GO Enrichment

Broňa Brejová

12.11.2020

Gene list analysis

Many analyses yield lists of genes. Examples:

- genes with positive selection in comparative genomics
- overexpressed or underexressed genes in expression analysis
- genes regulated by a specific transcription factor

Some of genes in a list will have a known function, others may be less studied

What to do with such a gene list?

- Look at several interesting candidates and study them in detail (bioinformatics / wet lab)
- Determine if the whole set is enriched in genes with some property
  – for example, genes under positive selection are often enriched for functions in immunity
  – this is caused by evolutionary pressure from pathogens

Example from Kosiol et al 2008

16,529 genes total

70 genes innate immune response (0.4% of all genes)

400 genes positive selection

8 genes positive selection + innate immune response (2% of pos. sel.)

Contingency table

	Pos.sel.	No pos.sel.	Total
Immunity	8 (n_{ip})	62	70 (n_i)
Not immunity	392	16067	16459
Total	400 (n_p)	16129	16529 (n)

Observations:

Innate immune response only a small fraction of pos.sel.

But large enrichemnt from 0.4% to 2%

Is it by chance (due to small numbers)?

Example from Kosiol et al 2008

	Pos.sel.	No pos.sel.	Total
Immunity	8 (n_{ip})	62	70 (n_i)
Not immunity	392	16067	16459
Total	400 (n_p)	16129	16529 (n)

Is enrichment due to chance?

Want p-value:

What would be a chance of obtaining such an enrichemnt if positive selection and role in innate immune response independent
(null hypothesis)

Null hypothesis

	Pos.sel.	No pos.sel.	Total
Immunity	8 (n_{ip})	62	70 (n_i)
Not immunity	392	16067	16459
Total	400 (n_p)	16129	16529 (n)

Urn with $n_i = 70$ white balls and $n - n_i = 16459$ black balls

Draw $n_p = 400$ balls from the urn

Denote by X the number of white balls in the selection

On average we expect $E(X) = n_p(n_i/n) = 1.7$

In reality we see $n_{ip} = 8$ pos. sel. genes with role in innate immunity

This is $4.7\times$ more

How likely is this by chance?

Null hypothesis

Urn with $n_i = 70$ white balls and $n - n_i = 16459$ black balls
Draw $n_p = 400$ balls from the urn
Denote by X the number of white balls in the selection

Variable X has **hypergeometric distribution**:

$$\Pr(X = n_{ip}) = \binom{n_i}{n_{ip}} \binom{n - n_i}{n_p - n_{ip}} \Big/ \binom{n}{n_p}$$

P-value is $\Pr(X \geq n_{ip}) = \Pr(X = n_{ip}) + \Pr(X = n_{ip} + 1) + \ldots$
Tail of the distribution

In our case $\Pr(X \geq 8) = 0.00028$

This is called **Hypergeometric** or Fisher's exact test
It can be approximated by χ^2 **test**

Multiple testing correction

Often we do many tests of the same type, for example

- Test 1000 genes for positive selection, select those with p-value ≤ 0.05
- Test enrichment of 1000 functional categories in a list of genes, selectthose with p-value ≤ 0.05

Problem: If each category has 5% chance of being there by chance, we expect 50 purely random results.

If the total number of positive tests was 100, half of them were false.

Multiple testing correction: lower threshold on p-value so that false positives do not constitute a large portion of results

Several techniques, e.g. FDR (false discovery rate)