TOWARDS IMPROVEMENT OF **RUMI**NANT BREEDING
THROUGH **GEN**OMIC AND EPIGENOMIC APPROACHES

Grant agreement number: 101000226

H2020 – Research and Innovation Action

Deliverable 8.3

Report on novel statistical tools and breeding strategies for balancing short/ long-term response and maintaining genetic diversity

Due date of deliverable: M30 – 11 2023

Actual submission date: M31 – 12 2023

WP concerned	WP8	
Lead partner	Han Mulder, WP leader Peer Berg, Task leader	Wageningen University and Research (WUR) Norwegian University of Life Sciences (NMBU)
Author(s)	Theo Meuwissen Xijiang Yu Mathias Schrauf Jeremie Vandenplas Yvonne Wientjes	NMBU NMBU WUR WUR WUR
Version	V1	
Dissemination level	- CO: Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°101000226.

Call: H2020-SFS-2020-2

Topic: SFS-13-2020 ● Genome and epigenome enabled breeding in terrestrial livestock

Start date of the project: June 1st, 2021 **Duration:** 60 months

End date of the project: May 31st, 2026 **Project ID:** 101000226

Table of content

1. Summary

Write a short summary of your Deliverable. This summary must be 2 pages maximum but very informative and must include the following elements:

Results in WP4 showed that the use of IBD (Identity by Descent) based genomic relationships in optimal contribution selection resulted in that (1) inbreeding rates were limited to a pre-defined value, (2) the highest ceiling of the selection response, reflecting maximum opportunities for long term response; and (3) high short term selection responses that were only exceeded by methods which severely compromised their response ceiling (long-term response) and genetic diversity.
The availability of dense markers has made it possible to use genome information as tools to monitor the genome structures of a population for future selection. Current computational achievements made it possible to explore the genome evolution and manipulate it by constraint with genome tools in a long-term course of selection.
Selection with breeding value using numerical, genomic, and IBD relationship matrices, namely ABLUP, GBLUP and IBLUP were explored. These three relationship matrices were also used as constraints for optimal contribution selection, namely AGBLUP (GBLUP with numerical relationship matrix **A** as constraint, and similar below), AABLUP, GGBLUP, IGBLUP and IIBLUP. Results of 20 generations of selection show that optimal contribution selection using **A** and **IBD** matrices have long-term benefits both on genetic progresses and maintaining the population diversity. The diversity maintained are reflected in less plus QTL loss, less inbreeding, less fixation of non-QTL loci, which are valuable for future challenges.
There is a wide range of genetic progress per inbreeding accumulation and per genetic potential loss, which shows the importance of the role of optimal contribution selection schemes. It is hence worthwhile to explore new optimal contribution selection algorithms for even better results.
To study possible solutions, we simulated populations of 1000 individuals subject to 50 generations of selection. We evaluated different genomic selection strategies to identify the ones which achieve the highest genetic gain in the long-term, better maintain genetic variance, and best conserve favourable de-novo mutations in the presence of additive and non-additive gene action. The genomic selection strategies were: truncation selection, which only focuses on short-term gain; and allele-reweighted selection, which upscales the effect of rare alleles in the breeding values. Two variants of allele-reweighted selection were implemented, one which softens the upscaling towards the last generations of the simulation (fixed time horizon) and another which maintains the same upscale strength throughout all generations (moving time horizon). Systematic differences between strategies were not observed for traits with epistatic gene action, although in general the purging of deleterious mutations was observed to be slow. For the trait under additive gene action, allele-reweighted selection with a fixed time-horizon obtained a higher long-term genetic gain than truncation selection while preserving a similar level of genetic variance. Meanwhile, allele-reweighted selection with a moving time-horizon obtained a similar long-term genetic gain than truncation selection while maintaining a higher proportion of standing genetic variance. These and similar strategies may contribute to the sustainable use of genomic selection.

Objectives:

Of your deliverable

WP 8 aims to develop and compare sustainable and socially acceptable breeding programs and management practices that would result in future generations of animals that are efficient and resilient to climate changes while maintaining genetic diversity. Tools and breeding programs developed will integrate genomic, epigenomic, and genome editing information. To reach this goal, WP8 will:

1. Investigate the impact of the historic evolution of breeding goals on trade-offs between production traits, fertility, and resilience, at the individual and genome-level for different cosmopolitan and local breeds in different environments; 2. Investigate the possible benefits of using trade-offs at the

genome level in balanced breeding strategies;
3. Propose optimal selection strategies to increase efficiency and resilience to climate changes;

4. Develop genomic prediction methods that better capture genetic variation due to genome edited variants, rare alleles, and de novo mutations;
5. Develop genomic selection strategies to balance short versus long term response, to manage the population genetic diversity and the overall genetic drift of the population for all the traits not improved by the selection, and to reduce the population genetic load at a predefined rate;

6. Develop statistical models that exploit epigenomic information to predict complex traits in uniform and contrasted environments.
The results of these objectives and of other WPs will be finally combined to propose strategies for adapting breeding objectives for future environmental conditions and conservation of genetic diversity within and among breeds.

Specifically this deliverable Reports on novel statistical tools and breeding strategies for balancing short/ long-term response and maintaining genetic diversity.

Rationale:

Describe the approach/methodology you chose to reach the objectives

The pedigree-based relationship matrix (**A**) is IBD based in that it tracks common ancestors in the pedigree and uses these to calculate IBD probabilities. But the **A** matrix does not account for all inbreeding since some of the inbreeding is due to genomic selection repeatedly selecting the same (superior) chromosome segments when selecting parents. Genomic based-IBD relationships can be calculated by linkage analysis which is an old technique known from QTL mapping, back when marker maps were sparse. However, in sparse QTL mapping this approach was applied to few QTL positions in shallow pedigrees. Meuwissen et al. (2010) developed linkage disequilibrium based multilocus iterative peeling (LDMIP), which can be applied to deep pedigrees and dense marker maps. But LDMIP is also too slow for current large-scale genomic data sets collected in practical national-scale-genomic selection schemes. The aim in Task 4.1 is to develop a new fast statistical tool for linkage analysis and thus for the calculation of IBD relationships in large breeding populations with dense marker maps.

Here we test the effect of IBD based relationships relative to pedigree or genomic relationships as commonly used in genetic evaluation, by stochastic simulation of a dairy cattle like breeding scheme for 20 generations to compare genetic gain, inbreeding and conversion of genetic variance to genetic gain.

While genomic selection has considerably increased the genetic gain for many breeding programs, consequences on diversity can be less desirable. This is particularly the case for rare alleles and de-novo mutations, as markers used in genomic selection are generally not strongly associated to rare alleles. Moreover, genomic selection allows for the selection of young individuals without records, thereby ignoring the effects of de-novo mutations. To study possible solutions, we simulated populations of 1000 individuals subject to 50 generations of selection. We evaluated different genomic selection strategies to identify the ones which achieve the highest genetic gain in the long-term, better maintain genetic variance, and best conserve favourable de-novo mutations in the presence of additive and non-additive gene action. The genomic selection strategies were: truncation selection, which only focuses on short-term gain; and allele-reweighted selection, which upscales the effect of rare alleles in the breeding values.

Teams involved:

Specify the list of RUMIGEN partners and other contributors that have worked on this Deliverable (name of the organisation).

Norwegian University of Life Sciences

Wageningen University and Research

2. Introduction

The general aim of livestock breeding programs is to improve future generations of livestock by selecting genetically superior males and females as parents of the next generation. Currently, most breeding programs have adopted genomic selection, resulting in an increase of short-term genetic improvement for traits under selection (Cole et al., 2020). However, there are indications that the higher short-term genetic improvement may come at the cost of a lower long- term genetic improvement due to an increased loss of genetic diversity (Doekes et al., 2018) and a smaller exploitation of de novo mutations (Mulder et al. 2019). Nowadays, several advances in DNA technologies can help us to unveil several aspects of complex traits inheritance. For example, recent advances in technology can provide whole genome sequence and genome-wide epigenetic information of an individual (Gonzalez-Recio, 2012). Furthermore, the advent of genome editing technologies has added the possibilities to alter individuals 'genomes (Bastiaansen et al., 2018). To date, it is unclear how these recent advances in DNA technologies could be combined with novel breeding strategies to result in socially acceptable breeding programs that lead to animals that are efficient and resilient to climate changes while maintaining genetic diversity.

3. Methods

Linkage analysis:

The linkage analysis technique developed for QTL mapping (Weller, 2001) predicted the probabilities of paternal versus maternal inheritance of chromosomal segments at (unknown) putative QTL positions. In the current genome-wide application of linkage analysis, the main difference is that we concentrate on known and informative marker positions, i.e. paternal versus maternal inheritance is assessed with certainty, and that these positions can differ from one parent-offspring pair to the next. For reasons of computational speed, positions in between these informative positions are imputed assuming that there are no double recombinations, which is reasonable due to the use of dense marker data.

Assessing the inheritance with certainty means that it is determined whether the paternally derived allele is transmitted to the offspring or the maternally derived allele. We developed the following algorithm to decide whether a marker position is informative about the paternal/maternal inheritance from parent to offspring, and if informative to decide on paternal/maternal inheritance. Assuming a SNP with alleles A and T, the steps of the algorithm are:

Step 1: The parent's genotype must be heterozygous for the marker position: AT (otherwise the marker genotype cannot be used to distinguish paternal from maternal inheritance).

Step 2: Given a AT parental genotype, it must be known whether the A allele is inherited from the grandsire (paternal) and the T allele from the granddam (maternal), or *vice versa*. This requires that at least one of the grandparents is genotyped and homozygous. E.g. if the grandsire has genotype AA, the A allele of the parent is paternal and its T allele is maternal.

Step 3: Given the known paternal (A) and maternal (T) allele of the parent, it must be known which allele was transmitted from the parent to its offspring. If the offspring genotype is homozygous, say

AA, the offspring inherited an A (paternal) allele from its parent. Alternatively, if the offspring's genotype is TT, this T allele was the maternal of its parent. Moreover, if the offspring genotype was heterozygous, say AT, and its other parent was homozygous, say TT, it inherited the paternal (A) allele from the first mentioned parent.

Step 4: non-informative marker genotypes: if a the paternal/maternal inheritance cannot be traced at a particular marker position, it will be assumed equal to that at its closest flanking marker position. This implies that, if a recombination occurred, it is assumed to have occurred in the middle between the informative flanking markers.

Step 5: in case of non-genotyped parents, grandparents or offspring, it is attempted to impute the missing genotypes. The non-genotyped animals are imputed based on genotypes of their offspring: heterozygous (say AT) if some of the offspring have opposite homozygous genotypes (some AA and some TT offspring); homozygous (say AA) if more than 10 offspring had genotype AA, without any TT offspring. In case this imputation does not resolve the missing genotypes, paternal / maternal inheritance probabilities are assumed to equal 50%.

Given the conditions described in steps 1-3, many SNPs will not be informative. However, due to the use of SNP-chip genotyping, we have dense marker genotypes and quite a few markers will be informative and provide flanking marker information on the paternal/maternal inheritance. Also, since few recombinations occur in a chromosome during the parent to offspring inheritance, relatively few informative marker positions suffice to trace the paternal/maternal inheritance over the chromosome. Hence, positions in between these informative positions are imputed to equal the paternal/maternal inheritance at its closest informative flanking marker position.

The linkage analysis based relationship matrix: $\mathbf{G_{LA}}$

When using dense marker information, the paternal/maternal inheritance pattern hardly changes from one marker to the next, since there are few recombinations. And since the calculation of an linkage analysis based relationship matrix is computationally costly, we only perform these calculations for every 10^{th} marker position, and subsequently average these IBD matrices across the marker positions. Since the inheritance patterns are assessed at the level of the paternal and maternal gametes, we set up a gametic relationship matrix $\mathbf{G_{LA}}$, i.e. the paternal and maternal gamete of each sire has its own entrance in the matrix. Thus, the matrix has 2N rows and columns, where N is the number of animals in the pedigree.

The algorithm for setting up the gametic $\mathbf{G_{LA}}$ matrix at marker position j is similar to that of the gametic relationship matrix based on pedigree. The animals are processed from old to young, and within animals its paternal gamete is processed first followed by its maternal gamete. The relationship of a gamete with itself is always 1, i.e. the diagonal of the IBD relationship consists of 1's and is not further considered in the algorithm below. Moreover, since the gametic relationship matrix (as any other relationship matrix) is symmetric, if we calculated the row elements for gamete i up to the diagonal (IBD(i,1:i-1)), the corresponding column elements (IBD(1:i-1,i)) are also known and filled in, which is denoted by IBD(i,1:i-1)= IBD(1:i-1,i). The $\mathbf{G_{LA}}$ at marker position j is obtained as:

$\mathbf{G_{LA\,j}} = \mathbf{I}$ (where $\mathbf{I}$ is an identity matrix)

For i = 1,2,…,N

For k = paternal, maternal

Step 1: if gamete ik was inherited from an unknown parent (e.g. paternal gamete from unknown sire), it is assumed a base gamete and assumed unrelated: $\mathbf{G_{LA\,j}}$(ik,1:ik-1)= $\mathbf{G_{LA\,j}}$(1:ik-1,ik)=0, where ik is the entrance of animal i gamete k into the $\mathbf{G_{LA\,j}}$ matrix.

Step 2: if gamete ik is inherited as the paternal allele of parent s, i.e. entry sp, its relationships are copied from entry sp: $\mathbf{G_{LA\,j}}$(ik,1:ik-1)= $\mathbf{G_{LA\,j}}$(1:ik-1,ik)= $\mathbf{G_{LA\,j}}$(sp,1:ik-1).

Step 3: if gamete ik is inherited as the maternal allele of parent s, i.e. entry sm, its relationships are copied from entry sm: $\mathbf{G_{LA\,j}}$(ik,1:ik-1)= $\mathbf{G_{LA\,j}}$(1:ik-1,ik)= $\mathbf{G_{LA\,j}}$(sm,1:ik-1).

Step 4: if the inheritance of gamete ik from parent s is unknown, i.e. paternal/maternal inheritance is 50/50: $\mathbf{G_{LA\,j}}$(ik,1:ik-1)= $\mathbf{G_{LA\,j}}$(1:ik-1,ik)= ½($\mathbf{G_{LA\,j}}$(sp,1:ik-1)+ $\mathbf{G_{LA\,j}}$(sm,1:ik-1)).

End of i and k loops.

As mentioned before, the $\mathbf{G_{LA\,j}}$ matrices are calculated for every 10th marker position and accumulated across the marker positions, in order to calculate the mean of $\mathbf{IBD_j}$ across the marker positions. The resulting gametic relationship matrix is called $\mathbf{G_{LA}}$ (without subscript j). The relationships between animals i and j are depicted by 4 elements in the $\mathbf{G_{LA}}$ matrix, namely the relationships between ip and jp, ip and jm, im and jp, and im and jm. For self-relationships of animal i these 4 relationships are 1, F_i, F_i and 1, respectively, where F_i is the linkage analysis based inbreeding coefficient of animal i. The $\mathbf{G_{LA}}$ relationship matrix at the animal level, i.e. 1 entry per animal, is obtained as half the sum of the aforementioned 4 elements of $\mathbf{G_{LA}}$. Hence, the self-relationship of animal i becomes 1+ F_i, as expected.

We tested the potential effect of IBD based relationship matrices by computing the true IBD based on known IBD markers in simulated pedigrees and compared with pedigree and genomic (van Raden) relationship matrices. See Annex A for details.

To study alternative methods for better utlising de-novo mutations and rare alleles we simulated breeding schemes with 1000 individuals for 50 generations, and compared random selection, truncation selection and allele re-weighted selection with either a fixed ot moving time horizon in the weighing factor. Following Wientjes et al. (2022), we simulated three traits: a trait with only additive effects (“additive”), a trait with additive and dominance effects (“dominant”), and a trait with additive, dominance, and epistatic effects (“epistatic”). For further details see Annex B.

4. Results

Result 1

In order to demonstrate the algorithm in the Methods section, a 10 generation pedigree was simulated following the breeding program in also used in Annex A, but no selection was assumed. Only a single chromosome was simulated which contained 3092 SNPs. In the simulated data, true relationships were assessed by allocating unique ID numbers to the founder alleles. Identity of founder allele IDs in any later generation thus indicated that this is a common ancestral allele and thus true inbreeding. The true IBD relationships (G_{IBD}(true)) were correlated to estimated IBD relationships based on linkage analysis using all generations’ genotypes (G_{LA}(1-10)), or only genotypes from the last 4 generations (G_{LA}(7-10)). In addition, the Genomic Relationship Matrix (**GRM**; VanRaden 2008, Method 1) and the pedigree based relationship matrix **A** were included n the comparison.

Table 1 shows the results of the comparisons between the methods. G_{LA}(1-10) relationships were almost perfectly correlated to G_{IBD}(true) relationships. But when only 4 generations were genotyped

the correlation between the relationships was a lot lower at 0.649. The A matrix has the poorest correlation with the true relationships, whilst the GRM relationships correlated quite well to the true relationships at 0.874. The regression coefficients of true relationships on estimated relationships were close to 1 for the IBD based methods (A and G_{LA}), which implies no inflation biases. The **GRM** had inflation bias, which implies that the GRM based relationships are more variable than the true relationships. This is also observed from the variances of the relationships. The GRM showed a mean relationship close to 0, which is because generation 10 allele frequencies were used as reference frequencies for the calculation of the **GRM**. Improved mean **GRM** relationships may be obtained when generation 0 allele frequencies had been used, but this was not attempted here. The use of generation 0 allele frequencies is not expected to affect the variance of GRM relationships, since it simply adds a constant, which accounts for the genetic drift since generation 0, to all **GRM** relationships.

Table 1. Relationships in generation 10 (correlations below diagonal; regression coefficients[1] above).

	A	G_{LA}(7-10)	G_{LA}(1-10)	GRM	**G_{IBD}(true)**
A	1.000	0.989	0.989	1.090	**1.009**
G_{LA}(7-10)	0.710	1.000	0.959	1.045	**0.955**
G_{LA}(1-10)	0.516	0.696	1.000	0.715	**1.033**
GRM	0.472	0.630	0.861	1.000	**0.776**
G_{IBD}(true)	**0.492**	**0.649**	**0.967**	**0.874**	**1.000**
Mean	0.066	0.066	0.078	-0.003	0.083
Var	0.005	0.009	0.018	0.026	0.020
[1]Regression of more variable on less variable relationship					

Result 2

All details are in Annex A. Main results are summarised in Table 2. Selection with breeding value using pedigree, genomic, and IBD relationship matrices, namely ABLUP, GBLUP and IBLUP were explored. These three relationship matrices were also used as constraints for optimal contribution selection, namely AGBLUP (GBLUP with numerical relationship matrix **A** as constraint, and similar below), AABLUP, GGBLUP, IGBLUP and IIBLUP. Results of 20 generations of selection show that optimal contribution selection have the highest genetic gain. Further, optimal contribution selection using **A** and **IBD** matrices have long-term benefits both on genetic progresses and maintaining the population diversity. The diversity maintained are reflected in less positive QTL alleles being lost, less inbreeding, less fixation at non-QTL loci, which are valuable for future challenges. IBD based matrices were superior in converting genetic variance into genetic gain, irrespective of whether this was measured relative to the maximal gain possible or inbreeding.

Table 2: $\Delta\overline{\text{TBV}}$ vs breeding ceiling and inbreeding changes for different breeding schemes. $\Delta\overline{\text{TBV}}$ and ceilings are of unit σ_0, the standard genic standard deviation of the founder population. The results were ordered on $\Delta\overline{\text{TBV}}/\Delta\overline{\text{Ceiling}}$.

Scheme	$\Delta\overline{\text{TBV}}$	**$\overline{\text{Ceiling}}$**	$\Delta\overline{\text{TBV}}/\Delta\overline{\text{Ceiling}}$ (%)	$\Delta\overline{\text{TBV}}/\Delta\bar{F}$	$\Delta\overline{\text{Ceiling}}/\Delta\overline{\text{TBV}}$
ABLUP	7.1	91.9	16.4	20.1	6.1
IBLUP	8.1	96.6	21.0	23.5	4.8
GGBLUP	14.0	72.9	22.5	27.5	4.4
GBLUP	8.4	97.9	22.6	25.6	4.4
AGBLUP	12.4	96.5	32.2	35.6	3.1

AABLUP	10.5	104.7	34.6	42.5	2.9
IGBLUP	11.4	105.6	38.9	52.3	2.6
IIBLUP	11.0	107.0	39.1	50.0	2.6

Result 3.

For full details see Annex B. We evaluated different genomic selection strategies to identify the ones which achieve the highest genetic gain in the long-term, better maintain genetic variance, and best conserve favourable de-novo mutations in the presence of additive and non-additive gene action. The genomic selection strategies were: truncation selection, which only focuses on short-term gain; and allele-reweighted selection, which upscales the effect of rare alleles in the breeding values. Two variants of allele-reweighted selection were implemented, one which softens the upscaling towards the last generations of the simulation (fixed time horizon) and another which maintains the same upscale strength throughout all generations (moving time horizon). Systematic differences between strategies were not observed for traits with epistatic gene action, although in general the purging of deleterious mutations was observed to be slow. For the trait under additive gene action, allele-reweighted selection with a fixed time-horizon obtained a higher long-term genetic gain than truncation selection while preserving a similar level of genetic variance. Meanwhile, allele-reweighted selection with a moving time-horizon obtained a similar long-term genetic gain than truncation selection while maintaining a higher proportion of standing genetic variance. These and similar strategies may contribute to the sustainable use of genomic selection.

Figure 1 highlights that allele re-weighting led to a larger genetic gain at a lower loss of genetic variance. However, there was no significant differences in the contribution of de-novo mutations to true breeding values between selection methods.

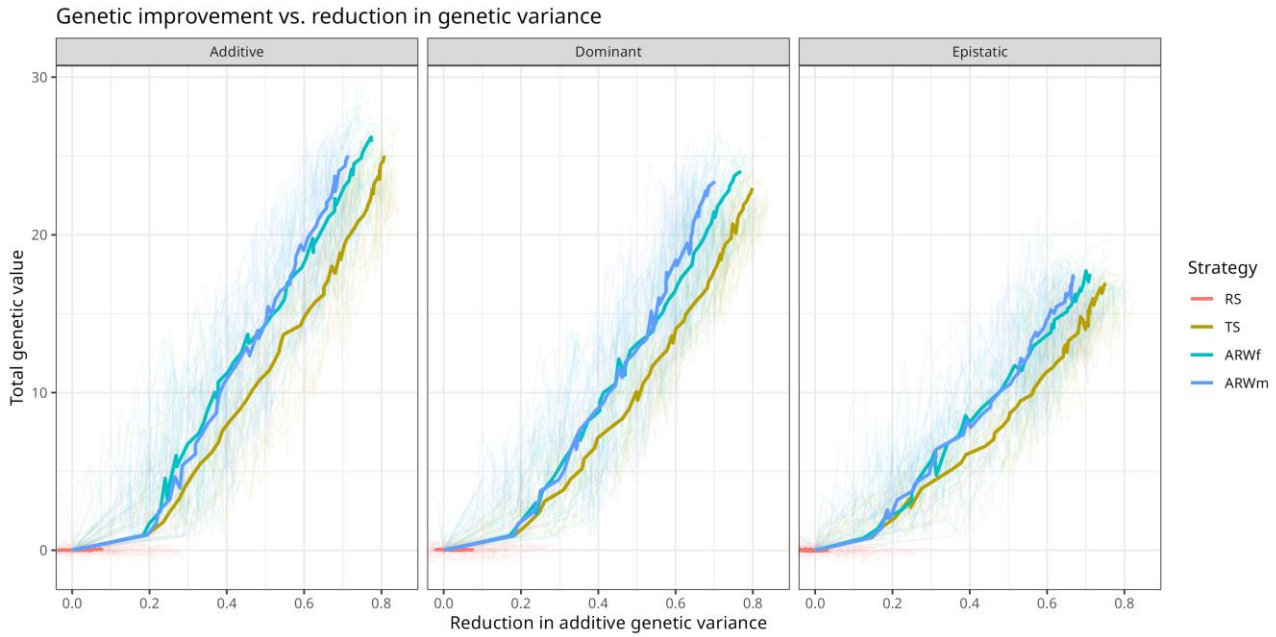

Figure 1: Reduction in additive genetic variance against total genetic value.

5. Conclusion

1. A computationally efficient linkage analysis algorithm was developed to calculate an IBD based relationship matrix $\mathbf{G}_{LA}$.

2. The $\mathbf{G_{LA}}$ bears some similarities to the predigree based **A** matrix, in that it assumes a founder population where individuals are assumed unrelated and non-inbred, but it uses genomic information to assess paternal versus maternal inheritance whereas **A** assumes 50/50 probabilities of maternal/paternal inheritance.
3. When all generations are genotyped, the $\mathbf{G_{LA}}$ shows a very high correlation of 0.97 with the true IBD relationships, but this accuracy drops substantially when some generations are not genotyped.
4. Calculation of the $\mathbf{G_{LA}}$. requires 3 generations of genotypes, although the genotypes of the ungenotyped grand-parents may be imputed from their offspring genotypes if they have more than 10 offspring, and one genotyped or imputed grandparent per parent suffices.
5. The GRM is also quite highly correlated to the true IBD relationships at a correlation of 0.87 and requires only genotypes in the last generation.
6. The GRM shows however inflation bias, i.e. the GRM relationships are substantially more variable than the true IBD relationships.
7. The $\mathbf{G_{LA}}$ relationships hardly show any inflation bias.
8. IBD based OCS selection are very effective on making genetic progresses whilst keeping the diversity of the population under selection. IBD based OCS have the highest genetic gain relative to both inbreeding and loss of potential future genetic gain.
9. The increased diversity maintained by OCS schemes are reflected in less positive QTL alleles being lost, less inbreeding, and less fixation at non-QTL loci.
10. Allele-reweighted selection, for the trait under additive gene action with a fixed time-horizon obtained a higher long-term genetic gain than truncation selection while preserving a similar level of genetic variance.
11. Allele-reweighted selection with a moving time-horizon obtained a similar long-term genetic gain than truncation selection while maintaining a higher proportion of standing genetic variance. These and similar strategies may contribute to the sustainable use of genomic selection.
12. Minor differences were found between truncation selection and allele re-weighted selection in the utilisation of de-novo mutations.

6. Annex

A. Alternative Relationship Matrices for Genomic Prediction and Managing Diversity
B. Long-term performance of genomic selection strategies with a focus on the use of rare alleles