Proc. Assoc. Advmt. Anim. Breed. Genet. 24:341-344

ASSESSING THE POTENTIAL OF PARENTAGE TESTING USING PORTABLE LONG READ SEQUENCING TECHNOLOGIES

E.M. Ross[1], H.J. Lamb[1], B.N. Engle[1], L.T. Nguyen[1] and B.J. Hayes[1]

[1] Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, QLD, 4072 Australia

SUMMARY

Parentage testing in cattle based on DNA markers generates pedigree information for predicting breeding potential, as well as for the inclusion of animals in breed specific stud books. Here we present a method that could potentially be used for parentage assignment of calves. We propose the use of the portable minION (Oxford Nanopore Technologies; ONT) sequencer to achieve the parental assignments. The method uses ultra-low-coverage sequence data and a combination of genetic distances taken from a genomic relationship matrix, and simple Mendelian inheritance patterns - if both parents are homozygous for the same allele at a loci, then the offspring must also be homozygous at that loci.

The method was tested by simulating base calls based of the read length, loci distribution, and error profile observed in a real ONT sequenced sample of Brahman (*Bos indicus*) tail hair. These variables were used to simulate the expected data that would be obtained from sequencing the genome of 1500 Brahman calves with 100,000 mapped reads each. The algorithm assigned 98.7% of calves to the same sire/dam pair as the DNA based pedigree information. This study suggests that ONT could be successfully used to perform parentage testing in cattle.

INTRODUCTION

Genotype based parental assignment in cattle is the most accurate way of pairing up bulls and cows with their progeny (e.g. McClure et al 2018). For cattle that are handled often, there is ample opportunity to sample tissue for DNA analysis and later make management decisions based on the results. For cattle in regions such as Northern Australia, cattle are rarely handled and so the time between taking a sample for DNA analysis and getting the parental assignment results that are needed to make management decisions is a limiting factor in the technologies use.

Portable sequencers are now available that can sequence long stretches of DNA in a single read. These have been applied in the field to address circumstances that require rapid turnaround of results, such as disease outbreaks (e.g O'Donnel *et al*. 2020). Here we suggest a new application: parentage testing of cattle. The implementation of parentage testing has the potential to be the first step towards crush-side-genotyping, where animals are genotyped and have genomic estimated breeding values calculated on farm, allowing for rapid management decisions to be implemented.

MATERIALS AND METHODS

SNP data. 2675 Brahman heifers, cows and bulls from a single property were genotyped with the Neogen TropBeef V2 array, with 50045 SNP (after quality control, with genotypes with QC score <0.6 set to missing, monomorphic SNP excluded and SNP with all heterozygous calls excluded). All animals genotypes were imputed to 611,000 SNP on the Bovine HD array (following further QC) using Eagle (Loh *et al*. 2016) for phasing and Minimac3 (Das *et al*. 2016) for imputation.

All of the dams and bulls (1175 in total) were used as the potential parental population against which each calf was tested.

ONT data. DNA was extracted from the tail hair of a Brahman heifer from Queensland Australia that was collected as part of routine industry genomic evaluation by a commercial supplier. DNA was extracted using the PureGene (Qiagen) DNA extraction kit. The DNA was quantified on a Qubit

4.0. A sequencing library was prepared from the DNA using the SQK-LSK109 kit from ONT. The library was sequenced on a flow cell (FLO-MIN106; ONT) for 72 hours. Guppy v4.2.2 was used to convert the raw data to fastq format. The fastq data was aligned to the ARS_UCD1.2 *Bos taurus* genome using minimap2 version 2.17 (Li 2018). Alignment positions for each read were extracted from the .sam output file. The observed sequence lengths and alignment locations were then used to simulate the expected SNP coverage in the test population of animals.

The error rates of the ONT data were taken from Lamb *et al.* (2021), where ONT data was mapped back to a reference assembly generated from the same animal. The frequency of each substitution error was then calculated for each of the four possible reference nucleotides.

ONT Simulation algorithm. For each read the start position of the alignment to the reference genome and the read length were used to calculate if that read is expected to overlap a SNP location in the 700K SNP data. If the read did not overlay a SNP the algorithm moved onto the next read. If the read did overlay a SNP location, the "true" genotype of the test animal at that location was taken. If the animal was heterozygous at that location one of the two alleles was randomly chosen as the allele that was sequenced. An error was then induced into the base call using the specific ONT error rates for each base, such that the error profile of the final base call reflected the error profile of the real ONT data. The SNP location and basecall was then output to be tested by the parentage calling algorithm. Additionally, an error rate of ten times the observed rate was also tested, by increasing the probability of each error by a factor of 10.

Parentage assignment algorithm. Each calf was individually assigned to a parental pair. To reduce the search space a two-stage parental algorithm was used. First the simulated ONT genotype calls for the test calf were merged with the imputed SNP array genotypes of all potential sires and dams (N=1175). For each test calf loci without simulated ONT coverage were removed from the matrix, approximately 30K SNP remained in the matrix. The genotype matrix was used to calculate a genomic relationship matrix (GRM) using the *A.mat* command (default settings) of the rrBLUP package v4.6.1 in R version 4.0.0. The relationship values between the test calf and each bull were extracted, and the 50 bulls with the highest relationship value to the calf were used in the next step of the parentage assignment. The same approach was taken to highly related cows (n = 50).

The 50 bulls and 50 cows that were selected for further parentage testing were combined into 2500 possible parental pairs (50 x 50). A minimum minor allele frequency (MAF) cutoff was used to avoid large numbers of loci being homozygous in the parental pair and calf by chance. Unless otherwise stated, a MAF of 0.4 was used to filter SNP, and SNP on unplaced scaffolds and the X chromosome were removed. Then, for each pair of potential parents, loci where both the bull and the cow were homozygous were identified. If the test calf simulated ONT data had coverage in this location then the loci was used to create a score. The score was initialized at 0. For every loci that was homozygous in the bull and the cow, and where the calf had matching simulated ONT data, a +1 was added to the 'match' score. Alternatively, for every loci that was homozygous in the bull and the cow, and where the calf had a different simulated ONT genotype, +1 was added to the 'non-match' score. The final parental score for the bull cow pair was returned as M/(M+Nx10), where M is the match score and N is the non-match score. The unmatched loci were weighted more highly than the matched loci because the likelihood of them appearing by chance in the true sire-dam-calf trio is expected to be very low, proportionate to the error rate of the sequence data. This was repeated for all 2500 bull/cow combinations. The highest score was used to identify the most likely parental pair. If another pair of animals was greater than 90% of the highest score, that parental pair was also reported. The most likely parental pair was then compared to the pedigree data of the animals.

RESULTS AND DISCUSSION

A known trio of animals consisting of a cow-bull-calf families were used to test the hypothesis that the highest level of concordance in homozygous sites would be between the true parents and

their offspring. An additional 18 animals from the same herd were used to assess the method. At each minimum MAF tested the highest proportion of concordant calls was the correct parental pair (Table 1). The second highest level of concordance at all MAF contained at least one of the true parents. The difference between the true parental pair and the next highest match increased as the minimum MAF increased (Table 1). Interestingly, when the GRM from the same calf was examined, the highest relationship to any bull was not to the true sire, but to one of his sons (a half sibling of the test calf). The difference between the top (correct) match and the next highest match increased with an increase in the MAF.

Table 1. Details of loci used to assign calves to sire and dam

Min MAF	N loci	Correct parental match[1]	Highest incorrect match[2]	Highest incorrect match (No parents)[3]
0	676430	97.84%	95.08%	93.97%
0.1	420021	95.48%	89.02%	86.65%
0.2	289094	93.82%	84.22%	80.35%
0.3	183804	92.45%	80.08%	75.02%
0.4	89926	91.42%	76.63%	70.72%

[1] Matches between the test calf and the true parents
[2] Matches between the test calf and the highest incorrect sire/dam pair (one parent can be correct)
[3] Matches between the test calf and the highest sire/dam pair where neither sire or dam is correct

The parental assignment algorithm described, which includes identifying the 50 most likely bulls and cows that for the parental pair of the calf being tested, was applied to simulated ONT base calls of 1500 calves. For the simulated reads only 11 (0.73%) calves had two or more potential sire/dam matches. When the error rate in the simulated reads was increased 10 fold, this number increased to 19 (1.26%).

When the parentage assignment results were compared to the pedigree information and historical mating records, the concordance was 98.7% and 98.6% for the low and high simulated error rates (Figure 1). After consultation with the producer, all but 5 of the calves were found to have errors in the pedigree data including misreporting of animal ID numbers. A number of the calves had been assigned to one of two potential bulls in the pedigree – with the DNA identifying the other as the true sire. Overall, after consultation with the producer, the level of agreement between the algorithm presented here and the pedigree information was 99.7% and 99.6%.

The best algorithm for parentage testing needs to play to the strengths of the technology being used. Here we opted for low pass sequencing, which allows many samples to be processed simultaneously. At the depth presented here approximately 100 samples could be processed on a flow cell, with an overall cost of under AU$20 per sample. This results in approximately 1 hour of sequencing per animal, which is a reasonable time to hold animal in yards while the parentage is determined, and clearly much shorter than the several weeks that other DNA parentage assignment tests take. Read alignment of the data can be performed in parallel with the sequencing on the device, and so does not add time to the test.

While using a GRM based approach with this data is possible, we observed that in at least some animals the highest relationship score of the potential bulls tested was not to the true sire, but rather to a half sibling of the test animal. Although the GRM is calculated on a reasonable number of loci (~30,000 loci), the genotype calls were constrained to information from one read, and hence heterozygous loci were randomly genotyped as homozygous. Consequently, the GRM alone was not sensitive enough to accurately differentiate between relationship levels of highly related individuals.

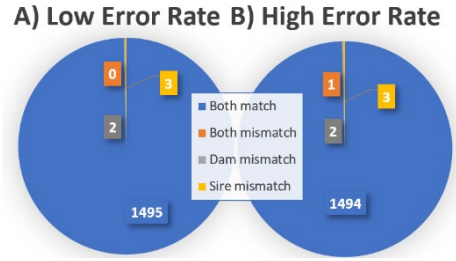

Figure 1. Concordance between parental pairs assigned based on simulated ONT sequencing data and producer curated pedigree information. N_{Calves} =1500; N_{Dams}=1034; N_{Sires}=141

One consideration is that the number of usable loci is directly affected by the genetic distance between the cow and the bull. Animals that are highly related are expected to share alleles by descent, therefore have a higher number of matching homozygous loci. In this population we observed 600-700 matching homozygous loci per bull/cow pair. The required amount of sequence data may differ based on the genetic diversity of the test animals.

The assignment of each calf took approximately 25 seconds of computational time, which includes the calculation of the GRM. The GRM calculation constitutes > 95% of the computational time. Potentially, the GRM based potential parents selection could be removed from the algorithm, and the costs in terms of computational time should be considered based on the size of the test population, which would typically be much smaller than what was tested here. One important consideration is that the number of scores that the homozygosity based test must calculate is equal to the product of the potential sires times potential dams. Hence, with 50 bulls and 50 cows the number of tests is 2500, while with 100 of each the number of tests is 10,000 (a four fold increase even though the population has only doubled). Hence, where the size of the potential sire/dam herd is large, some reduction in the number of animals being tested is likely to save significant computation time. While the computational time for the test is minimal, and the sequencing and bioinformatics analysis can be completed in ~ 1 hour per animal. Laboratory methods (DNA extraction, library preparation) have not been examined here, research into the optimisation of those approaches is being undertaken with promising results (Mason and Botella 2020; Gowers *et al*. 2019).

CONCLUSIONS

Here we present an algorithm for parentage testing from data obtained from ONT sequencing. When tested on simulated ONT data of 1500 calves, the accuracy of parental assignment (compared to curated pedigree information) was 99.7%. This work is the first step towards using ONT data to perform on-farm parentage assignment of cattle.

REFERENCES

Das S., Forer L., Schönherr S., Sidore C., Locke A.E. et al. (2016) *Nat. Genet.* **48:** 1284.
Gowers G., Vince O.; Charles J., Klarenberg I.; Ellis T., Edwards A. (2019) *Genes*. **10**: 902.
Lamb H.J., Hayes B.J., Nguyen L.T. and Ross E.M. (2020) *Genes* **11**:1478.
Lamb H.J., Hayes B.J., Nguyen L.T., Engle B.N. and Ross E.M. (2021) *Proc. Assoc. Advmt. Anim. Breed. Genet:* **24**: *submitted.*
Li, H (2018) *Bioinformatics* **34**, 3094-3100.
Loh P.R., Palamara P.F. and Price AL. (2016) *Nat Genet.* **48**:811.
Mason, M.G., Botella, J.R. (2020) *Nat Protoc* **15**: 3663.
O'Donnell V., Grau F., Mayr G., Samayoa T., Dodd K., *et al*. (2020*) J. Clin. Microbiol.* **58**:e01104