Bioinformatics, 31(22), 2015, 3706–3708
doi: 10.1093/bioinformatics/btv412
Advance Access Publication Date: 14 July 2015
Applications Note

Genetic and population analysis

Generation of sequence-based data for pedigree-segregating Mendelian or Complex traits

Biao Li, Gao T. Wang and Suzanne M. Leal*

Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 10, 2015; revised on June 12, 2015; accepted on July 7, 2015

Abstract

Motivation: There is great interest in analyzing next generation sequence data that has been generated for pedigrees. However, unlike for population-based data there are only a limited number of rare variant methods to analyze pedigree data. One limitation is the ability to evaluate type I and II errors for family-based methods, due to lack of software that can simulate realistic sequence data for pedigrees.

Summary: We developed RarePedSim (*Rare*-variant *Pedi*gree-based *Sim*ulator), a program to simulate region/gene-level genotype and phenotype data for complex and Mendelian traits for any given pedigree structure. Using a genetic model, sequence variant data can be generated either conditionally or unconditionally on pedigree members' qualitative or quantitative phenotypes. Additionally, qualitative or quantitative traits can be generated conditional on variant data. Sequence data can either be simulated using realistic population demographic models or obtained from sequence-based studies. Variant sites can be annotated with positions, allele frequencies and functionality. For rare variants, RarePedSim is the only program that can efficiently generate both genotypes and phenotypes, regardless of pedigree structure. Data generated by RarePedSim are in standard Linkage file (.ped) and Variant Call (.vcf) formats, ready to be used for a variety of purposes, including evaluation of type I error and power, for association methods including mixed models and linkage analysis methods.

Availability and Implementation: bioinformatics.org/simped/rare

Contact: sleal@bcm.edu

1 Introduction

Currently, there is great interest in analyzing next generation sequence (NGS) data generated for pedigrees. For Mendelian traits, gene identification is far from complete (Antonarakis and Beckmann, 2006) and performing linkage analysis can greatly increase the ability to identify causal variants (Ott *et al.*, 2015). For complex traits, pedigree-based studies differ from population-based design, i.e. better control of population stratification (Bansal *et al.*, 2010) and potentially more powerful to detect associations, particularly for rare variants, due to causal allele enrichment (Ott *et al.*, 2011). However, only a limited number of methods have been developed to analyze pedigree sequence data, e.g. extended FBAT multimarker test (De *et al.*, 2013), RV-TDT (He *et al.*, 2014), SEQLinkage (Wang *et al.*, 2015). One bottleneck is the lack of available software to realistically generate rare variant pedigree data, to evaluate type I error, to compare statistical power of association and linkage methods and aid in the design of studies through the evaluation of sample size for sufficient power.

Although there are a number of programs that simulate common variants for pedigrees, they are not useful for NGS association and

linkage studies. Currently, only SimRare (Li *et al.*, 2012) and SeqSIMLA (Chung and Shih, 2013) can simulate sequence-based genotype and phenotype data for rare variants. However, their focus is on generating data for unrelated individuals. Although SeqSIMLA can also generate sequence data for pedigrees, it is strictly limited to complex traits and a three-generational family structure with a fixed number of individuals per generation.

RarePedSim was developed to simulate region/gene-level genotype data and phenotypes for complex and Mendelian pedigrees, regardless of structure. To evaluate type I error, variant data can be generated using Mendelian segregation independent of the phenotype data. For evaluating power, where the trait is linked/associated with causal variant(s), variant data are generated conditional on the phenotypes. For Mendelian traits, mode of inheritance, penetrance and phenocopies are incorporated in the model; for complex traits an odds ratio model is implemented, while for quantitative traits a linear model is used. Phenotype data for Mendelian, complex and quantitative traits can also be generated conditional on the simulated variant data. RarePedSim output files are in standard Linkage and VCF formats. RarePedSim can generate data with allelic and locus heterogeneity as well as with errors and missing genotypes.

2 Description

2.1 Generation of variant data

RarePedSim uses forward-time simulation (Peng and Liu, 2011) to generate region/gene-based variant data that incorporates haplotype information. Forward-time simulation is superior to other methods, e.g. coalescent, since it can incorporate realistic evolutionary scenarios that consist of multi-stage demography with varying population sizes, variant-specific selection coefficients and multi-locus fitness effects. Variant data for populations, including African and European can be generated using state-of-the-art demographic and purifying selection models (Gazave *et al.*, 2013). The simulated variant data are diallelic and each site is represented by: minor allele frequency (MAF), selection coefficient and genomic position, which are used to generate haplotypes. Variant data information can also be obtained from exome or genome sequence data from genetic epidemiological studies e.g. NHLBI-Exome Sequencing Project or databases e.g. ExAC.

2.2 Generation of pedigree segregating data for Mendelian Trait

For Mendelian trait studies, phenotypes are usually known for pedigree structures. Thus, to generate variants which are linked with the phenotype status, of which only some may be causal, genotypes are simulated conditional on observed phenotypic data by incorporating a user-specified penetrance (complete or reduced) model and variant data information, e.g MAF, rate of recombination and purifying selection coefficients. To simulate genotypes, the pedigree is first split into nuclear family block(s) connected by joining individuals (JIs), who are non-founders with offspring. Starting with the founders, within each nuclear family block, the conditional genotypic likelihoods are calculated for each pedigree member. Then based upon the joint likelihoods of the JIs the conditional genotype likelihoods are updated for all pedigree members and genotypes are assigned starting with the JIs. Additionally, phenotypes can be generated for all pedigree members conditional on simulated variants and user-specified penetrance models.

2.3 Generation of pedigree segregating data for Complex Trait

Within a region which is associated with a phenotype, generated variants can increase or decrease risk, modulate quantitative phenotypes or have no effect on phenotype (non-causal). For a qualitative phenotype a logistic odds ratio (LOGIT) model is used. For a quantitative trait (QT) linear mean-shift model (LNR) is incorporated, where baseline QT follows a standard Gaussian distribution and contribution of functional variants alters the QT distribution by a shift in the mean. To efficiently simulate genotype data conditional on observed phenotypes, the users can specify either a LOGIT or LNR model. The same strategy as described for Mendelian trait is used to simulate genotypes conditional on phenotypes for complex trait, with penetrance derived from the effect size of causal variants (odds ratio in LOGIT model and mean shift in LNR model). Additionally, phenotypes can be generated conditional on genotypes and disease/phenotypic model (LOGIT or LNR), where for each individual the probability of being affected is obtained for a qualitative trait or the continuous probability distribution for a quantitative trait and either an affection status or a QT value is assigned, respectively.

2.4 Software overview

Benchmarks were performed on a 64-bit Linux machine with Intel Core i7-4770 k 3.5 GHz CPU and 16GB RAM. Using a population demographic model described by Gazave *et al.* (2013) it took 1.7 minutes to simulate a pool of 1 308 000 haplotypes of 1800 bp in length. Sampling from the haplotype pool, conditional on disease phenotypes it took per replicate 2.9 seconds to generate two susceptibility genes for a 57 member autosomal dominant pedigree with reduced penetrance and 3.2 s using a LOGIT fixed effect model to generate a disease associated gene for 2000 complex trait trios.

One of important applications of RarePedSim is to perform power analysis. For examples, a gene with 65 causal rare variant sites (average MAF = 0.14%) was generated: 1) for a Mendelian disease, variants were generated using an autosomal dominant model with 45% penetrance and 5% phenocopy rate for 562 nuclear families with 2-6 offspring and 1235 affected individuals in total and 2) for a complex trait, with 1% prevalence, variants were generated under a LOGIT model using an odds ratio of 2.5 to simulate data for 2000 trios. For each study 1000 replicates were generated. The Mendelian trait data was analyzed using parametric linkage analysis by implementing SEQLINKAGE and the estimated power was 78%. For the trios, association analysis was performed using RV-TDT and the estimated power was 65%.

For additional documentation, tutorials, examples and mathematical derivations of simulation method, see the RarePedSim URL http://bioinformatics.org/simped/rare

3 Conclusion

RarePedSim is written in Python and C++ and provides a user-friendly command-line interface. The program provides key functionalities to generate region/gene-based genotype pedigree data conditional or unconditional on phenotypes. The generated data are ready to be used for evaluating type I and II errors for both family-based association methods, including mixed models, and parametric and nonparametric linkage methods. RarePedSim can be used by researchers to efficiently generate pedigree variant data without having to develop special software. Ultimately RarePedSim will facilitate the development of novel methods to analyze rare variant

pedigree data for a variety of genetic etiologies and aid in the design of studies.

Funding

National Institute of Health (grants numbers DC011651, DC003594 and HG006493).

Conflict of Interest: none declared.

References

Antonarakis,S.E. and Beckmann,J.S. (2006) Mendelian disorders deserve more attention. *Nat. Rev. Genet.*, **7**, 277–282.

Bansal,V. *et al.* (2010) Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.*, **11**, 773–785.

Chung,R.H. and Shih,C.C. (2013) SeqSIMLA: a sequence and phenotype simulation tool for complex disease studies. *BMC Bioinformatics*, **14**, 199.

De,G. *et al.* (2013) Rare variant analysis for family-based design. *PLoS ONE*, **8**, e48495.

Gazave,E. *et al.* (2013) Neutral genomic regions refine models of recent rapid human population growth. *Proc. Natl. Acad.Sci.*, **111**, 757–762.

He,Z. *et al.* (2014) Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am. J. Hum. Genet.* **94**, 33–46.

Li,B. *et al.* (2012) SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. *Bioinformatics*, **28**, 2703–2704.

Ott,J. *et al.* (2011) Family-based designs for genome-wide association studies. *Nat. Rev. Genet.*, **12**, 465–474.

Ott,J. *et al.* (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* **16**, 275–284.

Peng,B. and Liu,X. (2011) Simulating sequences of the human genome with rare variants. *Hum. Hered.*, **70**, 287–291.

Wang,G.T. *et al.* (2015) Collapsed haplotype pattern method for linkage analysis of next-generation sequencing data. *Eur. J. Hum. Genet.* doi: 10.1038/ejhg.2015.64.