Original Paper

Human Heredity

Hum Hered 2005;60:119–122
DOI: 10.1159/000088914

Received: July 25, 2005
Accepted: August 8, 2005
Published online: October 13, 2005

SimPed: A Simulation Program to Generate Haplotype and Genotype Data for Pedigree Structures

Suzanne M. Leal[a] Kai Yan[a] Bertram Müller-Myhsok[b]

[a]Department of Molecular and Human Genetic, Baylor College of Medicine, Houston, Tex., USA;
[b]Computational Genetics Group, Max-Planck-Institute of Psychiatry, Munich, Germany

Key Words
Simulation · Pedigree structure · Type I error · Empirical p values

Abstract
With the widespread availability of SNP genotype data, there is great interest in analyzing pedigree haplotype data. Intermarker linkage disequilibrium for microsatellite markers is usually low due to their physical distance; however, for dense maps of SNP markers, there can be strong linkage disequilibrium between marker loci. Linkage analysis (parametric and nonparametric) and family-based association studies are currently being carried out using dense maps of SNP marker loci. Monte Carlo methods are often used for both linkage and association studies; however, to date there are no programs available which can generate haplotype and/or genotype data consisting of a large number of loci for pedigree structures. SimPed is a program that quickly generates haplotype and/or genotype data for pedigrees of virtually any size and complexity. Marker data either in linkage disequilibrium or equilibrium can be generated for greater than 20,000 diallelic or multiallelic marker loci. Haplotypes and/or genotypes are generated for pedigree structures using specified genetic map distances and haplotype and/or allele frequencies. The simulated data generated by SimPed is useful for a variety of purposes, including evaluating methods that estimate haplotype frequencies for pedigree data, evaluating type I error due to intermarker linkage disequilibrium and estimating empirical p values for linkage and family-based association studies.

Introduction

Dense maps of markers are currently available for both linkage studies and family-based association studies [1–4]. A number of programs allow for the generation of genotype data, with the markers in linkage equilibrium conditional or unconditional on the disease phenotype (i.e. SIMLINK [5], ALLEGRO [6], SLINK [7]) or only generate genotype data unconditional on the disease phenotype (i.e. SIMULATE [8, 9], MERLIN [10]). Although the ALLEGRO program can generate genotype data for a large number of marker loci, it is limited to simulating genotype data for small to medium sized pedigree structures. MERLIN, SLINK, SIMLINK and SIMULATE can all generate genotype data for large pedigree structures. While MERLIN and SIMULATE are both able to generate genotype data for a large number marker loci, the SIMLINK and SLINK programs are both limited in the number of markers for which they can generate genotype data. Besides being able to generate genotype data the SLINK program [10] can also generate haplotype data where the markers are in linkage disequilibrium either conditional or unconditional on the disease phenotype. However, the SLINK program is restricted to generating

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2005 S. Karger AG, Basel
0001–5652/05/0602–0119$22.00/0

Accessible online at:
www.karger.com/hhe

Suzanne M. Leal, PhD
Baylor College of Medicine, Department of Molecular and Human Genetics
One Baylor Plaza, N1619.01
Houston, TX 77025 (USA)
Tel. +1 713 798 4011, Fax +1 713 798 5373, E-Mail sleal@bcm.tmc.edu

data for a very small number of marker loci. For restricted pedigree structures (i.e. 3 generations with 3 sibships with each sibship containing 2–10 individuals and nuclear families) the SIMLA program [11] can generate marker data in linkage disequilibrium or linkage equilibrium. The markers generated can either be in linkage equilibrium or disequilibrium with the simulated susceptibility locus, and the marker loci and the susceptibility locus may be generated either linked or unlinked to each other. The SIMLA program simulates one susceptibility locus for each pedigree and uses specified penetrances to determine the affection status of the pedigree members. Due to the limitation on pedigree structure and simulation of the susceptibility locus, for most situations it is not possible to generate data for ascertained pedigree structures using the SIMLA program; it would therefore, for example not usually be possible to generate data using SIMLA to estimate empirical p values.

Currently SimPed is the only available program that can generate haplotypes and/or genotypes for a large number of marker loci regardless of the pedigree structure. Previously, microsatellite markers were employed for nearly all linkage studies [12, 13]. In most cases, for genome scan data consisting of microsatellite data, the amount of intermarker linkage disequilibrium between marker loci is negligible and it is usually valid to carry out linkage analysis assuming that the markers are in linkage equilibrium. For this situation it is usually appropriate to carry out simulation studies with the marker loci in linkage equilibrium. Currently many linkage studies carry out genome scans using dense maps [1, 4] of single nucleotide polymorphism (SNP) marker loci instead of using microsatellite marker loci. Additionally, family based association studies are carried out using dense maps of SNP marker throughout the genome or within candidate regions [2, 14]. For these dense maps of SNP marker loci there can be high levels of intermarker linkage disequilibrium. In order to simulate data which is representative of these dense maps of marker loci, it is important to be able to generate a large number of marker loci in linkage disequilibrium or a mixture of markers in linkage disequilibrium and linkage equilibrium for pedigree data. In order to meet this need the SimPed program was developed.

The data generated by the SimPed program can be used for a variety of analysis purposes. A few examples of how the data generated by SimPed can be used include: evaluating methods that estimate haplotype frequencies for pedigree data, assessing type I error due to intermarker linkage disequilibrium and estimating empirical p values for linkage and family-based association studies.

Methods

The SimPed program generates haplotype and/or genotype data for pedigree structures unconditional on the disease/quantitative trait status. Haplotype and/or genotype data can be generated either for the autosomes or the X chromosome. The pedigrees for which haplotype and/or genotype data is generated may be very large (>2,500 individuals) and may contain multiple consanguinity and marriage loops. Haplotypes and their frequencies are user-specified and can be either estimated from the investigator's data or from other sources such as the International HapMap project (www.hapmap.org) [14]. For genotype data, allele frequencies must be provided which can either be estimated from the user's data or obtained from public sources. It is possible to generate haplotypes or genotypes or a combination of haplotypes and genotypes for >20,000 marker loci, thus making it possible to simulate an entire chromosome or genome worth of marker loci. These loci can either be diallelic markers (e.g. SNPs) or multiallelic markers (e.g. microsatellites). Intermarker recombination or genetic map distances can be incorporated into the simulation of the haplotype and/or genotype data. The user provides intermarker recombination fractions or genetic map distances obtained from genetic maps [1, 3] or through interpolation. If no genetic map is available for the markers of interest, SNP marker loci can be ordered based upon their sequence-based physical map position and then interpolated onto a genetic map – for example, the Rutgers Combined Linkage-Physical Map [3] or the DeCode genetic map [13, 15].

The user must provide the SimPed program with two files. One file contains the pedigree structure(s) in standard linkage format (e.g. GENEHUNTER [16]) with or without a disease/quantitative trait locus. Additional column(s) in this file denote for which marker loci data is available. The parameter file contains information on genetic map distances/intermarker recombination fractions, haplotype and allele frequencies, and the number of replicates to be simulated. It is possible to efficiently specify genetic map distances and haplotype and allele frequencies for thousands of marker loci due to the format of the parameter file. The SimPed program is flexible, and it is possible to acquire haplotypes/genotypes for only a subset of family members or make unknown the genotypes for a subset of marker(s) for specified family members.

The program can be used to simulate data for large pedigrees; for example, both haplotype and genotype data was generated for a 6-generational pedigree with 2,827 members of which 472 family members were founders. The SimPed program was also used to generate haplotypes and genotypes for a pedigree with 11 consanguinity loops.

The SimPed program generates haplotype and/or genotype data for pedigrees as follows. For the autosomes all of the founders within the pedigree are assigned two haplotypes and/or two alleles conditional on the user specified frequencies for all of the marker loci. Once assignment is completed, each founder has two haplotypes. Starting at the top of the pedigree structure, the first offspring of the founder is randomly assigned one of the founder's haplotypes. The allele at the first marker from this haplotype is assigned to the offspring. It is then determined, based upon the genetic map, whether a recombination event has occurred between the first and second marker loci. If with probability θ a recombination event has occurred, then at the second marker locus the allele from the founder's other haplotype is assigned to the offspring. If a recombination event has not occurred with probability $(1 - \theta)$ then the allele from the founder's same haplotype is assigned at the second marker locus. This procedure is repeated until alleles for all markers' loci have been assigned from one founder to their offspring. The process is then repeated, this time assigning alleles to the offspring from their other parent. This procedure varies slightly for the simulation of marker loci on the X chromosome. Since males are hemizygous all founder males are allocated one haplotype and/or allele conditional on the specified frequencies for all of the marker loci. Once assignment is complete for all marker loci the haplotype is duplicated, since the standard LINKAGE pedigree file format is for males to be homozygous for all genotypes on the X chromosome. The haplotypes for the X chromosome for the founder females are determined using the same method as was applied for the autosomes. For non-founder males it is decided where recombination events occurred between the two maternal haplotypes as was done for the autosomes. Once the haplotype for the non-founder male is determined it is then duplicated. For female non-founders one paternal haplotype is assigned and the maternal haplotype is determined in exactly the same way as it was accomplished for the autosomes. In this manner, the haplotypes flow down the pedigree tree as all non-founders are assigned haplotypes conditional on parental haplotypes. Once all individuals within the pedigree have been assigned haplotypes, for those individuals/marker loci for which it was specified that they are unavailable, the genotypes are made unknown (i.e. 0 0).

Results and Discussion

Benchmarks were carried out for a variety of pedigree structures. For each pedigree structure both haplotype and genotype data were generated. For the haplotype data 50 diallelic marker loci defining a total of 20 haplotypes were generated. Additionally for each pedigree structure 50 diallelic marker loci were generated. In order to compare the speed of SIMPED with SLINK [7], genotype and haplotype data was generated for three diallelic marker loci, since it is not possible to generate data for 50 marker loci using SLINK. The pedigrees structures for which data was generated (table 1) included a small three-generational pedigree with 16 pedigree members (pedigree 1), a pedigree with 11 consanguinuity loops (pedigree 2) and a pedigree with 2,827 members (pedigree 3). For each analysis 1,000 replicates were generated on a computer with a Xeon 3.0 GHz processor and 4 GB of RAM running under Red Hat (v9.0) Linux operating system.

The SimPed program runs extremely quickly (table 1); for example, for 50 marker loci haplotype type data was generated for the 16-member pedigree 1 in 0.73 seconds and genotype data was generated in 0.75 seconds. To generate genotype data for 50 marker loci for the same pedigree structure it took SIMULATE [8, 9] 1.29 seconds. Consanguineous pedigree 2 took 2.16 seconds for SimPed to generate haplotype data for 1,000 replicates. Due to its size it took considerably longer to generate data for pedigree 3; haplotype data was generated for 1,000 replicates in less than 4 minutes. For the pedigree 1 it took SimPed 0.13 seconds to generate 1,000 replicates for both haplotype and genotype data for the three marker loci. To generate data for three marker loci for pedigree 1 it took SLINK 4.8 and 8.9 seconds to simulate genotype and haplotype data, respectively.

Table 1. Benchmarks for three pedigree structures. For the haplotype data, 20 haplotypes were generated consisting of 50 marker loci and for the genotype data 50 marker loci were generated

Pedigree ID	Number of generations	Number of individuals	Number of founders	Number of consanguinity loops	Haplotype (seconds)	Genotype (seconds)
1	3	16	4	0	0.73	0.75
2	7	52	11	11	2.16	2.31
3	6	2,827	472	0	221.06	231.72

For each benchmark 1,000 replicates were generated.

The SimPed program generates both genotype and haplotype data unconditional on the disease status or quantitative trait. The data generated by the program can be used for a variety of purposes including the evaluation of type I error and the estimation of empirical p values. When estimating empirical p values, the data is generated under the null hypothesis. For a linkage study the null hypothesis is that the disease locus is unlinked (θ = 0.0) to the map of marker loci. For an association study the null hypothesis is that there is no linkage disequilibrium between the disease locus and the map of marker loci. Usually when genome-wide empirical p values are estimated, marker loci are generated with the same allele frequencies and genetic distance as those marker used in the linkage study. Many replicates of the marker data are generated using the same analyses and phenotype definitions that were implemented in the original linkage study. For example, if three diagnostic schemes were used in the analysis and both parametric and non-parametric methods were carried out using GENEHUNTER [16], the GENEHUNTER program would be employed to analyze the simulated data in the exact same way it was implemented for the analysis of the original data set. It is then observed under no linkage what proportion of replicates had a resulting statistic (e.g. LOD score, NPL score) that is equal to or greater than the statistic that was observed in the original linkage study. When estimating small p values a large number of replicates must be analyzed in order to provide an accurate estimate. Although it is possible to generate genotype data for a large number of marker loci for virtually any pedigree structure, it is not possible to simulate haplotype data with existing programs for a large number of marker loci. Currently only MERLIN [10] and ILINK of the FASTLINK/LINKAGE package [17] implements linkage analysis allowing for intermarker linkage disequilibrium; however, only MERLIN can perform analysis for a large number of markers and additionally ILINK can only perform parametric linkage analysis. The MERLIN program can carry out the analysis incorporating intermarker linkage disequilibrium for a variety of analysis methods including parametric and non-parametric linkage analysis, variance components analysis and regression-based linkage analysis. The SimPed program is useful to calculate empirical p values for the resulting statistics from the MERLIN program since haplotype data can be simulated. The capability of being able to model intermarker linkage disequilibrium is also extremely important when evaluating empirical p value when data analysis is performed using programs which do not incorporate intermarker linkage disequilibrium in the analysis (e.g. FASTLINK/LINKAGE [17], GENEHUNTER [16], ALLEGRO [6], SIMWALK2 [18, 19]), since it is well known that missing parental genotypes can increase type I error in the presents of intermarker linkage disequilibrium [20]. An underestimation of empirical p values can occur if marker genotypes are analyzed that have been generated under the false assumption of intermarker equilibrium, thus leading to an underestimation of type I error rates. In addition to evaluating empirical p values for linkage studies, the SimPed program can be used for empirical p values estimation when studies are carried out using family-based association methods.

Availability of Software

The SimPed program is written in C. The source code, complied versions which can be run under Linux, Unix or Windows operating systems, user manual and sample data sets are available at http://www.hgsc.bcm.tmc.edu/genemapping.

Acknowledgement

This work was supported by the National Institute of Health Grant R01-DC03594.

References

1 Murray SS, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG, Rubano T, Kermani BG, Fan JB, Chee MS, Hansen MST: A highly informative SNP linkage panel for human genetic studies. Nat Methods 2004;1:113–117.

2 Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, Yang G, Webster T, Cawley S, Walsh P, Jones KW, Mei R: Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 2004;1:109–111.

3 Kong X, Murphy K, Raj T, He C, White PS, Matise TC: A combined linkage-physical map of the human genome. Am J Hum Genet 2004;75:1143–1148.

4 Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, ChenW, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SP, Jones KW: Large-scale genotyping of complex DNA. Nat Biotechnol 2003;21:1233–1237.

5 Boehnke M: Estimating the power of a proposed linkage study: a practical computer simulation approach. Am J Hum Genet 1986;39:513–527.

6 Gudbjartsson DF, Jonasson K, Frigge ML, Kong A: Allegro, a new computer program for multipoint linkage analysis. Nat Genet 2002;25:12–13.

7 Weeks DE, Ott J, Lathrop GM: A general simulation program for linkage analysis. Am J Hum Genet 1994;47(suppl):A204.

8 Terwilliger JD, Speer M, Ott J: Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 1993;10:217–224.

9 Terwilliger JD, Ott J: Handbook of Human Genetic Linkage. Baltimore, Johns Hopkins University Press, 1994.

10 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101.

11 Bass MP, Martin ER, Hauser ER: Software for Simulation Studies of Complex Traits: SIMLA. Am J Hum Genet 2002;71(suppl):A2341.

12 Broman KW, Murray JC, Sheffield VC, White RL, Weber JL: Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet. 1998;63:861–869.

13 Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K: A high-resolution recombination map of the human genome. Nat Genet 2002;31:241–247.

14 The International HapMap Consortium: The International HapMap Project. Nature 2003;426:789–796.

15 Nievergelt CM, Smith DW, Kohlenberg JB, Schork NJ: Large-scale integration of human genetic and physical maps. Genome Res 2004;14:1199–1205.

16 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: A unified multipoint approach. Am J Hum Genet 1996;51:1347–1363.

17 Cottingham RWJ, Indury RM, Schaffer AA: Faster sequential genetic linkage computations. Am J Hum Genet 1993;53:252–263.

18 Weeks DE, Sobel E, O'Connell JR, Lange K: Computer programs for multilocus haplotyping of general pedigrees. Am J Hum Genet 1995;56:1506–1507.

19 Sobel E, Lange K: Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 1996;58:1323–1337.

20 Huang Q, Shete S, Amos CI: Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. Am J Hum Genet 2004;75:1106–1112.