Proc. Assoc. Advmt. Anim. Breed. Genet. 18:239-242

SNP ORIGIN BIAS ON POPULATION STRUCTURE ANALYSIS: AN AUSTRALIAN BEEF CATTLE CASE STUDY

L. R. Porto Neto[1,2] and W. Barendse[1]

Cooperative Research Centre for Beef Genetic Technologies
[1] CSIRO Livestock Industries, Queensland Bioscience Precinct, 306 Carmody Road, St. Lucia 4067, Australia.
[2] The University of Queensland, School of Animal Studies, St. Lucia 4072, Australia.

SUMMARY

The use of single nucleotide polymorphism (SNP) in cattle molecular genetics studies has increased in the last few years by several factors including the identification of new markers and the development of new genotyping technologies. Most of the cattle SNP markers were developed by comparison of a Hereford genome sequence to a sequence of an animal of a different breed, leading to different breed of origin of the SNP markers. In this Australian case study we analysed 302 SNP markers of two different origins (Brahman and Holstein) in a population study including eight cattle breeds. We demonstrate that the breed of origin of the marker can potentially bias this analysis, showing that it is important to find a balance between the origin of the markers and the composition of the population being studied.

INTRODUCTION

Single nucleotide polymorphisms (SNP) are the most common molecular markers in the genome of an organism. Their use has increased through the development of high throughput genotyping platforms that have a low cost per genotype. In cattle the vast majority of SNP markers were identified by comparison of a Hereford genome sequence with sequences of another taurine or Brahman animal (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/). The breed of the DNA sequences, which was compared to the Hereford genome sequence to identify the marker, is called the breed of origin of the SNP marker in this study. A recent study comparing the minor allele frequencies (MAF) of different breeds showed that common SNP (MAF > 0.2) identified originally in taurine breeds often have MAF < 0.10 when evaluated in zebu animals (The Bovine HapMap Consortium, 2009).

Considering that differences in allelic frequency between populations forms the basis of population structure analysis and the knowledge that the SNP origin can bias the allelic frequency, we studied the influence of SNP origin on population structure analysis in a sample of Australian beef cattle.

MATERIAL AND METHODS

The animals used in this study and the genotyping has been reported previously (Barendse *et al.* 2009). There were 179 animals of eight cattle breeds used in this analysis. They were the taurine dairy Holstein (HOL, n=25), taurine meet Hereford (HFD, n=24), Murray Grey (MGY, n=16), Shorthorn (SHN, n=24) and Angus (ANG, n=25); the composite Belmont Red (BEL, n=21) and Santa Gertrudis (SGT, n=25); and the zebu Brahman (BRM, n=19). These animals were genotyped and quality control measures implemented as reported previously. Briefly, 9260 SNP, distributed in all chromosomes, were genotyped using the MegAllele™ Genotyping Bovine 10k SNP Panel (Hardenbol *et al.* 2005) by ParAllele Inc. on an Affymetrics GeneChip. Animals with more than 10% of missing data, and then loci with more than 10% of missing data were excluded from the analysis.

SNP markers. SNP originating in Brahman and Holstein were compared. Holstein were the most common taurine breed that was compared to the Hereford and Brahman was the only zebu breed used in SNP discovery. Among the approximately 300 Brahman markers genotyped only 151 were polymorphic in our population. One hundred and fifty one markers from Holstein were then selected by numeric order, distributed in most chromosomes (Table 1). Only SNP polymorphic in at least one breed were used.

Table 1. Distribution of Brahman (BRM) and Holstein (HOL) markers per cattle chromosome.

Chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	X
BRM	9	4	2	2	9	8	3	10	3	7	11	1	4	6	5	9	8	0	4	1	0	7	2	14	0	0	0	1	0	3
HOL	4	5	8	11	4	4	6	5	2	7	7	5	6	5	6	3	1	6	6	2	4	4	4	1	2	4	2	5	1	0

*two Holstein markers and one Brahman marker were not assign to a chromosome.

Population analysis. The minor allelic frequency (MAF) of the markers were observed and grouped into Brahman, composite and taurine subpopulations for comparison. The population stratification was evaluated using the STRUCTURE Software 2.2 (Pritchard *et al.* 2000) and visualized using *Distruct* (Rosenberg 2004). Three runs were performed for each of K=2 to 8 with burn in of 20,000 and 100,000 MC iterations without previous knowledge of breed assignment. The data shown is the analysis of one representative run. A major cluster was considered a cluster that contains more than 50% of the individuals of a breed. We observed the ability of the markers to determine subpopulations (clusters) that were compatible with the breed designations. To determine if a cluster was a good representation of a breed, the number of individuals of pure taurine or Brahman with more than 0.85 genetic composition assigned to the main cluster of their respective breed was counted and the results between the sets of markers were compared.

RESULTS AND DISCUSSION

Allelic frequency analysis. Seventy-one Brahman derived markers had lower (<0.10) MAF in the taurine group than in the Brahmans (23) and composites (21) (Figure 1). A similar trend in the opposite direction was also observed in the Holstein markers, Brahmans having a higher percent of MAF < 0.10 (61) but the difference to the other groups was not as pronounced as for the Brahman markers (Taurines 38 and Composites 44).

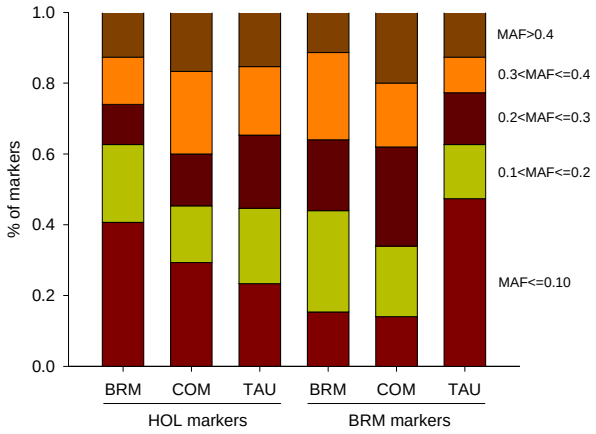

Figure 1. Deciles of minor allele frequency by genetic group and SNP origin

Population stratification analysis. To determine the number of underlying groups in the data, we determine K, the number of subpopulations. The biggest initial increase in likelihood was to K=2 and then a progressively smaller increases until an optimal K=7 for the combined set of markers (data not shown). There was not much difference in likelihood between K of a small number in the range 2-8. We found that with K=2, instead of a taurine vs zebu split, the two groups were meat taurine vs dairy taurine plus zebu, with the Belmont Red and Santa Gertrudis showing near equal contributions for the two groups regardless of SNP origin. For larger K (Fig 2) the groups split progressively into more breeds until with K=8, the Santa Gertrudis and Belmont Red did not only appear as composite taurine and zebu, but as separated groups. Throughout the process, Angus and Murray Grey were indistinguishable, as would be expected given their known history as well as from the principal components analysis for these animals (Barendse *et al.* 2009).

To determine what effect SNP origin had on the STRUCTURE plots we analysed each marker set separately. Most of the discrimination in the combined panel came from the Brahman markers, the increase in likelihood was 5 times greater with the Brahman SNP than the Holstein SNP (data not shown). At K=3, the Brahman panel clearly separated the Holstein from the other beef breeds and at K=8 was able to put most animals correctly into their breeds despite or perhaps because of having more loci with lower MAF (Table 2). The Holstein panel was less discriminatory than the Brahman panel and its analysis led to more than one set of assignments with approximately equal likelihood. The combined set of loci performed as well as the Brahman set, as would be expected given the greater information content in the Brahman SNP. It is worth noting that the Brahman SNP performed worse on Brahman and the Holstein SNP performed worse in classifying Holstein as 100% of that breed group.

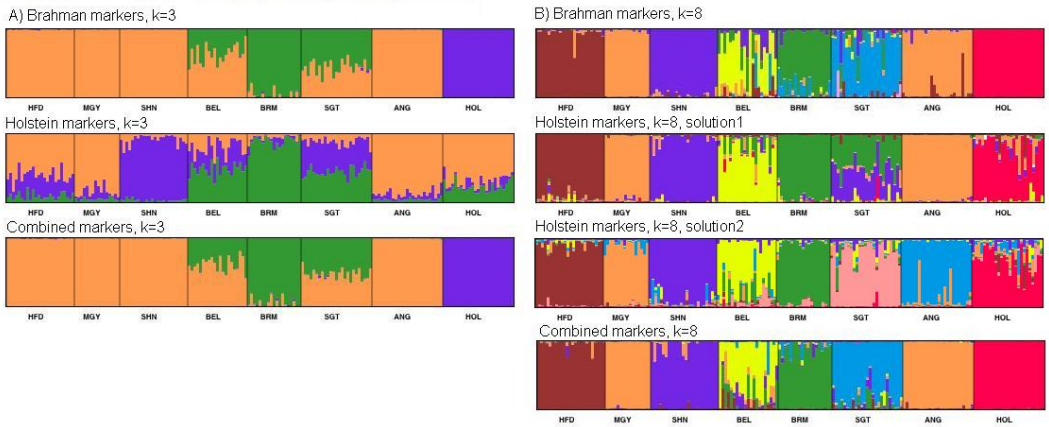

Figure 2. Population structure determined by Brahman, Holstein and Combined set of SNP markers from K=3 (A) and K=8 (B)

The poor performance of Brahman SNP in classifying Brahman and Holstein SNP in classifying Holstein is clearly of interest and suggests why the Holstein SNP have performed worse than the Brahman SNP in the STRUCTURE analysis (Table 2). The Brahman SNP have the highest MAF in the Brahman, as would be expected, and the Holstein SNP have the highest MAF in the Holstein. The Holstein SNP have higher MAF in the taurine breeds, the taurine breeds tend to have similar allele frequencies and so are more difficult to separate from each other using the Holstein panel. On the contrary, the Brahman SNP have greater differences between breeds, they have a more U-shaped distribution, and so they have greater power to discriminate between

breeds. Combining the two set of markers did not materially influence the discriminatory power of the panel of SNP and it is doubtful whether a many more SNP would be able to bring greater clarity in discrimination.

Table 2. Number of individuals of a particular breed that have at least 85% of its genetic composition on its respective breed main cluster at K=8

	N of animals	Brahman markers	Holstein markers[1]		Combined set
		Number of animals in the main cluster (%)	Number of animals in the main cluster (%)		Number of animals in the main cluster (%)
HFD[2]	24	23 (0.96)	19 (0.79)	21 (0.87)	22 (0.92)
MGY	16	16 (1.00)*	9 (0.56)	16 (1.00)*	16 (1.00)*
SHN	24	23 (0.96)	16 (0.67)	21 (0.87)	19 (0.79)
ANG	25	21 (0.84)*	20 (0.80)	25 (1.00)*	25 (1.00)*
HOL	25	25 (1.00)	6 (0.24)	11 (0.44)	25 (1.00)
BRM	19	8 (0.42)	14 (0.74)	18 (0.95)	14 (0.74)

[1] The Holstein markers lead to two different clustering, which are analysed separately.
[2] HFD Hereford, MGY Murray Grey, SHN Shorthorn, ANG Angus, HOL Holstein, BRM Brahman.
* The MGY and the ANG shared the same main cluster.

CONCLUSIONS

A relatively small number of SNP can be used to reconstruct the genetic divisions of breeds within cattle. However, an ascertainment bias in the origin of the SNP would generate spurious conclusions in the degree of similarity between breeds that is not consistent with other molecular work. As the ancestries of cattle are well known, these biases can be seen for what they are. In less well known species, a bias in ascertainment could generate groups that do not reflect the true links between groups.

ACKNOWLEDGMENTS

We thank J.W. Kijas for discussion on using STRUCTURE and *Distruct* analyses. LRPN is supported by an Endeavour International Postgraduate Research Scholarship, a University of Queensland International Student Living Allowance and a Beef CRC scholarship; WB is supported by CSIRO and a Beef CRC Research Grant.

REFERENCES

Barendse W., Harrison B.E., Bunch R.J., Thomas M.B., Turner L.B. (2009) *Bmc Genom* **10:**178.
The Bovine HapMap Consortium (2009) *Science* **324:**528.
Hardenbol P., Yu F.L., Belmont J., MacKenzie J., Bruckner C., Brundage T., Boudreau A., Chow S., Eberle J., Erbilgin A., Falkowski M., Fitzgerald R., Ghose S., Iartchouk O., Jain M., Karlin-Neumann G., Lu X.H., Miao X., Moore B., Moorhead M., Namsaraev E., Pasternak S., Prakash E., Tran K., Wang Z.Y., Jones H.B., Davis R.W., Willis T.D., Gibbs R.A. (2005) *Gen. Res.* **15**:69.
Pritchard J.K., Stephens M. & Donnelly P. (2000) *Genetics* **155:**945.
Rosenberg N.A. (2004) *Mol. Ecol. Notes* **4**:137.