ECLIPSE User's Guide

October 30, 2003

This directory contains two different programs. The eclipse3 program calculates the likelihood of the observed phenotypes for trios of individuals under the twenty-five relationships which are a combination of full sibs, half-sibs, MZ twins and unrelated. Eclipse3 also includes eighteen additional relationships which are distinct orderings of two full sibs with a niece, two full sibs with an aunt, two full sibs with a grandparent, two full sibs with a parent, two parents with an offspring, and two full sibs with a first cousin. These relationships are listed in Table 1. The eclipse2 program calculates phenotype likelihoods for pairs of individuals under the relationships: full sibs, half sibs, monozygous twins, unrelated individuals, parent-offspring, grandparent-grandchild, avuncular pair, and first cousin pair. Both of the programs can also compute the posterior probability of genotyping error at each locus.

As a default, these programs use Kosambi's map function to convert genetic distance to recombination fraction, although Haldane's map function is available by changing the marker-map.h file and recompiling.

To compile the programs use the commands:

- make eclipse3
- make eclipse2
- make clean

which compiles with the cxx compiler. An alternate approach is to compile using g++, which results in a slower running program:

- g++ -o eclipse3 eclipse3.cpp
- g++ -o eclipse2 eclipse2.cpp

Or replace g++ with the C++ compiler of your choice (no guarantees that it will work). To run the programs, you'll need to specify a *config* file. The default is "config" in the current directory. Alternately, one can be specified with the -c option. To run the programs:

- eclipse2 [-c configfile] [-p PostErrProb]
- eclipse3 [-c configfile] [-p PostErrProb]

The configuration file and data formats are the same for both programs and are described below.

Table 1: Relationships among trios considered for likelihood computations and the number of distinct orderings among ordered trios.

Relationship	ECLIPSE output	Number of Distinct Orderings
Full sibs	`full`	1
Half sibs (one common parent)[a]	`half`	1
Unrelated individuals	`unrel`	1
MZ twins (repeated samples)	`mono`	1
Two full sibs + half sib	`full_hlf[1-3]`	3
Half sibs (two different parents)[a]	`half[1-3]`	3
MZ twins + sib	`mz_full[1-3]`	3
MZ twins + unrelated	`mz_half[1-3]`	3
Two full sibs + unrelated individual	`full_unr[1-3]`	3
Two half sibs + unrelated	`half_unr[1-3]`	3
MZ twins + unrelated	`mz_unrel[1-3]`	3
Two full sibs + niece	`full_niec[1-3]`	3
Two full sibs + aunt	`full_aunt[1-3]`	3
Two full sibs + grandparent	`full_gp[1-3]`	3
Two full sibs + cousin	`full_cous[1-3]`	3
Two full sibs + parent	`full_par[1-3]`	3
Two parents + offspring	`par_off[1-3]`	3

[a] Shown in Figure 1.

1 The Configuration File

A *config* file must be specified or the working directory must contain a file named "config", which contains the file names for the locus file, the pedigree file and error file. It also specifies the locations for output and other options. The *config* file should be arranged as follows:

1. The first line should contain the name of the marker map file.
2. The second line should contain the name of the pedigree file.
3. The third line should contain the name of the error rate file.
4. The fourth, fifth and sixth lines should contain the names for the three output files: for warnings, for posterior genotyping error probability and for relationship likelihoods.
5. The seventh line gives the option to calculate only the likelihood of the relationship (the forward portion of the Baum algorithm). The line should read "forward" to calculate only the likelihoods or "backward" for both the likelihoods and the posterior probabilities of genotyping error. Even if the "forward" option is chosen, a filename must be specified for posterior probability output.
6. The eighth line gives the option to compare only individuals within a family. Specify "family" for comparisons only within a family or "all" for all trio (pair) comparisons.
7. The nineth line specifies the error model used in the computations (0-3). These models are described in detail below. In general, only error models 0 and 2 should be used for trios of individuals when more than two alleles exist per locus.

For example,

```
sample.loc
sample.ped
err.in
warnfile
dumpfile
logL
backward
all
0
```

2 The Map File

The map file (e.g. sample.loc) contains the marker characteristics, including map positions and allele names and frequencies. Each marker should be represented in this file, in order of their physical location.

1. The first entry for each marker is the chromosome number. The sex chromosomes must be labeled X or Y to distinguish them from the autosomal chromosomes and must be input AFTER all the autosomal chromosomes.
2. The second entry should be the position of the marker in cM. This position must be a positive number.
3. The third entry should contain the number of alleles at that marker, which should be no more than 99.
4. Finally, the name of each allele, which can be alpha-numeric, should be followed by its frequency.

For example,

```
1 0.0 3 101 0.18 102 0.55 103 0.27
1 5.0 4 201 0.01 202 0.07 203 0.86 204 0.06
1 7.0 3 301 0.32 302 0.61 303 0.07
1 13.0 5 401 0.14 402 0.02 403 0.79 404 0.01 405 0.04
X 0.0 3 501 0.3 502 0.5 503 0.2
Y 0.0 3 601 0.3 602 0.5 603 0.2
```

specifies four markers on chromosome 1 (with 3,4,3 and 5 alleles) and one marker on both the X and Y chromosomes. These same markers may also be specified as follows:

```
1 0.0 3
101 0.18
102 0.55
103 0.27
```

```
1 5.0 4
201 0.01
202 0.07
203 0.86
204 0.06

1 7.0 3
301 0.32
302 0.61
303 0.07

1 13.0 5
401 0.14
402 0.02
403 0.79
404 0.01
405 0.04

X 0.0 3
501 0.3
502 0.5
503 0.2

Y 0.0 3
601 0.3
602 0.5
603 0.2
```

3 The Pedigree File

The pedigree file (sample.ped) contains both the pedigree information and marker data for each individual. Only individuals with marker data should be included in this file.

1. The pedigree file should contain one line for each individual.
2. The first two positions should be filled by the individuals identifiers. The first of which should be the family's name, while the second one is the individuals name. The names can be letters or numbers.
3. The next two positions should be filled with the individuals father's and mother's individual identifiers, respectively. If these are unknown, these positions should be filled with zeros (0). Data on an individual's parents need not be included in the pedigree file and should not be included if no marker data is available.
4. The fifth position should be filled with the individuals sex: MALE = 1, FEMALE = 2, UNSPECIFIED = 0. (Note that these conventions can be changed by changing the MALE and FEMALE global variables in the `eclipse2_const.h` and `eclipse3_const.h` files.)

5. The remaining positions should be filled by the 2m (m= number of markers) alleles of the individuals genotype, in order by marker. If the genotype is unknown, then 0's may fill the positions. Males should be homozygous for every marker on the X and Y chromosomes. Females should have 0's in the Y chromosome marker positions.

For example:

```
f10_1 i900 i800 i801 0 101 101 202 202 301 301 404 404 501 501 603 603
f10_1 i901 i800 i801 1 102 103 201 203 302 303 401 405 501 502   0   0
f10_1 i902 i800 i802 1 101 102   0   0 301 302   0   0 501 503   0   0
f10_2  950  850  851 2 101 101 202 202 301 301 404 404 502 503   0   0
f10_2  951  852  851 2 102 103 201 203 302 303 401 405 502 502 601 601
```

4 The Error Rate File

The error file should specify the error rate parameter(s) used in the computations. The format of the error file and the interpretation of error rate depends on the error model being used.

Error models 0,1 and 2 have a single error rate parameter, which is discussed further below. In these cases, the error file should contain the list of error rates. If the "backward" option is specified the error rate, 0.0, will always be evaluated.

For example:

```
0 0.001 0.01 0.05
```

When error model 3 is specified, two error rates must be specified for every computation: an allelic error rate and an allelic loss rate. In this case, each line contains a pair of parameters. If the "backward" option is specified, the error rate, (0.0, 0.0), will always be evaluated.

For example:

```
0 0
0 0.001
0.001 0
0.01 0.01
0.05 0.01
```

Further discussion of the error models is made below. Note that the genotyping error rates have different interpretations in each genotyping error model.

5 The Output Files

For more technical details of the computations, see (Sieberts et al 2002). Both output files contain quantities computed for each of the considered relationships. For eclipse3, in which trios of individuals are considered jointly, 37 relationships are considered.

1. `full`: three full sibs

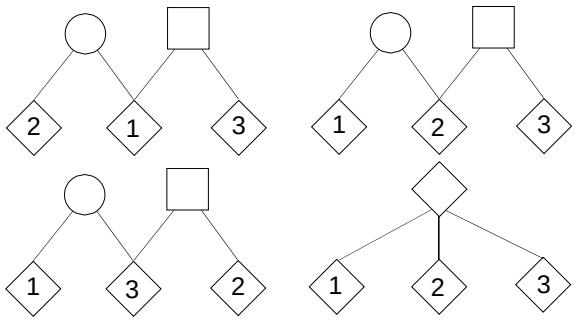

Figure 1: Four possible half sib relationships. The pedigree in the lower-right quadrant shows three half-sibs with a common parent (half); each of the remaining three pedigrees shows three half-sibs with two different parents (from the top left: half1, half2, and half3).

2. `half`: three half sibs with the same common parent (figure 1)
3. `unrel`: three unrelated individuals
4. `mz`: three monozygous twins
5. `full_half1`: the first individuals is half sibs with the remaining two full sibs.
6. `full_half2`: the second individuals is half sibs with the remaining two full sibs.
7. `full_half3`: the third individuals is half sibs with the remaining two full sibs.
8. `half1`: the first individual is half sibs with each of the other two (see figure 1).
9. `half2`: the second individual is half sibs with each of the other two (see figure 1).
10. `half3`: the third individual is half sibs with each of the other two (see figure 1).
11. `mz_full1`: the first individual is full sibs to two monozygous twins
12. `mz_full2`: the second individual is full sibs with two monozygous twins
13. `mz_full3`: the third individual is full sibs with two monozygous twins
14. `mz_half1`: the first individual is half sibs with two monozygous twins
15. `mz_half2`: the second individual is half sibs with two monozygous twins
16. `mz_half3`: the third individual is half sibs with two monozygous twins
17. `full_unrel1`: the first individual is unrelated to two full sibs
18. `full_unrel2`: the second individual is unrelated to two full sibs
19. `full_unrel3`: the third individual is unrelated to two full sibs
20. `half_unrel1`: the first individual is unrelated to two half sibs
21. `half_unrel2`: the second individual is unrelated to two half sibs
22. `half_unrel3`: the third individual is unrelated to two half sibs
23. `mz_unrel1`: the first individual is unrelated to two monozygous twins
24. `mz_unrel2`: the second individual is unrelated to two monozygous twins
25. `mz_unrel3`: the third individual is unrelated to two monozygous twins
26. `full_niece1`: two full sibs (individuals 2 and 3) with their niece (1)
27. `full_niece2`: two full sibs (individuals 1 and 3) with their niece (2)
28. `full_niece3`: two full sibs (individuals 1 and 2) with their niece (3)
29. `full_aunt1`: two full sibs (individuals 2 and 3) with their aunt (1)
30. `full_aunt2`: two full sibs (individuals 1 and 3) with their aunt (2)
31. `full_aunt3`: two full sibs (individuals 1 and 2) with their aunt (3)
32. `full_gp1`: two full sibs (individuals 2 and 3) with their grandparent (1)
33. `full_gp2`: two full sibs (individuals 1 and 3) with their grandparent (2)
34. `full_gp3`: two full sibs (individuals 1 and 2) with their grandparent (3)
35. `full_cous1`: two full sibs (individuals 2 and 3) with their first cousin (1)
36. `full_cous2`: two full sibs (individuals 1 and 3) with their first cousin (2)

37. `full_cous3`: two full sibs (individuals 1 and 2) with their first cousin (3)
38. `full_par1`: two full sibs (individuals 2 and 3) with their parent (1)
39. `full_par2`: two full sibs (individuals 1 and 3) with their parent (2)
40. `full_par3`: two full sibs (individuals 1 and 2) with their parent (3)
41. `par_off1`: two parents (individuals 2 and 3) with an offspring (1)
42. `par_off2`: two parents (individuals 1 and 3) with an offspring (2)
43. `par_off3`: two parents (individuals 1 and 2) with an offspring (3)

When individuals are considered pairwise in eclipse2, many fewer relationship are considered. The relationships are full sibs (`full`), half sibs (`half`), unrelated individuals (`unrel`), monozygous twins (`mono`), parent-offspring (`P-O`), grandparent-grandchild (`GP-GC`), avuncular pair (`avun`), and a first cousin pair (`FC`).

5.1 warnfile

The warning files differ slightly for eclipse2 and eclipse3. For both programs, the warning file begins with warnings for individuals with invalid genotypes. These invalid genotypes are ignored and treated as missing. Next the file contains some basic pedigree warnings including that males are homozygous for X and Y chromosome markers and that females have no data on the Y chromosome. For pairs of individuals with the same father, warnings are made for males that do not share all Y chromosome alleles and for females that do not have at least one allele in common at each X-chromosome marker.

For eclipse3, the file then contains a list of trio comparisons which are reported full sibs, but are more likely another relationship. For each error rate, the most likely relationship is listed with the log-likelihood difference (most likely relationship minus reported relationship). The same output format is used for the list of comparisons which are reported as a pair of full sibs with a half sib (relationships 5-7), half sibs (relationships 2 and 8-10), and parent offspring (relationships 38-43). Next a listing of trio comparisons, in which the reported relationship among the individuals is unrelated or unreported, but the most likely relationship is not unrelated, is given. For each error rate, the most likely relationship is listed along with the log-likelihood difference between the top two relationships.

For eclipse2, the file contains a list of reported full sib pairs that are more likely another relationship. For each error rate, the most likely relationship is listed with the log-likelihood difference (most likely relationship minus reported relationship). The same output format is used for the list of reported half sib pairs and the list of reported parent-offspring pairs. Finally, the list of unrelated or unreported individual pairs, which are likely related, is given. For each error rate, the most likely relationship is listed along with the log-likelihood difference between the top two relationships.

Using the *-p PostErrProb* option, the programs will also output markers at which the genotypes among a trio (or pair) have posterior genotyping error greater than *PostErrProb* for the reported relationship. In the case when the reported relationship is not the most likely relationship, probable genotyping errors are reported for both relationships. In this case the reported relationship is marked with (*) and the most likely relationship is marked with (**). For eclipse3, posterior probabilities are only output when the reported relationship is 1, 2, 5-10, or 38-43. For eclipse2, posterior probabilities are only output when the reported relationship is full sibs, half sibs, monozygous twins or parent-offspring.

5.2 dumpfile

The dumpfile outputs differ for eclipse2 and eclipse3. First we discuss eclipse3. For each error probability and at each marker, the dump file contains the $\log_{10}$ of cumulative probability of the data up to, but not including, the current marker for each of the relationships, in the order listed above. The next line contains the probability that the observed phenotype at the marker is correct under each relationship. Because of the large numbers of IBD and Markov states and relationships posterior probabilities of IBD are not reported in eclipse3, however the ``out_impl.cpp'' file can easily be changed to output these quantities if desired.

In eclipse2, when many fewer relationships and IBD states are considered, some additional quantities are output. For each error rate and marker, the output is as follows:

1. The marker number (starting at zero) and the marker position (possibly shifted by 0.1 cM).
2. For each of the four IBD states (sharing none, maternal, paternal, both), the probability of the IBD state at the marker, given all the data up to and including the marker under *full sibs*.
3. The $\log_{10}$ of the likelihood up to, but not including, the current marker under *full sibs*.
4. For the two half sib IBD states (sharing none or one), the probability of the IBD state at the marker, given all the data up to and including the marker under *half sibs*.
5. The $\log_{10}$ of the likelihood up to, but not including, the current marker under *half sibs*.
6. The $\log_{10}$ of the likelihood up to, but not including, the current marker under *unrelated*.
7. The $\log_{10}$ of the likelihood up to, but not including, the current marker under *monozygous twins*.
8. The $\log_{10}$ of the likelihood up to, but not including, the current marker under *parent-offspring*.
9. For the two GP-GC IBD states (sharing none or one), the probability of the IBD state at the marker, given all the data up to and including the marker under *GP-GC*.
10. The $\log_{10}$ of the likelihood up to, but not including, the current marker under *grandparent-grandchild*.
11. For each of the four avuncular Markov states, the probability of the state at the marker, given all the data up to and including the marker under *aunt-niece*.
12. The $\log_{10}$ of the likelihood up to, but not including, the current marker under *avuncular pair*.
13. For each of the seven first cousin Markov states, the probability of the state at the marker, given all the data up to and including the marker under *first cousins*.
14. The $\log_{10}$ of the likelihood up to, but not including, the current marker under first *cousins*.

15. For each of the four IBD states, the probability of the IBD state given all the marker data, under *full sibs*.

16. For each of the two half sib IBD states, the probability of the IBD state given all the marker data, under *half sibs*.

17. For each of the two GP-GC IBD states, the probability of the IBD state given all the marker data, under *GP-GC*.

18. For each of the four avuncular Markov states, the probability of the state given all the marker data, under *avuncular pair*.

19. For each of the seven FC Markov states, the probability of the state given all the marker data, under *first cousins*.

20. For each of the eight relationships (sib, halfsib, unrelated, MZ twins, P-O, GP-GC, avuncular and FC), the posterior probability that the observed phenotype is the true genotype, at the marker.

Thus, the posterior probability of no genotyping error under each relationship, which are probably of highest interest, are the last eight quantities on each line.

5.3 logL

The "logL" output file contains the $\log_{10}$ of the phenotypes given the relationship, for each pair of individuals at each error rate. Thus, likelihood ratios can be obtained by $10^{(P(rel1)-P(rel2))}$. Only the autosomal markers are used in likelihood computations, although the sex chromosomes are used for some basic checks.

6 The Error Models

There are four different error models from which to choose. In eclipse3, when three individuals are considered jointly, models 0 and 2 are feasible for multi-allelic markers, while models 1 and 3 should only be used with markers few (2 or 3) alleles. In eclipse2, all models are appropriate. The model assumptions are described below. Note that the genotyping error rate specified in the error file, ϵ (or ϵ_1 and ϵ_2 for model 3), has different interpretations in each of the different error models.

6.1 Model 0

Model 0 assumes that genotyping error occur in at most one allele per locus per individual. In other words, at each locus at least one of the two alleles must be correct. The genotype is correct with probability $1-\epsilon$, where ϵ is the genotyping error rate which is specified in the error file. The resulting probability of the observed phenotype given the true underlying genotype

depends on whether the true genotype is homozygous or heterozygous:

$$P(Y_n^{(i)} = y | X_n^{(i)} = (a,a), \epsilon) = \begin{cases} 1-\epsilon & y = (a,a) \\ \frac{\epsilon}{m-1} & y = (a,b) \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

When the true genotype is heterozygous:

$$P(Y_n^{(i)} = y | X_n^{(i)} = (a,b), \epsilon_S) = \begin{cases} 1-\epsilon_S & y = (a,b) \\ \frac{\epsilon_S}{2(m-1)} & y = (a,a), (a,c), (b,b), \text{ OR } (b,c) \\ 0 & \text{otherwise} \end{cases} , \quad (2)$$

This model is appropriate for all types of markers.

6.2 Model 1

The previous model was a simplification of model 1. Model 1 makes no restrictions on the number of alleles that are in error. Now, ϵ is the error rate per allele. Once again, the probability of the observed phenotype given the true underlying genotype depends on whether the true genotype is homozygous or heterozygous:

$$P(Y_n^{(i)} = y | X_n^{(i)} = (a,a), \epsilon_K) = \begin{cases} (1-\epsilon_K)^2 & y = (a,a) \\ 2\frac{\epsilon_K}{m-1}(1-\epsilon_K) & y = (a,b) \\ \left(\frac{\epsilon_K}{m-1}\right)^2 & y = (b,b) \\ 2\left(\frac{\epsilon_K}{m-1}\right)^2 & y = (b,c) \end{cases} . \quad (3)$$

When the true genotype is heterozygous,

$$P(Y_n^{(i)} = y | X_n^{(i)} = (a,b), \epsilon_K) = \begin{cases} (1-\epsilon_K)^2 + \left(\frac{\epsilon_K}{m-1}\right)^2 & y = (a,b) \\ \frac{\epsilon_K}{m-1}(1-\epsilon_K) & y = (a,a) \text{ OR } y = (b,b) \\ \frac{\epsilon_K}{m-1}(1-\epsilon_K) + \left(\frac{\epsilon_K}{m-1}\right)^2 & y = (a,c) \text{ OR } y = (b,c) \\ \left(\frac{\epsilon_K}{m-1}\right)^2 & y = (c,c) \\ 2\left(\frac{\epsilon_K}{m-1}\right)^2 & y = (c,d) \end{cases} , \quad (4)$$

This model is appropriate in eclipse2 for all types of markers and in eclipse3 when few alleles exist at each marker.

6.3 Model 2

Model 2, while less realistic than the other models, has computational properties which make it useful. This model assumes that given a genotyping error has occurred, the observed phenotype occurs according to its frequency in the population. As in model 0, the error rate, ϵ, is the

probability of a genotyping error per locus per individual (in contrast to a rate per allele). The probability of a phenotype given the true genotype is

$$P(Y_n^{(i)} = y | X_n^{(i)} = (a,b), \epsilon_R) = \begin{cases} (1-\epsilon_R) + \epsilon_R P(Y_n^{(i)} = (a,b)) & y = (a,b) \\ \epsilon_R P(Y_n^{(i)} = y) & y \neq (a,b) \end{cases} \quad (5)$$

where $P(Y_n^{(i)} = (a,b))$ is the proportion of the genotype (a,b) in the population.

This model is the least computationally intense and is appropriate in all situations.

6.4 Model 3

Model 3 is the most complex and probably the most realistic of the genotyping error models. This model requires the specification of two error rate parameters, ϵ_1 and ϵ_2. This is an extension of model 1, in which allelic mistyping occurs with rate ϵ_1. Model 3 also incorporates allelic loss, which could be due to amplification failure, for example. Allelic loss occurs at a rate of ϵ_2. When the true genotype is homozygous, the probability of the observed marker phenotype given the true genotype is

$$P(Y_n^{(i)} = y | X_n^{(i)} = (a,a), \epsilon_1, \epsilon_2) = \begin{cases} (1-\epsilon_2)(1-\epsilon_1)^2 + \epsilon_2(1-\epsilon_1) & y = (a,a) \\ 2\frac{\epsilon_1}{m-1}(1-\epsilon_2)(1-\epsilon_1) & y = (a,b) \\ (1-\epsilon_2)\left(\frac{\epsilon_1}{m-1}\right)^2 + \frac{\epsilon_1\epsilon_2}{m-1} & y = (b,b) \\ 2(1-\epsilon_2)\left(\frac{\epsilon_1}{m-1}\right)^2 & y = (b,c) \end{cases} . \quad (6)$$

When the true genotype is heterozygous, it is

$$P(Y_n^{(i)} = y | X_n^{(i)} = (a,b), \epsilon_1, \epsilon_2) = \begin{cases} (1-\epsilon_2)\left((1-\epsilon_1)^2 + \left(\frac{\epsilon_1}{m-1}\right)^2\right) & y = (a,b) \\ \frac{\epsilon_1}{m-1}(1-\epsilon_2)(1-\epsilon_1) + \frac{\epsilon_2}{2}(1-\epsilon_1) + \frac{\epsilon_1\epsilon_2}{2(m-1)} & y = (a,a) \text{ OR } y = (b,b) \\ (1-\epsilon_2)\left(\frac{\epsilon_1}{m-1}(1-\epsilon_1) + \left(\frac{1-\epsilon_1}{m-1}\right)^2\right) & y = (a,c) \text{ OR } y = (b,c) \\ (1-\epsilon_2)\left(\frac{1-\epsilon_1}{m-1}\right)^2 + \epsilon_2\frac{\epsilon_1}{m-1} & y = (c,c) \\ 2(1-\epsilon_2)\left(\frac{1-\epsilon_1}{m-1}\right)^2 & y = (c,d) \end{cases} , \quad (7)$$

where a, b, c, and d are unique alleles.

As with model 1, this model should only be used with eclipse3 when few alleles exist per marker. With eclipse2, this model can be used with any type of markers.

7 Missing Data

In eclipse3, when data is missing on two or more individuals at a marker, then the probability of the data at the marker given the IBD state, $P(Y_m|I_m)$, is reported as 1 regardless of whether data on the third individual is missing as well. This means that quantities such as the probability of the data (the likelihoods) are correct up to a constant. Thus the likelihood ratios are correct, but the raw probabilities are correct only up to a constant. Similar computational methods are used in eclipse2 when marker data is missing for one of the individuals.

References

[1] Sieberts SK, Wijsman EM, Thompson EA (2002) Relationship inference from trios of individuals, in the presence of typing error. Am J Hum Genet 70:170–180.