Pedigree Generation for Analysis of Genetic Linkage and Association

M.P. Bass, E.R. Martin, and E.R. Hauser

Pacific Symposium on Biocomputing 9:93-103(2004)

PEDIGREE GENERATION FOR ANALYSIS OF GENETIC LINKAGE AND ASSOCIATION

M.P. BASS, E.R. MARTIN, E.R. HAUSER

Department of Medicine, Center for Human Genetics, 595 LaSalle St., Box 3445,
Duke University Medical Center, Durham, NC 27710, USA
meredyth, emartin, bhauser@chg.duhs.duke.edu

We have developed a software package, SIMLA (simulation of linkage and association), which can be used to generate pedigree data under user-specified conditions. The number and location of disease loci, disease penetrances, marker locations, and marker disequilibrium with a disease locus and with other markers can be controlled. In addition, the pedigree size and availability of genotype data may also be specified, and a number of rules for family ascertainment are available. Estimates for power and type I errors can be evaluated under a variety of conditions, as needed by the user. We developed this simulation program because there are no publicly available programs to simulate variable levels of both recombination and linkage disequilibrium (LD) in general pedigrees. Genetic researchers are routinely applying both tests of linkage and family-based tests of association in the search for complex disease genes, and a plethora of different statistical approaches are available. Thus there is a need for the flexible statistical simulation program that we describe. This is the only program that we are aware of that allows simulation of linkage and association for multiple markers in extended pedigrees, nuclear families or in sets of unrelated cases and controls. Furthermore, the program not only allows for variable levels of LD among markers but also between markers and disease loci. SIMLA can simulate the complex and variable levels of LD that have been observed at close markers across the genome and allows for realistic simulation of complex relationships between markers. The program will be useful for studying and comparing existing statistical tests, for developing new genetic linkage and association statistics, planning sample sizes for new studies, and interpreting genetic analysis results.

1 Introduction

Genetic analysis is concerned more and more with the search for genes that play a role in very complicated disease pathways. For complex diseases it is the case that a single gene may act in concert with additional genes or environmental exposures. Issues faced by the researcher searching for complex disease genes include locus heterogeneity, low penetrances, phenotypic variation, and the presence of phenocopies, to name a few of the difficulties encountered. With such complex diseases under study, it is important to understand how different genetic analysis statistics will behave under varying conditions.

Further, in designing test statistics to detect disease genes for complex traits, it is crucial to be able to evaluate their performance under controlled situations. This includes estimating type I errors under the null hypothesis, assessing the power under conditions representative of alternative hypotheses, and determining optimal sample size for a given study design. More generally, it is useful to assess the statistical distribution of a statistic under study, for instance, to verify whether a set of observed means or variances is in agreement with expected values.

Simulations are invaluable in comparing results between statistics that are designed to perform similar tests. Simulation provides experimental conditions that allow the user to understand under which conditions one test may differ from another. For instance, one test may be preferable for detecting small genetic effects in homogeneous data sets, and another may be better in the presence of large sibships with multiple affected individuals.

We have developed the software package SIMLA (simulation of linkage and association) with the goal of allowing the researcher a great amount of flexibility in specifying test conditions. The user selects a number of parameters, including the number of replicates, the size of a data set, a map of one or more markers and the location and prevalence of up to ten disease loci. SIMLA is unique among simulation packages in that the researcher may specify varying levels of both linkage and linkage disequilibrium (LD) among markers and between markers and disease loci, thus enabling simultaneous studies of linkage and association in extended pedigrees. By allowing the user to specify the level of LD among markers, this program allows the user to model the complex patterns of LD often observed in real data. Output consists of data sets of pedigrees that conform to the user's selected inclusion criteria. These output files can then be used as input into various genetic analysis packages.

This software has already been used to test power and type I errors for the Ordered Subset Analysis software package, a software package designed to assess linkage in the presence of genetic heterogeneity using covariate information [1]. SIMLA was instrumental in discovering and correcting a bias with the PDT statistic[2, 3]. It has also been used to test the geno-PDT statistic, an extension of the PDT that analyzes transmitted vs. non-transmitted pairs of alleles rather than single alleles[4].

2 System and Availability

SIMLA was implemented using C++ on the Solaris 7 Unix operating system. Please contact us if you require an alternate system for running SIMLA. Downloads with detailed documentation and example input for running SIMLA are available on the Center for Human Genetics web site: http://wwwchg.duhs.duke.edu. Follow the link to CHG Software. Registration is required for future notification regarding program modifications or upgrades. Contact information is not used for any other purpose.

3 Algorithm

The pedigrees created by SIMLA are based on a common structure (Figure 1). The proband is the first individual in Generation III (III-1). In all pedigrees, there are four founders and three sibships. All the sibships, both within a family and across the data set, consist of the same number of individuals, ranging from two to ten siblings. Multiple disease genes can be specified for a data set, with subsets of families linked to each one. However, only one bi-allelic disease locus segregates within any one family. The proband is always affected, while the affection status of all other individuals is determined by their disease genotype and the penetrances specified in the SIMLA input parameter file.

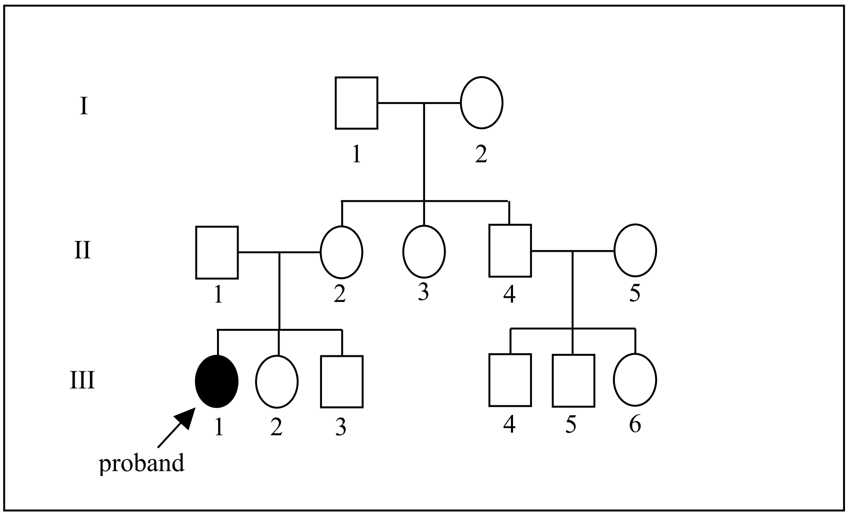

Figure 1. A standard pedigree as created by the SIMLA program. All pedigrees in a data set have the same number of individuals. The individual marked as the proband (III:1) is always affected, and the disease status of all remaining relatives is determined at random, based on disease genotype and penetrances. Sibships may be specified to contain from 2 to 10 people, and there are always three sibships generated for each family.

Assigning disease status for pedigree members is accomplished first by dropping the set of "blank" founder chromosomes through the family, with each parental chromosome equally likely to be transmitted. The proband will have received a segment that originated from a founder (I-1 or I-2) and a segment from his/her married-in parent (II-1). Given that the proband is affected, the disease genotype is assigned in accordance with specified disease allele frequencies and penetrances. All pedigree members with the same founder chromosome as the proband will receive the corresponding disease allele. The remaining six founder chromosomes are assigned a disease allele consistent with given disease allele frequencies. At this point, all of the people in the pedigree have a disease genotype assigned. Affection status is then determined based on user-specified penetrances (Figure 2).

If the resulting pedigree meets user-specified ascertainment criteria (an affected sibpair, for example), then marker genotypes are assigned to all individuals starting with the founders. In the case of no LD, founder chromosomes, which are blank except at the disease locus, are assigned alleles at each marker independently based on frequencies entered in the parameter file. When LD is desired between one or more markers and a disease locus, then frequencies for all possible haplotypes are specified in the parameter file. Two frequencies are given for each haplotype conditional on the presence or absence of the disease allele on the chromosome segment. Founders are assigned alleles for these markers as a set, based on designated haplotype frequencies. Remaining markers are assigned independently based on specified marker allele frequencies.

It is the conditional haplotype frequencies that determine the extent of LD between the markers and the disease locus and among the markers themselves. SIMLA offers a great deal of flexibility to model patterns of LD. Haplotype frequencies based on observed data may be entered. Blocks of LD may be simulated, where the user selects a subset of markers to be in LD with a disease locus while markers not selected are in linkage equilibrium with the disease locus. It is even possible to specify LD only among the markers and none between the markers and the disease locus.

For example, two markers A and B, each with two alleles 1 and 2, would have four possible allele combinations, or haplotypes. For each haplotype, two conditional frequencies must be designated, one in the presence of the disease allele and one in the absence of the disease allele. Table 1 shows possible sets of parameter values leading to three cases. The first case is complete LD between marker A and the disease locus but none between marker B and either locus. In this case, marker B could be left out of the haplotype set in the parameter file, though it is shown here for illustration. The second case shows LD between each marker and the disease locus as well as LD among the markers. The last case demonstrates LD only among the markers and none between the markers and the disease locus. Lewontin's D' statistic[5] is used to quantify the extent and direction of LD seen between each pair of loci for each situation, but any measure of LD could be used.

Once founder genotypes are assigned, chromosomes are passed down from parents to children in Mendelian fashion, allowing for cross-overs along the chromosome segment. Cross-over events occur based on recombination fractions between each locus along the map. Since the disease genotype has been determined for all family members at this point, it can be considered fixed for all family members. Accordingly, individual marker genotype assignments for children move out in either direction from the disease locus, allowing for chance recombination events.

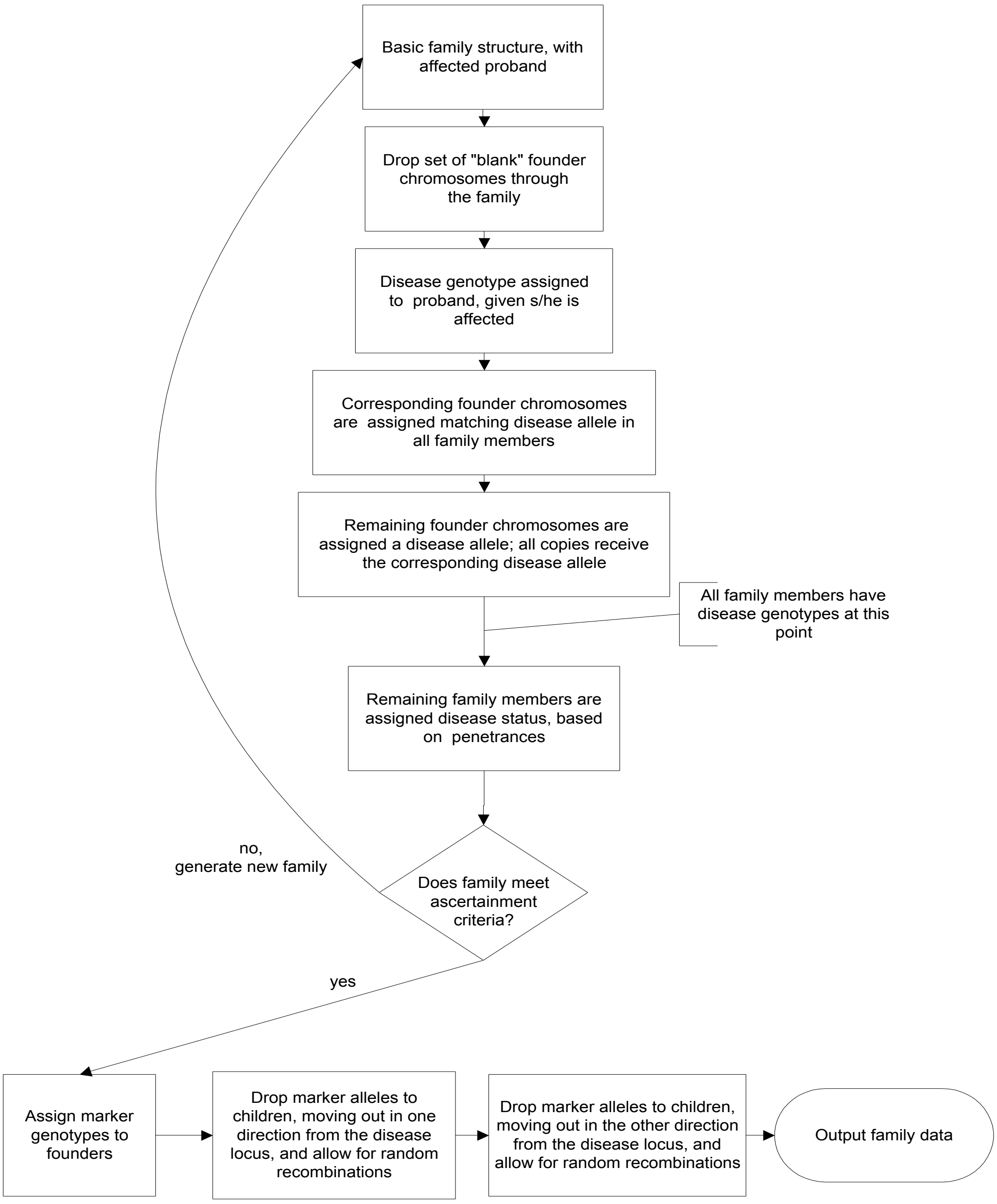
Basic family structure, with affected proband
Drop set of "blank" founder chromosomes through the family
Disease genotype assigned to proband, given s/he is affected
Corresponding founder chromosomes are assigned matching disease allele in all family members
Remaining founder chromosomes are assigned a disease allele; all copies receive the corresponding disease allele
All family members have disease genotypes at this point
Remaining family members are assigned disease status, based on penetrances
Does family meet ascertainment criteria?
no, generate new family
yes
Assign marker genotypes to founders
Drop marker alleles to children, moving out in one direction from the disease locus, and allow for random recombinations
Drop marker alleles to children, moving out in the other direction from the disease locus, and allow for random recombinations
Output family data

Figure 2. Family generation algorithm

Table 1. Example disequilibrium frequencies for two bi-allelic markers A and B showing varying levels of LD with a disease locus and with each other. Lewontin's D' measure of LD is given for allele A1 with the disease allele, B1 with the disease allele, and A1 with B1. N is the normal allele, and Dx is the disease allele for the disease locus.

	Complete LD between A1 and Disease allele		LD between A1 and Disease allele and between markers		Complete LD between A1 and B1; no LD with disease locus	
	N	Dx	N	Dx	N	Dx
A1-B1	0	0.5	0.5	0.5	0.5	0.5
A1-B2	0	0.5	0	0.5	0	0
A2-B1	0.5	0	0.5	0	0	0
A2-B2	0.5	0	0	0	0.5	0.5
D'(A1-Dx)	1		1		0	
D'(B1-Dx)	0		-1		0	
D'(A1-B1)	0		-1		1	

After marker genotypes have been assigned and passed down through the family, pedigree members who have been defined as "missing" have their genotypes cleared. The user indicates which individuals in the pedigree should not have marker data available, and these people are reassigned "0 0" genotypes for all markers in all pedigrees generated.

4 Implementation

The only input required by the SIMLA program is a parameter file. This parameter file contains details regarding data set generation and family size, including the number of replicates, the number of families per data set, the size of the sibships, and the proportion of families linked to one disease locus versus another (Table 2). The researcher may choose to ascertain families with an affected proband, an affected proband sibpair, an affected cousin pair, an affected parent-child pair, an affected proband sibpair and affected cousin sibpair, at least one discordant sibling pair, or a user-specified list of affected or unaffected individuals. The number of disease loci can range from 1 to 10, and the user determines the allele frequencies, penetrances, and locations relative to the markers on the map. The user also specifies the number of markers with corresponding allele frequencies and genetic location. A genotyping error rate may be set for any marker, where an allele may be replaced at random in a specified proportion of allele assignments. As discussed above, LD between markers and a disease locus, or among markers, can be specified with conditional haplotype frequencies. If desired, a separate output file can be created with one or more covariates for each family in a data set. These variables are normally distributed with a standard deviation of 1 and a mean set by the user. Detailed documentation on creating the SIMLA parameter file is included

Table 2. Description of variables used as input for the SIMLA program.

Parameters	**Description**
General variables:	
fams	Number of families per data set
gen_ped	Name of pedigree file
inc_code	Inclusion (ascertainment) code
index_list	List of individuals not genotyped
nrep	Number of replicates (data sets)
num_clear	Number of individuals with no data
pdt_flag	Whether to create PDT datfile
pdt_dat	Name of PDT datfile
sibsize	Number of sibs in a sibship
siblink_flag	Whether to create SIBLINK files
siblink_ped	Name of SIBLINK files
units	Map units (Haldane or Kosambi)
vars	Number of covariates
Disease gene variables:	
ndloc	Number of disease genes
chr	Chromosome
dx_name	Name of the locus
f0, f1, f2	Penetrances (for 0, 1, and 2 disease alleles)
freq	Allele frequencies
mloc	Map location
prop_list	Proportion of families linked to each disease gene
Marker variables:	
ntloc	Total number of markers
alleles	Number of alleles at a marker
chr	Chromosome
err	Amount of genotyping error
freq	Allele frequencies
mkr_flag	Whether a marker is in LD
mloc	Map location
name	Marker name
ord	Relative order in the map

with the download of the package from our web site http://wwwchg.duhs.duke.edu, as is an example parameter file. SIMLA uses this parameter file to create a post-MAKEPED LINKAGE[6] style pedigree file for each replicate generated. Alternatively, sets of unrelated cases and controls could be sampled from the simulated data.

There are also flags available in the parameter file that indicate whether to run SIBLINK, which performs non-parametric linkage analysis of affected sibpairs, and PDT, which performs a valid test of linkage and association in extended pedigrees. Either of these programs could be used to perform genetic analyses of real or simulated data, and both are freely available on the Center for Human Genetics web site.

SIMLA is able to create data set replicates in a reasonable amount of time. Table 3 gives a sample listing of simulation times to create data set replicates under varying situations. To give an idea of the simulation complexity, the number of replicates, the number of disease genes, disease allele frequencies, penetrances, number of covariates and data set sizes are listed. All simulations ascertained families with exactly one affected proband sibship and all were run using a Solaris 8 workstation.

We have demonstrated the use of SIMLA for a complex trait by assessing the correlation between association and linkage statistics [7]. We have described the bias of an existing association statistic in a late onset disease due to the lack of parental genotypes and determined that a new statistic correctly handles this situation [8]. We have used SIMLA to describe power and type I error for a novel linkage statistic for analyzing complex traits in conjunction with covariate information [9] and for an extension of an existing association statistic [4]. Thus SIMLA allows for assessment of test statistics in complicated study designs as well as in identifying powerful follow-up studies.

Table 3. Sample listing of computation times for SIMLA on a Solaris 8 workstation.

Reps	Families per data set	Disease Allele Frequency	Prevalence	Ascertainment Criterion	Sibship size	Time to complete (hr:min)
500	400	0.06,0.06	0.048	Affected sibpair	2	3:49
1000	500	0.25	0.0059	Exactly 1 Affected sibpair	3	0:49
2000	250	0.15	0.0046	Affected sibpair	3	1:56

5 Discussion

Our goal in creating SIMLA was to develop a tool that could be used to answer a variety of questions of interest to those developing methods for genetic analysis and conducting studies of complex disorders. SIMLA was designed to aid in assessing the effectiveness of both linkage and association statistics in family data. Accordingly, SIMLA can be used to determine optimal study design, sample size, and power under user-defined conditions. Unlike many other simulation packages available, such as SIMLINK [10, 11], SLINK [12, 13], and SIMULATE , SIMLA does not require a pre-existing data set on which to perform calculations. This program is unique because it allows the researcher to simulate complicated patterns of LD as well as simple linkage. The POWERFBAT program [14] will also generate LD in nuclear families, but SIMLA extends this capability to larger pedigrees. Its strength lies in the flexibility afforded the user to simulate conditions common in the search for causes of complex disorders, such as missing data, locus heterogeneity, and phenocopies.

SIMLA allows the researcher to understand how results might appear under various conditions. For instance, how might a multipoint LOD score curve appear when there are subsets of families linked to two disease loci on the same chromosome? Or, how much power is lost when parental data are not available? The answers to these and other questions can be critical in planning ascertainment or deciding which regions merit follow-up, given preliminary results.

Though we have designed SIMLA to be highly flexible for the user, we are planning a number of enhancements to further this flexibility. At this time, SIMLA considers one chromosome or chromosomal region at a time. We are extending SIMLA to simulate the entire genome. While it is possible with the current version of SIMLA to simulate markers as if they were unlinked from other markers by inserting very large genetic distances between them, our goal is to streamline this for the user.

Another limitation is that while we consider genetic heterogeneity models, only one susceptibility gene segregates through any one family. Our goal in future versions is to enable multiple genes to act through a single pedigree. Thus, we could incorporate epistatic models into the parameter file. It would also be of interest to simulate disease genes that exhibit parent-of-origin effects, as seen with imprinting, and to emulate quantitative as well as qualitative disease traits.

SIMLA will permit study of complex genetic analysis problems. It can provide insights into issues of power and sample size, as well as aid in interpretation of observed results. By adding complexity to simulation models, we anticipate that SIMLA will provide greater understanding of the linkage and association statistics that are available, and the relationship between linkage and association statistics in a wide variety of study designs for detection and localization of complex genetic traits.

Acknowledgments

We are grateful for the generous support provided by the National Institutes of Health (Grants RO1 MH59528 and RO1 AG20135), the Neurosciences Education and Research Foundation, and the Morris K. Udall Center of Research Excellence (Grant PO1 NS39764-02).

References

1. E.R.Hauser, M.P.Bass, E.R.Martin, R.M.Watanabe, W.L.Duren, and M.Boehnke, "Power of the ordered subset method for detection and localization of genes in linkage analysis of complex traits" *Am. J. Hum. Genet.* 69, 529 (2001)
2. E.R.Martin, S.A.Monks, L.L.Warren, and N.L.Kaplan, "A test for linkage and association in general pedigrees: the pedigree disequilibrium test" *Am. J. Hum. Genet.* 67, 146 (2000)
3. E.R.Martin, M.P.Bass, and N.L.Kaplan, "Correcting for a potential bias in the pedigree disequilibrium test" *Am. J. Hum. Gen.* 68, 1065 (2001)
4. E.R.Martin, M.P.Bass, and E.R.Hauser, "A genotype-based association test for general pedigrees: The geno-PDT" *Am. J. Hum. Genet.* 71, 2365A (2002)
5. R.C.Lewontin, "The interaction of selection and linkage. I. General considerations; heterotic models" *Genetics* 49, 49 (1964)
6. G.M.Lathrop, J.M.Lalouel, C.Julier, and J.Ott, "Strategies for multilocus linkage analysis in humans" *Proc. Natl. Acad. Sci. U. S. A* 81, 3443 (1984)
7. E.R.Martin, M.P.Bass, and E.R.Hauser, "Correlation between linkage and association tests in families." *Am. J. Hum. Genet.* 69, 511 (2001)
8. E.R.Martin, M.P.Bass, E.R.Hauser, and N.L.Kaplan. "Accounting for linkage in family-based tests of association with missing parental genotypes" Am J Hum Genet (2003)
9. E.R.Hauser, R.M.Watanabe, W.L.Duren, M.P.Bass, C.D.Langefeld, and M.Boehnke. "Ordered subset analysis in genetic linkage mapping of complex traits" Genetic Epidemiology (submitted) (2003)
10. M.Boehnke, "Estimating the power of a proposed linkage study: a practical computer simulation approach" *Am. J. Hum. Genet.* 39, 513 (1986)
11. L.M.Ploughman and M.Boehnke, "Estimating the power of a proposed linkage study for a complex genetic trait" *Am. J. Hum. Genet.* 44, 543 (1989)

12. J.Ott, "Computer simulation methods in human linkage analysis" *Proceedings of the National Academy of Science, USA* 86, 4175 (1989)

13. D.E.Weeks, J.Ott, and G.M.Lathrop, "SLINK: A general simulation program for linkage analysis" *Am. J. Hum. Genet.* 47, A204 (1990)

14. N.M.Laird, S.Horvath, and X.Xu, "Implementing a unified approach to family-based tests of association" *Genet. Epidemiol.* 19 Suppl 1, S36 (2000)