BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 20 2012, pages 2703–2704
doi:10.1093/bioinformatics/bts499

Genetics and population analysis

Advance Access publication August 22, 2012

SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits

Biao Li[1,2], Gao Wang[1] and Suzanne M. Leal[1,2,*]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza and [2]Departments of Bioengineering and Statistics, Rice University, 6100 Main Street, Houston, TX 77030, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Currently, there is great interest in detecting complex trait rare variant associations using next-generation sequence data. On a monthly basis, new rare variant association methods are published. It is difficult to evaluate these methods because there is no standard to generate data and often comparisons are biased. In order to fairly compare rare variant association methods, it is necessary to generate data using realistic population demographic and phenotypic models.

Result: SimRare is an interactive program that integrates generation of rare variant genotype/phenotype data and evaluation of association methods using a unified platform. Variant data are generated for gene regions using forward-time simulation that incorporates realistic population demographic and evolutionary scenarios. Phenotype data can be obtained for both case–control and quantitative traits. SimRare has a user-friendly interface that allows for easy entry of genetic and phenotypic parameters. Novel rare variant association methods implemented in R can also be imported into SimRare, to evaluate their performance and compare results, e.g. power and Type I error, with other currently available methods both numerically and graphically.

Availability: http://code.google.com/p/simrare/

Contact: sleal@bcm.edu

Received on June 1, 2012; revised on July 19, 2012; accepted on August 8, 2012

1 INTRODUCTION

Currently, there is great interest in detecting complex trait rare variant associations. Next-generation sequencing technologies, e.g. Illumina HiSeq, ABI 454 and Roche SOLiD, have made it possible to cost-effectively identify rare variants through the generation of exome and whole-genome sequence data. It has been demonstrated that for complex traits testing individual rare variants is grossly underpowered (Gorlov *et al*., 2008). Therefore, a large number of rare variant association methods, which aggregate variants across a region, which is usually a gene, have been developed. However, comparing the power of these methods by reviewing the literature is difficult, because there is no regularity in data generation and sometimes to make a particular method appear more powerful than others, data are generated in an unrealistic manner. Although variant data can be obtained for example from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) and The National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (Tennessen *et al*., 2012), due to limited samples sizes, it is not possible to generate variant data which is reflective of the true distribution of rare variants (e.g. singletons and doubletons). It is therefore crucial to be able to incorporate evolutionary history to generate large samples of variant data.

The SimRare program was developed to evaluate Type I and II errors of rare variant complex trait association methods by generating both variant and phenotype data using realistic models. SimRare, which has a graphical interface, is written in Python/C++ and can incorporate R scripts for the evaluation of novel and existing methods.

2 DESCRIPTION

2.1 Generation of variant data

SimRare can generate variant data using several existing population genetics and demographic models (e.g. Boyko *et al*., 2008; Kryukov *et al*., 2009) or parameters for other models can be specified. Variant data are generated using forward-time simulation (Peng and Kimmel, 2005; Peng and Liu, 2010). This implementation is superior to other methods (Liu *et al*., 2008) in that it incorporates a realistic evolutionary model, which includes demography, variable mutation rates, recombination, multi-locus fitness effect and locus-specific selection models including purifying selection. Using these population demographic models, it is possible to generate variant data, for different populations, e.g. European and African. Additionally, haplotypes that have been generated using other simulation methods can be implemented. Variant data can also be generated based on the spectrum of variant frequencies estimated from exome or whole-genome sequence data such as the NHLBI-Exome Sequencing Project.

2.2 Generation of phenotype data

To evaluate Type I, it is only necessary to generate data under the null hypothesis, i.e. trait and variants are not associated; therefore it suffices to generate variant data for a region, unconditional on the phenotype. To evaluate Type II errors/power, it is necessary to generate trait data conditional on the variant data within a region. Either conditional or unconditional on the generated variant data, SimRare can generate phenotype data for both case–control and quantitative traits using for a wide range of genetic etiologic assumptions: including penetrance-prevalence model, liability threshold model, population attributable risk and linear model. The underlying genetic mode of inheritance can be specified, e.g. additive and dominant. For quantitative

*To whom correspondence should be addressed.

traits, a random or an extreme trait sample can be generated. When an extreme quantitative trait sample is generated, the ranges can be specified, e.g. upper 90% and lower 10%. All sample sizes can be user-specified. Causal variants within a 'gene' region can either have an effect which is unidirectional e.g. all variants increase disease risk or are bidirectional, e.g. can either decrease or increase quantitative trait values. A proportion of the variants can also be non-causal. For causal variants, the magnitude of the effect can either be fixed, i.e. all causal variants have the same effect size or variable, effect size is determined by frequency or selection coefficients.

2.3 Analysis of data and evaluation of Type I and II errors

It is possible to analyze the generated data using a wide range of complex trait rare variant association methods, e.g. combined multivariate and collapsing method (Li and Leal, 2008), weighted sum statistic (Madsen and Browning, 2009), variable threshold (Price *et al*., 2010) and Kernel-based adaptive cluster (Liu and Leal, 2010). Additionally, any newly developed rare variant association method written in R can be evaluated using SimRare. For case–control studies and randomly ascertained quantitative trait studies, qualitative and quantitative analysis is performed, respectively. While for an extreme quantitative trait sample, the data can be dichotomized and qualitative analysis performed or the quantitative trait values can be analyzed. For rare variant association methods where significance cannot be evaluated analytically, *P*-values can be obtained empirically through permutation. For the evaluation of Type I and II error, the user determines the number of permutations and/or replicates which should be generated.

2.4 Software overview

SimRare, a stand-alone executable software, is compiled for Windows, Mac and Linux operating systems. Installation of R is only required if it is desired to evaluate novel methods. SimRare consists of three major modules: simulation of sequence-based genotype data, generation of phenotype data and evaluation of association methods. SimRare has a user-friendly interface. Output from SimRare can either take the form of tables or graphs, e.g. QQ plots for Type I errors and bar graphs to display power. Additionally, simulated data can be written to external files in standard linkage format.

Benchmark tests were performed on a 64-bit machine with Intel 3.2 GHz CPU and 8 GB RAM installed with the Ubuntu Linux operating system. First 100 haplotype pools were generated using a population demographic model described by Kryukov *et al*. (2009) model; the time to generate each pool was 1.8 min. The generated haplotype pools can be used for all subsequent power analyses. The total computation time for the power analysis was 18.4 min when 1000 replicates were generated each with 1000 case and 1000 controls and power was evaluated using the combined multivariate and collapsing method (Li and Leal, 2008).

3 DISCUSSION

SimRare, as a simulation platform with integrated modules of data generation and statistical analysis, can be widely used at many different levels for rare variant association studies. SimRare can be used to impartially assess the power and robustness of novel and existing rare variant association methods under a broad range of scenarios. These features should greatly aid researchers in the development of rare variant association methods. SimRare can also assist genetic epidemiological study designs by allowing the power to be tested for a wide variety of scenarios. For instance, for a population-based case–control exome-sequence study SimRare can be used determine the necessary sample size to achieve adequate power under a variety of different phenotypic and genetic models.

SimRare will greatly increase the ability and speed in which researchers can develop novel methods. SimRare allows researchers to create methods without having to develop special software to generate data in order to evaluate their methods. It also facilitates an easy way to compare their method to existing methods, because evaluation can be performed on consistent datasets and additionally it is not necessary to write software to implement existing methods for comparison purposes. Ultimately, we expect this universally functional software equipped with user-friendly interface to aid genetic epidemiologists and statistical geneticist in the development of methods and the design of sequence-based genetic association studies.

Funding: This work was supported by the National Institute of Health (NIH) – National Institute of Heart Lung and Blood (NHLBI) grant number HL102926, NIH-National Center on Minority Health and Health Disparities (NCMHD) grant number MD005964 and NIH-National Human Genome Research Institute (NHGRI) grant number HG006493.

Conflict of Interest: none declared.

REFERENCES

Boyko,A.R. *et al*. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*., **4**, e1000083.

Gorlov,P.I. *et al*. (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet*., **82**, 100–112.

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

Kryukov,G.V. *et al*. (2009) Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci. USA*., **106**, 3871–3876.

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.

Liu,Y. *et al*. (2008) A survey of genetic simulation software for population and epidemiological studies. *Hum. Genomics*, **3**, 79–86.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*., **5**, e1000384.

Peng,B. and Kimmel,M. (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**, 3686–3687.

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.

The 1000 Genomes Project Consortium(2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.