EVALUATION OF POOLED WHOLE GENOME SEQUENCING (POOL-SEQ) TO RECOVER KNOWN GWAS SIGNALS (GENE EFFECTS)

A. Mohamed, L. Porto-Neto, A. Reverter and J. Kijas

CSIRO Agriculture and Food, Queensland Bioscience Precinct, St Lucia, QLD, 4067
Australia

SUMMARY

Whole-genome sequencing (WGS) of pools of individuals (Pool-Seq) provides a cost-effective method for genome-wide association studies (GWAS), and offers an alternative to sequencing of individuals that remains cost prohibitive. Pool-Seq is being increasingly used in population genomic studies in both model and non-model organisms. In this paper, the ability of Pool-Seq to recover known GWAS signals was evaluated. Existing GWAS data for 2,112 animals with 729K SNPs were obtained and pooled to simulate data obtained from a pooled WGS approach. Traditional GWAS results was compared with the absolute allele frequency difference (dAF) metric suitable for use with Pool-Seq data. Specifically, we tested the ability of dAF scans to recover known GWAS signals for two different traits with large and moderate gene effects. Pools of different sizes (50, 100 and 200 individuals per pool) were also compared. The results showed the ability of the absolute allele frequency difference (dAF) approach to recover known GWAS peaks obtained by traditional SNP association and recommended the use of a pool size of 100 individuals for DNA pooling.

INTRODUCTION

Recent advances in next generation sequencing (NGS) technologies have tremendously changed genetic research by increasing the number of known molecular markers in both model and non-model organisms such as: single nucleotide polymorphisms (SNPs) (Ellegren 2014). Despite these technical advances, genotyping large numbers of individuals with thousands of SNPs remains costly for large genome-wide association studies (GWAS). In this context, determination of allele frequencies from whole genome sequencing of pooled DNA samples has been suggested as a cost-effective alternative to individual genotyping (Sham *et al.* 2002). Many studies have successfully adopted this approach by comparing allele frequencies between cases and controls in both model and non-model organisms. For example, Abraham *et al.* (2008) performed a genome-wide (case-control) association study to understand Alzheimer's disease in human through the use of DNA pooling and highly significant association with late-onset Alzheimer's disease (LOAD) was observed at the *APOE* locus. To test for loci selected during domestication in chicken, Rubin *et al.* (2010) compared domesticated species to a wild population and identified one domestication-specific adaptation in the thyroid-stimulating hormone receptor (*TSHR*) gene. Pool genome-wide association study (Pool-GWAS) was also used to examine female abdominal pigmentation in *Drosophila melanogaster*. Candidate single-nucleotide polymorphisms (SNPs) near the pigmentation genes *tan* and *bric-à-brac 1* were identified when the allele frequencies in pools of light and dark females were compared (Bastide *et al.* 2013). Moreover, in Atlantic salmon Pool-Seq was used to investigate age at maturation in both wild and domesticated salmon where Ayllon *et al.* (2015) performed a genome wide association study using a pool sequencing approach (20 individuals per pool) of male salmon returning to rivers as sexually mature and revealed that 138 SNPs were significantly associated with sea age at puberty, 74 (48%) of these significant SNPs were located in a region on chromosome 25. More recently, Pool-Seq approach has been successfully deployed to identify genes for the timing of reproduction in Atlantic herring (Martinez Barrio *et al.* 2016).

In this paper, we used existing cattle SNP chip data obtained from individual animals and the associated GWAS results (Porto-Neto *et al.* 2014), to evaluate the power of the pool-seq approach. Using absolute allele frequency difference (dAF) , the ability to recover known GWAS signals was assessed after varying i) number of individuals per pool and ii) trait architecture. Outcomes of this analysis will assist in the design of experiments that seek to use pool-Seq as an alternative to traditional GWAS methodologies.

MATERIALS AND METHODS

Porto-Neto *et al.* (2014) performed a genome-wide association study using 2,112 Brahman cattle with 729,068 SNP genotypes per individual and analysed ten traits related to tropical conditions. Data were retrieved and re-analysed for two different traits, Coat Colour (colour) and rectal temperature (temperature). The first of these was selected to represent traits with large gene effects (colour), while the second exhibits genes of moderate effects. Plink software was used to make a subset of the data for 100 (top and bottom 50), 200 (top and bottom 100) and 400 (top and bottom 200) individuals from the 2,112 animals (representing pool sizes of 50, 100 and 200, respectively) using the --make-bed and --keep functions. Those individuals were assigned into two phenotypes for the GWAS case/control test and two clusters for the delta allele frequency test.

For traditional SNP association approach, an association (GWAS case/control scenario) test was implemented in Plink using the --assoc and --pheno functions. P values of all SNPs were obtained and –logP values were visualised as Manhattan plot generated in the R statistical computing environment.

For absolute allele frequency difference (dAF) approach, allele frequencies were calculated in Plink using the --freq and --within functions. Differences in allele frequencies were calculated for each SNP in the 2 clusters. A Manhattan plot of the absolute values of dAF of all SNPs was generated in R.

A significance threshold ($-\log P \geq 5$) was applied to filter the SNPs and the corresponding absolute values of delta AF of those significant SNPs were retrieved. This threshold was chosen in order to capture enough data for valid comparison and was used previously in GWAS analysis (for example) Cui *et al.* (2016). In order to compare the two approaches, -logP and delta AF values for 1) all significant SNPs and 2) SNPs under peaks were plotted in genomic order. Also simple linear regression was applied and R^2 values were obtained to test the correlation of the results obtained from both approaches for each trait in each pool size used.

RESULTS AND CONCLUSION

Traditional SNP association (GWAS case/control) identified SNPs significantly associated with the two traits under investigation for each of the pool sizes used. Strong GWAS signal(s) were identified in chromosomes 6, 7 and 13 for colour. On the other hand, multiple peaks in many chromosomes were identified for temperature (Figure 1). These findings are consistent with the results in the Porto-Neto *et al.* study.

Absolute allele frequency difference (dAF) results were obtained for the two traits and were compared with the GWAS results. The same GWAS signals were recovered using dAF in each of the two traits for each pool (Figure 1). For example, dAF values of the 32 significant SNPs of the major GWAS peak chromosome 13 in the trait colour showed the same trend as their corresponding –log P values (Figure 2).

Linear regression was used to test the correlation of the results obtained from both approaches for each trait in each pool size. Pool size of 50 showed the least R^2 values in each of the two traits, while there were very small increases in R^2 values (0.02 and 0.01 in colour and temperature, respectively) from pool size 100 to 200 (Table 1). For each trait, number of significant SNPs increased by increasing the pool size, with the pool size of 50 yielding the smallest number of significant SNPs (Table 1).

Table 1. A summary table of the results of comparing SNP association and delta allele frequency approaches for two traits in the Brahman cattle using 3 different pool sizes

Cattle trait	Pool size	No. of sig. SNPs (-log P ≥ 5)	R^2 value
Colour	50	95	0.04
	100	638	0.7
	200	4,349	0.72
Temperature	50	147	0.00007
	100	951	0.74
	200	1,323	0.75

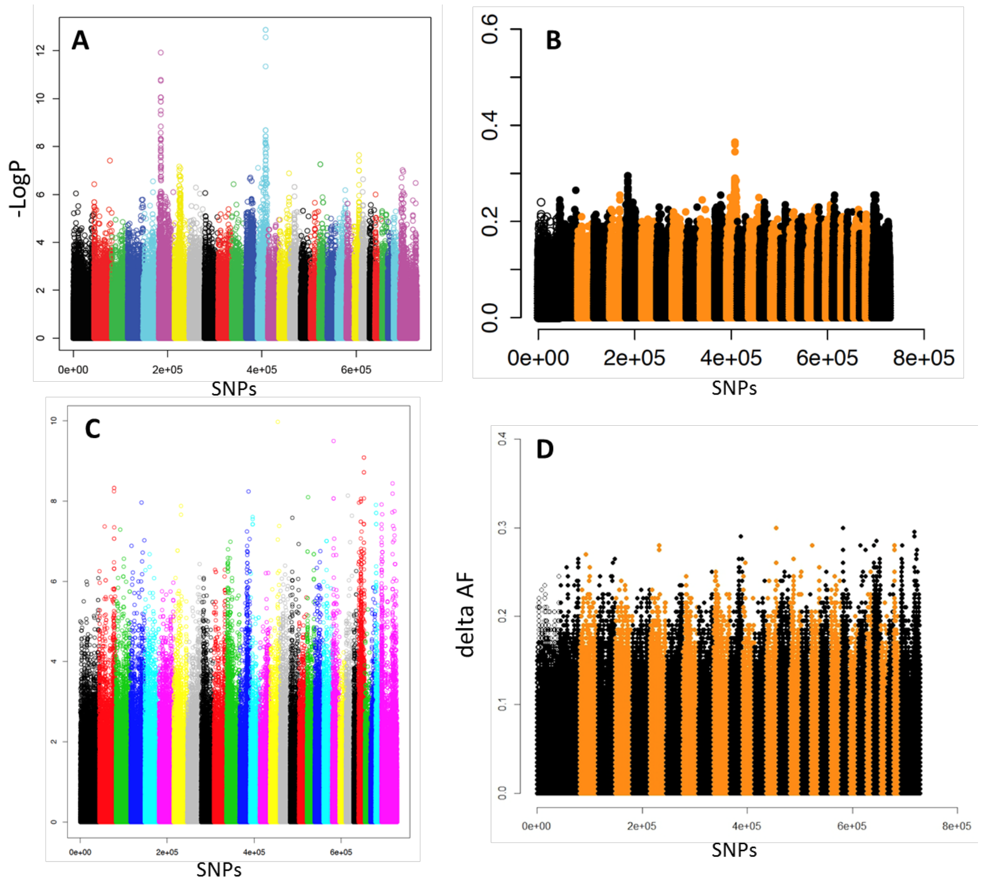

Figure 1. Comparison between GWAS SNP association and absolute allele frequency difference (dAF) approaches using a pool size of 100 (for example) revealed the ability of dAF to recover the same GWAS signals. A and B are Manhattan plots of –log P and absolute values of dAF values, respectively for colour while C and D are Manhattan plots of –log P and absolute values of dAF values, respectively for temperature

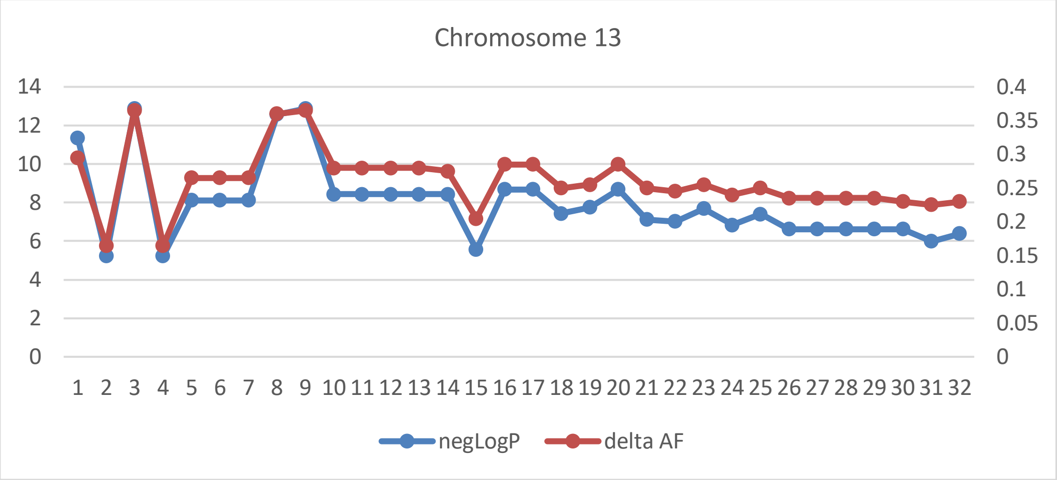

Figure 2. Zoom-in on the 32 significant SNPs (-log P ≥ 5) on a GWAS peak (chromosome 13) for the trait colour showing absolute delta allele frequency values (red) following the same trend as the –log P values (blue)

In conclusion, the absolute allele frequency difference (dAF) approach recovered the same GWAS signals obtained by traditional SNP association approach, for all the two traits under investigation. However, comparing the results from three different pool sizes suggested the use of pool size of 100 individuals for DNA pooling. These results confirm that, for traits controlled by a small number of major genes, the pool-Seq approach is likely to have the power to identify associations using the dAF metric. This opens the possibility to collect samples from only the phenotypic extremes within a population, before searching for associated genomic regions using a simple analytical approach and a modest research budget.

REFERENCES

Abraham R., Moskvina V., Sims R., *et al*. (2008). *BMC Med. Genomics*.**1**:44.
Ayllon F., Kjærner-Semb E., Furmanek T., *et al*. (2015). *PLoS Genet*. **11**:11.
Bastide H., Betancourt A., Nolte V., *et al*. (2013). *PLoS Genet.* **9**:6.
Cui, Z., Luo, J., Qi, C., *et al.* (2016). *BMC Genomics*. **17**:946.
Ellegren, H. (2014). *Trends Ecol. Evol*. **29**, 51–63.
Martinez Barrio, A., Lamichhaney, S., Fan, G., *et al*. (2016). *Elife* **5**, e12081.
Porto-Neto L.R., Reverter A., Prayaga K.C., *et al*. (2014). *PLoS ONE*. **9**:11.
Rubin, C. J., Zody M C., Eriksson J. *et al*. (2010). *Nature* **464**, 587–591.
Sham P., Bader J.S., Craig I., *et al*. (2002). **3**: 862-871.