Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods

Chunyu Chen,¹ Juan P. Steibel, and Robert J. Tempelman

Department of Animal Science, Michigan State University, East Lansing, Michigan 48824

ORCID ID: 0000-0002-7833-6730 (R.J.T.)

ABSTRACT A currently popular strategy (EMMAX) for genome-wide association (GWA) analysis infers association for the specific marker of interest by treating its effect as fixed while treating all other marker effects as classical Gaussian random effects. It may be more statistically coherent to specify all markers as sharing the same prior distribution, whether that distribution is Gaussian, heavy-tailed (BayesA), or has variable selection specifications based on a mixture of, say, two Gaussian distributions [stochastic search and variable selection (SSVS)]. Furthermore, all such GWA inference should be formally based on posterior probabilities or test statistics as we present here, rather than merely being based on point estimates. We compared these three broad categories of priors within a simulation study to investigate the effects of different degrees of skewness for quantitative trait loci (QTL) effects and numbers of QTL using 43,266 SNP marker genotypes from 922 Duroc–Pietrain F₂-cross pigs. Genomic regions were based either on single SNP associations, on nonoverlapping windows of various fixed sizes (0.5–3 Mb), or on adaptively determined windows that cluster the genome into blocks based on linkage disequilibrium. We found that SSVS and BayesA lead to the best receiver operating curve properties in almost all cases. We also evaluated approximate maximum *a posteriori* (MAP) approaches to BayesA and SSVS as potential computationally feasible alternatives; however, MAP inferences were not promising, particularly due to their sensitivity to starting values. We determined that it is advantageous to use variable selection specifications based on adaptively constructed genomic window lengths for GWA studies.

KEYWORDS genome-wide association; hierarchical Bayesian; variable selection

RECENT developments in genotyping technology have made single nucleotide polymorphism (SNP) genotype marker panels, based on thousands, and now millions, of markers, available for many livestock species (Wiggans *et al.* 2013; Kemper *et al.* 2015). Genome-wide association (GWA) analyses have been increasingly used to help pinpoint regions containing potential causal variants or quantitative trait loci (QTL) for economically important phenotypes based on fitting SNP markers as covariates. An increasingly popular inferential approach for GWA is based on fitting phenotypes as a joint linear function of all markers using mixed-model procedures such as those invoked in the popular EMMAX

procedure (Kang *et al.* 2010) and other similar procedures (Lippert *et al.* 2011; Zhou and Stephens 2012). Jointly accounting for all SNP effects when inferring upon a specific SNP marker of interest generally improves precision and power, while also accounting for potential population structure (Kang *et al.* 2008).

Now GWA inferences in EMMAX and related procedures are based on treating the effect of the SNP marker of interest as fixed, with all other marker effects as normally distributed random effects, noting that this process is repeated in turn for every single marker. These “fixed effects” hypothesis tests are based on generalized least squares (GLS) inference, with *P*-values being subsequently adjusted for the total number of markers or tests. Goddard *et al.* (2016) have recently pointed out the paradox with treating markers as fixed for inference but then otherwise as random to account for population structure for inference on association with other markers. Random-effects modeling with all SNP effects treated as random, including the one of inferential interest, is

Copyright © 2017 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.117.202259>

Manuscript received March 26, 2017; accepted for publication June 19, 2017; published Early Online June 21, 2017.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.202259/-/DC1.

¹Corresponding author: 1205 Anthony Hall, Michigan State University, East Lansing, MI 48824. E-mail: chench57@msu.edu

synonymous with shrinkage-based inference. Shrinkage or posterior inference has been demonstrated to facilitate reliable inference without any formal requirements for multiple comparison adjustments (Stephens and Balding 2009; Gelman *et al.* 2012). However, with SNP markers treated as identically and independently distributed variables from a Gaussian distribution, the resulting shrinkage from random-effects modeling can be too “hard,” particularly with greater marker densities (Hayes 2013). Subsequently, this random-effects test has been deemed to be far too conservative in various applications, as further demonstrated by Gualdrón Duarte *et al.* (2014).

Prior specifications that are sparser than Gaussian priors may be more important for GWA since they more likely better characterize the true genetic architecture of most traits relative to Gaussian priors (de los Campos *et al.* 2013). Sparser specifications have already been popularized in whole genome prediction (WGP), such as the Student's *t*-distribution used in BayesA (Meuwissen *et al.* 2001) and stochastic search and variable selection (SSVS) (George and McCulloch 1993; Verbyla *et al.* 2009). Both specifications generally lead to far less shrinkage of large effects yet greater shrinkage of small effects compared to a Gaussian prior. In particular, the use of variable selection procedures facilitate the determination of posterior probabilities of association (PPA), whose control may be far more effective in maximizing both sensitivity and specificity of GWA (Fernando *et al.* 2017) compared to frequentist-based inferences, which require adjustments for multiple testing, such as with EMMAX. Another common inferential strategy in GWA is to simply report the percent of variance explained by a marker or marker region (Fan *et al.* 2011; Tizioto *et al.* 2015; Wolc *et al.* 2016). However, point estimates of marker effects or percentage of variation explained, by themselves, do not provide formal evidence of association.

Most sparse prior WGP models have been implemented using Markov chain Monte Carlo (MCMC), which can be computationally expensive. Approximate analytical approaches based on the expectation–maximization (EM) algorithm to provide approximate maximum *a posteriori* (MAP) estimates of SNP effects have been developed to address computational limitations in these sparse prior WGP models (Meuwissen *et al.* 2009; Hayashi and Iwata 2010; Sun *et al.* 2012; Chen and Tempelman 2015). Strategies for estimating/tuning hyperparameters for MAP inference have been proposed, including those proposed by Kärkkäinen and Sillanpää (2012), Knürr *et al.* (2013), and Chen and Tempelman (2015), the latter adapting the average information restricted maximum likelihood (AIREML) algorithm for estimating hyperparameters in BayesA and SSVS specifications. These MAP implementations should also be assessed for their efficacy in GWA studies.

A pragmatic first objective in GWA is to pinpoint narrow genomic regions containing QTL rather than to specifically identify the QTL themselves, even though the latter is the

ultimate goal. That is, a large number of SNP markers in a region surrounding a typically untyped QTL might be in high linkage disequilibrium (LD) with the QTL and with each other, thereby thwarting precise inference on the causal QTL. Different GWA methods may differ in the number of SNP markers inferred to have an association within a genomic region with, for example, EMMAX tending to draw associations with more SNP markers in LD with a QTL compared to use of SSVS (Guan and Stephens 2011; Goddard *et al.* 2016).

Increasingly, more GWA studies are based on inferences involving joint tests on all of the SNP markers within a narrow genomic region, recognizing that single SNP marker associations may be fraught by low statistical power or problems with multicollinearity or both (Fernando *et al.* 2017). Some GWA studies have been based on using several arbitrary window sizes based on either nonoverlapping (Wolc *et al.* 2012, 2016; Moser *et al.* 2015) or sliding windows (Schmid and Yang 2008). Because of the arbitrariness of fixed-window sizes, whether defined by the number of SNP markers or by physical length in base pairs, it is possible to split a large LD block into two or more separate windows, thereby making such a division seemingly sub-optimal for GWA. Substantially different window lengths have been used in different studies. For example, a 5-SNP window was used for GWA based on 51,385 SNP markers in pigs (Fan *et al.* 2011), whereas a 250-kb window was used for 287,854 SNPs from the Wellcome Trust Case Control Consortium human data (Wellcome Trust Case Control Consortium 2007; Moser *et al.* 2015), and a 1-Mb window was used for a 24,425-SNP marker panel in chickens (Wolc *et al.* 2012). Dehman *et al.* (2015) recently proposed an approach to adaptively cluster windows of SNP markers of varying sizes based on LD relationships. That is, they performed spatially constrained hierarchical clustering of SNPs by minimizing a distance measure derived from Ward's criterion based on LD r^2 between SNP markers. They surmised that this procedure would estimate a suitable specification of genomic windows within each chromosome using a modified version of the gap statistic. This method has been implemented in the R package BALD (Dehman and Neuvial 2015).

We had three primary objectives in this study. One was to examine the potential benefits of using sparser priors (*i.e.*, BayesA and SSVS) relative to classical (*i.e.*, based on normality) random-effects specifications and strategies for GWA under a wide range of simulated architectures. A second objective was to assess whether the choice of different fixed genomic window sizes (specifically 0.5, 1, 2, and 3 Mb) vs. adaptively inferred window sizes based on LD clustering could affect GWA performance. A final objective was to assess the relative merit of approximate MAP approaches to theoretically exact yet computationally intensive MCMC approaches based on sparse prior specifications. Our assessments are based upon SNP marker genotypes and actual and simulated phenotypes on F₂ pigs deriving from a Duroc–Pietrain cross.

Materials and Methods

The hierarchical linear model

All analyses in this article are based on a hierarchical linear model which is characterized by the classical mixed-model specification:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e}. \quad (1)$$

Here $\mathbf{y}$ is an $n \times 1$ vector of phenotypes, $\mathbf{X}$ is a known $n \times p$ incidence matrix connecting $\mathbf{y}$ to the $p \times 1$ vector of unknown fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}$ is a known $n \times m$ matrix of genotypes connecting $\mathbf{y}$ to the $m \times 1$ vector of unknown random SNP marker effects $\mathbf{g}$, and $\mathbf{e}$ is the random error vector. We also assume throughout that $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, whereas $\mathbf{g}|\sigma_g^2, \mathbf{D} \sim N(0, \mathbf{D}\sigma_g^2)$ for $\mathbf{D}$ being a diagonal matrix of augmented data or variables (Chen and Tempelman 2015; Tempelman 2015). The prior specification on these diagonal elements is used to distinguish each of the competing models, as described later.

For pedagogical reasons, we assume one record per individual, although extensions to repeated records per individual are possible. An equivalent genomic animal (*i.e.*, subject) effects model (VanRaden 2008) to Equation 1 can then be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \mathbf{e}, \quad (2)$$

with $\mathbf{a} = \mathbf{Z}\mathbf{g}$ and all other terms defined previously as in (1) such that, conditionally on $\mathbf{D}$,

$$\text{var}(\mathbf{a}) = \text{var}(\mathbf{Z}\mathbf{g}) = \mathbf{Z} \text{var}(\mathbf{g})\mathbf{Z}' = \mathbf{Z}\mathbf{D}\mathbf{Z}'\sigma_g^2. \quad (3)$$

If $m \gg n$ it is generally computationally more tractable to work with the linear mixed model in Equation 2, based on the random-effects specification in Equation 3. One can then backsolve for the estimate of $\mathbf{g}$ that would be identical to those using a linear mixed model directly based on Equation 1 (Stranden and Garrick 2009).

Models

In the simplest model, which we denote as ridge regression (RR), there is no such data augmentation (*i.e.*, $\mathbf{D} = \mathbf{I}$), such that the elements of $\mathbf{g}$ are marginally distributed as independent normal (de los Campos *et al.* 2013). Sparser distributional specifications on $\mathbf{g}$ can be constructed as mixtures of normal densities (Andrews and Mallows 1974) by simply specifying prior distributions on functions of the diagonal elements of $\mathbf{D}$. Suppose that $\mathbf{D} = \text{diag}\{\tau_j\}_{j=1}^m$ with $\tau_j \sim \chi^{-2}(\nu_\tau, \nu_\tau)$; then it can be demonstrated that, marginally, elements of $\mathbf{g}$ are identically and independently distributed as a scaled Student's t with scale parameter σ_g^2 and degrees of freedom ν_τ (Chen and Tempelman 2015). This model is typically referred to as BayesA (Meuwissen *et al.* 2001). Alternatively, if $\mathbf{D} = \text{diag}\{\tau_j + (1 - \tau_j)/c\}_{j=1}^m$, where $\tau_j \sim \text{Bernoulli}(\pi_\tau)$; $\tau_j = 0, 1$ and $c \gg 1$, then the resulting model is Bayes SSVS in the spirit of George and McCulloch (1993). As a side note, we use $c = 1000$ for all SSVS analyses in this article.

As a final stage in each of the competing hierarchical models (RR, BayesA, and SSVS), we specify convenient conjugate priors wherever possible. For example, scaled inverted chi-square priors for variance components, *i.e.*,

$$p(\sigma_e^2 | \nu_e, s_e^2) \propto (\sigma_e^2)^{-\left(\frac{\nu_e}{2} + 1\right)} e^{-\frac{\nu_e s_e^2}{2\sigma_e^2}} \quad (4)$$

and

$$p(\sigma_g^2 | \nu_g, s_g^2) \propto (\sigma_g^2)^{-\left(\frac{\nu_g}{2} + 1\right)} e^{-\frac{\nu_g s_g^2}{2\sigma_g^2}}. \quad (5)$$

Whereas we specify a Beta prior on π_τ in SSVS, *i.e.*,

$$p(\pi_\tau | \alpha_0, \beta_0) \propto \pi_\tau^{\alpha_0} (1 - \pi_\tau)^{\beta_0}. \quad (6)$$

As we explain later, we arbitrarily specify ν_τ as known ($\nu_\tau = 2.5$), although conceptually it could also be estimated (Yang *et al.* 2015). We assume throughout that $p(\boldsymbol{\beta}) \propto 1$ as $p(\boldsymbol{\beta})$ is typically diffuse, although extensions to more informative specifications should be obvious. Furthermore, for all analyses in this article, we specify Gelman's noninformative prior (Gelman 2006) for σ_e^2 in Equation 4 based on $\nu_e = -1$ and $s_e^2 = 0$, and for σ_g^2 in Equation 5 based on $\nu_g = -1$ and $s_g^2 = 0$. Furthermore, as per Yang and Tempelman (2012), we specify $\alpha_0 = 1$ and $\beta_0 = 9$.

Joint posterior density

Given the specifications above, the joint posterior density can be written as in Equation 7:

$$p(\boldsymbol{\beta}, \mathbf{g}, \boldsymbol{\tau}, \sigma_e^2, \sigma_g^2, \theta_\tau | \mathbf{y}) \propto \left(\prod_{i=1}^n p(y_i | \boldsymbol{\beta}, \mathbf{g}, \sigma_e^2) \right) \left[\prod_{j=1}^m p(g_j | \sigma_g^2, \tau_j) p(\tau_j | \theta_\tau) \right] p(\boldsymbol{\beta}) p(\sigma_g^2 | \nu_g, s_g^2) p(\sigma_e^2 | \nu_e, s_e^2) p(\theta_\tau). \quad (7)$$

Note that $p(\tau_j | \theta_\tau)$ specifies the $\chi^{-2}(\nu_\tau, \nu_\tau)$ density under BayesA (*i.e.*, $\theta_\tau \equiv \nu_\tau$), whereas $p(\tau_j | \theta_\tau)$ specifies the Bernoulli(π_τ) density under SSVS (*i.e.*, $\theta_\tau \equiv \pi_\tau$). Furthermore, $p(g_j | \sigma_g^2, \tau_j)$ is Gaussian with null means under all three competing models but with variance $\sigma_g^2 \tau_j$ under BayesA and variance $\sigma_g^2 [\tau_j + (1 - \tau_j)/c]$ under SSVS. For RR, $\tau_j = 1 \forall j$ such that $p(g_j | \sigma_g^2, \tau_j = 1)$ is Gaussian with common variance $\sigma_g^2 \forall j$.

Algorithms

MCMC: The MCMC sampling strategies that we use here for BayesA are similar to those provided in Yang and Tempelman (2012) and Yang *et al.* (2015). However, since our parameterization is slightly different, we present the full conditional densities of interest for implementing BayesA in Supplemental Material, [File S1](#). For similar reasons, we also provide the full conditional densities for SSVS in [File S1](#), as our model differs from the model also labeled as SSVS in the genomic prediction work of Verbyla *et al.* (2009), whereas it is

virtually identical to the model presented in the seminal SSVS article by George and McCulloch (1993).

Maximum a posterior estimation: The complete details on our MAP procedure for both BayesA and SSVS are found in Chen and Tempelman (2015). Given that our application involved $m \gg n$, we conducted MAP-based inference on an equivalent subject-centric model using Equation 2 rather than based on a SNP-effects model as in Equation 1. Details on backsolving from a subject-centric model to provide estimates of SNP effects are provided in File S1. For pedagogical reasons, however, we work directly from the SNP-effect model (1) in our subsequent developments. Conditional on $\mathbf{D}$, the posterior variance-covariance matrix of $\mathbf{g}$, or equivalently its prediction error variance-covariance (PEV) matrix from a frequentist viewpoint, can be written as: $\text{var}(\mathbf{g}|\mathbf{y}, \mathbf{D}, \sigma_e^2, \sigma_g^2, \theta_\tau) = \mathbf{C}^{\text{gg}|\mathbf{D}}$. This expression can be derived from the inverse of the mixed-model coefficient matrix as:

$$\begin{bmatrix} \mathbf{C}^{\beta\beta|\mathbf{D}} & \mathbf{C}^{\beta\mathbf{g}|\mathbf{D}} \\ \mathbf{C}^{\mathbf{g}\beta|\mathbf{D}} & \mathbf{C}^{\mathbf{g}\mathbf{g}|\mathbf{D}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X}\sigma_e^{-2} & \mathbf{X}'\mathbf{Z}\sigma_e^{-2} \\ \mathbf{Z}'\mathbf{X}\sigma_e^{-2} & \mathbf{Z}'\mathbf{Z}\sigma_e^{-2} + \mathbf{D}^{-1}\sigma_g^{-2} \end{bmatrix}^{-1}. \quad (8)$$

That is, $\mathbf{C}^{\text{gg}|\mathbf{D}}$ is the random-by-random portion of the inverse coefficient matrix in Henderson's mixed-model equations, conditional on $\mathbf{D}$, σ_e^2 , σ_g^2 , and θ_τ ($\theta_\tau \equiv \nu_\tau$ for BayesA or $\theta_\tau \equiv \pi_\tau$ for SSVS). As noted earlier, values for hyperparameters such as σ_e^2 , σ_g^2 , and θ_τ required for Equation 8 can be determined using the REML or marginal maximum likelihood estimation strategies as described by Chen and Tempelman (2015), noting that we choose to fix ν_τ in BayesA as indicated earlier.

It can be readily demonstrated (Sorensen and Gianola 2002) that asymptotically $\text{MAP}(\mathbf{g}) \approx E(\mathbf{g}|\mathbf{y})$, whereby $\text{MAP}(\mathbf{g})$ can be iteratively determined using EM based on the Newton-Raphson method for maximization (M-) steps interwoven with expectation (E-) steps on elements of $\mathbf{D}$ (Chen and Tempelman 2015). Under RR, $\mathbf{D} = \mathbf{I}$ such that $\mathbf{C}^{\text{gg}} = \mathbf{C}^{\text{gg}|\mathbf{D}} \approx \text{var}(\mathbf{g}|\mathbf{y})$ represents the posterior variance-covariance matrix of $\mathbf{g}$ conditional on σ_e^2 , σ_g^2 , and θ_τ . In fact, $\text{MAP}(\mathbf{g})$ is synonymous with BLUP($\mathbf{g}$) under RR. Furthermore, $\mathbf{C}^{\text{gg}}$ is synonymous with the $\mathbf{g}$ component of the observed information matrix of the joint conditional posterior density of β and $\mathbf{g}$. This posterior density is formally defined in Equation 9:

$$p(\beta, \mathbf{g}, |\sigma_e^2, \sigma_g^2, \theta_\tau | \mathbf{y}) \propto \left[\prod_{i=1}^n p(y_i | \beta, \mathbf{g}, \sigma_e^2) \right] \times \left[\prod_{j=1}^m \int_{R_{\tau_j}} p(g_j | \sigma_g^2, \tau_j) p(\tau_j | \theta_\tau) d\tau_j \right]. \quad (9)$$

With $\mathbf{D} = \mathbf{I}$, there is no uncertainty on τ_j such that the integration in Equation 9 is not necessary, with $\mathbf{C}^{\text{gg}}$ being di-

rectly obtainable for RR using Equation 8. However, for BayesA and SSVS, uncertainty in $\mathbf{D}$ needs to be integrated out as per Equation 9. An indirect strategy for asymptotically providing $\mathbf{C}^{\text{gg}}$ for BayesA and SSVS is based on the strategy proposed by Louis (1982), with details provided in File S1. We subsequently use elements of $\mathbf{C}^{\text{gg}}$ to asymptotically determine key components of $\text{var}(\mathbf{g}|\mathbf{y})$ for both single SNP- and window-based GWA testing using MAP under all three models, noting again that MAP and BLUP are synonymous under RR.

Conducting GWA analyses

Single SNP marker associations: We subsequently describe how we conducted GWA inference for single SNP associations based on the algorithms (MCMC vs. MAP) and models (RR, BayesA, and SSVS). With respect to inference on association on SNP j , EMMAX is conceptually based on subsetting out Equation 1 as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{z}_j g_j + \mathbf{Z}_{-j} \mathbf{g}_{-j} + \mathbf{e}. \quad (10)$$

That is, $\mathbf{Z}$ is partitioned into column j , $\mathbf{z}_j$, being the genotypes for SNP j and all other remaining columns in $\mathbf{Z}_{-j}$. In EMMAX, g_j is actually treated as fixed whereas $\mathbf{g}_{-j}$ is treated as classically random; i.e., characterized by a Gaussian prior distribution. Writing $\mathbf{W}_j = [\mathbf{X} \quad \mathbf{z}_j]$ and $\mathbf{V}_{-j} = \mathbf{Z}_{-j} \mathbf{Z}_{-j}' \sigma_g^2 + \mathbf{I} \sigma_e^2$, the GLS estimator $\hat{g}_j$ of g_j , using all other markers to account for population structure, is the last element of the product $(\mathbf{W}_j' \mathbf{V}_{-j}^{-1} \mathbf{W}_j)^{-1} \mathbf{W}_j' \mathbf{V}_{-j}^{-1} \mathbf{y}$. Furthermore, the corresponding SE $\text{se}(\hat{g}_j)$ is determined by the square root of the last diagonal element of $(\mathbf{W}_j' \mathbf{V}_{-j}^{-1} \mathbf{W}_j)^{-1}$. The test statistic or fixed effects z -score for the EMMAX test can then be simply written as:

$$z_f = \frac{\hat{g}_j}{\text{se}(\hat{g}_j)}, \quad (11)$$

which is assumed to be $N(0, 1)$ under H_0 : $g_j = 0$. The “expedited” approach (Kang *et al.* 2010) in EMMAX, which we consider in this article, is based on approximating $\mathbf{V}_{-j}$ with $\mathbf{V} = \mathbf{Z}\mathbf{Z}' \sigma_g^2 + \mathbf{I} \sigma_e^2$ for inference of association on all SNP $j = 1, 2, \dots, m$; furthermore, σ_g^2 and σ_e^2 estimated only once using REML in an initial analysis that treats all SNP marker effects as random. A GWA test for a particular SNP marker j using EMMAX then essentially involves treating its effect jointly as both fixed and random by replacing $\mathbf{Z}_{-j} \mathbf{g}_{-j}$ with $\mathbf{Z}\mathbf{g}$ on the right side of Equation 10, implying that this double counting of g_j as both fixed and random should be trivial with large m .

A classical shrinkage or random-effects test for H_0 : $g_j = 0$ is based on treating all SNP effects, including a marker j of particular interest, as having a Gaussian prior such that the point estimate of the SNP-substitution effect is based on fitting Equation 1 or, equivalently, backsolving from fitting Equation 3, as demonstrated by Strandén and Garrick (2009) and also in File S1. A corresponding test statistic (z_r) can be based on dividing $\tilde{g}_j$, the BLUP of g_j , by the square root of its PEV where $\text{PEV}(\tilde{g}_j) = \text{var}(\tilde{g}_j - g_j)$ from a frequentist perspective. From a Bayesian perspective, the corresponding test statistic can be interpreted as a

posterior z -score (Gelman *et al.* 2012) since $\tilde{g}_j$ is analogous to a posterior mean [*i.e.*, $\tilde{g}_j = E(g_j|\mathbf{y}, \hat{\sigma}_e^2, \hat{\sigma}_g^2) \approx E(g_j|\mathbf{y})$], whereas the PEV is analogous to a posterior variance with $PEV(\tilde{g}_j) = \text{var}(g_j|\mathbf{y}, \hat{\sigma}_e^2, \hat{\sigma}_g^2) \approx \text{var}(g_j|\mathbf{y})$. We refer to this inference strategy as RRBLUP. It is important to indicate, nevertheless, that these RRBLUP inferences are empirically Bayesian (Robinson 1991) since these posterior means and variances are typically conditioned upon REML estimates of σ_e^2 and σ_u^2 . The posterior z -score (Gelman *et al.* 2012) can then equivalently derived from both frequentist and Bayesian perspectives as indicated in Equation 12:

$$z_r = \frac{\tilde{g}_j}{\sqrt{PEV(\tilde{g}_j)}} = \frac{E(g_j|\mathbf{y})}{\sqrt{\text{var}(g_j|\mathbf{y})}}. \quad (12)$$

Now Gualdrón Duarte *et al.* (2014), with a proof provided later by Bernal Rubio *et al.* (2016), determined that the fixed effects or EMMAX z -score, z_f in Equation 11, could be equivalently derived by treating all markers as classically random, but by dividing the corresponding BLUP $\tilde{g}_j$ for marker j by the square root of its frequentist definition of variance $\text{var}(\tilde{g}_j)$, as characterized by classical mixed-model theory (Searle *et al.* 1992) in Equation 13:

$$\text{var}(\tilde{g}_j) = \text{var}(g_j) - PEV(\tilde{g}_j) = \sigma_g^2 - PEV(\tilde{g}_j). \quad (13)$$

In other words, one can rewrite the fixed-effects test provided in both its frequentist (numerator = $\tilde{g}_j$) and Bayesian [numerator = $E(g_j|\mathbf{y})$] representations as in Equation 14:

$$z_f = \frac{\tilde{g}_j}{\sqrt{\sigma_g^2 - PEV(\tilde{g}_j)}} = \frac{E(g_j|\mathbf{y})}{\sqrt{\sigma_g^2 - \text{var}(g_j|\mathbf{y})}}. \quad (14)$$

Note that the computation using Equation 14 is far more tractable than that implied with Equation 11. That is, Equation 14 only requires computing BLUP of $\mathbf{g}$ and its corresponding PEV in one single-step determination for all m tests, whereas Equation 11 implies m different mixed-model analyses, each one in turn explicitly treating a different SNP-marker effect as fixed.

We perceive no computationally tractable fixed-effects test analogous to EMMAX that we could adapt for MAP based on sparser priors (*e.g.*, BayesA and SSVS). For BayesA, for example, this would entail treating the marker of interest j as fixed with all other markers treated as scaled Student's t -distributed. However, a posterior or random-effects z -score test can be constructed using the MAP estimate of g_j as the numerator and its asymptotic posterior SE as the denominator, noting that MAP and the posterior mean of $\mathbf{g}$ should approach each other asymptotically. Details on deriving those asymptotic SEs (*i.e.*, based on deriving $\mathbf{C}^{gg}$) for use in Equation 11 for these sparse prior specifications are provided in [File S1](#) such that we refer to these two corresponding GWA inference strategies as MAP-BayesA and MAP-SSVS.

For SSVS-based, single-SNP inferences using MCMC, we based inferences on the PPA for SNP marker j (*i.e.*, PPA _{j}) as in Equation 15:

$$PPA_j = \frac{\sum_{l=1}^N \tau_{j(l)}}{N}. \quad (15)$$

Here N denotes the number of MCMC cycles saved for posterior inference and $\tau_{j(l)}$ is a binary draw from the full conditional distribution of τ_j at MCMC cycle l . We denote this GWA method as MCMC-SSVS.

Since there is no variable selection inherent with BayesA under MCMC, we based single SNP inferences on a Bayesian analog to a P -value using

$$\hat{p}_j = 2 \min \left[\frac{\sum_{l=1}^N I(g_{j(l)} > 0)}{N}, 1 - \frac{\sum_{l=1}^N I(g_{j(l)} > 0)}{N} \right] \quad (16)$$

(Bello *et al.* 2010), where the indicator variable $I(\cdot) = 1$ if the condition within the argument is true and 0 otherwise. We denote this particular GWA method as MCMC-BayesA.

Windows-based associations: Window-based extensions to all of the above tests were also developed, some based on work previously presented above. Suppose that window k , $k = 1, 2, 3, \dots, K$ contains n_k markers such that $\mathbf{Z}$ can be partitioned accordingly into $\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2 \ \dots \ \mathbf{Z}_K]$ with $\mathbf{Z}_k$ having n_k columns, implying then that window k contains n_k SNP markers. Similarly, the vector $\mathbf{g}$ is partitioned accordingly; *i.e.*, $\mathbf{g} = [\mathbf{g}'_1 \ \mathbf{g}'_2 \ \dots \ \mathbf{g}'_K]'$ such that $\mathbf{g}_k$ is of dimension $n_k \times 1$. Recall that we denoted $\mathbf{C}^{gg} = PEV(\tilde{\mathbf{g}})$. For our proposed windows-based test, the key components of $\mathbf{C}^{gg}$ can be partitioned into K different blocks along the block diagonal; *i.e.*, $\mathbf{C}_1^{gg}, \mathbf{C}_2^{gg}, \dots, \mathbf{C}_K^{gg}$, where $\mathbf{C}_k^{gg}$ is of dimension $n_k \times n_k$. The extension to a joint fixed-effects or EMMAX-like test on n_k markers in window k involves the following extension of Equation 14:

$$\chi_f^2 = \tilde{\mathbf{g}}_k (\mathbf{I}_{n_k} \sigma_g^2 - \mathbf{C}_k^{gg})^{-1} \tilde{\mathbf{g}}_k. \quad (17)$$

That is, it can be readily demonstrated, extending results from Bernal Rubio *et al.* (2016), that χ_f^2 is chi-square distributed with n_k degrees of freedom under H_0 : $\mathbf{g}_k = 0$. The corresponding extension to a joint classical random-effects or RRBLUP test on window k is provided in Equation 18,

$$\chi_r^2 = \tilde{\mathbf{g}}_k (\mathbf{C}_k^{gg})^{-1} \tilde{\mathbf{g}}_k, \quad (18)$$

which would also be considered to be chi-square distributed with n_k degrees of freedom under H_0 : $\mathbf{g}_k = 0$. Similarly, one could use Equation 18 to construct the same tests for MAP-BayesA and MAP-SSVS, but basing the $\mathbf{C}^{gg}$ on the corresponding asymptotic posterior variance-covariance matrices as derived in [File S1](#).

For windows-based inference using MCMC-SSVS, we simply compute the PPA for window k (i.e., PPA_k) in Equation 19, following that presented in Fernando *et al.* (2017):

$$PPA_k = \frac{\sum_{l=1}^N \left[I \left(\sum_{j=1}^{n_k} \tau_{kj(l)} \right) > 0 \right]}{N}. \quad (19)$$

Here, $\tau_{kj(l)}$ defines a binary draw from the full conditional distribution of τ_j for SNP marker j located within window k drawn during MCMC cycle l . Note then that $I \left(\sum_{j=1}^{n_k} \tau_{kj(l)} \right) > 0$ is equal to 1 when any of the draws of $\tau_{kj(l)}$ within window k are equal to 1.

For windows-based GWA inference under MCMC-BayesA, we propose inferring upon the posterior probability of the proportion (q_w) of the genetic variance explained by the markers in a genomic window relative to the total genetic variance as proposed by Fernando and Garrick (2013) and determined in the following manner. First note that the genotypic value that is attributed to a genomic window k is defined as in Equation 20:

$$\mathbf{a}_k = \mathbf{Z}_k \mathbf{g}_k. \quad (20)$$

Then the variance explained by the window is defined as

$$\sigma_{a_k}^2 = \frac{\mathbf{a}_k' \mathbf{a}_k}{n_k} - \left(\frac{\mathbf{1}_{n_k}' \mathbf{a}_k}{n_k} \right)^2. \quad (21)$$

Similarly, the total genetic variance is computed as

$$\sigma_a^2 = \frac{\mathbf{a}' \mathbf{a}}{m} - \left(\frac{\mathbf{1}_m' \mathbf{a}}{m} \right)^2. \quad (22)$$

Hence, the proportion of genetic variance that is explained by a marker in window k is defined as

$$q_k = \frac{\sigma_{a_k}^2}{\sigma_a^2}. \quad (23)$$

Suppose that we deem genomic windows that explain $>1\%$ of the total genetic variance as being of potential interest. Hence, a variable selection modification of MCMC-BayesA can simply be based on the proportion of MCMC samples for which the genetic variance (q_k) for window k exceeds 0.01 (Fernando and Garrick 2013). One advantage of this approach is that it can be applied to any MCMC analyses based on a model where variable selection is not explicitly specified.

Data

Simulation study: To compare the various models (RR, BayesA, and SSVS) and algorithms (MAP vs. MCMC), we simulated data based on the Michigan State University Pig

Resource Population (MSUPRP) raised at the Michigan State University Swine Teaching and Research Farm, East Lansing, MI (Edwards *et al.* 2008). We specifically started with the SNP markers chosen for analysis by Gualdrón Duarte *et al.* (2014), which included 928 Duroc-Pietrain F_2 crosses. Roughly 1/3 of these pigs were directly genotyped using the Illumina Porcine SNP60 BeadChip (60K), whereas the remaining F_2 animals were genotyped using a lower density 9K set but these genotypes were subsequently imputed to the 60K set (Gualdrón Duarte *et al.* 2013). Edits excluded animals with $>10\%$ of their SNP markers missing, excluding SNP markers with $>10\%$ of animals missing genotypes for those markers, and excluding SNPs with a minor allele frequency <0.01 (Gualdrón Duarte *et al.* 2014). Some adjacent markers were in complete LD with each other. To circumvent multicollinearity issues, particularly its role in generating multimodality in the MCMC-generated posterior densities for some SNP markers (Calus *et al.* 2015), we randomly deleted one SNP within an adjacent pair in complete LD with each other before further analyses. After invoking this edit, 922 pigs and 43,266 SNPs remained. The original data source is provided as [File S3](#).

To simulate different but representative genetic architectures, we generated QTL effects from three different gamma densities with demonstrably different values of shape (γ), ranging from an effectively oligogenic density ($\gamma = 0.18$), which effectively specifies relatively much fewer QTL with large effects; to an effectively polygenic Gaussian density ($\gamma = 3.00$), where most QTL have intermediate effects with symmetrically small and large effects on either side. A third intermediate value ($\gamma = 1.48$) was also chosen. A good illustration of the gamma density of QTL effects based on these three different specifications for γ is provided in Figure 1. Note that this range in γ values for QTL effects has been reported for various traits in livestock based on previous empirical work (Hayes and Goddard 2001).

In addition to the distribution of QTL effects, we conjectured that the number of QTL (n_{qtl}) may also influence GWA performance such that we considered $n_{\text{qtl}} = 30, 90$, or 300. Hence, we simulated 10 replicated populations under each of the $3 \times 3 = 9$ different scenarios pertaining to the three different values for each of γ and of n_{qtl} . Each of the 90 simulated data sets were based on utilization of the 43,266 SNP marker genotypes on the $n = 922$ MSUPRP F_2 pigs as previously described. Within each data set, allelic substitution effects, $\mathbf{g}_{\text{qtl}}$, were simulated for each of the n_{qtl} randomly chosen SNP markers from the corresponding gamma distribution having shape γ , with a randomly chosen half of those effects multiplied by -1 as per Meuwissen *et al.* (2001). The corresponding genotypes $\mathbf{Z}_{\text{qtl}}$ for QTL on these animals were then a $n \times n_{\text{qtl}}$ subset of the SNP genotype matrix $\mathbf{Z}$ such that the cumulative genetic merit or true breeding values was determined as $\mathbf{u}_{\text{TRUE}} = \mathbf{Z}_{\text{qtl}} \mathbf{g}_{\text{qtl}}$. Phenotypes for animals were generated based on a heritability of 0.45 as estimated for 13th-week, 10th-rib back fat from this same data set. Only

Figure 1 Distribution of QTL effects under a gamma distribution for different specifications of shape (magenta curve, $\gamma = 0.18$; blue curve, $\gamma = 1.48$; and red curve, $\gamma = 3.00$).

the remaining (*i.e.*, non-QTL) marker genotypes $\mathbf{Z}_{-qtl}$ were used for all simulation study analyses.

In the simulation study, all parameters excluding ν_τ in BayesA were estimated using both MCMC and MAP. For MCMC, we ran 200,000 iterations, discarding the first 100,000 iterations as burn-in and basing inference on saving every 10 of the remaining 100,000 cycles for a total of 10,000 samples from the posterior density. Using MAP, estimation of variance components ($\boldsymbol{\theta}$) for BayesA and SSVS was based on

a convergence criterion of $\frac{[\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k-1)}]^T [\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k-1)}]}{[\hat{\boldsymbol{\theta}}^{(k)}]^T [\hat{\boldsymbol{\theta}}^{(k)}]} < 10^{-6}$. Based

on our previous experience (Chen and Tempelman 2015), we recognized that the specification of starting values in MAP-SSVS and MAP-BayesA was important for genomic prediction accuracy and, hence, likely important for GWA inferences as well. Strategies for specifying starting values for σ_g^2 , σ_e^2 , $\mathbf{g}$, and $\boldsymbol{\tau}$ may pragmatically involve using REML and RRBLUP inferences as in Chen and Tempelman (2015) since RRBLUP is not computationally intensive. For MAP-BayesA, starting values were based on REML estimates $\hat{\sigma}_{g(\text{REML})}^2$ and $\hat{\sigma}_{e(\text{REML})}^2$

using $\sigma_{g(0)}^2 = \frac{\nu_g - 2}{\nu_g} \hat{\sigma}_{g(\text{REML})}^2$ for σ_g^2 , $\mathbf{g}_0 = \text{BLUP}(\mathbf{g}) = \{\mathbf{g}_{0j}\}_{j=1}^m$

for $\mathbf{g}$, and $\tau_{(0)j} = \left(\frac{g_{0j}^2}{\sigma_{g(0)}^2} + \nu_g \right) / (\nu_g - 1)$ for $\tau_j, j = 1, 2, \dots, m$,

based on the posterior expectation derived from its full conditional density. For MAP-SSVS, the corresponding starting values were $\sigma_{g(0)}^2 = \hat{\sigma}_{g(\text{REML})}^2 / \pi_0$ for σ_g^2 with the starting value $\pi_{\tau(0)}$ for $\boldsymbol{\tau}$ based, in turn, on starting values for τ_j (*i.e.*, SNP-specific PPA) which were determined in the following manner. First of all, EMMAX-based P -values for each SNP were converted to local false discovery rate (IFDR) estimates

using the R package *ashr* (Stephens 2017). It has been demonstrated that these IFDR estimates, in turn, can be used to approximate PPA using $\text{PPA} \approx 1 - \text{IFDR}$ (Stephens 2017). These approximate PPA values were then chosen as the starting values for τ_j in MAP-SSVS. In turn, these starting values for τ_j were used to derive the starting value for π_0 in MAP-SSVS using the posterior expectation from its full conditional density, *i.e.*, $\pi_0 = (\alpha_0 + \sum_{j=1}^m \tau_j) / (\alpha_0 + \beta_0 + m)$. Upon convergence of variance components using the AIREML procedure outlined in Chen and Tempelman (2015), convergence of MAP-based solutions to $\mathbf{g}$ were based on the same criteria.

Single SNP marker inferences were based on the procedures outlined previously; *i.e.*, for MAP by comparing z_r in Equation 12 for the random-effect tests for RRBLUP, MAP-BayesA, and MAP-SSVS and z_f for the EMMAX test in Equation 11 to a standard normal distribution. Furthermore, the estimates of PPA and Bayesian P -values provided in Equations 15 and 16 were respectively used for GWA under MCMC-SSVS and MCMC-BayesA. Since the remaining genotypes $\mathbf{Z}_{-qtl}$ did not include the simulated QTL, a SNP marker was declared a true positive if a QTL was located between that marker and its closest SNP neighbor on either side.

Window-based inference was based on the procedures outlined previously; *i.e.*, for MAP by computing χ_r^2 in Equation 18 for the random effect tests using RRBLUP, MAP-BayesA, and MAP-SSVS and χ_f^2 for fixed-effects test in Equation 17 under EMMAX. These test statistics were compared to a chi-square distribution with degrees of freedom n_k . Furthermore, GWA was based on the PPA that $q_k > 0.01$ as provided in Equation 23 for MCMC-BayesA and on the PPA for MCMC-SSVS as provided in Equation 19.

For windows-based inference, four alternative fixed-window sizes were chosen: 0.5, 1, 2, or 3 Mb. The genome map used was the *Sus Scrofa* build 10.2 (http://www.ensembl.org/Sus_scrofa/Info/Index). Also, as per Moser *et al.* (2015), two different within-chromosome starting positions (starting at location 0 or 0.25 Mb for window sizes 0.5, starting at 0 or 0.5 Mb location for window sizes 1 Mb, starting at 0 or 1 Mb location for window sizes 2 Mb, and starting at 0 or 1.5 Mb location for window sizes 3 Mb) for each chromosome were chosen to partly counteract the chance effect of different LD patterns being associated with nonoverlapping windows. Finally, adaptive window sizes based on clustering SNP by LD r^2 were also determined using the BALD R package (Dehman and Neuvial 2015), using the procedure described by Dehman *et al.* (2015).

The relative performance of all methods and models were based on receiver operating characteristic (ROC) curves. In an ROC curve, the true positive rate is plotted against the false positive rate (FPR) for each competing method (Metz 1978). We were more specifically interested in the partial area under the curve up until an FPR = of 5% (pAUC05) so as to not include somewhat irrelevant ROC regions with low levels of specificity (Ma *et al.* 2013). A perfect classifier would have a pAUC05 of $0.05 \times 1 = 0.05$, whereas a random classifier

Table 1 Overall mean relative (random classifier = 1) partial areas under an ROC curve up until an FPR of 5% (pAUC05) for different methods on single-SNP associations

	Methods					
	MCMC-SSVS	MCMC-BayesA	EMMAX	MAP-SSVS	MAP-BayesA	RRBLUP
Mean pAUC05	2.61 ^{*,†}	2.52 [†]	2.77 [*]	2.69 ^{*,†}	2.76 [*]	2.73 [*]

Values not sharing the same symbol have different ($P < 0.05$) relative pAUC05. Mean estimates are based on 10 replicates per each of nine populations of a three-by-three factorial on a number (30, 100, or 300) of QTL and shape parameter (0.18, 1.48, or 3.00) for gamma distribution of QTL effects.

would have a pAUC05 of $0.05^2/2 = 0.00125$. We subsequently rescaled all pAUC05 measures by 0.00125^{-1} such that a random classifier is rescaled to a relative pAUC05 = 1. We used the R package ROCR (Sing *et al.* 2005) to obtain replicate-specific ROC curves and pAUC05 for each of the 10 replicated data sets for each method and window specification within each n_{qtl} and γ combination. For each window specification, specific comparisons between methods were based on using the logarithm of pAUC05 as the response variable in a mixed-model ANOVA with methods, n_{qtl} , and γ and all of their interactions included as fixed effects, and population replicate (nested within n_{qtl} and γ) as a random effect blocking factor. For windows-based inferences based on fixed-window sizes, replicate-specific pAUC05 values were averaged over the two different starting positions, as previously noted. Mean $\log(\text{pAUC05})$ estimates were back-transformed (*i.e.*, antilogged) to the original scale for reporting. Overall marginal means were separated using Tukey's test, whereas comparisons between methods were sliced out using ANOVA *t*-tests for each value of n_{qtl} or γ if the corresponding interaction between these factors with methods were significant ($P < 0.05$). We also conjecture that window size might actually influence the power of detecting QTL using the same method; therefore, we conducted separate tests comparing pAUC05 for each of the different window sizes, including adaptively chosen windows based on BALD, separately within each method.

MSUPRP data: We also compared all models and algorithms on 13-week, 10th-rib back fat (mm) within the MSUPRP data as per Gualdrón Duarte *et al.* (2014). Sex, contemporary group, and age of slaughter were treated as fixed effects (*i.e.*, β). We compared each of the six competing methods, computing either PPA or *P*-values in the same manner as in the simulation study. For MCMC-BayesA and MCMC-SSVS, we ran a total of 1 million MCMC iterations based on 500,000 burn-in iterations and 500,000 iterations post-burn-in, saving every 10 iterations such that posterior inference was based on 50,000 random draws from the posterior distribution. Since we did not know the true positions of the causal QTL for this trait, GWA inferences were compared between the various methods, based on PPA for MCMC-BayesA and MCMC-SSVS, *P*-values for RRBLUP, MAP-BayesA, and MAP-SSVS, and Bonferroni adjusted *P*-values for EMMAX. Note that no adjustments for multiple testing were invoked for *P*-values determined using the shrinkage-based procedures (RRBLUP, MAP-BayesA, and MAP-SSVS) as per Gelman *et al.* (2012),

whereas a Bonferroni adjustment based on the number of markers, or number of genomic windows for windows-based analyses, was invoked for EMMAX.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results

Simulation study

Overall mean comparisons between methods for pAUC05 based on single SNP inferences are provided in Table 1, noting that two-way interactions were not detected ($P > 0.05$) between methods with γ or with n_{qtl} . There was no evidence of a sizable difference between any of the methods given that pAUC05 ranged from 2.52 to 2.77 times that for a random classifier, although MCMC-BayesA did rank lowest.

For fixed 1-Mb window sizes (Table 2), the two-way interactions between method and γ and between method and n_{qtl} were both significant ($P < 0.0001$). Therefore, methods were compared separately for each different value of γ and of n_{qtl} . Nevertheless, MCMC-SSVS and MCMC-BayesA had the largest pAUC05 ($P < 0.05$) for each different value of γ and of n_{qtl} as well as overall. EMMAX generally followed MCMC-SSVS and MCMC-BayesA with MAP-SSVS, MAP-BayesA, and RRBLUP being the worst-performing methods. Most notably, these latter three methods generally did worse than a random classifier (*i.e.*, pAUC05 < 1) except for MAP-SSVS at $n_{\text{qtl}} = 30$.

Table 3 highlights the comparisons between the various methods using the adaptive window sizes inferred by BALD. Here, the two-way interaction between the method and n_{qtl} was important ($P < 0.05$), whereas the two-way interaction between method and γ was not; hence, we just compared different methods within each different value of n_{qtl} . As with the 1-Mb window inferences, MCMC-SSVS and MCMC-BayesA had the highest pAUC05, followed by EMMAX within each different value of n_{qtl} , such that these same rankings were found overall as well. Again, we found that MAP-SSVS, MAP-BayesA, and RRBLUP had lower pAUC05 compared to a random classifier, except for MAP-SSVS when $n_{\text{qtl}} = 30$.

We were also interested in pAUC05 comparisons between different window length specifications. Recognizing that the interaction between method and window length was important in our joint analysis involving all simulated data

Table 2 Least squares mean relative (random classifier = 1) partial areas under an ROC curve up until an FPR of 5% (pAUC05) for different methods for associations based on genomic windows of length 1 Mb

Factors	Methods					
	MCMC-SSVS	MCMC-BayesA	EMMAX	MAP-SSVS	MAP-BayesA	RRBLUP
Shape γ						
0.18	2.82*	2.75*	1.78†	0.74‡,a	0.63‡,a	0.48‡,a
1.48	4.22*	4.16*	2.54†	0.69‡,a	0.38‡,a	0.28**,a
3	4.63*	5.01*	2.47†	0.67‡,a	0.40‡,a	0.24**,a
n_{qtl}						
30	6.81*	7.28*	3.63†	1.69‡	0.67‡,a	0.31**,a
90	3.89*	3.78*	2.14†	0.61‡,a	0.47‡,a	0.37‡,a
300	2.08*	2.08*	1.42†	0.33‡,a	0.31‡,a	0.29‡,a
Overall	3.81*	3.86*	2.23†	0.70‡,a	0.46‡,a	0.32**,a

Comparisons are made within different specifications of shape parameter (γ) for gamma distribution of QTL and number of QTL (n_{qtl}). Values not sharing the same symbol within a row have different ($P < 0.05$) relative pAUC05 within the row. Mean estimates based on 10 replicates per each of nine populations of a three-by-three factorial on a number (30, 100, or 300) of QTL and shape parameter (0.18, 1.48, or 3.00) for gamma distribution of QTL effects.

† Indicates the corresponding method is not better than a random classifier.

sets, we choose to focus on window length comparisons separately within each of MCMC-SSVS, MCMC-BayesA, and EMMAX (Table 4), given that all other methods performed worse than a random classifier with windows-based inference. For EMMAX, single SNP inferences have significantly larger pAUC05 compared to inferences based on the longer genomic windows (2 and 3 Mb), with inference based on adaptively determined windows using BALD and shorter genomic windows (0.5 and 1 Mb) being intermediate in their performance. Conversely, for both MCMC-BayesA and MCMC-SSVS, single-SNP inference had the lowest pAUC05, whereas adaptively determined window selection based on BALD yielded the highest pAUC05 with fixed-window inferences being intermediate in their performance. In fact, the best overall performance was based on using the two MCMC-based methods with adaptively determined windows with a pAUC05 being over five times greater than that of a random classifier.

MSUPRP data

Manhattan plots based on single SNP associations for 13-week, 10th-rib back fat (mm) in MSUPRP are provided in Figure 2. The statistically most-significant marker identified by EMMAX was SNP label ALGA0104402 ($P = 2.36 \times 10^{-10}$) at location 136.0844 Mb in chromosome 6, marking the same location identified as being most significantly associated with this trait by Gualdrón Duarte *et al.* (2014). Another 11 nearby statistically significant markers ranged in location from 132.60 to 138.24 Mb, with one marker (SNP label MARC0035827) at 122.36 Mb on chromosome 6 being also statistically significant using EMMAX. For MCMC-SSVS, the marker (SNP label ALGA0122657) located at 136.0786 Mb on chromosome 6 had the highest PPA of 0.487 and was adjacent to the most significant marker ALGA0104402 as identified by EMMAX. MCMC-SSVS also inferred its second largest PPA = 0.227 with SNP marker ALGA0104402. Hence, the top two SNP markers identified by MCMC-SSVS and EMMAX were the same, albeit that their order of importance was reversed. Although the most significant single-SNP asso-

ciations were also determined within this same region for each of the four other methods, their levels of significance were clearly not important except perhaps for MAP-SSVS, which started to approach statistical significance with SNP label MARC0035827 ($P = 0.08$).

For windows-based inference, we focused on the adaptively chosen window strategy based on LD using BALD (Figure 3). For EMMAX, the most significant window ($P = 9.36 \times 10^{-8}$) ranged from 134.17 to 134.75 Mb on chromosome 6. Although this region did not contain any markers that were statistically significant based on single-SNP-based inferences, it was very close to a marker (SNP label ASGA0029653) at 134.14 Mb that was deemed to be statistically significant in Figure 2. Four other windows on chromosome 6 were also significant, covering regions 129.70–131.35 Mb, 132.87–134.14 Mb, 135.19–136.84 Mb, and 136.97–137.32 Mb. These windows included some statistically significant or nearly significant markers based on single SNP inferences in Figure 2. Using MCMC-SSVS, the most significant window (window 909) covered 135.19–136.84 Mb with a PPA = 0.722; this window also contained the most significant markers based on single-SNP inferences using EMMAX and MCMC-SSVS in Figure 2. Window 905 had the second highest PPA = 0.477 and ranged in location from 132.87–134.14 Mb with all other windows having smaller PPA (< 0.2). An LD heatmap of the genomic region containing both windows is provided in Figure 4, indicating that some SNP markers in window 905 are in relatively high LD with markers in window 909. These two windows also had the highest PPA under MCMC-BayesA being 0.459 and 0.553, respectively. For RRBLUP, MAP-BayesA, and MAP-SSVS, no window was deemed to be statistically significant ($P > 0.05$).

Discussion

The objectives of our study were multifaceted in that we wished to very broadly address the impact of (a) prior specifications on marker effects, (b) single marker associations vs. associations based on different specifications for genomic

Table 3 Least squares mean relative (random classifier = 1) partial areas under a ROC curve up until an FPR of 5% (pAUC05) for different methods for associations based on genomic windows adaptively chosen by the BALD software package

n_{qtl}	Methods					
	MCMC-SSVS	MCMC-BayesA	EMMAX	MAP-SSVS	MAP-BayesA	RRBLUP
30	8.83*	9.03*	3.57†	1.76‡	0.87%, ^a	0.29**, ^a
90	5.34*	4.98*	2.13†	0.8‡, ^a	0.50%, ^a	0.38%, ^a
300	3.89*	3.17*	1.30†	0.66‡, ^a	0.62‡, ^a	0.58‡, ^a
Overall	5.68*	5.22*	2.15†	0.97‡, ^a	0.65%, ^a	0.40**, ^a

Comparisons are made within different specifications of number of QTL (n_{qtl}). Values not sharing the same symbol within a row have different ($P < 0.05$) relative pAUC05 within the row. Mean estimates based on 10 replicates per each of nine populations of a three-by-three factorial on a number (30, 100, or 300) of QTL and shape parameter (0.18, 1.48, or 3.00) for gamma distribution of QTL effects.

^a Indicates the corresponding method is not better than a random classifier.

windows, and (c) of computationally tractable but analytical approximations for GWA inference based on sparse priors. Although our simulation study was based on genotypes derived from a specific population (MSUPRP), a wide variety of potential genetic architectures were constructed on top of that framework based on different degrees of skewness of a gamma distribution via alternative specifications of the shape parameter (γ) for QTL effects as well as different numbers (n_{qtl}) of QTL.

Most GWA studies have been conducted using single-SNP inferences (Goddard and Hayes 2009; Visscher *et al.* 2012; Goddard *et al.* 2016). In this specific context, we determined that the difference in pAUC05 between all methods was relatively small and unimportant even though MCMC-BayesA had significantly lower pAUC05 and hence worse GWA performance. However, for all windows-based analyses, MCMC-BayesA and MCMC-SSVS had significantly greater pAUC05 than all other methods across all combinations of γ and n_{qtl} , regardless of window size and whether these window sizes were fixed or adaptively inferred based on LD using the BALD software package. Conceptually, MCMC-BayesA might have even outperformed MCMC-SSVS for windows-based GWA as our comparisons may have been influenced by the arbitrariness of using 1% as a threshold for percentage of total genetic variance explained by a window when determining the PPA under MCMC-BayesA. That is, proper specification of such a threshold is likely to be density dependent. Admittedly, a BayesB like implementation (Meuwissen *et al.* 2001) could have captured the best features (*i.e.*, variable selection and heavy-tailed priors) of both BayesA and SSVS. EMMAX typically ranked third, whereas MAP implementations of BayesA and SSVS as well as RRBLUP did much more poorly for windows-based association. The latter was not too surprising since previously Gualdrón Duarte *et al.* (2014) also determined that RRBLUP was extremely conservative for GWA in this same data set. Furthermore, this liability of RRBLUP has been noted by others including Hayes (2013). We noted that the median and mean lengths for windows adaptively chosen by BALD software were 0.59 and 0.91 Mb (Figure S1A in File S2), respectively, such that it was reasonable to expect adaptively chosen windows to lead to a GWA performance closest to inferences either based on the 0.5 or 1 Mb fixed-window sizes, as we observed for the two MCMC-based procedures.

What was initially surprising to us was that the pAUC05 for the analytical shrinkage-based procedures, namely RRBLUP, MAP-SSVS, and MAP-BayesA, under windows-based inference, was often worse than that of a random classifier (*i.e.*, pAUC05 < 1). This, at first, seemed counterintuitive to us. Hence, we briefly investigated a scenario where the number of SNP markers per window was fixed to be 10 rather than basing window sizes on a fixed physical distance. Basing genomic windows on a fixed number of SNPs has been a strategy also considered elsewhere (Zhang *et al.* 2016). In our particular case, the average length of a 10-SNP window was 0.51 Mb, such that one might anticipate that inference based on 10-SNP marker windows might be comparable to using inference based on fixed 0.5-Mb-length windows. Nevertheless, we determined that 10-SNP, windows-based inference led to a ROC performance that was at least as good as a random classifier for each of RRBLUP, MAP-SSVS, and MAP-BayesA (Figure S2 in File S2), conversely to what we observed previously to windows based on any fixed physical distance. This contrast in pAUC05 performance between fixed physical distance and fixed number of markers could be explained as follows. For the vast majority of windows based on either scenario (fixed number of markers or fixed physical distance), the P -values for the chi-square tests of these shrinkage-based procedures were very large (*i.e.*, $P > 0.85$). With inference based on a fixed number of SNP markers per window and random assignment of QTL to these markers, it was reasonable to expect that the pAUC05 of any of these procedures should be at least as large as a random classifier. However, with inference based on fixed physical distance in Mb or even adaptively determined based on LD relationships, the number of SNP markers and hence the degrees of freedom for each window-specific chi-square test was highly variable, ranging from 1 to 35 with 0.5 Mb windows, for example. Hence, regions with few markers are more likely to have smaller P -values than regions with many markers by nature of a greater penalty incurred with a larger degrees of freedom chi-square test statistic. Furthermore, lower P -value regions with fewer markers are also less likely to contain a QTL because of random assignment of QTL to markers throughout the genome, such that regions with the smallest P -values would more likely include a greater-than-expected number of false positive results relative to a random classifier.

Table 4 Least squares mean relative (random classifier = 1) partial areas under a ROC curve up until an FPR of 5% (pAUC05) between different window sizes within each of EMMAX, MCMC-BayesA, and MCMC-SSVS

EMMAX		MCMC-BayesA		MCMC-SSVS	
Window	pAUC05	Window	pAUC05	Window	pAUC05
Single SNP	2.76*	Single SNP	2.52‡	Single SNP	2.61‡
0.5 Mb	2.43*,†	0.5 Mb	3.77†	0.5 Mb	3.65†
1 Mb	2.23†,‡	1 Mb	3.86†	1 Mb	3.81†
2 Mb	1.95‡	2 Mb	3.93†	2 Mb	3.94†
3 Mb	1.85‡	3 Mb	3.93†	3 Mb	4.04†
Adaptive	2.15†,‡	Adaptive	5.22*	Adaptive	5.67*

Values not sharing the same symbol within a column have different ($P < 0.05$) relative pAUC05 within the column. Mean estimates based on 10 replicates per each of nine populations of three-by-three-by-five factorials on a number (30, 100, or 300) of QTL, shape parameter (0.18, 1.48, or 3.00) for gamma distribution of QTL effects, and genomic region size (single SNP, 0.5, 1, 2, 3 Mb, or adaptively determined) for GWA.

One possible strategy to mitigate this problem is through the use of a likelihood ratio test for the variance component, characterizing the variance attributable to markers within a window can be considered for EMMAX or the MAP-based approaches as then the degrees of freedom for that test do not depend on the number of markers in that window (Wu *et al.* 2010; Wang *et al.* 2013). Gualdrón Duarte *et al.* (2014) present details for such a likelihood ratio test; nevertheless, this approach requires one to refit the entire model each time a particular window is being tested and hence can be computationally challenging.

We specifically determined that adaptive-window specifications based on BALD worked best for both MCMC-BayesA and MCMC-SSVS with significantly higher mean pAUC05 than inferences based on fixed-window lengths or single-SNP markers. In fact, there was no evidence of differences in pAUC05 between GWA associations based on windows of constant sizes ranging from 0.5 to 3 Mb when using either MCMC-BayesA or MCMC-SSVS. Hence, adaptive-window clustering based on LD measures seems to be an important factor to consider when partitioning genomic windows, at least for Bayesian sparse prior specifications.

We have previously established that starting values are important for MAP-SSVS and MAP-BayesA (Chen and Tempelman 2015); in fact, we then demonstrated that starting marker effects at null values were very suboptimal, even though that is a common strategy for genomic-prediction methods based on the use of the EM algorithm (Meuwissen *et al.* 2009; Kärkkäinen and Sillanpää 2012). As we adapted in this study, a practical strategy is to base starting values on RRBLUP and genomic REML, although we worried as to how suboptimal that might be, recognizing MAP estimates are asymptotic, *i.e.*, $\text{MAP}(\mathbf{g}) \rightarrow E(\mathbf{g}|\mathbf{y})$ only as $n \rightarrow \infty$ and such that $n \gg m$. To further assess whether starting values based on RRBLUP and genomic REML estimates might lead to suboptimal GWA inferences, we also based starting values for MAP-SSVS and MAP-BayesA on posterior mean estimates derived from their MCMC counterparts, focusing only, however,

on single-SNP and adaptive-window inference. We recognize that this would not be a practical MAP strategy as once MCMC-based inferences are obtained, then asymptotic MAP-based inferences would not have any extra value. As anticipated from our previous genomic-prediction work (Chen and Tempelman 2015), using MCMC-based starting values for MAP-SSVS leads to a larger pAUC05 compared to the use of RRBLUP or genomic REML starting values, except for no evidence of a difference at $n_{\text{qtl}} = 300$ (Table S5 in File S2). However, for adaptively determined windows, even MAP-SSVS inferences based on MCMC-based starting values were no better than a random classifier, except for when $n_{\text{qtl}} = 30$. Similar results for comparing different sets of starting values (MCMC-BayesA vs. BLUP) for MAP-BayesA are provided in Table S6 in File S2. These supplementary results further illustrate how precarious is the use of MAP-based procedures for Bayesian-regression GWA analyses; again, we would believe the sensitivity of MAP to starting values would only be greater with the use of high density marker panels.

As our GWA inference for MCMC-SSVS was based on PPA [*i.e.*, $\text{Prob}(\tau_j = 1|\mathbf{y})$], it might seem reasonable to specify GWA inference for MAP-SSVS in a similar manner; *i.e.*, using the E-step values of τ_j at convergence as estimates of PPA. However, we noted that these E-step values uniformly drifted either toward 0 or 1 such that there were never any intermediate estimates of PPA. A comparison of PPA based on τ_j for $\text{Prob}(\tau_j = 1|\mathbf{y})$ for MCMC-SSVS vs. the E-step values of τ_j at convergence on the MSUPRP data are provided in Figure S3A in File S2. Also, recall that the MAP procedure is sensitive to starting values and that starting values for MAP-SSVS were based on RRBLUP, as this might be a pragmatic and reasonable strategy in most cases. If we had based starting values on, say, their MCMC-SSVS posterior means, one would notice a different assortment of converged E-step values of τ_j compared to what we observed with RRBLUP starting values as we demonstrate with the MSUPRP data in Figure S3B in File S2.

Recall that for MAP-SSVS, we based starting values for the SNP-specific PPA on estimated lFDRs using the R package *ashr* since there is presumably a close relationship between them, *i.e.*, $\text{PPA} \approx 1 - \text{lFDR}$ (Stephens 2017). This procedure converts *P*-values to lFDRs such that we based lFDR determinations from the *P*-values computed under EMMAX. This begged the question as to whether PPA could be simply based on lFDR processing of EMMAX *P*-values. However, upon comparing $1 - \text{lFDR}$ estimated from the EMMAX *P*-values to PPA estimated using MCMC-SSVS of the MSUPRP data, it appeared that there was not generally very good agreement between the two sets of PPA estimates, except for the some near-zero PPA and the largest PPA estimated using both procedures (Figure S4 in File S2).

We also wondered if the strategy for computing window-based PPA could be simplified further from that presented in Fernando *et al.* (2017) and used in this article (*i.e.*, Equation 19) to that suggested by Moser *et al.* (2015) who simply summed SNP-specific PPA (*i.e.*, based on Equation 15) within

Figure 2 Manhattan plots for single-SNP analysis on 13th-week, 10th-rib back fat in Duroc–Pietrain F_2 cross ($n = 922$ pigs) based on different methods: (A) EMMAX, (B) MCMC-SSVS, (C) MCMC-BayesA, (D) RRBLUP, (E) MAP-BayesA, and (F) MAP-SSVS. The green horizontal lines indicate the threshold for declaring statistical significance for various methods including a Bonferroni adjusted Type I error rate of 5% for Panel (A) and a non-adjusted Type I error rate of 5% for Panels (C), (D), (E), and (F).

a window to determine the window-based PPA. One should anticipate that the approach of Moser *et al.* (2015) should lead to a higher estimated PPA. We compared the two PPA determination approaches for pAUC05 in the simulation study and noted that there was significant interaction between the PPA determination approach with γ and n_{qtl} but no significant interaction involving window size; hence we compared the two strategies within each value of γ and n_{qtl} averaged across window length (Table S7 in File S2). The only significant difference in pAUC05 occurred with $\gamma =$

3 and $n_{qtl} = 300$, for which the approach of Fernando *et al.* (2017) led to a higher pAUC05. Nevertheless, since point estimates of pAUC05 were always larger using the approach from Fernando *et al.* (2017) we would recommend their approach from Equation 19 for the determination of windows-based PPA. An excellent analytical discussion on control of false positives in GWA using PPA is further provided in Fernando *et al.* (2017).

We did not estimate ν_r using either the procedures outlined in Yang *et al.* (2015) for MCMC-BayesA or provided in

Figure 3 Manhattan plots for genomic window-based associations on 13th-week, 10th-rib back fat in Duroc–Pietrain F_2 cross ($n = 922$ pigs) based on different methods: (A) EMMAX, (B) MCMC-SSVS, (C) MCMC-BayesA, (D) RRBLUP, (E) MAP-BayesA, and (F) MAP-SSVS under adaptive-window inference. The green horizontal lines indicate the threshold for declaring statistical significance for various methods including a Bonferroni adjusted Type I error rate of 5% for Panel (A) and a non-adjusted Type I error rate of 5% for Panels (C), (D), (E), and (F).

Chen and Tempelman (2015) for MAP-BayesA, primarily because of the extremely poor MCMC mixing for sampling this hyperparameter and its poor convergence in MAP-BayesA. A typical specification for v_τ in BayesA is four or five (Colombani *et al.* 2013; Perez and de los Campos 2014). The specification of $v_\tau = 2.5$ that we chose for this article was based in part on results from Yang *et al.* (2015) and Nadaf *et al.* (2012) who determined that lower specifications of v_g (i.e., heavier tails) could lead to higher genomic-prediction accuracies when using BayesA. To assess this further,

we compared MCMC-BayesA using $v_\tau = 2.5$ vs. $v_\tau = 5$ for pAUC05 based on the BALD-derived adaptive-window inference. In general, the use of $v_\tau = 2.5$ yielded a higher mean pAUC05 than $v_\tau = 5$ except for a nonsignificant difference at $n_{\text{qtl}} = 300$ (Table S4 in File S2). For large-scale empirical analyses whereby hyperparameter inference seems daunting, researchers should consider conducting analyses based on a finite number of hyperparameter specifications, choosing those specifications that lead to the best cross-validation prediction accuracy. Similar arguments could be made for

Figure 4 LD (r^2 metric) heatmap for genomic region containing windows 905–909 on chromosome 6 as adaptively determined by BALD software. Blue dots are starting and ending points for window 905, whereas purple dots are starting and ending points for window 909. Black dots are the three markers at 133.9292, 136.0786, and 136.0844 Mb that are top three SNPs by MCMC-SSVS. The blue oval is used to highlight a pocket of higher r^2 measured SNP markers in window 905 and 909.

choosing the key hyperparameters in other Bayesian regression models. It is worth noting that even though we ran our MCMC algorithm for 1 million iterations, the mixing of the MCMC chain was still rather poor as it pertained to inference on other hyperparameters. For example, for MCMC-BayesA, the effective sample size (ESS) for σ_g^2 was estimated to be 66.33, whereas for SSVS the ESS was 61.03 for σ_g^2 and 53.48 for π_τ .

It should be apparent that given that MCMC-SSVS is a natural variable selection model, it might be favored over MCMC-BayesA, which is not a natural variable selection model. Our strategy for computing the proportion of genetic variance explained by each window and determining the posterior probability that that percentage exceeds an arbitrary threshold (1% in our analyses) is based on the strategy presented by Fernando and Garrick (2013). The flexibility of MCMC modeling allows posterior probabilities (*i.e.*, PPA) of this nature to be computed. However, one should be wary of the impact of the threshold since it obviously should depend upon marker density. That is, if the threshold is set too high, then sensitivity is lost. Based on the results from both the simulation study and real data analysis, we demonstrated that random-effects modeling can also be a powerful tool for GWA as long as the suitable priors, *i.e.*, in our case sparser priors, are used. Other variable selection implementations popularized in WGP including BayesB (Meuwissen *et al.* 2001) or BayesR (Erbe *et al.* 2012; Moser *et al.* 2015) could be considered as well.

Our MSUPRP application was interesting in that we discovered that SNP markers in two different blocks can be in high LD even when they are not adjacent to each other. However, we would quickly note that these strange LD patterns may be due to genome assembly errors in the pig genome (Groenen 2016) with particular issues having been identified in the chromosome 6 region (Warr *et al.* 2015), which contained the strongest associations in our study. This may somewhat complicate strategies for single SNP-specific or even window-based inference. We also recognize that there is a movement toward the use of multi-SNP haplotype modeling which may improve GWA performance (Cuyabano *et al.* 2014). Our adaptive window-based strategy seems to improve the performance of GWA relative to single SNP or fixed-window length inference, although, conceivably, there may be other better ways to group SNPs. With marker densities well beyond 50K, the adaptive-window strategy might not be viable since it requires the computation and storage of a matrix of LD r^2 values between every SNP marker within a chromosome before clustering analyses can be used to partition the genome into windows. Fernando *et al.* (2017) also suggested that PPA based on Bayesian GWA analyses similar to our MCMC-SSVS be based on whether nonzero associations were found, not only in that marker's resident window but also in either of the two flanking windows. Their strategy was based on fixed-window sizes such that it may be worthwhile to consider their flanking strategy in the context of adaptively chosen window sizes. We conjecture that if LD structure is appropriately used to partition the genome, the use of such flanking windows might not be necessary; however, this should be a topic for future research. It is also important to note that the comparisons in this article are context specific in terms of the genomic LD relationships germane to an F_2 cross in pigs. This cross naturally leads to a higher pairwise LD between adjacent SNP markers than what might be found in outbreeding populations, and most notably humans. Different LD patterns would naturally change the relative comparisons between single-SNP vs. windows-based inferences as well as the relative number and sizes of adaptively chosen windows based on LD relationships. Hence future investigation of our approaches in other populations is strongly warranted.

In summary, we found Bayesian variable selection to be a promising strategy for GWA when combined with window-based inference. Nevertheless, it seems prudent that window selection be carefully chosen using rules based on LD information rather than predetermined constant physical window lengths (in Mb) for genomic regions. Also, recently proposed analytical approaches for Bayesian regression models should be discouraged for GWA studies.

Acknowledgments

We are indebted to the group of Cathy Ernst for making this data available. We are also grateful to Jose-Luis Gualdrón Duarte and Yeni Bernal Rubio for their assistance in preparing

the data. This project was supported by Agriculture and Food Research Initiative Competitive grants no. 2011-77015-30338 and 2004-35604-14580 from the United States Department of Agriculture National Institute of Food and Agriculture.

Literature Cited

- Andrews, D. F., and C. L. Mallows, 1974 Scale mixtures of normal distributions. *J. R. Stat. Soc. B* 36: 99–102.
- Bello, N. M., J. P. Steibel, and R. J. Tempelman, 2010 Hierarchical Bayesian modeling of random and residual variance-covariance matrices in bivariate mixed effects models. *Biom. J.* 52: 297–313.
- Bernal Rubio, Y. L., J. L. Gualdrón Duarte, R. O. Bates, C. W. Ernst, D. Nonneman *et al.*, 2016 Meta-analysis of genome-wide association from genomic prediction models. *Anim. Genet.* 47: 36–48.
- Calus, M. P., J. Vandenplas, and J. Ten Napel, 2015 Ever-growing data sets pose (new) challenges to genomic prediction models. *J. Anim. Breed. Genet.* 132: 407–408.
- Chen, C., and R. J. Tempelman, 2015 An integrated approach to empirical Bayesian whole genome prediction modeling. *J. Agric. Biol. Environ. Stat.* 20: 491–511.
- Colombani, C., A. Legarra, S. Fritz, F. Guillaume, P. Croiseau *et al.*, 2013 Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCpi methods for genomic selection in French Holstein and Montbeliarde breeds. *J. Dairy Sci.* 96: 575–591.
- Cuyabano, B. C., G. Su, and M. S. Lund, 2014 Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* 15: 1171.
- de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193: 327–345.
- Dehman, A., and P. Neuvial, 2015 BALD: Blockwise approach using linkage disequilibrium information. R package version 0.2.1. <http://www.math-evry.cnrs.fr/logiciels/bald>.
- Dehman, A., C. Ambroise, and P. Neuvial, 2015 Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics* 16: 148.
- Edwards, D. B., C. W. Ernst, N. E. Raney, M. E. Doumit, M. D. Hoge *et al.*, 2008 Quantitative trait locus mapping in an F2 Duroc x Pietrain resource population: II. Carcass and meat quality traits. *J. Anim. Sci.* 86: 254–266.
- Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95: 4114–4129.
- Fan, B., S. K. Onteru, Z. Q. Du, D. J. Garrick, K. J. Stalder *et al.*, 2011 Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. *PLoS One* 6: e14726.
- Fernando, R., and D. Garrick, 2013 Bayesian methods applied to GWAS, pp. 237–274 in *Genome-Wide Association Studies and Genomic Prediction*, edited by C. Gondro, J. van der Werf, and B. Hayes. Humana Press, New York.
- Fernando, R., A. Toosi, A. Wolc, D. Garrick, and J. Dekkers, 2017 Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. *J. Agric. Biol. Environ. Stat.* 22: 172–193.
- Gelman, A., 2006 Prior distributions for variance parameters in hierarchical models (comment on an article by Browne and Draper). *Bayesian Anal.* 1: 515–533.
- Gelman, A., J. Hill, and M. Yajima, 2012 Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* 5: 189–211.
- George, E. I., and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* 88: 881–889.
- Goddard, M. E., and B. J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10: 381–391.
- Goddard, M. E., K. E. Kemper, I. M. MacLeod, A. J. Chamberlain, and B. J. Hayes, 2016 Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. Biol. Sci.* 283: 20160569.
- Groenen, M. A., 2016 A decade of pig genome sequencing: a window on pig domestication and evolution. *Genet. Sel. Evol.* 48: 23.
- Gualdrón Duarte, J. L., R. O. Bates, C. W. Ernst, N. E. Raney, R. J. Cantet *et al.*, 2013 Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genet.* 14: 38.
- Gualdrón Duarte, J. L., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney *et al.*, 2014 Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics* 15: 246.
- Guan, Y., and M. Stephens, 2011 Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* 5: 1780–1815.
- Hayashi, T., and H. Iwata, 2010 EM algorithm for Bayesian estimation of genomic breeding values. *BMC Genet.* 11: 3.
- Hayes, B., 2013 Overview of statistical methods for genome-wide association studies (GWAS), pp. 149–169 in *Genome-Wide Association Studies and Genomic Prediction*, edited by C. Gondro, J. van der Werf, and B. Hayes. Humana Press, New York.
- Hayes, B., and M. E. Goddard, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33: 209–229.
- Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709–1723.
- Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42: 348–354.
- Kärkkäinen, H. P., and M. J. Sillanpää, 2012 Back to basics for Bayesian model building in genomic selection. *Genetics* 191: 969–987.
- Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain *et al.*, 2015 Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet. Sel. Evol.* 47: 29.
- Knürr, T., E. Läärä, and M. J. Sillanpää, 2013 Impact of prior specifications in a shrinkage-inducing Bayesian model for quantitative trait mapping and genomic prediction. *Genet. Sel. Evol.* 45: 24.
- Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson *et al.*, 2011 FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8: 833–835.
- Louis, T. A., 1982 Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. B* 44: 226–233.
- Ma, H., A. I. Bandos, H. E. Rockette, and D. Gur, 2013 On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat. Med.* 32: 3449–3458.
- Metz, C. E., 1978 Basic principles of ROC analysis. *Semin. Nucl. Med.* 8: 283–298.
- Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829.

- Meuwissen, T. H., T. R. Solberg, R. Shepherd, and J. A. Woolliams, 2009 A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.* 41: 2.
- Moser, G., S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray *et al.*, 2015 Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* 11: e1004969.
- Nadaf, J., V. Riggio, T. P. Yu, and R. Pong-Wong, 2012 Effect of the prior distribution of SNP effects on the estimation of total breeding value. *BMC Proc.* 6: S6.
- Perez, P., and G. de los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198: 483–495.
- Robinson, G. K., 1991 That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* 6: 15–32.
- Schmid, K., and Z. Yang, 2008 The trouble with sliding windows and the selective pressure in BRCA1. *PLoS One* 3: e3746.
- Searle, S. R., G. Casella, and C. E. McCulloch, 1992 *Variance Components*. Wiley, New York.
- Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer, 2005 ROCR: visualizing classifier performance in R. *Bioinformatics* 21: 3940–3941.
- Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.
- Stephens, M., 2017 False discovery rates: a new deal. *Biostatistics* 18: 275–294.
- Stephens, M., and D. J. Balding, 2009 Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* 10: 681–690.
- Stranden, I., and D. J. Garrick, 2009 Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* 92: 2971–2975.
- Sun, X., L. Qu, D. J. Garrick, J. C. Dekkers, and R. L. Fernando, 2012 A fast EM algorithm for BayesA-like prediction of genomic breeding values. *PLoS One* 7: e49157.
- Tempelman, R. J., 2015 Statistical and computational challenges in whole genome prediction and genome-wide association analyses for plant and animal breeding. *J. Agric. Biol. Environ. Stat.* 20: 442–466.
- Tizioto, P. C., J. F. Taylor, J. E. Decker, C. F. Gromboni, M. A. Mudadu *et al.*, 2015 Detection of quantitative trait loci for mineral content of Nelore longissimus dorsi muscle. *Genet. Sel. Evol.* 47: 15.
- VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414–4423.
- Verbyla, K., B. Hayes, P. Bowman, and M. Goddard, 2009 Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res.* 91: 307–311.
- Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang, 2012 Five years of GWAS discovery. *Am. J. Hum. Genet.* 90: 7–24.
- Wang, X., N. J. Morris, X. Zhu, and R. C. Elston, 2013 A variance component based multi-marker association test using family and unrelated data. *BMC Genet.* 14: 17.
- Warr, A., C. Robert, D. Hume, A. L. Archibald, N. Deeb *et al.*, 2015 Identification of low-confidence regions in the pig reference genome (Sscrofa 10.2). *Front. Genet.* 6: 338.
- Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
- Wiggans, G. R., T. A. Cooper, C. P. Van Tassell, T. S. Sonstegard, and E. B. Simpson, 2013 Technical note: characteristics and use of the Illumina BovineLD and GeneSeek genomic profiler low-density bead chips for genomic evaluation. *J. Dairy Sci.* 96: 1258–1263.
- Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan *et al.*, 2012 Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. *Anim. Genet.* 43: 87–96.
- Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan *et al.*, 2016 Mixture models detect large effect QTL better than GBLUP and result in more accurate and persistent predictions. *J. Anim. Sci. Biotechnol.* 7: 7.
- Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock *et al.*, 2010 Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* 86: 929–942.
- Yang, W., and R. J. Tempelman, 2012 A Bayesian antedependence model for whole genome prediction. *Genetics* 190: 1491–1501.
- Yang, W., C. Chen, and R. J. Tempelman, 2015 Improving the computational efficiency of fully Bayes inference and assessing the effect of misspecification of hyperparameters in whole-genome prediction models. *Genet. Sel. Evol.* 47: 13.
- Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal, 2016 Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. *Front. Genet.* 7: 151.
- Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44: 821–824.

Communicating editor: M. J. Sillanpää