Program Manual for SimPed

Introduction

SimPed is a simulation program that quickly generates haplotype and/or genotype data unconditional on the disease/quantitative trait status for pedigrees of virtually any size and complexity. The program can generate data for a large number of either diallelic or multiallelic marker loci (> 20,000 loci). Haplotypes may be generated for pedigree structures employing user-specified haplotypes and their haplotype frequencies. Additionally, genotype data may be generated under the assumption of Hardy-Weinberg equilibrium by means of user-provided allele frequencies. A third option is to generate a mixture of haplotypes and genotypes. The intermarker recombination fractions and map distances are also user-specified.

SimPed is extremely flexible. The user can enter all intermarker recombination fractions/genetic map distances, the intermarker recombination fractions/genetic map distances may be all set equal, or a repeating pattern of recombination fractions/genetic maps distances may be used. Likewise, when genotype data is to be generated, allele frequencies may be provided for every marker loci, the same allele frequencies may be used for all marker loci, or a pattern of allele frequencies may be used for the marker loci. For haplotypes, it is possible to enter all groups of markers in linkage disequilibrium and their haplotype frequencies, or groups of markers in linkage disequilibrium may be entered; these "blocks" may then be repeated with the "blocks" in linkage equilibrium. Additionally, for a series of haplotypes, genotypes, or a mixture of haplotypes and genotypes a pattern may be repeated. For example, if the user wishes to generate 1,000 marker loci which alternate between three markers in linkage equilibrium and seven markers in linkage disequilibrium, the user can simply enter the allele frequencies for the three markers and the haplotypes and their frequencies for the seven marker loci in linkage disequilibrium and then have the SimPed program repeat the generation of this set of 10 marker loci 100 times in order to produce genotypes for 1,000 marker loci.

Input Files

SimPed accepts pedigree data in standard linkage format with or without a disease/quantitative trait locus. Additional columns in the pedigree file denote whether or not genotype data should be generated for a particular individual or marker locus. A second parameter file must be provided which gives information on the number of replicates to be generated, genetic map distances, haplotypes and haplotype frequencies and/or markers and their allele frequencies.

Creating the Pedigree File

The pedigree file is in standard LINKAGE format, contains information on the pedigree structure and on which individuals/markers genotypes should be generated. Pedigree and individual IDs may consist of number(s), letter(s) or a combination of both. If an individual is a founder, then the ID for both mother and father must be "0". The name of the pedigree file is provided by the user.

Format for the pedigree file

- o Column 1: pedigree identifier
- o Column 2: individual's ID
- o Column 3: the ID of the individual's father
- o Column 4: the ID of the individual's mother
- o Column 5: sex (1=male, 2=female)
- o Column 6-N: affection status, quantitative traits (these columns may also be omitted)
- o Column N+: a sequence of 0's and 1's to identify whether a given marker should be generated or not. A "0" indicates that no genotype data should be generated. If all marker genotypes for a particular individual will either be generated or made unknown, then a single "1" or "0" may be used to denote this. If a subset of markers is to be made unknown for an individual, then a string of "0's" and "1's" must be used for that individual.

Creating the Parameter File

A second file, the parameter file, is necessary to perform the analysis. The name of the parameter file is user-specified. The input for the parameter file is as follows:

- o Line 1: The name of the pedigree file and the output file

- o Line 2: Three seeds for the random number generator each <30,000

- o Line 3: The number of columns which contain information on the trait in the pedigree file. This is the number of columns between the sex column (column 5) and column(s) for the marker genotypes

- o Line 4: the number of replicates desired

- o Line 5 the total number of marker loci, followed by the number of times that a pattern of haplotype, genotype, or a combination of haplotype and genotype data should be repeated. Please note that, when specifying either haplotype frequencies or marker allele frequencies, it is also possible to repeat a pattern for a set group of marker loci. For example, if 16 diallelic markers are generated in linkage equilibrium where the minor allele frequencies for the markers are 0.1, 0.4, 0.1, 0.4, …., 0.1, 0.4 it is not necessary to enter all of the allele frequencies. The marker allele frequency for the first two markers may be given starting on line 8, and it may be denoted on line 5 that 16 markers should be generated and the pattern should be repeated 8 times. Alternatively, the user may specify on line 5 that 16 markers will be generated and the pattern will be generated once, and on line 8 it may be indicated that the allele frequencies for two marker loci are specified and the pattern should be repeated 8 times. Likewise, for haplotypes, if the same set of haplotypes are to be repeated it may either be denoted on line 5 or on line 8. A detailed description is given on pages 3-5 and in the sample exercises. If a series of haplotypes and genotypes should be repeated then this must be specified on line 5 – for example, four marker loci that are in linkage disequilibrium that are followed by a two

marker loci in linkage equilibrium, i.e. haplotype data followed by genotype data. It should be noted that the total number of marker loci divided by the number of times the pattern should be repeated must be a whole number. For example, the user provides eight haplotype frequencies for three markers and then the genotype frequencies for two marker loci, and then wants to generate 100 marker loci; the user would insert a "20" in the parameter file since the pattern for these five marker loci will be repeated 20 times and haplotype and genotype data will be generated for 100 marker loci.

o Line 6: The user specifies whether recombination fractions or map distance will be provided: "1" for recombination fractions, "2" Kosambi genetic map distances in Morgans or "3" Haldane genetic map distances in Morgans.

o Line 7: the first number indicates the number of intermarker recombination fractions/genetic map distances that will be provided by the user. If a "1" is entered and then one recombination fraction/genetic map distance is provided, this will be the intermarker recombination fraction/genetic map distance used between all marker loci. Another possibility is for the user to provide all of the intermarker recombination fractions/genetic map distances. Under this circumstance the user would enter as the first number on line 7 the number of intermarker recombination fractions/genetic map distances, which is equivalent to the number of marker loci-1. For example, if seven marker loci are being generated, if the user provides all intermarker recombination fractions/genetic map distances then the user would provide a "6" as the first number on line 7 and then specify 6 intermarker recombination fractions; i.e., line 7 would read "6 .01 0.03 0.05 0.05 0.02 0.01". A third option is that a number which is >1 and < (total number of marker loci – 1) may be provided. In this case the intermarker recombination fractions/genetic map distance will be repeatedly used by the program. For example if "7" is placed on line 5 for the total number of marker loci, and "2 0.01 0.05" is placed on line 7, the program will use as the intermarker recombination fractions/genetic maps distances 0.01 0.05 0.01 0.05 0.01 0.05.

♣For the following lines it is possible to provide information for either haplotypes or genotypes. Either all haplotype or all genotype data may be provided, or data can alternate between genotype and haploytpe data.

o Line 8:"1" for haplotypes or "2" for genotypes is provided.

If a "1" is provided, it should be followed by the number of haplotypes, the number of marker loci, and the number of times the pattern of haplotypes "blocks" should be repeated. By repeating patterns of haplotypes "blocks" it is possible to generate "blocks" of marker loci with the same haplotype frequencies one after the other. It should be noted that the markers within the blocks are in linkage disequilibrium with each other but the blocks themselves are in linkage equilibrium.

If a "2" is provided then genotype data will be generated for the marker loci. The "2" on line 8 is followed by the number of marker loci for which allele frequencies are provided and the number of times this pattern of marker allele frequencies should be repeated

For example, if haplotypes are being generated for seven marker loci which consist of 10 haplotypes, and this pattern is only used once, then on line 8 would appear "1 10 7 1". However, if instead genotypes are to be generated for seven marker loci then on line 8 "2 7 1" is inserted.

♣If on line 8 a "1" is provided, haplotype will be entered as follows:

- o Line 9: the frequency for each haplotype
  0.05 0.4 0.05 0.05 0.1 0.05 0.05 0.15 0.05 0.05   << Haplotype frequencies

- o Line 10+: The alleles for each haplotype on a separate line. For SNP markers there will be only two alleles for each marker loci; however, in the case of multiallelic marker loci, the allele numbers can be 1…n.
  For the example for seven marker loci forming 10 haplotypes, the lines 10-19 are as follows:
  1 1 1 1 1 1 1 1 1 1 << Haplotypes
  1 1 2 1 1 1 1 1 1 1
  2 2 2 1 1 1 1 1 1 1
  2 2 2 1 1 1 1 2 2 2
  2 2 2 1 1 1 1 1 1 2
  2 2 1 1 1 1 1 1 1 2
  2 1 1 1 1 1 1 1 1 2
  2 2 2 2 1 2 2 2 2 2
  1 2 1 2 1 1 1 1 1 1
  1 2 2 2 2 1 1 1 1 2

♣ If on line 8 a "2" is specified, then genotypes will be entered as follows:

- o Line 9+: on each line is provided the number of alleles and n-1 allele frequencies for each marker locus. It should be noted that, based on the provided allele frequencies for each locus, genotypes will be generated for the pedigree data under the assumption of Hardy-Weinberg equilibrium.
  For example, when genotypes for seven marker loci are to be generated the data on lines 9-15 are as follows:
  2 0.8 <<Number of alleles, allele frequency
  2 0.3 << Number of alleles, allele frequency
  2 0.4 << Number of alleles, allele frequency
  5 0.2 0.2 0.2 0.2 <<Number of alleles, allele frequency
  3 0.7 0.1 <<Number of alleles, allele frequency
  2 0.1 <<Number of alleles, allele frequency
  4 0.25 0.1 0.3 <<Number of alleles, allele frequency

From line 8 onward it is possible to provide information on multiple haplotypes, genotypes or a combination of both. For example, if the user wishes to generate three marker loci that form six haplotypes, and then genotype data for six markers followed by two sets of haplotype data for two markers in linkage disequilibrium, the data could be entered in the parameter file as follows from line 8+:

    1 6 3 1 <<Haplotypes, 6 haplotypes, for 3 marker loci, pattern repeated once
    0.4 0.1 0.1 0.1 0.1 0.2 <<Haplotype frequencies
    2 2 2 <<Haplotypes
    2 2 1
    2 1 1
    2 1 2
    1 1 2
    1 1 1
    2 6 1 <<Genotypes to be generated, six marker loci, pattern repeated once
    2 0.1 << # of alleles, allele frequencies
    2 0.5 << # of alleles, allele frequencies
    2 0.1 << # of alleles, allele frequencies
    2 0.5 << # of alleles, allele frequencies
    2 0.1 << # of alleles, allele frequencies
    2 0.5 << # of alleles, allele frequencies
    1 4 2 1 << Haplotypes, four haplotypes, for two marker loci, pattern repeated once
    0.1 0.25 0.25 0.4 << haplotype frequencies
    1 1 << Haplotypes
    1 2
    2 1
    2 2
    1 4 2 1 << Haplotypes, four haplotypes, for two marker loci, pattern repeated once
    0.1 0.25 0.25 0.4 << haplotype frequencies
    1 1 << Haplotypes
    1 2
    2 1
    2 2

However, it can be seen that some of the genotype and the haplotype data is repeated; alternatively, the data for the 13 marker loci could be entered as follows:
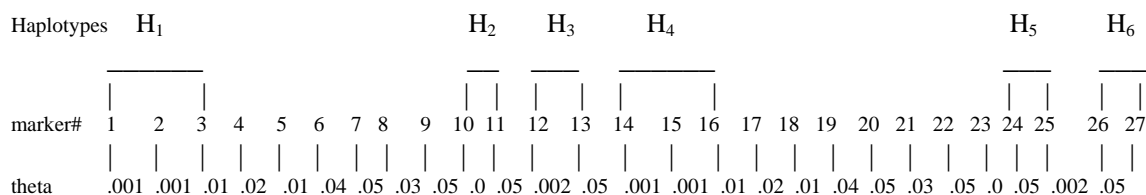
```
1 6 3 1  << haplotypes, six haplotypes, for three marker loci, pattern repeated once
0.4 0.1 0.1 0.1 0.1 0.2 << haplotype frequencies
2 2 2 << haplotypes
2 2 1
2 1 1
2 1 2
1 1 2
1 1 1
2 2 3 < genotypes, two marker loci, pattern repeated three times
2 0.1 << number of alleles, allele frequency
2 0.5 << number of alleles, allele frequency
1  4  2 2 << haplotypes, four haplotypes, two marker loci, pattern repeated twice
0.1 0.25 0.25 0.4
1 1
1 2
2 1
2 2
```

If this pattern of haplotypes for three markers, genotypes for six markers, haplotypes for two markers and haplotypes for two markers needs to be repeated to generate genotype data for 1,300 marker loci, the parameter file would be written as below. Note that the pattern is also repeated for the intermarker recombination fractions for the 13 marker loci; although 13 markers have only 12 intermarker recombination fractions, one additional recombination fraction is needed between the last marker that forms the pattern and the first marker when the pattern starts again. Figure 1 provides a diagramic overview of the map of markers that will be simulated. Due to space constraints only the first 27 markers of the 1,300 markers that will be generated are shown. Figure 1 displays each marker numbered 1 to 27 with the intermarker recombination fractions (labeled theta). For each pattern there are three haplotypes that will be generated. The first haplotype to be generated, H_1 consists of three marker loci. For H_1 six possible haplotypes will be generated with the frequencies that are provided in the below parameter file. Markers 4-9 are in linkage equilibrium. These markers are followed by two more haplotype H_2 and H_3. Each of these haplotypes consists of two marker loci. Both H_2 and H_3 have the same four haplotype frequencies which are provided in the below parameter file. The pattern is then repeated starting with marker 14.

Figure 1

6

pedin.dat pedfile.out << input file, output file

      23333 19999 27843 << three seeds for the random number generator

      0 << number of columns for the trait locus

      100,000 << number of replicates

      1300 100 << 1300 marker loci, pattern repeated 100 times

      1 << recombination fractions are used

      13 0.001 0.001 0.01 0.02 0.01 0.04 0.05 0.03 0.05 0.0 0.05 0.002 0.05

      1 6 3 1   << haplotypes, six haplotypes, for three marker loci, pattern repeated once

      0.4 0.1 0.1 0.1 0.1 0.2 << haplotype frequencies

      2 2 2 << haplotypes

      2 2 1

      2 1 1

      2 1 2

      1 1 2

      1 1 1

      2 2 3 < <genotypes, two marker loci, pattern repeated three times

      2 0.1 << number of alleles, allele frequency

      2 0.5 << number of alleles, allele frequency

      1 4 2 2 << haplotypes, four haplotypes, two marker loci, pattern repeated twice

      0.1 0.25 0.25 0.4

      1 1

      1 2

      2 1

      2 2

Below is another example where patterns of haplotypes and genotypes are used. In the example above, one pattern of haplotypes was followed by a pattern of genotypes. However, it is possible to repeat a pattern of genotypes followed by another pattern of genotypes and likewise repeat a pattern of haplotypes followed by another pattern of haplotypes. For the example following 1,300 marker loci will be generated; however, patterns for 20 marker loci will be specified and these marker loci will be in both linkage equilibrium and disequilibrium.

pedin.dat pedfile.out << input file, output file
27766 27969 29812 << three seeds for the random number generator
0 << number of columns for the trait locus
100,000 << number of replicates
1300 65 << 1300 marker loci, pattern repeated 65 times
1 << recombination fractions are used
20 0.0 0.0 0.1 0.002 0.002 0.05 0.0 0.07 0.001 0.06 0.0 0.1 0.1 0.1 0.02 0.1 0.04 0.05 0.03 0.05
1 6 3 2 << haplotypes, six haplotypes, for three marker loci, pattern repeated two times
0.4 0.1 0.1 0.1 0.1 0.2 << haplotype frequencies
2 2 2 << haplotypes
2 2 1
2 1 1
2 1 2
1 1 2
1 1 1
1 4 2 3 << haplotypes, four haplotypes, for two marker loci, pattern repeated three times
0.7 0.1 0.1 0.1 << haplotype frequencies
2 2 << haplotypes
2 1
1 2
1 1
2 2 2 < <genotypes, two marker loci, pattern repeated two times
2 0.1 << number of alleles, allele frequency
2 0.5 << number of alleles, allele frequency
2 2 2 < <genotypes, two marker loci, pattern repeated two times
2 0.2 << number of alleles, allele frequency
2 0.7 << number of alleles, allele frequency

All examples are given using pedigree 1

Pedigree 1: Nuclear pedigree with body mass index (BMI) for individuals 1-6.
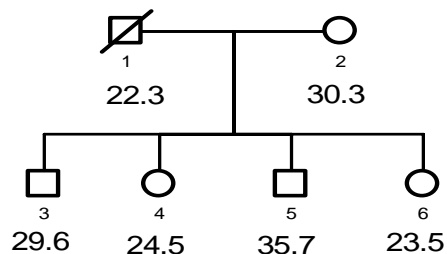
Pedigree file - *pedin1.pre* for example 1: Haplotypes consisting of 10 marker loci will be generated for pedigree 1. Information on the quantitative trait BMI will not be included in the pedigree file. Since individual 1 is deceased, no haplotype will be generated for this individual. For all other individuals in this pedigree, haplotype data will be generated for each of the 10 marker loci.

```
1   1   0   0   1   0
1   2   0   0   2   1
1   3   1   2   1   1
1   4   1   2   2   1
1   5   1   2   1   1
1   6   1   2   2   1
```

Pedigree file - *pedin2.pre* for example 2: The haplotypes to be generated for pedigree 1 will consist of 10 marker loci. However, this time the quantitative trait BMI will be included. In addition, for individual 4 genotypes for markers 5 and 8 will not be generated, and for individual 6 markers 3, 9 and 10 will not be generated.

```
1   1   0   0   1   22.3   0
1   2   0   0   2   30.3   1
1   3   1   2   1   29.6   1
1   4   1   2   2   24.5   1 1 1 1 0 1 1 0 1 1
1   5   1   2   1   35.7   1
1   6   1   2   2   23.5   1 1 0 1 1 1 1 1 0 0
```

Pedigree file - *pedin3.pre* for example 3: For pedigree 1, genotypes will be generated for 20 marker loci that are in linkage equilibrium. Information on the quantitative trait BMI will not be included in the pedigree file. Since individual 1 is deceased, no genotypes will be generated for this individual. For all other individuals in this pedigree, genotype data will be generated for each of the 20 marker loci. Please note that there is no difference in the pedigree file for this example and example 1. Note that information on haplotypes (marker loci in linkage disequilibrium) or genotypes (for marker loci in linkage equilibrium) are all provided in the parameter file.

```
1   1   0   0   1   0
1   2   0   0   2   1
1   3   1   2   1   1
1   4   1   2   2   1
1   5   1   2   1   1
1   6   1   2   2   1
```

Pedigree file - *pedin4.pre* for example 4: For pedigree 1 a total of 220 marker loci will be generated. Both haplotype and genotype data will be generated for this pedigree. The quantitative trait BMI will be included in the pedigree file. No genotype data will be generated for individual 1 who is deceased.

1	1	0	0	1	22.3	0
1	2	0	0	2	30.3	1
1	3	1	2	1	29.6	1
1	4	1	2	2	24.5	1
1	5	1	2	1	35.7	1
1	6	1	2	2	23.5	1

Pedigree file – *pedin5.pre*: If more than one family needs to be generated, then the pedigree structures are placed in the same pedigree file. Note that it does not matter if individuals within a pedigree are numbered sequentially.

3	7	0	0	1	0
3	8	0	0	2	1
3	1	0	0	1	0
3	2	7	8	2	0
3	3	1	2	1	1
3	4	1	2	2	1
3	5	1	2	1	1
3	9	1	2	1	1
3	6	1	2	2	1
1	1	0	0	1	0
1	2	0	0	2	1
1	3	1	2	1	1
1	4	1	2	2	1
1	5	1	2	1	1
1	6	1	2	2	1

Creating a parameter file for example 1 (input1.dat)

The first line of the ***input1.dat*** file for example 1 contains the name of the pedigree file ***pedin1.pre*** and the output file name ***pedfile1.pre***. The second line contains three seeds each < 30,000. Since there is no trait locus included in the ***pedin1.pre*** file for example 1, a "0" is placed on line 3. The number of replicates to be generated is placed on line 4. On line 5 is placed the total number of marker loci which will be generated and then a "1" since for this example only one set of haplotypes will be generated, i.e. the pattern will only be generated once. For this example all intermarker recombination fractions are provided; therefore, on line 6 a "1" is entered. The number of intermarker recombination fractions and their values are denoted on line 7; for this example there are 10 marker loci, and therefore nine intermarker recombination

fractions are provided. For this example haplotypes will be generated, so a "1" is placed on line 8 followed by the "8 10 1", since eight haplotypes consisting of 10 marker loci will be generated and this set of haplotypes will only be generated once. On line 9 is placed the frequency of each haplotype. On lines 10- 17 are the observed alleles for each haplotypes.

pedin1.pre pedfile1.pre << name of pedigree file, name of output file
23221 1601 21001   << three random seeds
0 << number of columns for affection status/quantitative trait
1000 <<number of replicates
10 1 << Total # of marker loci, # of times pattern to be repeated
1 << "1" recomb fraction, "2" Kosambi map distance & "3" Haldane map distance
9 0.001 0.0 0.005 0.002 0.0 0.01 0.002 0.0 0.001 << # of θ values, θ values
1 8 10 1 << "1" for haplotypes, # of haplotypes, # of marker loci, # of times pattern repeated
0.35 0.15 0.10 0.08 0.08   0.08 0.08 0.08 << the frequency for each haplotype
1 1 1 1 1 1 1 1 1 1   < << observed alleles for each haplotype
2 2 2 2 1 1 1 1 1 1
1 1 2 2 2 2 2 1 1 2
1 2 1 2 1 2 1 2 1 2
1 1 1 1 2 2 2 2 2 1
2 2 2 1 1 1 1 2 2 2
1 1 2 2 1 1 2 2 1 1
2 2 2 1 1 2 2 2 2 1

Creating a parameter file for example 2 (input2.dat)

On the first line of the parameter file, *input2.dat* for example 2 the names of the pedigree file *pedin2.pre* and the output file *pedfile2.pre* specified. On the second line, three seeds each < 30,000 are given. In the pedigree file *pedin2.pre* there is one quantitative trait locus, so therefore a "1" is placed on line 3. The number of replicates is placed on line 4; for this example, 10,000 replicates will be generated. On line 5 a "10" denotes the total number of marker loci and the "1" signifies that pattern will be repeated only once. A "2" is placed on line 6, since Kosambi genetic map distances will be provided in Morgans. On line 7 is placed the number of Kosambi map distances "9" followed by nine intermarker Kosambi map distances. On line 8 appears "1 10 10 1", the "1" denotes that haplotypes will be provided, the first "10" is for the ten haplotypes that will be generated and the second "10" denotes these haplotypes will consist of ten marker loci the "1" indicates that these ten markers that form the haplotypes will only be generated once, i.e. the pattern will not be repeated. On line 9 is placed the frequency of each haplotype. On lines 10-19 are the observed alleles for each haplotypes. Note that for this example, markers 2 and 9 have > 2 alleles.

pedin2.pre pedfile2.pre << name of pedigree file, name of output file
21998 26002 11981 << three random seeds
1 << number of columns for affection status/quantitative trait
10000 << number of replicates
10 1<< Total # of marker loci, # of time pattern to be repeated
2 << "1"recomb fraction, "2" Kosambi map distance & "3" Haldane map distance
9 0.001 0.02 0.005 0.02 0.001 0.001 0.002 0.01 0.001 <<# map dist. and their values in Morgans
1 10 10 1 << "1" for haplotypes, # of haplotypes, # of marker loci, # of times pattern repeated
0.30 0.25 0.10 0.10 0.05 0.04 0.04 0.04 0.04 0.04 << the frequency for each haplotype
1 3 1 1 1 1 1 1 4 1 < << observed alleles for each haplotype
2 2 2 2 1 1 1 1 1 1
1 1 2 2 2 2 2 1 1 2
1 2 1 2 1 2 1 2 3 2
1 1 1 1 2 2 2 2 2 1
2 3 2 1 1 1 1 2 2 2
1 1 2 2 1 1 2 2 4 1
2 2 2 1 1 2 2 2 2 1
2 2 1 2 2 2 1 2 3 2
2 3 1 2 1 1 2 2 4 1

Creating a parameter file for example 3 (input3.dat)

The *input3.dat* file for example 3 contains on the first line the name of the pedigree file *pedin3.pre* and the name of the output file *pedfile3.pre*. The second line contains three seeds each < 30,000. Since there is no trait locus included in the *pedin3.pre* file for example 3, a "0" is placed on line 3. The number of replicates to be generated is placed on line 4. On line 5 is placed the total number of marker loci "20" followed by a "2" since allele frequency will be provided for 10 marker loci and the set of 10 marker loci will be generated twice. For this example, recombination fractions will be entered in the parameter file; therefore on line 6 a "1" is entered. The intermarker recombination fractions are denoted on line 7; for this example 10 intermarker recombination fractions are provide and these intermarker recombination fractions will be repeated to provide 19 intermarker recombination fractions. It should be noted that it also possible to enter 19 intermarker recombination fractions. Since for the first set of marker loci genotypes will be generated a "2" is placed on line 8 followed by "3" since allele frequencies are provided for three marker loci and a second "2" follows to denote that the allele frequencies provided will be repeated twice to generate genotype data for six marker loci. The allele frequency for the first allele of the six marker loci is as follows: 0.3, 0.4, 0.2, 0.3, 0.4 and 0.2. On line 11 is given the information "2 2 1", which specifies that genotype data will be generated for two marker loci and that this pattern will be generated only once. It should be noted that these two marker loci which will be generated have more than two alleles (i.e. three and six)   On line 15 appears "2 1 2", denoting genotypes will be generated, allele frequency data is provided for one marker loci and the pattern will be repeated twice. Therefore two markers will be generated each with their first allele having the frequency 0.5. The first allele frequency for the 20 marker loci is as follows: 0.3, 0.4, 0.2, 0.3, 0.4, 0.2, 0.7, 0.1, 0.5, 0.5, 0.3, 0.4, 0.2, 0.3, 0.4, 0.2, 0.7, 0.1, 0.5 and 0.5.

pedin3.pre pedfile3.pre << name of pedigree file, name of output file
1344 2673 12228 << three random seeds
0 << number of columns for affection status/quantitative trait
1000 << number of replicates
20 2   << Total # of marker loci, # of times the pattern will be repeated
1 << "1" recomb fraction, "2" Kosambi map distance & "3" Haldane map distance
10 0.01 0.05 0.05 0.09 0.01 0.01 0.02 0.05 0.01 0.05 << intermarker recombination fractions
2 3 2 << "2" genotype data will be provided, # of marker loci, # of times pattern repeated
2 0.3 << marker 1, with 2 alleles and the allele freq. for the first allele
2 0.4 << marker 2, with 2 alleles and the allele freq. for the first allele
2 0.2 << marker 3, with 2 alleles and the allele freq. for the first allele
2 2 1 << 2" genotype data will be provided, # of marker loci, # of times pattern repeated
3 0.7 0.1 << marker 7, with 3 alleles and the allele freq. for the first two alleles
6 0.1 0.1 0.1 0.1 0.1 << marker 8, with 6 alleles and the allele freq. for the first five alleles
2 1 2 << 2" genotype data will be provided, # of marker loci, # of times pattern repeated
2 0.5 marker << marker 9, with 2 alleles and the allele freq. for the first allele

Creating a parameter file for example 4 (input4.dat)

For example 4 a mixture of haplotype and genotype data will be generated. For the ***input4.dat*** file, the first line contains the names of the pedigree file ***pedin4.pre*** and the output file ***pedfile4.pre***. The second line contains three seeds for the random number generator each < 30,000. In the pedigree file ***pedin4.pre*** there is one quantitative trait locus, so therefore a "1" is placed on line 3. The number of replicates to be generated is placed on line 4 (i.e. 1000). On the next line is the total number of marker loci that will be generated; therefore, a "220" is placed on line 5 which is followed by a "10" which denotes that information will be provided for 22 marker loci and this pattern will be repeated in order to generate data for 220 marker loci. For this example recombination fractions are provided; thus on line 6 a "1" is entered. On line 7 the number of intermarker recombination fractions which are provided is specified. A total of 22 intermarker recombination fractions are provided. These intermarker recombination fractions will be used 10 times in order to provide the necessary 219 intermarker recombination fractions necessary for the simulation study. Please note that although $10 \times 22 = 220$, since the last intermarker recombination fraction is not necessary it will not be used.

First, genotype data will be generated for six marker loci. On line 8 is specified "2 2 3", the first "2" denotes that genotype data will be generated, the second "2 " denotes that allele frequencies will be specified for two marker loci and the last "3" denotes that the pattern should be repeated three times. Line 8 is followed by two lines, on each line is the marker allele frequency for each of the two diallelic marker loci, where the 2 denotes that the marker has two alleles and the second number is the allele frequency for the first allele. Since this pattern is to be repeated three times as specified on line 8, the allele frequencies for the first allele of the first six marker loci is 0.4, 0.5, 0.4, 0.5, 0.4 and 0.5. For the next set of marker loci haplotypes are to be generated. The first set of haplotypes contains five marker loci which are in linkage disequilibrium. On line 11 is denoted "1 8 5 2", Where the "1" specifies that haplotypes will be generated, the "8" denotes that eight haplotypes will be generated, the "5" denotes the

haplotypes are for five marker loci and the "2" denotes that this pattern will be repeated twice. On line 12, eight haplotype frequencies are specified which refer to the frequencies of each of the eight haplotypes configurations found in lines 13-20. Note that for five marker loci it is possible to have 32 different haplotype configurations; here only eight of these possible haplotypes are specified since the frequency for the other possible 24 haplotypes is 0. Since this pattern is to be repeated twice, haplotype data will be generated for the five marker loci in linkage disequilibrium and then for another set of five marker loci in linkage disequilibrium with the same haplotype frequencies. However these two "blocks" of marker loci will be in linkage equilibrium with each other. Lastly additional haplotype data will be generated. On line 21 is indicated "1 3 2 3" to denote that haplotype data will be generated ("1") that will consist of three haplotypes ("3") that are formed by two marker loci ("2") and this pattern will be repeated three times ("3"). On line 22 the haplotype frequency for the three haplotypes that are denoted on line 23-25 is specified. Here there are a possible four haplotypes for the two marker loci, but one of the haplotypes' 2 1 frequency is 0 and therefore will not be generated. The haplotypes for two markers is repeated three times, each set of haplotype "blocks" have the same haplotype frequency and although each of the two markers within a "block" are in linkage disequilibrium each of the "blocks" of marker loci are in linkage equilibrium with each other. The haplotype and allele frequencies have been specified for 22 marker loci. However, it is requested that 220 markers be generated on line 5; therefore, this pattern of haplotype and genotype frequencies (as denoted by the allele frequencies) will be repeated ten times.

pedin4.pre pedfile4.pre << name of pedigree file, name of output file
2312 1678 13247   << three random seeds
1 << number of columns for affection status/quantitative trait
1000 << number of replicates
220 10 << Total # of marker loci, # of times pattern should be repeated
1 << "1" recomb fraction, "2" Kosambi map distance & "3" Haldane map distance
22 0.05 0.07 0.05 0.1 0.03 0.04 0.0 0.0 0.0 0.0 0.05 0.001 0.0 0.0 0.0 0 0.07 0.0 0.05 0.0 0.03 0.0
2 2 3 << "2" genotype data, # of marker loci, # of times the pattern should be repeated
2 0.4 << marker 2, with 2 alleles and the allele freq. for the first allele
2 0.5 << marker 3, with 2 alleles and the allele freq. for the first allele
1 8 5 2 << "1" for haplotypes, number of haplotypes, number of marker loci
0.35   0.15   0.10   0.08   0.08   0.08   0.08   0.08   << the frequency for each haplotype
1 1 1 1 1 < << observed alleles for each haplotype
2 2 2 2 2
1 1 2 2 2
1 2 1 2 1
1 1 1 1 2
2 2 2 1 1
1 1 2 2 1
2 2 2 1 1
1 3 2 3
0.6 0.3 0.1
2 2
1 2
1 1

Executing SimPed

The SimPed program can be run by giving the command **SimPed** followed by the name of the parameter file. For example in order to run the analysis for example 1 the following command would be given:

> simped input1.dat

```
**********************************
*          Program   SimPed              *
**********************************
```

Usage: simped parameter_file

Constants in effect:
  Maximum number of pedigrees   100
  Maximum number of individuals   3000
  Maximum number of loci   10000
  Maximum number of haplotypes   1000
  Maximum number of alleles at a locus   20

Input pedigree file: pedin1.pre
Output file: pedfile1.pre

Shown below are the **output** files for examples 1-4. If necessary for further analysis, the **output** file can be processed using MAKEPED to create a post makeped file (e.g. *pedfile.dat*) which contains information on first offspring identifier, paternal sibling identifier, maternal sibling identifier and proband status.

Example 1: *pedfile1.pre* **(first 2 replicates only)**
```
1 1 0 0 1  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 0 0 2  2 2 2 2 2 2 1 1 1 1 2 1 2 1 2 2 2 2 2 1 2
1 3 1 2 1  1 2 1 2 1 2 1 1 1 1 1 2 1 2 1 2 1 2 1 2 1 1
1 4 1 2 2  1 2 2 2 1 2 2 1 1 1 2 1 1 1 2 2 1 2 2 2
1 5 1 2 1  1 2 2 2 1 2 2 1 1 1 2 1 1 1 2 2 1 2 2 2
1 6 1 2 2  1 2 1 2 1 2 1 1 1 1 1 2 1 2 1 2 1 2 1 1
2 1 0 0 1  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 2 0 0 2  1 1 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 1 1 2
2 3 1 2 1  2 1 2 1 2 2 1 2 1 2 1 2 1 2 2 1 2 1 2 2
2 4 1 2 2  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 5 1 2 1  2 1 2 1 2 2 1 2 1 2 1 2 1 2 2 1 2 1 2 2
2 6 1 2 2  2 1 2 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1
```

Example 2: *pedfile2.pre* **(first 2 replicates only)**
```
1 1 0 0 1 22.3  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 0 0 2 30.3  1 1 2 3 1 1 2 1 1 1 2 1 1 1 2 1 3 4 2 1
1 3 1 2 1 29.6  2 1 2 2 2 1 2 2 1 1 1 2 1 1 1 2 1 3 1 2
1 4 1 2 2 24.5  1 1 3 3 1 1 1 1 0 0 1 1 1 1 0 0 4 4 1 1
1 5 1 2 1 35.7  1 1 3 2 1 1 1 2 1 1 1 2 1 1 1 2 4 3 1 2
1 6 1 2 2 23.5  1 1 3 3 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0
2 1 0 0 1 22.3  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 2 0 0 2 30.3  1 1 1 1 2 1 2 1 2 2 2 2 2 1 2 1 2 2 1
2 3 1 2 1 29.6  2 1 3 1 2 2 1 2 1 2 1 2 1 2 2 2 1 2 1 2 2
2 4 1 2 2 24.5  2 1 2 1 2 1 2 1 0 0 1 2 1 2 0 0 1 2 1 1
2 5 1 2 1 35.7  2 1 3 1 2 2 1 2 1 2 1 2 1 2 2 2 1 2 1 2 2
2 6 1 2 2 23.5  2 1 2 1 0 0 2 1 1 2 1 2 1 2 2 2 0 0 0 0
```

Example 3: *pedfile3.pre* **(first 2 replicates only)**
```
1 1 0 0 1  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 0 0 2  2 2 1 2 2 2 2 2 2 1 1 2 3 3 4 6 1 1 2 1 1 2 2 2 2 1 2 1 1 1 1 1 3 1 6 6 2 1 1 1
1 3 1 2 1  2 2 1 1 2 2 1 2 1 2 1 1 2 3 4 4 2 1 2 2 2 2 2 2 1 1 2 1 1 1 2 1 3 1 2 6 1 1 1 1
1 4 1 2 2  2 2 2 1 2 2 1 2 2 2 2 1 1 3 5 4 1 1 1 2 1 1 2 2 1 2 2 2 1 1 2 1 3 3 2 6 1 2 1 1
1 5 1 2 1  2 2 1 2 2 2 1 2 1 1 1 2 2 3 4 6 2 1 2 1 2 2 2 2 1 1 2 1 1 1 2 1 3 1 2 6 1 1 1 1
1 6 1 2 2  2 2 2 2 2 2 1 2 2 1 2 2 1 3 5 6 1 1 1 1 2 2 2 1 1 2 1 2 1 2 1 1 1 4 6 2 1 2 1
2 1 0 0 1  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 2 0 0 2  2 2 1 2 2 2 1 2 1 2 2 2 1 1 6 2 1 1 1 2 2 2 1 2 2 1 1 2 2 2 2 1 3 1 2 6 1 2 2 1
2 3 1 2 1  2 2 2 1 2 2 1 1 2 1 2 2 1 1 5 6 1 1 1 1 2 2 2 1 2 1 2 2 1 2 2 1 1 1 3 6 1 2 1 1
2 4 1 2 2  2 2 2 1 2 2 1 1 2 2 2 2 1 1 5 2 1 1 1 2 2 2 2 2 2 1 2 2 1 2 2 2 1 3 3 2 1 1 2 2
2 5 1 2 1  2 2 2 1 2 2 2 1 2 1 2 2 1 1 5 6 2 1 2 1 2 2 1 1 1 2 2 1 2 2 2 2 2 3 6 2 2 1 2 2
2 6 1 2 2  2 2 2 2 2 2 2 2 2 2 2 2 1 1 6 2 2 1 2 2 2 2 1 2 1 1 2 2 1 2 2 1 1 1 3 6 1 2 1 1
```

Example 4: *pedfile4.pre* **(first 2 replicates and 22 markers only)**
```
1 1 0 0 1 22.3  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 0 0 2 30.3  1 2 2 1 2 1 1 1 2 2 1 1 1 1 1 1 2 1 2 1 2 1 1 1 1 2 2 2 2 1 2 1 2 2 1 2 2 2 1 1 2 1
1 3 1 2 1 29.6  2 1 1 2 1 2 1 1 1 2 2 1 1 1 1 1 2 2 2 2 2 2 1 1 2 1 1 2 2 2 1 2 2 2 2 2 2 1 2 2 1 1 2 2
1 4 1 2 2 24.5  2 2 2 1 1 1 1 1 1 1 2 2 1 1 1 1 1 2 2 2 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 2 2 2 2 2 2 1 1 2 1
1 5 1 2 1 35.7  2 1 1 2 1 2 1 1 2 2 1 1 1 1 1 1 2 2 2 2 2 2 1 1 2 1 1 2 2 2 1 2 2 2 2 2 2 1 2 2 1 1 2 1
1 6 1 2 2 23.5  2 2 1 1 1 1 1 1 2 2 2 1 1 1 1 1 2 1 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 1 2 2 2 2 2 2 1 1 2 1
2 1 0 0 1 22.3  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 2 0 0 2 30.3  2 2 1 1 2 2 1 2 2 2 1 1 2 1 2 1 2 2 2 1 2 1 2 1 2 1 1 1 1 1 1 1 2 1 2 2 2 2 1 1 2 1 2 2 2 2
2 3 1 2 1 29.6  1 2 2 1 1 2 1 1 1 2 2 1 1 2 1 2 1 2 1 2 1 1 1 1 1 1 1 2 1 2 1 1 2 2 2 2 2 1 1 2 2 2 2 2 2
2 4 1 2 2 24.5  1 2 1 1 1 2 1 2 1 2 2 1 1 2 1 2 1 2 1 2 1 1 1 1 2 1 2 1 2 1 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2
2 5 1 2 1 35.7  1 2 1 1 2 2 1 2 1 2 1 2 2 1 1 1 1 1 2 1 2 1 2 2 1 2 1 2 1 2 1 2 1 2 2 2 2 2 1 2 1 2 2 2 2
2 6 1 2 2 23.5  2 2 2 1 1 2 1 1 1 1 2 2 1 1 2 1 2 1 2 1 2 1 2 1 1 1 1 2 1 2 1 2 1 2 1 2 2 2 2 2 2 1 2 2 2 2 2
```

16

Technical Information

The SimPed program is written in C. The Source code for SimPed as well as executable files are available to run under LINUX, UNIX and Windows. The program allows for dynamic allocation of memory and the executable files are distributed with constants set to the values shown on page 15 (i.e. section entitled "Executing SimPed"). The parameter values can be increased or decreased as necessary by the user and the program can be recompiled. A number of freely available C compilers including "cc" and "gcc" can be used to recompile the SimPed program when necessary.

 The command to recompile SimPed using "gcc" is:
> gcc –lm simped.c –o simped

The reference for the SimPed program is:

Leal SM, Yan K, Muller-Myhsok B. SimPed: a simulation program to generate haplotype and genotype data for pedigree structures. Hum Hered, 60 (2), 119-22 (2005).