QC ANALYSES OF SNP ARRAY DATA: EXPERIENCES FROM A LARGE POPULATION OF DAIRY SIRES WITH 23.8 MILLION DATA POINTS

K.R. Zenger[1,2], M.S. Khatkar[1,2], B.Tier[1,4], M.Hobbs[1,2], J.A.L. Cavanagh[1,2], J. Solkner [1,2], R.J. Hawken[1,3], W. Barris[1,3], H.W. Raadsma[1,2]

[1]Co-operative Research Centre for Innovative Dairy Products-CRC IDP, [2]ReproGen – Centre for Advanced Technologies in Animal Genetics and Reproduction, Faculty of Veterinary Science, The University of Sydney, Camden, Australia.[3] CSIRO Livestock Industries, Brisbane, Australia, [4] AGBU, University of New England, Armidale, Australia

SUMMARY

The use of a high throughput SNP genotyping platform with 15,380 bovine SNP assays, across 1546 dairy bulls resulted in a data set of approximately 23.8 M SNP data points. Stringent control measures based around low polymorphic content, sample failure, deviation from HWE, low call rate, non-Mendelian inheritance, tri-allelic SNP, and incompatible clustering of data, resulted in removal of 4321 SNPs. The majority (2973) were due to low polymorphic content (MAF<0.01), with the remaining features consistent with assay quality. Despite the need for removal of some SNP, repeatability of SNP call rate was extremely high (>99%) across repeat samples, and between platforms. SNP technology has now matured where comprehensive genome-wide analyses can be conducted in cattle with a high degree of robustness.

INTRODUCTION

Molecular data promise greater understanding of the biological processes those generate the phenotypic variation in livestock populations. Recent advances in technology now permit the DNA of whole genomes to be amplified and assessed for large numbers of single nucleotide polymorphic (SNP) markers simultaneously and at reasonable cost. Such quantities of genotypes have stimulated new approaches to the analysis of phenotypic variation and the prediction of genetic merit. With such large numbers of SNPs used in genetic prediction, the possibility of error rates arising from non-concordant SNPs increases. Before implementing any genetic evaluation approach the SNP data needs to be thoroughly verified through a series of stringent tests. This paper describes many of the necessary procedures needed to identify aberrant SNP data in large experimental populations.

MATERIALS AND METHODS

Samples and SNPs: Samples and genotyping was described by Raadsma et al. (2007, these proceedings). DNA was extracted from semen provided by Genetics Australia, from 1546 bulls born between 1951 and 2001. The bulls were selected to maximize genetic variation and depth of pedigree. Genotyping was conducted on the ParAllele (now Affymetrix) platform incorporating 15,380 SNPs based on 10,643 SNPs from the bovine genome project and an additional 4,737 custom SNPs from the IBISS database (Hawken *et al.* 2007). Data integrity was measured using anonymous duplicate samples and SNPs within runs and between runs. In total there were 23 DNA replicates, four of which were measured 4 times, and 212 duplicate SNPs throughout the entire procedure. In addition, results from a previous genotyping platform (Illumina - Golden Gate assay) were compared on common DNA (372) and SNPs (272) and agreement between these two platforms was 99.2%. Platform repeatability of duplicate samples was > 99% for either platform.

Assay quality control: In addition to a measure of accuracy provided by ParAllele, these 23.8

million genotypes were examined for veracity in a number of ways. Assays and or animals were excluded from further analysis if any of the following conditions occurred;

i. No genotypic variation observed (MAF < 0.01),
ii. SNPs with less than 75% bull coverage,
iii. Significant deviation (P < 0.000001) from Hardy-Weinberg equilibrium (HWE) at a population level (not including X chromosome SNPs),
iv. Significant deviation (frequency > 0.05) from Mendelian inheritance within large (>10) half-sib families,
v. Duplicate concordance of samples < 80%,
vi. Weighted accuracy of replicate SNPs < 0.8 (values < 1 indicate poor results),
vii. Affymetrix SNP call rate < 95% in > 200 individuals,
viii. Tri-allelic or aberrant loci (determined from clustering raw normalized SNP data).

Individual and SNP evaluations using PCA: Principal Component Analysis was also used to interrogate the data to identify outlier SNPs or individuals in the population. The program GENETIX Version 4.05 (Belkhir *et al.* 1996–2004; http://www.univ-montp2.fr/~genetix/genetix/intro.htm) was used, incorporating four main principal components of variation.

RESULTS AND DISCUSSION

Assay quality controls: Table 1 shows the number of SNPs detected from the final dataset (15,036 unique SNPs) in each independent test as a result of the data integrity measures.

Table 1. Number of SNPs detected for each test.

Criteria	Number of SNPs detected
SNPs with MAF < 0.01	2973
Incomplete dataset coverage (< 75%)	339
Deviation from HWE (P < 0.000001)	810
Pedigree non-Mendelian inheritance	382
Concordance < 80	20
Replicate < 0.8	530
SNP call rate < 95% in > 200 bulls	258
Tri-allelic SNPs	199
Incompatible clustering of raw data	392

A total of 4321 unique SNPs (28.7%) were removed from the dataset following data integrity measures. Many non-concordant SNPs identified may have been a result of assay limitations (eg., SNP location, DNA quality and quantity, probe design), and hence may work in other genotyping platforms (see below).

Aberrant genotype clusters: Normalised 'raw' allele signals were plotted for each assay that deviated from HWE. Many of our assays showed signs of non-canonical clustering of data points (e.g. Fig. 1). Previous published studies demonstrate non-canonical clusters attributable to genotyping miscalls, genomic deletions, and hemizygous or double null genotypes which give rise to new clusters (eg., Carlson et al. 2006). Our data indicate that most of our aberrant SNPs are assay artefacts, resulting from close proximity of flanking SNPs to one another. The ParAllele assays use molecular inversion probe (MIP) technology and we speculate that the binding of the probe to its target in one assay can be affected by the allele present at the neighbouring SNP.

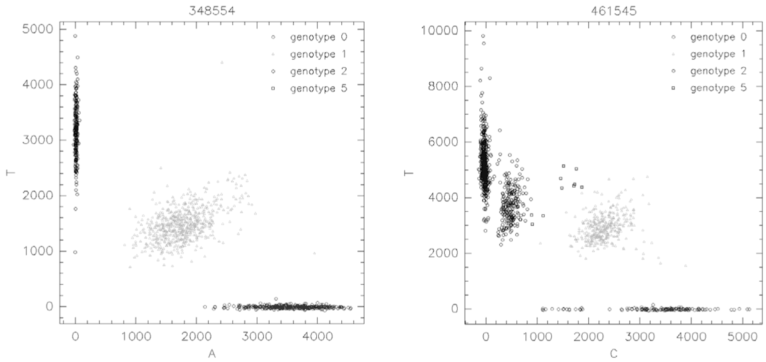

Figure 1. Scatter plots of allele signals from two assays, 348554 (showing three canonic clusters of points) and 461545 (showing additional unexpected clusters).

Tri-allelic loci: Based on normalised 'raw' signal clustering, we identified 199 assays which showed evidence of an additional 'unexpected' allele. For example, bi-allelic A/C assay 348563 has A, C and T alleles (Fig. 2), which produces 6 clouds of points corresponding to the 6 possible tri-allelic genotypes (AA, AC, AT, CC, CT, TT). Since the genotype scoring algorithm only recognizes A/C bi-allelic alleles, these additional clusters have been incorrectly scored to fit into the three cluster coordinates (Fig. 2).

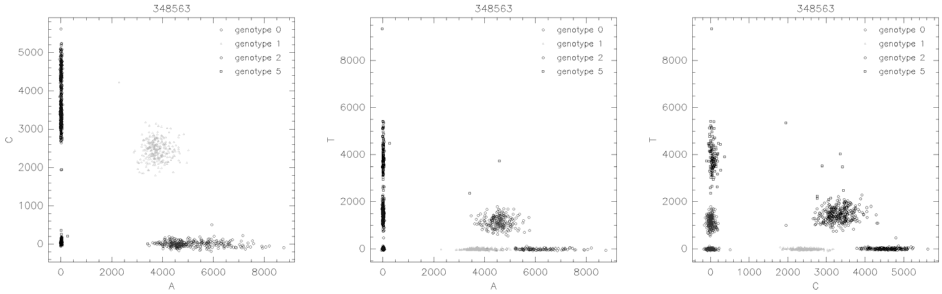

Figure 2. Scatter plots of allele signals from assay 348563. This assay is designed to measure A/C polymorphism and so calls of 0,1,2 and 5 assay correspond to genotypes AA, AC, CC and 'no call'.

Population and SNP evaluation using PCA: The four principal components contributed a total of 9.07% of the variation among SNP between individuals in the dataset (3.34%, 2.90%, 1.45% and 1.37% respectively). The first two components accounted for most of this variation (~70%). When examining the clustering of individuals and SNPs, two distinct groupings were generated based on the first component (3.34%; Fig. 3). Further examination revealed that these outlier individuals had spurious genotypes for a sub-set of loci. DNA data indicate that these individuals had lower quality DNA, which had the effect of producing false genotypes for sensitive loci in the assay. Removal of the susceptible loci and or samples with low quality DNA resulted in a single population and SNP cluster. The second major component (2.90%) indicated that variation was primarily associated with year of birth and country of origin. This result is in agreement with that of Zenger *et al.* (2007) who showed mild genetic sub-division between old Australian born bulls (pre-1985) and those breed after the influx of foreign semen primarily from North America.

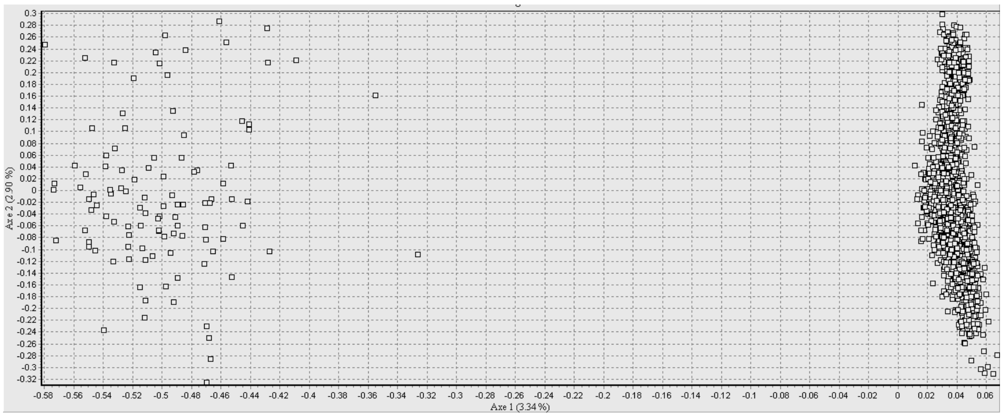

Figure 3. Clustering of individuals using the first two PCA components.

CONCLUSIONS

High stringency control measures for data integrity, resulted in removal of approximately 30% of the genotyping data prior to use in genetic prediction methods. It is essential that the data is ‘clean’ since aberrant genotypes can affect the calculation and genetic prediction of animals. We show here the step-wise procedures used to ensure high data integrity in large datasets.

ACKNOWLEDGMENTS

We thank Genetics Australia for semen samples and ADHIS for pedigree data.

REFERENCES

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. and Bonhomme, F. (1996-2004). GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France).

Carlson, C.S., Smith, J.D., Stanaway, I.B., Rieder, M.J. and Nickerson, D.A. (2006). *Hum Mol Genet* **15:** 1931.

Hawken, R. J., Barris, W. C., McWilliam, S. M. and Dalrymple, B. P. (2004). An interactive bovine in silico SNP database (IBISS). *Mamm Genome* **15:** 819.

Raadsma, H.W., Zenger, K.R., Khatkar, M.S., Crump, R., Moser, G., Solkner, J. , Cavanagh, J.A.L., Hawken, R.J., Hobbs, M., Barris, W., Nicholas, F.W. and Tier, B. (2007) *Proc. Assoc. Advmt. Anim. Breed. Genet.* ***17:*** 231.

Zenger, K.R., Khatkar, M.S., Cavanagh, J.A., Hawken, R.J. and Raadsma, H.W. (2007). *Anim Genet* **38:** 7.