GSE Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

Inference about quantitative traits under selection: a Bayesian revisitation for the post-genomic era

Daniel Gianola[1*], Rohan L. Fernando[2] and Chris C. Schön[3]

Abstract

Background: Selection schemes distort inference when estimating differences between treatments or genetic associations between traits, and may degrade prediction of outcomes, e.g., the expected performance of the progeny of an individual with a certain genotype. If input and output measurements are not collected on random samples, inferences and predictions must be biased to some degree. Our paper revisits inference in quantitative genetics when using samples stemming from some selection process. The approach used integrates the classical notion of fitness with that of missing data. Treatment is fully Bayesian, with inference and prediction dealt with, in an unified manner. While focus is on animal and plant breeding, concepts apply to natural selection as well. Examples based on real data and stylized models illustrate how selection can be accounted for in four different situations, and sometimes without success.

Results: Our flexible "soft selection" setting helps to diagnose the extent to which selection can be ignored. The clear connection between probability of missingness and the concept of fitness in stylized selection scenarios is highlighted. It is not realistic to assume that a fixed selection threshold t holds in conceptual replication, as the chance of selection depends on observed and unobserved data, and on unequal amounts of information over individuals, aspects that a "soft" selection representation addresses explicitly. There does not seem to be a general prescription to accommodate potential distortions due to selection. In structures that combine cross-sectional, longitudinal and multi-trait data such as in animal breeding, balance is the exception rather than the rule. The Bayesian approach provides an integrated answer to inference, prediction and model choice under selection that goes beyond the likelihood-based approach, where breeding values are inferred indirectly.

Conclusions: The approach used here for inference and prediction under selection may or may not yield the best possible answers. One may believe that selection has been accounted for diligently, but the central problem of whether statistical inferences are good or bad does not have an unambiguous solution. On the other hand, the quality of predictions can be gauged empirically via appropriate training-testing of competing methods.

Background

Quantitative genetics explains and describes inheritance of complex traits such as many diseases in humans and animals or agriculturally relevant targets, e.g., yield and product quality in maize or dairy cattle. It focuses on statistical quantities, e.g., allelic and haplotype frequencies, locus effect sizes, means, variances and covariances between individuals or traits. Theory-derived parameters

*Correspondence: gianola@ansci.wisc.edu

[1] Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI, USA
Full list of author information is available at the end of the article

like heritability and genetic correlations and linkage disequilibrium are of interest as well. Although they are interpretable, these parameters may not provide a meaningful mechanistic explanation of genetic systems and represent abstractions. Interaction is key in biological processes, both at the biochemical (e.g., cycles) and population levels, e.g., the shifting balance theory of evolution [1, 2]. Yet, quantitative genetics heavily relies on the additive genetic model, both in its pre- and post-genomics versions. Actually, this model is a crucial tool in the armamentarium of animal and plant breeders and also plays a role in the "polygenic scores" used for prediction in human medicine [3]. Hence, learning well the parameters of additive models is important.

All unknown quantities are inferred using finite samples of experimental or observational data. Quantitative genetic models have been informed by phenotypes and genealogies [4–7] and more recently by molecular markers [8–11]. Also, it is increasingly feasible to obtain (often expensively) joint measures of the genome, epigenome, metabolome, proteome, metagenome, behavior, robustness, resilience and sustainability from samples of individuals. Such inputs are fed to prediction and decision machines used for artificial selection, field fertilization, animal management and disease treatment algorithms. However, data are often not representative of a target population because of selection schemes in animals and plants, and expensive measurements are seldom taken at random. Likewise, in medical trials, individuals not meeting certain criteria or "culling levels" are excluded and there may be a non-random dropout of accepted patients, i.e., some abandon the study due to treatment effects. Selection schemes distort inference when estimating differences between treatments or genetic associations between traits, or degrade prediction of outcomes, e.g., the expected performance of the progeny of an individual with a certain genotype. If input and output measurements are not taken on random samples, inferences and predictions may be biased to some extent.

The preceding problem is not novel in quantitative genetics. Ideally, properties of estimators and predictors should be studied relative to a setting representing the selection or dropout process occurring, which is not an easy task. There are many potential scenarios: selection may be by truncation of a distribution, it may be disruptive with two tails of the distribution selected, or aimed at stabilizing a population near some optimum [12]. However, there are situations in which it is impossible or awkward to model the selection process in a simple manner. For instance, if individuals are heterogeneous, related through complex pedigree loops or possess unequal amounts of information, classical balanced-data selection index formulae [13–15] or stylized treatments for parameter estimation are not entirely adequate.

With the advent of genomics and "big" post-genomic data, distortions produced by selection may have been exacerbated. Best linear unbiased prediction (BLUP) evolved into GBLUP, with "G" denoting "genomic" and Bayesian linear regression models emerged as competing prediction machines [10, 16–20]. Although the data used in genome-based models is rarely random, the "selection problem" has been seldom discussed in depth from a theoretical perspective. Earlier studies [21–23] pointed out that if a process (e.g., selection) that leads to "missing data" depends on observed data only and on parameters that are "separate" from those of the statistical model employed for analysis, selection is said to be ignorable. However, when data is available on "survivors" only, the selection process and its parameters must be considered in the analysis for appropriate inference or prediction. In between, there is a plethora of scenarios.

This paper revisits inference of quantitative genetic unknowns when using samples stemming from a selection process. Our approach integrates classical notions of fitness with that of missing data. The treatment is fully Bayesian, with inference and prediction dealt with, in an unified manner. The focus is on animal and plant breeding but concepts apply to natural selection as well.

Bayesian setting

Let $\mathbf{y}$ and θ be vectors of observable variables and unknown quantities, respectively. In genetic contexts, θ may include genomic and epigenomic site effects, genetic and environmental components of (co) variance, nuisance location parameters, latent quantities such as breeding values and yet to be observed (future) phenotypes. One seeks to learn θ from the observed $\mathbf{y}$. In a Bayesian treatment all elements of θ are viewed as randomly varying, reflecting aleatory or causal uncertainty [24] and are assigned some prior distribution with density $p(\theta|Hyp, M)$, where Hyp denotes known hyperparameters under some model M, e.g., a linear regression with a certain linear structure and distributional assumptions; model M', say, may assume a thick-tailed residual distribution, while M may have a non-linear part and treat the residuals as Gaussian.

Without natural or artificial selection, the joint density of $\mathbf{y}$ and θ under model M is:

$$p(\mathbf{y}, \theta|Hyp, M) = p(\mathbf{y}|\theta, M)p(\theta|Hyp, M), \quad (1)$$

where $p(\mathbf{y}|\theta,M)$ is the density of the data-generating distribution under M, with θ fixed. The posterior density of θ is

$$p(\theta|\mathbf{y},Hyp,M) = \frac{p(\mathbf{y}|\theta,M)p(\theta|Hyp,M)}{\int p(\mathbf{y}|\theta,M)p(\theta|Hyp,M)d\theta} = \frac{p(\mathbf{y}|\theta,M)p(\theta|Hyp,M)}{p(\mathbf{y}|Hyp,M)}. \quad (2)$$

The denominator $p(\mathbf{y}|Hyp,M)$ is the Bayesian marginal density of the data, that is, the reciprocal of the integration constant of the posterior density under M. The latter depends on $\mathbf{y}$ and on M, but not on θ, which has been averaged out using $p(\theta|Hyp,M)$ as weight function [25]. Often, the target of the analysis is a component of θ, e.g., the vector of breeding values. Partitioning the entire parameter vector into θ_i and θ_{-i}, where θ_{-i} is θ without θ_i, the marginal posterior density of sub-vector θ_i is $p(\theta_i|\mathbf{y},Hyp,M) = \int p(\theta|\mathbf{y},Hyp,M)d\theta_{-i}$ [25]. Bayesian computations are typically done via Monte Carlo sampling procedures.

Fitness depending on observed data

Various formulations of fitness functions appear in [12, 26–29]. Suppose natural or artificial selection operate on phenotypes through a fitness function $H(\mathbf{y}|\varphi,\theta)$, where φ is a parameter vector that may be distinct from θ and that does not enter into the process generating observed data $\mathbf{y}$. The fitness function may depend on phenotypes linearly or non-linearly and is proportional to the probability that an individual possessing some observed attributes will reproduce or survive.

Let $p(\theta,\varphi|Hyp,M)$ be the joint prior density of all parameters. The post-selection joint density of $\mathbf{y}$, θ and φ is:

$$p_s(\mathbf{y},\theta,\varphi|Hyp,M) = \frac{H(\mathbf{y}|\varphi,\theta)p(\mathbf{y}|\theta,M)p(\theta,\varphi|Hyp,M)}{\iiint H(\mathbf{y}|\varphi,\theta)p(\mathbf{y}|\theta,M)p(\theta,\varphi|Hyp,M)d\mathbf{y}d\theta d\varphi} = \frac{H(\mathbf{y}|\varphi,\theta)p(\mathbf{y}|\theta,M)p(\theta,\varphi|Hyp,M)}{\iiint H(\mathbf{y}|\varphi,\theta)p(\mathbf{y},\varphi,\theta|Hyp,M)d\mathbf{y}d\theta d\varphi} = \frac{H(\mathbf{y}|\varphi,\theta)p(\mathbf{y}|\theta,M)p(\theta,\varphi|Hyp,M)}{\overline{H}} \quad (3)$$

Here, $\overline{H}$ is "Bayesian mean fitness"; it does not involve $\mathbf{y}$, θ and φ since these vectors have been averaged out, so $\overline{H}^{-1}$ acts as an integration constant. Note that $\frac{H(\mathbf{y}|\varphi,\theta)}{\overline{H}}$ is the relative fitness associated with $\mathbf{y}$ at fixed φ,θ; values conferring lower fitness are assigned less density following selection. Without selection, $H(\mathbf{y}|\varphi,\theta)$ is constant and $H(\mathbf{y}|\varphi,\theta) = \overline{H}$ for all $\mathbf{y}$.

To illustrate, suppose that there is a single known parameter μ and that phenotypes are distributed independently as $N(\mu,\sigma^2)$. Suppose that selection is such that only phenotypes under a threshold t are observed and that sample size is N; here, $\varphi = t$. The joint density of the observations, after selection, is:

$$p_s(\mathbf{y}|t,\mu,\sigma^2) = \frac{p(\mathbf{y}|\mu,\sigma^2)}{\int_{-\infty}^{t} p(\mathbf{y}|\mu,\sigma^2)d\mathbf{y}} = \frac{p(\mathbf{y}|\mu,\sigma^2)}{\prod_{i=1}^{N}\Pr(z_i < t)} = \frac{p(\mathbf{y}|t,\mu,\sigma^2)}{\overline{H}(t,\mu,\sigma^2)}, \quad (4)$$

where $z_i = \frac{t_i-\mu}{\sigma}$ and $\overline{H}$ is the mean fitness, given μ and σ^2. If a prior distribution F were assigned to μ, Bayesian mean fitness would be $E_F\left[\overline{H}(t,\mu,\sigma^2)\right] = \overline{H}(t,\sigma^2)$, where the outer expectation indicates that the values of μ are averaged out using distribution F as mixing process. See [23] for an example of a discrete fitness function applied to the "cow setting" of [30], where fitness is equivalent to the probability of observing a certain pattern.

Therefore, the joint posterior density under selection is:

$$\begin{aligned} p_s(\theta,\varphi|\mathbf{y},Hyp,M) &\propto p_s(\mathbf{y},\theta,\varphi|Hyp,M) \\ &\propto H(\mathbf{y}|\varphi,\theta)p(\mathbf{y}|\theta,M)p(\theta,\varphi|Hyp,M) \\ &\propto H(\mathbf{y}|\varphi,\theta)p(\mathbf{y}|\theta,M)p(\varphi|\theta,M)p(\theta|Hyp,M), \end{aligned} \quad (5)$$

where $p(\varphi|\theta,M)$ is the conditional (given θ) prior density of φ. If φ and θ are independent a priori and if $H(\mathbf{y}|\varphi,\theta) = H(\mathbf{y}|\varphi)$, i.e., fitness does not depend on θ, one has

$$p_s(\theta,\varphi|\mathbf{y},Hyp,M) \propto [H(\mathbf{y}|\varphi)p(\varphi)][p(\mathbf{y}|\theta)p(\theta|Hyp,M)], \quad (6)$$

implying that θ and φ are also independent a posteriori. Therefore,

$$p_s(\theta|\mathbf{y},Hyp,M) = p(\theta|\mathbf{y},Hyp,M), \quad (7)$$

the posterior density of θ without selection. Thus, if selection is based on a fitness function (linear or non-linear) $H(\mathbf{y}|\varphi)$ of the observed data that does not depend on parameters θ, and if θ and φ are a priori independent, the selection process is ignorable from a Bayesian

perspective. The posterior distributions before and after selection are exactly the same, in agreement with [22, 31, 32].

The preceding argument is at the root of the oft-made claim that BLUP remains "unbiased if the history of the selection process is represented in the data used", that is, if all records on which selection decisions were based are included in the analysis. The claim is not correct, at least from the frequentist perspective. Suppose $\mathbf{y} = \mathbf{g} + \mathbf{e}$ is an n-dimensional vector in R^n, where $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G})$ is a vector of genetic effects and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$ is a vector of environmental deviates, with $\mathbf{g}$ independent of $\mathbf{e}$. Selection is such that $\mathbf{y} \in S$ and S is the sampling space constrained by selection. If $\mathbf{G}$ and $\mathbf{R}$ are known, the posterior distribution of the genetic effects is $\mathbf{g}|\mathbf{G}, \mathbf{R}, \mathbf{y} \sim N\left(\widehat{\mathbf{g}}, \left[\mathbf{R}^{-1} + \mathbf{G}^{-1}\right]^{-1}\right)$, where the posterior expectation is $\widehat{\mathbf{g}} = \mathbf{G}[\mathbf{R} + \mathbf{G}]^{-1}\mathbf{y}$ [31]. Following Eq. (7), the posterior distribution is unaffected by selection for any $\mathbf{y}$. Now, $\widehat{\mathbf{g}}$ is also a BLUP in a frequentist context and, in the absence of selection ($\mathbf{y} \in R^n$), $E(\widehat{\mathbf{g}}) = \mathbf{G}\left[\mathbf{R}^{-1} + \mathbf{G}^{-1}\right]^{-1} E(\mathbf{y}) = 0$; hence, $E(\widehat{\mathbf{g}}) = E(\mathbf{g})$ and the predictor is unbiased in the frequentist Hendersonian sense. Furthermore, $Var(\widehat{\mathbf{g}}) = \mathbf{G}[\mathbf{R} + \mathbf{G}]^{-1}\mathbf{G}$. However, if $\mathbf{y} \in S$,

$$E_s(\widehat{\mathbf{g}}) = \mathbf{G}\left[\mathbf{R}^{-1} + \mathbf{G}^{-1}\right]^{-1} E_s(\mathbf{y}) \tag{8}$$

and

$$Var_s(\widehat{\mathbf{g}}) = \mathbf{G}\left[\mathbf{R}^{-1} + \mathbf{G}^{-1}\right]^{-1} Var_s(\mathbf{y}) \mathbf{G}\left[\mathbf{R}^{-1} + \mathbf{G}^{-1}\right]^{-1}, \tag{9}$$

since selection modifies the distribution of $\mathbf{y}$. Clearly, the distribution $\widehat{\mathbf{g}} \sim N\left(\mathbf{0}, \mathbf{G}[\mathbf{R} + \mathbf{G}]^{-1}\mathbf{G}\right)$ is modified by the selection process. While the Bayesian treatment allows ignoring selection, a frequentist analysis requires finding the sampling distribution after selection. BLUP would be unbiased by selection only if it could be shown that $E_s(\widehat{\mathbf{g}}) = E_s(\mathbf{g})$.

In a Bayesian treatment, the predictive distribution under selection of yet to be observed phenotypes $\mathbf{y}_f$ is

$$p_s(\mathbf{y}_f|\mathbf{y}, Hyp, M) = \int p_s(\mathbf{y}_f|\mathbf{y}, \theta, M) p_s(\theta|\mathbf{y}, Hyp, M) d\theta. \tag{10}$$

If the process of generating future observations is unaltered by selection, $p_s(\mathbf{y}_f|\mathbf{y}, \theta, Hyp, M) = p(\mathbf{y}_f|\mathbf{y}, \theta, Hyp, M)$. Since the posterior distribution is unaffected by selection based on observed data, (10) can be re-written as:

$$\begin{aligned} p_s(\mathbf{y}_f|\mathbf{y}, Hyp, M) &= \int p(\mathbf{y}_f|\mathbf{y}, \theta, M) p(\theta|\mathbf{y}, Hyp, M) d\theta \\ &= p(\mathbf{y}_f|\mathbf{y}, Hyp, M). \end{aligned} \tag{11}$$

Therefore, the predictive distribution is also unaltered by selection based on observed data. Prediction of future phenotypes can be carried out as if selection had not occurred. Hence, inferences from posterior or predictive probabilities are unaffected by this type of selection.

The same holds true for the posterior distribution of the model when treated as uncertain. Let there be K mutually exclusive and exhaustive competing models $M_1, M_2, \ldots, M_K$ with prior probabilities $P_1, P_2, \ldots, P_K$, and parameters $\theta_1, \theta_2, \ldots, \theta_K$, respectively. Under selection, the posterior probability assigned to model k is:

$$P_s(M_k|\mathbf{y}, Hyp) = \frac{p_s(\mathbf{y}|M_k, Hyp) P_k}{\sum_{k=1}^{K} p_s(\mathbf{y}|M_k, Hyp) P_k}; k = 1, 2, \ldots, K. \tag{12}$$

Above, $p_s(\mathbf{y}|M_k, Hyp)$ is the marginal distribution of the data under model k with selection. Furthermore,

$$p_s(\mathbf{y}|M_k, Hyp) = \int p_s(\mathbf{y}|\theta_k, M) p_s(\theta_k|Hyp, M_k) d\theta_k. \tag{13}$$

Using the reasoning employed for the predictive distribution, if $p_s(\mathbf{y}|\theta_k, M) = p(\mathbf{y}|\theta_k, M)$, then

$$p_s(\mathbf{y}|M_k, Hyp) = \int p(\mathbf{y}|\theta_k, M) p(\theta_k|Hyp, M) d\theta_k = p(\mathbf{y}|M_k, Hyp). \tag{14}$$

Using Eq. (14) in Eq. (12)

$$\begin{aligned} P_s(M_k|\mathbf{y}, Hyp) &= \frac{p(\mathbf{y}|M_k, Hyp) P_k}{\sum_{k=1}^{K} p(\mathbf{y}|M_k, Hyp) P_k} \\ &= P(M_k|\mathbf{y}, Hyp); k = 1, 2, \ldots, K. \end{aligned} \tag{15}$$

Therefore, the posterior distribution of the "model random variable" is also unaffected by selection and model choice can be carried out ignoring the selection process.

Fitness based on observed and missing data

Inference under selection in quantitative genetics was formalized by Im et al. [22] by adapting missing data theory [21, 33] to maximum likelihood estimation of genetic parameters. In a study of the trajectory of genetic variance over time in a population undergoing selection, Sorensen et al [23] used such theory from a Bayesian perspective. To motivate this, let traits A and B be

measured in n individuals, respectively, but 25% lack phenotypes for trait B. If the records in this portion are missing at random (e.g., it is expensive to record B and a random choice is made), there is no selection. What about if there is a non-random basis for the pattern of observed data? A connection between missingness and fitness is made below from a Bayesian perspective. The dependence on model M will be suppressed in the notation hereinafter.

A classical example dating back to 1959 [30], where selection is based on observed data only, serves to introduce the notion of missing data. Suppose two dairy cows have a first lactation record; only one of the cows will be allowed to produce a second lactation. Without selection, each of the cows would have milk records for each of the two lactations and the "complete" data would be $\mathbf{y} = (y_{11}, y_{21}, y_{12}, y_{22})'$, where i denotes cow and j record number. If $y_{11} > y_{21}$, the observed data would be $\mathbf{y}_{obs} = (y_{11}, y_{21}, y_{12})'$; if $y_{11} \leq y_{21}$ then $\mathbf{y}_{obs} = (y_{11}, y_{21}, y_{22})'$. A vector $\mathbf{r}$ describing the pattern of "missingness" is part of the information on the problem. The two patterns are $\mathbf{r} = (1, 1, 1, 0)'$ and $\mathbf{r} = (1, 1, 0, 1)'$, respectively, where "1" is observed and "0" is missing. The complete data vector of phenotypes is $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{miss})'$ where $\mathbf{y}_{miss}$ includes the records that would have been observed if selection had not taken place.

Under selection, the joint density of all data (complete data and missingness pattern) and of the parameters is:

$$p_s(\mathbf{y}, \mathbf{r}, \theta, \varphi|Hyp) = \Pr(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{miss}, \theta, \varphi) p(\mathbf{y}_{obs}, \mathbf{y}_{miss}|\theta) p(\theta, \varphi|Hyp), \tag{16}$$

where φ are parameters of the missing data (selection) process and $\Pr(\mathbf{r}|\mathbf{y}, \theta, \varphi)$ is the conditional probability of observing the pattern. $\Pr(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{miss}, \theta, \varphi)$ is equivalent to the fitness function $H(\mathbf{y}|\varphi, \theta)$ employed in Eq. (3) above and gives probabilities of "survival" ("death"), conditionally on phenotypes, observed and unobserved, and model parameters. Integrating with respect to $\mathbf{y}_{miss}$

$$\begin{aligned} p_s(\mathbf{y}_{obs}, \mathbf{r}, \theta, \varphi|Hyp) &= \int \Pr(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{miss}, \theta, \varphi) p(\mathbf{y}_{obs}, \mathbf{y}_{miss}|\theta) p(\theta, \varphi|Hyp) d\mathbf{y}_{miss} \\ &= \int \Pr(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{miss}, \theta, \varphi) p(\mathbf{y}_{miss}|\mathbf{y}_{obs}, \theta) p(\mathbf{y}_{obs}|\theta) p(\theta, \varphi|Hyp) d\mathbf{y}_{miss}. \end{aligned} \tag{17}$$

A rearrangment of the preceding equation leads to:

$$\begin{aligned} &p_s(\mathbf{y}_{obs}, \mathbf{r}, \theta, \varphi|Hyp) \\ &\quad \propto p(\mathbf{y}_{obs}|\theta) p(\theta, \varphi|Hyp) \\ &\quad \left[\int \Pr(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{miss}, \theta, \varphi) p(\mathbf{y}_{miss}|\mathbf{y}_{obs}, \theta) d\mathbf{y}_{miss}\right]. \end{aligned} \tag{18}$$

The term in brackets is the expected fitness after averaging over the conditional distribution of the missing data, given $\mathbf{y}_{obs}$, and θ, which we will denote as $H(\mathbf{y}_{obs}, \varphi, \theta)$. Since $p(\theta, \varphi|Hyp) = p(\theta|Hyp)p(\varphi|\theta)$ it follows that the posterior density under selection is:

$$\begin{aligned} p_s(\theta, \varphi|\mathbf{y}_{obs}, \mathbf{r}, Hyp) &\propto p(\mathbf{y}_{obs}|\theta) p(\theta|Hyp) H(\mathbf{y}_{obs}, \varphi, \theta) p(\varphi|\theta) \\ &\propto p(\theta|\mathbf{y}_{obs}, Hyp) H(\mathbf{y}_{obs}, \varphi, \theta) p(\varphi|\theta). \end{aligned} \tag{19}$$

The preceding equation indicates that, in general, the posterior density under selection differs from the posterior density without selection. Correct Bayesian inference needs to take into account the selection process by formulating a selection model (i.e., defining a fitness function), as well as prior knowledge of φ , deterministic or probabilistic. A selection model or $H(\mathbf{y}_{obs}, \varphi, \theta)$, must be specified, and a prior distribution of φ elicited or the value of this parameter specified *ex ante*.

A special case is when the fitness function does not involve missing data and parameters θ, so $\Pr(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{miss}, \theta, \varphi) = \Pr(\mathbf{r}|\mathbf{y}_{obs}, \varphi)$. Here, the integral in (18) produces:

$$\begin{aligned} H(\mathbf{y}_{obs}, \varphi, \theta) &= \int \Pr(\mathbf{r}|\mathbf{y}_{obs}, \varphi) p(\mathbf{y}_{miss}|\mathbf{y}_{obs}, \theta) d\mathbf{y}_{miss} \\ &= \Pr(\mathbf{r}|\mathbf{y}_{obs}, \varphi), \end{aligned} \tag{20}$$

since $p(\mathbf{y}_{miss}|\mathbf{y}_{obs}, \theta)$ integrates to 1. Then, Eq. (19) becomes:

$$p_s(\theta, \varphi|\mathbf{y}_{obs}, \mathbf{r}, Hyp) \propto p(\theta|\mathbf{y}_{obs}, Hyp) \Pr(\mathbf{r}|\mathbf{y}_{obs}, \varphi) p(\varphi|\theta). \tag{21}$$

Furthermore, if φ and θ are independent a priori $p(\varphi|\theta) = p(\varphi)$, then:

$$\begin{aligned} p_s(\theta, \varphi|\mathbf{y}_{obs}, \mathbf{r}, Hyp) &\propto p(\theta|\mathbf{y}_{obs}, Hyp) \Pr(\mathbf{r}|\mathbf{y}_{obs}, \varphi) p(\varphi) \\ &\propto p(\theta|\mathbf{y}_{obs}, Hyp) p(\varphi|\mathbf{y}_{obs}, \mathbf{r}), \end{aligned} \tag{22}$$

where

$$p(\varphi|\mathbf{y}_{obs}, \mathbf{r}) = \frac{\Pr(\mathbf{r}|\mathbf{y}_{obs}, \varphi) p(\varphi)}{\int \Pr(\mathbf{r}|\mathbf{y}_{obs}, \varphi) p(\varphi) d\varphi} \tag{23}$$

is the posterior density of fitness function parameter φ. Expression (22) implies that θ and φ are also independent a posteriori. Hence, $p_s(\theta|\mathbf{y}_{obs}, \mathbf{r}, Hyp) = p(\theta|\mathbf{y}_{obs}, Hyp)$, so Bayesian inference about θ can be carried out as if selection had not taken place, irrespective of the pattern of missingness created by selection on $\mathbf{y}_{obs}$. In other words: if the conditional probability of observing a phenotype ("fitness") depends on observed data but not on missing data and on θ, and if parameters θ and φ are independent *a priori* (note that Im et al. [22] used the term "distinct" in their likelihood-based treatment), selection can be ignored. The two conditions, however, represent strong assumptions. Equation (19) indicates that even if θ and φ are assigned independent prior distributions, selection cannot be ignored in inference any time that the fitness function involves θ, i.e., if it has the form $H(\mathbf{y}_{obs}, \varphi, \theta)$ as opposed to $H(\mathbf{y}_{obs}, \varphi)$.

In spite of the strong assumptions required, most animal breeding programs employ statistical methods that ignore selection coupled with data sets that do not reflect the entire history of the selection process. Clearly, there are enormous difficulties in representing the type of selection undergoing in populations of animals. Generations overlap, animals have unequal amounts of information and of relatedness to other animals, and it is not always transparent why certain individuals are kept as parents even when recording is complete and meticulous. A cautionary view may be that inference of breeding values is always distorted (often loosely referred to as "bias") to some extent due to various factors that cannot be accounted for statistically. Viewing estimates as being free from selection bias is perhaps naïve.

Selection and partially observed data

Consider (19) placing focus on the fitness function $H(\mathbf{y}_{obs}, \varphi, \theta)$ defined in (20). The distribution $[\mathbf{y}_{miss}|\mathbf{y}_{obs}, \theta]$ follows from the statistical model assumed. For example, if the joint distribution of $\mathbf{y}_{miss}$ and $\mathbf{y}_{obs}$, given θ, is normal, the conditional distribution is normal as well, with mean vector and covariance matrix readily derived from theory. Next, one would need to assume a selection model representing the distribution of the missing data pattern ($\mathbf{r}$). Four examples are presented below to illustrate concepts, motivated by those described in [22] and adapted to a Bayesian perspective.

Example 1: selection based on records not available for analysis

Country B buys frozen semen of m out of n bulls ($m < n$) in country A. The m bulls chosen exceed a minimum threshold of "performance" t based on information provided by country A. Country B develops a breeding program using such m bulls and collects records. Only records from country B are available for analysis. The performances in the two countries are regarded as distinct traits [34], a concept that has been employed in global dairy cattle breeding.

Assuming conditional (given θ) independence between records, the data generating model for the m records observed in country B is:

$$p(\mathbf{y}_{obs}|\theta) = \prod_{i=1}^{m} p(y_{iB}|\theta), \tag{24}$$

where y_{iB} is the performance of bull i ($i = 1, 2, \ldots, m$). Following [22], the conditional probability of selection involves two binary indicator variables, r_{iA} and r_{iB}, where the values 0 and 1 mean "missing" and "observed", respectively. For $i = 1, 2, \ldots, n$, where n (number of bulls in country A),

$$\begin{aligned} \Pr(r_{iA} = 0|\mathbf{y}_{obs}, \mathbf{y}_{miss}) = 1 \text{ and } \Pr(r_{iB} \\ = 1|\mathbf{y}_{obs}, \mathbf{y}_{miss}) \\ = 1_{(t,\infty)}(y_{iA}). \end{aligned} \tag{25}$$

The value $r_{iA} = 0$ for all i means that all records from the exporting country (A) are not available ("missing"); $r_{iB} = 1$ means that records on bull i are observed in country B only if selection threshold t is exceeded by such bull in country A.

The density of all data (observed and missing) and of $\mathbf{r}$, is:

$$\begin{aligned} p(\mathbf{y}_{obs}, \mathbf{y}_{miss}, \mathbf{r}|\theta) = &\prod_{i=1}^{m} p(y_{iA}, y_{iB}|\theta) 1_{(t,\infty)}(y_{iA}) \\ &\prod_{i=m+1}^{n} p(y_{iA}, y_{iB}|\theta) 1_{(-\infty,t)}(y_{iA}). \end{aligned} \tag{26}$$

Missing data includes all n records from country A, and the $n-m$ records that would have been observed in country B but were not because the corresponding bulls did not perform over threshold t in A. Integrating Eq. (26) with respect to the missing data yields:

$$\begin{aligned} &p\left(\mathbf{y}_{obs},\mathbf{r}|\theta\right)\\ &=\left\{\prod_{i=1}^{m}\int p\left(y_{iA},y_{iB}|\theta\right)1_{(t,\infty)}\left(y_{iA}\right)dy_{iA}\right\}\left\{\prod_{i=m+1}^{n}\iint p\left(y_{iA},y_{iB}|\theta\right)1_{(-\infty,t)}\left(y_{iA}\right)dy_{iA}dy_{iB}\right\}\\ &=\prod_{i=1}^{m}p\left(y_{iB}|\theta\right)\left[\prod_{i=1}^{m}\Pr\left(y_{iA}>t|y_{iB},\theta\right)\prod_{i=m+1}^{n}\Pr\left(y_{iA}<t|\theta\right)\right]. \end{aligned} \tag{27}$$

Using observed data does not make use of the information on the selection process provided by the two terms between brackets.

Here, $\varphi=t$ (known) is the only parameter that governs the missing data process. The posterior density after accounting for selection is:

$$\begin{aligned} &p\left(\theta|\mathbf{y}_{obs},\mathbf{r},Hyp\right)\\ &\propto\prod_{i=1}^{m}p\left(y_{iB}|\theta\right)p\left(\theta|Hyp\right)\left[\prod_{i=1}^{m}\Pr\left(y_{iA}>t|y_{iB},\theta\right)\prod_{i=m+1}^{n}\Pr\left(y_{iA}<t|\theta\right)\right]\\ &\propto p\left(\theta|\mathbf{y}_{obs},Hyp\right)H\left(\mathbf{y}_{obs},t,\theta\right), \end{aligned} \tag{28}$$

where $H\left(\mathbf{y}_{obs},t,\theta\right)$ is the term in brackets.

This simple selection scheme produces an analytically intractable problem. In the absence of selection, let performances in countries A and B have the bivariate distribution:

$$\begin{bmatrix} y_{iA}\\ y_{iB}\end{bmatrix}|\mu_A,\mu_B,\Sigma\sim N\left(\begin{bmatrix}\mu_A\\ \mu_B\end{bmatrix},\begin{bmatrix}\sigma_A^2 & \rho\sigma_A\sigma_B\\ \rho\sigma_A\sigma_B & \sigma_B^2\end{bmatrix}\right);\ i=1,2,\ldots n. \tag{29}$$

Assume such vectors are mutually independent, with μ_A and μ_B being country means, ρ the correlation coefficient between performances, and σ_A^2 and σ_B^2 the variances in countries A and B, respectively. Suppose all parameters are known, save for μ_B, the mean performance in the importing country. The conditional distribution $[y_{iA}|y_{iB},\theta]$ is:

$$y_{iA}|y_{iB},\theta\sim N\left(\mu_{A.B}=\mu_A+b\left(y_{iB}-\mu_B\right),v_{A.B}=\sigma_A^2\left(1-\rho^2\right)\right); \tag{30}$$

above, $b_{=}\dfrac{\rho\sigma_A\sigma_B}{\sigma_B^2}$ is the regression of performance in A on that in B. Using (28) and assigning a flat prior to μ_B, the posterior density is:

$$\begin{aligned} p\left(\mu_B|\mathbf{y}_{obs},\mathbf{r},Hyp\right)&\propto\exp\left[-\frac{\sum_{i=1}^{m}\left(y_{iB}-\mu_B\right)^2}{2\sigma_B^2}\right]\prod_{i=1}^{m}[1-\Phi_{i,A.B}(t)]\prod_{i=m+1}^{n}\Phi_{i,A}(t)\\ &\propto\exp\left[-\frac{\sum_{i=1}^{m}\left(y_{iB}-\mu_B\right)^2}{2\sigma_B^2}\right]\prod_{i=1}^{m}[1-\Phi_{i,A.B}(t)], \end{aligned} \tag{31}$$

where $\Phi_{i,A.B}$ is the distribution function of Eq. (30), which depends on μ_B. The expression involving $\Phi_{i,A}$, the distribution function of performance in country A, is absorbed by the integration constant. If selection is ignored, the posterior distribution would be proportional to the

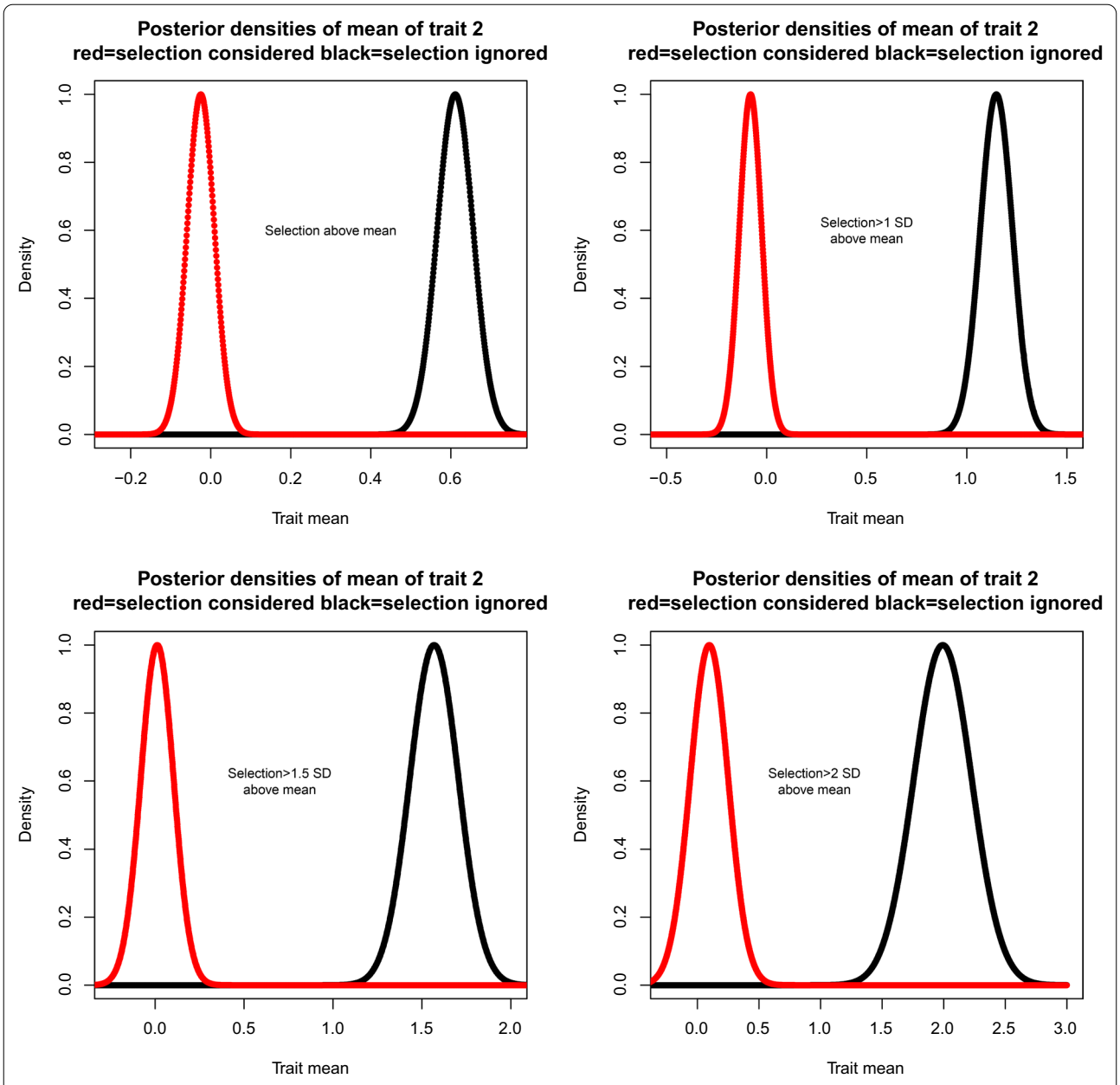

Fig. 1 Posterior distribution: selection for correlated traits at various intensities

Gaussian kernel above with mean (mode) $\widehat{\mu}_B = \dfrac{\sum_{i=1}^{m} y_{iB}}{m}$. Under selection, however, the posterior mean cannot be written in closed form and locating the mode requires an iterative procedure. If $\sigma_A^2 = \sigma_B^2 = 1$ and $\mu_A = 0$, for instance, the posterior density takes the form:

$$p(\mu_B|\mathbf{y}_{obs}, \mathbf{r}, Hyp) = \frac{\exp\left\{-\frac{1}{2}\left[m\left(\mu_B - \widehat{\mu}_B\right)^2 - 2f(\mu_B)\right]\right\}}{\int \exp\left\{-\frac{1}{2}\left[m\left(\mu_B - \widehat{\mu}_B\right)^2 - 2f(\mu_B)\right]\right\} d\mu_B} \tag{32}$$

where

$$f(\mu_B) = \sum_{i=1}^{m} \log\left[1 - \Phi_i\left(\frac{t - \rho\left(y_{iB} - \mu_B\right)}{\sqrt{1-\rho^2}}\right)\right]. \tag{33}$$

To illustrate how accounting for selection may lead to correct Bayesian inference, we simulated $n = 1000$ pairs from a bivariate standard normal distribution with $\rho = 0.8$. Selection operated on trait A by picking individuals with phenotypes that were above the mean or 1, 1.5 and 2 standard deviations over the mean. Such selection produced samples of sizes $m = 494, 155, 53$ and 18 on which performance for trait B was available. The only parameter treated as unknown was μ_B with a flat prior attached to it. If selection were ignored, the posterior distribution would be $\mu_B|y_B \sim N\left(\widehat{\mu}_B, m^{-1}\right)$; if selection is accounted for, the posterior density is as in (32). On the one hand, Fig. 1 shows that ignoring selection grossly overstated the true value of the parameter: 0. On the other hand, the true value was assigned appreciable density in the "correct" posterior distributions, irrespective of the selection intensity applied. Note that the fitness function (missing data process) employed corresponds exactly to how selection was simulated. An incorrect formulation of the selection process would have probably produced distorted inferences. The example illustrates that a proper Bayesian analysis may capture true parameter values in situations of non-ignorable selection when the latter is modeled properly.

Example 2: pre-selected samples

The example is motivated by a situation in animal and plant breeding that has taken place during the genomic era and that it produces what is called "pre-selection bias". It was studied by Patry and Ducrocq [35] in a dairy cattle setting but from a different perspective to the one employed in our paper. In New Zealand dairy cattle, Winkelman et al. [36] proposed a method of genetic evaluation that combined a Gaussian kernel that is constructed using genomic information with features of single-step BLUP methodology. The procedure did not use the notions of missing data or of fitness functions. With real records, they found that the proposed methodology delivered a smaller predictive bias and a higher predictive correlation than a previously used procedure that "blended" pedigree and genomic information. The correlation improved 1–2% for production traits, but negligibly for traits such as fertility or longevity. In a simulation study [37], produced 15 generations of selection. At that point, parents were preselected with various degrees of intensity using either parental averages, a genome-based choice, or at random. Subsequently, they estimated genomic breeding values of preselected animals with a single-step BLUP procedure that excluded all the information from pre-culled individuals. They were not able to detect bias in the evaluations of such animals. Wang et al. [38] considered the impact of genomic selection on estimates of variance components obtained from using different similarity kernels, and used simulation and real data from broilers. When genotyping was at random, estimates obtained with a single-step model did not exhibit bias in the simulated data sets; otherwise, estimates had a marked bias. The impact of such bias on estimated breeding values was noticeable. It is unclear to what extent the results from these three studies generalize to more general forms of selection, as consideration of a general prescription was not addressed. Simulations provide "local" guidance only: results may change drastically if the assumptions adopted or the structure of the data are varied. These researchers, however, seemed aligned with the view that accounting for the history of the selection process as completely as possible can attenuate the impact of selection on inference and prediction, rendering selection quasi-ignorable.

When genomic selection began to be applied [10], decisions had to be made on individuals (e.g., bulls) to be genotyped for single nucleotide polymorphisms (SNPs). Due to the high cost of SNP chips, not all candidates for selection could be genotyped; a similar situation occurs now with next-generation DNA sequences or with expensive epigenomic or metabolomic measurements. Suppose that m out of n ($m < n$) dairy bulls that possess pedigree-based estimates of breeding value are chosen for genotyping, with "genomic breeding values" estimated as if the m bulls were randomly sampled. A "pre-selection bias" is expected to accrue since the m bulls chosen may not be representative of the current population. Can the distortion in inference be tempered analytically?

For illustration, a wheat yield data set examined in several studies and available in the BGLR package was used [39–41]. The dataset spans $n = 599$ inbred lines of wheat, each genotyped for $p = 1279$ binary markers that denote presence or absence of an allele at a locus. The target phenotype was wheat grain yield in each line planted in "environment 1". The dataset also includes a pedigree-based additive relationship matrix, $\mathbf{A}$, of size 599×599 and several of the lines are completely inbred. A genomic relationship matrix [17] among lines was built as $\mathbf{G} = \mathbf{X}\mathbf{X}'/p$ where $\mathbf{X}$ was the 599×1279 matrix of marker codes (0,1) with each column centered and standardized. Using the entire dataset, i.e., without any selection practiced, genetic (pedigree or genome based) and residual variance components were estimated via maximum likelihood. The random effects models used were $\mathbf{y} = \mathbf{a} + \mathbf{e}$ and $\mathbf{y} = \mathbf{g} + \mathbf{e}^*$ in pedigree-based and genome-enabled analyses, respectively, where $\mathbf{a}$ and $\mathbf{g}$ are pedigree and genomic breeding values to be learned, with $\mathbf{e}$ and $\mathbf{e}^*$ being model residuals. We posed $\mathbf{a}|\sigma_a^2 \sim N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right)$, $\mathbf{e}|\sigma_e^2 \sim N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$ as mutually independent, and likewise for $\mathbf{g}|\sigma_g^2 \sim N\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right)$, $\mathbf{e}^*|\sigma_{e^*}^2 \sim N\left(\mathbf{0}, \mathbf{I}\sigma_{e^*}^2\right)$; the $\sigma^{2\prime}s$ are

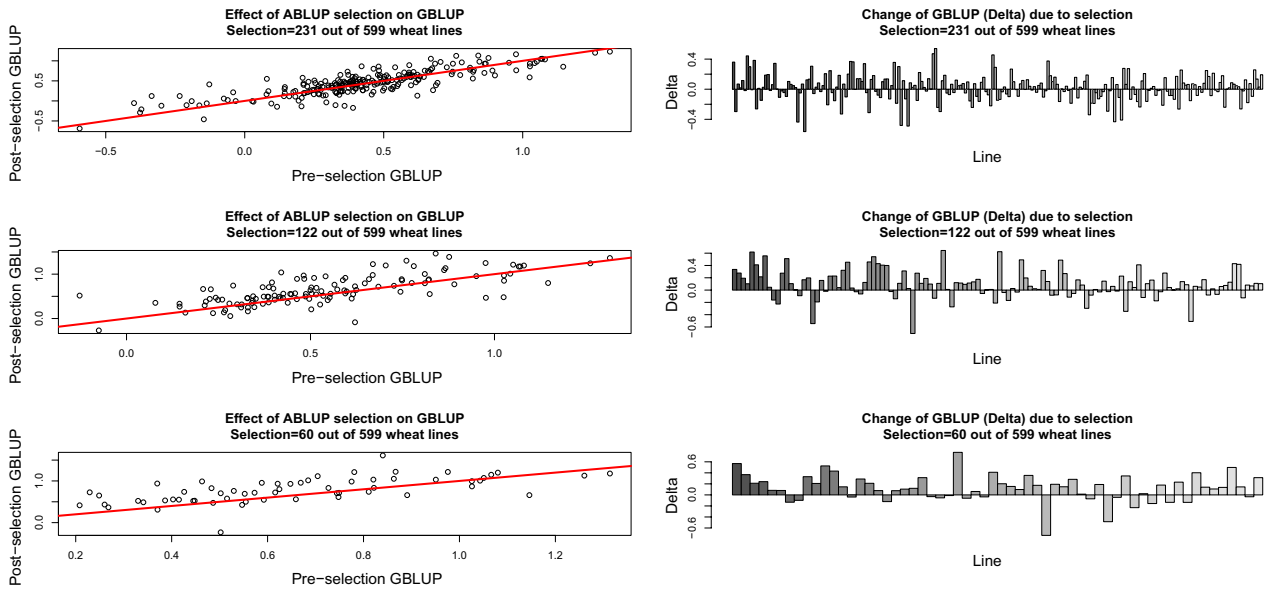

Fig. 2 Effect of selection on posterior means of pedigree (ABLUP) and genome-based (GBLUP) breeding values

variance components. Using the maximum likelihood estimates of variances as true values, best linear unbiased predictions of $\mathbf{a}$ and $\mathbf{g}$ were calculated, $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{g}}$, respectively. Under a Bayesian framework the posterior distributions of the pedigree and genomic breeding values are $\mathbf{a}|\mathbf{y},\siga^2,\sigma_e^2\sim N\left(\widehat{\mathbf{a}},\mathbf{C}_a\right)$ and $\mathbf{g}|\mathbf{y},\sigma_g^2,\sigma_{e^*}^2\sim N\left(\widehat{\mathbf{g}},\mathbf{C}_g\right)$, with variance components treated as known hyper-parameters. Here,

$$\widehat{\mathbf{a}}=\left(\mathbf{I}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1}\right)^{-1}\mathbf{y} \text{ and } \widehat{\mathbf{g}}=\left(\mathbf{I}+\frac{\sigma_{e^*}^2}{\sigma_g^2}\mathbf{G}^{-1}\right)^{-1}\mathbf{y}, \tag{34}$$

give the posterior expectations and

$$\mathbf{C}_a=\left(\mathbf{I}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1}\right)^{-1}\sigma_e^2 \text{ and } \mathbf{C}_g=\left(\mathbf{I}+\frac{\sigma_{e^*}^2}{\sigma_g^2}\mathbf{G}^{-1}\right)^{-1}\sigma_{e^*}^2, \tag{35}$$

are the posterior covariance matrices. In a frequentist setting, the posterior means correspond to $BLUP(\mathbf{g})$ and $BLUP(\mathbf{a})$, whereas $\mathbf{C}_a$ and $\mathbf{C}_g$ are interpreted as prediction error covariance matrices.

Using the 599 values in $\widehat{\mathbf{a}}$, we selected lines with pedigree breeding value estimates larger than the threshold $t=0.20$, 0.40 or 0.60, resulting in 231, 122 and 60 "top" lines, respectively. The posterior distributions of $\mathbf{g}$ were calculated before and after selection (ignoring the missing data process but using the same variance components) and compared. As depicted in Fig. 2, the analysis based on selected lines tended to overstate estimates of genomic breeding values relative to those obtained without selection. Ignoring selection introduces a selection "bias" that is impossible to evaluate because the true breeding values are unknown, except when data are simulated. Letting M_a and M_g denote pedigree and genome-based models, respectively, note that

$$E\left(\widehat{\mathbf{a}}|\mathbf{a},M_a\right)=\left(\mathbf{I}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1}\right)^{-1}E\left(\mathbf{y}|\mathbf{a}\right)=\left(\mathbf{I}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1}\right)^{-1}\mathbf{a}, \tag{36}$$

$$E\left(\widehat{\mathbf{g}}|\mathbf{g},M_g\right)=\left(\mathbf{I}+\frac{\sigma_{e^*}^2}{\sigma_g^2}\mathbf{G}^{-1}\right)^{-1}E\left(\mathbf{y}|\mathbf{g}\right)=\left(\mathbf{I}+\frac{\sigma_{e^*}^2}{\sigma_g^2}\mathbf{G}^{-1}\right)^{-1}\mathbf{g}. \tag{37}$$

Hence, both $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{g}}$ have an "epistemic" bias [24]. Such bias differs from the notion used by Henderson [42, 43], who defined "prediction unbiasedness" as $E\left(\widehat{\mathbf{a}}\right)=E(\mathbf{a})$ under M_a or $E\left(\widehat{\mathbf{g}}\right)=E\left(\mathbf{g}\right)$ under M_g, i.e., posterior means (BLUP) are unbiased for the mean of the prior distributions, but not for the estimands $\mathbf{a}$ and $\mathbf{g}$. Selection introduces an additional distortion relative to "true" breeding values, pedigree or genome-defined.

How does one account for the distortion in inference? Our representation of selection will follow the protocol employed in the example. Threshold t is the only parameter governing selection here. Let "*sel*" and "*nsel*" denote selected and unselected individuals, respectively. Following (28), the posterior density of the genomic breeding values after accounting for selection and assuming that $[\widehat{\mathbf{a}}|\mathbf{y},\mathbf{g}] = [\widehat{\mathbf{a}}|\mathbf{g}]$ (i.e., given $\mathbf{g}$, $\widehat{\mathbf{a}}$ is independent of $\mathbf{y}$), one has:

$$\begin{aligned} &p(\mathbf{g}|\mathbf{y}_{obs},\mathbf{r},Hyp) \\ &\propto \prod_{i=1}^{m} p(y_i|g_{sel,i})p(\mathbf{g}|Hyp)\left[\prod_{i=1}^{m}\Pr\left(\widehat{a}_{sel,i}>t|\mathbf{g}_{sel}\right)\prod_{i=m+1}^{n}\Pr\left(\widehat{a}_{nsel,i}<t|\mathbf{g}_{nsel}\right)\right] \\ &\propto \left[\prod_{i=1}^{m} p(y_i|g_{sel,i})p(\mathbf{g}_{sel}|Hyp)\right]\times\left[\prod_{i=1}^{m}\Pr\left(\widehat{a}_{sel,i}>t|g_{sel,i}\right)\prod_{i=m+1}^{n}\Pr\left(\widehat{a}_{nsel,i}<t|g_{nsel,i}\right)\right]p(\mathbf{g}_{nsel}|\mathbf{g}_{sel},Hyp). \end{aligned} \tag{38}$$

The first term in brackets is the posterior density of the genomic breeding values calculated from selected data only but ignoring selection. The joint fitness function [second term in brackets in (38)] assumes that the choice of an individual for genotyping is based only on whether or not t is exceeded, independently of what happens with other individuals, but conditionally on the unknown genomic breeding value of the individual in question. Integrating (38) with respect to $\mathbf{g}_{nsel}$ produces

$$\begin{aligned} &p(\mathbf{g}_{sel}|\mathbf{y}_{obs},\mathbf{r},Hyp) \\ &\propto p(\mathbf{g}_{sel}|\mathbf{y}_{obs},Hyp)\prod_{i=1}^{m}\Pr\left(\widehat{a}_{sel,i}>t|g_{sel,i}\right) \\ &\times\int\prod_{i=m+1}^{n}\Pr\left(\widehat{a}_{nsel,i}<t|g_{nsel,i}\right)p(\mathbf{g}_{nsel}|\mathbf{g}_{sel},Hyp)d\mathbf{g}_{nsel}. \end{aligned} \tag{39}$$

Since unselected individuals are not genotyped, there is no information available for writing $p(\mathbf{g}_{nsel}|\mathbf{g}_{sel},Hyp)$, which is Gaussian with mean $\mathbf{G}_{nsel,sel}\mathbf{G}^{-1}_{sel,sel}\mathbf{g}_{sel}$ and covariance matrix $\mathbf{G}_{nsel,nsel}-\mathbf{G}_{nsel,sel}\mathbf{G}^{-1}_{sel,sel}\mathbf{G}_{sel,nsel}$. Hence the integral in (39) does not convey information on $\mathbf{g}_{sel}$ and is treated as a constant. The preceding is an important matter and may adversely affect the ability of accounting for selection. Finally,

$$p(\mathbf{g}_{sel}|\mathbf{y}_{obs},\mathbf{r},Hyp)\propto\exp\left[-\frac{1}{2\sigma^2_{e^*}}(\mathbf{g}_{sel}-\widetilde{\mathbf{g}}_{sel})'\widetilde{\mathbf{C}}^{-1}_{g,sel}(\mathbf{g}_{sel}-\widetilde{\mathbf{g}}_{sel})+f_{sel}(\mathbf{g}_{sel})\right], \tag{40}$$

where,

$$\widetilde{\mathbf{g}}_{sel}=\widetilde{\mathbf{C}}_{g,sel}\mathbf{y}_{sel}, \tag{41}$$

$$\widetilde{\mathbf{C}}_{g,sel}=\left[\mathbf{I}_{sel,sel}+\frac{\sigma^2_{e^*}}{\sigma^2_g}\mathbf{G}^{-1}_{sel,sel}\right]^{-1}, \tag{42}$$

and

$$f_{sel}(\mathbf{g}_{sel})=\sum_{i=1}^{m}\log\left[\Pr\left(\widehat{a}_{sel,i}>t|g_{sel,i}\right)\right]. \tag{43}$$

We note that it might be possible to approximate the distribution $[\mathbf{g}_{nsel}|\mathbf{g}_{sel},Hyp]$ by making an imputation from pedigree information, as in single-step methods [44]; however, this is a technical matter beyond the scope of our paper.

Simplifying assumptions are required in order to proceed. A canonical case is one where individuals are independently and identically distributed. Without selection, let

$$\begin{bmatrix} g_i \\ \widehat{a}_i \end{bmatrix}\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix},\begin{bmatrix} \sigma^2_g & \rho_{g\widehat{a}}\sigma_g\sigma_{\widehat{a}} \\ \rho_{g\widehat{a}}\sigma_g\sigma_{\widehat{a}} & \sigma^2_{\widehat{a}} \end{bmatrix}\right);\ i=1,2,\ldots,n, \tag{44}$$

where $\rho_{g\widehat{a}}$ is the expected correlation between unknown genomic breeding value and pedigree-based posterior mean (BLUP here) and the σ^2 are variance parameters. In real applications $\rho_{g\widehat{a}}$ actually varies over candidates due to unequal amounts of information. In the simplest case, $\sigma_{\widehat{a}}=\sqrt{Var(h^2_a y_i)}=h^2_a\sqrt{\sigma^2_a+\sigma^2_e}$, where σ^2_a and σ^2_e pertain to the pedigree-based model and $h^2_a=\frac{\sigma^2_a}{\sigma^2_a+\sigma^2_e}$ is heritability. Then, $E(\widehat{a}_i|g_i)=\rho_{g\widehat{a}}\frac{\sigma_{\widehat{a}}}{\sigma_g}g_i$, and $Var(\widehat{a}_i|g_i)=h^4_a(\sigma^2_a+\sigma^2_e)\left(1-\rho^2_{g\widehat{a}}\right)$ for all i. Dispersion parameter estimates were $\sigma^2_a=0.2859, \sigma^2_e=0.5761, h^2_a=0.3316$ and $\sigma^2_g=0.5315$. Since there is no information on $\rho_{g\widehat{a}}$, using the unselected data we crudely estimated $\rho_{\widehat{g}\widehat{a}}$ at 0.82 and took $\rho_{g\widehat{a}}=0.75$ for the example. In order to account somehow for the fact that individuals were not identically distributed, the

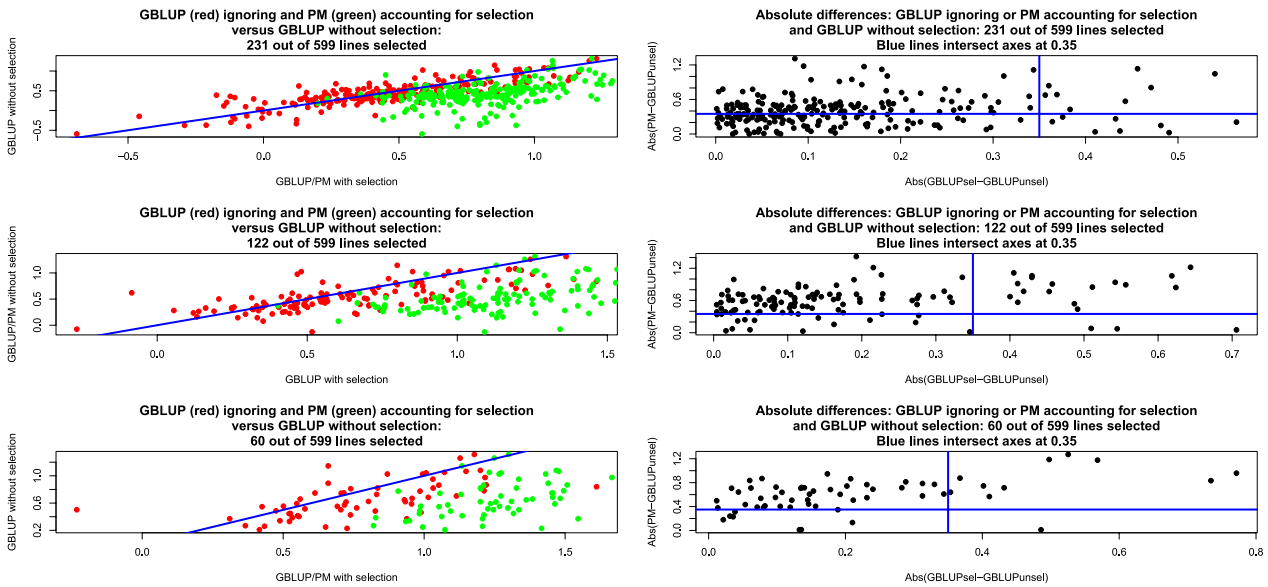

Fig. 3 Posterior means without, ignoring or accounting for selection

following modifications of the previous formulae were made using BLUP theory: $\sigma_{\widehat{a}} \rightarrow \sqrt{\sigma^2_{\widehat{a}_i}}$; $\sigma_g \rightarrow \sqrt{\sigma^2_{g_i}}$ and $Var(\widehat{a}_i|g_i) = \sigma^2_{\widehat{a}_i}\left(1 - \dfrac{\sigma^2_{\widehat{a}_i}}{\sigma^2_{g_i}}\right)$ for $i = 1, 2, \ldots, n$. Here, for example, $\rho_{g_i\widehat{a}_i}$ is the correlation value specific to individual i and σ_{g_i} is the square root of the appropriate diagonal element of $\mathbf{G}\sigma^2_g$. The posterior density of genomic breeding values after selection was therefore:

$$p(\mathbf{g}_{sel}|\mathbf{y}_{obs}, \mathbf{r}, Hyp) \propto \exp\left\{-\frac{1}{2\sigma^2_{e^*}}(\mathbf{g}_{sel} - \widetilde{\mathbf{g}}_{sel})'\widetilde{\mathbf{C}}^{-1}_{g,sel}(\mathbf{g}_{sel} - \widetilde{\mathbf{g}}_{sel}) + \sum_{i=1}^{m}\log[1 - \Phi(z_i)]\right\}, \quad (45)$$

where

$$z_i = \frac{t - \rho_{g_i\widehat{a}_i}\dfrac{\sigma_{\widehat{a}_i}}{\sigma_{g_i}}g_i}{\sqrt{\sigma^2_{\widehat{a}_i}\left(1 - \dfrac{\sigma^2_{\widehat{a}_i}}{\sigma^2_{g_i}}\right)}}. \quad (46)$$

$$f_{sel}(\mathbf{g}_{sel}) = \sum_{i=1}^{m}\log\left[1 - \Phi(z_i)\right]. \quad (47)$$

The posterior distribution cannot be recognized and Markov chain Monte Carlo sampling may be considered for inference of $\mathbf{g}_{sel}$. A candidate-generating distribution in a Metropolis scheme [25, 45] could be

$$\mathbf{g}^*_{sel}|\mathbf{y}, \sigma^2_a, \sigma^2_{e^*} \sim N\left(\widetilde{\mathbf{a}}_{sel}, \left(\mathbf{I}_{sel} + \frac{\sigma^2_e}{\sigma^2_a}\mathbf{A}^{-1}_{sel,sel}\right)^{-1}\sigma^2_e\right), \quad (48)$$

where $\widetilde{\mathbf{a}}_{sel} = \left(\mathbf{I}_{sel} + \frac{\sigma^2_e}{\sigma^2_a}\mathbf{A}^{-1}_{sel,sel}\right)^{-1}\mathbf{y}_{sel}$. If $\mathbf{g}^*_{prop}$ is a draw from the proposal distribution, the probability of moving from state $\mathbf{g}^{now}$ to $\mathbf{g}^*_{prop}$ is $\alpha = \min\left[R\left(\mathbf{g}^{now}, \mathbf{g}^*_{prop}\right), 1\right]$, where

$$R\left(\mathbf{g}^{now}, \mathbf{g}^*_{prop}\right) = \exp\left[-\frac{1}{2\sigma^2_{e^*}}Q\left(\mathbf{g}^{now}, \mathbf{g}^*_{prop}\right) + \Delta_{sel}\left(\mathbf{g}^{now}, \mathbf{g}^*_{prop}\right)\right], \quad (49)$$

and

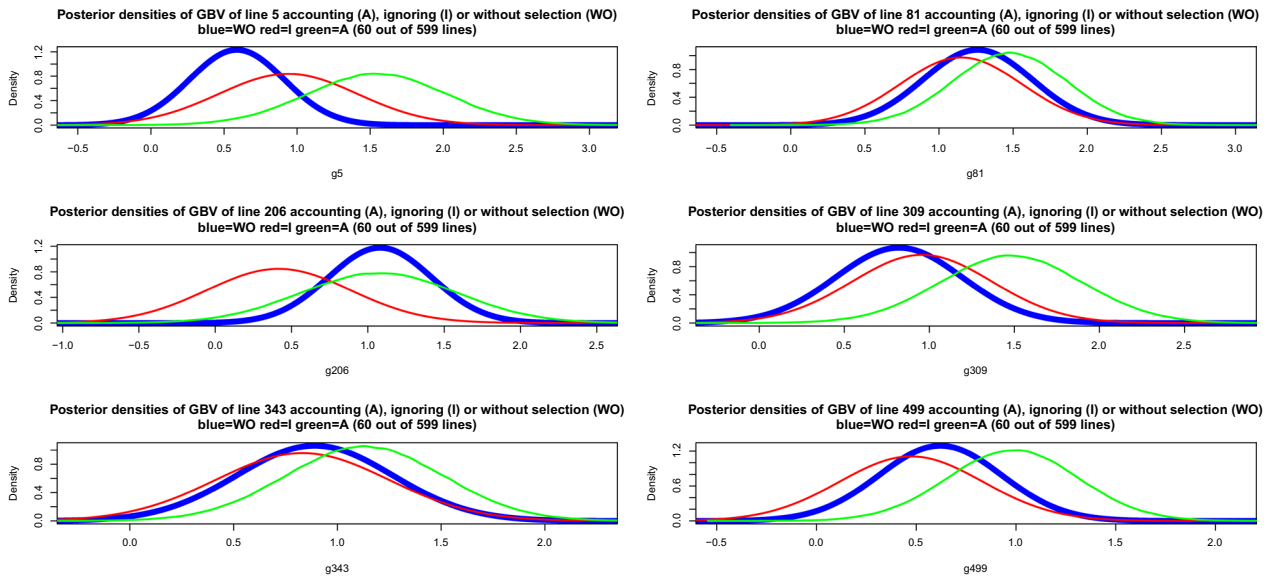

Fig. 4 Posterior densities of six lines without, ignoring or accounting for selection

$$Q\left(\mathbf{g}^{now}, \mathbf{g}^{*}_{prop}\right) = \left(\mathbf{g}^{*}_{prop} - \widetilde{\mathbf{g}}_{sel}\right)' \widetilde{\mathbf{C}}^{-1}_{g,sel}\left(\mathbf{g}^{*}_{prop} - \widetilde{\mathbf{g}}_{sel}\right) - \left(\mathbf{g}^{now} - \widetilde{\mathbf{g}}_{sel}\right)' \widetilde{\mathbf{C}}^{-1}_{g,sel}\left(\mathbf{g}^{now} - \widetilde{\mathbf{g}}_{sel}\right), \tag{50}$$

$$\Delta_{sel}\left(\mathbf{g}^{now}, \mathbf{g}^{*}_{prop}\right) = \sum_{i=1}^{m} \log \frac{1 - \Phi(z_i^{now})}{1 - \Phi\left(z_i^{prop}\right)}. \tag{51}$$

The next state in the chain is given by the rule (U is an uniform random deviate):

$$\mathbf{g}^{new} = \begin{cases} \mathbf{g}^{*}_{prop} \text{ if } U \leq R\left(\mathbf{g}^{now}, \mathbf{g}^{*}_{prop}\right) \\ \mathbf{g}^{now} \text{ otherwise} \end{cases}. \tag{52}$$

A Metropolis sampler with Eq. (48) as a proposal-generating process was used to estimate the posterior distribution having a density as given in Eq. (45) in scenarios where genotypes were available only for individuals whose $\widehat{\mathbf{a}}$ values exceeded 0.2,0.4 and 0.6, producing 231, 122 and 60 selected lines, respectively. The posterior distribution of genomic breeding values of the 599 lines prior to selection was $\mathbf{g}|\mathbf{y}, \sigma_g^2, \sigma_{e^*}^2 \sim N\left(\widehat{\mathbf{g}}, \mathbf{C}_g\right)$, as presented earlier. We also computed posterior distributions of the genomic breeding values ignoring the selection process from the data in selected lines. Metropolis sampling was done by running three chains (one per selection intensity) of length 100,000 each. After diagnostic tests [45, 46], a conservative burn-in period of 20,000 samples was adopted. Subsequently, a single long-chain of 480,000 iterations was run, with 380,000 samples used for inference, per setting. Figure 3 (left panels) depicts scatter plots of posterior means of genomic breeding values in the absence of selection (GBLUP without selection, y-axis) versus either GBLUP ignoring selection or posterior means accounting for selection (x-axis). Ignoring selection tended to overstate the estimates of genomic breeding values calculated without selection and from a larger sample ($n = 599$ versus $n = 231, 122, 60$ in the selection schemes). Accounting for selection produced estimated genomic breeding values (posterior means denoted as PM in the plots) that were even further away from those calculated without selection. The three panels at the right of Fig. 3 show larger differences between posterior means without selection (GBLUPunsel) and with selection accounted for (PM in the y-axis) than between GBLUPunsel and GBLUPsel, i.e., with the selection process ignored. Our way of accounting for selection produced larger absolute errors (taking GBLUPunsel as reference) than when selection was ignored, i.e., GBLUPsel. Posterior densities of the genomic breeding values of lines 5, 81, 206, 309, 343 and 499 are presented in Fig. 4; the data with selection pertain to the setting were 60 lines had been selected out of the 599 candidates. Densities ignoring selection (in red) tended to match better those obtained without selection (in blue) than the densities obtained with a selection model incorporated into

inference (green). However, there was much uncertainty within each of the settings, leading to overlap, although the "green" density function was centered further right along the x-axis than the "blue" or "red" densities.

The following messages can be extracted from the example. First, paradoxically, our attempt at accounting for selection seemed to distort inference on the selected lines beyond what was obtained by ignoring selection: there was a noticeable overstatement of estimated breeding values. Second, it was not easy to account for selection even when the process leading to missing data was known. For instance, part of the information on selection had to be ignored because of the inability of writing the conditional distribution of genomic breeding values of unselected individuals, given those of the selected ones. That may be what "single-step" methods (e.g., [44]) implicitly do by including ungenotyped (but pedigreed) individuals in the analysis. Third, we employed variance parameters estimated prior to selection. This action was chosen because of the impossibility of obtaining sufficiently precise estimates given the small sizes of the selected samples of lines. Finally, at least in small samples it is always difficult to disentangle the impact of non-randomness from that of noise, in view of the uncertainty remaining after analysis, irrespective of whether selection is ignored or modelled explicitly.

Example 3: multiple-trait sequential selection

A multiple-trait sequential selection scenario is described using two traits as an illustration, for simplicity. There is a set of S candidates (e.g., lines) with phenotypes $y_{11}, y_{12}, \ldots, y_{1S}$ available for trait 1, where the first subscript denotes the trait. A subset S^+ of the candidates is chosen to be measured for a second trait; the complementary subset S^- contains candidates that have phenotypes for trait 1 but not for trait 2. The dataset presented for analysis contains only the pairs $\{y_{1i}, y_{2i}\}$ in S^+; pairs in S^- are missing. Here, $\mathbf{y}_{obs} = \left(\mathbf{y}'_{1S^+}, \mathbf{y}'_{2S^+}\right)'$ and $\mathbf{y}_{miss} = \left(\mathbf{y}'_{1S^-}, \mathbf{y}'_{2S^-}\right)'$.

To define the distribution of the vector $\mathbf{r}$ describing the missing data pattern, an assumption about the selection process must be made. It will be assumed that a candidate is measured for trait 2 if its phenotype for trait 1 is larger than some $\varphi = y_{1,\min}$, a "minimum threshold of performance" for trait 1. Following [22], the conditional probability of selection ("fitness") for candidate i is:

$$\Pr\left(r_i = 1|\mathbf{y}_{obs}, \mathbf{y}_{miss}, \theta, \varphi\right) = I_{(\varphi = y_{1,\min}, \infty)}\left(y_{1i}\right); \; i = 1, 2, \ldots, S, \tag{53}$$

where θ are the unknown model parameters; $I_{(y_{1,\min}, \infty)}\left(y_{1i}\right) = 1$ if $y_{1i} > y_{1,\min}$ and 0 if $y_{1i} \leq y_{1,\min}$. If, given θ, pairs $\{y_{1i}, y_{2i}\}$ are mutually independent over individuals, the complete dataset and $\mathbf{r}$ have the joint density:

$$\begin{aligned} &p\left(\mathbf{y}_{obs}, \mathbf{y}_{miss}, \mathbf{r}|\varphi, \theta\right) \\ &= \prod_{i \epsilon S^+} p\left(y_{1i}, y_{2i}|\theta\right) \prod_{i \epsilon S^-} p\left(y_{1i}, y_{2i}|\theta\right) I_{(-\infty, \varphi)}\left(y_{1i}\right) \\ &= \prod_{i \epsilon S^+} p\left(y_{1i}, y_{2i}|\theta\right) \prod_{i \epsilon S^-} p\left(y_{2i}|y_{1i}, \theta\right) p\left(y_{1i}|\theta\right) I_{(-\infty, \varphi)}\left(y_{1i}\right). \end{aligned} \tag{54}$$

Integrating out the missing data, i.e., $\{y_{1i}, y_{2i}\}$ in S^-, yields:

$$\begin{aligned} &p\left(\mathbf{y}_{obs}, \mathbf{r}|\varphi, \theta\right) \\ &= \prod_{i \epsilon S^+} p\left(y_{1i}, y_{2i}|\theta\right) \prod_{i \epsilon S^-} \int \left[\int p\left(y_{2i}|y_{1i}, \theta\right) dy_{2i}\right] I_{(-\infty, \varphi)}\left(y_{1i}\right) p\left(y_{1i}|\theta\right) dy_{1i} \\ &= \prod_{i \epsilon S^+} p\left(y_{1i}, y_{2i}|\theta\right) \prod_{i \epsilon S^-} \Pr\left(y_{1i} < \varphi|\theta\right). \end{aligned} \tag{55}$$

The posterior density is therefore:

$$\begin{aligned} &p\left(\theta|\mathbf{y}_{obs}, \mathbf{r}, Hyp\right) \\ &\propto \prod_{i \epsilon S^+} p\left(y_{1i}, y_{2i}|\theta\right) p\left(\theta|Hyp\right) \prod_{i \epsilon S^-} \Pr\left(y_{1i} < \varphi|\theta\right) \\ &= \frac{p\left(\theta|\mathbf{y}_{obs}, Hyp\right) \exp\left\{\sum_{i \epsilon S^-} \log\left[\Pr\left(y_{1i} < \varphi|\theta\right)\right]\right\}}{\int p\left(\theta|\mathbf{y}_{obs}, Hyp\right) \exp\left\{\sum_{i \epsilon S^-} \log\left[\Pr\left(y_{1i} < \varphi|\theta\right)\right]\right\} d\theta}. \end{aligned} \tag{56}$$

The wheat dataset was employed again to give a numerical illustration. The 599 lines have records in each of four distinct environments. To represent a scenario where selection does not occur, we fitted a four-variate linear model with the performances of the lines in different environments treated as distinct traits [34]. The model was:

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_3\\ \mathbf{y}_4\end{pmatrix}=\begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\\ \mathbf{g}_3\\ \mathbf{g}_4\end{pmatrix}+\begin{pmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{pmatrix}, \tag{57}$$

where $\mathbf{y}_i$ $(i=1,2,3,4)$ is a 599×1 vector of phenotypes for trait i, the $\mathbf{g}_i's$ are genomic breeding values for the trait (marked with the 1279 markers) and δ_i is a model trait-specific residual vector. Prior assumptions were:

$$\begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\\ \mathbf{g}_3\\ \mathbf{g}_4\end{pmatrix}|\mathbf{G}_0\sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix},\mathbf{G}_0\otimes\mathbf{G}\right), \tag{58}$$

$$\begin{pmatrix}\mathbf{r}_1\\ \mathbf{r}_2\\ \mathbf{r}_3\\ \mathbf{r}_4\end{pmatrix}|\mathbf{R}_0\sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix},\mathbf{R}_0\otimes\mathbf{I}_{599}\right), \tag{59}$$

where $\mathbf{G}$ is the genomic relationship matrix among the 599 lines; $\mathbf{G}_0$ is a 4×4 between-trait genomic covariance matrix and $\mathbf{R}_0$ is a 4×4 residual covariance matrix; residuals were independent between individuals, but a full covariance structure within individuals was posed. The four traits correspond to different locations so environmental correlations are expected to be null. However, we specified an unstructured $\mathbf{R}_0$ to account for correlations that are potentially created by non-additive genetic effects not accounted for in our additive genetic model, but known to exist for wheat yield. The two covariance matrices were estimated using a crude maximum likelihood procedure with ad-hoc adjustments to ensure positive-definiteness (a less crude but more involved algorithm would have given estimates that would not require any such adjustment). The estimates were subsequently taken as known and treated as hyper-parameters. The matrices (rounded values) were:

$$\mathbf{G}_0=\begin{bmatrix}0.831 & -0.319 & -0.247 & -0.350\\ -0.319 & 0.750 & -0.195 & -0.213\\ -0.247 & -0.195 & 0.757 & -0.176\\ -0.350 & -0.213 & -0.176 & 0.752\end{bmatrix}, \tag{60}$$

and

$$\mathbf{R}_0=\begin{bmatrix}0.830 & -0.225 & -0.289 & -0.202\\ -0.225 & 0.872 & -0.330 & -0.317\\ -0.289 & -0.3330 & 0.918 & -0.352\\ -0.202 & -0.317 & -0.352 & 0.895\end{bmatrix}. \tag{61}$$

The matrix of phenotypic correlations between traits (Ψ), $\mathbf{V}_0=\mathbf{G}_0+\mathbf{R}_0$, was:

$$\Psi=\begin{bmatrix}1 & -0.329 & -0.321 & -0.334\\ -0.329 & 1 & -0.3162 & -0.322\\ -0.321 & -0.3162 & 1 & -0.318\\ -0.334 & -0.322 & -0.318 & 1\end{bmatrix}.$$

The adjustment for positive-definiteness produced estimates of phenotypic correlations (negative and similar between pairs of traits). Traits turned out to be negatively correlated at genetic, residual and phenotypic levels. The correlation values should not be interpreted inferentially as the adjustment was done solely to facilitate calculation and for illustrative purposes.

Prior to selection, each line had a grain yield in each environment. The data available for analysis post-selection (phenotypes centered and standardized) included the lines exceeding the performance thresholds $t_1=0$, $t_2=0.6$ and $t_3=1.3$ units above the mean in environments 1, 2 and 3, respectively, so only lines with performance above the minimum values for the three traits had records in environment 4. Only 10 such lines met the "culling levels", so phenotypes presented to the hypothetical analyst consisted of a 10×4 matrix, lines in rows and traits in columns. Ignoring selection, the posterior distribution (given $\mathbf{G}_0$ and $\mathbf{R}_0$) of the genomic breeding values using the selected dataset is multivariate normal, with the mean vector:

$$\widetilde{\mathbf{g}}_{sel}=\begin{pmatrix}\widetilde{\mathbf{g}}_{1,sel}\\ \widetilde{\mathbf{g}}_{2,sel}\\ \widetilde{\mathbf{g}}_{3,sel}\\ \widetilde{\mathbf{g}}_{4,sel}\end{pmatrix}=\left(\mathbf{G}_0\otimes\mathbf{G}_{sel,sel}\right)\left(\mathbf{G}_0\otimes\mathbf{G}_{sel,sel}+\mathbf{R}_0\otimes\mathbf{I}_{10}\right)^{-1}\begin{pmatrix}\mathbf{y}_{1,sel}\\ \mathbf{y}_{2,sel}\\ \mathbf{y}_{3,sel}\\ \mathbf{y}_{4,sel}\end{pmatrix}, \tag{62}$$

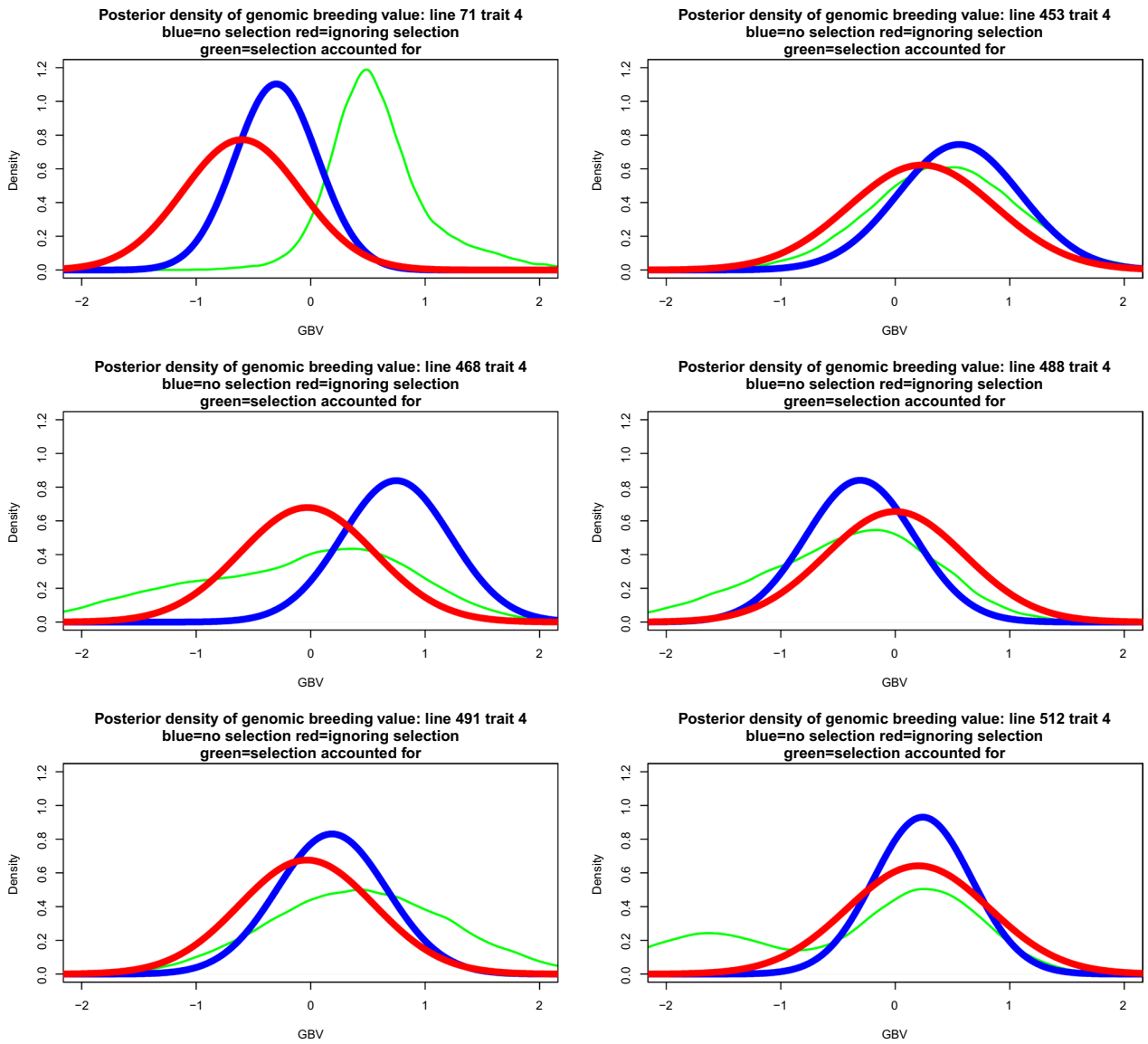

Fig. 5 Posterior densities of six lines without, ignoring or accounting for selection: tetra-variate analysis

and covariance matrix

$$\mathbf{C}_{g,sel} = \left(\mathbf{G}_0 \otimes \mathbf{G}_{sel,sel}\right) \left[\mathbf{I} - \left(\mathbf{G}_0 \otimes \mathbf{G}_{sel,sel} + \mathbf{R}_0 \otimes \mathbf{I}_{10}\right)^{-1} \left(\mathbf{G}_0 \otimes \mathbf{G}_{sel,sel}\right)\right]. \quad (63)$$

Above, $\widetilde{\mathbf{g}}_{i,sel}$ is a 10×1 vector of posterior means of genomic breeding values in the ten selected lines for trait i; $\mathbf{G}_{sel,sel}$ is the 10×10 genomic relationship matrix between selected lines, and $\mathbf{y}_{i,sel}$ is their vector of phenotypes.

Under the multivariate selection scheme adopted, Eq. (56) is expressible as:

$$
\begin{aligned}
&p(\theta|\mathbf{y}_{obs},\mathbf{r},Hyp)\\
&\propto \prod_{i\epsilon S^+} p(y_{1i},y_{2i},y_{3i},y_{4i}|\theta)p(\theta|Hyp)\prod_{i\epsilon S^-}\Pr\left(y_{1i}<t_1,y_{2i}<t_2,y_{3i}<t_3|\theta\right)\\
&\propto \prod_{i\epsilon S^+} p(y_{1i},y_{2i},y_{3i},y_{4i}|\theta_{sel})p(\theta_{sel}|Hyp)\\
&\times \prod_{i\epsilon S^-}\Pr\left(y_{1i}<t_1,y_{2i}<t_2,y_{3i}<t_3|\theta_{nsel}\right)p(\theta_{nsel}|\theta_{sel},Hyp)\\
&\propto p(\theta_{sel}|\mathbf{y}_{obs},Hyp)\prod_{i\epsilon S^-}\Pr\left(y_{1i}<t_1,y_{2i}<t_2,y_{3i}<t_3|\theta_{nsel}\right)p(\theta_{nsel}|\theta_{sel},Hyp).
\end{aligned}
\tag{64}
$$

Above,

$$
p(\theta_{sel}|\mathbf{y}_{obs},Hyp)\propto\exp\left[-\frac{1}{2}(\mathbf{g}_{sel}-\widetilde{\mathbf{g}}_{sel})'\mathbf{C}_{g,sel}^{-1}(\mathbf{g}_{sel}-\widetilde{\mathbf{g}}_{sel})\right],
\tag{65}
$$

$$
\begin{aligned}
&\Pr\left(y_{1i}<t_1,y_{2i}<t_2,y_{3i}<t_3|\theta_{nsel}\right)\\
&=\int_{-\infty}^{t_1}\int_{-\infty}^{t_2}\int_{-\infty}^{t_3}\phi\left(\mathbf{y}_{i,nsel}|\mathbf{g}_{i,nsel},\begin{bmatrix}R_{11}&R_{12}&R_{13}\\R_{21}&R_{22}&R_{23}\\R_{31}&R_{32}&R_{33}\end{bmatrix},t_1,t_2,t_3\right)d\mathbf{y}_{i,nsel}\\
&=\Phi\left(\mathbf{y}_{i,nsel}|\mathbf{g}_{i,nsel},\mathbf{R}_0,t_1,t_2,t_3\right);\ i\epsilon S^-,
\end{aligned}
\tag{66}
$$

where $\phi(.)$ is the density of a trivariate normal distribution with the mean vector $\mathbf{g}_{i,nsel}=\left(g_{1i,nsel},g_{2i,nsel},g_{2i,nsel}\right)'$, R_{ij} is an appropriate element of $\mathbf{R}_0$, and $\Phi(.)$ is a trivariate normal distribution function. Further, $p(\theta_{nsel}|\theta_{sel},Hyp)$ is the density of a multivariate normal distribution with mean vector

$$
\mathbf{m}_{n|s}=\left(\mathbf{G}_0\otimes\mathbf{G}_{nsel,sel}\right)\left(\mathbf{G}_0^{-1}\otimes\mathbf{G}_{sel,sel}^{-1}\right)\mathbf{g}_{sel}=\left(\mathbf{I}_4\otimes\mathbf{G}_{nsel,sel}\mathbf{G}_{sel,sel}^{-1}\right)\mathbf{g}_{sel},
\tag{67}
$$

and covariance matrix

$$
\mathbf{V}_{n|s}=\mathbf{G}_0\otimes\mathbf{G}_{nsel,nsel}-\mathbf{G}_0\otimes\mathbf{G}_{nsel,sel}\mathbf{G}_{sel,sel}^{-1}\mathbf{G}_{sel,nsel}
\tag{68}
$$

Collecting terms,

$$
p(\theta|\mathbf{y}_{obs},\mathbf{r},Hyp)\propto\exp\left[-\frac{1}{2}(Q_{sel}+Q_{nsel})+\Delta(\mathbf{y}_{nsel},\mathbf{g}_{nsel})\right]
\tag{69}
$$

where

$$
Q_{sel}=(\mathbf{g}_{sel}-\widetilde{\mathbf{g}}_{sel})'\mathbf{C}_{g,sel}^{-1}(\mathbf{g}_{sel}-\widetilde{\mathbf{g}}_{sel}),
\tag{70}
$$

$$
Q_{nsel}=(\mathbf{g}_{nsel}-\mathbf{m}_{n|s})'\mathbf{V}_{n|s}^{-1}(\mathbf{g}_{nsel}-\mathbf{m}_{n|s}),
\tag{71}
$$

$$
\Delta(\mathbf{y}_{nsel},\mathbf{g}_{nsel})=\sum_{i\epsilon S^-}\log\left[\Phi(\mathbf{y}_{i,nsel}|\mathbf{g}_{i,nsel},\mathbf{R}_0,t_1,t_2,t_3)\right].
\tag{72}
$$

To estimate the posterior distribution, a Metropolis algorithm was tailored as in Example 2. Here, the proposal distribution (with dimension 599) was a multivariate normal distribution with mean vector equal to:

$$
\begin{bmatrix}\mu_1\\\mu_2\\\mu_3\\\mu_4\end{bmatrix}=\begin{bmatrix}G_{11}\mathbf{A}(G_{11}\mathbf{A}+2R_{11}\mathbf{I}_{599})^{-1}\mathbf{y}_1\\G_{22}\mathbf{A}(G_{22}\mathbf{A}+2R_{22}\mathbf{I}_{599})^{-1}\mathbf{y}_2\\G_{33}\mathbf{A}(G_{33}\mathbf{A}+2R_{33}\mathbf{I}_{599})^{-1}\mathbf{y}_3\\G_{44}\mathbf{A}(G_{44}\mathbf{A}+2R_{44}\mathbf{I}_{599})^{-1}\mathbf{y}_4\end{bmatrix},
\tag{73}
$$

and block-diagonal covariance matrix

$$
\mathbf{D}=\oplus_{i=1}^4 G_{ii}\mathbf{A}\left[\mathbf{I}_{599}-(G_{ii}\mathbf{A}+2R_{ii}\mathbf{I}_{599})^{-1}G_{ii}\mathbf{A}\right].
\tag{74}
$$

The proposal distribution used an overdispersed covariance matrix. We ran four separate chains: two had 2500 iterations each, and the third and fourth ones had 5000 and 40,000 iterations, respectively. The *R* metric [46] indicated that convergence had been attained at about iteration 1500. Conservatively, the last 500 iterations of chains 1 and 2 were saved for inference; likewise, 3000 and 38,000 iterations were saved from chains 3 and 4, respectively. Posterior distributions of the genomic breeding values of the 10 selected lines were estimated from the 42,000 samples available for inference. Figure 5 displays posterior densities of 6 out of the 10 selected lines. Ideally, we would expect the blue and green curves to separate from the red curve (ignoring selection). Because selection was so severe (10 out of an initial 599), the three Bayesian analyses displayed great uncertainty, especially the one where the selection process was accounted for.

This example suggests that, even when the selection or dropout process is known, accounting for it may not be a fruitful process.

Example 4: nor-optimal selection is not ignorable

Often, selection is of a stabilizing form and aimed at moving the population towards some optimum value. A model for such selection was introduced by [26] and used later by [29, 47]. Let $\mathbf{y} = \left[\begin{array}{cc}\mathbf{y}_1' & \mathbf{y}_2'\end{array}\right]'$ be a vector of random variables; some of its components may be unobservable. Without selection, assume:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} \sim N\left(\mathbf{m} = \begin{bmatrix}\mathbf{m}_1\\ \mathbf{m}_2\end{bmatrix}, \mathbf{V} = \begin{bmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{bmatrix}\right), \tag{75}$$

where the $\mathbf{m}$'s are mean vectors and the $\mathbf{V}$'s are covariance matrices. Selection operates on $\mathbf{y}_1$ through the Gaussian fitness function

$$H(\mathbf{y}_1) = \exp\left[-\frac{1}{2}(\mathbf{y}_1 - \lambda)'\boldsymbol{\Gamma}^{-1}(\mathbf{y}_1 - \lambda)\right], \tag{76}$$

where λ is a vector-valued "optimum" and the positive-definite matrix Γ describes the sharpness of multivariate selection. The fitness function is symmetric about λ and has a maximum value of 1 when $\mathbf{y}_1 = \lambda$; values of $\mathbf{y}_1$ "far away" from λ confer lower fitness. In a single-variable situation, fitness takes the form $H(y) = \exp[-\frac{(y-\lambda)^2}{2\gamma}]$. A smaller γ implies a sharper decay in fitness when y deviates from λ; larger values denote a more gentle selection, with no selection at all if $\gamma = \infty$. Write

$$H(\mathbf{y}_1) = \exp\left[-\frac{1}{2}(\mathbf{y} - \lambda_0)'\boldsymbol{\Gamma}_0^-(\mathbf{y} - \lambda_0)\right], \tag{77}$$

where $\lambda_0' = \left[\begin{array}{cc}\lambda' & \mathbf{0}\end{array}\right]$; $\mathbf{0}$ is a null vector with order equal to that of $\mathbf{y}_2$ and $\boldsymbol{\Gamma}_0^- = \begin{bmatrix}\boldsymbol{\Gamma}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$. The density of $\mathbf{y}$ after selection acting on $\mathbf{y}_1$ (ignoring dependence on parameters in the notation) is:

$$p_s(\mathbf{y}_1, \mathbf{y}_2) = \frac{H(\mathbf{y}_1)p(\mathbf{y}_1, \mathbf{y}_2)}{\int\int H(\mathbf{y}_1)p(\mathbf{y}_1, \mathbf{y}_2)d\mathbf{y}_1 d\mathbf{y}_2} = \frac{H(\mathbf{y}_1)}{\overline{H}}p(\mathbf{y}_1, \mathbf{y}_2); \tag{78}$$

equivalently, the density in "survivors" is:

$$p_s(\mathbf{y}_1, \mathbf{y}_2) \propto H(\mathbf{y}_1)p(\mathbf{y}_1, \mathbf{y}_2). \tag{79}$$

Using Eq. (77), the post-selection density becomes:

$$p_s(\mathbf{y}_1, \mathbf{y}_2) \propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y} - \lambda_0)'\boldsymbol{\Gamma}_0^-(\mathbf{y} - \lambda_0) + (\mathbf{y} - \mathbf{m})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{m})\right]\right\}. \tag{80}$$

Combining the two quadratic forms on $\mathbf{y}$ in Eq. (80) yields

$$\begin{aligned}&(\mathbf{y} - \lambda_0)'\boldsymbol{\Gamma}_0^-(\mathbf{y} - \lambda_0) + (\mathbf{y} - \mathbf{m})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{m})\\ &= (\mathbf{y} - \mathbf{m}_s)'\left(\boldsymbol{\Gamma}_0^- + \mathbf{V}^{-1}\right)(\mathbf{y} - \mathbf{m}_s) + (\lambda_0 - \mathbf{m})'\boldsymbol{\Gamma}_0^-\left(\boldsymbol{\Gamma}_0^- + \mathbf{V}^{-1}\right)^{-1}\mathbf{V}^{-1}(\lambda_0 - \mathbf{m})\end{aligned} \tag{81}$$

Above,

$$\mathbf{m}_s = \left(\boldsymbol{\Gamma}_0^- + \mathbf{V}^{-1}\right)^{-1}(\boldsymbol{\Gamma}_0^-\lambda_0 + \mathbf{V}^{-1}\mathbf{m}) \tag{82}$$

Since the second component of Eq. (81) does not involve $\mathbf{y}$, (80) can be written as:

$$p_s(\mathbf{y}_1, \mathbf{y}_2) \propto exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{m}_S)'\left(\boldsymbol{\Gamma}_0^- + \mathbf{V}^{-1}\right)(\mathbf{y} - \mathbf{m}_S)\right\}. \tag{83}$$

Hence, the joint distribution of $\mathbf{y}$ remains normal in survivors to selection but with different parameters. Therefore, the post-selection distribution is:

$$[\mathbf{y}]_s \sim N[\mathbf{m}_s, \mathbf{V}_s = (\boldsymbol{\Gamma}_0^- + \mathbf{V}^{-1})^{-1}]. \tag{84}$$

Note that

$$\mathbf{V}_s = (\boldsymbol{\Gamma}_0^- + \mathbf{V}^{-1})^{-1} = \mathbf{V}(\mathbf{I} + \boldsymbol{\Gamma}_0^-\mathbf{V})^{-1} \tag{85}$$

where $(\mathbf{I} + \boldsymbol{\Gamma}_0^-\mathbf{V})^{-1}$ is related to the "coefficient of centripetal selection" S [29]. In a scalar situation $(\mathbf{I} + \boldsymbol{\Gamma}_0^-\mathbf{V})^{-1} = \frac{\gamma}{V+\gamma} = 1 - S$, gives the fraction of variance (V) remaining after selection and $S = \frac{V}{V+\gamma}$ measures the fraction removed by nor-optimal selection.

We use a canonical setting to show that selection is not ignorable, i.e., selection must be modelled for appropriate inference. Assume that the mean and the additive genetic and environmental variance components prior to selection are known. The setting is $y = u + e$, where $u \sim N(0, h^2)$ and $e \sim N(0, 1 - h^2)$ are independently distributed standardized breeding values and environmental effects, respectively, and h^2 is trait heritability. The marginal distribution of phenotypes is $y \sim N(0, 1)$. Without selection,

$$\begin{bmatrix} y \\ u \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & h^2 \\ h^2 & h^2 \end{bmatrix} \right), \tag{86}$$

$E(u|y) = h^2 y$, and $Var(u|y) = h^2(1-h^2)$. After a round of phenotypic selection towards λ with sharpness γ, the joint distribution of breeding values and phenotypes remains bivariate normal. Post-selection,

$$E_s \begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} \lambda S \\ h^2 \lambda S \end{bmatrix}, \tag{87}$$

and

$$Var_s \begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} 1-S & h^2(1-S) \\ h^2(1-S) & h^2(1-h^2 S) \end{bmatrix}, \tag{88}$$

respectively, where $S = \dfrac{1}{1+\gamma}$. The additive variance is a fraction $1-h^2 S$ of the genetic variation prior to selection. Post-selection, the best predictor [42, 48] of u is:

$$E_s(u|y) = h^2 \lambda S + \frac{h^2(1-S)}{1-S}(y - \lambda S) = h^2 y, \tag{89}$$

and

$$Var_s(u|y) = h^2(1-h^2 S) - \frac{h^4(1-S)^2}{1-S} = h^2\left(1-h^2\right). \tag{90}$$

The parameters of the conditional distribution $[u|y]$ are as prior to selection, so the latter is ignorable from the point of view of learning u from y.

Suppose now that the model is $y_i = \mu + u_i + e_i$ with $u \sim N(0, h^2)$ and $e \sim N(0, 1-h^2)$ as before; h^2 is known but μ is unknown. Given a sample of size n, without selection, the posterior distribution of μ after assigning a flat prior to the latter parameter is $\mu|\mathbf{y} \sim N\left(\bar{y}, n^{-1}\right)$ where $\bar{y} = \dfrac{\sum_{i=1}^n y_i}{n}$ is the maximum likelihood estimator of μ [25]. The analyst, however, is presented with a sample of m individuals known to belong to a population subjected to stabilizing selection towards λ with coefficient of selection S. After selection, the marginal distribution of the phenotypes is:

$$N(\mu_s = \mu(1-S) + \lambda S, V_s = (1-S)). \tag{91}$$

If a flat prior is assigned to μ_s, then $\mu_s|\mathbf{y} \sim N\left(\dfrac{\sum_{i=1}^m y_i}{m}, (1-S)m^{-1}\right)$. What is the posterior distribution of μ? Changing variables $\mu_s \longrightarrow \mu$,

$$\begin{aligned} p_s(\mu|\mathbf{y}) &\propto (1-S) p_s(\mu_s|\mathbf{y}) \\ &\propto \exp\left\{ -\frac{m}{2(1-S)} \left[\mu(1-S) + \lambda S - \frac{\sum_{i=1}^m y_i}{m} \right]^2 \right\} \\ &\propto \exp\left\{ -\frac{m(1-S)}{2} \left[\mu - \frac{\bar{y} - \lambda S}{1-S} \right]^2 \right\}. \end{aligned}$$

The preceding implies that the posterior distribution of μ after selection changes to $[\mu|\mathbf{y}]_s \sim N\left[\dfrac{\bar{y} - \lambda S}{1-S}, \dfrac{1}{n(1-S)}\right]$. In short, the missing data process must be considered from the point of view of inferring the mean of the base population, but can be ignored for learning the additive genetic value u.

Discussion

Selection is a central theme in evolutionary and applied quantitative genetics [12]. Yet, textbooks and papers place focus on stylized models, with less emphasis on parameter inference using data from real selection processes. The literature from animal breeding has special relevance because their data derive mostly from farm records of performance with incomplete reporting and follow-up, especially of events leading to non-randomly missing observations. In this section, some landmark papers on the topic are discussed and their messages are contrasted with our results.

A large part of the animal breeding literature in the first six or seven decades of the 20th century reported randomized selection experiments, typically possessing a small scale and insufficient resolution or replication [49]. By virtue of design, the analysis of these experiments was not too challenging, statistically speaking, and much work centered on the assessment of the expected variability of response to selection in unreplicated experiments [50]. Perhaps the first formal attempt at addressing distortion in inference from observational data generated by a type of culling employed in animal breeding was [30]. Suppose $\mathbf{y}_0, \mathbf{y}_1 \in S_1$ and $\mathbf{y}_2 \in S_2$ are observable data derived from a sequential selection of individuals, i.e., $\mathbf{y}_0 \rightarrow \mathbf{y}_1 \longrightarrow \mathbf{y}_2$ where S_1 and S_2 are sampling spaces modified by selection, while $\mathbf{y}_0$ has unrestricted space. Based on $\mathbf{y}_0$, then $\mathbf{y}_1$ is observed, and given $\mathbf{y}_0$ and $\mathbf{y}_1$, data $\mathbf{y}_2$ are collected. These authors showed bias of linear ordinary least squares estimators of production differences

between, e.g., ages of cow or of time trends when there was sequential selection. The least squares estimators examined were: (a) difference between averages of second lactation records ($\mathbf{y}_1$) and of all first records $\mathbf{y}_0$ (gross comparison), and (b) between second and first lactation averages, but only for cows that had the two records of production (paired comparison). In the gross comparison, missing data for second records were those for cows with lower "fitness" due to having lower first production records. In the paired comparison, the missing data were all records (first and second lactation) of cows not given an opportunity to have a second lactation. In the absence of selection, a multivariate normal joint distribution was assumed by [30], with the fixed parameter vector θ (fixed effects and variance-covariance components) inferred by maximum likelihood. Henderson [30] noted that if all available records are used in the analysis, the maximum likelihood estimator of age effects, i.e., a location parameter, would not contain bias even if selection is ignored in the analysis. Under normality and with a general but known covariance structure, the maximum likelihood estimator is generalized least squares, not ordinary least squares. Their paper did not address inference of unobserved producing abilities or of random genetic effects such as breeding values, which are factors underpinning the non-diagonal covariance matrix structure.

More generally, let $[\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2; \theta]$ represent the joint distribution (does not need to be Gaussian) of the $\mathbf{y}$ vectors, indexed by some parameter θ. Apart from numerical issues, the maximum likelihood estimates are straightforward to obtain because of the automaticity of the method. Since the conditional distributions $[\mathbf{y}_1|\mathbf{y}_0; \theta]$ and $[\mathbf{y}_2|\mathbf{y}_0, \mathbf{y}_1; \theta]$ hold for any $\mathbf{y}_0$ and $\mathbf{y}_1$, the joint density can be written as $p(\mathbf{y}_0|\theta)p(\mathbf{y}_1|\mathbf{y}_0; \theta)p(\mathbf{y}_2|\mathbf{y}_0, \mathbf{y}_1; \theta) = p(\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2; \theta)$. Hence, the likelihood function (any part of the joint density involving θ) is unaffected by selection and the non-random process can be ignored when computing maximum likelihood estimates of θ if such sequential mechanism takes place. However, the asymptotic properties of the estimators are affected by selection, since Fisher's expected information must be computed by taking expectations with respect to $[\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2; \theta, \mathbf{y}_1 \in S_1, \mathbf{y}_2 \in S_2]$ instead of $[\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2; \theta]$; a selection model must be adopted for calculation of expected second derivatives. This was done by [51] for estimation of repeatability of production in dairy cattle.

Maximum likelihood estimates are typically biased in finite samples even without selection, e.g., the maximum likelihood estimator of residual variance in linear regression models has a downwards bias. In an early study, Rothschild et al. [52] examined bias of estimates of genetic parameters (based on variance and covariance components) by simulating first and second lactation records in dairy cattle. The selection scheme was stylized: 50 % of 5000 progeny of 100 bulls was allowed to have second records, and individuals were chosen either at random or were those with the highest first record of production; the scheme was replicated 200 times. The simulation did not detect bias of estimates of heritability of first and second records, or of genetic and phenotypic correlations between the two lactations. However, the mean squared error of the estimates was larger under selection than under a random choice, illustrating that the finite sample distribution of maximum likelihood estimates is affected by selection. In other words, ignorability of selection cannot be claimed without qualification: it does not necessarily imply bias removal or unaffected sampling distribution of estimates. Ignorability means that the likelihood function can be constructed as if selection had not taken place.

Results in [30] carry beyond normality because composition of a joint density as a sequence of conditional densities follows directly from probability theory. However, selection is often based on unobserved "externalities" as well as on data available for analysis, as discussed in our paper. For instance, in a clinical trial, a patient may abandon the study due to some unrecorded event, e.g., tooth ache, that may be treatment-related, so some component of θ affects a missing process involving unobserved data. The notion of ignorability of missingness of data was formalized in [21] and adapted to likelihood-based inference in animal breeding by Im et al. [22]. The main messages of such work are: if the probability of missingness (our fitness function is an equivalent metric) depends on observed data only, or if it involves parameters that are "distinct" from θ, the selection process can be ignored from the point of view of locating the maximizer of the likelihood.

Apart from genetic and environmental parameters, animal and plant breeders also seek estimates of unobservable breeding values, i.e., quantities that vary over individuals and that are not construed as parameters in classical inference. However, results in [30] do not apply without qualification to inference of unobservable (but realized) random variables (called "prediction" of a random vector $\mathbf{u}$), since estimation and prediction are treated distinctly in frequency and likelihood-based approaches. Actually, the mixed model equations algorithm was derived in their paper by joint maximization with respect to fixed effects and $\mathbf{u}$, of a "penalized" likelihood function under Gaussian assumptions and known dispersion structure, incorrectly interpreted as a classical likelihood. The $\mathbf{u}$-solution of the estimating equations was later shown to correspond to the BLUP, and for many years was wrongly referred to as "maximum likelihood

estimator" of **u.** Since the the penalized likelihood is proportional $[\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, \mathbf{u}; \theta]$ one can also argue that the penalized maximum likelihood estimator of **u** (that is, what we would call BLUP) could be calculated ignoring the sequential selection process $\mathbf{y}_0 \rightarrow \mathbf{y}_1 \longrightarrow \mathbf{y}_2$.

Perhaps the bias issue is what motivated Henderson et al. [30] to study BLUP under the form of selection described by Pearson [53], who had shown how selection operating upon a multivariate normal distribution altered its mean vector and covariance matrix. A paper [54] noted that a special case of Pearson–Henderson selection is the truncation scheme in textbooks of quantitative genetics [55]. Under Pearson's model and Gaussian assumptions, the first and second moments of the joint distribution of a set of random variables (observed or latent) after selection, can be arrived at analytically; formulae apply to a single cycle of selection only, as multivariate normality is destroyed post-selection. Assuming the dispersion structure was known [43], derived conditions under which Pearsonian selection could be ignored in the computation of BLUP. Essentially, he considered linear predictors, $\mathbf{L}'\mathbf{y}$ of an unobservable random vector **u**, such that $E_s(\mathbf{L}'\mathbf{y}) = E_s(\mathbf{u})$, where s denotes Pearsonian selection. In this class of predictors, Henderson [43] searched for the **L** matrix that produced minimum variance of prediction error. Two of his results have had a marked influence on animal breeding modeling. One was that, if $E(\mathbf{L}'\mathbf{y}) = \mathbf{0}$ (location invariance), the selection process could be ignored, with BLUP computed as if selection had not taken place. The other one, was that if selection had been based on the linear combination $\mathbf{L}'\mathbf{u}$, by treating fixed effects levels of other random vectors in the model (e.g., contemporary groups) as fixed effects one could arrive at unbiased predictors of **u**.

Henderson's treatment of selection was discussed critically by, e.g., [11] and by [56, 57]. The frequentist setting in [43] assumed that the **L** matrix was constant (and, therefore, the incidence matrices in the linear model) over conceptual repeated sampling. This assumption is not reasonable, as a conceptual replication with the same distribution of observations over subclasses could not be expected to occur with unbalanced field data collected over a large number of contemporary groups and several years. The most widely cited result was that BLUP is unaffected by selection if the criterion used for ranking ($\mathbf{L}'\mathbf{y}$) has a probability distribution that is translation invariant. This requirement is violated in animal breeding any time that a model with fixed "genetic group effects" is used, a routine modeling strategy, e.g., in beef cattle evaluation; in this case $\mathbf{L}'\mathbf{y}$ has a non-null expectation. The $\mathbf{L}'\mathbf{u}$ development is logically difficult to follow. For instance, selection based on $\mathbf{L}'\mathbf{u}$ requires knowledge of **u**, so there, predicting this latter vector would not be needed. The study of Schaeffer [58] discussed pros and cons of treating a large number of contemporary groups as fixed effects, the strategy recommended by Henderson to eliminate biases due to non-random associations between breeding values of bulls (used over many herds) and farm effects, which the latter interpreted as a special case of selection based on $\mathbf{L}'\mathbf{u}$. Furthermore, Schaeffer [58] observed that if the number of individuals per contemporary group is large, treating their effects either as fixed or random is inconsequential. However, the well-known James-Stein theoretical result indicates that shrinkage of fixed effects estimates can yield smaller mean squared of estimation if the number of contemporary groups is large [59]. Hence, Henderson's prescription is not on solid ground.

Many papers, mainly using simulation, observed that use of "full" relationship matrices in the statistical model could account in some sense for selection, even in situations where missing data lead to non-ignorable selection. In our approach, unless genetic relatedness enters explicitly into the fitness function, there is no transparent theoretical reason for such expectation, other than the benefit stemming from a correct specification of the covariance matrix. As discussed, Henderson et al. [30] observed that ordinary least-squares, assuming independent and identically distributed observations, produced biased estimates of fixed effects under sequential selection, while generalized least-squares did not. In the "gross comparison" of the cow example, all records are used, so selection is based entirely on observed data. However, least squares assumes that first and second lactation records are conditionally independent, inducing a likelihood that is proportional to the product (of densities) $p(\mathbf{y}_0|\theta)p(\mathbf{y}_1|\theta)p(\mathbf{y}_2|\theta)$, so all data used for selection decisions are included but all covariances are ignored. On the other hand, the generalized least-squares estimator derives from a likelihood that is proportional to $p(\mathbf{y}_0|\theta)p(\mathbf{y}_1|\mathbf{y}_0,\theta)p(\mathbf{y}_2|\mathbf{y}_1,\mathbf{y}_0,\theta)$. Thus, what a genetic relationship matrix does is to represent extant dependencies properly. Apart from an appropriate model specification, the more complete the pedigree (or genomic) information is, the better the dependencies are modelled. The issue here is one of proper specification of the likelihood, i.e., some effects treated as random with an appropriate covariance structure producing shrinkage of least-squares solutions. Henderson et al. [30] referred to this matter as "incomplete repeatability" of records. Ignoring data relevant to the selection decisions does produce distortion in inference, since the fitness function contains information about the unknown θ that is not ignorable.

Harville [60] examined selection using the following setting. Without selection, the data have a distribution with density function $p(\mathbf{y}|\theta)$ and with the sample space of $\mathbf{y}$ unrestricted in any manner. Selection is such that only data in a restricted space S is observed, i.e., when $\mathbf{y}\in S$. Under selection, observations appear with density

$$p_s(\mathbf{y}|\theta) = \frac{p(\mathbf{y}|\theta)}{\Pr\left(\mathbf{y}\in S|\theta\right)}; \mathbf{y}\in S, \tag{92}$$

so the posterior density under selection, with the prior density being $p(\theta)$, is:

$$p_s(\theta|\mathbf{y}) = \frac{\frac{1}{\Pr\left(\mathbf{y}\in S|\theta\right)}p(\mathbf{y}|\theta)p(\theta)}{\int \frac{1}{\Pr\left(\mathbf{y}\in S|\theta\right)}p(\mathbf{y}|\theta)p(\theta)d\theta} \propto \frac{1}{\Pr\left(\mathbf{y}\in S|\theta\right)}p(\theta|\mathbf{y}). \tag{93}$$

Since $\Pr\left(\mathbf{y}\in S|\theta\right)$ depends on θ, the selection process cannot be ignored. A well known special case of this type of selection (as noted, also a special case of Pearsonian selection) is the classical truncation model of quantitative genetics. To illustrate this, suppose that, prior to selection, observations are identically and independently distributed as $N(\mu,\sigma^2)$ with the variance σ^2 known but μ unknown. Selection is such that n observations exceeding a known threshold t are presented to the analyst and a flat prior is assigned to μ. The posterior density after selection is:

$$p_s(\mu|\mathbf{y},t) \propto \prod_{i=1}^{n} \frac{\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right]}{\Pr\left(y_i > t|\mu,\sigma^2\right)}, \tag{94}$$

which is the product of n truncated normal density functions. Using standard algebra [25]

$$p_s(\mu|\mathbf{y},t) \propto \frac{\exp\left[-\frac{n}{2\sigma^2}(\mu-y)^2\right]}{\left[1-\Phi\left(\frac{t-\mu}{\sigma}\right)\right]^n}, \tag{95}$$

where $\Phi(.)$ is the normal distribution function. The posterior density of μ under selection does not have a closed form.

A terminology employed in classification problems, i.e., "hard" versus "soft" [61] may be useful to contrast the treatments of selection employed by [43, 54, 60] with the one we used in the present paper. In "hard selection", the sampling space S implies fixed constraints (e.g., culling levels) defining a simplex, inside of which data are observed. In a "soft selection model", the fitness or probability of selection depend on arguments that can include fixed hyper-parameters, unknown parameters, observed and unobserved data. The latter corresponds to the missing data treatment examined in [21, 22]. The more realistic and flexible setting in "soft selection" may lead to a diagnosis of the extent to which selection can be ignored. It is not realistic to assume that a fixed selection threshold t holds in conceptual replication. The chance of selection depends on varying observed and unobserved data, and on unequal amounts of information over individuals, aspects that the "soft" selection representation addresses.

There does not seem to be a general prescription to accommodate potential distortions due to selection. In structures that combine cross-sectional, longitudinal and multi-trait data such as in animal breeding, balance is the exception rather than the rule. In plant breeding, datasets are more structured and are often outcomes of designed randomized trials. However, such experiments may involve multiple-environments and years and multiple traits. The missing data or fitness treatment presented here may be also pertinent to data from designed experiments, where missing observations also occur in incomplete block layouts, and the missingness may not be random. The Bayesian approach, together with our treatment of selection, offer an integrated answer to inference, prediction and model selection [25, 62, 63] and goes beyond the likelihood-based approach, where breeding values are inferred indirectly. In the Bayesian treatment, the fitness function can include data, parameters and breeding values, as the latter are members of the vector of unknowns, although assigned distinct prior distributions. Bayesian methods produce automatic measures of uncertainty even under selection and the posterior distribution of the fitness function can be estimated using draws from its posterior distribution.

Is it always fruitful to account for selection by introducing a fitness function or by modeling the missing data process? Modeling selection through a fitness function is not without pitfalls. An incorrect specification of fitness may deteriorate inferences beyond those obtained by ignoring selection altogether. Accounting for selection may not reduce uncertainty about breeding values, as we found in our examples with real data, where the missing data process assumed, univariate or multivariate, patterned exactly the protocols constructed. More generally, since the quality of inferences cannot be assessed unambiguously (one does not know the "true value" of parameters), it is risky to assert that inferences are good or bad. Formal model comparison may shed some light. For instance, a Bayes factor analysis may reveal that accounting for selection provides a more plausible description of the data than ignoring selection.

Lastly, since animal and plant breeders are interested in predicting future phenotypes, a predictive assessment may be the most appropriate gauge for constructing and

calibrating models representing competing forms of describing the selection process. For example, Gianola and Schön [64] address several ways of carrying-out cross-validation, directly or indirectly. From a Bayesian perspective, Fong and Holmes [65] argue that the marginal density of the data (denominator of Bayes theorem) is "equivalent" to exhaustive leave-k out cross-validation averaged all possible values of k when a log-posterior predictive distribution is used as scoring rule for competing models. However, their theoretical results depend on the notion of "exchangeability", i.e., that permutation of indexes of observations does not alter the analysis. This concept does not apply to quantitative genetics settings, since, for example, if a parent is individual i, say, the analysis would change drastically if it is permuted with grand-children j. Another example from dairy cattle breeding is as follows: if a cow is m, its production record cannot be exchanged with bull n, with thousands of progeny. Obviously, a bull cannot be milked and a cow can seldom produce such a large progeny group of individuals.

Conclusions

We reviewed and extended theory for analyzing quantitative genomics data stemming from cryptic or structured selection processes. The Bayesian approach provided an integrated approach to inference and prediction under selection, but may or may not yield the best possible predictions, as each problem is essentially unique. One may believe that selection has been accounted for meticulously, but the central question of whether inferences are good or bad does not have an answer. As pointed out by Mark Twain: "*It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so*". It may well be that some statistical learning procedure that ignores quantitative genetics theory and non-randomness ends up as the best prediction machine. Since the most celebrated prediction and classification machines do not make claims about lack of bias or minimum variance of structural parameters, e.g., connection strengths of deep neural networks, these two concepts largely driving the Hendersonian era (at least in animal breeding) may gradually lose relevance. On the one hand, Bayesians may experience a certain *schadenfreude*[1] if this were to occur. On the other hand, it is possible to attain empirically unbiased predictions via calibration of the machines. A positive finding may not help to understand the state of nature, but it may enhance the progress of agriculture.

[1] A certain pleasure derived by someone from another person's misfortune.

Author contributions
DG developed the main structure of the paper and designed the examples used for illustrating the concepts. RLF and C-CS provided critical insights and comments. All authors read and approved the manuscript.

Funding
DG and CCS acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG; Grant No. SCHO 690/4-1). Work was partially supported by the Wisconsin Agriculture Experiment Station.

Availability of data and materials
The wheat data set is publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI, USA. [2]Department of Animal Science, Iowa State University, Ames, IA, USA. [3]Department of Plant Breeding, Technical University of Munich, Freising, Germany.

Received: 19 April 2022 Accepted: 26 October 2022
Published online: 02 December 2022

References

1. Wright S. Evolution in Mendelian populations. Genetics. 1931;16:97–159.
2. Wright S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Proceedings of the 6th international congress of genetics: 24–31 August 1932, Ithaca; 1932. p. 356–66.
3. Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, Visscher PM. From basic science to clinical application of polygenic risk scores: a primer. JAMA Psychiatry. 2021;78:101–9.
4. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb. 1918;52:399–433.
5. Malécot G. Les Mathématiques de l'hérédité. Paris: Masson et Cie; 1948.
6. Kempthorne O. The correlation between relatives in a random mating population. Proc R Soc Lond B Biol Sci. 1954;143:103–13.
7. Henderson CR. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 1976;32:69–83.
8. Weller JI. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics. 1986;42:627–41.
9. Fernando FL, Grossman M. Marker assisted selection using best linear unbiased prediction. Genet Sel Evol. 1989;21:467–77.
10. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
11. Gianola D, Rosa GJM. One hundred years of statistical developments in animal breeding. Annu Rev Anim Biosci. 2015;3:19–56.
12. Walsh B, Lynch M. Evolution and selection of quantitative traits. New York: Oxford University Press; 2018.
13. Fairfield Smith H. A discriminant function for plant selection. Ann Eugenics. 1936;7:240–50.
14. Hazel LN. The genetic basis for constructing selection indexes. Genetics. 1943;28:476–90.
15. Céron-Rojas JJ, Crossa J. Linear selection indices in modern plant breeding. Cham: Springer International Publishing AG; 2018.

16. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.
17. Van Raden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
18. Heslot N, Yang HP, Sorrells ME, Jannink JL. Genomic selection in plant breeding: a comparison of models. Crop Sci. 2012;52:146–60.
19. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193:327–45.
20. Gianola D. Priors in whole genome regression: the Bayesian alphabet returns. Genetics. 2013;194:573–96.
21. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
22. Im S, Fernando RL, Gianola D. Likelihood inferences in animal breeding under selection: a missing data theory viewpoint. Genet Sel Evol. 1989;21:399–414.
23. Sorensen DA, Fernando RL, Gianola D. Inferring the trajectory of genetic variance in the course of artificial selection. Genet Res. 2001;77:83–94.
24. Spiegelhalter D. The art of statistics: how to learn from data. London: Penguin; 2019.
25. Sorensen D, Gianola D. Likelihood, Bayesian, and MCMC methods in quantitative genetics. New York: Springer-Verlag; 2002.
26. Haldane JBS. The measurement of natural selection. Caryologia. 1954;6(480–7):1.
27. Fisher RA. The genetical theory of natural selection. 2nd ed. Springfield: Dover; 1958.
28. Crow JF, Kimura M. An introduction to population genetics theory. New York: Harper & Row; 1970.
29. Latter BDH. Selection in finite populations with multiple alleles. II. Centripetal selection, mutation, and isoallelic variation. Genetics. 1970;66:165–86.
30. Henderson CR, Kempthorne O, Searle RS, von Krosigk M. Estimation of environmental and genetic trends from records subject to culling. Biometrics. 1959;15:192–218.
31. Gianola D, Fernando RL. Bayesian methods in animal breeding theory. J Anim Sci. 1986;63:217–44.
32. Fernando RL, Gianola D. Statistical inferences in populations undergoing selection or non-random mating. In: Gianola D, Hammond K, editors. Advances in statistical methods for genetic improvement of livestock. Heidelberg: Springer-Verlag; 1990. p. 437–53.
33. Little RJA, Rubin DB. Statistical analysis with missing data. 1st ed. New York: Wiley; 1987.
34. Falconer DS. The problem of environment and selection. Am Nat. 1952;86:293–8.
35. Patry C, Ducrocq V. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J Dairy Sci. 2011;94:1011–20.
36. Winkelman AM, Johnson DL, Harris BL. Application of genomic evaluation to dairy cattle in New Zealand. J Dairy Sci. 2015;98:659–75.
37. Jibrila I, Ten Napel J, Vandenplas J, Veerkamps RF, Calus MPL. Investigating the impact of preselection on subsequent single-step genomic BLUP evaluation of pre-selected animals. Genet Sel Evol. 2020;52:42.
38. Wang L, Janss LL, Madsen P, Henshall J, Huang CH, Marois D, et al. Effect of genomic selection and genotyping strategy on estimation of variance components in animal models using different relationship matrices. Genet Sel Evol. 2020;52:31.
39. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198:483–95.
40. Gianola D, Fariello MI, Naya H, Schön CC. Genome-wide association studies with a genomic relationship matrix: a case study with wheat and Arabidopsis. G3 (Bethesda). 2016;6:3241–56.
41. Gianola D, Cecchinato A, Naya H, Schön CC. Prediction of complex traits: robust alternatives to best linear unbiased prediction. Front Genet. 2018;9:195.
42. Henderson CR. Sire evaluation and genetic trends. J Anim Sci. 1973;1973:10–41.
43. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;31:423–49.
44. Fernando RL, Dekkers JCM, Garrick DA. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol. 2014;46:50.
45. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. Boca Raton: Chapman and Hall/CRC Press; 2013.
46. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with discussion). Stat Sci. 1992;7:457–72.
47. Bulmer MG. The genetic variability of polygenic characters under optimizing selection, mutation and drift. Genet Res. 1972;19:17–25.
48. Fernando RL, Gianola D. Optimal properties of the conditional mean as a selection criterion. Theor Appl Genet. 1986;72:822–5.
49. Robertson A. Selection experiments in laboratory and domestic animals. In: Proceedings of the 30th annual meeting of the European Federation of Animal Science, 21–22 July 1979; Harrogate; 1979.
50. Hill WG. Design and efficiency of selection experiments for estimating genetic parameters. Biometrics. 1971;27:293–311.
51. Curnow RN. The estimation of repeatability and heritability from records subject to culling. Biometrics. 1961;7:553–66.
52. Rothschild MF, Henderson CR, Quaas RL. Effects of selection on variances and covariances of simulated first and second lactations. J Dairy Sci. 1979;62:996–1002.
53. Pearson K. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philos Trans R Soc A. 1903;200:1–66.
54. Gianola D, Im S, Fernando RL. Prediction of breeding values under Henderson's selection model: a revisitation. J Dairy Sci. 1988;71:2790–8.
55. Falconer DS, Mackay TFC. Introduction to quantitative genetics. Harlow: Pearson Education Limited; 1996.
56. Thompson R. Sire evaluation. Biometrics. 1979;35:339–53.
57. Gianola D, Fernando RL, Im S, Foulley JL. Likelihood estimation of quantitative genetic parameters when selection occurs: models and problems. Genome. 1989;31:768–77.
58. Schaeffer LR. Necessary changes to improve animal models. J Anim Breed Genet. 2018;135:124–31.
59. Weigel KA, Gianola D, Tempelman RJ, Matos CA, Chen IHC, Wang T, et al. Improving estimates of fixed effects in a mixed linear model. J Dairy Sci. 1991;74:3174–82.
60. Harville DA. Bayesian inference is unaffected by selection: fact or fiction? Am Stat. 2022;76:22–8.
61. Wahba G. Soft and hard classification by reproducing kernel Hilbert space methods. Proc Natl Acad Sci USA. 2002;99:16524–30.
62. Box GEP, Tiao GC. Bayesian inference in statistical analysis. Reading: Addison-Wesley; 1973.
63. Bernardo JM, Smith AFM. Bayesian theory. Chichester: Wiley; 1994.
64. Gianola D, Schön CC. Cross-validation without doing cross-validation in genome-enabled prediction. G3 (Bethesda). 2016;6:3107–28.
65. Fong E, Holmes CC. On the marginal likelihood and cross-validation. Biometrika. 2020;107:489–96.

Publisher's Note