Myriads Manual v 1.2

Antonio Carvajal-Rodríguez

Departamento de Bioquímica, Genética e Inmunología

Universidad de Vigo, Vigo 36310, Spain

Email: myriads@uvigo.es

Web: http://myriads.webs.uvigo.es

Contents

VERSION

Current version is 1.2. In this version:

1) New FWER and FDR controlling procedures have been added. The new FWER methods are Bonferroni and Hommel. The FDR methods are Benjamini & Yekutieli (BY) and the two-stage linear step-up procedure from Benjamini, Krieger & Yekutieli (BKY).

2) A permutation based method, maxT, has been implemented for two-sample *t*-test with equal or unequal variances.

3) New input format only required when maxT option is chosen. If the maxT option is selected then the data should be in the new format (matrix with genes in rows and the samples in columns , see below the corresponding section).

4) Option for selecting a subsample from the data matrix, performing the analyses, and saving the subsample data matrix to a file.

5) New option (not applicable to maxT) for transforming the given *p*-values to *p*-values obtained from a standard normal scale.

THE PROGAM

The Myriads software performs three different tasks, namely, multiple testing adjustment, detection of dependency and *p*-value simulation (or gene expression like data simulation). The simulation mode does not require input file.

Download

The program can be downloaded from

http://myriads.webs.uvigo.es

Ready-to-use executables

Windows: Myriads1.2.exe

Linux: Myriads1.2

Compiling the program manually (linux/MAC)

Go to the source folder that should include a file called Makefile, and just type make.

INPUT

The program can be used interactively to compute just the number of significant tests after SGoF correction. In this case, the program asks the user to introduce the number of tests and the number of values below the nominal level α, and then computes the number of values that remain significant after correction.

However, for obtaining adjusted *p*-values and/or performing different multiple testing corrections, a list of *p*-values or normalized gene expression data (for maxT algorithm) is required.

***p*-values**

For multiple testing correction, Myriads requires a list of *p*-values with SGoF-like input file format (Fig. 1). The input file should have an integer number indicating the total number of tests and two columns with pairs of identifiers and *p*-values. The identifier can be a number or a character string. The list of *p*-values does not need to be sorted, indeed, is required to be not if the dependency test is performed. The input file name is assumed by default to be PvalMyriads.dat but any other name can be given by the user.

There is no limit on the number of *p*-values other than computer memory.

```
7
1       0.003
id2     5e-3
gen3    0.998
id4     0.34
5       0.01
6       0.004
7       0.445
```

Fig. 1. Myriads p-values input file example.

Normalized gene expression data

A different input format is required for performing the step-down maxT procedure (Westfall and Young 1993). The normalized gene expression data format in Myriads consists in a text file, the first row has information about the class or response measurement (1 or 2) and the remaining rows have data, one row per gene. The first column contains the gene name/ID and the remaining columns represent the different samples. Fig 2 represents the expression data of 3 genes (1, id2 and Gen3). The first row contains the class values with sample size 5 and 4 for the first and second class respectively.

	1	1	1	1	1	2	2	2	2
1	-1.5088	2.0596	1.8461	-1.921	-0.1459	1.2070	0.1395	0.7120	-0.4932
Id2	1.0025	1.8061	0.2639	2.1447	-0.1298	0.7901	-0.3681	1.723	0.9699
Gen3	-0.8563	0.2216	0.8952	-1.8183	0.6688	0.3350	0.2892	-1.6688	-0.5058

Fig. 2. Myriads normalized gene expression data input file example.

Missing values should be coded as NA, the corresponding row is skipped.
The default name for the gene expression input file is MaxTfile.txt but the user can give any desired name.

Subsampling

Under the command line and with the maxT option selected (-*maxT* 1) there is the possibility of analysing a data subsample by means of the tag -*sub ini end* so that the subsample from *ini* to *end*, both included, is considered and the subsequent tests and correction are performed only over the subsample data. Also, the data subsample is written to a file. Note that this option can be used to generate a desired subsample data file (without waiting for maxT) simply by setting -maxt 1 with just one permutation and the desired subsample. If *ini* ≥ *end* the whole sample is analysed.

METHODS

In the flow chart of Fig. 3 the different methods available in Myriads are given jointly with their adequacy for different experimental settings.

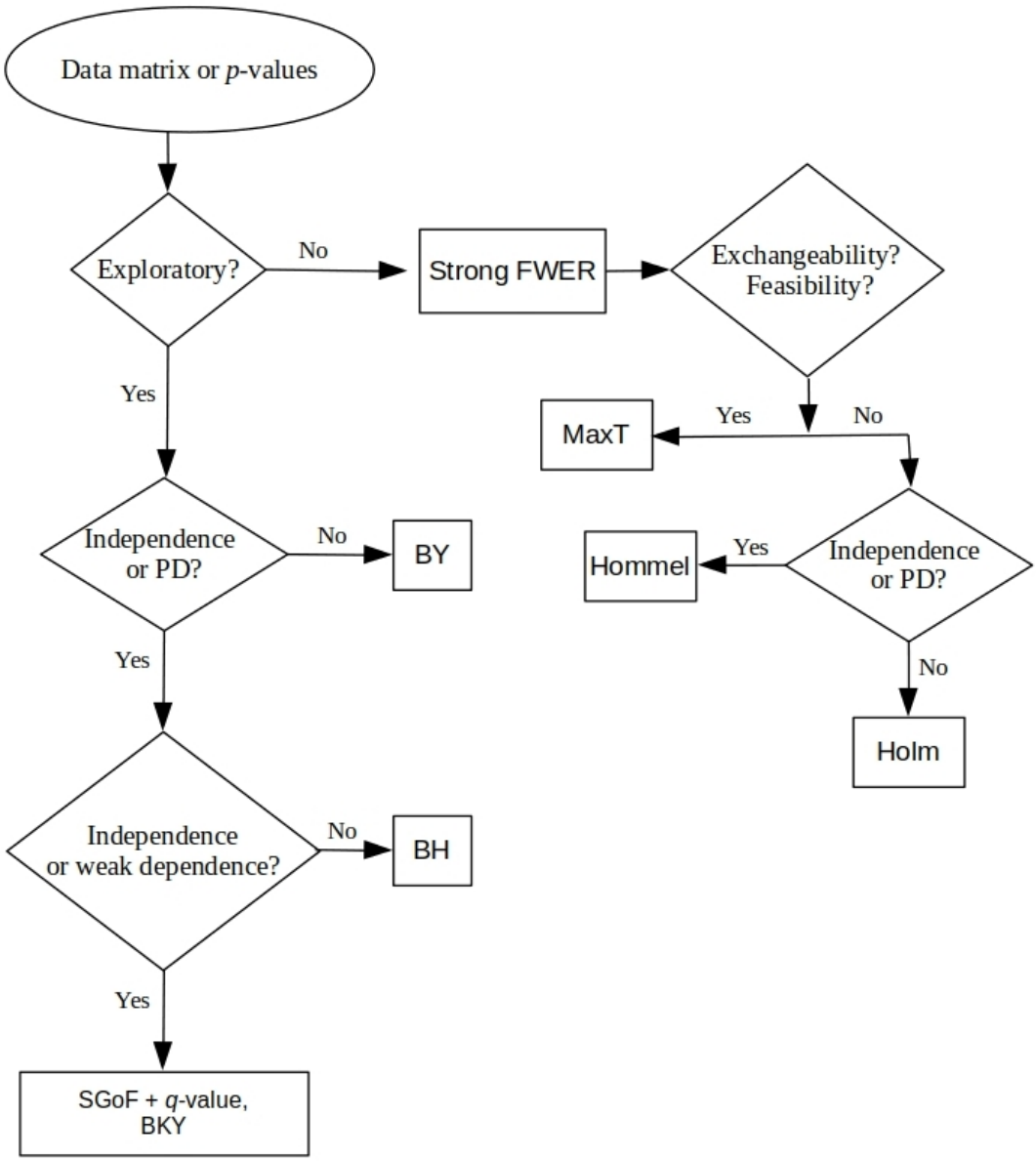

Fig. 3. Decision flowchart for multiple testing procedures. PD: positive dependence.

FWER methods

Myriads performs the following FWER controlling methods: Bonferroni, Holm (sequential Bonferroni), Hommel and maxT, that control FWER in the strong sense and SGoF that controls FWER in the weak sense.

MaxT

The permutation-based maxT method is optional. The method is implemented for two-sample *t*-tests with equal or unequal variances and is based in the algorithm given in

(Ge et al. 2003). The results are stored in a folder called MaxT within the Myriads output folder. The method can be selected via the command line or the option 12 in the general menu (Fig. 4). It permits setting a tuning constant (default is 0) which sums to the standard error in the *t* statistic as in (Tusher et al. 2001). It performs equal/unequal variance two/one tail *t*-tests. In the menu version it only performs the two tailed and the right tailed version, the command line version allows two tailed and left and right tailed versions. Finally, the user is asked to let the default $B = 0.5\times 10^6$ permutations or choose the desired number and to input the data file name or use the default one which is MaxTfile.txt.

For sample sizes n_1 and n_2, the maximum number of permutations between classes is $P = n!/(n_1!n_2!)$ where $n = n_1 + n_2$. If $B \geq P$ or $P \leq 10^4$ or $n \leq 14$ the maximum number P of permutations is performed otherwise B +1 permutations are executed so that the minimum adjusted *p*-value that can be obtained is $1/P$ or $1/(B+1)$.

The maxT method can also be selected via command line arguments with the tag -*maxt* 1; similarly equal or unequal variance tests can be selected by the tag -*eqvar* and twotail test by -*twotail* or right or left tail by -*onetailR*. The tuning constant s0 can be settled by the tag -*s0* and the number of permutations by the tag -*B*. So, the line

./Myriads1.2 -maxT 1

is equivalent to the default value arguments

./Myriads1.2 -maxT 1 -file MaxTfile.txt -eqvar 0 -twotail 1 -onetailR 1 -s0 0 -sub 0 0 -B 500000.

See the command line section for details.

FDR methods

Myriads perform the following FDR controlling methods: Benjamini & Hochberg (BH, Benjamini and Hochberg 1995), Benjamini & Yekutieli (BY, Benjamini and Yekutieli 2001), Benjamini, Krieger and Yekutieli (BKY, Benjamini et al. 2006) adaptive method, the sliding linear model (SLIM, Wang et al. 2011) and the BonferroniSEV modified from (Li et al. 2017) to incorporate SLIM π_0 estimation.

PROGRAM USAGE

Interactive menu options

When called without arguments, the program offers a menu (Fig. 4)

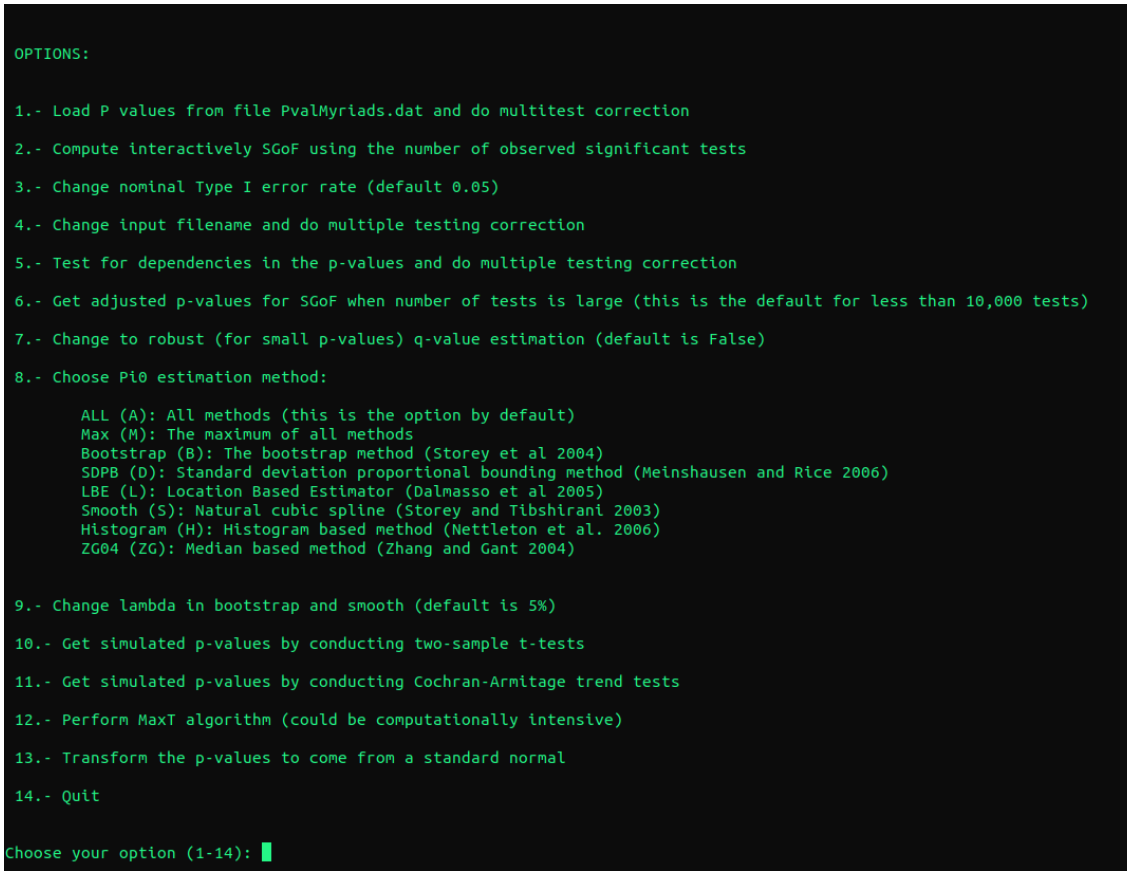

Fig. 4. Myriads general menu.

Choosing the option 1 from the menu executes the default parameter values (see below).

If the program is called with any argument, the menu is skipped and the command line execution mode is performed.

Command line options for multiple testing correction

We also can skip the menu and run the default options from the command line by passing the tag *-default*. Thus, the calling

Myriads -default

is equivalent to the first option of the menu and to the line with arguments:

Myriads -simulation 0 -inpath . -outpath ./Myriads_output -inputfile PvalMyriads.dat -outputfile Myriads -SL 0.05 -epi0 A -grid 5 –robust 0 -dependence 0 -lag 1 -minblock 50 -light 0 -AdjustLarge 0 -normalize 0 -maxT 0 -eqvar 0 -twotail 1 -onetailR 1 -s0 0 -sub 0 0 -B 500000

Note that if the tag -*default* is utilized, every argument after it will be ignored. This means that the line

-dependence 1 -default

is the same as

-dependence 1

but the line

-default -dependence 1

is the same as

-default (i.e. -default -dependence 0)

The utility of -*default* is to skip the menu without requiring the specification of any argument.

The explanation of each argument is as follows:

-*simulation* <INTEGER> Simulation (1|2) or analysis (0) mode. By default is 0.

-*inpath* <STRING> Specifies the path to the input file (default is the working directory represented by a dot).

-*outpath* <STRING> Specifies the path to the output file (default is ./Myriads_output).

-*input* <STRING> Specifies the input file name (default is PvalMyriads.dat).

-*output* <STRING> Specifies the output file name.

-*SL* <DOUBLE> The significance level (default 0.05).

-*epi0* <DOUBLE> Choose the π_0 estimation method (option 8 from the menu, default is A).

-*grid* <DOUBLE> Change lambda in bootstrap and smooth (option 9 from the menu).

-*robust* <BOOLEAN> Change (if value is 1) *q*-value estimation to robust (option 7 from menu).

-*dependence* <BOOLEAN> Include the dependence test in the analysis (1) or not (0, default).

-*lag* <INTEGER> The lag for the autoregression in the dependence analysis (default is 1).

-*minblock* <INTEGER> The minimum block size for the dependence test (default is 50).

-*light* <BOOLEAN> Activates (1) the light output mode (default is 0).

-*AdjustLarge* <BOOLEAN> Activates (1) the adjusted *p*-values for SGoF when the number of tests is large (option 6 from menu).

-*normalize* <BOOLEAN> Transforms (1) the list of given *p*-values to *p*-values obtained from a standard normal scale (Efron and Hastie 2016). The default is 0. This option only works over a list of *p*-values and it is not applied to the maxT procedure if defined but to the other multiple testing correction methods.

-*maxT* <BOOLEAN> If 1 performs the maxT procedure. The maxT procedure can be computationally intensive depending on the number of tests and permutations. By default is 0 (maxT is not performed).

-*eqvar* <BOOLEAN> This argument has effect only if maxT is performed. If 1 performs equal variance two sample *t*-tests otherwise performs unequal variance. By default is 0 (unequal variance).

-*twotail* <BOOLEAN> This argument has effect only if maxT is performed. If 1 performs two tail tests. By default is 1.

-*onetailR* <BOOLEAN> This argument has effect only if maxT is performed and if twotail is 0. If onetailR is set to 1 it performs right tailed tests otherwise left tailed. By default is 1.

-*s0* <DOUBLE> This argument has effect only if maxT is defined. The *t* tests are computed as (ave1-ave2) / (S + s0) where S is the pooled standard deviation and s0 can be a small positive constant to ensure that variance is independent of gene expression (Tusher et al. 2001). By default s0 = 0.

-*sub* <INTEGER INTEGER> This argument has effect only if maxT is performed. It defines the range from *ini* to *end* (both included) of the subsample to be analysed. If for example *ini* = n_1 /2 and *end* = n_2, the first sample is read from n_1 /2 to n_1 instead from 1 to n_1, while the second sample is fully included. When *ini* < *end* the given subsample is written to a file. If *ini* ≥ *end* there is no subsampling and the complete samples (n_1 and n_2) are considered. By default is 0 0 (no subsampling).

-*B* <INTEGER> The number of permutations for maxT. By default is 500000.

Interactive menu options for *p*-value simulation

The simulation menu is displayed (Fig. 5) when the user chooses the options 10 or 11 from the general menu.

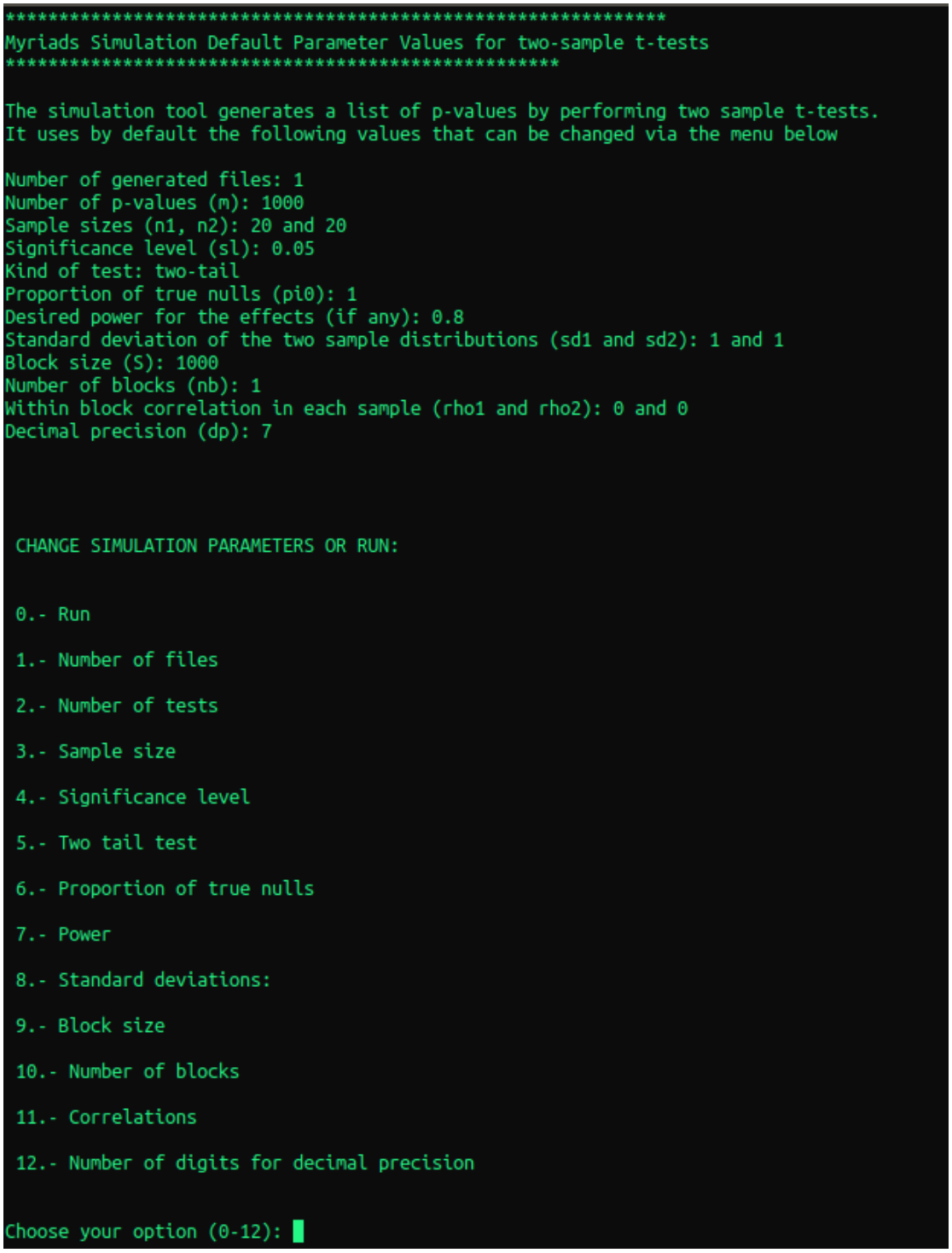

Fig. 5. Myriads simulation menu for option 10 from the general menu.

A similar menu is obtained when the option 11 is chosen from the general menu.

From the simulation menu, choosing the option 0 without any change, executes the simulation default parameter values (see below).

Command line options for *p*-value simulation

The simulation model can also be executed from the command line simply by putting 1 (*t*-tests) or 2 (Cochran-Armitage tests) in the tag -simulation when running the command line option as explained above.

t-tests

The calling

Myriads -simulation 1

is equivalent to choosing the option 0 of the menu and is also equivalent to the line with arguments:

Myriads -simulation 1 -data 0 –twotail 1 -numfiles 1 -numtests 1000 -n1 20 -n2 20 -blocksize 1000 –numblocks 1 -SL 0.05 -pi0 1 -sd1 1 -sd2 1 -rho1 0 -rho2 0 -power 0.8 -randepos 0 -dir Myriads_Sims_t -effcorr 0 -fixseed 0

The explanation of the new arguments is as follows:

-*data* <BOOLEAN> . This option is only available under the -simulation 1. When set to 1 generates gene expression like data instead of *p*-values. The data is stored under the data format already explained (Fig.2).

- *twotail* <BOOLEAN> Two tail (1) or one tail (0) *t* test (default is 1).

- *numfiles* <INTEGER> The number of *p*-value files to generate (default is 1).

- *numtests* <INTEGER> The number of *p*-values in each file (default is 1000).

- *shift* <INTEGER> Default is 0.The output file name is MyriadsPval_number.txt. The number is obtained as number=number of file + shift;

-*n1* <INTEGER> Sample size 1 for the two sample *t*-test (default is 20).

-*n2* <INTEGER> Sample size 2 for the two sample *t*-test (default is 20).

- *blocksize* <INTEGER> The size of each correlated block if any (default is 1000).

- *numblocks* <INTEGER> The number of blocks N_b. If N_b is not passed as argument, the program automatically computes N_b = (numtests / blocksize) +$\mathbf{1}_{(\text{numtests \%blocksize} \geq \text{minblock})}$, that defines the maximum possible number of blocks with size blocksize,

plus one more block, with size the residue of numtests /blocksize, in the case that the residue have the minimum acceptable block size (50 by default).

- *pi0* <DOUBLE> Proportion of true nulls (default is 1, the complete null).
- *sd1* <DOUBLE> Standard deviation from the distribution of sample 1 (default is 1).
- *sd2* <DOUBLE> Standard deviation from the distribution of sample 2 (default is 1).
- *rho1* <DOUBLE> Within-block correlation in sample 1 (default is 0).
- *rho2* <DOUBLE> Within-block correlation in sample 2 (default is 0).
- *power* <DOUBLE> Power for the two-sample *t*-test (default is 0.8).
- *randepos* <BOOLEAN> Positions with effects (in sample 2) are randomly distributed (1) or (0) they are consecutive within the block (default is 0).
- *effcorr* <BOOLEAN> The correlation in sample 2 affects only to the positions with effects (1) or (0) the correlation affects the consecutive positions within block independently of the effects (default is 0).

-*dir* <STRING> Specifies the output directory name (default is Myriads_Sims).

-*fixseed*< INTEGER > If positive it fixes the seed for the Monte Carlo simulations to that value, so that the result for different runs with equal arguments should be the same (default is 0, so that the seed is randomly generated at each program execution).

Cochran-Armitage tests

The calling

Myriads -simulation 2

is equivalent to choosing the option 0 of the simulation menu and is also equivalent to the line with arguments:

Myriads -simulation 2 –twotail 1 -numfiles 1 -numtests 1000 - -blocksize 1000 –numblocks 1 -SL 0.05 -pi0 1 -rho1 0 -power 0.8 -randepos 0 -dir Myriads_Sims_CA -fixseed 0 -maf 0.1 -prevalence 0.1 -risk2 3 -phi 0.5 -diseasemodel 0.5

The meaning of most arguments is the same as for the other simulation case. However, note that sample sizes and standard deviations are not considered when the simulation=2 mode is called. This is because sample sizes are automatically computed based on the desired power. Also, only one correlation value is necessary. Additionally,

new arguments are needed for the case-control simulation. The explanation of the new arguments is as follows:

- *maf* <DOUBLE> Minor allele frequency (default is 0.1).
- *prevalence* <DOUBLE> Prevalence of disease (default is 0.1).
- *risk2* <DOUBLE> Homozygote genotype relative risk (default is 3).
- *phi* <DOUBLE> Proportion of cases (default is 0.5, i.e. same sample size for cases and controls).
- *diseasemodel* <DOUBLE> Recessive (0), additive (0.5) or dominant (1) disease model (default is 0.5).
- *rho1* <DOUBLE> Within-block correlation in the SNPs (default is 0).
- *power* <DOUBLE> Power for the two-sample Cochran-Armitage test (default is 0.8).
- *randepos* <BOOLEAN> Positions with effects are randomly distributed (1) or (0) they are consecutive within the block (default is 0).

Run Myriads in MacOsX or Linux

From the console or terminal type

./Myriads

jointly with the desired arguments. If there are no arguments the menu will appear. If we want to skip the menu and run the default arguments we just call

./Myriads -default

which run the Myriads default options for multiple testing correction as already explained.

Run Myriads in Windows

Double click or go to the command prompt (cmd.exe) and type

Myriads

for running the program with the interactive menu. If you want to skip the menu and run the default arguments just type

Myriads -default

You can also access the Run command by pressing the Windows logo key +r

then drag and drop the .exe file from your folder and add the desired arguments, e.g. if Myriads.exe is in the folder Myriads then after drag and drop you will have

C:\ Myriads\Myriads.exe

now add the desired arguments and then hit the Intro key. For example:

C:\Myriads\Myriads.exe -inputfile pval1.txt -dependence 1

DEPENDENCE TEST

Myriads incorporate an autocorrelation test based on the generalized Durbin-Watson (D-W) statistic (Ali, 1987; Vinod, 1973). The test permits to identify strong dependencies in the *p*-values and estimate the minimum detectable block size of correlated values. The lowest the detected block size the strongest the correlation in the data.

For performing the dependence test the user just need to choose the option 5 in the general menu or add the corresponding arguments in the command line:

Myriads -dependence 1

See the Supplementary Document (Carvajal-Rodriguez, 2017) for detailed information about the algorithm.

SIMULATION

As already indicated, Myriads incorporates the option for simulating *p*-values by conducting a two sample *t*-test or a Cochran-Armitage case-control test.

For performing the simulation the user just need to choose the options 10 (*t*-test) or 11 (CA test) in the menu or add the corresponding arguments in the command line:

Myriads -simulation 1 (or Myriads -simulation 2)

See the corresponding Command Line Options section for the different arguments available. Detailed information of the simulation steps follows.

Two sample *t*-test

Myriads incorporates the option for simulating normalized gene expression data or *p*-values by conducting a two sample *t*-test. The simulation steps are as follows:

1.- The data to generate consist in normalized gene expression values. There are *m* values for each individual (library, culture or tissue) and two samples with user-defined

sizes of n_1 and n_2 individuals respectively. Therefore, we have $n_1 + n_2$ copies for the expression value of each gene. These genes can be differentially expressed or not. If a gene is differentially expressed we say it has an effect. Sample 1 is a control group without effects while sample 2 may incorporate a given proportion of genes with effects. The desired proportion of true nulls (genes without effects in sample 2) is controlled by the user via the parameter π_0. Therefore, the number of effects will be $(1- \pi_0)m$.

2.- The genes can be dependent. Suppose we desire a set of correlated values with a given mean μ and standard deviation σ. The correlation matrix is obtained based on the user-defined correlation values ρ_1 and ρ_2, for sample 1 and 2 respectively. The data structure is organized in blocks of user-defined size. These blocks can correspond to genes without and/or with effects. Each block is independent of the other. The data within each block is correlated following a multivariate normal distribution $N(\mu, \Sigma)$ where μ is the mean vector and Σ is the covariance matrix. The leading diagonal containing all 1's and the off-diagonal entries are ρ (ρ_1 or ρ_2 depending on the sample). The correlated variables are obtained as $Z_r = L \times Z$ where L is the Cholesky decomposition of the correlation matrix and Z a vector of independent standard normal variables. The desired final variables correspond to $N_r = \sigma Z_r + \mu$.
Negative correlation is allowed. However, if the block size S is higher than 2, the negative correlation $\rho = -r$ is incorporated in the covariance matrix as follows:

$S = \{s_{ij}\}$; $i < j$: if $j < S$ then $s_{ij} = r$ else ($j = S$) then $s_{ij} = -r$; $i > j$: $s_{ij} = s_{ji}$.

3.- The size of the effects is defined by introducing the preferred power for the test. Under the user desired power (default 0.8), with two sample *t*-tests and nominal α-level, the effect size is determined as follows:
Given a power of 1-B, if the distribution with no effects is $N(\mu_1, \sigma_1)$, the desired mean for the alternative distribution $N(\mu_2, \sigma_2)$ is obtained as

$$\mu_2 = \mu_1 + [Q(1- \alpha/2) - Q(B)] \times sqrt\{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2\}$$

where B is the type 2 error, and Q is the standard normal quantile function.
Note that we use the normal quantile instead of the inverse *t* distribution function. In consequence, the obtained mean-effect is slightly down-biased when the sample size

is small although the approximation is good under medium-high sample sizes. This lower-than-expected mean-effect would produce slightly lower power. For example, 0.1634 instead 0.1788 (1.5% less) under n_1= n_2=5 due to an effect size of 0.458 instead of 0.5 caused by the down-biased mean.

4.- Once the necessary parameters has been introduced, the program performs a number *m* of independent two-sample two-tail *t* tests for obtaining the corresponding *p*-values. We distinguish two cases depending on if the variances for the null and effects distributions are the same (default option) or not.

Equal variances

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

S_1^2 and S_2^2 are the unbiased estimators of the variances of the two samples and n_1+ n_2 - 2 are the degrees of freedom.

Unequal variances (Welch's t-test)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_W}$$

$$S_W = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

S_1^2 and S_2^2 are the unbiased estimators of the variances of the two samples. The degrees of freedom are calculated as

$$df = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2/(n_1 - 1) + \left(S_2^2/n_2\right)^2/(n_2 - 1)}$$

5.- Finally, depending on the command line tag -*data*, by default the obtained *p*-values are written to a file following the Myriads (SGoF-like) format. The simulator allows obtaining several files in the same run. In this case, the position of the effects is the

same in every file (but the effects are different). If distinct positions are desired then it is necessary to generate just one file per run and perform as many runs as needed.
If the tag *data* is set to 1 then no *p*-values are computed but the generated data is written in the gene expression format (see Fig. 2).

Cochran-Armitage trend test

Myriads incorporates the option for simulating *p*-values from case-control GWAS by conducting a Cochran-Armitage trend test (Armitage, 1955; Zheng and Gastwirth, 2006). The test is performed separately for each SNP. The data for each SNP with alleles 0 and 1 can be represented as a contingency table of genotype counts (e.g., 0/0, 0/1 and 1/1).

Table 1. Genotype counts and disease model for the case-control studies

Genotype	00	01	11	Sample Size
Disease model	0	*x*	1	
Case	n_{10}	n_{11}	n_{12}	n_1
Controls	n_{20}	n_{21}	n_{22}	n_2
Total	N_0	N_1	N_2	N

Let a case-control experiment with sample sizes n_1 and n_2 for cases and controls, respectively. The genotypes 00, 01 and 11 are noted with indexes *g*:{0,1 2}, respectively. So n_{10} is the count of genotype 00 in cases and n_{20} is the count of genotype 00 in controls and in general we have n_{1g} counts for genotype *g* in cases and n_{2g} in controls. The total count by genotypes in the pooled sample is N_g and the absolute total $N = n_1 + n_2 = N_0 + N_1 + N_2$ (Table 1).
The association trend test statistic can be computed as (Zheng and Gastwirth, 2006)

$$Z^2 = \frac{\left\{\sum_{i=0}^{2} w_i[(1-\varphi)n_{1i} - \varphi n_{2i}]\right\}^2}{N V_1(\varphi)}$$

where w_i represents the disease model {0, *x*, 1}, $\varphi = n_1/N$ is the proportion of cases, and NV_1 (φ) is the variance estimation where

$V_1(\varphi) = \varphi(1-\varphi)V_{C1}$ with

$$V_{C1}=\sum_{i=0}^{2} w_i^2\left(\frac{N_i}{N}\right)-\left[\sum_{i=0}^{2} w_i\left(\frac{N_i}{N}\right)\right]^2$$

The test follows a chi-square distribution with one degree of freedom under the null hypothesis of no association (Zheng and Gastwirth, 2006). The alternative hypothesis considers that a statistical association exists between the disease trait and the genetic marker (SNP). The risk of disease increases with the number of '1' alleles.
The flow of the simulation is as follows:

1.- Input

The power (1 - *B*): default (0.8) or user-defined.

The disease prevalence (*K*): default (0.1) or user-defined.

The relative risk γ_2: default (3) or user-defined.

The disease model (*x*): default additive (0.5) or user-defined (*x* = 0 for recessive, and *x* = 1 for dominant).

The proportion of cases (φ): default (0.5) or user-defined.

The proportion of true nulls (π_0): for some SNPs, the genotypes can have different frequencies in cases versus controls (SNP-disease association). The desired proportion of true nulls (proportion of SNPs with no association) is by default 1 or alternatively, controlled by the user via the parameter π_0. Therefore, the number of SNPs with association will be $(1-\pi_0)m$.

The correlation (ρ): the frequency of the SNPs can be correlated. The data structure is organized in blocks of correlated values of user-defined size (default is one block of size *m* and no correlation). The algorithm for generating correlated blocks is similar to the two sample *t*-tests already explained. Thus, after generating independent numbers from the standard normal distribution, we end with one or more correlated blocks where each variable Z_r belongs to the multivariate standard normal with correlation ρ.

2.- *Disease model*

The additive disease model (0, *x* = 0.5, 1) with penetrance $(f_0, f_1, f_2) = (f_0, 2f_0, 3f_0)$ is assumed by default. In general, the penetrance is $(f_0, f_1, f_2) = (f_0, \gamma_1 f_0, \gamma_2 f_0)$.

So, define $f_1 = \gamma_1 f_0$, $f_2 = \gamma_2 f_0$. The relative risk γ_1 depends on the disease model as follows:

Recessive (x = 0): $f_1 = f_0$ so $\gamma_1 f_0 = f_0 \Rightarrow \gamma_1 = 1$.

Dominant (x = 1): $f_1 = f_2$ so $\gamma_1 f_0 = \gamma_2 f_0 \Rightarrow \gamma_1 = \gamma_2$.

Additive (x = 0.5): $f_1 = (f_0 + f_2) / 2$ and $f_1 = \gamma_1 f_0$ so $\gamma_1 f_0 = f_0(1+ \gamma_2)/2 \Rightarrow \gamma_1 = (1+ \gamma_2)/2$

Thus, given a disease model x, prevalence K, and relative risk γ_2, after the population genotypes g are computed (see below) we obtain $f_0 = K / (g_0 + \gamma_1 g_1 + \gamma_2 g_2)$.

3.- *Sample size*

For a given power, the sample size depends on the genotype frequencies in case and controls which would be different in different SNPs. Thus, we should estimate the sample size that would be enough for achieving the desired power under a range of SNP frequencies.

In the case of recessive and additive disease models, the minor allele frequency (MAF) requires the highest sample size (Zheng and Gastwirth 2006) and so, under these models we use p = MAF as disease allele frequency, for computing the sample size. Under the dominant model, p = 0.5 requires higher sample size and we use this frequency if the dominant model is defined by the user.

With disease allele frequency p and assuming HWE, the population genotype frequencies are $(g_0, g_1, g_2) = \{(1-p)^2, 2(1-p)p, p^2\}$. Now, as indicated above $f_0 = K / (g_0 + \gamma_1 g_1 + \gamma_2 g_2)$ and $f_1 = \gamma_1 f_0$, $f_2 = \gamma_2 f_0$.

The genotype frequencies in cases are (p_0, p_1, p_2) with $p_i = f_i g_i / K$ and in controls (q_0, q_1, q_2) with $q_i = (1-f_i)gi / (1-K)$.

Therefore, given the power 1-B we obtain the sample size $N = n_1 + n_2$ as

$$N = \{Q(1- \alpha/2)[\sigma_1^2(1-\varphi) + \mu_1^2]^{1/2} + Q(1-B)\sigma_1(\varphi)\}^2 \ / \ \mu_1^2$$

where Q is the standard normal quantile function and

$$\mu_1 = \varphi(1-\varphi)\sum_{i=0}^{2} w_i(p_i - q_i)$$

$$\sigma_1^2(\varphi) = \varphi(1-\varphi)^2 V_1 + \varphi^2(1-\varphi)V_2$$

with

$$V_1 = \sum_{i=0}^{2} w_i^2 p_i - \left[\sum_{i=0}^{2} w_i p_i\right]^2$$

$$V_2 = \sum_{i=0}^{2} w_i^2 q_i - \left[\sum_{i=0}^{2} w_i q_i \right]^2$$

The sample size in cases will be $n_1 = \varphi N$ and in controls $n_2 = N - n_1$.

4.- *Disease allele frequency and genotype frequencies*

From herein, the procedure is performed for each of the *m* SNPs. We obtain *m* Z_r normal deviates following the same steps implemented for the simulation of *t*-tests. Then we apply the univariate normal CDF to each Z_r for deriving the corresponding probability p_r. However, since we are interested only in minor SNPs having at least a minor allele frequency MAF, we take

$p = max(1-p_r, \text{MAF})$ if $p_r \geq 0.5$ or

$p = max(p_r, \text{MAF})$ otherwise,

as the risk allele frequency in the population.

Assuming HWE, the population genotype frequencies are $(g_0, g_1, g_2) = \{(1-p)^2, 2(1-p)p, p^2\}$.

Now, if the SNP has no effect i.e. it belongs to the π_0 proportion, then we assume that the expected genotype frequencies in cases and controls are the population ones, so $p_i = q_i = g_i$.

Alternatively, if the SNP is in the 1- π_0 proportion, then the genotype frequencies in cases are $p_i = f_i g_i / K$ and in controls $q_i = (1-f_i) g_i / (1-K)$.

5.- *Case and control counts*

We get the count for cases $(n_{10}, n_{11}, n_{12}) = multinomial(n_1, p_0, p_1, p_2)$ and for controls $(n_{20}, n_{21}, n_{22}) = multinomial(n_2, q_0, q_1, q_2)$.

6.- *p-values*

For each SNP we perform the test Z^2 (Sasieni 1997; Slager and Schaid 2001) using the multinomial counts. The list of *p*-values is written to a file following the Myriads (SGoF-like) format.

As before, the simulator allows obtaining several files in the same run. In this case, the SNPs with association are the same in every file (but the frequencies would be

different unless we fix the random seed). If we desire distinct SNP positions among files then it is necessary to generate just one file per run and perform as many runs as required.

OUTPUT

The program provides with various output files in the folder 'Myriads_output' or any other folder defined by the user (via console argument). Myriads produces an html file and a text file with extension ods (OpenDocument Spreadsheet) that provide the full list of original significant *p*-values, the adjusted *p*-value after each method (FWER and FDR ones) and the estimated *q*-values.
In addition (from version 1.2), Myriads provides a separate summary for the FWER and FDR-based correction methods.

FWER multiple testing

The summary file for the FWER-based methods is called by default FWER_Summ_Myriads.txt or FWER_Summ_Glb.txt if the input file was Gbl.txt. This summary includes the tests remaining significant after the different FWER-based adjustments.
Adjusted Bonferroni and sequential Bonferroni (SB) *p*-values are computed following (Yekutieli and Benjamini 1999). Adjusted Hommel *p*-values are computed following (Meijer et al. 2019). When the number of tests is below 10,000 (a myriad), the adjusted *p*-values for SGoF are computed following (Castro-Conde and de Uña-Álvarez 2015). For higher number of tests, the metatest *p*-values are given by default, and the adjusted *p*-values, when computed, follow the approximation developed in (Carvajal-Rodríguez 2018).

FDR multiple testing

Adjusted BH, BKY, BonSEV and BY *p*-values are computed following (Yekutieli and Benjamini 1999). Values under the SLIM column correspond to the *q*-value obtained using the SLIM estimate for the proportion of true nulls.

MaxT

When maxT is run Myriads produces 3 output files. First, is a summary with the results, for example the file MyriadsMaxT_Summ_Golub_W.txt gives the summary for the Golub data (Golub et al. 1999; Dudoit et al. 2002) after performing the unequal variance (Welch) tests. The output indicates the number of permutations performed and includes columns for the row id, test, *p*-value and adjusted *p*-value after maxT. The number of rows correspond to the number of significant raw *p*-values (those ≤ α) .
A second file including all test and raw *p*-values, e.g. 3051 for the Golub data, MyriadsTestandpval_Golub_W.txt and finally a third file, MyriadsTPval_Golub_W.txt, including just the list of raw *p*-values in the Myriads input format. If the test is the equal variance one, the postfix _W is not included in the file name.

Dependence test

If the dependence test is performed, another file called MyriadsDepRes.txt is added to the previous ones. If dependence was detected, the content of this file includes the rejection level, the minimum block size for which the test was significant and the number of significant blocks at that size. An example of the output is as follows:

```
Input file: PvalMyriads.dat
Myriads dependence analysis output:
----------

Minimum allowed block size = 20
REJECTION LEVEL = 0.00066
Zcrit = 3.40655
Number of p-values = 3170
Block size = 160
Number of blocks = 19
Number of significant blocks = 1
Lag = 1
Test = 3.58074

Sat Aug 12 16:36:43 2017
```

Light output

The light mode has not been updated with the new methods and offers the same output as for the previous version 1.1. This is the output obtained when the –light argument is set to 1. In this case the only file obtained is by default called MyriadsFast.txt, although, both the folder and the file name, can be changed by the

corresponding arguments as we have seen in the command line options section. The light mode is only available from the command line and only under the non-simulation mode. It is useful when we desire to analyze many files so that the output of each is accumulated in a line as follows:

FILE	#m	m0	#SIGT	PI0	BH	BSEV	SB	SGOF	SLIM	QVAL	Qmin	DEP	BlockSize
file.txt	3170	3170	606	0.69	94	102	2	423	124	158	0	0	0
file2.txt	3170	3170	306	0.99	0	0	0	0	0	0	0.61	1	160

….

Most of the headers are self-explanatory. The number under m_0 is a guess for the number of true nulls computed as *m truepi0* where *truepi0* is 1 by default but can be user-defined by means of the argument -truePI0. The later can be useful is we are analyzing simulated files under known conditions, for benchmarking purposes.
The number below the different methods refers to the number of *p*-values detected as significant after the adjustment, or in the SLIM and QVAL cases the number of *q*-values that are below the pre-defined α-level. Qmin is the lowest *q*-value divide by the PI0 estimate. If the dependence test is not performed or no dependence was detected (DEP 0 in any case) the block size is 0, on the contrary, if dependence is detected (DEP 1), the minimum significant block size is given.

Simulation output

If the simulation mode is executed Myriads produce as many files as indicated by the tag -*numfiles*. Each file has as many *p*-values as indicated by the tag -*numtests*. The default output folder name is Myriads_Sims_ this name can be changed by the argument -*dir*. The format of the output files coincide with the input file format for the analysis of *p*-values (Fig. 1). The name of each file is MyriadsPval_1.txt, MyriadsPval_2.txt etc until the number of files indicated in -*numfiles* is reached. However, for the simulation mode 1 (normalized gene expression data and *t*-tests) there is the possibility of obtaining the normalized expression data instead of the *p*-values. This is attained via the command line tag -*data* set to one (non default option), then no *p*-values are computed but the generated data is written in the gene expression format (see Fig. 2). The obtained data is saved in a Myrids_Data folder with,

for example, the file name DataMatrix_n1_27_n2_38_1.txt for the case when the sample sizes are n_1 = 27 and n_2 = 38 and the simulated file is the first or the unique.

DEFINITIONS

Test of hypotheses

A rule for deciding whether to accept or reject a hypothesis. The hypothesis under test is called the main or null hypothesis. The boundary for deciding between hypotheses is called the critical value of the test.

Type I error

The rejection of a true null hypothesis.

Nominal Type I error rate α

The user-supplied upper-bound for the type I error rate (Greenland 2019). It is the type I error expected using the (a priori defined) critical value of the test.

Type II error

The acceptance of a false null hypothesis.

Power of a test

Is the probability that the test reject the null hypothesis when the null is false. In the context of binary classification, e.g. medical testing, it is also called sensitivity or true positive rate. The power is one minus the probability of a type II error.

Observed *p*-value

Is the probability, under the null hypothesis H_0, that a test statistic would be equal to or more extreme than the observed value. More formally, given a statistical model A and a tested hypothesis H_0, the observed *p-value* is the probability that the test statistic be equal or larger than its observed value in the current sample realization if every model assumption were correct, including H_0 (Greenland et al. 2016).

Complete null hypothesis

When all null hypotheses are true.

Control in the weak sense

Control in the weak sense occurs when the type I error rate is controlled at the specified level only under the complete null hypothesis.

Control in the strong sense

Control in the strong sense occurs when the type I error rate is controlled at the specified level for any combination of true and false null hypotheses.

Family-Wise Error Rate (FWER)

Is the probability of at least one type I error in the set of tests i.e. FWER = Pr($V \geq 1$) where V is the number of false rejections. Controlling the FWER is a conservative strategy which means that in general the probability of rejecting the null hypotheses is below the nominal α level.

False Discovery Proportion (FDP)

FDP is the (unobserved) proportion of false rejections V among total rejections R, FDP = V / R, with the convention of FDP = 0 when $R = 0$.

False Discovery Rate (FDR)

FDR is the FDP averaged over all possible experimental replicates. More technically, is the expected FDP unconditional on the occurrence of rejections.

Positive False Discovery Rate (pFDR)

The expected proportion of false rejections V among the rejections R conditioned on and least one rejection.

REFERENCES

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57:289–300.

Benjamini Y, Krieger A, Yekutieli D. 2006. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93:491–507.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29:1165–1188.

Carvajal-Rodríguez A. 2018. Myriads: P-value-based multiple testing correction. *Bioinformatics* 34:1043–1045.

Castro-Conde I, de Uña-Álvarez J. 2015. Adjusted p-values for SGoF multiple test procedure. *Biom. J.* 57:108–122.

Dudoit S, Fridlyand J, Speed TP. 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *J. Am. Stat. Assoc.* 97:77–87.

Efron B, Hastie T. 2016. Computer age statistical inference. Cambridge University Press

Ge Y, Dudoit S, Speed TP. 2003. Resampling-based multiple testing for microarray data hypothesis. *Test* 12:1–44.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.

Greenland S. 2019. Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values. *Am. Stat.* 73:106–114.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31:337–350.

Li D, Xie Z, Zand M, Fogg T, Dye T. 2017. Bon-EV: an improved multiple testing procedure for controlling false discovery rates. *BMC Bioinformatics* 18:1.

Meijer RJ, Krebs TJP, Goeman JJ. 2019. Hommel's procedure in linear time. *Biom. J.* 61:73–82.

Sasieni PD. 1997. From genotypes to genes: doubling the sample size. *Biometrics*:1253–1261.

Slager SL, Schaid DJ. 2001. Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Hum. Hered.* 52:149–153.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* 98:5116–5121.

Wang H-Q, Tuominen LK, Tsai C-J. 2011. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics* 27:225–231.

Westfall PH, Young SS. 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. New York: Wiley

Yekutieli D, Benjamini Y. 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Inference* 82:171–196.

Zheng G, Gastwirth JL. 2006. On estimation of the variance in Cochran–Armitage trend tests for genetic association using case–control studies. *Stat. Med.* 25:3150–3159.