QUALITY CONTROL FOR OVINE SNP50 BEADCHIP GENOTYPES

K. G. Dodds, B. Auvray, N. Pickering and J. C. McEwan

AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand

SUMMARY

We use a dataset of Ovine SNP50 BeadChip genotypes to investigate quality control issues for genomic data. A number of criteria are investigated: % loci scored per animal, % animals scored per locus, deviations from Hardy Weinberg, comparison with animal information (gender, parentage, breed), reproducibility for replicate samples, and unusual allelic ratios. These checks can be used to clean the dataset for its endpoint analysis.

INTRODUCTION

An ovine single nucleotide polymorphism (SNP) chip able to assay over 50,000 SNPs has recently been developed by Illumina in collaboration with the International Sheep Genomics Consortium (ISGC; www.sheephapmap.org). Quality control is an essential step in the analyses of such data. This is especially important for early studies using a genotyping platform, as often the SNPs have not been independently verified following ascertainment from sequencing data. We discuss some quality control procedures and give examples of their application to Ovine SNP50 BeadChip data.

MATERIALS AND METHODS

Methods. A number of criteria are investigated. Individual criteria or combinations of them may lead to the rejection of subsets of the data.

Genotyping success rates. Simple summary statistics of success (a genotype reported) rates by animal and locus are calculated.

Consistency between animal replicates. Replicated assays of the same animal were compared.

Consistency with recorded gender. SNPs on the X chromosome should show a pattern consistent with recorded gender. Samples from females have many heterozygous calls across the X chromosome. Samples from males should not be heterozygous for loci in the non-pseudoautosomal region (the lower portion of the X chromosome), although allowance is made for genotyping errors and mis-positioned loci.

Consistency with recorded pedigree. Animals with one or both parents also genotyped were checked to see if their SNP results were consistent with those of their putative parents. The checks account for (or discard) SNPs which are inherited in an X-linked manner.

Consistency with recorded breed. Principal components were calculated from the genomic relationship matrix which in turn was calculated using the first method of VanRaden (2008). The principal components were plotted against each other with breed denoted.

Validation of SNP position by linkage mapping. Most of the SNPs have been positioned on v1.0 of the ovine sequence assembly (www.livestockgenomics.csiro.au/sheep/oar1.0.php). These assembly positions can be checked by linkage mapping the SNPs in an appropriate resource, in this case the international mapping flock (IMF; see below). A series of mapping steps was used to allow an initial validation of the SNP positions – these methods are likely to find only gross errors in position, e.g. assigned to the wrong chromosome. The first step was an approximate (for speed) linkage analysis against loci on the Maddox *et al.* (2001) map for the assigned chromosome. This analysis used only the last generation and assumed phases that gave the strongest linkage. Loci with a lod>2.5 were assumed to be correctly assigned to chromosome. For those that remained, the

same procedure was used against the Maddox *et al.* (2001) framework map loci for the other chromosomes, with a lod>4 being used as evidence for linkage. Remaining loci were then analysed in the full IMF pedigrees with Cri-map (Lander and Green 1987). Two-point lod scores were calculated for loci with more than 10 informative meioses (the others having insufficient information for detecting linkage) against loci on the assigned chromsome. Those with a lod>3 were assumed to be correctly assigned.

Deviations from Hardy-Weinberg equilibrium. A chi-squared test statistic for Hardy-Weinberg equilibrium was calculated for each SNP within each breed. An animal was assigned to a breed if it was recorded as being more than 75% of that breed. Quantile-quantile (QQ) plots were used to aid determining which loci showed extreme values, as these plots allow one to visually account for effects of population substructure and multiple testing.

Allelic ratios and relative intensity. Illumina report normalised intensities (denoted X and Y) for the two alleles assayed for a SNP. Plots of allele frequency (Y/(X+Y)) against genome position were created for each sample genotyped. The smoothed $\log_2$ relative intensity was also plotted, where the intensity was calculated as $r = \sqrt{X^2 + Y^2}$, and then calculated as the ratio to the mean value for all animals in the analysis for that locus.

Animals. The locus mapping procedure used the International Mapping Flock (IMF) pedigrees (Maddox *et al.* 2001). Other procedures are illustrated using all or parts of a multi-breed set of animals that are part of an Ovita-funded programme investigating the relationship between locus genotypes and traits of economic importance. This resource, comprising 2785 animals, was sourced from a number of research and breeder flocks in New Zealand. They were predominantly derived from Romney, Coopworth, Perendale and Texel breeds.

Genotypes. Genotyping was undertaken by Illumina for both sets of animals using their Ovine SNP50 Beadchip. The IMF animals were included as part of the HapMap project of the ISGC. The Ovita project involved 2865 samples, with 20 animals being run in duplicate and 60 Illumina controls. Forty animals failed the genotyping, including both samples from one that was duplicated. The chip assayed 59,454 potential loci, with genotype results being reported for 53,903 (90.7%) of these. A further 338 loci had intensity (X and Y) values reported, but not genotypes.

RESULTS AND DISCUSSION

Genotyping success rates. There were 48,944 (90.8%) loci scored for all 2839 successful samples. The distributions of the intensities of 3629 loci indicated that there may have been a nearby polymorphism, creating difficulties for scoring; 1606 of these had less than 95% samples scored. There were only 2 other loci scored this poorly. The success rates for each sample and for each locus, classified as above, are shown in Figure 1. Illumina also provide a quality score for each result, and these can be used to highlight potential genotype, locus or DNA sample problems.

Consistency between animal replicates. Of the 19 duplicated animals genotyped, there were no differences in the scored genotypes, and on average only 26 loci were scored in one sample and not in its duplicate. Gross inconsistencies may have indicated mislabelling or incorrect transfers between DNA stocks, plates and chips. Minor inconsistencies would reflect the repeatability of the genotyping process. Comparing inconsistencies by SNP may indicate problematic SNPs.

Consistency with recorded gender. Four animals (2 males and 2 females) had X chromosome genotypes inconsistent with their recorded gender, later found to be due to mislabelled samples.

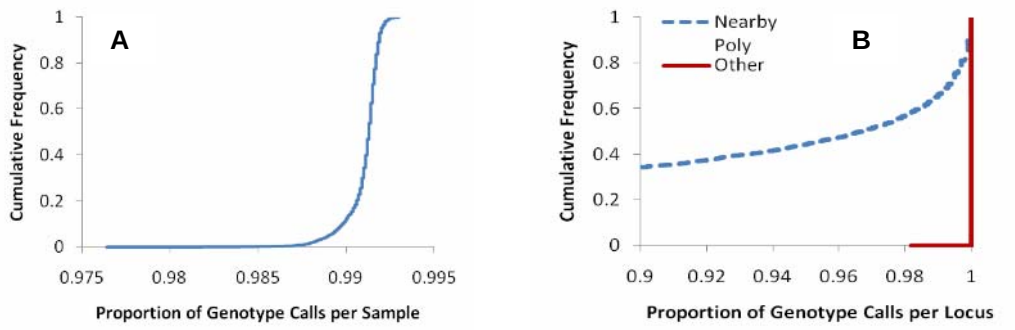

Figure 1. Genotype success rate: (A) per sample; (B) per locus.

Consistency with recorded pedigree. Of the 1302 animals with one parent also genotyped, 194 had fully concordant genotypes, 1025 had less than 30 discordant genotypes, while 83 had more than 1000 discordant genotypes. There were 8 animals with both parents genotyped; 5 of these had less than 20 discordant genotypes, while the other three had 450-500 discordant genotypes.

Consistency with recorded breed. The first two principal components are shown in Figure 2. The animals designated as being at least 90% of a particular breed tend to cluster in the same region of the figure.

Validation of SNP position by linkage mapping. Results from applying the mapping strategy to chromosome 26 are shown in Table 1. One of those that mapped elsewhere was linked to X chromosome markers. This locus was also noted to show an X-locus clustering pattern in allele intensities.

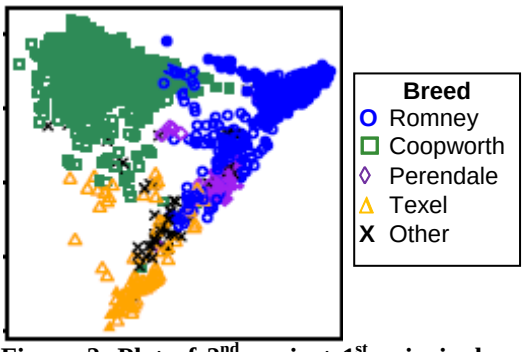

Figure 2. Plot of 2nd against 1st principal component. Closed symbols are used if the main breed component > 90%.

Table 1. Mapping Ovine SNP50 Beadchip loci from chromosome 26

	Number or %
Loci scored	917
Mapped to chromosome	90.3%
Low information	8.2%
Unmapped	1.3%
Mapped elsewhere	0.2%

Deviations from Hardy-Weinberg equilibrium. A QQ plot for the most common breed (Romney) is shown in Figure 3.SNPs with nearby polymorphism or appearing X-linked are denoted as class 2. Loci (not X-inherited) with high chi-squared values are candidates for further investigation.

Allelic ratios and relative intensity. Figure 4 shows a typical plot of allelic ratios and relative intensity for one chromosome of one animal. The frequency of one of the alleles is denoted by +, while the log relative intensity is shown by a solid line. These plots allow detection of chromosomal features of interest (Gibbs and Singleton 2006). Regions with low heterozygosity and normal intensity reflect identical by descent (inbred) regions (e.g., central region of Figure 4). Regions with low heterozygosity and low intensity reflect chromosomal deletions. Regions with high intensity and allele frequencies around 1/3 and 2/3, as well as 0 or 1, reflect chromosomal duplications. There were no obvious chromosomal abnormalities in these data.

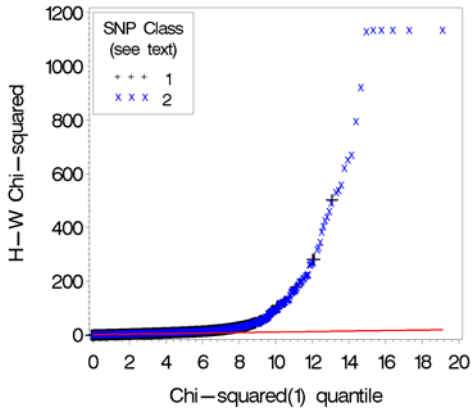

Figure 3. QQ Plot of Hardy-Weinberg chi-squared test statistics for Romney for those loci assigned to autosomes.

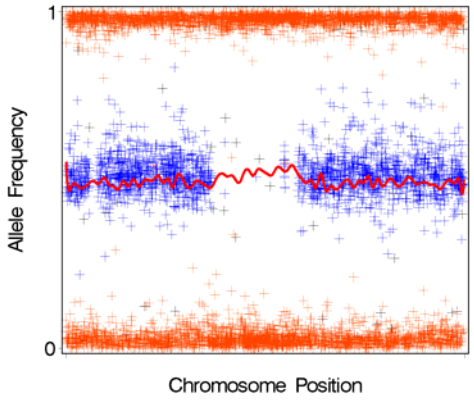

Figure 4. Plot of allelic ratios (+) and relative intensity (line) for one chromosome of one animal.

CONCLUSIONS

A variety of quality control procedures enable the screening of SNP-chip data before final analysis. These procedures often highlight data that require further checking. The process is often iterative between screening unsuitable markers and unsuitable animals.

ACKNOWLEDGMENTS

This research was supported by Ovita Ltd. We also thank Marylinn Munson and Kimberly Geizen (Illumina) for genotyping; ISGC for development of the ovine chip and for allowing us to use their genotyping results from the IMF; the many New Zealand breeders who provided DNA samples and Shannon Clarke, Dianne Hyndman and Nadia McLean for sample preparation.

REFERENCES

Gibbs, J.R. and Singleton, A. (2006) *PLoS Genetics* **2**:e150.

Lander, E. S., and Green, P. 1987. *Proc Natl Acad Sci U S A* **84:** 2363.

Maddox, J.F., Davies, K.P., Crawford, A.M., Hulme, D.J., Vaiman, D., Cribiu, E.P., Freking, B.A., Beh, K.J., Cockett, N.E., Kang, N., Riffkin, C.D., Drinkwater, R., Moore, S.S., Dodds, K.G., Lumsden, J.M., van Stijn, T.C., Phua, S.H., Adelson, D.L., Burkin, H.R., Broom, J.E., Buitkamp, J., Cambridge, L., Cushwa, W.T., Gerard, E., Galloway, S.M., Harrison, B., Hawken, R.J., Hiendleder, S., Henry, H.M., Medrano, J.F., Paterson, K.A., Schibler, L., Stone, R.T. and van Hest, B. (2001) *Genome Res.* **11**:1275.

VanRaden, P.M. (2008) *J. Dairy Sci.* **91**:4414.