Maintaining long-term genetic gain using a segment-stacking algorithm

Kira Villiers (ORCID: 0000-0002-6410-3477)[1,*], Kai P. Voss-Fels (ORCID: 0000-0003-0782-366X)[2,1], Eric Dinglasan (ORCID: 0000-0002-4824-5977)[1], Bertus Jacobs[3], Lee Hickey (ORCID: 0000-0001-6909-7101)[1], Ben J. Hayes (ORCID: 0000-0002-5606-3970)[1]

[1] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, QLD, Australia.

[2] Department of Grapevine Breeding, Hochschule Geisenheim University, Geisenheim, Germany.

[3] LongReach Plant Breeders Management Pty Ltd, Lonsdale, SA, Australia.

[*] To whom correspondence should be addressed.

Contact address: k.villiers@uq.edu.au

Abstract

Loss of genetic diversity in elite crop breeding pools can severely limit long-term genetic gains, and limits gain for new traits that are becoming important as the climate changes, such as heat tolerance. Introgression of specific traits and pre-breeding germplasm is an alternative to introduce new diversity, however, introgression is typically slow and challenging. Here we use a genetic algorithm (GA) to select sets of parents which between them contain chromosome segments with very high segment breeding values for the target trait, as an alternative to truncation selection on genomic estimated breeding values (GEBVs). Repeated recurrent selection and intercrossing for a yield trait, beginning with a founder population of wheat genotypes and guided by the GA, was compared to repeated recurrent selection and to optimal contribution selection (OCS) on yield GEBV. After 100 generations of selection and intercrossing, truncation selection had exhausted the genetic diversity, while considerable diversity remained in the GA population. In some situations, particularly where the number of parents crossed each generation was relatively small and the number of progeny per cross was large, gain from the GA exceeded that from truncation selection. Compared to OCS, selection under the GA maintained more useful diversity, i.e. diverse segments with high yield GEBV, which allowed it to maintain higher rates of gain for longer. These results indicate that GA-based parent selection could be a promising method to maintain diverse alleles with higher genetic merit in breeding populations with little additional effort to a regular genomic selection program.

Keywords: Long-term selection; genomic selection; genetic algorithm; breeding simulation

Key message (30 words or less): A segment-stacking algorithm can outperform optimal contribution selection and truncation selection in long-term crop breeding programs, particularly with small populations, while also maintaining more useful diversity.

Intro

A breeding program can be seen as a targeted reduction in the diversity of a population. That is, undesirable alleles are culled from the pool, and beneficial alleles are increased in frequency. However, a breeding program with no diversity can produce no genetic gains, because every genotype in the population is identical with regards to its allelic constitution and represents the peak of the population's genetic merit. A significant challenge for elite plant breeding programs is to maintain or increasing rates of gain without exhausting their breeding program's genetic diversity.

Maintaining or improving rates of gain is a necessary task to both ensure long term gains and to enable selection for new traits, such as heat tolerance as the climate changes. Crop breeding programs are increasingly turning to genomic selection (GS) to increase genetic gains. GS involves ranking selection candidates or candidate parents for parent selection on genomic estimated breeding values (GEBVs) calculated from genotypes marker effects estimated from a reference population (Meuwissen et al. 2001). Genomic selection can provide very high rates of gain in some species as it enables early selection of breeding individuals (Voss-Fels et al. 2019). The introduction of GS has boosted genetic gain in both animal and plant breeding programs (García-Ruiz et al. 2016; Guinan et al. 2023; Scott et al. 2021; Wolc et al. 2015; Crossa et al. 2014; Cooper et al. 2014; Bernardo 2021; Hickey et al. 2017).

An emerging strategy to integrate GS in crop breeding programs is rapid recurrent selection on GEBVs, in which breeding cycles can be iterated very quickly due to the ability to identify and cross the best individuals, as early as generation F1. However, truncation approaches to selecting the 'best' that cull all but the top ranked individuals cause dramatic loss of diversity and often lead to rapid inbreeding in the breeding pool (Lin et al. 2016; Cowling et al. 2019). In dairy cattle, for example, the rate of inbreeding has doubled as a result of the introduction of genomic selection (Scott et al. 2021).

Potentially, the loss of diversity caused by selecting this way could be mitigated by introgressing genetically distant germplasm. However, truncation selection is not well suited to introgression scenarios. Diverse or landrace germplasm often lacks important agronomic traits compared to elite varieties that have been developed through generations of targeted selection. The diverse germplasm's mediocre GEBVs for key traits would see it and its descendants culled from the breeding program, even though they may carry beneficial alleles not yet present in elite varieties. The benefits of the diverse germplasm are 'masked' by the whole-genome GEBV scores (Gorjanc et al. 2016). A whole-genome perspective loses the locus-specific information that could distinguish the germplasm's usefulness based on individual alleles.

Optimal contribution selection (OCS) is a strategy for balancing genetic gain and genetic diversity in animal and plant breeding programs (Wray and Goddard 1994; Meuwissen 1997; Woolliams et al. 2015; Lin et al. 2017; Gorjanc et al. 2018). In OCS, breeders can choose how to balance coancestry, a measure of diversity, against a genetic value measure of choice, e.g. the trait GEBV. This produces a Pareto frontier of OCS strategies for different goals or priorities. Used with genotyping and GEBVs to calculate the value of candidates, it can also be used as the selection method in a rapid recurrent selection program. OCS programs determine genetic contribution proportions from each potential parent, or can be adapted to provide mate allocations (Kinghorn 2011). OCS was designed for animal breeding, but is also seeing increasing use in plant breeding. Different plant breeding program designs based on OCS have been proposed (Cowling et al. 2017; Cowling et al. 2019), and it has been applied in forestry (Kerr et al. 2015), common bean breeding (Saradadevi et al. 2021), and spring canola breeding (Cowling et al. 2023) projects.

Segment stacking or "ultimate genotype" building approaches are not widely applied in plant breeding at present. In stacking approaches, genetic gain relies on beneficial recombination events causing the accumulation of good alleles or good segments in the same genotype (Dekkers and Hospital 2002; Servin et al. 2004). The minimum population sizes to guarantee fixation of desirable alleles or segments grows exponentially with the number of alleles or segments (Hospital 2003; Servin et al. 2004). Therefore, segment stacking does not seem suited to use together with GS, which uses thousands of markers. However, by bundling markers into chromosome segments, and selecting on those segment haplotypes, this kind of selection becomes feasible.

For example, Daetwyler et al. (2015) simulated selection for the plant with the best component haplotype segments in a haploid-doubling based breeding program. The selection method increased both genetic gain and diversity of the population. Kemper et al. (2012) investigated a similar selection method based on targeting the "ultimate genotype", the highest-scoring genotype possible from segments in the selection. They developed a genetic algorithm (GA) to select an initial set of bulls (theirs was a dairy example) with high potential gain in economic value. Their study then tested segment stacking strategies in simulation on this initial set and found the stacking techniques were quite effective at realising the ultimate genotype, though they assumed free recombination between segments and no recombination within segments. They did not, however, test the performance of the GA used to select the parental genotypes over repeated generations of intercrossing and selection.

Here, we investigate the segment stacking strategy implemented in Kemper et al. (2012)'s GA for performance in crop breeding programs for achieving genetic gain, and maintaining diversity, when applied over successive generations.

We performed a simulation study, starting with based on real wheat founder genotypes, that tested the performance of recurrent rounds of truncation selection, OCS, and Kemper et al. (2012)'s GA over a long time horizon on both genetic gain and genetic diversity. We hypothesised that the segment stacking approach described by Kemper et al. (2012) could maximise long term genetic gains in plant breeding programs, and could produce competitively viable gains in the shorter term when only a small number of founder lines are available for the breeding program. We also hypothesised this approach could be competitive with or even outperform OCS, with regards to maintaining genetic diversity and ensuring long-term gains for complex traits.

Methods

Base population

A set of 16,565 diverse elite wheat germplasm lines, phased and genotyped across 5112 SNPs on 21 chromosomes (with a total genome length of 2994 centimorgans), were provided by breeding company LongReach Plant Breeders. A set of real additive effect effects for the 5112 SNPs, for grain yield, were also provided (SNP yield effects were from best linear unbiased estimates (BLUEs) on data from 186 field trials between 2012 and 2018). In all simulations, these were assumed to represent true and exact effects that remained unchanged over generations, so selection was carried out on true breeding values. The local GEBVs of chromosome segments for GA selection were calculated for each chromosome segment haplotype as the sum of the marker effects of all markers on that haplotype in that segment:

$$\text{local GEBV} = \sum_{k \in m} h_k e_{1k} + (1 - h_k) e_{0k}$$

where m is the set of SNP markers in the segment of the variant, h_k is the allele of one marker (either 0 or 1), and e_{0k} and e_{1k} are the additive marker effects of alleles 0 and 1 on marker k respectively. The number of chromosome segments was fixed at 105 (1/5th of each chromosome by length) for consistency for all simulations.

Two base populations of 50 founders were prepared from the diverse germplasm. One set (GA-selected founders, abbreviated GA_f, GA described below) was selected by the GA – i.e., the 50 genotypes with the highest potential stacked local GEBV. The other set (GS-selected founders, abbreviated GS_f) was selected by truncation selection – i.e., the 50 genotypes with the highest GEBVs. The genomic relationships among the base population genotypes are shown in Figure 1.

Simulating progeny

All offspring beyond the base populations were generated using simulation. To generate the genomes of offspring, recombination between a parents' gametes was simulated as occurring with uniform density across each chromosome, using the Haldane mapping function. No mutation was simulated. The simulation was implemented with genomicSimulation (Villiers et al. 2022). Local GEBV were calculated for each progeny and each chromosome segment as described above.

For direct candidate-selection methods, like truncation selection and Kemper et al. (2012)'s GA, the half diallel of 50 x 49 / 2 = 1,225 crosses was simulated to create a pool of offspring, then selection was applied, to select the parents for the next generation out of this pool (Figures 2a, 2c). For the mate-selection approach of OCS, one offspring was produced from each mate pair suggested by OCS (Figure 2b). The selected subset then became the parents of the next generation. This was continued for 100 generations.

As seen in Figure 2), two parameters besides the selection method were varied between simulation trials: the number of progeny p produced from each cross in the diallel (1, 10, or 100; this parameter does not apply in OCS conditions because of the lack of diallel of crosses) and the number of genotypes n to be selected to progress to the next generation (2, 10, 25, 50 or 100), so that every generation after the founders had n members. A factorial design (Table 1) was used to cover all possible pairs of these two parameters.

Selection and mating strategies

Three selection strategies were compared in the simulation, truncation selection (Figure 2a), optimum contribution selection (Figure 2b) and Kemper et al. (2012)'s GA for segment stacking (Figure 2c). We investigated the genetic gain and genetic diversity over generations using both base populations (GA or truncation selection selected).

For truncation selection, the top n offspring by GEBV of the diallel were selected each generation.

The tool AlphaMate (Gorjanc and Hickey 2018) was used to perform OCS on the parents set each generation, and return a mating plan that was then implemented to produce the n offspring to be progressed to the next generation. A grid search with a coarseness of 5 degrees of OCS target angles, followed by binary search to a coarseness of 0.5 degrees, was performed for each value of n, to find the OCS target angles to 0.5 degree accuracy that best matched the following comparison criteria (see Table 1): the initial degree of coancestry of the 100-progeny-per-cross truncation condition of the same n (100 progeny per cross in order to have no limit on number of offspring that can be selected from the same cross, to match OCS' similar lack of limitation); the initial degree of coancestry of the 10-progeny-per-cross GA condition of the same n (10 progeny per cross because 100 progeny per cross GA conditions do not exist for n of 50 or 100 (see Table 1)); and initial rate of

GEBV increase of the 10-progeny-per-cross GA condition of the same n. These provided opportunities to compare GA and truncation selection with OCS conditions that theoretically shared similar weighting of diversity and genetic gain.

Kemper et al. (2012)'s GA aims to maximise the score of the "ultimate genotype" that can be constructed out of an n-member subset of its input genotypes. The ultimate genotype of a subset is defined as the genotype heterozygous at each chromosome segment in the genome for the best two haplotype segments that originate in separate members of the subset.

Choosing a subset that contains the segments to produce the highest-scoring ultimate genotype is a highly combinatorial problem. The number of possible subsets grows rapidly with the size of the population and the size of the desired subset, such that it is completely infeasible to explore all possible subsets to find the optimum. Even the task of choosing 50 founders from a base breeding population of 200 has over 10^{47} possible subsets of 50. This is too large of a task to solve analytically. The GA is a suitable approach because it is an evolutionary search strategy, not guaranteed to find the subset with the globally maximal ultimate genotype score, but able to refine good choices of subsets into better ones to produce a 'good enough' solution in a defined number of iterations.

The GA takes as input a table of the local GEBVs of the two haplotypes of each input genotype at each of a set of defined segments. It gives as output the set of n of the input genotypes (parents) that assemble the highest ultimate genotype and so have the highest stacking potential of the sets it has explored. Formally, the fitness function of the GA is:

$$\text{Maximise} \sum_{g \in C} \sum_{s \in S} a_{gs}\, x_{gs}, \text{ subject to } \sum_{g \in C} a_{gs} = 2 \text{ for all } s \in S \text{ and with } |C| = n$$

where g is a genotype in C, the current n-member subset being evaluated; S is the set of chromosome segments on which the potential founders are scored; a_{gs} are a collection of binary decision variables used to select the two genotypes containing the two best haplotypes for each position, to contribute to the ultimate genotype score; and x_{gs} is the higher of the genotype g's two haplotype scores for segment s. That is, the fitness function searches for a subset C that maximises the GEBV of the ultimate genotype given the heterozygosity constraint that for each chromosome segment two haplotypes with the highest local GEBV in the population should be selected. Preliminary trials showed that running the GA with an internal population size of 1000 chromosomes for 2500 generations reliably led to acceptable convergence of the fitness function. Further information on the design of the GA is available in Kemper et al. (2012)'s Supplementaries.

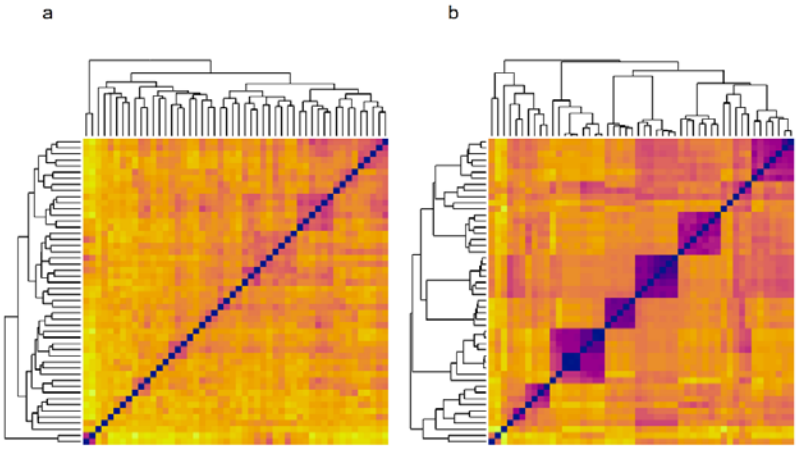

Fig. 1: Summary of the base population wheat genotypes. (a) Heatmap of Roger's genetic distance between GA-selected founding genotypes (that is, the 50 genotypes selected by Kemper et al. (2012)'s genetic algorithm (GA) from a large diverse population). Dark colours indicate a low genetic distance between pairs of genotypes. (b) Corresponding heatmap for genetic distance between GS-selected founding genotypes (that is, the 50 genotypes selected by truncation selection from the large diverse population).

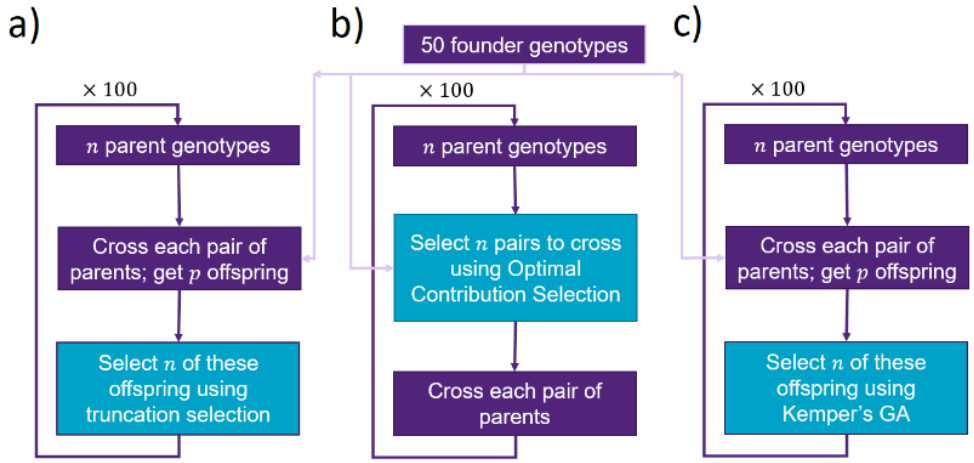

Fig. 2: Structure for the three compared simulation experiments. (a) Truncation selection simulation structure involves simulating a cross producing full-sibling progeny between every possible pair of founder genotypes, followed by selecting of the resulting pool of offspring and installing these genotypes as the founders of the next generation. (b) OCS (optimal contribution selection) simulation structure involves using AlphaMate (Gorjanc and Hickey 2018) to identify mate allocations from the set of founder genotypes. The mating plan is then followed to produce the next generation. The change-of-order of the selection stage in the cycle is because OCS selects parent contributions, rather than choosing individual genotypes. Grid search with a coarseness of 2.5 degrees was used to identify the target degree for AlphaMate (that is, the balance of genetic gain vs coancestry) that is matched to the initial rates of gain and coancestry increase of truncation selection and GA selections. (c) GA selection simulation structure is the same as the truncation selection procedure, but selection is done using Kemper et al. (2012)'s genetic algorithm (GA).

Table 1: Factorial design of simulation experiments. The three selection conditions (GA selection (that is, selection using Kemper et al. (2012)'s genetic algorithm), truncation selection, and OCS (optimal contribution selection)) are shown as major row headers. Each GA or truncation selection condition can be simulated with 1, 10, or 100 progeny per cross ("ppc"), and a population size of 2, 10, 25, 50 or 100. Because producing one progeny per cross from the one possible cross between two parents only produces one offspring, which is below the replacement rate needed for the next generation, the 1 progeny per cross and population size 2 combination is not possible. The GA implementation's output format was

incompatible with offspring pools of more than 99 999 members, so the 100 progeny per cross and population size 50, and the 100 progeny per cross and population size 100 could not be simulated with the GA. The OCS software, AlphaMate (Gorjanc and Hickey 2018), did not have a setting to restrict the number of uses of a particular parent in its crossing plan, so the progeny per cross parameter could not be controlled. Instead, grid search with a coarseness of 5 degrees, followed by binary search to a coarseness of 0.5 degrees, was used to identify the target degree for AlphaMate that matched truncation selection and GA selections' initial rates of gain and initial coancestry. The identified target degree for each comparative OCS condition in shown in the table.

		Select 2	Select 10	Select 25	Select 50	Select 100
GA	1 ppc		X	X	X	X
	10 ppc	X	X	X	X	X
	100 ppc	X	X	X		
Truncation	1 ppc		X	X	X	X
	10 ppc	X	X	X	X	X
	100 ppc	X	X	X	X	X
OCS	Match truncation initial coancestry		10°	30°	30°	35°
	Match GA initial gain		45°	75°	80°	86°
	Match GA initial coancestry		70°	78°	82.5°	86°

Summary and analysis

Simulation of each condition was run independently 20 times for all but the largest GA scenarios (10 progeny per cross and a population size of 50 or 100), which, due to runtime issues, were only replicated 10 times. Four summary statistics were calculated for each generation of each simulation trial: mean GEBV, maximum possible haplotype stack score, mean coancestry, and a gene diversity index from Nei (1973). The results show these summary measures averaged across replications.

The first two summary measures were used to judge rate of genetic gain. The first was the mean GEBV of the population at a given generation. The second was the partial yield GEBV of the "ultimate haplotype" possible from that population. This ultimate haplotype was constructed by choosing, ,at each marker, the allele present in the population with the highest contribution to the yield score. It can be interpreted as $0.5 \times$ the upper limit on the GEBV of this population's descendants, assuming a completely additive model of marker effects and no introduction of fresh diversity.

The third summary statistic, the mean coancestry of the population at a given generation, is calculated as twice the mean of the values in the genomic relationship matrix, including the diagonal, with the reference population (base) being the founder set, GA_f or GS_f (Gorjanc and Hickey 2018; Isik, Holland, and Maltecca 2017).

Finally, the last statistic was the mean of Nei (1973)'s gene diversity index H_e among a given generation's population, calculated as follows:

$$\frac{1}{|C|}\sum_{g\in C} H_e = \frac{1}{|C|}\sum_{g\in C}\left(1-\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{p}_\mathrm{i}^2\right) = \frac{1}{|C|}\sum_{g\in C} 2\mathrm{p}_\mathrm{A}\mathrm{p}_\mathrm{T}$$

where g is a genotype in C, the population of selected genotypes of a given generation. The simplification is possible thanks to all markers being SNPs (thus $k = 2$), and the assumption of Hardy–Weinberg equilibrium.

Results

Truncation selection versus GA for genetic gain and maintaining genetic diversity

The base populations selected by truncation selection and the GA were quite distinct sets, with the populations selected by the GA being considerably more diverse (Figures 1a and 1b). After selecting founders, both recurrent selection using the GA and recurrent truncation selection were carried out on the two base populations (Figures 3 and 4). GA selection samples more evenly from the available genetic diversity. Conversely, truncation selection strongly favours offspring of only a few founders. Compare the more uniform proportion of first-generation offspring of each founder chosen under GA selection on either founder population, to the favouring of the offspring of a few founders under truncation selection (Figures 3a and 3b).

Figure 4 compares the average GEBV of the population at each generation under the aforementioned four simulation conditions: GS-selected base population and recurrent GA selection; GS-selected base population and recurrent truncation selection; GA-selected base population and recurrent GA selection; and GA-selected base population and recurrent truncation selection. Selecting the initial founder population using the GA produces higher GEBV scores in the long-term: under either simulation strategy, rates of gain are steeper. Recurrent truncation selection on the more diverse base population surpasses the mean GEBV of truncation on the less-diverse starting population after 12 generations. Recurrent GA selection produces higher long-term results than recurrent truncation selection. They begin to outperform the truncation strategy on the same base population after 70 generations under breeding program parameters of 10 progeny per cross, and selecting 25 genotypes each generation.

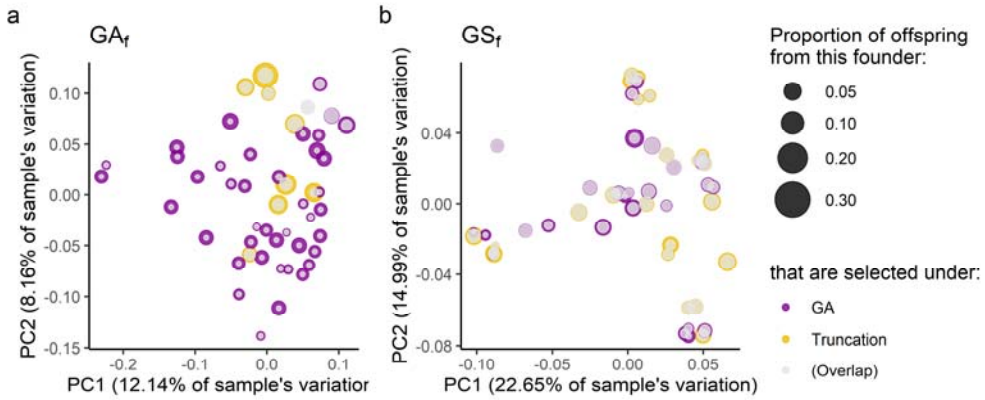

Fig. 3: Founder population diversity and proportion of lines selected by truncation and genetic algorithm-guided selection. (a) Principal component plot of GA_f, the founding population of 50 genotypes selected by Kemper et al. (2012)'s genetic algorithm (GA) from a large diverse population, with the size of the scatter points corresponding to the proportion of the first generation selection that has that founder as a parent. Proportions of first generation offspring selected by Kemper et al. (2012)'s GA and of first generation offspring selected by truncation selection are shown in different colours. (b) Corresponding principal component plot of GS_f, the founding population of 50 genotypes selected by truncation selection from the large diverse population.

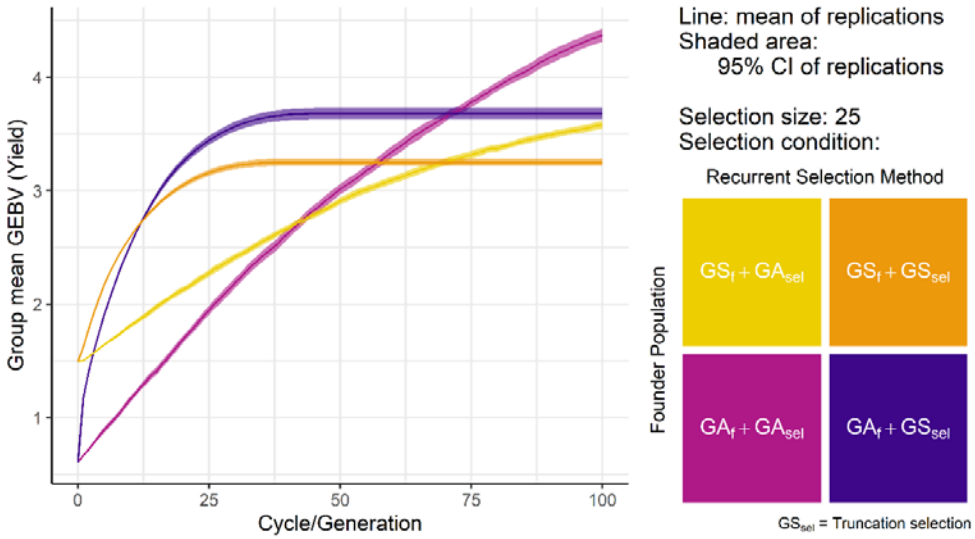

Fig. 4: **Trends in simulation results, for different founder population diversity** Each generation's mean and 95% confidence interval of yield GEBV, comparing conditions having GA-selected (GA_f) and GS-selected (GS_f) founder sets and having GA recurrent selection or truncation (GS) recurrent selection. The simulation runs in this plot use the parameters 10 progeny per cross and recurrent selection size of 25.

Even if the set of available germplasm is not as diverse as the population from which the base populations were drawn, as little as a single generation of GA selection boosts the long-term performance of the population. One of the simulation conditions shown in Figure 5a) is a 'hybrid' selection condition, in which the first generation is selected with the GA, but following generations are selected by truncation. The 'hybrid' selection results show that even one generation of GA in the early stages of a program is sufficient to provide a dramatically higher GEBV plateau, and the plateau is achieved no more slowly than other truncation conditions.

Figure 5a) is a plot of trends in genetic gain over the 100 generations, comparing GA and truncation selection over different numbers of progeny per cross. Figure 5b) shows the corresponding trends in genetic diversity, and Figure 5c) the trends in ultimate haplotype score. Figure 5a) shows the same general profiles as Figure 4: while recurrent truncation selection rounds result in very high initial rates of gain, the yield GEBV soon reaches a plateau and sees no improvement thereafter. 5b) demonstrates that the plateau occurs when all the diversity of the population is exhausted, so no further improvement is possible. Meanwhile, GA selection conditions show substantially slower initial rates of gain but also significantly lower rates of loss of diversity, leaving them at the end of 100 generations with populations with higher GEBVs and a retained stock of diversity they can use to continue improving into the future. From Figure 5c), it is visible that the initial high rates of gain of truncation selection cause the loss of much useful diversity (seen in falling ultimate haplotype scores). The GA maintains more genetic diversity over a longer time period.

Figures 6) and 7) provide summaries of the same three summary statistics as Figure 5) at specific points in time in the simulation runs, across the different combinations of selection size and progeny per cross parameters. Figure 6) shows the summary after 10 generations of selection, and Figure 7) the summary after 100 generations. GA selection benefits from more progeny per cross, but while in early generations truncation selection likewise shows higher GEBVs in conditions with more progeny per cross, truncation selection conditions with more progeny per cross have lower GEBVs than their few-progeny counterparts in later generations, after the plateau is reached. While truncation selection benefits from higher population sizes for achieving higher rates of genetic gain, GA

selection benefits, down to a point, from smaller population sizes. When the selection size is 2, both selection conditions exhaust their diversity too fast to see much genetic gain. But GEBVs from GA selection are higher when the size of the selection is 10 than when it is 25, and higher when it is 25 than when it is 50 or 100. Larger selection sizes however allow the GA to maintain more diversity than smaller ones. By generation 100, all genetic diversity with truncation selection was exhausted, but all GA conditions with population sizes greater than 2 maintain diversity under Nei (1973)'s index, and had higher ultimate haplotype scores, suggesting useful diversity to use to continue to build improvements.

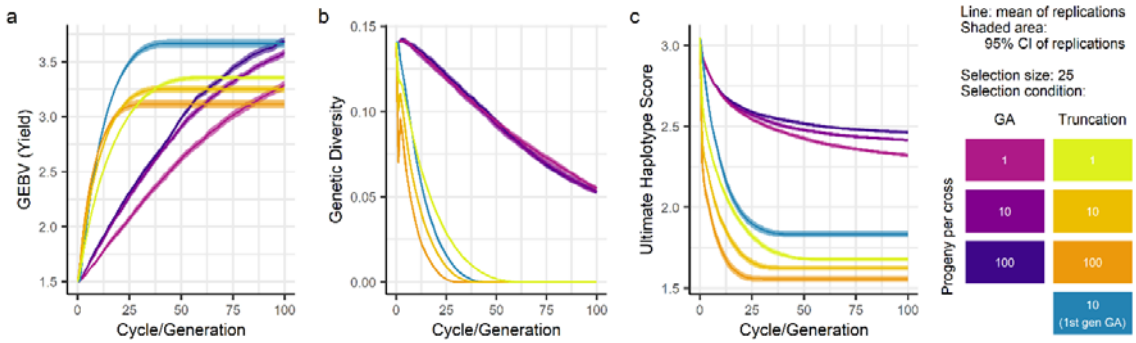

Fig. 5: Comparison of truncation selection and Kemper et al. (2012)'s genetic algorithm (GA) for segment stacking. (a) Average progeny GEBV over generations. (b) Genetic diversity over generations (Nei (1973)'s expected heterozygosity measure). (c) The score of the ultimate haplotype stack possible from that population. Results are from simulation with a recurrent selection size of 25. Represented in the plots are the following conditions: selection by GA with 1, 10 or 100 progeny per cross; truncation selection with 1, 10 or 100 progeny per cross; and a 'hybrid' condition with 10 progeny per cross, where selection is done by GA in the first generation and by truncation for every generation thereafter.

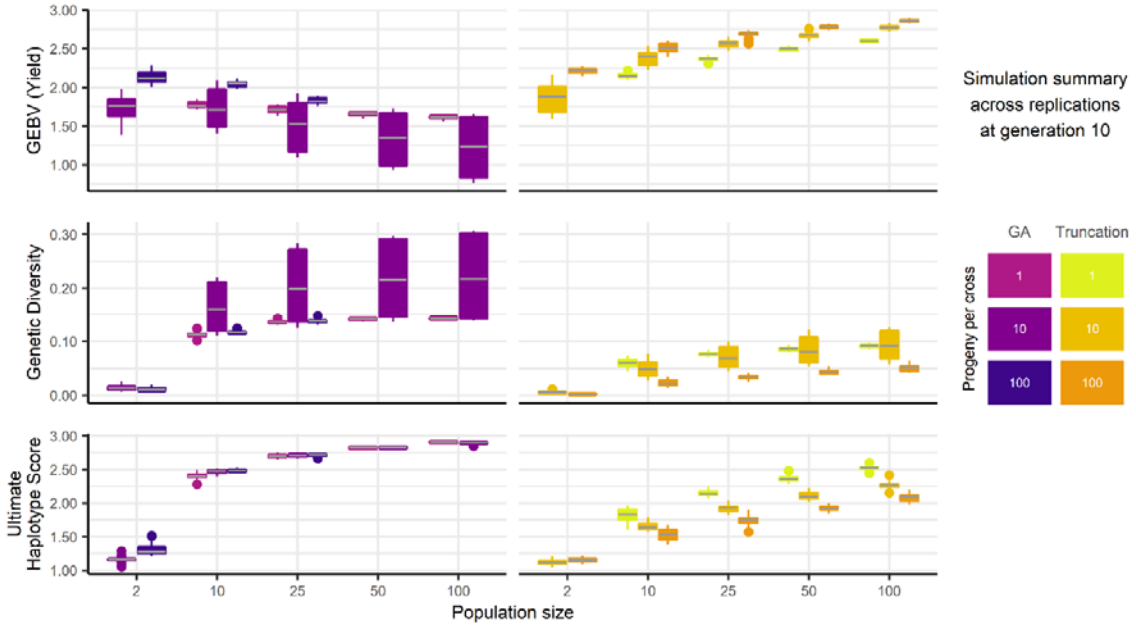

Fig. 6: Comparison of Kemper et al. (2012)'s genetic algorithm (GA) and truncation selection after 10 generations, across all combinations of population size and progeny per cross parameters in the factorial design.

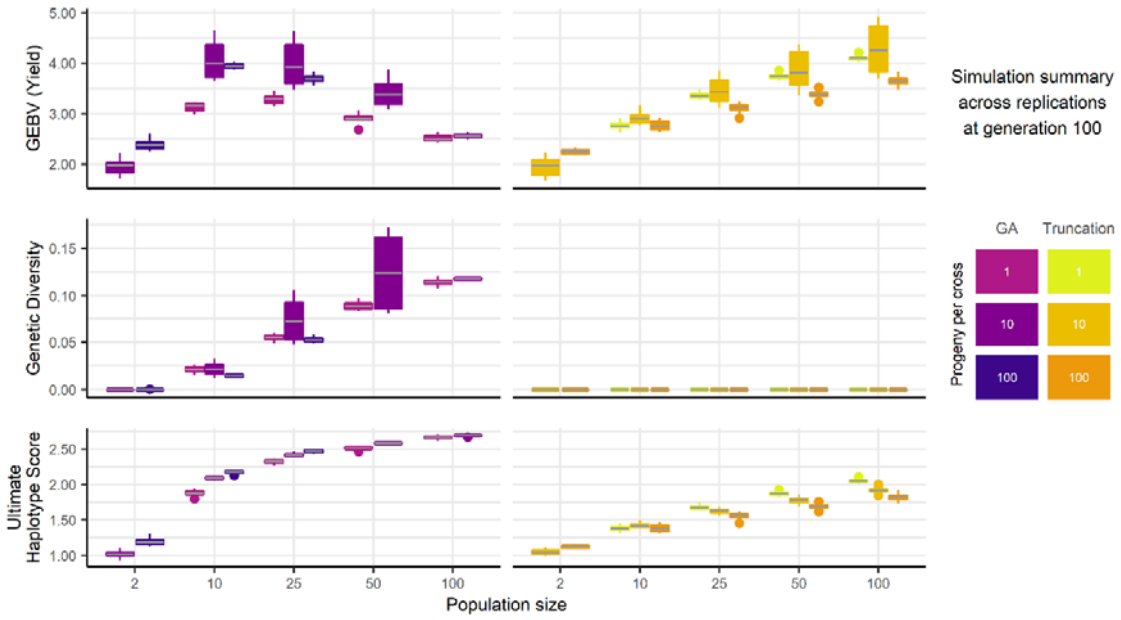

Fig. 7: Comparison of Kemper et al.'s genetic algorithm (GA) and truncation selection after 100 generations, across all combinations of population size and progeny per cross in the factorial design. For all truncation selection scenarios, genetic diversity was exhausted (0) by generation 100.

Optimum contribution selection versus GA for genetic gain and maintaining genetic diversity

The OCS condition matched to truncation selection's initial rate of coancestry increase shows the same trends in GEBV, coancestry, and ultimate haplotype scores as truncation selection, but on a different (lower) scale. It loses useful diversity much faster than the corresponding truncation selection condition, even while maintaining slightly lower coancestry, so its increase in GEBV before its plateau is approximately half of what truncation selection achieves (Figure 8). Both the OCS condition matched to the GA's initial rate of coancestry increase and the one matched to the GA's initial rate of genetic gain show the same trends in mean GEBV over the generations as the GA, but they have a lower rate of increase and after 100 generations have not caught up to truncation selection's plateaued score (Figure 8a). In addition, the coancestry trends from OCS flatten out more than the GA's trend (Figure 8b). The explanation for their lower improvements in GEBV compared to the GA can be seen in Figure 8c), where it is shown that they lose useful diversity slower than truncation selection but much faster than the GA. By the end of the 100 generations, they have lower haplotype genotype scores than truncation selection, so do not have the potential to surpass the truncation selection GEBV plateau, even if more generations were simulated.

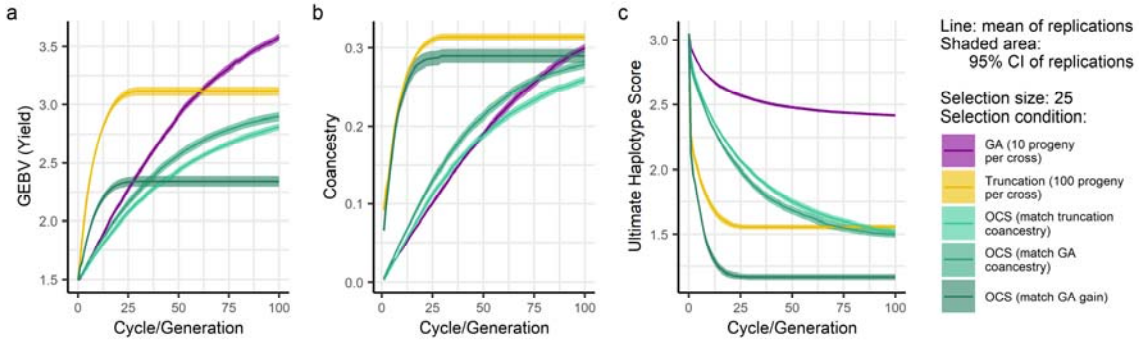

Fig. 8: Trends in simulation results, comparing Kemper et al.'s genetic algorithm (GA), truncation selection, and optimal contribution selection (OCS). (a) Trends in each generation's mean yield GEBV. (b) Trends in each generation's genetic diversity (Nei (1973)'s expected heterozygosity measure). (c) Trends in each generation's ultimate haplotype score. Results are from simulation with a recurrent selection size of 25. Represented in the plots are the following conditions: truncation selection with 100 progeny per cross; selection by GA with 10 progeny per cross; OCS with target angle chosen to match the initial increase in coancestry of the truncation condition; OCS with target angle chosen to match the initial increase in coancestry of the GA condition; and OCS with target angle chosen to match the initial gain in yield GEBV score of the GA condition.

Discussion

Our results demonstrate that targeting individual chromosome segments as a selection method, as implemented by Kemper et al. (2012)'s GA, maintains long term improvements and can provide greater long-term gains than either recurrent truncation selection or OCS, for the simulation conditions tested. The GA benefits most from small populations of parents selected and mated each year, and is frequently outperformed with regard to genetic gain even over 100-generation timescales for larger population sizes.

In our results, an appropriate mate selection strategy (OCS or the segment-stacking GA) can maintain genetic variation in crop breeding programs for many generations, even if the starting point is a narrow pool of elite germplasm. In the situations tested, the GA was able to achieve greater genetic gains than OCS, for OCS constrained to achieve the same coancestry as the GA (e.g. Figure 8a). The measure of coancestry incorporates genetic information, but only represents a single global score capturing the genetic difference of a pair of individuals, without accounting for the individual contribution of different molecular variants to the quantitative trait of interest. Meanwhile, the GA considers which chromosome segments differ between individuals and what their individual effect on the trait of interest is, thereby identifying useful diversity with regards to increasing the genetic merit of the target trait (Figure 8c).

Gorjanc et al. (2016) discuss the problem of 'masking' in pre-breeding programs, where desirable chromosome segments (e.g. chromosome segments with favourable effects on the target trait) are contained in the genome of lines with an overall low GEBV, and so are hidden to single-score selection measures. The GA selection method addresses this masking problem by targeting segments as selection units directly. Even if a candidate genotype contains only one good segment, the GA still has the chance to select it, if that good segment is very high-performing or good segments at that chromosome location are rare in the breeding pool.

A distinguishing feature of the GA in our results was how much genetic diversity was maintained compared to other selection methods. Truncation selection conditions saw a sharp initial decline in diversity, while the trend of GA conditions' genetic diversity measures was a much shallower, near-linear decrease. This may be because the GA's subsets can carry along some mediocre candidate genotypes because they have one or few good segments, genotypes which would not be selected by truncation selection. The other, mediocre, segments of such genotypes will likely be different to the higher-value segments of other genotypes, which will add to measurements of genetic diversity.

Our time horizon in simulation was 100 generations. Although this seems very long, it is worth pointing out with the advent and widespread adoption of speed breeding (Watson et al. 2018), with which up to seven generations of wheat can be turned over each year, 100 generations equates to just 15 years.

The GA has been shown to, even in one generation of application, identify useful recombinants from the offspring pool and ensure that good segments that would otherwise have been lost under truncation selection remain in the population (eg. Figure 5). It could therefore be well-suited to early

generations after the introduction of new diversity to a breeding program, to facilitate combining new and old alleles, especially when new alleles are carried by individuals with comparatively low overall GEBV. It may also perform well in tasks involving breeding for multiple traits (e.g. stacking disease resistances), breeding for a new complex trait, or targeted exploration of gene bank germplasm resources.

The GA's task is to identify a subset of candidates that contain, between them, chromosome segments to construct an optimal "ultimate genotype". There could be several subsets of candidates that construct similar-scoring ultimate genotypes but share no or very few members, if different candidate genotypes carry different combinations of good chromosome segments. This would create a multi-peak landscape for the GA to search. Furthermore, each candidate contributes a different amount to the ultimate genotype score of different subsets based on the segments provided by other genotypes in the subset, which results in noise in the fitness landscape. Among evolutionary algorithms, GAs perform particularly well on complex or noisy fitness landscapes (Mitchell 1998).

As an evolutionary algorithm, a GA iteratively improves its optimal subset guesses to get a close to optimal subset in a reasonable time frame. While the globally optimal subset may be a slightly better set of chromosome segments, distributed through the selection, the genetic improvement of the breeding program, especially in the short- and medium- term, depends more on how many useful chance recombination events that bring good segments together into the same genotype can occur in the allotted time and cost space.

The runtime for simulating selection with the GA was significantly slower than simulating truncation selection or OCS. In the condition with 100 progeny per cross and a recurring selection size of 10 (resulting in an offspring pool of 4500 offspring), one generation of GA selection required 14 minutes compared to the 2 seconds of truncation selection. The GA's execution time is proportional to the size of the pool of candidates. Reducing the GA's running parameters (number of iterations and number of subsets under consideration per iteration) could reduce the disparity, but GA selection will still likely require more computing power than truncation selection.

The relative performance of the GA improved in both the long and short term with higher numbers of progeny per cross. This is not because GA selects many similar genotypes. Full- or half-siblings, no matter how high their GEBVs, are likely to get that GEBV from many of the same high-scoring segments, and therefore the GA will not select them together. Higher numbers of progeny per cross do however mean greater chances of desired recombination between favourable alleles occurring.

Improvements in performance from the GA with more progeny per cross are diminishing, with the first tenfold increase from 1 to 10 progeny per cross improving the GA's performance more than the second tenfold increase from 10 to 100 progeny per cross (Figure 5a). A balance between the GA's performance improvement and the increased genotyping costs and diminishing returns of more offspring per cross must exist. But overall, its performance suggests the GA will be most powerful in species where large numbers of progeny of the same cross can be generated and selected between, and affordable genotyping platforms exist.

Larger selection sizes mean it is possible to retain a wider variety of genotypes with more diverse and useful alleles in the population. Both truncation selection and OCS perform better with larger selection sizes, because they cannot over-cull as strongly and so will maintain more diversity to be turned into genetic gain. However, the genetic gain achieved by the GA over the time horizons tested was relatively higher for smaller populations, for example 10 or 25 parents each generation (Figure 7). This is likely because the GA has an indefinite future horizon. In a bottleneck scenario, its

ultimate genotype's segments must come from a limited number of candidates, which makes it more likely that recombination events will occur that bring those good segments together, and so the GA's observed rate of gain will be higher. Given instead a large selection size, the GA can pick exceptional segments isolated in otherwise low-value genotypes. It then requires more rare recombination events to stack the segments and see genetic gain. This behaviour is visible in Figure 7: the gap between the mean generation-100 GEBV and the score of the ultimate stacked inbred genotype (2 × the ultimate haplotype score) is larger for larger population sizes.

The effect of increasing the number of chromosome segments for the GA is very similar to the effect of increasing the recurrent selection size. Having more chromosome segments allows the GA to be more fine-grained and ambitious in the stacking events required to create the ultimate genotype, and therefore results in very slow rates of improvement of the population. This has been observed in trial simulations using larger numbers of smaller chromosome segments (results not shown).

In this, the GA is quite arbitrary as a selection method. Selection pressure is roughly controlled by the number of genotypes it is set to select, and the number of segments into which the genome is divided. A method for optimising these parameters for particular time frames or desired rates of gain has not yet been developed but could potentially greatly enhance the efficiency of the GA.

The limitations of our simulations are worth pointing out. SNP effects were assumed to be true effects - additive effects that held constant over time. That is, it was assumed markers perfectly tracked the QTL over 100 generations of simulation; the controlling QTL had only additive effects; the positions and order on the chromosome of all markers were perfectly known; and for the purposes of calculating local GEBVs for the GA, phase could be perfectly determined. QTL effects could be simulated, however this has the issue that the number of QTL and the distribution of their effects would have to be assumed. In a real-world situation, the estimates of marker effects could be iteratively updated with phenotype data collected in each new season's trials. Economic equivalences between different conditions and parameter choices were also not considered.

Further, no mutation was simulated, even though beyond 20 generations of artificial selection, the contribution of mutation to genetic variance becomes quite visible (Dudley and Lambert 2004; Barton and Keightley 2002; Keightley 2004). If marker effects that capture the new mutation's value or updated SNP arrays were available, truncation selection and OCS conditions would be able to improve past their plateaus. However, simulating mutation is quite challenging because the distribution of mutation effects is poorly defined. Future studies will investigate the effect of mutation and erosion of linkage disequilibrium between markers and QTL over time.

One of the assumptions built into the simulation tool was that crossovers occur with uniform frequency along the length of each chromosome. This means the effects of linkage disequilibrium on chances of recombination at the desired points were not investigated. The GA also assumes the chance of recombination between segments is uniform across all segments: though recombination in simulation occurred within segments as well as between segments, the performance of the GA when its uniformity-of-recombination assumption was properly challenged was not tested. Building into the GA's objective function some measure of this probability, for example from LD block structures, could improve the validity of its suggestions for real organisms.

Conclusion

The GA has been shown to, even in one generation of application, identify useful recombinants and ensure that some good segments that would otherwise have been lost under truncation selection

remain in the breeding population. It could therefore be well-suited to early generations of an elite breeding program after the introduction of new diversity, in resistance breeding for stacking multiple resistances together, or in pre-breeding programs that explore genetic resources from gene banks. Simulation results have also shown that the segment stacking approach does an excellent job of maintaining the diversity of a breeding population and in the long-term achieves genetic gain comparable to or exceeding in some cases truncation selection and OCS. This study therefore shows this segment stacking approach is a promising avenue for further investigation.

List of References

Barton NH & Keightley PD (2002). Understanding quantitative genetic variation. *Nat. Rev. Genet.* 3: 11–21. https://doi.org/10.1038/nrg700

Bernardo R (2021). Predictive breeding in maize during the last 90 years. *Crop Science*, 61:2872–2881. https://doi.org/10.1002/csc2.20529

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D & Graham G (2014). Predicting the future of plant breeding. Complementing empirical evaluation with genetic prediction. *Crop Pasture Sci 65*(4):311. https://doi.org/10.1071/cp14007

Cowling WA, Castro-Urrea FA, Stefanova KT, Li L, Banks RG, Saradadevi R, Sass O, Kinghorn BP, Siddique KHM (2023). Optimal Contribution Selection Improves the Rate of Genetic Gain in Grain Yield and Yield Stability in Spring Canola in Australia and Canada. *Plants* 12(2):383. https://doi.org/10.3390/plants12020383

Cowling WA, Li L, Siddique KHM, Henryon M, Berg P, Banks RG, Kinghorn BP (2017). Evolving gene banks: improving diverse populations of crop and exotic germplasm with optimal contribution selection, *Journal of Experimental Botany*, (68)8:1927–1939, https://doi.org/10.1093/jxb/erw406

Cowling WA, Li L, Siddique KHM, Banks RG & Kinghorn BP (2019). Modeling crop breeding for global food security during climate change. *Food and Energy Security, 8*(2), e00157. https://doi.org/10.1002/fes3.157

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, & Mathews K (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity (Edinb). 112*(1):48-60. https://doi/10.1038/hdy.2013.16

Daetwyler HD, Hayden MJ, Spangenberg GC, & Hayes BJ (2015). Selection on Optimal Haploid Value Increases Genetic Gain and Preserves More Genetic Diversity Relative to Genomic Selection, *Genetics*, *200*(4):1341–1348, https://doi.org/10.1534/genetics.115.178038

Dekkers JCM & Hospital F (2002). The use of molecular genetics in the improvement of agricultural populations. *Nat. Rev. Genet., 3*:22-32, https://doi.org/10.1038/nrg701

Dudley JW & Lambert RJ (2004). 100 Generations of selection for oil and protein in corn. *Plant Breed. Rev. 24*(1):79-110. https://doi.org/10.1002/9780470650240.ch5

García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, & Tassell CPV (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences*, 113(28):E3995-E4004. https://doi.org/10.1073/pnas.1519061113

Guinan FL, Wiggans GR, Norman HD, Dürr JW, Cole JB, Van Tassell CP, Misztal I, & Lourenco D (2023). Changes in genetic trends in US dairy cattle since the implementation of genomic selection. *Journal of Dairy Science, 106*(2):1110-1129, https://doi.org/10.3168/jds.2022-22205

Gorjanc G, Gaynor RC, & Hickey JM (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics, 131*(9):1953-1966. https://doi.org/10.1007/s00122-018-3125-3

Gorjanc G & Hickey JM (2018). AlphaMate: a program for optimising selection, maintenance of diversity, and mate allocation in breeding programs. *Bioinformatics*, 34(19):3408-3411, https://doi.org/10.1093/bioinformatics/bty375

Gorjanc G, Jenko J, Hearne SJ, & Hickey JM (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics, 17*(1):30. https://doi.org/10.1186/s12864-015-2345-z

Hickey J, Chiurugwi T, Mackay I, Powell W & Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat Genet* **49**:1297–1303. https://doi.org/10.1038/ng.3920

Hospital F (2003). Marker-assisted breeding. H.J. Newbury (Ed.), Plant Molecular Breeding, Blackwell Scientific Publishers, London, pp. 30-56.

Isik F, Holland J, & Maltecca C (2017). Genomic Relationships and GBLUP. *Genetic Data Analysis for Plant and Animal Breeding*. Springer, Cham. https://doi.org/10.1007/978-3-319-55177-7_11

Keightley PD (2004). Mutational variation and long-term selection response. *Plant Breed. Rev. 24*(1):227-247. https://doi.org/10.1002/9780470650240.ch10

Kemper KE, Bowman PJ, Pryce JE, Hayes BJ, & Goddard ME (2012). Long-term selection strategies for complex traits using high-density genetic markers. *Journal of Dairy Science, 95*(8):4646-4656. https://doi.org/10.3168/jds.2011-5289

Kerr RJ, McRae TA, Dutkowski GW, & Tier B (2015). Managing the rate of increase in average co-ancestry in a rolling front tree breeding strategy. *Journal of Animal Breeding and Genetics, 132*(2):109-120. https://doi.org/10.1111/jbg.12157

Kinghorn BP (2011). Automated Mate Selection Analyses. *Proc. Assoc. Advmt. Anim. Breed. Genet., 19*:15-22.

Lin Z, Cogan NOI, Pembleton LW, Spangenberg GC, Forster JW, Hayes BJ, & Daetwyler H D (2016). Genetic Gain and Inbreeding from Genomic Selection in a Simulated Commercial Breeding Program for Perennial Ryegrass. *The Plant Genome, 9*(1), https://doi.org/10.3835/plantgenome2015.06.0046

Meuwissen TH (1997). Maximizing the response of selection with a predefined rate of inbreeding. *J Anim Sci., 75*(4):934-40. https://doi.org/10.2527/1997.754934x

Meuwissen TH, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157*(4):1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Mitchell M (1998). *An introduction to genetic algorithms* (1st MIT Press paperback ed.). MIT Press.

Nei M (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA 70:3321–3323.* https://doi.org/10.1073/pnas.70.12.3321

Saradadevi R, Mukankusi C, Li L, Amongi W, Mbiu JP, Raatz B, Ariza D, Beebe S, Varshney RK, Huttner E, Kinghorn B, Banks R, Rubyogo JC, Siddique KHM, & Cowling WA (2021). Multivariate genomic analysis and optimal contributions selection predicts high genetic gains in cooking time, iron, zinc, and grain yield in common beans in East Africa. *Plant Genome*, 14:e20156. https://doi.org/10.1002/tpg2.20156

Servin B, Martin OC, Mézard M, Hospital F (2004). Toward a theory of marker-assisted gene pyramiding. *Genetics, 168*(1):513-23. https://doi.org/10.1534/genetics.103.023358

Scott BA, Haile-Mariam M, Cocks BG, & Pryce JE (2021). How genomic selection has increased rates of genetic gain and inbreeding in the Australian national herd, genomic information nucleus, and bulls. *Journal of Dairy Science 104*(11):11832-11849, https://doi.org/10.3168/jds.2021-20326

Voss-Fels KP, Cooper M, Hayes BJ (2019). Accelerating crop genetic gains with genomic selection. *Theor Appl Genet. 132*(3):669-686. https://doi.org/10.1007/s00122-018-3270-8

Wolc A, Zhao HH, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Stricker C, Habier D, Fernando RL, Garrick DJ, Lamont SJ, and Dekkers JCM (2015). Response and inbreeding from a genomic selection experiment in layer chickens. *Genet Sel Evol* **47**:59. https://doi.org/10.1186/s12711-015-0133-5

Woolliams JA, Berg P, Dagnachew BS, & Meuwissen THE (2015). Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics, 132*(2):89-99. https://doi.org/10.1111/jbg.12148

Wray NR & Goddard ME (1994). Increasing long-term response to selection. *Genetics, Selection, Evolution : GSE, 26*:431-451, https://doi.org/10.1186/1297-9686-26-5-431

Statements and Declarations

Funding

authors are grateful for funding from the Australian Research Council, Linkage Project LP170100317, “Fastack - evolutionary computing to stack desirable alleles in wheat”.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. The first draft of the manuscript was written by KV and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability

The simulated datasets generated during the current study are available from the corresponding author on reasonable request.