Supplementary material

1 Data preprocessing details

The founder dataset, consisting of 192 individuals and 2591 SNPs, was cleaned and preprocessed:

1. Remove markers (223) with unknown position in the genetic map.
2. Remove markers (46) and individuals (0) with more than 20% missing values.
3. Recode markers as number of copies of minor allele (0,1,2).
4. Impute missing values using Beagle 3.3.2 (Browning and Browning, 2007, 2009) through the R package synbreed (version 0.10-5) (Wimmer *et al.*, 2012). Beagle uses a hidden Markov model (HMM) to reconstruct missing values based on flanking markers.
5. Filter redundant markers (291). Markers were considered redundant if they had the same map position and the same allele was observed in all individuals. From each set of redundant markers, only one was retained.
6. Spread remaining markers mapped to same position at 0.1 cM intervals in arbitrary order.

This procedure retained 2031 polymorphic SNPs and all 192 individuals.

2 Genomic optimal contributions selection

2.1 Genomic inbreeding control

Here we formally derive how the inbreeding rate ΔF relates to SNP allele frequencies in the population and the changes of these frequencies over time, and to the GOCS constraint C_{t+1}. First we express C_{t+1} in terms of SNP allele frequencies starting from its definition

$$C_{t+1} = \frac{\mathbf{c_t^\intercal G_t c_t}}{2}.$$

Here, $\mathbf{c_t}$ is a vector of assigned contributions (under optimization) and $\mathbf{G_t}$ is the realized genomic relationship matrix of the selection candidates in generation t, defined as

$$\mathbf{G_t} = \frac{\mathbf{Z_t Z_t^\intercal}}{2\sum_{j=1}^{m} p_j(1-p_j)}$$

where m is the number of markers, p_j is the reference allele frequency of the jth marker in the population of selection candidates and $\mathbf{Z_t}$ is the centered marker matrix of the selection candidates:

$$\mathbf{Z_t} = \mathbf{X_t} - 2\mathbf{P_t}$$

where $\mathbf{X_t}$ is the original marker matrix containing reference allele counts (0/1/2) and $\mathbf{P_t}$ is a matrix whose jth column contains the current allele frequency p_j of the jth marker:

$$\mathbf{P_t} = \begin{bmatrix} p_1 & p_2 & \cdots & p_m \\ p_1 & p_2 & \cdots & p_m \\ \vdots & \vdots & \ddots & \vdots \\ p_1 & p_2 & \cdots & p_m \end{bmatrix}.$$

As such, the values of $\mathbf{Z_t}$ represent reference allele counts relative to the population mean and each of its columns sums to zero. It follows that (Woolliams *et al.*, 2015)

$$\mathbf{Z_t^\intercal c_t} = 2 \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_m \end{bmatrix}$$

where Δ_j is the expected allele frequency change when mating the individuals from the selection population according to the assigned contributions $\mathbf{c_t}$. Therefore

$$\mathbf{c_t^\intercal Z_t Z_t^\intercal c_t} = 4 \sum_{j=1}^{m} \Delta_j^2$$

and thus

$$C_{t+1} = \frac{\mathbf{c_t^\intercal G_t c_t}}{2} = \frac{\mathbf{c_t^\intercal Z_t Z_t^\intercal c_t}}{4 \sum_{j=1}^{m} p_j(1-p_j)} = \frac{1}{2m} \frac{\mathbf{c_t^\intercal Z_t Z_t^\intercal c_t}}{H_t} = \frac{2}{mH_t} \sum_{j=1}^{m} \Delta_j^2$$

where $H_t = \frac{1}{m} \sum_{j=1}^{m} 2p_j(1-p_j)$ is the expected heterozygosity in the selection population.

Now we also express the inbreeding rate ΔF in terms of the SNP allele frequencies and their changes, starting from its definition:

$$\Delta F = \frac{F_{t+1} - F_t}{1 - F_t}$$

with (for the SNP marker panel used)

$$F_t = \frac{1}{m} \sum_{j=1}^{m} p_j^2 + (1-p_j)^2 \quad \text{and} \quad 1 - F_t = H_t.$$

It follows that

$$\begin{aligned}
\Delta F_{IBS} &= \frac{1}{mH_t} \left[\sum_{j=1}^{m} (p_j + \Delta_j)^2 + (1 - p_j - \Delta_j)^2 - \sum_{j=1}^{m} p_j^2 + (1-p_j)^2 \right] \\
&= \frac{1}{mH_t} \sum_{j=1}^{m} p_j^2 + 2p_j\Delta_j + \Delta_j^2 + (1-p_j)^2 - 2(1-p_j)\Delta_j + \Delta_j^2 - p_j^2 - (1-p_j)^2 \\
&= \frac{1}{mH_t} \sum_{j=1}^{m} 2\Delta_j^2 + 2p_j\Delta_j - 2(1-p_j)\Delta_j \\
&= \frac{1}{mH_t} \sum_{j=1}^{m} 2\Delta_j^2 + 4p_j\Delta_j - 2\Delta_j \\
&= \frac{2}{mH_t} \sum_{j=1}^{m} \Delta_j^2 + \frac{2}{mH_t} \sum_{j=1}^{m} \Delta_j(2p_j - 1) \\
&= \frac{\mathbf{c_t^\intercal G_t c_t}}{2} + \frac{2}{mH_t} \sum_{j=1}^{m} \Delta_j(2p_j - 1) \\
&= C_{t+1} + \frac{2}{mH_t} \sum_{j=1}^{m} \Delta_j(2p_j - 1)
\end{aligned}$$

As such, the inbreeding rate ΔF_{IBS} defined for SNP markers is the sum of the GOCS constraint $C_{t+1} = \mathbf{c_t^\intercal G_t c_t}/2$ and an additional term $\frac{2}{mH_t} \sum_{j=1}^{m} \Delta_j(2p_j - 1)$ that is not constrained by GOCS.

2.2 Selection procedure

The optimal contributions selection (OCS) strategy was originally presented in an animal breeding context by Meuwissen (1997) where it is required that the contributions of males and females both sum to 1/2. In our plant breeding scheme such constraint does not apply and all contributions must simply sum to one, i.e. $\sum \mathbf{c_t} = 1$. This slightly changes the optimization formulas of GOCS based on Lagrangian multipliers to

$$\mathbf{c_t} = \frac{\mathbf{G_t^{-1}}(\mathbf{GEBV_t} - \lambda)}{2\lambda_0}$$

$$\lambda = \frac{\mathbf{1^\intercal G_t^{-1} GEBV_t} - 2\lambda_0}{\mathbf{1^\intercal G_t^{-1} 1}}$$

$$\lambda_0^2 = \frac{\mathbf{GEBV_t^\intercal}(\mathbf{G_t^{-1}} - \frac{\mathbf{G_t^{-1} 1 1^\intercal G_t^{-1}}}{\mathbf{1^\intercal G_t^{-1} 1}})\mathbf{GEBV_t}}{8C_{t+1} - \frac{4}{\mathbf{1^\intercal G_t^{-1} 1}}}.$$

Any negative contributions are eliminated by setting the most negative value to zero and iteratively re-optimizing the remaining contributions. Following Meuwissen (2002) a minimum and/or maximum contribution, c_{min} and c_{max}, respectively, may be imposed. As a special case we set $c_{min} = c_{max} = 1/n$ to select n individuals with equal contribution. To deal with these additional constraints, contributions exceeding the maximum value are truncated and those individuals with a too low contribution are discarded. In each step of the algorithm, the following rules are applied to adjust the contributions, after which the remaining ones are re-optimized:

1. Discard the individual with the most negative contribution, if any, by fixing its contribution to zero.

2. Else, if any contribution exceeds the imposed maximum c_{max}, truncate the largest contribution to c_{max} and exclude the corresponding individual from the optimization. In addition, all individuals that were previously discarded, if any, are re-included in the optimization.

3. Else, if any selected individual has a contribution below the imposed minimum c_{min}, discard the individual with the smallest positive contribution.

Meuwissen (1997) explained in an appendix how to extend the formulas to optimize the remaining contributions $\mathbf{c_o}$ when some have already been fixed to $\mathbf{c_f}$, in our case to either zero or $c_{max} = 1/n$. Adjusting the formulas for our plant breeding scheme yields

$$\mathbf{c_o} = \frac{\mathbf{G_{oo}^{-1}}(\mathbf{GEBV_o} - 2\lambda_0 \mathbf{G_{of}} \mathbf{c_f} - \lambda)}{2\lambda_0}$$

$$\lambda = \frac{\mathbf{1^\intercal G_{oo}^{-1}}(\mathbf{GEBV_o} - 2\lambda_0 \mathbf{G_{of}} \mathbf{c_f}) - 2\lambda_0 s}{\mathbf{1^\intercal G_{oo}^{-1} 1}}$$

$$\lambda_0^2 = \frac{1}{4} \frac{\mathbf{GEBV_o^\intercal\ P\ GEBV_o}}{K + L - M - N}$$

where

$$s = 1 - \sum c_f$$

$$\mathbf{P} = \mathbf{G_{oo}^{-1}} - \frac{\mathbf{G_{oo}^{-1}} \mathbf{1} \mathbf{1}^\intercal \mathbf{G_{oo}^{-1}}}{\mathbf{1}^\intercal \mathbf{G_{oo}^{-1}} \mathbf{1}}$$

$$K = 2C_{t+1} - \mathbf{c_f}^\intercal \mathbf{G_{ff}} \mathbf{c_f}$$

$$L = \mathbf{c_f}^\intercal \mathbf{G_{fo}} \mathbf{P} \mathbf{G_{of}} \mathbf{c_f}$$

$$M = \frac{s^2}{\mathbf{1}^\intercal \mathbf{G_{oo}^{-1}} \mathbf{1}}$$

$$N = \frac{2s \mathbf{1}^\intercal \mathbf{G_{oo}^{-1}} \mathbf{G_{of}} \mathbf{c_f}}{\mathbf{1}^\intercal \mathbf{G_{oo}^{-1}} \mathbf{1}}$$

Here, $\mathbf{GEBV_o}$ is a vector of genomic estimated breeding values of the individuals under optimization, and $\mathbf{G_{xy}}$ is the genomic relationship matrix restricted to rows $\mathbf{x}$ and columns $\mathbf{y}$. Applying these formulas optimizes the remaining contributions $\mathbf{c_o}$ to maximize the expected genetic gain $\mathbf{c_o}^\intercal \mathbf{GEBV_o}$ while constraining

$$C_{t+1} = \frac{\mathbf{c_t}^\intercal \mathbf{G_t} \mathbf{c_t}}{2} = \frac{1}{2}[\mathbf{c_o}^\intercal \mathbf{G_{oo}} \mathbf{c_o} + 2\mathbf{c_o}^\intercal \mathbf{G_{of}} \mathbf{c_f} + \mathbf{c_f}^\intercal \mathbf{G_{ff}} \mathbf{c_f}]$$

to the target inbreeding rate $C_{t+1} = \Delta F_{target}$, with $\sum \mathbf{c_t} = \sum (\mathbf{c_o} + \mathbf{c_f}) = 1$. It may happen that, in a certain step of the iterative heuristic, this constraint cannot be satisfied for the remaining individuals. In such case, we assign the remaining contributions by minimizing the corresponding realized genomic relationship, in order to approach the requested constraint value as closely as possible. Formulas for the latter optimization problem are also obtained with Lagrangian multipliers by minimizing $\mathbf{c_t}^\intercal \mathbf{G_t} \mathbf{c_t}/2$ with $\sum \mathbf{c_t} = 1$:

$$\mathbf{c_o} = \mathbf{G_{oo}^{-1}} (\frac{1}{2}\lambda - \mathbf{G_{of}} \mathbf{c_f})$$

$$\lambda = \frac{2(1 - \sum \mathbf{c_f} + \mathbf{1}^\intercal \mathbf{G_{oo}^{-1}} \mathbf{G_{of}} \mathbf{c_f})}{\mathbf{1}^\intercal \mathbf{G_{oo}^{-1}} \mathbf{1}}$$

3 Objective function normalization

The applied set selection strategy maximizes a weighted index

$$F(S) = (1 - \alpha) \cdot V(S) + \alpha \cdot D(S)$$

that balances average breeding value $V(S)$ with a diversity component $D(S)$. To obtain a fair balance for weights $\alpha \in [0, 1]$, both components are normalized, leading to maximization of

$$F^*(S) = (1 - \alpha) \cdot V^*(S) + \alpha \cdot D^*(S).$$

We follow the Pareto minimum-based upper-lower-bound approach described by Marler and Arora (2005) where

$$V^*(S) = \frac{V(S) - V(S_D^*)}{V(S_V^*) - V(S_D^*)}$$

and

$$D^*(S) = \frac{D(S) - D(S_V^*)}{D(S_D^*) - D(S_V^*)}.$$

Table S1: Used R packages and versions.

Package	Version	Reference(s)
BGLR	1.0.4	de los Campos and Pérez (2015)
coda	0.17-1	Plummer *et al.* (2006)
rrBLUP	4.3	Endelman (2011)
synbreed	0.10-5	Wimmer *et al.* (2012)
hypred	0.5	Technow (2014)
gdata	2.17.0	Warnes *et al.* (2015)
Hmisc	3.16-0	Harrell *et al.* (2015)
rJava	0.9-7	Urbanek (2015)
setRNG	2013.9-1	Gilbert (2014)

Here, S_V^* and S_D^* are the selections with the highest achievable breeding value and diversity, respectively. The former, S_V^* is easily constructed by selecting the n candidates with the highest individual breeding value, just as in standard GS. On the other hand, S_D^* is approximated through a preliminary optimization procedure, using the same parallel tempering algorithm that is eventually applied to maximize the normalized weighted index $F^*(S)$—allowing a maximum of three seconds between subsequent improvements. The applied normalization procedure is thus fully automated and dynamically adapts to the provided data.

4 Supplementary figures and tables

The following pages contain supplementary Table S1 and Figures S1–S5.

References

Browning, B. L. and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics **84**: 210–223.

Browning, S. R. and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics **81**: 1084–1097.

de los Campos, G. and P. Pérez, 2015 *BGLR: Bayesian Generalized Linear Regression*. R package version 1.0.4.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome **4**: 250–255.

Gilbert, P., 2014 *setRNG: Set (Normal) Random Number Generator and Seed*. R package version 2013.9-1.

Harrell, F. E. J., with contributions from Charles Dupont, and many others., 2015 *Hmisc: Harrell Miscellaneous*. R package version 3.16-0.

Marler, R. T. and J. S. Arora, 2005 Function-transformation methods for multi-objective optimization. Engineering Optimization **37**: 551–570.

Meuwissen, T., 1997 Maximizing the response of selection with a predefined rate of inbreeding. Journal of animal science **75**: 934–940.

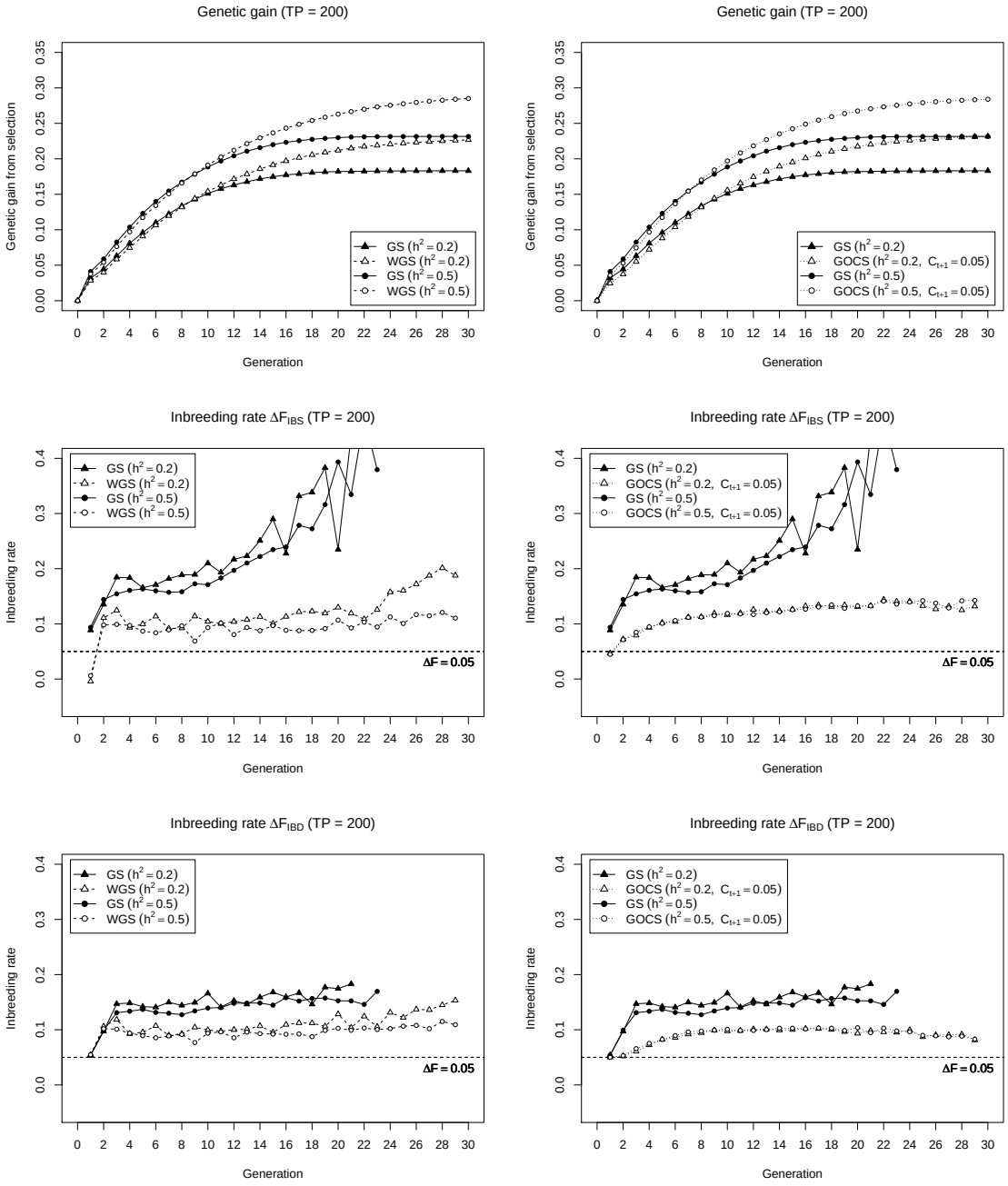

Figure S1: Weighted genomic selection and genomic optimal contributions selection. Cumulative genetic gain (top) and inbreeding rate (IBS: middle; IBD: bottom) for weighted genomic selection (WGS; left) and genomic optimal contributions selection (GOCS; right) as compared to standard genomic selection (GS). Results are reported for a low ($h^2 = 0.2$) and high ($h^2 = 0.5$) heritability with a small initial training population (TP = 200) and are averages of 200 simulation runs. The inbreeding rates are reported until at least half of the simulation runs have lost all variability for the SNP marker panel used.

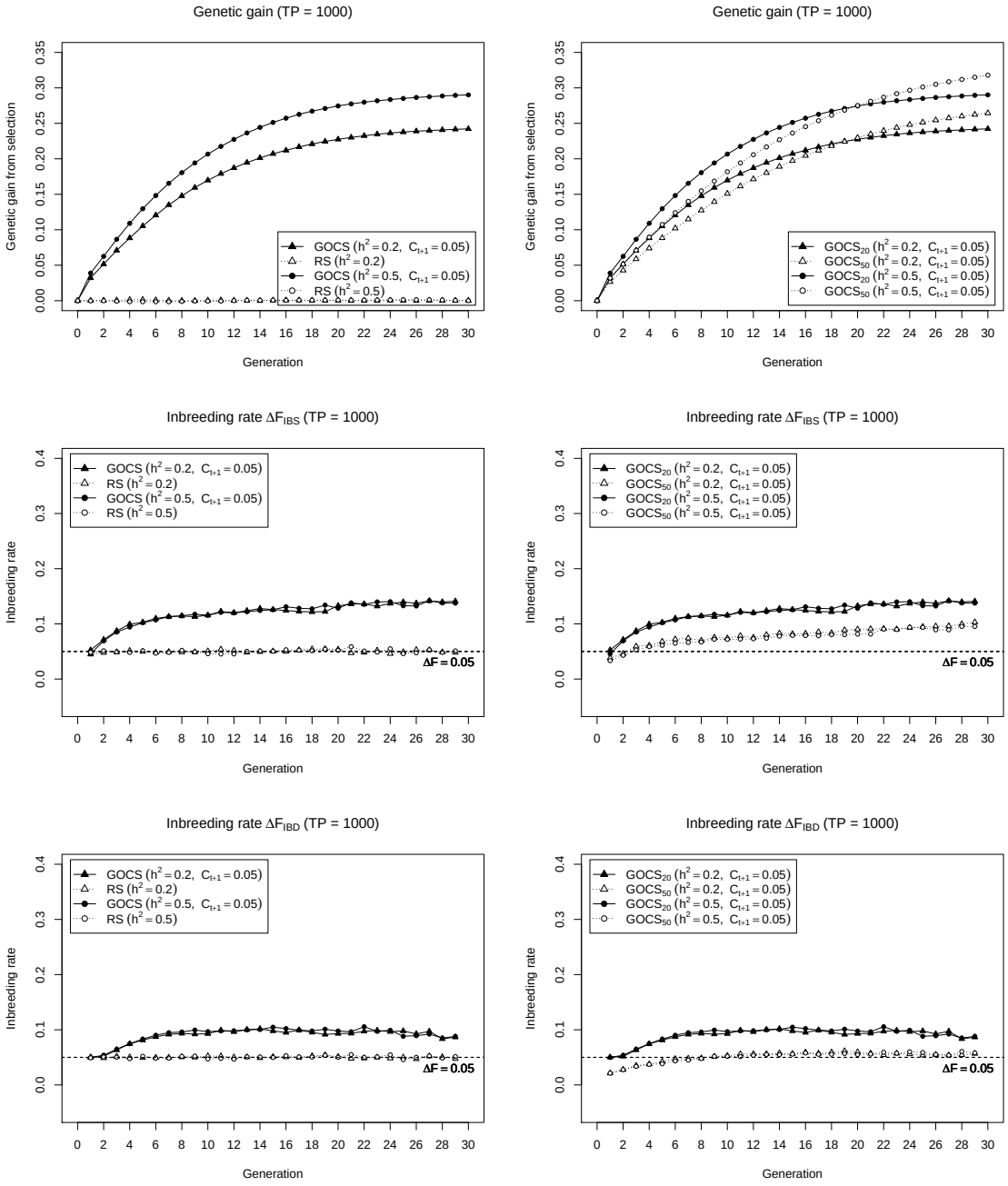

Figure S2: Influence of selection on inbreeding rate realized by GOCS. Cumulative genetic gain (top) and inbreeding rate (IBS: middle; IBD: bottom) in case no genomic selection is performed (RS; left), i.e. where 20 individuals are chosen randomly in each cycle, and for genomic optimal contributions selection with a larger selection consisting of 50 individuals ($GOCS_{50}$; right), as compared to GOCS with the default selection size (20). Results are reported for a low ($h^2 = 0.2$) and high ($h^2 = 0.5$) heritability with a large initial training population (TP = 1000) and are averages of 200 simulation runs.

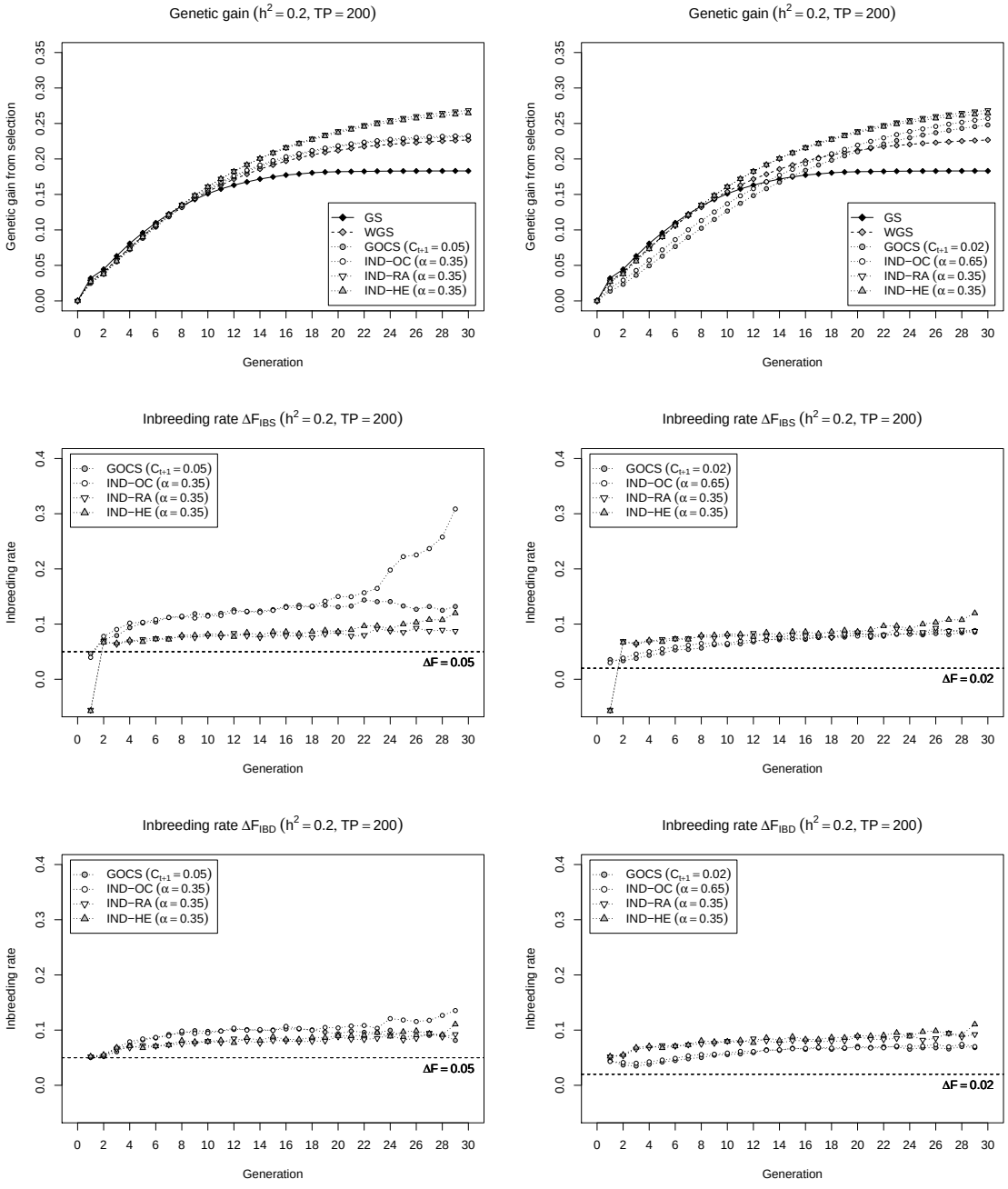

Figure S3: Current and new selection strategies in a unified optimization framework. Cumulative genetic gain (top) and inbreeding rate (IBS: middle; IBD: bottom) of selection strategies that maximize a weighted index containing breeding value and a diversity measure chosen to control inbreeding (IND-OC, IND-HE) or to avoid loss of rare alleles (IND-RA). Results for GS, WGS, and GOCS are provided as a reference. For clarity, inbreeding rates of GS and WGS are omitted. Two scenarios were considered to set the parameters C_{t+1} and α: maintain the same short-term gain as WGS (left), or achieve a similar inbreeding rate ΔF_{IBS} (right). Results are reported for a low heritability ($h^2 = 0.2$) with a small initial training population (TP = 200) and are averages of 200 simulation runs.

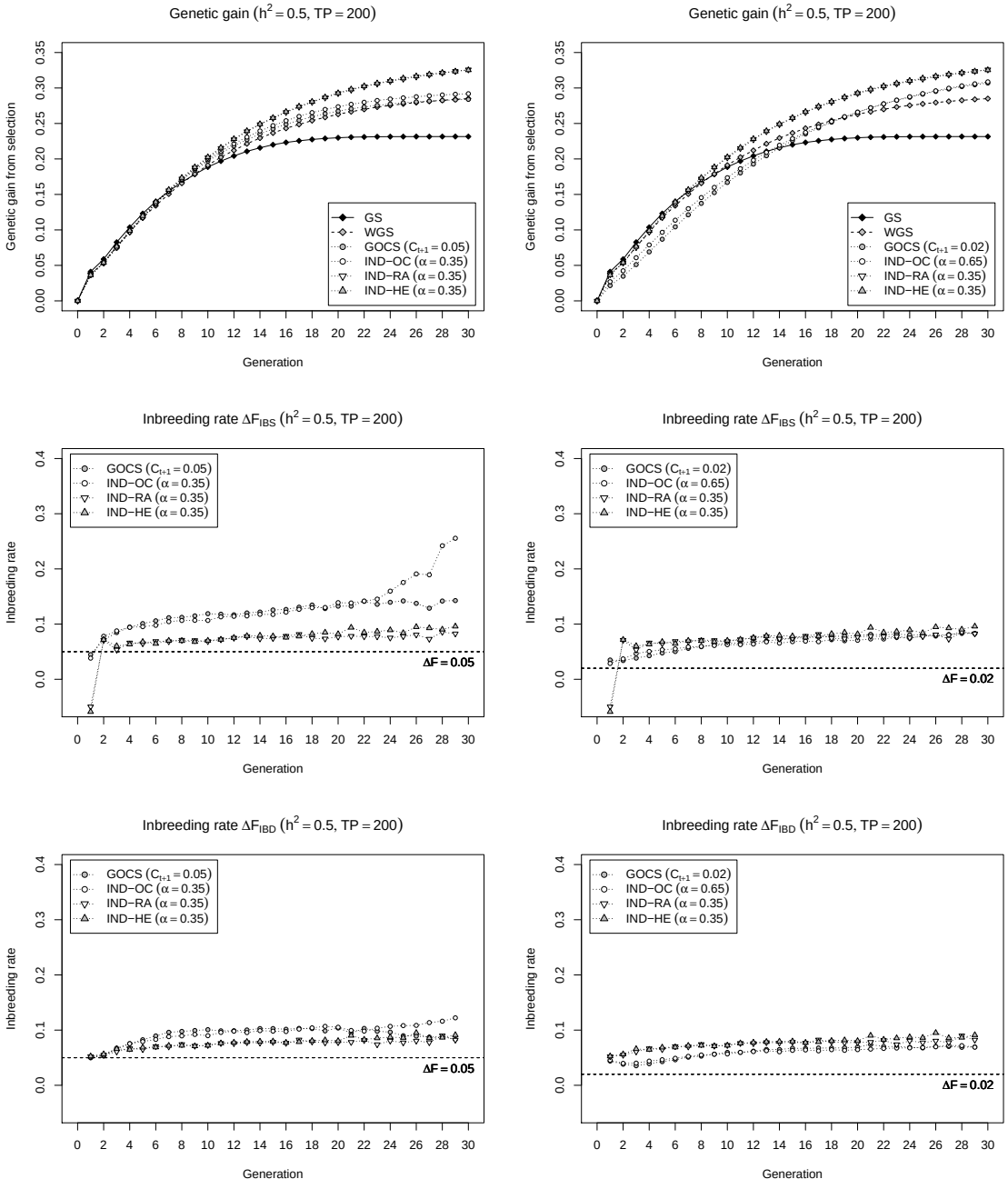

Figure S4: Current and new selection strategies in a unified optimization framework. Cumulative genetic gain (top) and inbreeding rate (IBS: middle; IBD: bottom) of selection strategies that maximize a weighted index containing breeding value and a diversity measure chosen to control inbreeding (IND-OC, IND-HE) or to avoid loss of rare alleles (IND-RA). Results for GS, WGS, and GOCS are provided as a reference. For clarity, inbreeding rates of GS and WGS are omitted. Two scenarios were considered to set the parameters C_{t+1} and α: maintain the same short-term gain as WGS (left), or achieve a similar inbreeding rate ΔF_{IBS} (right). Results are reported for a high heritability ($h^2 = 0.5$) with a small initial training population (TP = 200) and are averages of 200 simulation runs.

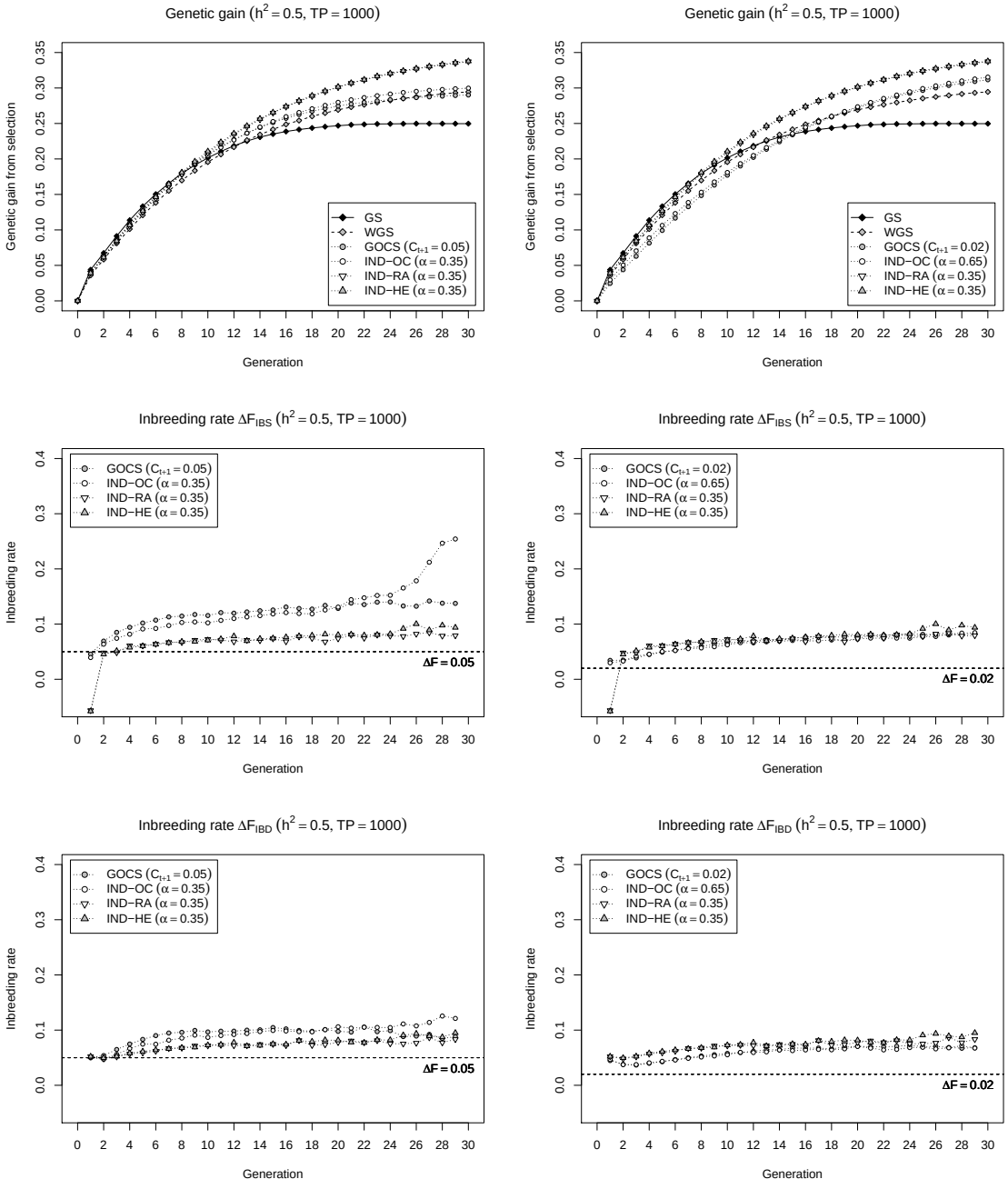

Figure S5: Current and new selection strategies in a unified optimization framework. Cumulative genetic gain (top) and inbreeding rate (IBS: middle; IBD: bottom) of selection strategies that maximize a weighted index containing breeding value and a diversity measure chosen to control inbreeding (IND-OC, IND-HE) or to avoid loss of rare alleles (IND-RA). Results for GS, WGS, and GOCS are provided as a reference. For clarity, inbreeding rates of GS and WGS are omitted. Two scenarios were considered to set the parameters C_{t+1} and α: maintain the same short-term gain as WGS (left), or achieve a similar inbreeding rate ΔF_{IBS} (right). Results are reported for a high heritability ($h^2 = 0.5$) with a large initial training population (TP = 1000) and are averages of 200 simulation runs.

Meuwissen, T., 2002 GENCONT: an operational tool for controlling inbreeding in selection and conservation schemes. Proceedings of 7th World Congr Genet Appl Livest Prod, Montpellier pp. 769–770.

Plummer, M., N. Best, K. Cowles, and K. Vines, 2006 CODA: Convergence diagnosis and output analysis for MCMC. R News **6**: 7–11.

Technow, F., 2014 *hypred: Simulation of Genomic Data in Applied Genetics*. R package version 0.5.

Urbanek, S., 2015 *rJava: Low-Level R to Java Interface*. R package version 0.9-7.

Warnes, G. R., B. Bolker, G. Gorjanc, G. Grothendieck, A. Korosec, T. Lumley, D. MacQueen, A. Magnusson, J. Rogers, and others, 2015 *gdata: Various R Programming Tools for Data Manipulation*. R package version 2.17.0.

Wimmer, V., T. Albrecht, H.-J. Auinger, and C.-C. Schoen, 2012 synbreed: a framework for the analysis of genomic prediction data using r. Bioinformatics **28**: 2086–2087.

Woolliams, J., P. Berg, B. Dagnachew, and T. Meuwissen, 2015 Genetic contributions and their optimization. Journal of Animal Breeding and Genetics **132**: 89–99.