Simulations of rate of genetic gain in dry bean breeding programs

Jennifer Lin
McGill University
Vivi Arief
The University of Queensland
Zulfi Jahufer
The University of Queensland
Juan Osorno
North Dakota State University
Phil McClean
North Dakota State University
Diego Jarquin
University of Florida
Valerio Hoyos-Villegas (valerio.hoyos-villegas@mcgill.ca)
McGill University https://orcid.org/0000-0003-1080-9148

Research Article

Keywords: optimal breeding scenario, breeding framework, selection strategy, selection efficiency, mass selection, pedigree, single seed descent, common bean, genetic gain, favorable allele fixation

Posted Date: March 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1442864/v1

Version of Record: A version of this preprint was published at Theoretical and Applied Genetics on January 20th, 2023. See the published version at https://doi.org/10.1007/s00122-023-04244-x.

Abstract

Dry beans (*Phaseolus vulgaris* L.) are a nutrient dense legume that is consumed by developed and developing nations around the world. The progress to improve this crop has been quite steady. However, with the continued rise in global populations, there are demands to expedite genetic gains. Plant breeders have been at the forefront at increasing yields in the common bean. As breeding programs are both time consuming and resource intensive, resource allocation must be carefully considered. To assist plant breeders, computer simulations can provide useful information that may then be applied to the real world. This study evaluated multiple breeding scenarios in the common bean and involved five breeding strategies, three breeding frameworks, and four different parental population sizes. In addition, the breeding scenarios were implemented in three different traits: days to flowering, white mold tolerance, and seed yield. Results from the study reflect the complexity of breeding programs, with the optimal breeding scenario varying based on trait being selected. Relative genetic gains per cycle of up to 8.69% for seed yield could be obtained under the use of the optimal breeding scenario. Principal component analyses revealed similarity between strategies, where single seed descent and the modified pedigree method would often aggregate. As well, clusters in the direction of the Hamming distance eigenvector are a good indicator of poor performance in a strategy.

Key Message

A reference study for breeders aiming at maximizing genetic gain in common bean. Depending on trait heritability and genetic architecture, conventional approaches may provide an advantage over other frameworks.

Introduction

With increasing global populations and the current implications of climate change, meeting demands for food security while instilling sustainable practices is imperative. In addition to providing high quality nutrients for both human and animal consumption, legumes are remarkably sustainable to grow. They can reduce greenhouse gas emissions and can improve soil fertility by increasing carbon and nitrogen content and availability (Stagnari et al. 2017). Dry beans are an important legume crop grown in many developing countries that greatly contribute to the energy and nutritional intake in low-income regions (Siddiq and Uebersax 2012; Stagnari et al. 2017). Rich in proteins, carbohydrates, fibers, vitamins, and minerals, dry beans offer health benefits that are unrivaled. Research has shown that dry beans contain soluble fibers that can lower serum cholesterol, which in turn improves coronary health. Dry beans are also excellent for metabolic control. They lead to miniscule increases in blood glucose and insulin, making them highly suitable for diabetic individuals. Due to the nutritional quality of dry beans, they may be also used to combat obesity (Geil and Anderson 1994).

Increasing dry bean yield is of importance for both developed and developing countries that rely on this legume. The main constraints to increasing yield are biotic and abiotic stresses. Breeding for tolerance to

drought stress, heat stress, cold stress, and low nutrient stress is important in particularly in areas with harsher growing conditions. Meanwhile, for biotic stresses, dry beans are susceptible to several diseases that can severely limit yield. In temperate growing regions, the most common diseases include common bacterial blight, halo blight, rust, and white mold. Some breeders are also interested in agronomic traits that may improve yield. For example, selecting for upright plant architecture can facilitate harvest and reduce vulnerability to disease, which can indirectly benefit yield (Soltani et al. 2016). When it comes to dry bean breeding, the market class must be taken into consideration. For certain market classes, enhancing yield may be difficult due to the yield component compensation, where some yield components are negatively correlated with each other (Adams 1967). In general, plant breeders will develop strategies that are applicable to their growing region and market class of choice. Traditionally, dry bean breeders have used early generation testing and visual selection to improve yield. However, these strategies have their limitations, namely in that yield testing is extremely costly and laborious. Thus, it may be worthwhile to delay yield testing until later generations (Kelly et al. 1998). Other traits of interest for improvement include those that are consumer driven. In developing countries, faster cooking time is desired since fuel is often in short supply. To fight malnutrition in low-income areas, breeding programs may focus on improving nutrient content, such as zinc and iron. In developed countries, canning quality is an important trait for improvement (Beaver and Osorno 2009).

An important goal in plant breeding is achieving genetic gain (ΔG), which is the rate of change in the mean of a trait being selected for in a population (Falconer, 1960; Moose and Mumm, 2008; Sun et al., 2011). The equation for genetic gain is as follows:

$$\Delta G = \frac{h^2 i \sigma_a}{L} \text{ [1]}$$

Where, h^2 refers to the narrow sense heritability, σ_a is the additive variance, *i* is the selection intensity, and *L* is the generation interval (Sun et al., 2011). Eq. 1 highlights the different factors that contribute to genetic gain. In the simulations however, genetic gain is calculated using Eq. 5, which is described later on. Typical dry bean breeding programs take up to 10 years and require extensive resources in the process. Due to the long-term commitment, the decisions that go into a breeding program must be carefully considered. For this purpose, plant breeders can make use of computer simulations to assist in decision making.

Due to the complexity of breeding programs, the breeder's equation is used as a basis for which the simulation studies are conducted. The data obtained from studies may be used to help breeders decide where emphasis should be placed when designing a breeding program. The goal of any breeding program is to maximize genetic gain in the shortest amount of time. The heritability of a trait will impact a breeding program. Traits with a higher heritability can result in greater genetic gain. The selection intensity will also impact the genetic gain.

The overall aim of this study was to explore the feasibility of driving genetic gain forward in the presence of multiple reported QTL segregating in the population. The specific objectives of this study were to: 1)

Simulate a baseline of selection strategies and determine differences in genetic gain performance; 2) Test whether changes in initial parental population size and trait heritability lead to increased genetic gain and percentage of fixed favourable alleles.; 3) Test if genomic selection and speed breeding outperform conventional breeding frameworks in terms of long-term genetic gain and allele fixation rate.

Materials And Methods

We used the QuLinePlus (Hoyos-Villegas et al., 2019) module in the QU-GENE platform to simulate the outcomes of different breeding frameworks, selection strategies and initial parental population sizes on three traits of varying heritability. The focus of this paper will be on stochastic modeling of genetic gain across two agronomic traits (yield and days to flowering) and a biotic stress (white mold tolerance). The three traits were chosen on known differing heritability levels examined with seed yield as low h^2, days to flowering as moderate h^2, and white mold tolerance as high h^2. Three breeding frameworks, Five selection strategies and three initial parental population sizes were considered. We provide a short description of each breeding scenario (selection strategies and frameworks) and their implementation in QuLinePlus.

Selection strategies

Several selection strategies for self pollinated crops were tested. Conventional selection strategies included bulk breeding, single seed descent, mass selection, the pedigree method, and the modified pedigree method. These conventional strategies relied solely on phenotypic selection.

Mass selection

Mass selection is the oldest form of crop improvement and was carried out by farmers long before the concepts of Mendelian genetics and the development of pure-lines were commonplace (Fehr, 1987). In mass selection, desirable plants are selected from an entire population and a sample of the seeds collected form the next generation of plants. This process is repeated for several generations until the multi-environment trial phase (Fig. 1a). The key purpose of mass selection is to improve the average of the baseline population (Acquaah 2009). However, this improvement is typically constrained by the genetic variability of the initial population. Mass selection may be used to develop varieties from a hybridized population. In this approach, undesirable plants are removed from the population. In some cases, mass selection is performed to purify lines. When deciding to use mass selection, the trait heritability should be considered, as high heritability traits are much more successful (Fehr, 1987).

Bulk breeding

Bulk breeding is a strategy that relies on natural selection in early generations to remove low performing genotypes (Fehr, 1987). Artificial selection is only conducted in later generations once a high amount of homozygosity is present in the F_2 derived lines. The process begins with the crossing of two parents and continues with the bulking of each segregating generation. Once sufficient homozygosity has been achieved, the plants will be assessed and those with the desired trait will be selected. Following this,

multi-environment testing will take place, and superior lines will be identified (Fig. 1b). One of the major criticisms of bulk breeding is that it promotes competition between genotypes, so there is a possibility that a desirable genotype is outcompeted by an undesirable genotype. Another concern is that some traits that persist due to natural selection have no agricultural benefit. Nevertheless, bulk breeding is still less labour intensive and cheaper than some other strategies and it allows plant breeders to make and assess more crosses (Acquaah 2009).

Single seed descent

Single seed descent is a method that attempts to achieve homozygosity in the shortest amount of time (Acquaah 2009). The objective is to advance as many F_2 plants as possible to the F_5 generation. This is done by taking one random seed from each plant to advance to the next generation until yield trials (Fig. 1c). Not only does this method require fewer resources, but it is also possible to advance multiple generations in a single year by using greenhouses and winter nurseries. Selection only takes place in later generations once adequate homozygosity is reached. Unlike bulk breeding, earlier generations do not undergo natural selection and each F_2 plant is equally represented, meaning each generation has more genetic diversity. The main disadvantage is that not every seed will germinate or be sampled from the population. Thus, not all F_2 individuals will be represented in the later generations.

Pedigree method

The pedigree method is a strategy whereby pedigree relationships are carefully recorded; thus, any individual plant can be easily traced back to an F_2 plant. The pedigree method differs from the previous methods in that artificial selection takes place in segregating populations. Selection occurs at each generation and begins at the F_2 generation. Individual F_2 plants that were selected are grown in rows, forming the F_3 generation. Each row can also be referred as a family. Individual plants within rows or even entire rows may be selected (Fig. 1d). This continues until there is an acceptable level of homozygosity (Fehr, 1987). A benefit to using the pedigree method is that thorough pedigree, and other valuable genetic information is now available to plant breeders through plant breeding software. Furthermore, the records may be used to better select lines that carry a desirable trait. The main concern with the pedigree method is that it is resource demanding. Record-keeping is time-consuming and progeny rows can take up lots of space (Acquaah 2009).

Modified Pedigree method

The modified pedigree is a method that takes into consideration the importance of inbreeding before making selections. This is because genetic variance will increase between lines, but decrease within lines (Brim 1966). Individual plant and row selections take place in the F_2 and F_4 generations, where plants are grown in their target growing region. This strategy also makes use of winter nurseries in the F_3 and F_5 generation, where selected lines are harvested in bulk (Fig. 1e). In short, the use of winter nurseries in the modified pedigree method saves time and resources, as harvesting plants in bulk is easier to manage. Meanwhile, it simultaneously allows plants to achieve homozygosity in less time (Acquaah 2009). This

method has most recently been used for breeding a rust resistant variety of black bean (Osorno et al. 2021).

Breeding Frameworks

In recent years, new proposed breeding frameworks have emerged, namely, speed breeding (Watson et al., 2018) and genomic selection (Meuwissen et al., 2001). Speed breeding can circumvent the developmental constraints in plants, thus reducing the total length of a breeding program and subsequently allowing for greater genetic gains per year. Genomic selection uses models that predict phenotypes based on all markers across a genome to select on genotypes. This allows for selection to take place before at the seedling stage. Using genomic selection, a plant breeder may genotype and cull poor performing lines, saving time and resources otherwise required to phenotype each line.

Speed Breeding

Speed breeding is a technique used to increase the rate of development in crops and as a result, decrease generation times (Watson et al., 2018). Methods in speed breeding typically involve lengthening the photoperiod, with 22 hours of light and 2 hours of dark. Speed breeding has been successfully implemented in a number of crop species, including wheat, barley, chickpea, canola, and pea (Watson et al., 2018). In dry beans, speed breeding may be used to advance plants from the crossing block to the F_4 generation in a single year, significantly cutting down the duration of a breeding program (Larsen et al. 2019). Speed breeding in dry beans can reduce the cycle time from 8 generations to 5 generations. Thus, the cycle length in speed breeding was reduced by 1.6 times. This was modification was done post-run using R.

Genomic selection

First described by Meuwissen et al., (2001), genomic selection involves estimating the effects of all molecular markers (e.g. single nucleotide polymorphisms, SNPs) and selecting on individuals based on their genomic estimated breeding value (GEBV) (Michel et al., 2016). To implement genomic selection, a training population was first generated and the genotypic and phenotypic information of each individual was combined with the training population. A prediction model was trained, validated, and then applied to a breeding population (Taylor, 2014). The model was then used to predict a GEBV for each individual in the testing population (Crossa et al., 2017). In QuLinePlus, individuals that had the same rare marker haplotype shared a common ancestor and had the same QTL allele. Markers were assumed at each QTL in LD with at least one marker (Goddard and Hayes, 2007; Nadeem et al., 2018).

QU-GENE simulation workflow and simulation files

Simulation was conducted to compare four different combinations of initial parents, three different traits, three breeding frameworks, and five selection strategies. Each simulation consisted of 10 cycles with 50 runs/cycle. A summary of the simulation criteria is displayed in Table 1. All the files required by the simulation can be found on the GitHub page (https://github.com/McGillHaricots/peas-

andlove/tree/master/Simulation-files). Figure 2 shows the workflow in QU-GENE for the simulation of conventional breeding, as well as the new proposed breeding frameworks, which require additional steps.

Table 1
Simulation criteria

Cycles	Runs	Parents	Traits	Environments	Framework	Strategies
10	50	15, 30, 60, 100	DF, WM, SY	Nursery, Winter Nursery, Field	Conventional, Speed breeding, Genomic selection	Mass selection, Bulk breeding, Single seed descent, Pedigree method, Modified pedigree method

Briefly, the file required by the QU-GENE engine is the .qug file, which contained the following: traits, environments, error variances, linkage map, QTLs, markers, populations, and diagnostics. In terms of traits, the three simulated traits were days to flowering, white mold tolerance, and seed yield. The simulation also involved three environments: nursery, winter nursery, and field. The error variances were based on within error variances and were calculated from narrow-sense heritabilities reported for each trait from literature sources. Heritability estimates obtained for each trait in each environment are summarized in Table 2. The linkage map, QTL, and markers described in a previous section were included in the .qug file. The diagnostic indicated that the file was error free and was able to be run in the QU-GENE engine.

Table 2
Narrow-sense heritability (h2) estimates for three traits in three environments.

Trait	Environment	h^2 estimate	Reference
DF	Nursery	0.67	(Singh et al., 1990)
	Winter nursery	0.68	(Nienhuis and Singh, 1988)
	Field	0.92	(Atuahene-Amankwa et al., 2004)
WM	Nursery	0.33	(Carneiro et al., 2011)
	Winter nursery	0.65	(Carvalho et al., 2013)
	Field	0.78	(Miklas et al., 2001)
SY	Nursery	0.21	(White and Singh, 1991)
	Winter nursery	0.29	(Mendes et al., 2008)
	Field	0.70	(Kolkman and Kelly, 2002)

DF: days to flowering (in days); WM: white mold tolerance (in disease incidence); SY: seed yield (in kg/hectare)

Since QU-GENE simulates error variances based on the per plant broad-sense heritability, it was necessary to calculate these values based on the per plot heritability estimates reported in the literature. While QU-GENE uses broad-sense heritability, narrow-sense heritability estimates were provided to QU-GENE, as only additive effects were considered among the QTL. The following equation was used to determine the per plant broad-sense heritability:

$$H^2_{perplant} = \frac{V_g}{V_g + \left(\frac{V_e}{y} \times n\right)} \quad [2]$$

where V*a* is the additive variance, the phenotypic variance $V_p = V_g + \frac{V_e}{y} \times n = \frac{1}{h^2_{perplot}}$, the error variance $V_e = V_p - V_g$, *n* is the plot size (number of plants), and *y* is the year.

The .qmp file included information on the selection strategies to be simulated. For each strategy, one cycle consisted of 8 generations, with selection occurring at different stages. As a closed system was being simulated, initial and final family sizes were the same. It included general information such as the number of strategies, the number of runs, and the number of cycles that were completed. It also included information specific to each selection strategy such as propagation type, generation advance method, number of replications, plot size, number of testing locations, and selection strategy (e.g. within family, among family, etc). The propagation type indicated how the selected individuals from the previous generation were to be propagated to generate the individuals in the current generation. This experiment only considered “self” (self-pollination) and “clone” (asexual) as the propagation type. Although *Phaseolus vulgaris* L. does not have an asexual reproduction route, the term “clone” was used in

QuLinePlus to migrate individuals from one breeding cycle to the next without changing their genetic composition. The term "clone" was also used to extract individuals sampled from the parent population into the first generation of the first breeding cycle. The generation advance method indicated how the selected plants were harvested. This experiment used the following generation advances: “pedigree”, “bulk”, and “superbulk”. “pedigree” meant plants were harvested individually, and each plant would result in a family in the next generation. “bulk” involved harvesting all plants in a family together, with no mixing of families. Finally, in “superbulk”, all plants were harvested to form one population regardless of family. Details for each strategy are shown in Supplementary Table 1.

The QuLinePlus module was used to simulate the selection strategies. It is capable of simulating both self-pollinating and cross-pollinating species (Hoyos-Villegas et al., 2019). Output files obtained from the QU-GENE engine were used as input files for QuLinePlus. As the simulations were performed remotely on servers provided by compute Canada (https://ccdb.computecanada.ca/). Access to remote servers required establishing a secure shell via the terminal on MacOS. To browse and manipulate files, the cloud storage browser, cyberduck (https://cyberduck.io/) was used.

Linkage map and QTLs

The common bean consensus linkage map reported by Galeano et al. (2011) was used for this study. It was developed from the recombinant inbred lines from three different Mesoamerican intra-gene pool linkage mapping populations. The consensus linkage map was made up of 1010 markers and had a map length of 2041 cM over 11 linkage groups. Each linkage group had an average of 91 markers. Since more markers could be identified through the combined from multiple segregating populations than can be obtained from a single population, and greater coverage can be achieved, this consensus map was selected for conducting the simulations. In contrast to a physical map, a genetic consensus map was used to provide QU-GENE with recombination fraction information for simulation.

A total of 38 QTL found in the literature were considered for this study. Specifically, 11 seed yield QTL, 8 white mold disease incidence QTL, and 19 days to flowering QTL were selected (Table 3). Seed yield QTL effect sizes ranged from − 36.91 to -197.46 kg/ha. QTL effect sizes for white mold disease QTL ranged from 3.16 to -7.2. QTL effect sizes in days to flowering ranged from 0.68 to -1.21 days. The reported QTL effect sizes were the additive genetic effects that could be attributed to having one of the alleles. In the simulation, it was assumed that having the alterative allele would lead to an equal but opposite effect. That is, if at locus A, the possible genotypes were AA, Aa, and aa, and allele A had an effect size of s, then it was assumed that AA would have effect size $2s$, Aa would have effect size 0, and aa would have effect size $-2s$.

Table 3
Description of QTLs used in the simulation

Trait	QTL name	Linkage group	Position (cM)	Effect size	Mapping population	Reference
Days to Flowering (Days)	DF41	4	167.11	0.68	DOR 364 × BAT 477	(Diaz et al., 2017)
	DF51	5	45.21	0.45	DOR 364 × BAT 477	(Diaz et al., 2017)
	DF52	5	56.71	0.49	DOR 364 × BAT 477	(Diaz et al., 2017)
	DF53	5	82.21	0.46	DOR 364 × BAT 477	(Diaz et al., 2017)
	DF54	5	105.21	0.43	DOR 364 × BAT 477	(Diaz et al., 2017)
	DF11a	11	96.51	-0.6	DOR 364 × BAT 477	(Diaz et al., 2017)
	DF11b	11	108.71	-0.49	DOR 364 × BAT 477	(Diaz et al., 2017)
	EM86	2	21.6	0.57	Bunsi × Newport	(Ender and Kelly, 2005)
	EM78	7	1.1	-0.6	Bunsi × Newport	(Ender and Kelly, 2005)
	EM550	7	13.6	-0.96	Bunsi × Newport	(Ender and Kelly, 2005)
	EM223	7	8.6	-1.21	Bunsi × Newport	(Ender and Kelly, 2005)
	DF121	1	51	0.02	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	DF122	1	62	-0.69	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	DF111	1	47	-0.62	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	DF13	1	19	0.12	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	DF112	1	40	0.03	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	DF123	1	59	-0.66	SER48 × Merlot	(Hoyos-Villegas et al., 2016)

Trait	QTL name	Linkage group	Position (cM)	Effect size	Mapping population	Reference
	DFmn1	1	16.9	-0.8	AN-37 × P02630	(Hoyos-Villegas et al., 2015)
	DFmn2	1	105.7	-0.8	AN-37 × P02630	(Hoyos-Villegas et al., 2015)
White Mold Severity (1–9 Score)	WM2010	3	91.5	-7.2	AN-37 × P02630	(Hoyos-Villegas et al., 2015)
	WM31	3	111.1	-4	AN-37 × P02630	(Hoyos-Villegas et al., 2015)
	DSI1	2	8	3.15	Bunsi × Newport	(Ender and Kelly, 2005)
	DSI2	2	21	-2.66	Bunsi × Newport	(Ender and Kelly, 2005)
	DSI3	5	27.7	3.16	Bunsi × Newport	(Ender and Kelly, 2005)
	DSI4	7	8.6	-4.17	Bunsi × Newport	(Ender and Kelly, 2005)
	DSI5	7	14.8	-4.01	Bunsi × Newport	(Ender and Kelly, 2005)
	DSI6	8	1.4	2.93	Bunsi × Newport	(Ender and Kelly, 2005)
Seed Yield (kg/ha)	Yd21	2	151.2	-46.88	DOR 364 × BAT 477	(Diaz et al., 2017)
	Yd71	7	35.1	-36.91	DOR 364 × BAT 477	(Diaz et al., 2017)
	Yd72	7	47.8	-97.3	DOR 364 × BAT 477	(Diaz et al., 2017)
	syMO14	3	113.7	-153.6	BK004-001 × H68-4	(Sandhu et al., 2018)
	syMO16a	7	10.6	-170.9	BK004-001 × H68-4	(Sandhu et al., 2018)
	syMO16b	8	0.5	-140.2	BK004-001 × H68-4	(Sandhu et al., 2018)
	SY10v1	10	41	-178.77	SER48 × Merlot	(Hoyos-Villegas et al., 2016)

Trait	QTL name	Linkage group	Position (cM)	Effect size	Mapping population	Reference
	SY3v3	3	53	-155.91	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	SY7v3	7	51	-197.46	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	SY7v4a	7	68	-178.85	SER48 × Merlot	(Hoyos-Villegas et al., 2016)
	SY7v4b	7	67	-97.54	SER48 × Merlot	(Hoyos-Villegas et al., 2016)

Model for genomic selection

The model used to determine the marker effects in genomic selection is shown in Eq. 3:

$y = X\beta + Zu + \epsilon$ [3]

where $\mathrm{u}\tilde{N}(0, K\sigma^2{}_u)$, y is the phenotypic value of a trait, X is the design matrix for the fixed effects β, **Z** is the design matrix for random effects u, and ϵ is the residual error. The R package 'rrBLUP' (Endelman, 2011) using the function mixed.solve was used to calculate the marker effects, or fixed effects β. The calculated marker effects were then input into the .qug file as locus effects. The training population consisted of the parental populations that were generated via SimuPop (Peng and Kimmel 2005). Thus, the size of the training population was 15, 30, 60, and 100, corresponding to the different parental population sizes for the different simulations.

Simulating Linkage Disequilibrium

Linkage disequilibrium (LD) can be defined as a non-random association between alleles found at different loci (Flint-Garcia et al. 2003). By default, QU-GENE will generate populations in Hardy-Weinberg equilibrium with little to no LD. This is an issue for simulating genomic selection since adequate LD is necessary for markers to be linked to QTL. Two methods for estimating LD make use of the parameters D' and r^2 (Oraguzie et al. 2007). For verifying LD measures, the r^2 parameter was used. For two loci, with alleles A and a at the first loci and allele B and b at the second loci, the allele frequencies can be expressed as P_A, P_a, P_B, and P_b, respectively. The resulting haplotype or allele pair will be AB, Ab, aB, and aB, with the respective haplotype frequencies, P_{AB}, P_{Ab}, P_{aB}, and P_{AB}. The difference between the haplotype frequencies that are observed and the frequencies that are expected can be written as:

$D_{AB} = P_{AB} - P_A P_B$ [4]

This difference is also known as the coefficient of linkage disequilibrium and is important for calculating D' and r^2. r^2 can be expressed as follows:

$r^2 = \frac{(D_{AB})^2}{P_A P_B P_a P_b}$ [5]

There are several factors that are responsible for the LD found in a population. Mutations create the polymorphisms that will be in LD. The reduction of intrachromosomal LD can be attributed to recombination. Meanwhile, independent assortment is the main cause for the breakdown of interchromosomal LD. Furthermore, population size can greatly influence LD. Small populations are subject to more genetic drift, which results in the fixation of alleles. The resulting loss of rare combinations of alleles will increase LD. Mating systems in a population can also impact LD. Selfing populations are less affected by recombination, since individuals are typically homozygous. As a result, species that undergo outcrossing generally experience a faster decay in LD compared to selfing species. LD can be generated from admixed populations, where genetically distinct populations intermate. In populations that undergo random mating, LD will decrease rapidly. Another factor that can influence LD is the drastic fall in population size or a bottleneck event, which results in genetic drift and consequently an increase in LD. Selection can also increase LD between the selected locus and any loci linked to it (Flint-Garcia et al. 2003).

To generate LD in our simulated populations, the forward-in-time simulation tool SimuPOP was used. SimuPOP is implemented in python. The program can be used to evolve a population over time *in silico*. By allowing a population to undergo natural selection via SimuPOP, populations with substantial LD could be obtained. The population generated from SimuPOP was converted to the QU-GENE format via R. Analysis of LD in the population was also performed in R, using the LD.Measures() function in the package 'LDcorSV' and an LD heatmap was generated using the function LDheatmaps() in the package LDheatmap. The initial and final LD patterns simulated can be found in Supplementary Fig. 1.

Handling simulation output data

QuLinePlus produces several output files that can be used to estimate the genetic gain, fixation of favourable alleles, Hamming distance, genetic variance, and effective population size. The .fit file reports the adjusted genotypic or fitness values for the population after each cycle. This is calculated using Eq. 5:

$F_{Ad} = \frac{F - TG_l}{TG_h - TG_l} \times 100$ [5]

where *F* is the fitness, TG_h is the highest target genotypic value, and TG_l is the lowest target genotypic value. The adjusted genetic gain can then be calculated as the difference from one cycle to the next, as shown in Eq. 6:

$\Delta G_{Ad} = F_{Ad(n)} - F_{Ad(n-1)}$ [6]

where ΔG_{AD} is the adjusted genetic gain, $F_{AD(n)}$ is the adjusted fitness value after n cycles and $F_{AD(n-1)}$ is the adjusted fitness value after n-1 cycles. The *.fix* file reports the percentage of fixed favourable and unfavourable alleles after each cycle. This can be used to determine the allele fixation rate. The *.ham* file reports the Hamming distance of the population after each cycle. In information theory, Hamming

distance is used as a measure of dissimilarity between two strings of the same length (Li et al. 2012; Wang et al. 2015). When applied to breeding programs for assessing individuals, the Hamming distance refers to the number of alleles that differ from the target genotype for all loci. A smaller Hamming distance would indicate an individual is closer to the target or ideal genotype, thus a lower value for the Hamming distance is more desirable. The *.var* file reports the additive genetic variance after each cycle. The reported values were converted to relative percentages where cycle 0 was used as a baseline and set to 100%. This parameter was used to assess the amount of genetic diversity in the population. The R packages 'dplyr' and 'ggplot2' were used to subset the data and generate plots.

Finally, a principal component analysis (PCA) was generated for each strategy to compare the following factors: genetic gain, Hamming distance, fixation of favourable alleles, genetic variance, and effective population size. The PCA plots were created using the '**ggplot2'** package in R.

Results

Genetic variance

The selection strategies and frameworks were first compared in terms of changes to genetic variance for the three simulated traits, days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Genetic variance was represented as a relative percentage, with cycle 0 defined as 100%. Differing numbers of initial parents were also compared for each trait. The analysis of variance (ANOVA) for additive genetic variance revealed that the strategy, framework, and number of parents were all statistically significant. As expected, relative genetic variance decreased over the five cycles. For days to flowering, as the number of initial parents increased, less relative genetic variance was maintained (Supplemental Fig. 2). Similar trends were observed for white mold tolerance (Supplemental Fig. 3) and seed yield (Supplemental Fig. 4). Genomic selection led to equal or greater genetic variance being maintained when compared to conventional breeding. Meanwhile, speed breeding resulted in lower genetic variance maintained compared to both conventional breeding and genomic selection. Interestingly, the use of genomic selection for seed yield resulted in maintenance of more genetic variance under the mass selection strategy, when compared to conventional breeding. For days to flowering, bulk breeding maintained the greatest amount of genetic variance for most scenarios. With 30 initial parents under genomic selection, the modified pedigree method maintained the most genetic variance. With 60 initial parents under genomic selection, mass selection maintained the most genetic variance.

For white mold tolerance, bulk breeding led to the greatest genetic variance maintained when the parental population size was 15. For parental population sizes of 30, 60, and 100, mass selection resulted in the most genetic variance maintained.

For seed yield, mass selection resulted in the most genetic variance being maintained for most scenarios. With 15 initial parents under conventional and speed breeding, bulk breeding led to the greatest genetic variance maintained.

Fixation of favourable alleles and Hamming distance

The fixation of favourable alleles was plotted over 10 cycles. Figures 3, 4 and 5 display the plots for the fixation of favourable alleles in days to flowering, white mold tolerance, and seed yield, respectively. For days to flowering (Fig. 3), as the parental population size increased, a lower percentage of alleles were fixed. Across all scenarios, the pedigree method had the fastest allele fixation rate. Mass selection had the slowest allele fixation rate and resulted in the fewest alleles being fixed. The scenario resulting in the greatest percentage of fixed alleles was single seed descent under genomic selection with 15 parents, where 93.68% of favourable alleles were fixed.

For white mold tolerance, multiple scenarios led to 100% of favourable alleles being fixed (Fig. 4). In general, as parental population size increased, a higher percentage of alleles were fixed. Under genomic selection with 100 initial parents, the pedigree method allowed for 100% of favourable alleles to be fixed in only 2 cycles. This scenario led to the greatest percentage of fixed alleles in the fewest cycles. Across all scenarios, the pedigree method had the fastest allele fixation rate.

For seed yield, a parental population size of 15 resulted in the greatest fixation of alleles (Fig. 5). The scenario resulting in the highest percentage of fixed favourable alleles was single seed descent under speed breeding with 15 initial parents, where 98.91% of alleles were fixed.

The plots for average Hamming distance are displayed in Supplemental Figs. 5, 6 and 7. Overall, Hamming distance had a general decreasing trend which eventually plateaued. For days to flowering (Supplemental Fig. 5), Hamming distance was higher in scenarios with larger parental population sizes, particularly for 60 and 100 parents. Across all scenarios, mass selection had the highest Hamming distance. This was especially pronounced under genomic selection when 30 and 100 parents were simulated. Conventional breeding, speed breeding, and genomic selection were all comparable, with minor differences. Under conventional and speed breeding, bulk breeding and single seed descent resulted in the lowest Hamming distance. Under genomic selection, the optimal strategy for Hamming distance depended on the parental population size. Bulk breeding, single seed descent, pedigree method, and modified pedigree method led to the smallest Hamming distance for the parental population sizes 15, 30, 60, and 100, respectively.

For white mold tolerance (Supplemental Fig. 6), larger parental population sizes produced smaller Hamming distances in the selected individuals. In addition, differences between the strategies were only observed with fewer initial parents. Across all scenarios, mass selection resulted in the largest Hamming distance. The three frameworks, conventional breeding, speed breeding, and genomic selection led to similar results. With 15 initial parents, bulk breeding allowed for the smallest Hamming distance. For 30 parents under conventional and speed breeding, all strategies, except for mass selection, led to the same Hamming distance. Under genomic selection with 30 parents, bulk breeding, single seed descent, and the modified pedigree method had the smallest Hamming distance. When the parental population size was

60 and 100, the strategies, except for mass selection, resulted in the same Hamming distance after 10 cycles.

For seed yield (Supplemental Fig. 7), a parental population size of 15 led to a smaller Hamming distance compared to larger parental population sizes. Similar to white mold tolerance, differences between the strategies were more noticeable with few initial parents. Mass selection consistently resulted in the largest Hamming distance across all scenarios. When comparing the Hamming distance observed in the final cycle, conventional breeding, speed breeding, and genomic selection produced similar results. It was noted that mass selection had a much larger Hamming distance under genomic selection than for the other frameworks. For 15 parents, single seed descent was the strategy that led to the smallest Hamming distance. For 30, 60, and 100 parents, the strategies, except for mass selection, resulted in the same Hamming distance.

Genetic gain

The relative genetic gain averaged across runs was determined for each cycle for the various simulation scenarios. Figure 6 displays the trend in genetic gain for the five strategies, as well as the cumulative genetic gain averaged across strategies when days to flowering was selected. The cumulative genetic gain was greater in conventional and speed breeding compared to genomic selection for all parental population sizes. Figure 7 displays genetic gain for white mold tolerance, while Fig. 8 shows the genetic gain plot for seed yield. A parental population size of 100 led to the greatest percent cumulative genetic gain, followed by 30, 15, and 60 initial parents.

For days to flowering genetic gain (Fig. 6), the initial parental population size of 100 resulted in a maximum of 50% cumulative genetic gain, while the parental population size of 60 led to a minimum of 36% cumulative genetic gain. Conventional and speed breeding resulted in greater cumulative genetic gains compared to genomic selection.

For white mold tolerance genetic gain (Fig. 7), a parental population of 30 led to the greatest cumulative genetic gain, followed by 15, 100, and 60. Interestingly, genomic selection resulted in similar cumulative gains to conventional and speed breeding when the parental population size was 30, 60, and 100. Meanwhile, genomic selection had much lower cumulative gains than conventional and speed breeding when 15 parents were used. The parental population size of 30 resulted in a maximum of 49% cumulative genetic gain. In contrast, the parental population size 15 led to a minimum of 37% cumulative genetic gain.

For seed yield genetic gain (Fig. 8), a larger parental population size resulted in greater cumulative genetic gains, with 100 parents leading to the highest cumulative genetic gains. In general, conventional and speed breeding led to higher cumulative genetic gains compared to genomic selection. The parental population size of 100 resulted in a maximum of 50% cumulative genetic gain. Meanwhile, the parental population size of 15 led to a minimum of 29% cumulative genetic gain.

The proportion of cumulative genetic gain was determined for each cycle when averaged across all strategies (Figs. 6–8). The proportions were determined for the simulation of days to flowering. By cycle five under the conventional framework, on average the various strategies had achieved between 91 and 96% of cumulative genetic gain. Meanwhile, for speed breeding, 91 to 96% of cumulative genetic gain was achieved within the first three cycles. Lastly, for genomic selection, 89 to 98% of the cumulative genetic gain was achieved in the first 6 cycles.

In the simulation for improving white mold tolerance, 83 to 97% of cumulative genetic gain was achieved by cycle 3 for under the conventional framework. Meanwhile, speed breeding led to 83 to 97% of cumulative genetic gains in the first 2 cycles. 93 to 96% cumulative gains were observed in genomic selection. Figure 9 shows the number of cycles required for 95% cumulative ΔG. On average, 3.31 cycles were necessary to achieve 95% cumulative ΔG. The scenario requiring the fewest cycles to obtain 95% cumulative ΔG was dependant on the trait. For days to flowering, the pedigree method under speed breeding with 60 parents required only 1.12 cycles to achieve 95% cumulative ΔG. For white mold tolerance, the pedigree method under speed breeding with 30 initial parents required 1.02. For seed yield, the pedigree method under speed breeding with 30 initial parents allowed for 95% cumulative ΔG to be obtained in 1.04 cycles.

The average ΔG per cycle was determined for all scenarios (Fig. 10). On average, 5.25% ΔG could be obtained per cycle. The scenario resulting in the greatest ΔG per cycle varied depending on the trait being selected. For days to flowering, single seed descent with 100 initial parents under speed breeding led to 8.45% ΔG per cycle. For white mold tolerance, bulk breeding with 15 initial parents under speed breeding resulted in 8.32% ΔG per cycle. For seed yield, single seed descent, pedigree method, and modified pedigree method with 100 initial parents under speed breeding each led to 8.69% ΔG per cycle.

Principal component analysis (PCA)

Principal component analyses were conducted to examine patterns in simulation outputs immediately after the first cycle, where each run is represented by a single point on the biplot. Results are shown in Fig. 11 with eigenvector loadings of various population and quantitative genetics statistics, such as effective population size, fixation of favorable alleles, hamming distance and genetic gain. These statistics were evaluated on the simulated populations under selection for three traits (days to flowering, white mold tolerance, and seed yield) with varying levels of initial parental population sizes, heritability, breeding frameworks and selection strategies.

PCA biplots also included genomic selection and speed breeding as breeding method alternatives to conventional, and all of these contained various selection strategies (bulk, mass, pedigree, modified pedigree and single seed descent). For days to flowering (Fig. 12a), a large linear-like cluster representing a parental population size of 100 can be observed to the right of the PCA plot between the eigenvectors for genetic gain and Hamming distance. At the extreme of the eigenvector for Hamming distance, was the cluster of runs for mass selection under genomic selection. There was a cluster for pedigree method with 100 parents under conventional breeding in the direction of the eigenvector of the genetic gain. At the

extreme of the eigenvector for effective population size, there was a cluster corresponding to bulk breeding with a parental population size of 100 under speed breeding. A cluster representing the pedigree method with 15 and 30 parents under speed breeding formed in the extreme of the eigenvector for the fixation of favourable alleles. Between the eigenvectors for fixed favourable alleles and genetic gain, a cluster corresponding to the pedigree method under genomic selection and speed breeding was found. A cluster representing both single seed descent and the modified pedigree method was located closer to the center of the plot along the axis of the genetic gain vector. Between the eigenvectors for fixed favourable alleles and effective population size, a cluster consisting of multiple strategies including mass selection, the pedigree method, single seed descent, and the modified pedigree method was found.

For white mold tolerance, the first two principal components accounted for 81.8% of the variance (Fig. 12b). Notably, there were fewer distinct clusters that formed, with most points concentrated in the center of the plot. To the extreme in the direction of the effective population size eigenvector, there was a line-like cluster representing the pedigree method under speed breeding. Between the eigenvectors for effective population size and fixed favourable alleles, there was a cluster consisting of single seed descent and the modified pedigree method under speed breeding. Between the eigenvectors for Hamming distance and effective population size, there were many points corresponding to mass selection. Points reflecting all the strategies were dispersed between the vectors for Hamming distance and genetic gain, with a larger parental population size concentrated towards the center of the plot. In the most extreme of the vector for genetic gain, there were many points representing the pedigree method with 15 and 30 parents under conventional breeding.

For seed yield (Fig. 12c), the two major principal components explained 72.3% of the variance. In the outermost region of the plot, there were a number of points representing bulk breeding with 100 parents under conventional breeding between the vectors for Hamming distance and effective population size. Towards the center of the plot, there were clusters for bulk breeding that corresponded to speed breeding and genomic selection, as well as mass selection. There was a distinct cluster for mass selection with 100 parents under genomic selection that was in the direction of the Hamming distance eigenvector. In the direction of the genetic gain eigenvector, there was a cluster corresponding to the pedigree method under conventional breeding. Meanwhile, there was a sparse cluster along the fixed favourable alleles eigenvector, which consisted of the pedigree method, single seed descent, and the modified pedigree method. More points representing single seed descent and the modified pedigree method with 100 parents were found in the center of the plot. In the extreme of the fixed favourable allele eigenvector were points corresponding to the pedigree method with 30 parents under speed breeding.

Discussion

Comparison of selection strategies

The selection strategies showed different responses for each scenario simulated and depended on the trait being selected. For the trait days to flowering, the scenario utilizing single seed descent led to the

highest genetic gain per cycle. For white mold tolerance, the breeding scenario using bulk breeding resulted in the greatest gain achieved for each cycle. For seed yield, the scenario producing to the greatest genetic gain per cycle relied upon single seed descent, the pedigree method, or the modified pedigree method. Interestingly, for all three traits, the pedigree method required fewer cycles until 95% cumulative genetic gain, meaning it may have been more efficient, but the genetic gains achieved were smaller.

Limited studies have been conducted in common beans to compare selection strategies. However, researchers have investigated the use of different selection strategies in soybean breeding. One particular study demonstrated that for the selection of yield, the highest performing lines were obtained via the pedigree method, while single seed descent produced the highest mean seed yield. The authors also found that bulk breeding was impractical for soybean breeding (Djukic et al. 2011). In contrast, a separate study conducted on soybean breeding found that bulk breeding was the most effective for obtaining the highest yielding individuals, while the pedigree method was ideal for less complex traits (Khosla 2019). The authors noted that bulk breeding was better suited to cases where breeding materials are abundant, and in cases with limited resources, pedigree may be the better choice. The results of our simulation study, which was conducted in the common bean, closely reflect previous findings in soybean breeding. Specifically, when it came to seed yield with few breeding materials, it was found that single seed descent, the pedigree method, and the modified pedigree method resulted in the greatest genetic gains. Empirically, Urrea and Singh (1994) tested the performance of mass selection, F_2-derived family selection and single seed descent in an interracial population from the cross between ICA Pijao x Pinto UI114. In their experiment, the authors found that the mean seed yield when testing lines derived from single seed descent was the lowest and highest when derived using the pedigree method. When dealing with traits of moderate heritability (canopy architecture), the authors found a higher frequency of desirable lines using the mass selection and single seed descent. The authors found that the mean of a relatively high heritability trait (days to maturity), was most shifted using single seed descent and mass selection, resulting in more early maturity lines. These results are consistent with our simulations and confirm that, methods that delay line derivation to late generations are more advantageous when dealing with traits of low heritability. In contrast, selection methods that focus on performing early generation selection can result in greater gains when heritability is moderate or high.

Comparison of breeding framework

Three different breeding frameworks were compared in this study. These included conventional breeding, speed breeding, and genomic selection. Conventional breeding was used as a baseline for the other two frameworks to determine how implementable they are in future breeding programs. Based on the results, speed breeding led to the greatest genetic gain achieved. It also led to the fixation of favourable alleles in the shortest time. Considering the breeder's equation, where L, the years per cycle, was greatly reduced, this outcome was to be expected. From the simulation, it was revealed that genomic selection had a similar performance to conventional frameworks. The effectiveness of genomic selection greatly depends on the prediction accuracy, as well as the time and costs saved by replacing phenotyping with genotyping. While prediction accuracies of genomic selection were determined, this study did not factor

the time and cost savings that would be associated with the use of genomic selection. Nonetheless, genomic selection performed on a level that was similar to conventional breeding. As the main advantage with genomic selection is the opportunity to circumvent phenotyping costs, breeders may find utilizing genomic selection to be worthwhile if they have the means to perform large-scale genotyping. However, depending on model accuracy erosion rate, breeding method, and trait of interest, genomic selection may underperform conventional frameworks. Breeders considering genomic selection may also need to consider the expenses associated with establishing the optimal training population, which may require more resources (Hickey et al., 2014). In terms of prediction accuracy, Taylor (2014) reported that GS is optimized when the training population is dynamic, where the progeny of the training population is combined with the training population. In addition, GS is expected to perform poorly if training takes place in one population, but GEBV are to be obtained for a reproductively isolated population. Finally, it was noted that GS becomes less effective in each advancing generation if a static training population is to be used for predicting traits that are difficult to phenotype.

Number of initial parents and crosses

Four different parental population sizes were investigated in this study. A full diallel crossing scheme was employed for each breeding scenario. Since a closed breeding system was simulated, the lines selected at the end of the cycle would be used as the parents of the next cycle. As a result, there was a need for fewer parents and more crosses. While this scheme was mainly used to accommodate the requirements of a closed breeding system, previous researchers have theorized that having more crosses with smaller populations is more effective. According to Bernardo 2003, Witcombe and Virk 2001 and Yonezawa and Yamagata 1978, the use of more crosses with smaller populations was more effective. At the F_2 generation, a breeder with limited resources has the option to create more crosses, each with smaller populations, or create fewer crosses, each with larger populations. This assumed that no prior knowledge on the crosses were available and was found to be true for any choice of parents. In practice, plant breeders will often have information, such as the pedigree and the performance of parents. The optimal choice of parents can typically be ascertained from general and specific combining abilities, and breeders can make decisions accordingly. For simulations, where parents are not thoroughly tested for general and specific combining ability, the inclusion of more parents may influence the effectiveness of the breeding program. The simulation study presented here considered four different parental population sizes. For two of the three traits analyzed, a larger parental population size resulted in higher %ΔG per cycle compared to smaller parental population sizes. Since a full diallel crossing scheme was implemented, there was a greater likelihood that a high performing cross was created and later selected for. For the trait white mold, the smallest parental population size led to the greatest %ΔG per cycle. The total cumulative genetic gain was higher with the use of 15 parents. Under 100 parents, the initial genetic gains were quite high, but gains dropped off very quickly within the first few cycles. The white mold simulation consisted of the fewest QTLs, and 100% of the favourable alleles were fixed within the first two cycles. Thus, selection for white mold was very efficient and it's likely that there was no genetic variance remained after the first couple of cycles in the scenario involving 100 parents. Based on the breeder's equation [1], the

amount of additive genetic variance will influence the genetic gain. As a result, after the first two cycles of selection under 100 parents, no additional genetic gain could be achieved.

Trait heritability and number of QTL

The three traits that were simulated had different heritability levels. Days to flowering was a high heritability trait, with a narrow sense heritability of 0.9. White mold tolerance had a moderate heritability of 0.6, while seed yield had a low heritability of 0.3. The traits also had differing numbers of QTLs, which were included based on certain criteria and available information in the literature. The aim of the study was to simulate breeding scenarios that would closely reflect breeding programs in real life. Thus, only QTLs with reported effects were included. This is unique from previous studies, in which QTL effects were randomly drawn from a normal distribution (Ali et al. 2020; Lorenz 2013; Wang et al. 2003). For all traits, the optimal framework was speed breeding. However, the optimal strategy and number of parents was dependant on the trait being selected. For white mold tolerance, the optimal number of parents was 15, while for seed yield and days to flowering, the optimal number of parents was 100. This may be due to the number of QTLs that were included in the simulation. For white mold tolerance, only 8 QTLs were considered. Selection was likely very efficient and in a closed system, little to no genetic gain could be achieved after the first few cycles. This is reflected in Fig. 10, where the cumulative genetic gain is much lower in the scenario with 100 parents. Days to flowering considered many QTL and seed yield had a lower heritability, meaning selection was likely not as efficient and the use of 100 parents was beneficial for obtaining high performing lines.

PCA plots revealed that the pedigree method often formed clusters in the direction of the eigenvector for the fixation of favourable alleles. This would suggest that the pedigree method had advantages over the other strategies. However, when considering genetic gain, the pedigree method was outperformed by single seed descent and bulk breeding for the simulation of days to flowering and white mold tolerance. This may be due to the efficiency of the pedigree method, which reduced genetic variance rapidly during early cycles of breeding. Other patterns observed from the PCA plots indicated that single seed descent and modified pedigree methods had similarities, as they would often cluster together. This was the case for most breeding scenarios when considering the genetic gain per cycle. The exception, however, was under genomic selection with 15 parents for white mold tolerance and seed yield, where single seed descent and modified pedigree differed significantly in terms of genetic gain per cycle. Lastly, mass selection would often cluster in the direction of the Hamming distance eigenvector. As higher values for a Hamming distance indicates a poor performing line, scenarios clustering in the direction of the Hamming distance eigenvector are likely to be more influenced by this factor.

Conclusions

Breeding programs are complex and may be influenced by many factors. Computer simulations provide the opportunity to investigate multiple breeding scenarios at the same time to evaluate their

effectiveness. The findings from this study show that the success of a breeding program is impacted by the strategy used, the chosen framework, and the parental population size. As well, the optimal breeding scenario depends on the trait being simulated. For a low heritability trait, a large parental population size produced the greatest genetic gain per cycle. For traits involving few QTL, use of a small parental population size is sufficient. In terms of the optimal strategy, single seed descent was most effective for days to flowering, while bulk breeding was ideal for the selection of white mold tolerance. Finally, for the improvement of seed yield, single seed descent, the pedigree method, and the modified pedigree method are all acceptable strategies to use. Some of the limitations in this study mainly involved the inclusion of variable numbers of QTLs for each trait. QU-GENE requires a genetic map rather than a physical map. As a result, QTLs identified as physical positions could not easily be converted to a genetic distance and thus were omitted. Although many reported QTL were used in this study, seed yield remains a complex trait comprised of many small effect QTLs that are difficult to detect, suggesting that QTLs included in this study represent a subsample of the total QTLs that contribute to the trait.

Abbreviations

GS: Genomic selection

GEBV: genomic estimated breeding value

Declarations

Funding

Natural Science and Engineering Council of Canada – Discovery Grants.

Conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Data and code are available at:

Datadryad (TBD)

Github : https://github.com/McGillHaricots/peas-andlove/tree/master/Simulation-files

Code availability

Not applicable

Authors' contributions

JL: Conducted the research and experiments, analyzed the data, wrote the manuscript.

VA: Assisted with research project guidance, reviewed the manuscript, and provided comments.

ZJ: provided input and reviewed the manuscript

JO: provided input and reviewed the manuscript.

PM: provided input and reviewed the manuscript.

DJ: provided input and reviewed the manuscript.

VHV: Led the project and conceived the idea, assisted with data interpretation, assisted with writing of the manuscript.

References

1. Acquaah G (2009) Principles of Plant Genetics and Breeding. Wiley
2. Adams M (1967) Basis of yield component compensation in crop plants with special reference to the field bean, Phaseolus vulgaris 1. Crop Sci 7:505–510
3. Ali M, Zhang L, DeLacy I, Arief V, Dieters M, Pfeiffer WH, Wang J, Li H (2020) Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. Crop J 8:866–877
4. Beaver JS, Osorno JM (2009) Achievements and limitations of contemporary common bean breeding using conventional and molecular approaches. Euphytica 168:145–175
5. Bernardo R (2003) Parental selection, number of breeding populations, and size of each population in inbred development. Theor Appl Genet 107:1252–1256
6. Brim CA (1966) A modified pedigree method of selection in soybeans 1. Crop Sci 6:220–220
7. Djukic V, Djordjevic V, Miladinovic D, Tubic S, Burton J, Miladinovic J (2011) Soybean breeding: comparison of the efficiency of different selection methods.Tur. Jour Agric(35):469–480

8. Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Gen 4:250–255. doi:10.3835/plantgenome2011.08.0024
9. Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of Linkage Disequilibrium in Plants. Annu Rev Plant Biol 54:357–374
10. Galeano CH, Fernandez AC, Franco-Herrera N, Cichy KA, McClean PE, Vanderleyden J, Blair MW (2011) Saturation of an Intra-Gene Pool Linkage Map: Towards a Unified Consensus Linkage Map for Fine Mapping and Synteny Analysis in Common Bean. PLoS ONE 6:e28135
11. Geil PB, Anderson JW (1994) Nutrition and health implications of dry beans: a review. J Am Coll Nutr 13:549–558
12. Kelly JD, Kolkman JM, Schneider K (1998) Breeding for yield in dry bean (Phaseolus vulgaris L.). Euphytica 102:343–356
13. Khosla GS, p (2019) Comparison of different breeding methods for developing superior genotypes in soybean. Agricultural Res J 56:628
14. Larsen J, Morneau E, Zhang B, Digweed Q, Page ER, Mylnarek JJ, Wally OSD (2019) Speed Breeding in Dry Beans
15. Li X, Zhu C, Wang J, Yu J (2012) Chapter six - Computer Simulation in Plant Breeding. In: Sparks DL (ed) Advances in Agronomy. Academic Press, pp 219–264
16. Lorenz AJ (2013) Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. G3 Genes|Genomes|Genetics 3:481–491
17. Oraguzie NC, Gardiner SE, Rikkerink EH, Silva HN (2007) Association mapping in plants. Springer
18. Osorno JM, Vander Wal AJ, Posch J, Simons K, Grafton KF, Pasche JS, Valentini G, Pastor-Corrales M (2021) A new black bean with resistance to bean rust: Registration of 'ND Twilight'. J Plant Registrations 15:28–36
19. Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21:3686–3687
20. Siddiq M, Uebersax MA (2012) Dry Beans and Pulses: Production, Processing and Nutrition. Wiley
21. Soltani A, Bello M, Mndolwa E, Schroder S, Moghaddam SM, Osorno JM, Miklas PN, McClean PE (2016) Targeted analysis of dry bean growth habit: Interrelationship among architectural, phenological, and yield components. Crop Sci 56:3005–3015
22. Stagnari F, Maggio A, Galieni A, Pisante M (2017) Multiple benefits of legumes for agriculture sustainability: an overview. Chem Biol Technol Agric 4:2
23. Taylor JF (2014) Implementation and accuracy of genomic selection. Aquaculture 420–421:S8–S14
24. Urrea CA, Singh SP (1994) Comparison of mass, F2-derived family, and single-seed-descent selection methods in an interracial population of common bean. Can J Plant Sci 74:461–464
25. Wang C, Kao W-H, Hsiao CK (2015) Using Hamming Distance as Information for SNP-Sets Clustering and Testing in Disease Association Studies. PLoS ONE 10:e0135918

26. Wang J, van Ginkel M, Podlich D, Ye G, Trethowan R, Pfeiffer W, DeLacy IH, Cooper M, Rajaram S (2003) Comparison of Two Breeding Strategies by Computer Simulation. Crop Sci 43:1764–1773
27. Witcombe J, Virk D (2001) Number of crosses and population size for participatory and classical plant breeding. Euphytica 122:451–462
28. Yonezawa K, Yamagata H (1978) On the number and size of cross combinations in a breeding programme of self-fertilizing crops. Euphytica 27:113–116

Figures

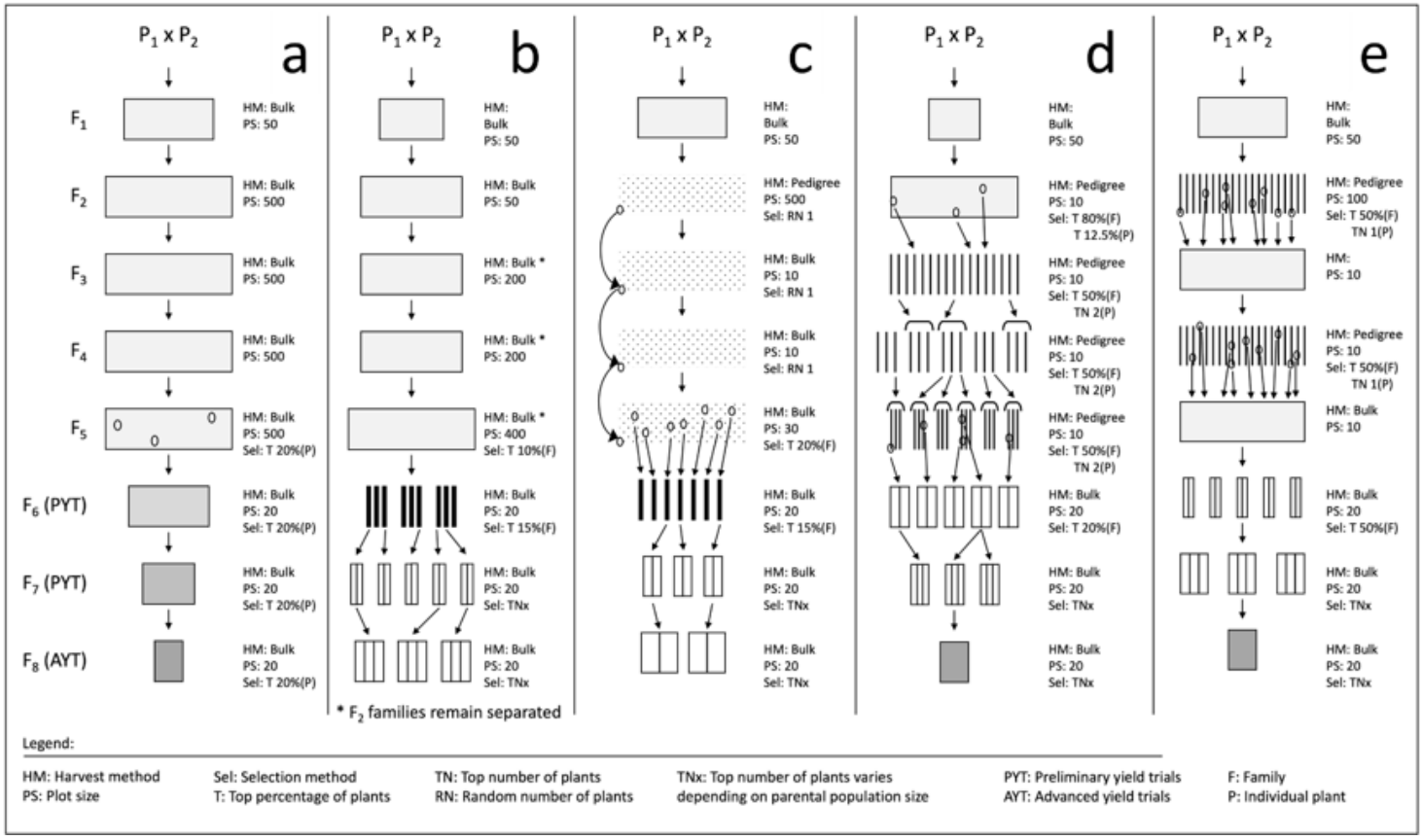

Figure 1

Selection strategies simulated in QuLinePlus. a) Mass selection, b) Bulk breeding, c) Single seed descent, d) Pedigree and e) Modified pedigree.

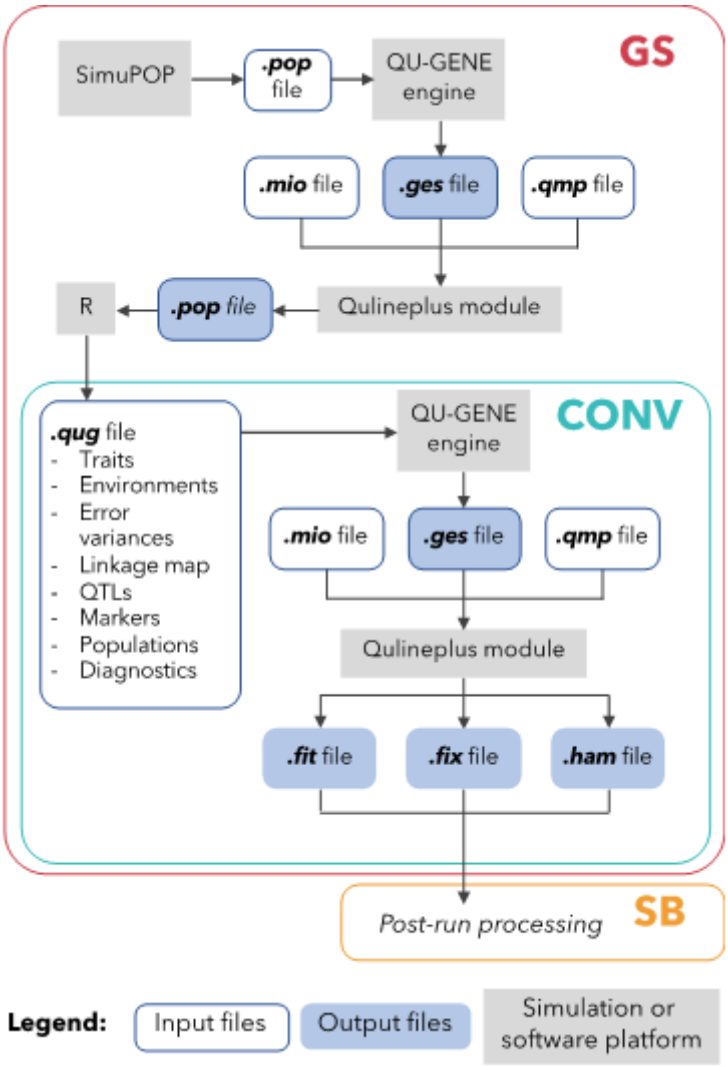

Figure 2

QU-GENE simulation workflow for simulation of breeding frameworks. Genomic selection (GS), conventional (CONV), and speed breeding (SB).

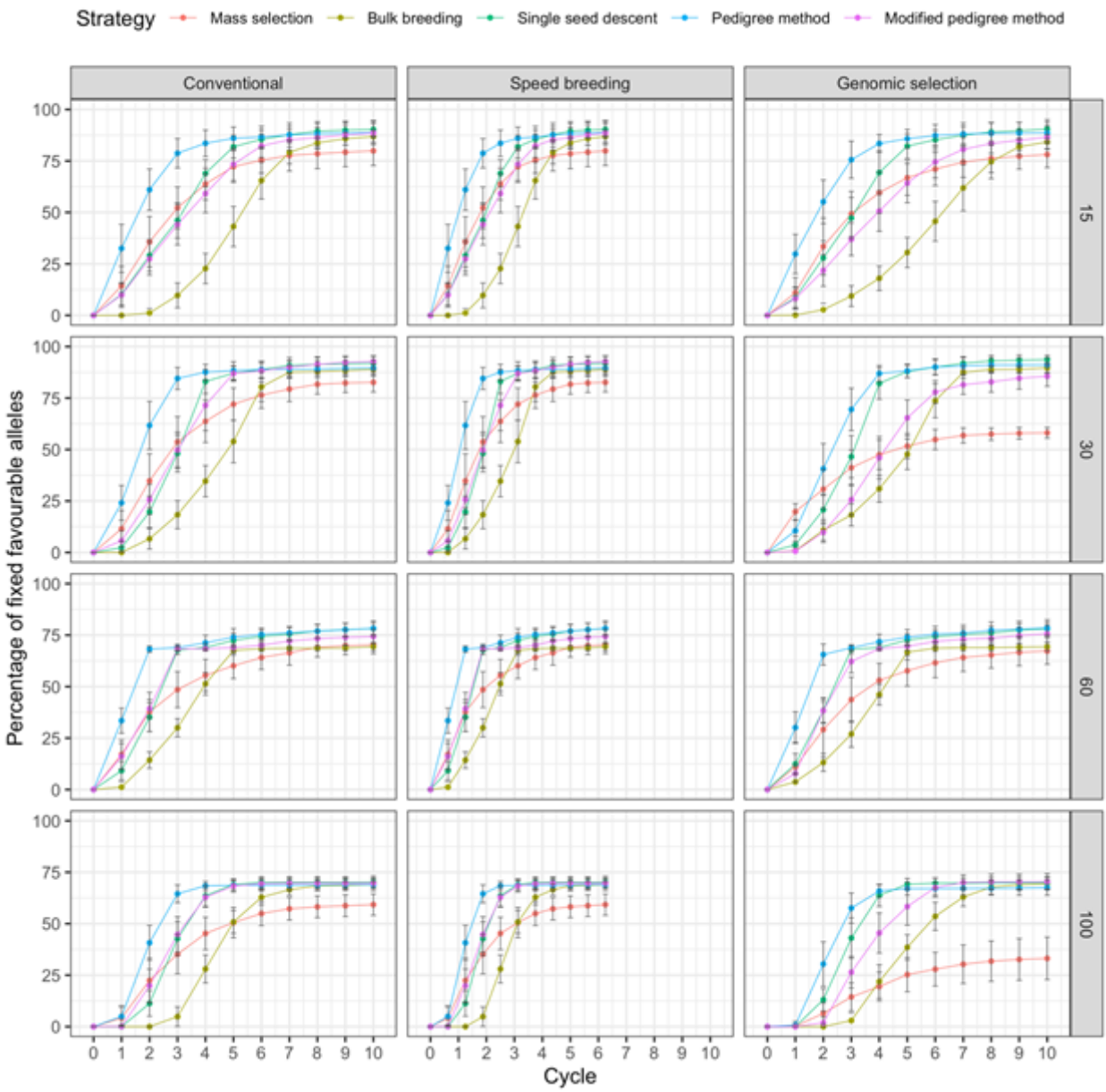

Figure 3

Comparison of five breeding strategies in terms of fixation of favourable alleles over 10 cycles of selection across 50 runs in a closed system. Selection for days to flowering was simulated with increasing numbers of initial parents displayed on the right and differing breeding frameworks shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

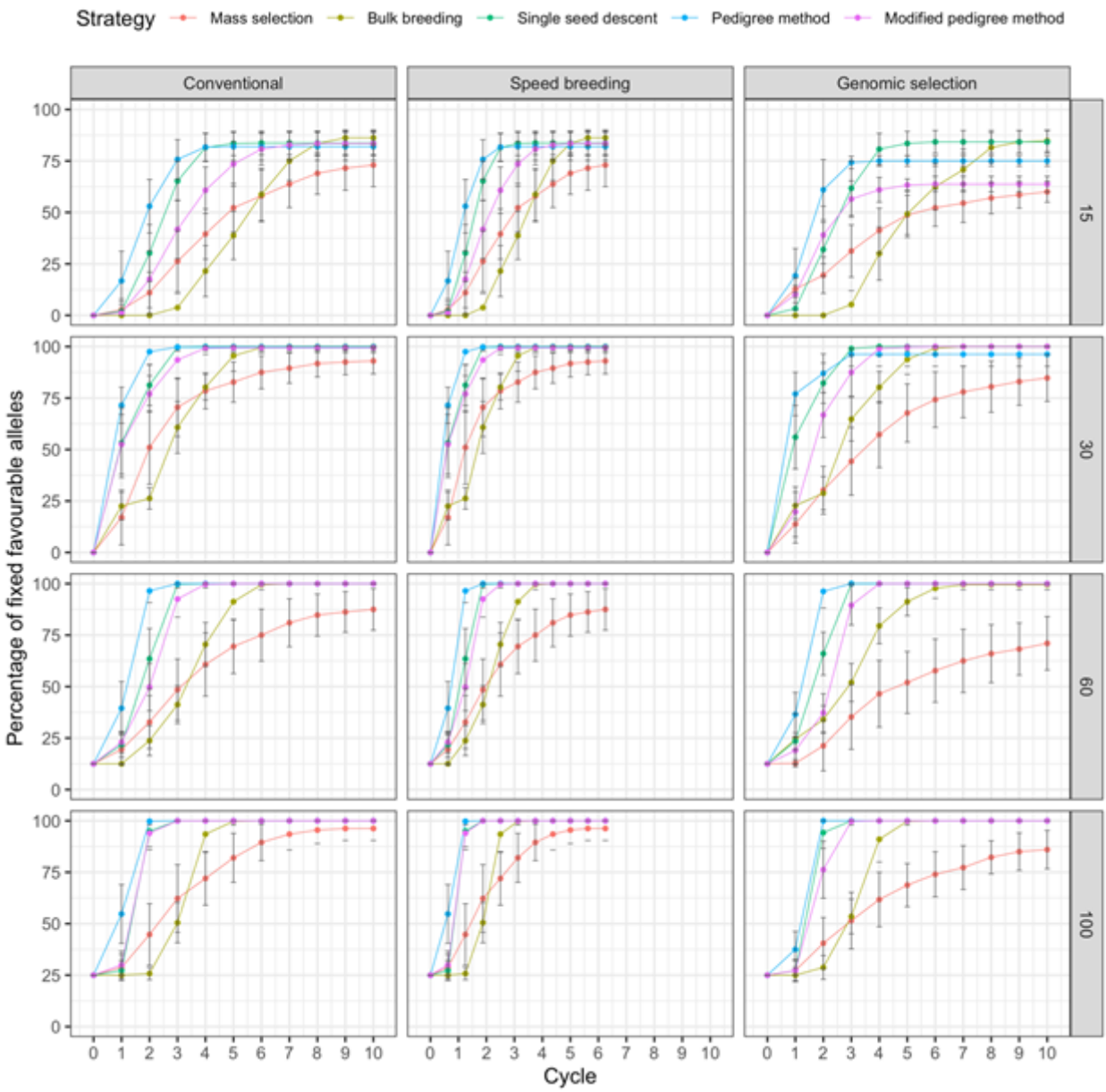

Figure 4

Comparison of five breeding strategies in terms of fixation of favourable alleles over 10 cycles of selection across 50 runs in a closed system. Selection for white mold tolerance was simulated with increasing numbers of initial parents displayed on the right and differing breeding frameworks shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

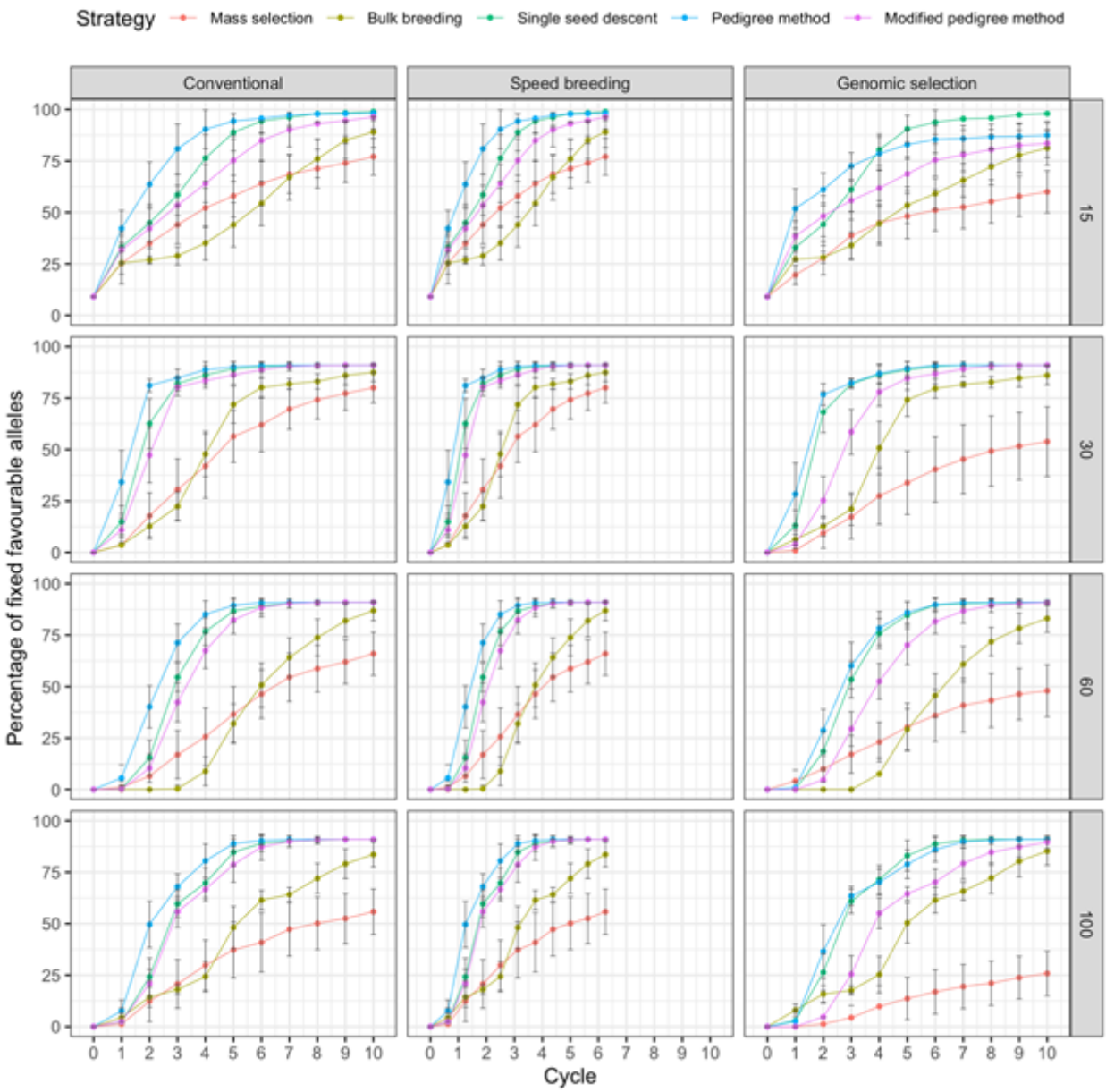

Figure 5

Comparison of five breeding strategies in terms of fixation of favourable alleles over 10 cycles of selection averaged across 50 runs in a closed system. Selection for seed yield was simulated with increasing numbers of initial parents displayed on the right and differing breeding frameworks shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

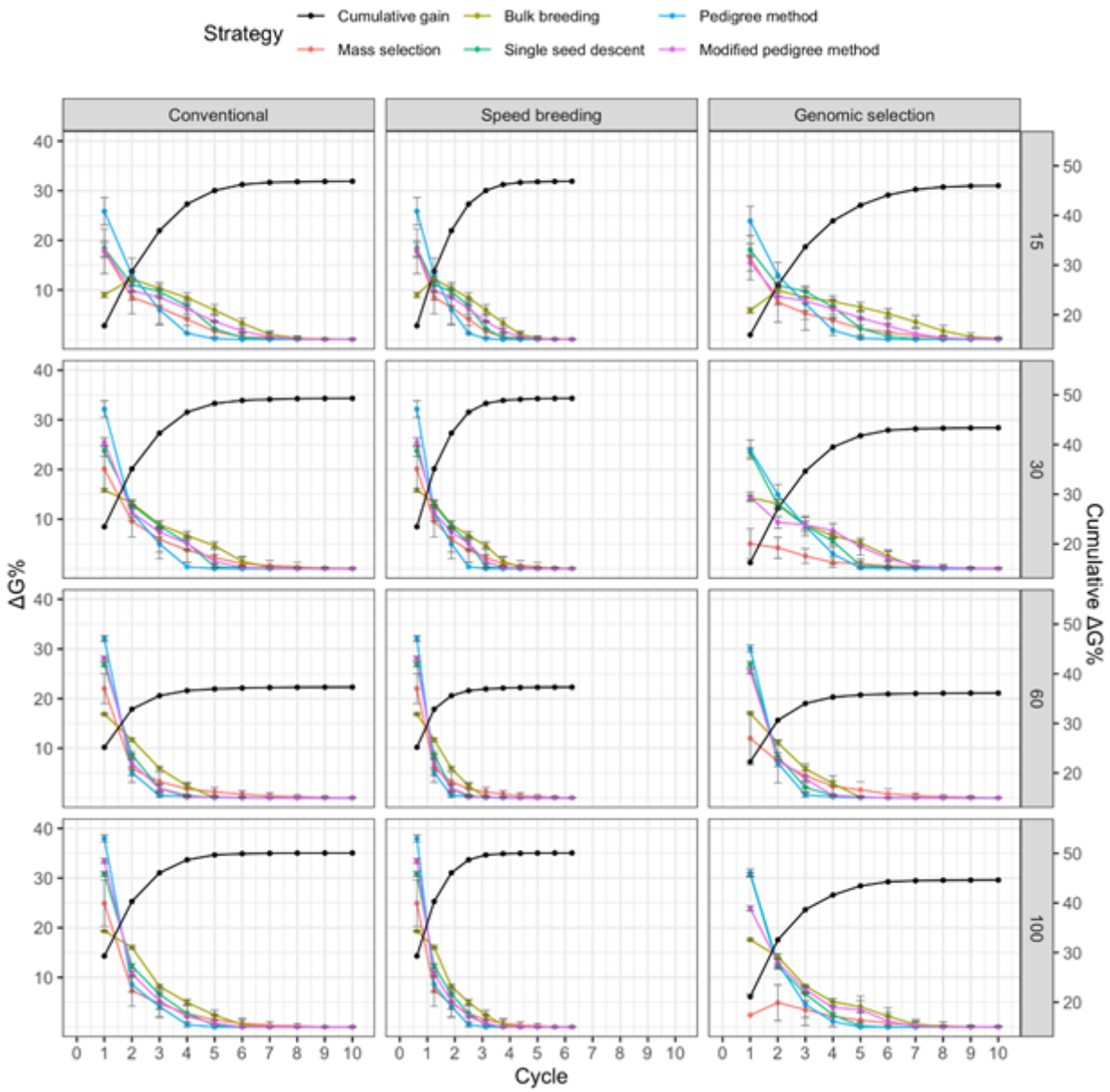

Figure 6

Comparison of five breeding strategies in terms of genetic gain per cycle over 10 cycles of selection averaged across 50 runs in a closed system. Selection for days to flowering was simulated with increasing numbers of initial parents displayed on the right and differing breeding frameworks shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Cumulative genetic gain averaged across strategies indicated in black on the right Y axis. Error bars indicate standard error.

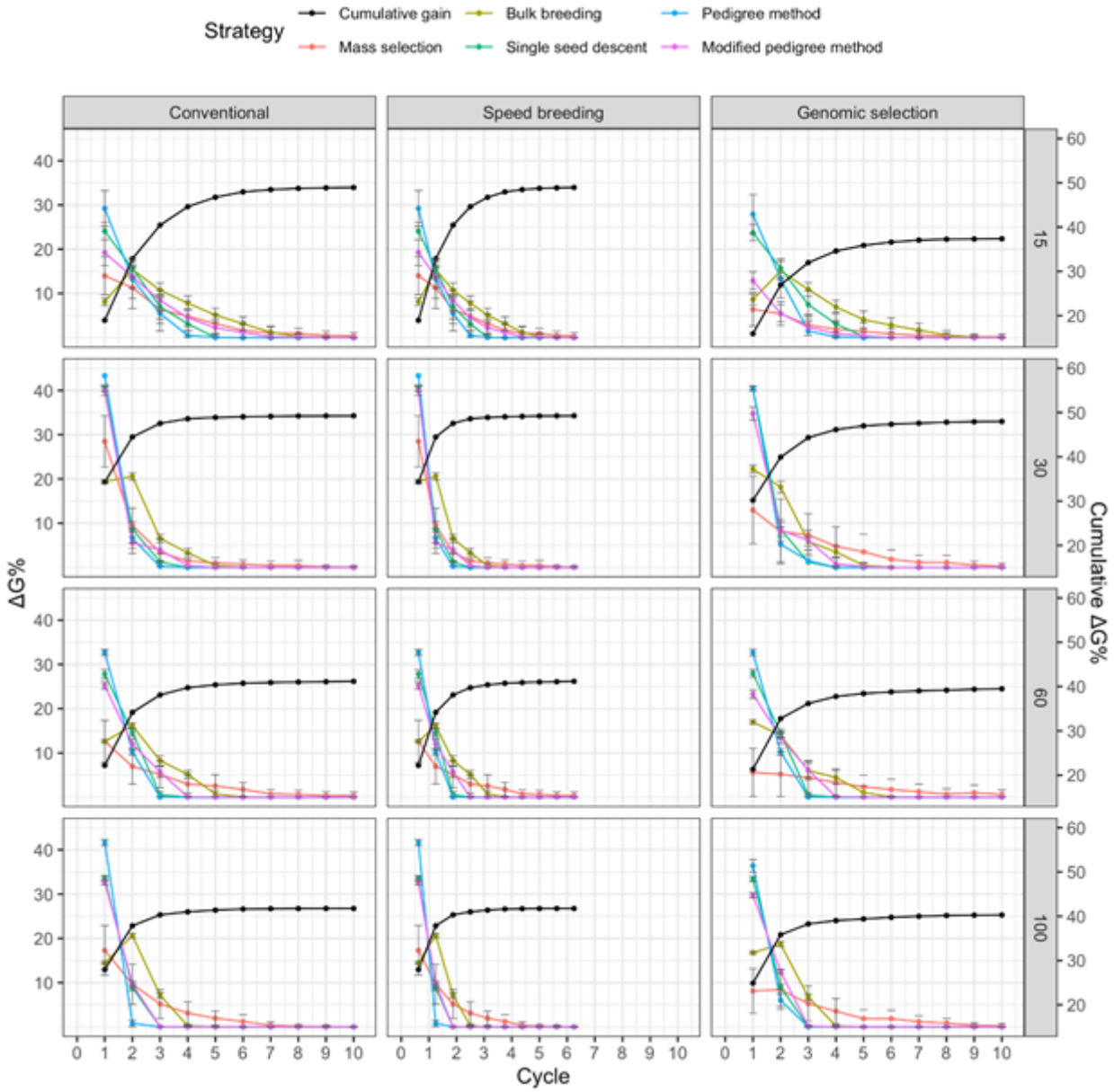

Figure 7

Comparison of five breeding strategies in terms of genetic gain over 10 cycles of selection averaged across 50 runs in a closed system. Selection for white mold tolerance was simulated with increasing numbers of initial parents displayed on the right and differing breeding frameworks shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Cumulative genetic gain averaged across strategies indicated in black on the right Y axis. Error bars indicate standard error.

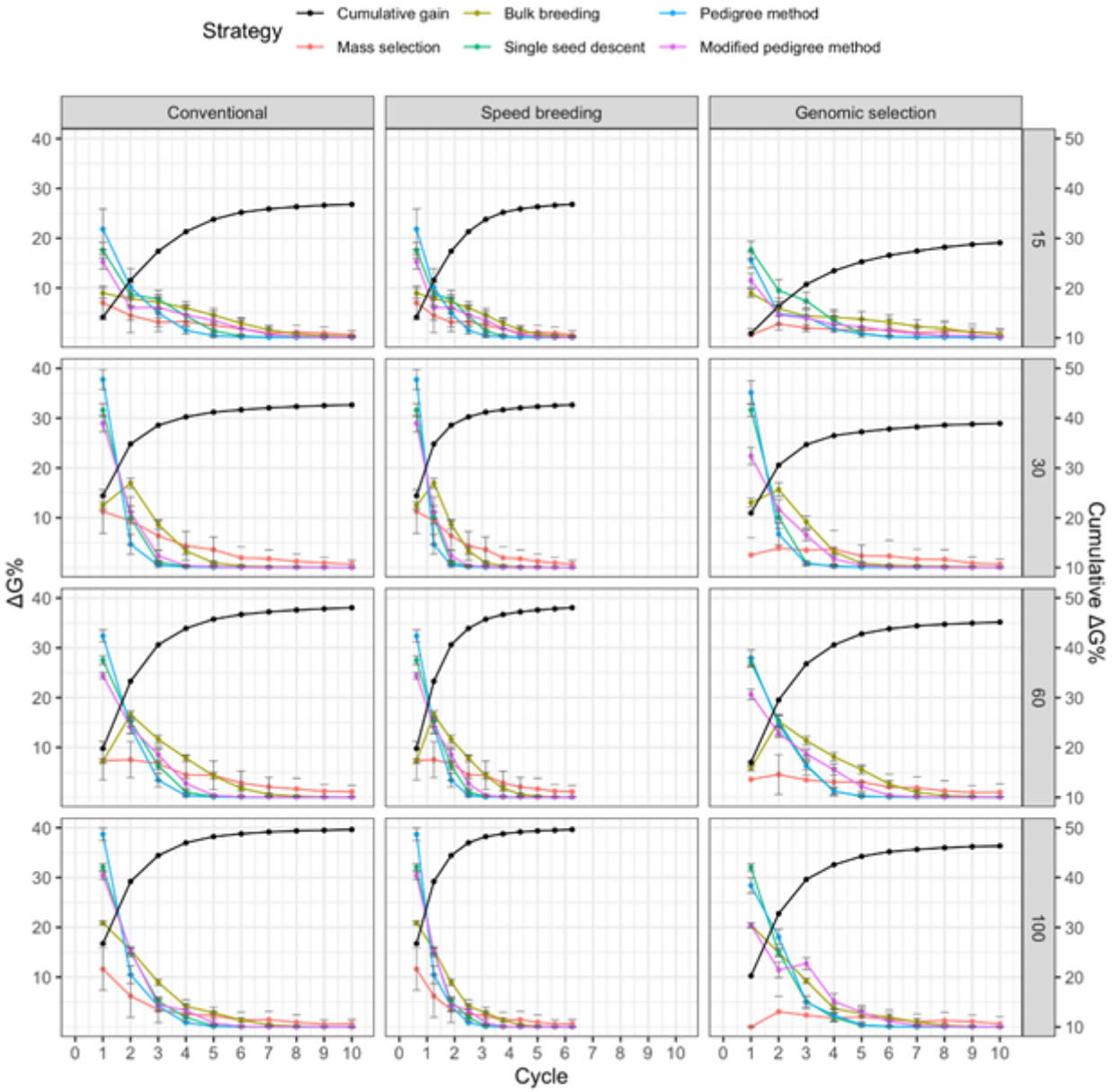

Figure 8

Comparison of five breeding strategies in terms of genetic gain over 10 cycles of selection averaged across 50 runs in a closed system. Selection for seed yield was simulated with increasing numbers of initial parents displayed on the right and differing breeding frameworks shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Cumulative genetic gain averaged across strategies indicated in black on the right Y axis. Error bars indicate standard error.

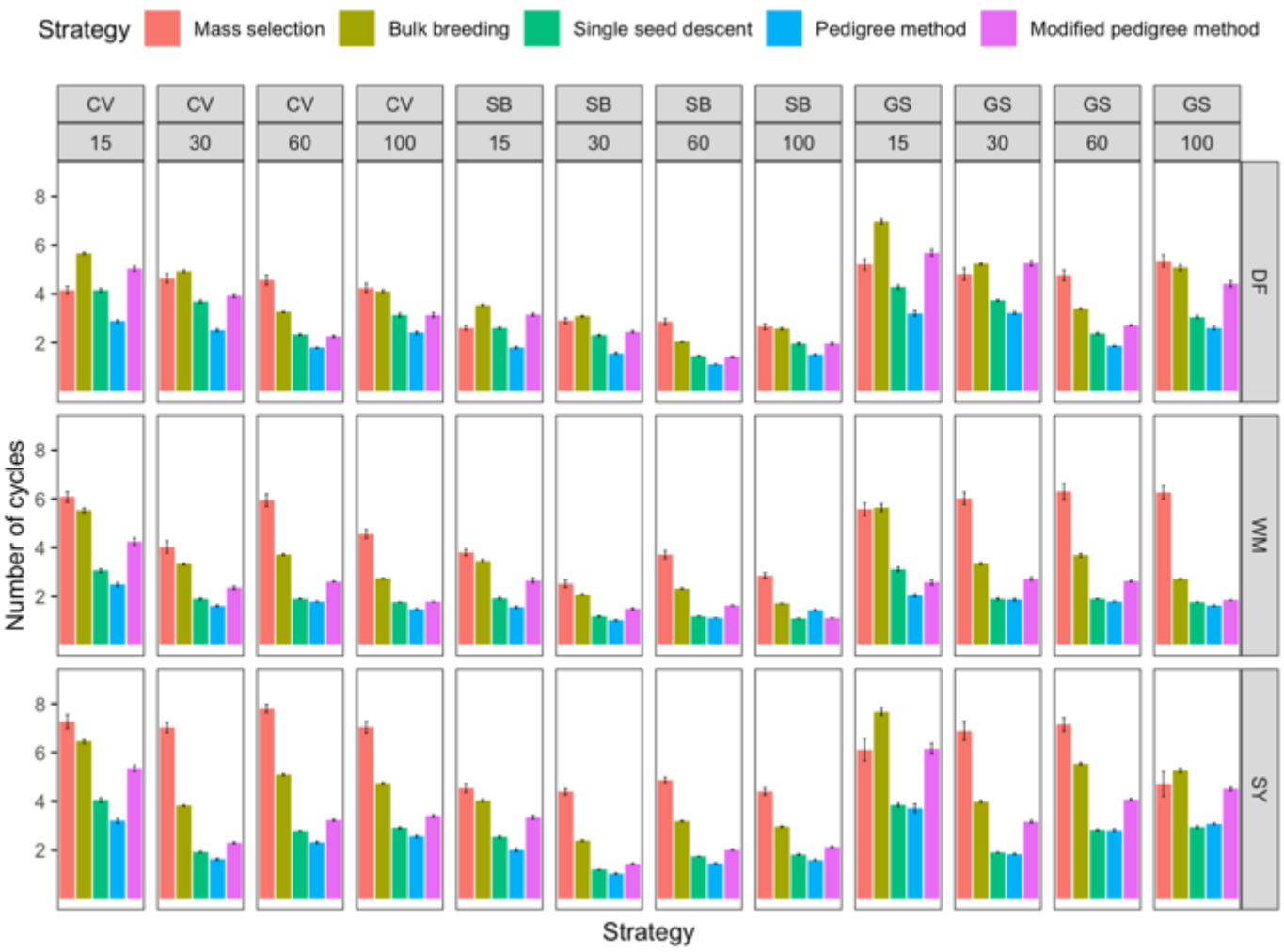

Figure 9

Comparison of five breeding strategies in terms of number of cycles until 95% cumulative of genetic gain for 10 cycles averaged over 50 runs in a closed system. Selected traits include days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Increasing numbers of initial parents displayed on the top along with different breeding frameworks. Breeding frameworks include conventional breeding (CV), speed breeding (SB), and genomic selection (GS). Coloured bars represent the breeding strategies, which include mass selection, bulk breeding, single seed descent, pedigree method, and modified pedigree method. Error bars indicate standard error.

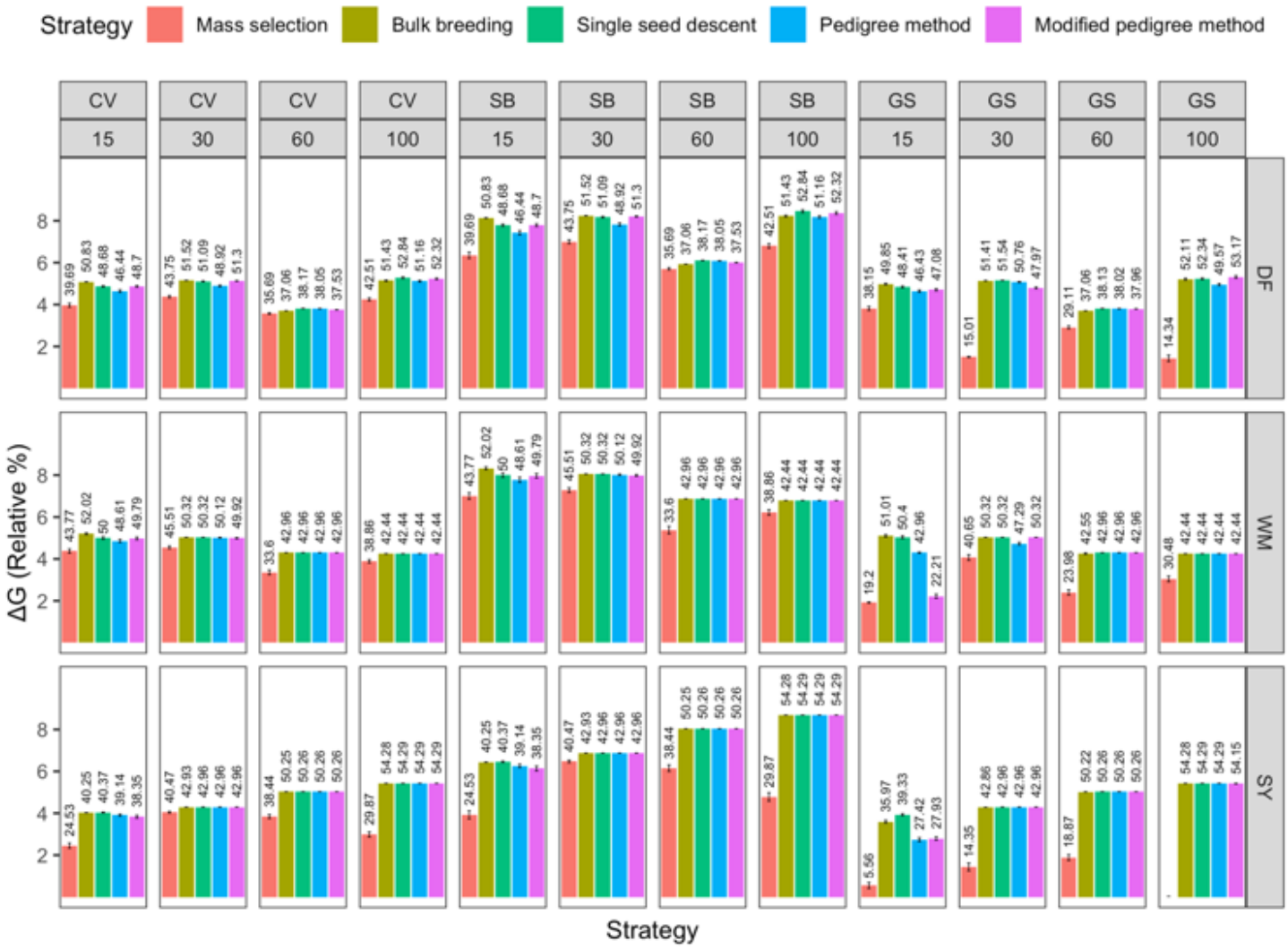

Figure 10

Comparison of five breeding strategies in terms of relative genetic gain per cycle across 10 cycles averaged over 50 runs in a closed system. Selected traits include days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Increasing numbers of initial parents displayed on the top along with different breeding frameworks. Breeding frameworks include conventional breeding (CV), speed breeding (SB), and genomic selection (GS). Coloured bars represent the breeding strategies, which include mass selection, bulk breeding, single seed descent, pedigree method, and modified pedigree method. Error bars indicate standard error. Values above bars indicate the total cumulative genetic gain (%) at the end of the simulation.

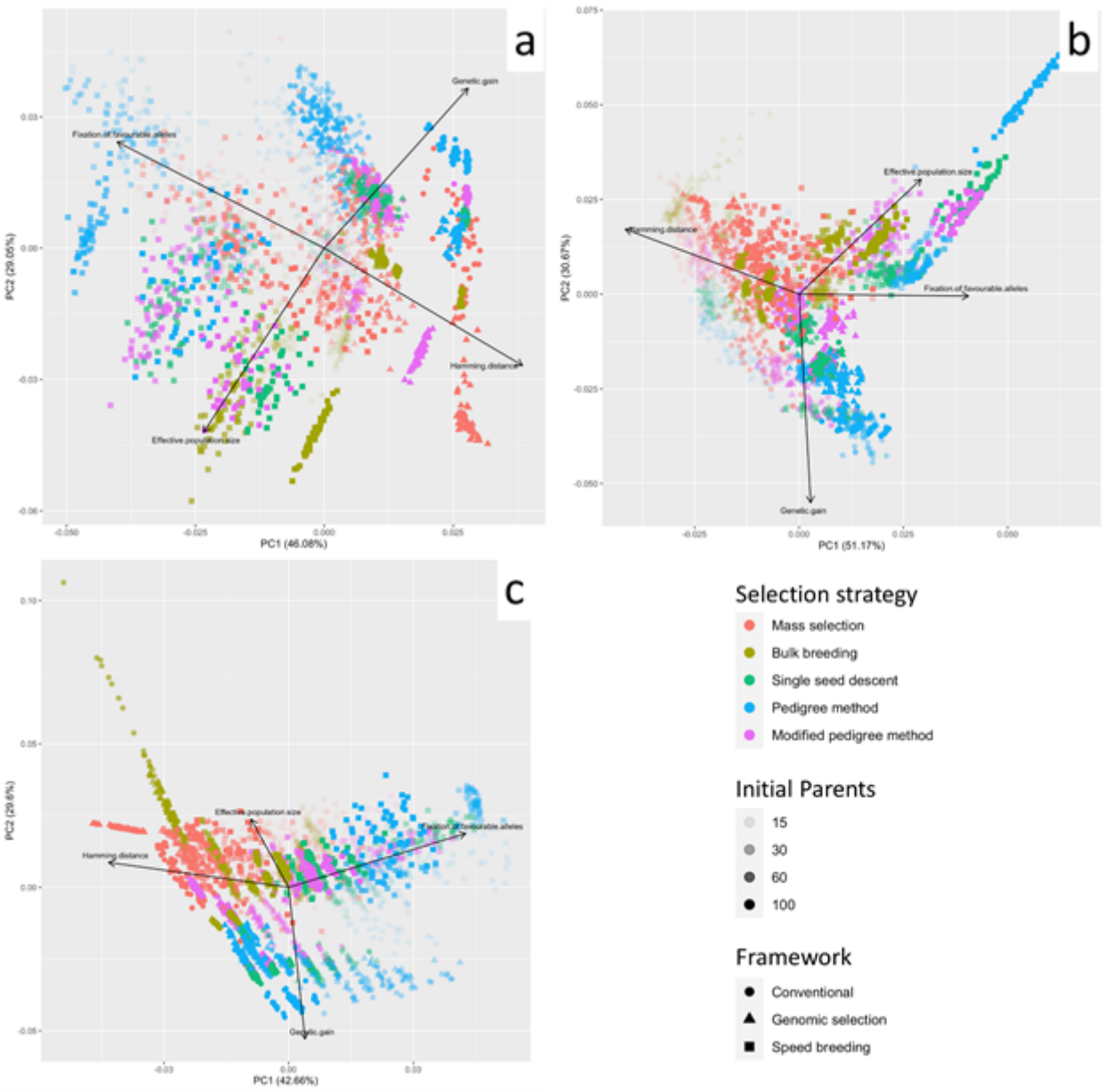

Figure 11

Principal component analysis (PCA) plot of genetic gain across five breeding strategies, three breeding frameworks and four initial parent population sizes. a) Days to flowering, b) white mold tolerance and c) seed yield were selected in simulated populations with increasing parental population sizes represented by different shapes. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, and modified pedigree method, and are distinguished by colour. Vectors specify the direction and strength of genetic gain variables. The first two principal axes explained 75.1% of the variance.

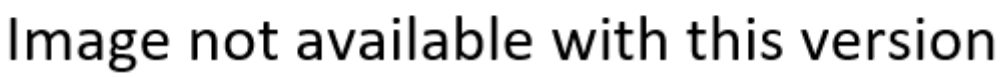

Figure 12

This image is not available with this version

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigures.docx
- Supplementarytables.docx