GENLIB: new function to simulate haplotype transmission in large complex genealogies

Mohan Rakesh[1], Hélène Vézina[2,3,4], Catherine Laprise[4,5], Ellen E Freeman[1,6], Kelly M Burkett[7], Marie-Hélène Roy-Gagnon[1]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada; [2]Projet BALSAC, Université du Québec à Chicoutimi, Chicoutimi, Canada; [3]Département des sciences humaines et sociales, Université du Québec à Chicoutimi, Chicoutimi, Canada; [4]Centre intersectoriel en santé durable, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada; [5]Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada; [6]Centre de Recherche, Hôpital Maisonneuve-Rosemont, Montréal, Québec, Canada; [7]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

Abstract

Summary

Founder populations with deep genealogical data are well suited for investigating genetic variants contributing to diseases. Here, we present a new function added to the genealogical analysis R package GENLIB, which can simulate the transmission of haplotypes from founders to probands along very large and complex user-specified genealogies.

Availability and implementation

The new function is available in the latest version of the GENLIB package (v1.1.6), available on the CRAN repository and from https://github.com/R-GENLIB/GENLIB. Stand-alone scripts for analyzing the output of the function can be accessed at https://github.com/R-GENLIB/simuhaplo_scripts.

Introduction

Founder populations have been utilized extensively in the study of Mendelian diseases because they can have higher incidence rates of rare autosomal-recessive genetic diseases due to drift effects, e.g.: Gaucher disease, Tay-Sachs disease, and cystic fibrosis in the Ashkenazi Jewish population (Charrow, 2004), or one of the over 30 identified autosomal recessive diseases with elevated frequency in the Finnish population (Norio, 2003; Pastinen, et al., 2001). In founder populations, affected individuals are more likely to have the causal mutation on longer haplotypes that are homozygous by recent decent, aiding in mutation discovery (Bourgain and Genin, 2005; Libiger and Schork, 2007). Some founder populations have extensive records allowing for reconstruction of deep and large genealogies (Falchi, et al., 2004; Liu, et al., 2007; Ober, et al., 2001; Vézina and Bournival, 2020). Gene-dropping simulations (Chen, et al., 2015; Maccluer, et al., 1986) can be performed within these genealogies, wherein ancestral genotypes are passed down a fixed pedigree structure. For example, allele-dropping was used to study mutation frequencies in the Hutterite (Chong, et al., 2012), and French-Canadian (Heyer, 1999) founder populations.

Gene-dropping is not limited to dropping specific alleles. Transmission of genomic regions, chromosomes, or even the entire genome can be simulated. This type of simulations can provide important information on the distribution of genomic sharing and the probability of sharing a specific genomic segment among close or distant relatives, and can identify specific founders and transmission paths responsible for the observed sharing. However, in very large genealogies, these gene-dropping simulations are computationally feasible only if one does not consider the allelic state of any specific locus, but rather only the positions of recombination events and the origin (founder) of the segments bounded by the crossovers (Cheng, et al., 2015). We have implemented such a gene-dropping simulation tool in the GENLIB genealogical analysis R package (Gauvin, et al., 2015; R Core Team, 2021). This new tool (named gen.simuhaplo) is fast even for large genomic regions and deep

genealogies with many individuals because it does not consider any alleles, mutations, or phenotypes. To our knowledge, it is the first user-friendly simulation tool that can perform gene-dropping simulations of long genomic segments in very large and complex genealogies, such as those available in the French-Canadian founder population, while allowing the ability to retrace all transmission paths.

Implementation and usage

The gen.simuhaplo function is written in C++ and uses the Rcpp package (Eddelbuettel and Francois, 2011), as well as the R core C API to interface with the R environment.

The function simulates meiosis using one of three possible models (see Supplementary Appendix 1 for details). The first is a Poisson process, which is a no-interference model of meiosis (Haldane, 1919). The second model is a count-location model (Karlin and Liberman, 1978; Karlin and Liberman, 1979), where the number of chiasma is sampled from a zero-truncated Poisson distribution (Risch and Lange, 1979; Sturt, 1976). This model accounts for the obligate chiasma phenomenon, i.e., requirement for at least one chiasma per tetrad (Fledel-Alon, et al., 2009). The third model is a stationary gamma process (Broman and Weber, 2000), which accounts for chromosomal interference. After the locations of the crossovers are obtained in Morgans they are converted from genetic distance to physical distance and a meiotic product is selected and transmitted. The user may provide a map to convert genetic distance to physical distance, or else the relationship between genetic and physical distance will be assumed to be linear across the length of the chromosome. The choice of model and the use of a genetic-physical map can alter the distribution of the lengths of segment identical-by-descent (IBD) (Caballero, et al., 2019).

After importing the genealogical data and creating a “genealogy” object within GENLIB, the gen.simuhaplo function can be called. First, a list of individuals is created such that any dependent individual is listed after their ancestors (i.e.: parents before children, and all ancestors before

descendants). Since a family tree is a directed acyclic graph, we can always perform such a topological sort. Then we iterate through the list, and, at each individual, we simulate meiosis in the parents and pass down the meiotic products. A meiotic product is created by first obtaining a list of chiasma positions (in base pairs, BP), then assuming that each chiasma will have a probability of 0.5 of appearing in the given meiotic product as a crossover. After selecting all the crossovers, we select the parental copy with which the meiotic product begins with probability of 0.5. We copy over the parental chromosome until reaching a crossover position, then we alternate to copying from the other parental chromosome. The chromosomes are stored as linked lists where each segment contains its end position in BP, and points to the next segment in the chromosome.

The output of the function is a text file containing the location of the breakpoints and the code for the founder haplotypes, from which the simulated haplotypes can be derived for all the specified probands. Optionally the function can output a second text file containing the haplotypes for all individuals along the inheritance paths. The optional second text file contains a recombination history that allows users to easily retrace genomic segments from probands up to internal ancestors (see Supplementary Appendix 2). An example of the output format is shown in Figure 1. Length of segments shared, or position shared on the simulated segment, can be obtained from this output even though no actual genotypes at specific markers in the segment are used in the simulations. If users need to simulate with genotype data, we provide a Perl script for converting the output into genotype data (see Supplementary Appendix 2). The user can provide haploid genotypes for the founders and the proband haplotypes can be converted into corresponding genotype data (phasing into haplotypes can then be ignored if desired). This could be used to distinguish between alleles shared identical by descent versus by state.

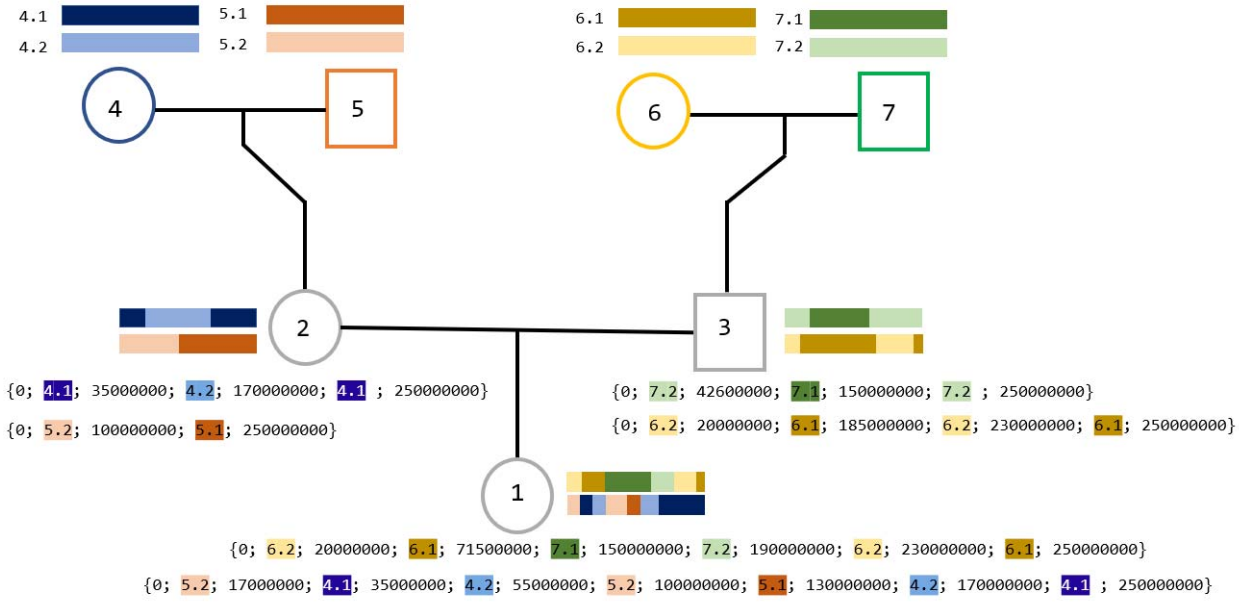

Figure 1. Example simulation of a hypothetical 250,000,000 BP segment. Each individual in the genealogy has a unique integer ID. This ID number is used to label founder chromosomes, e.g.: founder "4" will have chromosomes labelled "4.1" and "4.2". All founder chromosomes are labelled in this manner. The function iterates through all individuals. For every non-founder individual, we simulate meiosis in both parents and pass down a selected meiotic product from each parent. The notation inside the curly braces demonstrates how the haplotypes appear in the text file output. Segments are identified by their founder of origin (ID #), and the boundary positions are recorded in BP.

Other available gene-dropping software are not designed to efficiently simulate transmission of genomic regions through large genealogies. This is due to software either handling only few alleles (making them unable to simulate large regions), tracking a large number of alleles (making them inefficient for large genealogies and many replicates), or having additional functionalities (e.g.,

handling phenotypes). A detailed comparison to other software is provided in the Supplementary Appendix 4.

The output haplotypes describe the true IBD origin for each of the continuous segments that make up the simulated region. Hence, results obtained with our simulation function can help characterize patterns of genomic sharing between specific individuals and overall in the population. The output can be used to describe the distribution of haplotype lengths in the population or for specific individuals, obtain the distribution of the length of IBD segments shared by a pair of individuals, or estimate the likelihood of an IBD segment greater than a given length being transmitted from any ancestor. Simulation results can also be used to compare different statistical methods, as illustrated in Burkett, et al., 2022 who used a beta version of the function (with limited functionalities) to compare genomic- and genealogical/coalescent-based inference of homozygosity by descent in two different pedigree structures from the French-Canadian Founder population. In Supplementary Appendix 3, we provide additional examples of application of the function, highlighting all available functionalities.

Conclusion

The gen.simuhaplo function combines the GENLIB R package's existing support for handling large genealogies to allow users to simulate inheritance of large genomic regions even in genealogies with hundreds of thousands of individuals. To our knowledge no other simulators with similar functionalities supports such large genealogies.

References

Bourgain, C. and Genin, E. Complex trait mapping in isolated populations: Are specific statistical methods required? *Eur J Hum Genet* 2005;13(6):698-706.

Broman, K.W. and Weber, J.L. Characterization of human crossover interference. *American Journal of Human Genetics* 2000;66(6):1911-1926.

Burkett, K.M.*, et al.* Correspondence Between Genomic- and Genealogical/Coalescent-Based Inference of Homozygosity by Descent in Large French-Canadian Genealogies. *Front Genet* 2022;12:808829.

Caballero, M.*, et al.* Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genetics* 2019;15(12):1-29.

Charrow, J. Ashkenazi Jewish genetic disorders. *Familial Cancer* 2004;3(3-4):201-206.

Chen, H.S.*, et al.* Genetic simulation tools for post-genome wide association studies of complex diseases. *Genetic Epidemiology* 2015;39(1):11-19.

Cheng, H., Garrick, D. and Fernando, R. Xsim: Simulation of descendants from ancestors with sequence data. *G3: Genes, Genomes, Genetics* 2015;5(7):1415-1417.

Chong, J.X.*, et al.* A population-based study of autosomal-recessive disease-causing mutations in a founder population. *American Journal of Human Genetics* 2012;91(4):608-620.

Eddelbuettel, D. and Francois, R. Rcpp: Seamless R and C plus plus Integration. *Journal of Statistical Software* 2011;40(8):1-18.

Falchi, M.*, et al.* A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia. *Am J Hum Genet* 2004;75(6):1015-1031.

Fledel-Alon, A.*, et al.* Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genetics* 2009;5(9):1-7.

Gauvin, H.*, et al.* GENLIB: an R package for the analysis of genealogical data. *BMC bioinformatics* 2015;16:160.

Haldane, J.B.S. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* 1919;8(4):299-309.

Heyer, E. One Founder / One Gene Hypothesis in a New Expanding Population : Saguenay (Quebec , Canada). *Human Biology* 1999;71(1).

Karlin, S. and Liberman, U. Classifications and comparisons of multilocus recombination distributions. *Proc Natl Acad Sci U S A* 1978;75(12):6332-6336.

Karlin, S. and Liberman, U. A natural class of multilocus recombination processes and related measures of crossover interference. *Advances in applied probability* 1979;11(3):479-501.

Libiger, O. and Schork, N.J. A simulation-based analysis of chromosome segment sharing among a group of arbitrarily related individuals. *European Journal of Human Genetics* 2007;15(12):1260-1268.

Liu, F., *et al.* A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. *Am J Hum Genet* 2007;81(1):17-31.

Maccluer, J.W., *et al.* PEDIGREE ANALYSIS BY COMPUTER-SIMULATION. *Zoo Biology* 1986;5(2):147-160.

Norio, R. The Finnish disease heritage III: The individual diseases. In, *Human genetics*. Springer Verlag; 2003. p. 470-526.

Ober, C., Abney, M. and McPeek, M.S. The genetic dissection of complex traits in a founder population. *Am J Hum Genet* 2001;69(5):1068-1079.

Pastinen, T., *et al.* Dissecting a population genome for targeted screening of disease mutations. *Human Molecular Genetics* 2001;10(26):2961-2972.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. http://www.R-project.org/

Risch, N. and Lange, K. An alternative model of recombination and interference. *Ann Hum Genet* 1979;43(1):61-70.

Sturt, E. A mapping function for human chromosomes. *Annals of Human Genetics* 1976;40(2):147-163.

Vézina, H. and Bournival, J.S. An overview of the BALSAC database: past developments, current state and future prospects. *Historical Life Course Studies* 2020;11(2):1-17.