Proc. Assoc. Advmt. Breed. Genet. 21: 249-252

SNP-BASED PARENTAGE IN AN AUSTRALIAN CATTLE INDUSTRIES CONTEXT: DOES ONE SIZE FIT ALL?

R.E. Lyons[1], S. Buttsworth[1], D. Waine[1] and M. Kelly[2]

[1]The University of Queensland, School of Veterinary Sciences, Gatton, QLD 4343, Australia.
[2]The University of Queensland, Centre for Animal Science, Brisbane, QLD 4062, Australia.

SUMMARY

Globally, there is a trend away from microsatellites or short tandem repeats (STRs) to single nucleotide polymorphisms (SNPs) on the basis of perceived advantages for genetic identification, traceability and assessment of parentage. However, the transition is not simple and here we highlight unique problems faced in adapting low cost SNP-based assays for Australian systems.

INTRODUCTION

Much has been written about the promises of SNP-based parentage verification in livestock and animal traceability across the supply chain (Heaton *et al*. 2002, Van Eenennaam *et al*. 2007, Baruch and Weller 2008). Advantages discussed includes abundance, amenability to high-throughput genotyping platforms and reproducibility across laboratories. The biallelic nature of the marker along with automation of bioinformatic analysis makes this process less prone to reader error. Unlike microsatellites, they are synergistic with existing genomics applications and hence more cost-effective for those breeds undertaking genomics-based breeding programs now or in the near future. Based upon modelling data and validation in taurine breeds the International Society of Animals Genetics (ISAG) recommended a set of 100 core SNP, and later added an additional set of 100 markers to increase the exclusion power in indicine and synthetic breeds.

Following development and optimisation of Sequenom SNP panels, we demonstrate that the ISAG-recommended core bovine SNP parentage panel is not sufficient to provide accurate parentage verification in many common Australia production systems. The objectives of this study are: (1) to demonstrate factors influencing effectiveness of the tests, (2) develop additional analyses to clearly identify, communicate and eliminate problems pre- and post-analysis, and (3) maximise accuracy and completeness of parentage verifications especially in large test cohorts.

MATERIALS AND METHODS

Samples and DNA Extraction. Commercial populations of Brahman or Brahman-cross animals were used as case studies for sire verification only. DNA was extracted and purified from hair follicles using customised protocols.

Genotyping. Genotyping was performed using iPLEX reagents and platinum protocols for high multiplex PCR, single base primer extension (SBE) and generation of mass spectra, as per the manufacturer's instructions (Sequenom, San Diego). SEQ1 iPLEX panels contained a total of 138 SNP including 95 ISAG core plus 4 ISAG additional SNP. The additional panel in SEQ2 consisted of 59 SNPs for a combined total of 197 markers genotyped and total of 97 ISAG core SNP. These new markers were developed to be informative in Brahman and Tropical Composite breeds. Mass spectra were analysed using TYPER software (Sequenom, San Diego) in order to generate genotype calls and allele frequencies. Some sires were genotyped using the custom GeneSeek Genomic Profiler low-density BeadChip (GGP-LD) with ~ 25,000 SNPs assayed per sample.

Post-genotyping data analysis. To identify issues of mislabelling or sampling errors, duplicate sample checks were performed by counting the number of discordant marker calls between two samples. Less than or equal to 5 discordant markers between genotypes across assays were considered likely to be from the same animal, indicative of sampling or testing errors and

requiring recollection to ensure accurate genotypes. Prior to requesting recollection of samples, putative duplicates were routinely checked via microsatellite analysis to ensure these samples did not represent closely related individuals (e.g. full sib). To date all cases have been confirmed as the same individual (n=19). Hence a discordance threshold of ≤ 5 SNP to represent potential duplicate samples appears a suitable value.

Populations and Primary Analysis. Batches represented small, medium and large multi-sire matings, hence denoting increasing degrees of complexity in sire assignment. The small batch contained 20 Brahman progeny and 5 sire candidates. The medium-sized batch included 173 Brahman crosses with 26 sire candidates. The large batch had 706 Brahmans originating from 3 properties with a total of 42 sire candidates (Table 1). Each batch was initially parent verified using the SEQ1 SNP data. Parentage analysis was via exclusion based on opposing homozygotes with strict criteria (exclusions ≤ 3). In the large herd, the SEQ2 SNP test was used to assess for increased accuracy of parentage assignment (exclusions ≤ 3). Any sire-progeny matches with a misconcordance rate > 3 SNP were not accepted.

Detection of potential sibs. Often in large extensive beef herds it is not possible for the breeder to supply all potential sires. Thus we tested the ability of SEQ2 to assign unqualified animals to sib groups for a set of 204 Brahman progeny with known sire information. This represented progeny of 29 sires with an average of 7 progeny per sire (min=1 and max=17). These animals representing a subset of the large multi-sire population discussed previously. The accuracy of detecting the sib families was assessed against the known sire to group potential sibs from a method using genomic relationship matrix (GRM) developed with SEQ1 and SEQ2 panels. The GRM matrix was formed using all animals in the genotyped parentage analysis as per VanRaden (2008). The subset of animals requiring allocation to sib families was then selected from within this full matrix. Potential sib groups were formed by successively adding animals to the sib group if their mean relationship with the current group of the new animal was greater than an empirically defined threshold. In the current analysis this threshold was varied from 0.12 through to 0.2 to examine the trade off in accuracy and number of animals assigned. In practice the threshold could be determined by analysis of animals with known parentage within the same parentage population.

RESULTS AND DISCUSSION

Broadly speaking, the process for the provision of large multi sire herd parentage analyses is as follows: owner provides hair samples to laboratory with a list of offspring and potential parents, lab staff prepare and genotype samples, data is analysed and results are returned to owner. The measure of success of a parentage verification is the proportion of all calves correctly assigned with no resubmission of samples required. However this is often not the case with a number animals remaining unresolved following the initial analysis. Unresolved cases may be due to incorrect sample submission, unrepresented sires/dams in the analysis, or the inclusion of genotyping errors with less than acceptable call rates. Data on true batches of increasing complexity are shown in Table 1 which illustrates a number of important considerations in deciding upon the test panel chosen.

The small commercial batch had 100% of progeny assigned to a sire. All sires are accounted for and present in the testing pool. For the medium sized commercial herd 82% of available progeny were matched to a sire despite 2 sires and 12 progeny failing to reach an acceptable SNP count (n ≥ 120 or ~ 85% of the total markers). These were unavailable for retesting and not included in the primary analysis. Upon consultation with the owner, one additional sire candidate was identified and genotyped. When this sire was included in a reanalysis, the overall total increased to 89% of available progeny matched to a sire which is comparable to that achieved through microsatellites.

Table 1. Case studies from commercial batches across small, medium and large populations

POPULATION	BREED	TEST	PROGENY (n)	SIRES (n)	Assigned (%)	Unassigned (%)	Recollect (%)
SMALL	Brahman	SEQ1	20	5	100	0	0
MEDIUM: analysis 1	Brahman X	SEQ1	173	24	82	11	7
MEDIUM: analysis 2	Brahman X	SEQ1	173	25	89	4	7
LARGE:Analysis 1	Brahman	SEQ1	706	42	57	35	8
LARGE:Analysis 2	Brahman	SEQ2	706	42	97	3	0

As shown in Table 1, the rates of successful assignment are affected by a number of factors including size and completeness of the animal data set provided at initial testing. To better identify the factors that may be leading to failures in assigning parents affecting results, pre- and post-analysis data assessment tools have been generated to identify and resolve issues in a timely fashion. For example, the duplicate genotype check before parentage verification has been invaluable in identifying problems such as transcriptional errors during sampling or laboratory error, and as such saves time and increases accuracy/confidence as recollects for suspect samples can be organised promptly. Similarly, and especially for large batches of sample, graphic representations such as that shown in Figure 1 can be helpful in demonstrating issues and corrective steps required to resolve the analyses.

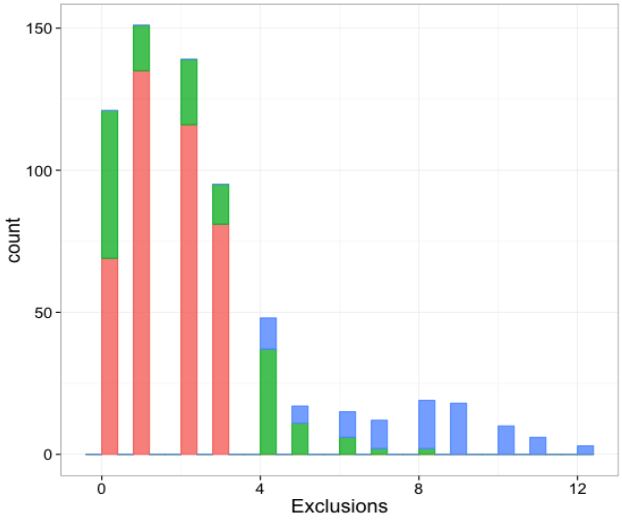

Figure 1. Representation of data quality and reasons for unassigned progeny in initial testing. Pink represents qualified animals, Green represents those requiring retesting, and Blue represents missing sires.

Noting the increased complexity of the large commercial batch, it would be recommended to future clients that all sires are genotyped using the GGP-LD test with progeny and dams on the lower density SEQ1/SEQ2 assay. Generally the number of sires is small in comparison to the total number of animals requiring genotyping and the marginal cost of genotyping the sires on a higher density panel is low and provides three immediate benefits: firstly the SNP array tends to be more accurate, secondly they have higher call rates and thirdly any particularly difficult to resolve cases (progeny) can be upgraded and testing across a much larger set of markers. While there is significant benefit in upgrading sires in particular to GGP-LD, the economic impost of genotyping all animals (sires and progeny) on this platform remains an impediment, and genotyping progeny on the smaller assays offsets this cost. Previous analyses have shown the importance of increasing marker counts in these large herds, as shown in the primary analysis in Table 1 where only a subset of markers representative of the SEQ1 test panel are used in parentage verification. While 97% of progeny matched to a sire using SEQ2 data for progeny (analysis 2), had the SEQ1 test option been chosen only 57% of progeny were resolved.

It is important to note that the 97% assignment for analysis 2 was achieved only after extensive consultation with owners. In the first iteration of the SEQ2 analysis, 79% of progeny were assigned to a sire which is still superior to that of the SEQ1 test.

With large multi-sire groups where herds may have been across multiple properties or extensively grazed, the nature of the enterprise makes complete mustering and collection of all sires logistically and economically difficult and it is not uncommon to have missing sires. While missing sires are evident through numbers of unassigned progeny, it is not immediately clear if one or more sires are missing. To this end, the use of GRM has been investigated to cluster progeny and hopefully give an estimate of missing sire numbers. *In silico* testing using CRC data for progeny (n = 204) was used to identify an appropriate stringency to allow clustering whist retaining relative accuracies. A minimum mean sib relationship of 0.13 was deemed the best balance between accuracy and percentage assigned. In practice this threshold would be determined by examining the mean relationship in animals within each data set where parentage is known. It is important to note that this is not intended to be a verification tool, but rather used as a guide only to estimate the likely number of sires not submitted and the likely groups of sibs. The benefits of returning estimated sire-groups for unresolved calf cases are considerable. Using this data and specifically animal groupings, producers may be able to identify the missing sires by paddock, birth/joining date or even physical characteristics of the calf groups.

Table 2. **Accuracy and assignment rates of GRM analyses for clustering unassigned progeny in parentage verification analyses**

	Stringency				
	0.12	**0.13**	**0.14**	**0.15**	**0.2**
Accuracy	85%	93%	93%	94%	100%
% assigned	60%	55%	50%	41%	19%

While there are very good arguments for the shift from microsatellites to SNP-based parentage and in the long term SNP will become the dominant mode of identification in parentage and traceability testing, the difficulties associated with this transition have often been understated and largely unreported. The Australian cattle industry with its large diversity of breeds and extensive production systems represents a sector where unique challenges to successful transition exist. Previous research has demonstrated that ISAG's core (100 SNP) panel can be inadequate for parentage testing for some breeds (Strucken et al. 2014) and our recent experiences suggest this is also the case for large *Bos indicus* cohorts in Australian production systems. Clear communication between laboratory and producers is essential including identification of challenges or constraints to achieving high levels of assignment. For example, tight deadlines for verification on consanguineous groups can be best met if sires and/or dams are collected and genotyped in advance of the progeny. Correct sampling techniques (clean and adequate amounts) is important. Improved results obtained over three seasons for a large Brahman herd from 71% assignment in year 1 and 2, to 89% in Year 3 were largely attributable to improved sampling technique on-farm. It is our view that implementation of these additional analyses at strategic points across the pipeline will further enhance rates of assignment.

REFERENCES

Baruch, E. and J. I. Weller (2008) *Animal Genetics* **39**: 474.
Heaton, M. P., G. P. Harhay, G. L. Bennett, *et al*. (2002) *Mammalian Genome* **13**: 272.
Strucken, E. M., B. Gudex, M. H. Ferdosi, *et al*. (2014) *Animal Genetics* **45**: 572.
Van Eenennaam, A. L., R. L. Weaber, D. J. Drake, *et al*. (2007) *J. Anim. Sci.* **85**: 3159.
VanRaden, P. M. (2008) *J. Dairy Sci.* **91**: 4414.