Bioinformatics, 36(7), 2020, 2295–2297
doi: 10.1093/bioinformatics/btz881
Advance Access Publication Date: 25 November 2019
Applications Note

Genetics and population analysis

SimRVSequences: an R package to simulate genetic sequence data for pedigrees

Christina Nieuwoudt[1], Angela Brooks-Wilson[2,3] and Jinko Graham[1,*]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A 1S6, [2]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 1L3 and [3]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on July 5, 2019; revised on November 12, 2019; editorial decision on November 20, 2019; accepted on November 22, 2019

Abstract

Summary: We present the R package SimRVSequences to simulate sequence data for pedigrees. SimRVSequences allows for simulations of large numbers of single-nucleotide variants (SNVs) and scales well with increasing numbers of pedigrees. Users provide a sample of pedigrees and SNV data from a sample of unrelated individuals.

Availability and implementation: SimRVSequences is publicly-available on CRAN https://cran.r-project.org/web/packages/SimRVSequences/.

Contact: jgraham@sfu.ca

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

1 Introduction

Family-based studies are comprised of pedigrees, which may be small or may include many individuals and span several generations. Pedigrees may belong to a population cohort, or they may be ascertained on the basis of disease-affected relatives. Family-based studies are increasingly deployed to identify causal rare variants (*c*RVs) for a disease because of their increased power (Wijsman, 2012). To simulate data for such studies we have developed the R package SimRVSequences.

Genetic association studies may fail to identify *c*RVs because of genetic heterogeneity, which occurs when different genetic mutations predispose to the same disease. The disease-predisposing variants may be restricted to a single gene, i.e. allelic heterogeneity, or they may occur on different genes, i.e. locus heterogeneity (Nussbaum *et al*., 2007). For example, there is evidence of locus heterogeneity in familial Alzheimer's disease since different loci residing on chromosomes 14, 19 and 21 predispose carriers to Alzheimer's disease (Tanzi *et al*., 1993). Cystic Fibrosis is known to exhibit allelic heterogeneity; to date, thousands of mutations in the CFTR gene have been associated with cystic fibrosis in carriers (Nussbaum *et al*., 2007). SimRVSequences allows users to model genetic heterogeneity and can simulate genetic heterogeneity in a user-specified pathway.

Several approaches have been proposed to simulate genetic sequence data for pedigrees; e.g. the R package sim1000G (Dimitromanolakis *et al*., 2019) and the C++ program RarePedSim (Li *et al*., 2015). Both of these programs reduce user burden by simulating founder haplotypes for pedigree data; however, neither can accommodate simulations over a large number of single-nucleotide variants (SNVs). SimRVSequences can accommodate genome-wide, exon-only simulations over a large number of SNVs and scales well with large numbers of pedigrees. To our knowledge, SimRVSequences is the only R package that simulates genetic sequence data conditionally given the *c*RV status of family members. This feature allows users to specify *c*RVs prior to simulating genetic data in pedigrees. For additional information regarding the comparability and features of sim1000G, RarePedSim and SimRVSequences please refer to Supplementary Material S3.

2 Description

SimRVSequences is a publicly-available, platform-independent R package for R $\geq$ 3.5.0, which can be obtained from the comprehensive R archive network. To simulate sequence data for family-based studies users provide: a sample of pedigrees, and SNV data from unrelated individuals representing the population of pedigree founders.

2.1 Pedigree data

SimRVSequences can accept pedigree data from any source provided that pedigrees are properly formatted. In particular, users who wish to simulate the segregation of known *c*RVs must specify the *c*RV status of each pedigree member. For ease of use, we note that pedigrees simulated by the R package SimRVPedigree are appropriately formatted for SimRVSequences and include the required genotypes at the disease locus. As described in Nieuwoudt *et al*. (2018), SimRVPedigree simulates pedigrees forward-in-time by way of a competing-risk model and makes use of age-specific hazard rates of

disease onset and death provided by the user. At the individual level, disease onset is influenced by the presence (or absence of) a *c*RV, by way of a proportional-hazards model. At the family level, SimRVPedigree models complex ascertainment so that simulated pedigrees are representative of sampled pedigrees. We emphasize that users are not required to use pedigrees simulated by SimRVPedigree. For information regarding pedigree formatting please refer to Supplementary Material S1.

2.2 SNV data

By convention, pedigree members can be classified as founders or non-founders; non-founders have both a mother and a father, while founders have neither. Users are required to provide haplotype data for pedigree founders in the form of SNV data. SimRVSequences can accept SNV data obtained from various sources provided they are properly formatted. To reduce user burden, SimRVSequences includes methods to (i) assist with exon-only simulations and (ii) import and format SNV data simulated by the freely-available software SLiM (Haller and Messer, 2017). Additionally, SimRVSequences provides users with pre-formatted, genome-wide, exon-only SNV data from the 1000 Genome Project (1000 Genomes Project, 2010). Users who prefer to import and format SNV data from VCF files may follow the instructions provided in Supplementary Material S1.

2.3 Pathway implementation and selection of founder haplotypes

After formatting the SNV data, the user must specify which mutations will be modelled as *c*RVs. We model genetic heterogeneity through the specification of a pool of *c*RVs from a user-defined pathway. Users may implement allelic heterogeneity among families by selecting *c*RVs in the same gene. For example, users who wish to simulate Cystic Fibrosis should select SNVs from the CFTR gene to be modelled as *c*RVs. A pedigree may segregate at most one *c*RV which is sampled with replacement from the user-specified pool of *c*RVs according to its relative frequency. Upon identifying the familial *c*RV, the haplotypes for each pedigree founder are sampled from the user-supplied distribution of haplotypes conditioned on the founder's *c*RV status at the familial disease locus.

2.4 Simulation of sequence data for non-founders

To simulate sequence data for non-founders we perform a conditional gene drop, which can be described as a two-part process. Given a parent-offspring pair, we first simulate the formation of gametes from the parental haplotypes. To accomplish this, we employ the model proposed by Voorrips and Maliepaard (2012) for recombination with chiasmata interference. Although the proposed model can be applied to tetraploids, in SimRVSequences we restrict attention to diploid organisms. Upon simulation of parental gametes, we next simulate the conditional inheritance of a gamete given the offspring's *c*RV status at the familial disease locus. To simulate genetic data in an entire pedigree, we repeat this process, forwards-in-time, for each parent-offspring pair to ensure the correct genetic variability among relatives. For additional details regarding the conditional gene-drop algorithm, please refer to Supplementary Material S1.

2.5 Example

SimRVSequences provides exon-only SNV data from the 1000 Genomes Project which has already been formatted for use. The function load_1KG is used to load the pre-formatted data by chromosome. For example, to load the SNV data for chromosome 1 we execute the command: load_1KG(chrom = 1).

The function sim_RVstudy is used to simulate SNV data for pedigrees. This function has two required arguments: ped_files and SNV_data. The user is expected to provide a sample of pedigrees to ped_files and SNV data from unrelated individuals representing the population of pedigree founders to SNV_data. After appropriately formatting these data objects they are supplied to sim_RVstudy as follows: sim_RVstudy(ped_files = example_peds, SNV_data = example_SNVdata). We note that sim_RVstudy also includes several optional arguments, which are discussed at length in Supplementary Material S1.

3 Discussion

Although gene-dropping can be computationally expensive, SimRVSequences stores and extracts haplotypes efficiently (Cheng *et al*., 2015; Voorrips and Maliepaard, 2012) for improved performance in R. Simulating the transmission of 178 430 rare SNVs across the 22 human autosomes for the disease-affected relatives in 1 pedigree requires approximately 0.5 seconds on a Windows OS with an i7-4790 @ 3.60 GHz and 12 GB of RAM. For 200 pedigrees the same simulation requires approximately 2 minutes. Due to data-allocation limitations in R, users who wish to simulate more than 250 000 rare SNVs may need to run chromosome-specific analyses.

To assess the validity of the simulated data, we performed analyses to check that the recombination frequency and genetic kinship matched theoretical expectations. Details of these analyses may be found in Supplementary Material S2.

We emphasize that, while SimRVSequences can model genetic heterogeneity among families, we assume that within a family genetic cases are due to a single causal SNV. Hence, SimRVSequences is not appropriate for users who wish to simulate the transmission of multiple causal SNVs in the same pedigree. Furthermore, SimRVSequences does not allow for simulation of sequence data in sex chromosomes.

4 Conclusion

SimRVSequences allows for efficient simulation of SNV data for a sample of pedigrees in R. Users may model genetic heterogeneity and incorporate pedigrees that include individuals with sporadic disease in simulation. Additionally, users may take advantage of several functions to import and format SNV data from various sources including pre-formatted, exon-only SNV data from the 1000 Genomes Project.

Acknowledgement

The authors thank undergraduate summer research student Wen Tian Wang (Simon Fraser University) for her assistance in processing the 1000 Genomes Project Data for use with SimRVSequences.

Funding

This work was supported in part by the Natural Science and Engineering Research Council of Canada (RGPIN-2018-04296), the Canadian Statistical Sciences Institute (CTRMS-342085-2014) and the Canadian Institutes of Health Research (MOP-130311).

Conflict of Interest: none declared.

References

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

Cheng,H. *et al*. (2015) XSim: simulation of descendants from ancestors with sequence data. *G3 (Bethesda)*, **5**, 1415–1417.

Dimitromanolakis,A. *et al*. (2019) sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, **20**, 1.

Haller,B.C. and Messer,P.W. (2017) SLiM 2: flexible, interactive forward genetic simulations. *Mol. Biol. Evol*., **34**, 230–240.

Li,B. *et al*. (2015) Generation of sequence-based data for pedigree-segregating Mendelian or complex traits. *Bioinformatics*, **31**, 3706–3708.

Nieuwoudt,C. *et al*. (2018) Simulating pedigrees ascertained for multiple disease-affected relatives. *Source Code Biol. Med*., **13**, 2.

Nussbaum,R.L. *et al.* (2007) *Patterns of Single-Gene Inheritance. Thompson & Thompson Genetics in Medicine*, 7th edn. Saunders/Elsevier, Philadelphia, pp. 115–149.

Tanzi,R. *et al.* (1993) Genetic heterogeneity of gene defects responsible for familial Alzheimer disease. *Genetica*, **91**, 255–263.

Voorrips,R.E. and Maliepaard,C.A. (2012) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*, **13**, 248.

Wijsman,E.M. (2012) The role of large pedigrees in an era of high-throughput sequencing. *Hum. Genet.*, **131**, 1555–1563.