

METHODOLOGY ARTICLE

Open Access

Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations

Jose L Gualdrón Duarte¹, Rodolfo JC Cantet¹, Ronald O Bates², Catherine W Ernst², Nancy E Raney² and Juan P Steibel^{2,3*}

Abstract

Background: Currently, association studies are analysed using statistical mixed models, with marker effects estimated by a linear transformation of genomic breeding values. The variances of marker effects are needed when performing the tests of association. However, approaches used to estimate the parameters rely on a prior variance or on a constant estimate of the additive variance. Alternatively, we propose a standardized test of association using the variance of each marker effect, which generally differ among each other. Random breeding values from a mixed model including fixed effects and a genomic covariance matrix are linearly transformed to estimate the marker effects.

Results: The standardized test was neither conservative nor liberal with respect to type I error rate (false-positives), compared to a similar test using Predictor Error Variance, a method that was too conservative. Furthermore, genomic predictions are solved efficiently by the procedure, and the *p*-values are virtually identical to those calculated from tests for one marker effect at a time. Moreover, the standardized test reduces computing time and memory requirements.

The following steps are used to locate genome segments displaying strong association. The marker with the highest $-\log(p\text{-value})$ in each chromosome is selected, and the segment is expanded one Mb upstream and one Mb downstream of the marker. A genomic matrix is calculated using the information from those markers only, which is used as the variance-covariance of the segment effects in a model that also includes fixed effects and random genomic breeding values. The likelihood ratio is then calculated to test for the effect in every chromosome against a reduced model with fixed effects and genomic breeding values. In a case study with pigs, a significant segment from chromosome 6 explained 11% of total genetic variance.

Conclusions: The standardized test of marker effects using their own variance helps in detecting specific genomic regions involved in the additive variance, and in reducing false positives. Moreover, genome scanning of candidate segments can be used in meta-analyses of genome-wide association studies, as it enables the detection of specific genome regions that affect an economically relevant trait when using multiple populations.

Keywords: Genome wide association, Marker variance, Pig genotype

* Correspondence: steibelj@msu.edu

²Department of Animal Science, Michigan State University, East Lansing, MI, USA

³Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA

Full list of author information is available at the end of the article

Background

The availability of high density genotypes of single nucleotide polymorphism (SNP) markers for plants and livestock species, in conjunction with phenotypic data for complex traits, allows the calculation of: 1) estimates of genomic breeding values (GEBVs) [1,2] for genomic evaluation [3], and 2) estimates of the effects of genomic regions associated with the genetic variability in genome wide association studies (GWAS) [2,4,5].

There is an increasing number of GWAS data sets analyzed by mixed models and multiple testing procedures [6], after fitting all individual effects of genomic regions into the model [4]. The model may be difficult to fit when both, the number of individuals and SNP effects, are large. We propose to use a linear transformation of genomic breeding values to estimate the marker effects from a simpler equivalent mixed model, and then testing those effects using a standardized test statistic that employs the variance (rather than prediction error variance) of the same effects.

The method of *genomic selection* proposed by Meuwissen et al. [7] to estimate GEBVs starts by fitting the SNP effects to a given data set. Next is to estimate GEBV of any individual using its genotype (SNP), by adding across the entire genome those solutions corresponding to the individual's SNP. The mixed model employed conveys vectors of fixed effects, and random effects of markers or SNPs (\mathbf{g}) assumed to be normally distributed with null mean and a covariance matrix proportional to the identity matrix times the variance of SNP effects ($\mathbf{I} \sigma_g^2$). Errors are assumed to be Gaussian, independent and identically distributed with null mean and covariance matrix $\mathbf{I} \sigma_e^2$. An equivalent mixed model discussed by Garrick [8] and Strandén [9] is fitted after the linear transformation $\mathbf{a} = \mathbf{Z} \mathbf{g}$ where \mathbf{a} is a random vector of breeding values, and \mathbf{Z} an incidence matrix that relates elements in \mathbf{a} to those in \mathbf{g} . Each column of \mathbf{Z} is associated with a given SNP and the elements are standardized by functions of SNP allele frequencies and by the total number of SNP. It is worth noting that the same \mathbf{Z} is used in our implementation of the model of Meuwissen et al. [7] to relate the vector of marker effects in \mathbf{g} to the data phenotypes. Moreover, GEBVs in the equivalent model have variance-covariance matrix $\mathbf{G} \sigma_A^2 = \mathbf{Z} \mathbf{Z}' \sigma_g^2$. The procedure requires that the variances are equal, i.e. $\sigma_A^2 = \sigma_g^2$. Once the equivalent model is fit, SNP effects are calculated by the transformation $\mathbf{g} = \mathbf{Z}' \mathbf{G}^{-1} \mathbf{a}$, and individual SNP effects in \mathbf{g} are divided by the square root of its variance ($\text{Var}(\mathbf{g}_j)$) to get the so called SNP_{ej} test statistics. We also provide a formula to calculate $\text{Var}(\mathbf{g}_j)$ without having to fit the model with SNP effects. The next step is to select genome segments that may be highly associated with the genetic variability of the trait for each chromosome. In doing so, we look for the SNP having the highest value

of minus the logarithm of the p -value throughout the chromosome. Once the SNP is located, a segment of one Mb to the left and one to the right is defined, and a relationship matrix is calculated using only the information from those markers. The relationship matrix is used as the proportional variance-covariance of the segment effects in a model that also includes fixed effects and random GEBVs. In a final step, the likelihood ratio is calculated to test the significance of the largest effect segment of each chromosome by comparing against a reduced model with fixed effects and GEBVs. The critical value (size of the test) is adjusted by the Bonferroni correction. The algorithm not only delivers genome wide associations and genomic predictions efficiently, but it also minimizes computing time and memory requirements. Moreover, the specific variance of the SNP effects is used in calculating the test, thus taking into account the amount of information of any given marker. Instead, other testing approaches rely on a prior variance or a constant estimate of the additive variance.

Methods

Dataset

The experimental population was raised at the Michigan State University Swine Teaching and Research Farm, East Lansing, MI [10]. Parents from the initial generation (F_0) were four Duroc boars mated to 15 Pietrain sows by artificial insemination. From all resulting F_1 animals, 50 females and 6 males (progeny of 3 F_0 sires) were selected as parents for the F_2 generation, by avoiding full or half sib matings. A total of 1,259 F_2 piglets were born alive from 142 litters out of 11 farrowing groups. Phenotypic data for growth, carcass merit and meat quality traits were collected for approximately 950 F_2 pigs (for more details refer to Edwards et al. [10,11]). Data used for the study were measures of the growth trait *13 week tenth rib backfat (mm)* (bf10_13wk). The trait was chosen as it displays a sizable heritability (0.42) and a normal distribution. Animal protocols were approved by the Michigan State University All University Committee on Animal Use and Care (AUF# 09/03-114-00).

Genotyping and data editing

DNA was isolated from white blood cells using standard procedures as previously described for this population [10]. Quantity and quality of DNA samples were determined using a Qubit fluorometer (Invitrogen by Life Technologies, Carlsbad, CA, USA). The experimental population was genotyped with two marker SNP panels. 1) 411 animals were genotyped (4 F_0 Duroc boars, 15 F_0 Pietrain sows, 6 F_1 males, 50 F_1 females and 336 F_2 pigs) with a commercial panel, the Illumina PorcineSNP60 beadchip (60 K) [12] and 2) 612 F_2 animals were genotyped with a second panel composed of a 9 K tagSNP

set referred to as the GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE) [13]. A set of 5,350 SNP out of $M = 62,163$, were eliminated from all analyses as their physical positions were unknown. Mendelian inconsistencies ($\leq 0.01\%$) were taken as missing genotypes, and 21 animals (1 F_1 and 20 F_2) with more than 10% of SNP missing were not used for any analysis. By similar considerations, 2,978 SNP were removed from the analyses as they had more than 10% missing data. Additionally, 9,877 SNP were excluded as their minor allele frequency (MAF) was below 0.01. This editing procedure followed that of Badke et al. [14] and Gualdrón et al. [15], and the program PLINKv1.07 [16] was used for the task. F_2 animals genotyped with the 9 K panel were imputed to 60 K following procedures discussed by Gualdrón et al. [15], by means of the software AlphaImpute [17], resulting in imputation accuracy of around 0.99 [15]. Genotypes imputed in the F_2 had a second editing procedure by $MAF < 0.01$, which excluded 759 virtually monomorphic SNP. The editing policies and genotype imputation resulted in a data set with records from 1002 pigs (F_0 , F_1 and F_2) having 44,055 SNP per animal.

Estimation of genomic relationship matrix

The genomic relationship matrix was estimated from observed and imputed high density (~44 K) SNP genotypes. Genotypes were expressed as allelic dosage [13,15], such that genotypes were entered into a marker matrix M of dimension ($n \times m$), where n is the number of animals and m the number of SNP, having elements in the interval $[0, 2]$, i.e. the count of the allele used as reference. In the sequel, we will use the sub index i to refer to the individual. Matrix M was standardized to matrix Z that has generic elements equal to

$$Z_{ij} = \frac{M_{ij} - 2p_j}{\sqrt{m(2p_j(1-p_j))}}$$

Elements of Z are then calculated by subtracting twice the frequency of the reference allele at the j th marker (p_j), to the corresponding element of M [18], and then dividing the resulting difference by the square root of the expected variance $2p_j(1-p_j)$ of each element in the column multiplied by the number of columns (m) in M . The allele frequency p_j was calculated from the F_0 generation (19 animals). The genomic relationship matrix was finally calculated as:

$$G = ZZ' \tag{1}$$

Prediction model

Using the genomic relationship matrix from equation (1), the centered animal model for genomic evaluation can be written as:

$$y = X\beta + a + e \tag{2a}$$

where y is the phenotypic vector containing the data on *13-week tenth rib backfat (mm)*, X is the incidence matrix that relates records to the fixed effects of sex in β , vector a contains the random breeding values such that $a \sim N(0, G\sigma_A^2)$, e is the random error vector such that $e \sim N(0, I\sigma_e^2)$, and I is the identity matrix. Variance components were estimated with REML using the regress version 1.3-10 R package [19].

Following Strandén et al. [9] an equivalent model to (2a) is

$$y = X\beta + Zg + e \tag{2b}$$

Every element in (2b) is defined as before except for the vector g of SNP effects. To show that (2a) and (2b) are equivalent models, we employ the fact that $a = Zg$. Then, the variances of a and g are related in the following manner:

$$G\sigma_A^2 = \text{Var}(a) = \text{Var}(Zg) = Z \text{Var}(g) Z' = ZZ'\sigma_g^2$$

Necessary conditions for models (2a) and (2b) to be equivalent (Henderson, 1984) are that $G = ZZ'$ and $\sigma_A^2 = \sigma_g^2$.

Variance of SNP effects

In this section, we describe the algorithm to calculate the variance of the estimated SNP effects g (i.e. $\text{Var}(\hat{g})$). The SNP effects were obtained from a linear transformation of breeding values in \hat{a} [4,9,20,21], as follows:

$$\begin{aligned} \text{BLUP}(\hat{g}) &= \hat{g} = \text{cov}(g, a') [\text{Var}(a)]^{-1} \hat{a} \\ &= \text{cov}(g, g') Z' G^{-1} (\sigma_A^2)^{-1} \hat{a} \\ &= \left(\frac{\sigma_g^2}{\sigma_A^2} \right) Z' G^{-1} \hat{a} = Z' G^{-1} \hat{a} \end{aligned} \tag{3}$$

The last step results from the fact that model equivalence involves $\sigma_A^2 = \sigma_g^2$. Now, from equation (3) $\text{Var}(\hat{g})$ is obtained as follows:

$$\text{Var}(\hat{g}) = \text{Var}(Z' G^{-1} \hat{a}) = Z' G^{-1} \text{Var}(\hat{a}) G^{-1} Z \tag{4}$$

Now, we know that the predictor error variance (PEV) of \hat{a} from model (2a) is equal to:

$$\text{PEV}(\hat{a}) = \text{Var}(a - \hat{a}) = C^{aa} = \text{Var}(a) - \text{Var}(\hat{a})$$

So that

$$\text{Var}(\hat{\mathbf{a}}) = \text{Var}(\mathbf{a}) - \mathbf{C}^{aa} = \mathbf{G} \sigma_A^2 - \mathbf{C}^{aa}$$

Matrix \mathbf{C}^{aa} results from inverting the coefficient matrix of the mixed model equations [22] such that:

$$\mathbf{C}^{aa} = \sigma_e^2 \left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{G}^{-1}\lambda \right)^{-1}, \lambda = \frac{\sigma_e^2}{\sigma_A^2}$$

Then, on replacing with the latter expression into $\text{Var}(\hat{\mathbf{a}})$ (displayed in (4)), we have:

$$\begin{aligned} \text{Var}(\hat{\mathbf{g}}) &= \mathbf{Z}'\mathbf{G}^{-1}(\mathbf{G} \sigma_A^2 - \mathbf{C}^{aa}) \mathbf{G}^{-1} \mathbf{Z} \\ &= \mathbf{Z}' \mathbf{G}^{-1} \mathbf{Z} \sigma_A^2 - \mathbf{Z}' \mathbf{G}^{-1} \mathbf{C}^{aa} \mathbf{G}^{-1} \mathbf{Z} \end{aligned} \quad (5)$$

Expression (5) results in a large matrix of dimension $(m \times m)$ with m the number of SNP. However, we only need its diagonal elements. Also notice that the first term in (5), $\mathbf{Z}' \mathbf{G}^{-1} \mathbf{Z}$, can be computed and stored to be reused for the different traits, whereas \mathbf{C}^{aa} has to be computed for each trait.

Standardization of SNP effects (SNP_{ej})

The estimated SNP effects in (3) were standardized by dividing with their corresponding $\text{Var}(\hat{\mathbf{g}}_j)$ obtained from (5) as follows:

$$\text{SNP}_{ej} = \frac{\hat{\mathbf{g}}_j}{\sqrt{\text{Var}(\hat{\mathbf{g}}_j)}} \quad (6)$$

P-values and genome screening

The p -values were assessed as 1 minus the cumulative probability density of the absolute value of SNP_{ej} , a number that was then multiplied by 2 so as to obtain:

$$p\text{-value}_j = 2(1 - \Phi(|\text{SNP}_{ej}|))$$

where $\Phi(x)$ is the cumulative density function of the normal distribution for the random variable x . When analyzing the trait *13 week tenth rib backfat (mm)*, the p -values for each SNP were plotted across the genome as $-\text{Log}_{10}(p\text{-value})$ using the physical position of the SNP in Mega-bases (Mb).

Standardization of SNP effects using the PEV of the marker

A second standardization of the j^{th} SNP effect (3) was performed using the PEV ($\hat{\mathbf{g}}_j$) as follows:

$$\text{SNP}_{epj} = \frac{\hat{\mathbf{g}}_j}{\sqrt{\text{Var}(\hat{\mathbf{g}}_j) - \text{Var}(\hat{\mathbf{g}}_j)}} \quad (7)$$

As discussed above, $\sigma_g^2 = \sigma_A^2$. The p -values and genome screening for SNP_{epj} were assessed and plotted in the same fashion as for SNP_{ej} .

Simulation

A plasmode simulation was performed to compare how the standardized values SNP_{ej} and SNP_{epj} affected the nominal size of the test for the effect to be equal to zero. Data on 928 animals with 44,055 SNP each were used for the study, and the 1018 SNP on chromosome 18 were reshuffled. Two scenarios were considered: 1) Dependency: rows of the genotype matrix were permuted for columns corresponding to SNP on chromosome 18, thus keeping Linkage Disequilibrium (LD) within chromosomes but breaking the relationship between genotypes and phenotypes for the 1018 SNP on the chromosome. 2) Independence: the genotype of any animal was permuted independently by marker (resulting in linkage equilibrium, or LE between markers) for those SNP on chromosome 18, and the relationship with the phenotype was broken too. For both scenarios model (2a) was fitted to the data, and two tests were calculated for each scenario: $\text{test1} = \text{SNP}_{ej}$ and $\text{test2} = \text{SNP}_{epj}$. Permutations were repeated 200 times per scenario, and in each permutation the \mathbf{G} matrix was calculated while fitting model (2a). As a result, the heritability of the trait was similar to the original heritability due to relationships in the other 17 chromosomes being kept intact, and p -values for those SNP (that are now non-associated) on chromosome 18 were obtained for the different tests. Under the null hypothesis and assuming independence (i.e., SNP are unlinked to the polymorphism controlling the trait), an approach that controls for type I error appropriately [23], the 1018 test p -values follow a uniform distribution. Consequently, to estimate the empirical quantiles of the distribution for the null hypothesis, we used a uniform density $U \sim (0, 1)$ to generate 200 replicated sets for the 1018 p -values.

SNP effects and tests obtained by a single marker model

The SNP effects were tested on a one by one basis. The model approach used for testing purposes is better known as "efficient mixed-model association" (EMMA) [24]. The model included fixed effects of sex and one-marker-at-a-time; random variable was the animal effect with variance-covariance equal to the genomic relationship matrix using all markers, which was calculated as described before. The R package rrBLUP [25] was used for fitting the different models and for calculating the tests and p -values.

Proportion of variance explained by segments with large effect

After the genome screen using model 2a, the SNP with the smallest p -values were selected to form SNP segments. These segments were defined by taking all SNP within one Mb upstream and one Mb downstream of the SNP with smallest p -value on each chromosome. The size of the segment was chosen using a criterion similar to the one employed by Hayes et al. [4]. The point of change in

the rate of decay in linkage disequilibrium in this population was about $r^2 = 0.2$ at 1 Mb (data not shown), which essentially would imply a minimal contribution to the additive variance from markers located beyond such distance. Moreover, segment sizes about two Mb have been reported to be significant in association studies [20,26-28]. The proportion of variance associated with each segment was estimated by building a genomic relationship matrix G_1 (as described in (1)) using all SNPs that belonged to the segment, whereas genomic relationship matrix G_2 was built using all remaining SNPs. The model fitted can be represented as:

$$y = X\beta + a_1 + a_2 + e \quad (8)$$

where a_1 is the vector of additive random effects associated with those SNP located in the segment, such that $a_1 \sim N(0, G_1 \sigma_{A_1}^2)$, and a_2 is the vector of additive random effects associated with all SNPs except those involved with a_1 , such that $a_2 \sim N(0, G_2 \sigma_{A_2}^2)$. Model (8) assesses the proportion of variance explained by the segment of interest (local variance) from the genome variance explained by all SNPs (global variance). The variances estimated in (8) were compared with those estimates from model (2a). Hayes et al. [4] used a similar model to assess the segment variance. Applying either model (8), or the approach of Hayes et al. [4] gave similar estimated variance components. In practice, the advantage of fitting model (8) is that G_2 is computed by subtracting from G the columns of Z related to the segment being tested. Let Z_s be a matrix having as columns those related to the segment being tested, then $G_2 = G - Z_s Z_s'$. On the contrary, in the model of Hayes et al. [4] G is different from segment to segment. Additionally, the calculation of G_1 and $Z_s Z_s'$ is fast and involves only those SNPs located in the segment.

To adjust the level of significance for multiple comparisons, a Bonferroni Correction (BC) was performed. In this context, if the pig genome is ~2800 Mb long and the average size of the segment is 2 Mb, there are 1400 segments along the genome with corresponding multiple tests. Thus, for $\alpha = 0.05$, the BC was equal to $0.05/1400 = 3.571429e^{-05}$ (adjusted α or critical value). Hence, in order to evaluate the significance of the segments, a second p -value for the Likelihood Ratio Test (p -value_{LRT}) was calculated to compare against BC. This p -value_{LRT} was assessed as 1 minus the distribution function of a chi-square (χ^2) random variable with 0.5 degrees of freedom [29,30] as follows:

$$p\text{-value}_{LRT} = 1 - \Omega(\text{LRT})$$

where $\Omega(x)$ is the distribution function of a random variable having the χ^2 as density, and LRT is the

Likelihood Ratio Test obtained by contrasting appropriate models.

Results

Genome screening

The p -values of the 44055 SNP were obtained as described in the Methods section. First, the p -values for SNP_{epj} , i.e. using $\text{Var}(\hat{g}_j)$, were plotted along the genome (Manhattan plot in Figure 1) to identify genomic positions that are associated with variation in *13-week tenth rib backfat (mm)*. Large peaks ($-\text{Log}_{10}(p\text{-value})$) above 5 can be seen at chromosomes 6 and 3, suggesting noticeable genetic variation for the trait. On the other hand, p -values for SNP_{epj} (i.e. standardized with prediction error variance) were very large, with a maximum $-\text{Log}_{10}(p\text{-value}^2)$ of 0.20. In essence, the pattern observed in Figure 2 is the result of dividing the non-standardized SNP effects by a constant. Specifically, the normalizing value was $[\text{Var}(g_j) - \text{Var}(\hat{g}_j)]$, with $\text{Var}(g_j) = 2.6768$. The use of the square root of the difference between those two values resulted in a practically constant denominator for the test-statistic that was equal to 2.66. Also, a look at Figure 2 suggests signals at chromosomes 1, 12, 14, and 18, a fact that is not observed in Figure 1. However, this might be an artefact of the constant denominator that tends to overestimate the true variability for some SNP, thus resulting in corresponding false positives across the genome.

In order to study the type I error rate of the two proposed tests we performed a plasmode simulation [31]. A plasmode is a dataset created from real data where some of the truth is known. In brief, our plasmode is a simulation that uses reshuffling in a portion of the data as explained in the methods section. We performed a simulation assuming independent SNP, and another one keeping the dependency between SNP (LD structure) intact. Simulation results were plotted into a Quantil-quantil plot graph (Figure 3) using the number $-\text{Log}(p\text{-value})$ for each case of standardization. First, the p -values for test1 (SNP_{epj}) obtained in the scenario under independent SNPs (scenario 2, LE) displayed an identical distribution of p -values when obtained by the reference distribution $U \sim (0, 1)$. In contrast, under dependency (scenario 1, LD) less extreme p -values were observed, a fact that was not reflected in a uniform distribution. This is a well known fact in human genetic epidemiology [32], where the implementation of the Bonferroni correction of p -values from associated SNP under the assumption of independence results in tests that are too conservative. On the other hand, for test 2 (SNP_{epj}) even p -values obtained for independent SNP (scenario 2, LE) displayed a distribution that was too conservative. Furthermore, the results from the dependent scenario (LD)

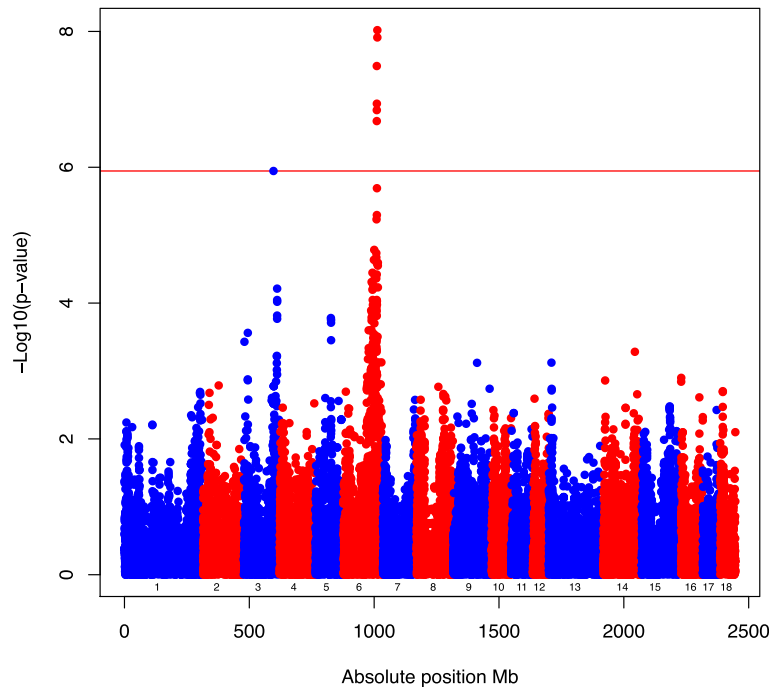


Figure 1 Manhattan Plot for trait 13-week tenth rib backfat (mm) by standardization SNP_{ej} . Genome screening for 44055 SNP using standardization $\text{Var}(\hat{g}_j)$. $-\log_{10}(p\text{-value})$ (y axis) versus the absolute SNP position in Mb (x axis). The red line represents a genome-wide significance threshold ($p < 1.1349 \times 10^{-6}$). Numbers from 1 to 18 represent the chromosome ID.

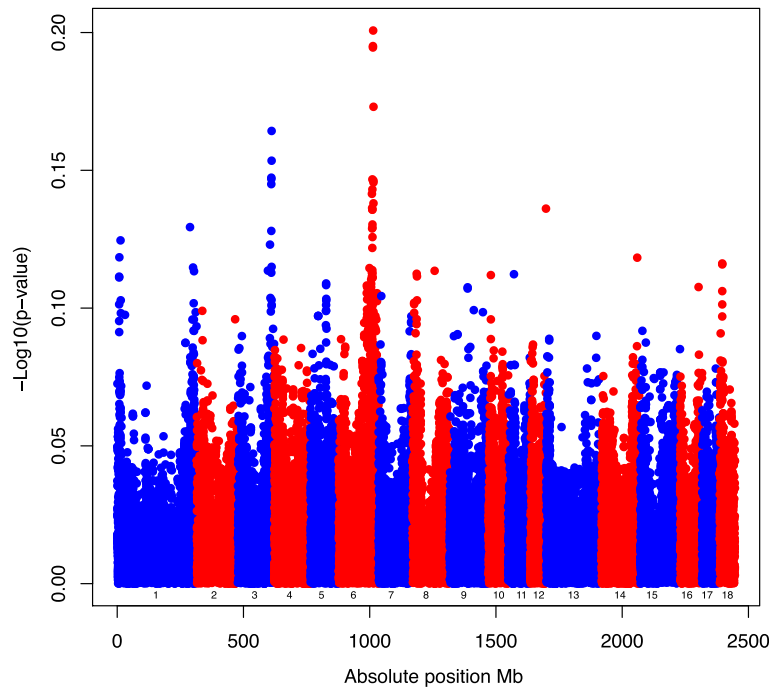


Figure 2 Manhattan Plot for trait 13-week tenth rib backfat (mm) by standardization SNP_{ej} . Genome screening for 44055 SNP using standardization $\text{PEV} = \text{Var}(\hat{g}) - \text{Var}(\hat{g}_j)$. $-\log_{10}(p\text{-value})$ (y axis) versus the absolute SNP position in Mb (x axis). Numbers from 1 to 18 represent the chromosome ID.

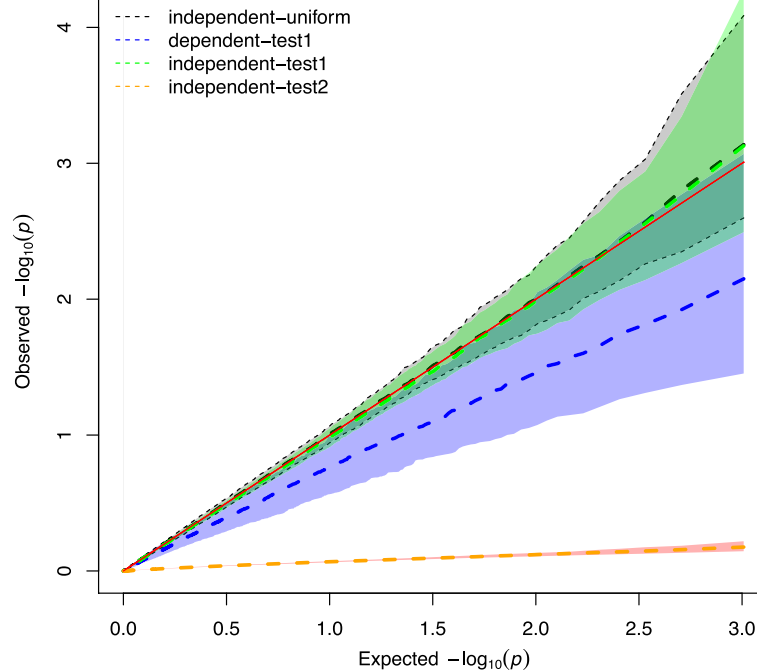


Figure 3 Quantil-quantil plot of the observed and expected $-\log_{10}(p)$ obtained by simulation. Reference distribution was an independent and uniform distribution $U \sim (0, 1)$ for 1018 p -values simulated (black dotted line). Test1(scenario1) = under dependent (LD) and standardization by $\text{Var}(\hat{g})$ (blue dotted line). Test1(scenario2) = under independent (LE) and standardization by $\text{Var}(\hat{g})$ (green dotted line). Test2 (scenario2) = under independent (LE) and standardization by PEV (orange dotted line). Each scenario has 1018 p -values permuted 200 times. Bands represent confidence intervals of 95% (blue band = test1(scenario1), green band = test1(scenario2), pink band = test2(scenario2)).

were even more conservative than those from the independent scenario (results not displayed in the Q-Q plot), thus indicating that the use of the square root of $\text{Var}(\hat{g}_j)$ as the denominator of the test-statistic results in a more powerful and not too liberal choice when compared to the use of the square root of $\text{PEV} = \text{Var}(\mathbf{g}_j) - \text{Var}(\hat{g}_j)$.

SNP effects and tests obtained by the marker model

The analyses of one SNP tested at a time using the EMMA procedure [24] resulted in p -values that were almost identical (Additional file 1) to those of SNP_{ej} (Additional file 2). The time taken to compute 44055 SNP tests one at a time was 84 minutes. In comparison, the algorithm used to fit model (2a) and to perform the tests of standardized effects took a total time of 29 minutes (CPU and memory: Quad-core 2.7GHz AMD Opteron 8384, 256 GB). This time includes the computation of the G matrix, the fit of the animal model, the back transformation to calculate the SNP effects, and the calculation of the standard errors that are needed to compute the test-statistics.

Tests of segment effects

We also compared the results from our proposed method to those obtained with a segment-based likelihood ratio test that has been used by animal breeders [4]. Due to computational demand, we only performed the LRT to

test for segment effects. Thus, the SNP with the smallest p -values (or highest $-\text{Log}_{10}(p)$ -values) on each chromosome were chosen, whereas no segments were tested using LRT for regions with SNP_{epj} resulting in exceedingly low p -values. The three segments from chromosomes with the smallest p -values are displayed in Table 1, and the remaining segments from the 15 other chromosomes are shown in the additional files (Additional file 3). All segments measured 2 Mb (1 Mb on each side of the SNP with the smallest p -value). The estimates of the variance components and the LogLikelihood obtained from model equation (8) were compared with those from model equation (2a). These results are displayed in Table 2.

Results from the LRT indicated that the segment on chromosome 6 was significant: p -value_{LRT-6} = $1.133459e^{-09}$, a number smaller than the critical 0.05 Bonferroni threshold for 1400 segments ($P_{\text{critical}} = 0.05/1400 = 3.571429e^{-05}$). On the contrary, the segments located on all other chromosomes were not significant. The proportion of variance explained by the segment from chromosome 6 ($-\text{Log}(p\text{-value}) = 8.02$) was 11% of the total variance, a fact that was reflected in a similar reduction of the estimated additive variance (σ_A^2) in model (8): $1.952 + 0.698 = 2.650$. This latter value is close to 2.678, i.e. the estimated value of σ_A^2 from model (2a) (see Table 2). For all other chromosomal segments,

Table 1 SNP selected by smallest p-value per chromosome

SNP-name	Chromosome	Position Mb	$-\log_{10}(p\text{-value})$	$ \hat{g} $
ALGA0104402	6	136.08	8.02	0.77
H3GA0010564	3	119.34	5.95	0.48
ALGA0032063	5	61.37	3.78	0.42
ALGA0081287	14	125.98	3.28	0.33
DRGA0011971	13	10.47	3.12	0.36
MARCO022304	9	94.99	3.12	0.42
ALGA0106422	16	111.82	2.90	0.28
ASGA0010464	2	62.15	2.79	0.30
ALGA0111088	8	88.01	2.77	0.48
ASGA0078865	18	10.72	2.70	0.49
ALGA0010607	1	302.88	2.69	0.43
MARCO082230	12	6.14	2.59	0.31
ALGA0045724	7	129.47	2.57	0.41
ASGA0092331	4	138.29	2.52	0.27
ASGA0070227	15	111.82	2.48	0.29
ASGA0077393	17	55.27	2.43	0.32
ASGA0045992	10	7.00	2.42	0.30
ALGA0060793	11	10.50	2.38	0.34

SNP name = SNP marker name, Position Mb = Marker physical position in Mega-Bases, $-\log_{10}(p\text{-value})$ = -Logarithm in base 10 of the smallest p-value, $|\hat{g}|$ = absolute value of the SNP effect estimated for the trait 13 week tenth rib backfat (mm).

the estimated value of σ_A^2 did not decrease to a significant amount.

Discussion

The main goal of this research was to develop a novel procedure to perform a rapid genome scan, or GWAS analysis, from a genomic evaluation. Moreover, the *sufficient statistics* of our methodology are: the Best Linear Unbiased Prediction (BLUP) of the breeding values from an animal model, G as the covariance matrix (or H for a single step evaluation [33]), Z as the standardized marker effects matrix, variance components, and C^{aa} . This setting makes the implementation extremely feasible after the genomic evaluation has been performed as discussed by Legarra et al. [33].

Variance of the SNP effect

First, the SNP effects \hat{g}_j were calculated by a linear transformation of \hat{a} using expression (3). Then, we calculated $\text{Var}(\hat{g}_j)$ using an expression derived from mixed model theory (see (4–5)). Next, we divided \hat{g}_j by the square root of $\text{Var}(\hat{g}_j)$ to standardize the effect, and referred the statistics as SNP_{ej} . The p-values for the tests of specific genome regions were calculated with a level of significance $-\log_{10}(p\text{-value}) = 5$. Additionally, Prediction Error

Table 2 Variance components and LogLikelihood for models with or without the segment

Seg-chromosome	6	3	5
SNP – $\log_{10}(p\text{-value})$	8.02	5.94	3.78
Lk_m1	-1227.938	-1227.938	-1227.938
Lk_m2	-1210.800	-1223.178	-1224.540
LRT	34.28	9.52	6.80
p-value _{LRT}	1.1×10^{-9}	6.5×10^{-4}	3.1×10^{-3}
VarE_m1	3.70	3.70	3.70
VarA_m1	2.68	2.68	2.68
VarE_m2	3.73	3.67	3.69
VarA_m2	1.95	2.42	2.55
segmVA	0.70	0.63	0.15
%segmVA	0.11	0.09	0.02

Seg-chromosome = Number of chromosome where segment is located, **m1** = model(2a) without the segment: $y = X\beta + a + e$, **m2** = model (8) with the segment $y = X\beta + a_1 + a_2 + e$, **SNP – $\log_{10}(p\text{-value})$** = -Logarithm in base 10 of the SNP p-value selected to create a segment, **Lk_m1** = -LogLikelihood for m1, **Lk_m2** = -LogLikelihood for m2, **LRT** = Likelihood Ratio Test for m1 and m2, **p-value_{LRT}** = p-value for LRT, **VarE_m1** = Error variance (σ_e^2) of m1, **VarA_m1** = Additive variance (σ_A^2) of m1, **VarE_m2** = Error variance (σ_e^2) of m2, **VarA_m2** = Additive variance (σ_A^2) of m2, **segmVA** = Additive variance segment ($\sigma_{A_s}^2$) of m2, **%segmVA** = Proportion in% of the total variance explained by the segment.

Variance ($PEV = \text{Var}(\hat{g}_j) - \text{Var}(\hat{g}_j)$) was employed for a second standardization, and it was called the SNP_{epj} statistic. After the analyses, we obtained higher p-values (maximum $-\log_{10}(p\text{-value}) = 0.20$) and detected stronger signals (higher peaks in the Manhattan plot) for SNP_{epj} than with SNP_{ej} . Furthermore, a simulation was carried out with the same structure of SNPs markers and animal data as in the current study, in order to compare the performance of empirical p-values of both standardized tests. The SNPs markers of chromosome 18 were reshuffled, and two scenarios were simulated: 1) Dependent genotypes (LD), and 2) Independent genotypes (LE). Neither scenario displayed a relationship with the phenotype, whereas both standardized tests were calculated at each scenario. The reference distribution for the p-values considered was the uniform. In the independent scenario (LE), standardization with $\text{Var}(\hat{g}_j)$ gave an empirical distribution of p-values that resembled the uniform density, but in the dependent scenario (LD) the SNP_{ej} performed conservatively. Instead, the standardization with $[\text{Var}(\hat{g}_j) - \text{Var}(\hat{g}_j)]$ produced conservative results in the independent scenario (LE), and very conservative tests in the dependent scenario (LD). In this context, standardizing SNP effects with $\text{Var}(\hat{g}_j)$ resulted in p-values that were closer to the simulated ones. Moreover, the performance of SNP_{ej} under LD was not too conservative, a scenario that could be extrapolated to the

genotypes in the current study. In addition, the p -values calculated using the EMMA procedure [24] were similar to those obtained with SNP_{ej} . These results suggest that SNP_{ej} behaves reasonably to control type I error rate or false positives. Also, the computing time for fitting model (2a) and then calculating (6) using expressions (3)-(5) was 2.5 to 3 times less than the computing time for the EMMA model.

In order to identify SNP with important phenotypic associations [34], the calculation of SNP effects \hat{g}_j from genomic breeding values \hat{a} [8,9,34] has been used in several studies [5,20,21]. In this context, the variance of SNP effects has been estimated using different approaches. Wang et al. [21] employed the classical definition of the variance of additive effects from quantitative genetics [35], so that the variance for each j th marker was obtained as follows: $\sigma_{A,j}^2 = \hat{g}_j^2 2p_j(1-p_j)$. Whereas, McClure et al. [20] proposed equating the variance of SNP effects to $\left(2 \sum p_j q_j\right)^{-1} \sigma_A^2$, and then normalizing the SNP effects with the square root of this estimated and constant variance. This test performed similar to SNP_{epj} (7), when the estimated SNP effects (\hat{g}_j) was divided by a constant denominator, a value almost equal to the prior variance 2.67, and resulted in a very conservative test.

In contrast, the advantage of the standardized test (SNP_{ej}) presented here was that each SNP effect was scaled by its own (and different) standard deviation rather than the use of a prior variance [20] or by the square of each specific SNP effect \hat{g}_j^2 [21] as variance. Furthermore, the computation of SNP_{ej} involves the same variance for the same SNPs markers and animals, i.e. $\sigma_g^2 = \sigma_A^2$, and the use of the standardized incidence matrix Z , a function of $2p_j(1-p_j)$, takes into account this latter quantity into SNP_{ej} . Additionally, the matrix Z uses the allele frequencies from the F_0 generation calculated with unrelated individuals, and a proper expected variance by marker (see Methods section). In addition, the test statistics SNP_{ej} that standardizes SNP effects produces a p -value, a result that is appealing to many researchers that are more familiar with the method of testing one SNP at the time rather than with the proportion of additive variance that is explained by a genomic region. A further advantage of the method is that detection of many false positives are avoided, and genome positions with sizeable effects are highlighted.

Candidate segment approach

Later in the research, genome segments that expressed higher signals were located. To this purpose, SNPs with the smallest p -values from SNP_{ej} (6) were selected, and

for each of these SNP a segment of 2 Mb long (1 Mb at each side) was created. The next step was to estimate the variance components and the Log-Likelihood from the centered animal models (2a) and (8). The latter model includes the random vector of SNP segments a_1 . Lastly, we compare the performance of both models. Hayes et al. [4] used a similar model to (8), although the random SNP effect was taken from the breeding value and fitted as a separate segment effect. We observed similar results from the use of either approach. The advantage of fitting model (8) is that matrix G is the same for all segments, so that it was calculated only once, and stored in memory for the calculations, whereas in the model of Hayes et al. [13] a different G has to be calculated for each segment. This implies an extended computing time and higher requirements of CPU memory to obtain similar results to those from model (8).

To evaluate the significance of the segments, the effects of each chromosome segment were tested by the Likelihood Ratio Test. The size of the test was adjusted by the Bonferroni correction. As a result, the segment located on chromosome 6 (physical position 135 Mb-137 Mb) was significant, and explained 11% of the trait total variance. Previous studies by Edwards et al. [10] and Choi et al. [36], using microsatellites and a small number of SNP, found significant regions (physical positions between 135 and 139 Mb) on chromosome 6 for *13 week tenth rib backfat* in the current population under study.

Additionally, forty eight markers between the physical position between 128 Mb and 139 Mb on chromosome 6 (<http://www.animalgenome.org/QTLdb/pig.html>), have been reported to be associated with the trait. Furthermore, recent studies showed the importance of chromosome 6 [37,38] in the expression of the trait. Therefore, our results confirm the presence of genetic variability in the trait from chromosome 6.

Conclusions

Fast genome screening of SNP effects linearly transformed from genomic breeding values is advantageous, as a by-product of genomic evaluations for different species of farm animals. Moreover, the standardized tests of SNP effects using their own variance ($\text{Var}(\hat{g}_j)$) developed in this study helps in detecting specific genomic regions involved in the additive variation of the trait and reducing false positive locations using less computing time. Additionally, genome segments of about 2 Mb formed by surrounding the SNP with the smallest p -values on each chromosome, and tested with a standardized test involving $\text{Var}(\hat{g}_j)$ and with the Bonferroni correction, could detect genome regions responsible for sizeable fractions of

the trait genetic variance. This methodology involving genome scan and candidate segment approach is a useful method for meta-analyses of genome-wide association studies, as it enables the detection of specific genome regions that affect an economically relevant trait when using multiple populations. Code and data to obtain and reproduce the results presented is publicly available at https://www.msu.edu/~steibelj/JP_files/GBLUP.html.

Additional files

Additional file 1: Highest – Log₁₀(p-values) on each chromosome for trait 13-week tenth rib backfat (mm) by standardization SNP_{ej} and EMMA. The blue and red circle represents highest – Log₁₀(p-values) on each chromosome by the standardization SNP_{ej} and efficient mixed-model association (EMMA) using rrBLUP, respectively.

Additional file 2: Dispersion plot of – Log₁₀(p-values) for trait 13-week tenth rib backfat (mm) by EMMA and standardization SNP_{ej}. Dispersion plot for 44055 –log₁₀ (p-values) by efficient mixed-model association (EMMA) using the rrBLUP R package in the x axis, and by the standardization SNP_{ej} in the y axis. Red straight line is the reference line 0–1.

Additional file 3: Variance components and LogLikelihood for models with or without the segment for all chromosomes. (Results for the 18 chromosomes).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JPS, RJCC, JLGD: performed and supervised statistical and simulation analyses and wrote the manuscript. ROB, CWE: designed the resource population and led collection of phenotypic data. CWE, NER: performed DNA extraction and coordinated genotyping with commercial laboratory. JPS, ROB, CWE: designed high density genotyping scheme. All authors read and approved the paper.

Acknowledgements

This project was supported by Agriculture and Food Research Initiative Competitive Grants no. 2010-65205-20342 and no. 2011-67015-30338 from the USDA National Institute of Food and Agriculture and by funding from the National Pork Board Grant no. 11-042. Partial funding was also provided by the US Pig Genome Coordinator. Computer resources were provided by the Michigan State University High Performance Computing Center (HPCC). JLGD and RJCC were funded by UBACyT 20020100100861 from Universidad de Buenos Aires, and PIP 11220120100621CO from CONICET (Argentina).

Author details

¹Departamento de Producción Animal, Facultad de Agronomía, UBA-CONICET, Buenos Aires, Argentina. ²Department of Animal Science, Michigan State University, East Lansing, MI, USA. ³Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA.

Received: 3 February 2014 Accepted: 7 July 2014

Published: 19 July 2014

References

1. Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C: Genomic selection and prediction in plant breeding. *J Crop Improv* 2011, **25**:239–261.
2. Goddard ME, Hayes BJ: Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* 2009, **10**:381–391.
3. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 2009, **92**:433–443.
4. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME: Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet* 2010, **6**:e1001139.
5. Kumar S, Garrick DJ, Bink MC, Whitworth C, Chagné D, Volz RK: Novel genomic approaches unravel genetic architecture of complex traits in apple. *BMC Genomics* 2013, **14**:393.
6. Zhou X, Stephens M: Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012, **44**:821–824.
7. Meuwissen TH, Hayes BJ, Goddard ME: Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 2001, **157**:1819–1829.
8. Garrick DJ: Equivalent mixed model equations for genomic selection. *J Bone Miner Res* 2007, **90**(Suppl):376 (Abstr.).
9. Strandén I, Garrick DJ: Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J Dairy Sci* 2009, **92**:2971–2975.
10. Edwards DB, Ernst CW, Tempelman RJ, Rosa GJM, Raney NE, Hoge MD, Bates RO: Quantitative trait loci mapping in an F2 Duroc x Pietrain resource population: I. Growth traits. *J Anim Sci* 2008, **86**:241–253.
11. Edwards DB, Ernst CW, Raney NE, Doumit ME, Hoge MD, Bates RO: Quantitative trait locus mapping in an F2 Duroc x Pietrain resource population: II. Carcass and meat quality traits. *J Anim Sci* 2008, **86**:254–266.
12. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu Z-L, Kerstens HH, Law AS, Megens H-J, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MA: Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* 2009, **4**:e6524.
13. Badke YM, Bates RO, Ernst CW, Schwab C, Fix J, Van Tassell CP, Steibel JP: Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genet* 2013, **14**:8.
14. Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP: Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics* 2012, **13**:24.
15. Gualdrón Duarte JL, Bates RO, Ernst CW, Raney NE, Cantet RJC, Steibel JP: Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genet* 2013, **14**:38.
16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, **81**:559–575.
17. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA: A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genet Sel Evol* 2012, **44**:9.
18. VanRaden PM: Efficient methods to compute genomic predictions. *J Dairy Sci* 2008, **91**:4414–4423.
19. Clifford D, McCullagh P: The regress function. *R News* 2006, **6**:6–10.
20. McClure MC, Ramey HR, Rolf MM, McKay SD, Decker JE, Chapple RH, Kim JW, Taxis TM, Weaber RL, Schnabel RD, Taylor JF: Genome-wide association analysis for quantitative trait loci influencing Warner-Bratzler shear force in five taurine cattle breeds. *Anim Genet* 2012, **43**:662–673.
21. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM: Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet Res (Camb)* 2012, **94**:73–83.
22. Henderson C: *Applications of Linear Models in Animal Breeding*. Guelph: University of Guelph; 1984.
23. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES: A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 2006, **38**:203–208.
24. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: Efficient control of population structure in model organism association mapping. *Genetics* 2008, **178**:1709–1723.
25. Endelman JB: Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome J* 2011, **4**:250.
26. Rangkasenee N, Murani E, Brunner RM, Schellander K, Cinar MU, Luther H, Hofer A, Stoll M, Witten A, Ponsuksilli S, Wimmers K: Genome-wide association identifies TBX5 as candidate gene for Osteochondrosis providing a functional link to cartilage perfusion as initial factor. *Front Genet* 2013, **4**:78.

27. Do DN, Ostersen T, Strathe AB, Mark T, Jensen J, Kadarmideen HN: **Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs.** *BMC Genet* 2014, **15**:27.
28. Fan Y, Xing Y, Zhang Z, Ai H, Ouyang Z, Ouyang J, Yang M, Li P, Chen Y, Gao J, Li L, Huang L, Ren J: **A further look at porcine chromosome 7 reveals VRTN variants associated with vertebral number in Chinese and Western pigs.** *PLoS One* 2013, **8**:e62534.
29. Liang K-Y, Self SG: **On the asymptotic behaviour of the Pseudolikelihood Ratio Test Statistic.** *J R Stat Soc Ser B* 1996, **58**:785–796.
30. Self SG, Liang K-Y: **Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions.** *J Am Stat Assoc* 1987, **82**:605–610.
31. Vaughan LK, Divers J, Padilla M, Redden DT, Hemant K, Pomp D, Allison DB: **The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies.** *Comput Stat Data Anal* 2009, **53**:1755–1766.
32. Klein RJ, Zeiss C, Chew EY, Tsai J, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Susan T, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H Polymorphism in age-related macular degeneration.** *Science (80-)* 2006, **308**:385–389.
33. Legarra A, Aguilar I, Misztal I: **A relationship matrix including full pedigree and genomic information.** *J Dairy Sci* 2009, **92**:4656–4663.
34. Sun X, Fernando RL, Garrick DJ, Dekkers JCM: **An iterative approach for efficient calculation of breeding values and genome-wide association analysis using weighted genomic BLUP.** *J Anim Sci* 2011, **89**(E–Suppl 2):e11.
35. Falconer D, Mackay T: *Introduction to quantitative genetics.* New York: Longman; 1996.
36. Choi I, Steibel JP, Bates RO, Raney NE, Rumph JM, Ernst CW: **Application of alternative models to identify QTL for growth traits in an F2 Duroc x Pietrain pig resource population.** *BMC Genet* 2010, **11**:97.
37. Fan B, Onteru SK, Du Z-Q, Garrick DJ, Stalder KJ, Rothschild MF: **Genome-wide association study identifies Loci for body composition and structural soundness traits in pigs.** *PLoS One* 2011, **6**:e14726.
38. Switonski M, Stachowiak M, Cieslak J, Bartz M, Grzes M: **Genetics of fat tissue accumulation in pigs: a comparative approach.** *J Appl Genet* 2010, **51**:153–168.

doi:10.1186/1471-2105-15-246

Cite this article as: Gualdrón Duarte *et al.*: Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics* 2014 **15**:246.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

