyze microarray expression data to sort genes based on how much they differ between organs, using an appropriate statistical model.
- Form pairs of groups based on the sorted differences:
  - $t > threshold$
  - $t < -threshold$

Summary of methodology (I)

- Analyze microarray expression data to sort genes based on how much they differ between organs, using an appropriate statistical model.
- Form pairs of groups based on the sorted differences:
  - $t > threshold$
  - $t < -threshold$
- Test for differential frequency of GO terms between the two groups, with adjustment for multiple testing.

Summary of methodology (II)

Repeat last two steps, changing value of “threshold”; e.g., if threshold takes values 15, 10, 5, we are selecting groups where differences between kidney and testis are progressively less extreme.

As we change the threshold that defines the two groups, we might find that different GO terms show different frequencies between the two groups; thus we can highlight the biological relevance of the statistical differences in gene expression data.

Analysis of kidney and testis data

- Subtracted background from foreground and obtained $\log_2$ experimental/control; set as missing all samples with flag -50; data normalized dividing each $\log_2$ ratio by median $\log_2$ ratio of the array.

Analysis of kidney and testis data

- Subtracted background from foreground and obtained $\log_2$ experimental/control; set as missing all samples with flag -50; data normalized dividing each $\log_2$ ratio by median $\log_2$ ratio of the array.
- To reduce unbalance among organs, only used genes where organ with smallest n had n at least 87.5% of of organ with largest n (for an organ with complete data: $0.875 = 21/24$). Left with 3939 genes (of these, 3830 had at most six missing values).

Fitting the linear model

Fit model

$$y_{ijkl} = \mu + dye_i + mouse_j + organ_k + error_{ijkl},$$

parameterized so that:

- large positive coefficient (or t) $\equiv$ gene much more expressed in kidney;
- large negative coefficient (or t) $\equiv$ gene much more expressed in testis;

Fitting the linear model

Fit model

$$y_{ijkl} = \mu + dye_i + mouse_j + organ_k + error_{ijkl},$$

parameterized so that:

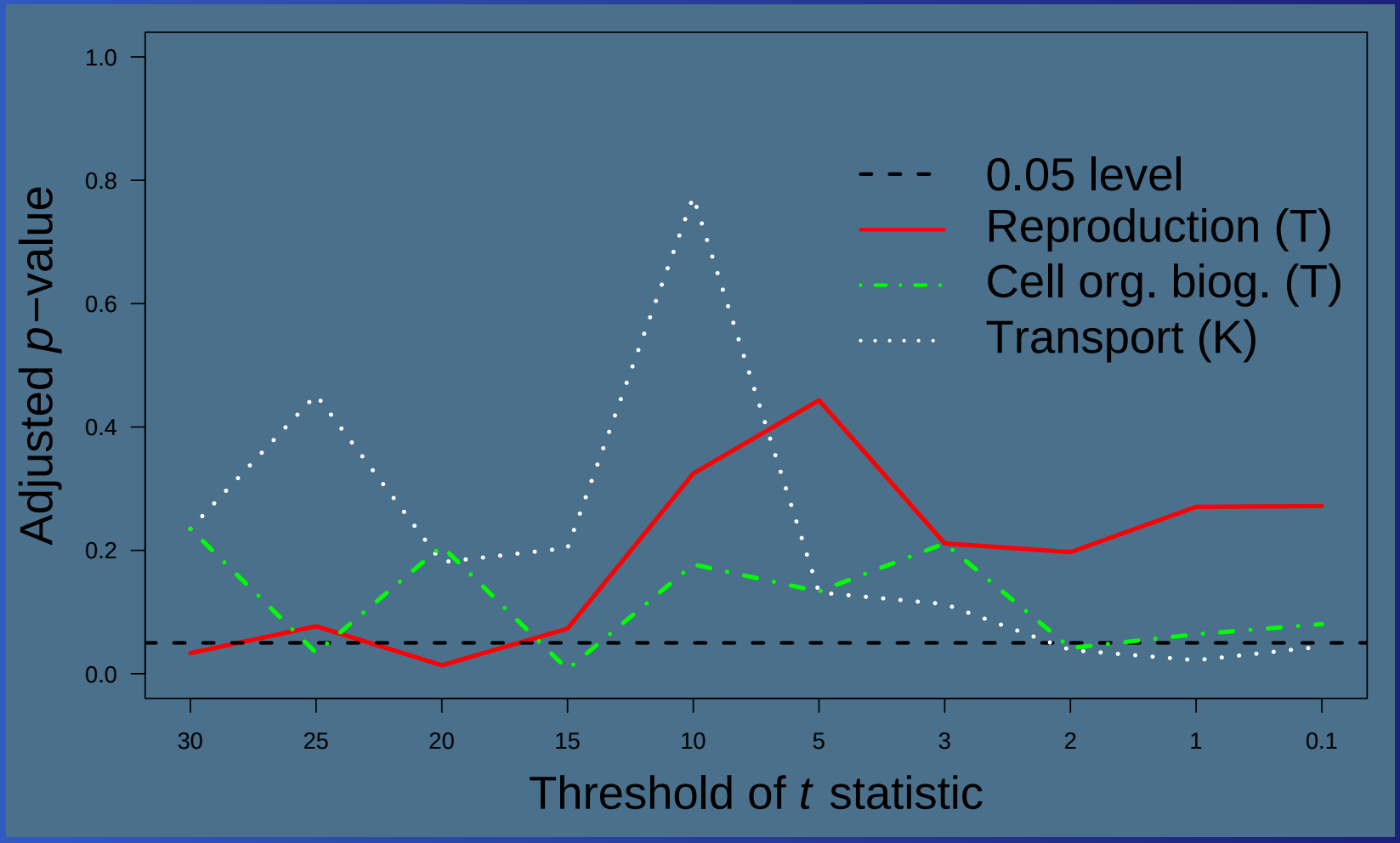
- large positive coefficient (or t) $\equiv$ gene much more expressed in kidney;
- large negative coefficient (or t) $\equiv$ gene much more expressed in testis;

Why t instead of p-value? a) Simpler; b) Approx. same d.f.; c) Threshold just a device to form groups.

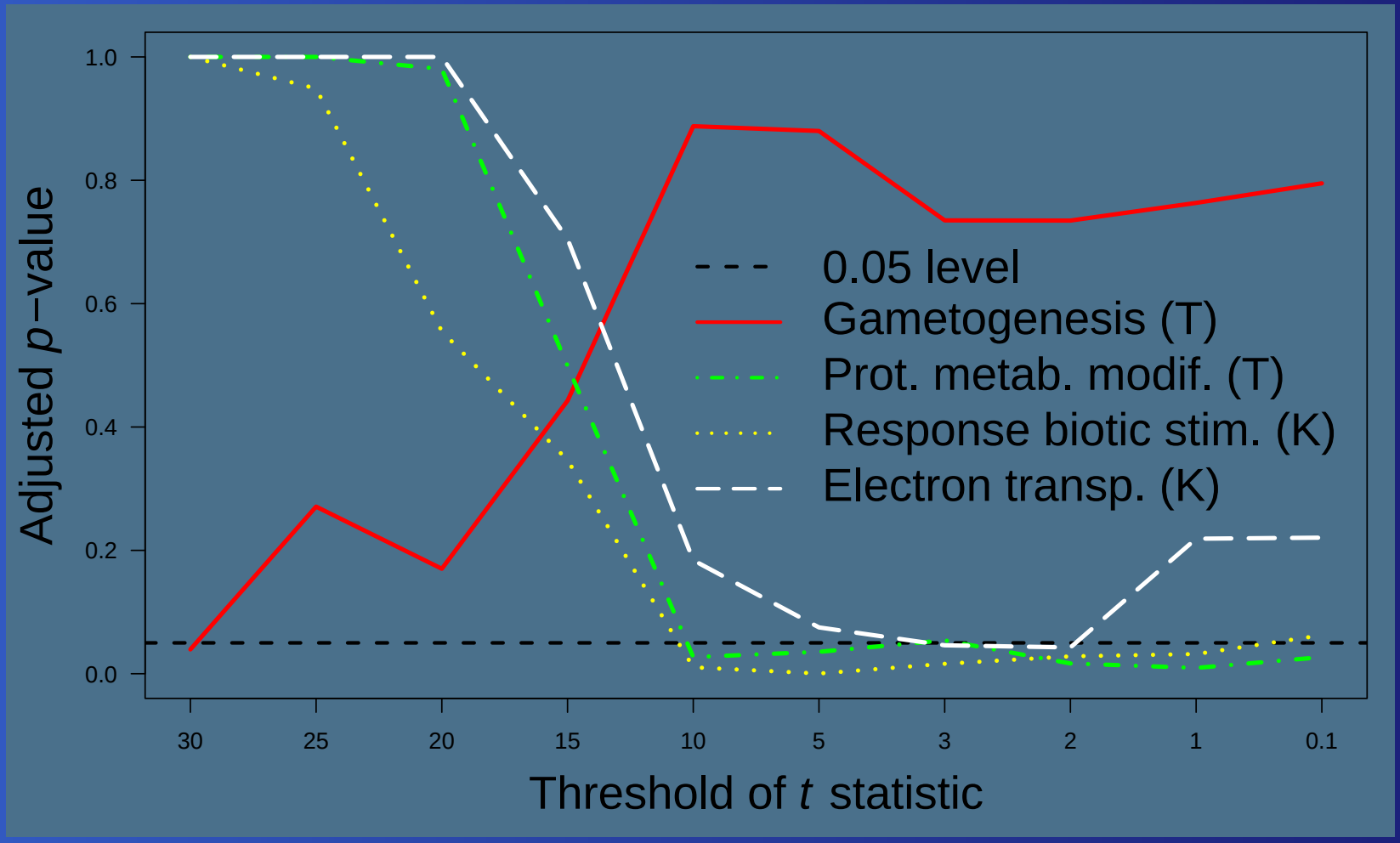
Threshold for *t*

Threshold	Size of groups; kidney:testis	max. test-wise *p*-value	FDR-adjusted *p*-value
30	88:63	$< 10^{-27}$	$< 10^{-25}$
25	133:118	$< 10^{-24}$	$< 10^{-22}$
20	230:257	$< 10^{-20}$	$< 10^{-18}$
15	371:469	$< 10^{-16}$	$< 10^{-15}$
10	654:752	$< 10^{-11}$	$< 10^{-10}$
5	1162:1175	0.000012	0.00018
3	1475:1408	0.0052	0.062
2	1649:1577	0.053	0.571
1	1823:1760	0.323	1
0.1	1985:1912	0.92	1

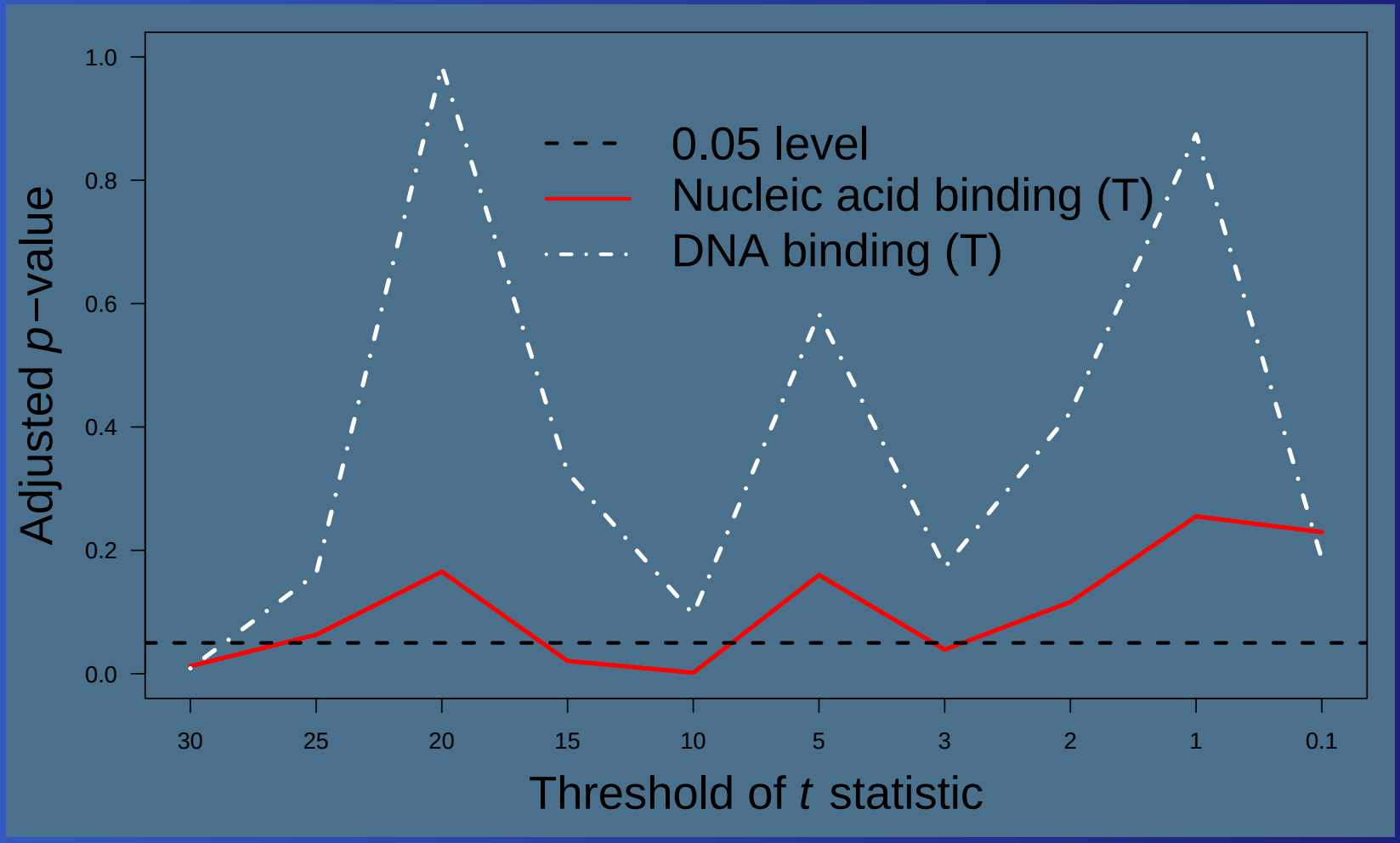
Biological process, level 3



Biological process, level 4



Molecular function, levels 3 & 4



Alternatives to steps 1 & 2

- Form groups based on *p*-value: no directionality, just genes differentially expressed vs. non-different (instead of genes more common in testis vs. more common in kidney).

Alternatives to steps 1 & 2

- Form groups based on *p*-value: no directionality, just genes differentially expressed vs. non-different (instead of genes more common in testis vs. more common in kidney).
  - Some results similar (e.g., transport);

Alternatives to steps 1 & 2

- Form groups based on *p*-value: no directionality, just genes differentially expressed vs. non-different (instead of genes more common in testis vs. more common in kidney).
  - Some results similar (e.g., transport);
  - Others different (e.g., oxidoreductase).

Alternatives to steps 1 & 2

- Form groups based on *p*-value: no directionality, just genes differentially expressed vs. non-different (instead of genes more common in testis vs. more common in kidney).
  - Some results similar (e.g., transport);
  - Others different (e.g., oxidoreductase).
- Three-organ comparisons using ANOVA (*F*-ratio or *p*-value).

Alternative to step 3

- As we decrease the threshold, our new groups include genes that were already included with larger thresholds.

Alternative to step 3

- As we decrease the threshold, our new groups include genes that were already included with larger thresholds.
- Alternatively:

Alternative to step 3

- As we decrease the threshold, our new groups include genes that were already included with larger thresholds.
- Alternatively:
  - form groups that are not overlapping;

Alternative to step 3

- As we decrease the threshold, our new groups include genes that were already included with larger thresholds.
- Alternatively:
  - form groups that are not overlapping;
  - form groups by a sliding window of fixed size of genes;

Alternative to step 3

- As we decrease the threshold, our new groups include genes that were already included with larger thresholds.
- Alternatively:
  - form groups that are not overlapping;
  - form groups by a sliding window of fixed size of genes;
- Currently exploring it; largest problem: small sample sizes and low power of Fisher's test.

Discussion: ease of use

- Procedure is easy to use:
  - First two steps can be implemented in any statistical package with programming facilities (e.g., R).
  - Third step: FatiGO tool available at

    `http://bioinfo.cnio.es/cgi-bin/tools/FatiGO/FatiGO.cgi`

Discussion: limitations (I)

- Many genes have no GO annotations:

Discussion: limitations (I)

- Many genes have no GO annotations:
  - In the examples, about 18% of the genes are annotated at level 4, and about 25% annotated at level 3.

Discussion: limitations (I)

- Many genes have no GO annotations:
  - In the examples, about 18% of the genes are annotated at level 4, and about 25% annotated at level 3.
  - Lack of annotation affects more strongly smaller groups.

Discussion: limitations (I)

- Many genes have no GO annotations:
  - In the examples, about 18% of the genes are annotated at level 4, and about 25% annotated at level 3.
  - Lack of annotation affects more strongly smaller groups.
  - Absence of annotation $\not\equiv$ absence of function.

Discussion: limitations (I)

- Many genes have no GO annotations:
  - In the examples, about 18% of the genes are annotated at level 4, and about 25% annotated at level 3.
  - Lack of annotation affects more strongly smaller groups.
  - Absence of annotation $\not\equiv$ absence of function.

These limitations should be of lesser importance in the future.

Discussion: limitations (II)

- No longer maintaining control of the Experiment Wise Type I Error Rate: process of examining significance of GO terms is repeated at each threshold.

Discussion: limitations (II)

- No longer maintaining control of the Experiment Wise Type I Error Rate: process of examining significance of GO terms is repeated at each threshold.
- Given exploratory nature of this heuristic method, low power of Fisher's exact test, and loss of power because of scarcity of annotations, we prefer to decrease Type II error rates.

Conclusion

Promising method to identify biological features of genes differentially expressed between/among organs or conditions (e.g., cancer vs. non-cancer): it incorporates information on the biological processes and molecular functions associated to the genes.

Acknowledgments

- Fátima Al-Shahrour for the FatiGO tool.
- Luis Lombardía and Ana Dopazo for answers about GENEPIX.