A COMPARISON OF RELATEDNESS ESTIMATES FROM SNP CHIP GENOTYPES AND FROM GENOTYPING-BY-SEQUENCING RESULTS

K.G. Dodds, J.C. McEwan, T.C. Van Stijn, R. Brauning and S.M. Clarke

AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand

SUMMARY

Estimates of genomic relatedness derived from either SNP chip (two different densities) or genotyping-by-sequencing (GBS) resources were compared in a group of 95 sheep. The estimates were highly correlated (*r* = 0.983-0.992 for relatedness between individuals) although GBS estimates were slightly higher than chip-based estimates. These results provide evidence that GBS is a useful technique for genomic studies.

INTRODUCTION

Genomic information is increasingly being used in animal breeding. Many livestock industries have SNP chips available at a range of densities and at a cost where they are being used in breeding programmes. The SNP chip results are used either directly or indirectly, often after imputation to a higher density, to estimate genomic relatedness between animals in breeding programmes. An alternative technology is to use genotyping-by-sequencing (GBS), based on sequencing a fraction of the genome, possibly at low depth (to reduce costs). GBS can be applied in species without extensive genomic resources (such as SNP chips and reference genome assemblies). Methods have been developed to estimate relatedness using GBS results (Dodds *et al.* 2015). Here we compare relatedness estimates in a sub-flock of 95 sheep genotyped using both genotyping technologies.

MATERIALS AND METHODS

Animals. A group of sheep that had previously been genotyped using SNP chips were chosen for GBS genotyping to allow a comparison of methods. This group were a set of 89 male and female progeny from a single cohort (born in 2014), 5 of their sires and a control sample; 80 of the progeny had their sire in these data. Two of the sires were Primera, two were predominantly Texel, and the other was predominantly Texel x Coopworth. The control animal was a Texel x Coopworth. The dams were unrecorded, but were a maternal type (predominantly Romney).

SNP chip genotypes. The set of animals had been previously genotyped. All animals except for 12 of the progeny had been genotyped with the Illumina ovine HD beadchip (Kijas *et al.* 2014). Although this chip assays over 600,000 SNPs, only 41,020 of those SNPs (referred to as 41k) are used here, being those that are also on the Illumina ovine SNP50 beadchip and which passed quality control (including being autosomal) on both chips using the criteria in Auvray *et al.* (2014). All progeny had been genotyped with a custom Illumina BovineLDplusovine SNP chip which assays 5283 ovine SNPs; this study used 4015 (referred to as 4k) of those SNPs, being those that were also on both the HD beadchip and the SNP50 beadchip, and which passed quality control. For some animals, genotypes for these SNPs from a higher density chip were used as the 4k genotypes.

GBS genotypes. The animals were genotyped by GBS using the methods described by Dodds *et al.* (2015) and based on the GBS protocol of Elshire *et al.* (2011). Briefly, DNA samples and a negative control were digested with *Pst*I; a different barcode adaptor was added to each sample, along with a common adapter. Samples were then combined and fragments in the range 150-500bp were selected and single-end sequenced on one lane of an Illumina HiSeq2500 resulting in approximately 2 million reads per sample. The resulting sequence reads were adapter-trimmed and then UNEAK (Lu *et al.* 2013) was used to detect variants (without the use of a reference genome) and report allele counts for each variant and sample.

Estimation of relatedness. Relatedness between each pair of individuals, and self-relatedness for each individual were estimated by the methods of Dodds *et al.* (2015) which accounts for the read depth in a genotype call. This included estimation from SNP chip genotypes, where the depth was taken to be infinite. This is then equivalent to the first method of vanRaden (2008), except that only SNPs with data for the individual or pair of individuals involved are used for that estimate (i.e., missing genotypes are not imputed). The allele frequencies used were taken as the sample allele frequencies using allele counts. For chip data the allele counts were the usual counts (e.g. AA has 2 A alleles). All SNPs reported by UNEAK were used for the GBS-based analysis. Methods are compared by correlation and by regressions of the differences on the means (Altman and Bland 1983) for each pair of methods. Standard errors for the regressions using pairs of individuals were calculated using the number of individuals rather than the number of pairs as an approximate method to account for the non-independence of the pairs.

RESULTS AND DISCUSSION

The GBS process resulted in calls for 68,293 SNPs with a mean read depth of 6.1. The 41k SNPs had 407 with a minor allele frequencies (MAF) of 0 in these data, and these were removed before further analysis. Summary statistics are shown in Table 1; for GBS, having at least one read at a SNP is taken as a call. Call rates were high for the chip data, but lower for GBS due to the randomness of the sequence reads. The MAFs were highest for the 4k chip, where SNPs were highly selected to be informative, and lowest for GBS where SNPs were not pre-selected.

Table 1. Summary statistics for the different genotyping methods

Marker set	Number of SNPs used	Mean call rate	Mean minor allele frequency	Mean inbreeding estimate	Mean relatedness
41k chip	40,613	99.96%	0.289	-0.037	-0.012
4k chip	4,014	99.37%	0.367	-0.035	-0.010
GBS	68,293	86.73%	0.225	0.058	-0.003

Table 2. Summary statistics for relatedness comparisons including correlations of the estimates and regressions of the differences (first marker type minus second marker type) on the means

Marker comparison	Relatedness	Number compared	Correlation (r)	Mean difference (SE)	Slope (SE)
41k – 4k	Self	83	0.844	-0.002 (0.002)	0.093 (0.065)
41k – GBS	Self	83	0.769	-0.095 (0.003)***	0.060 (0.080)
4k – GBS	Self	95	0.662	-0.094 (0.003)***	-0.068 (0.093)
41k – 4k	Between	3403	0.992	-0.001 (0.002)	-0.012 (0.014)
41k – GBS	Between	3403	0.989	-0.008 (0.002)***	-0.055 (0.016)**
4k – GBS	Between	4465	0.983	-0.007 (0.002)**	-0.047 (0.019)*

* P<0.05, ** P<0.01, *** P<0.001

Comparisons of relatedness estimates are shown in Figure 1 and Tables 1 and 2. In general, the estimates appear to be quite similar across methods. GBS produced higher (P<0.001) inbreeding estimates and they were less consistent with the chip estimates than the two chip results were with each other. The breeding design for the progeny set tends to involve breed crosses, so we would expect inbreeding to be low (with low variation) in the flock. The differences in inbreeding between GBS and chips appeared to be uniform over the observed range; the regression slopes for the differences were not significant. One possible reason for GBS giving higher inbreeding estimates is that the SNPs have not been pre-selected, and in particular are likely to include non-autosomal SNPs. This could elevate the results for males, as they would appear homozygous for X-linked and Y chromosome markers. The inbreeding in the male progeny was higher than in the females, but by only a small amount (0.005, SE = 0.006, NS). These regions would be expected to have around half the average read depth (in males) and the method of estimating inbreeding adjusts for un-observed heterozygosity with low depth (assuming autosomal markers), which would dampen any increase in estimated inbreeding due to these regions.

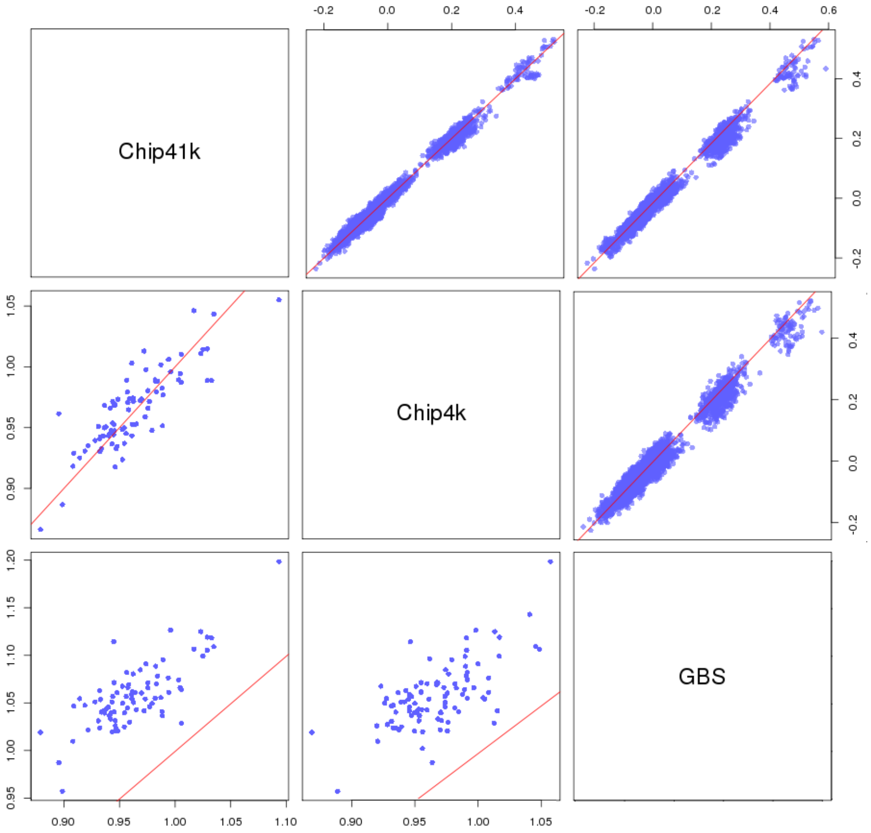

Figure 1. Comparison of relatedness estimates using different genotyping methods. Plots below the diagonal are for self-relatedness of individuals and those above the diagonal are for relatedness between all pairs of individuals. Diagonal labels show the method for the horizontal axis in that column and vertical axis in that row. Lines of equality are also drawn

The relatedness values were all highly correlated (Figure 1, Table 2). Once again, GBS produced higher (P<0.01) values overall, but only by a small amount (0.007 or 0.008 on average). There was also a significant (P<0.05) slope for these two comparisons, meaning that there was a larger difference between GBS-based estimates and chip-based estimates for higher values of relatedness. The relatedness estimates form three main groups. The group with higher values are mainly sire-progeny pairs, but there are also pairs from within the progeny group, presumably full-sibs. The middle group contains a pair of sires, while all other pairs are within the progeny group, presumably half-sibs.

The relatedness estimates average close to zero, a by-product of estimating allele frequencies within the dataset, rather than having ancestral frequencies (Yang *et al.* 2010). As GBS SNPs were not pre-selected, and the methods gave similar estimates, it suggests that there is not a large ascertainment bias on the chips, in terms of estimating relatedness. It is also interesting to note that the estimates appear to be similarly correlated for low values compared to high values of relatedness. This suggests that the rankings of relatedness when the estimates are negative are still meaningful (pairs with more negative values are less related than pairs with negative values close to zero).

One of the main reasons for estimating relatedness in agricultural species is to allow genomic selection, for example these estimates can be used directly in a GBLUP model. Having the relatedness estimates for the three methods correlate well suggests that they would perform similarly for genomic prediction, but further work is needed to verify this. For example, it is generally accepted that at least 10,000 SNPs are needed for genomic prediction, suggesting that the high correlation (0.992) between the 4k and 41k sets seen here may not be enough to guarantee satisfactory predictions from the 4k set. If GBS is to be adopted in resources were many individuals have been genotyped with SNP chips, there will need to be an investigation on how to combine relatedness estimates from different methods as has been required for combining pedigree and genomic-based relatedness (Aguilar *et al.* 2010).

We have shown that there is good agreement between relatedness estimates from GBS and from SNP chips, especially in terms of their correlation. There were some small differences in the mean levels of relatedness, so that adjustments would be required if combining data using different methods. It would be useful if this comparison could be extended to genomic relatedness estimation across divergent breeds and also to examine different GBS protocols, i.e. different enzymes, to check the robustness of these results. In summary, GBS is a promising method for genomic analyses using relatedness estimates and can be rapidly deployed, even for species with poor genomic resources.

ACKNOWLEDGEMENTS

This work was funded by MBIE (NZ) through the “Genomics for Production & Security in a Biological Economy” programme. The Illumina ovine HD BeadChip was developed by the International Sheep Genomics Consortium with FarmIQ (http://www.farmiq.co.nz/), a joint NZ Primary Growth Partnership programme. The animals involved in this work were part of the FarmIQ project and thanks is extended to Focus Genetics for access to this resource.

REFERENCES

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. and Lawlor T.J. (2010) *J. Dairy Sci.* **93**:743.

Altman D.G. and Bland J.M. (1983) *J. Royal Stat. Soc. Ser. D (The Statistician)* **32**:307.

Auvray B., McEwan J.C., Newman S.A.N., Lee M. and Dodds K.G. (2014) *J. Anim. Sci.* **92**:4375.

Dodds K.G., McEwan J.C., Brauning R., Anderson R.A., Van Stijn T.C., Kristjánsson T. and Clarke S.M. (2015) *BMC Genomics* **16**:1047.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S. and Mitchell S.E. (2011) *PLoS ONE* **6**:e19379.

Kijas J.W., Porto-Neto L., Dominik S., Reverter A., Bunch R., McCulloch R., Hayes B.J., Brauning R., McEwan J. and the International Sheep Genomics Consortium (2014) *Anim. Genet.* **45**:754.

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S. and Costich D.E. (2013) *PLoS Genet.* **9**:e1003215.

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

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E. and Visscher P.M. (2010) *Nat. Genet.* **42**:565.