A statistical approach to the comparison of lists of differentially expressed genes

Marta Blangiardo[1] and Sylvia Richardson[1]

[1] *Department of Epidemiology, Public Health and Primary Care, Imperial College London (UK) – email: m.blangiardo@imperial.ac.uk*

Problem presentation

In the microarray framework researchers often are interested in the comparison of two or more similar experiments, which involve different treatment/exposures, tissues, or species. The aim is finding some common denominators between these experiments in the form of a maximal list of genes that are differentially expressed in both (all) the experiments and from which to start further investigations.

Problem presentation

In the microarray framework researchers often are interested in the comparison of two or more similar experiments, which involve different treatment/exposures, tissues, or species. The aim is finding some common denominators between these experiments in the form of a maximal list of genes that are differentially expressed in both (all) the experiments and from which to start further investigations.

Ideally, such a problem should involve the joint re-analysis of the two (all) experiments, but it is not always easily feasible (e.g different platforms), and in any case computationally demanding.

Problem presentation

In the microarray framework researchers often are interested in the comparison of two or more similar experiments, which involve different treatment/exposures, tissues, or species. The aim is finding some common denominators between these experiments in the form of a maximal list of genes that are differentially expressed in both (all) the experiments and from which to start further investigations.

Ideally, such a problem should involve the joint re-analysis of the two (all) experiments, but it is not always easily feasible (e.g different platforms), and in any case computationally demanding.

Alternatively, a natural approach is to consider the measures of differential expression for the two (all) experiments and compute the intersection of the lists. However, some of the genes in the maximal intersection list can be due to chance.

Outline of the work

- We propose a permutation based test for assessing whether the size of the common list is higher than expected by chance under the hypothesis of independence of the measures of differential expression.

Outline of the work

- We propose a permutation based test for assessing whether the size of the common list is higher than expected by chance under the hypothesis of independence of the measures of differential expression.

- We present some limitations of this approach and use a Bayesian model to overcome the problem.

Outline of the work

- We propose a permutation based test for assessing whether the size of the common list is higher than expected by chance under the hypothesis of independence of the measures of differential expression.

- We present some limitations of this approach and use a Bayesian model to overcome the problem.

- Some applications are shown, both on simulated and on real data.

List structure

Suppose we have two experiments, each reporting a measure (e.g. p value) of differential expression on a probability scale:

Small p value $\Longrightarrow$ MOST differentially expressed

p value nearer 1 $\Longrightarrow$ NOT differentially expressed

Experiment A	Experiment B
p_{A1}	p_{B1}
p_{A2}	p_{B2}
$\ldots$	$\ldots$
p_{An}	p_{Bn}

If we simply consider a cut off on the measure of differential expression and count the number of differentially expressed genes in common, we do not take into account the number of genes in common by chance.

2×2 table

For each threshold q:

		Exp B		
		DE	$\overline{DE}$	
Exp A	DE	$O_{11}(q)$	$O_{1+}(q) - O_{11}(q)$	$O_{1+}(q)$
	$\overline{DE}$	$O_{+1}(q) - O_{11}(q)$	$n - O_{+1}(q) - O_{1+}(q) + O_{11}(q)$	$n - O_{1+}(q)$
		$O_{+1}(q)$	$n - O_{+1}(q)$	n

The number of genes in common by chance is calculated as:

$$E(O_{11}(q) \mid H_0) = \frac{O_{1+}(q) \times O_{+1}(q)}{n}$$

The number of genes observed in common is $O_{11}(q)$

Stone's test

The idea comes from Stone (1988), who proposed a test for investigating the excess environmental risks around putative sources.
He partitions the area of interest in N sub-areas and for each sub-area he calculates the number of cases of a disease (O_i) and the number of expected cases (E_i) using some reference population.
Ordering the sub-areas on the basis of their distance from the source he calculates the ratio of the partial sum of the number of observed cases to the partial sum of the number of expected, by increasing distance from the source:

$$T(q^*) = max_{1 \leq q \leq N} \frac{\Sigma_{i=1}^{q} O_i}{\Sigma_{i=1}^{q} E_i}$$

This is relating to the radius around the putative source within which the observed relative risk is maximized.

Ratio

We propose to calculate the maximum of the observed to expected ratio:

$$T(q^*) = max_q T(q) = \frac{O_{11}(q^*)}{E(O_{11}(q^*) \mid H_0)} \qquad \text{where} \qquad T(q) = \frac{O_{11}(q)}{E(O_{11}(q) \mid H_0)}$$

It is the maximal deviation from the underneath independence model.

The list of these $O_{11}(q^*)$ genes can be extracted for further biological investigations.

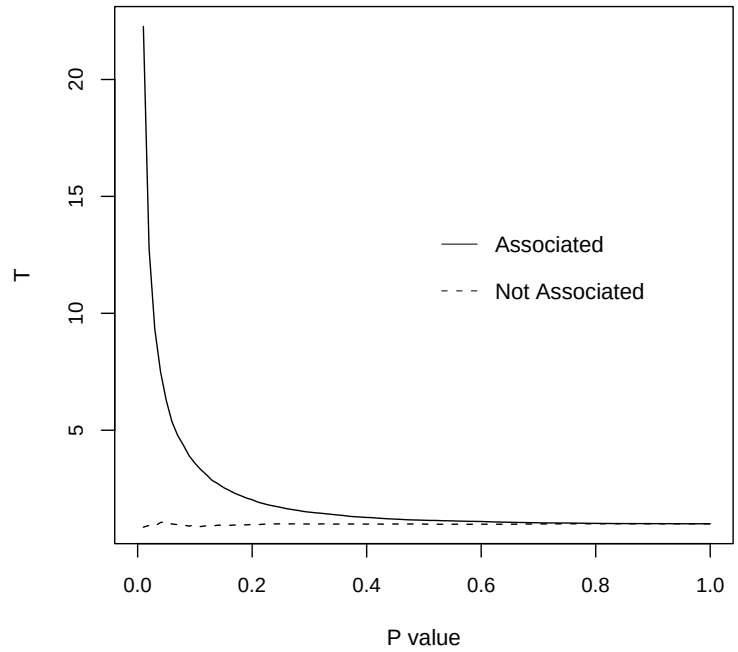
Ratio

We propose to calculate the maximum of the observed to expected ratio:

$$T(q^*) = max_q T(q) = \frac{O_{11}(q^*)}{E(O_{11}(q^*) \mid H_0)} \qquad \text{where} \qquad T(q) = \frac{O_{11}(q)}{E(O_{11}(q) \mid H_0)}$$

It is the maximal deviation from the underneath independence model.

The list of these $O_{11}(q^*)$ genes can be extracted for further biological investigations.



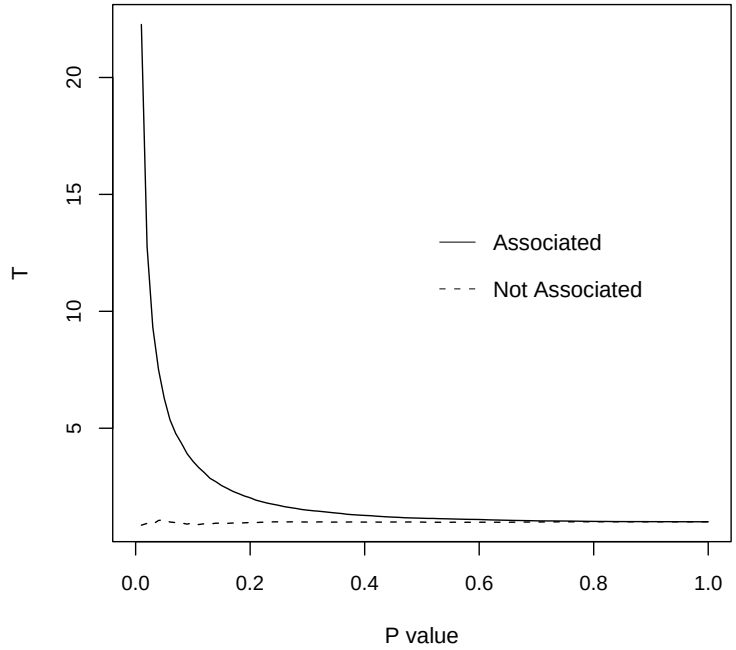
Ratio

We propose to calculate the maximum of the observed to expected ratio:

$$T(q^*) = max_q T(q) = \frac{O_{11}(q^*)}{E(O_{11}(q^*) \mid H_0)} \qquad \text{where} \qquad T(q) = \frac{O_{11}(q)}{E(O_{11}(q) \mid H_0)}$$

It is the maximal deviation from the underneath independence model.

The list of these $O_{11}(q^*)$ genes can be extracted for further biological investigations.



	Not associated	Associated
$T(q^*)$	1.1	22.27
q^*	0.05	0.01
$O_{1+}(q^*)$	200	76
$O_{+1}(q^*)$	180	65
$O_{11}(q^*)$	19	55

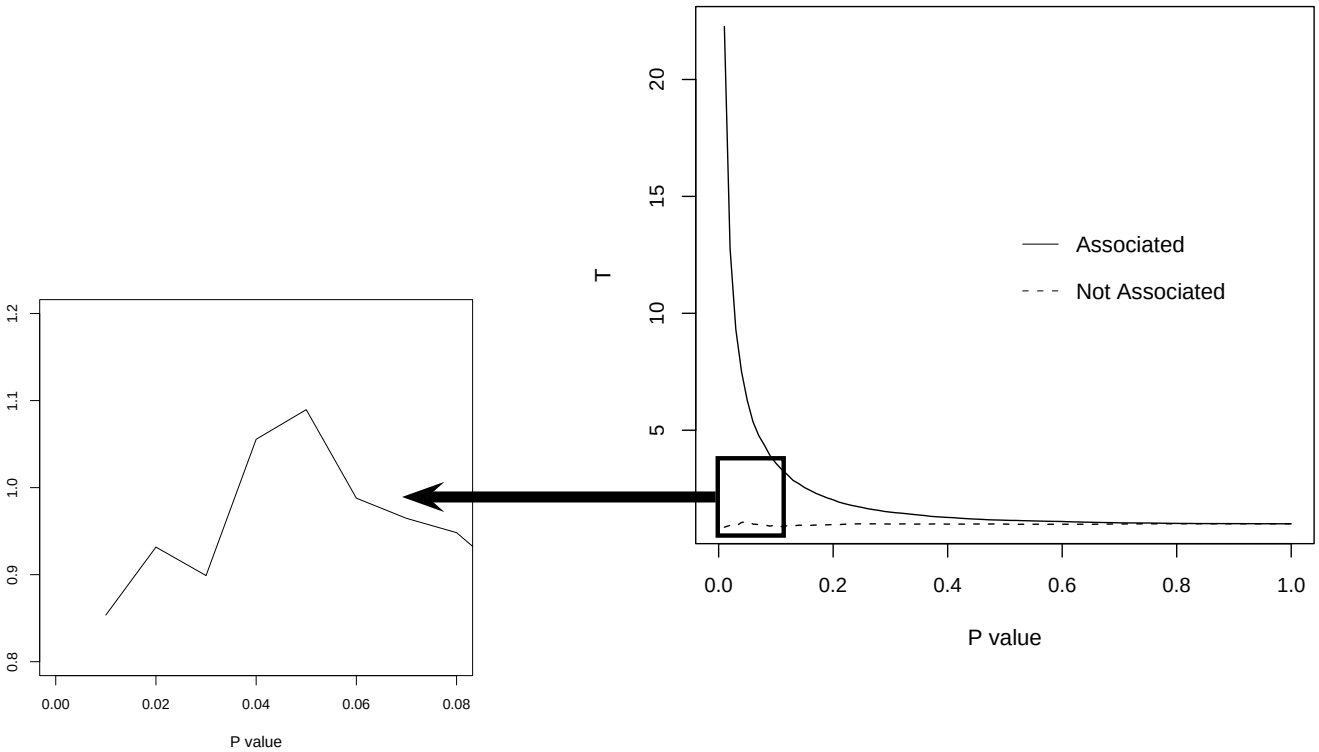
Ratio

We propose to calculate the maximum of the observed to expected ratio:

$$T(q^*) = max_q T(q) = \frac{O_{11}(q^*)}{E(O_{11}(q^*) \mid H_0)} \qquad \text{where} \qquad T(q) = \frac{O_{11}(q)}{E(O_{11}(q) \mid H_0)}$$

It is the maximal deviation from the underneath independence model.

The list of these $O_{11}(q^*)$ genes can be extracted for further biological investigations.



	Not associated	Associated
$T(q^*)$	1.1	22.27
q^*	0.05	0.01
$O_{1+}(q^*)$	200	76
$O_{+1}(q^*)$	180	65
$O_{11}(q^*)$	19	55

Modelling $O_{11}(q)$

Given a threshold q and fixed margins, $O_{11}(q)$ has a hypergeometric distribution:

$$O_{11}(q) \sim Hyper(O_{1+}(q), O_{+1}(q), n)$$

$$P(O_{11}(q) \mid O_{1+}(q), O_{+1}(q), n, H_0) = \frac{\begin{pmatrix} O_{1+}(q) \\ O_{11}(q) \end{pmatrix} \begin{pmatrix} n - O_{1+}(q) \\ O_{+1}(q) - O_{11}(q) \end{pmatrix}}{\begin{pmatrix} n \\ O_{1+}(q) \end{pmatrix}}$$

hence the distribution of the ratio $T(q)$ is:

$$T(q) \propto Hyper(O_{1+}(q), O_{+1}(q), n)$$

However, the distribution of $T(q^*)$ is not easily obtained, since the tables are not independent (nested in each other).
We take advantage of the empirical distribution for $T(q^*)$ obtained through simulation.

Permutation test

We perform a permutation test of $T(q^*)$ under the null hypothesis of independence between the two experiments.

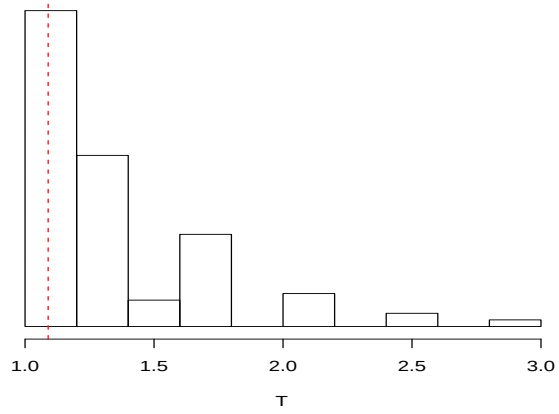
In this test, the measures of probability of one list are randomly permuted several times, while the ones for the other list are keeping fixed. For the s^{th} permutation $(s = 1, \ldots, S)$:

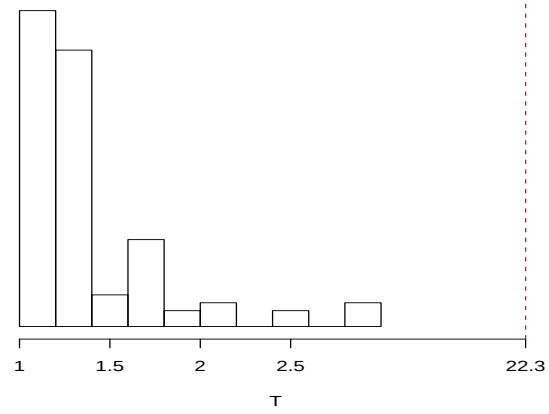
Experiment A	Experiment B
p_{A1}	$p_{Bs(1)}$
p_{A2}	$p_{Bs(2)}$
$\ldots$	$\ldots$
p_{An}	$p_{Bs(n)}$

Any relationship between the two lists is destroyed.

At each permutation, a statistic $T(q^*)_s$ is calculated and the sampling permutation distribution under the condition of independence is built.

An empirical p value is used to evaluate how the observed $T(q^*)$ is far from the null distribution.

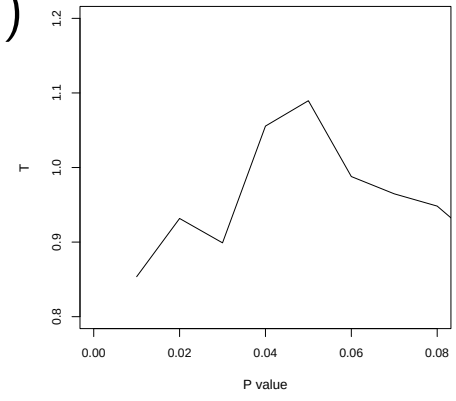




Permutation test [cont'd]

Limitations of the test:

- The uncertainty of the margins is not taken into account
- The size of the list of genes in common can be very small (typically when the total number of differentially expressed genes is small) and this can cause an instability in the estimate of T(q)



We propose a Bayesian model treating also the margins as random variables.

Bayesian model

Starting from the 2×2 table, we specify a multinomial distribution of dimension 3 for the vector of joint frequencies

		Exp B		
		DE	$\overline{DE}$	
Exp A	DE	$O_{11}(q)$	$O_{1+}(q) - O_{11}(q)$	$O_{1+}(q)$
	$\overline{DE}$	$O_{+1}(q) - O_{11}(q)$	$n - O_{+1}(q) - O_{1+}(q) + O_{11}(q)$	$n - O_{1+}(q)$
		$O_{+1}(q)$	$n - O_{+1}(q)$	n

$$Multi(\mathbf{O} \mid \boldsymbol{\theta}, n) \quad \propto \quad \theta_1^{O_{11}(q)} \theta_2^{[O_{1+}(q) - O_{11}(q)]} \theta_3^{[O_{+1}(q) - O_{11}(q)]} \times (1 - \Sigma_{i=1}^{3} \theta_i)^{[n - O_{+1}(q) - O_{+1}(q) + O_{11}(q)]}$$

The vector of parameters θ is modelled as a Dirichlet:

$$\boldsymbol{\theta} \sim Di(0.25, 0.25, 0.25, 0.25)$$

Bayesian model [cont'd]

The derived quantity of interest is the ratio of the probability that a gene is in common, to the probability that a gene is in common by chance:

$$R(q) = \frac{\theta_{O_{11}(q)}}{\theta_{O_{1+}(q)} \times \theta_{O_{+1}(q)}} \tag{1}$$

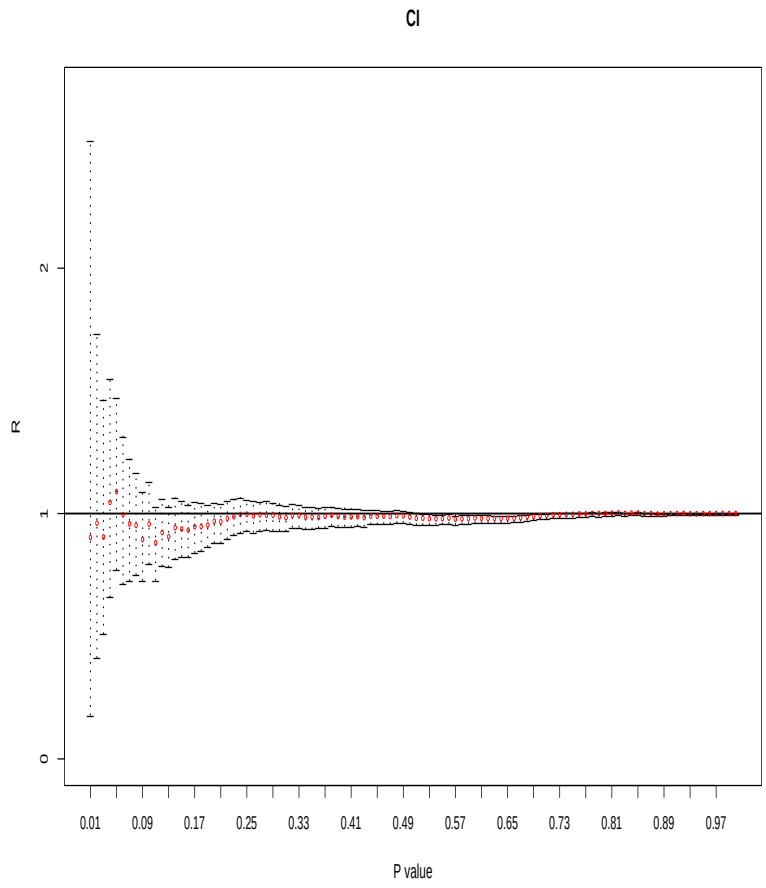
Since the model is conjugated, the posterior distribution for $\boldsymbol{\theta}$ is again Dirichlet:

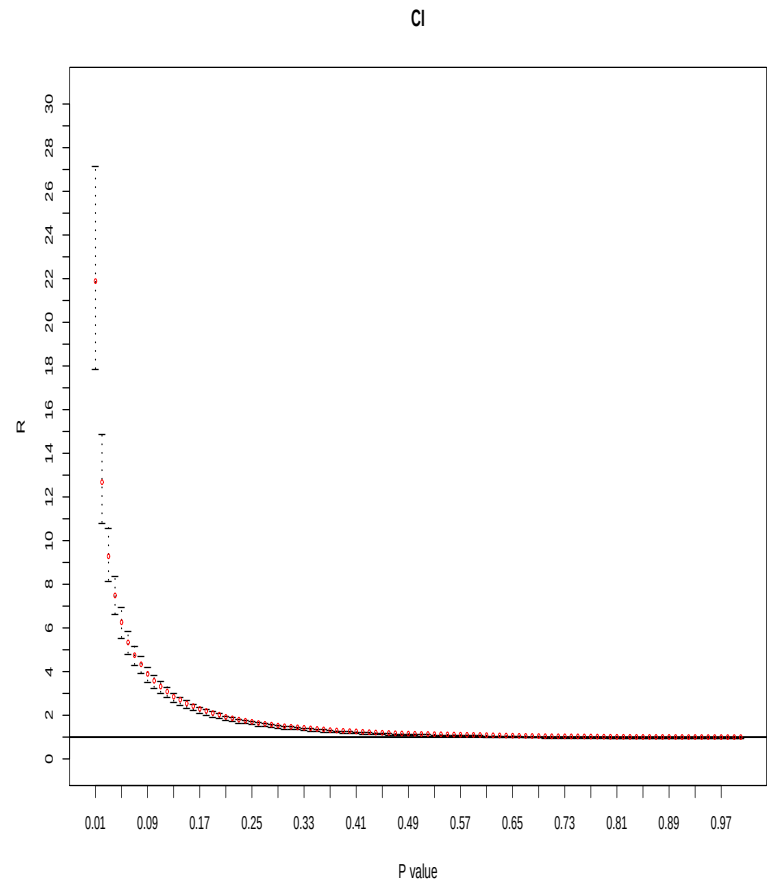
$$\begin{aligned} \boldsymbol{\theta} \quad \sim \quad & Di(O_{11}(q) + 0.25, [O_{1+}(q) - O_{11}(q)] + 0.25, \\ & [O_{+1}(q) - O_{11}(q)] + 0.25, [n - O_{1+}(q) - O_{+1}(q) + O_{11}(q)] + 0.25) \end{aligned}$$

We can obtain a sample from the posterior distribution of the derived quantity $R(q)$
Credibility intervals (CI) at 95% level can be estimated for each threshold q.

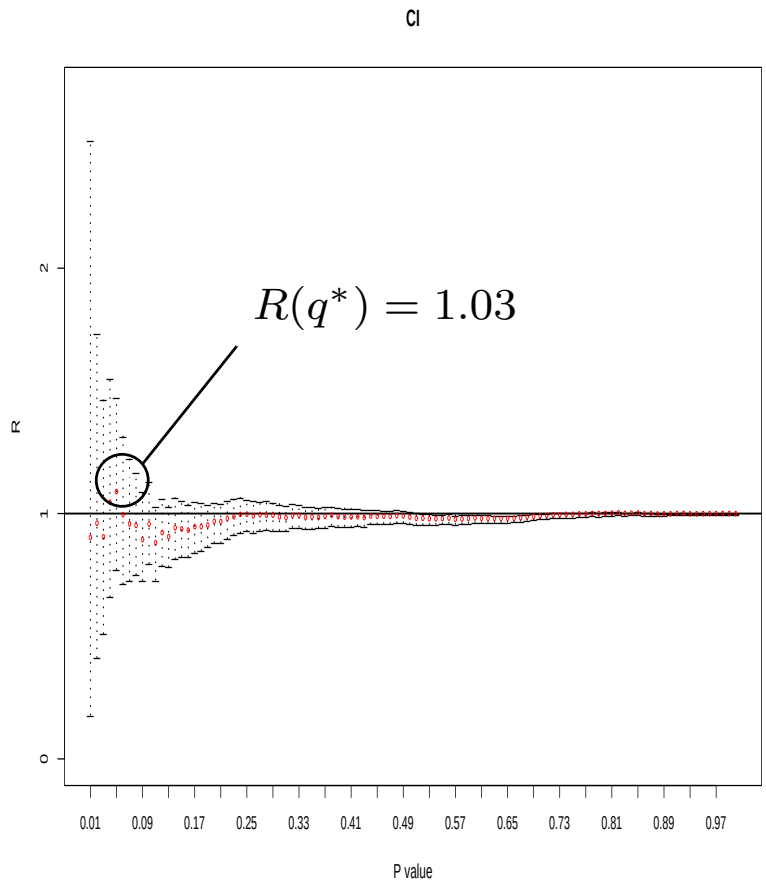
- We can calculate the maximum of R as $R(q^*) = \max_q R(q)$ considering the credibility intervals which do not include 1.

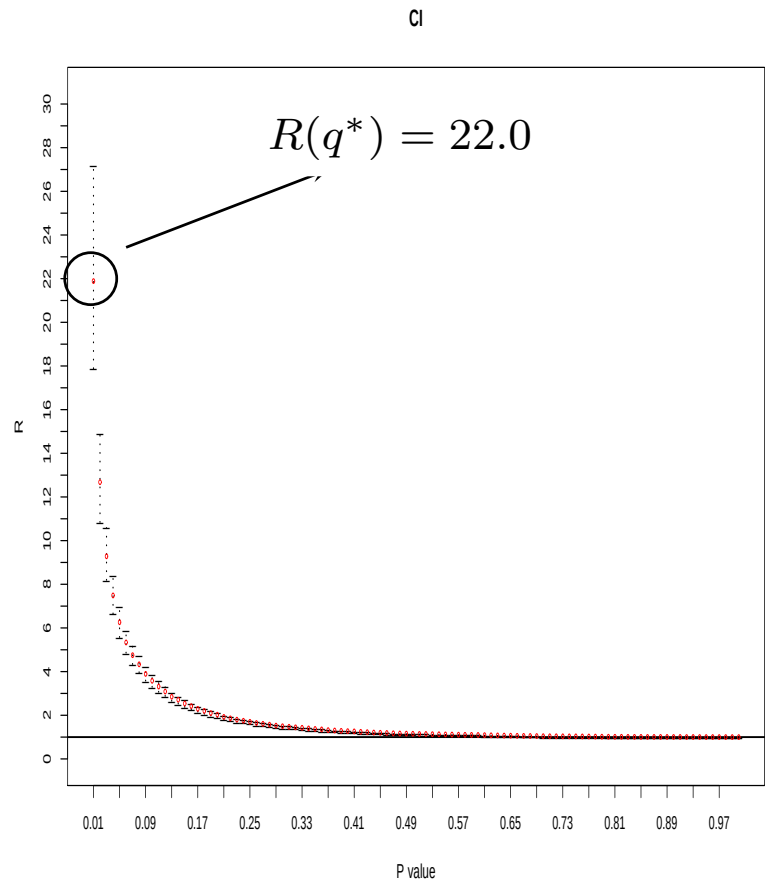
Results from the Bayesian model



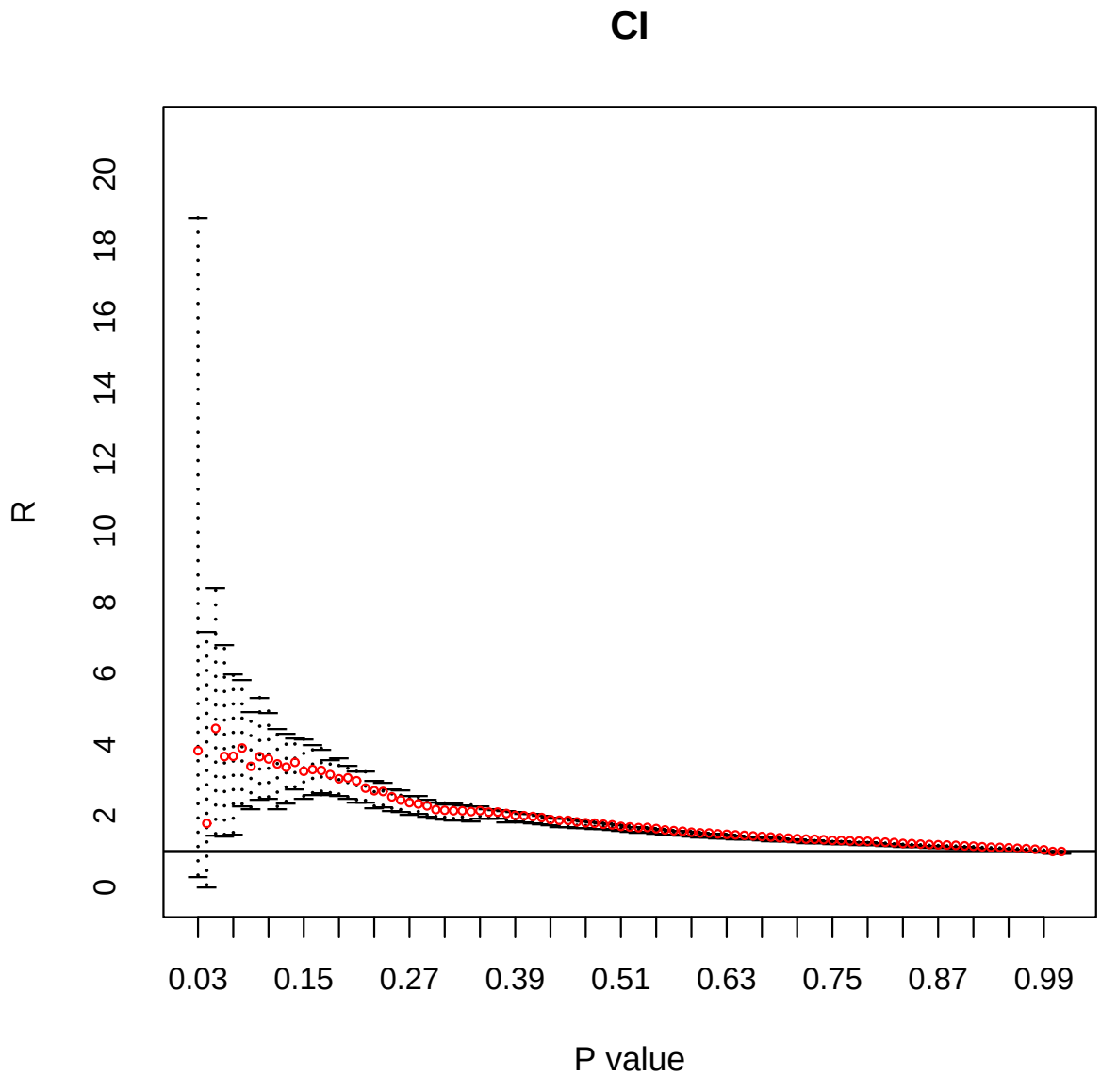


Results from the Bayesian model

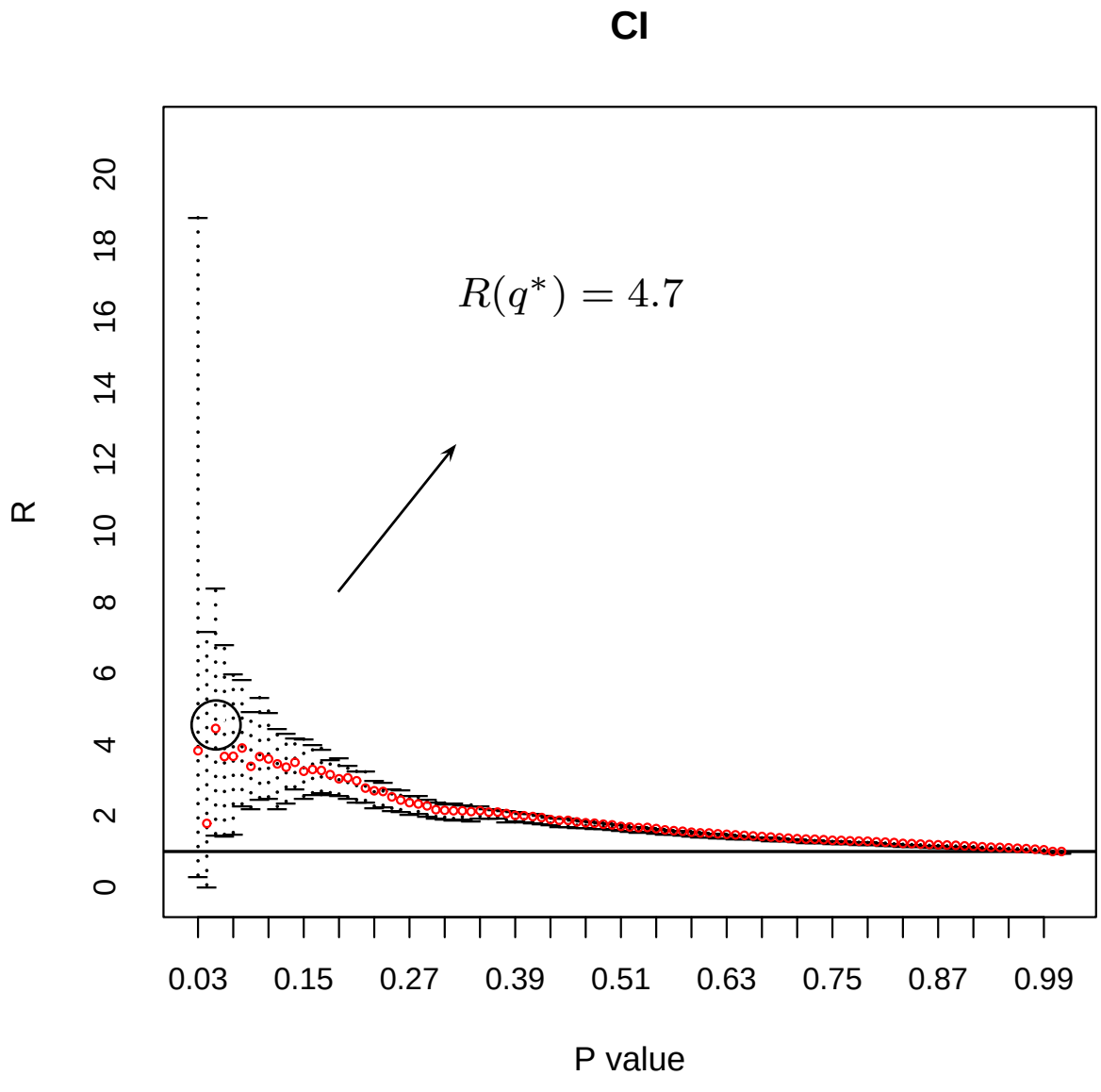




Results from the Bayesian model[2]



Results from the Bayesian model[2]



Animal experiment: Hi Fat Diet vs IRS2

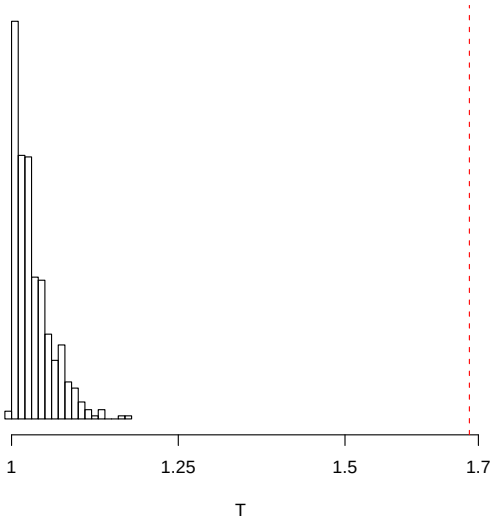
IRS2 data Expression measures are obtained for muscle of 8 weeks old mice, for two conditions (Wild Type mice and mice with a Knocked Out gene), through Affymetrix microarray. The knocked out genes is related to the pathology which occurs when the normal amount of insulin secreted by the pancreas is not able to help the body utilizing blood glucose.

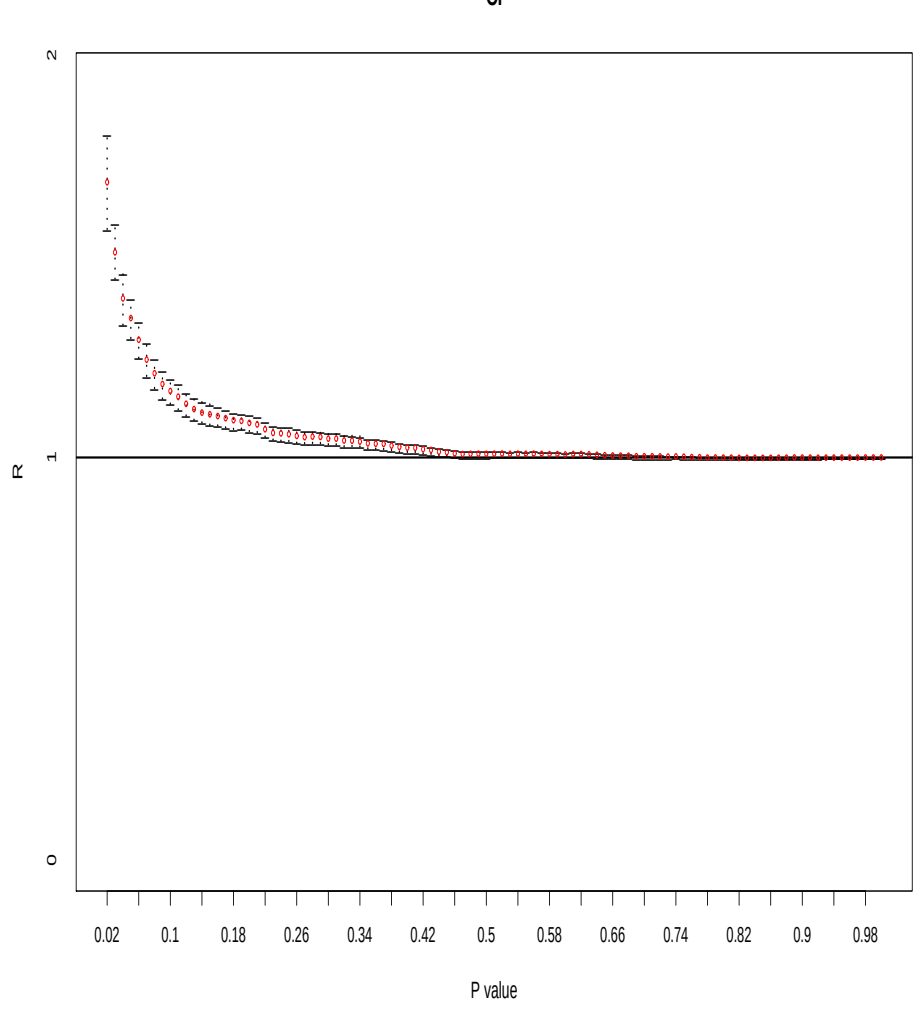
Hi Fat Diet Expression measures are obtained for 2 months old mice, for the same tissue and for two conditions related to the diet (Hi Fat Diet and Normal Fat Diet).

We consider the biological processes level. For both the experiments a list of p values is returned.

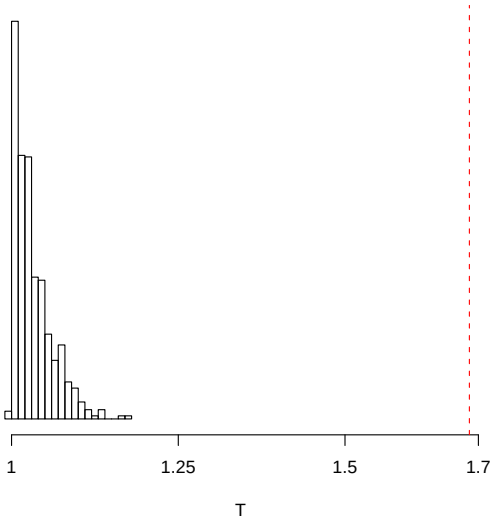
Our interest is finding genes changing due to presence of diet and knock out gene.

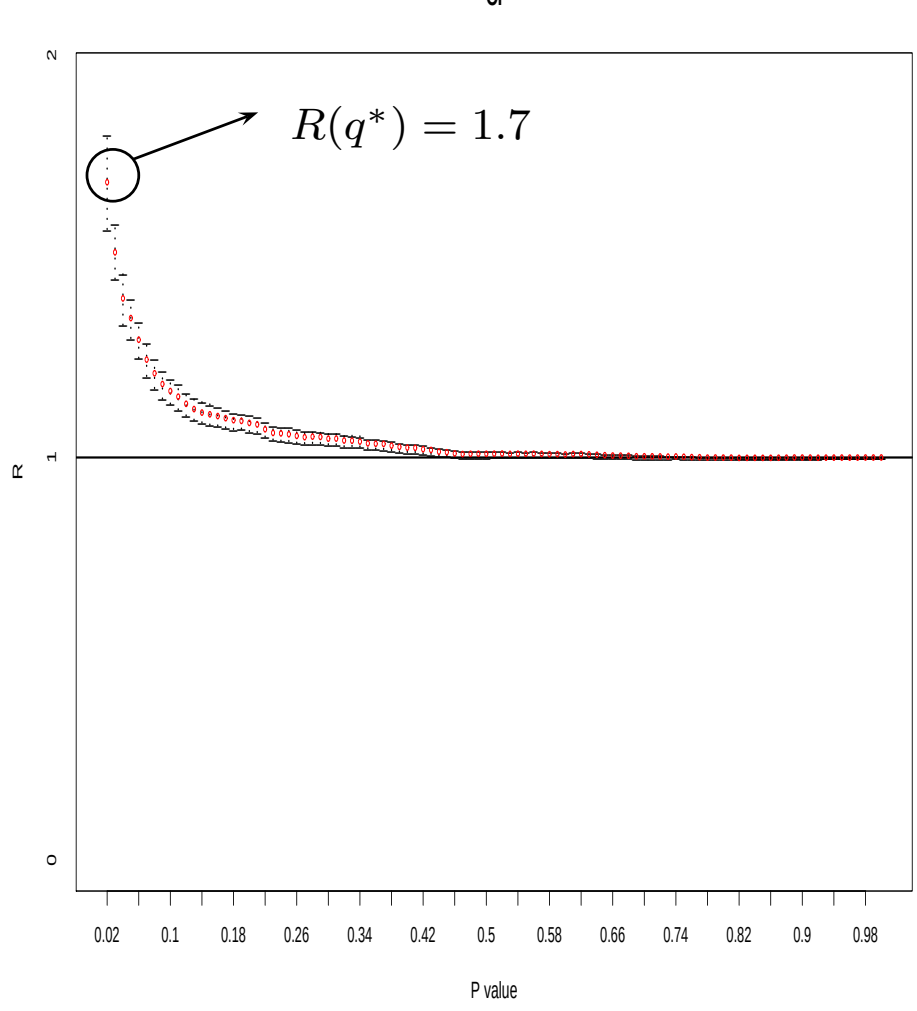
Animal experiment: Results



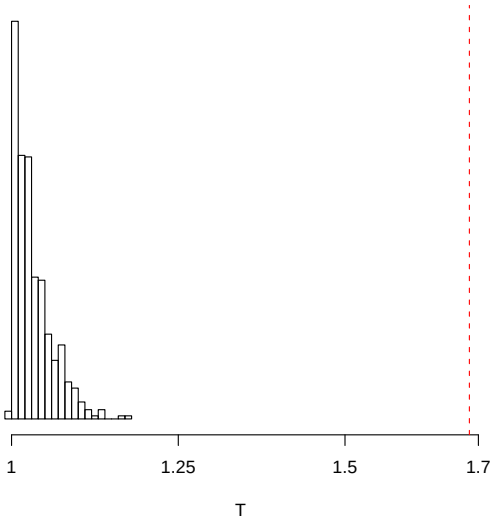


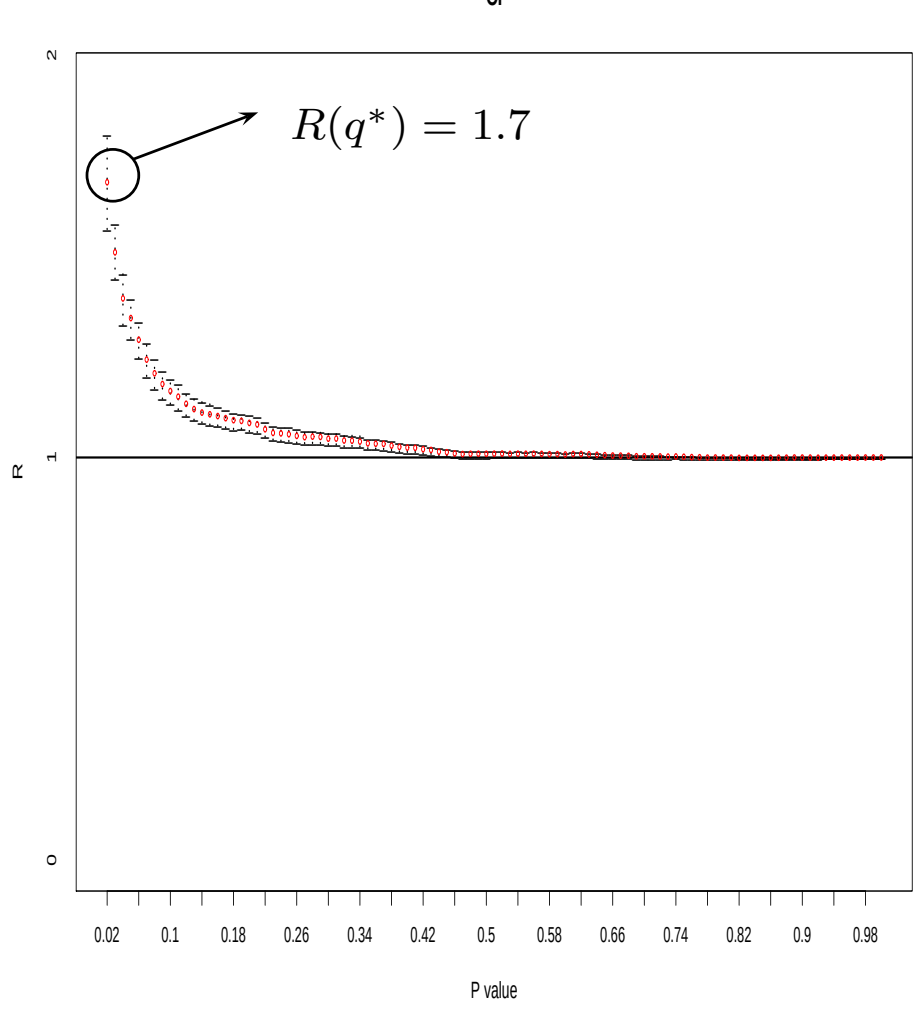
Animal experiment: Results





Animal experiment: Results





$R(q^*)$	1.7
q^*	0.01
$O_{1+}(q^*)$	586
$O_{+1}(q^*)$	1109
$O_{11}(q^*)$	318

Discussion

- We proposed a simple procedure to evaluate if two lists of differentially expressed genes are associated
- The permutation based test allows to have a first look under a model in which the marginal frequencies are fixed.
- The Bayesian model permits to enlarge the scenario, introducing variability on all the components
- It is very flexible and can be adapted to several comparisons, at different levels (e.g. gene level, biological processes level) and for different problems (e.g. between species comparison, between platform comparison)