Optimization of sampling designs for pedigrees and association studies

Olivier David, Arnaud Le Rouzic, Christine Dillmann

Optimization of sampling designs for pedigrees and association studies

Olivier David[1,*], Arnaud Le Rouzic[2], and Christine Dillmann[3]

[1]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
ORCID ID: 0000-0002-1389-7158
[2]Université Paris-Saclay, CNRS, IRD,
Évolution, Génomes, Comportement, Écologie,
91198, Gif-sur-Yvette, France
ORCID ID: 0000-0002-2158-3458
[3]Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE - Le Moulon,
91190, Gif-sur-Yvette, France
ORCID ID: 0000-0002-9025-341X
*email: Olivier.David@inrae.fr

SUMMARY: In many studies, related individuals are phenotyped in order to infer how their genotype contributes to their phenotype, through the estimation of parameters such as breeding values or locus effects. When it is not possible to phenotype all the individuals, it is important to properly sample the population to improve the precision of the statistical analysis. This article studies how to optimize such sampling designs for pedigrees and association studies. Two sampling methods are developed, stratified sampling and D optimality. It is found that it is important to take account of mutation when sampling pedigrees with many generations: as the size of mutation effects increases, optimized designs sample more individuals in late generations. Optimized designs for association studies tend to improve the joint estimation of breeding values and locus effects, all the more as sample size is low and the genetic architecture of the trait is simple. When the trait is determined by few loci, they are reminiscent of classical experimental designs for regression models and tend to select homozygous individuals. When the trait is determined by many loci, locus effects may be difficult to estimate, even if an optimized design is used.

KEY WORDS: Bayesian statistics, genetic algorithm, high-dimensional statistics, optimal designs, quantitative genetics

1. Introduction

Quantitative genetics studies the inheritance of quantitative traits such as height or weight, which vary continuously. Experimental studies in quantitative genetics use phenotypic and genetic data to address various fundamental or applied issues such as the selection of plants or animals in breeding programs (Jannink et al., 2010), the identification of individuals at increased risk of disease (Vazquez et al., 2012), the evaluation of genetic resources stored in gene banks (Crossa et al., 2016), the inference of the genetic architecture of traits in association studies (Barton et al., 2007, Chap. 14; Rau et al., 2019), the study of the genetic mechanisms underlying the response to selection (Durand et al., 2010, 2015) or the inference of the strength of evolutionary forces such as drift, mutation or selection (Gillespie, 1998, Chap. 5; Barton et al., 2007, Chap. 19; David et al., 2020). Such studies require a sampling step and the choice of a sampling strategy when it is not possible to phenotype all the individuals of the studied population (Cochran, 1977).

Several studies have investigated how to optimize sampling designs for genomic prediction (Isidro et al., 2015; Rincent et al., 2017; Akdemir and Isidro-Sánchez, 2019). In such experiments, the population is genotyped with genome-wide markers. The individuals belonging to a candidate set are available for phenotyping, while the other individuals cannot be phenotyped because they may be dead, for example. Since phenotyping is costly, some individuals are sampled in the candidate set and are phenotyped. The genetic and phenotypic data are then used to predict the breeding values of target individuals. Designing such surveys consists of choosing the individuals to be phenotyped in the candidate set in order to predict the target breeding values as precisely as possible.

Sampling designs for genomic prediction have been generated using stratified sampling (SS) (Isidro et al., 2015). This method first divides the population into homogeneous subpopulations by clustering the genetic data. Individuals are then sampled in each subpopulation. This method accounts for the population structure and ensures a high genetic diversity in the sample. Alternatively, sampling designs for genomic prediction have been optimized using optimal design theory (Atkinson and Donev, 1992; Chaloner and Verdinelli, 1995; Isidro et al., 2015; Rincent et al., 2017; Akdemir and Isidro-Sánchez, 2019). In this approach, a criterion that quantifies the merit of a design is first defined, for example, a criterion that quantifies the variance of the estimators of interest. Efficient designs are then generated by optimizing this criterion over an appropriate class of designs. Design criteria may be optimized using combinatorial optimization methods such as genetic algorithms or

simulated annealing (Reeves, 1993; David et al., 2000; Van Groenigen, 2000; Müller et al., 2004; Amzal et al., 2006; Brus and Heuvelink, 2007; Butler et al., 2014; Del Moral and Doucet, 2014; Akdemir and Isidro-Sánchez, 2019). In these studies, optimized designs generally performed better than simple random sampling (SRS) and improved the precision of genomic prediction.

Quantitative genetics includes other types of experiments than genomic prediction, for example, surveys involving pedigrees or association studies. It is not clear how to sample pedigrees, if more individuals should be sampled in early or in late generations, and how mating system and mutation influence sampling designs. The experimental setup of an association study is similar to that of genomic prediction, however, their objective is to estimate the effects of loci on traits rather than breeding values. Likewise, it is not clear how to sample individuals in association studies and how the genetic architecture of the trait influences sampling designs.

This article develops sampling methods for pedigrees and association studies. Observations are assumed to follow a classical infinitesimal model, in which the phenotype can be partitioned into independent genetic and environmental components (Section 2). Two sampling methods are developed, SS and D optimality (DO) (Section 3). A real maize experiment is used to study how to sample pedigrees, and the influence of mating system and mutation on sampling (Section 4). Finally, simulations are carried out to study how to sample individuals in association studies, and the influence of sample size and the genetic architecture of the trait on sampling (Section 5).

2. Models

We consider a population of N diploid individuals. This population contains a candidate set $\mathcal{C}$, whose individuals are available for phenotyping for some quantitative trait. A sample of $n < N$ individuals to be phenotyped is chosen in $\mathcal{C}$. Let $\boldsymbol{X}$ be the $n \times N$ incidence matrix whose entry in row j and column i satisfies $\boldsymbol{X}_{ji} = 1$ if the jth observation is collected on individual i, and $\boldsymbol{X}_{ji} = 0$ otherwise. For example, if the second observation is collected on individual 69, then $\boldsymbol{X}_{2\,69} = 1$ and $\boldsymbol{X}_{2\,i} = 0$ for $i \neq 69$.

Trait values are assumed to follow an infinitesimal model (Wray, 1990; Gillespie, 1998; Lynch and Walsh, 1998; Barton et al., 2017). They are assumed to be the sum of a genetic component, referred to as the breeding value, and an environmental component. They are observed with measurement errors: for $1 \leq j \leq n$, the jth observation $\boldsymbol{y}_j$ is

$$\boldsymbol{y}_j = \mu + \sum_{i=1}^{N} \boldsymbol{X}_{ji}\boldsymbol{u}_i + e_j + \xi_j,$$

where μ is the grand mean, $\boldsymbol{u}_i$ is the breeding value of individual i, e_j is an environmental effect, and ξ_j is a measurement error. The residuals $\boldsymbol{\varepsilon}_j = e_j + \xi_j$ are assumed to be independent and normally distributed with mean 0 and variance $\sigma^2_{\boldsymbol{\varepsilon}}$, that is, $\boldsymbol{\varepsilon}_j \sim \mathcal{N}\left(0, \sigma^2_{\boldsymbol{\varepsilon}}\right)$. They are assumed to be independent of breeding values, that is, there are no genotype-environment interactions.

2.1 *Modeling breeding values using the pedigree*

The survey is assumed to involve some genetic data in addition to the phenotypic data. These data may consist of a pedigree dataset, that indicates the parents and the date of birth of each individual. Without loss of generality, dates are expressed here in generations, like in Wray (1990). This is more appropriate for populations with discrete generations, but any other time unit such as day or week could be used instead. The only restriction is that parents must appear in earlier time units than their offspring. The pedigree is assumed to have $G + 1$ generations, denoted by $0, \ldots, G$.

In order to study the sensitivity of designs to mutation, breeding values are modeled taking account of mutation. The mutations arising at generation 0 are not taken into account since they are considered as being part of the alleles present at this generation. New mutations are assumed to arise independently in the individuals of generation 1 and of subsequent generations. The alleles present at generation 0 and new mutations are assumed to have additive effects on breeding values (Wray, 1990): $\boldsymbol{u}_i = \sum_{g=0}^{G} \boldsymbol{\alpha}_{gi}$, where $\boldsymbol{\alpha}_{0i}$ is the component of the breeding value of individual i that is the result of alleles inherited from ancestors born in generation 0, and for $g \geq 1$, $\boldsymbol{\alpha}_{gi}$ is the component that is the result of mutations arising in generation g. The vector $\boldsymbol{\alpha}_0$ of genetic effects inherited from generation 0 is assumed to be normally distributed (Wray, 1990; Lynch and Walsh, 1998; Barton et al., 2017):

$$\boldsymbol{\alpha}_0 \,|\, \sigma^2_0 \sim \mathcal{N}\left(0, \sigma^2_0\, \boldsymbol{A}_0\right),$$

where σ^2_0 is the additive genetic variance, and $\boldsymbol{A}_0$ is an $N \times N$ kinship (or relationship, similarity) matrix. This matrix can be computed in various ways. For the sake of simplicity, it is assumed here that $\boldsymbol{\alpha}_{0i}$ is under the control of a single locus with many alleles and that it is the sum of the contributions from the two alleles carried by individual i, that is, there is no dominance (Gillespie, 1998, Chap. 5). Alleles are transmitted from parents to offspring by Mendelian inheritance. The entry $\boldsymbol{A}_{0ii'}$ is then equal to twice the coefficient of kinship between individuals i and i' (Wright, 1922; Gillespie, 1998, Chap. 5). This coefficient is the probability that two alleles, one sampled from individual i and one sampled from individual i' independently, are identical by descent. It is also assumed here that the individuals of generation 0 are not related and are not inbred, but this assumption can easily be relaxed if information on their relatedness and their inbreeding is available.

For $g \geq 1$, the vector $\boldsymbol{\alpha}_g$ of mutation effects is assumed to be normally distributed (Wray, 1990):

$$\boldsymbol{\alpha}_g \,|\, \sigma_m^2 \sim \mathcal{N}\left(0, \sigma_m^2\, \boldsymbol{A}_g\right),$$

where σ_m^2 is the mutational variance, and $\boldsymbol{A}_g$ is an $N \times N$ kinship matrix. Since $\boldsymbol{\alpha}_{gi}$ is the result of mutations arising at generation g, $\boldsymbol{\alpha}_{gi} = 0$ if individual i was born before generation g, with the convention that all the elements of $\boldsymbol{A}_g$ involving this individual are equal to 0, that is, $\boldsymbol{A}_{gii'} = 0$ for $1 \leq i' \leq N$. Since new mutations are transmitted from parents to offspring by Mendelian inheritance and are assumed to have additive effects, the elements of $\boldsymbol{A}_g$ involving individuals born after generation $g-1$ are equal to twice their coefficient of kinship. Since mutations are assumed to arise independently in gametes, those arising at generation g do not cause any inbreeding in the individuals born at generation g.

Since mutations arise independently over time, the vector of breeding values is normally distributed with mean 0 and conditional variance matrix $\mathrm{Var}\left(\boldsymbol{u} \,|\, \sigma_{\boldsymbol{\varepsilon}}^2, \gamma_0^2, \gamma_m^2\right) = \sigma_{\boldsymbol{\varepsilon}}^2\, \boldsymbol{V}$, where

$$\boldsymbol{V} = \gamma_0^2\, \boldsymbol{A}_0 + \gamma_m^2 \sum_{g=1}^{G} \boldsymbol{A}_g,$$

and the variance components γ_0^2 and γ_m^2 are equal to $\gamma_0^2 = \sigma_0^2/\sigma_{\boldsymbol{\varepsilon}}^2$ and $\gamma_m^2 = \sigma_m^2/\sigma_{\boldsymbol{\varepsilon}}^2$.

2.2 *Modeling breeding values using genotyping data*

Alternatively, the genetic data may consist of a genotyping dataset, giving the genotype of all the individuals at L biallelic loci. Let $\boldsymbol{M}$ be the $N \times L$ matrix that specifies which alleles each individual carries at each locus. Specifically, the genotype $\boldsymbol{M}_{il}$ of individual i at locus l is set to 1 (respectively, -1) if individual i carries two copies of allele 1 (respectively, 2) at locus l, and to 0 if individual i is heterozygous at locus l (VanRaden, 2008). The columns of $\boldsymbol{M}$ are often standardized (VanRaden, 2008; Rincent et al., 2017; Gianola et al., 2020). As explained in the next paragraph, they are divided here by the square root of the average diagonal element of $\boldsymbol{M}\boldsymbol{M}^T$, where T denotes transposition, yielding the matrix $\boldsymbol{Z} = \boldsymbol{M}/\sqrt{z}$, where $z = \mathrm{trace}\left(\boldsymbol{M}\boldsymbol{M}^T\right)/N$.

Alleles are assumed to have additive effects on breeding values, that is, there is no dominance and no epistasis:

$$\boldsymbol{u}_i = \sum_{l=1}^{L} \boldsymbol{Z}_{il}\, \boldsymbol{\beta}_l,$$

where $\boldsymbol{\beta}_l$ is a parameter that quantifies the effect of locus l. As L increases, the genetic architecture of the trait is more complex. The vector $\boldsymbol{\beta}$ of locus effects is assumed to be normally distributed (Gianola et al., 2020):

$$\boldsymbol{\beta} \,|\, \sigma_{\boldsymbol{\beta}}^2 \sim \mathcal{N}\left(0, \sigma_{\boldsymbol{\beta}}^2\, \boldsymbol{I}_L\right),$$

where $\sigma_{\boldsymbol{\beta}}^2$ is the additive genetic variance and $\boldsymbol{I}_L$ is the identity matrix of size L. Thus, the vector of breeding values is normally distributed with mean 0 and conditional variance matrix $\mathrm{Var}\left(\boldsymbol{u} \,|\, \sigma_{\boldsymbol{\varepsilon}}^2, \gamma_{\boldsymbol{\beta}}^2\right) = \sigma_{\boldsymbol{\varepsilon}}^2\, \boldsymbol{V}$, where

$$\boldsymbol{V} = \gamma_{\boldsymbol{\beta}}^2\, \boldsymbol{Z}\boldsymbol{Z}^T$$

and the variance component $\gamma_{\boldsymbol{\beta}}^2$ is equal to $\gamma_{\boldsymbol{\beta}}^2 = \sigma_{\boldsymbol{\beta}}^2/\sigma_{\boldsymbol{\varepsilon}}^2$. Thanks to the scaling of the columns of $\boldsymbol{Z}$ mentioned above, the prior variance of breeding values is approximately equal to $\sigma_{\boldsymbol{\varepsilon}}^2\gamma_{\boldsymbol{\beta}}^2$, which facilitates the specification of the prior distribution of $\gamma_{\boldsymbol{\beta}}^2$. The columns of $\boldsymbol{Z}$ are not centered here so that the grand mean μ represents the average trait value of a heterozygous individual at all the loci.

2.3 *Prior distributions*

This article uses a Bayesian framework. A vague normal prior distribution is placed on the intercept μ: $\mu \,|\, \sigma_{\boldsymbol{\varepsilon}}^2 \sim \mathcal{N}\left(\widetilde{\mu}, \sigma_{\boldsymbol{\varepsilon}}^2\, \widetilde{v}\right)$, where $\widetilde{\mu}$ is a known constant and $\widetilde{v} \to +\infty$. The residual variance is assumed to follow a prior distribution with mean $\widetilde{\sigma}_{\boldsymbol{\varepsilon}}^2$. Variance components are assumed to follow prior distributions with means $\widetilde{\gamma}_0^2$, $\widetilde{\gamma}_m^2$ and $\widetilde{\gamma}_{\boldsymbol{\beta}}^2$. The prior means $\widetilde{\gamma}_0^2$ and $\widetilde{\gamma}_{\boldsymbol{\beta}}^2$ could range from 0.25 to 4, leading to heritabilities ranging from 0.2 to 0.8 (Johnson and Barton, 2005; Gillespie, 1998, Chap. 5; Barton et al., 2007, Chap. 14). The prior mean $\widetilde{\gamma}_m^2$ could range from 0.0001 to 0.04 (Wray, 1990; Barton et al., 2007, Chap. 14). The unconditional prior variance of breeding values is $\mathrm{Var}\left(\boldsymbol{u}\right) = \widetilde{\sigma}_{\boldsymbol{\varepsilon}}^2\, \mathrm{E}\left(\boldsymbol{V}\right) = \widetilde{\sigma}_{\boldsymbol{\varepsilon}}^2\, \boldsymbol{V}(\widetilde{\gamma}_0^2, \widetilde{\gamma}_m^2)$ when the genetic data are pedigree data, or $\mathrm{Var}\left(\boldsymbol{u}\right) = \widetilde{\sigma}_{\boldsymbol{\varepsilon}}^2\, \boldsymbol{V}(\widetilde{\gamma}_{\boldsymbol{\beta}}^2)$ when the genetic data are genotyping data.

3. Sampling methods

The objective of the sample survey is to estimate the breeding values of the individuals belonging to some target set $\mathcal{T}$, or to estimate locus effects. Without loss of generality, the population is assumed to be the union of $\mathcal{C}$ and $\mathcal{T}$. Choosing the design of the survey consists of choosing the sample, that is, n individuals in $\mathcal{C}$ to be phenotyped.

3.1 *Stratified sampling*

The population may be sampled using SS. First, the individuals of the candidate set are divided into n clusters using a hierarchical clustering. When the genetic data are pedigree data, the dissimilarity $\delta_{ii'}$ between individuals i and i' is equal to $\delta_{ii'} = \sqrt{\boldsymbol{V}_{ii}\boldsymbol{V}_{i'i'} - \boldsymbol{V}_{ii'}}$, where $\boldsymbol{V} = \boldsymbol{V}(\widetilde{\gamma}_0^2, \widetilde{\gamma}_m^2)$. When the genetic data are genotyping data, it is the Euclidian distance calculated from $\boldsymbol{M}$ (Isidro et al., 2015). The design is then generated by sampling one individual in each cluster independently and uniformly.

3.2 *D optimality*

3.2.1 *D criteria.* Sampling designs may be generated by maximizing the D criterion for breeding values (Atkinson and Donev, 1992, Chap. 6; Chaloner and Verdinelli, 1995; Verdinelli, 2000). This criterion quantifies estimation errors. When breeding values are estimated by their posterior mean, the matrix of squared errors is equal to $\{\boldsymbol{u} - \mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y})\}\{\boldsymbol{u} - \mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y})\}^T$. Since the true breeding values are unknown and since observations are not available when the survey is designed, squared errors are averaged over the possible values of these variables, yielding the matrix

$$\int \{\boldsymbol{u} - \mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y})\}\{\boldsymbol{u} - \mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y})\}^T p(\boldsymbol{u}, \boldsymbol{y}) d\boldsymbol{u} d\boldsymbol{y},$$

where $p(\boldsymbol{u}, \boldsymbol{y})$ is the joint density of $\boldsymbol{u}$ and $\boldsymbol{y}$. This matrix is equal to the average posterior variance matrix of breeding values:

$$\int \{\boldsymbol{u} - \mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y})\}\{\boldsymbol{u} - \mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y})\}^T p(\boldsymbol{u} \mid \boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{u} d\boldsymbol{y} = \int \mathrm{Var}(\boldsymbol{u} \mid \boldsymbol{y})\, p(\boldsymbol{y}) d\boldsymbol{y}.$$

The D criterion for breeding values is based on the determinant of this matrix:

$$D_{\boldsymbol{u}} = \ln[\det\{\int \mathrm{Var}(\boldsymbol{u} \mid \boldsymbol{y})\, p(\boldsymbol{y}) d\boldsymbol{y}\}].$$

When variance components have low prior variance, it is easier to compute, since it is approximately equal to (online Appendix A)

$$D_{\boldsymbol{u}} \approx \ln\{\det(\boldsymbol{V}^{-1} + \boldsymbol{X}^T \boldsymbol{Q} \boldsymbol{X})\},$$

where $\boldsymbol{Q} = (\boldsymbol{I}_n - \boldsymbol{J}_n/n)$, $\boldsymbol{J}_n$ is the square matrix of ones of size n, and $\boldsymbol{V}$ is equal to $\boldsymbol{V}(\widetilde{\gamma}_0^2, \widetilde{\gamma}_m^2)$ or to $\boldsymbol{V}(\widetilde{\gamma}_{\boldsymbol{\beta}}^2)$ depending on the experimental setup. This expression assumes that the prior variance matrix of breeding values is invertible, which is often the case. It may be singular when, for example, two individuals are identical twins (or have the same genotype at all the loci), when the individuals born at generation 0 have been selected, or when $L < N$ (Henderson, 1985, 1986; VanRaden, 2008; Fernando et al., 2016).

Alternatively, sampling designs may be generated by maximizing the D criterion for locus effects, which is equal to

$$D_{\boldsymbol{\beta}} = \ln[\det\{\int \mathrm{Var}(\boldsymbol{\beta} \mid \boldsymbol{y})\, p(\boldsymbol{y}) d\boldsymbol{y}\}].$$

When the prior variance of $\gamma_{\boldsymbol{\beta}}^2$ is low, this criterion is approximately equal to (online Appendix A)

$$D_{\boldsymbol{\beta}} \approx \ln\{\det(\gamma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_L + \boldsymbol{Z}^T \boldsymbol{X}^T \boldsymbol{Q} \boldsymbol{X} \boldsymbol{Z})\}.$$

When variance components have low prior variance, maximizing $D_{\boldsymbol{u}}$ or $D_{\boldsymbol{\beta}}$ is equivalent to maximizing the following version of the D criterion (online Appendix A):

$$D = \ln\{\det(\boldsymbol{I}_n + \boldsymbol{Q} \boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^T \boldsymbol{Q})\}.$$

Thus, the D criteria for estimating all of the breeding values or all of the locus effects can be optimized simultaneously.

3.2.2 *Genetic algorithm.* The D criterion can be maximized using the genetic algorithm STPGA (Akdemir et al., 2015; Akdemir and Isidro-Sánchez, 2019). This algorithm makes a population of designs evolve so that D values increase in the population over iterations. At each iteration, designs with large D values are selected in the population and produce offspring that forms the next generation. Offspring may not be identical to their parents because of mutation and recombination. Inefficient designs that were visited recently may be removed from the next generation (Tabu search). Designs generated using regression methods may be introduced in the next generation.

This algorithm can be implemented using the function GenAlgForSubsetSelectionNoTest of the `R` package `STPGA` (Akdemir, 2018). In this article, each design of the initial population is generated by SS independently (Section 3.1). Maximizing D using `STPGA` is computationally more efficient than maximizing $D_{\boldsymbol{u}}$ since D involves a matrix of dimension $n \times n$ only; it is more efficient than maximizing $D_{\boldsymbol{\beta}}$ when $n < L$.

4. Sampling designs for pedigrees

4.1 *Explicit result*

When the genetic data are pedigree data and when the prior variances of γ_0^2 and γ_m^2 are low, if the breeding values of the selected individuals are not correlated a priori and have maximal prior variance among $\mathcal{C}$, then the design is D-optimal over the set of all designs given $\mathcal{C}$ and n (online Appendix B).

4.2 *Simulation study*

A simulation study was carried out to check that `STPGA` can optimize the D criterion. Its `R` code can be found in Supporting Information. In this study, the population consisted of a theoretical pedigree comprising 80 individuals, which reproduced by outcrossing and which were divided into 10 unrelated families (see online Appendix C for more information). The candidate and target sets both comprised all the individuals. Sample size was either equal to 10 or to 70. Breeding values were modeled using the pedigree. In total, 80 design searches were carried out with different values of variance components in cases when optimal/efficient designs were known. For example, when $n = 10$, sampling the individuals of the last generation was optimal (Section 4.1). `STPGA` performed rather well since it always generated designs as efficient as the reference designs, except in two cases.

4.3 *Sampling a maize pedigree*

4.3.1 *Experimental setup.* The influence of mating system and mutation on sampling was studied using a real maize population. To study the genetic mechanisms underlying the response to selection, a divergent selection experiment for maize flowering time was carried out in France (Durand et al., 2015). It yielded four early and four late subpopulations for flowering time. For the sake of simplicity, we only considered one early subpopulation and one late subpopulation here, that were derived from a seed lot of the maize inbred line F252 by successively selecting and selfing the earliest and the latest plants at each generation. Each subpopulation had a single and different ancestor. Both subpopulations formed the target set that comprised 163 plants. These plants were dead and could not be phenotyped. However, they had been selfed and the resulting seeds could be sown to produce plants that could be phenotyped. These seeds, one for each plant of the target set, formed the candidate set of plants that could be phenotyped, which comprised 163 plants. The population included a total of $N = 326$ plants that were distributed in $G + 1 = 17$ generations (Figure 1). It was structured since it comprised the early and late subpopulations. The pedigree of the population and some genotyping data were available, while other genotyping data remained to be collected. Scientists were interested in estimating both breeding values and locus effects. We studied how to sample $n = 50$ plants in the candidate set.

In order to study the influence of mating system, a theoretical pedigree with outcrossing was considered in addition to the real pedigree with selfing. In this scenario, each plant that was not a founder of the pedigree had two parents. The first one was its parent in the original pedigree with selfing. The second one was an additional plant placed in generation 0, which differed between subpopulations. Thus, all the plants of generations 1 to 16 of a given subpopulation had the same second parent (Figure 2).

4.3.2 *Methods.* Kinship matrices were generated using the pedigree. The prior means of variance components took two possible values: $\widetilde{\gamma}_0^2 = 0.25$ or $\widetilde{\gamma}_0^2 = 4$, $\widetilde{\gamma}_m^2 = 0.0001$ or $\widetilde{\gamma}_m^2 = 0.04$.

The explicit result of Section 4.1 did not allow the construction of optimal designs in this application. For each combination of values of variance components, a design was generated by maximizing the D criterion using `STPGA`, and 100 designs were generated by SS (see online Appendix D for more information). In addition, 100 designs were generated by SRS.

Sampling methods were compared using the D criterion. This criterion assumed that the target set consisted of the whole population, which was not the case in this application. To better assess designs, the maximal and average variance criteria were also computed: $MV_{\boldsymbol{u}} = \max_{i\in\mathcal{T}}\{\int \mathrm{Var}\left(\boldsymbol{u}_i \mid \boldsymbol{y}\right) p(\boldsymbol{y}) d\boldsymbol{y}\}$, $A_{\boldsymbol{u}} = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \int \mathrm{Var}\left(\boldsymbol{u}_i \mid \boldsymbol{y}\right) p(\boldsymbol{y}) d\boldsymbol{y}$, where $|\mathcal{T}|$ was the cardinal of $\mathcal{T}$. These criteria took the target set into account. As they were based on posterior variances, they quantified estimation errors (Section 3.2). Their computation assumed that variance components had low prior variance (online Appendix D).

4.3.3 *Results.* Designs were sensitive to genetic variances and mating system. As the mutational variance increased, SS and DO sampled more plants in late generations (Figures 1, 2 and online Figures D.1-D.4). In addition, this effect was stronger when plants reproduced by outcrossing or when the prior mean of the additive genetic variance component $\widetilde{\gamma}_0^2$ was low. For example, when plants reproduced by selfing and $\widetilde{\gamma}_0^2 = 0.25$, DO sampled more plants in early (respectively, late) generations when $\widetilde{\gamma}_m^2 = 0.0001$ (respectively, $\widetilde{\gamma}_m^2 = 0.04$) (Figure 1).

SS and DO tended to select individuals whose breeding values were weakly correlated a priori and had large prior variances (online Tables D.1-D.3). For example, when plants reproduced by selfing and when $\widetilde{\gamma}_m^2 = 0.0001$, they sampled more plants in early generations, presumably because the correlation between breeding values was high in late generations.

SS and DO had better D, $MV_{\boldsymbol{u}}$ and $A_{\boldsymbol{u}}$ values than SRS, except for the $A_{\boldsymbol{u}}$ criterion when plants reproduced by selfing, $\widetilde{\gamma}_0^2 = 0.25$ and $\widetilde{\gamma}_m^2 = 0.0001$ (Figure 3, online Figures D.5 and D.6). SS had a lower efficiency variability than SRS. SS and DO efficiencies were larger in the presence of selfing.

5. Sampling designs for association studies

5.1 *Explicit result*

When the genetic data are genotyping data, when the prior variance of $\gamma_{\boldsymbol{\beta}}^2$ is low, and when $L \leq n$, if at each locus, half of the selected individuals has genotype -1 and the other half has genotype 1, and if locus effects are not correlated a posteriori (when $L > 1$), then the design is D-optimal over the set of all designs given $\mathcal{C}$ and n (online Appendix B). Locus effects are

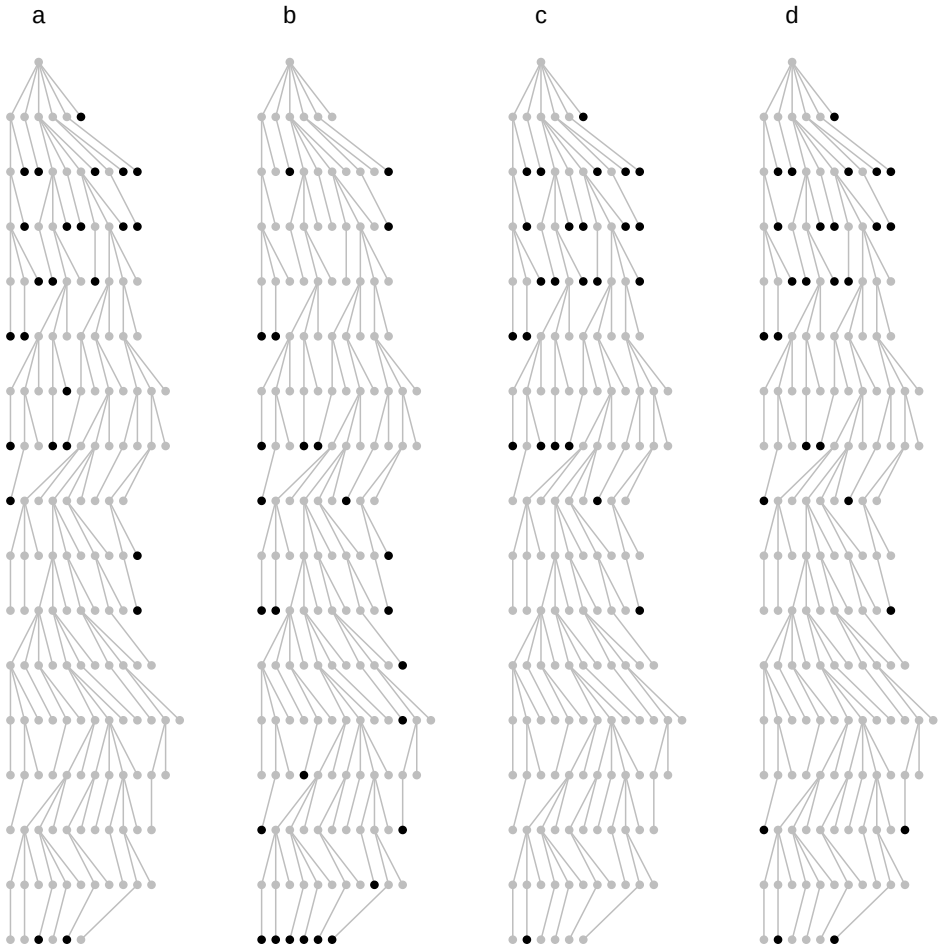

Figure 1. Sampling designs for the early maize subpopulation when plants reproduce by selfing. Each line corresponds to a generation, with generation 0 (respectively, 16) at the top (respectively, bottom) of the pedigree. Vertices are plants, and edges connect parents and their offspring. The candidate set comprises the tips of ancestral lines. Sampled (respectively, not sampled) plants are in black (respectively, gray). The designs were generated by *D*O with parameter values $\widetilde{\gamma}_0^2 = 0.25$ and $\widetilde{\gamma}_m^2 = 0.0001$ (a), $\widetilde{\gamma}_0^2 = 0.25$ and $\widetilde{\gamma}_m^2 = 0.04$ (b), $\widetilde{\gamma}_0^2 = 4$ and $\widetilde{\gamma}_m^2 = 0.0001$ (c), $\widetilde{\gamma}_0^2 = 4$ and $\widetilde{\gamma}_m^2 = 0.04$ (d). The corresponding designs for the late subpopulation can be found in online Appendix D.

not correlated a posteriori when the matrix $\boldsymbol{Z}^T\boldsymbol{X}^T\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Z}$ is diagonal, that is, when the genotypes of the selected individuals at different loci are not correlated.

5.2 *Simulation study*

5.2.1 *Experimental setup.* The influence of sample size and the genetic architecture of the trait on sampling was studied by means of simulations. The `R` code of this study can be found in Supporting Information. For the sake of simplicity, the experimental setup had a smaller size than real association studies: the population was composed of 16 individuals, genotyped at 18 loci. The candidate and target sets both consisted of the whole population.

The population was divided into 4 subpopulations comprising 4 individuals. Allele frequencies varied between subpopulations (see online Appendix E for more information). Each locus was assumed to be at Hardy-Weinberg equilibrium in each subpopulation. Genotypes were simulated independently between individuals and between loci, that is, there was no linkage disequilibrium.

Sample size took three possible values: $n = 4$, $n = 8$ or $n = 12$. The trait was assumed to be affected either by $L = 2$ loci or by all the loci ($L = 18$). The prior mean of the additive genetic variance component $\widetilde{\gamma}_{\boldsymbol{\beta}}^2$ was either equal to 0.25 or to 4.

5.2.2 *Methods.* For each sample size, 100 designs were generated by SRS. For each L value and each sample size, 100 designs were generated by SS. When $L = 2$, and when $n = 4$ or 8, *D*O designs were generated using the explicit result of

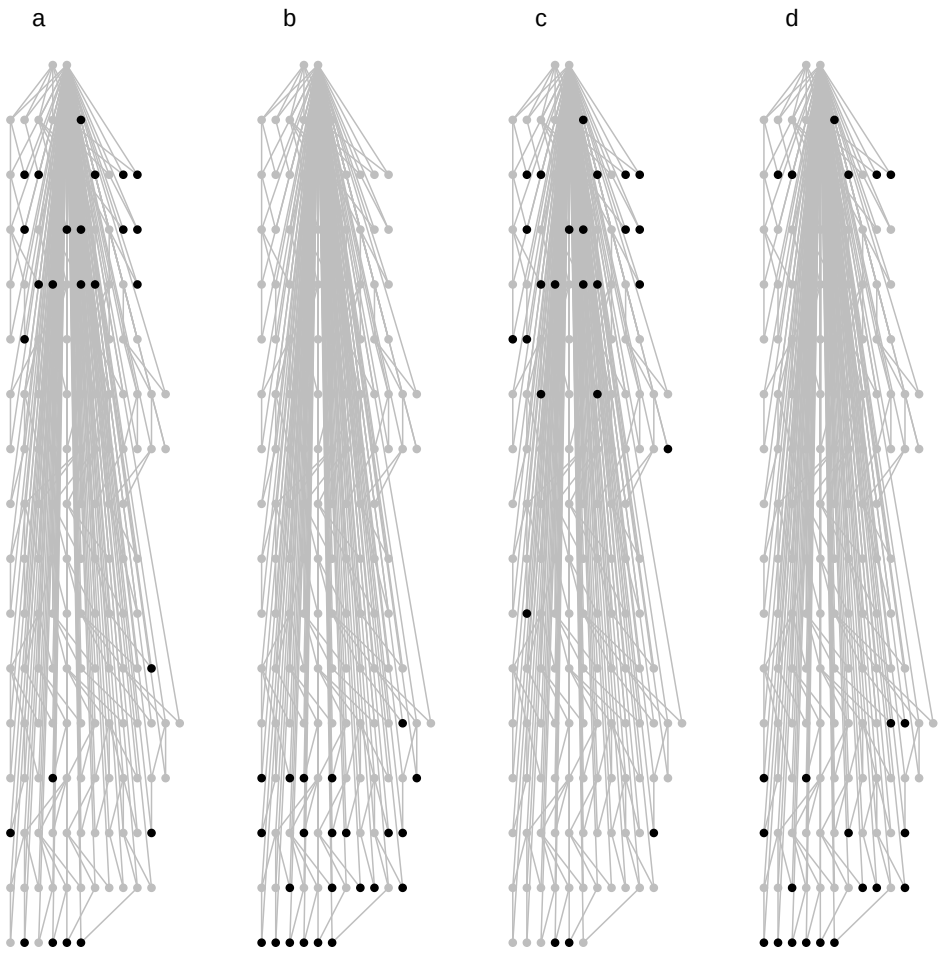

Figure 2. Sampling designs for the early maize subpopulation when plants reproduce by outcrossing. Each line corresponds to a generation, with generation 0 (respectively, 16) at the top (respectively, bottom) of the pedigree. Vertices are plants, and edges connect parents and their offspring. The candidate set comprises the tips of ancestral lines. Sampled (respectively, not sampled) plants are in black (respectively, gray). The designs were generated by *D*O with parameter values $\widetilde{\gamma}_0^2 = 0.25$ and $\widetilde{\gamma}_m^2 = 0.0001$ (a), $\widetilde{\gamma}_0^2 = 0.25$ and $\widetilde{\gamma}_m^2 = 0.04$ (b), $\widetilde{\gamma}_0^2 = 4$ and $\widetilde{\gamma}_m^2 = 0.0001$ (c), $\widetilde{\gamma}_0^2 = 4$ and $\widetilde{\gamma}_m^2 = 0.04$ (d). The corresponding designs for the late subpopulation can be found in online Appendix D.

Section 5.1. For the other parameter values, *D*O designs were generated by optimizing the *D* criterion using `STPGA` (see online Appendix E for more information).

Sampling methods were compared using the *D* criterion. To better quantify estimation errors for breeding values and locus effects, the maximal and average variance criteria ($MV_{\boldsymbol{u}}$, $A_{\boldsymbol{u}}$, $MV_{\boldsymbol{\beta}}$ and $A_{\boldsymbol{\beta}}$) were also computed for these parameters. For example, the $A_{\boldsymbol{u}}$ criterion was equal to $A_{\boldsymbol{u}} = \frac{1}{N}\sum_i \int \mathrm{Var}\left(\boldsymbol{u}_i \,|\, \boldsymbol{y}\right) p(\boldsymbol{y}) d\boldsymbol{y}$. The computation of these criteria assumed that the additive genetic variance component had low prior variance (online Appendix E).

Estimation errors when $n = 4$ and $\widetilde{\gamma}_{\boldsymbol{\beta}}^2 = 4$ were also quantified for each sampling method by simulating trait values. The mean squared error of estimates and the mean correlation between estimated and true parameter values were estimated for breeding values and locus effects (online Appendix E).

5.2.3 *Results.* Unlike SRS, SS and *D*O always sampled individuals in all the subpopulations (Figure 4, online Figures E.1 and E.2). When $L = 2$ and $n = 4$, *D*O sampled the genotypes $(1, 1)$, $(1, -1)$, $(-1, 1)$ and $(-1, -1)$ (online Figure E.1). When $L = 2$ and $n = 8$, it sampled these genotypes twice. It was not very sensitive to $\widetilde{\gamma}_{\boldsymbol{\beta}}^2$ (Figure 4 and online Figure E.2).

Optimized designs tended to improve the estimation of both breeding values and locus effects (Figures 5, 6, online Figures E.3-E.6 and online Table E.1). As sample size increased and the genetic architecture of the trait was more complex, the advantage of optimized designs over SRS diminished. Criterion values were less dispersed for SS than for SRS. The performance of design methods was not very sensitive to $\widetilde{\gamma}_{\boldsymbol{\beta}}^2$ (online Figures E.3 and E.4).

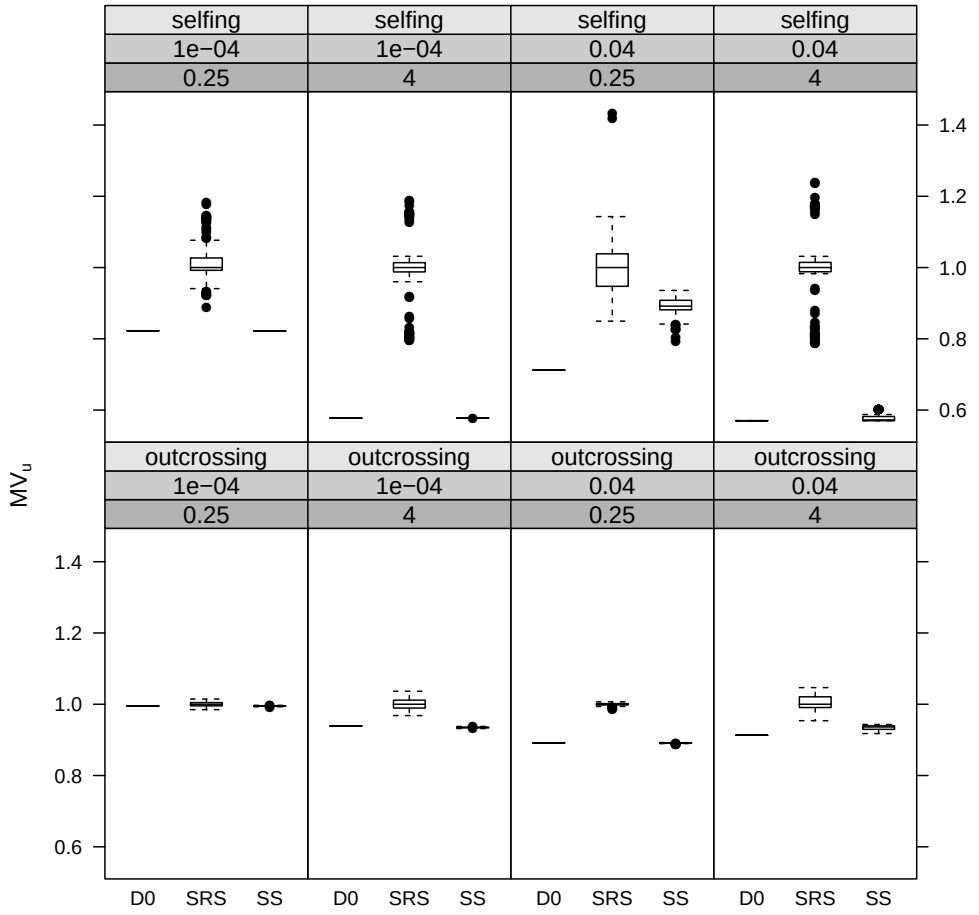

Figure 3. Efficiency of designs for sampling a maize population. Sampling methods are SRS, SS or DO. The prior mean of the additive genetic variance component $\widetilde{\gamma}_0^2$ is either equal to 0.25 or to 4 (lower strip above each plot). The prior mean of the mutational variance component $\widetilde{\gamma}_m^2$ is either equal to 0.0001 or to 0.04 (middle strip above each plot). Plants reproduce either by selfing or by outcrossing (upper strip above each plot). Designs are compared using the $MV_{\boldsymbol{u}}$ criterion, whose values are divided by the median $MV_{\boldsymbol{u}}$ values of SRS.

However, there were a few cases when optimized designs were less efficient. When $L = 18$, all the sampling methods showed similar estimation errors for locus effects, with correlations between estimated and true parameters lower than 0.3 (Figure 6, online Figure E.6 and online Table E.1). In this case, estimation errors were lower for breeding values than for locus effects, with correlations between estimated and true parameters larger than 0.6 (online Table E.1).

SS was less efficient than SRS when $L = 2$ and $n > 4$ (Figures 5, 6 and online Figures E.3-E.6). Unlike DO, SS tended to avoid sampling some genotypes several times by sampling individuals in different clusters, which was not very efficient when $L = 2$ and $n > 4$.

6. Discussion

6.1 *Sampling designs for pedigrees*

This article investigated how to sample pedigrees. Mating system and mutation influenced designs through the prior variance matrix of breeding values. In the maize application, it was preferable to sample more individuals in early generations when the mutational variance component was low and when individuals reproduced by selfing, because individuals from late generations provided similar information since their breeding values were highly correlated. On the contrary, when the mutational variance component was larger, breeding values were less correlated in late generations so that more individuals could be sampled in these generations. It is often important to take account of mutation when analyzing long-term experiments, as, for example,

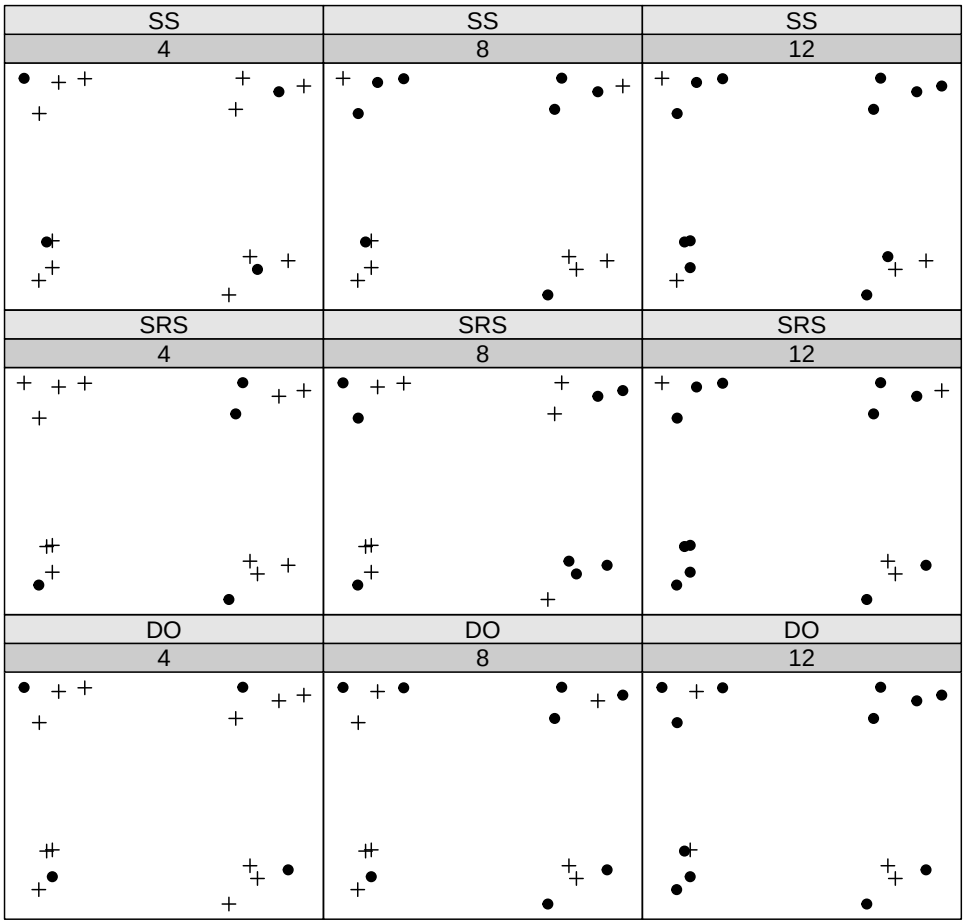

Figure 4. Sampling designs for the theoretical association study. The genetic data are represented using a multidimensional scaling. There are 4 subpopulations, which are located near plot corners. The trait is determined by 18 loci. Sample size is 4, 8 or 12 (lower strip above each plot). Sampling methods are SRS, SS or DO with $\widetilde{\gamma}^2_{\boldsymbol{\beta}} = 4$ (upper strip above each plot). Dots (respectively, crosses) represent sampled (respectively, not sampled) individuals.

in selection experiments where mutation can explain a continued response (Wray, 1990; Durand et al., 2010). Our results show that it is also useful to take account of mutation at the design stage.

In the association study, the additive genetic variance component did not greatly influence sampling designs, in agreement with earlier results (Akdemir et al., 2015). On the contrary, in the maize application, it influenced designs, with low values strengthening the influence of mutation. This is because when breeding values are modeled using the pedigree, variance components determine the relative importance of standing genetic diversity, mutation and information provided by the data.

6.2 *Sampling designs for association studies*

This article investigated how to sample populations in association studies. Optimized designs tended to improve the joint estimation of breeding values and locus effects, all the more as sample size was low (Akdemir and Isidro-Sánchez, 2019). Both objectives were consistent when all breeding values and all locus effects were of interest, but they may be inconsistent when the survey is aimed at inferring some particular breeding values or locus effects.

When the genetic architecture of the trait was rather simple, optimized designs were reminiscent of classical experimental designs. Since they tended to sample homozygous individuals, they were similar to optimal designs for linear regression (Atkinson and Donev, 1992, Chap. 6). Since they tended to reduce the correlation between genotypes at different loci, they were similar to orthogonal factorial designs (Atkinson and Donev, 1992, Chap. 7; Kobilinsky et al., 2017). This is because our model can be considered as a linear model in which loci play the role of independent variables or factors.

When the genetic architecture of the trait was rather complex, optimized designs were not more efficient than SRS for estimating locus effects, and the correlation between estimated and true locus effects was low. This is why the number of

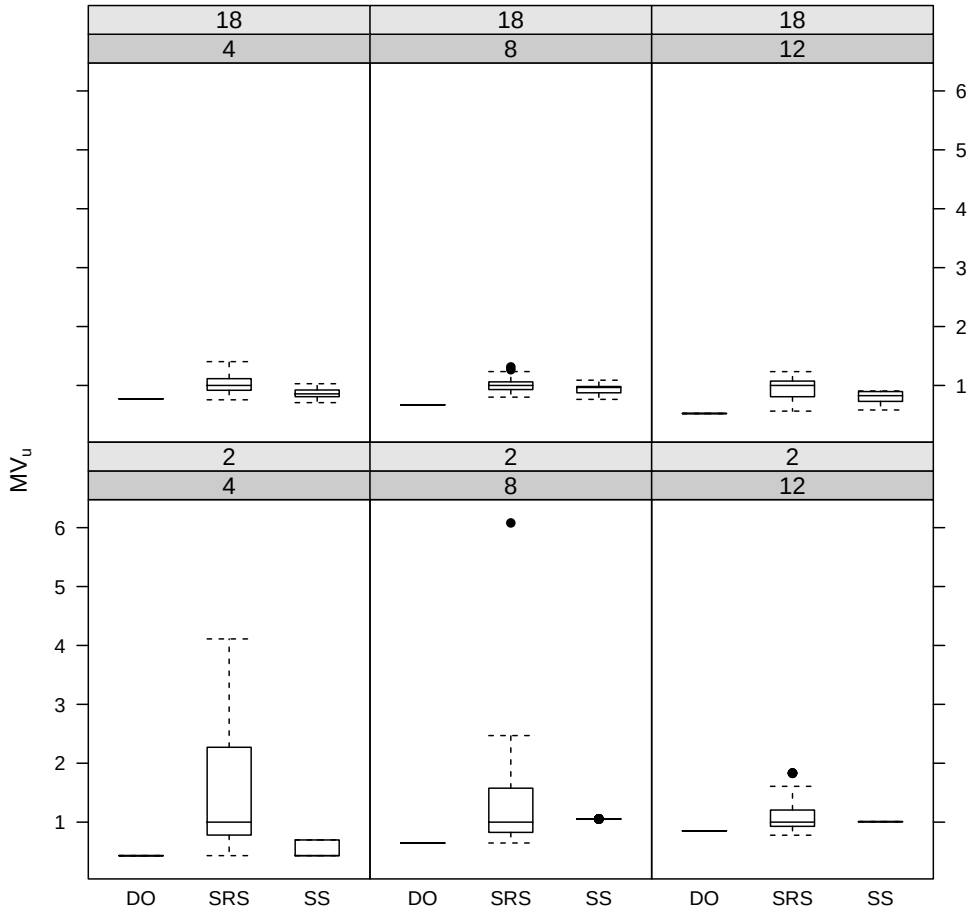

Figure 5. Efficiency of sampling methods in the theoretical association study. Sampling methods are SRS, SS or *D*O. The prior mean of the additive genetic variance component $\widetilde{\gamma}^2_{\boldsymbol{\beta}}$ is equal to 4. The trait is determined by 2 or 18 loci (upper strip above each plot). Sample size is 4, 8 or 12 (lower strip above each plot). Designs are compared using the $MV_{\boldsymbol{u}}$ criterion, whose values are divided by the median $MV_{\boldsymbol{u}}$ values of SRS.

loci was kept not much larger than population size in our association study. On the contrary, optimized designs were more efficient than SRS for estimating breeding values in this case, and estimation errors were lower for breeding values than for locus effects. Locus effects were difficult to estimate because the number of model parameters was larger than sample size. Our results raise the issue of identifying when high-dimensional regression methods provide reliable results (Verzelen, 2012; Giraud, 2014, Chap. 1).

6.3 *Stratified sampling*

This article developed an SS method for sampling related individuals. This method does not require an optimization step. It can be used with both informative or weakly-informative prior distributions for variance components when kinship is modeled using the pedigree. In previous SS methods, the population was divided into large subpopulations and several individuals were sampled in each subpopulation (Isidro et al., 2015). The population was divided here into smaller clusters and one individual was sampled in each cluster. This may more effectively capture structures in the population and reduce the efficiency variability of sampling designs. However, efficiency variability may remain high with our method when clusters are large, as, for example, when sample size is much smaller than population size.

SS performed rather well in our applications. This may be because populations were structured in these examples (Isidro et al., 2015). SS was not very efficient when the genetic architecture of the trait was simple and when sample size was not low. In this case, SS tended to sample more heterozygous individuals than *D*O to increase the genetic diversity in the sample, which was not optimal. However, including heterozygous individuals in the sample could be useful to check model assumptions, in particular the absence of dominance (Atkinson and Donev, 1992, Chap. 20).

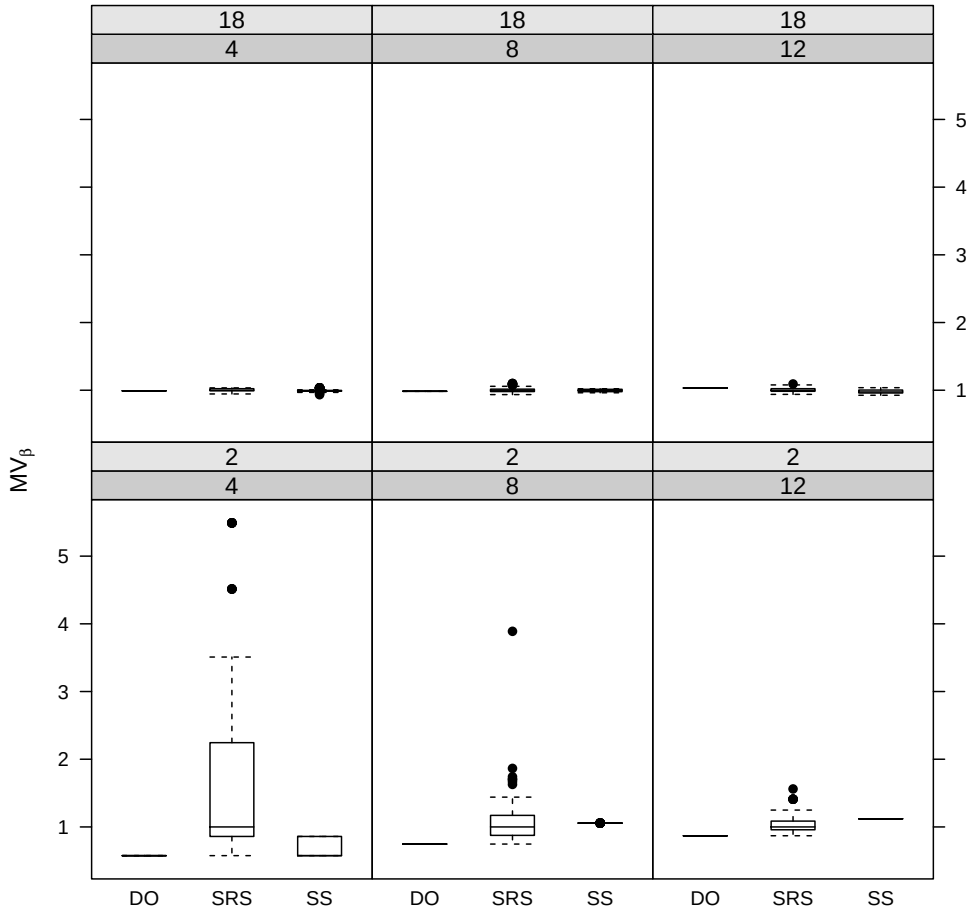

Figure 6. Efficiency of sampling methods in the theoretical association study. Sampling methods are SRS, SS or *D*O. The prior mean of the additive genetic variance component $\widetilde{\gamma}^2_{\boldsymbol{\beta}}$ is equal to 4. The trait is determined by 2 or 18 loci (upper strip above each plot). Sample size is 4, 8 or 12 (lower strip above each plot). Designs are compared using the $MV_{\boldsymbol{\beta}}$ criterion, whose values are divided by the median $MV_{\boldsymbol{\beta}}$ values of SRS.

6.4 *D optimality*

This article also implemented optimal design theory using the D criterion. This criterion quantifies the precision of the analysis. It is more relevant when experimenters are interested in estimating all the breeding values or all the locus effects. When the objective is to estimate some particular breeding values, locus effects or contrasts, criteria that quantify precision for these parameters are more appropriate (Rincent et al., 2017; Akdemir and Isidro-Sánchez, 2019). In the maize application, the D criterion remained relevant even if the target set did not include all the plants, because the candidate and target sets were both large and sampled the pedigree in a similar manner (Akdemir and Isidro-Sánchez, 2019). The D criterion allowed the derivation of explicit results, which helped to assess the genetic algorithm, to understand the properties of optimal designs, and to study the link between designs for estimating breeding values and designs for estimating locus effects.

Optimal design theory requires an optimization step. The D criterion reduces the corresponding computing time since its computation only involves the determinant of an $n \times n$ matrix. The optimizations carried out in this article typically lasted a few minutes. On the contrary, other criteria, such as the CD (coefficient of determination), PEV (prediction error variance), $A_{\boldsymbol{u}}$ or $MV_{\boldsymbol{u}}$ criteria, require the inversion of an $N \times N$ matrix for their evaluation when $L \geq N$ (Rincent et al., 2017; Akdemir and Isidro-Sánchez, 2019). Computing times ranging from a few seconds to a few days have been reported for the optimization of these criteria (Isidro et al., 2015; Rincent et al., 2017).

*D*O was implemented here in the presence of prior information on variance components. This made the design criterion easier to compute. Prior information on variance components is sometimes available: it can rely on general estimates of phenotypic variance or estimates of heritability from earlier parent-offspring studies in the same or in close populations or species (Butler et al., 2014). For example, in the maize application, some phenotypic data have already been collected and

could be used to specify prior distributions for variance components (Durand et al., 2010, 2015). Our implementation of *DO* is expected to work better when some reliable prior information on variance components is available. When this is not the case, it would be more satisfactory to more effectively take the uncertainty on variance components into account, but the design criterion would then have a more complicated expression (Bueno Filho and Gilmour, 2007; Chaloner and Verdinelli, 1995), or to implement a sequential approach.

Acknowledgements

This study was funded by the French BASC project Itemaize. The authors are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing resources. The authors thank the associate editor and the three referees for their helpful comments. They also thank Gail Wagman for improving the English of the article.

This is the accepted version of the following article:
David, O., Le Rouzic, A., Dillmann, C. (2021) Optimization of sampling designs for pedigrees and association studies. Biometrics,
which has been published in final form at https://doi.org/10.1111/biom.13476. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy (http://www.wileyauthors.com/self-archiving).

Data Availability Statement

The data that support the findings in this paper are available from the corresponding author upon reasonable request.

References

Akdemir, D. (2018). *STPGA: Selection of Training Populations by Genetic Algorithm.* R package version 5.2.1.

Akdemir, D. and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Scientific reports* **9,** 1–15.

Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution* **47,** 38.

Amzal, B., Bois, F. Y., Parent, E., and Robert, C. P. (2006). Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association* **101,** 773–785.

Atkinson, A. C. and Donev, A. N. (1992). *Optimum experimental designs.* Oxford University Press, Oxford.

Barton, N., Etheridge, A., and Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology* **118,** 50–73.

Barton, N. H., Briggs, D. E. G., Eisen, J. A., Goldstein, D. B., and Patel, N. H. (2007). *Evolution.* Cold Spring Harbor Laboratory Press, NY.

Brus, D. J. and Heuvelink, G. B. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* **138,** 86–95.

Bueno Filho, J. S. S. and Gilmour, S. G. (2007). Block designs for random treatment effects. *Journal of statistical planning and inference* **137,** 1446–1451.

Butler, D. G., Smith, A. B., and Cullis, B. R. (2014). On the design of field experiments with correlated treatment effects. *Journal of Agricultural, Biological and Environmental Statistics* **19,** 539–555.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science* **10,** 273–304.

Cochran, W. G. (1977). *Sampling Techniques.* Wiley, third edition.

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., Vikram, P., Sansaloni, C., Petroli, C., Akdemir, D., et al. (2016). Genomic prediction of gene bank wheat landraces. *G3: Genes, Genomes, Genetics* **6,** 1819–1834.

David, O., Monod, H., and Amoussou, J. (2000). Optimal complete block designs to adjust for interplot competition with a covariance analysis. *Biometrics* **56,** 389–393.

David, O., van Frank, G., Goldringer, I., Rivière, P., and Turbet Delof, M. (2020). Bayesian inference of natural selection from spatiotemporal phenotypic data. *Theoretical Population Biology* **131,** 100–109.

Del Moral, P. and Doucet, A. (2014). Particle methods: An introduction with applications. *ESAIM: Proceedings* **44,** 1–46.

Durand, E., Tenaillon, M. I., Raffoux, X., Thépot, S., Falque, M., Jamin, P., Bourgais, A., Ressayre, A., and Dillmann, C. (2015). Dearth of polymorphism associated with a sustained response to selection for flowering time in maize. *BMC Evolutionary Biology* **15,** 103.

Durand, E., Tenaillon, M. I., Ridel, C., Coubriche, D., Jamin, P., Jouanne, S., Ressayre, A., Charcosset, A., and Dillmann, C. (2010). Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds. *BMC Evolutionary Biology* **10,** 2.

Fernando, R. L., Cheng, H., and Garrick, D. J. (2016). An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. *Genetics Selection Evolution* **48,** 80.

Gianola, D., Fernando, R. L., and Schön, C.-C. (2020). Inferring trait-specific similarity among individuals from molecular markers and phenotypes with Bayesian regression. *Theoretical Population Biology* **132,** 47–59.

Gillespie, J. H. (1998). *Population genetics: a concise guide*. JHU Press.

Giraud, C. (2014). *Introduction to high-dimensional statistics*, volume 138. CRC Press.

Henderson, C. R. (1985). Best linear unbiased prediction using relationship matrices derived from selected base populations. *Journal of Dairy Science* **68,** 443–448.

Henderson, C. R. (1986). Estimation of singular covariance matrices of random effects. *Journal of Dairy Science* **69,** 2379–2385.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics* **128,** 145–158.

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics* **9,** 166–177.

Johnson, T. and Barton, N. (2005). Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **360,** 1411–1425.

Kobilinsky, A., Monod, H., and Bailey, R. A. (2017). Automatic generation of generalised regular factorial designs. *Computational Statistics & Data Analysis* **113,** 311–329.

Lynch, M. and Walsh, J. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Assocs. Inc., Sunderland, MA.

Müller, P., Sansó, B., and De Iorio, M. (2004). Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association* **99,** 788–798.

Rau, A., Flister, M., Rui, H., and Auer, P. L. (2019). Exploring drivers of gene expression in the cancer genome atlas. *Bioinformatics* **35,** 62–68.

Reeves, C. R. (1993). *Modern heuristic techniques for combinatorial problems*. John Wiley & Sons, Inc.

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics* **130,** 2231–2247.

Van Groenigen, J. W. (2000). The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma* **97,** 223–236.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91,** 4414–4423.

Vazquez, A. I., de los Campos, G., Klimentidis, Y. C., Rosa, G. J., Gianola, D., Yi, N., and Allison, D. B. (2012). A comprehensive genetic approach for improving prediction of skin cancer risk in humans. *Genetics* **192,** 1493–1502.

Verdinelli, I. (2000). A note on Bayesian design for the normal linear model with unknown error variance. *Biometrika* **87,** 222–227.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics* **6,** 38–90.

Wray, N. R. (1990). Accounting for mutation effects in the additive genetic variance-covariance matrix and its inverse. *Biometrics* **46,** 177–186.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist* **56,** 330–338.

Supporting Information

Web Appendices, Tables, Figures, and Codes referenced in Sections 3, 4 and 5, are available with this paper at the Biometrics website on Wiley Online Library.