BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 16 2014, pages 2377–2378
doi:10.1093/bioinformatics/btu296

Genetics and population analysis

Advance Access publication April 28, 2014

Power analysis and sample size estimation for sequence-based association studies

Gao T. Wang[1], Biao Li[1], Regie P. Lyn Santos-Cortez[1], Bo Peng[2] and Suzanne M. Leal[1,*]

[1]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine and [2]Department of Bioinformatics and Computational Biology, The University of Texas, M D Anderson Cancer Center, Houston, TX 77030, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Statistical methods have been developed to test for complex trait rare variant (RV) associations, in which variants are aggregated across a region, which is typically a gene. Power analysis and sample size estimation for sequence-based RV association studies are challenging because of the necessity to realistically model the underlying allelic architecture of complex diseases within a suitable analytical framework to assess the performance of a variety of RV association methods in an unbiased manner.

Summary: We developed SEQPower, a software package to perform statistical power analysis for sequence-based association data under a variety of genetic variant and disease phenotype models. It aids epidemiologists in determining the best study design, sample size and statistical tests for sequence-based association studies. It also provides biostatisticians with a platform to fairly compare RV association methods and to validate and assess novel association tests.

Availability and implementation: The SEQPower program, source code, multi-platform executables, documentation, list of association tests, examples and tutorials are available at http://bioinformatics.org/spower.

Contact: sleal@bcm.edu

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

Received on October 8, 2013; revised on March 31, 2014; accepted on April 23, 2014

1 INTRODUCTION

Power analysis is one of the most crucial steps in designing complex trait genetic association studies. The aim is to determine the power to detect an association for a specified sample size or to estimate the sample size for a given power for a variety of genetic models and statistical methods. Over the past decade, genome-wide association studies using single nucleotide polymorphism markers have been highly successful in the study of complex diseases, with power analysis aided by software packages such as Genetic Power Calculator (Purcell *et al.*, 2003) and CaTS (Skol *et al.*, 2006). Currently, there is growing interest in detecting complex trait rare variant (RV) associations using next-generation sequencing (NGS) data. Many sequence-based RV association methods have been developed to jointly analyze multiple SNVs by aggregating them across a region, e.g. a gene. Complexities arise, making power and sample size estimations for these RV association methods challenging. The allelic architecture of RVs is dependent on population genetic parameters, which are difficult to realistically model. Additionally, within a region, the genetic effect sizes of variants are not uniform and non-causal variants as well as variants with bidirectional effects, e.g. protective and detrimental, can be present. Also there can be genotyping error and missing data. This genetic complexity makes estimation of sample size and power for RV association studies challenging. Owing to mathematical intractability, application of theoretical power analysis is limited. Empirical power studies are also hindered by a lack of consistent implementation of many existing RV association methods.

To perform power analysis, SEQPower uses sequence data generated via advanced genetic simulation technique as well as data derived from real-world NGS studies. It generates qualitative and quantitative trait (QT) data based on observed variants in the simulated genetic region. Analytic and empirical power analysis can be performed using a large variety of RV association methods.

2 METHODS

2.1 Simulation of DNA sequence and disease phenotype data

Owing to the uncertainty of RV architecture, many simulation methods have been proposed to generate DNA sequence data using Wright's formula, empirical Bayesian estimates, coalescent, forward-time and resampling from real-world data. SEQPower generates sequencing data using forward-time simulation while incorporating demographic and natural selection parameters (Peng and Liu, 2010) or extrapolated minor allele frequency (MAF) spectra based on data from the The National Heart, Lung, and Blood Institute Exome Sequencing Project (ESP) (Tennessen *et al.*, 2012). It is also possible to provide SEQPower with other simulated or real-world NGS datasets. Variants with a potential contribution to disease traits are annotated based on (i) frequencies, (ii) selection coefficients or (iii) functional annotations, such as scores from variant effect prediction algorithms or measures of nucleotide conservation. The joint effects of a genomic region on phenotypes are reflected as either disease status or QT value. The impact of each variant can be modeled as the logit of odds, population-attributable risk or the difference in mean QT. Three study designs, case–control, extreme QTs and randomly ascertained quantitative phenotypes, are available. Additionally, variant data can be modified to mimic NGS platforms as well as exome genotyping array data (e.g. 'exome chip'). More details on

*To whom correspondence should be addressed.

data simulation methods can be found in Supplementary Table S2 and in the online documentation.

2.2 Analytic power and sample size calculations

Analytic power analysis can be performed for several basic models and statistical methods. The fundamental idea behind RV association methods is to compare the difference in cumulative MAF between cases and controls, or the difference in mean QT values between wild-type individuals and those with alternative alleles. For the case–control design, we calculate case and control MAF under the Bayes' law. Given the simulated MAF spectrum and effect size at each variant site *i*, the case–control status-specific genotype frequency is calculated as $p(g_i|status) = \frac{p(g_i)f_i}{p(status)}$ where $p(g_i)$ is the population genotype frequency, f_i is penetrance, $p(status)$ is disease prevalence (K) in cases and $1\text{-}K$ in controls. To perform analytic power analysis for the Combined Multivariate and Collapsing (CMC) method (Li and Leal, 2008), for a genetic region with M variants, cumulative MAF for cases or controls can be calculated as $p = 1 - \prod_i^M (1 - p_i)$ and power for detecting the difference between p_{case} and $p_{control}$ can be calculated (Fleiss *et al.*, 1980). To perform analytic power analysis for the Burden of Rare Variants (BRV) (Auer *et al.*, 2013) method, a 2×2 contingency table of expected counts for minor (N_1) and major (N_2) alleles in cases as well as in controls (N'_1 and N'_2) is constructed, and a χ^2 test is applied. For QTs, the expected mean shift is the joint effect of variants across the region. Denoting each set of causal variants as V, and the corresponding V^C as the set variant sites that are homozygous for the wild-type allele, the probability to observe such set of variants in the samples is $\prod_{i \in V} p_i \prod_{j \in V^C} (1 - p_j)$ with effect size $\sum_{i \in V} \lambda_i$ where λ_i is the effect size of variant *i*. Then a linear regression-based goodness-of-fit test can be constructed to perform power and sample size estimates. Compared with empirical analysis, the SEQPower analytic framework provides efficient sample size estimates. The analytic framework also allows for modeling using simulated data; however, multiple replicates are necessary to compute the average sample size or power to adjust for the randomness in generating variant sites and their effect sizes.

2.3 Empirical power comparisons

SEQPower can also perform empirical power analysis, which is more flexible than analytic power analysis and sample size estimation. Empirical power analysis is available for a large variety of study designs, disease models and association tests. Power is estimated by the proportion of successes (e.g. $P \leq 0.05$) of the total number of independent replicates. Details of association tests in SEQPower can be found in Supplementary Table S1, and several power analysis examples are provided in Supplementary Material. Although for empirical analysis it is not feasible to directly calculate sample size, it is possible to create a grid search using a small number of replicates to find the approximate sample size. Because of computational burden, sample size estimation is best suited for RV association methods for which asymptotic *P*-values can be obtained [CMC, Sequence Kernel Association Test (Wu *et al.*, 2011)]. Methods for which *P*-values must be obtained empirically through permutation [Kernel-based Adaptive Clustering (Liu and Leal, 2010), Variable Threshold (Price *et al.*, 2010)], it can be computationally intensive to calculate power and sample sizes for small significant levels (e.g. $\alpha = 2.5 \times 10^{-6}$ for exome-wide NGS association studies), despite implementation of adaptive permutation (which reduces the number of permutations used to estimate non-significant *P*-values). SEQPower also has a mechanism to incorporate user-provided R scripts for assessment of type I error and power of novel association methods.

2.4 Performance

SEQPower is written in C++ and Python. It compiles and runs on most Unix/Linux systems and Mac workstations. We recently performed a power analysis using variant frequencies for European–Americans obtained from the NHLBI-ESP exome variant server (Tennessen *et al.*, 2012). For a sample size of 1000 cases and 1000 controls the power to detect an association for $\alpha = 2.5 \times 10^{-6}$ using the BRV method was evaluated for all genes within the exome with at least two variants sites, i.e. 19 044 genes. Using logistic regression, it took 6.0 min for analytic and 14.6 h for empirical power analyses on an Intel i7-3770 Quad Processor.

3 DISCUSSION

SEQPower is a practical tool for investigators to design adequately powered RV association studies. Although the true underlying genetic model is unknown, using a range of models, gene sizes, allelic architectures, effect sizes, etc., an investigator can determine whether a study has adequate power to at least detect associations with some of the genes involved in disease etiology. It is also an important tool to benchmark and comprehensively evaluate RV association tests, and also aid in the development of new statistical methodologies.

ACKNOWLEDGEMENTS

The authors would like to thank Ulrike Peters, Benjamin Neale and Dajiang Liu for their many useful comments and suggestions.

Funding: This work was supported by National Institutes of Health grants (HL102926, MD005964, HG006493 and HG005859) and additional grant support was provided by the MD Anderson Cancer Center (CA016672).

Conflict of Interest: none declared.

REFERENCES

Auer,P.L. *et al.* (2013) Testing for rare variant associations in the presence of missing data. *Genet. Epidemiol.*, **37**, 529–538.

Fleiss,J.L. *et al.* (1980) A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, **36**, 343–346.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Liu,D.J. and Leal,S.M. (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, **6**, e1001156.

Peng,B. and Liu,X. (2010) Simulating sequences of the human genome with rare variants. *Hum. Hered.*, **70**, 287–291.

Price,A.L. *et al.* (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, **86**, 832–838.

Purcell,S. *et al.* (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, **19**, 149–150.

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.

Tennessen,J.A. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.