Published November 21, 2014

Numerical analysis of intensity signals resulting from genotyping pooled DNA samples in beef cattle and broiler chicken[1]

A. Reverter,*[2] J. M. Henshall,† R. McCulloch,* S. Sasazaki,*‡ R. Hawken,§ and S. A. Lehnert*

*CSIRO Food Futures Flagship and CSIRO Animal, Food and Health Sciences, 306 Carmody Road, St. Lucia, Brisbane, Queensland 4067, Australia; †CSIRO Food Futures Flagship and CSIRO Animal, Food and Health Sciences, Chiswick, Armidale, New South Wales 2350, Australia; ‡Laboratory of Animal Breeding and Genetics, Graduate School of Agricultural Science, Kobe University, Kobe 657-8501, Japan; and §Cobb-Vantress Inc., 4703 U.S. Highway 412 East, Siloam Springs, AR 72761-1030

ABSTRACT: Pooled genomic DNA has been proposed as a cost-effective approach in genomewide association studies (GWAS). However, algorithms for genotype calling of biallelic SNP are not adequate with pooled DNA samples because they assume the presence of 2 fluorescent signals, 1 for each allele, and operate under the expectation that at most 2 copies of the variant allele can be found for any given SNP and DNA sample. We adapt analytical methodology from 2-channel gene expression microarray technology to SNP genotyping of pooled DNA samples. Using 5 datasets from beef cattle and broiler chicken of varying degrees of complexity in terms of design and phenotype, continuous and dichotomous, we show that both differential hybridization (M = green minus red intensity signal) and abundance (A = average of red and green intensities) provide useful information in the prediction of SNP allele frequencies. This is predominantly true when making inference about extreme SNP that are either nearly fixed or highly polymorphic. We propose the use of model-based clustering via mixtures of bivariate normal distributions as an optimal framework to capture the relationship between hybridization intensity and allele frequency from pooled DNA samples. The range of M and A values observed here are in agreement with those reported within the context of gene expression microarray and also with those from SNP array data within the context of analytical methodology for the identification of copy number variants. In particular, we confirm that highly polymorphic SNP yield a strong signal from both channels (red and green) while lowly or nonpolymorphic SNP yield a strong signal from 1 channel only. We further confirm 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 multiple regression model with linear and quadratic components can be developed with high prediction accuracy. We conclude that when these approaches are applied to the estimation of allele frequencies, the resulting estimates allow for the development of cost-effective and reliable GWAS.

Key words: cattle, genomewide association study, pooled DNA, poultry

J. Anim. Sci. 2014.92:1874–1885
doi:10.2527/jas2013-7133

INTRODUCTION

Pooling DNA samples can provide a cost-effective approach to increase power in genomewide association studies (**GWAS**; Sham et al., 2002). However, the estimation of SNP allele frequencies in a pooled DNA sample requires a numerical procedure that exploits the relative intensity signal of the 2 alternate alleles.

According to Craig et al. (2005), allelic frequencies are approximated using a *k*-correction method such that $f = A/(A + kB)$, in which k is a correction

[1]We acknowledge the Cooperative Research Center for Beef Genetic Technologies for provision of individual cattle genotyping data and thank Paul Williams, Nick Corbet, and Dom Niemeyer for blood sample collections from cattle populations. We thank Dr Ian Braithwaite, who contributed pregnancy data on beef cattle, and pedigree farm technical crew and laboratory staff at Cobb-Vantress Inc. for the collection and processing of chicken samples. The authors are grateful to Yutao Li and Sonja Dominik for their review of this manuscript.

[2]Corresponding author: tony.reverter-gomez@csiro.au

Received September 8, 2013.
Accepted February 6, 2014.

Table 1. Description of the 5 datasets used in this study

Dataset	Species	Samples	Description
DATA1	Bovine	3	Bovine Proof of Concept: A single DNA sample and a pool of 2 and a pool of 5 samples were genotyped to explore the emerging clusters of intensity signals.
DATA2	Bovine	24	Bovine Stature: Samples from 76 individuals to generate 24 pools each with 7 samples from a genotyped population of 1,193 Santa Gertrudis cows pooled according to stature.
DATA3	Bovine	69	Bovine Pregnancy Status: Samples from 959 age-matched cows were pooled according to pregnancy status (644 pregnant and 315 nonpregnant).
DATA4	Chicken	12	Chicken Proof of Concept: Thirty-five individually genotyped chickens were pooled in groups of 5, 10, or 20 and with 2 blood volumes and 2 technical replicates.
DATA5	Chicken	103	Chicken Feed Efficiency: One hundred three pools were genotyped from 2,007 chickens pooled according to their average feed efficiency performance.

factor and A and B represent the intensity signals from the 2 alleles in the SNP. The authors devised a pooling test statistic as a function of the number of individuals in the pool, the SD of the technical replicates, and the number of replicates. The approach was successfully used by Pearson et al. (2007) and general issues regarding the feasibility of GWAS using pooled DNA samples was explored by the same authors in Szelinger et al. (2011).

From a different perspective, Brohede et al. (2005) proposed a polynomial-based algorithm to estimate allele frequencies and its optimality was later ascertained by Anantharaman and Chew (2009) concluding that the algorithm is highly accurate and reproducible, especially when a reference sample is used to estimate parameters of the polynomial.

More 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 confirmed that pooling DNA from individuals within groups is superior to pooling DNA across groups.

The aim of this paper was to conduct an initial examination of the value of analyzing intensity signals from SNP chips based on pooled DNA samples from beef cattle and broiler chicken. Analytical approaches include model-based clustering and polynomial regression of signal intensities.

MATERIALS AND METHODS

Blood samples were collected from commercial herds and flocks under the guidance of the local committees for the care and use of animals. Cattle blood samples were collected under approval number A6/2011 of the Commonwealth Scientific and Industrial Research Organisation Brisbane Animal Ethics Committee, chicken blood samples were collected following the Cobb-Vantress Inc. Animal Welfare Policy.

Data and Edits

We used 5 datasets with varying number of samples from 3 to 103. All samples were genotyped for approximately 50,000 SNP designed for bovine or chicken DNA. Table 1 lists and briefly describes the structure of the 5 datasets. Further details are provided next.

1. ***DATA1 – Bovine Proof of Concept.*** To explore the pattern of clusters in the fluorescent intensity signals that can be expected from SNP data using DNA from pooled blood samples, we designed a simple experiment comprising 3 bovine samples genotyped using the BovineSNP50 V2 array chip (Illumina Inc., San Diego, CA). For this initial, proof of concept experiment, DNA prepared from a single blood sample and a pool of 2 and a pool of 5 blood samples was tested. For pooled blood DNA preparation, 200 μL of whole blood from each animal was combined and mixed by inverting the tubes. Subsequently, DNA was extracted from a subsample of the blood mixture with the DNeasy Blood and Tissue Kit (Qiagen Inc., Hilden, Germany).
2. ***DATA2 – Bovine Stature.*** Blood samples from 76 individual cows where used to create 11 pools. Ten pools contained equal volumes of blood from 7 individuals and 1 pool contained 6 samples. Genomic DNA was extracted from the pooled blood as described above and genotyped using the BovineSNP50 V2 array chip (Illumina). Individuals within a pool were selected according to their stature so that individuals with similar height were pooled together. To allow for the measurement of technical variation, 1 of the pools was replicated. These criteria resulted in 12 pools, which were further subjected to 2 treatments based on the number of freeze–thaw cycles. In the first treatment, the DNA was extracted from whole blood frozen and thawed twice, while in the second treatment, the DNA was extracted whole blood frozen and thawed 3 times. Importantly, these 76 individuals were part of a larger population of 1,193 cows previously individually genotyped with the BovineSNP50 V2 chip (Illumina) previously reported by Henshall et al. (2012).

3. ***DATA3 – Bovine Pregnancy Status.*** Blood samples from a total of 959 3-yr-old Santa Gertrudis cows at their first rebreeding opportunity were pooled according to pregnancy status. The animals were unrelated to animals in DATA2 and were part of a commercial cow herd located in Queensland and bred by natural mating and subjected to once-yearly pregnancy testing performed at the time of weaning the previous year's calves. All cows were lactating at the time of pregnancy testing, which was performed by palpation of the reproductive tract. Cows that had not reconceived by natural mating while suckling their first calf received the designation of "nonpregnant." Cows in which pregnancy was detected by palpation were called "pregnant." There were 644 pregnant and 315 nonpregnant cows and 69 pools were created and DNA was extracted and genotyped using the BovineSNP50 V2 chip (Illumina). On average, there were 20.51 DNA samples in a pool and these ranged from 1 (3 pools) to 25 (6 pools).
4. ***DATA4 – Chicken Proof of Concept.*** Thirty-five broiler chickens were individually genotyped using the Illumina ChickenSNP60 chip (Illumina). The chip contains 57,636 SNP markers from a whole-genome panel developed by Groenen et al. (2009). Blood from the 35 chickens previously genotyped was pooled in groups of 5, 10, or 20, and DNA was extracted from blood pools and genotyped. For the DNA extraction, 2 blood volumes were explored, 20 and 50 μL, and 2 technical replicates performed to use a total of 12 samples (i.e., 3 pool sizes × 2 blood volumes × 2 replicates).
5. ***DATA5 – Chicken Feed Efficiency.*** A total of 2,007 individual chicken blood samples were used to make 103 blood pools according to their average feed efficiency (**FE**). On average, there were 19.5 individuals in each pool (range: 13 to 23) and individuals within a pool were from the same management group ($n = 6$) and sex ($n = 2$). The original data represented 80 sire families and contained 776 males and 1,231 females with an average (SD) FE of 0.00 (99.66) and –1.04 (97.15), respectively.

Intensity Signals in the Context of SNP Genotype Data

The allele specific intensity signals can be explored by means of the scatter plot of the M values (green minus red intensity signals) in the y axis against the A values (average of green and red intensity signals) in the x axis. The base-2 logarithmic scale is used throughout. Originally coined by Dudoit et al. (2002) in the context of gene expression data, these plots are typically used to check for the need to further normalize that data and, most importantly, to identify genes differentially expressed. In the context of SNP genotype data from truly biallelic SNP and individual samples, the intensity signals are supposed to be either perfect green (e.g., genotype AA) or perfect red (e.g., genotype BB) or perfect yellow (e.g., 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.

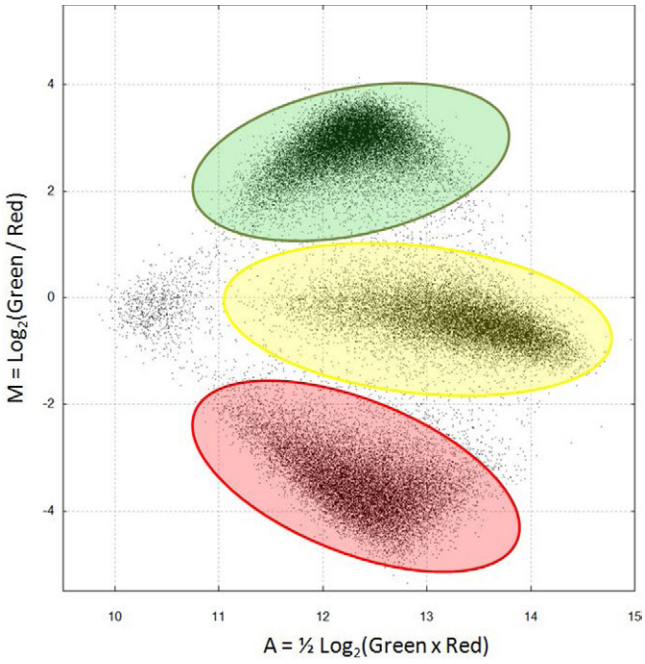

Figure 1. Rationale for the use of the scatter plot of the M values (green minus red intensity signals) against the A values (average of green and red intensity signals) in the context of genotyping pools of DNA. Each point in the scatter represents a SNP. Three clusters are clearly distinguishable from green (most individuals in the DNA pool having genotype AA for these SNP), to yellow (genotype AB), to red (genotype BB). See online version for figure in color.

Figure 1 illustrates the rationale for the use of the scatter plot of M and A values in the context of genotyping pools of DNA. Each point in the scatter represents a single SNP. From top to bottom, 3 distinct clusters can be identified: 1) The uppermost cluster, or "green" cluster, corresponds to the SNP for which the green signal predominates (i.e., the "A" allele is more common than the "B" allele) and most individuals in the DNA pool have genotype AA for these SNP. 2) The middle cluster, or "yellow" cluster, corresponds to the SNP for which neither signal, green or red, predominate resulting in the emission of a yellow signal. Most individuals in the DNA pool have genotype AB for these SNP. 3) The bottom cluster, or "red" cluster, corresponds to the SNP for which the red signal predominates (i.e., the "A" allele is less common than the "B" allele) and most individuals in the DNA pool have genotype BB for these SNP.

Model-Based Clustering via Bivariate Mixture Models

Model-based clustering via mixture of distributions has been proposed by a number of authors to analyze microarray gene expression data in a uni- and bivariate fashion (see for instance Reverter et al. [2006] and references therein). In the present study, for each SNP in *i*, the paired data points in M_i and A_i were assumed to be independent observation from a *p*-component mixture model (or clusters) with probability density function:

$$f\begin{pmatrix} M_i \\ A_i \end{pmatrix} = \sum_{j=1}^{p} \pi_j \varphi_j \left(\begin{pmatrix} M_i \\ A_i \end{pmatrix}; \grave{\mathbf{i}}_j, \mathbf{V}_j \right),$$

in which

$$\varphi_j \left(\begin{pmatrix} M_i \\ A_i \end{pmatrix}; \grave{\mathbf{i}}_j, \mathbf{V}_j \right)$$

denotes a bivariate normal density function with 2-dimensional mean vector $\boldsymbol{\mu}_j$ and a 2 × 2 covariance matrix $\mathbf{V}_j$, and π_j are the mixing proportions constrained to be nonnegative and sum to unity. In the present study, we explored $p = 3$ and $p = 5$ clusters depending on the number of DNA samples in the pools. In all cases, parameters of the mixture model were estimated using the EMMIX software (McLachlan et al., 2002).

Multiple Regression Models

For DATA2 (Bovine Stature), DATA4 (Chicken Proof of Concept), and DATA5 (Chicken Feed Efficiency) for which SNP genotypes were available on pools as well as on the individual DNA samples comprising the pools, we used the PROC REG procedure (SAS Inst. Inc., Cary, NC) to analyze the frequency of the first allele (*p*) in a multiple regression model that included the effects of M_i and A_i with linear and quadratic components. The regression model was as follows:

$$p_i = \beta_0 + \beta_1 M_i + \beta_2 A_i + \beta_3 M_i^2 + \beta_4 A_i^2 ,$$

in which p_i is the frequency of the first allele for the *i*th SNP and obtained from genotyping individual DNA samples, M_i and A_i are the intensity signal metrics defined earlier, and the β correspond to the estimated partial regression coefficients.

Finally, the regression equation resulting from DATA2 (Bovine Stature) was used to estimate the p_i in DATA3 (Bovine Pregnancy Status) for which only pools were available.

Genomewide Association Studies

With the first allele frequencies estimated as previously described we performed GWAS for DATA3 (Bovine Pregnancy Status) and DATA5 (Chicken Feed Efficiency).

For the GWAS of DATA3, we adapted the methodology described by Macgregor et al. (2006) for case-control samples and with "pregnant" vs. "nonpregnant" as our contrast. In particular, the difference in allele frequency between pregnant and nonpregnant pools was tested using the following test statistic:

$$T = \frac{\left(\tilde{p}_P - \tilde{p}_N\right)^2}{\tilde{V} + 2\,\mathrm{var}(e_{\mathrm{pool-1}})} \sim \chi_1^2 .$$

In the expression above, $\tilde{p}_P$ and $\tilde{p}_N$ denote the first allele frequency estimated for pregnant and nonpregnant pools, respectively. The binomial sampling variance in $\tilde{V}$ is estimated by $\tilde{V} = \tilde{p}_P(1-\tilde{p}_P)/2n_P + \tilde{p}_N(1-\tilde{p}_N)/2n_N$, in which n_P and n_N denote the number of pregnant and nonpregnant pools, respectively. Finally, $\mathrm{var}(e_{\mathrm{pool-1}})$ is the variance of the pool-specific error in allele frequency estimation and computed over all SNP as follows:

$$\mathrm{var}(e_{\mathrm{pool-1}}) = \frac{1}{2}\mathrm{mean}\left[\left(\tilde{p}_P - \tilde{p}_N\right)^2 - \tilde{V}\right] .$$

For the GWAS of DATA5 (Chicken Feed Efficiency), the additive effect of each SNP was estimated based on logistic regression of estimated allele frequencies in pools on phenotype as measured by the average FE of the individuals in each pool. Logistic regression analyses for the pools were performed using analytical methodology described in Henshall et al. (2012) and estimated SNP effects from pooled DNA samples were compared with those obtained from individual DNA samples on the available 2,007 chicken.

For the GWAS on individual DNA samples, the effect of each SNP was estimated in turn using the following mixed model equations:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{S}_i a_i + \mathbf{e}$$

in which **y** represents the vector of FE measures across the 2,007 chicken, **X** is the incidence matrix relating fixed effects in β with observations in **y**, **Z** is the incidence matrix relating random additive polygenic effects in *u* with observations in **y**, $\mathbf{S}_i$ is the vector of genotypes for the *i*th SNP across all chicken, a_i represents the additive effect of the *i*th SNP, and **e** is the vector of random residual effects. Fixed effects included in the model were contemporary group and sex with $n = 6 \times 2 = 12$ levels. We used Qxpak5 (Pérez-Enciso and Misztal, 2011) to estimate SNP additive effects and test their significance.

Following Bolormaa et al. (2013), the false discovery rate (**FDR**) was calculated as

$$\text{FDR} = P\left[1-(S/T)\right]/(S/T)(1-P)$$

In which P is the P-value tested (e.g., 0.0001), S is the number of SNP that were significant at the P-value tested, and T is the total number of SNP tested.

RESULTS AND DISCUSSION

Summary Statistics for M and A Values

Preliminary analyses of DATA2 (Bovine Stature) across all SNP revealed no significant association (P-value = 0.3108) of the treatment effect (i.e., 2 versus 3 rounds of freeze–thaw cycles) on the red to green relative intensity signals. Similarly, for DATA4 (Chicken Proof of Concept) neither pool size (i.e., 5 vs. 10 vs. 20; P-value = 0.9573) nor volume (20 vs. 50 μL; P-value = 0.9899) nor technical replicate (P-value = 0.8790) were significant sources of variation in the relative intensity signals.

Table 2 provides the number of records, SNP, and summary statistics for the M and A intensity signal metrics across the 5 datasets. The number of records (N) is the product of the number of SNP times the number of samples genotyped. Across datasets, consistent M values were observed: centered at 0 and with a SD averaging 2.18 (SD range from 1.99 to 2.35 for DATA5 and DATA1, respectively). Values for the A metric were also consistent across datasets averaging 12.78 and with a CV ranging from 5.8% for DATA1 and DATA4 to 6.8% for DATA2 and DATA3.

Importantly, the range of M and A values observed here are in agreement with those reported within the context of gene expression microarray. See for instance the early work of Dudoit et al. (2002) and Bolstad et al. (2003) for 2-channel cDNA and 1-channel oligonucleotide array data, respectively. Our M and A values also agree with those from SNP array data within the context of analytical methodology for the identification of copy number variants (recently reviewed by Li and Olivier [2013]).

Plots of M and A Values and Mixtures of Distributions

Figure 2 shows the plots of M and A values resulting from the analysis of DATA1 (Bovine Proof of Concept) along with the estimated distributions of the mixture models. When only the DNA of 1 individual is genotyped the M and A plot shows 3 tight clusters (Fig. 2A) corresponding to the 3 possible genotypes: an upper cluster of positive M values capturing 40.8% of the SNP according to the mixture model and presumably all with homozygous AA genotype, a middle cluster of intermediate (i.e., around 0) M values capturing 29.1% of the SNP according to the mixture models and presumably all with heterozygous AB genotype, and a lower cluster of negative M values made of the remaining 30.1% of the SNP presumably with homozygous BB genotype.

Table 2. Summary statistics for the M[1] and A[1] intensity signal metrics across the 5 datasets

Dataset	SNP	N	Metric	Mean[2]	SD[2]	Minimum[2]	Maximum[2]
DATA1	54,606	163,818	M	–0.00	2.35	–5.35	4.18
			A	12.67	0.74	9.64	14.96
DATA2	47,762	1,146,288	M	0.05	2.17	–6.71	5.67
			A	13.14	0.90	8.49	15.50
DATA3	47,844	3,301,236	M	0.02	2.33	–8.15	6.80
			A	12.97	0.88	7.04	15.25
DATA4	49,756	597,072	M	–0.02	2.06	–6.19	5.07
			A	12.66	0.74	9.53	14.76
DATA5	57,589	5,931,667	M	–0.22	1.99	–6.94	4.55
			A	12.47	0.75	7.91	14.64

[1]Metrics: M stands for minus and is computed from the difference between green and red intensity signals; A stands for average and is computed from the average of green and red intensity signals. Intensity signals are expressed in base-2 logarithmic units.

[2]Units are fluorescent intensity units in base-2 log-scale.

The expectation of a clear distinction between the 3 clusters observed in Fig. 2A is what SNP genotype calling algorithms exploit in the mapping of raw allele A and allele B intensities from each SNP into the 3 genotype calls: AA, AB, or BB. See for instance the work of Ritchie et al. (2011) and Chai et al. (2010) respectively for Illumina (Illumina Inc., San Diego, CA) and Affymetrix SNP chips (Affymetrix Inc., Santa Clara, CA).

When samples from 2 individuals are pooled and genotyped, the resulting plot of M and A values shows five distinct clusters (Fig. 2B) corresponding to observing 0 to 4 copies of the variant allele, B. This would be equivalent to genotyping a biallelic SNP (still with alleles A and B) on a DNA from a tetraploid individual where the 5 possible genotypes (percent of SNP according to the mixture models in brackets) would be AAAA (25.2%), AAAB (11.5%), AABB (22.3%), ABBB (23.7%), and BBBB (17.3%). Finally, when 5 samples are pooled and genotyped, the clusters get diffuse (Fig. 2C) with monomorphic SNP occupying the extremes in the scale of M values.

Importantly, in all 3 cases, 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. This finding anticipates the importance of using not only the relative intensity signal of each channel (red and green) captured by M values but also the abundance of both signals captured by the average in A values.

While there is not a precise reason as to why the cluster of intermediate M values is associated with higher A values, it is tempting to speculate that highly polymorphic SNP yield a strong signal from both channels (red and green) while lowly or nonpolymorphic SNP yield a strong signal from 1 channel only. When averaging is made in the computation of A values, highly polymor-

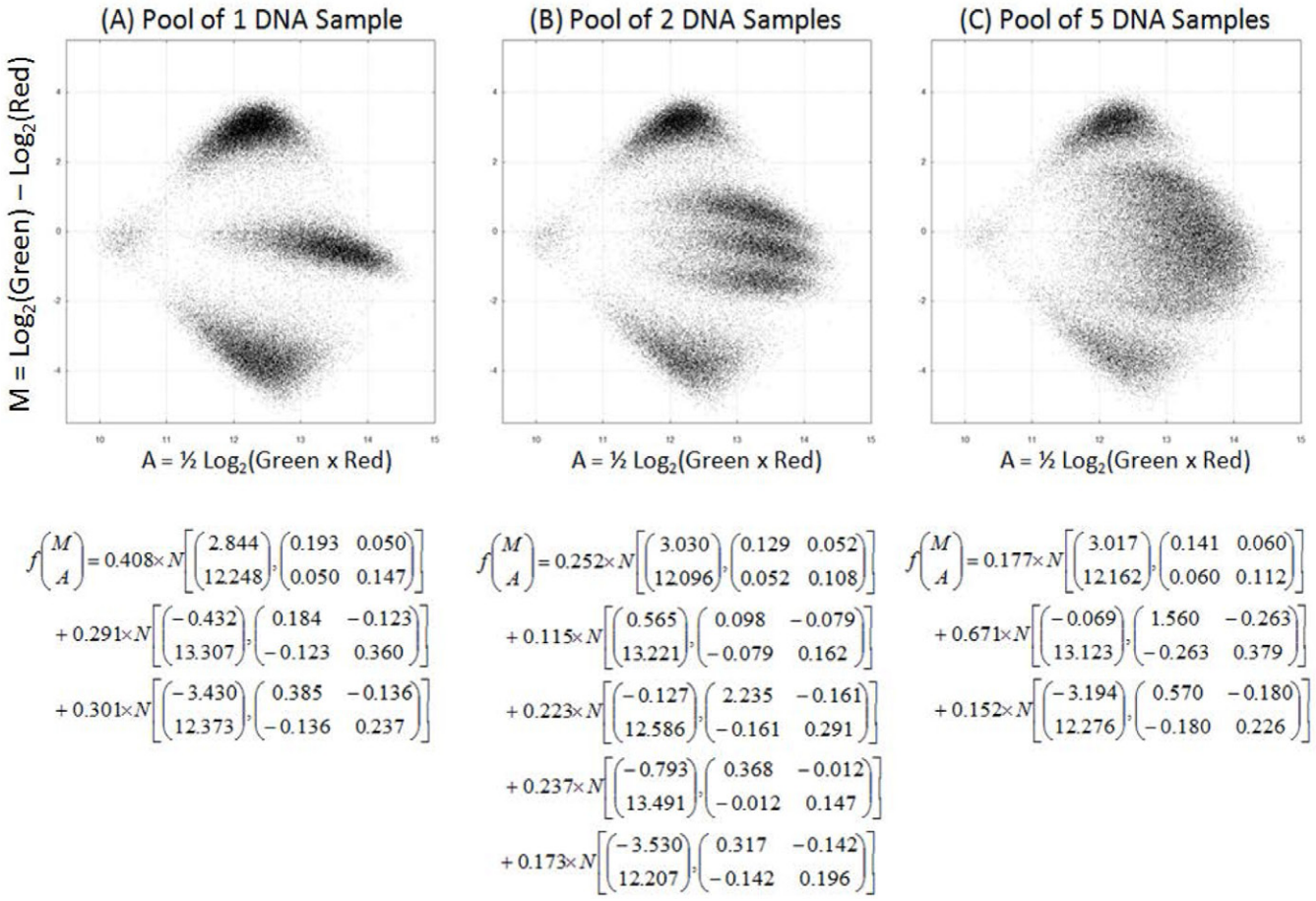

Figure 2. Scatter plots of M values (green minus red intensity signals) against A values (average of green and read intensity signals) and parameters of the model-based clustering via mixtures of distributions for the three samples of DATA1 – Proof of Concept: (A) a single DNA sample, (B) a pool of 2 DNA samples, and (C) a pool of 5 DNA samples.

phic SNP show twice as much average signal than lowly polymorphic SNP and this doubling in the average signal is reflected by a difference of 1 in the base-2 log-scale. Supporting this speculation, the estimates of the mixtures of distribution show a difference of 1 between the estimate of the mean A values (approximately 12.3 intensity units) for the upper and lower clusters and the estimate of the mean A values (approximately 13.3) for the middle cluster. Quite significantly, this pattern of 1 unit difference can be observed in Fig. 5A of Ritchie et al. (2011) with an example of signals from a good quality array.

Estimation of Allele Frequency

Figure 3 shows the M and A plots resulting from the analyses of DATA2 (Bovine Stature) and DATA4 (Chicken Proof of Concept). Every point in these plots represents a single SNP and its location in the M and A coordinates corresponds to the average across all replicates. Overlaid in these plots are the SNP first allele frequencies (p_i) estimated from genotyping the individual DNA samples and color coded from red to yellow to green for low, intermediate, and high p_i, respectively. These plots illustrate the strong relationship that exists between the p_i and the M and A values resulting from genotyping pools and confirming the expectation of intermediate M values (i.e., near 0) corresponding to highly polymorphic SNP, while extreme M values (i.e., either extreme positive or extreme negative) correspond to lowly polymorphic SNP.

When the p_i were modeled as a function of the M and A values, we estimated the multiple regression equations given in Table 3. The goodness of fit, as explained by the R^2, indicated that over 78% (and a maximum of 92.4% for DATA4) of the variation in p_i can be explained by linear and quadratic components of M and A values. These R^2 are similar to those reported by Brohede et al. (2005), which averaged 90.4 and 95.9% for biological and technical replicates, respectively. They also contrast with the 96% of the corrected relative allele signal methodology recently reported by Teumer et al. (2013) with technical replicates and using the Birdseed2 genotype calling algorithm (Korn et al., 2008).

Consistent across the 3 datasets (DATA2, DATA4, and DATA5), we observed a significant ($P < 0.001$) negative β_1 (the partial regression coefficient associated with the linear component of the M values) indicating

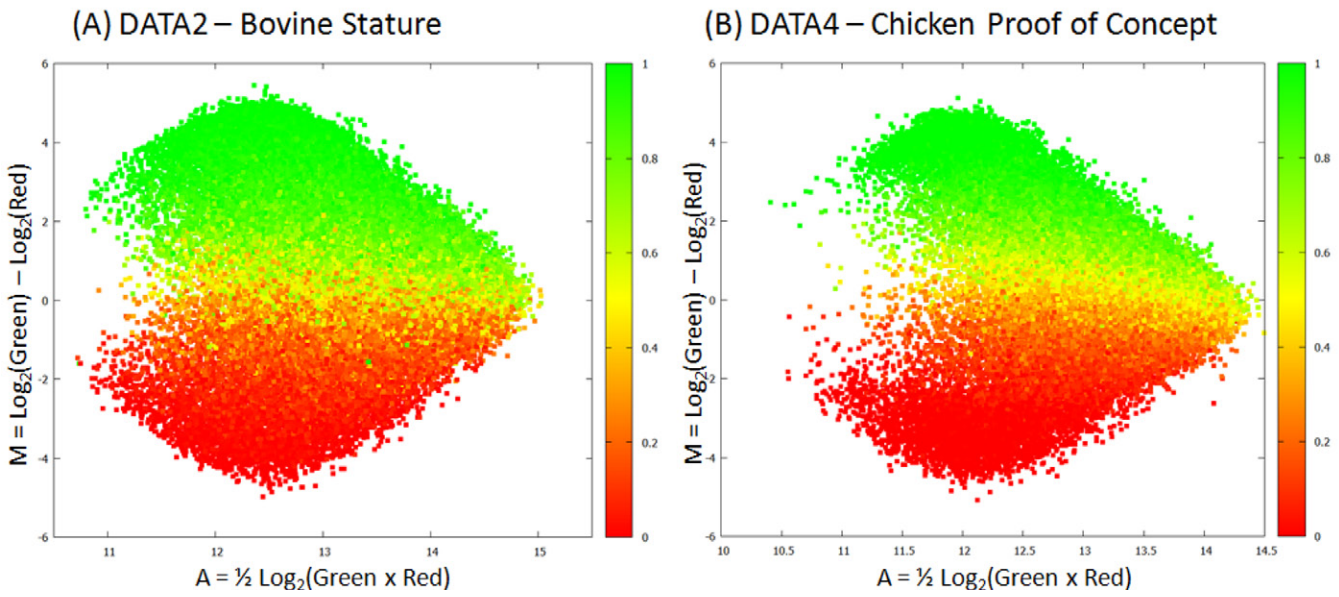

Figure 3. Scatter plots of the M values (green minus red intensity signals) against the A values (average of green and red intensity signals) for (A) DATA2 – Bovine Stature and (B) DATA4– Chicken Proof of Concept and with overlaid estimates of first allele frequency from red (low allele frequency) to green (high allele frequency) based on genotyping of individual samples. See online version for figure in color.

that as the M value increases the frequency of the first allele decreases. Similarly, a consistent and significant ($P < 0.001$) positive β_2 (the partial regression associated with the linear component of the A values) was estimated indicating that as A values increases the frequency of the first allele also increases. However, this linear increase is offset by the significant ($P < 0.001$) and negative estimate of β_4 (the partial regression associated with the quadratic component of the A values).

Furthermore, when this polynomial was used to predict the p_i from the pools in DATA3 (Bovine Pregnancy Status) the results allowed us to undertake a GWAS for pregnancy rate (Fig. 3).

Genomewide Association Studies

The multiple regression equation obtained with DATA2 (Bovine Stature) was used to estimate the p_i from DATA3 (Bovine Pregnancy Rate) and these were then used to perform a GWAS to identify loci associated with pregnancy rate. Figure 4 shows the distribution along the genome of *P*-values (Manhattan plot) of the 47,844 SNP included in the GWAS as well as the Q-Q plot of *P*-values.

The peaks of significance in the Manhattan plot are less clear than what it is usually found from the analysis of individual DNA samples. Nevertheless, the number of SNP found to be significantly associated with pregnancy rate at the nominal *P*-values of 0.01, 0.001, and 0.0001 was 943 (FDR = 50.24%), 321 (FDR = 14.82%), and 145 (FDR = 3.29%), respectively. When the genomic region of the 145 associated SNP ($P < 0.0001$) was surveyed we found that 34 of them (Table 4) were related to genes reported in the recent review of Fortes et al. (2013a) as associated with fertility in cows. Prominent among these genes are *PENK* and *IGF2*. The preproenkephalin gene (*PENK*) is in the region of BTA14 reported for its pleiotropic potential in cattle (Fortes et al., 2013b) and where *PLAG1* gene is encoded. Karim et al. (2011) showed that a mutation on *PLAG1* affecting stature in cattle also changed the expression of *PENK* in fetal tissues. The insulin-like growth factor 2 gene (*IGF2*) is an imprinted gene, expressed only from the paternal allele, and exerts its effects by binding the IGF1 receptor (Baker et al., 1993). The role of IGF1 pathway genes and their association with age of puberty in cattle has been documented by Fortes et al. (2012).

The application of the equation to estimate allele frequencies for DATA5 (Chicken Feed Efficiency) allowed us to perform GWAS to identify loci associated with FE using the 103 DNA pools. Figures 5A and 5B

Table 3. Parameter estimates of the multiple regression of first allele frequency on M[1] and A[1] intensity signal metrics[2]

Dataset	β_0	β_1	β_2	β_3	β_4	R^2, %
DATA2	–5.586	–0.141	0.933	0.003	–0.036	78.4
DATA4	–0.054 NS	–0.154	0.129	–0.003	–0.007	92.4
DATA5	–5.808	–0.153	1.021	–0.000 NS	–0.042	80.2

[1]Metrics: M stands for minus and is computed from the difference between green and red intensity signals; A stands for average and is computed from the average of green and red intensity signals. Intensity signals are expressed in base-2 logarithmic units.

[2]The regression model is as follows: $p_i = \hat{a}_0 + \hat{a}_1 M_i + \hat{a}_2 A_i + \hat{a}_3 M_i^2 + \hat{a}_4 A_i^2$. Parameter estimates not significantly different from 0 ($P > 0.05$) are indicated by (NS). (NS = not significant.)

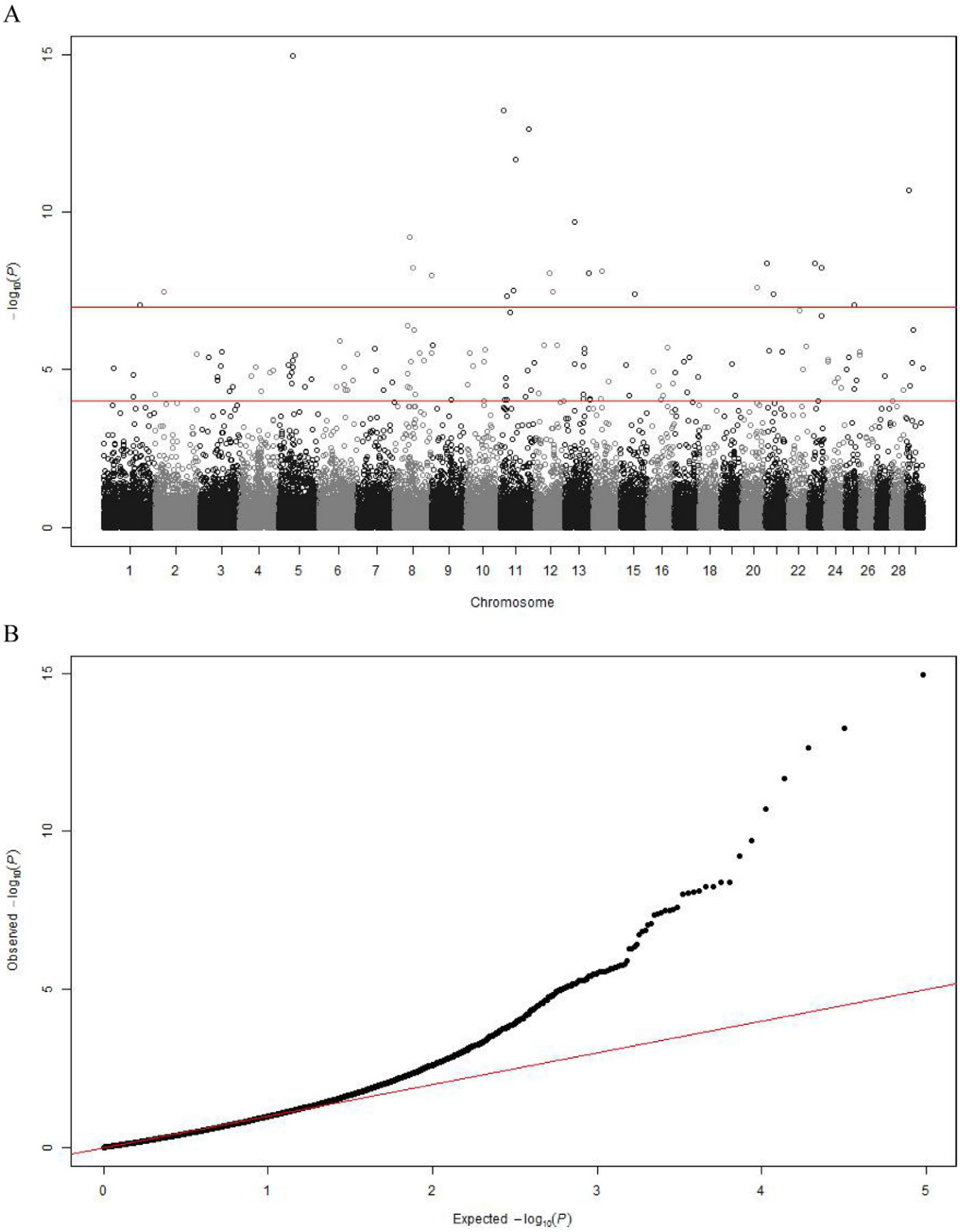

Figure 4. Results from the genomewide association studies (GWAS) of DATA3 – Bovine Pregnancy Status. (A) Manhattan plot of the distribution of *P*-values of SNP in association with pregnancy rate. The horizontal axis represents the SNP location alongside the 30 bovine chromosomes (with chromosome 30 being the X chromosome) and the vertical axis gives the $-\log_{10}(P\text{-value})$. Horizontal lines correspond to nominal thresholds of $-\log_{10}(P\text{-value})$ of 4 and 7 capturing significant SNP listed in Table 4. (B) The Q-Q plot of *P*-values from GWAS of pregnancy rate. See online version for figure in color.

Table 4. Identity of SNP associated with pregnancy rate ($P < 0.0001$) in the present study and with genomic region and candidate genes reported in the literature as influencing bovine female fertility phenotypes

SNP	BTA[1]	Mb	Gene	Reference
Hapmap50687-BTA-41950	1	30.8	*GBE1*	Cole et al. (2011)
BTB-01211220	1	92.3	*NLGN1*	Cole et al. (2011)
BTB-00077049	2	1.7	*ARHGEF4*	Hawken et al. (2012)
BTA-68622-no-rs	3	91.8	*BSND*	Cole et al. (2011)
ARS-BFGL-NGS-78389	4	49.1	*SLC26A3*	Sahana et al. (2010)
BTB-00202925	4	93.9	*TSPAN33*	Sahana et al. (2010)
ARS-BFGL-NGS-25578	5	28.1	*ACVRL1*	Hawken et al. (2012)
BTB-00226316	5	44.2	*MIR2427*	Hawken et al. (2012)
ARS-BFGL-NGS-6160	6	118.6	*PSAPL1*	Hawken et al. (2012)
ARS-BFGL-NGS-55438	8	61.1	*MELK*	Hawken et al. (2012)
BTA-24875-no-rs	8	112.5	*RAB14*	McClure et al. (2010)
ARS-BFGL-NGS-20827	8	112.8	*ALLC*	McClure et al. (2010)
ARS-BFGL-NGS-108654	11	29.9	*MSH6*	Hawken et al. (2012)
Hapmap56532-rs29016027	11	42.8	*BCL11A*	Hawken et al. (2012)
ARS-BFGL-NGS-36039	11	49.7	*KCMF1*	Holmberg and Andersson-Eklund (2006)
BTA-31432-no-rs	12	11.4	*MTRF1*	Lien et al. (2000)
ARS-BFGL-NGS-103125	13	29.7	*CDNF*	Holmberg and Andersson-Eklund (2006)
ARS-BFGL-NGS-86040	13	31.1	*PTER*	Hawken et al. (2012)
ARS-BFGL-BAC-839	13	57.3	*PHACTR3*	Sahana et al. (2010)
BTA-32994-no-rs	13	57.8	*SLMO2*	Sahana et al. (2010)
BTB-02067445	13	58.9	*PMEPA1*	Sahana et al. (2010)
ARS-BFGL-BAC-13199	13	59.9	*FAM209B*	Sahana et al. (2010)
BTB-01779799	14	25.3	*PENK*	Hawken et al. (2012)
BTB-00569940	14	49.7	*EIF3H*	Schnabel et al. (2005)
BTA-38885-no-rs	16	41.1	*FASLG*	Hawken et al. (2012)
Hapmap27883-BTA-154035	17	37.5	*FSTL5*	Hawken et al. (2012)
Hapmap32084-BTA-147824	21	6.3	*CERS3*	Hawken et al. (2012)
ARS-BFGL-NGS-69151	23	41.0	*DTNBP1*	Hawken et al. (2012)
BTA-110818-no-rs	24	29.4	*CDH2*	Hoglund et al. (2009)
Hapmap43304-BTA-59744	25	25.2	*IL21R*	McClure et al. (2010)
BTB-00905776	25	26.6	*SEZ6L2*	McClure et al. (2010)
Hapmap43264-BTA-41979	27	26.5	*WRN*	Hawken et al. (2012)
Hapmap49260-BTA-66294	29	9.96	*TMEM126A*	Hawken et al. (2012)
ARS-BFGL-NGS-29984	29	50.1	*IGF2*	Cobanoglu et al. (2005)

[1]BTA = *Bos taurus* autosomal chromosome.

show the Manhattan plot of the distribution of *P*-values of SNP in association with FE using individual and pooled DNA samples. Table 5 provides the number of significant SNP and FDR at 5 nominal *P*-values. Using pooled DNA samples, the number of significant SNP at the nominal *P*-values of 0.01, 0.001, and 0.0001 was 895 (FDR = 46.16%), 178 (FDR = 23.40%), and 30 (FDR = 13.92%), respectively. As expected, at any given significance threshold, the GWAS of individual DNA samples yielded more SNP and hence lower FDR than the GWAS of DNA pools (Table 5; Fig. 5C). A total of 1,852, 428, and 76 SNP were found to be significant in both GWAS at *P*-values of 0.05, 0.01, and 0.001, respectively (Table 5). The SNP effects estimated from pools were highly correlated with those estimated from individual DNA samples and this correlation increases with the significance threshold (Fig. 5D). At *P*-values of 0.05, 0.01, and 0.001, the correlation between SNP effects were 0.915, 0.923, and 0.935, respectively.

Table 5. Number of significant SNP and false discovery rate (FDR) at various *P*-value thresholds from the analysis of DATA5 – Chicken Feed Efficiency using individual and pooled DNA samples, and number of significant SNP in the overlap

	Individual samples		Pooled samples		Overlap
P-value	SNP	FDR, %	SNP	FDR, %	SNP
0.05	4,490	43.7	3,288	61.6	1,852
0.01	1,489	27.3	895	46.2	428
0.001	293	14.2	178	23.4	76
0.0001	76	5.5	30	13.9	13
0.00001	15	2.8	4	10.4	3

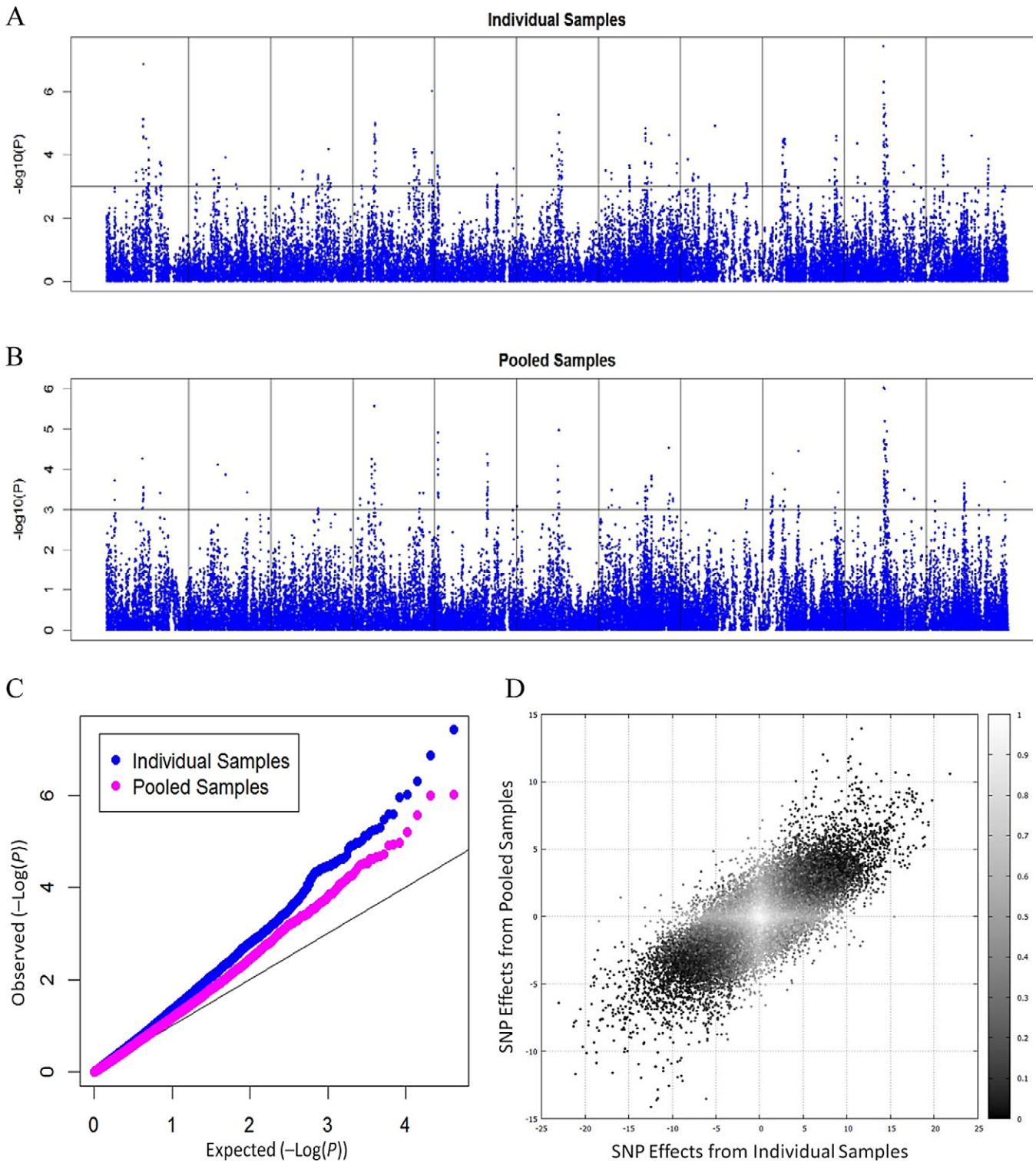

Figure 5. Results from the genomewide association studies (GWAS) of DATA5 – Chicken Feed Efficiency. (A) Manhattan plot of the distribution of *P*-values of SNP in association with feed efficiency using individual DNA samples. (B) Manhattan plot of the distribution of *P*-values of SNP in association with feed efficiency using pooled DNA samples. For these Manhattan plots, chromosomes have been randomly shuffled and vertical lines placed at equally spaced intervals to assist with the visual comparison. Horizontal lines correspond to nominal thresholds of $-\log_{10}$(*P*-value) 3.0. (C) The Q-Q plots of *P*-values from GWAS using pooled and individual DNA samples. (D) Scatter plot of the SNP effects estimated using individual DNA samples (*x* axis) and pooled DNA samples (*y* axis). Overlaid in the scatter is the *P*-value of the significance of the association of each SNP to feed efficiency, averaged across the 2 analyses, individual and pooled DNA samples. See online version for figure in color.

Concluding Remarks

With the declining cost of genotyping technologies, the search for cost-effective alternatives such as genotyping pools of DNA becomes less imperative. However, in the context of animal breeding and genetics, there are still situations where DNA pooling will remain an attractive proposition in the foreseeable future. Examples of such situations are found in aquaculture and in broiler chicken operations that result in large contemporary groups and where the phenotype of interest is expensive to measure such as residual feed intake or disease resistance. Another example is when phenotypes are collected routinely and in unpedigreed animals such as commercial beef cows raised in extensive conditions.

The present study represents an 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 2 channels (red and green) and the SNP allele frequencies and show how this relationship can be formally explored by means of mixtures of distributions and polynomial equations. When these approaches are applied to the estimation of allele frequencies, the resulting estimates allow for the development of cost-effective and reliable GWAS.

LITERATURE CITED

Anantharaman, R., and F. T. Chew. 2009. Validation of pooled genotyping on the Affymetrix 500 k and SNP6.0 genotyping platforms using the polynomial-based probe-specific correction. BMC Genet. 10:82.

Baker, J., J.-P. Liu, E. J. Robertson, and A. Efstratiadis. 1993. Role of insulin-like growth factors in embryonic and postnatal growth. Cell 75:73–82.

Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes, W. Barendse, Y. Zhang, C. M. Reich, B. A. Mason, R. J. Bunch, B. E. Harrison, A. Reverter, R. M. Herd, B. Tier, H.-U. Graser, and M. E. Goddard. 2013. Accuracy of prediction of genomic breeding values for residual feed intake, carcass and meat quality traits in *Bos taurus*, *Bos indicus* and composite beef cattle. J. Anim. Sci. 91:3088–3104.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. 2003. A comparison of normalization methods for high density oligonuclotide array data based on variance and bias. Bioinformatics 19:185–193.

Brohede, J., R. Dunne, J. D. McKay, and G. N. Hannan. 2005. PPC: An algorithm for accurate estimation of SNP allele frequencies in small equimolar pools of DNA using data from high density microarrays. Nucleic Acids Res. 33:e142.

Chai, H. S., T. M. Therneau, K. R. Bailey, and J. P. Kocher. 2010. Spatial normalization improves the quality of genotype calling for Affymetrix SNP 6.0 arrays. BMC Bioinf. 11:356.

Cobanoglu, O., P. J. Berger, and B. W. Kirkpatrick. 2005. Genome screen for twinning rate QTL in four North American Holstein families. Anim. Genet. 36:303–308.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. Van Tassell, J. Yang, S. W. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. BMC Genomics 12:408.

Craig, D. W., M. J. Huentelman, D. Hu-Lince, V. L. Zismann, M. C. Kruer, A. M. Lee, E. G. Puffenberger, J. M. Pearson, and D. A. Stephan. 2005. Identification of disease causing loci using an array-based genotyping approach on pooled DNA. BMC Genomics 6:138.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statist. Sinica 12:111–140.

Fortes, M. R. S., K. L. DeAtley, S. A. Lehnert, B. M. Burns, A. Reverter, R. J. Hawken, G. Boe-Hansen, S. S. Moore, and M. G. Thomas. 2013a. Genomic regions associated with fertility traits in male and female cattle: Advances from microsatellites to high-density chips and beyond. Anim. Repro. Sci. 141:1–19.

Fortes, M. R. S., K. Kemper, S. Sasazaki, A. Reverter, J. E. Pryce, W. Barendse, R. Bunch, R. McCulloch, B. Harrison, S. Bolormaa, Y. D. Zhang, R. J. Hawken, M. E. Goddard, and S. A. Lehnert. 2013b. Evidence for pleiotropism and recent selection in the PLAG1 region in Australian beef cattle. Anim. Genet. 44:636–647.

Fortes, M. R. S., Y. Li, E. Collis, Y. Zhang, and R. J. Hawken. 2012. The IGF1 pathway genes and their association with age of puberty in cattle. Anim. Genet. 44:91–95.

Groenen, M. A., P. Wahlberg, M. Foglio, H. H. Chen, H. J. Megens, R. P. Crooijmans, F. Besnier, M. Lathrop, W. M. Muir, G. K. Wong, I. Gut, and L. Andersson. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 19:510–519.

Hawken, R. J., Y. D. Zhang, M. R. S. Fortes, E. Collis, W. C. Barris, N. J. Corbet, P. J. Williams, G. Fordyce, R. G. Holroyd, J. R. W. Walkley, W. Barendse, D. J. Johnston, K. C. Prayaga, B. Tier, A. Reverter, and S. A. Lehnert. 2012. Genome-wide association studies of female reproduction in tropically adapted beef cattle. J. Anim. Sci. 90:1398–1410.

Henshall, J. M., R. J. Hawken, S. Dominik, and W. Barendse. 2012. Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples. Genet. Sel. Evol. 44:12.

Hoglund, J. K., B. Guldbrandtsen, G. Su, B. Thomsen, and M. S. Lund. 2009. Genome scan detects quantitative trait loci affecting female fertility traits in Danish and Swedish Holstein cattle. J. Dairy Sci. 92:2136–2143.

Holmberg, M., and L. Andersson-Eklund. 2006. Quantitative trait loci affecting fertility and calving traits in Swedish dairy cattle. J. Dairy Sci. 89:3664–3671.

Karim, L., H. Takeda, L. Lin, T. Druet, J. A. Arias, D. Baurain, N. Cambisano, S. R. Davis, F. Farnir, B. Grisart, B. L. Harris, M. D. Keehan, M. D. Littlejohn, R. J. Spelman, M. Georges, and W. Coppieters. 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nat. Genet. 43:405–413.

Korn, J. M., F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. 40:1253–1260.

Li, W., and M. Olivier. 2013. Current analysis platforms and methods for detecting copy number variation. Physiol. Genomics 45:1–16.

Lien, S., A. Karlsen, G. Klemetsdal, D. I. Vage, I. Olsaker, H. Klungland, M. Aasland, B. Heringstad, J. Ruane, and L. Gomez-Raya. 2000. A primary screen of the bovine genome for quantitative trait loci affecting twinning rate. Mamm. Genome 11:877–882.

Macgregor, S., P. M. Visscher, and G. Montgomery. 2006. Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates. Nucleic Acids Res. 34:e55.

McClure, M. C., N. S. Morsci, R. D. Schnabel, J. W. Kim, P. Yao, M. M. Rolf, S. D. McKay, S. J. Gregg, R. H. Chapple, S. L. Northcutt, and J. F. Taylor. 2010. A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. Anim. Genet. 41:597–607.

McLachlan, G. J., R. W. Bean, and D. Peel. 2002. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18:413–422.

Pearson, J. V., M. J. Huentelman, R. R. Halperin, W. D. Tembe, S. Melquist, N. Homer, M. Brun, S. Szelinger, K. D. Coon, V. L. Zismann, J. A. Webster, T. Beach, S. B. Sando, J. O. Aasly, R. Heun, F. Jessen, H. Kolsch, M. Tsolaki, M. Daniilidou, E. M. Reiman, A. Papassotiropoulos, M. L. Hutton, D. A. Stephan, and D. W. Craig. 2007. Identification of the genetic basis of complex disorders by use of pooling-based genome-wide single-nucleotide-polymorphism association studies. Am. J. Hum. Genet. 80:126–139.

Pérez-Enciso, M., and I. Misztal. 2011. Qxpak.5: Old mixed model solutions for new genomics problems. BMC Bioinf. 12:202.

Reverter, A., A. Ingham, S. A. Lehnert, S. H. Tan, Y. H. Wang, A. Ratnakumar, and B. P. Dalrymple. 2006. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. Bioinformatics 22:2396–2404.

Ritchie, M. E., R. Liu, B. S. Carvalho, Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), and R. A. Irizarry. 2011. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinf. 12:68.

Sahana, G., B. Guldbrandtsen, C. Bendixen, and M. S. Lund. 2010. Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. Anim. Genet. 41:579–588.

Schnabel, R. D., T. S. Sonstegard, J. F. Taylor, and M. S. Ashwell. 2005. Whole-genome scan to detect QTL for milk production, conformation, fertility and functional traits in two US Holstein families. Anim. Genet. 36:408–416.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA pooling: A tool for large-scale association studies. Nat. Rev. Genet. 3:862–871.

Szelinger, S., J. V. Pearson, and D. W. Craig. 2011. Microarray-based genome-wide association studies using pooled DNA. Methods Mol. Biol. 700:49–60.

Teumer, A., F. D. Ernst, A. Wiechert, K. Uhr, M. Nauck, A. Petersmann, H. Völzke, U. Völker, and G. Homuth. 2013. Comparison of genotyping using pooled DNA samples (allelotyping) and individual genotyping using the affymetrix genome-wide human SNP array 6.0. BMC Genomics 14:506.