

Supplementary Materials

Assessing genuine parents-offspring trios
for genetic association studies

Teo et al.

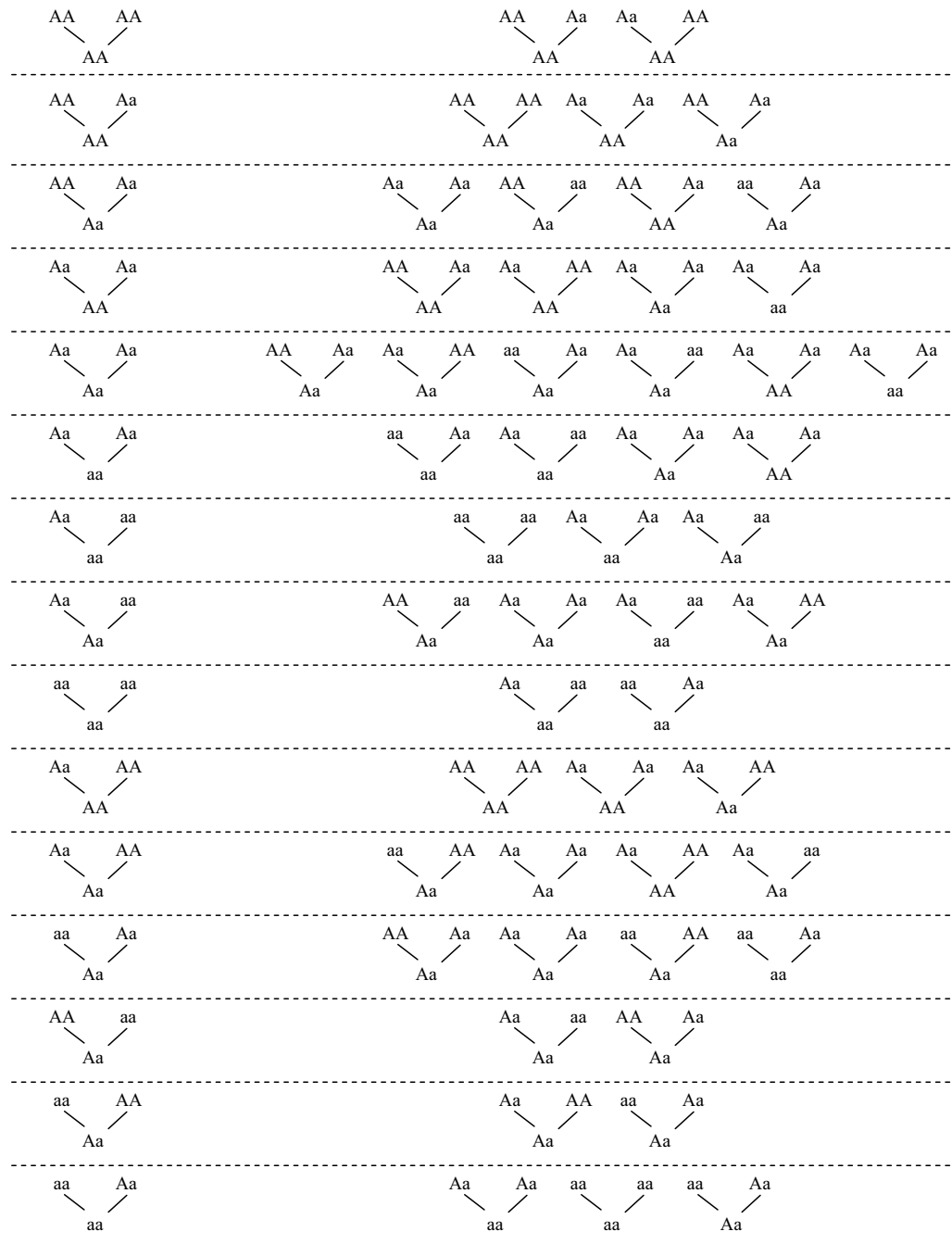
L1: Genotype configurations

In order to assess the conditional likelihood $\Pr(x_{il}|g_{il} = 1, m_{il}, z_i)$, every possible combination of trio genotypes that is consistent with the observed genotypes in the presence of one genotyping error needs to be considered. We allow two kinds of genotyping errors to occur: (i) an error involving only a single allele (i.e. $CC \rightarrow CT$); (ii) an error involving both alleles (i.e. $CC \rightarrow TT$). In the diagrams below, the three nodes in a triangle of relationship represent the putative father (top left), putative mother (top right) and offspring (bottom). A solid line links each of the parental node to the offspring if the putative parent-offspring relationship is true; a dashed line links each of the parental node to the offspring if the putative parent-offspring relationship has been misspecified. We represent the two alleles of a biallelic SNP as A and a .

TRUE FATHER-MOTHER-OFFSPRING TRIO
(Pedigree Consistent)

Observed Genotypes

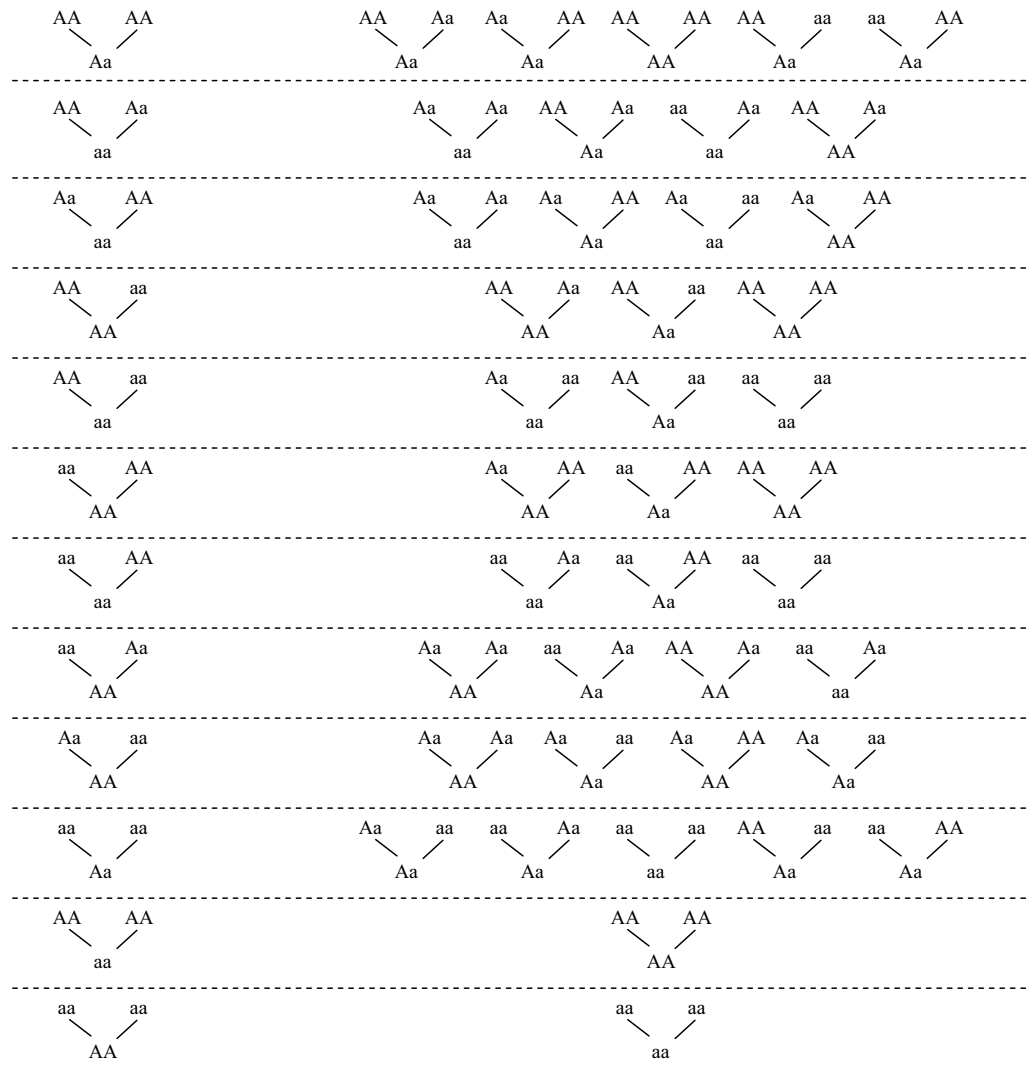
Possible Genotypes (1 error removed)



TRUE FATHER-MOTHER-OFFSPRING TRIO
(Pedigree Inconsistent)

Observed Genotypes

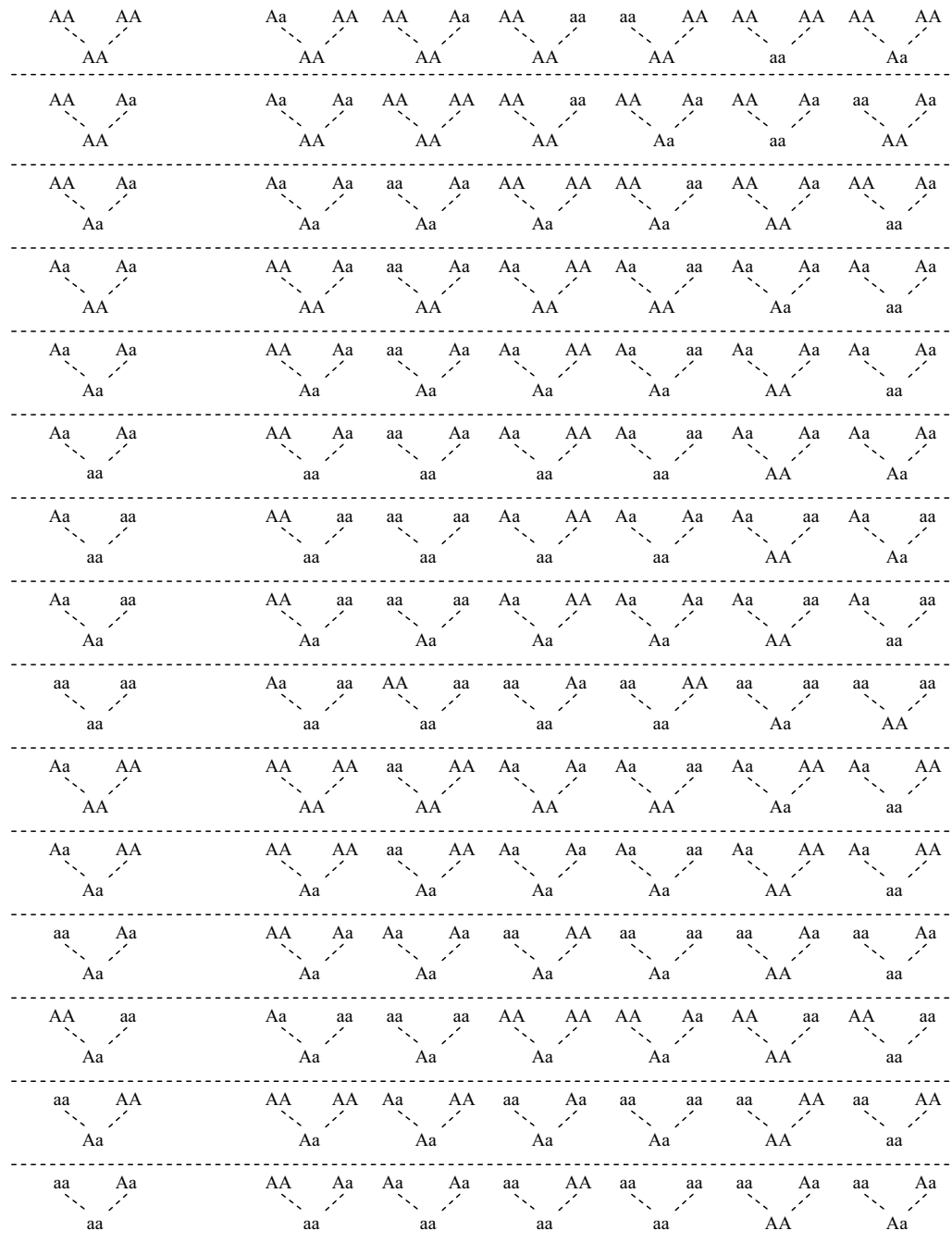
Possible Genotypes (1 error removed)



3 UNRELATED INDIVIDUALS TRIO
(Pedigree Consistent)

Observed Genotypes

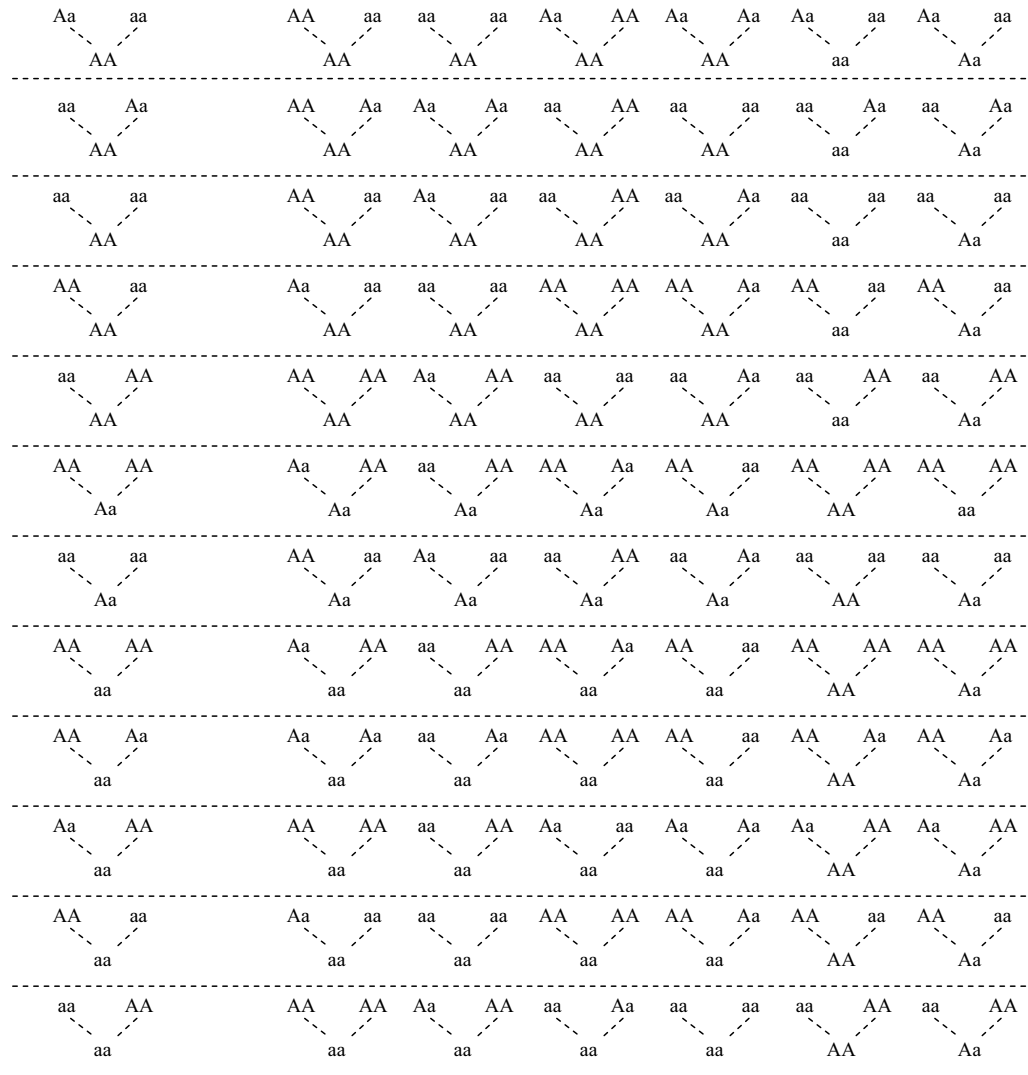
Possible Genotypes (1 error removed)



3 UNRELATED INDIVIDUALS TRIO
(Pedigree Inconsistent)

Observed Genotypes

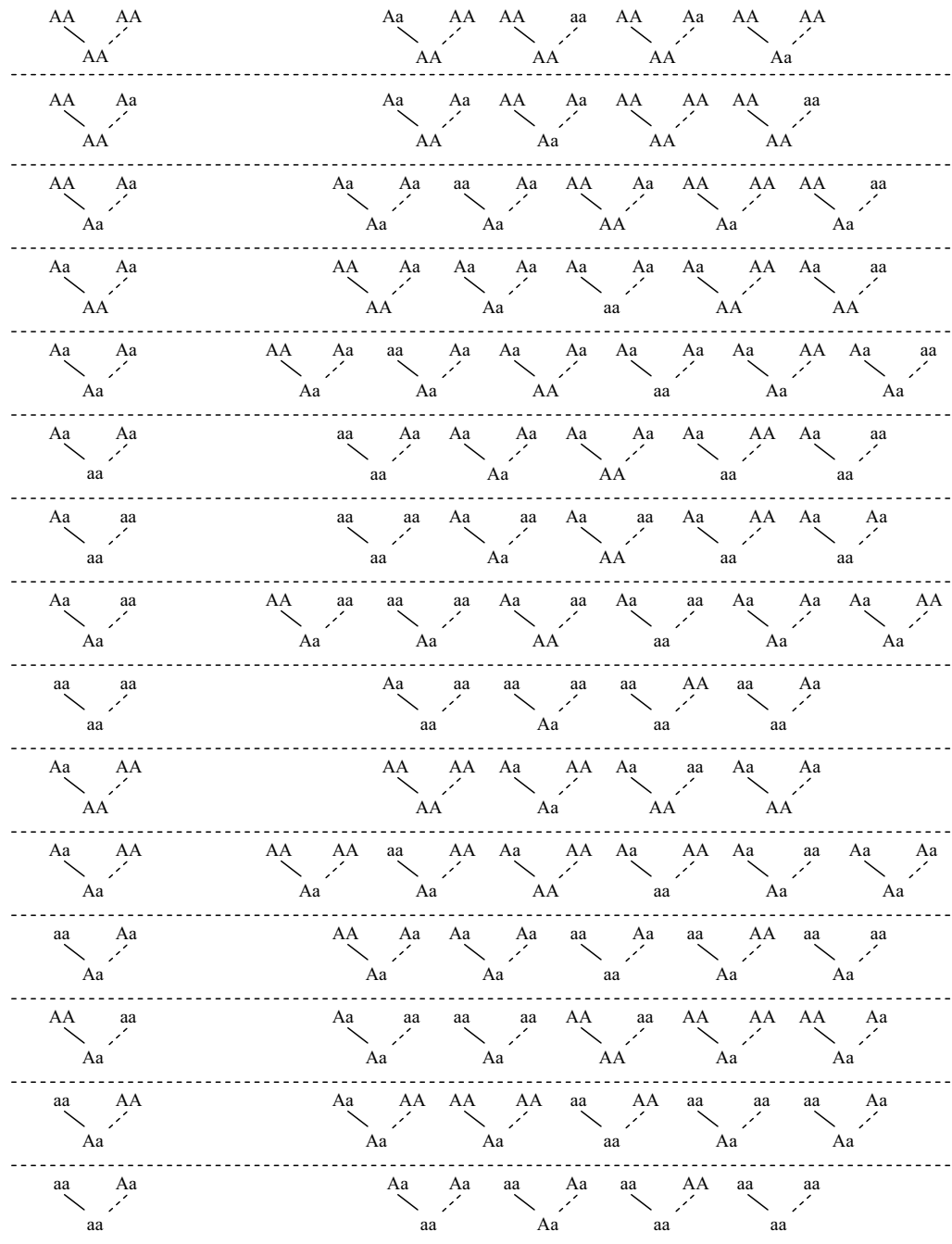
Possible Genotypes (1 error removed)



FATHER-OFFSPRING MISSPECIFIED MOTHER TRIO
(Pedigree Consistent)

Observed Genotypes

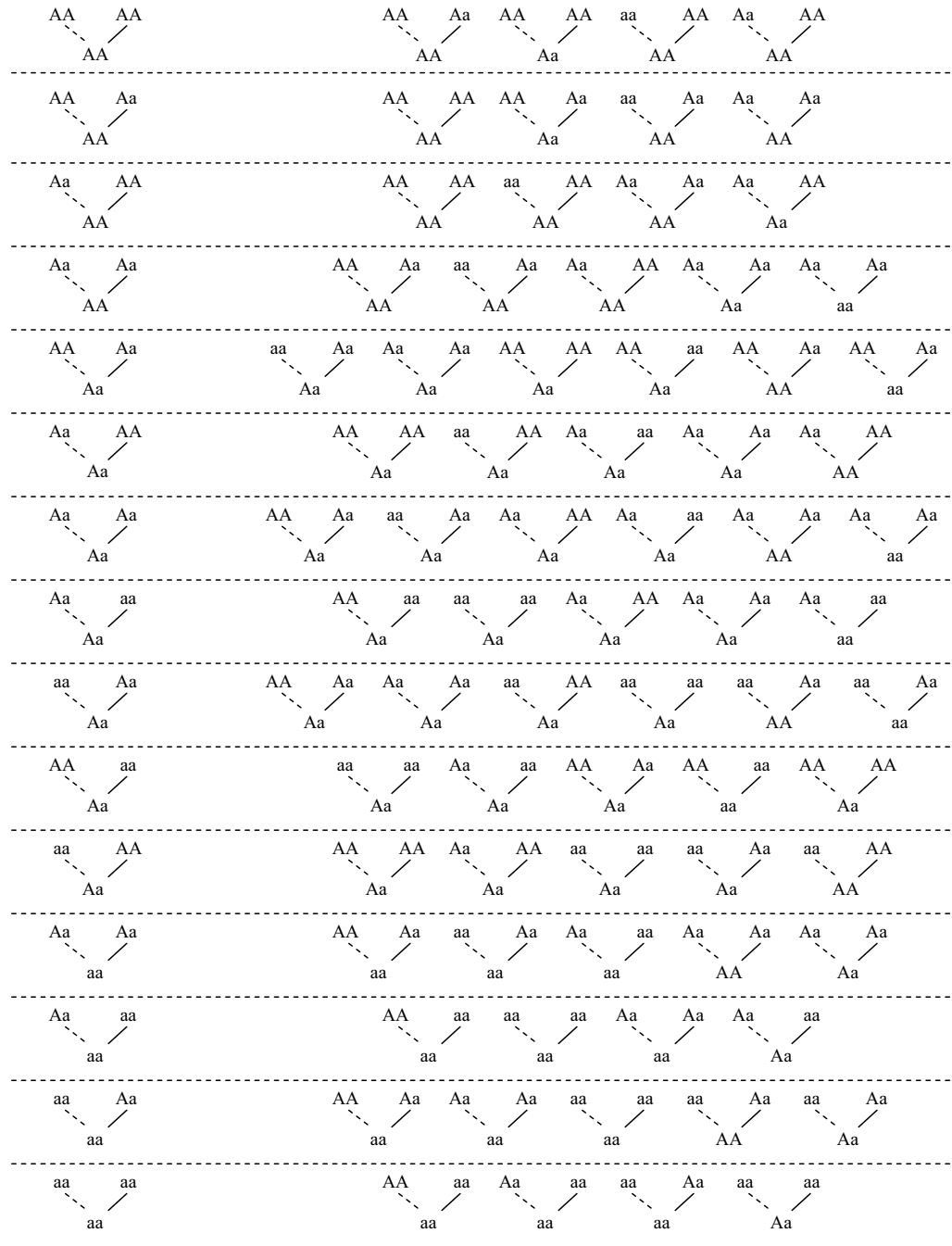
Possible Genotypes (1 error removed)



MOTHER-OFFSPRING MISSPECIFIED FATHER TRIO
(Pedigree Consistent)

Observed Genotypes

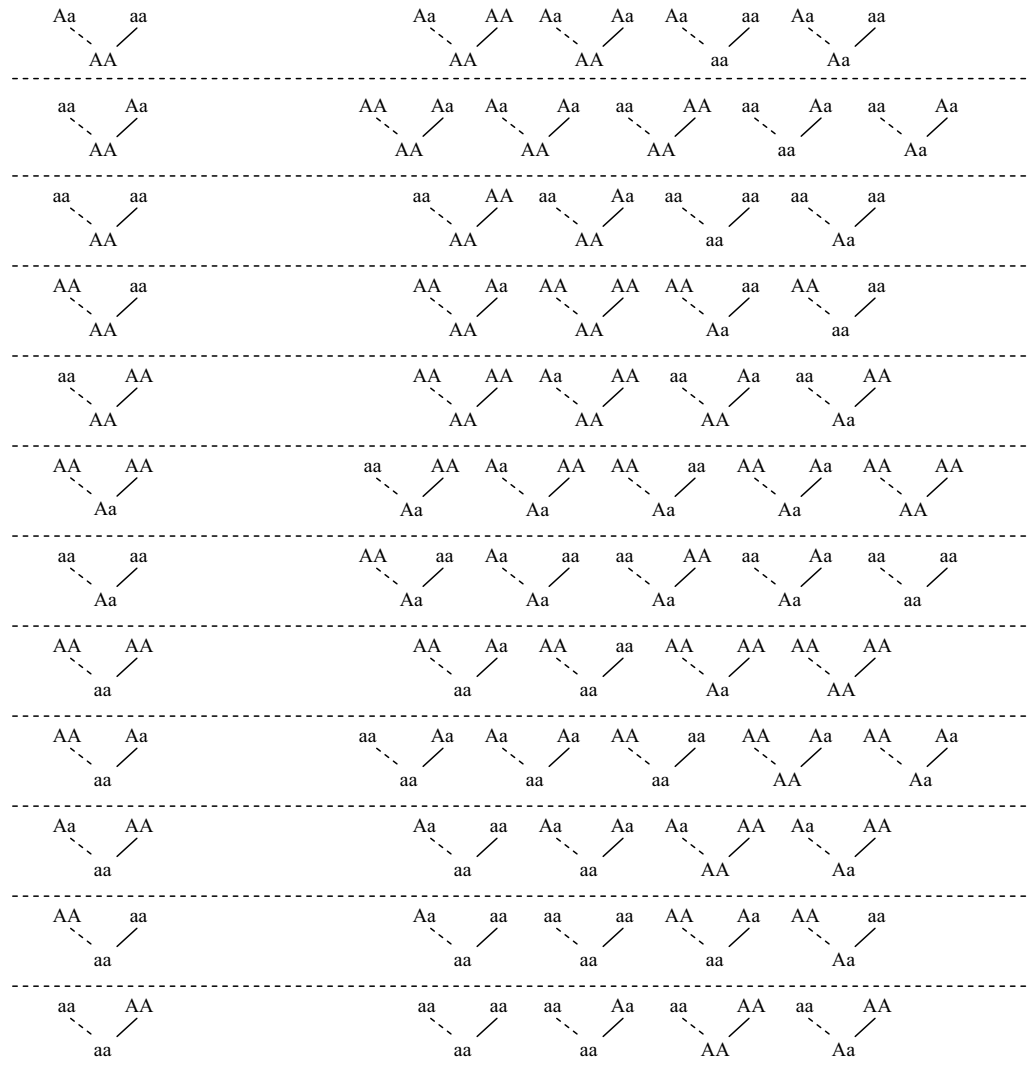
Possible Genotypes (1 error removed)



MOTHER-OFFSPRING MISSPECIFIED FATHER TRIO
(Pedigree Inconsistent)

Observed Genotypes

Possible Genotypes (1 error removed)



L2: Interpretation of available programs

In order to provide an objective comparisons between *Nucl3ar* and the available programs for inferring pedigree relationships, we provide a detailed description of how we operated and interpreted the output from these programs.

Prest

For each trio, *Prest* compares each of the two putative parent-offspring relationships separately. For each parent-offspring pair, the program estimates the probabilities of sharing 0, 1 and 2 alleles identity-by-descent (IBD), and also evaluates the two-tail significances for the AIBS and IBS tests. As the AIBS test has higher power than that of the IBS test, we use the significance of the AIBS test as an objective function and evaluate the performance of *Prest* at various thresholds. A trio is assigned as a true trio if the significance levels from the tests for both putative father-offspring and mother-offspring relationships are above the threshold. Otherwise it is assigned as a false trio. While it was recommended that the estimated IBD probabilities be used for testing parent-offspring relationship, we found that it was non-trivial to interpret objectively the estimated IBD probabilities against the null probabilities of (0, 1, 0) for true parent-offspring relationship. The significance thresholds used are 0.4, 0.3, 0.2, 0.1, 0.05 and 0.02.

Relcheck

Relcheck performs all possible (three) pairwise comparisons between the individuals in each trio and for each pair, calculates the number of markers at which the pair has been successfully typed and share 0, 1 or 2 alleles identical-by-state (IBS). The log-likelihood ratio (effectively the LOD score) for each of five possible relationships (monozygotic twins, parent-offspring, full siblings, half siblings, unrelated) is also calculated, and this is relative to the relationship giving the maximum likelihood. Thus the most likely relationship has a LOD score of zero, and the remaining four relationships have negative LOD scores. We subtract the LOD of the parent-offspring from the maximum of the LOD for the remaining four relationships, and this thus provides the means for assessing how likely the parent-offspring relationship is, relatively to the most likely relationship. We then set various thresholds for this difference, and a trio is assigned to be true only if both putative father-offspring and mother-offspring pairs have differences less than or equal to the designated threshold. The thresholds used are 0, 0.25, 0.5, 0.8 and 1.0. As the LODs are \log_{10} likelihood ratios, if the difference is greater than a threshold

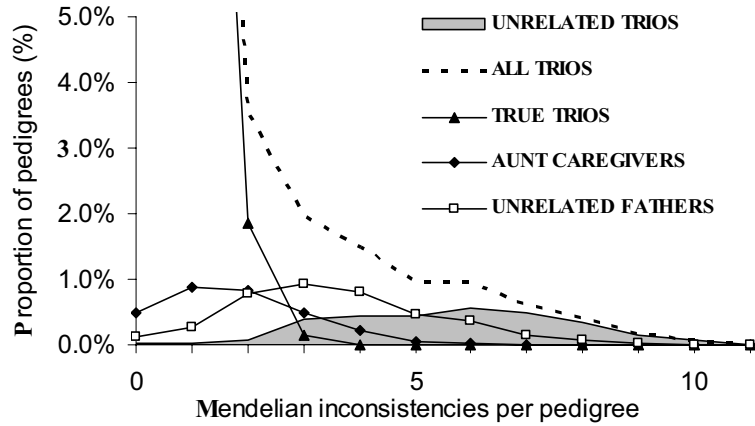
of x , it means that the most likely relationship (excluding parent-offspring) is at least 10^x more likely than the parent-offspring relationship.

Relpair

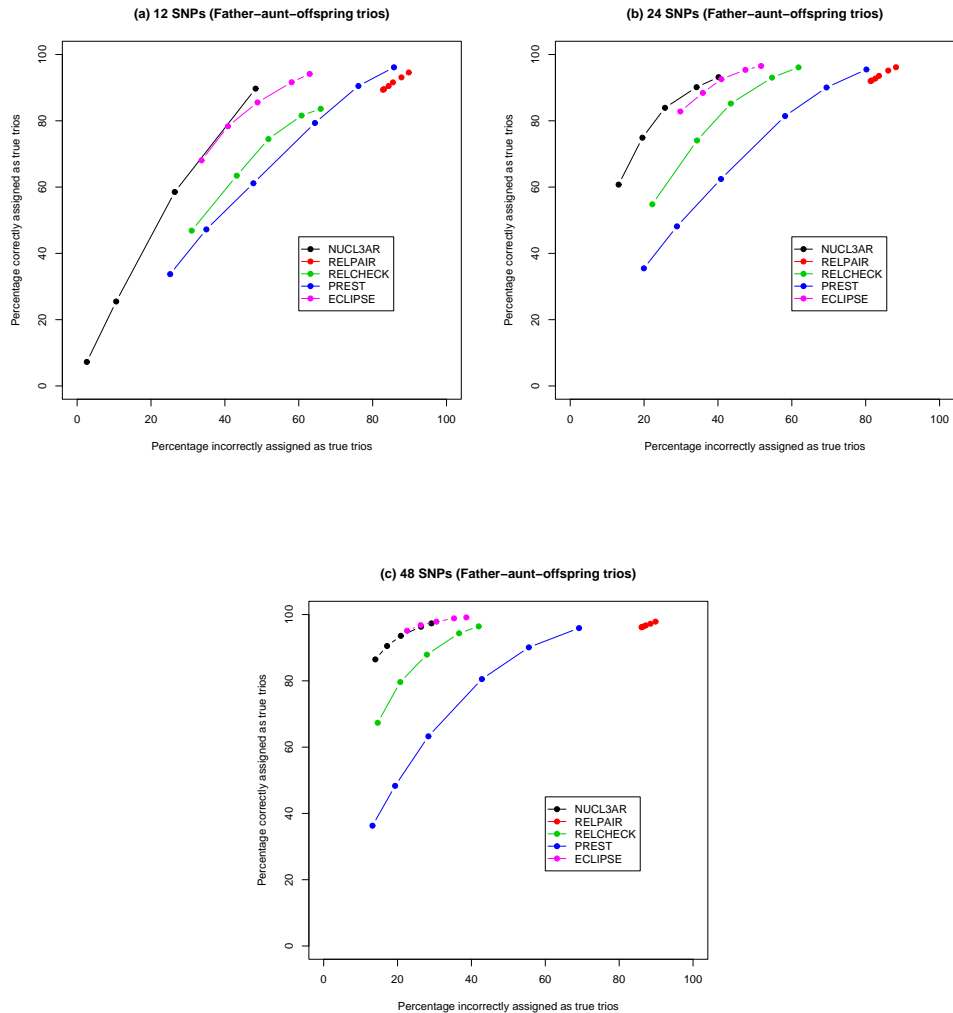
For each trio, *Relpair* (version 2.0) compares each of the two putative parent-offspring relationships separately for nine relationships (parent-offspring, grandparent/grandchild, avuncular, first cousins, monozygotic twins, full sibs, half sibs, unrelated pairs), and effectively calculates the likelihood ratio between the putative relationship (parent-offspring in our case) and inferred relationship (the most likely relationship out of the nine considered). For each of the nine relationships, a scaled likelihood ratio between zero and one is also calculated, relative to the most likely relationship which has a scaled likelihood ratio of one, and larger values indicate relationships that are more likely. In our situation of evaluating true/false trios, we consider only the scaled likelihood ratio for the parent-offspring relationship, and assign a trio as true if both father-offspring and mother-offspring have scaled likelihood ratios larger than some threshold. The thresholds we consider are 0.7, 0.8, 0.9, 0.95, 0.99 and 0.999.

Eclipse3

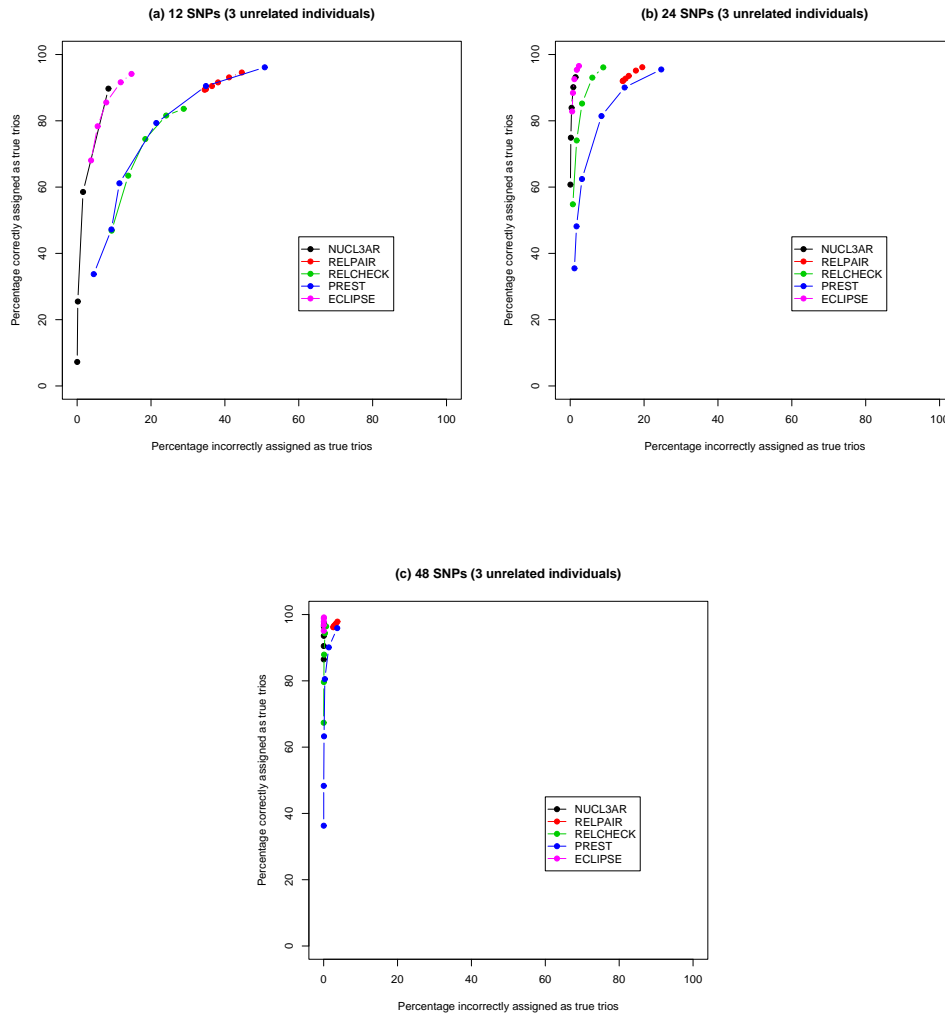
Eclipse3 considers the information from analysing all three individuals in a trio jointly, and assesses the (\log_{10})likelihoods of 43 possible relationships between three individuals. For a detailed description of the possible relationships, we recommend referring to the manual provided by the author on the method. For each trio, we subtract the log-likelihood of the parents-offspring relationship from the maximum of the remaining log-likelihoods, which represents the most likely relationship excluding the parents-offspring relationship. A trio is assigned to be true only if the difference is less than or equal to various thresholds, defined as 0, 0.25, 0.5, 0.8 and 1.0. The interpretation of the difference is similar to that of *Relcheck*.



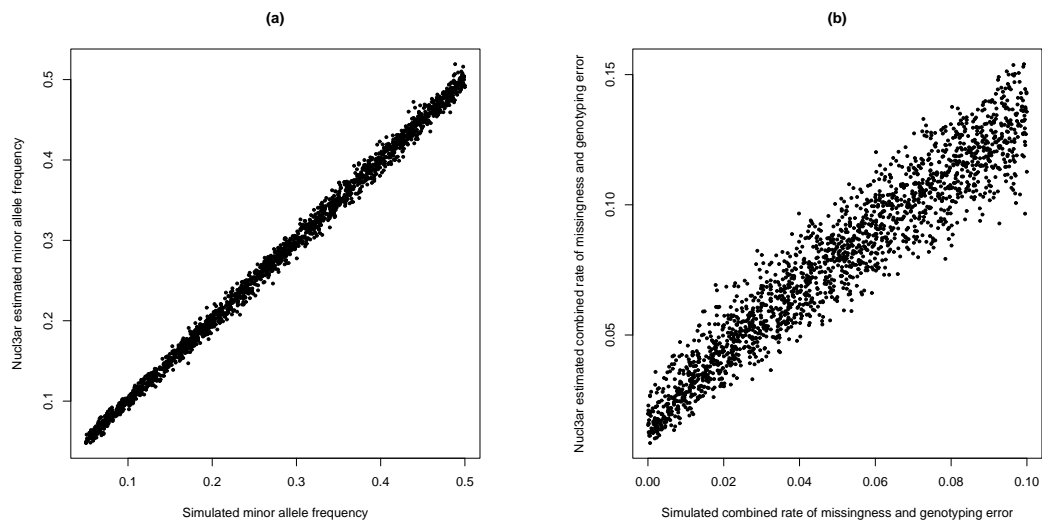
S1: Mendelian inconsistencies in a simulated dataset of family trios with a range of relationship misspecifications. Using SimPed (Leal et al. 2005), we simulated a dataset of 10,000 trios genotyped for 24 unlinked biallelic neutral markers each with minor allele frequency of 0.25. Random genotyping errors (at a rate of 2%) were inserted and 2% of the data were removed to simulate missing genotypes. Rates of relationship misspecification were: 4% paternal discrepancy (misspecified father); 3% caregive aunts (maternal aunts as mothers); 3% swapped cases (trios of unrelated individuals). Although trios with misspecified relationships have increased rates of Mendelian inconsistencies, a mixture of misspecification types, rates and random genotyping error creates a long tail - rather than a distinct cut-off between trios with and without misspecification. This leaves the user with a choice between statistical power (excluding some true trios) and bias (retaining misspecified trios).



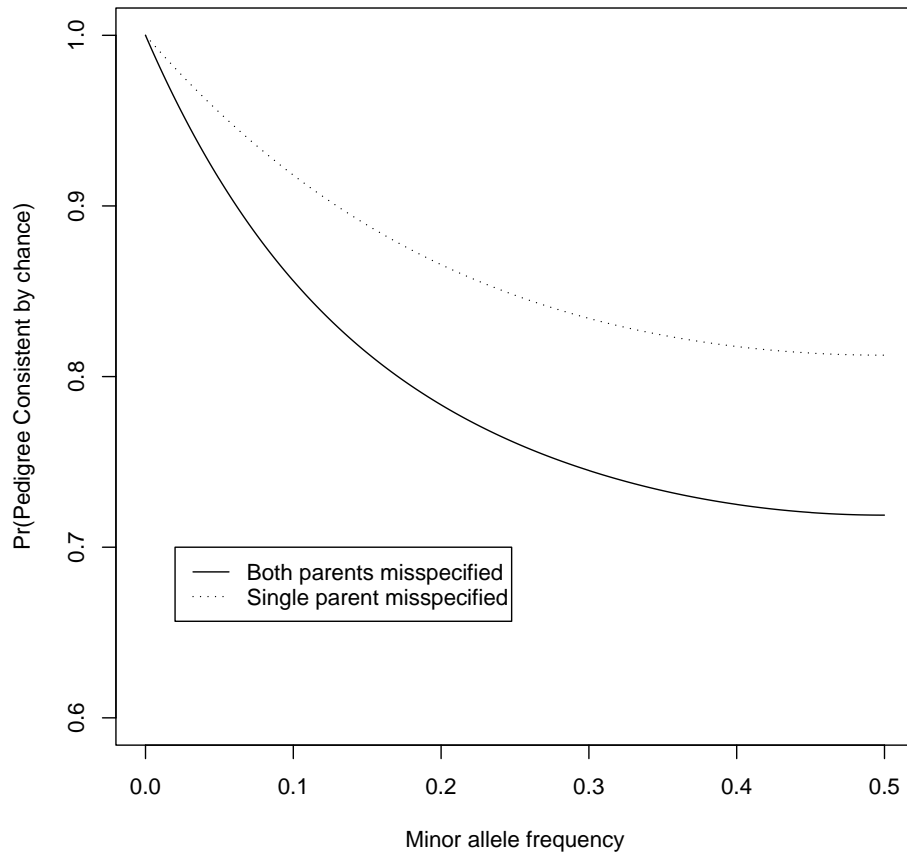
S2: Percentages of correct and incorrect trio assignment as true, for (a) 12 SNPs; (b) 24 SNPs; (c) 48 SNPs. The x-axes show the percentages of the data simulated as father-maternal aunt-offspring trios which have been incorrectly assigned as true trios. The y-axes show the percentages of the data simulated as true trios which have been correctly assigned as true trios. The curves are obtained by considering the various thresholds: *Nucl3ar* (black - maximum posterior probability, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95), *Relpair* (red - 0.7, 0.8, 0.9, 0.95, 0.99, 0.999), *Relcheck* (green - 0, 0.25, 0.5, 0.8, 1.0), *Prest* (blue - 0.4, 0.3, 0.2, 0.1, 0.05, 0.02), *Eclipse3* (pink - 0, 0.25, 0.5, 0.8, 1.0).



S3: Percentages of correct and incorrect trio assignment as true, for (a) 12 SNPs; (b) 24 SNPs; (c) 48 SNPs. The x-axes show the percentages of the trio data simulated with both parents misspecified (effectively 3 unrelated individuals) which have been incorrectly assigned as true trios. The y-axes show the percentages of the data simulated as true trios which have been correctly assigned as true trios. The curves are obtained by considering the various thresholds: *Nucl3ar* (black - maximum posterior probability, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95), *Relpair* (red - 0.7, 0.8, 0.9, 0.95, 0.99, 0.999), *Relcheck* (green - 0, 0.25, 0.5, 0.8, 1.0), *Prest* (blue - 0.4, 0.3, 0.2, 0.1, 0.05, 0.02), *Eclipse3* (pink - 0, 0.25, 0.5, 0.8, 1.0).



S4: Comparisons of parameters used for simulating the datasets under the four trio relationships against the *Nucl3ar* estimations for: (a) minor allele frequency for each SNP; (b) combined rate of missingness and genotyping error. Plots shown are for simulations performed with 48 SNPs.



S5: Plot of the probability of observing a pedigree consistent genotype configuration for a false trio by chance, against the minor allele frequency. The solid line indicates the relationship if the underlying trio relationship has both putative parents misspecified, while the dashed line indicates the relationship when there is a single misspecified parent.