Grashei *et al. Genet Sel Evol (2018) 50:26*
https://doi.org/10.1186/s12711-018-0397-7

Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

Using genomic relationship likelihood for parentage assignment

Kim E. Grashei[1,2*], Jørgen Ødegård[1,2] and Theo H. E. Meuwissen[2]

Abstract

Background: Parentage assignment is usually based on a limited number of unlinked, independent genomic markers (microsatellites, low-density single nucleotide polymorphisms (SNPs), etc.). Classical methods for parentage assignment are exclusion-based (i.e. based on loci that violate Mendelian inheritance) or likelihood-based, assuming independent inheritance of loci. For true parent–offspring relations, genotyping errors cause apparent violations of Mendelian inheritance. Thus, the maximum proportion of such violations must be determined, which is complicated by variable call- and genotype error rates among loci and individuals. Recently, genotyping using high-density SNP chips has become available at lower cost and is increasingly used in genetics research and breeding programs. However, dense SNPs are not independently inherited, violating the assumptions of the likelihood-based methods. Hence, parentage assignment usually assumes a maximum proportion of exclusions, or applies likelihood-based methods on a smaller subset of independent markers. Our aim was to develop a fast and accurate trio parentage assignment method for dense SNP data without prior genotyping error- or call rate knowledge among loci and individuals. This genomic relationship likelihood (GRL) method infers parentage by using genomic relationships, which are typically used in genomic prediction models.

Results: Using 50 simulated datasets with 53,427 to 55,517 SNPs, genotyping error rates of 1–3% and call rates of ~ 80 to 98%, GRL was found to be fast and highly (~ 99%) accurate for parentage assignment. An iterative approach was developed for training using the evaluation data, giving similar accuracy. For comparison, we used the Colony2 software that assigns parentage and sibship simultaneously to increase the power of the likelihood-based method and found that it has considerably lower accuracy than GRL. We also compared GRL with an exclusion-based method in which one of the parameters was estimated using GRL assignments.This method was slightly more accurate than GRL.

Conclusions: We show that GRL is a fast and accurate method of parentage assignment that can use dense, non-independent SNPs, with variable call rates and unknown genotyping error rates. By offering an alternative way of assigning parents, GRL is also suitable for estimating the expected proportion of inconsistent parent–offspring genotypes for exclusion-based models.

Background

In the field of animal genetics, low-density single nucleotide polymorphisms (SNPs), microsatellites, and amplified fragment length polymorphisms (AFLP) have long been the preferred types of genomic data for parentage assignment due to their low cost [1–3]. In practice, the foundation of parentage assignment rests on exclusion- and likelihood-based methods [4]. Exclusion-based methods rely on their ability to exclude false parent–offspring combinations when the offspring's candidate parents' genotypes violate Mendel's laws. These methods are often used due to their ease of interpretation, but the number of expected exclusions depends on allele frequencies in the population and on genotype call rates and error rates [5]. Exclusion-based methods also require more loci than likelihood-based methods since only genotypes with Mendelian inconsistencies are used [6]. Likelihood-based methods often calculate

*Correspondence: kim.erik.grashei@aquagen.no; kim.grashei@nmbu.no
[1] AquaGen AS, P.O. Box 1240, NO-7462 Trondheim, Norway
Full list of author information is available at the end of the article

the likelihood ratio (LR) of the genotype of the offspring, which is the probability of the offspring's genotype given the genotypes of the candidate parents, relative to the probability of observing the genotype in the population by chance. The LR statistic effectively gives more weight to rare alleles. Different loci are typically assumed independent, such that total LR is multiplied over all loci. Likelihood-based methods have higher power than exclusion-based methods, but their interpretation is more complicated. Both likelihood- and exclusion-based models usually assume known and homogenous genotype error rates and independent loci, and do not account for variation in genotype call rates [5, 7, 8], which are all important assumptions when working with high-density SNP data. For dense SNP chip data, the assumption of independent inheritance among loci is not realistic (i.e., alleles are inherited on large DNA segments), which may lead to inflated LR values when using conventional likelihood-based methods.

Parentage can also be assigned and tested by using realized genomic relationships. The interrelationship between parents governs the expected inbreeding in offspring, as well as parent–offspring relationships. Realized genomic relationships assess the average genomic similarity across loci and do not assume independence of the loci. Increasing the number of markers in the calculations, increases the precision of the genomic relationships. Our aim was to study whether genomic relationships can be used to perform computationally fast and accurate parentage testing with high-density SNP data.

Methods

Residual genomic relationships

Estimates of genomic relationships require large numbers of loci [9], and their expectation is proportional to the genetic covariance between individuals. The proposed method for parentage testing is developed for trio parentage testing, i.e. using a single offspring and two parental candidates. The method uses genomic relationships estimated by VanRaden's first method [10], in which the genomic relationship between two individuals is calculated as follows:

$$r_{ij} = \frac{\sum_{t=1}^{c} (m_{it} - 2p_t)(m_{jt} - 2p_t)}{2\sum_{t=1}^{c} p_t(1 - p_t)}, \qquad (1)$$

where r_{ij} is the genomic relationship between individuals i and j, m_{it} and m_{jt} are the genotypes (coded 0, 1 or 2 for the alternative homozygous, the heterozygous, and the homozygous reference genotypes, respectively) for individuals i and j at locus t, p_t is the allele frequency in the population at locus t, and c is the number of loci (i.e. SNPs). Genomic relationships can be calculated even for extremely dense genomic data (even up to full sequence), and do not assume independence of the loci. Figure 1 shows the relationships in a trio consisting of an offspring and two (candidate) parents.

We used Eq. (1) to estimate the genomic interrelationships between parents and offspring, i.e., the relationship of the offspring with itself ($r_{O,O}$), relationships of the two parent candidates with themselves (r_{P_1,P_1} and r_{P_2,P_2}), relationships of the offspring with both parent candidates (r_{O,P_1} and r_{O,P_2}), and relationships between the parent candidates (r_{P_1,P_2}), see Fig. 1.

Expected genomic relationships of an offspring with its true parents (TP) are [11]:

$$E\left(r_{O,P1}|TP\right) = 0.5\left(r_{P1,P1} + r_{P1,P2}\right),$$

$$E\left(r_{O,P2}|TP\right) = 0.5\left(r_{P2,P2} + r_{P1,P2}\right).$$

In other words, the relationship of an offspring with a parent is the average of the genomic relationship of the parent with itself and the relationship between the two parents. The expected relationship of the offspring with itself is [12]:

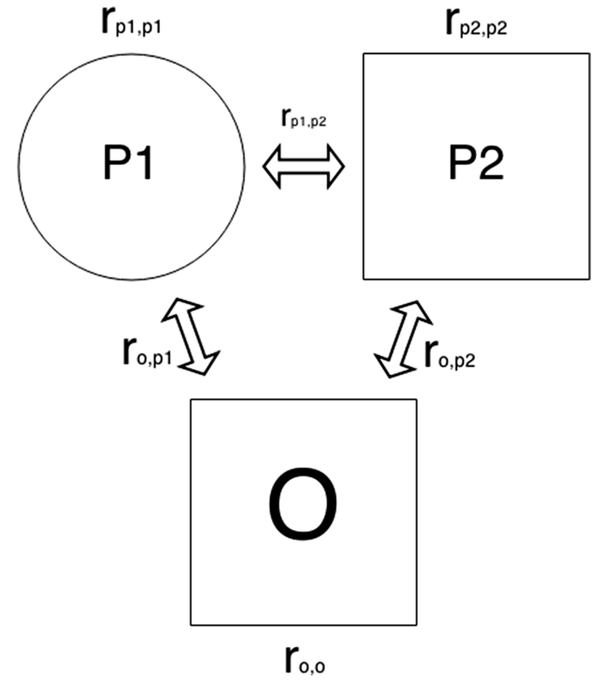

Fig. 1 A trio of offspring (O), first parent (P1) and second parent (P2). The variables near the arrows indicate genetic relationships between individuals, while the variables over P1 and P2, and below O, are the individuals' genetic relationships to themselves, respectively. Sexes are included in the figure but are not used by the GRL method

$$E\left(r_{O,O}|TP\right) = 1 + 0.5r_{P1,P2},$$

where $0.5r_{P1,P2}$ is the expected inbreeding coefficient of the offspring. Three residual relationships are defined as differences between actual and expected genomic relationships:

$$e_{O,P1} = r_{O,P1} - E\left(r_{O,P1}|TP\right),$$

$$e_{O,P2} = r_{O,P2} - E\left(r_{O,P2}|TP\right),$$

$$e_{O,O} = r_{O,O} - E\left(r_{O,O}|TP\right).$$

Inbreeding is accounted for when using the above residuals, as well as the direction of the relationships. For example, using the offspring as a candidate parent, and/or using a true parent as the offspring, will result in large residuals, i.e., realized relationships that deviate substantially from the expectations of a true parent–offspring trio.

Genomic relationship likelihood (GRL)

The above residual relationships are used to calculate a genomic relationship log-likelihood using a multivariate normal density function, assuming:

$$\mathbf{e} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\mathbf{e} = \begin{bmatrix} e_{O,P1} \\ e_{O,P2} \\ e_{O,O} \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$ is a vector of the overall means for the residuals for true parent–offspring trios. In the absence of genotyping errors, the residuals are expected to be approximately normally distributed around zero ($\mathbf{e} \sim N(0, \boldsymbol{\Sigma})$), see [Additional file 1: Figure S1]. The central limit theorem states that the sum of many independently and identically distributed variates will be approximately normally distributed. The variates in Eq. (1) may be considered as originating from a common (albeit unknown) distribution, but not all are independent (i.e., the effective number of loci is lower than the actual number of loci). Still, given a substantial number of loci distributed over the entire genome (i.e., most of the loci are indeed independent), genomic relationships (summed over all variates) are still likely to approach a normal distribution (see [13], Theorem 27.4). Plotting the residual relationships for true parent–offspring trios revealed that they were approximately normally distributed [see Additional file 1: Figures S1, S2 and S3].

Since genotyping errors can occur in real data (and the expected residual relationship may thus deviate from 0), parameters of the distribution of residual relationships were estimated using an iterative method (see Section "Estimation of model parameters" below). Matrix $\boldsymbol{\Sigma}$ is the 3×3 (co)variance matrix of the three residual variates in true parent–offspring trios and was also estimated using the iterative method. The genomic relationship likelihood (GRL) was defined as:

$$\mathrm{GRL} = -\frac{1}{2}(\mathbf{e} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{e} - \boldsymbol{\mu}),$$

which is proportional to the natural logarithm of a multivariate normal density function. Based on (iteratively assigned) parent–offspring trios, a threshold for acceptable GRL values can be defined. In this study, we assumed that a parent–offspring trio had to have a GRL value that was within the highest 99% of the known parent–offspring GRL values, thus accepting a false negative rate of 1%.

Difference between the top two trios (ΔGRL)

To reduce the false positive rate and increase the true negative rate, the value of ΔGRL was also assessed based on:

$$\Delta\mathrm{GRL} = \mathrm{GRL}_1 - \mathrm{GRL}_2,$$

where GRL_1 (GRL_2) is the (second) highest GRL value achieved for an offspring across all candidate parent–offspring trios. This is analogous to the Δ statistic used in Marshall et al. [7], with more details in Appendix 1.

In datasets where both parents of an offspring are present and no other relatives are available, ΔGRL will typically be very high, since no other realistic trio exists. When other close relatives of the offspring are included among the candidate parents, ΔGRL may be lower due the potential existence of multiple "likely" false parent candidates, e.g. uncles, aunts, grandparents, siblings or descendants of the offspring. High relatedness to the offspring alone is not sufficient to obtain a high value for GRL_2 since the method accounts for interrelationships of the whole trio. For example, if the parent candidates consist of one true parent and one full-sib of the offspring, interrelationships of the trio will typically be inconsistent because of the high relationship between the two parental candidates, although the relationships of the offspring with itself and with the parent candidates may be "normal" (these should be elevated if the relationship among the two parent candidates is high). In cases where a parent is missing but many other close relatives of the offspring are present, GRL_1 can, in rare cases, exceed the threshold for GRL_1-values, but then ΔGRL will typically be low, since multiple highly-related candidate parents are present. Thus, thresholds for assignment must be set for both GRL_1 and ΔGRL.

Estimation of model parameters

Estimation of the GRL-parameters, i.e. $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the GRL threshold, is undertaken with an iterative method which is briefly described below. The ΔGRL threshold was set to 6.9, which implies that the best parent pair should be at least 1000 ($= e^{6.9}$) times more likely than the second-best parent pair. See Section 2 in Additional file 2: for more details.

Step 1: allele dropping

Random matings between individuals from the dataset are performed in silico to produce simulated offspring. For simplicity, all loci are assumed to be inherited independently. The simulated trios are then used to obtain initial estimates of the GRL parameters. A smaller subset of the loci may be used in this step.

Step 2: assignment iteration

Trios are initially assigned using the GRL method based on the parameters estimated in Step 1. The method relies on the presence of true trios (albeit unknown) in the data. Parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are then re-estimated using the newly assigned trios from evaluation data, and then used as the basis of the next assignment iteration. Iteration stops when the number of assignments is smaller than in the previous iteration. Thus, the GRL training procedure iteratively assigns trios while (re-)estimating the GRL-parameters until no more trios can be assigned. See Section 1 in Additional file 2: for more information about the training procedure.The parameter estimates obtained in the second-to-last iteration are considered optimal. To limit the number of plausible trios to test, only individuals with a relationship larger than 0.25 with an offspring were considered as potential parents, i.e. $r_{O,P1} > 0.25$ and $r_{O,P2} > 0.25$. The GRL threshold is not re-estimated in this step.

When pre-defined parameter estimates are used, the assignment process starts without estimating parameters. This is equivalent to running only the second-to-last iteration of Step 2.

Simulation study

A simulation study was conducted to investigate the strengths and weaknesses of the GRL method. QMSim [14] was used to produce simulated datasets. The initial size of the historical population was set to 500 and remained constant for 5000 generations to achieve mutation/drift equilibrium. In generation 5001, the population size was reduced to 300, of which 100 were males and 200 were females. Twenty chromosomes were simulated, each 1 Morgan long, and the number of SNPs was set such that approximately 54,000 SNPs (53,427 to 55,517) with a minor allele frequency higher than 0.05 existed in the population. The SNP mutation rate was set to 0.00003, assuming a recurrent mutation model (i.e. only two possible alleles exist). After the historical population, a recent population was simulated over five generations, with 1000 individuals per generation (5000 individuals in total). These were produced by random mating of 100 sires and 200 dams per generation, with one sire mated with two dams and each mating resulting in five recorded offspring. Of these, the last two generations were used in the parentage assignment tests. Fifty repetitions of the QMSim simulations were performed to produce 50 datasets. Genotype errors (1 and 3%) and call rates (80–100%) were added using a custom script written in the Python programming language, allowing both erroneous and missing genotypes among individuals, see Section 2 in Additional file 2: for more information.

The GRL method was programmed in the C++ programming language that emphasizes parallel processing. The program was run in a Linux cluster environment using multiple CPU. Tests were run using the training procedure on all (evaluation) datasets. In addition, pre-estimated parameters were obtained from some of the runs with training. The datasets were not divided into offspring and parents, and thus all true offspring and parents had the potential to be assigned parents both correctly (offspring only) and incorrectly (parents and offspring).

There are three possible outcomes of the assignment process: (1) 'Correct', meaning correct assignment of true parents to the unknown offspring (parents must be present), (2) 'Incorrect', meaning wrong candidate parents were assigned and (3) 'No-assign', meaning no assignment was made. These were quantified for each analysis.

Comparison with a conventional likelihood-based method

To compare GRL with other methods, we analyzed five of the simulated datasets, arbitrarily chosen from all 50 datasets, by using the Colony2 software V2.0.6.3 [15]. Colony2 uses a likelihood-based method that jointly assigns both sibship and parentage based on a simulated annealing process [16, 17]. This increases the assignment power compared to methods that use a single unknown individual (the offspring) and one or two candidate parents. Colony2 was run using a 1% genotype error (true and assumed). In addition, the following settings were chosen: (1) do not update allele frequencies, (2) assume no inbreeding, (3) no sibship scaling, (4) no sibship prior, (5) short run length, (6) use the pairwise likelihood score (PLS) and (7) allelic dropout rate set to zero for all markers. The 'ParentPairs'-file produced by Colony2 was used to check accuracy of assignments. Any assignments for which mother, father or both were missing, or for which the assignment probability reported by Colony2 was less than 0.5, were categorized as a "No-assign". Suggested parent pairs with at least one incorrect parent were categorized as "Incorrect" assignments and pairs with both parent candidates correct were categorized as "Correct" assignments.

Comparison with an exclusion-based method: the binomial exclusion method

We developed an exclusion-based method in which one of the parameters was estimated using GRL-assigned trios using custom scripts written in the R programming language. Exclusion ratios (ER) for the GRL-assigned trios were calculated as the ratio of the number of exclusions for a trio and the number of loci for which all three individuals in the trio had called genotypes. We used a binomial distribution as a basis for the new assignments, i.e. $E \sim Bin(n, p)$, where E is the number of trio exclusions, n (number of trials) is the number of calls for the trio, and p (success probability) is the median ER from the GRL assigned trios.

To limit the number of trios for binomial exclusion assignment, we used the same parent–offspring genomic relationship threshold that we used for the GRL assignments, i.e. $r_{O,P1} > 0.25$ and $r_{O,P2} > 0.25$. Assignment was done in a similar manner as with GRL, using both a confidence cutoff and a Δ-score. For more information, see Section 3 in Additional file 2:. We refer to this method as the binomial exclusion method (BEM) in the text.

Results

Assignment results using Colony2 are shown in Fig. 2, and the analogous GRL- and BEM results are shown in Figs. 3 and 4. The most noticeable differences in results between GRL- and BEM are shown in Figs. 5 and 6. Here, both methods used training estimates from a dataset with a 3% genotype error, while the true error was 1%. Results that were similar between GRL and BEM are shown in Figures S4, S5, S6, S7, S8, S9, S10 and S11 [see Additional file 3: Figures S4, S5, S6, S7, S8, S9, S10 and S11]. In Figures S4 (GRL) and S5 (BEM), parameters were pre-estimated at a 3% genotype error (true and assumed). Figures S6 (GRL) and S7 (BEM) show the results for a true error of 3% and an assumed error of 1%. Figures S8 (GRL) and S9 (BEM) show the results for training with a 1% error rate, and Figures S10 (GRL) and S11 (BEM) for training with a 3% error rate. Total results over all datasets are shown in Table S1 [see Additional file 4: Table S1].

The Colony2 software was tested using a 1% true genotype error rate (assumed and true). When parents are available, Colony2 had a correct assignment rate of 22.4%, a no-assign rate of 75.4% and an incorrect assignment rate of 2.2%. For individuals without parents, the incorrect assignment rate climbed to 14.7% and the (correct) no-assign rate climbs to 85.3% (see Fig. 2).

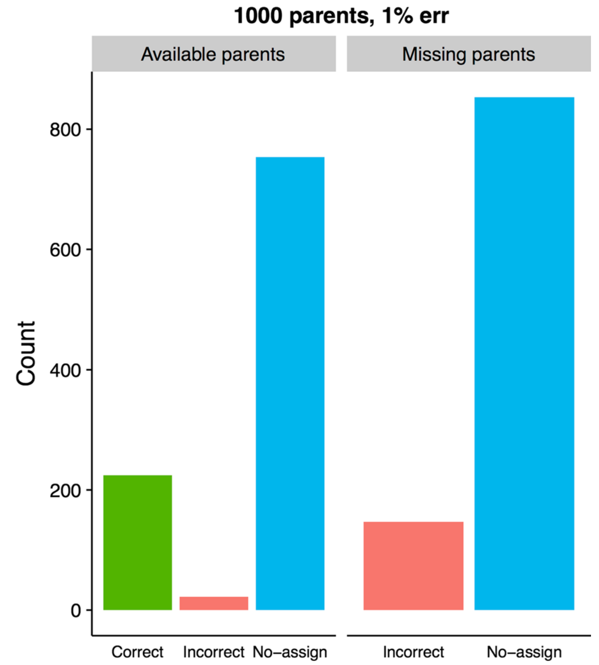

Fig. 2 Assignment results from Colony2 for individuals with (left panel) and without (right panel) available parents in the dataset. Results from five simulated datasets are averaged. The true and assumed genotype error rate was 1% for all datasets

Figures 3 and 4 show the comparison between GRL and BEM when parameter estimates from an arbitrarily chosen dataset were used. When parents were available in the dataset and the genotype error rate (true and assumed) was 1%, using GRL resulted in 99.5% of the individuals being correctly assigned both parents (Fig. 3), while 99.9% were assigned correctly with the (GRL-trained) BEM (Fig. 4). In both cases, no individuals with parents in the dataset were assigned incorrect parent pairs. When parents were not available, the incorrect assignment rate for GRL climbed to 0.01% for both 1% and 3% genotype error rates (Fig. 3 and Additional file 3: Figure S4).

The most notable difference in results between GRL and BEM was found for a true genotype error rate of 1% when parameter estimates were from a dataset with a 3% error rate (Figs. 5 and 6). Here, GRL did not assign any trios. However, BEM assigned all trios correctly when parents were available, but incorrectly assigned 1.0% of the trios when parents were not available. When the true and assumed genotype error rates were reversed (i.e. a true error rate of 3% and an incorrectly assumed error rate of 1%), neither method assigned any

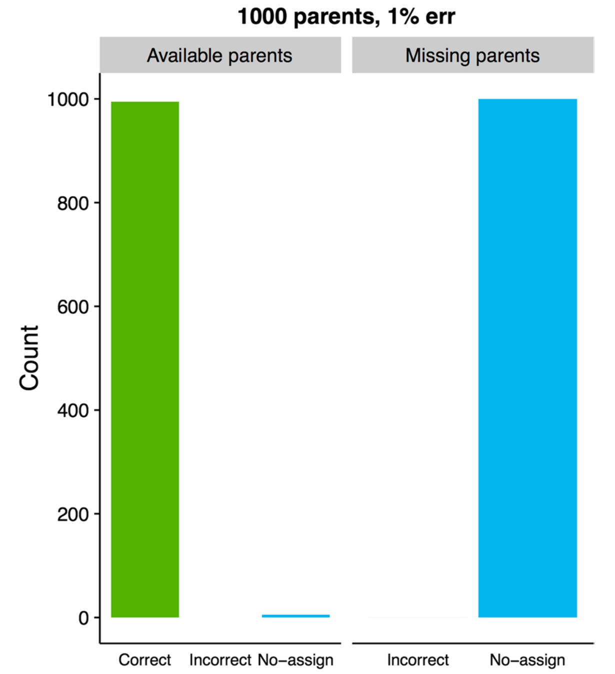

Fig. 3 Assignment results using GRL at a 1% genotype error rate (true and assumed) for individuals with (left panel) and without (right panel) available parents in the dataset. Results from 50 simulated datasets are averaged. Parameters were pre-estimated using one arbitrarily chosen dataset with a 1% genotype error

trios, while the GRL method incorrectly assigned 0.02% trios, both when parents were available and when they were missing [see Additional file 3: Figures S6 and S7] and [see Additional file 4: Table S1].

An alternative to assuming a set of predefined parameters is to estimate these by using the evaluation data directly. Averaged results for each dataset are shown in Figures S8 and S9 [see Additional file 3: Figures S8 and S9] (1% genotype error) and in Figures S10 and S11 [see Additional file 3: Figures S10 and S11] (3% genotype error). These results are very similar to the results shown in Figs. 3 and 4 (1% true and assumed error rates), and Figures S4 and S5 [see Additional file 3: Figures S4 and S5] (3% true and assumed error rates).

Discussion

Parentage assignment is mostly performed using likelihood-based models with microsatellites [2, 7], low-density SNPs [1] or exclusion-based models [18]. However, assignments methods often impose idealized assumptions, such as known age, generation and gender of all individuals, a limited number of known parental candidates, independent markers, little or no inbreeding, no stratification of the population or sample, no biased sampling of individuals, Hardy–Weinberg equilibrium (HWE) and little or no variation in genotype error or call rates within and between samples. For GRL and BEM, we perfomed assignments with unknown age, generation and gender, with no assumption as to independence of markers, HWE, inbreeding, family size or family composition, and with dense (SNP) markers, closely related individuals and varying genotype error and call rate. Colony2 assumes HWE, independent markers and no inbreeding.

GRL

Residual relationships were approximately normally distributed even when genotype errors were present [see Additional file 1: Figures S2 and S3], but with different expectations compared to genomic data without genotype errors [see Additional file 1: Figure S1].

It did not appear to be a problem that the parent and offspring generations were unknown when using GRL and BEM. High accuracies were achieved, although individuals had numerous close relatives that were eligible as parent candidates, such as the true parents, full- and half-sibs, own offspring, uncles/aunts and nieces/nephews. Similar results were obtained when the genotype error was increased to 3%, which was used to show that the GRL and BEM work even when the genotype error rate has quite extreme values. These properties may be useful for populations with large sibling groups, such as in fish, poultry or pigs, when generations cannot be clearly differentiated, or when the genotype error or call rates vary a lot.

A strength of the GRL training procedure is that no reference dataset with known pedigree is required for training and that the training is only partly done by simulation (allele-dropping). As long as there is a sufficient number of true (but unknown) trios present for assignment, the training can proceed. The method requires a pre-defined ΔGRL threshold (i.e. the minimum acceptable value). The ΔGRL is (the log of) the odds for correct assignment, given that the correct trio is among the two best trios (this is nearly always the case if true parents are present). In this study, the threshold was set to 6.9, i.e., the best trio should be at least $e^{6.9} = 1000$ times more likely than the second-best trio. Relaxing this assumption will increase both the true and false positive assignment rates of the model, while setting a stricter threshold will have the opposite effect.

In some cases, the iterative training method may fail because the initial iteration results in no assignments. This may be caused by two factors: (1) the number of loci used in the allele-dropping simulation step may be set too high (giving too idealized parent–offspring relationships compared with evaluation data), or (2) there are no

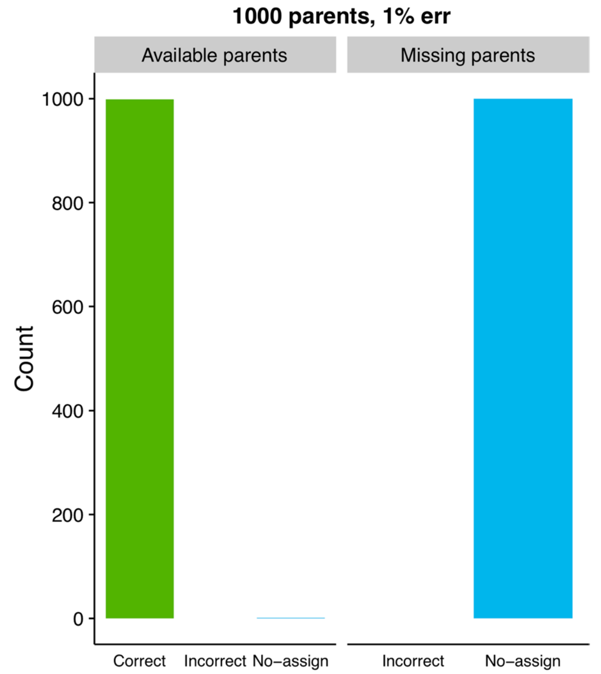

Fig. 4 Assignment results using BEM at a 1% genotype error rate (true and assumed) for individuals with (left panel) and without (right panel) available parents in the dataset. Results from 50 simulated datasets are averaged. Parameters were pre-estimated using the GRL-assignments from one arbitrarily chosen dataset with a 1% genotype error

true trios present in the evaluation dataset. If reducing the number of SNPs used in the allele-dropping step does not start the iteration process, the latter may be the case. During training, there is no need to estimate or assume a genotype error rate with the GRL method, as long as the training procedure is done using the evaluation dataset.

Exclusion using parent–offspring duos (i.e. offspring and a single candidate parent) or trios is a relatively simple method for parentage assignment, by identifying incorrect parents by genotypes that violate the laws of Mendelian inheritance ("exclusion genotypes"). The GRL method is a fundamentally different approach and can be used to estimate exclusion-based parameters in true parent–offspring trios (assigned by GRL). Assignment of a single parent to an offspring is also possible using a similar method as for trios, but this was not explored in this study. The training-based GRL has the advantage that it requires no prior assumption with respect to genotype error rate or expected number of exclusions.

Binomial exclusion method

Estimation of the p-parameter for the BEM was done using trios that were assigned using GRL. An alternative to using GRL-assignments is using a training dataset with genotyped trios and known pedigree. Such a training dataset would need to have a similar genotype error rate as the evaluation dataset since having a discrepancy between the true and assumed genotype error rate could lead to decreased accuracy [see Additional file 4: Table S1]. Since pedigree information is not always reliable, we prefer to use GRL assignments (preferably using a relatively big dataset) for parameter estimation.

Comparing GRL and the binomial exclusion method with Colony2

The GRL and BEM resulted in much more accurate assignments of parents than Colony2. Parameters for Colony2 were chosen to minimize running time, so assignment accuracy may be improved by adjusting the parameters, but at the expense of time and/or computing resources required to perform the analysis. Colony2 incorrectly assumes that marker loci are independently distributed, while GRL and BEM do not. This is likely the main reason for the poor results obtained with Colony2 on these relatively dense marker datasets.

Comparing GRL with the binomial exclusion method

Using BEM resulted in a slightly higher accuracy than GRL when the genotype error assumption was correct, or when GRL-parameters were estimated using the evaluation data (Figs. 3 and 4) and [see Additional file 3: Figures S4, S5, S8 and S9]. However, when pre-estimated model parameters are used, assuming a too high genotype error rate will lead to some false assignments with BEM (Fig. 6), and assignment failure for the GRL method (Fig. 5). Thus, GRL can be used when it is crucial to minimize the false-positive rate. Assuming a too low genotype error rate resulted in both methods failing to correctly assign any trios, but GRL had a small fraction (0.016%) of false assignments while BEM did not [see Additional file 4: Table S1]. Although the success parameter (p, see Methods) of BEM was estimated using already GRL-assigned trios, the results indicate that the two methods are somewhat complementary and can be used together to increase overall assignment accuracy.

When the assumed genotype error rate was correct (Figs. 3 and 4) and [see Additional file 3: Figures S4 and S5] or when the evaluation dataset was used to estimate parameters [see Additional file 3: Figures S8, S9, S10 and S11], nearly all the individuals were assigned correctly and there were hardly any false assignments with either method. Thus, parameters should be estimated using the available data whenever possible, which should be the case in most situations.

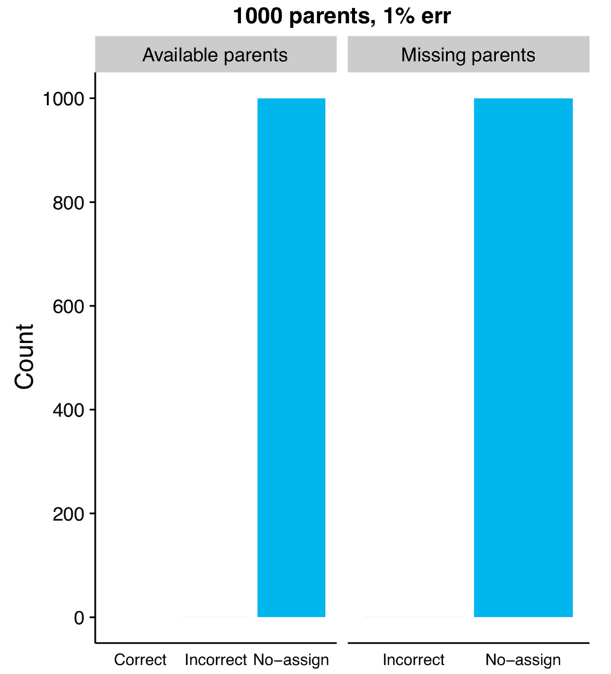

Fig. 5 Assignment results using GRL for a 1% true genotyping error rate but using parameter estimates from a dataset with 3% genotype errors. Individuals with (left panel) and without (right panel) available parents are present in the dataset. Results from 50 simulated datasets are averaged

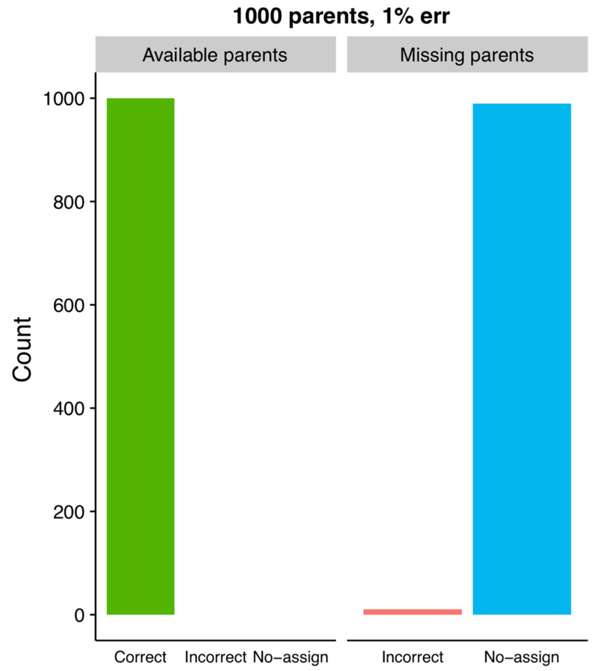

Fig. 6 Assignment results using BEM for a 1% true genotyping error rate but using parameter estimates using GRL-assignments from a dataset with 3% genotype errors. Individuals with (left panel) and without (right panel) available parents are present in the dataset. Results from 50 simulated datasets are averaged

Using GRL with clones or duplicated DNA

A possible novel use for the GRL method is analysis of genomic data that contain possibly duplicated genomes (e.g., by sampling of clones in plants or monozygotic twins in animals, or by duplicated sampling of DNA from the same individual). Using traditional likelihood-based or exclusion-based methods, duplicated samples/clones should be removed prior to the analysis, as these may be assigned as their own parents. For the GRL method, duplication of offspring genotypes is not a problem since GRL looks at patterns in parent–offspring relationships rather than the likelihood of each single genotype. For example, if clones of a non-inbred offspring are inserted as one or both putative parents, the GRL method would expect the offspring to be highly inbred, which will not match the observed relationship of the offspring with itself, and thus yields a low GRL value. However, duplication of parental genotypes will inevitably lead to assignment failure, since two or more trios will appear equally likely.

Conclusions

The GRL method is a promising trio parentage assignment method which is well suited to perform parentage assignment with high accuracy on high-density SNP datasets. GRL can be applied with success on datasets with high and/or unknown genotype error rates, highly dependent marker loci, closely-related individuals, inbreeding and in some cases clones. Estimation of the GRL parameters can be done without having a pre-existing reference dataset with known parent–offspring trio combinations. In addition, GRL can be used for training of exclusion-based methods.

Additional files

Additonal file 1: Figures S1, S2 and S3. Residual relationships plotted for all true trios from the 50 datasets. This file contains three figures (Figures S1, S2 and S3). Residual densities for offspring to itself (top panel), offspring to real mother (mid panel) and offspring to real father (bottom panel) are shown as a continuous line in all Figs. 50,000 values were sampled from the normal distribution using the means and variances of the residuals as parameters, shown as a dashed line in each panel. Figure S1 shows results in which there is no genotype error or call rate variance, Figure S2 in which there is 1% genotype error and a ~ 80 to 100% call rate and Figure S3 in which there is a 3% genotype error and a ~ 80 to 100% call rate.

Additonal file 2. Supplementary material. This file contains three sections with extended information about the GRL training procedure, call rate and genotype error simulation, and the binomial exclusion method (BEM), respectively.

Additonal file 3: Figures S4, S5, S6, S7, S8, S9, S10 and S11. Assignment results using GRL or BEM for individuals with (left panel) and without (right panel) available parents in the dataset. This file contains eight figures in which assignment results from 50 simulated datasets are averaged. Parameters were pre-estimated using one arbitrarily chosen dataset in Figures S4, S5, S6 and S7, while training was performed on each evaluation dataset in Figures S8, S9, S10 and S11. Figures S4, S6, S8 and S10 show results using GRL, while Figures S5, S7, S9 and S11 show results using BEM. Figures S4 and S5 show results when there is a 3% genotype error (true and assumed), Figures S6 and S7 have pre-esimated parameters from a dataset with a 1% genotype error, while the (true) evaluation genotype error is 3%. Figures S8, S9, S10 and S11 use training on each evaluation dataset, both at 1% (Figures S8 and S9) and 3% (Figures S10 and S11) genotype errors. In all figures, the call rates are ~ 80 to 100%.

Additonal file 4: Table S1. Summary table of total number of correct, incorrect and non-assigned trios with or without parents and genotype errors for all 50 datasets. Genotype error: either 1% or 3%, and with assumption of genotype error in parenthesis (only applicable for models that are pre-trained). Available parents: all individuals with parents available for assignment in the dataset (Yes) or where all parents are missing (No). Correct: Number of correctly assigned individuals over all 50 datasets (only applicable when parents are available). Incorrect: Number of incorrectly assigned individuals over all 50 datasets. No-assign: Number of individuals that could not be assigned parents over all 50 datasets.

Authors' contributions

KEG wrote the software, performed the study and drafted the manuscript. JO conceived the GRL method, coordinated the whole study and contributed in writing and revising the manuscript. THEM helped finalize the theory behind the training portion of the GRL method as well as revising the manuscript critically. All authors read and approved the final manuscript.

Author details

[1] AquaGen AS, P.O. Box 1240, NO-7462 Trondheim, Norway. [2] Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway.

Acknowledgements

The research leading to these results has received funding from The Research Council of Norway through both the NAERINGSPHD (Project No. 251664) and the HAVBRUK2 (Project No. 245519) programs, as well as the breeding company AquaGen AS. The authors thank Thore Egeland for helpful comments on an early version of the draft and Jinliang Wang for providing support for the Colony2 software. We also wish to thank the editors and reviewers, especially reviewer 2 whose comments lead to a significant increase in the quality of the end result.

Competing interests

KEG and JO are employed by AquaGen AS. AquaGen has applied for a patent regarding the use of the GRL methodology in parentage assignment.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

The research leading to these results has received funding from The Research Council of Norway through the research programs NAERINGSPHD (Project No. 251664) and the HAVBRUK2 (Project No. 245519), and AquaGen AS.

Appendix

Mathematical foundation for the GRL method

In this article, only the hypothesis of true parents is used for the GRL method:

$$H_1 : \textit{Both candidate parents are the true parents of the child}.$$

We assume $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathbf{x}$ is the vector of residual genomic relationships, i.e. it holds the residual values for trio assignments. We define x_1 as being the most probable trio, while x_2 is the second most probable trio, that is $P(x_1|H_1) \geq P(x_2|H_1)$.

The difference $\Delta_{\mathrm{GRL}} = \mathrm{GRL}_1 - \mathrm{GRL}_2$, where GRL_1 and GRL_2 refer to the best and the second best trio candidates, respectively, can be shown to be identical to the natural logarithm of the probability of observing x_1 given H_1 divided by the probability of observing x_2 given H_1. Since $\mathbf{x}$ is assumed to be normally distributed, the multivariate normal probability density function used is:

$$f(\mathbf{x}|H_1) = \frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

where $\mathbf{x}$ is the 3×1 vector of genomic residuals, $\boldsymbol{\mu}$ is 3x1 vector of expected residuals and $\boldsymbol{\Sigma}$ is the 3x3 covariance matrix. If we define $\frac{L_1}{L_2} = \frac{f(x_1|H_1)}{f(x_2|H_1)}$ (i.e. how many times more likely is x_1 given H_1 compared to x_2 given H_1), we find that:

$$\frac{f(\boldsymbol{x}_1|H_1)}{f(\boldsymbol{x}_2|H_1)} = \frac{\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu})}}{\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu})}}$$
$$= \frac{e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})}}{e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})}}.$$

If we take the natural logarithm of this ratio we get:

$$\begin{aligned}
\ln\left[\frac{f(x_1|H_1)}{f(x_2|H_1)}\right] &= \ln\left[\frac{e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})}}{e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})}}\right] \\
&= \ln\left(e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})}\right) \\
&\quad - \ln\left(e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})}\right) \\
&= -\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}) \\
&\quad - \left(-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})\right) \\
&= \mathrm{GRL}_1 - \mathrm{GRL}_2 = \Delta_{\mathrm{GRL}}.
\end{aligned}$$

The above formula shows that Δ_{GRL} has a logarithmic point probability ratio expectation. We can compare this to the $\Delta_{Marshall}$ test statistic which is defined as in [7], that is:

$$\text{LOD} = \ln(LR) = \ln\left(\frac{P(data|H_1)}{P(data|H_2)}\right),$$

where H_2 is defined as:

$$H_2 : \textit{Two random individuals are the true parents of the child}.$$

LOD_1 is defined to be the LOD-score of the most likely trio, while LOD_2 is the second most likely trio. Then:

$$\Delta_{\text{Marshall}} = \text{LOD}_1 - \text{LOD}_2 = \ln(LR_1) - \ln(LR_2).$$

Both $P(data|H_1)$ and $P(data|H_2)$ can be written as follows:

$$P(data|H_1) = \prod_{t=1}^{c} P_t\left(g_C|g_F, g_M, H_1\right) * P_t\left(g_F\right) * P_t\left(g_M\right),$$
$$P(data|H_2) = \prod_{t=1}^{c} P_t\left(g_C|H_2\right) * P_t\left(g_F\right) * P_t\left(g_M\right),$$

where $P_t\left(g_C|g_F, g_M, H_1\right)$ is the probability of observing the offspring genotype given the father and mother genotypes under H_1 at locus t, $P_t\left(g_F\right)$ is the probability of observing the father genotype at locus t, $P_t\left(g_M\right)$ is the probability of observing the mother genotype at locus t, $P_t\left(g_C|H_2\right)$ is the probability of observing the offspring genotype under H_2 at locus t and c is the number of loci. Since $LR = \frac{P(data|H_1)}{P(data|H_2)}$, we can simplify LR to be:

$$LR = \frac{P(data|H_1)}{P(data|H_2)} = \prod_{t=1}^{c} \frac{P_t\left(g_C|g_F, g_M, H_1\right) * P_t\left(g_F\right) * P_t\left(g_M\right)}{P_t\left(g_C|H_2\right) * P_t\left(g_F\right) * P_t\left(g_M\right)}$$
$$= \prod_{t=1}^{c} \frac{P_t\left(g_C|g_F, g_M, H_1\right)}{P_t\left(g_C|H_2\right)}.$$

Since LR_1 is the likelihood ratio of the most likely trio and LR_2 is the likelihood ratio of the second most likely trio (defined above), we can write LR_1 and LR_2 as:

$$LR_1 = \frac{P(data_1|H_1)}{P(data_1|H_2)} = \prod_{t=1}^{c} \frac{P_t\left(g_C|g_{F_1}, g_{M_1}, H_1\right)}{P_t\left(g_C|H_2\right)},$$

and

$$LR_2 = \frac{P(data_2|H_1)}{P(data_2|H_2)} = \prod_{t=1}^{c} \frac{P_t\left(g_C|g_{F_2}, g_{M_2}, H_1\right)}{P_t\left(g_C|H_2\right)},$$

where g_{F_1} and g_{M_1} are the genotypes of the father and mother in the most likely trio at locus t, respectively, and g_{F_2} and g_{M_2} are the genotypes of the father and mother at locus t in the second most likely trio, respectively. Since the same offspring is used in both trios, g_C is the same for both LR_1 and LR_2 for locus t.

Inserting LR_1 and LR_2 into the $\Delta_{Marshall}$-formula above we get:

$$\Delta_{Marshall} = \ln(LR_1) - \ln(LR_2) = \ln\left(\frac{LR_1}{LR_2}\right)$$
$$= \ln\left(\frac{\prod_{t=1}^{n} \frac{P_t\left(g_C|g_{F_1}, g_{M_1}, H_1\right)}{P_t\left(g_C|H_2\right)}}{\prod_{t=1}^{n} \frac{P_t\left(g_C|g_{F_2}, g_{M_2}, H_1\right)}{P_t\left(g_C|H_2\right)}}\right)$$
$$= \ln\left(\prod_{t=1}^{n} \frac{P_t\left(g_C|g_{F_1}, g_{M_1}, H_1\right)}{P_t\left(g_C|g_{F_2}, g_{M_2}, H_1\right)}\right)$$
$$= \ln\left(\frac{P\left(\mathbf{g}_C|\mathbf{g}_{F_1}, \mathbf{g}_{M_1}, H_1\right)}{P\left(\mathbf{g}_C|\mathbf{g}_{F_2}, \mathbf{g}_{M_2}, H_1\right)}\right),$$

where the explanation for g_C, g_{F_1}, g_{M_1}, g_{F_2}, g_{M_2} is the same as above, while $\mathbf{g}_C$, $\mathbf{g}_{F_1}$, $\mathbf{g}_{M_1}$, $\mathbf{g}_{F_2}$ and $\mathbf{g}_{M_2}$ are the genotypes for the offspring (or child), for most probable father and mother and for the second most probable father and mother, respectively, over all loci in vector-notation. The $\Delta_{Marshall}$ method only uses the probability of observing the child genotypes given that F_1 and M_1, or F_2 and M_2 are the true parents. The fact that the information in the H_1 hypothesis is not used makes the $\Delta_{Marshall}$ method similar to Δ_{GRL}, we see this when the two method definitions are compared:

$$\textit{GRL method} : \Delta_{GRL} = \ln\left[\frac{f(\boldsymbol{x}_1|H_1)}{f(\boldsymbol{x}_2|H_1)}\right],$$

$$\textit{Marshall method} : \Delta_{Marshall} = \ln\left(\frac{P\left(\mathbf{g}_C|\mathbf{g}_{F_1}, \mathbf{g}_{M_1}, H_1\right)}{P\left(\mathbf{g}_C|\mathbf{g}_{F_2}, \mathbf{g}_{M_2}, H_1\right)}\right).$$

Both methods produce an estimated logarithmic ratio of the probability that C is the child of the two most probable parent candidates versus the probability that C is the child of the two second most probable parent candidates, hence the results produced by the two methods can be considered analogous.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 16 October 2017 Accepted: 4 May 2018
Published online: 18 May 2018

References

1. Heaton MP, Leymaster KA, Kalbfleisch TS, Kijas JW, Clarke SM, McEwan J, et al. SNPs for parentage testing and traceability in globally diverse breeds of sheep. PLoS One. 2014;9:e94851.
2. Waldbieser GC, Bosworth BG. A standardized microsatellite marker panel for parentage and kinship analyses in channel catfish, *Ictalurus punctatus*. Anim Genet. 2013;44:476–9.
3. Campbell D, Duchesne P, Bernatchez L. AFLP utility for population assignment studies: analytical investigation and empirical comparison with microsatellites. Mol Ecol. 2003;12:1979–91.
4. Jones AG, Small CM, Paczolt KA, Ratterman NL. A practical guide to methods of parentage analysis. Mol Ecol Resour. 2010;10:6–30.
5. Morrissey MB, Wilson AJ. The potential costs of accounting for genotypic errors in molecular parentage analyses. Mol Ecol. 2005;14:4111–21.
6. Strucken EM, Lee SH, Lee HK, Song KD, Gibson JP, Gondro C. How many markers are enough? Factors influencing parentage testing in different livestock populations. J Anim Breed Genet. 2016;133:13–23.
7. Marshall TC, Slate J, Kruuk LE, Pemberton JM. Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol. 1998;7:639–55.
8. Purfield DC, McClure M, Berry DP. Justification for setting the individual animal genotype call rate threshold at eighty-five percent. J Anim Sci. 2016;94:4558–69.
9. Goddard ME, Hayes BJ, Meuwissen TH. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011;128:409–21.
10. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
11. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Harlow: Longman group Ltd; 1996.
12. Malécot G. Les mathématiques de l'hérédité. Paris: Masson; 1948.
13. Billingsley P. Probability and measure. 3rd ed. New York: Wiley; 1995.
14. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25:680–1.
15. Jones OR, Wang J. COLONY: a program for parentage and sibship inference from multilocus genotype data. Mol Ecol Resour. 2010;10:551–5.
16. Wang J, Santure AW. Parentage and sibship inference from multilocus genotype data under polygamy. Genetics. 2009;181:1579–94.
17. Wang J. Computationally efficient sibship and parentage assignment from multilocus marker data. Genetics. 2012;191:183–94.
18. Hayes BJ. Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J Dairy Sci. 2011;94:2114–7.