

Appendix

Data Simulation

Simulating a dataset involves (i) generating genotypes for each of n individuals; (ii) calculating the risk value for each individual; and (iii) generating event status for each individual. We use the following notation

- i index for individual ; n individuals
- g index for gene ; G genes
- ρ overall proportion of events in the population
- f_g frequency of minor allele variant for gene g
- K_{gi} number of minor alleles for gene g that subject i has, $K_{ij}=0,1$ or 2
- α_g log odds ratio for gene g

(i) Genotype

For each gene, g , the two alleles are generated independently as binary random variables with probability f_g of being the minor variant E and probability $1-f_g$ of being the major variant e .

(ii) Risk

For each subject his risk is calculated as $risk_i = \exp(S_i / (1 + \exp S_i))$ where

$$S_i = \alpha_0 + \sum_g^G \alpha_g K_{gi}$$

Here,

$$\alpha_0 = \log \rho / (1 - \rho) + \sum_g^G \log DLR_g^-$$

where DLR_g^- is the negative diagnostic likelihood ratio for gene g , defined as h_g/l_g in the following table. Using the marginal frequencies in the table and the constraints that $\log(c_g l_g / h_g d_g) = \alpha_g$ and $\log(a_g l_g / b_g h_g) = 2\alpha_g$, one calculates the value of DLR_g^- for each gene.

Genotype	Case Count	Control Count	Total
EE	a_g	b_g	nf_g^2
Ee	c_g	d_g	$2nf_g(1-f_g)$
ee	h_g	l_g	$n(1-f_g)^2$
Total	$n\rho$	$n(1-\rho)$	n

Program

Programs to simulate datasets and calculate summary statistics shown in Tables 2 and 3 use the freely available R statistical package. Within R, install the package `mgrp` available from the Diagnostics and Biomarkers Statistical Center software website (<http://labs.fhcr.org/pepe/dabs/software.html>). The key command is `simulateRiskPerf`.

To generate a single row of Table 2 with 50 genes (`nog=50`), common odds ratio=1.05 (`or=1.05`), risk allele frequency of 5% (`f=.05`), setting the population event rate ρ to 10% (`p=0.1`), the number of subjects to 100,000 (`n=1e+05`), and using a high risk threshold of 20% (`highRisk=.2`), the following command was run to yield the corresponding row of Table 2:

```
simulateRiskPerf(nog=50,or=1.05,f=.05,p=0.1,n=1e+05,highRisk=0.2)
```

Options to vary the output and to set the random number seed are also available. Full documentation is provided with the software.