ACCURACY OF GENOMIC PREDICTION IN BRAHMAN CATTLE USING SIMULATED GENOTPYES FROM LOW-COVERAGE NANOPORE SEQUENCING

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

[1]Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, QLD 4067, Australia

SUMMARY

Rapid, on-farm genotyping may be an alternative to SNP chip genotyping for genomic selection in certain agriculture industries. This study aimed to assess the accuracy of genomic breeding values, estimated from simulated Oxford Nanopore derived genotypes. Oxford Nanopore Technologies' (ONT) single nucleotide sequencing and genotyping accuracy was calculated from real sequencing runs of cattle DNA, and used to alter 50K SNP array genotypes in a population of 868 Brahman heifers. Genomic breeding values for age of first corpus luteum (an indicator of age of puberty, were estimated from the simulated ONT genotypes. The accuracies were compared to accuracies calculated using the original SNP array genotypes. Simulated ONT genotypes representing as little as 4 X sequencing coverage were able to generate accuracies not statistically different to SNP chip genotype accuracies.

INTRODUCTION

Genomic selection (GS) first described by Meuwissen *et al.* (2001), is a technique widely used in agriculture, which uses genomic information to predict the genomic estimated breeding value (GEBV) of an individual for key traits. Typically, single nucleotide polymorphism (SNP) arrays are used to cost effectively genotype tens-of-thousands of SNPs, spread evenly across the genome, for genomic selection. Given a sufficiently large reference population of genotype and phenotype data, the GEBV can be accurately predicted from the SNP genotypes.

Turnaround time has limited the use of SNP genotyping and GS in Australia's northern beef industry, where cattle are often only handled once a year. With Queensland, the Northern Territory and Western Australia accounting for 62% of Australia's national beef herd, the difficulty of adopting GS in northern Australia represents a significant loss of potential productivity. We previously proposed a solution to this problem, namely crush-side genotyping (Lamb *et al.* 2020). Crush-side genotyping describes the use of ONT's MinION sequencer to rapidly, genotype cattle on-farm as they pass through the crush. A major limitation to the technology, is its high error rate. Improvements in flow cell chemistry and base calling algorithms has seen the error rate steadily decrease in recent years. However, the current error rate (between 5-8%) is still significantly higher than that of SNP array genotyping. The objective of this study was to ascertain the effect of ONT sequencing errors on the accuracy of genomic estimated breeding values in Brahman cattle.

MATERIALS AND METHODS

Ethics. All analysis was performed using phenotypes and DNA samples previously collected with approval by the J.M. Rendel Laboratory Animal Experimental Ethics Committee (CSIRO, Queensland) as approvals TBC107 (1999 to 2009) and RH225-06 (2006 to 2010).

Nanopore Sequencing Error Rates. To determine ONT sequence error rates, ONT sequence data (approximately 8 X coverage) from a Brahman cow sequenced on MinION R9 flow cells was aligned against the Brahman genome (assembled from data from the same animal ; Ross 2019) using minimap2 (Li 2018) with the default settings for ONT alignment. Samtools mpileup (version 1.2, Li *et al.* 2009) was used to create a genome wide mpileup of the reads aligned to the reference genome. A maximum read depth of 50 was used to avoid chimeric repeats or ambiguously aligned regions of the genome. The number of single nucleotide mismatches for each locus across the genome was calculated from the mpileup using R. The error rates were reported as percentages of mismatches for each nucleic acid, given the total number of observations of nucleotides at all reference sites of a particular nucleic acid. For example, adenosine-guanine errors are the number of Guanine mismatches divided by all observations at reference adenosine sites.

Nanopore Genotyping Error Rates. A subset of reads, representing 4 X, 6 X, 8 X, 10 X and 18 X coverage from a second Brahman cow sequenced on the MinION, were then aligned using Minimap2 to the *Bos taurus* reference genome. Reference assembly UMD 3.1.1 was used to ensure reference loci and strand direction matched between sequencing and SNP chip genotypes. Samtools and BCFtools were used with a probability threshold (P value) of 1 for SNP discovery and a phred scaled base accuracy threshold (Q score) of 7, to genotype loci on the BovineSNP50 BeadChip (Illumina, San Diego, CA). Three methods (variable allele count, set ratio and minimum allele count) for assigning genotypes from the sequence were examined. The variable allele count method grouped loci by total coverage, and used a separate minimum allele count for each group to verify a genotype (Figure 1). This method was hypothesised to better distinguish between sequencing noise and heterozygous genotypes at higher coverages. The set ratio method called a particular observation as a likely true genotype if the allele was observed in greater than 10% of total observations at that loci. Finally, the minimum allele count method called a true genotype if a particular allele was observed more than twice no matter the total coverage. Any loci with more than two different alleles observed were considered incorrect genotype calls. All genotypes were then compared to the SNP chip genotypes to calculate genotyping accuracy, as well as the percentage of missing calls (loci with less than 2X coverage).

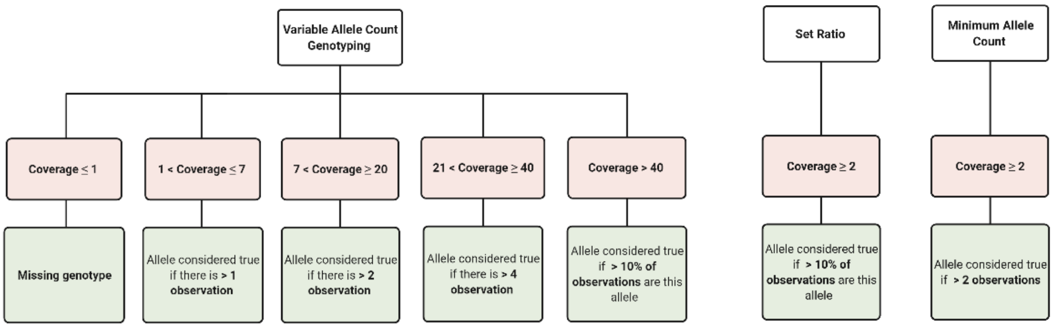
Variable Allele Count Genotyping
Coverage ≤ 1
Missing genotype
1 < Coverage ≤ 7
Allele considered true if there is > 1 observation
7 < Coverage ≥ 20
Allele considered true if there is > 2 observation
21 < Coverage ≥ 40
Allele considered true if there is > 4 observation
Coverage > 40
Allele considered true if > 10% of observations are this allele
Set Ratio
Coverage ≥ 2
Allele considered true if > 10% of observations are this allele
Minimum Allele Count
Coverage ≥ 2
Allele considered true if > 2 observations

Figure 1. Genotyping method. Three different SNP genotyping methods were used to call variable loci

Simulating Nanopore Genotypes and Genomic Breeding Value Prediction. The cattle used in this experiment represent a subset of the Northern Breeding Project population, established by the Cooperative Research Centre for Beef Genetic Technologies. Phenotypes and management history for this herd have been extensively documented (Johnston *et al.* 2009; Engle *et al.* 2019). Records from a subset of 868 Brahman heifers was taken, including management history and age of first corpus luteum (AGECL), as determined using ultrasound scanning. The 868 heifers were also genotyped using the BovineSNP50 BeadChip (Illumina, San Diego, CA; Hawken *et al.* 2012).

Herd of origin, management cohort and birth month were concatenated into a single factor: contemporary group, which was modelled as a fixed effect (Engle *et al.* 2019). As only a Brahman subset was used in this study *Bos indicus* content was excluded as a covariate.

Genomic best linear unbiased prediction (GBLUP) was used to calculate GEBVs for AGECL using the univariate model:

$$y = XB + Zu + e$$

Where y is the vector of phenotypes, X is a design matrix allocating phenotypes to fixed effects, B is a vector of the fixed effect contemporary group, Z is a matrix of SNP genotypes and u is a vector of additive SNP effects.

The genotyping error rate for each coverage was used to randomly select a number of SNP genotypes in Z to alter. The calculated Nanopore sequencing error rate was then used to simulate errors at these loci consistent with the Nanopore error profile. The percentage of missing genotypes was also used to introduce missing SNPs.

To calculate the GEBV accuracy for AGECL 5-fold cross validation was used, with each validation population representing 20% of the total population (n = 868). Validation animals were included in the G matrix but coded with missing phenotypes. The package MTG2 (Lee and van der Werf 2016) was used for the predictions and the accuracy was calculated using $acc = r(GEBV, AGECL_{res})/\sqrt{h^2}$ where $h^2 = 0.55$. The 95% confidence interval was used to compare accuracies across the different simulations.

Two scenarios were simulated when calculating the accuracy of the GEBVs. The first simulation represented a scenario where, all animals, both reference and validation populations, were genotyped using ONT. This was simulated by simulating ONT errors in all animals. The second simulation represented, the more realistic situation where the reference population was SNP chip genotyped, while the validation population was genotyped using ONT. This was simulated by inducing errors into only animals in the validation population.

RESULTS AND DISCUSSION

Cytosine and thymine were found to have the lowest sequencing accuracies with 0.84% and 0.83% of bases at cytosine and thymine loci being inaccurately sequenced. The sequencing error rate revealed that for each nucleotide there was a single nucleic acid which was significantly more likely to be incorrectly called than the other nucleic acids (Table 1). For example, errors at adenosine loci were three time more likely to be called as guanine than either cytosine or thymine.

Table 1. Nanopore sequencing error rates. The distribution of substitution errors observed in Nanopore sequencing data mapped to the reference genome built from the same animal

		Reference Nucleotide[1]			
		A	**C**	**T**	**G**
Alternate Nucleotide[2]	**A**	NA	18.03%	68.74%	16.34%
	C	17.85%	NA	13.13%	65.83%
	G	65.70%	13.27%	NA	17.83%
	T	16.46%	68.70%	18.13%	NA

[1] The observed nucleotide in the reference genome

[2] The nucleotide observed in the mapped Nanopore reads

The minimum allele count method performed best at high coverages while the set ratio method had better genotype calling accuracies at lower coverages. Despite this the variable allele count method still outperformed the other two methods across all coverages (Table 2). At 18 X coverage the maximum genotyping accuracy achieved was 93.89%, in order to further increase the genotyping accuracy methods to disseminate between systematic sequencing errors, such as methylation, may still be required. Strand bias for example, could be used to filter out methylation signals to increase the accuracy of genotyping.

Table 2. Nanopore genotyping accuracies and percentage of missing genotypes for various coverages

	Coverage				
	4	**6**	**8**	**10**	**18**
Percentage of loci not called[1]	41.2%	9.5%	4.4%	4.1%	0.6%
Accuracy (Variable allele count)[2]	84.5%	87.4%	89.7%	91.4%	93.9%
Accuracy (Minimum allele count)[3]	66.9%	74.8%	81.5%	86.8%	93.7%
Accuracy (Set ratio)[4]	84.4%	87.1%	89.2%	90.3%	93.0%

[1] Loci which did not meet the minimum depth criteria (>2 reads) for genotyping

[2] Variable SNP calling criteria were used based on the sequencing depth at each loci (See Figure 1)

[3] Alleles were called as present if observed more than 2 times

[4] Alleles were called as present if they comprised more than 10% of observed alleles at that locus

The genotyping errors observed (Table 2) also supported the ratios of nucleotide sequencing errors (Table 1), for example, at homozygous adenosine loci (AA) for 10 X coverage, 95.5% of loci were called correctly as AA or TT (the reverse compliment), while 3.3% of loci were called incorrectly as AG or GA. The other 12 genotype combinations shared the remaining 1.2% of AA loci evenly. This supports the earlier findings that A-G errors are more than three times more common in Nanopore sequencing than A-C or A-T. This pattern was observed in the results across all genotype combinations and could be leveraged to further increase the accuracy of Nanopore genotyping by incorporating a more stringent threshold for calling a genotype which corresponds to the most error prone nucleic acid given the reference loci. Using the AA example above, this would mean increasing the threshold for a guanine genotype call at an adenosine reference locus to decrease incorrect AG/GA genotype calls.

The GEBV accuracy of AGECL from the SNP chip genotypes was 0.39 ± 0.03 which is not statistically different to the accuracy reported by Engle *et al.* (2019), although removing tropical composites from the herd (effectively decreasing the reference population by 1,000 animals) likely describes the difference in average accuracy. At coverage as low as 4 X, there was no difference between the SNP chip accuracy and the simulated Nanopore genotype accuracies (Figure 2). Another study using Nanopore sequence data to predict genomic breeding values in cattle for three other traits: body condition score, hip height and body weight also reported accurate genomic predictions were possible from 4 X sequencing coverage without imputation (Lamb *et al.* 2021). This demonstrates accurate genomic prediction from Nanopore data is possible for a range of desirable traits.

A difference between the 95% confidence interval in the two different genotyping scenarios (reference and validation versus validation only) can be seen at 4 X coverage. However, this difference appears to decrease at higher coverages, likely due to the overall increase in genotyping accuracy.

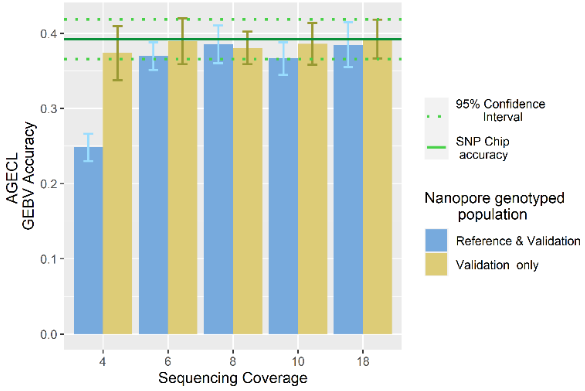

Figure 2: GEBV accuracies for AGECL calculated from 33k genotyped loci. Genotypes were either directly observed in the SNP array data or had the error profile observed in SNP calling from ONT data simulated in the dataset. ONT errors were either simulated in both the reference and validation population or only in the validation population to represent two different sequencing scenarios

CONCLUSIONS

Here, we have demonstrated genotyping accuracies as high as 85% are achievable with just over 4 X Nanopore sequencing coverage. Using a SNP chip genotyped reference population, simulated Nanopore genotypes generated GEBV accuracies that were not significantly different ($P > 0.05$) from accuracies achieved using entirely SNP chip genotypes. This suggests ONT genotyping at low coverages can provide comparable GEBV accuracies to traditional SNP chip genotyping.

ACKNOWLEDGMENTS

The authors would like to acknowledge the numerous contributions from the staff and students of the Hayes' lab group, in particular Mrs. Christie Warburton. We would also like to acknowledge the efforts of all individuals who have contributed to the collection and maintenance of the Beef CRC phenotype and genotype data. Financial contributions from Meat and Livestock Australia (project codes B.STU.2001 and L.GEN.1808) are also gratefully acknowledged.

REFERENCES

Engle, B.N., Corbet, N.J., Allen, J.M., Laing, A.R., Fordyce, G., McGowan, M.R., Burns, B.M., Lyons, R.E., Hayes, B. (2019) *J. Anim. Sci.* **97**, 90-100.

Hawken, R.J., Zhang, Y.D., Fortes, M.R., Collis, E., Barris, W.C., Corbet, N.J., Williams, P.J., Fordyce, G., Holroyd, R.G., Walkley, J.R., Barendse, W., Johnston, D.J., Prayaga, K.C., Tier, B., Reverter, A., Lehnert, S.A. (2012) *J. Anim. Sci.* **90**, 1398-410.

Johnston, D.J., Barwick, S.A., Corbet, N.J., Fordyce, G., Holroyd, R.G., Williams, P.J., Burrow, H.M. (2009) *Anim. Prod. Sci.* **49**, 399-412.

Lamb, H.J., Hayes, B.J., Nguyen, L.T., Ross, E.M. (2020) *Genes (Basel)* **11**,

Lamb, H.J., Hayes, B.J., Randhawa, I.A.S, Nguyen, L.T., Ross, E.M. (2021) *BioRxiv*

Lee, S.H., van der Werf, J.H. (2016) *Bioinformatics* **32**, 1420-2.

Li, H. (2018) *Bioinformatics* **34**, 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data Processing, S (2009) *Bioinformatics* **25**, 2078-9.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001) *Genetics* **157**, 1819-1829.

Ross, E. (2019) 'Characterisation of the Brahman Genome' Meat and Livestock Australia Limited.