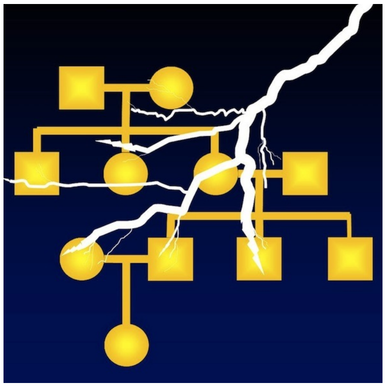
STORM:
Software for Testing Hypotheses of Relatedness and Mating Patterns



ver 2.0
(Updated January 31, 2012)

Written By:

Tim Frasier

Department of Biology & Forensic Sciences Program
Saint Mary's University
923 Robie St.
Halifax, NS B3H 3C3
Canada
Tel: (902) 491-6382
E-mail: timothy.frasier@smu.ca

Contents

1 Overview

STORM is a program written in C designed to use Monte Carlo simulations to test a variety of hypotheses regarding relatedness, mating, and/or fertilization patterns. Most options consist of two steps: (1) obtaining results from your observed data; and (2) using Monte Carlo simulations to obtain the "expected" results under the null hypothesis. By conducting many simulations (e.g. 1,000), the results of these simulations provide the expected distribution of results, which can then be compared to the observed values and used to estimate p-values. The details for each set of analyses are described below. There are many strengths to testing hypotheses in this fashion. First is that they allow testing of complex hypotheses that would be difficult to properly test using typical statistical tests. Second, they greatly increase the power of testing hypotheses by building the null distribution from the same data, structure, and model as the observed data. Thus, the characteristics of this distribution (e.g. the mean, standard deviation, etc.) are solely based on *your* data, and not on a general probability distribution that may or may not be a good representation of your true null expectations.

Monte Carlo simulations are reliant on random number generators (or 'pseudo-random' number generators), and STORM uses random number generators from the GNU Scientific Library (GSL, Galassi *et al.* 2006).

1.1 Citing STORM

The citation for STORM is:

Frasier TR (2008) STORM: software for testing hypotheses of relatedness and mating patterns. *Molecular Ecology Resources* **8**: 1263-1266.

1.2 Caution!

Monte Carlo simulations are an appropriate method to test hypotheses only when the number of possible combinations of your data greatly outnumber the number of iterations that are performed. If the number of iterations you perform outnumbers the possible combinations of your data then you will get the same scenario replicated multiple times in your simulations, and thus get erroneous estimates of your p-values.

As an example, suppose that I have genotyped 100 individuals. Within that data set I have identified two groups of individuals, each group containing three individuals, and the individuals within these groups are always found together. I suspect that they are relatives, and I want to use STORM to test if they are more/less related than expected based on chance. If I just use the data from the two groups ($N = 6$ individuals) then there are only 10 different ways to group them into to groups of three. Therefore, if I perform 1,000 iterations, I will have repeated each scenario $\sim$10 times, making my estimated p-value completely bogus. What would be appropriate in this case would be to make a third "dummy" group containing the other 94 genotypes. This way the individuals can be shuffled to and from this dummy group throughout the iterations, and you can get reliable p-value estimates from the simulations for your two groups of interest (because the number of ways to group 6 individuals into two groups of three out of a pool of 100 individuals is much greater than 1,000). Indeed, this is the best way to test the null hypotheses that the two observed groups of three represent random groups of individuals (from your sampled population) with respect to relatedness.

The take-home message is to consider your sample size when conducting your analyses and deciding how many Monte Carlo iterations to perform for hypothesis testing. Ensure that you have your analyses set up in such a way that the randomization processes (described in detail below) will properly test your null hypothesis and will not (likely) result in duplicates during the simulation process.

1.3 Stop the Program!

If you need to stop the program at any time, press CTRL-C. This is the generic "kill" command.

1.4 Changes From Previous Versions

Version 1.0 (the first release of the program) was released in 2008.

Version 1.1 was released in 2009. This changed involved a complete re-writing of the program, but no change in functionality. The original program was written in such a way that entire genotype files were read into the RAM, and each calculation created a new (and large) array that took up even more RAM. Thus, the size of data sets that could be analyzed with STORM were limited by the amount of RAM available. The code was re-written to be more efficient, so that the RAM only contains a small amount of information at any one time. With this change it should be possible to analyze data set of any size with STORM.

Version 2.0 was released in 2012. The major change was the addition of three new functions/analyses. The version was changed to 2.0 rather than 1.2 to represent the significant changes (additions) to functionality that this update represents. The first new function is that STORM can now estimate allele frequencies for you (based on your genotypes) and format them appropriately for subsequent analyses. This change should reduce some headaches for the users. The second new function allows users to test hypotheses of the relatedness of individuals within groups *relative to one member of the group.* Instead of comparing all pair-wise relatedness values within a group (which is still an option). This function compares the relatedness of all individuals in a group to one specific individual within each group (such as the relatedness of males relative to a focal female in a mating group). A simulation function was added as well to test hypotheses regarding this scenario. The last added function is a means to test a new hypothesis of allele inheritance pattern; specifically one where fetal loss occurs due to similarity of the offspring genotype to that of mom. Further details of these new functions can be found below.

2 Legal Stuff

I tested STORM extensively throughout its development, and subsequently tested it using several data sets upon completion and updating. It appears to be functioning properly, and I am currently using it to analyze data from projects that I am working on. However, I cannot guarantee that it does not contain any errors, or that its results will always be reliable.

3 Disclaimer

I have tried to include a copious amount of notes in the code so that population geneticists interested in writing their own programs could hopefully make their way through it and figure out exactly what is going on. However, error messages to the user are sparse. I have included a few, but the tendency is for the program to just kick you out when an error occurs. With LINUX or MAC operating systems STORM will usually give you a brief error message, but with Windows it will just close the program. So, if you get kicked out, chances are that there was a problem with the data you tried to enter. Note that with Windows it will kick you out when it is done running, but your outfile should appear shortly, indicating that it ran successfully. If an outfile is not generated, then there has been an error with the program.

Please let me know if you have any questions, comments, or suggestions.

4 Installation

STORM can be run on LINUX, MAC, or Windows operating systems. With LINUX and MAC systems, you should run the program from the command line, even though it may appear to be working properly when you click on the icon. For some reason (I don't know why yet) it does not work properly when you click on the icon in these systems. However, clicking on the icon works fine with Windows. For all three systems you can download and install an executable file (and associated files). You can also download the source code and compile it yourself on any system.

Rather than including all installation instructions (for all systems) here, I have instead included a README text file within each zipped package that you download. Therefore, you should read the README file to see how to install the program on your operating system.

5 Infiles

STORM is essentially several separate programs combined together into one main menu. This design made it easier to develop, and also makes it easier to trouble-shoot problems. However, it is a bit inefficient, and results in the user having to enter the information about their data each time they want to conduct an analysis in STORM. So, you will have to enter your infile information for each analysis that you perform, and the required infiles are different for each analysis. Of course, your results will be completely incorrect (and you may not know it!) if your infiles are not prepared correctly. All infiles should be in tab-delimited text format, and all file names (both infile and outfile, and including extensions) must be 40 characters or less.

5.1 Allele Frequency File

STORM can estimate allele frequencies for you, and format them appropriately for subsequent analyses. Or, you can generate the necessary file yourself. The allele frequency file will have one column for each locus, with each column containing the allele frequencies for that locus. The loci MUST be in the same order as they occur in your genotypes. For example, if your alleles are ordered as 1, 2, 3, etc. in your genotypes, then the first allele frequency listed for that locus must be the frequency of allele 1, the second will be the frequency of allele 2, and so on. A zero (0) needs to be present at the top of each column, and you cannot include locus names. ALL columns must have values for the same number of rows. This means that the number of rows will equal the number of alleles in your most polymorphic locus. You MUST fill these rows with integers > 1 as place holders for those loci with fewer alleles. As example frequency file is shown below for four (4) loci. The first locus has two alleles, the second has four alleles, the third has five alleles, and the fourth has three alleles. In this case, 5 was used as the placeholder.

0	0	0	0
0.2	0.1	0.1	0.4
0.8	0.2	0.3	0.1
5	0.5	0.15	0.5
5	0.2	0.2	5
5	5	0.25	5

5.2 Genotype File

Each analysis requires that you enter your genotype file(s). This is the basic format that you will typically have your data in anyway, with a column of individual IDs, followed by two columns for each locus you used (one column for each allele). Individual IDs MUST be numeric (they cannot contain text!). For all analyses except those under the internal relatedness heading (IR), your alleles can be in base pairs. For the IR analyses, your alleles will have to be ordered as 1, 2, 3, etc. within each locus. For example, your smallest (in base pairs) allele must be coded as 1, the next allele coded as 2, and so on. You can do all analyses with the alleles in this ordered format, but the IR analyses are the only ones that *require* that your data are in this format. I have made a simple MicroSoft Excel conversion spreadsheet that you can download from my website that you can paste your data into, and it will convert your genotypes into this ordered format (double-check the conversion though!). Examples are below for both formats. You cannot include a row of locus names in your genotype file, and ALL missing data MUST be recorded as zero (0). Examples are below for five individuals genotyped at three loci.

5.2.1 For all analyses except IR

1001	120	122	203	203	98	100
1002	120	120	0	0	100	100
1003	118	122	201	203	100	102
1004	118	118	201	201	98	100
1005	120	120	203	205	98	104

5.2.2 Same Data Coded for IR Analysis (this format can be used for all analyses)

1001	2	3	2	2	1	2
1002	2	2	0	0	2	2
1003	1	3	1	2	2	3
1004	1	1	1	1	1	2
1005	2	2	2	3	1	4

5.3 Group File

For analyses of relatedness within groups, you will need to supply a file that is one column containing the number of individuals within each group. Therefore, if you have two groups, the file will contain two numbers. If you have three groups, the file will have three numbers, and so on. The trick is that THESE NUMBERS NEED TO BE CUMULATIVE. For example, suppose that you are analyzing three groups; the first group has 4 individuals, the second group has 3 individuals, and the third group has 4 individuals. Your group file would then look like this:

4
7

11

6 Conducting Analyses

6.1 Internal Relatedness (IR) of Individuals

6.1.1 What You Will Need

1. The number of loci you used
2. The number of alleles at your most polymorphic locus
3. Your allele frequency file
4. The number of individuals in your individual/genotype file
5. Your individual/genotype file
   (a) This file should contain the IDs and genotypes of the individuals for which you want to calculate IR. Note that in this file alleles must be coded as 1, 2, 3, etc. within each locus as described above.

6.1.2 What the Program Does

This analysis calculates the internal relatedness (IR) for each individual in your individual/genotype file. The approach used is that described in Amos et al. (2001). The equation is:

$$\frac{2H-\Sigma f_i}{2N-\Sigma f_i}$$

where:
H is the number of loci at which the reference individual is homozygous;
N is the number of loci at which the reference individual is genotyped; and
f_i is the frequency of the ith allele in the genotype.

As an example, we'll use the data below. These are also the name example infiles, so you can double-check the program by making sure that you get these same results (there will be minor differences due to rounding). Note that the loci with missing data are excluded from the analysis.

Allele Frequency File ("frequencies"):

0	0
0.1	0.2
0.3	0.4
0.5	0.4
0.1	5

Genotype File ("irgenotypes"):

1	1	1	1	1
2	2	0	3	2
3	1	2	2	2
4	3	3	0	2
5	0	1	2	3
6	1	4	3	0
7	2	2	2	2
8	4	4	1	2
9	2	4	2	3
10	1	1	3	3

Results:

Individual	$2H$	$2N$	Σf_i	IR
1	4	4	0.6	1.0
2	0	2	0.8	-0.667
3	2	4	1.2	0.286
4	2	2	1.0	1.0
5	0	2	0.8	-0.667
6	0	2	0.2	-0.111
7	4	4	1.4	1.0
8	2	4	0.8	0.375
9	0	4	1.2	-0.429
10	4	4	1.0	1.0
Average				**0.279**

6.2 Internal Relatedness (IR) of Simulated Offspring

6.2.1 What You Will Need

1. The number of alleles in your most polymorphic locus
2. The number of loci you used
3. Your allele frequency file
4. The number of males in your "reproductive males" file
5. Your "reproductive males" file
   (a) This is a file containing the IDs and genotypes (formatted with the alleles ordered as 1, 2, 3, etc. as described above) for the males that you want to use as your "male gene pool" for this analysis.
6. The number of females in your "reproductive females" file.
7. Your "reproductive females" file
   (a) This is a file containing the IDs and genotypes (formatted with the alleles ordered as 1, 2, 3, etc. as described above) for the females that you want to use as your "female gene pool" for this analysis.
8. How many iterations you want to perform
9. How many offspring you want to generate in each iteration

6.2.2 What the Program Does

This analysis generates the IR values expected from your gene pool if mating is random with respect to parental relatedness. It does this by sampling males and females from their respective gene pools (with replacement), generating offspring based on Mendelian inheritance, and calculating the average IR values for those generated offspring. The results can then be imported into Excel, or your favourite spreadsheet software, and used to generate your expected distribution of IR values, and to estimate your *p*-value.

As an example, suppose that you have parent-offspring data for 50 offspring. You calculate the IR value for those offspring, and get an average IR value of -0.0425. This seems surprisingly low to you, so you want to compare it to expectations if mating and/or fertilization is random with respect to parental relatedness. You now make your respective male and female gene pool files (based on the identified parents of the offspring) and generate 1,000 iterations of 50 offspring each, so that each iteration is one realization of the expected IR value for *your* data set. The resulting file is the average IR for each iteration. You put this into Excel, and find that only six of the iterations have an average IR value lower than your observed value of -0.0425. The interpretation is that the average IR value for the observed offspring is significantly lower than expected from this gene pool, and the *p*-value is < 0.007 (e.g. you observed fewer than 7 values equal to, or less than, your observed value out of 1,000 iterations). Now your project just became more complicated (and interesting!): is this pattern due to pre- or post-copulatory mate choice, inbreeding avoidance, mate incompatibility, or errors in your data set!!??
Example infiles for this analysis are supplied with the program, but it is difficult to present an example scenario here. The example files are:

1. *frequencies* - the allele frequency file. It is the same one as used for the IR analysis. There are 2 loci, with 4 alleles in the most variable locus.

2. *fathers* - a file containing the genotypes of 6 males, typed at 2 loci.

3. *mothers* - a file containing the genotypes of 6 females typed at 2 loci.

Note that you will get a result of “nan” (in LINUX) or “-1.#INDO0” (in Windows) in those cases where IR could not be calculated due to missing genotypes in the parents (i.e. if one or both parents are missing data at each locus so that an offspring genotype cannot be generated at any loci). Of course, this should be a problem in a real data set, but it does occur with this very small example data set.

6.3 Homozygosity By Loci (HL) of Individuals

6.3.1 What You Will Need

1. The number of loci you used
2. The number of alleles at your most polymorphic locus
3. Your allele frequency file
4. The number of individuals in your individual/genotype file
5. Your individual/genotype file
   (a) A file containing the IDs and genotypes of the individuals for which you want to calculate HL. Note that in this file alleles can be in base pairs or ordered.

6.3.2 What the Program Does

This analysis calculates the homozygosity by loci (HL) for each individual in your individual/genotype file. The approach is that described in Aparicio et al. (2006). The equation is:

$$\frac{\Sigma E_h}{\Sigma E_h + \Sigma E_j}$$

where:
E_h is the expected heterozygosity for the loci at which the individual is homozygous; and
E_j is the expected heterozygosity for the loci at which the individual is heterozygous.

As an example, we'll use the data below. These are the same as for the IR analysis, and are also the example infiles, so you can double-check the program by making sure that you get these same results (there will be minor differences due to rounding). Note that the loci with missing data are excluded from the analysis.

Allele Frequency File ("frequencies"):

0	0
0.1	0.2
0.3	0.4
0.5	0.4
0.1	5

Genotype File ("irgenotypes"):

1	1	1	1	1
2	2	0	3	2
3	1	2	2	2
4	3	3	0	2
5	0	1	2	3
6	1	4	3	0
7	2	2	2	2
8	4	4	1	2
9	2	4	2	3
10	1	1	3	3

Expect heterozygosity (H_E) for locus 1 = 1 - ($0.1^2 + 0.3^2 + 0.5^2 + 0.1^2$) = **0.64**.
Expect heterozygosity (H_E) for locus 2 = 1 - ($0.2^2 + 0.4^2 + 0.4^2$) = **0.64**.

Results:

Individual	ΣE_h	$\Sigma E_h + \Sigma E_j$	HL
1	1.28	1.28	1.00
2	0	0.64	0.00
3	0.64	1.28	0.50
4	0.64	0.64	1.00
5	0	0.64	0.00
6	0	0.64	0.00
7	1.28	1.28	1.00
8	0.64	1.28	0.50
9	0	1.28	0.00
10	1.28	1.28	1.00
Average			**0.500**

6.4 Homozygosity By Loci (HL) of Simulated Offspring

6.4.1 What You Will Need

1. The number of alleles in your most polymorphic locus
2. The number of loci you used
3. Your allele frequency file
4. The number of males in your "reproductive males" file
5. Your "reproductive males" file
   (a) This is a file containing the IDs and genotypes (formatted with the alleles ordered as 1, 2, 3, etc. as described above) for the males that you want to use as your "male gene pool" for this analysis.
6. The number of females in your "reproductive females" file.
7. Your "reproductive females" file
   (a) This is a file containing the IDs and genotypes (formatted with the alleles ordered as 1, 2, 3, etc. as described above) for the females that you want to use as your "female gene pool" for this analysis.
8. How many iterations you want to perform
9. How many offspring you want to generate in each iteration

6.4.2 What the Program Does

This analysis generates the HL values expected from your gene pool if mating and/or fertilization patterns are random with respect to homozygosity. It does this by sampling males and females from their respective gene pools (with replacement), generating offspring based on Mendelian inheritance, and calculating the average HL for those generated offspring. The results can then be imported into Excel, or your favourite spreadsheet software, and used to generate the expected distribution of HL values and to estimate *p*-values.

As an example, suppose that you have parent-offspring data for 50 offspring. You calculate the HL value for those offspring, and get an average HL value of 0.123. This seems surprisingly low to you, so you want to compare it to expectations if mating and/or fertilization patterns are random with respect to homozygosity. You now make your respective male and female gene pool files (based on the identified parents of the offspring) and generate 1,000 iterations of 50 offspring each, so that each iteration is one realization of the expected HL value for *your* data set. The resulting file is the average HL for each iteration. You put this into Excel, and find that only two of the iterations have an average HL lower than your observed value of 0.123/ The interpretation is that the average HL value for the observed offspring is significantly lower than expected from this gene pool, and the *p*-value is < 0.003 (you observed fewer than three values equal to or lower than your observed value, out of 1,000 iterations). As with the interpretation of the IR data described above, this pattern may be due to several factors such as pre- or post-copulatory inbreeding avoidance, mate incompatibility, differential survival of offspring, or errors in your data set (among other things).

Example infiles for this analyses are supplied with the program, but it is difficult to present an example scenario here. So you can just use those files to ensure that the program is running correctly. The example files are:

1. *frequencies* - the allele frequency file. It is the same one as used for the IR analysis. There are 2 loci, with 4 alleles in the most variable locus.

2. *fathers* - a file containing the genotypes of 6 males, typed at 2 loci.

3. *mothers* - a file containing the genotypes of 6 females typed at 2 loci.

Note that you will get a result of “nan” (in LINUX) or “-1.#INDO0” (in Windows) in those cases where IR could not be calculated due to missing genotypes in the parents (i.e. if one or both parents are missing data at each locus so that an offspring genotype cannot be generated at any loci). Of course, this should be a problem in a real data set, but it does occur with this very small example data set.

6.5 Pairwise Relatedness

6.5.1 What You Will Need

1. The number of loci you used
2. The number of alleles in your most polymorphic locus
3. Your allele frequency file
4. The number of individuals in your individual/genotype file
5. Your individual/genotype file (formatted with alleles in base pairs or ordered)

6.5.2 What the Program Does

This analysis calculates the pairwise relatedness values for a list of individuals. Therefore, if your infile contains data for 5 individuals, the output will be $[n(n-1)]/2 = 10$ pairwise comparisons. The calculation is based on the method described in Li et al. (1993), with each locus weighted using the method described in Lynch & Ritland (1999) and Van de Casteele et al. (2001). This method was chosen out of the many available approaches for calculating relatedness because it is unbiased, it is never undefined, and it consistently performs well in a variety of situations (and often out-performs all other estimators) (Van de Casteele et al. 2001; Wang 2002; Krützen et al. 2003). The equation for relatedness at each locus (l) is:

$$r_{xy}(l) = \frac{S_{xy}-S_o}{1-S_o}$$

Where $S_{xy} = 1$ when genotype $x = ii$ and genotype $y = ii$, or when $x = ij$ and $y = ij$. $S_{xy} = 0.75$ when $x = ii$ and $y = ij$ or vice versa. $S_{xy} = 0.5$ when $x = ij$ and $y = ik$. $S_{xy} = 0$ when $x = ij$ and $y = kl$, where i, j, k, and l represent different alleles.

$$S_o = \sum_{i=1}^{n} p_i^2(2-p_i)$$

Where p_i is the population allele frequency for allele i. Multilocus relatedness values can be obtained by multiplying r_{xy} for each locus by the weight for that locus, summing this value across loci, and then dividing by the sum of all weights used. The weight for each locus is calculated based on the number of alleles in each locus.

$$w(l) = \frac{n_j-1}{\sum n_j-1}$$

Where n_j is the number of alleles at locus j.

So, the total relatedness is:

$$r_{xy} = \frac{1}{W}\sum w(l)r_{xy}(l)$$

Where W is the sum of the weights for all loci used.

Note that in theory (and in the example below) weights only need to be calculated for each locus, and these weights will be the same across all pairwise comparisons (because they only depend on the characteristics of the loci being used). However, this is not the case with real data, where some individuals may be missing data at some loci. Therefore, STORM calculates different weights for each pairwise comparison that are based only on the loci used for that comparison, rather than all loci in the data set.

Let's walk through an example using the example infiles. We will use the file “frequencies” from above, and the file “pairwise”, with a list of 5 individuals genotyped at 2 loci.

“frequencies”

```
0    0
0.1  0.2
0.3  0.4
0.5  0.4
0.1  5
```

“pairwise”

```
1  1  2  2  2
2  2  2  2  3
3  3  3  1  1
4  1  2  2  3
5  1  3  2  2
```

S_o for locus 1 = 0.19 + 0.153 + 0.375 + 0.019 = 0.566
S_o for locus 2 = 0.072 + 0.256 + 0.256 = 0.584

Weight for locus 1 = 3/6 = 0.5
Weight for locus 2 = 2/6 = 0.333

Total weights (W) = 0.833

Results:

Pair	S_{xy} locus 1	S_{xy} locus 2	r_{xy} locus 1	r_{xy} locus 2	r-value
1 & 2	0.75	0.75	0.424	0.399	0.414
1 & 3	0	0	-1.30	-1.40	-1.34
1 & 4	1	0.75	1	0.399	0.760
1 & 5	0.5	1	-0.152	1	0.309
2 & 3	0	0	-1.30	-1.40	-1.34
2 & 4	0.75	1	0.424	1	0.654
2 & 5	0	0.75	-1.30	0.399	-0.621
3 & 4	0	0	-1.30	-1.40	-1.34
3 & 5	0.75	0	0.424	-1.40	-0.305
4 & 5	0.5	0.75	-0.152	0.399	0.0683

Note that you will get a result of “nan” (in LINUX) or “-1.#INDO0” (in Windows) in those cases where IR could not be calculated due to missing genotypes in the parents (i.e. if one or both parents are missing data at each locus so that an offspring genotype cannot be generated at any loci). Of course, this should be a problem in a real data set, but it does occur with this very small example data set.

6.6 Calculate the Relatedness of Mating Pairs

6.6.1 What You Will Need

1. The number of loci you used
2. The number of alleles at your most variable locus
3. Your allele frequency file
4. The number of individuals in your individual/genotype file
5. You individual/genotype file (formatted with alleles in base pairs, or ordered)

This analysis differs from the “Pairwise Relatedness” option because this analysis does not calculate ALL pairwise relatedness values, but just the relatedness values for the pairs of individuals that you supply. Thus, your genotype file must contain the individuals arranged as the specific pairs for which you want to calculate relatedness. For example, suppose that you have 5 mating pairs with males given odd numbers (1, 3, 5, 7, 9), and females given even numbers (2, 4, 6, 8, 10). Your mating pairs have been identified as 1 & 2, 3 & 4, 5 & 6, 7 & 8, and 9 & 10. Your genotype file would then look like the example below (using the ordered allele format), with the individuals of each mating pairs occurring together (sequentially) in the file, and your results would be 5 relatedness values (one for each pair).

1	1	1	1	1
2	2	2	2	2
3	1	3	2	2
4	3	3	1	3
5	2	3	2	3
6	1	3	2	3
7	2	2	1	1
8	1	2	1	1
9	3	3	1	3
10	2	2	1	3

6.6.2 What the Program Does

The program calculates the relatedness of each supplied mating pair using the calculations described in the “Calculating Pairwise Relatedness” section. A full work-through will not be given again here, and that section should be referred to for details. The example files supplied are the “frequencies” file used in all other examples (containing data for 2 loci, with 4 alleles at the most polymorphic locus), and a file containing the five mating pairs above (“mpairs”). Please make sure that you get results similar to those calculated below.

“frequencies”	
0	0
0.1	0.2
0.3	0.4
0.5	0.4
0.1	5

“mpairs”				
1	1	1	1	1
2	2	2	2	2
3	1	3	2	2
4	3	3	1	3
5	2	3	2	3
6	1	3	2	3
7	2	2	1	1
8	1	2	1	1
9	3	3	1	3
10	2	2	1	3

The relatedness values would then be:

Pair 1	-1.34
Pair 2	-0.305
Pair 3	0.309
Pair 4	0.654
Pair 5	-0.381
Average	**-0.213**

Note that you will get a result of “nan” (in LINUX) or “-1.#INDO0” (in Windows) in those cases where IR could not be calculated due to missing genotypes in the parents (i.e. if one or both parents are missing data at each locus so that an offspring genotype cannot be generated at any loci). Of course, this should be a problem in a real data set, but it does occur with this very small example data set.

6.7 Calculate the Relatedness of Randomly Generated Mating Pairs

6.7.1 What You Will Need

1. The number of loci you used
2. The number of alleles in your most variable locus
3. Your allele frequency file
4. The number of males in your "reproductive males" file
5. Your "reproductive males" file. This is a file containing the IDs and genotypes of the males that you want to use as your male gene pool for this analysis.
6. The number of females in your "reproductive females" file
7. Your "reproductive females" file. This is a file containing the IDs and genotypes of the females that you want to use as your female gene pool for this analysis.
8. How many iterations you want to perform
9. How many mating pairs you want to generate in each iteration

6.7.2 What the Program Does

This analysis creates the distribution of expected relatedness values of mating pairs in your data set if mating is random with respect to relatedness. It does this by sampling your male and female gene pools (with replacement) to generate the given number of mating pairs. The relatedness of these mating pairs is then calculated, along with the average relatedness among all pairs within each iteration. This list of average relatedness values (one for each iteration) can be used to create a distribution of expected relatedness values, and to estimate p-values.

As an example, suppose that you have genotype data for 50 mating pairs. You calculate the relatedness of those mating pairs, and get an average r-value of -0.0121. You suspect that individuals are avoiding mating with close relatives, which is resulting in this low r-value, so you want to compare it to expectations if mating is random with respect to relatedness. You now make your respective male and female gene pool files (based on the identified pairs) and generate 1,000 iterations of 50 mating pairs each, so that each iteration is one realization of the expected r-value for *your* data set. The resulting file is the average r-value for each iteration. You find that 15 of the iterations have an average r-value equal to or lower than your observed value of -0.0121. The interpretation is that the average r-value of the observed mating pairs is significantly lower than expected from this gene pool, and the p-value is < 0.016 (e.g. you observed fewer than 16 values equal to or lower than your observed value, out of 1,000 iterations). This result suggests that mating is NOT random with respect to relatedness in your population, and that identified mating pairs are less related than expected.

6.8 Calculate Within-Group Relatedness (all vs all)

6.8.1 What You Will Need

1. The number of loci you used
2. The number of alleles at your most variable locus
3. Your allele frequency file
4. How many individuals you have in your genotype file
5. How many pairwise comparisons you have in total.
   This is figured out by calculating the number of pairwise comparisons for each of your groups, and then summing across groups. The calculations for the number of pairwise comparisons is $[n(n-1)/2$. For the example infiles (see below) there are 11 individuals in the genotype file, and there are three groups. The first group has 4 individuals, the second group has 3 individuals, and the third group ahs 4 individuals. The total number of pairwise comparisons is then:

   $$[4(4-1)]/2 + [3(3-1)]/2 + [4(4-1)]/2 = 15$$

6. Your genotype file.
   This file should contain individuals IDs and genotypes for each individual. The order of individuals in the genotype file should represent their organization in groups. For example, if you have 4 individuals in the first group, these should be the first 4 genotypes in the genotype file; if you have 3 individuals in the second group, then these should be the next three genotypes in your genotype file...and so on.
7. How many groups you have
8. Your group File
   This is the file that contains a list of the number of individuals in each group, and THESE NUMBERS ARE CUMULATIVE. For example, if you have 3 groups, with 4 individuals in the first group, 3 individuals in the second group, and 4 individuals in the third group, then your group file should look like this:

   ```
   4
   7
   11
   ```

6.8.2 What the Program Does

The program calculates the average pairwise relatedness values within each group, and averages that across all groups. Relatedness is calculated as described in the "Calculating Pairwise Relatedness" section, which should be referred to for details. As an example, we will work through the example files. You should run the program with the example files to ensure that you get the same result. In this example we have 11 individuals, 4 in group #1, 3 in group #2, and 4 in group #3.

Example Files

"frequencies" – allele frequency data for 2 loci with 4 alleles at the most polymorphic locus.

```
0     0
0.1   0.2
0.3   0.4
0.5   0.4
0.1   5
```

“group.txt” – a file containing the genotypes of 11 individuals. Individuals 1001, 1002, 1003, and 1004 make up group #1; individuals 1005, 1006, and 1007 make up group #2; and individuals 1008, 1009, 1010, and 1011 make up group #3.

```
1001  1  1  1  2
1002  2  2  2  2
1003  1  2  3  3
1004  2  0  1  3
1005  3  4  2  2
1006  1  2  1  1
1007  1  1  3  3
1008  1  1  1  1
1009  1  1  1  2
1010  1  2  1  1
1011  1  2  1  2
```

“infile” – the group file for these data. Because we have 3 groups, with 4 individuals in the first group, 3 individuals in the second group, and 4 individuals in the third group, this file looks like this:

```
4
7
11
```

The results should then be:

Pairs	r-value	Average within-group r-values
1001 & 1002	-0.623	
1001 & 1003	0.174	
1001 & 1004	-0.202	
1002 & 1003	0.414	
1002 & 1004	-1.40	
1003 & 1004	-0.202	Group #1 **-0.306**
1005 & 1006	-1.34	
1005 & 1007	-1.34	
1006 & 1007	-0.307	Group #2 **-0.995**
1008 & 1009	0.760	
1008 & 1010	0.654	
1008 & 1011	0.414	
1009 & 1010	0.414	
1009 & 1011	0.654	
1010 & 1011	0.760	Group #3 **0.609**
Total Average		**-0.231**

Note that your results file will not show the data for all pairwise comparisons, but instead just contains the average within-group relatedness values for each group, as well as for the overall average within-group r-value.

Note that you will get a result of “nan” (in LINUX) or “-1.#INDO0” (in Windows) in those cases where IR could not be calculated due to missing genotypes in the parents (i.e. if one or both parents are missing data at each locus so that an offspring genotype cannot be generated at any loci). Of course, this should be a problem in a real data set, but it does occur with this very small example data set.

6.9 Calculate Within-Group Relatedness with Individuals Shuffled Between Groups)

6.9.1 What You Will Need

1. The number of loci you used
2. The number of alleles at your most variable locus
3. Your allele frequency file
4. How many individuals you have in your genotype file
5. How many pairwise comparisons you have in total.
   This is figured out by calculating the number of pairwise comparisons for each of your groups, and then summing across groups. The calculations for the number of pairwise comparisons is $[n(n-1)/2$. For the example infiles (see below) there are 11 individuals in the genotype file, and there are three groups. The first group has 4 individuals, the second group has 3 individuals, and the third group ahs 4 individuals. The total number of pairwise comparisons is then:

   $$[4(4-1)]/2 + [3(3-1)]/2 + [4(4-1)]/2 = 15$$

6. Your genotype file.
   This file should contain individuals IDs and genotypes for each individual. The order of individuals in the genotype file should represent their organization in groups. For example, if you have 4 individuals in the first group, these should be the first 4 genotypes in the genotype file; if you have 3 individuals in the second group, then these should be the next three genotypes in your genotype file...and so on.
7. How many groups you have
8. Your group File
   This is the file that contains a list of the number of individuals in each group, and THESE NUMBERS ARE CUMULATIVE. For example, if you have 3 groups, with 4 individuals in the first group, 3 individuals in the second group, and 4 individuals in the third group, then your group file should look like this:

   ```
   4
   7
   11
   ```

9. How many iterations you want to perform.

6.9.2 What the Program Does

This analysis creates the distribution of expected within-group average relatedness values if group composition is random with respect to relatedness. It does this by shuffling individuals between groups, while keeping each group size constant, and calculating the average within-group relatedness values (one for each iteration). These can be used to create a distribution of expected relatedness values, and to estimate p-values.

As an example, suppose that you have genotype data for 100 individuals, organized as 10 groups with each group containing 10 individuals. You calculate the observed within-group relatedness and get an average r-value of 0.565. You suspect that the social structure of this population is based on matrilines, and therefore that the observed groups represent mothers and their offspring, grand-offspring, etc., which is resulting in this high r-value. Now you want to compare this observed value to what would be expected if the groups represent random associations of individuals with respect to relatedness. You now generate 1,000 iterations, each of 10 groups with 10 individuals each, so that each iteration is one realization of the expected r-value for *your* data set. The resulting file is the average r-value for each group, as well as the total average, for each iteration. You put this into Excel, and find that 1 of the iterations had an r-value as high or higher than your observed value of 0.565. The interpretation is that the observed average within-group r-value is significantly higher than expected in this population if groups represent random associations of individuals within respect to relatedness, and the p-value is < 0.002 (you observed fewer than two values as high, or higher, than your observed value out of 1,000 iterations).

You can run this analysis with the example files below to ensure that it is working properly.

"frequencies" – the allele frequency file. There are 2 loci, with 4 alleles in the most polymorphic locus.

"group.txt' – a file containing the genotypes of 11 individuals genotyped at 2 loci. They are organized into 3 groups, with 4 individuals in the first group, 3 individuals in the second group, and 4 individuals in the third group (for a total of 15 pairwise comparisons in total).

"infile" - the group file for these data.

The results will be a tab-delimited text file with the average within-group relatedness for each group, as well as across all groups, for each iteration. For example, with the data above, the out file might look like this:

Iteration	**Group 1**	**Group 2**	**Group 3**	**Average**
1	-0.0067	0.023	0.001	0.0058
2	0.013	-0.034	0.067	0.0153
etc.				

Note that this approach allows you to simultaneously test both global hypotheses (i.e. is realness within groups higher/lower than expected overall?), as well as hypotheses regarding each group (e.g. are individuals in Group 1 more/less related than expected?).

6.10 Calculate Within-Group Relatedness (relative to one group member)

Sometimes, it may be desirable to compare the relatedness of all individuals within a group *relative to one particular member* rather than across all individuals. Some such scenarios may include if there is a dominant or "focal" individual within each group. This new version of STORM provides a means to do this. The approach is similar to that described above, with a few modifications. The main one being that you now need two genotype files: one containing the genotypes of the "focal" individuals (1 for each group); and another containing the genotypes for the rest of the individuals included in the analyses. Note that these genotypes must be in the appropriate order, meaning that the first individual in the "reference" genotypes is the focal individual for group #1, and the first individuals in the "regular" genotype file represent the other individuals in group #1.

6.10.1 What You Will Need

1. The number of alleles at your most variable locus
2. The number of loci you used
3. Your allele frequency file
4. The number of individuals in your "regular" genotype file.
   These are the individuals and genotypes that you want to compare to the focal individuals.
5. Your "regular" genotype file.
   This is a file containing the ID and genotypes for the individuals that you want to compare to the focal individuals. The first individuals should correspond to the individuals in group #1, and so on. You will specify how many individuals are in each group later.
6. The total number of groups that you have.
7. Your group File
   This is the file that contains a list of the number of **non-focal** individuals in each group, and THESE NUMBERS ARE CUMULATIVE. For example, if you have 3 groups, with 3 individuals in the first group, 2 individuals in the second group, and 3 individuals in the third group, then your group file should look like this:
   3
   5
   8
8. Your focal individual genotype file.
   This is a file that contains the IDs and genotypes of the focal individuals (one individual per group).

6.10.2 What the Program Does

Your "group" file will tell STORM how many non-focal individuals are in each group. The program will calculate the average relatedness of everyone within each group *relative to the focal individual.* The output will be one row for each group, with each row containing the average within-group relatedness value relative to the focal individual. Relatedness is calculated as described in the "Calculating Pairwise Relatedness" section, which should be referred to for details. As an example, we will work through the example files. You should run the program with the example files to ensure that you get the same result. In this example we have 3 groups, with one focal individual in each group and 3 non-focal individuals in group #1, 2 non-focal individuals in group #2, and 3 non-focal individuals in group #3, for a total of 8 non-focal individuals.

Example Files

"frequencies" – allele frequency data for 2 loci with 4 alleles at the most polymorphic locus.

```
0    0
0.1  0.2
0.3  0.4
0.5  0.4
0.1  5
```

"non-focal.txt" – a file containing the genotypes of the 8 non-focal individuals. Individuals 1002, 1003, and 1004 make up group #1; individuals 1006 and 1007 make up group #2; and individuals 1009, 1010, and 1011 make up group #3.

```
1002  2  2  2  2
1003  1  2  3  3
1004  2  0  1  3
1006  1  2  1  1
1007  1  1  3  3
1009  1  1  1  2
1010  1  2  1  1
1011  1  2  1  2
```

"group" – the group file for these data. Because we have 3 groups, with 3 non-focal individuals in the first group, 2 non-focal individuals in the second group, and 3 non-focal individuals in the third group, this file looks like this:

```
3
5
8
```

"focal.txt" – a file containing the genotypes of the 3 focal individuals. Individual 1001 is the focal individual for group #1; individual 1005 is the focal individual for group #2; and individual 1008 is the focal individual for group #3.

```
1001  1  1  1  2
1005  3  4  2  2
1008  1  1  1  1
```

The results should then be:

Pairs	r-value	Average within-group r-values
1001 & 1002	-0.623	
1001 & 1003	0.174	
1001 & 1004	-0.202	Group #1 **-0.217**
1005 & 1006	-1.34	
1005 & 1007	-1.34	Group #2 **-1.34**
1008 & 1009	0.760	
1008 & 1010	0.654	
1008 & 1011	0.414	Group #3 **0.609**
Total Average		**-0.317**

Note that your results file will not show the data for all pairwise comparisons, but instead just contains the average within-group relatedness values for each group, as well as for the overall average within-group r-value.

Note that you will get a result of "nan" (in LINUX) or "-1.#INDO0" (in Windows) in those cases where IR could not be calculated due to missing genotypes in the parents (i.e. if one or both parents are missing data at each locus so that an offspring genotype cannot be generated at any loci). Of course, this should be a problem in a real data set, but it does occur with this very small example data set.

6.11 Calculate Within-Group Relatedness (relative to one group member) Shuffling Individuals Between Groups

This program randomly shuffles non-focal individuals between groups, while keeping the number and size of groups constant, and calculates the average within-group relatedness of the non-focal individuals relative to the focal individual for each group. This function is useful for testing hypotheses about observed relatedness patterns of individuals within groups relative to a focal individual.

6.11.1 What You Will Need

1. The number of alleles at your most variable locus
2. The number of loci you used
3. Your allele frequency file
4. The number of individuals in your "regular" genotype file.
   These are the individuals and genotypes that you want to compare to the focal individuals.
5. Your "regular" genotype file.
   This is a file containing the ID and genotypes for the individuals that you want to compare to the focal individuals. The first individuals should correspond to the individuals in group #1, and so on. You will specify how many individuals are in each group later.
6. The total number of groups that you have.
7. Your group File
   This is the file that contains a list of the number of **non-focal** individuals in each group, and THESE NUMBERS ARE CUMULATIVE. For example, if you have 3 groups, with 3 individuals in the first group, 2 individuals in the second group, and 3 individuals in the third group, then your group file should look like this:
   3
   5
   8
8. The number of iterations you want to perform
9. Your focal individual genotype file.
   This is a file that contains the IDs and genotypes of the focal individuals (one individual per group).

6.11.2 What the Program Does

This analysis creates the distribution of expected within-group average relatedness values, relative to a focal individual in each group, if group composition is random with respect to relatedness. It does this by shuffling non-focal individuals between groups while keeping the focal individual, group size, and group number constant, and calculating the average within-group relatedness values (one for each iteration). These can be used to create a distribution of expected relatedness values, and to estimate p-values.

As an example, suppose that you are studying a species where multiple males try to copulate with single females during the mating season. You think that perhaps individuals are only mating with particularly genetically dissimilar mates. Suppose that you have data from 10 such "groups", with each group containing 1 female (the "focal" individual) and 9 males. You calculate the observed within-group relatedness relative to the focal female, and get an average r-value of -0.232. This suggests that you are correct, and that the males within these mating groups are particularly dissimilar from the females. Now you want to compare this observed value to what would be expected if the groups represent random associations of individuals with respect to relatedness. You now generate

1,000 iterations, each of 10 groups with 9 non-focal individuals each, so that each iteration is one realization of the expected r-value for *your* data set. The resulting file is the average r-value for each group, as well as the total average, for each iteration. You put this into Excel, and find that 3 of the iterations had an r-value as low or lower than your observed value of -0.232. The interpretation is that the observed average within-group r-value is significantly lower than expected in this population if groups represent random associations of individuals within respect to relatedness to the focal individual, and the p-value is < 0.004 (you observed fewer than four values as low, or lower, than your observed value out of 1,000 iterations).

You can run this analysis with the example files below to ensure that it is working properly.

"frequencies" – the allele frequency file. There are 2 loci, with 4 alleles in the most polymorphic locus.

"non-focal.txt' – a file containing the genotypes of 8 non-focal individuals genotyped at 2 loci. They are organized into 3 groups, with 3 individuals in the first group, 2 individuals in the second group, and 3 individuals in the third group.

"group" - the group file for these data.

"focal.txt' – a file containing the genotypes of the 3 focal individuals genotyped at 2 loci.

The results will be a tab-delimited text file with the average within-group relatedness (relative to the focal individual) for each group, as well as across all groups, for each iteration. For example, with the data above, the out file might look like this:

Iteration	**Group 1**	**Group 2**	**Group 3**	**Average**
1	0.323895	-1.344027	-0.297191	-0.439108
2	-0.136934	-0.622873	0.368944	-0.130288
etc.				

Note that this approach allows you to simultaneously test both global hypotheses (i.e. is realness within groups higher/lower than expected overall?), as well as hypotheses regarding each group (e.g. are individuals in Group 1 more/less related than expected?).

6.12 An Explanation of Allele Inheritance

Prior to explaining exactly what the program is doing for the different analyses of allele inheritance, I thought it would be important to explain the differences between the two approaches used by STORM. Briefly, there are two primary hypotheses regarding biased allele inheritance patterns, both of which have substantial data behind them. Unfortunately, these two different processes are not always distinguished from one another in the literature, and they have also both been referred to as "mate incompatibility" or "genetic incompatibility". The first involves increased rates of fetal loss when offspring are too genetically similar to the mother, which results in a breakdown in self/non-self recognition and subsequent fetal abortion (e.g. Ober et al. 1998). The result is that *observed/surviving* offspring represent a biased sample of those fertilizations where the offspring inherited a paternal allele making the zygote dissimilar from the mother. The second mechanism involves increased fertilization/pregnancy success between genetically dissimilar gametes (Birkhead et al. 2004; Evans & Marshall 2005; Dziminski et al. 2008), which presumably results from selection for heterozygous offspring (Brown et al. 1997). For species with internal fertilization, it is generally not known if this results from pre-fertilization processes (e.g. cryptic sperm choice by females), or from differential mortality of zygotes (Olsson et al. 1999).

Although these two processes of genetic incompatibility have similarities, they result in different predictions for genetic signatures in surviving offspring (see **Figure 1**). STORM provides a means to test for both of these patterns within a given data set. The two mechanisms are titled "Divergent Alleles" and "Fetal Loss" to describe genetic compatibility based on heterozygosity and self/non-self recognition, respectively, throughout the rest of the documentation.

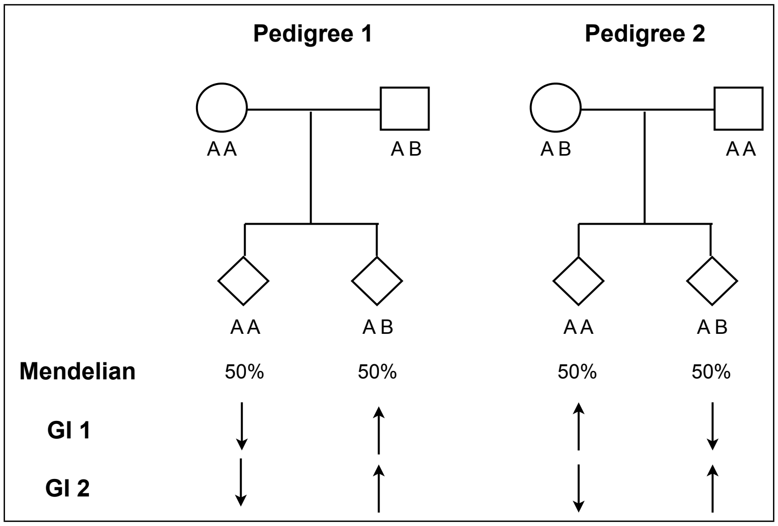


Figure 1: *Example pedigrees showing expectations in the genotypes of surviving offspring under hypotheses of Mendelian inheritance, genetic incompatibility based on self/non-self recognition (GI 1), and genetic incompatibility based on heterozygosity (GI 2). Arrows indicate either an increase or decrease relative to Mendelian expectations. Pedigree 1 represents a scenario where both processes result in the same expectations, and pedigree 2 represents a scenario where the expectations are different. Again, the mechanism of GI 1 results in offspring genetically dissimilar from the mother, whereas the mechanism of GI 2 results in heterozygous offspring, regardless of their similarity to the mother.*

6.13 Calculate Allele Inheritance of Observed Offspring: Divergent Alleles (AI_{DA})

6.13.1 What You Will Need

1. The number of alleles at your most variable locus
2. The number of loci you used
3. Your allele frequency file
4. How many offspring you have in your offspring file.
   This is the number of offspring for which you want to calculate allele inheritance.
5. Your offspring genotype file.
   This is a file containing the IDs and genotypes of the offspring for which you want to calculate allele inheritance.
6. How many pairs you have in your mating pairs file.
7. Your mating pairs file.
   This file should contain the IDs and genotypes of the mating pairs. The mating pairs should be in the same order as the offspring in the offspring file. So, the first mating pair in the mating pairs file should be the parents of the first offspring in the offspring file, and so on. In this file **the mother's genotype should be first, and the father's genotype second!!**. Thus, if you have three mating pairs genotyped at one locus, the format should be:

Mom#1	100	100
Dad#1	102	102
Mom#2	100	102
Dad#2	100	100
Mom#3	102	104
Dad#3	100	100

6.13.2 What the Program Does

This function calculates the proportion of the time that a paternal allele different from the maternal allele is inherited by the offspring. Just based on Mendelian inheritance, the expectation of this value is 50% or, 0.5. Results can range from zero to one. The trick for this calculation is that most loci will be uninformative. For example, all loci for which the father is homozygous are uninformative because there is not the "option" for offspring to inherit different alleles from the father. Similarly, loci are uninformative when the mother and father are heterozygous for for different alleles because in this case the offspring *must* inherit a paternal allele that is different than the maternal allele. All of the possible parental and resulting offspring allelic combinations are given below, along with how they are used in the analysis. In all cases the paternal allele in the offspring is listed first, and the maternal allele listed second.

Father	ii^1	ii^1	ij	ij	ij	ii^1	ij^2
Mother	ii	ij	ii	ij	ik	jj	kl
Offspring	ii	ii	ii^3	ii^3	ii^3	ij	ik
		ij	ji^4	ij^4	ik^5		il
				ji^4	ji^4		jk
				jj^3	jk^5		jl

[1] Indicates loci that are uninformative due to the father being homozygous.
[2] Indicates loci that are uninformative due to parents being heterozygous for different alleles.
[3] Indicates informative scenarios where offspring inherited a paternal allele that is the same as the maternal allele.

[4] Indicates informative scenarios where offspring inherited a paternal allele that is different than the maternal allele.
[5] Indicates scenarios that are uninformative due to the offspring inheriting a maternal allele not present in the father, thus there is not an "option" for the offspring to inherit paternal alleles that are the same as the maternal alleles.

Note that there are an equal number of informative scenarios resulting in the offspring inheriting a paternal allele the same as, or different, than the maternal allele; thus the expectations based solely on informative scenarios results in the typical expected Mendelian ratio of 0.5. Only informative scenarios are used in the calculation, and loci with any individual of the mother-offspring-father triad missing data are excluded from the calculation.

It is widely recognized that metrics of genetic characteristics are more informative (and more accurately estimate the value of interest) when they are weighted appropriately based on the characteristics of the loci used (e.g. Coltman et al. 1999; Amos et al. 2001; Aparicio et al. 2006). Although it may initially appear that the calculation of allele inheritance (AI_{DA}) itself is not directly dependent on the allele frequencies (because it only depends on which alleles are in the parents), the number of loci that are informative in a mother-father-offspring triad will be influenced by the variability of the loci used. Thus, because AI_{DA} is a combination of both the number of informative loci and the allele inheritance patterns at those loci, it makes sense to weight AI_{DA} values based on the characteristics of the loci used. Moreover, it also makes intuitive sense that AI_{DA} values with a strong signal from relatively uninformative loci should be weighted more than the same value from more informative loci. Weighting loci based on their expected heterozygosity (H_E, Nei 1978), has proven to be appropriate under many circumstances (Queller and Goodnight 1989; Aparicio et al. 2006). Here, AI_{DA} is weighted by the inverse of the average expected heterozygosity, which produces the desired weighting pattern of a given AI_{DA} value being weighted more for relatively uninformative loci than for more informative loci. This results in the total calculation for AI_{DA} being:

$$AI_{DA} = (L_p \ / \ L) \ / \ \{(\sum_{l=1}^{l} H_E \ / \ l\}$$

Where:

L_p is the number of informative loci where a paternal allele different than the maternal allele was inherited.

L is the number of informative loci.

$\sum H_E$ is the sum of expected heterozygosities for all typed loci.

l is the number of typed loci.

In words, the numerator is the proportion of informative loci for which a paternal allele different than the maternal allele was inherited (ranging from 0 to 1), and the denominator is the average heterozygosity for the typed loci.

Note that a result of "999" will be returned for those individuals for which there were not any informative scenarios, and thus allele inheritance could not be calculated. These individuals are then ignored when the program calculates averages.

As an example, let's walk through this calculation for the example in files that are provided. The example files included are:

“frequencies” – allele frequency data for 2 loci with 4 alleles at the most polymorphic locus.

```
0      0
0.1    0.2
0.3    0.4
0.5    0.4
0.1    5
```

“ogenotypes” – A file containing the genotypes of 10 offspring genotyped at 2 loci.

```
1    1   1   1   1
2    2   0   3   2
3    1   2   2   2
4    3   3   0   2
5    0   1   2   3
6    1   4   3   0
7    2   2   2   2
8    4   4   1   2
9    2   4   2   3
10   1   1   3   3
```

“parents” – A file containing the parent for these 10 offspring.

```
1    1   1   1   1   Mom for offspring #1
2    1   2   1   1   Dad for offspring #1
3    0   0   2   2   Mom for offspring #2
4    2   2   2   3   Dad for offspring #2
5    1   2   1   2   Mom for offspring #3
6    1   2   2   3   Dad for offspring #3
7    2   3   1   2   Mom for offspring #4
8    2   3   0   0   Dad for offspring #4
9    1   1   2   3   Mom for offspring #5
10   1   1   1   2   Dad for offspring #5
11   1   4   0   0   Mom for offspring #6
12   3   4   3   3   Dad for offspring #6
13   1   2   2   2   Mom for offspring #7
14   2   2   2   2   Dad for offspring #7
15   3   4   1   1   Mom for offspring #8
16   3   4   2   2   Dad for offspring #8
17   1   2   2   3   Mom for offspring #9
18   4   4   2   3   Dad for offspring #9
19   1   1   2   3   Mom for offspring #10
20   1   1   3   3   Dad for offspring #10
```

The calculation would then be:

Expected heterozygosity (H_E) for locus 1 = 1 - (0.1^2 + 0.3^2 + 0.5^2 + 0.1^2) = **0.64**
Expected heterozygosity (H_E) for locus 2 = 1 - (0.2^2 + 0.4^2 + 0.4^2) = **0.64**

Offspring	L	L_p	$\sum H_E$	l	AI_{DA}
1	1	0	1.28	2	0
2	1	1	0.64	1	1.56
3	2	1	1.28	2	0.781
4	1	0	0.64	1	0
5	0	NA	0.64	1	999
6	0	NA	0.64	1	999
7	0	NA	1.28	2	999
8	1	0	1.28	2	0
9	1	1	1.28	2	1.56
10	0	NA	1.28	2	999
Average					**0.650**

6.14 Calculate Allele Inheritance of Simulated Offspring: Divergent Alleles

6.14.1 What You Will Need

1. The number of alleles at your most variable locus

2. The number of loci you used

3. Your allele frequency file

4. How many pairs you have in your mating pairs file.

5. Your mating pairs file.
   This file should contain the IDs and genotypes of the mating pairs. The mating pairs should be in the same order as the offspring in the offspring file. So, the first mating pair in the mating pairs file should be the parents of the first offspring in the offspring file, and so on. In this file **the mother's genotype should be first, and the father's genotype second!!**. Thus, if you have three mating pairs genotyped at one locus, the format should be:
   ```
   Mom#1   100   100
   Dad#1   102   102
   Mom#2   100   102
   Dad#2   100   100
   Mom#3   102   104
   Dad#3   100   100
   ```

6. How many iterations you want to perform

6.14.2 What the Program Does

This function calculates the expected AI_{DA} values for your offspring under Mendelian expectations. It does this by keeping all mating pairs constant, but generating simulated offspring from these pairs based on Mendelian inheritance. The AI_{DA} is then calculated for these offspring, and an overall average AI_{DA} is calculated. If this process is repeated many times (e.g. 1,000), then the expected distribution of AI_{DA} values can be generated, which can then be used to interpret your data and estimate p-values.

As an example, suppose that you have 50 offspring with identified parents, and the average AI_{DA} for these offspring is 0.989. You suspect that fertilization are only successful between genetically dissimilar gametes. To test this hypothesis you perform 1,000 iterations, with each iteration generating 50 offspring from *these same* mating pairs. Thus, each iteration is one realization of the expected AI_{DA} value for *your* data under Mendelian expectations. Suppose that you do not observe any iterations with an average AI_{DA} value greater than your observed value of 0.989. The interpretation is that the observed offspring inherited paternal alleles that were different than the maternal alleles significantly more often than expected based solely on Mendelian inheritance, and the p-value is < 0.001 (you observed fewer than 1 value higher than your observed value, out of 1,000 iterations). It would appear that a large portion of eggs are fertilized by sperm that are particularly genetically dissimilar.

The example files provided for this analysis are:

"frequencies" − allele frequency data for 2 loci with 4 alleles at the most polymorphic locus.
```
0     0
0.1   0.2
0.3   0.4
0.5   0.4
0.1   5
```

“parents” – A file containing the parent for these 10 offspring.

1	1	1	1	1	Mom for offspring #1
2	1	2	1	1	Dad for offspring #1
3	0	0	2	2	Mom for offspring #2
4	2	2	2	3	Dad for offspring #2
5	1	2	1	2	Mom for offspring #3
6	1	2	2	3	Dad for offspring #3
7	2	3	1	2	Mom for offspring #4
8	2	3	0	0	Dad for offspring #4
9	1	1	2	3	Mom for offspring #5
10	1	1	1	2	Dad for offspring #5
11	1	4	0	0	Mom for offspring #6
12	3	4	3	3	Dad for offspring #6
13	1	2	2	2	Mom for offspring #7
14	2	2	2	2	Dad for offspring #7
15	3	4	1	1	Mom for offspring #8
16	3	4	2	2	Dad for offspring #8
17	1	2	2	3	Mom for offspring #9
18	4	4	2	3	Dad for offspring #9
19	1	1	2	3	Mom for offspring #10
20	1	1	3	3	Dad for offspring #10

6.15 Calculate Allele Inheritance of Observed Offspring: Fetal Loss (AI_{FL})

6.15.1 What You Will Need

1. The number of alleles at your most variable locus
2. The number of loci you used
3. Your allele frequency file
4. How many offspring you have in your offspring file.
   This is the number of offspring for which you want to calculate allele inheritance.
5. Your offspring genotype file.
   This is a file containing the IDs and genotypes of the offspring for which you want to calculate allele inheritance.
6. How many pairs you have in your mating pairs file.
7. Your mating pairs file.
   This file should contain the IDs and genotypes of the mating pairs. The mating pairs should be in the same order as the offspring in the offspring file. So, the first mating pair in the mating pairs file should be the parents of the first offspring in the offspring file, and so on. In this file **the mother's genotype should be first, and the father's genotype second!!**. Thus, if you have three mating pairs genotyped at one locus, the format should be:

Mom#1	100	100
Dad#1	102	102
Mom#2	100	102
Dad#2	100	100
Mom#3	102	104
Dad#3	100	100

6.15.2 What the Program Does

This function calculates the proportion of the time that a paternal allele is inherited that results in a fetal genotype dissimilar from the mother. The trick for this calculation is that most loci will be uninformative. For example, all loci for which the father is homozygous are uninformative because there is not the "option" for offspring to inherit different alleles from the father. Similarly, loci are uninformative when the mother and father are heterozygous for for different alleles because in this case the offspring *must* inherit a paternal allele that results in a genotype dissimilar from the mother. All of the possible parental and resulting offspring allelic combinations are given below, along with how they are used in the analysis. In all cases the paternal allele in the offspring is listed first, and the maternal allele listed second.

Father	ii^1	ii^1	ij	ij	ij	ii^1	ij^2
Mother	ii	ij	ii	ij	ik	jj	kl
Offspring	ii	ii	ii^3	ii^4	ii^5	ij	ik
		ij	ji^4	ij^3	ik^3		il
				ji^3	ji^5		jk
				jj^4	jk^4		jl

[1] Indicates loci that are uninformative due to the father being homozygous.
[2] Indicates loci that are uninformative due to parents being heterozygous for different alleles.
[3] Indicates informative scenarios where offspring inherited a paternal allele that results in a genotype similar to the mother.
[4] Indicates informative scenarios where offspring inherited a paternal allele that results in a genotype dissimilar to that of the mother.
[5] Indicates scenarios that are uninformative because the mother has passed on an allele

where it is impossible for either paternal allele to result in a genotype the same as the mother.

Note that there are an equal number of informative scenarios resulting in the offspring inheriting a paternal allele the same as, or different, than the maternal allele; thus the expectations based solely on informative scenarios results in the typical expected Mendelian ratio of 0.5. Only informative scenarios are used in the calculation, and loci with any individual of the mother-offspring-father triad missing data are excluded from the calculation.

It is widely recognized that metrics of genetic characteristics are more informative (and more accurately estimate the value of interest) when they are weighted appropriately based on the characteristics of the loci used (e.g. Coltman et al. 1999; Amos et al. 2001; Aparicio et al. 2006). Although it may initially appear that the calculation of allele inheritance (AI_{FL}) itself is not directly dependent on the allele frequencies (because it only depends on which alleles are in the parents), the number of loci that are informative in a mother-father-offspring triad will be influenced by the variability of the loci used. Thus, because AI_{FL} is a combination of both the number of informative loci and the allele inheritance patterns at those loci, it makes sense to weight AI_{FL} values based on the characteristics of the loci used. Moreover, it also makes intuitive sense that AI_{FL} values with a strong signal from relatively uninformative loci should be weighted more than the same value from more informative loci. Weighting loci based on their expected heterozygosity (H_E, Nei 1978), has proven to be appropriate under many circumstances (Queller and Goodnight 1989; Aparicio et al. 2006). Here, AI_{FL} is weighted by the inverse of the average expected heterozygosity, which produces the desired weighting pattern of a given AI value being weighted more for relatively uninformative loci than for more informative loci. This results in the total calculation for AI_{FL} being:

$$AI_{FL} = (L_p \;/\; L) \;/\; \{(\sum_{l=1}^{l} H_E \;/\; l\}$$

Where:

L_p is the number of informative loci where a paternal allele was inherited that results in a genotype dissimilar to that of the mother.

L is the number of informative loci.

$\sum H_E$ is the sum of expected heterozygosities for all typed loci.

l is the number of typed loci.

In words, the numerator is the proportion of informative loci for which a paternal allele was inherited that results in a genotype dissimilar to that of the mother (ranging from 0 to 1), and the denominator is the average heterozygosity for the typed loci.

Note that a result of “999” will be returned for those individuals for which there were not any informative scenarios, and thus allele inheritance could not be calculated. These individuals are then ignored when the program calculates averages.

As an example, let’s walk through this calculation for the example in files that are provided. The example files included are:

“frequencies” − allele frequency data for 2 loci with 4 alleles at the most polymorphic locus.

```
 0     0
 0.1   0.2
 0.3   0.4
 0.5   0.4
 0.1   5
```

“ogenotypes” – A file containing the genotypes of 10 offspring genotyped at 2 loci.

```
 1   1  1  1  1
 2   2  0  3  2
 3   1  2  2  2
 4   3  3  0  2
 5   0  1  2  3
 6   1  4  3  0
 7   2  2  2  2
 8   4  4  1  2
 9   2  4  2  3
10   1  1  3  3
```

“parents” – A file containing the parent for these 10 offspring.

```
 1   1  1  1  1   Mom for offspring #1
 2   1  2  1  1   Dad for offspring #1
 3   0  0  2  2   Mom for offspring #2
 4   2  2  2  3   Dad for offspring #2
 5   1  2  1  2   Mom for offspring #3
 6   1  2  2  3   Dad for offspring #3
 7   2  3  1  2   Mom for offspring #4
 8   2  3  0  0   Dad for offspring #4
 9   1  1  2  3   Mom for offspring #5
10   1  1  1  2   Dad for offspring #5
11   1  4  0  0   Mom for offspring #6
12   3  4  3  3   Dad for offspring #6
13   1  2  2  2   Mom for offspring #7
14   2  2  2  2   Dad for offspring #7
15   3  4  1  1   Mom for offspring #8
16   3  4  2  2   Dad for offspring #8
17   1  2  2  3   Mom for offspring #9
18   4  4  2  3   Dad for offspring #9
19   1  1  2  3   Mom for offspring #10
20   1  1  3  3   Dad for offspring #10
```

The calculation would then be:

Expected heterozygosity (H_E) for locus 1 = 1 - (0.1^2 + 0.3^2 + 0.5^2 + 0.1^2) = **0.64**
Expected heterozygosity (H_E) for locus 2 = 1 - (0.2^2 + 0.4^2 + 0.4^2) = **0.64**

Offspring	L	L_p	$\sum H_E$	l	AI_{FL}
1	1	0	1.28	2	0
2	1	1	0.64	1	1.56
3	1	0	1.28	2	0
4	1	1	0.64	1	1.56
5	1	0	0.64	1	0
6	1	0	0.64	1	0
7	0	NA	1.28	2	999
8	1	1	1.28	2	1.56
9	1	0	1.28	2	0
10	0	NA	1.28	2	999
Average					**0.586**

6.16 Calculate Allele Inheritance of Simulated Offspring: Fetal Loss

6.16.1 What You Will Need

1. The number of alleles at your most variable locus
2. The number of loci you used
3. Your allele frequency file
4. How many pairs you have in your mating pairs file.
5. Your mating pairs file.
   This file should contain the IDs and genotypes of the mating pairs. The mating pairs should be in the same order as the offspring in the offspring file. So, the first mating pair in the mating pairs file should be the parents of the first offspring in the offspring file, and so on. In this file **the mother's genotype should be first, and the father's genotype second!!**. Thus, if you have three mating pairs genotyped at one locus, the format should be:
   ```
   Mom#1   100   100
   Dad#1   102   102
   Mom#2   100   102
   Dad#2   100   100
   Mom#3   102   104
   Dad#3   100   100
   ```
6. How many iterations you want to perform

6.16.2 What the Program Does

This function calculates the expected AI_{FL} values for your offspring under Mendelian expectations. It does this by keeping all mating pairs constant, but generating simulated offspring from these pairs based on Mendelian inheritance. The AI_{FL} is then calculated for these offspring, and an overall average AI_{FL} is calculated. If this process is repeated many times (e.g. 1,000), then the expected distribution of AI_{FL} values can be generated, which can then be used to interpret your data and estimate p-values.

As an example, suppose that you have 50 offspring with identified parents, and the average AI_{FL} for these offspring is 1.23. You suspect that there is a high rate of fetal loss in your population due to the genetic similarity of mothers and their fetuses. To test this hypothesis you perform 1,000 iterations, with each iteration generating 50 offspring from *these same* mating pairs. Thus, each iteration is one realization of the expected AI_{FL} value for *your* data under Mendelian expectations. Suppose that you observe two iterations with an average AI_{FL} value greater than your observed value of 1.23. The interpretation is that the observed offspring inherited paternal alleles that result in genotypes genetically dissimilar from the mother significantly more often than expected based solely on Mendelian inheritance, and the p-value is < 0.001 (you observed fewer than 1 value higher than your observed value, out of 1,000 iterations). It would appear that there is a high rate of fetal loss due to genetic incompatibility in your population.

The example files provided for this analysis are:

"frequencies" – allele frequency data for 2 loci with 4 alleles at the most polymorphic locus.

```
0     0
0.1   0.2
0.3   0.4
0.5   0.4
0.1   5
```

“parents” – A file containing the parent for these 10 offspring.

1	1	1	1	1	Mom for offspring #1
2	1	2	1	1	Dad for offspring #1
3	0	0	2	2	Mom for offspring #2
4	2	2	2	3	Dad for offspring #2
5	1	2	1	2	Mom for offspring #3
6	1	2	2	3	Dad for offspring #3
7	2	3	1	2	Mom for offspring #4
8	2	3	0	0	Dad for offspring #4
9	1	1	2	3	Mom for offspring #5
10	1	1	1	2	Dad for offspring #5
11	1	4	0	0	Mom for offspring #6
12	3	4	3	3	Dad for offspring #6
13	1	2	2	2	Mom for offspring #7
14	2	2	2	2	Dad for offspring #7
15	3	4	1	1	Mom for offspring #8
16	3	4	2	2	Dad for offspring #8
17	1	2	2	3	Mom for offspring #9
18	4	4	2	3	Dad for offspring #9
19	1	1	2	3	Mom for offspring #10
20	1	1	3	3	Dad for offspring #10

7 References

Amos W, Worthington Wilmer J, Fullard K, Burg TM, Croxall JP, Bloch D, Coulson T (2001) The influence of parental relatedness on reproductive success. *Proceedings of the Royal Society of London* Series B **268**: 2021-2027.

Aparicio JM, Ortego J, Cordero PJ (2006) What should we weight to estimate heterozygosity, alleles or loci? *Molecular Ecology* **15**: 4659-4665.

Birkhead TR, Chaline N, Biggins JD, Burke T, Pizzari T (2004) Nontransitivity of paternity in a bird. *Evolution* **58**: 416-420.

Brown JL (1997) A theory of mate choice based on heterozygosity. *Behavioral Ecology* **8**: 60-65.

Coltman DW, Pilkington JG, Smith JA, Pemberton JM (1999) Parasite-mediated selection against inbred Soay sheep in a free-living, island population. *Evolution* **53**: 1259-1267.

Dziminski MA, Roberts JD, Simmons LW (2008) Fitness consequences of parental compatibility in the frog *Crinia georgiana*. *Evolution* **62-4**: 879-886.

Evans JP, Marshall DJ (2005) Male-by-female interactions influence fertilization success and mediate the benefits of polyandry in the sea urchin *Heliocidaris erythrogramma*. *Evolution* **59**: 106-112.

Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, Rossi F (2006) GNU Scientific Library Reference Manual—revised second edition (v 18). Network Theory Ltd., Bristol UK.

Krützen M, Sherwin WB, Connor RC, Barré LM, Van de Casteele T, Mann J, Brooks R (2003) Contrasting relatedness patterns in bottlenose dolphins (*Tursiops* sp.) with different alliance strategies. *Proceedings of the Royal Society of London* Series B **270**: 497-502.

Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity* **43**: 45-52.

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1753-1766.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583-590.

Ober C, Hyslop T, Elias S, Weitkamp LR, Hauck WW (1998) Human leukocyte antigen matching and fetal loss: results of a 10 year prospective study. *Human Reproduction* **13**: 33-38.

Olsson M, Pagel M, Shine R, Madsen T, Doums C, Gullberg A, Tegelström H (1999) Sperm choice and sperm competition: suggestions for field and laboratory studies. *Oikos* **84**: 172-175.

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution* **43**: 258-275.

Van de Casteele T, Galbusera P, Matthysen E (2001) A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology* **10**: 1539-1549.

Wang J (2002) An estimator for pairwise relatedness using molecular markers. *Genetics* **160**: 1203-1215.