Jurg Ott / 8 August 2011
Rockefeller University New York
Email: ott@rockefeller.edu
http://linkage.rockefeller.edu/ott/

Documentation to SIMULATE program

Originally written by Joe Terwilliger [1], SIMULATE is a computer program to simulate genotypes in family members for a map of linked markers unlinked to a given affection status locus. Output from this program is in SLINK format and is ready for analysis with UNKNOWN, ISIM, LSIM, or MSIM of the SLINK package. The program is written in Free Pascal for Windows (32 bit and 64 bit) and Linux and comes in the following three versions:

(1) SIMULATE is essentially the original version, in which all marker genotypes are generated based on population marker characteristics (allele frequencies, etc.) and recombination fractions between them.

(2) SIMULATE2 assumes that for founder individuals with known genotypes (i.e., they are "typed") the original (observed) genotypes are provided. These genotypes will not be modified (generated) in the course of the simulations. For founders with unknown genotypes ("untyped"), marker genotypes will be generated as in the *SIMULATE* program. However, prior to running *SIMULATE2*, users may want to impute such genotypes from the original data so that few founders have unknown genotypes. The two program versions have somewhat different input requirements. For example, the random number generator is different -- *SIMULATE* requires 3 seeds and *SIMULATE2* requires only one.

(3) SIMULATE3 is an extension of SIMULATE in that markers must be SNPs (have only two alleles) and may be correlated (in linkage disequilibrium) with each other (see below how these correlations are generated). Also, this program version incorporates an updated random number generator.

LOCUS TYPES

All locus types from the LINKAGE programs can be simulated with the following exceptions:

- At affection status loci to be simulated, only one liability class is permitted (at affection status loci, which are not to be simulated, any number of liability classes is permitted).
- At quantitative trait loci, only one trait per locus is permitted.
- If the first locus is an affection status type of locus (common situation), the program will expect an affection status phenotype in the pedigree file and simulate the marker map independent of this affection status locus. However, if the first locus is not affection status, then it will be simulated as well and the program will just simulate a linked map of markers with no disease present.

The *SIMULATE2* program can only handle marker loci and optionally a trait locus as the first locus.

INPUT FILES

This program requires three input files as follows (exceptions for SIMULATE3 are given in a separate section below):

1. **SIMDATA.DAT** a Standard LINKAGE parameter file (datafile), specifying the map of markers (chromosome order = file order). Note that, as in SLINK, the last input line is not relevant. However, the input line before last specifies recombination fractions between loci.

2. **SIMPED.DAT** a post-MAKEPED LINKAGE pedigree file with an additional line at the top specifying the following numbers:

- Number of pedigrees
- Number of individuals in pedigree 1, 2, etc. Terminate this input line when you do not want to use the option described in the next bullet. There must be no trailing characters after this list of numbers except when the optional numbers below are given.The number of individuals specified must be identical with the number of input lines per pedigree in the LINKAGE pedigree file, that is, a duplicated individual in a loop is counted as two individuals. In the pedigree file, the individuals must be listed sequentially (in the order of increasing ID number). When a loop is broken by the MAKEPED program, individuals are not usually in sequential order, but they must be brought into sequential order before the pedigree file is acceptable to the SIMULATE program.
- (optional, only for SIMULATE, not for SIMULATE2) A sequence of 0's and 1's, one such number for each locus, where a 1 indicates that all founder individuals will be assigned genotypes with consecutive allele numbers, 1/2, 3/4, 5/6, etc. (thus making all founders heterozygous), and a 0 indicates that founders' genotypes will be simulated as usual. Warning: if this option is chosen, sequential numbers will be assigned to all founders at a given locus, irrespective of the actual numbers of alleles at this locus. Thus, it is the responsibility of the user to ensure that no allele numbers will be assigned that exceed the number of alleles at a given locus. Do not use this option unless you have a specific reason for doing so.

For subsequent input lines (one line per individual), different rules apply to the SIMULATE and SIMULATE2 programs:

SIMULATE. The fields for id's, sex, and probands are as in standard LINKAGE pedigree files. However, since this program simulates all marker loci, you must only have one digit per marker in the pedigree file (no marker genotypes, or `0 0`, as in SLINK; see exception for SIMULATE2 below) to tell whether that marker is to be simulated or left unknown in that individual. A 0 means that marker should be untyped, and a 1 means it should be typed (simulated). See below for locus order and types of loci. That is, each marker may be designated as being known or unknown, not only each individual as in SLINK.

SIMULATE2. Depending on whether an individual is a founder (no parents in pedigree) or non-founder, and whether genotypes are known or not, marker genotypes are coded as follows (after an initial optional affection status locus, only marker genotypes are permitted). Note that the coding scheme below is analogous to that in SLINK but different from that in SIMULATE.

- Founder, known (observed) genotypes: The actual genotypes (two alleles) must be provided. Missing genotypes at some markers are indicated by `0 0`).
- Founder, unknown genotypes: Enter `0 0` for each unknown genotype. Alternatively (when all genotypes are unknown), the first (and only) genotype code may be -1. Any numbers following an initial -1 will be ignored.
- Non-founder, known (simulated) genotypes: A code of `1 1` for each known genotype. Alternatively (when all genotypes are known), the first (and only) genotype code may be -2.
- Non-founder, unknown genotypes: A code of `0 0` for each unknown genotype. Alternatively (when all genotypes are unknown), the first (and only) genotype code may be -1.

3. **PROBLEM.DAT** A file containing the following numbers:

- 3 integer numbers between 1 and 30323 as seeds for the random number generator
- The desired number of replicates of your pedigree set. There must be no trailing characters after this number except when the optional item mentioned below is furnished.
- (optional) The number of runs carried out with SIMULATE or SIMULATE2 in a batch application (see Batch Runs below).

At the end of the simulation, the program writes a new seed to PROBLEM.DAT such that SIMULATE/2 may be called repeatedly and each time continue with an updated seed.

OUTPUT FILES

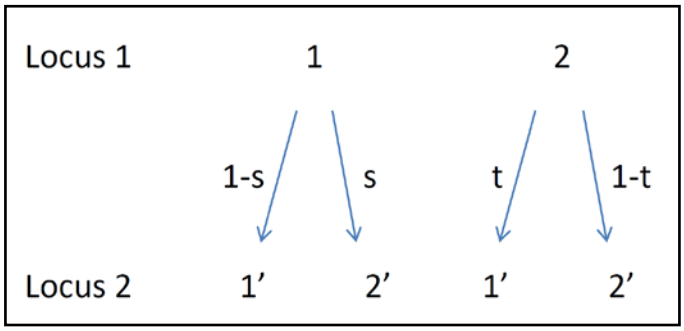
The program creates the following output files:

- PEDFILE.DAT for analysis by the companion programs to SLINK
- SIMOUT.DAT summarizes the parameters used
- PROBLEM.DAT is rewritten and contains an updated seed and, optionally, the number of runs incremented by 1 (see Batch Runs below).

Markov model for correlated SNPs

A simple model for correlated SNPs is to assume that SNP (i – 1) only depends on SNP i. While such a dependency may not be completely realistic, it can nonetheless serve a useful role in bioinformatics investigations on genomes [2]. Generating such SNP genotypes for each individual may be accomplished as follows.

Consider two adjacent SNPs, locus 1 (alleles *1* and *2*) and locus 2 (alleles *1'* and *2'*). Let p = P(*1*) be the minor allele frequency at locus 1, and q = P(*1'*) be the minor allele frequency at locus 2, and assume that the two alleles in a genotype are in a fixed order; for now we only consider the first allele in a given genotype. We want to make alleles at locus 2 depend on those at locus 1 in the manner indicated on the left. Thus, for example, if an individual has allele *1* at locus 1, then his probability of having allele *1'* at locus 2 is



$$P(1'|1) = 1 - s.$$

This model has three parameters: p, s, and t, with allele frequencies at locus 2 being given by

$$q = p(1 - s) + (1 - p)t.$$

The joint distribution of alleles at two loci may be constructed based on the conditional probabilities introduced above, for example, P(1', 1) = P(1'|1)P(1) = $(1 - s)p$. This leads to the following table:

Locus 2	Locus 1		
	1	*2*	total
1'	$(1 - s)p$	$t(1 - p)$	q
2'	sp	$(1 - t)(1 - p)$	$1 - q$
total	p	$1 - p$	1

For given p and q, there is one free parameter (degree of freedom, df) left, which we conveniently choose to be a measure of linkage disequilibrium between the SNPs. Thus, we want to express the square of the correlation coefficient, r^2, and the scaled disequilibrium parameter, D', in terms of p, q, s, and t. To obtain r^2, we attach respective values of x = 1 and x = 0 to alleles *1* and *2*, and values of y = 1 and y = 0 to alleles *1'* and *2'*. The correlation coefficient is defined as

$$r = [E(xy) - E(x)E(y)]/\sqrt{[V(x)V(y)]},$$

with E denoting expectation and V denoting variance, for example,

$$V(x) = E(x^2) - E^2(x) = p - p^2 = p(1 - p).$$

With this, we obtain the squared correlation coefficient in terms of the parameters of table 1 above as

$$r^2 = [P(1, 1') - pq]^2/[p(1 - p)q(1 - q)] = p(1 - s - q)^2/[(1 - p)q(1 - q)].$$

Similarly, assuming a positive correlation between the two SNPs, we find the scaled linkage disequilibrium as

$$D' = [P(1, 1') - pq]/\min[p(1-q), (1-p)q] = p(1-s-q)/\min[p(1-q), (1-p)q].$$

When allele frequencies at loci 1 and 2 are the same, $q = p$, then the above formulas simplify to

$$r = 1 - s/(1-p) \text{ and } D' = 1 - s/(1-p) = r.$$

To generate (predict) an individual's genotype at locus 2 given his genotype at locus 1, as noted above, we treat the two alleles at a genotype as independent entities, that is, we assume Hardy-Weinberg equilibrium (HWE) and generate each allele separately. For example, for the first allele in genotypes at loci 1 and 2, allele *1'* (locus 2) will be obtained with probability $P(1'|1) = 1 - s$ if the first allele at locus 1 is *1*, and with probability $P(1'|2) = t$ if the first allele at locus 1 is *2*. Thus, we need to express s and t in terms of p, q, and r (only r^2 and not D' will be used here to express linkage disequilibrium). Simple algebraic manipulations lead to

$$s = 1 - q - \frac{\sqrt{r^2 p(1-p)q(1-q)}}{p} \qquad \text{and}$$

$$t = q - \frac{\sqrt{r^2 p(1-p)q(1-q)}}{1-p}.$$

For $q = p$, these expressions simplify to

$$s = (1-p)(1-r) \text{ and } t = p(1-r).$$

For a given individual, it is convenient to successively generate alleles at all loci to form a very long haplotype along a chromosome, and then to generate a second haplotype at all loci. The two haplotypes will then define the genotypes of this individual. This procedure has been implemented in the *Simulate* program as *Simulate3*.

Input files for SIMULATE3

File **simdata.dat** as described above.

File **simped.dat** as described above except that the top line must contain (1) the number of pedigrees, followed by (2) the numbers of individuals in each pedigree, and (3) the square of the correlation coefficient, r^2, as a measure of linkage disequilibrium between adjacent SNPs. If (3) is missing, $r^2 = 0$ is assumed. Note that, as above, this input file is post-Makeped, and individuals must be sequentially ordered by their IDs.

File **problem.dat** holding only one number, that is, the number of replicates desired.

File **seed.txt** holding a positive integer seed.

An example for each of these files is contained in the distribution package.

REFERENCES

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

2 Waterman MS: Introduction to computational biology: Maps, sequences and genomes, ed 1. Boca Raton FL, Chapman and Hall/CRC, 1995.