SMARTER

SMAll RuminanTs breeding for Efficiency and Resilience

Research and Innovation action: H2020 – 772787
Call: H2020-SFS-2017-2
Type of action: Research and Innovation Action (RIA)
Work programme topic: SFS-15-2016-2017
Duration of the project: 01 November 2018 – 31 October 2022

A report for optimum contribution to manage diversity at critical regions and to assist mating design to maximise heterozygosity and expression of heterosis and evaluation of level inbreeding rate across genomic regions and their impact on performance

Ricardo Pong-Wong

Roslin Institute, University of Edinburgh

* Deliverable leader – Contact: ricardo.pong-wong@roslin.ed.ac.uk

DELIVERABLE D5.4

Workpackage N°5

Due date: M36

Actual date: 10/12/2021

Dissemination level: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 772787

About the SMARTER research project

SMARTER will develop and deploy innovative strategies to improve Resilience and Efficiency (R&E) related traits in sheep and goats. SMARTER will find these strategies by: i) generating and validating novel R&E related traits at a phenotypic and genetic level ii) improving and developing new genome-based solutions and tools relevant for the data structure and size of small ruminant populations, iii) establishing new breeding and selection strategies for various breeds and environments that consider R&E traits.

SMARTER with help from stakeholders chose several key R&E traits including feed efficiency, health (resistance to disease, survival) and welfare. Experimental populations will be used to identify and dissect new predictors of these R&E traits and the trade-off between animal ability to overcome external challenges. SMARTER will estimate the underlying genetic and genomic variability governing these R&E related traits. This variability will be related to performance in different environments including genotype-by-environment interactions (conventional, agro-ecological and organic systems) in commercial populations. The outcome will be accurate genomic predictions for R&E traits in different environments across different breeds and populations. SMARTER will also create a new cooperative European and international initiative that will use genomic selection across countries. This initiative will make selection for R&E traits faster and more efficient. SMARTER will also characterize the phenotype and genome of traditional and underutilized breeds. Finally, SMARTER will propose new breeding strategies that utilise R&E traits and trade-offs and balance economic, social and environmental challenges.

The overall impact of the multi-actor SMARTER project will be ready-to-use effective and efficient tools to make small ruminant production resilient through improved profitability and efficiency.

Table des matières

1 Summary

This document presents a report on several aspects that still need to be sorted out before Optimum Contribution Selection (OCS) can be fully integrated with genomic data to enhance its potential to manage the genetic diversity of commercial populations. We derived a new formulation of the OCS which may improve its practical properties. The value of the genomic relationship matrices as indicator of the genetic diversity was assessed. We showed that some of the genomic relationship matrices (GRMs) may have some properties which are inconsistent with our understanding of how genetic diversity arises. Finally, we quantified the trade-off between extending the OCS to improve the management of genetic diversity with any potential loss of response to selection. We showed that the improved approach has little impact on the genetic gain of the trait under selection.

2 Introduction

Genomic Prediction methods use high dense genotyping in the genetic evaluation to improve the accuracy of the predictions. The most popular method is the genomic Best Linear Unbiased predictor (GBLUP) (Garrick 2007; VanRaden 2008) as it has the practical convenience that its implementation is similar to traditional pedigree-based BLUP, but the Numerator Relationship Matrix (NRM) is replaced by a Genomic Relationship Matrix (GRM) calculated with dense genotype information. The highly beneficial effect of genomic selection (GS) on improving accuracy of genetic predictions has prompted its rapid uptake by commercial breeding companies across most livestock species. However, the higher accuracy from GS is likely to promote a higher rate of inbreeding, which may lead to greater loss of genetic variance, inbreeding depression and increased accumulation of deleterious mutations. Nevertheless, the rapid integration of genomic prediction is also bringing some opportunities for better management of the genetic diversity in commercial populations.

An effective approach to control the rate of inbreeding is Optimum Contribution Selection (OCS) which maximises genetic progress while controlling the rate of inbreeding to a given value preset by the breeder. OCS can benefit from the availability of genomic data, more specifically from the calculation of better relationship matrices calculated using dense SNP genotyping arrays (i.e., GRM) that can enhance OCS, resulting in better control of inbreeding and greater genetic gain. Additionally, genomic information offers opportunities to go beyond the standard scope of OCS towards a more customised approach to manage genetic diversity. SNP information can be used to calculate GRMs specific for a region of the genome and then OCS can be applied including separate restrictions on the rate of inbreeding of these regions. This will allow a prioritisation of areas of the genome that need to have a stronger control, and thereby better control of the genetic diversity in them. All of these opportunities arise from the availability of genomic data, which have been collected for the only purpose of performing genetic evaluation.

However, before the genomic data can be used to enhance the full potential of OCS, there are still some issues which need to be sorted out. Although several methods for performing OCS have been proposed (Meuwissen 1997; Kinghorn *et al.* 2002; Pong-Wong and Woolliams 2007; Pong-Wong and Woolliams 2018), there is still room for improvement in order to make them more computationally efficient or more flexible to handle the new opportunities arising in this area. Additionally, it is a general belief that the GRMs are better estimates of the relationship matrices as they improve the accuracy of genomic evaluation, and several methods to calculate them are available. But, although they have proven to be useful for genomic prediction, their value as indicator of genetic relationships among individuals is more uncertain as their usage can lead to inconsistent estimates of the relationship between individuals. Finally, the newly extension of the OCS to enhance its potential to manage diversity may have some conflict with genetic gain and reduce efficacy of selection to produce better animals.

In this report, we performed several studies to address these issues which may affect the integration of OCS with genomic data. We studied a new reformulation of the OCS with the aim to produce more algorithms with better practical properties. We also studied the value of the GRM as an indicator of genomic relationship between the individuals of the population. Finally, we quantified the trade-off, which may be between better management of the genetic diversity and genetic gain.

3 Extending development of optimum contribution selection to manage genetic diversity

3.1 Background

Optimum contribution selection (OCS) is an effective tool for controlling the rate at which coancestry increases in close managed populations. However, despite the potential benefit of OCS, its practical uptake remains low. A reason for this is the limited number of methods implementing OCS and their scope to accommodate for practical conditions. OCS has been implemented using four different approaches based on: (i) relaxed parameter space (Meuwissen 1997), (ii) evolutionary algorithms (Kinghorn *et al.* 2002), (iii) semidefinite programming (Pong-Wong and Woolliams 2007) and (iv) quadratic programming (Pong-Wong and Woolliams 2018). The method based on relaxed parameter space (also known as the Lagrange multiplier method) is fast but the way it deals with the constraints to ensure that the solutions are valid may result in suboptimal solutions (i.e. the solution may not be the best). Additionally, this method allows for only one constraint on coancestry and its modification to add more than one is not trivial. Methods based on evolutionary algorithms are flexible but their convergence cannot be ensured/tested. The methods based on semidefinite programming and quadratic programming guarantee that the solution is optimum (i.e. the solution is the best, given the constraints) and they are flexible allowing to accommodate for multiple constraints on inbreeding. The computational efficiency of the OCS implementation using the quadratic programming approach can be substantially superior to when using semidefinite programming, especially when several constraints in coancestry are to be included in the OCS.

Whilst the quadratic programming is the most desirable approach for implementing the OCS due to its optimality properties and its computational efficiency, there are limitations which still require to be sorted out, one of which is how the constraint in minimum contribution is handled.

The desired behaviour of the OCS when including constraint on minimum contribution would be for the optimisation to determine if a candidate is selected or not, and being assigned a contribution of at least a given magnitude if selected. Practical examples for having a restriction as above would be the case when (i) female candidates contribute to only one offspring/litter in the next generation, (ii) having too few offspring allocated to a male being impractical or (iii) mating allocation to maximise heterozygosity in the offspring. However such assumption on minimum contribution implies that the space of valid contributions would be discontinuous (i.e. 0 if not selected, and within the range [minval: maxval] if selected).

However, this discontinuity in the space of valid solution cannot be handled by the method proposed by Pong-Wong and Woolliams (2018) and the inclusion of a constraint in minimum contribution would mean that the candidate will always be selected and given a contribution of at least the minimum assigned. This problem can be overcome by using an approach similar to the one implemented in the OCS method proposed by Meuwissen (1997). The contributions are optimised without restriction on minimum contribution, and if the magnitude of optimised contributions for some candidates are lower than their minimum required, these contributions are fixed to 0 or its minimum based on some *ad hoc* rules and the optimisation is redone with the remaining candidates. Such recurrent way of re-optimising the contribution when they are outside the valid range of a constraint has shown to lead to suboptimal solutions which may not necessarily maximise the objective function (Pong-Wong and Woolliams 2007).

Here we proposed a novel formulation of the OCS problem based on mixed integer programming (MIP) in order to account for the discontinuity of the valid parameter space.

3.2 Theory of optimum contribution and notation

Let *n* candidates be available for selection and their sex described with the incidence vectors **s** and **d**, where s_i = 1 if candidate *i* is a male and 0 otherwise and d_i = 1- s_i. Their estimated breeding values are in the vector **g**. Let **c** be the vector of the candidates' genetic contributions, where c_i represents half the proportion of offspring from candidate *i.* The values for c_i ranges between [0:0.5] and the sum of contributions with a sex group sums to 0.5.

The expected genetic and inbreeding level in the offspring generation is equal to **c'g** and **c'Gc**/2, respectively (WOOLLIAMS AND THOMSON 1994), where **G** is the relationship matrix for the group of candidates, which can be estimated using either pedigree, high dense marker information or both (NEJATI-JAVAREMI *et al.* 1997; VILLANUEVA *et al.* 2005; VANRADEN 2008). Furthermore, genomic information also allows to calculate relationship matrices for specific regions of the genome so their expected inbreeding can be estimated for the region.

Hence, OCS aims at optimising the candidates' genetic contribution to maximise genetic response while restricting the average level of inbreeding in the offspring generation to increase at a rate lower or equal to a value preset by the breeder (MEUWISSEN 1997; GRUNDY *et al.* 1998). In genetic conservation programmes, the OCS is implemented by replacing the objective function to maximise genetic gain by one which minimises the average loss of genetic diversity in the population (e.g. (CARA *et al.* 2011)). Furthermore, as **G** can be calculated at specific regions of the genome, the OCS can be implemented to include separate restrictions on the rate of inbreeding of these regions. This would allow for a more customised approach to manage genetic diversity prioritizing areas of the genome which require stronger control of the remaining diversity (GÓMEZ-ROMANO *et al.* 2016).

Following GÓMEZ-ROMANO *et al.* (2016) a general OCS formulation considering the genetic diversity of several regions of the genome separately would be the optimisation of **c** to:

Minimise: $h(\mathbf{c})$ (1)

s.t: $\mathbf{s}'\mathbf{c} = 0.5$

$\mathbf{d}'\mathbf{c} = 0.5$

$\mathbf{c} \geq \underline{\mathbf{m}}$

$\mathbf{c} \leq \overline{\mathbf{m}}$

$\frac{\mathbf{c}'\mathbf{G}_j\mathbf{c}}{\mathbf{2}} \leq F_j^*, j = 1, p$

where $h(\mathbf{c})$ is either -**c'g** or **c'Gc**/2, when the goal is to maximise genetic gain or minimise a given coancestry (overall or a specific region), respectively. The first two restrictions ensure that the contribution within sex group sums to 0.5; the next two imposes the restriction on the candidates' minimum ($\underline{\mathbf{m}}$) and maximum ($\overline{\mathbf{m}}$) contribution; and the last ones are restrictions on the average coancestry to be applied separately to *p* different regions of the genome (and it may include the average across the whole region), where $\mathbf{G}_j$ is the relationship for region *j* and F_j^* is its maximum average coancestry to be allowed.

3.3 A quadratic programming approach to OCS

Pong-Wong and Woolliams (2018) showed that the OCS problem with several restrictions of the rate of inbreeding as described in (1) can be solved very efficiently using quadratic programming. The Lagrangian function $\mathcal{L}\big(\mathbf{c},\lambda_s,\lambda_d,\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}},\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}},\boldsymbol{\lambda}_j\big)$ associated to (1) is:

$$\mathcal{L}\big(\mathbf{c},\lambda_s,\lambda_d,\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}},\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}},\boldsymbol{\lambda}_j\big) = h(\mathbf{c}) - \lambda_s(\mathbf{s}'\mathbf{c}-0.5) - \lambda_d(\mathbf{d}'\mathbf{c}-0.5) - \boldsymbol{\lambda}'_{\underline{\boldsymbol{m}}}\big(\mathbf{c}-\underline{\boldsymbol{m}}\big) + \boldsymbol{\lambda}'_{\overline{\boldsymbol{m}}}(\mathbf{c}- \overline{\boldsymbol{m}}) + \textstyle\sum_{j=1}^{p}\lambda_j\big(\mathbf{c}'\mathbf{G}_{\boldsymbol{j}}\mathbf{c}/2-\mathrm{F}_{\boldsymbol{j}}^*\big) \tag{2}$$

where λ_s, λ_d, $\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}}$, $\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}}$, $\boldsymbol{\lambda}_j$ are Lagrangian multipliers, with size 1, 1, *n*, *n*, and *p*, respectively. If minimum and maximum contribution are not constrained, they are removed from the problem and the fourth and fifth terms of $\mathcal{L}\big(\mathbf{c},\lambda_s,\lambda_d,\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}},\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}},\boldsymbol{\lambda}_j\big)$ disappears.

Hence, the Karush-Kuhn-Tucker (KKT) optimality conditions for (1) are:

$$\begin{aligned}
\nabla_{\mathbf{c}}h(\mathbf{c}) - \lambda_s\mathbf{s} - \lambda_d\mathbf{d} - \boldsymbol{\lambda}_{\underline{\boldsymbol{m}}} + \boldsymbol{\lambda}_{\overline{\boldsymbol{m}}} + \Sigma_{j=1}^{p}\big(\lambda_j\mathbf{G}_{\boldsymbol{j}}\mathbf{c}\big) &= \mathbf{0}\\
0.5-\mathbf{s}'\mathbf{c} &= 0\\
0.5-\mathbf{d}'\mathbf{c} &= 0\\
\mathbf{y}_{\underline{\boldsymbol{m}}} - \mathbf{c} + \underline{\mathbf{m}} &= \mathbf{0}\\
\mathbf{y}_{\overline{\boldsymbol{m}}} + \mathbf{c} - \overline{\mathbf{m}} &= \mathbf{0}\\
\big[\boldsymbol{y}_j + \mathbf{c}'\mathbf{G}_{\boldsymbol{j}}\mathbf{c}/2 - F_j^*\big] &= 0_j, j = 1, p\\
\boldsymbol{\Lambda}_{\underline{\boldsymbol{m}}}\mathbf{Y}_{\underline{\boldsymbol{m}}}\mathbf{e} &= \mathbf{0}\\
\boldsymbol{\Lambda}_{\overline{\boldsymbol{m}}}\mathbf{Y}_{\overline{\boldsymbol{m}}}\mathbf{e} &= \mathbf{0}\\
\boldsymbol{\Lambda}_{\boldsymbol{j}}\mathbf{Y}_{\mathrm{j}}\mathbf{e} &= \mathbf{0}
\end{aligned} \tag{3}$$

with $\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}}$, $\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}}$, $\boldsymbol{\lambda}_j$, $\mathbf{y}_{\underline{\boldsymbol{m}}}$, $\mathbf{y}_{\overline{\boldsymbol{m}}}$, $\boldsymbol{y}_j$ are ≥0. The vectors $\mathbf{y}_{\underline{\boldsymbol{m}}}$, $\mathbf{y}_{\overline{\boldsymbol{m}}}$ and $\mathbf{y}_j$ are slack variables associated to the inequality constraints, $\boldsymbol{\Lambda}_{\boldsymbol{x}}$ and $\mathbf{Y}_{\boldsymbol{x}}$ are diagonal matrices containing the values of $\boldsymbol{\lambda}_{\boldsymbol{x}}$ and $\boldsymbol{y}_{\boldsymbol{x}}$ in their diagonal and **e** is a vector of ones.

Defining $\mathrm{R}(\boldsymbol{\theta})$ to be the nine LHS terms of the KKT optimality conditions, the optimum solution for (1) is the roots of $\mathrm{R}(\boldsymbol{\theta})$, and it may be searched iteratively using Newton-Raphson (NR). However, the standard NR approach may lead to unfeasible solutions not fulfilling one or more of the restrictions, so Pong-Wong and Woolliams (2018) proposed to use an interior point algorithm based on the Mehrotra's predictor-corrector algorithm to solve (1). Their results showed that such approach is computationally efficient as well as it leads to the optimum solution given the objective function.

3.4 A mixed integer programming formulation of the OCS to better account the constraint in minimum contribution

3.4.1 Algorithm to maximise genetic gain/genetic diversity accounting for several restriction on coancestry

As stated before, adding a restriction on minimum contribution in (1) assumes that the candidate will be selected. A general assumption for the restriction on $\underline{\mathbf{m}}$, would be for the optimisation: (i) to determine if a candidate is to be selected or not; and (ii) for those which end being selected, to assign a contribution greater or equal than $\underline{m_i}$ (i.e. assuming that the space of valid contribution for a candidate is discontinuous being: 0 if not selected, and a value in the range $[\underline{m_i}, \overline{m_i}]$ if selected).

To do so, we introduce a new set of binary variables in vector $\tilde{\mathbf{c}}$ indicating the selection status for each candidate, where $\tilde{c}_i$ is equal to 1 if individual i is selected as parent (i.e. c_i>0) and 0 otherwise (i.e. c_i =0). Then the equations representing the restrictions on minimum and maximum contribution for candidate *i* would be: $c_i \geq \underline{m_i} * \tilde{c}_i$ and $c_i \leq \tilde{c}_i * \overline{m_i}$. Because $\tilde{c}_i$ takes values 0 or 1, the minimum and maximum contribution would be 0 when the candidate is not selected (i.e. $\tilde{c}_i$=0), and between $[\underline{m_i}, \overline{m_i}]$ if selected (i.e. $\tilde{c}_i$=1). Additionally to ensure that c_i takes value of 0 or 1, an additional constraint is added as: $c_i * (c_i - 1) = 0$.

Hence the MIP formulation for the OCS would the optimisation of ($\mathbf{c}$, $\tilde{\mathbf{c}}$) to**:**

Minimise: $h(\mathbf{c})$ (4)

s.t: $\mathbf{s}'\mathbf{c} = 0.5$

$$\mathbf{d}'\mathbf{c} = 0.5$$

$$\tilde{\mathbf{c}} * (\tilde{\mathbf{C}} - \mathbf{I}) = \mathbf{0}$$

$$\mathbf{c} \geq \underline{\mathbf{m}}\tilde{\mathbf{C}}$$

$$\mathbf{c} \leq \overline{\mathbf{m}}\tilde{\mathbf{C}}$$

$$\frac{\mathbf{c}'\mathbf{G}_j\mathbf{c}}{\mathbf{2}} \leq F_j^*, j = 1, p$$

where $\tilde{\mathbf{C}}$ is a diagonal matrix where its diagonal values are $\tilde{\mathbf{c}}$. The third constraint is a new one to ensure that the solution of $\tilde{\mathbf{c}}$ is binary.

The Lagrangian function for (4) is:

$$\mathcal{L}(\mathbf{c}, \tilde{\mathbf{c}}, \lambda_s, \lambda_d, \boldsymbol{\lambda}_{\tilde{\mathbf{c}}}, \boldsymbol{\lambda}_{\underline{\boldsymbol{m}}}, \boldsymbol{\lambda}_{\overline{\boldsymbol{m}}}, \lambda_j) = h(\mathbf{c}) - \lambda_s(\mathbf{s}'\mathbf{c} - 0.5) - \lambda_d(\mathbf{d}'\mathbf{c} - 0.5) - \tilde{\mathbf{c}}'(\tilde{\mathbf{C}} - \mathbf{I})\boldsymbol{\lambda}_{\tilde{\mathbf{c}}} - \boldsymbol{\lambda}'_{\underline{\boldsymbol{u}}}(\mathbf{c} - \underline{\mathbf{m}}\tilde{\mathbf{C}}) + \boldsymbol{\lambda}'_{\overline{\boldsymbol{u}}}(\mathbf{c} - \overline{\mathbf{m}}\tilde{\mathbf{C}}) + \sum_{j=1}^{p} \lambda_j(\mathbf{c}'\mathbf{G}_j\mathbf{c}/2 - \mathrm{F}_j^*)$$

Hence the KKT conditions for optimality are:

$$\begin{aligned}
\nabla_{\mathbf{c}}\mathcal{L}\left(\mathbf{c},\tilde{\mathbf{c}},\lambda_s,\lambda_d,\boldsymbol{\lambda}_{\tilde{\mathbf{c}}},\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}},\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}},\lambda_j\right)=\nabla_{\mathbf{c}}h(\mathbf{c})-\ \lambda_s\mathbf{s}-\ \lambda_d\mathbf{d}-\boldsymbol{\lambda}_{\underline{\boldsymbol{u}}}+\ \boldsymbol{\lambda}_{\overline{\boldsymbol{u}}}+\Sigma_{j=1}^{p}\left(\lambda_j\mathbf{G}_{\boldsymbol{j}}\mathbf{c}\right) &= \mathbf{0}\\
\nabla_{\tilde{\mathbf{c}}}\mathcal{L}\left(\mathbf{c},\tilde{\mathbf{c}},\lambda_s,\lambda_d,\boldsymbol{\lambda}_{\tilde{\mathbf{c}}},\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}},\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}},\lambda_j\right)=-\left(2\tilde{\mathbf{C}}-\mathbf{I}\right)\boldsymbol{\lambda}_{\tilde{\mathbf{c}}}+\boldsymbol{\Lambda}_{\underline{\boldsymbol{m}}}\underline{\mathbf{u}}-\boldsymbol{\Lambda}_{\overline{\boldsymbol{m}}}\overline{\mathbf{u}} &= \mathbf{0}\\
0.5-\mathbf{s}'\mathbf{c} &= 0\\
0.5-\mathbf{d}'\mathbf{c} &= 0\\
-\left(\tilde{\mathbf{C}}-\mathbf{I}\right)\tilde{\mathbf{c}} &= \mathbf{0}\\
\mathbf{y}_{\underline{\boldsymbol{m}}}-\mathbf{c}+\underline{\mathbf{u}}\tilde{\mathbf{C}} &= \mathbf{0}\\
\mathbf{y}_{\overline{\boldsymbol{m}}}+\mathbf{c}-\overline{\mathbf{u}}\tilde{\mathbf{C}} &= \mathbf{0}\\
\Sigma_{j=1}^{p}\left(\boldsymbol{y}_j+\mathbf{c}'\mathbf{G}_{\boldsymbol{j}}\mathbf{c}/2-2\mathrm{F}_{\boldsymbol{j}}^{*}\right) &= 0_j, j=1,p\\
\left(\lambda_{\underline{m}}*y_{\underline{m}}\right)_i &= 0_i, i=1,n\\
(\lambda_{\overline{m}}*y_{\overline{m}})_i &= 0_i, i=1,n\\
\left(\lambda_j*y_j\right) &= 0_j, j=1,p\\
\left(\boldsymbol{\lambda}_{\underline{\boldsymbol{m}}},\boldsymbol{y}_{\underline{\boldsymbol{m}}}\right) &\geq \mathbf{0}\\
(\boldsymbol{\lambda}_{\overline{\boldsymbol{m}}},\boldsymbol{y}_{\overline{\boldsymbol{m}}}) &\geq \mathbf{0}\\
\left(\lambda_j,y_j\right) &\geq 0, j=1,p
\end{aligned}$$

where the $\boldsymbol{\lambda}_x$, and $\mathbf{y}_x$ are the Lagrangian multipliers and slack variables similar as in the case of OCS (1).

3.4.2 Algorithm for optimising mating to maximise heterozygosity/minimise coancestry

Similarly, the MIP approach can be used to optimise mating strategies to maximise heterozygosity (given that the candidates have been selected and their contributions assigned).

Let assume n_s and n_d being the number of selected males and females, so the number of possible matings is $n_m = n_s * n_d$. The number of matings needed for each candidate is already assigned and stored in the vector $\mathbf{o}$. The variable **m** is a vector of size n_m with row *ij* containing the mating status between male *i* and female *j* (1 if assigned, and 0 otherwise). **P** is an incidence matrix (size (n_m x ($n_s + n_d$)) indicating the male and female which are involved in a given mating. Finally, **g** is a vector with row *ij* containing the expected homozygosity or average coancestry of offspring.

Hence, the optimisation of the mating to increase heterozygosity in the offspring can be formulated as the optimisation of **m** to:

minimise $\mathbf{g}'\mathbf{m}$ (5)

s.t. $\mathbf{P}'\mathbf{m} = \mathbf{o}$

$\mathbf{m} * (\mathbf{M} - \mathbf{I}) = \mathbf{0}$

where $\mathbf{M}$ is a diagonal matrix with the values of its diagonal being $\mathbf{m}$. The first restriction ensures that the optimisation assigns the correct number of mating expected for each candidate, and the second is to force binary (0/1) solution for $\mathbf{m}$.

3.4.3 Implementation issues and final remarks

The extension of the quadratic programming provides a more versatile reformulation of the OCS problem and it would facilitate its practical implementation in commercial populations. However, while the inclusion of integer variables can be easily formulated as a quadratic programming, the solving of such problem are less trivial and finding efficient algorithms to ensure convergence to the true optimum is still an issue within researchers in the area of optimisation. Probably the IMP OCS proposed here may be of practical use until a stable approach to optimise problem with integer solutions is implemented.

4 The value of genomic relationship matrices to estimate levels of inbreeding

4.1 Background

The beneficial effect of using genomic information in the genetic prediction to improve the accuracy of the genomic estimated breeding values (GEBV) has prompted a rapid intake of the methodology across most commercial livestock species. Genomic information is incorporated in the evaluation through the genomic relation matrix (GRM), calculated with genomic information, and used it to replace the Numerator Relationship Matrix (NRM) in the standard BLUP analysis. Hence, GRMs are believed to be better estimate of the true relationships among individuals, and given that the diagonals of the NRM are equal to 1 plus the inbreeding coefficients for the corresponding individuals, it has been generally accepted that the diagonals of the GRM are 1 plus the realized inbreeding level for the corresponding individuals. However, there are several methods for calculating GRM (e.g. (LI AND HORVITZ 1953; VANRADEN 2008; YANG *et al.* 2011)), and they can result in very different outcomes and the correlations between these estimators vary greatly and can even be negative (e.g. (ZHANG *et al.* 2015; KARDOS *et al.* 2016)). Thus, there is still an unresolved debate on which are the best measures of inbreeding when estimated using genomic information.

Here, we present a summary of the study carried out to compare genomic inbreeding coefficients obtained from using different methods to calculate the GRM. For a more detailed description of the full study see the article given in appendix 1, which is already published in Genetic Selection Evolution (VILLANUEVA *et al.* 2021).

4.2 Methods for calculating GRM and predictions of their expected level of inbreeding

Here we studied the estimated genomic inbreeding of five different methods to calculate the GRM: F_{NEJ} (NEJATI-JAVAREMI *et al.* 1997), $F_{L\&H}$ (LI AND HORVITZ 1953), F_{VR1} (VANRADEN 2008), F_{VR2}

(VanRaden 2008) and F_{YAN} (Yang *et al.* 2011). Genomic inbreeding is defined as the diagonal of the GRM minus one, so they can calculated as:

$$F_{NEJ} = \frac{\sum_{k=1}^{S}(\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij_k})/2}{S} - 1$$

$$F_{L\&H} = \frac{SF_{NEJ} - \sum_{k=1}^{S}[1-2p_{k(0)}(1-p_{k(0)})]}{S-\sum_{k=1}^{S}[1-2p_{k(0)}(1-p_{k(0)})]}$$

$$F_{VR1} = \frac{\sum_{k=1}^{S}(x_k - 2p_{k(0)})^2}{2\sum_{k=1}^{S} p_{k(0)}(1-p_{k(0)})} - 1$$

$$F_{VR2} = \frac{1}{S}\sum_{k=1}^{S}\frac{(x_k - 2p_{k(0)})^2}{2p_{k(0)}(1-p_{k(0)})} - 1$$

$$F_{YAN} = \frac{1}{S}\sum_{k=1}^{S}\frac{x_k^2 - (1+2p_{k(0)})x_k + 2p_{k(0)}^2}{2p_{k(0)}(1-p_{k(0)})}$$

where S is the number of SNPs in the chip array; I_{ij_k} is the allelic similarity of alleles *i* and *j* at SNP *k,* being 1 if the individual is homozygote and 0 otherwise; x_k is the genotype score at SNP *k* equal to 0, 1, 2 for genotypes AA, AB and BB, respectively; and $p_{k(0)}$ is the frequency of the reference allele B, at the base reference population, *0*.

Hence, the average genomic inbreeding is a reflexion of the changes in genotype frequencies in the population (regardless it is due to drift or selective forces). Then the expected average value can be estimated based on a single SNP model. Assuming that mating is random so genotype frequencies will be in HWE, the deterministic prediction of the expected average population genomic inbreeding time *t* calculated with the GRM x, ($E(Fx_t)$) is:

$$E(Fx_t) = \sum_{g=AA,AB,BB} freq(g_t) * Fx_{g_0} \qquad (6)$$

where $freq(g_t)$ is the HWE frequency of genotype g at time *t*; Fx_{g_0} is the genomic inbreeding of method x, for an individual with genotype g, calculated assuming the initial allele frequency at the reference base population.

Since the prediction is based on a single SNP model, F_{VR1} and F_{VR2} will have the same prediction.

4.3 Expected genomic inbreeding as an indicator of loss of genetic variance

Under the infinitesimal model, the average inbreeding in the population is an indicator of the proportion of the genetic variance loss in the population relative to the initial starting variance in the base population. Hence, a fully inbred population would have an inbreeding of 1 and all its initial genetic variance has been lost. Here we assessed the behaviour of the different estimator of genomic inbreeding as an indicator of change in genetic variance in the population.

Assuming HWE genotype frequency, the genetic variance accounted by a SNP will be 2p(1-p), so the maximum variance would be when the SNP allele frequency is 0.5 and it declines as the frequency deviates from 0.5. Then, the change in genetic variance across time is a direct reflexion of changes in the SNP allele frequencies, with the genetic variance decreasing as the minor allele frequency becomes smaller but increasing if the frequency is moved closer towards 0.5. Hence, if the genomic inbreeding is an indicator of change in genetic variance, the average genomic inbreeding should increase as the allele frequencies are moved towards zero or one (i.e, starting MAF_0 > final MAF_t) and becoming 1 when the SNP is fixed (indicating that all the starting genetic variance was lost). Conversely, when the allele is moved closer to 0.5, an increase in genetic variance is expected so the genomic inbreeding should become negative (a situation which does not happen when the inbreeding is calculated using pedigree information, assuming the infinitesimal model). Additionally, as inbreeding indicates change in variance relative to the reference base population, the upper limit of genomic inbreeding should be 1 (as it is not possible to lose more than 100% the starting variance) but its lower limit may be –α, reflecting that the starting allele frequency could have been infinitesimally small and it was moved towards 0.5 at time t (i.e. it is possible to gain an infinitesimal large proportion of initial variance if it started being very small).

Here, the value of the different estimates of genomic inbreeding was assessed based on its behaviour as an indicator of change in genetic variance. For it, the expected genomic inbreeding was predicted using (6) across the whole range of starting frequencies (p_0) and the final frequencies at time t (p_t).

The prediction on the expected genomic inbreeding for three different methods for calculating GRM are shown in Figure 1. The heatmap showed that the E($F_{L\&H}$) can take values between]–α , 1], with positive values of F occurring when MAF_0 > MAF_t and negative when MAF_t is driven towards 0.5. Hence, $F_{L\&H}$ seems to be a valid indicator of change in genetic variance. However, this is not the case for F_{VR} and F_{YAN}, as they may yield estimates which can be greater than 1. This is an invalid estimate as it is inconsistent that a population can lose more genetic variance than they initially started with. Additionally, F_{YAN} shows that, in average, a population is not expected to have gain in genetic variance even when frequency is driven towards 0.5.

The most striking trends are observed for F_{VR} (where the average genomic inbreeding can be negative, positive between [0:1] and > 1). Firstly, E(F_{VR}) can be > 1, wrongly implying that the population has lost more variance that it initially started with. Secondly, the situations where E(F_{VR}) < 0 happen when MAF_0 > MAF_t (i.e. the estimate indicates that the genetic variance has increased, but in fact, the population is losing variance). Thirdly, a significant proportion of the situations when E(F_{VR}) > 0 occurs when the frequency has been moved towards 0.5 (i.e. the estimate indicates that genetic variance is being lost , but in fact the genetic variance has increased). Furthermore, for all situations where p_t > 2*p_o, the estimates lead to the invalid result of E(F_{VR}) > 1 (i.e. genetic variance may be increasing but the estimate suggests that the population has lost even more variance than what it started with).

The trends observed with the deterministic prediction on the different estimates of genomic inbreeding were validated using a real data from a small and highly inbred pig population. The analysis on this population showed that for genomic regions where all SNPs were fixed (i.e. total loss of the genetic variance), E(F_{VR}) can lead to wrong conclusions that genetic variance is increasing. The differences in the magnitude and sign of these estimates of genomic inbreeding has a large impact when using them to calculate inbreeding depression. The results from the real pig population showed that the estimates of inbreeding depression obtained with the different methods can be of opposite sign for as many as 40% of all estimates. The results from the real dataset can be seen in the published article given in appendix 1.

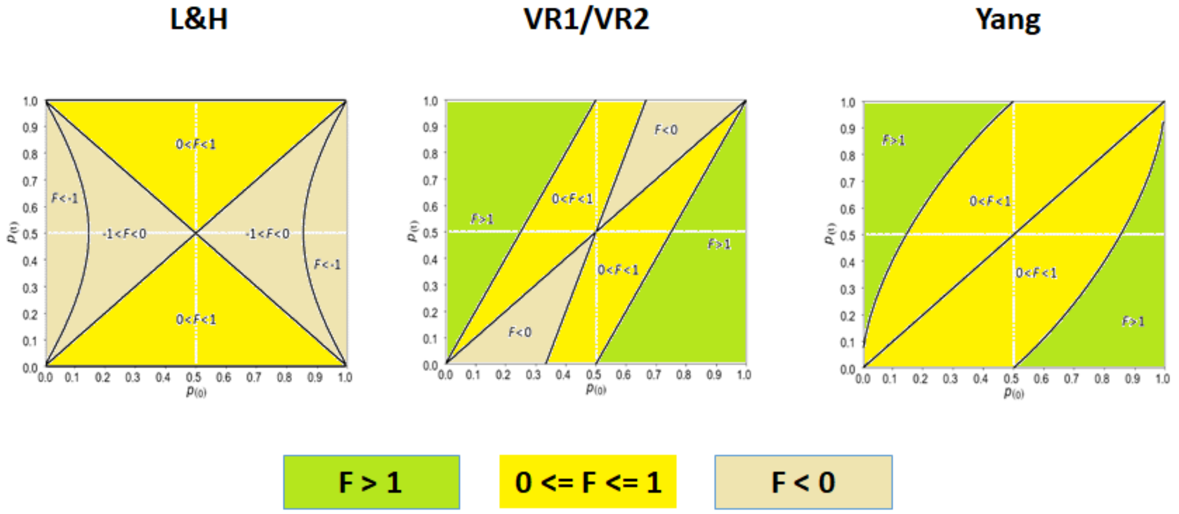

Figure 1. Heatmap of the prediction on expected genomic inbreeding coefficient when using $F_{L\&H}$, F_{VR} and F_{YAN}, across the whole range of starting and final allele frequencies.

4.4 Conclusions

Deterministic predictions of average genomic inbreeding showed that the estimates can be substantially different depending of the method used to calculate the GRM. $F_{L\&H}$ provided estimates which are consistent to what is expected given the changes in genetic variance, with $F_{L\&H}$ < 0 when the population has gained variance and $F_{L\&H}$ > 0 when it has lost variance. However, this was not the case for F_{VR} and F_{YAN}, where estimated genomic inbreeding greater than one can be observed. Moreover, for a large proportion of scenarios, F_{VR} can predict that the population is gaining variance when in fact is losing or predict loss of variance when in fact is gaining. The conclusions were tested using real data, which confirmed the inconsistent estimates of genomic inbreeding observed with the deterministic predictions

5 Changes in Allele Frequencies When Different Genomic Coancestry Matrices Are Used for Maintaining Genetic Diversity

5.1 Background

The implementation of genomic prediction to improve accuracy of the GEBVs requires a large-scale high density genotyping in the candidates to selection. This opens a great opportunity for implementing an OCS, which will come at virtually no extra cost as the genomic information would be available.

A genomic OCS (gOCS) would use the GRM when adding the restriction on the rate of inbreeding to be allowed. Several methods have been proposed to calculate GRMs (e.g. (LI AND HORVITZ 1953; VANRADEN 2008; YANG *et al.* 2011)), which have shown great potential to improve the accuracy of GEBV in the genetic evaluation. However, their value to restrict rate of inbreeding with gOCS requires further studies.

GÓMEZ-ROMANO *et al.* (2016) discussed that the GRM used in the OCS would affect the performance of the scheme to preserve genetic variation. After a close inspection of the different methods, they speculated that an OCS using the GRM calculated with the method proposed by LI AND HORVITZ (1953) would drive the SNPs towards intermediate gene frequency, whilst using a GRM calculated with the methods from VANRADEN (2008) would promote to maintain the original frequencies. They went further by predicting that using the GRM from LI AND HORVITZ (1953) would reduce the chances of fixing rare alleles as the OCS would push them towards more intermediate frequencies. More recently, MORALES-GONZÁLEZ *et al.* (2020) showed that the choice of the GRM for the gOCS impacted on the selected candidates and their assigned contributions. More specifically, they showed that the gOCS using the LI AND HORVITZ (1953)'s GRM would maximise the expected heterozygosity in the offspring generation, compared to when using a GRM from VANRADEN (2008) or YANG *et al.* (2011). Such results are yet to be tested assuming a multiple generation scheme.

Here, we present a summary of the study carried out to compare the behaviour of gOCS when using different GRM matrices. We assessed the impact of GRM on pattern of change in gene frequency as an indicator of the changes in genetic variance. For a more detailed description of the full study see the article given in appendix 2, which it is already published in Genes (MORALES-GONZÁLEZ *et al.* 2021).

5.2 Methods

A simulation study was carried out to assess the impact of the GRM used in a gOCS scheme aiming at maximising the maintenance of genetic diversity of a small population across 50 generations.

A population with a genome in linkage disequilibrium (LD) was simulated using a mutation-drift equilibrium approach, where a population with several chromosomes with thousands of SNPs was allowed to evolve with mutations creating new variants and drift driving them to be lost/fixed or increased in frequency. After a large number of generations the population ended with segregating loci in LD. The population was expanded and a set of the animals were randomly chosen to be the base population where the conservation programme started. Thereafter, two sets of over 55,000 segregating SNPs were selected to be the array of genotyped and ungenotyped SNPs: the set of known SNPs were used to manage the population and the set of genotyped was used to evaluate the performance of the conservation scheme.

At a given generation, OCS was applied to optimise the contribution of candidates with the objective of maximising the genetic variance retained in the population. The relationship matrix used in the OCS was a GRM calculated with either the method proposed by Li and Horvitz (1953) (SO_L&H) or with the method 2 proposed by VanRaden (2008) (SO_VR). The GRM was calculated using the set of genotyped SNPs, either using all genotyped SNPs or after filtering for a MAF > 0.05 or MAF >0.25. The effect of the size of the managed population (n=20 and 100) was also evaluated. The conservation programme was carried out for 50 generations.

The criteria of comparison to evaluate the effect of the GRM used in the OCS were Kullback–Leibler (KL) between current and initial frequency (at generation 0). Heterozygosity, distribution of allele frequencies at generation 50 and fixation rate were also used as criteria of comparison.

5.3 Results

Figure 2 shows the heterozygosity and KL deviation criterion from the conservation schemes using SO_L&H and SO_VR to calculate the GRM used in the OCS scheme. Overall, the heterozygosity tended to decrease as the generation progressed, but the reduction was at a much greater rate when using the SO_VR GRM. The scheme using SO_L&H with no filtering or filtering for MAF > 0.05 to calculate the GRM was very efficient in retaining the expected heterozygosity over the course of the scheme and only starting to decrease in the last third of the conservation programme. However, those schemes with the lowest reduction on the average heterozygosity also had the highest KL deviation criterion at generation 50, which indicates that they had the highest divergence (although their expected heterozygosity changes less across the scheme).

Figure 3 and 4 show the histogram for the allele frequency at the start and at end of the conservation scheme (gen 0 and gen 50), calculated on the set of genotype and ungenotyped

SNPs, respectively. At generation 50, a larger proportion of SNPs at intermediate frequency (i.e. close to 0.5) was observed when using the SO_L&H GRM than when using SO_VR. This trend was more accentuated in the set of genotyped SNPs, which corresponds to the set used to calculate the GRMs. However, it is also interesting to observe that the scheme with SO_L&H tended to have slightly higher proportion of SNPs which were fixed.

The results from this study tend to confirm the speculation from GÓMEZ-ROMANO *et al.* (2016) that the OCS using the SO_L&H GRM would drive SNPs towards intermediate frequencies, but it rejects the hypothesis that using the SO_L&H GRM reduces the chances of losing rare alleles due to fixation.

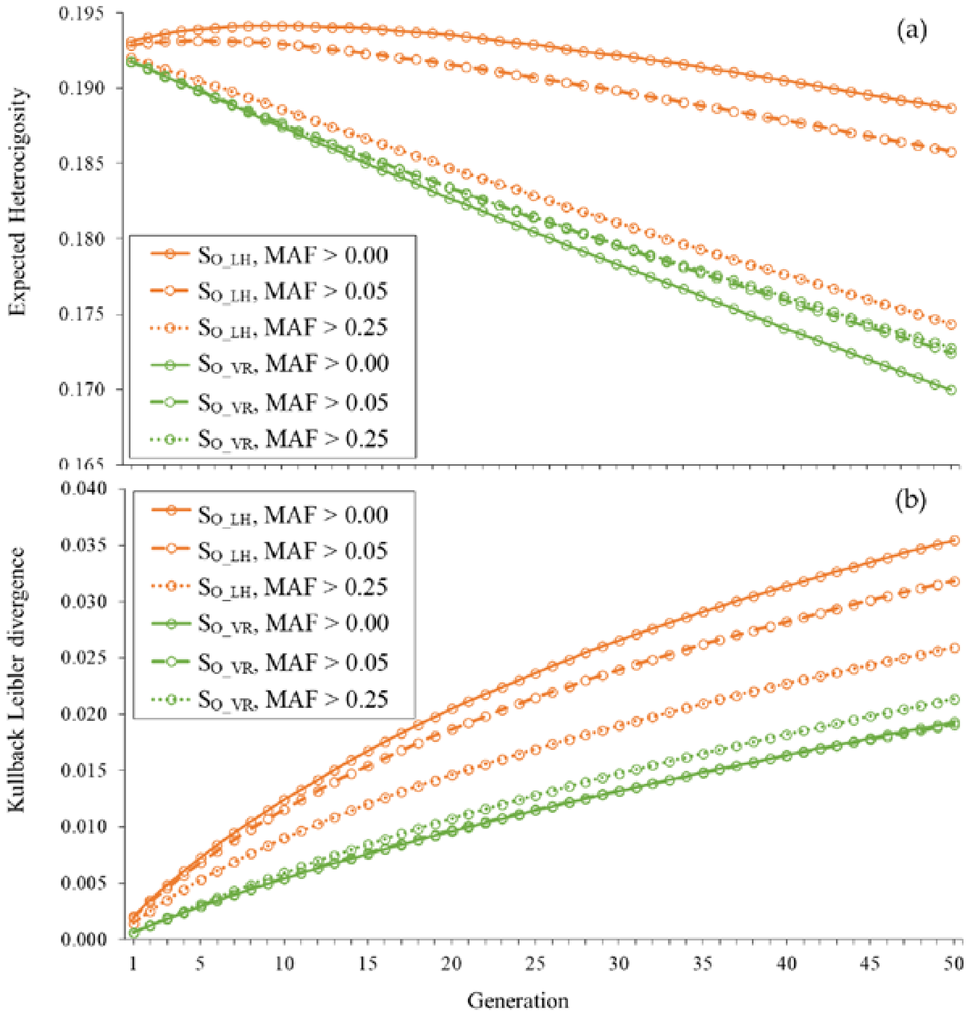

Figure 2. Expected heterozygosity (a) and Kullback-Leibler divergence (b) for unobserved loci across generations when contributions are optimised using Li and Horvitz (SO_LH) and VanRaden (SO_VR) coancestry matrices computed with SNPs with MAF > 0.00, MAF > 0.05 and MAF > 0.25 in a population of 100 individuals.

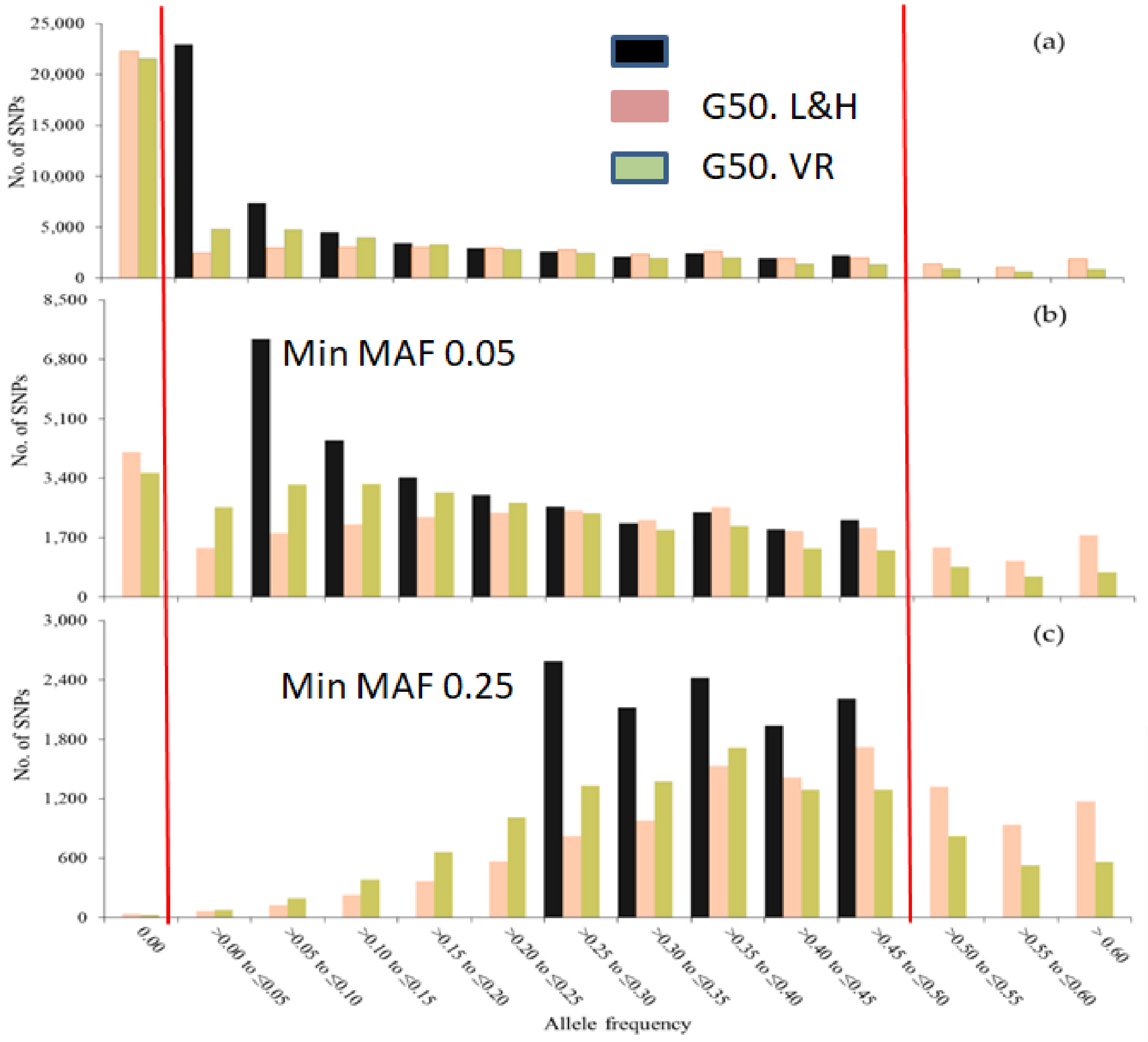

Figure 3. Histogram of allele frequencies at generation 0 and 50, for the set of genotyped SNPs when OCS was done using the GRM calculated with the Li and Horvitz and the VanRaden methods. GRMs were also calculated using all genotyped SNPs, or those after filtering retaining those with MAF > 0.05 and MAF> 0.25. Results are for a population of 100 individuals.

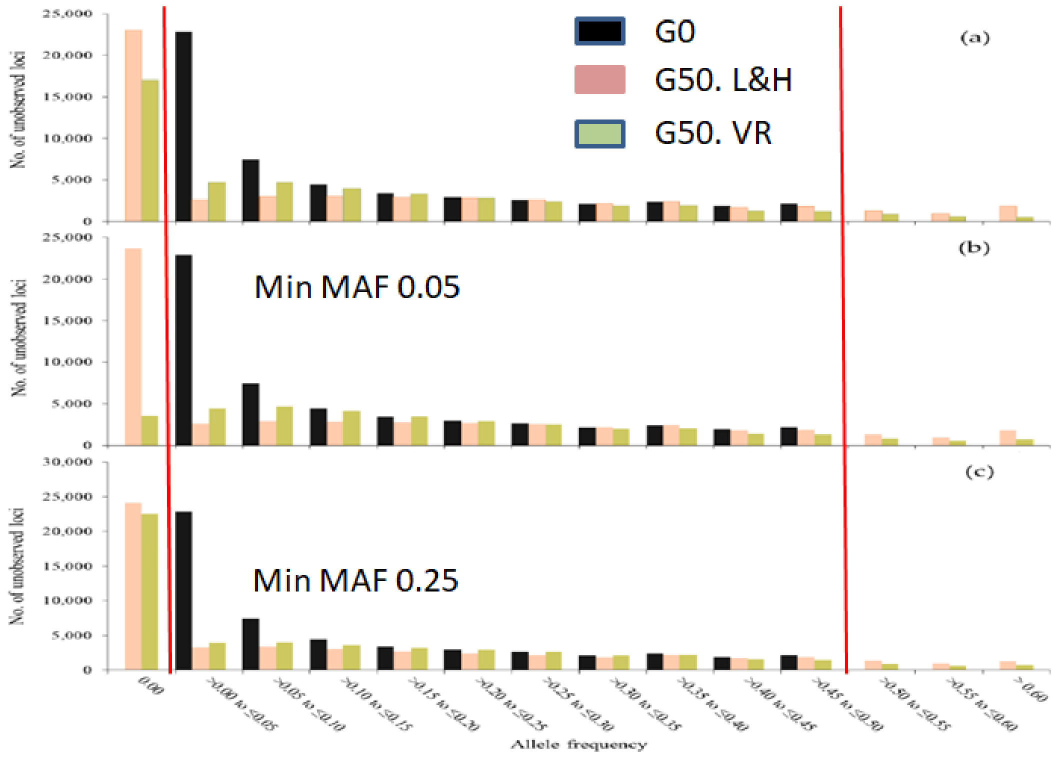

Figure 4. Histogram of allele frequencies at generation 0 and 50, for the set of genotyped SNPs when OCS was done using the GRM calculated with the Li and Horvitz and the VanRaden methods. GRMs were also calculated using all genotyped SNPs, or those after filtering retaining those with MAF > 0.05 and MAF> 0.25. Results are for a population of 100 individuals.

5.4 Conclusions

The behaviour of the OCS is affected by the choice of the GRM used to restrict the rate of genomic inbreeding. The OCS using SO_L&H GRM resulted in a more divergent population, relative to the initial one. In average, it tends to promote SNPs to move towards intermediate frequency, but it does not reduce the probability of losing alleles in individual SNP with rare alleles.

6 Effect of including several restriction of inbreeding on the OCS

6.1 Background

The large-scale genotyping of candidates for the purpose of improving genetic evaluation, has opened a great opportunity for implementing a gOCS which can go beyond its original scope towards a more customised management of the genetic diversity (which will come at virtually no extra cost as the genomic information would be available). GÓMEZ-ROMANO *et al.* (2016) showed that the OCS can be extended to add separate restrictions on the rate of inbreeding to be allowed on different genomic regions (by using GRM specific for the regions on interest). Treating these regions independently in the OCS would allow to prioritise areas of the genome in need of a stronger control, and thereby better retention of their available genetic diversity. GÓMEZ-ROMANO *et al.* (2016) tested the impact of including several restrictions on genetic diversity in conservation programme where the overall objective is to maintain the overall genetic diversity, so the objectives and restriction were, someway, similar. Situations where the objective is to maximise genetic gain would probably prove to be more challenging.

Theoretically, it is possible to use gOCS to control the ΔF at every region of the genome separately, but in practice, the number of separate restrictions may be more limited. Breeding populations, and especially those which require careful management of the genetic diversity, would probably be of small to medium size. Increasing the complexity of the optimisation procedure by adding extra restrictions, would affect the feasibility of finding valid solutions making the OCS fail. In a more optimistic scenario, adding several separately constraints on ΔF would likely have an impact on the selection response, resulting in lower genetic gain for the trait under selection.

In this study, we tested the effect of adding several separate restrictions on genetic diversity to the OCS, when the objective is to maximise genetic gain.

6.2 Methods

6.2.1 Simulation of the gene pool of the reference population in linkage disequilibrium

The gene pool of the population to be used as reference in the genomic prediction was simulated by creating a founder population in LD and, thereafter, expanded it to create a larger population still representative of the smaller one, but with less closely related individuals. This final expanded gene pool population was then used to sample the population for each replicate.

In the first step, the founder population in LD was simulated using a mutation-drift-equilibrium algorithm as suggested by MEUWISSEN *et al.* (2001). Briefly, an initial population of *N* individuals is allowed to reproduce, with each individual producing two offspring (one male and one female). Their genome is composed of several chromosomes with biallelic loci mutating at a given rate. As the population develops across the generations, new mutations appear which are lost or increased in their frequency due to drift. After a large number of generations, the resulting population reaches an equilibrium with a genome containing segregating linked loci in LD. The simulation can be tuned to yield a specific LD pattern by adjusting the population size and mutation rate parameters. This population in equilibrium will be referred to here as the founder population. In the second step, the founder population is allowed to reproduce by further extra generations with a low expansion rate and no mutation rate, and individuals of the last generation are taken as the gene pool population. By sampling the individuals to be used in each replicate from a much enlarged gene pool population, it allows independence between replicates while ensuring that they share a similar LD pattern.

In order to simulate the genome with similar LD pattern as a typical commercial sheep population, the initial population to create the LD (step 1) was composed of 100 individuals (50 males and 50 females). The genome consisted of 26 autosomal chromosomes of 1 Morgan, each with 1,000,000 loci (all fixed to one allele) with their mutation rate set at 10^{-7}. After 10,000 generations, over 9,000 loci were segregating at different frequency in each chromosome (around 250,000 segregating SNPs were simulated across the whole genome). Individuals at generation 10,000 were considered to be the founder population. For the expansion step, the founder population was further reproduced by five extra generations at a 4X expansion rate (i.e. a male/female was randomly mated with several mates to produce 8 offspring each). Finally, 10,000 individuals from generation 10,005 were selected to form the gene pool. In order to further reduce close relationships among individuals from the gene pool, both the LD creation and expansion steps for each chromosome were done independently. This approach means that a given pair of individuals could have a half sib relationship at a given chromosome only but not for the rest.

6.2.2 Selection scheme and optimum contribution selection

To test the impact of the number of restriction, we sampled a population from the gene pool, simulated their breeding values and phenotypes. Thereafter, their GEBV were calculated using GBLUP and then several OCS analyses were performed with different numbers of separate restrictions added to the OCS.

Genetic Architecture and Population Structure: The genotypes of *n* individuals (half of each sex) were sampled from the genepool for the 26 chromosomes. Thereafter 1,100 segregating loci per chromosome were randomly sampled and 1,000 were assigned as SNPs from the chip panel and 100 assigned as QTLs (i.e. the number of SNPs and QTLs across the whole genome were 26,000 and 2,600, respectively).

The QTLs were assumed to be fully additive, and their QTL effects were sampled from a standardised normal distribution and the true breeding value (TBV) for a given individual was calculated as the sum of the all QTL effects. Then the variance of the TBV was calculated and the SNP effects were rescaled to achieve the desirable genetic variance and the TBV recalculated with the correct QTL effect. Thereafter an environmental deviation was sampled and added to the TBV.

Genomic evaluation was carried out using GBLUP(GARRICK 2007). The GRM used in the evaluation was calculated using VR2 method (VANRADEN 2008) with the 26,000 SNPs assigned to the SNP chip.

Optimum Contribution Selection: Several OCS analyses were carried out in the dataset, varying the number of separate restrictions on genetic diversity. Once the contributions were maximised, the genetic gain and the expected rate of inbreeding for each independent region was calculated.

The OCS was carried out using the quadratic programming algorithm proposed by PONG-WONG AND WOOLLIAMS (2018) which allows to include several restrictions and solves the problem in efficient computational manner. The simple OCS was including one single restriction corresponding to the overall genetic variance, with other OCS including restrictions on several regions plus the restriction of the overall diversity.

In order to include a given restriction on diversity, a GRM specific to the region of interest was calculated using only the SNPs which were within the region of interest. The method for calculating the GRM to be used to restrict loss on diversity was based on the method from LI AND HORVITZ (1953). This was done because of a previous study, whose results (given in section 3) showed that this matrix was more consistent as indicator of genetic diversity remaining in the population.

6.2.3 Scenarios compared

The population was simulated assuming a trait controlled by a totally additive effect. The genetic variance was assumed to be 20 and the environmental variance 80 (h2= 0.2). We compared the effect of population size assuming 200, 400, 600, and 1,000 candidates available for selection. The rate of inbreeding tested were 0.005, 0.01, 0.02, 0.03, 0.04 and 0.05.

The OCS was carried out including one restriction on global diversity (covering all 26 chromosomes) and then we performed other OCS scenarios adding the global diversity plus incremental number of specific region, until we included 9 extra separated restrictions on diversity (each covering 2 chromosomes). Note that the constraint on global diversity represents the average value across all the genome including the nine regions which were included when increasing the number of constraint in the OCS.

6.3 Results

The results of the optimisation when including a single constraint on the global inbreeding for all population size and rate of the constraint are shown in Table 1. In all cases considered here, the optimisation succeeded to find a solution which fulfils the constraint on the global ΔF used in the optimisation. As expected, the genetic gain was related to the size of the population and the strength of the constraint: the expected gain was higher with bigger population and higher rate of inbreeding allowed.

Table 1: Performance of the OCS when including only the constraint on global F: expected gain and number of genomic regions which failed the global constraint on F.

Constraint on global ΔF	Gain	Extra regions	Regions with F greater than the global restriction used in the optimisation			
			Number of regions		Extra F (%)	
			Mean	Max	Mean	Max
		n=200				
0.005	2.75	9	5.08	7	30.55	61.84
0.01	3.28	9	5.20	8	19.12	32.20
0.02	3.83	9	5.14	8	14.41	20.65
0.03	4.15	9	5.20	8	12.04	16.62
0.04	4.36	9	5.28	8	10.63	13.57
0.05	4.53	9	5.46	8	9.53	11.08
		n=400				
0.005	3.28	9	5.26	8	29.96	39.04
0.01	3.76	9	5.36	9	19.11	11.96
0.02	4.28	9	5.26	9	14.11	8.94
0.03	4.57	9	5.22	8	12.06	10.74
0.04	4.78	9	5.22	8	10.53	10.64
0.05	4.91	9	4.94	8	10.41	12.40
		n=600				
0.005	3.64	9	5.50	8	32.70	173.85
0.01	4.06	9	5.28	8	21.55	116.37
0.02	4.54	9	5.30	8	15.28	65.31
0.03	4.84	9	5.32	8	13.17	40.53
0.04	5.05	9	5.40	9	11.68	26.87
0.05	5.18	9	5.34	9	10.66	20.01
		n=1000				
0.005	4.84	9	5.42	7	20.78	49.78
0.01	5.33	9	5.64	8	14.64	32.48
0.02	5.63	9	5.56	9	12.77	26.84
0.03	5.83	9	5.44	9	11.58	26.63
0.04	5.98	9	5.42	9	10.63	27.87

However, when inspecting the nine genomic regions separately, the OCS allowed for some of them to have a ΔF greater than the value allowed for the global F. In average, 5.3 regions (out of 9) were found to show level of inbreeding higher that the global ΔF allowed in the optimisation. The excess loss of diversity on these regions averaged around 15% extra ΔF than the global one, but this value was as high as 50% with 2 outlier scenarios where a region showed a ΔF more than double the value achieved for the global average value. The results obtained in this study confirm and quantify the common belief that OCS with a restriction on the global diversity would yield results where some genomic regions will have a higher loss in genetic diversity. Hence considering the diversity of different genomic regions separately and including them separately on the OCS should provide a better management of the genetic diversity.

The results of the OCS including the global diversity constraint plus extra ones associated to the diversity of specific genomic regions are shown in Figure 5. In all scenarios across several population size and degree of strength on the constraints, all regions had rate for inbreeding lower or equal to the one assigned during the optimisation. An interesting result is that adding these extra constraints has little detrimental effect of the expected genetic gain. In average there is a reduction on the gain when adding extra constraints, but the quantity was relative small to be of practical concern.

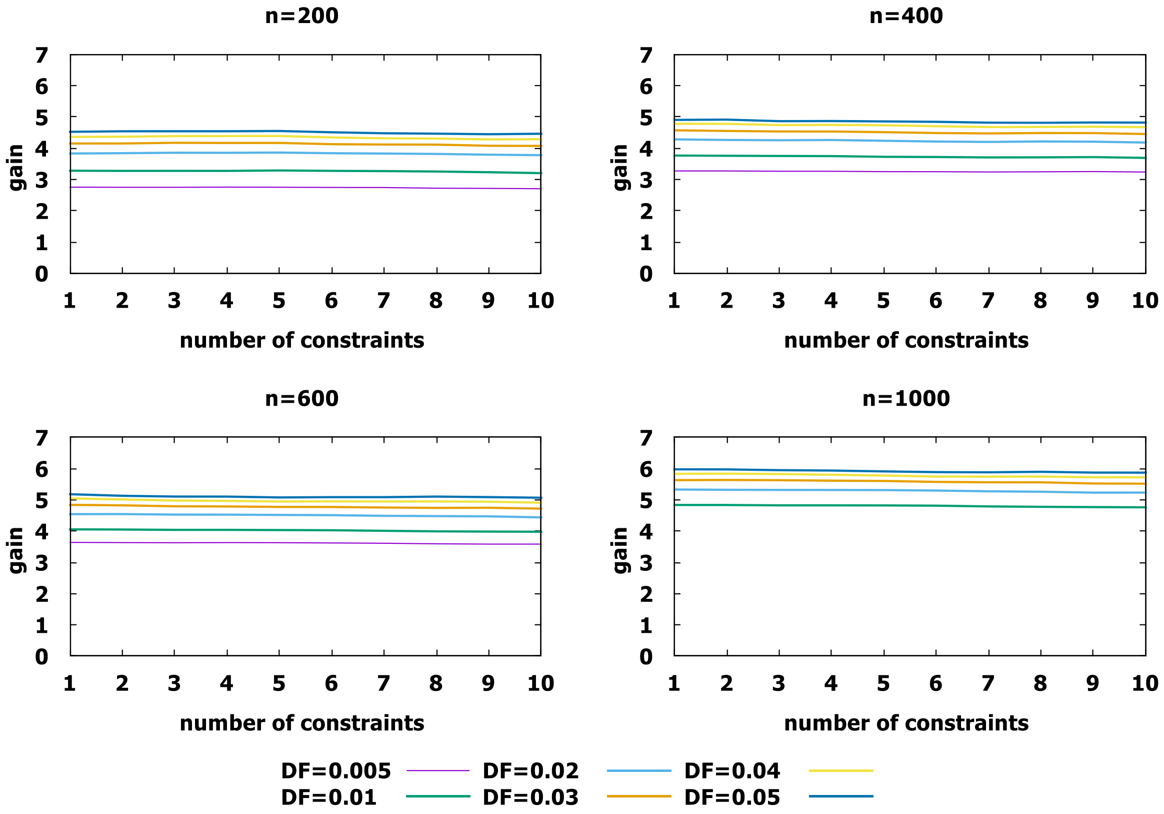

Figure 5. Expected genetic gain of the OCS when including extra constraints on diversity. The results are for four population sizes (200, 400, 600 and 1,000 candidates) and 6 level of restriction given to each restriction (0.05, 0.1, 0.2, 0.3, 0.4 and 0.05).

6.4 Final remarks

The results from this study confirmed the general belief that an OCS where genetic diversity is managed by restricting the global inbreeding would allow for some regions of the genome to exhibit a greater loss of the genetic diversity, supporting the proposal that a better approach would be to consider diversity of genomic regions separately and included them into the OCS independently in different constraints. Our results showed that this is feasible with very little detrimental effect of the genetic gain, making an attractive alternative for the management of commercial close populations.

7 Conclusions

Here we provided a new reformulation of the OCS in order to improve some practical behaviour of the OCS. The value of the different GRMs as indicator of the level of genetic diversity in the population was assessed. We showed that some of these matrices have behaviours which may lead to inconsistent interpretation of how level of genetic diversity is evolving across the time in a given population. We also showed that some new enhancement of the OCS to improve the management of diversity have little or no detrimental effect on the selection response.

8 Deviations or delays

Three weeks of delay. Delay does not have impact in further works in WP5

9 References

Cara, M. A. R., J. Fernández, M. A. Toro and B. Villanueva, 2011 Using genomic wide information to minimize the loss of diversity in conservation programmes. J Anim Breed Genet 128.

Garrick, D. J., 2007 Equivalent mixed model equations for genomic selection. Journal of Animal Science 85**:** 376-376.

Gómez-Romano, F., B. Villanueva, J. Fernández, J. A. Woolliams and R. Pong-Wong, 2016 The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. Genetics Selection Evolution 48**:** 2.

Grundy, B., B. Villanueva and J. Woolliams, 1998 Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. Genetics Research 72**:** 159-168.

Kardos, M., H. R. Taylor, H. Ellegren, G. Luikart and F. W. Allendorf, 2016 Genomics advances the study of inbreeding depression in the wild. Evolutionary applications 9**:** 1205-1218.

Kinghorn, B. P., S. A. Meszaros and R. D. Vagg, 2002 Dynamic tactical decision systems for animal breeding, pp. 07 in *7th World Congress on Genetics Applied to Livestock Production*.

Li, C. C., and D. G. Horvitz, 1953 Some methods of estimating the inbreeding coefficient. Am J Hum Genet 5**:** 107-117.

Meuwissen, T. H. E., 1997 Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75**:** 934-940.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157**:** 1819-1829.

Morales-González, E., J. Fernández, R. Pong-Wong, M. Á. Toro and B. Villanueva, 2021 Changes in Allele Frequencies When Different Genomic Coancestry Matrices Are Used for Maintaining Genetic Diversity. Genes 12**:** 673.

Morales-González, E., M. Saura, A. Fernández, J. Fernández, R. Pong-Wong *et al.*, 2020 Evaluating different genomic coancestry matrices for managing genetic variability in turbot. Aquaculture 520**:** 734985.

Nejati-Javaremi, A., C. Smith and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75**:** 1738-1745.

Pong-Wong, R., and J. A. Woolliams, 2007 Optimisation of contribution of candidate parents to maximise genetic gain and restricting inbreeding using semidefinite programming - (Open Access publication). Genetics Selection Evolution 39**:** 3-25.

Pong-Wong, R., and J. A. Woolliams, 2018 A general quadratic programming method for the optimisation of genetic contributions using interior point algorithm, pp. 714 in *Proceedings of the World Congress on Genetics Applied to Livestock Production*.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91**:** 4414-4423.

Villanueva, B., A. Fernández, M. Saura, A. Caballero, J. Fernández *et al.*, 2021 The value of genomic relationship matrices to estimate levels of inbreeding. Genetics Selection Evolution 53**:** 42.

Villanueva, B., R. Pong-Wong, J. Fernandez and M. A. Toro, 2005 Benefits from marker-assisted selection under an additive polygenic genetic model. Journal of Animal Science 83**:** 1747-1752.

Woolliams, J., and R. Thomson, 1994 A theory of genetic contributions, pp. 127-134, edited by C. Smith, J. S. Gavora, J. Chesnais, W. Fairfull, J. P. Gibson *et al.* Organising Committee, Guelph, Canada.

Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher, 2011 GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics 88**:** 76-82.

Zhang, Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund and G. Sahana, 2015 Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16**:** 88.

10 Appendix

10.1 Appendix 1 Article from Villanueva *et al.* (2021).

10.2 Appendix 2 Article from MORALES-GONZALEZ et al. (2021).

Villanueva *et al. Genet Sel Evol* (2021) 53:42
https://doi.org/10.1186/s12711-021-00635-0

RESEARCH ARTICLE

Open Access

The value of genomic relationship matrices to estimate levels of inbreeding

Beatriz Villanueva[1*], Almudena Fernández[1], María Saura[1], Armando Caballero[2], Jesús Fernández[1], Elisabeth Morales-González[1], Miguel A. Toro[3] and Ricardo Pong-Wong[4]

Abstract

Background: Genomic relationship matrices are used to obtain genomic inbreeding coefficients. However, there are several methodologies to compute these matrices and there is still an unresolved debate on which one provides the best estimate of inbreeding. In this study, we investigated measures of inbreeding obtained from five genomic matrices, including the Nejati-Javaremi allelic relationship matrix (F_{NEJ}), the Li and Horvitz matrix based on excess of homozygosity ($F_{L\&H}$), and the VanRaden (methods 1, F_{VR1}, and 2, F_{VR2}) and Yang (F_{YAN}) genomic relationship matrices. We derived expectations for each inbreeding coefficient, assuming a single locus model, and used these expectations to explain the patterns of the coefficients that were computed from thousands of single nucleotide polymorphism genotypes in a population of Iberian pigs.

Results: Except for F_{NEJ}, the evaluated measures of inbreeding do not match with the original definitions of inbreeding coefficient of Wright (correlation) or Malécot (probability). When inbreeding coefficients are interpreted as indicators of variability (heterozygosity) that was gained or lost relative to a base population, both F_{NEJ} and $F_{L\&H}$ led to sensible results but this was not the case for F_{VR1}, F_{VR2} and F_{YAN}. When variability has increased relative to the base, F_{VR1}, F_{VR2} and F_{YAN} can indicate that it decreased. In fact, based on F_{YAN}, variability is not expected to increase. When variability has decreased, F_{VR1} and F_{VR2} can indicate that it has increased. Finally, these three coefficients can indicate that more variability than that present in the base population can be lost, which is also unreasonable. The patterns for these coefficients observed in the pig population were very different, following the derived expectations. As a consequence, the rate of inbreeding depression estimated based on these inbreeding coefficients differed not only in magnitude but also in sign.

Conclusions: Genomic inbreeding coefficients obtained from the diagonal elements of genomic matrices can lead to inconsistent results in terms of gain and loss of genetic variability and inbreeding depression estimates, and thus to misleading interpretations. Although these matrices have proven to be very efficient in increasing the accuracy of genomic predictions, they do not always provide a useful measure of inbreeding.

Background

Inbreeding, i.e. the mating of individuals related by ancestry, is a fundamental concept in many areas of biology, including animal and plant breeding [1], human genetics [2, 3], and evolutionary [4] and conservation biology [5]. Inbreeding results in a reduction of genetic diversity, as it increases homozygosity at the expense of heterozygosity. This increase in homozygosity in turn increases the incidence of homozygous recessive defects and decreases population means for many quantitative traits (i.e., inbreeding depression), particularly those related to fitness [6, 7].

*Correspondence: villanueva.beatriz@inia.es
[1] Departamento de Mejora Genética Animal, INIA, Ctra. de La Coruña, km 7.5, 28040 Madrid, Spain
Full list of author information is available at the end of the article

The level of inbreeding of an individual is measured by the inbreeding coefficient, which was defined by Wright as the correlation between homologous alleles of the two gametes that unite to form the individual [8], and later by Malécot as the probability that two homologous alleles at a given locus are identical-by-descent [9]. The inbreeding coefficient also gives the proportion by which the heterozygosity of an individual is reduced by inbreeding [10] and, thus, the proportional loss of genetic variation. Classically, the inbreeding coefficient of an individual has been determined based on its pedigree. However, the pedigree-based inbreeding coefficient provides only expected proportions of the genome that are identical-by-descent.

The level of inbreeding has also been estimated from molecular data, such as those contained in high-density single nucleotide polymorphism (SNP) arrays. Genomic inbreeding coefficients can be more accurate than pedigree-based measures because they capture the variation due to Mendelian sampling and therefore can differentiate among individuals with the same pedigree (e.g. [11]). Genomic measures also allow us to differentiate inbreeding at specific regions of a genome, which is not possible with pedigree-based inbreeding.

Several methods have been proposed to calculate inbreeding coefficients using genomic data, including methods based on continuous runs of homozygosity (e.g. [11, 12] and methods applied on a SNP-by-SNP basis (e.g. [13–16]). Some of the latter measures come from matrices that are used to obtain genomic predictions in animal breeding. In this context, best linear unbiased predicted (BLUP) evaluations are replaced by genomic BLUP (GBLUP) evaluations, in which the numerator relationship matrix (NRM) is substituted by one of several genomic relationship matrices (GRM) [15, 16]. Given that the diagonals of the NRM equal 1 plus the inbreeding coefficients for the corresponding individuals, it has been generally accepted that the diagonals of the GRM are 1 plus the realized inbreeding level for the corresponding individuals. These genomic measures of inbreeding have been widely used [11, 17–47]. However, they can result in very different outcomes and the correlations between these estimators vary greatly and can even be negative, e.g. [27, 35]. Thus, there is still an unresolved debate on which are the best measures of inbreeding.

In this study, we compared genomic inbreeding coefficients that were obtained from different SNP-by-SNP methods to understand their relationship with traditional definitions of inbreeding. First, we describe different coefficients based on genomic information at the individual level. Second, we derive expectations at the population level for the different coefficients based on a single locus model. These expectations are then used to explain the patterns of the coefficients computed based on thousands of SNP genotypes across the genome in a highly inbred pig population.

Methods

Inbreeding coefficients obtained from genomic data

Individual inbreeding coefficients were obtained from the diagonal elements of five different genomic relationship matrices. These coefficients have been widely used in the literature, but under different names (see Table 1) and there is no consensus about the nomenclature. Here, the name chosen for each coefficient

Table 1 Summary of the names given to different genomic inbreeding coefficients in the literature

Nomenclature used in this paper	Nomenclature used in the literature	References
F_{NEJ}	F_{PH}	[19]
	F_{M}	[20]
	F_{MOL}	[33]
	Homozygosity	[21]
	F_{HOM}	[28, 35]
	$\mathrm{HOM}_{\mathrm{SNP}}$	[37]
	SNP-Similarity*	[29]
	SIM*	[47]
$F_{L\&H}$	F_{h} or F_{H}	[11, 25, 40]
	F_{snp}	[26]
	F_{HOM}	[27, 33, 36, 41, 42, 46, 55]
	F_{ExHOM}	[35]
	F_{PLINK}	[31]
	F_{IS}	[45]
	F_{EH}	[34]
	LHR	[22]
	L&H*	[47]
F_{VR1}	F_{GRM}	[19, 41]
	F_{GRM1}	[35]
	F_{VR}	[34]
	F_{G}	[17]
	VR1*	[47]
F_{VR2}	*F*hatI, F^{I}	[18, 42, 54]
	F_{GRM}	[27, 33]
	$\mathrm{F}_{\mathrm{GRM2}}$	[35]
	VR2*	[47]
F_{YAN}	*F*hatIII, F^{III}	[18, 42, 54]
	F_{alt}	[11, 40]
	GRM_F, F_{GRM}	[21, 28, 31]
	F_{UNI}	[27, 35, 36, 41, 55]
	F_{grm}	[31]
	SNP-Yang*	[29]
	YAN*	[47]

[*] Self-relationship or self-coancestry

makes reference to the authors who first proposed or formulated it explicitly, to the best of our knowledge. We compared the following coefficients:

1. F_{NEJ}: inbreeding coefficient computed from the diagonal elements of the allelic relationship matrix of Nejati-Javaremi et al. [14] as:

$$F_{NEJ} = \frac{\sum_{k=1}^{S}(\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij_k})/2}{S} - 1,$$

where I_{ij_k} is the identity of the two alleles (i and j) of the individual at SNP k, which takes the value of 1 if the two alleles are identical and 0 if they are not. Note that F_{NEJ} is simply the proportion of SNPs that are homozygous for the individual and thus it does not distinguish between identity-by-state (IBS) and identity-by-descent (IBD) [48].

2. $F_{L\&H}$: inbreeding coefficient based on the relationship matrix that describes deviations from Hardy–Weinberg proportions, computed as:

$$F_{L\&H} = \frac{SF_{NEJ} - \sum_{k=1}^{S}[1 - 2p_{k(0)}(1 - p_{k(0)})]}{S - \sum_{k=1}^{S}[1 - 2p_{k(0)}(1 - p_{k(0)})]},$$

where $p_{k(0)}$ is the frequency of the reference allele (allele B) of SNP k in the base (reference) population [13]. $F_{L\&H}$ estimates the deviation of the observed frequency of homozygotes (AA and BB) from that expected in the base population under Hardy–Weinberg proportions. Thus, it corrects for the homozygosity that was present in the base population and expresses molecular inbreeding in terms of IBD [42, 48, 49].

3. F_{VR1}: inbreeding coefficient computed from the diagonal elements of the genomic relationship matrix obtained according to VanRaden's method 1 [15], as follows:

$$F_{VR1} = \frac{\sum_{k=1}^{S}(x_k - 2p_{k(0)})^2}{2\sum_{k=1}^{S} p_{k(0)}(1 - p_{k(0)})} - 1,$$

where x_k is the genotype of the individual for SNP k, coded as 0, 1 or 2 for genotypes AA, AB and BB, respectively, and $p_{k(0)}$ is as defined for $F_{L\&H}$. F_{VR1} is based on the variance of additive genetic values and provides a measure relative to frequencies of the reference allele in the base population. However, F_{VR1} differs from $F_{L\&H}$ in that with F_{VR1} homozygous genotypes are weighted by the inverse of their allele frequency and, thus, rare homozygous genotypes contribute more to the inbreeding measure than common homozygous genotypes [35].

4. F_{VR2}: inbreeding coefficient computed from the diagonal elements of the genomic relationship matrix obtained according to VanRaden's method 2 [15] as follows:

$$F_{VR2} = \frac{1}{S}\sum_{k=1}^{S}\frac{(x_k - 2p_{k(0)})^2}{2p_{k(0)}(1 - p_{k(0)})} - 1,$$

where x_k and $p_{k(0)}$ are as for F_{VR1}. F_{VR2} is similar to F_{VR1} but the summation across markers is made differently, such that the weight given to rare alleles is even greater. In F_{VR2}, the contribution of each SNP is divided by its own variance, whereas in F_{VR1} the contributions of all SNPs are divided by the same denominator [35].

5. F_{YAN}: inbreeding coefficient computed from the diagonal elements of the genomic relationship matrix of Yang [16] as follows:

$$F_{YAN} = \frac{1}{S}\sum_{k=1}^{S}\frac{x_k^2 - \left(1 + 2p_{k(0)}\right)x_{k_i} + 2p_{k(0)}^2}{2p_{k(0)}(1 - p_{k(0)})},$$

where x_k and $p_{k(0)}$ are as for F_{VR1}. This coefficient is based on the correlation between uniting gametes [16, 42] and also gives more weight to homozygotes for the minor allele than to homozygotes for the major allele [40]. However, it has a lower sampling variance than the previous coefficients [18, 35] because it accounts for the sampling error associated with each SNP [16, 28].

The coefficients that depend on allele frequencies, i.e. $F_{L\&H}$, F_{VR1}, F_{VR2}, and F_{YAN}, need to be computed using the initial frequencies; i.e. those in the base population. Note that F_{NEJ} is equivalent to F_{VR1}, F_{VR2} and F_{YAN} when base population allele frequencies equal 0.5 [29].

Expected genomic inbreeding coefficients at the population level: a single locus model

Expected values for $F_{L\&H}$, F_{VR1}, F_{VR2} and F_{YAN} at the population level were derived based on a single SNP model. Let $p_{(0)}$ be the frequency of allele B in the base population. After t generations, the frequency will have changed to $p_{(t)}$ due to random drift and selection, among other reasons. Assuming random mating, we can expect that genotype frequencies within a generation are in Hardy–Weinberg equilibrium. Thus, the expected F for a group of individuals from generation t can be obtained as:

$$E(F) = [freq(\mathrm{AA})F_{\mathrm{AA}} + freq(\mathrm{AB})F_{\mathrm{AB}} + freq(\mathrm{BB})F_{\mathrm{BB}}],$$

where $freq(\mathrm{AA}) = (1 - p_{(t)})^2$, $freq(\mathrm{AB}) = 2p_{(t)}(1 - p_{t(t)})$ and $freq(\mathrm{BB}) = p_{(t)}^2$, and F_{XY}

is the inbreeding coefficient for an individual with genotype XY, which is computed using the initial frequency $p_{(0)}$, as described in the previous section. To assess the impact of initial and current allele frequencies on expected values of the evaluated inbreeding coefficients, the latter were evaluated for the whole range of values for $p_{(0)}$ and $p_{(t)}$.

Evaluation of genomic inbreeding in a population of Guadyerbas pigs

Results from the single locus model were evaluated in a population of Iberian pigs, with thousands of SNPs used to compute the inbreeding coefficients across the genome and at specific genomic regions.

Pig samples and SNP genotypes

The data used were from a herd of Guadyerbas Iberian pigs. The Guadyerbas strain is one of the most ancient surviving Iberian strains. It is highly inbred and in serious danger of extinction. The strain originated from four males and 20 females [50] and was conserved from 1944 until 2011 as a genetically isolated population. Accurate and complete genealogy was available from when the herd was first established (about 25 generations) and included 1178 animals born from 197 sires and 467 dams.

DNA samples were available for 86 males and 141 females born in the herd between 1992 and 2011 and were genotyped with the Illumina PorcineSNP60 BeadChip v1. SNP positions in the genome were based on the genome assembly Sscrofa 11.1. After quality control, as described in Saura et al. [20], 219 animals and 47,120 SNPs remained. In Iberian pigs, the generation interval is about three years, and thus for analysis of genomic inbreeding, we considered six cohorts of animals born in successive periods of three years, starting from year 1994 (Table 2).

Table 2 Number of genotyped animals per cohort and sex in the Guadverbas population

Cohort	Birth year range	Males	Females
1	1994–1996	13	18
2	1997–1999	10	42
3	2000–2002	24	18
4	2003–2005	8	19
5	2006–2008	8	7
6	2009–2011	19	29
Total		82	133

Patterns of genomic inbreeding coefficients

Genomic coefficients were obtained for all genotyped pigs using the SNPs that segregated in cohort 1 (17,951 SNPs). The frequencies used to calculate $F_{L\&H}$, F_{VR1}, F_{VR2}, and F_{YAN} were those for cohort 1 (i.e. this cohort was considered to be the base population). Patterns of inbreeding across the genome were determined using sliding windows of 35 SNPs (average length of 4.25 Mb) that were moved one SNP at a time (17,339 windows). For each window, the average F was computed in order to reduce the noisiness of single-locus estimates and to clarify the graphical representations [51–53]. For the coefficients that depend on allele frequencies ($F_{L\&H}$, F_{VR1}, F_{VR2} and F_{YAN}), the formulae were applied within each window. Finally, values were averaged across individuals.

Inbreeding depression

The behavior of the different genomic inbreeding coefficients will have consequences when they are used to estimate the rate of inbreeding depression across the genome. In order to investigate this, we performed a genome scan for inbreeding depression for the number of piglets born alive in the Guadyerbas population, using all genotyped sows with records born in the six cohorts (109 sows and 265 litter records) and the sliding window approach. The animals and phenotypic data used and the model fitted are described in detail in Saura et al. [26]. Briefly, inbreeding depression was estimated by regressing the number of piglets born alive on F assuming a linear model. Fixed effects included the combination of season of farrowing and farrowing facilities, parity, strain of boar, and the linear regression on F. Random effects included additive genetic, permanent environmental, and residual effects. The variance–covariance matrix of additive genetic effects was assumed to be the pedigree-based numerator relationship matrix. Three measures of F ($F_{L\&H}$, F_{VR2} and F_{YAN}) computed using genotypes for all genotyped sows with phenotypic data born from cohort 1 to cohort 6, were used as covariates to estimate inbreeding depression.

Results

Range of values and interpretation of the genomic inbreeding coefficients

The inbreeding coefficients investigated differ in the range of values that they can contain and, with the exception of F_{NEJ}, their ranges depend on the allele frequency in the base population $p_{(0)}$. Coefficient F_{NEJ} ranges from 0 to 1 because it is the proportion of homozygous SNPs. At the individual level, values for $F_{L\&H}$ range from $-\infty$ to 1, and those for F_{VR1}, F_{VR2} and F_{YAN} range from -1 to ∞ (Figs. 3, 4, and 5, in Zhang et al. [27]). When all SNP genotypes are homozygous, $F_{L\&H}$ equals 1 and when all are heterozygous, it ranges from $-\infty$ to -1. F_{VR1} and F_{VR2} cover the entire range (from -1 to ∞) both when all SNP genotypes are homozygous or heterozygous. Finally, when all SNP genotypes are homozygous, F_{YAN} ranges from 0 to ∞ and when they all are heterozygous, F_{YAN} equals -1. Thus, values for $F_{L\&H}$, F_{VR1}, F_{VR2} and F_{YAN} can be outside the permitted ranges for probabilities and correlations. Nevertheless, as inbreeding coefficients, they can still be interpreted as the proportional loss or gain in variability (heterozygosity) relative to the variability in the base population, with a negative value indicating that variability has been gained and a positive value that variability has been lost. It is also possible to gain more than 100% of the initial variability but it is not possible to lose more than 100%. A value equal to 1 indicates that all the variability that was present in the base population has been lost but a value greater than 1 indicates that more variability than what existed initially has been lost, which does not make sense.

Expected values of genomic inbreeding coefficients based on the single locus model

Expected values for $F_{L\&H}$, F_{VR1} (or F_{VR2}) and F_{YAN} based on the single locus model for the whole range of starting ($p_{(0)}$) and current ($p_{(t)}$) frequencies are shown in Fig. 1. Note that for a single locus model $E(F_{VR1}) = E(F_{VR2}) = E(F_{VR})$.

The expected value for $F_{L\&H}$ (Fig. 1a) ranged from $-\infty$ and 1. When the frequency of the minor allele increases (i.e. $p_{(t)} > p_{(0)}$) towards 0.5, $E(F_{L\&H})$ becomes negative, which indicates that some variability has been gained. This makes sense given that the maximum variability occurs when the frequency is 0.5. Given that the upper limit of $E(F_{L\&H})$ is 1, when using this coefficient, one never expects more variability to be lost than the variability that initially existed. $E(F_{L\&H})$ takes the value of 1 when the SNP becomes fixed, which is equivalent to all the variability being lost.

The expected value for F_{VR} based on the diagonals of VanRaden's GRM is within the range [0, 1] for some combinations of $p_{(0)}$ and $p_{(t)}$, but for many other combinations it is outside this range (Fig. 1b). In fact, $E(F_{VR})$ ranges from -1 to ∞. This means that $E(F_{VR})$ can indicate that some variability has been gained but this gain can never be greater than 100% of the initial variability, as the lower limit is -1. It also means that $E(F_{VR})$ can indicate that more than 100% of the initial variability is lost, as it can take values higher than 1 (up to ∞).

In the right panel of Fig. 1b, the grid of initial and current frequencies is divided in regions where $E(F_{VR})$ is <0, between 0 and 1, or >1. When the frequency of the minor allele is doubled (i.e. $p_{(t)} = 2p_{(0)}$) but still lower than 0.5, $E(F_{VR}) = 1$, which means that 100% of the variability has been lost in the current generation when, in fact, variability has increased. For instance, if $p_{(0)}$= 0.25 and $p_{(t)}$= 0.5, $E(F_{VR})$ indicates that all the initial variability has been lost, although the maximum variability is reached at a frequency of 0.5. When the frequency of the minor allele more than doubles (i.e. $p_{(t)} > 2p_{(0)}$), $E(F_{VR})$ becomes >1 (for instance, for $p_{(0)} = 0.1$ and $p_{(t)} = 0.3$, $E(F_{VR}) = 2.2$), which indicates that more than 100% of the initial variability has been lost, which is unreasonable. When the initial frequency of the minor allele is lower than 0.33 and decreases, then $E(F_{VR}) < 0$, which indicates that variability has increased relative to its initial value. This is also the case when the minor allele is lost ($p_{(t)} = 0$). Thus, although variability in the current generation is lower than in the initial generation in these cases, $E(F_{VR})$ incorrectly indicates that some variability has been gained.

On the one hand, although for a particular individual in the population, F_{YAN} can be negative (up to -1), contrary to $E(F_{VR})$, $E(F_{YAN})$ is never smaller than 0 (it ranges from 0 to ∞; Fig. 1c), which indicates that the level of heterozygosity cannot become larger than the level that existed initially, which is unreasonable. On the other hand, and as for $E(F_{VR})$, $E(F_{YAN})$ can be greater than 1, implying that more heterozygosity than what existed initially can be lost. In addition, although increasing the frequency of the minor allele towards 0.5 increases variability, $E(F_{YAN})$ can indicate a decrease in variability. For instance, when $p_{(0)}$= 0.1 and remains at 0.1 in the current generation, $E(F_{YAN}) = 0$. However, if the frequency increases to 0.2, $E(F_{YAN})$ becomes greater than 0 ($E(F_{YAN}) = 0.11$), which indicates that some variability has been lost. And, if it increases to 0.5 (in theory a value at which the variability is maximum), $E(F_{YAN})$ becomes greater than 1 ($E(F_{YAN}) = 1.78$), which indicates that more than 100% of the initial variability was lost.

When the initial frequency ($p_{(0)}$) is set to 0.5, the expected value for $F_{L\&H}$, F_{VR} and F_{YAN} is the same regardless of the current frequency ($p_{(t)}$) (see Additional file 1: Figure S1). In this scenario, these three coefficients range from 0 (when $p_{(t)}$ remains at 0.5) to 1 (when the SNP becomes fixed; i.e. $p_{(t)} = 0$ or 1).

Figure 2 shows the same profiles as in Fig. 1, but with the reference allele driven to a frequency of 0 (Fig. 2a), 0.5 (Fig. 2b) or 1 (Fig. 2c) in the current generation. Note that there is some redundancy in Fig. 2a, c since fixation of the major allele is equivalent to loss of the minor allele. For any value, $E(F_{L\&H}) = 1$ when the SNP becomes fixed ($p_{(t)} = 0$ or 1), regardless of the initial frequency, as expected (Fig. 2a, c). However, when the major allele is lost (Fig. 2a) or when the minor allele is fixed (Fig. 2c), both $E(F_{VR})$ and $E(F_{YAN})$ take values greater than 1 (in fact their upper limit is ∞). Losing the minor allele leads to negative values for $E(F_{VR})$ when $p_{(0)} < 1/3$ and its limit

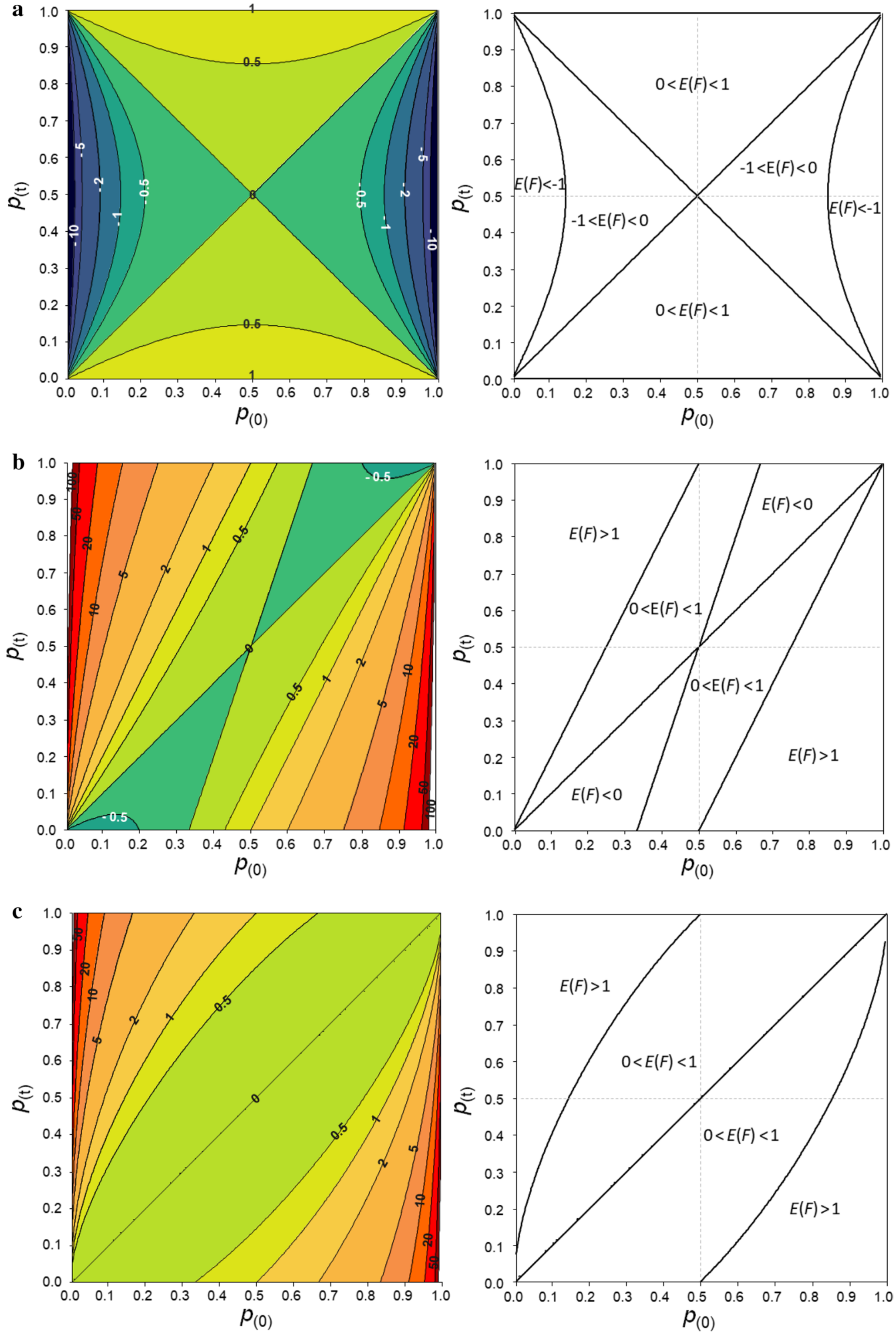

Fig. 1 Expected inbreeding coefficient based on excess of homozygosity ($F_{L\&H}$) (**a**) and expected inbreeding coefficients computed from the diagonal elements of the genomic relationship matrices of VanRaden (methods 1 and 2; $F_{VR} = F_{VR1} = F_{VR2}$) (**b**) and of Yang ($F_{YAN}$) (**c**) as a function of starting and current allele frequencies at a single locus. On the right, the grid of initial and current frequencies is divided in regions where the expected value of F is < -1, < 0, between -1 and 0, between 0 and 1, or > 1

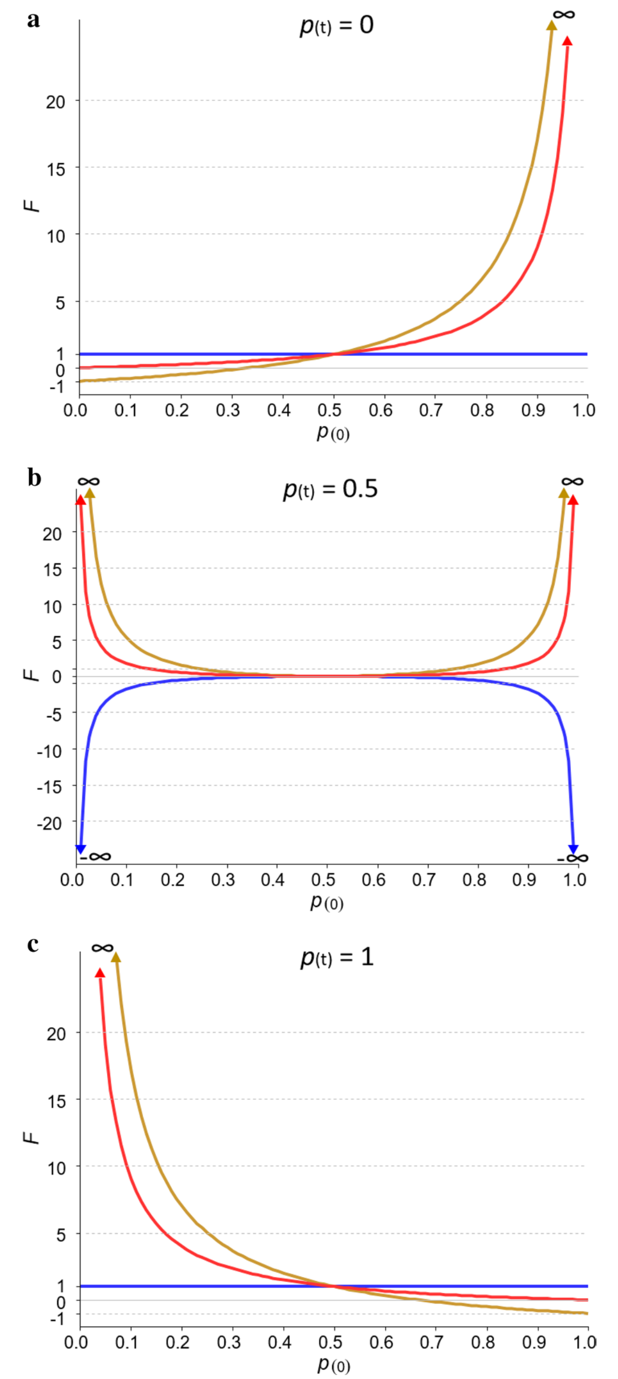

Fig. 2 Expected $F_{L\&H}$, F_{VR} ($F_{VR1} = F_{VR2}$) and F_{YAN} in the generation in which the reference allele was lost (**a**), driven to a frequency of 0.5 (**b**), or fixed (**c**) relative to the initial frequency ($p_{(0)}$). $F_{L\&H}$: blue line, F_{VR}: brown line, F_{YAN}: red line

is − 1 (Fig. 2a). Another way of looking at this is that fixing the major allele leads to negative values for $E(F_{VR})$ when $p_{(0)} > 2/3$, and its limit is also − 1 (Fig. 2c). In these scenarios, the value of $E(F_{YAN})$ remains equal to 1. It is interesting to note that when $p_{(t)}$= 0.5, $E(F_{YAN})$, and to a lesser extent $E(F_{VR})$, behave as a mirror image of $F_{L\&H}$ (Fig. 2b).

In summary, expected values for $F_{L\&H}$, F_{VR1}, F_{VR2} and F_{YAN} depend on frequency changes. When the inbreeding coefficient is interpreted as an indicator of loss or gain of variability, $F_{L\&H}$ gives sensible values but F_{VR1}, F_{VR2}, and F_{YAN} do not. In fact, $E(F_{L\&H})$ follows the trend of loss or gain in heterozygosity due to changes in allele frequencies. When the minor allele frequency (MAF) decreases (i.e. when heterozygosity decreases relative to that in a reference base population), $E(F_{L\&H})$ increases. However, $E(F_{VR1})$ and $E(F_{VR2})$ can lead us to think that: (i) more than 100% of the initial variability is lost; and, even worse, (ii) variability has increased when in reality it has decreased or vice versa. $E(F_{YAN})$ also leads to inconsistent results since it never indicates that variability has increased, but it can indicate that more than 100% of the initial variability is lost.

Patterns of genomic inbreeding in the population of Guadyerbas pigs

Summary statistics for the different inbreeding coefficients, computed both at the individual level and at the regional (window) level, are in Table 3 for the first and last cohorts. Average values for each coefficient at the individual and regional levels were practically the same but those at the regional level varied much more than those at the individual level, particularly for cohort 6. The proportion of homozygous loci (F_{NEJ}) increased by 5% from cohort 1 to cohort 6. Coefficient F_{NEJ} had a much higher average and a lower standard deviation than the other coefficients. Coefficients that are weighted by the initial frequencies (i.e. $F_{L\&H}$, F_{VR1}, F_{VR2} and F_{YAN}) were on average less than 0 for cohort 1 (about − 0.1) and became positive (up to ~ 0.2) for cohort 6.

Pairwise correlations between coefficients computed both at the individual and regional (window) level are in Fig. 3. Correlations at the individual animal level (which are averages across the genome) ranged from 0.4 to 1 for cohort 1 and from 0.7 to 1 for cohort 6. As expected, the correlation between F_{NEJ} and $F_{L\&H}$ was 1. Correlations higher than 0.9 were also found between F_{YAN} and F_{NEJ}, F_{YAN} and $F_{L\&H}$, F_{YAN} and F_{VR1}, and F_{VR1} and F_{VR2}. The lowest correlations were between F_{VR2} and F_{NEJ} and between F_{VR2} and $F_{L\&H}$, but these correlations increased from ~ 0.4 for cohort 1 to ~ 0.7 for cohort 6, which could be due to the loss of rare alleles over time but also to random fluctuations. At the regional genomic level, changes in frequencies can be more exaggerated, which results in lower correlations between coefficients than at the individual animal level, particularly for those involving VanRaden's coefficients.

Table 3 Mean, standard deviation (SD) and minimum and maximum values for the different genomic inbreeding coefficients when computed at the individual animal or genomic region level in cohorts 1 and 6 of the Guadyerbas population

Cohort		Individual level				Regional level			
		Mean	SD	Min	Max	Mean	SD	Min	Max
1	F_{NEJ}	0.616	0.024	0.580	0.674	0.615	0.084	0.300	0.968
	$F_{L\&H}$	− 0.095	0.070	− 0.198	0.071	− 0.095	0.122	− 0.517	0.373
	F_{VR1}	− 0.095	0.076	− 0.230	0.119	− 0.095	0.122	− 0.517	0.372
	F_{VR2}	− 0.088	0.108	− 0.279	0.211	− 0.088	0.106	− 0.450	0.340
	F_{YAN}	− 0.088	0.053	− 0.165	0.061	− 0.088	0.106	− 0.450	0.341
6	F_{NEJ}	0.669	0.025	0.631	0.743	0.669	0.124	0.307	1.000
	$F_{L\&H}$	0.056	0.070	− 0.052	0.268	0.056	0.353	− 2.939	1.000
	F_{VR1}	0.120	0.079	− 0.014	0.417	0.111	0.352	− 0.853	2.717
	F_{VR2}	0.175	0.108	0.023	0.609	0.172	0.558	− 0.876	4.118
	F_{YAN}	0.090	0.076	− 0.014	0.364	0.089	0.215	− 0.465	1.306

The pattern of homozygosity clearly varied across chromosomes and across regions within chromosomes (Fig. 4). For several genomic regions, SNPs that were still segregating in cohort 1 became fixed in cohort 6 (see for example, *Sus scrofa* (SSC) chromosomes 4, 8, 13, 14 and 17).

Figure 5 compares the patterns of the different coefficients across the genome for cohort 6. Here, we only consider the SNPs that segregated in cohort 1. In general, the patterns differed a lot between coefficients. It is interesting to note that, in general, the patterns for F_{VR1} and F_{VR2} were mirror images of those for $F_{L\&H}$. One particularly striking result is that in regions where SNPs had become fixed (see also Fig. 4), $F_{L\&H}$ was equal to 1 whereas F_{VR1} and F_{VR2} were negative with large absolute values. Two very clear examples are the region between 43 and 56 Mb on SSC4 and the region between 58 and 82 Mb on SSC14. In both these regions, the initial frequency of the minor allele was very low ($p_{(0)} \leq 0.1$), and the allele was already lost in cohort 6 ($p_{(t)} = 0$). At all positions within these regions, $F_{L\&H}$ was equal to 1, while F_{VR1} and F_{VR2} became negative (about − 0.8 in the SSC4 region and ranging from − 0.9 to − 0.6 in the SSC14 region), which incorrectly suggests that some variability was gained, and F_{YAN} was low (about 0.1 in the SSC4 region and ranging from 0.1 to 0.2 in the SSC14 region). These observations agree with the expectations described above and lead us to conclude that $F_{L\&H}$ is a much more valuable measure of change in variability than F_{VR1} or F_{VR2}. For the regions where all variability was lost, $F_{L\&H}$ is expected to indicate that this is the case, but both F_{VR1} and F_{VR2} indicate that variability was gained.

In addition, there are some regions for which the variability increased from cohort 1 to cohort 6, as F_{NEJ} was lower in the latter (Fig. 4), e.g. the regions between 102 and 112 Mb on SSC3, between 41 and 72 Mb on SSC6, and between 81 and 97 Mb on SSC13. In all these cases, $F_{L\&H}$ did indeed show this increase in variability since it became negative. However, F_{VR1} and F_{VR2} were again like mirror images of $F_{L\&H}$, while F_{YAN} was positive but close to 0. These observations also agree with expectations. For instance, the average $p_{(0)}$ and $p_{(t)}$ in the SSC13 region were 0.31 and 0.40, respectively. With this change in frequency, $E(F_{L\&H})$ varied from − 1 to 0, while the expected values for F_{VR1}, F_{VR2} and F_{YAN} were all between 0 and 1. Also remarkable are the high peaks observed for F_{VR2}. In some regions on SSC2 and SSC17, F_{VR2} reached a value as high as 4. In these regions (between 35 and 38 Mb on SSC2 and between 33 and 34 Mb on SSC17), there are SNPs with rare alleles ($p_{(0)} < 0.1$) which had a high increase in frequency ($p_{(t)} > 0.3$), and under these circumstances, F_{VR2} is expected to reach very high positive values (Fig. 1b), while $F_{L\&H}$ is expected to become negative (Fig. 1a).

With VanRaden's and Yang's coefficients, and in particular F_{VR2}, a higher inbreeding coefficient is assigned to an individual that is homozygous for a rare allele than to an individual that is homozygous for a common allele. Thus, F_{VR1}, F_{VR2}, and F_{YAN} put a greater weight on SNPs that have a low MAF. Based on this, in addition to the scenario considered so far, in which all the SNPs segregating in cohort 1 (MAF > 0) were used to calculate the inbreeding coefficients, we analyzed two additional scenarios with different MAF thresholds in cohort 1: (i) using only the common variants (here defined as SNPs with MAF > 0.05); and (ii) using only the very common variants (here defined as SNPs with MAF > 0.25). This allowed us to determine how the differences between coefficients were affected by MAF.

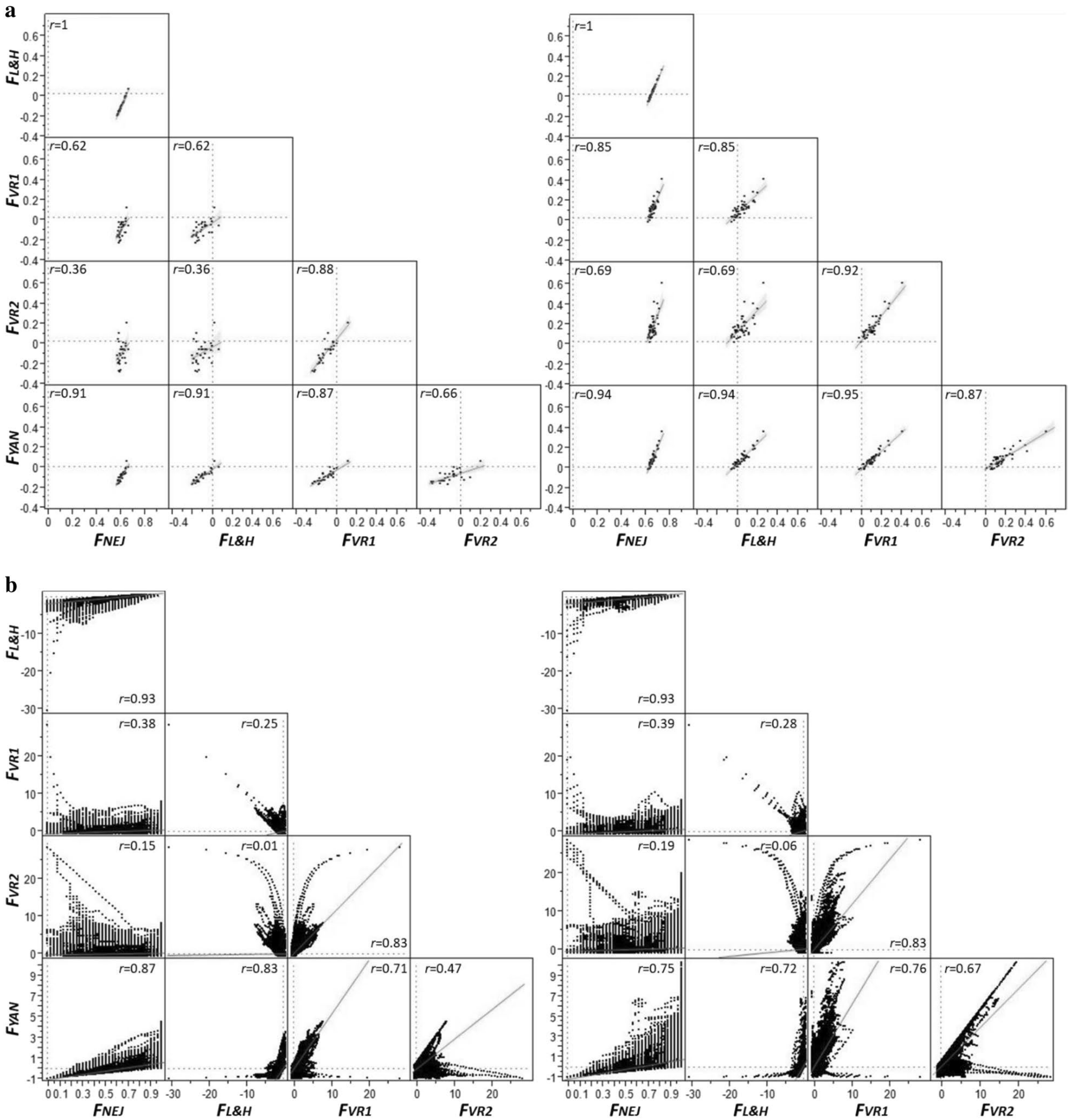

Fig. 3 Scatter plots for inbreeding coefficients F_{NEJ}, $F_{L\&H}$, F_{VR1}, F_{VR2} and F_{YAN} in the Guadyerbas population when computed at the individual animal (**a**) or genomic region (**b**) level in cohorts 1 (left panels) and 6 (right panels), and the corresponding correlation coefficients (*r*)

Figure 6 shows the patterns of each coefficient computed using only SNPs with a MAF > 0.05 or > 0.25 for three chromosomes. When only SNPs with a MAF higher than 0.05 in cohort 1 were used, some of the strong peaks previously obtained disappeared, in particular for F_{VR2} (Fig. 6 versus Fig. 5). Using an even stricter MAF filter (MAF > 0.25) led to very similar patterns for all inbreeding coefficients (Fig. 6). In fact, pairwise correlations between coefficients increased considerably compared to those shown in Fig. 3. When only SNPs with a MAF higher than 0.25 were used, all correlations were higher than 0.95, both in cohorts 1 and 6. SNPs with a MAF higher than 0.05 and higher than 0.25 represented 92% (16,532 SNPs) and 54%

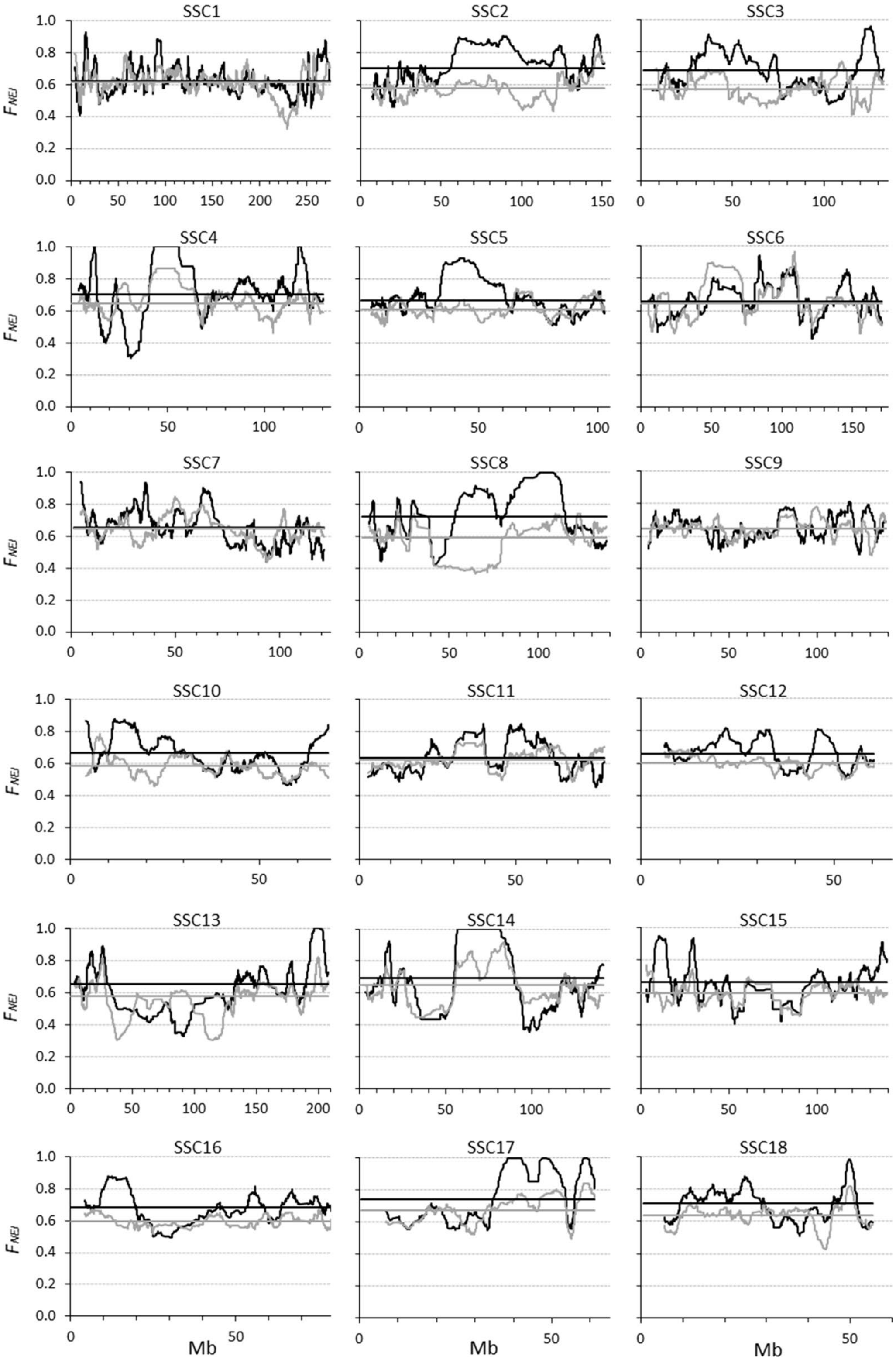

Fig. 4 Evolution of the proportion of the genome that becomes homozygous (i.e. F_{NEJ}) from cohort 1 (grey lines) to cohort 6 (black lines) for the different chromosomes (SSC) in the Guadyerbas population when using SNPs with non-zero minor allele frequencies. The horizontal lines represent averages across the genome

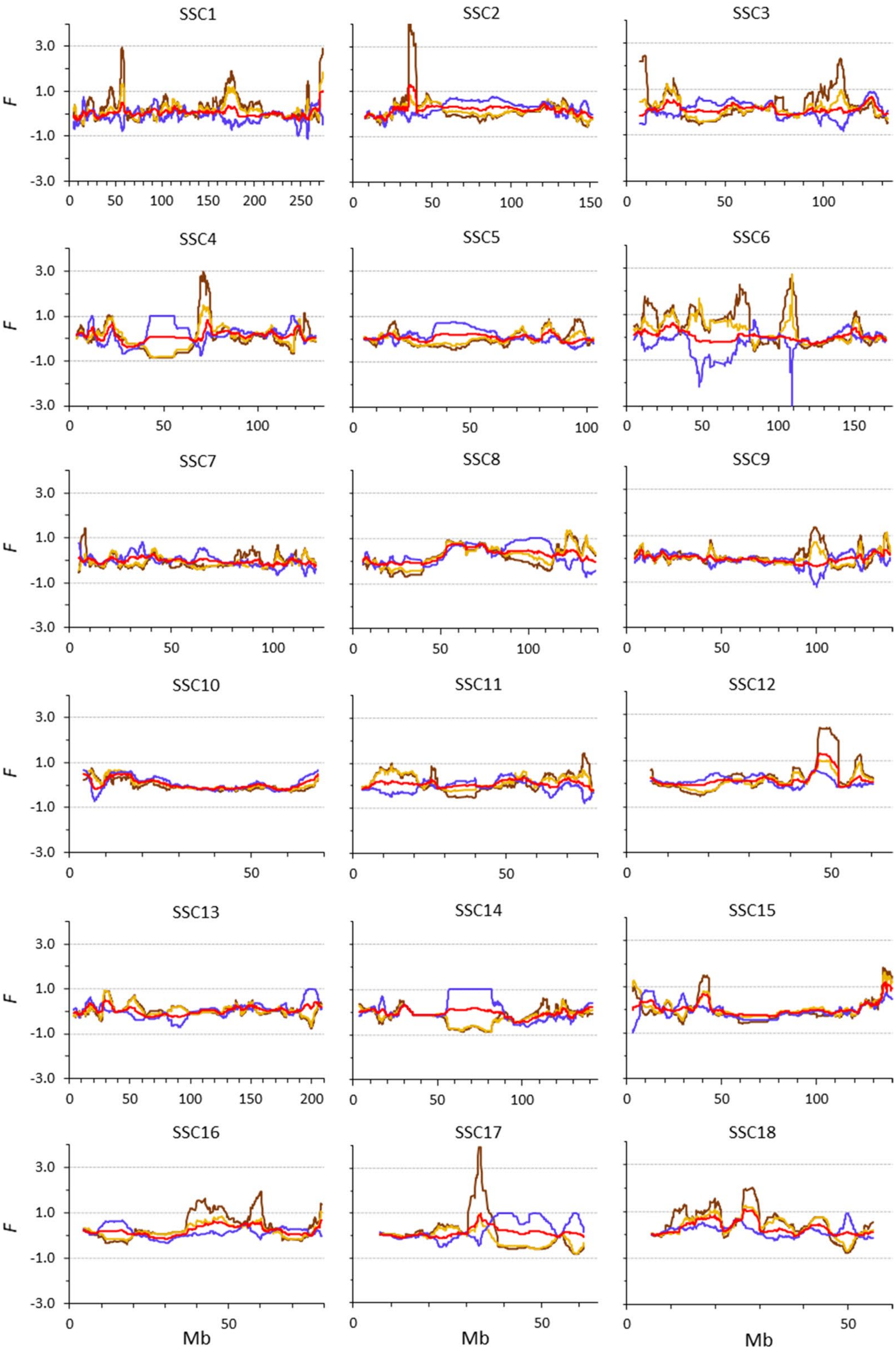

Fig. 5 Patterns of different measures of genomic inbreeding ($F_{L\&H}$ blue line, F_{VR1} light brown line, F_{VR2} dark brown line, F_{YAN} red line) in cohort 6 for different chromosomes (SSC) in the Guadyerbas population when using SNPs with non-zero minor allele frequencies in cohort 1

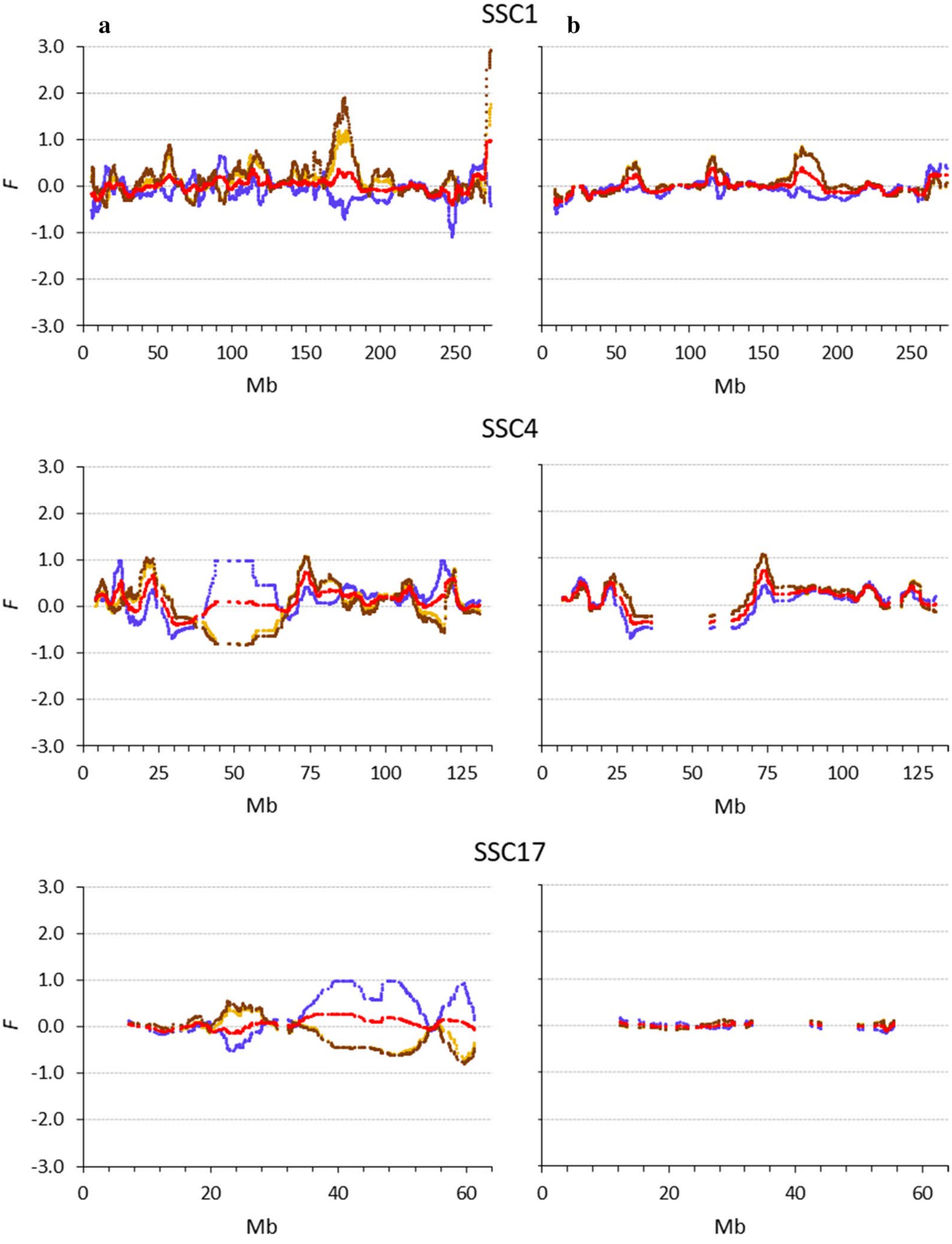

Fig. 6 Patterns of different measures of genomic inbreeding ($F_{L\&H}$ blue line, F_{VR1} light brown line, F_{VR2} dark brown line, F_{YAN} red line) in cohort 6 for chromosomes 1, 4, and 17 in the Guadyerbas population when using SNPs with minor allele frequencies > 0.05 (**a**) or > 0;25 (**b**) in cohort 1

(9,716 SNPs), respectively, of the total number of segregating SNPs in cohort 1. Note that SNP density greatly decreased in some regions when SNPs were filtered on MAF, resulting in the discontinuities seen in Fig. 6. These results show that the inconsistencies described earlier for F_{VR1}, F_{VR2}, and F_{YAN} occur when there are SNPs with a low MAF, and in practice such SNPs exist. Removing loci with rare alleles would defeat the rationale behind the coefficients that intentionally give more weight to rare alleles.

Consequences for the estimation of inbreeding depression

For each of the three measures of F, the patterns of the rate of inbreeding depression (i.e. the regression coefficient, b) for all chromosomes are shown in Additional file 2: Figure S2. Across the whole genome, the estimates of b differed substantially between the methods used to compute F. In some regions within chromosomes, estimates of b were very similar across methods but in other regions they differed greatly, not only in magnitude but also in sign. As an illustration, Fig. 7 shows selected regions within chromosomes for which the conclusions on the magnitude and sign of the rate of inbreeding depression differ substantially. In the regions from 50 to 70 Mb and from 98 and 109 Mb on SSC6, estimates of b were close to 0 when using $F_{L\&H}$ and F_{VR2} but clearly different from 0 when using F_{YAN}. However, in other regions (e.g. from 90 to 113 Mb on SSC8, from 20 to 24 Mb on SSC10, from 50 to 65 Mb on SSC14, and from 56 to 60 Mb on SSC17), $F_{L\&H}$ and F_{YAN} led to estimates of b that were of the same sign but opposite to estimates obtained when using F_{VR2}. In the region from 7.5 to 11 Mb on SSC18, the sign of the estimate of b obtained with $F_{L\&H}$ was opposite to that obtained with F_{VR2} and F_{YAN}.

Pairwise correlations between estimates of rates of inbreeding depression computed with $F_{L\&H}$, F_{VR2}, and F_{YAN} are in Additional file 3: Figure S3. Across the genome, correlations involving estimates based on F_{VR2} ranged from ~ 0.4 to 0.5, whereas the correlation between estimates based on $F_{L\&H}$ and F_{YAN} was high (0.84). About 40% of the estimates of b in Additional file 3: Figure S3 were of opposite sign when based on $F_{L\&H}$ and F_{VR2}, and this percentage decreased to ~ 27% when estimates were based on F_{VR2} and F_{YAN} and to ~ 15% when using $F_{L\&H}$ and F_{YAN}. This reinforces the idea that care should be taken when interpreting estimates of inbreeding depression obtained with different measures of genomic inbreeding.

Discussion

The inbreeding coefficient has been defined as a probability [9] or as a correlation [8], and thus its legitimate range is between 0 and 1 or between − 1 and 1, respectively. Another interpretation of the inbreeding coefficient, which we have used here, is in terms of loss or gain of variability relative to a reference base population. Under this interpretation, on the one hand, a negative value (even a value lower than − 1) makes sense and means that some variability has been gained. On the other hand, a value higher than 1 means that more variability than that initially existing has been lost, which is not reasonable.

Using a single locus model, we provided expectations for different genomic inbreeding coefficients that have been widely used in the literature. These expectations help to understand the patterns of these coefficients when they are computed using thousands of SNPs in a real population. Except for F_{NEJ}, none of the genomic coefficients considered here (i.e. those depending on allele frequencies) match with Malécot's or Wright's definition of the inbreeding coefficient as a probability or correlation, respectively, since their values can be outside

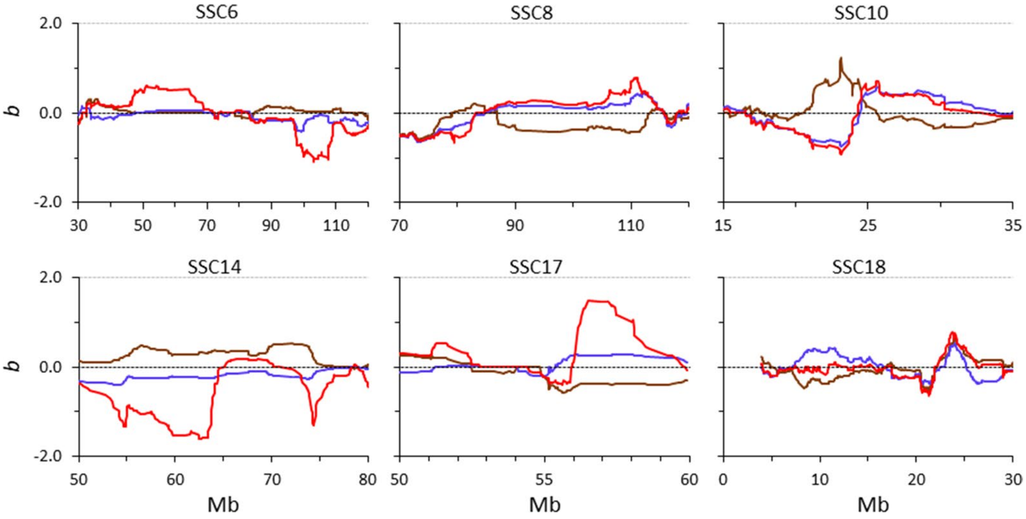

Fig. 7 Patterns of the rate of inbreeding depression (b) for number of piglets born alive in the Guadyerbas population when computed using different measures of genomic inbreeding ($F_{L\&H}$ blue line, F_{VR2} brown line, F_{YAN} red line) for specific regions of six chromosomes. All genotyped sows with phenotypic data that were born from cohort 1 to cohort 6 were included in the analyses

the legitimate ranges [47]. In fact, at the individual animal level, $F_{L\&H}$ can range from $-\infty$ to 1 and F_{VR1}, F_{VR2}, and F_{YAN} from -1 to ∞. At the population level (see the section on expected genomic coefficients under a single locus model above), the ranges are the same as at the individual level, except for F_{YAN}, which can range from 0 to ∞. When these coefficients are interpreted as indicators of whether variability is gained or lost over generations, $F_{L\&H}$ leads to sensible results but F_{VR1}, F_{VR2}, and F_{YAN} do not. This also has consequences when estimating the rate of inbreeding depression for specific genome regions since different measures of inbreeding can lead to very different results.

Although $F_{L\&H}$ is not a probability or a correlation, this measure of inbreeding is useful for determining whether variability is lost or gained. The largest variability (heterozygosity) for biallelic loci occurs at allele frequencies equal to 0.5. When a rare allele increases its frequency towards 0.5, $F_{L\&H}$ indicates that variability is gained, as expected. In addition, this measure of inbreeding never indicates that more variability than what existed in the initial generation was lost. In contrast, F_{VR1}, F_{VR2}, and $F_{L\&H}$ also do not match with a definition of inbreeding based on the proportion of variability lost or gained. In fact, for some $p_{(0)}$ and $p_{(t)}$ combinations, these three coefficients can indicate that variability is lost when the MAF increased towards 0.5, and this loss can be even higher than 100% of the initial variability. Moreover, F_{VR1} and F_{VR2} can take values that indicate that heterozygosity in the current generation is higher than what existed in the initial generation, although some heterozygosity has actually been lost. This does not occur with $F_{L\&H}$ at the population level since $E(F_{L\&H})$ is never negative, i.e. it always indicates that heterozygosity decreases although, in theory, it could increase.

One of the advantages of using genomic rather than pedigree data to measure inbreeding is the possibility of investigating the pattern of inbreeding along the genome. Here, we compared the patterns of each genomic coefficient computed from thousands of SNPs obtained with the Illumina PorcineSNP60 BeadChip v1 in a population of Iberian pigs. This population is highly inbred, with an estimated effective population size as low as 20 [54, 55]. The behaviour of each inbreeding coefficient observed with real data was well explained by the expectations developed. In the six generations (i.e. from cohort 1 to cohort 6), many SNPs became fixed (see Fig. 4). This loss of variability was captured by F_{NEJ} (which is simply the proportion of homozygous SNPs) and was also clearly reflected in the value for $F_{L\&H}$ ($F_{L\&H} = 1$). However, the negative values obtained for F_{VR1} and F_{VR2} for these regions indicate that variability in cohort 6 was higher than in cohort 1. For regions where the variability increased from cohort 1 to cohort 6, $F_{L\&H}$ had negative values (which reflects reality), while F_{VR1}, F_{VR2}, and F_{YAN} had positive values (which do not reflect the reality). Although $E(F_{YAN})$ predicts that variability can be gained in some circumstances, this occurs when, in fact, variability has been lost. In general, F_{VR1} and F_{VR2} behave similarly, although values for F_{VR2} are more extreme. Values for F_{YAN} lie between those for $F_{L\&H}$ and those for F_{VR1} and F_{VR2}, and generally, are close to 0. Given these results, it is clear that, when specific genome regions are targeted to control inbreeding, the choice of the genomic coefficient used should be done with care.

We have analysed the behaviour of five genomic measures of inbreeding that have been widely used in the literature. However, other measures have been proposed (see review by Kardos et al. [32]). For instance, both the PLINK [56] and GCTA [18] software provide a modification of $F_{L\&H}$ (their F^{II}, here referred to as $F_{L\&H2}$). Although to our knowledge, $F_{L\&H2}$ is not widely used, it is interesting to note that the difference between $F_{L\&H}$ and $F_{L\&H2}$ is equivalent to the difference between F_{VR1} and F_{VR2} in that it only differs in how the summation over SNPs is carried out. This is clearly illustrated by the patterns of these coefficients obtained for the Guadyerbas pig population (see Additional file 4: Figure S4). The patterns for $F_{L\&H}$ were, in general, mirror images of the patterns for F_{VR1} and those for $F_{L\&H2}$ were, in general, mirror images of patterns for F_{VR2}. Another widely used measure of genomic inbreeding, which we have not considered here, is the coefficient F_{ROH} based on continuous runs of homozygosity (ROH) [11, 12]. Contrary to the coefficients considered here, which are computed on a SNP-by-SNP basis, F_{ROH} is computed on a segment basis and has the advantages that (i) its values range from 0 to 1 (F_{ROH} is the proportion of the genome that is in ROH); and (ii) it can distinguish between distant (based on short ROH) and recent (based on long ROH) inbreeding. Its ability to detect inbreeding depression has been proven in multiple studies [3, 11, 12, 19, 21, 25, 26, 28, 32, 36, 39, 40, 42, 57]. However, the exact definition of F_{ROH} varies across studies, depending on the choice of the parameters to define a ROH (e.g. number of heterozygous genotypes permitted in a ROH, minimum SNP density required, maximum distance allowed between two consecutive homozygous SNPs, and minimum number of SNPs). Because of this, population-wide expected values for F_{ROH} are difficult to derive.

The pedigree-based numerator relationship matrix, NRM [58] has been used very extensively for many years to estimate the genetic covariance between individuals that are genetically evaluated via best linear unbiased prediction (BLUP). With the advent of genomic evaluations [59], the NRM has been replaced by more precise realised relationship matrices, which has led to an increase in the accuracy of predicted breeding values (e.g. [60]). Replacing NRM with GRM has also led to two other applications. The first application was the object of our study. Given that self-relationships in the NRM are expected to be equal to 1 plus the individual's inbreeding coefficient, genomic inbreeding coefficients have been also obtained from the diagonals of the GRM. However, as we have shown here, this is not always justified. In the ideal situation, with an infinite number of independent loci and absence of migration, mutation, and selection, the average allele frequencies remain constant over generations and all measures, except F_{NEJ}, are expected to produce unbiased estimates of the inbreeding coefficient (IBD) relative to a base population that is in Hardy–Weinberg equilibrium [42, 61]. However, in more realistic situations, the proposed genomic estimators of inbreeding can result in very different outcomes. The second application of the GRM is based on the fact that the NRM is equal to twice the matrix of coancestry coefficients and, as such, it has been used to optimize contributions of breeding candidates by applying the optimal contribution method (OC) for maintaining genetic variability and avoiding inbreeding depression in genetic conservation programs [47, 62, 63]. In this context, GRM have been used in OC, replacing NRM. de Cara et al. [64] and Gomez-Romano et al. [65] showed that the use of the coancestry matrix computed from Nejati-Javaremi's GRM in OC resulted in a higher level of genetic diversity (measured as expected heterozygosity) being maintained than when using the NRM in OC. Morales-González et al. [47] showed that the amount of genetic variability retained was higher when using Nejati-Javaremi's or Li and Horvitz's matrices in OC than when using VanRaden and Yang's GRM, although the latter were also efficient in controlling the loss of genetic diversity. Thus, in the context of optimizing contributions for maintaining diversity, VanRaden and Yang's GRM are useful. In fact, it has been suggested [66] that although the use of VanRaden and Yang's GRM in OC results in less variability being maintained, they could lead to allele frequencies that are closer to those in the original population (i.e. allele frequencies would tend to be unchanged), which can be an objective in conservation programs, particularly in ex situ conservation programs, where the final aim is reintroduction to the wild [67].

It has been suggested that the use of whole-genome sequence data could produce improved genomic inbreeding coefficient estimates [23] because it captures the many variants with rare alleles, which may not be included in the SNP panels due to their ascertainment bias. However, including a higher proportion of variants with rare alleles is expected to lead to even more inconsistent results than those shown here when using F_{YAN} and F_{VR1}, and particularly F_{VR2}.

Under the infinitesimal model, the NRM is a matrix of covariances of breeding values but, importantly, it is also twice the matrix of coancestry coefficients, with self-coancestries on the diagonal. Given that the relationship between self-coancestry (f) and inbreeding (F) coefficients is $f = 0.5(1 + F)$, the NRM provides estimates of F. GRM are also covariance matrices that have proven to work very well in genomic predictions. However, although F_{NEJ} and $F_{L\&H}$ correctly indicate when variability is lost or gained, this is not the case with F_{VR1}, F_{YVR2}, and F_{YAN}.

Conclusions

Except for F_{NEJ} (which ranges from 0 to 1), values for the genomic coefficients investigated here are outside the ranges of Malécot's and Wright's definitions of coefficient of inbreeding. When using a third interpretation of inbreeding in terms of loss or gain of variability, $F_{L\&H}$ gives sensible values but F_{VR1}, F_{VR2}, and F_{YAN} do not. In fact, the expectations derived here at the population level show some inconsistencies for these three coefficients. These include indications that (i) more variability than what initially existed can be lost (F_{VR1}, F_{VR2}, and F_{YAN}); (ii) variability has decreased when in reality it has increased (F_{VR1}, F_{VR2}, and F_{YAN}); (iii) variability has increased when in reality it has decreased (F_{VR1} and F_{VR2}); and (iv) it is not possible to gain more variability than what existed initially (F_{YAN}). The expectations developed here clearly explain the different patterns of these coefficients obtained for a highly inbred pig population when using thousands of SNP genotypes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-021-00635-0.

Additional file 1: Figure S1. Expected $F_{L\&H}$, F_{VR} ($F_{VR1} = F_{VR2}$) and F_{YAN} at a given current frequency ($p_{(t)}$) when the initial frequency of the reference allele ($p_{(0)}$) is set to 0.5. $F_{L\&H}$: blue line, F_{VR}: brown line, and F_{YAN}: red line.

Additional file 2: Figure S2. Patterns of the rate of inbreeding depression for number of piglets born alive (b) when computed using different measures of genomic inbreeding ($F_{L\&H}$: blue line, F_{VR2}: brown line, and F_{YAN}: red line) for each chromosome of the Guadyerbas genome. All genotyped sows with phenotypic data that were born from cohort 1 to cohort 6 were included in the analyses.

Additional file 3: Figure S3. Scatter plots for rates of inbreeding depression for number of piglets born alive (b) when computed using $F_{L\&H}$, F_{VR2} or F_{YAN} against each other, and corresponding correlation coefficients (*r*). All genotyped sows with phenotypic data that were born from cohort 1 to cohort 6 were included in the analyses. Values indicated with different colors correspond to regions presented in Fig. 7.

Additional file 4: Figure S4. Patterns of different measures of genomic inbreeding ($F_{L\&H}$: blue line, $F_{L\&H2}$: grey line, F_{VR1}: light brown line,

F_{VR2}: dark brown line, and F_{YAN}: red line) in cohort 6 for each chromosome of the Guadyerbas genome when using SNPs with a MAF higher than 0 in cohort 1.

Acknowledgements
We are grateful to Jaime Rodrigáñez, Luis Silió and the staff of the CIA 'Dehesón del Encinar' (Oropesa, Toledo) for their excellent job in maintaining the Guadyerbas herd for decades. We thank the 'Centro de Supercomputación de Galicia' (Cesga) for their computing support and two anonymous referees for helpful comments on the manuscript.

Authors' contributions
BV and RPW conceived and designed the study. AF performed the analyses and constructed the figures. BV wrote the first draft of the manuscript. All authors contributed to the discussion of results and the edition of the manuscript revision. All authors read and approved the final manuscript.

Funding
This work was carried out with funding from the Ministerio de Ciencia e Innovación, Spain (CGL2016-75904-C2, PID2020-114426GB-C22), the European Commission Horizon 2020 (H2020) Framework Programme through grant agreement no 727315 MedAID project (Mediterranean Aquaculture Integrated Development), Xunta de Galicia (ED431C 2020-05) and Fondos FEDER. R. Pong-Wong is funded by the European Union's Horizon 2020 research and innovation program under the Grant Agreement n°772787 (SMARTER) and the Biotechnology and Biological Sciences Research Council through Institute Strategic Programme Grant funding (BBS/E/D/30002275).

Availability of data and materials
The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Ethic Committee approval was not obtained for this study because no new animals were handled in this experiment. Analyses were performed using data previously collected with approval by the INIA Scientific Ethic Committee. Animal manipulations were performed according to the Spanish Policy for Animal Protection RD1201/05, which meets the European Union Directive 86/609 about the protection of animals used in experimentation.

Consent for publication
Not applicable.

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

Author details
[1]Departamento de Mejora Genética Animal, INIA, Ctra. de La Coruña, km 7.5, 28040 Madrid, Spain. [2]Centro de Investigación Mariña, Universidade de Vigo, Departamento de Bioquímica, Genética E Inmunología, Campus de Vigo, 36310 Vigo, Spain. [3]Departamento de Producción Agraria, ETSI Agrónomos, Universidad Politécnica de Madrid, 28040 Madrid, Spain. [4]Genetics and Genomics, The Roslin Institute and the R(D)SVS, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK.

Received: 17 July 2020 Accepted: 19 April 2021
Published online: 01 May 2021

References

1. Walsh B, Lynch M. Evolution and selection of quantitative traits. Oxford: Oxford University Press; 2018.
2. McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy B, Esko T, et al. Evidence of inbreeding depression in human height. PLoS Genet. 2012;8:e1002655.
3. Yengo L, Wray NR, Visscher PM. Extreme inbreeding in a European ancestry sample from the contemporary UK population. Nat Commun. 2019;10:3719.
4. Roff DA. Evolutionary quantitative genetics. New York: Chapman & Hall; 1997.
5. Frankham R, Ballou JD, Briscoe DA. Introduction to conservation genetics. 2nd ed. Cambridge: Cambridge University Press; 2010.
6. Falconer DS, MacKay TFC. Introduction to quantitative genetics. 4th ed. Harlow: Pearson Longman; 1996.
7. Caballero A. Quantitative genetics. Cambridge: Cambridge University Press; 2020.
8. Wright S. Systems of mating. Genetics. 1921;6:111–78.
9. Malécot G. Les mathématiques de l'hérédité. Paris: Masson et Cie; 1948.
10. Wright S. Coefficients of inbreeding and relationships. Am Nat. 1922;56:330–9.
11. Keller MC, Visscher PM, Goddard ME. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics. 2011;189:237–49.
12. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008;83:359–72.
13. Li CC, Horvitz DG. Some methods of estimating the inbreeding coefficient. Am J Hum Genet. 1953;5:107–17.
14. Nejati-Javaremi A, Smith C, Gibson JP. Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci. 1997;75:1738–45.
15. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
16. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–9.
17. VanRaden PM, Olson KM, Wiggans GR, Cole JB, Tooker ME. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 2011;94:5673–82.
18. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
19. Bjelland DW, Weigel K, Vukasinovic N, Nkrumah JD. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. J Dairy Sci. 2013;96:4697–706.
20. Saura M, Fernández A, Rodríguez MC, Toro MA, Barragán C, Fernández AI, Villanueva B. Genome-wide estimates of coancestry and inbreeding in a closed herd of Iberian pigs. PLoS One. 2013;8:e78314.
21. Pryce JE, Haile-Mariam M, Goddard ME, Hayes BJ. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet Sel Evol. 2014;46:71.
22. Wang J. Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. J Evol Biol. 2014;27:518–30.
23. Eynard SE, Windig JJ, Leroy G, van Binsbergen R, Calus MPL. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genet. 2015;16:24.
24. Howard JT, Haile-Mariam M, Pryce JE, Maltecca C. Investigation of regions impacting inbreeding depression and their association with the additive genetic effect for United States and Australia Jersey dairy cattle. BMC Genomics. 2015;16:813.
25. Kardos M, Luikart G, Allendorf FW. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity (Edinb). 2015;115:63–72.
26. Saura M, Fernández A, Varona L, Fernández AI, de Cara MAR, Barragán C, Villanueva B. Detecting inbreeding depression in reproductive traits in Iberian pigs using genome-wide data. Genet Sel Evol. 2015;47:1.
27. Zhang Q, Calus MPL, Guldbrandtsen B, Lund MS, Sahana G. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genet. 2015;16:88.
28. Bérénos C, Ellis PA, Pilkington JG, Pemberton JM. Genomic analysis reveals depression due to both individual and maternal inbreeding in a free-living mammal population. Mol Ecol. 2016;25:3152–68.
29. Eynard SE, Windig JJ, Hiemstra SJ, Calus MPL. Whole-genome sequence data uncover loss of genetic diversity due to selection. Genet Sel Evol. 2016;48:33.

30. Garbe JR, Prakapenka D, Tan C, Da Y. Genomic inbreeding and relatedness in wild panda populations. PLoS One. 2016;11:e0160496.
31. Huisman J, Kruuk LEB, Ellis PA, Clutton-Brock T, Pemberton JM. Inbreeding depression across the lifespan in a wild mammal population. Proc Natl Acad Sci USA. 2016;113:3585–90.
32. Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study of inbreeding depression in the wild. Evol Appl. 2016;9:1205–18.
33. Mastrangelo S, Tolone M, Di Gerlando R, Fontanesi L, Sardina MT, Portolano B. Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds. Animal. 2016;10:746–54.
34. Brito LF, Kijas JW, Ventura RV, Sargolzaei M, Porto-Neto LR, Cánovas A, et al. Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. BMC Genomics. 2017;18:229.
35. Solé M, Gori AS, Faux P, Bertrand A, Farnir F, Gautier M, et al. Age-based partitioning of individual genomic inbreeding levels in Belgian Blue cattle. Genet Sel Evol. 2017;49:92.
36. Yengo L, Zhu Z, Wray NR, Weir BS, Yang J, Robinson MR, et al. Detection and quantification of inbreeding depression for complex traits from SNP data. Proc Natl Acad Sci USA. 2017;114:8602–7.
37. Doekes HP, Veerkamp RF, Bijma P, Hiemstra SJ, Windig JJ. Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2018. Genet Sel Evol. 2018;50:15.
38. Baes CF, Makanjuola BO, Miglior F, Marras G, Howard JT, Fleming A, et al. Symposium review: the genomic architecture of inbreeding: How homozygosity affects health and performance. J Dairy Sci. 2019;102:2807–17.
39. Clark DW, Okada Y, Moore KHS, Mason D, Pirastu N, Gandin I, et al. Associations of autozygosity with a broad range of human phenotypes. Nat Commun. 2019;10:4957.
40. Nietlisbach P, Muff S, Reid JM, Whitlock MC, Keller LF. Non-equivalent lethal equivalents: models and inbreeding metrics for unbiased estimation of inbreeding load. Evol Appl. 2018;12:266–79.
41. Alemu SW, Kadri NK, Harland C, Charlier C, Faux P, Caballero A, et al. An evaluation of inbreeding measures using a whole genome sequenced cattle pedigree. Heredity (Edinb). 2020;126:410–23.
42. Caballero A, Villanueva B, Druet T. On the estimation of inbreeding depression using different measures of inbreeding from molecular markers. Evol Appl. 2020;14:416-28.
43. Legarra A, Aguilar I, Colleau JJ. Methods to compute genomic inbreeding for ungenotyped individuals. J Dairy Sci. 2020;103:3363–7.
44. Makanjuola BO, Miglior F, Abdalla EA, Maltecca C, Schenkel FS, Baes CF. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. J Dairy Sci. 2020;103:5183–99.
45. McGivney BA, Han H, Corduf LR, Katz LM, Tozaki T, MacHugh DE, et al. Genomic inbreeding trends, influential sire lines and selection in the global Thoroughbred horse population. Sci Rep. 2020;10:466.
46. Meuwissen THE, Sonesson AK, Gebregiwergis G, Woolliams JA. Management of genetic diversity in the era of genomics. Front Genet. 2020;11:880.
47. Morales-González E, Saura M, Fernández A, Fernández J, Pong-Wong R, Cabaleiro S, et al. Evaluating different genomic coancestry matrices for managing genetic variability in turbot. Aquaculture. 2020;520:734985.
48. Toro MA, Barragán C, Óvilo C, Rodrigáñez J, Rodríguez C, Silió L. Estimation of coancestry in Iberian pigs using molecular markers. Conserv Genet. 2002;3:309–20.
49. Toro MA, Villanueva B, Fernández J. Genomics applied to management strategies in conservation programmes. Livest Sci. 2014;166:48–53.
50. Toro MA, Rodrigañez J, Silió L, Rodríguez MC. Genealogical analysis of a closed herd of black hairless Iberian Pigs. Conserv Biol. 2000;14:1843–51.
51. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 2005;15:1468–76.
52. Engelsma KA, Veerkamp RF, Calus MPL, Bijma P, Windig JJ. Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. J Anim Breed Genet. 2012;129:195–205.
53. Kleinman-Ruiz D, Villanueva B, Fernández J, Toro MA, García-Cortés LA, Rodríguez-Ramilo ST. Intra-chromosomal estimates of inbreeding and coancestry in the Spanish Holstein cattle population. Livest Sci. 2016;185:34–42.
54. Saura M, Tenesa A, Woolliams JA, Fernández A, Villanueva B. Evaluation of the linkage-disequilibrium method for the estimation of effective population size when generations overlap: an empirical case. BMC Genomics. 2015;16:922.
55. Santiago E, Novo I, Pardiñas AF, Saura M, Wang J, Caballero A. Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. Mol Biol Evol. 2020;37:3642–53.
56. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
57. Kardos M, Nietlisbach P, Hedrick PW. How should we compare different genomic estimates of the strength of inbreeding depression? Proc Natl Acad Sci USA. 2018;115:E2492–3.
58. Henderson CR. Application of linear models in animal breeding. Guelph: University of Guelph Press; 1984.
59. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
60. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.
61. Toro MA, García-Cortés LA, Legarra A. A note on the rationale for estimating genealogical coancestry from molecular markers. Genet Sel Evol. 2011;43:27.
62. Villanueva B, Pong-Wong R, Woolliams JA, Avendaño S. Managing genetic resources in commercial breeding populations. In: Simm G, Villanueva B, Sinclair KD, Townsend S, editors. Farm animal genetic resources. BSAS Occasional Publication No. 30. Nottingham: Nottingham University Press; 2004. p. 113–32.
63. Fernández J, Toro MA, Caballero A. Fixed contributions designs vs. minimization of global coancestry to control inbreeding in small populations. Genetics. 2003;165:885–94.
64. de Cara MAR, Fernandez J, Toro MA, Villanueva B. Using genome-wide information to minimize the loss of diversity in conservation programmes. J Anim Breed Genet. 2011;128:456–64.
65. Gómez-Romano F, Villanueva B, de Cara MAR, Fernandez J. Maintaining genetic diversity using molecular coancestry: the effect of marker density and effective population size. Genet Sel Evol. 2013;45:38.
66. Gómez-Romano F, Villanueva B, Fernández J, Woolliams JA, Pong-Wong R. The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. Genet Sel Evol. 2016;48:2.
67. Saura M, Pérez-Figueroa A, Fernández J, Toro MA, Caballero A. Preserving population allele frequencies in ex situ conservation programs. Conserv Biol. 2008;22:1277–87.

Publisher's Note

genes

MDPI

Article

Changes in Allele Frequencies When Different Genomic Coancestry Matrices Are Used for Maintaining Genetic Diversity

Elisabeth Morales-González [1,*], Jesús Fernández [1], Ricardo Pong-Wong [2], Miguel Ángel Toro [3] and Beatriz Villanueva [1]

[1] Departamento de Mejora Genética Animal, INIA, Ctra. de La Coruña, km 7.5, 28040 Madrid, Spain; jmj@inia.es (J.F.); villanueva.beatriz@inia.es (B.V.)

[2] Genetics and Genomics, The Roslin Institute and R(D)SVS of the University of Edinburgh, Midlothian EH25 9RG, Roslin, UK; ricardo.pong-wong@roslin.ed.ac.uk

[3] Departamento de Producción Agraria, Universidad Politécnica de Madrid, 28040 Madrid, Spain; miguel.toro@upm.es

* Correspondence: morales.elisabeth@inia.es

Abstract: A main objective in conservation programs is to maintain genetic variability. This can be achieved using the Optimal Contributions (OC) method that optimizes the contributions of candidates to the next generation by minimizing the global coancestry. However, it has been argued that maintaining allele frequencies is also important. Different genomic coancestry matrices can be used on OC and the choice of the matrix will have an impact not only on the genetic variability maintained, but also on the change in allele frequencies. The objective of this study was to evaluate, through stochastic simulations, the genetic variability maintained and the trajectory of allele frequencies when using two different genomic coancestry matrices in OC to minimize the loss of diversity: (i) the matrix based on deviations of the observed number of alleles shared between two individuals from the expected numbers under Hardy–Weinberg equilibrium (θ_{LH}); and (ii) the matrix based on VanRaden's genomic relationship matrix (θ_{VR}). The results indicate that the use of θ_{LH} resulted in a higher genetic variability than the use of θ_{VR}. However, the use of θ_{VR} maintained allele frequencies closer to those in the base population than the use of θ_{LH}.

Keywords: genetic diversity; allele frequencies; genomic coancestry matrix; optimal contributions

Citation: Morales-González, E.; Fernández, J.; Pong-Wong, R.; Toro, M.Á.; Villanueva, B. Changes in Allele Frequencies When Different Genomic Coancestry Matrices Are Used for Maintaining Genetic Diversity. *Genes* **2021**, *12*, 673. https://doi.org/10.3390/genes12050673

Academic Editor: Nico M. Van Straalen

Received: 31 March 2021
Accepted: 29 April 2021
Published: 29 April 2021

1. Introduction

Genetic diversity is a prerequisite for populations to be able to face future environmental changes and to ensure long-term survival [1]. Thus, a common objective in genetic conservation programs is to minimize the loss of genetic variability. This can be achieved using the Optimal Contributions (OC) method that optimizes the contributions of candidates to the next generation by minimizing the global coancestry [2–4]. It has been demonstrated that OC maximizes genetic diversity measured as expected heterozygosity [5], which is proportional to the additive genetic variance of quantitative traits [6]. Controlling the loss of genetic diversity also keeps the inbreeding rate under control and therefore the risk of inbreeding depression.

A different objective in genetic conservation programs can be to maintain allele frequencies to preserve the uniqueness of a particular population, since current frequencies are the result not only of genetic drift, but also of previous selection processes [7–9]. Selection and drift can lead to a given allele responsible for a desirable trait at a high frequency. Moreover, trying to move the frequency to intermediate values to increase genetic variability would remove the uniqueness of the population. Thus, changes in the genetic composition of populations may be undesirable, particularly when dealing with ex situ conservation programs where the final aim is the reintroduction to the wild [9].

When the OC method is applied using pedigree information to compute coancestries, both objectives (maximum heterozygosity and maintenance of allele frequencies) are achieved [9], but this is not the case when coancestries are computed from molecular marker data. Previous studies have shown that using a coancestry matrix (θ) computed from large numbers of single nucleotide polymorphisms (SNPs) in OC is more efficient for maintaining diversity than using the pedigree-based coancestry matrix [10–12]. However, given that the highest expected heterozygosity is obtained at intermediate allele frequencies, a consequence of applying OC using a θ based on SNP genotypes is that the genetic composition of the population is modified [9–11,13,14].

Different genomic coancestry matrices have been proposed for being used in OC [10,11,15–17]. They include the matrix that describes deviations of the observed numbers of alleles shared by two individuals from the expected numbers under Hardy–Weinberg equilibrium [18], and those obtained from genomic relationship matrices currently used in genomic predictions [19,20]. In a recent study, Morales-González et al. [16] have shown that the expected heterozygosity retained through OC was higher when using the matrix proposed by Li and Horvitz [18] than when using different genomic relationship matrices (i.e., the VanRaden's matrices based on Method 1 and 2 [19] and the Yang's matrix [20]). However, as mentioned above, the genomic θ used in OC will have an impact not only on the diversity maintained, but also on the trajectory of the change in allele frequencies. Gómez-Romano et al. [21] suggested that while OC using a genomic coancestry matrix that simply measures the proportion of alleles shared by two individuals [22] and that correlates perfectly with Li and Horvitz's matrix favors solutions that tend to move allele frequencies towards 0.5, OC using VanRaden's matrices would lead to solutions that tend to keep allele frequencies closer to those in the original population (i.e., allele frequencies would tend to be unchanged). This has been recently confirmed by Meuwissen et al. [17] in the context of OC aimed at maximizing genetic gain through selection while restricting the increase in inbreeding (i.e., restricting the loss of genetic diversity).

In general, populations under conservation programs are small and genetic drift leads to a loss of diversity and changes in allele frequencies. The magnitude of these drift effects depends on the effective population size (N_e) which can be estimated from genomic coancestry. However, Toro et al. [23] have recently questioned the meaning of N_e when genomic matrices are used in OC. In particular, when optimal management is carried out using marker information, genetic diversity can increase in the initial generations implying negative estimates of N_e. Moreover, in the long term, N_e does not attain an asymptotic value, but it shows an unpredictable behavior. Their findings were based on OC using Nejati-Javaremi´s matrix [22] and it is unclear if they hold when other genomic coancestry matrices are used.

The objective of this study was to evaluate, through computer simulations, the genetic variability maintained and the trajectory of allele frequencies when different genomic coancestry matrices are used in OC. Estimates of N_e obtained from the change in heterozygosity computed from different genomic matrices were also compared.

2. Materials and Methods

Scenarios simulated involved the management of populations through the OC method using two different genomic coancestry matrices, for 50 discrete generations. Management started from a base population with family structure. The same base population was used for the 100 replicates run and it was created in two steps. Firstly, a population at mutation-drift equilibrium was generated. Secondly, the population was expanded in order to have enough individuals for sampling the 100 replicates (see below, in Section 2.1). The simulations were carried out with our own Fortran 90 codes.

2.1. Generation of the Base Population

The simulation of the base population was done in two steps to simulate a realistic amount of linkage disequilibrium and to ensure independency among the replicates. The first step was to generate a population in LD using a mutation-drift equilibrium approach. For this, 10,000 discrete generations of random mating for a population of 100 individuals (50 males and 50 females) were simulated. Using a larger population size would have generated an unrealistically low LD. Sires and dams were sampled with replacement and were mated at random. Each mating produced 2 offspring (1 of each sex). Thus, N_e was equal to 100. The genome was composed of 20 chromosomes of 1 Morgan each. Two types of biallelic loci (SNP and unobserved loci) were simulated and they differed simply in their subsequent use. SNP loci were used for computing the genomic coancestry matrices used in the management of the population that started after the base population was created. The unobserved loci were used for measuring diversity and changes of allele frequencies, and for estimating N_e across generations. Thus, the effect of different management strategies (i.e., using different genomic coancestry matrices) can be evaluated in the rest of the genome and not only on the loci used in the management (i.e., it is sometimes done using SNPs). A total of 500,000 SNPs and 500,000 unobserved loci were simulated per chromosome. At the initial generation, all loci were fixed. The mutation rate per locus and generation (μ) was 2.5×10^{-6} for all loci. The number of new mutations per generation was sampled from a Poisson distribution with mean $2N_e n_c \mu n_l$,, where n_c is the number of chromosomes (i.e., 20) and n_l is the total number of loci per chromosome (i.e., 1,000,000). Mutations were then randomly distributed across individuals, chromosomes and loci, switching allele 1 to allele 2 and vice versa. When generating the gametes, the number of crossovers per chromosome was drawn from a Poisson distribution with mean equal to 1. Crossovers were randomly distributed without interference. At the end of the process, the expected heterozygosity measured at both types of loci had stabilized (mutation-drift equilibrium). The second step consisted of expanding this population so we could sample the individuals to be used at the first generation of each replicate. The population was expanded during 4 generations with the aim of having enough individuals to sample 100 different replicates. During the 4 generations of expansion, each individual was randomly allocated to 8 different mates and each mating produced 1 offspring. In this way, the number of individuals in the population was multiplied by 4 each generation. After these 4 generations, the population was composed by 25,600 individuals and constituted the base population (t = 0). There were a total of 56,017 SNPs and 55,840 unobserved loci still segregating in t = 0. The expected heterozygosity (H_e) computed with all loci (SNPs and unobserved loci) still segregating was 0.1811 and the linkage disequilibrium (measured as r^2, the squared correlation between pairs of loci) between consecutive loci was 0.131.

2.2. Management Strategies

Management was performed on populations of two different sizes (N = 20 and N = 100 individuals, half of each sex) using the OC method across 50 generations. Population size was kept constant across generations. The founder individuals for each replicate were randomly sampled from the base population. Note that, given that the set of individuals sampled in t = 0 differs across replicates, the number of segregating loci can also differ. In most scenarios (see below, at the end of this section), all loci segregating in t = 0 were used for managing the population, for measuring diversity and changes of allele frequencies, and for estimating N_e.

The problem to be solved in the OC method is related to the allocation of contributions, i.e., the number of offspring each candidate should produce the next generation. The pursued strategy is to minimize the global coancestry weighted by those contributions, i.e., minimize $\mathbf{c}'\theta\ \mathbf{c}$, where $\mathbf{c}$ is a $N \times 1$ vector of proportions of offspring left by each candidate (i.e., the vector of solutions), N is the number of candidates and θ is the coancestry matrix. A restriction was imposed in the optimization such as the sum of the contributions of males and females is the same and equal to $\frac{1}{2}$, i.e., $\mathbf{Q}'\mathbf{c} = \frac{1}{2}\ \mathbf{1}$, where $\mathbf{Q}$ is a ($N \times 2$) known incidence matrix indicating the sex of the candidates with 0s and 1s, and 1 is a (2 × 1) vector of ones. The optimization problem was solved using Lagrangian multipliers [2,24]. Note that with this approach, $\mathbf{c}$ can contain negative values for some candidates. The contribution of candidates with $c_i < 0$ was then set to 0 and the optimization was repeated with the remaining candidates until all elements of $\mathbf{c}$ were non-negative. Finally, the contribution of individual i (c_i), which is a proportion, was converted to a number of offspring by multiplying c_i by $2N$ and rounding to the nearest integer but ensuring that the number of offspring of each sex equals to $N/2$. Each parent was randomly allocated to different mates (among the selected individuals) to produce its offspring.

Two management strategies were investigated, and they differed in the genomic coancestry matrix used in the optimization of contributions. Under strategy $\mathrm{S_{O\_LH}}$, the coancestry matrix used was matrix $\boldsymbol{\theta}_{\mathbf{LH}}$ which describes the excess in the observed number of alleles shared by two individuals relative to the expected number under Hardy–Weinberg equilibrium [18,25]. Specifically, the coancestry coefficient between individuals i and j was computed as

$$f_{LH(i,j)} = \frac{\sum_{k=1}^{S} f_{OBS(i,j)k} - S + 2\sum_{k=1}^{S} p_k(1-p_k)}{2\sum_{k=1}^{S} p_k(1-p_k)} \tag{1}$$

where $f_{OBS(i,j)}$ is the proportion of alleles shared by individuals i and j, S is the number of SNPs and p_k is the frequency of the reference allele (allele B) of SNP k in t = 0. Under strategy $\mathrm{S_{O\_VR}}$, the coancestry matrix used was matrix $\boldsymbol{\theta}_{\mathbf{VR}}$ which is based on the genomic relationship matrix obtained from VanRaden's method 2 [19]. Specifically, the coancestry coefficient between individuals i and j was computed as

$$f_{VR(i,j)} = \frac{1}{2S}\sum_{k=1}^{S} \frac{(x_{ki} - 2p_k)\left(x_{kj} - 2p_k\right)}{2p_k(1-p_k)} \tag{2}$$

where x_{ki} is the genotype of individual i for SNP k, coded as 0, 1 or 2 for genotypes AA, AB and BB, respectively, and p_k is as defined for f_{LH}.

In most scenarios, both coancestry matrices were computed every generation using all SNPs that were segregating in t = 0. However, we analyzed two additional scenarios where two different minor allele frequency (MAF) thresholds were imposed for the SNPs to be used to compute the coancestry matrices: (i) using only SNPs with MAF > 0.05; and (ii) using only SNPs with MAF > 0.25. The first threshold (MAF > 0.05) was considered because it is commonly applied when analyzing real data to reduce the number of potential genotyping errors. The second threshold (MAF > 0.25) was considered to explore the influence of rare alleles on the performance of the coancestry matrices investigated. It is known that with VanRaden's method rare alleles contribute more to the coancestry coefficient than common alleles [21,26]. It is, thus, interesting to determine how the differences between management strategies $\mathrm{S_{O\_LH}}$ and $\mathrm{S_{O\_VR}}$ vary in the different MAF scenarios. Management in these additional scenarios was performed for 50 generations.

Furthermore, as a benchmark, we simulated a strategy (strategy S_E) where the contributions of all candidates were equalized (i.e., all individuals contributed with two offspring to the next generation). This is the simplest management strategy that has been proposed to maintain genetic diversity by increasing N_e. It should be noticed that when dealing with populations in which the relationships between individuals are homogeneous (all equally related), this strategy leads to a N_e close to $2N$.

2.3. Parameters Evaluated

Management strategies were compared in terms of the genetic variability retained and the trajectory of the allele frequencies across generations for the SNPs and for the unobserved loci. Moreover, strategies were compared in terms of the number of individuals selected to produce the next generation (N_S) and the number of loci still segregating in a given generation, both for SNPs and for unobserved loci. The amount of genetic variability retained was measured as the expected heterozygosity (H_e) computed as $1 - \sum_{k=1}^{L}\sum_{l=1}^{2} p_{kl}^2$, where L is the number of loci (SNPs or unobserved loci) and p_{lk} is the frequency of allele l of locus k.

In order to evaluate the 'distance' between frequencies in a given generation t and frequencies in t = 0, we used the Kullback–Leibler (KL) divergence criterion, which measures how different is a particular distribution from a reference distribution [27], which here is the distribution of allele frequencies in t = 0. The KL divergence between current frequencies and frequencies in t = 0 was computed as

$$KL = \sum_{k=1}^{L}\sum_{l=1}^{2} p'_{kl} log \frac{p'_{kl}}{p_{kl}}, \quad (3)$$

where p_{kl} is the frequency of allele l of locus k in t = 0, and p'_{kl} is the corresponding frequency in the current generation (t > 0). The summation over alleles included only alleles with $p'_{kl} > 0$.

Finally, N_e was estimated from the change in heterozygosity in SNP loci. Thus, N_e in generation t was computed as $N_e = 1/2\ \Delta H_e$, where ΔH_e equals $H_{e(t-1)} - H_{e(t)}/H_{e(t-1)}$. All results presented are averages over the 100 replicates.

3. Results

3.1. Expected Heterozygosity and Kullback–Leibler Divergence for Populations of Size N = 100

For populations of size N = 100, and using all the SNPs segregating in t = 0, strategy S_{O_LH} led to higher genetic variability (measured as H_e) than strategy S_{O_VR} (Table 1) and the difference between both strategies increased across generations. In particular, H_e was about 1%, 4% and 11% higher with S_{O_LH} than with S_{O_VR} in t = 1, 10 and 50, respectively. With S_{O_LH}, H_e even slightly increased in the initial generations while with S_{O_VR}, H_e decreased from the start. Moreover, H_e obtained with strategy S_{O_VR} was very similar to H_e obtained with strategy S_E. Table 1 also shows that S_{O_VR} maintained allele frequencies closer to those in the base population than S_{O_LH} given that the KL values for S_{O_LH} were ≥ 100% higher than for S_{O_VR}. The differences in KL between both strategies increased across generations. Moreover, at later generations, S_{O_VR} was slightly more efficient in maintaining the initial frequencies than S_E, a strategy that is expected to maximize N_e and, thus, to minimize genetic drift.

Table 1. Expected heterozygosity (H_e, in %) and Kullback–Leibler divergence for unobserved loci ($KL \times 10^2$), number of selected candidates (N_S), and number of SNPs (S) and unobserved loci (U) segregating across generations (t) when contributions are equalized (S_E) and when they are optimized using Li and Horvitz (S_{O_LH}) and VanRaden (S_{O_VR}) coancestry matrices computed with SNPs with MAF > 0.00 in a population of 100 individuals.

	S_E					S_{O_LH} *					S_{O_VR} *				
t	H_e	KL	N_S	S	U	H_e	KL	N_S	S	U	H_e	KL	N_S	S	U
1	19.17	0.06	100	51,035	50,894	+0.14	+0.14	−39	−2239	−2246	0.00	0.00	0	+8	+18
2	19.12	0.12	100	49,873	49,737	+0.21	+0.23	−36	−3206	−3229	0.00	0.00	0	−22	0
3	19.07	0.18	100	48,852	48,729	+0.28	+0.30	−35	−3792	−3847	0.00	0.00	0	−61	−52
4	19.03	0.24	100	47,946	47,828	+0.35	+0.37	−35	−4182	−4261	0.00	0.00	−1	−113	−101
5	18.98	0.30	100	47,108	47,003	+0.41	+0.43	−33	−4384	−4499	0.00	−0.01	−1	−162	−157
10	18.73	0.57	100	43,777	43,691	+0.68	+0.68	−30	−4731	−4975	0.00	−0.03	−2	−399	−401
15	18.51	0.82	100	41,311	41,217	+0.89	+0.86	−28	−4523	−4855	−0.01	−0.06	−5	−595	−587
20	18.27	1.06	100	39,313	39,229	+1.08	+0.99	−26	−4152	−4567	−0.01	−0.09	−6	−714	−720
30	17.82	1.50	100	36,231	36,140	+1.40	+1.16	−24	−3329	−3896	+0.01	−0.18	−9	−906	−899
40	17.38	1.90	100	33,854	33,759	+1.67	+1.24	−22	−2517	−3215	+0.03	−0.26	−11	−995	−970
50	16.95	2.28	100	31,940	31,848	+1.92	+1.27	−21	−1786	−2594	+0.05	−0.35	−12	−1081	−1036

* S_{O_LH} and S_{O_VR} values are those deviated from S_E. Standard errors (computed across replicates) ranged from 4.91×10^{-5} to 9.54×10^{-5} for H_e and from 0.16×10^{-5} to 7.39×10^{-5} for KL.

The use of both matrices (θ_{LH} and θ_{VR}) in OC also led to different numbers of individuals selected as parents of the next generation (N_S). In particular, with S_{O_LH}, between 10% and 30% fewer individuals were selected than with S_{O_VR} (Table 1). In fact, with the latter, almost all individuals were selected in all generations up to t = 10. The difference in N_S entailed a difference in the number of loci that remained segregating across generations that was much higher with S_{O_VR} than with S_{O_LH} (Table 1), particularly in the initial generations. As for H_e and for *KL*, strategies S_{O_VR} and S_E led to very similar values of N_S.

Table 2 shows the evolution across generations of the average frequency of the minor allele in t = 0. This average frequency was practically constant with S_E and slightly decreased with S_{O_VR}. However, with S_{O_LH}, it increased from ~1% in t = 1 to 16–19% in t = 50. Thus, it is clear that S_{O_LH} leads average frequencies upward (ultimately towards 0.5) and S_{O_VR} tends to maintain them. As expected, these patterns were more evident for the SNPs than for the unobserved loci.

Table 2. Average frequency of the minor allele in generation 0 (× 10^2) across generations (t) for SNPs and unobserved loci when contributions are equalized (S_E) and when they are optimized using Li and Horvitz (S_{O_LH}) and VanRaden (S_{O_VR}) coancestry matrices in a population of 100 individuals.

	SNPs			Unobserved Loci		
t	S_E	S_{O_LH}	S_{O_VR}	S_E	S_{O_LH}	S_{O_VR}
0	13.45	13.45	13.45	13.39	13.39	13.39
1	13.44	13.68	13.45	13.39	13.60	13.40
2	13.44	13.81	13.45	13.39	13.72	13.40
3	13.44	13.94	13.45	13.38	13.82	13.39
4	13.44	14.06	13.44	13.38	13.93	13.39
5	13.44	14.17	13.44	13.38	14.02	13.39
10	13.44	14.67	13.41	13.38	14.44	13.36
15	13.45	15.08	13.37	13.39	14.77	13.33
20	13.44	15.42	13.32	13.39	15.05	13.29
30	13.44	15.96	13.23	13.39	15.46	13.23
40	13.45	16.36	13.12	13.39	15.75	13.15
50	13.45	16.67	13.01	13.40	15.98	13.07

Figures 1 and 2 show the frequency (f) distribution also for minor alleles in t = 0 in this generation and after 50 generations of management, using different sets of SNPs to compute coancestries. When using all SNPs segregating in t = 0, the distributions for SNPs and unobserved loci were very similar (Figures 1a and 2a). However, when using only SNPs with MAF > 0.05 or MAF > 0.25, the distribution for SNPs was greatly affected. When using SNPs with MAF > 0 or MAF > 0.05 (Figure 1a,b), a greater number of SNPs was fixed with S_{O_LH} than with S_{O_VR} across generations (see class f = 0.00). However, more loci (SNPs and unobserved loci) with low frequencies ($0.00 < f \leq 0.15$) were observed with S_{O_VR} than with S_{O_LH} and more loci with higher frequencies (f > 0.4) were observed with S_{O_LH} than with S_{O_VR}. Thus, although more alleles are fixed with S_{O_LH}, those that are kept segregating increase their frequency, while with S_{O_VR} the frequencies tend to be maintained. The highest difference between SNPs and unobserved loci was found when only SNPs with MAF > 0.25 were used to estimate the coancestry matrices (Figures 1c and 2c). These differences are due to the fact that no MAF filtering was done for the unobserved loci.

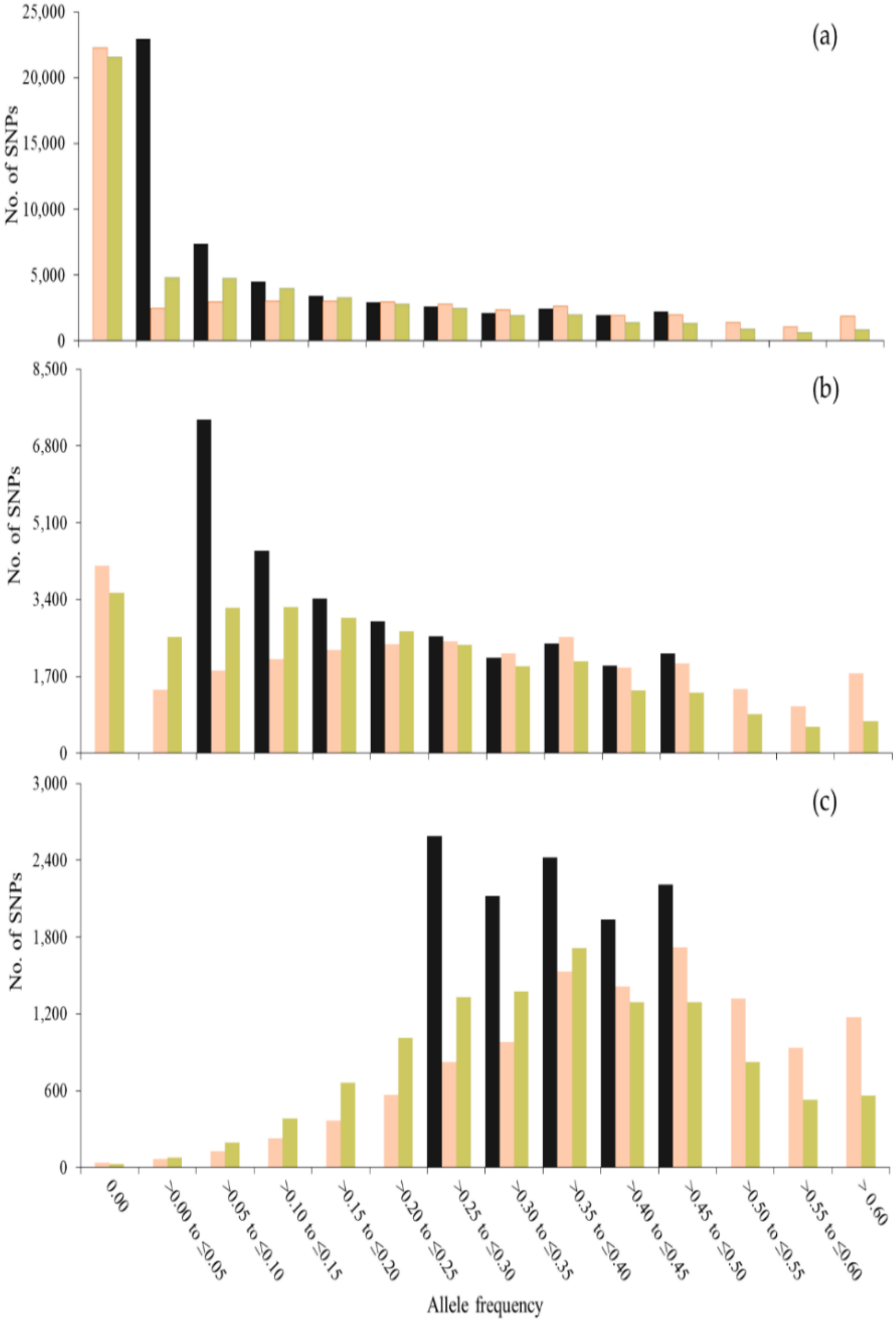

Figure 1. Number of SNPs for each class of allele frequency of the allele that was minor at generation 0 (gray bars) and the frequency of this allele after 50 generations, when contributions are optimized using Li and Horvitz (S_{O_LH}, in orange) and VanRaden (S_{O_VR}, in green) coancestry matrices computed with SNPs with MAF > 0.00 (**a**), MAF > 0.05 (**b**) and MAF > 0.25 (**c**) in a population of 100 individuals.

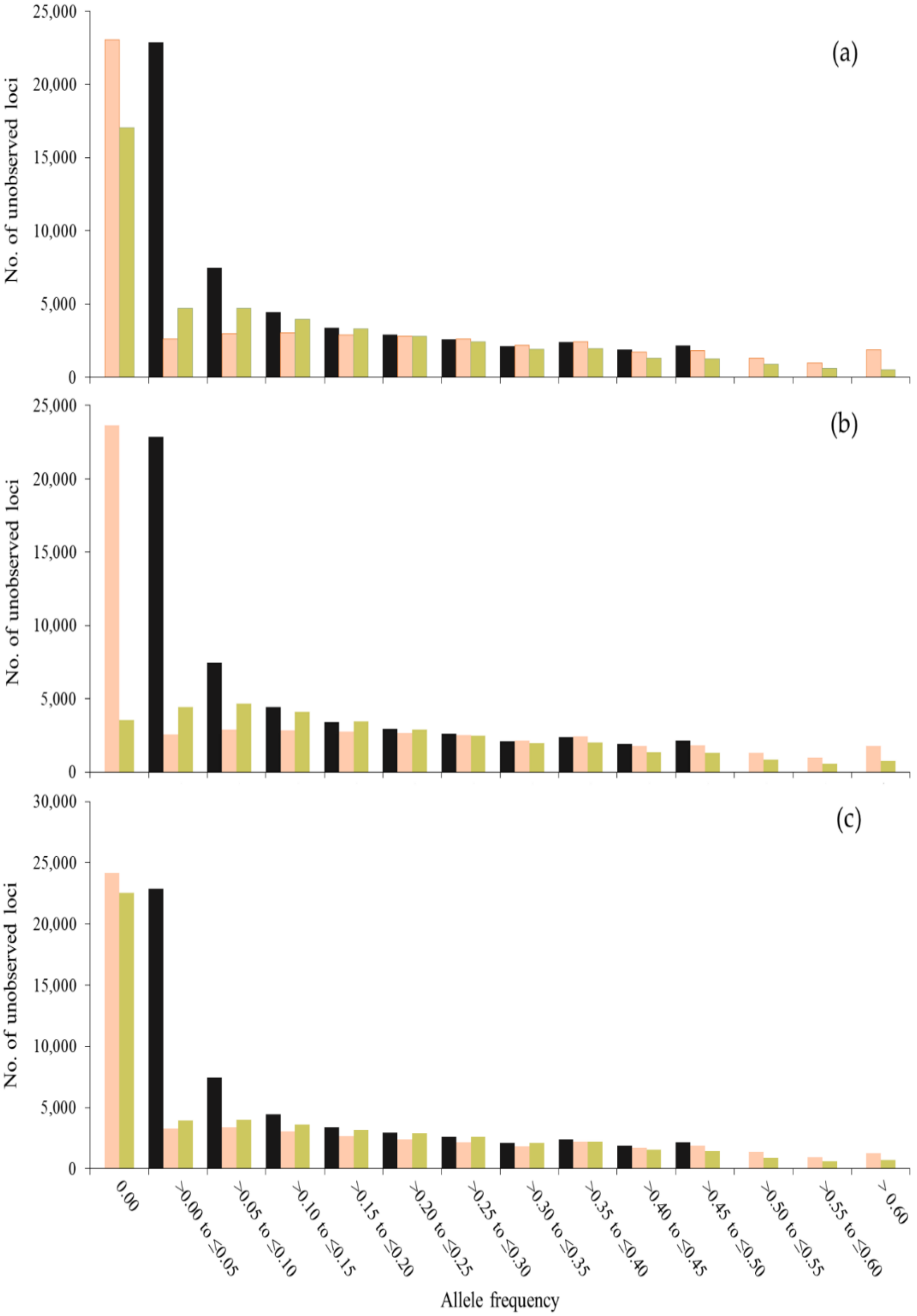

Figure 2. Number of unobserved loci for each class of allele frequency of the allele that was minor at generation 0 (gray bars) and the frequency of this allele after 50 generations, when contributions are optimized using Li and Horvitz (S_{O_LH}, in orange) and VanRaden (S_{O_VR}, in green) coancestry matrices computed with SNPs with MAF > 0.00 (**a**), MAF > 0.05 (**b**) and MAF > 0.25 (**c**) in a population of 100 individuals.

Figure 3 shows the trajectories of H_e and KL across generations for unobserved loci under strategies S_{O_LH} and S_{O_VR} using the three different sets of SNPs. The heterozygosity maintained with S_{O_LH} decreased as the MAF criterion chosen for the SNPs used to estimate coancestries becomes more restrictive given that the number of SNPs used decreased. In fact, the small increase in H_e observed in the initial generations when using all SNPs (MAF> 0.00) was not observed when using only the SNPs with MAF > 0.05 or MAF > 0.25. In parallel, the KL divergence with S_{O_LH} also decreased when increasing the severity of the restriction imposed on the SNPs used. However, with S_{O_VR}, the changes observed in H_e and KL when using a different set of SNPs were very small.

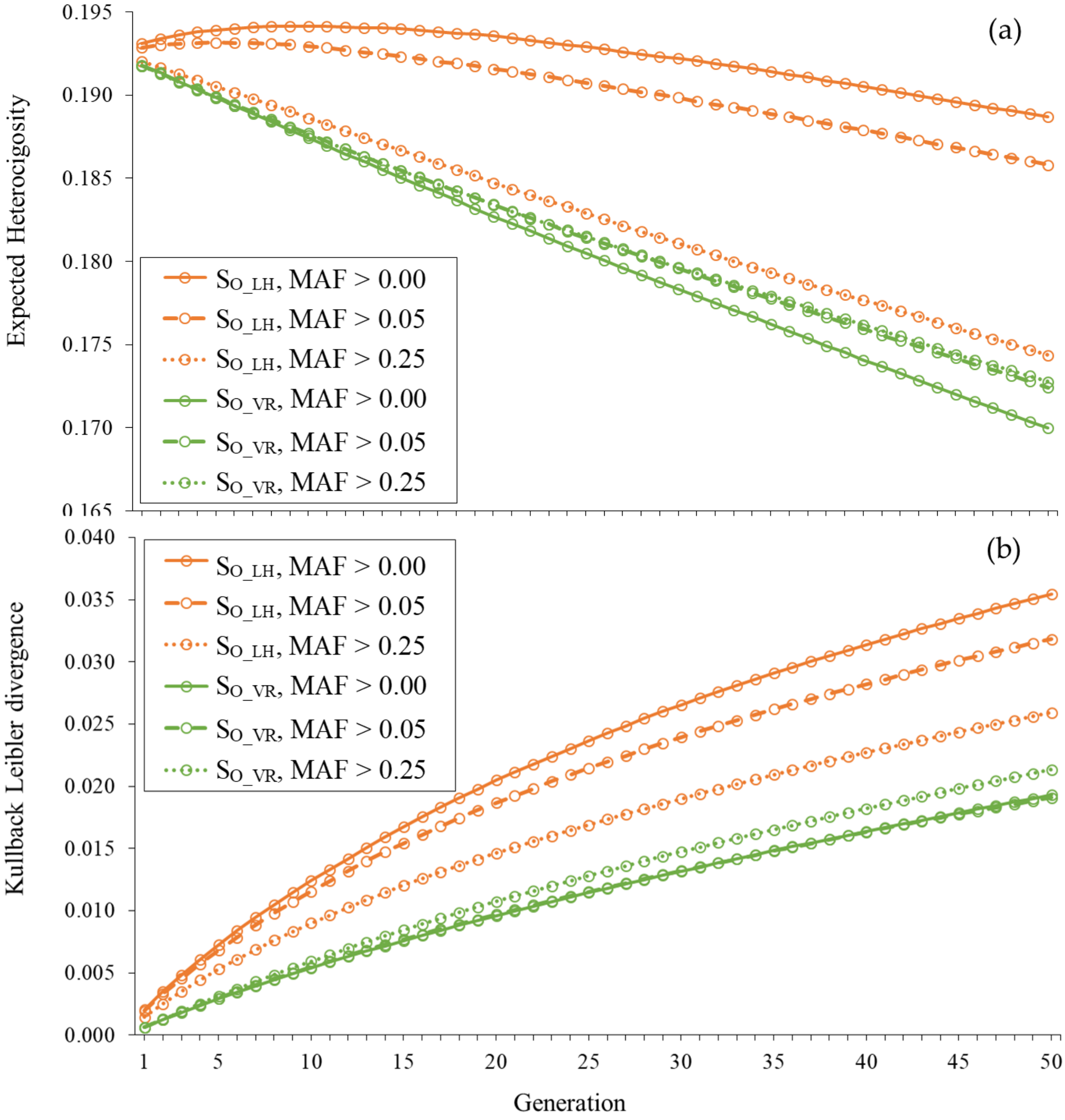

Figure 3. Expected heterozygosity (**a**) and Kullback–Leibler divergence (**b**) for unobserved loci across generations when contributions are optimized using Li and Horvitz (S_{O_LH}) and VanRaden (S_{O_VR}) coancestry matrices computed with SNPs with MAF > 0.00, MAF > 0.05 and MAF > 0.25 in a population of 100 individuals.

3.2. Expected Heterozygosity and Kullback–Leibler Divergence for Populations of Size N = 20

Table 3 shows results from the different strategies (S_E, S_{O_LH} and S_{O_VR}) for populations of size N = 20, when all SNPs segregating in t = 0 were used in the management. Similar to the results found for populations of N = 100, (i) S_{O_LH} led to higher H_e than S_{O_VR} and S_E; and (ii) S_{O_VR} maintained allele frequencies closer to those in t = 0 than S_{O_LH}. However, differences among strategies were smaller for populations of N = 20. For instance, for N = 20, H_e in t = 10 was less than 1% higher when managing with S_{O_LH} than when managing with S_{O_VR}, while for N = 100 this percentage was about 4%. For KL, the highest difference between strategies was 0.0027 units with N = 20 and 0.0127 units with N = 100. However, with N = 20, contrary to what happened with N = 100, S_{O_LH} managed to keep frequencies closer to the initial frequencies than S_E in the last generations ($t \geq 30$).

Table 3. Expected heterozygosity (H_e, in %) and Kullback–Leibler divergence for unobserved loci ($KL \times 10^2$), number of selected candidates (N_S), and number of SNPs (S) and unobserved loci (U) segregating across generations (t) when contributions are equalized (S_E) and when they are optimized using Li and Horvitz (S_{O_LH}) and VanRaden (S_{O_VR}) coancestry matrices computed with SNPs with MAF > 0.00 in a population of 20 individuals.

	S_E					S_{O_LH} *					S_{O_VR} *				
t	H_e	KL	N_S	S	U	H_e	KL	N_S	S	U	H_e	KL	N_S	S	U
1	23.35	0.27	20	38,995	38,955	+0.04	+0.05	−1	−193	−233	+0.03	0.00	0	+31	+134
2	23.06	0.52	20	37,093	37,050	+0.06	+0.07	−1	−275	−335	+0.01	0.00	0	+52	+155
3	22.76	0.76	20	35,522	35,472	+0.10	+0.09	−1	−356	−410	−0.02	+0.01	0	−12	+104
4	22.48	0.99	20	34,166	34,119	+0.07	+0.11	−1	−390	−442	−0.02	−0.01	0	−16	+94
5	22.19	1.20	20	33,016	32,978	+0.08	+0.13	−1	−456	−528	−0.03	0.00	0	−69	+37
10	20.79	2.17	20	28,782	28,692	+0.17	+0.18	−1	−533	−563	−0.07	−0.03	−1	−269	−62
15	19.52	3.00	20	25,844	25,763	+0.24	+0.17	−1	−497	−563	−0.03	−0.07	−1	−400	−206
20	18.33	3.75	20	23,512	23,434	+0.37	+0.13	−1	−336	−424	−0.01	−0.12	−1	−429	−247
30	16.02	5.13	20	19,854	19,795	+0.79	−0.02	−2	+81	−59	+0.04	−0.25	−2	−469	−337
40	14.03	6.26	20	17,044	17,002	+1.15	−0.16	−1	+545	+377	+0.18	−0.43	−2	−432	−309
50	12.32	7.23	20	14,853	14,811	+1.39	−0.27	−1	+787	+592	+0.19	−0.52	−2	−433	−322

* S_{O_LH} and S_{O_VR} values are those deviated from S_E. Standard errors (computed across replicates) ranged from 1.15×10^{-4} to 3.37×10^{-4} for H_e and from 10×10^{-4} to 1.72×10^{-4} for KL.

In populations of size N = 20, individuals are more closely related than in populations of size N = 100 and the genetic variability is smaller. Thus, most (if not all) individuals were selected to be parents of the next generation with all management strategies across generations. It should be noted that the number of loci segregating in t = 0, when management started, was substantially smaller when simulating populations of size N = 20. In order to investigate if the differences observed between N = 20 and N = 100 are a consequence of the different number of loci segregating in t = 0, a scenario with N = 100 starting with the same number of SNPs as in the scenario with N = 20 (about 40,000 SNPs) was simulated. The results indicate that the differences between scenarios with different N were due to the population size and not to the different number of loci (results not shown).

3.3. Effective Population Size

Table 4 shows estimates of N_e across generations for the different scenarios simulated. For N = 100, estimates of N_e were around 200 individuals under strategies S_E and S_{O_VR}. This is the expected value for N_e when contributions are equalized since N_e is approximately equal to $2N$. However, under strategy S_{O_LH}, estimates of N_e were unreasonable as they took negative values in the initial generations. In later generations, N_e became positive but did not reach a stable value. For N = 20, estimates under strategies S_E and S_{O_VR} were around 40 individuals, as expected. Estimates of N_e under strategy S_{O_LH} were between 6% and 50% higher than under strategy S_E.

Table 4. Effective population size (N_e) across generations (t) when contributions are equalized (S_E) and when they are optimized using Li and Horvitz (S_{O_LH}) and VanRaden (S_{O_VR}) coancestry matrices in populations of different sizes (N).

	N = 100			N = 20		
t	S_E	S_{O_LH}	S_{O_VR}	S_E	S_{O_LH}	S_{O_VR}
1	188.21	−111.90	195.55	36.92	42.27	40.40
5	199.07	−855.78	197.46	36.78	41.24	34.31
10	191.56	−5777.32	193.05	38.54	40.81	41.77
15	203.50	1855.71	194.54	36.65	45.41	43.18
20	202.62	1033.03	201.52	40.61	47.25	40.02
25	190.44	636.00	209.85	40.20	47.08	42.02
30	193.58	670.07	209.79	36.45	53.03	38.57
35	193.30	524.97	206.03	33.41	50.28	44.62
40	204.95	601.67	212.53	36.94	47.91	49.68
45	207.44	703.31	205.00	37.52	48.50	40.09
50	206.86	481.08	213.02	41.99	46.20	38.53

4. Discussion

Using computer simulations, this study has compared two different management strategies in terms of two important criteria in genetic conservation programs, i.e., genetic diversity (H_e) maintained and changes in allele frequencies. Both strategies optimize contributions for maintaining diversity but differ in the genomic coancestry matrix used in the optimization (θ_{LH} in strategy S_{O_LH} and θ_{VR} in strategy S_{O_VR}). Moreover, as a benchmark, the simplest management strategy proposed to maintain genetic diversity that implies equalizing the contributions of all candidates (strategy S_E) was evaluated.

The changes in allele frequencies were evaluated using the *KL* divergence criterion. The greater the value of *KL*, the greater the divergence of frequencies with respect to the frequencies in the base population. When the strategies were compared using the *KL* criterion, it was clear that strategy S_{O_LH} gives higher values than strategy S_{O_VR}, indicating that the latter is able to maintain allele frequencies closer to the original frequencies (lower *KL* values). On the other hand, with strategy S_{O_LH}, the population evolves differently as it pushes frequencies towards 0.5 and thus changes the genetic composition of the population more than strategy S_{O_VR}.

Pushing frequencies towards 0.5 as strategy S_{O_LH} does leads to higher genetic variability when measured as expected heterozygosity. Thus, the hypothesis raised by Gómez-Romano et al. [21] that using matrix θ_{LH} in OC designed for maintaining genetic diversity better achieves the objective (i.e., higher H_e) than using matrix θ_{VR}, but using the latter maintains allele frequencies closer to the initial frequencies, is confirmed. This was observed both in populations with N = 20 and in populations with N = 100 although the differences between both strategies were smaller with N = 20. This is because individuals in the smaller populations are more closely related and there are less options to choose among individuals and strategies behave more similarly.

Saura et al. [9] showed that the use of the pedigree-based coancestry matrix in OC maintained allele frequencies close to those of the initial population. This is related to the high levels of N_e obtained when minimizing pedigree coancestry (close to 2N), leading to reduced drift and little departures to the original frequencies. Additionally, several studies [10,12] have shown that OC based on pedigrees leads to less maintained genetic diversity than the use of genomic coefficients based on Nejati-Javaremi´s matrix [22]. This is due to the fact that genomic data provide realized estimates of coancestry, while pedigree data provide expected values. Therefore, results under the management of populations with OC using the pedigree-based coancestry matrix would be similar to those under S_{O_VR}.

Strategy S_{O_VR} was only slightly more efficient for maintaining frequencies than strategy S_E. This strategy tends to reduce the change in allele frequencies, which implies a reduced genetic drift [17]. The magnitude of drift is minimized when N_e equals approximately 2N, and it is well known that, when managing the population using pedigree information (as said before), this is achieved by equalizing contributions [6,28]. The small advantage of $\mathrm{S_{O\_VR}}$ in terms of maintaining frequencies over $\mathrm{S_E}$ arises from the fact that the former uses realized relationships and detects real differences between individuals while $\mathrm{S_E}$ assumes homogeneous relationships. Contrarily, $\mathrm{S_{O\_LH}}$ does not minimize drift but maximizes H_e by shifting frequencies towards 0.5. Thus, results from $\mathrm{S_{O\_LH}}$ are quite different to those obtained under $\mathrm{S_E}$ in terms of the number of selected candidates and their optimal contributions.

Given that strategy $\mathrm{S_{O\_LH}}$ brings the frequencies towards 0.5, H_e increased in the initial generations and this led to negative estimates of N_e in the largest population (N = 100). As generations go by, N_e becomes positive but with unrealistic very high values without attaining an asymptotic value. This was also observed by Toro et al. [23] who questioned the meaning of N_e when genomic coancestry matrices are used in OC. They showed an unpredictable behavior for N_e when using the similarity genomic matrix of Nejati-Javaremi et al. [22], which has a correlation of 1 with the θ_{LH} matrix used here [5,16,29]. However, our results show that when using θ_{VR} in OC, estimates of N_e were close to the expected value when equalizing contributions (approximately 2N). As has been discussed above, the results from strategy $\mathrm{S_{O\_VR}}$ were very similar to those from strategy $\mathrm{S_E}$ given that both tend to minimize drift. For the smallest population considered (N = 20), estimates of N_e were close to 2N not only with $\mathrm{S_{O\_VR}}$ but also with $\mathrm{S_{O\_LH}}$. In such a small population, there are fewer options to choose among individuals and most of them are selected to contribute (Table 3). Thus, the three strategies investigated led to similar results.

Strategy $\mathrm{S_{O\_LH}}$ led to higher H_e but also to a higher loss of segregating loci than strategy $\mathrm{S_{O\_VR}}$. In the largest population (N = 100), the percentage of alleles lost for unobserved loci at t = 1 was 13% and 9% with $\mathrm{S_{O\_LH}}$ and $\mathrm{S_{O\_VR}}$, respectively (Table 1). The difference in both management strategies in terms of the number of alleles lost could be due to a different number of individuals selected to contribute to the next generation that was lower with $\mathrm{S_{O\_LH}}$. It must be emphasized that the mean coancestry of each individual with all the candidates (including the individual), i.e., the marginal of the coancestry matrix, is a useful concept for understanding the different numbers selected with both strategies. This is because the marginal of the coancestry matrix is a measure of the 'relevance' of each individual, in terms of the degree of genetic information shared with the rest, and the optimal solutions will depend on all relationships between candidates. Its value is the same for all candidates when considering θ_{VR}. Then, all candidates are equally useful and should be selected as it was observed minimizing the global coancestry through OC using θ_{VR} (strategy $\mathrm{S_{O\_VR}}$). However, when considering θ_{LH}, the average coancestry of individuals *AA* (homozygous for the minor allele) is lower than that of individuals *BB* (homozygous for the major allele), since individuals *AA* harbor genetic information that is underrepresented (i.e., they carry the rarer allele) and should be favored for selection and contributions. Therefore, OC using θ_{LH} minimize the objective function when selecting the same number of *AA* and *BB* candidates. This leads to an increase in the frequency of allele *A* (actually to 0.5 in a single generation in this example with only one locus) while frequencies stay unchanged when using θ_{VR}.

Fernández et al. [13] claimed that OC management using coancestry matrices based on allele sharing moves frequencies to intermediate values and reduces the probability of losing alleles. In fact, these authors observed that strategies that maximize heterozygosity, by managing contributions from parents, keep levels of allelic diversity as high as strategies that maximize allelic diversity itself. Their results were obtained when applying OC using the similarity genomic matrix of Nejati-Javaremi et al. [22], calculated with up to 40 multiallelic markers, but the same could be expected when using θ_{LH} given that correlation between both matrices is 1. However, we have obtained solutions which maintain genetic diversity (H_e) but result in a higher number of fixed loci and this could be due to the different numbers of markers used in both studies.

To understand these contrasting results, we carried out extra simulations to compare observed with expected values for the number of fixed loci under both management strategies (i.e., S_{O_LH} and S_{O_VR}). In this extra scenario, a population with $N = 20$ individuals was managed during four generations, with different numbers of SNPs used for the calculation of the coancestry matrices (20 and 1000). A single chromosome was simulated. The expected number of fixed SNPs (ES_f) was estimated using the solutions that came out of each optimization before generating the offspring, following Fernández et al. [13]. Thus, ES_f was computed as $\sum_{k=1}^{2}\prod_{i=1}^{N} prob_{ki}$, where $prob_{ki}$ is the probability of individual i not transmitting allele k. If parent i carries a unique type of allele (that is, homozygous for the h allele) and leaves descendants, $prob_{ki}$ is 0 if $k = h$ and 1 if $k \neq h$. If it carries two different alleles (that is, heterozygous), the probability is $prob_{ki} = (0.5)^{c_i}$, where c_i is the number of offspring to be contributed by parent i. ES_f value can be averaged then across loci. Table 5 shows that expected and observed numbers of SNPs becoming fixed each generation were close. When using only 20 SNPs, even though only seven–eight individuals are selected with S_{O_LH}, the expected (observed) number of SNPs that become fixed is lower than with S_{O_VR}. However, when the number of SNPs used was increased, the trend reversed and the expected (and observed) number of fixed SNPs becomes lower for S_{O_VR} than for S_{O_LH}, even when the number of selected individuals increases for S_{O_LH}. The explanation for this performance could be that, with many markers, S_{O_LH} is able to find a solution with higher mean H_e by keeping loci with high MAF and allowing SNPs with rare alleles to become fixed.

Table 5. Number of selected candidates (N_S) and expected (ES_f) and observed number of fixed SNPs (S_f) across generations (t) when contributions are optimized using Li and Horvitz's (S_{O_LH}) and VanRaden's (S_{O_VR}) coancestry matrices computed with two different number of SNPs (S), for a population of 20 individuals.

		S_{O_LH}			S_{O_VR}		
t	***S***	N_S	ES_f	S_f	N_S	ES_f	S_f
1	20	7	0.3	0	20	0.3	0
2		7	0.7	0	13	0.8	1
3		8	0.8	0	13	1.4	1
4		8	0.9	0	12	1.7	1
1	1000	15	21.7	21	20	17.6	18
2		16	38.9	37	19	34.6	33
3		15	54.6	52	19	50.9	47
4		15	68.6	64	18	66.3	60

The results show that the differences in maintained diversity (H_e) and divergence from the original frequencies (KL) between strategies S_{O_LH} and S_{O_VR} decreased when using only SNPs with a minimum MAF (MAF > 0.05 or MAF > 0.25) for computing the coancestry matrices. As mentioned above, S_{O_LH} promotes the contribution of individuals carrying rare alleles, as their coancestries with the rest of the population are smaller, and thus increases the frequencies of rare alleles. When the minimum MAF permitted increases, the number of rare alleles decreases, and the differences between the average coancestries between pairs of individuals decrease. In such situation, S_{O_LH} does not prioritize too much the contributions from any individual and leads to solutions that imply a higher number of candidates selected. Consequently, the results are closer to those obtained with strategy S_{O_VR}. Moreover, when using only SNPs with high MAF in $t = 0$ (i.e., initial frequencies are close to 0.5), the performance of S_{O_VR} (i.e., keeping those initial frequencies) is similar to the performance of S_{O_LH} (moving them to intermediate values). These observations are in agreement with results from Morales-González et al. [16] and Villanueva et al. [29], who found that the correlation between VanRaden's and Li and Horvitz's coefficients increases with increasing the MAF of the SNPs used.

Here, we have optimized contributions of parents for minimizing the loss of variability and then changes in frequencies have been evaluated. On the other hand, Saura et al. [9] optimized contributions of parents for minimizing changes in allele frequencies and then the loss of genetic variability was evaluated. An alternative to both approaches could be to consider simultaneously the control of variability and the allele frequency changes. Similar to the OC algorithm designed for maximizing genetic gain while restricting the rate of inbreeding [2,3,24] or for maximizing the phenotypic level for a trait of interest while restricting the loss in variability when creating base populations [30], one could develop an algorithm for minimizing the loss of variability while restricting the change in frequencies or, alternatively, for minimizing frequency changes while restricting the loss of variability. The specific objective would depend on the particular interest of the managers of the program. This kind of approach was followed by Fernández et al. [31] in the context of optimizing the sampling strategy for establishing a gene bank. In particular, they developed an algorithm that simultaneously allows targeting frequencies for alleles at a particular locus while controlling the genetic diversity of other unlinked loci.

It could be also possible to combine both coancestry matrices (θ_{LH} and θ_{VR}) in the objective function when the specific objective differs across genomic regions (i.e., in some regions the interest may be to maintain diversity, and in other regions the interest may be to maintain frequencies). Maintaining diversity may be of interest for regions associated with inbreeding depression for fitness-related traits and also for regions that harbor loci involved in general resistance to diseases (e.g., the major histocompatibility complex, MHC) as a high level of genetic diversity is desirable to ensure that the population can deal with potential new disease challenges [21]. Maintaining frequencies may be of interest in regions containing loci that have been under natural or artificial selection, and one wants to keep the genetic progress obtained. Gómez-Romano et al. [21] showed that the OC method using a matrix equivalent to θ_{LH} is efficient in maintaining H_e in specific regions and simultaneously restricts the loss of H_e in the rest of the genome. Their approach could be extended to include the use of θ_{VR} for minimizing the change in allele frequencies in some genomic regions. However, it has to be kept in mind that the higher the number of different parameters to be controlled, or the more regions to be treated differently, the lower the control of each objective one can expect.

In a conservation program, the maintenance of genetic variability throughout the genome is the general aim because usually there is no information available on the relevance of each genome region and the current or future use of the genetic variability present in particular regions. Therefore, it is better to conserve as much diversity as possible because if alleles are lost in a population, they will be no longer available. However, this strategy can lead to the maintenance or even an increase in the frequency of deleterious alleles. Different methods have been proposed to avoid this when using the OC method, including (i) selection of the best sib from the group of offspring generated by the selected parents [28] and (ii) combining selection with inbred matings [14] to allow for some kind of purging. Sonesson et al. [32] also proposed a model in which they tried to eliminate a disease from a population in different scenarios by explicitly performing selection against this condition. Currently, genomics can provide information on deleterious variability and the loci determining the occurrence of the disease [33], so a strategy where selection is made against these deleterious alleles [17], while you restrict the loss of variability in the rest of the genome, could be possible.

The amount of genetic variability retained was measured as the expected heterozygosity (H_e). However, other measures such as allelic diversity can be used [13,34]. Allelic diversity is essential from an evolutionary perspective, since the limit of selection response is determined by the initial number of alleles [35,36]. It is worth noting that strategy $\mathrm{S_{O\_VR}}$ would be more efficient than strategy $\mathrm{S_{O\_LH}}$, not only to maintain allele frequency but also to maintain diversity when this is measured as the number of unobserved loci segregating. It is thus clear that the coancestry matrix to be used in OC when managing a particular genetic conservation program would be case specific.

Finally, it is worth mentioning that further work is needed to explore how the relaxation of some of the assumptions implicit in our simulations could affect the results obtained. Extra work would be necessary to investigate schemes with overlapping generations, variable population size over the management time frame, and different degrees of relatedness between the founders.

5. Conclusions

When applying strategy S_{O_LH}, more H_e is maintained than when applying strategy S_{O_VR} given that S_{O_LH} moves allele frequencies towards 0.5. However, S_{O_VR} maintained allele frequencies closer to those of the initial generation and more loci segregating than S_{O_LH}. Therefore, considering that conservation programs generally aim to increase genetic diversity, but it is also important to maintain population uniqueness, the choice of which genomic coancestry matrix is used in management may depend on which of these two goals is more important for each particular case. When a subset of SNPs with MAF > 0.05 or MAF > 0.25 is used to estimate coancestry matrices, the differences between both strategies in terms of both H_e and KL were reduced. The differences between strategies were smaller for populations of smaller sizes given that in a smaller population it is more difficult to differentiate between individuals.

Author Contributions: Conceptualization, J.F., R.P.-W. and B.V.; methodology, E.M.-G., J.F. and B.V.; software, E.M.-G. and J.F.; formal analysis, E.M.-G.; writing—original draft preparation, E.M.-G., J.F. and B.V.; writing—review and editing, M.Á.T. and R.P.-W.; supervision, B.V.; project administration, J.F. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results has received funding from the Ministerio de Ciencia, Innovación y Universidades, Spain (grant CGL2016-75904-C2-2-P). R. Pong-Wong is funded by the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n°772787 (SMARTER) and the Biotechnology and Biological Sciences Research Council through Institute Strategic Programme Grant funding (BBS/E/D/30002275).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The codes to perform the simulations are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare that there are no competing interests with respect to the authorship or publication of this article.

References

1. Frankham, R.; Ballou, J.D.; Briscoe, D.A.; Ballou, J.D. *Introduction to Conservation Genetics*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2002.
2. Meuwissen, T. Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* **1997**, *75*, 934–940. [CrossRef]
3. Grundy, B.; Villanueva, B.; Woolliams, J.A. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. *Genet. Res.* **1998**, *72*, 159–168. [CrossRef]
4. Fernández, J.; A Toro, M.; Caballero, A. Fixed contributions designs vs. minimization of global coancestry to control inbreeding in small populations. *Genetics* **2003**, *165*, 885–894. [CrossRef]
5. Caballero, A.; Toro, M.A. Interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genet. Res.* **2000**, *75*, 331–343. [CrossRef] [PubMed]
6. Falconer, D.S.; Mackay, T.F.C. Introduction to quantitative genetics. In *Introduction to Quantitative Genetics*, 4th ed.; Longman: Harlow, UK, 1996.
7. Lacy, R.C. Should we select genetic alleles in our conservation breeding programs? *Zoo Biol.* **2000**, *19*, 279–282. [CrossRef]
8. Frankham, R. Genetic adaptation to captivity in species conservation programs. *Mol. Ecol.* **2008**, *17*, 325–333. [CrossRef] [PubMed]
9. Saura, M.; Pérez-Figueroa, A.; Fernández, J.; Toro, M.A.; Caballero, A. Preserving population allele frequencies in ex situ conservation programs. *Conserv. Biol.* **2008**, *22*, 1277–1287. [CrossRef] [PubMed]
10. De Cara, M.A.R.; Fernández, J.; Toro, M.A.; Villanueva, B. Using genome-wide information to minimize the loss of diversity in conservation programs. J. *Anim. Breed. Genet.* **2011**, *128*, 456–464. [CrossRef]

11. De Cara, M.; Ángeles, R.; Villanueva, B.; Toro, M.Á.; Fernández, J. Using genomic tools to maintain diversity and fitness in conservation programmes. *Mol. Ecol.* **2013**, *22*, 6091–6099. [CrossRef]
12. Gómez-Romano, F.; Villanueva, B.; De Cara, M.Á.R.; Fernández, J. Maintaining genetic diversity using molecular coancestry: The effect of marker density and effective population size. *Genet. Sel. Evol.* **2013**, *45*, 38. [CrossRef]
13. Fernández, J.; Toro, M.A.; Caballero, A. Managing Individuals' Contributions to Maximize the Allelic Diversity Maintained in Small, Conserved Populations. *Conserv. Biol.* **2004**, *18*, 1358–1367. [CrossRef]
14. De Cara, M.A.R.; Villanueva, B.; Toro, M.A.; Fernández, J. Purging deleterious mutations in conservation programs: Combining optimal contributions with inbred mattings. *Heredity* **2013**, *110*, 530–537. [CrossRef] [PubMed]
15. Eynard, S.E.; Windig, J.J.; Hiemstra, S.J.; Calus, M.P.L. Whole-genome sequence data uncover loss of genetic diversity due to selection. *Genet. Sel. Evol.* **2016**, *48*, 33. [CrossRef] [PubMed]
16. Morales-González, E.; Saura, M.; Fernández, A.; Fernández, J.; Pong-Wong, R.; Cabaleiro, S.; Martínez, P.; Martín-García, A.; Villanueva, B. Evaluating different genomic coancestry matrices for managing genetic variability in turbot. *Aquaculture* **2020**, *520*, 734985. [CrossRef]
17. Meuwissen, T.H.E.; Sonesson, A.K.; Gebregiwergis, G.; Woolliams, J.A. Management of Genetic Diversity in the Era of Genomics. *Front. Genet.* **2020**, *11*, 880. [CrossRef]
18. Li, C.C.; Horvitz, D.G. Some methods of estimating the inbreeding coefficient. *Am. J. Hum. Genet.* **1953**, *5*, 107–117.
19. VanRaden, P.M. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [CrossRef]
20. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.D.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010**, *42*, 565–569. [CrossRef]
21. Gómez-Romano, F.; Villanueva, B.; Fernández, J.; Woolliams, J.A.; Pong-Wong, R. The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. *Genet. Sel. Evol.* **2016**, *48*, 1–17. [CrossRef]
22. Nejati-Javaremi, A.; Smith, C.; Gibson, J.P. Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* **1997**, *75*, 1738–1745. [CrossRef]
23. Toro, M.A.; Villanueva, B.; Fernández, J. The concept of effective population size loses its meaning in the context of optimal management of diversity using molecular markers. *J. Anim. Breed. Genet.* **2019**, *137*, 345–355. [CrossRef]
24. Woolliams, J.A.; Berg, P.; Dagnachew, B.S.; Meuwissen, T.H.E. Genetic contributions and their optimisation. *J. Anim. Breed. Genet.* **2015**, *132*, 89–99. [CrossRef]
25. Toro, M.; Barragán, C.; Óvilo, C.; Rodrigañez, J.; Rodriguez, C.; Silió, L. Estimation of coancestry in Iberian pigs using molecular markers. *Conserv. Genet.* **2002**, *3*, 309–320. [CrossRef]
26. Forni, S.; Aguilar, I.; Misztal, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* **2011**, *43*, 1. [CrossRef] [PubMed]
27. Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, New York, NY, USA, 1997.
28. Fernández, J.; Caballero, A. Accumulation of deleterious mutations and equalization of parental contributions in the conservation of genetic resources. *Heredity* **2001**, *86*, 480–488. [CrossRef] [PubMed]
29. Villanueva, B.; Fernández, A.; Saura, M.; Caballero, A.; Fernández, J. The value of genomic relationship matrices for estimating inbreeding. *Genet. Sel. Evol.* **2021**. under review.
30. Fernã¡ndez, J.; Toro, M.Ã.; Sonesson, A.K.; Villanueva, B.; Fernández, J. Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress. *Front. Genet.* **2014**, *5*, 414. [CrossRef] [PubMed]
31. Fernández, J.; Roughsedge, T.; Woolliams, J.A.; Villanueva, B. Optimization of the sampling strategy for establishing a gene bank: Storing PrP alleles following a scrapie eradication plan as a case study. *Anim. Sci.* **2006**, *82*, 813–821. [CrossRef]
32. Sonesson, A.K.; Janss, L.L.; Meuwissen, T.H. Selection against genetic defects in conservation schemes while controlling inbreeding. *Genet. Sel. Evol.* **2003**, *35*, 1–16. [CrossRef] [PubMed]
33. Charlier, C.; Coppieters, W.; Rollin, F.; Desmecht, D.; Agerholm, J.S.; Cambisano, N.; Carta, E.; Dardano, S.; Dive, M.; Fasquelle, C.; et al. Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat. Genet.* **2008**, *40*, 449–454. [CrossRef]
34. Caballero, A.; Rodríguez-Ramilo, S.T. A new method for the partition of allelic diversity within and between subpopulations. *Conserv. Genet.* **2010**, *11*, 2219–2229. [CrossRef]
35. James, J.W. The founder effect and response to artificial selection. *Genet. Res.* **1970**, *16*, 241–250. [CrossRef] [PubMed]
36. Hill, W.G.; Rasbash, J. Models of long term artificial selection in finite population. *Genet. Res.* **1986**, *48*, 41–50. [CrossRef] [PubMed]