Estadística para Bioinformática

…experiencia de un usuario no-estadista

Corrección por múltiples comparaciones: valores p ajustados.

- 1. RNA-seq – p-value
- 2. Differentially expressed genes - FDR
- 3. Epigenetics – Fischer test
- 4. GO Terms – Hypergeometric test

Identifying Differentially Expressed Genes

- Individual Hypotheses Testing

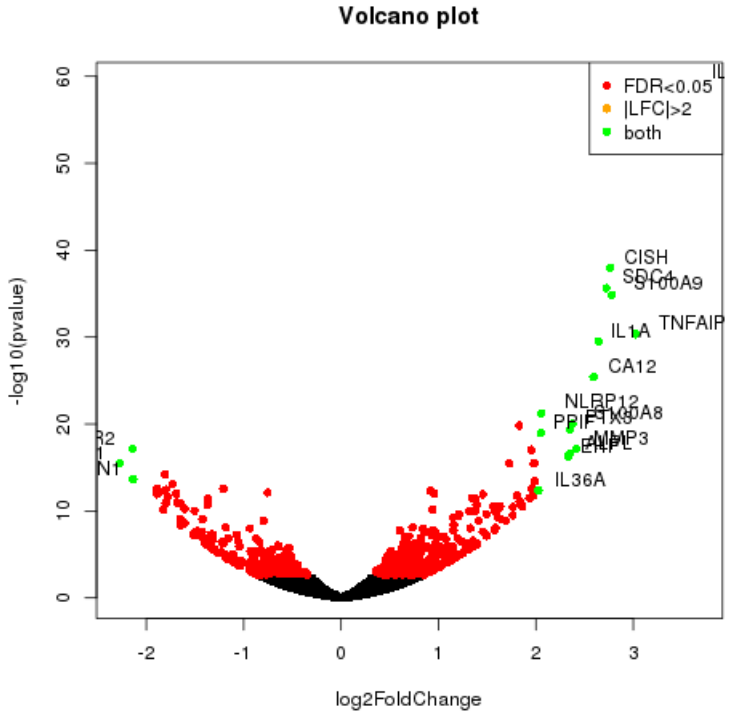
For each gene use a significance test of 0.05 level

H_0: Gene is similarly expressed

H_1: Gene is differently expressed

A t-statistic is calculated for comparing gene expression mean between the control and treatment groups.

RNA-set output

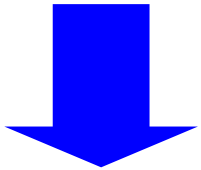


P-value?

Thousands of genes in an experiment

Thousands of hypotheses are tested

Probability of type I error (false discovery) committed increases sharply



Multiplicity problem!

Are you sure all the tests are independent?

Multiple Testing Matters! Is actually essential

Genomics = Lots of Data = Lots of Hypothesis Tests

A typical RNA-seq experiment might result in performing 6,000/25,000 separate hypothesis tests. If we use a standard p-valuecut-off of 0.05, we'd expect ? genes to be deemed "significant" by chance.

$$P(\text{Making an error}) = \alpha$$

$$P(\text{Not making an error}) = 1 - \alpha$$

$$P(\text{Not making an error in m tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in m tests}) = 1 - (1 - \alpha)^m$$

Experiments in finding people with paranormal powers:
Joseph Rhine (1950)

1000 people guessed the sequence of 10 cards: red or black?

12 persons guessed 9 of 10 cards, two of them all 10 cards

All these “physics” in further experiments did’t confirm their paranormal abilities

What did really happen?

Probability to guess all 10 cards = $\left(\frac{1}{2}\right)^{10} \approx 0.00098$

Probability to guess 9 cards = $10\left(\frac{1}{2}\right)^{10} \approx 0.0098$

Probability to guess 9 or all 10 cards = $11\left(\frac{1}{2}\right)^{10} \approx 0.0107$

Chances to find a "psychic"
among 100 persons = $1-(1-0.0107)^{100} \approx 0.66$

Chances to find a "psychic"
among 1000 persons = $1-(1-0.0107)^{1000} \approx 0.9998$

How to avoid false discovery

During m independent statistic test with α significance level, the probability of at least one false discovery should be

$$1-(1-\alpha)^m < 0.05$$

$$\alpha = 1-(1-0.05)^{1/m} \approx \frac{0.05}{m}$$

Bonferroni correction: during m independent statistic tests only those results are significant, for which

$$p < \frac{0.05}{m}$$

What Does Correcting for Multiple Testing Mean?

- When people say "adjusting p-values for the number of hypothesis tests performed" what they mean is ***controlling the Type I error rate***

- Very active area of statistics - many different methods have been described

- Although these varied approaches have the same goal, they go about it in fundamentally different ways

The False Discovery Rate (FDR) criterion

Benjamini and Hochberg (95) :

R = # rejected hypotheses = # discoveries

V of these may be in error = # false discoveries

The error (type I) in the entire study is measured by

$$Q = \frac{V}{R} \qquad R > 0$$
$$= 0 \qquad R = 0$$

i.e. the proportion of false discoveries among the discoveries (0 if none found)

$$FDR = E(Q)$$

Does it make sense?

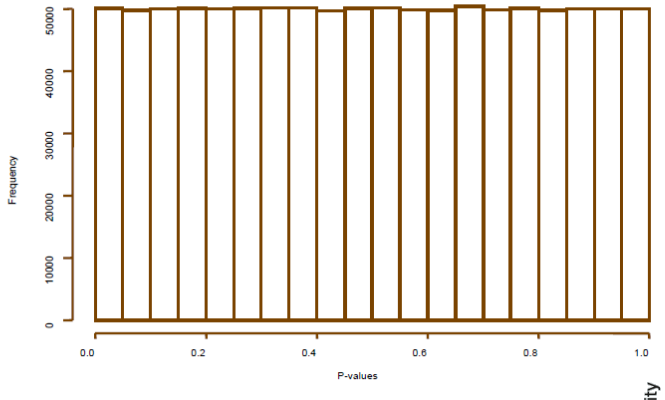
What's a q-value?

- q-value is defined as the minimum FDR that can be attained when calling that "feature" significant (i.e., expected proportion of false positives incurred when calling that feature significant)

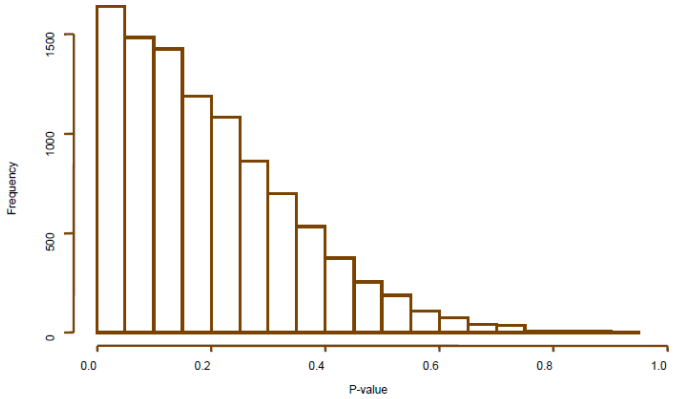
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)

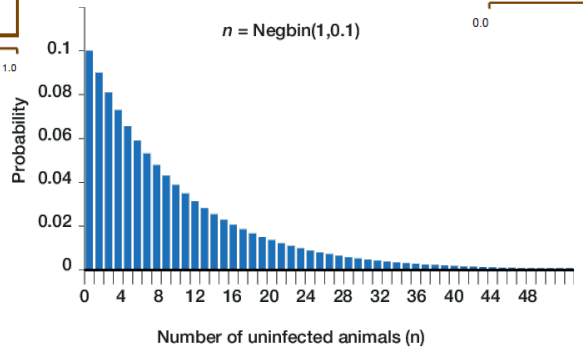
- Thus, in an array study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives

- Under the null hypothesis p-values are expected to be uniformly distributed between 0 and 1



Under the alternative hypothesis p-values are skewed towards 0

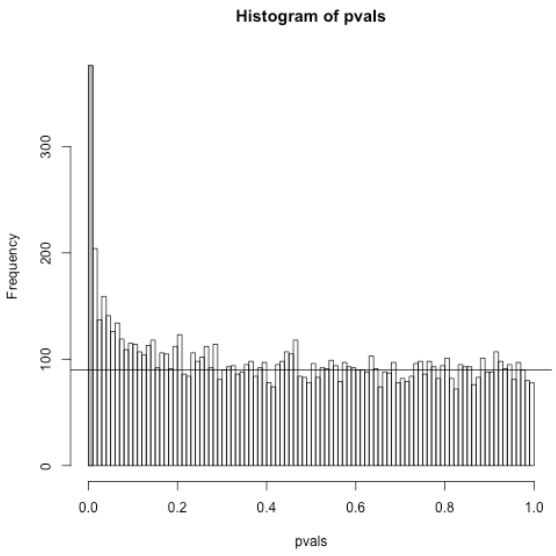


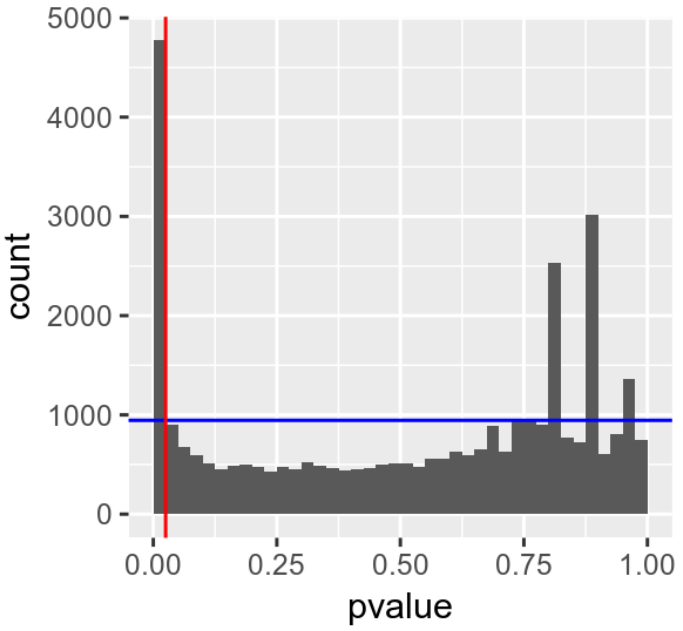


Negative binomial distribution

FDR

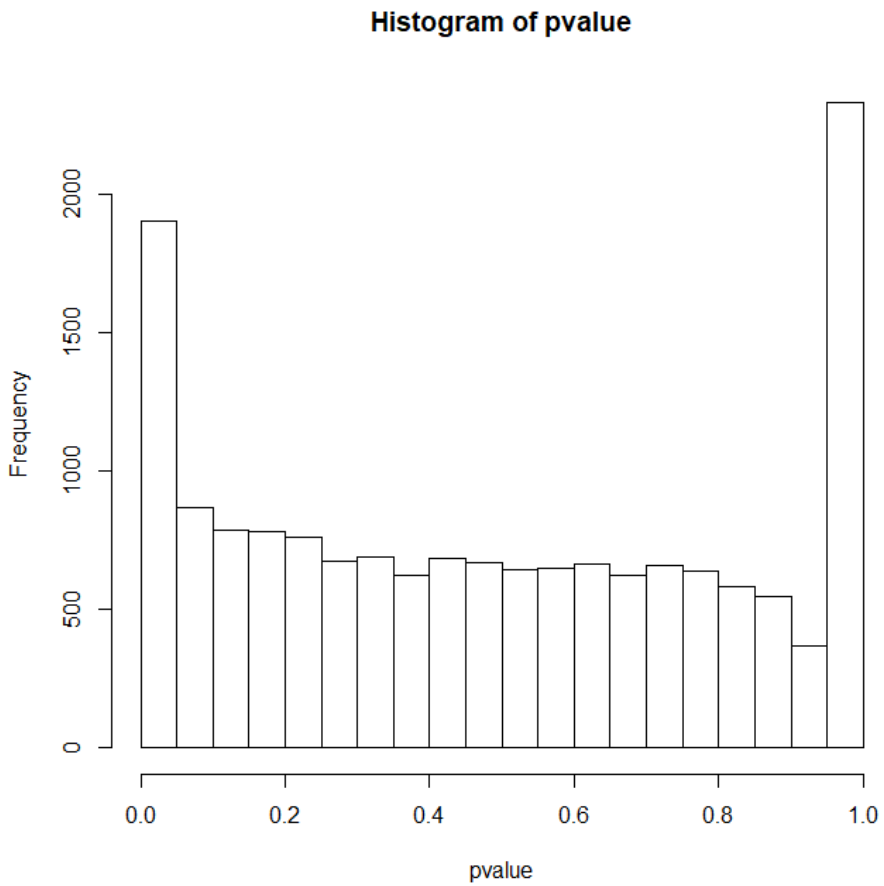
False discovery rate (FDR) is the expected proportion of Type I errors among the rejected hypotheses

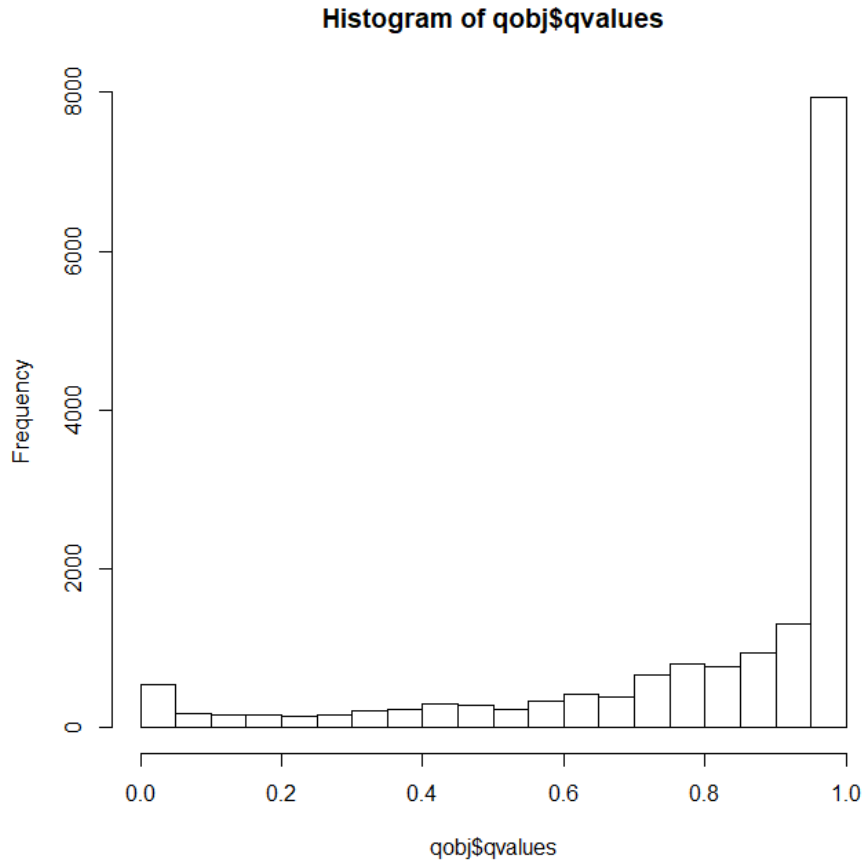




Example

```
#Install qvalue package
if (!requireNamespace("BiocManager", quietly = TRUE))
   install.packages("BiocManager")

BiocManager::install("qvalue")


#call package
library(qvalue)

#add your file
pvalue <- scan("C:\\Users\\Usach\\Desktop\\Pvalues.csv")
head(pvalue)

#visualise the data distribution
hist(pvalue)

#run q_value

qobj <- qvalue(pvalue, fdr.level=0.05)
summary(qobj)
hist(qobj$qvalues)
plot(qobj)

#save pdf
pdf("plot_q_values.pdf")
plot(qobj)
dev.off()
```


Histogram of pvalue
Frequency
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
pvalue


Histogram of qobj$qvalues
Frequency
0
2000
4000
6000
8000
0.0
0.2
0.4
0.6
0.8
1.0
qobj$qvalues

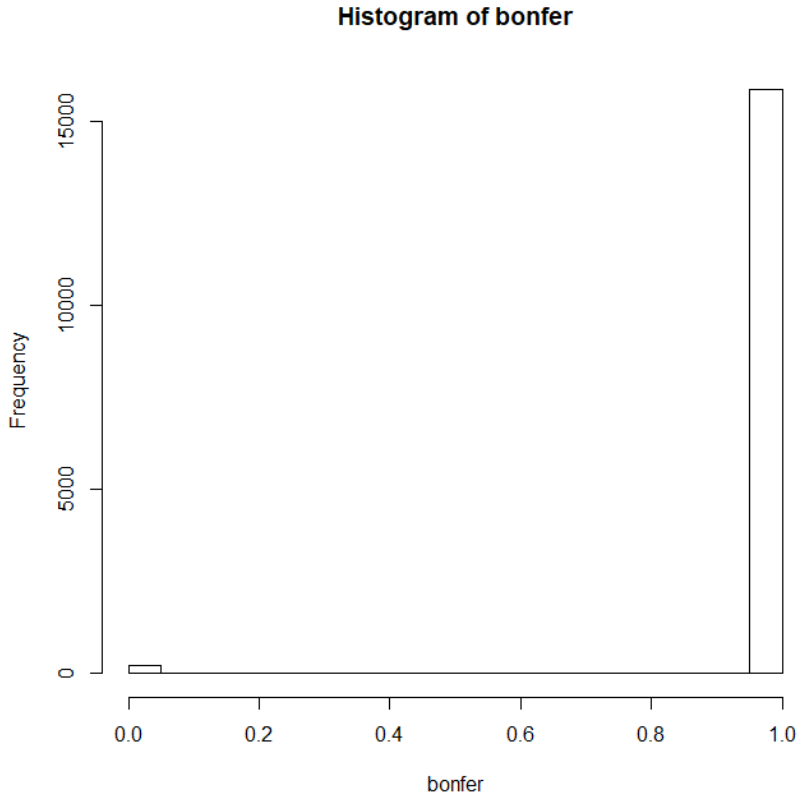
Critisism of FDR approach

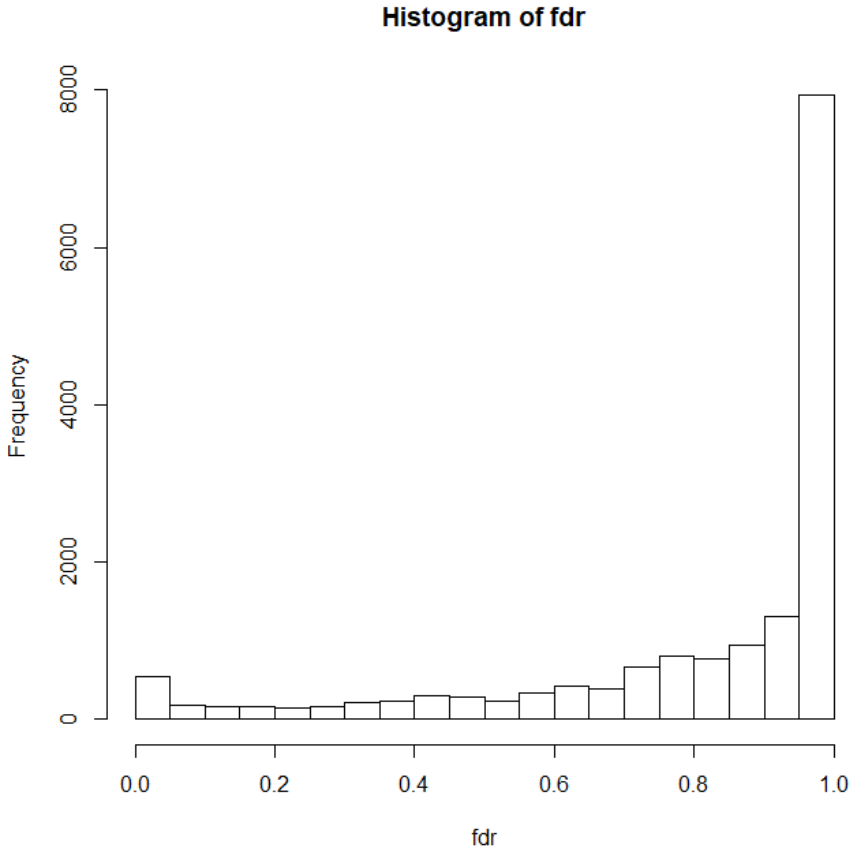
- It is ”possible to cheat”:
  - You can choose what p-values to use to alter the distribution and the outcome

Bonferroni vs FDR

Bonferroni: Very simple method for ensuring that the overall Type I error rate of α is maintained when performing m independent hypothesis tests

```
bonfer <- p.adjust(pvalue, method = 'bonferroni', n = length(pvalue))
head(bonfer)
summary(bonfer)
hist(bonfer)

fdr <- p.adjust(pvalue, method = 'fdr', n = length(pvalue))
head(fdr)
summary(fdr)
hist(fdr)
matplot(pvalue, fdr)
matplot(pvalue,bonfer)
```


Histogram of bonfer
Frequency
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
bonfer


Histogram of fdr
Frequency
0
2000
4000
6000
8000
0.0
0.2
0.4
0.6
0.8
1.0
fdr

Epigenetics – Fisher test/chi-square

What is Fisher's Exact Test of Independence?
Fisher's Exact Test of Independence is a statistical test used when you have two nominal variables and want to find out if **proportions** for one nominal variable are different among values of the other nominal variable. For experiments with small numbers of participants (under around 1,000), Fisher's is more accurate than the chi-square test or G-test.

A **chi-square test for independence** compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
- A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.
- A **very large chi square test statistic** means that the data does not fit very well. In other words, there isn't a relationship.

When to use the ‘contingency table’?

FISCHERS EXACT TEST CONTINGENCY TABLE

	Positive	*Negative*
Chemo-sensitive	*x*	*x*
Chemo-resistance	*x*	*x*

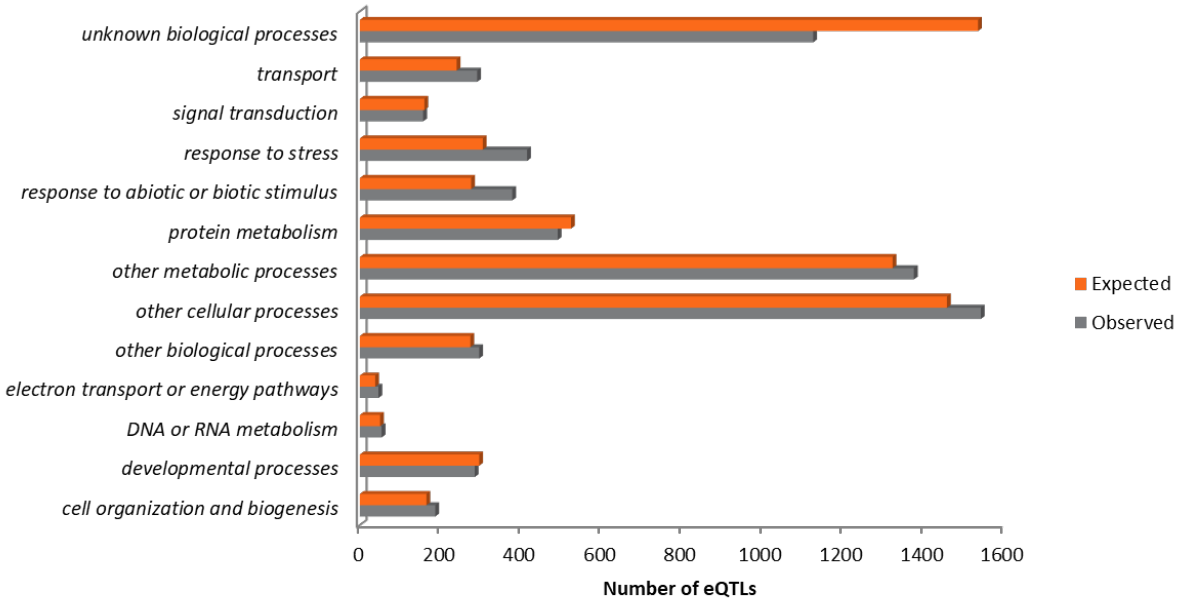
Figure 24: shows a *2 x 2* contingency table for Fischer’s exact test. The values (*x*) is entered for the number of patients that show a chemo-sensitive or chemo-resistant response against a positive or a negative IHC score for a particular staining localization.

- 2 treatments x 2 accessions/strains
- 2 variables x 2 targets

Ejemplo

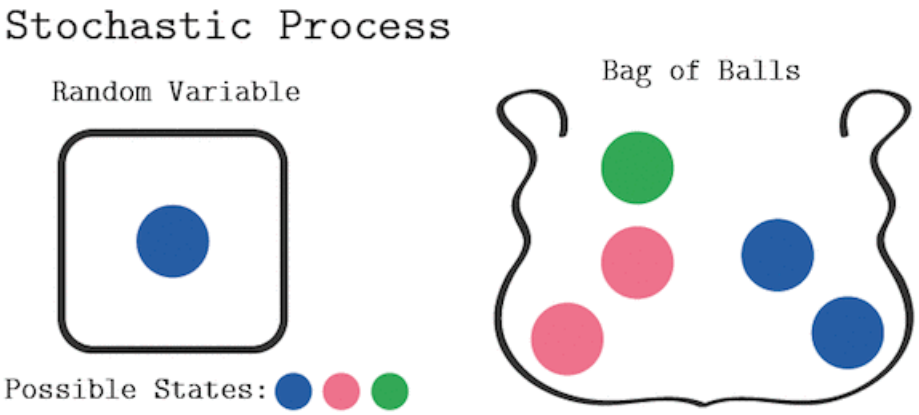
	sRNA_1	sRNA_2
Planta Sin bicho	10	10
Planta Con bicho	10	30

Gene Ontology Term enrichment
Observed vs expected



Hypergeometric test

Binomial Test
50/50



Hypergeometric test
1/50

Example

Universe: 18840 balls total
red balls in the universe: 6680

Sample: 382 balls
total red balls in the sample: 160

I would like to estimate if the percentage of red balls in my sample is significantly different from the percentage of reds in universe

dhyper(x, m, n, k, log = FALSE)

x, q	vector of quantiles representing the number of white balls drawn without replacement from an urn which contains both black and white balls.
m	the number of white balls in the urn.
n	the number of black balls in the urn.
k	the number of balls drawn from the urn.
p	probability, it must be between 0 and 1.

dhyper(160, 6680, 12160, 382, log = FALSE)
P = 0,013

Evidence for directional allelic effect