A COMPARISON OF GENOTYPES GENERATED BY INFINIUM BEADCHIPS AND A TARGETED GENOTYPE BY SEQUENCING METHOD IN CATTLE

A.J. Chamberlain[1], P. Beatson[4], B.A. Mason[1], C. Prowse-Wilkins[1], C.M. Reich[1], C. Schrooten[5], B.J. Hayes[1,2], J. Tibbits[1,3], and M. Hayden[1]

[1] Department of Economic Development, Jobs, Transport and Resources, Bundoora, Victoria, Australia; [2] La Trobe University, Bundoora, Victoria, Australia; [3] Melbourne University, Victoria, Australia; [4]CRV, Hamilton, New Zealand; [5]CRV, 6800 AL, Arnhem, the Netherlands

SUMMARY

The cost of SNP genotyping is of major importance in genomic selection programs. One possibility for reducing the expense of applying genomic selection would be to take advantage of recent dramatic decreases in the cost of genome sequencing, by using genotyping-by-sequencing (GBS) techniques to provide cheaper genotypes. This paper presents a GBS method that can target individual variants and therefore any SNP of interest. Comparing array and GBS genotypes from 471 individuals for 5119 SNP we show that with GBS we can achieve sample call rates of 93%, as compared with 95% for arrays, and that genotypes called from the GBS are 98% concordant with those from SNP arrays. With further refinement of the custom reference we will be able to achieve higher call rates and genotyping accuracy.

INTRODUCTION

Genomic selection has now been widely implemented in dairy cattle industries worldwide, predominantly for the selection of sires. Genomic selection uses SNP genotypes to estimate SNP effects in a reference population, such reference populations have been established within most dairy countries. These SNP effects are used along with SNP genotypes of individual animals to make genomic predictions of breeding value in selection populations. Hence the cost of SNP genotyping is of major importance in genomic selection programs. Array based genotyping methods have been widely used to genotype cattle, particularly the Illumina Infinium 7K, 50K and HD beadchips. These are robust and accurate genotyping platforms, however the costs required to genotype an animal can be too high, preventing many dairy farmers from using such tests for heifer selection for example (Pryce, Hayes 2012).

Next generation sequencing has dramatically decreased in cost over the past decade and so many have turned to GBS techniques to provide cheaper genotypes for genomic selection programs. Many GBS techniques rely on the cleavage of DNA with restriction enzymes to generate a pool of DNA fragments, which are sequenced to enable SNP discovery and genotype calling (Elshire *et al.* 2011). These methods provide SNP that are randomly located across the genome. However much work has now been done to identify the most informative variants to use for genomic selection, and this work will continue into the future. Therefore, GBS techniques that can target individual variants would be more informative and flexible than those that use random variants. The GBS technique presented here is one such method. In this paper we present results demonstrating the performance and accuracy of this targeted GBS technique on 479 Holstein cows for 9102 SNP.

MATERIALS AND METHODS

Probes were designed to the flanking sequencing of 9102 target SNP, 5119 of which form part of the Illumina Infinium BovineSNP50 beadchip. These probes were used to capture DNA fragments containing the target sites from 479 bovine DNA samples in a method similar to that of (Shen *et al.* 2013). The products were PCR amplified using indexed primers that provided compatibility with sequencing on the HiSeq2000 genome analyser platform, and sequenced using single read chemistry.

An informatics pipeline was created to perform the following steps: 1) Sequence reads were trimmed of adapter sequence and poor quality bases (qscore < 20) using in-house scripts. 2) The quality filtered reads were aligned using BWA v0.7.7 (Li, Durbin 2009) to a custom reference genome (described below) allowing 4 mismatches and performing an exhaustive search for each read. 3) Samtools v0.1.19 (Li *et al.* 2009) mpileup tool with an input file listing all target SNP sites was used to create vcf files for all samples which in turn were used to create allele counts at all 9102 target SNP sites. 4) Allele counts were used to call genotypes, where the total count must be 6 or greater and a heterozygote had to have a minimum minor allele frequency of 0.167 (1 in 6 counts). Where the total allele count was <6, the genotype was set to NC (no call). 5) The genotypes (in UMD3.1 forward format) were then converted to TOP-TOP format (http://www.illumina.com/documents/products/technotes/technote_topbot.pdf).

471 of the DNA samples were also genotyped with the Illumina Infinium BovineSNP50 beadchip as per manufacturer's instructions, with genotype calls output in TOP-TOP format.

Starting contigs, consisting of the SNP and it's flanking sequence captured by the probes, were created for each SNP. Additional SNP within those sequences, discovered in the 1000 bull genomes (Daetwyler *et al.* 2014) run3.0 dataset, along with phased genotypes of all Holstein and Jersey animals, were used to create known haplotypes for each contig. Starting contigs were then edited to reflect the new haplotypes and new contigs created. All contigs were then combined into the custom reference genome which was used in the above informatics pipeline.

RESULTS AND DISCUSSION

Custom reference. The custom reference consisted of 27,918 contigs for 9102 target SNP. Target SNP had up to 17 SNP within the flanking sequence, and up to 53 different haplotypes per target. For targets with such large numbers of flanking SNP and so many different haplotypes, using standard reference genomes to align reads would result in reduced numbers of reads mapping as many reads would exceed the 4 allowed mismatches. This in turn would result in inaccurate genotype calls. Therefore custom reference building is essential, and must be revised periodically to incorporate new haplotypes within the population being genotyped.

GBS performance. On average 1.9 million reads were generated per library, of which 1.2 million reads passed quality filtering and trimming. On average 82% of all reads for each sample mapped to the custom reference genome. The rank ordered distribution of log10(read counts) between samples (Figure 1a) and between SNP across samples (Figure 1b) showed each sample had relatively uniform representation and that the majority of SNP were evenly covered. The method is very flexible and so the assay can be changed to include different or additional SNP. Therefore in the future SNP that failed, ie had very low read count, would be removed from the assay and new SNP that are deemed important would be included. This dataset presented here targets 9102 SNP, however this could be increased to very large numbers if need be, however proportionally more sequence reads would be required. The informatics pipeline for each sample took on average 4 hours to run, using less than 2 gigabases of memory.

Table 1. Average call rate for GBS and Infinium genotypes and concordance between the two for SNP and samples, where N is the number of SNP or samples and SD is the standard deviation

	N	GBS call rate (SD)	Infinium call rate (SD)	Concordance (SD)
SNP	5119	92.51% (0.175)	99.96% (0.002)	97.40% (0.069)
Sample	471	92.51% (0.073)	95.05% (0.067)	97.74% (0.047)

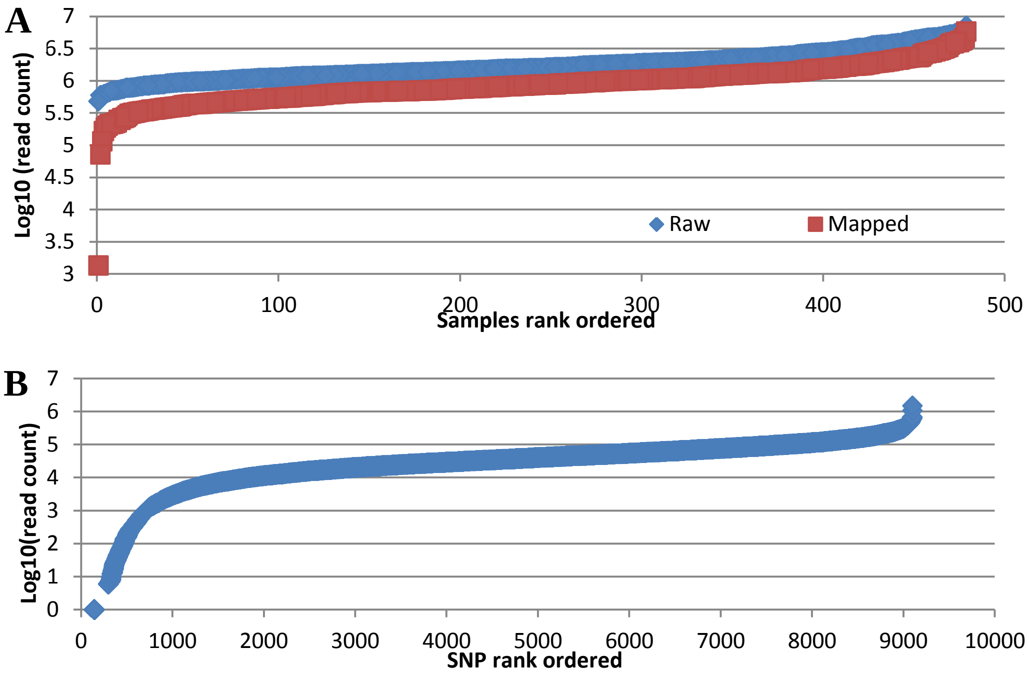

Figure 1. Distribution of reads amongst A) 479 samples (rank ordered) for raw and mapped reads and B) 9102 SNP (rank ordered) for mapped reads only. Read counts are expressed as log$_{10}$(read count).

GBS and Infinium Beadchip concordance. We calculated the call rate and concordance for the targeted GBS and infinium genotypes for each sample for the 5,119 SNP common to both assays (Table 1). The average call rate for the 471 samples was 93% and 95% for the targeted GBS and infinium assays, respectively. The concordance between assays was 98% (Table 1). The majority of samples had both high call rates and high concordance (Figure 2a). Only one sample failed to generate genotype calls in the targeted GBS assay. In a separate experiment, we observed a correlation between the proportion of reads mapping and the amount of DNA used in the targeted GBS assay. This experiment showed a consistent high proportion of reads mapped could be achieved with an input amount of >400 ng DNA (Figure 3). Where samples have high proportion of reads mapped call rates are maximised.

While the average genotype concordance between the targeted GBS and Infinium assays was high (97%, Table 1), there was a subset of SNP with low concordance (Figure 2b). Upon closer investigation it was found that the target region for those SNP had additional variants, either other SNP or indels, that were previously unknown. As these variants were not represented in our custom reference, they often caused the number of mismatches to exceed the limit of 4 specified in the alignment. This resulted in not all of the reads associated with the target loci being mapped, and therefore incorrect genotype calling. Further work is currently being undertaken to discover new variants to update the custom genome and improve genotype calling accuracy.

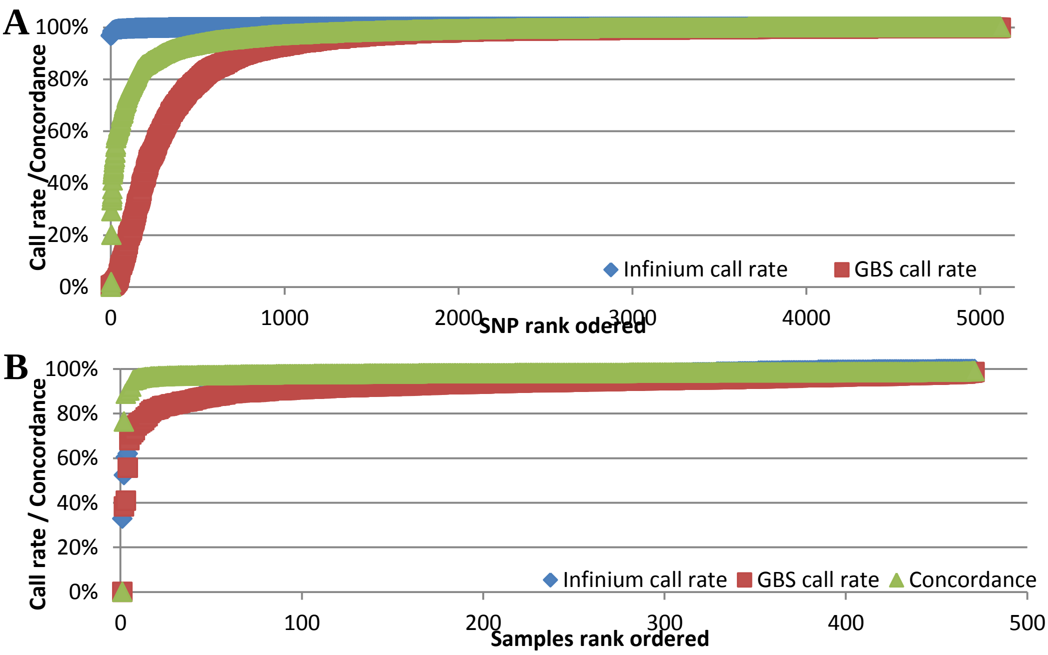

Figure 2. Infinium and GBS call rates as well as concordance between the two genotypes for A) samples (rank ordered) and B) SNP (rank ordered).

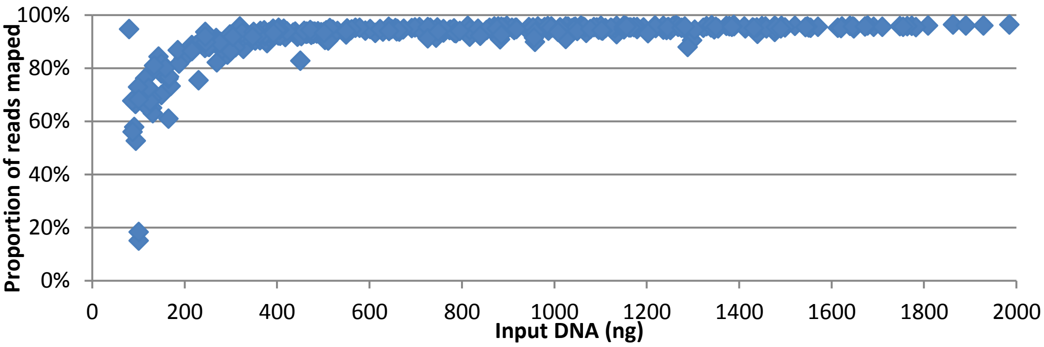

Figure 3. Relationship between the amount of input DNA in nanograms and the proportion of reads mapped to the custom genome for the GBS assay.

REFERENCES

Daetwyler, H.D., Capitan, A., Pausch, H., *et al.* (2014) Nature Genet. **46**:858.
Elshire, R.J., Glaubitz, J.C., Sun, Q., *et al.* (2011) *PLoS One.* **6**:e19379.
Li, H., Durbin, R. (2009) Bioinformatics **25**:1754.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J. *et al.* (2009) Bioinformatics **25**:2078.
Pryce, J., Hayes, B. (2012) Animal Production Science **52**:180.
Shen, P., Wang, W., Chi, A.-K., Fan, Y., Davis, R., Scharfe, C. (2013) Genome Medicine **5**:50.