BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 2 2009, pages 272–273
doi:10.1093/bioinformatics/btn616

Genome analysis

Programs for calculating the statistical powers of detecting susceptibility genes in case–control studies based on multistage designs

Nobutaka Kitamura[1,*], Kouhei Akazawa[1], Akinori Miyashita[2], Ryozo Kuwano[2], Shin-ichi Toyabe[1], Junichiro Nakamura[1], Norihito Nakamura[1], Tatsuhiko Sato[1] and M. Aminul Hoque[3]

[1]Department of Medical Informatics, Niigata University Medical and Dental Hospital, Niigata 951-8520, [2]Genome Science Branch, Center of Bioresource-Based Researches, Brain Research Institute, Niigata University, Niigata 951-8585, Japan and [3]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

Received on June 5, 2008; revised on November 14, 2008; accepted on November 24, 2008

Advance Access publication November 28, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: A two-stage association study is the most commonly used method among multistage designs to efficiently identify disease susceptibility genes. Recently, some SNP studies have utilized more than two stages to detect disease genes. However, there are few available programs for calculating statistical powers and positive predictive values (PPVs) of arbitrary n-stage designs.

Results: We developed programs for a multistage case–control association study using R language. In our programs, input parameters include numbers of samples and candidate loci, genome-wide false positive rate and proportions of samples and loci to be selected at the k-th stage ($k = 1, \ldots, n$). The programs output statistical powers, PPVs and numbers of typings in arbitrary n-stage designs. The programs can contribute to prior simulations under various conditions in planning a genome-wide association study.

Availability: The R programs are freely available for academic users and can be downloaded from
http://www.med.niigata-u.ac.jp/eng/resources/informatics/gwa.html

Contact: nktmr@m12.alpha-net.ne.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genome-wide association (GWA) case–control studies using genetic polymorphism data, including data on single nucleotide polymorphisms (SNPs) and microsatellite markers, aim to identify genes involved in common human diseases. Some multistage GWA case–control studies have been proposed as powerful and cost-effective study designs. The most commonly used method among the multistage designs is a two-stage design. In the designs, all markers in the first stage are genotyped and statistical tests are conducted for each marker. Only significant markers selected in the first stage remain for the following stages. At the end of the last stage, the overall evidence for association is evaluated by replication-based analysis (RBA) (Satagopan *et al*., 2002) or joint analysis (JA) (Skol *et al*., 2006).

Recently, some SNP-based studies have utilized more than two stages of multistage designs (Kathiresan *et al*., 2007; Prentice *et al*., 2006). However, there are few available programs for multistage designs. Skol *et al*. (2006) proposed JA by new statistics of a linear combination weighted by the square root of the proportion of subjects and presented their power calculation for one- and two-stage designs on their website, but they have not reported about more than two stages. In previous studies, statistical power was used to evaluate the performance of different designs. On the other hand, positive predictive values (PPVs) are more practical for identifying disease susceptibility genes implicated in complex traits. However, there is no program for statistical powers, PPVs and numbers of typings of multistage designs for arbitrary $n > 2$.

In this article, we present programs for calculating the indicators of arbitrary n-stage designs using disease model parameters (prevalence, disease-associated allele frequency and genetic relative risk) or 2×3 contingency table data as input parameters by R language (Crawley, 2005). The programs can output the indicators for allelic models, additive models, dominant models and recessive models under various conditions.

2 METHODS

2.1 Statistical background for detecting risk SNPs

We extended Skol's power-evaluation methods of two-stage designs (Skol *et al*., 2006) to generalized n-multistage designs (n being an arbitrary natural number). Let $\pi_{s,k}$ be a proportion of the sample size at the k-th stage and let $\pi_{m,k}$ be a proportion of loci to be selected at the k-th stage ($k = 1, \ldots, n$). In RBA, statistical tests are conducted by statistics for difference in ratio of an expected disease-associated allele rate in cases (p') and that in controls (p). Critical values for the k-th stage of RBA ($C_{r,k}$) can be obtained by calculating $\Phi^{-1}\left[1-\pi_{m,k}/2\right]$ under the null hypothesis, and powers of k-stage designs of RBA are calculated by multiplying cumulative distribution functions of a normal distribution of each stage under the alternative hypothesis.

*To whom correspondence should be addressed.

Test statistics of JA are represented as the $\sqrt{\pi_{s,k}}$-weighted sum of statistics of each stage. Since the statistics of JA and those of previous stages are not independent from those definitions, critical values of JA ($C_{j,k}$) can be calculated iteratively by solving the equation of n-multiple integrals under the null hypothesis using the Newton–Raphson method. Powers of JA can be calculated with $C_{j,k}$ under the alternative hypothesis.

PPVs can be calculated by the number of true positive loci (tp_k) and false positive loci (fp_k). Let N be the number of samples in cases and controls, M be the number of candidate loci and m be the number of true loci. tp_k is given as $m \times$ power of RBA or JA, and fp_k is given as $\left(M-m\right)\times\prod_{k=1}^{n}\pi_{m,k}$.

Therefore, PPV at the k-th stage is calculated as $tp_k \big/ \left(tp_k + fp_k\right)$.

The number of typings of one-stage design is given as $2MN$.

The number of typings of n-stage design is

$$2MN\pi_{s,1}+\sum_{k=2}^{n}\left(2MN\pi_{s,k}\prod_{l=1}^{k-1}\pi_{m,l}\right) \quad (n\geqslant 2).$$

2.2 Algorithm and programs for multistage designs

The algorithm for the powers and PPVs and numbers of typings of multistage designs consists of the following three modules: a module for internal parameters, a module for RBA and a module for JA.

Let a candidate locus be a base pair of allele A and allele a and let a penetrance of aa be f_0. The relative risk to f_0 in a human population of Aa and that of AA will be f_1/f_0 and f_2/f_0, respectively. By using disease model parameters, such as prevalence (Prev) and disease-associated allele rate (p_{po}) as input parameters, f_0, f_1 and f_2 in the allelic model, in which the numbers of each allele per cases and controls are counted, is calculated as follows:

$$f_0=\frac{\mathrm{Prev}}{p_{\mathrm{po}}^2\times\left(f_2/f_0\right)+2\times p_{\mathrm{po}}\times\left(1-p_{\mathrm{po}}\right)\times\left(f_1/f_0\right)+\left(1-p_{\mathrm{po}}\right)^2},$$

$$\frac{f_1}{f_0}=\frac{f_2}{f_1}=\mathrm{GRR}\ (\text{genotype relative risk}),$$

$$\frac{f_2}{f_0}=\left(\frac{f_1}{f_0}\right)^2=\mathrm{GRR}^2.$$

Therefore, p' and p are calculated as follows:

$$p'=\frac{p_{\mathrm{po}}^2\times f_2+p_{\mathrm{po}}\times\left(1-p_{\mathrm{po}}\right)\times f_1}{\mathrm{Prev}},$$

$$p=\frac{p_{\mathrm{po}}^2\times\left(1-f_2\right)+p_{\mathrm{po}}\times\left(1-p_{\mathrm{po}}\right)\times\left(1-f_1\right)}{\left(1-\mathrm{Prev}\right)}.$$

In additive models, dominant models and recessive models, those genotype rates in cases and controls are calculated according to each definition (refer to Supplementary Material).

The programs for calculating the indicators were made by using R language (version 2.7.1). In the program, calculation of the equation of n-multiple integrals uses the 'pmvnorm' function. This function computes the distribution function of the multivariate normal distribution for arbitrary limits and correlation matrices based on algorithms by Genz (1992). We prepared the R package and the web user interface (WUI) for calculating them in our website (see Availability).

2.3 Simulation algorithm

Assume the following conditions for calculating the powers and PPVs and the numbers of typings in one- to five-stage designs: $N=1000$, $M=500\,000$, $m=5$, genome-wide false positive rate (α_{genome}) $=0.05$ and odds ratio (OR) $=1.5$, which is calculated by Prev $=0.1$, disease-associated allele frequency $=0.4$ and GRR $=1.44$, and the genetic model is set to be an allelic model. The samples were equally divided into each stage. In addition, $\pi_{m,k}$ are set to equal proportions to adjust the product of them to 0.0001 so that the number of remaining loci in the final stage by each stage design is equal.

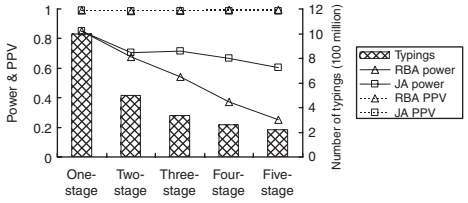

Fig. 1. Powers and PPVs and numbers of typings by RBA and JA in one- to five-stage designs.

3 RESULTS

Figure 1 shows the powers and PPVs and the numbers of typings of RBA and JA in one- to five-stage designs. The powers of JA were always higher than those of RBA. In RBA, with an increase in the number of stages, the numbers of typings and the powers monotonically decreased. However, in JA, the powers decreased more slowly than in RBA. The powers of the three-stage design were higher than the powers of two-, four- and five-stage designs. On the other hand, PPVs of RBA and those of JA both exhibited plateau behaviors in a high range of PPVs.

4 DISCUSSION

Multistage case–control association designs have the distinct advantage of considerable reduction in the number of typings, while maintaining high powers for many practical projects to detect disease-related genes.

We made programs for multistage designs by an arbitrary number of stages by using package R language. In this study, the properties of multistage designs of RBA and JA were investigated by comparisons of one- to five-stage designs, and it was shown that the powers by JA are larger than those by RBA in any number of stages and that three-stage designs are superior to two-, four- and five-stage designs in powers under the condition that the samples are equally divided into each stage (refer to Supplementary Material).

Factors affecting powers include N, M, $\pi_{s,k}$, $\pi_{m,k}$ and study designs. However, there is no simple way to predict the powers by various study designs. Therefore, our programs would be beneficial in study communities at the planning stage of GWA studies.

Funding: This research was supported by grants-in-aid for scientific research [Based research (B)] by the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Conflict of Interest: none declared.

REFERENCES

Crawley,M.J. (2005) *Statistics: An Introduction Using R*. John Wiley & Sons, England.

Genz,A. (1992) Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Stat.*, **1**, 141–150.

Kathiresan,S. *et al*. (2007) A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.*, **8**, http://www.biomedcentral.com/1471-2350/8/S1/S17 (last accessed on December 7, 2008).

Prentice,R.L. and Lihong,Qi. (2006) Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation. *Biostatistics*, **7**, 339–654.

Satagopan,J.M. *et al*. (2002) Two-stage designs for gene-disease association studies. *Biometrics*, **58**, 163–170.

Skol,A.D. *et al*. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.*, **38**, 209–213.