Proceedings, 10th World Congress of Genetics Applied to Livestock Production

AccurAssign, software for accurate maximum-likelihood parentage assignment

***D Boichard*[1], L Barbotte[2], L Genestout[2]**
[1]INRA, UMR1313 Gabi, 78350 Jouy en Josas, France ; [2]LABOGENA-DNA, 78350 Jouy en Josas, France

ABSTRACT: Knowing pedigrees is essential for selection, conservation, and management of animal populations. AccurAssign is a software implementing the main principles of parentage assignment by maximum likelihood which has been developed and used at LABOGENA-DNA. As compared to the simple exclusion approach, AccurAssign exhibits several advantages, such as accounting for genotyping errors, ranking potential parents, and avoiding incorrect matching when the correct parents are not included in the dataset. The main features of the software are presented. Sensitivity and specificity are estimated for a simulated population genotyped with a varying number of SNP and with parents genotyped or not.
Keywords: Parentage assignment; Maximum likelihood; Software

Introduction

Management of animal populations requires information about pedigrees. However, controlled matings are not always possible (eg, in natural or free-range populations) or are time-consuming. Having an efficient way to *a posteriori* identify actual parentage also allows for flexible mating systems such as insemination with mixed semen or optimal rearing systems such as collective tanks for fish. For several years, parentage assignment with genetic markers has become an increasingly popular method and will become more so with the decreasing cost of genotyping.

Most current assignment methods rely on parentage exclusion. A candidate parent can be excluded when it does not share one or several markers with its offspring. This approach is fast and efficient, especially with massive marker datasets (eg, Hayes, 2011), but it presents several limitations: it must allow for some incompatibilities due to genotyping errors, on an arbitrary basis; it does not rank the retained candidates in an objective way; it strongly relies on a complete availability of parent genotypes and does not provide any rule in case of their absence. Methods based on likelihood, such as implemented in Cervus (Marshall et al, 1998; Kalinowski et al 2007), can overcome these difficulties but they are often computationally demanding. In a high throughput environment such as a large-scale genotyping laboratory, a fast and efficient method is needed. AccurAssign was developed for that purpose.

Materials and Methods

Likelihood expression

As we are interested only in the sire-dam-offspring trio, the likelihood formula is straightforward and fast to compute. A similar approach was also used independently by Melony et al (2012). For a given marker, the contribution to the likelihood is derived in a straightforward way from Mendelian rules and is given in table 1 for a homozygous offspring and in table 2 for a heterozygous offspring. For instance, the probability of observing an AA or AB offspring from two AB parents is 0.25 and 0.5, respectively. To cope with potential genotyping errors, any incompatibility is given a small non-zero probability. The true error probability varies according to the genotyping method. But its value is not critical for parent assignment, as soon as it is small enough to penalize incompatibilies but different from zero to avoid exclusion based on a single marker incompatibility. The lower this probability, the more compatible markers are needed to counterbalance this penalty.

Table 1. Likelihood contribution of one marker when the offspring is homozygous (eg, AA).

Sire \ Dam	AA	AC	CC	Missing
AA	1	0.5	e	f_A
AC	0.5	0.25	e	0.5 f_A
CC	e	e	e	e
Missing	f_A	0.5 f_A	e	f_A^2

f_A: frequency of allele A
C: any allele different from A
e: genotyping error probability

Table 2. Likelihood contribution of one marker when the offspring is heterozygous (eg, AB).

Sire /Dam	AA	AB	AC	BB	BC	CC	Missing
AA	e	0.5	e	1	0.5	e	f_B
AB	0.5	0.5	0.25	0.5	0.25	e	0.5 $(f_A + f_B)$
AC	e	0.25	e	0.5	0.25	e	0.5 f_B
BB	1	0.5	0.5	e	e	e	f_A
BC	0.5	0.25	0.25	e	e	e	0.5 f_A
CC	e	e	e	e	e	e	e
Missing	f_B	0.5 $(f + f_B)$	0.5 f_B	f_A	0.5 f_B	e	2 f_A f_B

f_A and f_B: frequencies of alleles A and B, respectively
C: any allele different from A and B

When a genotype is missing for a given candidate parent, the likelihood uses allelic frequencies of the parental population. Finally, for a given individual, the log-likelihood is obtained for each sire-dam couple by summing the log-contribution of each combination over all markers, assuming markers are independent from each other. When parents from only one sex are genotyped, the log-likelihood is computed using allelic frequencies as for missing genotypes.

Likelihood versus exclusion approach

To illustrate the difference between exclusion and likelihood approaches, let us consider an example with one individual with the same genotype AB at 10 loci. Two couples of parents are tested. In the first couple, the sire is AA and the dam is BB at all 10 loci, whereas in the second couple, both sire and dam are AB at all loci. With the exclusion approach, both couples are ranked equally with 100% compatibility. With the likelihood approach, the first couple is about 1000 times more likely than the second, because the likelihood of couple one is $1^{10} = 1$ whereas it is 0.5^{10} for the second couple.

Let us consider now a similar example, but for one locus, the offspring's genotype is AA. The exclusion approach ranks the second couple first with 100% compatibility whereas the first couple has one incompatibility (ie 90% compatibility). Assuming a 1% genotyping error rate, the likelihood values are $L1=1^9 \times 0.01 = 0.01$, and $L2=0.5^{11}$ rounded to 0.0005. Therefore, with the likelihood approach, the first couple is ranked first and is 20 times more likely than the second one, in spite of one incompatibility. In that particular case, the contribution of all compatible markers is large enough to counterbalance the impact of the genotyping error.

Ranking of assigned parents

Each tested couple is characterized by its log-likelihood. The various couples can be ordered from the highest likelihood to the lowest. Several potential couples can be retained when their probabilities are not too different. A couple P1 which log-likelihood differs by x from the one of couple P2 is e^x more likely than P2. For instance, if x=2.3, P1 is 10 times more likely than P2. Typically, x-values between 2 and 3 can be chosen as a threshold to retain candidate couples of parents.

Assignment power

The empirical power of the design can be estimated by simulating offspring from couples randomly sampled from the list of potential sires and dams. Genotyping error and missing genotype rates are accounted for in the simulation. Assignment power is then estimated by the ratio of the number of correctly assigned couples with the highest likelihood value (rank=1) over the number of simulated offsprings.

Potential parents missing from the dataset

When one or more parents are not genotyped, an additional difficulty is to avoid selecting an incorrect one by default, just because it shows the highest likelihood among the candidate parents. A statistical test can be built on the average Mendelian transmission probability defined as p=exp (LogV/m) where LogV is the log-likelihood and m the number of markers. A convenient way is to build the empirical distributions of the Mendelian probabilities, under the assumption that the individual is an offspring of the couple (H1 hypothesis) or is not its offspring (H0 hypothesis).

The H1 distribution is obtained by simulating an offspring for a number of couples randomly sampled among the candidates and by computing the log-likelihood for each sampled couple. To generate the H0 distribution, a true offspring is obtained as in H1, but another (incorrect) couple is sampled at random to compute its log-likelihood. During the simulations, the genotyping error probability and the rate of missing genotypes are accounted for.

These two distributions can be used to derive confidence thresholds of the Mendelian probability values, as done by Melony et al (2012). Let us denote h1 the 1% quantile of the H1 distribution and h0 the 99% quantile of the H0 distribution. Let us consider the best couple with Mendelian probability p: if $h0<h1$, we can conclude that the couple is acceptable if $p>h1$, that it is not the right one if $p<h0$, and that the result is ambiguous if $h0<p<h1$. Of course, if H0 and H1 distributions overlap and $h0>h1$, results are more difficult to interpret due to a lack of information. However, in the latter case, if $p>h0>h1$, the couple is acceptable, whereas it should be disregarded if $p<h0$, even if $p>h1$.

The same approach is used with only one tested parent. Because H0 and H1 distributions are different if one parent or a couple is tested, the comparison is carried out by comparing p with the appropriate distributions.

The same approach can be used with the number of incompatible markers. H1 is the distribution of the number of incompatible markers for the correct couple (which requires that the genotyping error rate be estimated first) and H0 the number of incompatible markers for an incorrect couple. A complementary test can be derived by comparing the observed incompatibility number with the H0 and H1 distributions.

Comparison with a mating design

Sometimes, a partial mating plan is available from the breeder. For instance, in poultry, matings occur in pedigree pens, with a limited number of males and females. In sheep flocks with only natural mating rams, a list of candidate rams is available. In practice, it is preferable to predict the parents without this information because limiting the search space may lead to wrong results where the information is incorrect. However, it is interesting for the laboratory as well as for the breeder to compare the predictions with the possibilities proposed by the breeder.

AccurAssign main features

Input files of AccurAssign are:

- A genotype file, including all individuals (parents and offspring)
- A parent file, including the list of candidate parents
- A parameter file, for file names and options
- Optionally, a mating plan file, with all couples or trios suggested by the breeder.

AccurAssign provides a number of output files:
- Characteristics of the markers: number of alleles, heterozygosity, exclusion probability with one parent or two parents (Garber and Morris, 1983; Jamieson and Taylor, 1997), identity probability for each marker individually and altogether, allelic frequencies in the parent and offspring populations, genotypic frequencies.
- Quality control results, based on missing genotyping information
- Assignment file: for each individual, all proposed couples with their log-likelihood, the log-likelihood difference with the best solution, the average Mendelian transmission probability, a proposed diagnosis ("SURE", "POSSIBLE", "DOUBTFUL"), the number of informative markers and the number of incompatible markers with the sire, the dam, and the couple, and the list of incompatible markers.
- The list of all incompatibilities, based on "SURE" results ranked first. These results help to detect unreliable markers and can be used to estimate genotyping error rates for each marker.
- The list of unassigned individuals
- The actual mating plan, with the number of offspring per couple
- A comparison with the proposed mating plan

AccurAssign allows the user to choose several options, including:
- The maximum number of couple proposals for one offspring in the output file
- The maximum difference in log-likelihood value as compared to the best couple for listing couples in the output file
- Genotyping error and missing genotype rates for the simulations
- Number of simulations to estimate assignment power.

Results

Example

Due to the deterministic approach, AccurAssign is fast and well adapted to the daily routine activity of a large genotyping lab. To speed up the process, a first screening is carried out to reduce the number of possible parents, before exploring all the corresponding couples. As an example, the analysis of 1686 trout born from 100 sires and 102 dams with 13 microsatellite markers lasted only 3 seconds on a laptop, including the 5000 simulations for the H0 and H1 distributions and 2000 simulations for assignment power (estimated at 95.2% with 2 genotyped parents). In that case, the total error rate was estimated to be 0.02. This low value for microsatellites results from the long experience of the lab with these markers. Quantile values h0 and h1 were 0.10 and 0.13, to be compared to the average Mendelian probability of the selected couples equal to 0.31. As a result, 1595 individuals (94.6%) were assigned to one couple, 63 to two possible couples, 13 to 3 couples, and 15 to 4-7 couples.

Assignment power estimated by simulation

A population was simulated with 50 sires and 100 dams per generation. Full and half-sib parents were simultaneously selected in order to generate highly related parents. The dams were not genotyped. When all sires were included in the analysis and assuming a genotyping error rate equal to 0, assignment power was higher than 99% with 50 biallelic markers with Minor Allele Frequency (MAF) 0.3, or 75 markers with MAF equal to 0.2. When the genotyping error rate increased to 2% (a high value for SNP), the assignment rate decreased from 99.5 to 97% with 50 markers and from 100 to 99.8% with 75 markers, with marker MAF equal to 0.3. The method was found to be very robust to the assumed (non-zero) genotyping error rate. More results are available in Barbotte et al (2012).

Assignment specificity

The specificity is defined by 1 minus the frequency of assignment to an incorrect sire. Similar simulations were carried out after removing 50% of the sires from the analysis. Again, the dams were not genotyped. The observed specificity was found to be low with less than 100 markers and only became satisfactory with more than 150 markers (table 3).

Table 3. Power and specificity of assignment according to the number of SNP (dams not genotyped, 50% sires not in the dataset, MAF=0.3)

# SNP	200	175	150	125	100	75	50
Power	100	100	99	99	99	99	91
Specificity	100	100	93	78	57	32	13

specificity=1 – correct sire assignment frequency

Conclusion

AccurAssign is a fast and accurate software for parentage assignment, based on a maximum likelihood approach. Developed for the needs of a routine genotyping laboratory, it provides various features in addition to assignment, especially to characterize the quality of the markers and the dataset. It is robust to genotyping error. It also provides a measure of the reliability of the assignment and a ranking between proposed couples

Literature Cited

Barbotte L., Genestout L., Fritz S., et al. (2012). 19èmes Rencontres Recherches Ruminants, 92, Paris, France.

Garber, R.A., and Morris J.W. (1983) In: Inclusion Probabilities in Parentage Testing (ed. by R.H. Walker), pp. 277–80.

Hayes, B.J. (2011). Journal of Dairy Science 94, Volume 94, 2114-2117.

Jamieson, A., and Taylor, St.C. S. (1997). Animal Genetics, 28, 397–400

Kalinowski S.T., Taper M.L., and Marshall T.C. (2007) Molecular Ecology 16, 1099–1106.

Marshall T.C., Slate J., Kruuk, L.E.B., et al (1998) Molecular Ecology 7, 639–655.

Melony, J.S., Dierens L., McWilliam S. et al (2012). Aquaculture Research, 2012, 1–10