PRELIMINARY ANALYSIS OF INTENSITY SIGNALS FROM SNP DATA BASED ON POOLED DNA SAMPLES IN BEEF AND POULTRY

A. Reverter[1], J. Henshall[2], R. McCulloch[1], R. Hawken[3] and S.A. Lehnert[1]

[1]CSIRO Food Futures Flagship, 306 Carmody Rd., Brisbane, QLD 4067, Australia
[2]CSIRO Animal, Food and Health Sciences, Chiswick, Armidale, NSW 2350, Australia
[3]Cobb-Vantress Inc., 4703 U.S. Highway 412 East, Siloam Springs, Arkansas 72761-1030, USA

SUMMARY

Pooled genomic DNA has been proposed as a cost-effective means of conducting genome-wide association studies (GWAS) as they reduce the number of genotyping assays required. However, algorithms for genotype calling of biallelic SNP are not adequate with pooled DNA samples because they assume the presence of two fluorescent signals, one for each allele, and operate under the expectation that, at most, two copies of the variant allele can be found for any given SNP and DNA sample. We adapted analytical methodology developed originally for two-channel gene expression microarray technology and applied it to SNP genotyping of pooled DNA samples in three datasets. We show that both differential hybridization (green minus red intensity signals) and abundance (average of both signals) provide useful information in the prediction of SNP allele frequencies. This is particularly true when making inference about extreme SNP that are either nearly fixed or highly polymorphic. We demonstrate the use of a model-based clustering method via mixtures of bi-variate normal distributions to capture the relationship between hybridization intensity values and SNP allele frequencies from pooled DNA samples. We further show that when the SNP allele frequencies are known, either because the individuals in the pools or from a closely related population are themselves genotyped, a polynomial regression model with linear and quadratic components can be developed with high prediction accuracy.

INTRODUCTION

According to Craig *et al.* (2005), SNP allelic frequencies are approximated using a correction factor for the ratio of the intensity of A and B probes corresponding to the two alleles. The authors proposed a pooling-test statistic which is a function of the number of individuals in the pool and the number and standard deviation of the replicates. The approach was successfully employed by Pearson *et al.* (2007) and general issues regarding the feasibility of GWAS using pooled DNA samples was recently and comprehensively reviewed by the same authors (Szelinger *et al.* 2011).

Brohede *et al.* (2005) proposed a so-called polynomial-based probe-specific (PPC) algorithm:

$$f(A_j) = \beta_0 + \beta_1 x_j + \beta_2 x_j^2$$

where x_j = $A_j/(A_j+B_j)$ and A_j and B_j are the observed signal intensity values for A and B alleles, respectively. Using the PPC approach in pooled DNA samples, Anantharaman and Chew (2009) concluded that the algorithm is highly accurate and reproducible especially when a suitable reference sample set is used to estimate the beta values for PPC.

Recently, Henshall *et al.* (2012) explored the value of logistic regression of genotype on phenotype to estimate the effect of SNP genotype from pooled DNA samples. Various pooling strategies were explored and pooled genotypes generated *in silico* as the frequencies of alleles in animals in the pool. The authors concluded that pooling DNA from individuals within groups was superior to pooling DNA across groups.

The aim of this paper is to conduct an initial examination of the value of analysing intensity signals from SNP data based on pooled DNA samples in beef and poultry. Analytical approaches include a model-based clustering method and a polynomial regression of signal intensities.

Table 1. Description of pooled DNA samples employed in this study

Dataset	Species	Chips	Description
DATA1	Bovine	3	Proof of Concept: One, two and five DNA samples are pooled and genotyped to explore the resulting signals.
DATA2	Bovine	24	Bovine Stature: 24 pools each with seven DNA samples from a genotyped population of 1,193 Santa Gertrudis cows.
DATA3	Chicken	12	Chicken Pools: 35 individually genotyped chickens are pooled in groups of 5, 10 or 20 and the pools genotyped.

MATERIALS AND METHODS

Data and edits. Three datasets were employed with varying number of SNP chips from 3 to 24 (Table 1). All chips contained ~50K SNP designed for bovine and chicken DNA.

DATA1 – Proof of Concept. In order to explore the pattern of clusters in the intensity signals that emerge from SNP data using pooled DNA samples, we designed a simple experiment comprising three 50K SNP chips and bovine samples. For this initial proof of concept, a single DNA sample as well as DNA samples from pooling two and five whole blood samples were analysed.

DATA2 – Bovine Stature. Blood samples from 76 cows where used to create 11 pools. Each pool contained equal amounts of whole blood from 7 individuals pooled according to their stature so that individuals with similar height were pooled together. To allow for the measurement of technical variation, one of the pools was replicated. These 12 pools were assembled from whole blood frozen and thawed once. The same pool design was then replicated, but with whole blood frozen and thawed twice. These 76 individuals were part of a larger population of 1,193 cows previously individually genotyped with the same SNP chip.

DATA3 – Chicken. The blood from 35 chickens individually genotyped using the Illlumina Chicken 60K SNP chip was pooled in groups of 5, 10 or 20, and DNA extracted. Technical replicates were performed to obtain a total of 12 chips.

MA-Plots in SNP genotype data. In MA-plots, the y-axis containing "M" (for Minus) or difference between green and red intensity signals was plotted against "A" (for Average) in the x-axis. The base-2 logarithmic scale was used throughout. These plots are often employed in the context of gene expression data to check for the need for further normalization of the raw intensity signals and, most importantly, to identify genes differentially expressed. In the context of SNP data from truly biallelic SNPs and individual samples, the intensity signals are supposed to be either perfect green (eg. genotype AA), or perfect red (eg. genotype BB) or perfect yellow (eg. genotype AB). However, when pooled samples are used deviations from "perfect" green, red or yellow are expected from any given SNP due to possible genotype differences among the samples.

Model-based clustering. Model-based clustering via mixture of distributions has been proposed by a number of authors to analyse microarray gene expression data (see for instance Reverter *et al.* (2006) and references therein). In the present study, the MA-paired values of each SNP were subjected to model-based clustering via a mixture of an *n*-component mixture of bi-variate normal densities. Parameters of the mixture were estimated using EMMIX (McLachlan *et al.* 2002).

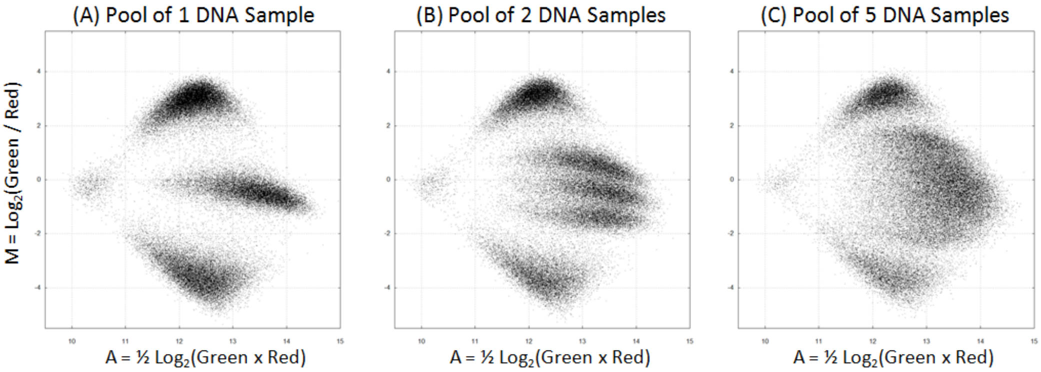
RESULTS AND DISCUSSION

Figure 1 shows the MA-Plots resulting from the analysis of DATA1 (Proof of Concept) along with the estimated distributions of the mixture models. When only the DNA of one individual is genotyped (Figure 1A) the MA-Plot shows three distinct clusters corresponding to the three possible genotypes: upper cluster for homozygous AA with positive M-values, middle cluster for heterozygous AB with intermediate M-values, and lower cluster for homozygous BB with negative M-values. When the DNA of two individuals is pooled and genotyped, the resulting MA-Plot (Figure 1B) shows five distinct clusters for 0 to 4 copies of the variant allele, B. Finally, when 5 DNA samples are pooled the clusters get diffuse with monomorphic SNPs occupying the extremes in the scale of M-values (Figure 1C). Importantly, in all three cases, one novel finding is that the clusters with intermediate M-values are associated with higher A-values and this is reflected in the estimated means for the distributions of the mixture models.

Figure 2 shows the MA-Plots resulting from the analyses of DATA2 (Bovine Stature; Figure 2A) and DATA3 (Chicken Pools; Figure 2B). Overlaid in these plots are the SNP first allele frequencies (FAF) estimated from genotyping the individual DNA samples and colour-coded from red to yellow to green for low, intermediate and high FAF, respectively. These plots anticipate the strong relationship between the FAF and the MA-values resulting from genotyping pools.

In particular, when the FAF was analysed as a function of the MA-values, the following second-degree polynomial was obtained (R^2 = 86%):

$$\text{FAF} = -0.655 - 0.154\text{M} + 0.211\text{A} + 0.0015\text{M}^2 - 0.0091\text{A}^2$$



$$f\begin{pmatrix} M \\ A \end{pmatrix} = 0.408 \times N\left[\begin{pmatrix} 2.844 \\ 12.248 \end{pmatrix}, \begin{pmatrix} 0.193 & 0.050 \\ 0.050 & 0.147 \end{pmatrix}\right]$$
$$+ 0.291 \times N\left[\begin{pmatrix} -0.432 \\ 13.307 \end{pmatrix}, \begin{pmatrix} 0.184 & -0.123 \\ -0.123 & 0.360 \end{pmatrix}\right]$$
$$+ 0.301 \times N\left[\begin{pmatrix} -3.430 \\ 12.373 \end{pmatrix}, \begin{pmatrix} 0.385 & -0.136 \\ -0.136 & 0.237 \end{pmatrix}\right]$$

$$f\begin{pmatrix} M \\ A \end{pmatrix} = 0.252 \times N\left[\begin{pmatrix} 3.030 \\ 12.096 \end{pmatrix}, \begin{pmatrix} 0.129 & 0.052 \\ 0.052 & 0.108 \end{pmatrix}\right]$$
$$+ 0.115 \times N\left[\begin{pmatrix} 0.565 \\ 13.221 \end{pmatrix}, \begin{pmatrix} 0.098 & -0.079 \\ -0.079 & 0.162 \end{pmatrix}\right]$$
$$+ 0.223 \times N\left[\begin{pmatrix} -0.127 \\ 12.586 \end{pmatrix}, \begin{pmatrix} 2.235 & -0.161 \\ -0.161 & 0.291 \end{pmatrix}\right]$$
$$+ 0.237 \times N\left[\begin{pmatrix} -0.793 \\ 13.491 \end{pmatrix}, \begin{pmatrix} 0.368 & -0.012 \\ -0.012 & 0.147 \end{pmatrix}\right]$$
$$+ 0.173 \times N\left[\begin{pmatrix} -3.530 \\ 12.207 \end{pmatrix}, \begin{pmatrix} 0.317 & -0.142 \\ -0.142 & 0.196 \end{pmatrix}\right]$$

$$f\begin{pmatrix} M \\ A \end{pmatrix} = 0.177 \times N\left[\begin{pmatrix} 3.017 \\ 12.162 \end{pmatrix}, \begin{pmatrix} 0.141 & 0.060 \\ 0.060 & 0.112 \end{pmatrix}\right]$$
$$+ 0.671 \times N\left[\begin{pmatrix} -0.069 \\ 13.123 \end{pmatrix}, \begin{pmatrix} 1.560 & -0.263 \\ -0.263 & 0.379 \end{pmatrix}\right]$$
$$+ 0.152 \times N\left[\begin{pmatrix} -3.194 \\ 12.276 \end{pmatrix}, \begin{pmatrix} 0.570 & -0.180 \\ -0.180 & 0.226 \end{pmatrix}\right]$$

Figure 1. MA-Plots and model-based clustering via mixtures of distributions for the three chips of DATA1 – Proof of Concept: A: a single DNA sample; B: A pool of two DNA; C: A pool of five DNA samples.

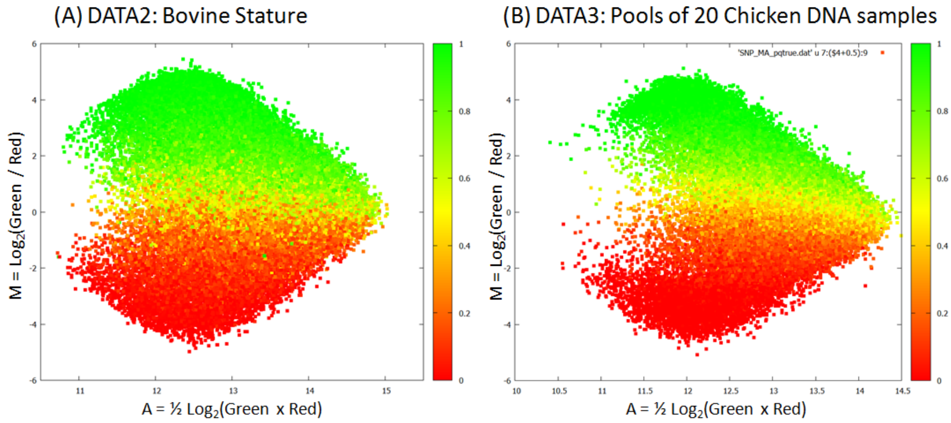


Figure 2. MA-Plots for DATA2 and DATA3 with overlaid estimates of first allele frequency from red (low frequency) to green (high frequency) based on genotypes of individual DNA samples.

CONCLUSIONS

The present study represents a first attempt to explore the numerical attributes of the intensity signals that should be considered when the intention is to genotype pools of DNA. We conclude that a strong relationship exists between the relative signal intensity of the two channels (red and green) and the SNP allele frequencies and show how this relationship can be formally ascertained by means of mixtures of distributions and polynomial equations. Further research is required to ascertain the extent to which model-based clustering and polynomial equations are suited to the use of pooled DNA samples to the development of application tools including estimation of family contributions to pools, SNP association to phenotypes and accurate genomic predictions.

ACKNOWLEDGMENTS

The work described under DATA2 was funded by a CSIRO Julius Career Award to Rachel J Hawken. The authors are grateful to Beef CRC project staff for collecting blood samples.

REFERENCES

Anantharaman R. and Chew F.T. (2009) *BMC Genet*. **10**:82.

Brohede J., Dunne R., McKay J.D. and Hannan G.N. (2005) *Nucl. Acids Res*. **33**:e142.

Craig D.W., Huentelman M.J., Hu-Lince D., Zismann V.L., Kruer M.C., Lee A.M., Puffenberger E.G., Pearson J.M. and Stephan D.A. (2005) *BMC Genomics* **6**:138.

Henshall J.M., Hawken R.J., Dominik S. and Barendse W. (2012) *Genet. Sel. Evol*. **44**:12.

McLachlan G.J., Bean R.W. and Peel D. (2002) *Bioinformatics* **18**:413.

Pearson J.V., Huentelman M.J., Halperin R.F., Tembe W.D., Melquist S., Homer N., Brun M., Szelinger S., Coon K.D., Zismann V.L., Webster J.A., Beach T., Sando S.B., Aasly J.O., Heun R., Jessen F., Kolsch H., Tsolaki M., Daniilidou M., Reiman E.M., Papassotiropoulos A., Hutton M.L., Stephan D.A. and Craig D.W. (2007) *Am. J. Hum. Genet*. **80**:126.

Reverter A., Ingham A., Lehnert S.A., Tan S.H., Wang Y., Ratnakumar A. and Dalrymple B.P. (2006) *Bioinformatics* **22**:2396.

Szelinger S., Pearson J.V. and Craig D.W. (2011) *Methods Mol. Biol*. **700**:49.