Replic2 – software for evaluating replication datasets

Jurg Ott 8 Feb 2025

Scenario

Assume you have two independent case-control datasets, e.g. males and females, that have been genotyped for M_1 and M_2 variants (SNPs), respectively. You select a subset of these, N_1 and N_2, as being particularly promising, for example, because they are significant or because they show odds ratios exceeding 10 each. Upon combining N_1 and N_2 into a single dataset (spreadsheet), you find that N_{12} of these variants are common between the two groups. How likely is this occurring just by chance, or can you conclude that group 2 represents a significant replication for group 1 by virtue of these N_{12} common variants?

Replic2 program

This question can be answered by simulating the following null situation: You randomly pick N_1 variants out of a larger set of M_1 variants for group 1 and analogously for group 2, then combine them to see how many random matches you find, N_{12rand}. You do this many times to obtain, for example, 100,000 such N_{12rand} numbers. The empirical significance level, p, is then given by the proportion of N_{12rand} values equal to or exceeding the observed number N_{12} of matches.

The easiest way to use the Replic2 program is to run it in a command box, either in Windows or Linux, followed on the command line by the name of an input file. For example, you type `Replic2 sample.in`, where the `sample.in` file contains the following lines:

```
1000 1000 Two large sample sizes
100 100   Two sets of SNPs selected from the large sample sizes
100000    Number of replicates for p-value calculation
3         Number of SNPs N12, common to the two groups...
5         ... repeat as often as desired
8
10
20
30
50
-1        Finish with "-1" or just end the input
```

The resulting output file will then list each number N_{12} of common SNPs and their associated p-values. Because we take the observed data as one of the null datasets, there will never be a zero p-value. For example, with 100,000 replicates, the smallest possible significance level is $p = 0.00001$, which may be interpreted as $p \leq 0.00001$.

The program currently works only for a relatively small number of variants. As time permits, I will modify the text (Pascal script) to accommodate larger numbers. The program is available on github.