The simulation of plant breeding scenarios in the common bean (*Phaseolus vulgaris L.*)

Jennifer Lin

Department of Plant Science

MacDonald campus of McGill University, Montreal

21111 Lakeshore, Sainte-Anne-de-Bellevue, Quebec H9X 3V9

February 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science

Table of contents

Acknowledgements

I would first like to thank my supervisor, Dr. Valerio Hoyos-Villegas, for the endless support and guidance throughout my graduate studies. Thank you for always being encouraging and accommodating, especially during the hardships I endured during the pandemic. I would not be where I am at without your constant encouragement. Thank you for believing in me.

I would also like to thank my committee member, Dr. Phillipe Seguin, who took the time to listen to my progress reports and provide helpful feedback. Your expertise was greatly appreciated.

Next, I would like to thank my lab members. Thank you for your support. I will always appreciate the constructive criticism you've given me before any of my presentations and workshops. Also, thank you for two amazing summers of working on the farm. Though it was laborious, I had a lot of fun thanks to you guys.

Finally, I'd like to thank my family and friends for getting me back on my feet anytime I had doubts about myself. I could not have done it without you.

Abstract

The common bean (*Phaseolus vulgaris*) or dry bean is a legume crop that many developing nations rely on for nutrients. As global populations rise, challenges with ensuring food security become exacerbated. Crop improvement of dry beans requires plant breeding, which can take up to 10 years. To ensure success in a breeding program, plant breeders must carefully consider the decisions they make, including phenotyping method, resource allocation, and choice of breeding strategy. Computer simulations can provide abundant information without the need for empirical studies. In this study, five conventional breeding strategies used for the selection of three traits with differing heritabilities were evaluated via computer simulation using the program QU-GENE. These conventional breeding strategies were then compared to new proposed plant breeding methods, genomic selection and speed breeding. Finally, the accuracy of genomic selection was evaluated.

Résumé

Le haricot commun (*Phaseolus vulgaris*), aussi appelé communément haricot sec, est une légumineuse jouant un rôle crucial dans l'alimentation de plusieurs pays en voie de développement de par son aspect nutritionnel. Avec l'augmentation de la population mondiale, plusieurs enjeux liés à la sécurité alimentaire seront exacerbés. L'amélioration génétique du haricot sec nécessite de longs cycles de sélection pouvant prendre jusqu'à dix années. Pour s'assurer du succès d'un programme d'amélioration, les sélectionneurs doivent effectuer les meilleurs choix quant à la méthode de phénotypage, le schéma d'allocation des ressources et la stratégie de sélection. Les simulations informatiques peuvent fournir des informations abondantes sans avoir besoin d'études empiriques. Dans cette étude, cinq stratégies de sélection conventionnelles utilisées pour la sélection de trois caractères avec des héritabilités différentes ont été évaluées par simulation informatique à l'aide du programme QU-GENE. Ces stratégies de sélection conventionnelles ont ensuite été comparées aux nouvelles méthodes de sélection végétale proposées, à la sélection génomique et à la sélection rapide. Enfin, la précision de la sélection génomique a été évaluée.

List of abbreviations

ΔG: Genetic gain

ANOVA: Analysis of variance

BL: Bayesian lasso

BRR: Bayesian ridge regression

cM: Centimorgan

CONV: Conventional breeding

CV: Conventional breeding

DF: Days to flowering

DHAP: Double haploid

E-bayes: empirical Bayes

GEBV: Genomic estimated breeding value

GEPRSS: Germplasm enhancement (),

GEXP: Genetic experiments

GS: Genomic selection

GWAS: Genome wide association study

GxE: Genotype by environment

HGPRSS: Half germplasm enhancement (),

HMSSLT: Half mass selection

HSRRS: Half-sib reciprocal recurrent selection

LD: Linkage disequilibrium

MAS: Marker assisted selection

MSSLY: Mass selection

Ne: Effective population size

NNET: Neural network

PCA: Principal component analysis

QTL: Quantitative trait loci

QU-GENE: Quantitative genetics

RF: Random Forest

RIL: Recombinant inbred line

RKHS: Reproducing kernel Hilbert space

rrBLUP: Random regression best linear unbiased prediction

SB: Speed breeding

SVM: Support vector machine

SY: Seed yield

TBV: True breeding value

TPE: Target population of environments

wBSR: Weighted Bayesian shrinkage regression

WM: White mold tolerance

List of figures

List of tables

List of equations

Chapter 1 Literature Review

1.1 Introduction

Rising global populations, unequal distributions of global food production, and the implications of climate change may have serious consequences for future food security. Malnutrition, in the form of undernutrition, nutrient deficiency, and obesity are issues that developed and developing countries alike continue to face. Undernutrition and nutrient deficiency are particularly problematic in impoverished regions around the globe. Individuals living in low-income areas that rely solely on mono cereal crops as a food source are at risk of inadequate protein intake. Meanwhile, populations in developed countries are at risk of malnutrition in the form of obesity, resulting from low quality nutrients and high intake of carbohydrates and saturated fats. Thus, emphasis should be placed on developing sustainable crops. Common beans, (*Phaseolus vulgaris*) are an important legume crop which numerous countries across the globe rely on for proteins, healthy carbohydrates, and other nutrients. Previous studies have shown that common beans offer a number of health benefits, including reduced risk of diabetes, heart disease, cancer, and obesity. As a nutritionally compact legume, dry beans have the potential to fight malnutrition. Dry bean breeding programs in Canada and the United States have tackled increasing dry bean yield, as well as resistance to biotic and abiotic stresses. Due to the complexity and lengthy duration of breeding programs, plant breeders must carefully consider each aspect that goes into their breeding programs, including selection methods, selection intensity, labour and land resources available, and genotyping and phenotyping tools at hand. Computer simulations, which have become popular in the last few decades, may be used to assist plant breeders in decision making. Simulations provide information that could not be obtained empirically. Softwares including AlphaSimR, DeltaGen, ADAM-Plant, and QU-GENE are capable of simulating breeding programs. The stochastic simulation platform QU-GENE, which is based on the *E(N:K)* model, offers ease and flexibility.

1.2 Pulses and food security

1.2.3 Global malnutrition

Worldwide populations have been projected to surpass 9.5 billion by 2050 and reach 11 billion by 2100 (UN, 2021). This unrestrained population growth, coupled with uneven global crop production and the pressing concerns with climate change, may mean serious food shortages in the near future. Adoption of sustainable food sources will be needed to ensure food security and combat malnutrition. Malnutrition is a serious global concern that comes in many forms, including undernutrition, nutrient deficiency, and obesity. In 2020, 194 million children were either too short or too thin for their respective age and height, while 38.9 million children were either obese or overweight. While every country in the world experiences at least one form of malnutrition, it is particularly devastating for impoverished nations. Protein malnutrition is especially problematic in developing countries. Many regions in sub-Saharan Africa rely on mono cereal crops to feed its populace. Thus, the inhabitants do not receive adequate protein in their diets. Malnutrition in developed countries must also be addressed. Diets that are disproportionately high in carbohydrates and saturated fats, while simultaneously low in quality proteins and essential micronutrients can lead to obesity. Thus, emphasis should be placed on increasing production of highly nutritious crops that are sustainable to grow.

1.2.2 Nutritional aspects of common beans

Common beans and other pulses have numerous health benefits. Pulses, which are categorized as dry edible seeds in the legume family, are low in fat and contain high levels of complex carbohydrates and proteins. Important minerals, such as zinc, iron, potassium, phosphorus, and selenium, can also be found in pulses. Furthermore, pulses are rich in folate, thiamin, niacin6, and other B vitamins (Rosegrant, 2003). Global organizations, such as the United Nations and

World Health Organization made efforts to promote the health benefits of pulses through World Pulses Day. American and Canadian individuals that regularly consume pulses were found to have better diet quality, with higher intakes of fibres, proteins, carbohydrates, and vitamins (Mitchell et al., 2009; Mudryj et al., 2012). There have also been some studies that point to an association between the consumption of pulses and the reduction of risk for cardiovascular disease, diabetes, and obesity. Pulses have a low glycemic index, which has been shown to decrease the risk of coronary heart disease in women. Subjects that were given a diet consisting of pulses for five weeks had greater glycemic control and produced more high-density lipoproteins. They were also predicted to have a greater decrease in waist circumference and eventually lose weight if they should remain on the diet (Mollard et al., 2012). Finally, some studies suggest that dietary pulses may reduce the risk of certain types of cancer. In higher quantities, some of the nutrients and bioactive components in pulses may protect against cancer (Mathers, 2002). In an Italian population, pulses were found to protect against pancreatic cancer (Polesel et al., 2010). Thus, common beans are a nutritionally dense crop with many health benefits that may be utilized to combat malnutrition.

1.3 History of common bean breeding

1.3.1 General characteristics of common beans

Common beans are an annual legume grown in both tropical and temperate climates. The common bean is diploid ($2n = 2x = 22$) with 11 chromosomes and a genome size of approximately 587 Mb (Schmutz et al., 2014). Common beans are sustainable to grow. They are capable of growing in soil that is poor in macro and micro-nutrients. By forming symbiotic relationships with nitrogen-fixing microbes at the root level, common beans are able to improve soil health by increasing nitrogen availability. In addition, as they continue to grow, carbon

exudates are released into the soil from their roots, which then alters the chemical properties of the soil favourably (Gogoi, Baruah, & Meena, 2018). Common beans have been grown for its dry edible seeds for thousands of years and are currently considered to be a staple crop across the world.

1.3.2 Domestication of dry beans and genetic implications

Phaseolus vulgaris, otherwise known as the common bean or dry bean, was first domesticated, likely more than once, in the Andes and Mesoamerica (Shree P. Singh et al., 1991; Chacón s et al., 2005). Domestication led to drastic changes in the morphology of the bean plants.

In addition to this, as a result of separate domestication events, the common bean has two distinct gene pools: the Andean gene pool and the Middle American gene pool. Due to multiple domestication events in the Mesoamerican region, the Middle American gene pool has greater genetic variation (Siddiq & Uebersax, 2012). The gene pools can be differentiated with phaseolin and allozymes analyses (P. Gepts, Osborn, Rashka, & Bliss, 1986; Koenig & Gepts, 1989). Within these gene pools, dry beans can be further classified into different races mainly based on morphological characteristics. There are four Mesoamerican races (Mesoamerica, Durango, Jalisco, and Guatemala) and three Andean races (Nueva Granada, Peru, and Chile). Previously, chloroplast DNA was used to further explore how common beans were domesticated. Results from the study support the hypothesis of a single domestication event for the Andean gene pool and multiple domestication events for the Mesoamerican gene pool (Chacón, Pickersgill, & Debouck, 2005). Modern varieties of dry beans come from one of these two gene pools. The black, navy, pinto, great northern, and small red market classes belong to the Mesoamerican gene pool. Meanwhile, the kidney and cranberry market classes belong to the Andean gene pool.

1.3.3 Dry bean breeding

Dry bean breeding programs in Canada and the United States have made substantial progress in improving biotic and abiotic tolerances, in addition to increasing yield. Breeding efforts have been focused on improving specific market classes locally. The objective of every breeding program is to improve yield, which is typically measured in kg/ha. In the United States, the rate of genetic gain reported for pinto beans was 13.9 kg/ha per year, and 17.4 kg/ha per year for navy beans. Dry bean breeding programs typically follow a general procedure, beginning with hybridization, followed by multiple rounds of generation advancement, during which selection takes place, and concluding with multi-location and multi-year field trials, in which the best genotypes are identified and released as a new variety (Siddiq & Uebersax, 2012). Despite the vast amount of genetic and phenotypic information available to plant breeders, there is still a gap in transferring this knowledge to breeding practices. Breeding programs are both time consuming and resource extensive, with each decision made having consequences for the outcome of the program. With the aid of genome wide association studies (GWAS), useful quantitative trait loci (QTL) or genes have been identified in the common bean. For many cereals, a common approach to selecting based on QTL is marker-assisted selection (MAS). However, MAS is not widely used in pulses due to difficulties in establishing marker-trait associations for useful markers and the high genotype by environment interactions present in many pulse crops (Kumar, Choudhary, Solanki, & Pratap, 2011). Thus, challenges still remain for accumulating desirable QTL and gene pyramiding multiple traits in new varieties (Assefa et al., 2019). These challenges may be addressed with the aid of computer simulations

1.4 Computer simulations in plant breeding

1.4.1 Simulation theory

Computer simulations have come into the spotlight in recent decades as a way to evaluate all possible conditions that one may face in practice. They allow current models to be tested in different scenarios, which may in turn increase confidence in said models. Simulation studies may be classified as either deterministic or stochastic. In deterministic simulations, the output obtained from one input will always be the same. Contrarily, stochastic simulations allow for randomness. The outputs are distributed around the true value, so they are considered to be probabilistic. In other words, the same input may result in different outputs. Computer simulations may be applied to four areas of plant breeding: comparison of breeding schemes, validating the effectiveness of gene mapping, crop modeling to link genotypes and phenotypes, and simulating entire breeding processes to accommodate gene-environment interactions (Li, Zhu, Wang, & Yu, 2012). A simulation study was previously conducted to assess two breeding strategies used in CIMMYT's wheat breeding program. The findings from the study indicated that the selected bulk method had 3.3% greater gains compared to the modified pedigree method (Jiankang Wang et al., 2003). Thus, computer simulations have become highly informative for deciding upon the best breeding strategy to use.

1.4.2 Plant breeding simulation platforms

Numerous plant breeding simulation platforms have been developed that currently available to plant breeders. These include ADAM-Plant, AlphaSimR, DeltaGen, Plabsoft, MBP, GREGOR, and GENEFLOW. Each program makes certain assumptions, which must be carefully considered when deciding whether it is suited for simulating a breeding program.

ADAM-Plant

ADAM-Plant is a stochastic simulation software extending from the animal breeding software, ADAM. It is applicable to self-pollinated and cross-pollinated crops and has the capacity to simulate overlapping generations. In addition, it considers genotype by environment interactions. Two genetic models are available: an infinitesimal model and a genomic model, where users must indicate markers and QTLs (Liu et al., 2019).

AlphaSimR

AlphaSimR is a stochastic simulation that generates founder haplotypes with linkage disequilibrium and allele frequency distributions matching user specific genetic model. Traits are simulated based on additive, dominance, epistatic, or GxE models. Meanwhile, a number of functions are available to simulate different selection schemes, including genomic selection (Gaynor, Gorjanc, & Hickey, 2021).

DeltaGen

DeltaGen is a plant breeding decision support application that can be implemented in the statistical software R. DeltaGen facilitates statistical analysis of field data with linear and mixed models that are integrated within its framework. DeltaGen allows for simulation of breeding strategies that are defined within the program. These strategies include half-sib, half-sib with progeny testing, among and within half-sib, etc. (Jahufer & Luo, 2018). A drawback of this program is that it does not allow for simulation of user defined breeding strategies.

Plabsoft

Plabsoft is a population genetics simulation program that is available as a package in R. it may be used to estimate allele frequencies, various genetic distances, and genetic diversity. It is also applications in plant breeding and is capable of simulating stages or even the entirety of a

breeding program. The genotypic value is estimated as the sum effects of a select number of loci. The software also consists of an algorithm that locates haplotype blocks. (Maurer, Melchinger, & Frisch, 2004). One of the criticisms it faces is the lack of a user-friendly interface.

MBP

MBP was developed to assist in hybrid maize breeding using double haploids. The software incorporates cost effectiveness estimates to allow users to make decisions based on available resources and materials. The genotypic variance is estimate from the general and specific combining ability of a test cross. Thus, MBP may be used to optimize the general combining ability given a restricted budget. The software can also output loss of genetic variance per year (Gordillo & Geiger, 2008). A concern with this program may be the capacity to simulate breeding schemes outside of double haploids.

GREGOR

GREGOR is a research and educational software that can simulate outcomes from different mating or selection schemes. The inputs are defined in three objects: population, traits, and marker list. The population can undergo specific mating or selection schemes, and the phenotype of the resulting population is estimated from the trait and marker list. (Tinker & Mather, 1993). While GREGOR is very straightforward to use, all inputs are simulated within the program and results are based on a hypothetical genome, which may not be reflective of reality.

GENEFLOW

GENEFLOW is a commercial software that may be used for plant breeding decision support. It uses an amalgamation of pedigree information, genotypic data, and phenotypic data to help users understand genetic relationships, trait inheritance, and population structures. It provides estimates for genetic diversity and gives information on gene-trait relationships.

1.4.3 QU-GENE

QU-GENE (QUantitative-GENEtics) is a software that can be used as a simulation platform for studying genetic models (Podlich and Cooper, 1998). It is versatile and can be used to investigate populations from a quantitative genetics standpoint, such as how different genotype-by-environment models can impact the performance of a genotype. The QU-GENE software is made up of two elements, an engine and a module. The engine essentially specifies a genetic model for the genotype-environment system. Meanwhile, the module is used to alter and examine genotype populations in the specified genotype-environment system. One of the benefits of QU-GENE is that the engine produces baseline information regarding the genotype-environment system, meaning that to conduct computer simulations, one only needs to focus on applying the module. Thus, it is possible to run a number of simulations using different modules in the same genotype-environment system. There are several modules that are available for use. These include mass selection (MSSLT), half mass selection (HMSSLT), half-sib reciprocal recurrent selection (HSRRS), double haploid (DHAP), germplasm enhancement (GEPRSS), half germplasm enhancement (HGPRSS), pedigree (PEDIGREE), and genetic experiments (GEXP) (Podlich and Cooper, 1998).

Before discussing how QU-GENE can be used for simulation experiments, it is important to understand the *E(N:K)* model, which is essentially the backbone of the QU-GENE program (Podlich and Cooper, 1998). The *E(N:K)* model makes use of linear statistical and landscape models by bringing together stochastic and deterministic elements. In the E(N:K) model, *E* stands for the number of different environment types in the genotype-environment system. The number of environment types and how often they occur then determines the target population of environments (TPE). Next, *N* stands for the number of genes involved in expression of traits

(Podlich and Cooper, 1998). Lastly, *K* stands for the average amount of epistasis in the genotype-environment system. Once the *E(N:K)* model has been chosen, the researcher can then indicate additional information to include. For example, information on the locations of the genes on the chromosomes, the number of traits that are affected by the genes, if there are interactions between the loci, the types of environments in which certain genes are expressed, and the heritability of traits (Podlich and Cooper, 1998). The main advantage to using QU-GENE over other simulation platforms is the flexibility The breeding strategies are user-defined, making it possible to compare even small differences between strategies. Another aspect is the extensive output provided by QU-GENE. The population files generated by QU-GENE contain allelic information for every individual. This allows for additional analysis to be conducted.

Due to the flexibility, accessibility, and user-friendliness of the platform, QU-GENE was used to simulate multiple breeding scenarios. This paper focuses on five conventional breeding strategies under three breeding frameworks with four different parental population sizes. Among the three frameworks is genomic selection, a novel selection method that relies on the prediction of phenotypes from genotypes via modeling. The effectiveness of this framework is considered in chapter 2, while its accuracy is investigated in chapter 3.

Chapter 2 Evaluation of breeding scenarios in the common bean with the use of simulations in QU-GENE

Abstract

The common bean is a nutritiously dense legume that is consumed by 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.

2.1 Introduction

2.1.1 Importance of dry beans

With ever 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, Maggio, Galieni, & Pisante, 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 & 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 & Anderson, 1994).

2.1.2 Traits for improvement

Increasing dry bean yield is of importance for both developed and developing countries that rely on this legume. The main hindrances 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 a number of 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, Kolkman, & Schneider, 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 & Osorno, 2009). The focus of this paper will be on yield-related traits and biotic stresses. More specifically, the three traits of differing heritability levels that were examined include seed yield, days to flowering, and white mold tolerance.

2.1.3 Dry bean yield

Enhanced crop yield is a result of improved cultivars, higher production inputs, suitable agronomic practices, and good growing conditions. In general, improved cultivars plays a major role in allowing for high crop productivity. Since dry beans growing conditions are rarely free from diseases, drought, insects, or extreme temperatures, breeding for seed yield often involves the accumulation of genes and QTL that improve yield, as well as genes that confer tolerance to abiotic and biotic stresses. For the purpose of accumulating genes for high yield, it is necessary to understand the underlying genetics that dictate seed yield. This can be accomplished by performing quantitative trait loci (QTL) analyses to identify regions in the genome that are associated with a high yield. Association mapping studies are preferable to bi-parental mapping

studies for the detection of QTL because greater resolutions can be obtained due to smaller linkage disequilibrium (LD) blocks. A number of studies have been conducted to identify QTL associated with seed yield in units of kg/hectare. In one study, nine QTL were identified in a population advanced from a cross between a commercial common bean variety and a wild common bean. These QTL were found on linkage groups 2, 3, 4, and 9, and together, they accounted for 9 to 21% of the variance with effect sizes ranging from 98 to 326 kg/ha (Blair, Iriarte, & Beebe, 2006). A recombinant inbred line (RIL) obtained from crossing two black bean cultivars revealed QTL on linkage groups 3, 5, 10, and 11 with additive effects ranging from 41 to 192 kg/ha. One of the QTL on group 10 and explained 28% of the variance (E. M. Wright & Kelly, 2011). In a study involving three half-sib populations obtained from small red bean crosses, four QTL were found that collectively explained 87.9% of the variance. The QTL found on linkage group 3 had the largest effect size, contributing 435 kg/ha (Hoyos-Villegas, Song, Wright, Beebe, & Kelly, 2016).

An important factor to consider is market class. Dry bean market classes include black bean, cranberry bean, great northern bean, red kidney bean, navy bean, pinto bean, and small red bean (Sinha, Hui, Evranuz, Siddiq, & Ahmed, 2010). These market classes vary in size and may be categorized as small seeded (<25 g 100 seed weight^{-1}), medium seeded (25 to 40g 100 seed weight^{-1}), and large seeded (>40 g 100 seed weight^{-1}). Dry beans sometimes exhibit yield component compensation, where seed yield is negatively correlated with seed weight (Paul Gepts et al., 1991). However, this phenomenon is influenced by the environment and is exacerbated when there is competition between plants due to limited resources (Westermann & Crothers, 1977).

2.1.4 Dry bean flowering time

Dry beans may be day-neutral, meaning that they will flower irrespective of photoperiod. Alternatively, they may be photoperiod-sensitive, whereby flowering is influenced by day length (J. W. White & Laing, 1989). Research into dry bean photoperiods has given more direction into developing cultivars that have increased production in their respective growing regions. Latitudes play a role in photoperiod sensitivity. Common bean genotypes grown in regions further from the equator are more likely to be day-neutral. Meanwhile, dry bean cultivars from countries located close to the equator show more variability. However, this variability may be due to differing preferences for selected traits. When considering the influence of latitudes, temperature must also be accounted for. Higher temperatures are correlated with increased photoperiod sensitivity. Genotypes originating from warmer regions are commonly day-neutral, while those originating from cooler locations are more often photoperiod-sensitive. Day-neutrality appears to be associated with increased seed yield, regardless of temperatures. Growing photo-sensitive genotypes in warmer environments leads to lower yields (J. W. White & Laing, 1989). Photoperiod and temperatures both impact the number of days to flowering, which has been positively correlated with yield-related traits, such as number of pods per plant (AlBallat & Al-Araby, 2019). Therefore, making selections for days to flowering may indirectly improve yield. Understanding the underlying genetics that control days to flowering may ease the breeding process. QTL analyses from inter-gene pool derived populations have revealed a number of QTL contributing to days to flowering. One study found three QTL on linkage groups 1, 2, and 8, which when combined, explained 85.5% of the phenotypic variation (Pérez-Vega et al., 2010). In another study, a QTL found on linkage group 1 explained 8.6 to 22.3% of the phenotypic variation. Meanwhile, a QTL found on linkage group 4 explained 7.1 to 14.3% of the

phenotypic variation (Mukeshimana, Butare, Cregan, Blair, & Kelly, 2014). A different study found a QTL on linkage group 1 that explained up to 18.96% of the phenotypic variation (González et al., 2016). Further studies have identified some candidate genes involved in days to flowering on chromosomes 1, 3, 5, 7, and 8. Two of these genes encode a putative 5′-nucleotidase SurE and a putative ubiquitin-conjugating enzyme E2, both of which are involved in plant growth. Another candidate gene, encoding an ATP binding/protein kinase, was thought to play a role in sensing light. Finally, a probable polygalacturonase gene may be responsible for pollen growth (Ates et al., 2018).

2.1.5 Dry bean white mold tolerance

Sclerotinia sclerotiorum Lib. de bary is a destructive fungal pathogen with disease incidences that are difficult to predict. Disease impact is highly contingent on environmental and agronomic conditions. Fungal growth escalates in humid conditions, this dense canopies, which accumulate moisture, promote white mold colonization (HAAS & BOLWYN, 1972). Total resistance to white mold does not exist in common beans. However, some dry bean cultivars display partial resistance to white mold, either though physiological tolerance or disease avoidance. Dry bean cultivars with upright architectures exhibit white mold avoidance and are less susceptible to infection due to more sunlight and air being able to infiltrate the canopy (Miklas, Johnson, Delorme, & Gepts, 2001). When selecting for white mold tolerance, breeders will typically introgress both physiological resistance and avoidance genes. A number of studies have reported QTL that contribute to white mold resistance. A large-effect QTL was previously identified on linkage group 7, which accounted for 38% of the phenotypic variation in straw test disease scores (Miklas et al., 2001). QTLs on linkage group 5 and 8 were later found to explain 10.7%

and 9.2% of the phenotypic variation for plot-based disease severity, respectively (Ender & Kelly, 2005). Linkage group 7 was also found to contain a QTL

In recent years, researchers have been able to identify candidate genes that may contribute to white mold resistance. Researchers have successfully narrowed down 9 meta-QTL regions from existing QTL studies and from new populations. Sources of genetic resistance were derived from Andean gene pool, the sister species, *P. coccineus*, and the navy bean ICA bunsi. Some of the candidate genes described were those involved with pathogen recognition and signal relaying, while others were involved with metabolism during abiotic and biotic stress. The authors also identified ethylene-responsive transcription factors that play a role in programmed cell death (Lucy Milena Diaz et al., 2018). Other candidate genes include those that encode leucine-rich repeat (LRR) proteins, as well as an EF-Tu receptor gene, and may also confer physiological resistance in dry beans (Oladzadabbasabadi, Mamidi, Miklas, Lee, & McClean, 2019). Additional candidate genes were discussed in a meta-QTL analysis, which revealed 37 different QTL, 20 identified through the straw test and 13 identified through field evaluations. Within the WM1.1 QTL, a candidate gene coding for a wall-associated receptor kinase protein is thought to be involved in recognizing pathogens invading the cell wall. Another candidate gene in this region is a coronatine-insensitive protein 1 (*COI 1)* believed to take part in the jasmonic acid signaling cascade during plant defense. Within the WM2.2 QTL region, the pathogenesis-related protein chalcone synthase (*ChS)* was identified. A candidate gene encoding a peroxidase was found on WM3.1, while a gene coding for an MYB domain protein was found on WM5.4. When selecting for white mold tolerance, breeders will typically introgress both physiological resistance and avoidance genes.

2.1.2 Progress in breeding for Quantitative traits in common bean breeding

Most traits of interest in plant breeding are quantitative and will display a measurable phenotype, such as plant height. The variation in a trait may be partially explained by regions in the genome known as QTL (Doerge, 2002). Environmental factors may also contribute to variation in a quantitative trait. QTL may have large or small effects. For example, Mendelian loci are discrete with large effects. Essentially, a single gene is responsible for trait. On the other hand, numerous small effect QTL may determine the phenotype. In these cases, detection of QTL comes with challenges.

2.1.6 The breeder's equation

An important concept in plant breeding is 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:

$$\Delta G = h^2 \times \sigma_a \times \frac{i}{L} \quad [2.1]$$

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). Due to the complexity of breeding programs, the breeder's equation is used as a basis for which the simulation studies were conducted. The data obtained from the study 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

2.1.7 Objectives

Improvement of dry beans continues to be a challenge amidst rising global populations. 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. Plant breeders can make use of computer simulations to assist in decision making. The simulation platform QU-GENE was used to simulate the outcomes of different breeding strategies and selection intensities.

The following hypotheses were tested:

1. Simulated breeding strategies (mass selection, bulk breeding, single seed descent, pedigree method, and the modified pedigree method) will significantly differ in terms of genetic gain, percentage of fixed favourable alleles, and Hamming distance
2. Higher initial parental population size and trait heritability will lead to increased genetic gain and percentage of fixed favourable alleles
3. New proposed methods for plant breeding (genomic selection and speed breeding) will outperform conventional breeding methods in terms of genetic gain, allele fixation rate, and Hamming distance

2.2 Methods

2.2.1 Breeding strategies and new proposed methods of plant breeding

There are a number of breeding strategies available to plant breeders. Well-known conventional breeding strategies include bulk breeding, single seed descent, mass selection, the pedigree method, and the modified pedigree method. These conventional strategies rely solely on phenotypic selection. In recent years, new proposed breeding methods have begun to emerge, namely, speed breeding and genomic selection. These methods have garnered more popularity in

the literature due to promises of enhancing genetic gains. 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 from all markers across a genome in order to select on genotypes. This allows for selection to take place before a plant has reached maturity. For example, using genomic selection, a plant breeder may genotype entire germplasms to select against poor performing lines. This saves the time and resources that would have been required to assess the phenotype of each germplasm accession.

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 then form the next generation of plants. This process is repeated for a number of generations until the multi-environment trial phase (Figure 2.1). 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 picked off and 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).

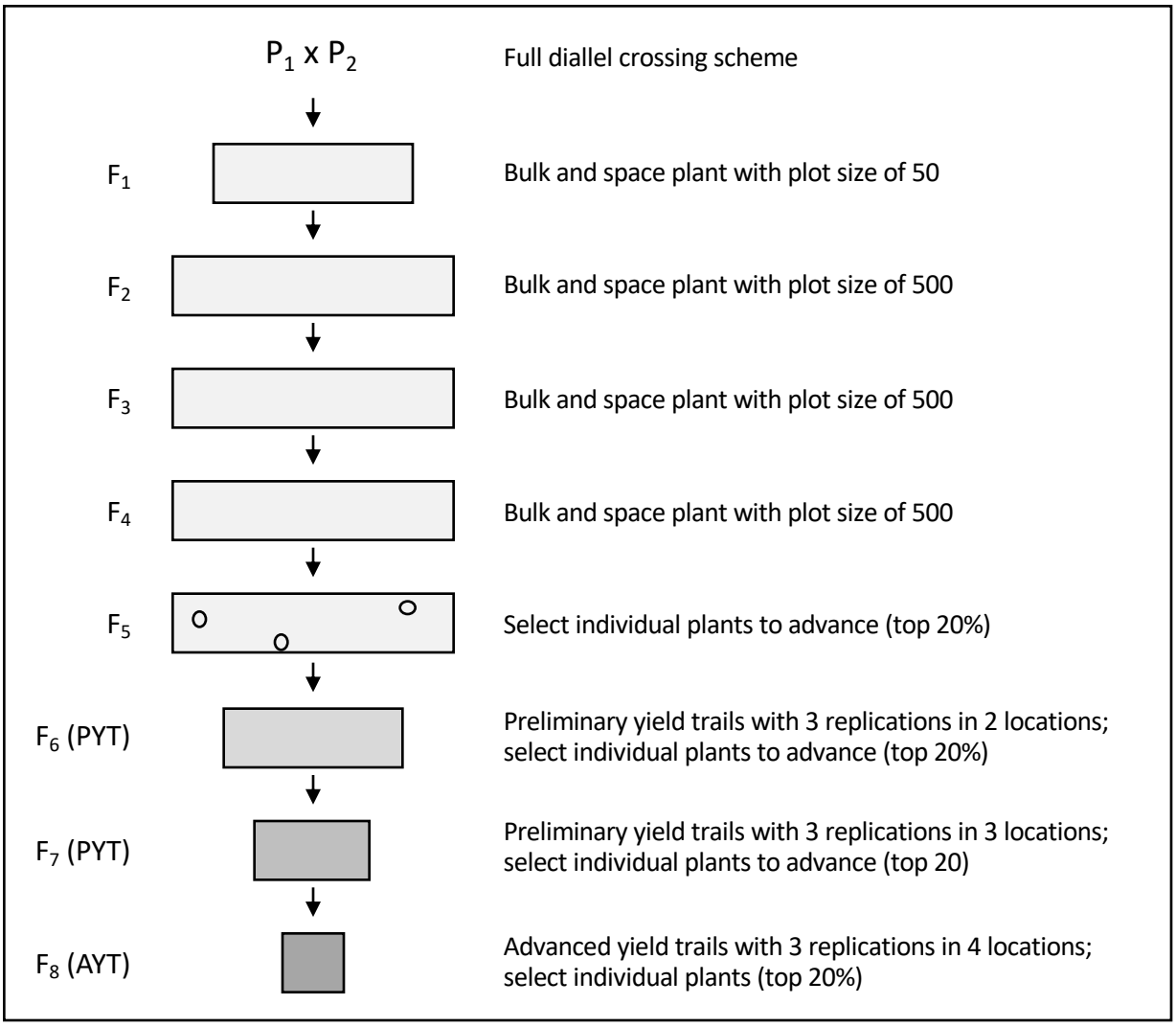

Figure 2.1: Mass selection breeding strategy

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 (Figure 2.2). 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).

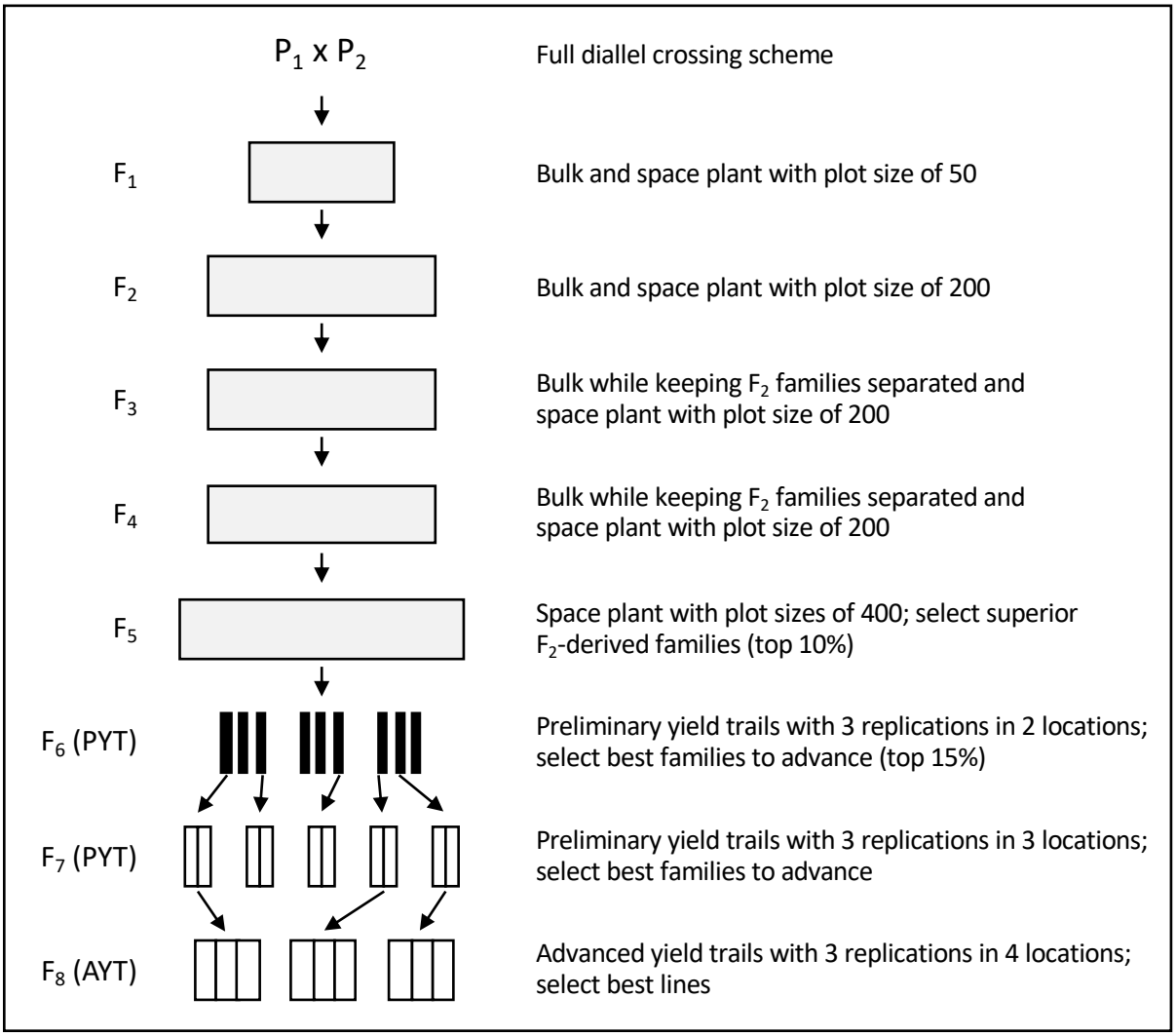

Figure 2.2: Bulk breeding strategy

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 (Figure 2.3). 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, so some F_2 plants will not be represented in the later generations (Acquaah, 2009).

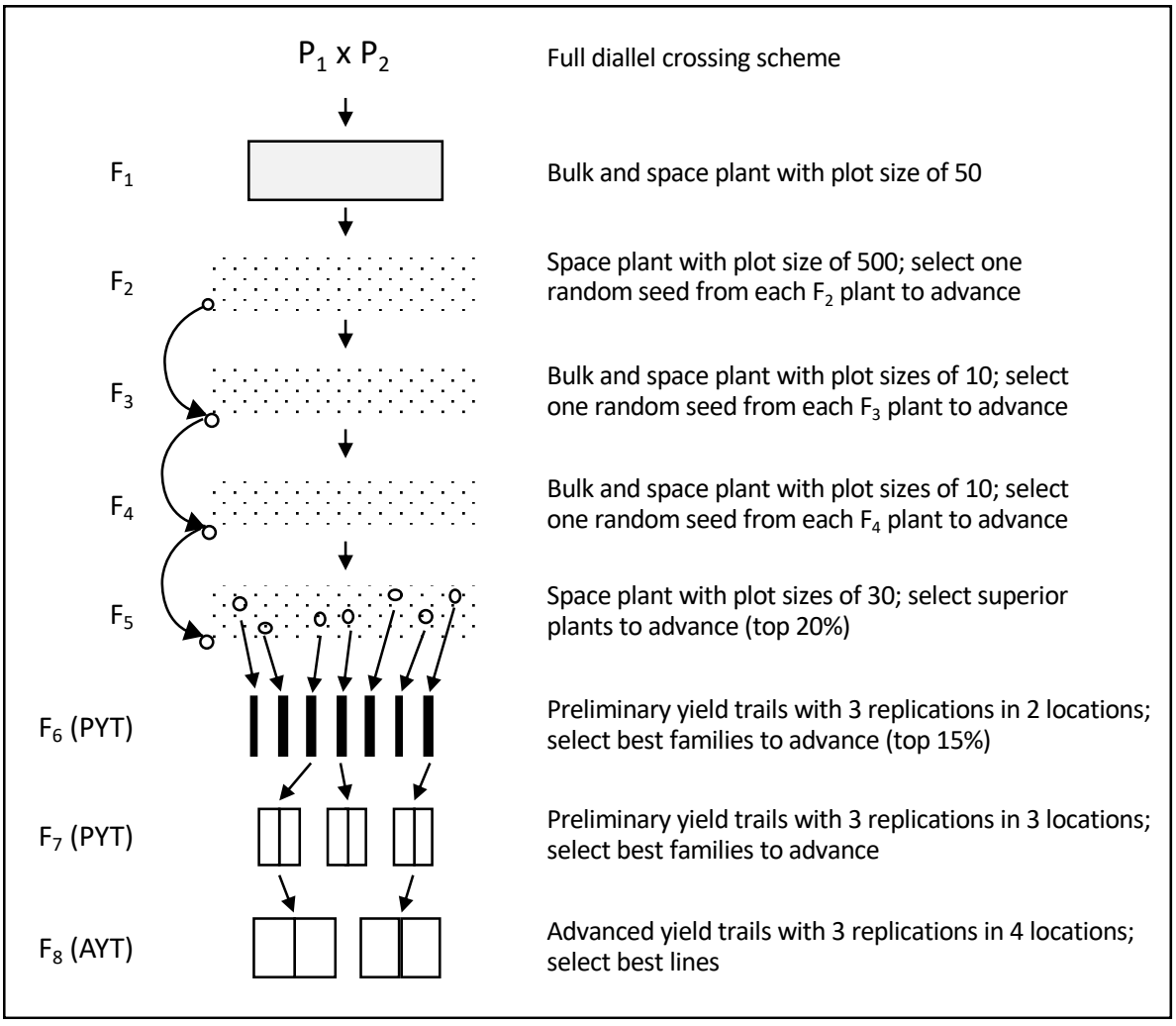

Figure 2.3: Single seed descent breeding strategy

Pedigree method

The pedigree method is a strategy whereby parent-progeny 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 in each generation begin 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 to as a family. Individual plants within rows or even entire rows may be selected (Figure 2.4). This continues until there is an acceptable level of homozygosity (Fehr, 1987). A benefit to using the pedigree method is that through record-keeping, valuable genetic information is now available to

plant breeders. 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).

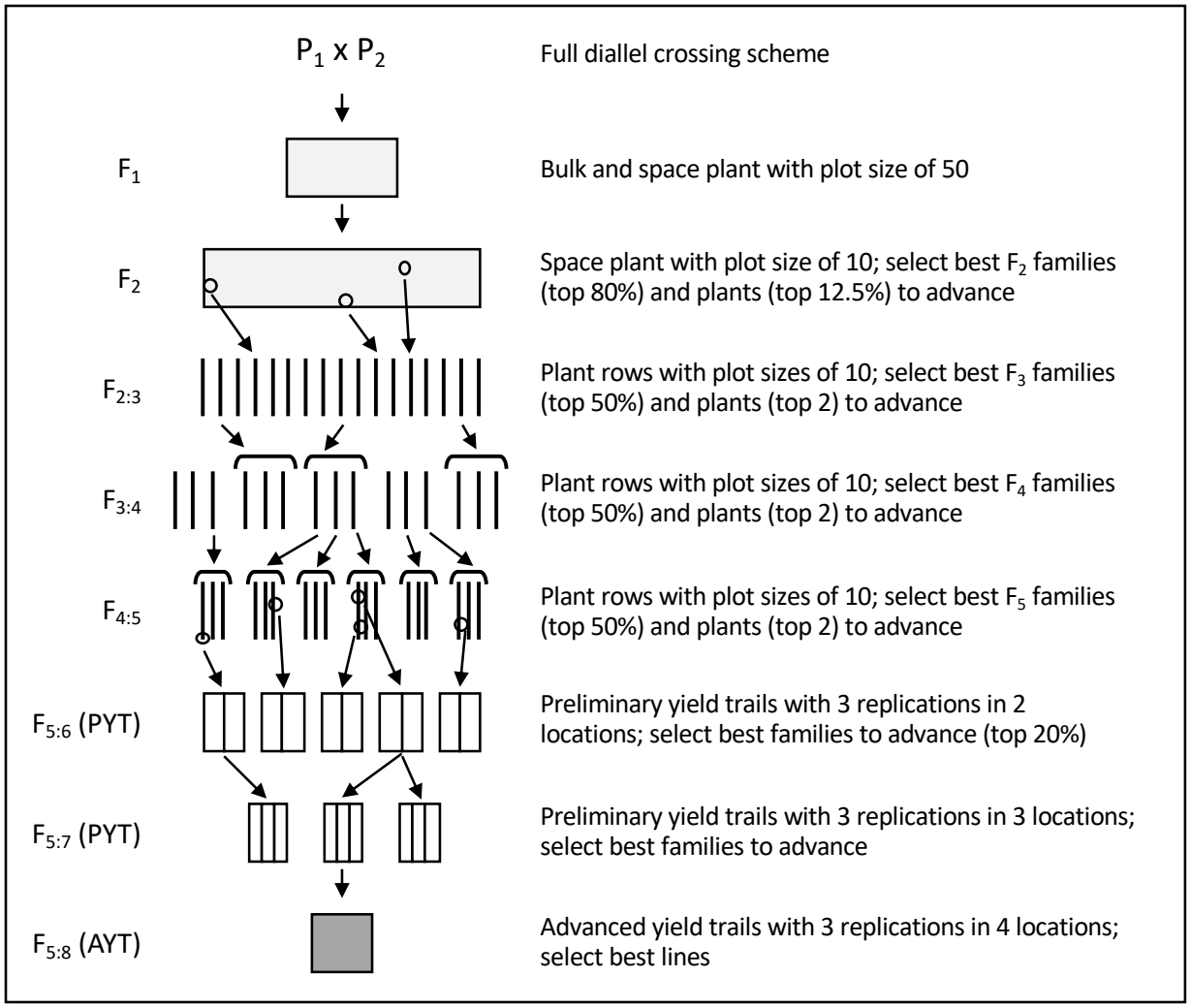

Figure 2.4: Pedigree method breeding strategy

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 (Figure 2.5). 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).

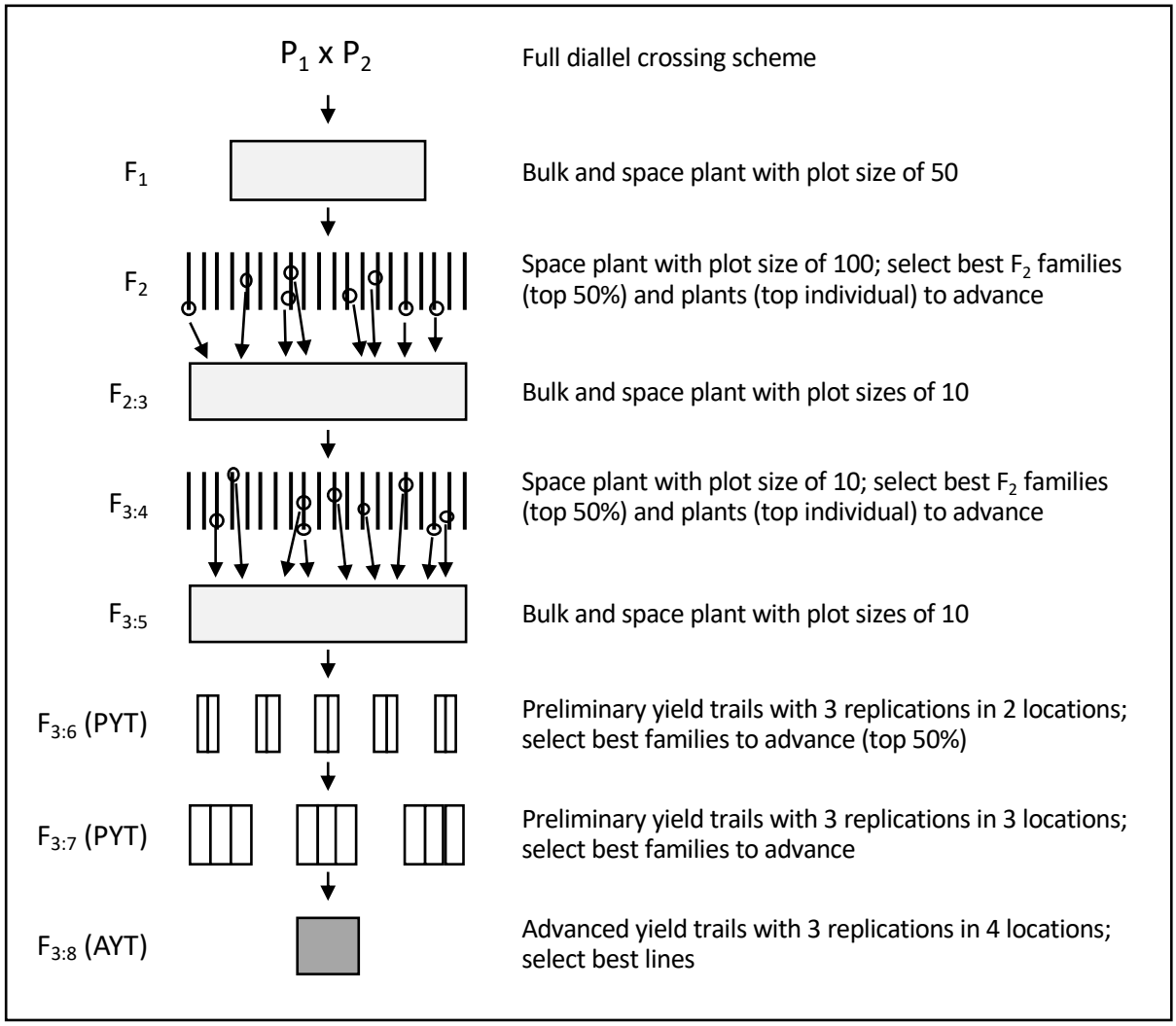

Figure 2.5: Modified pedigree method breeding strategy

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).

Genomic selection

First described by Meuwissen et al., (2001), genomic selection (GS) involves estimating the effects of all molecular markers and selecting on individuals based on their genomic estimated breeding value (GEBV) (Michel et al., 2016). Figure 2.6 shows a schematic for how GS is conducted. With a high number of markers, certain alleles will be correlated with a positive effect on a quantitative trait. The large number of markers also ensures that each QTL will be in LD with at least one marker (Goddard and Hayes, 2007; Nadeem et al., 2018). Markers that are close in proximity may be joined together as a haplotype. Individuals that have the same rare marker haplotype likely share a common ancestor and will have the same QTL allele (Meuwissen et al., 2001). To carry out GS, a training population is first created. The genotypic and phenotypic information of each individual is combined in the training population. A model is then "trained" on population, validated, and then applied to a testing population (Taylor, 2014). It is important to note that individuals in the testing population have not been phenotyped, only genotyped. The model will then predict a GEBV for each individual in the testing population (Crossa et al., 2017). GS is advantageous in that it can save time. Since only genotypic information is required for selection, individuals can be genotyped during early stages of development.

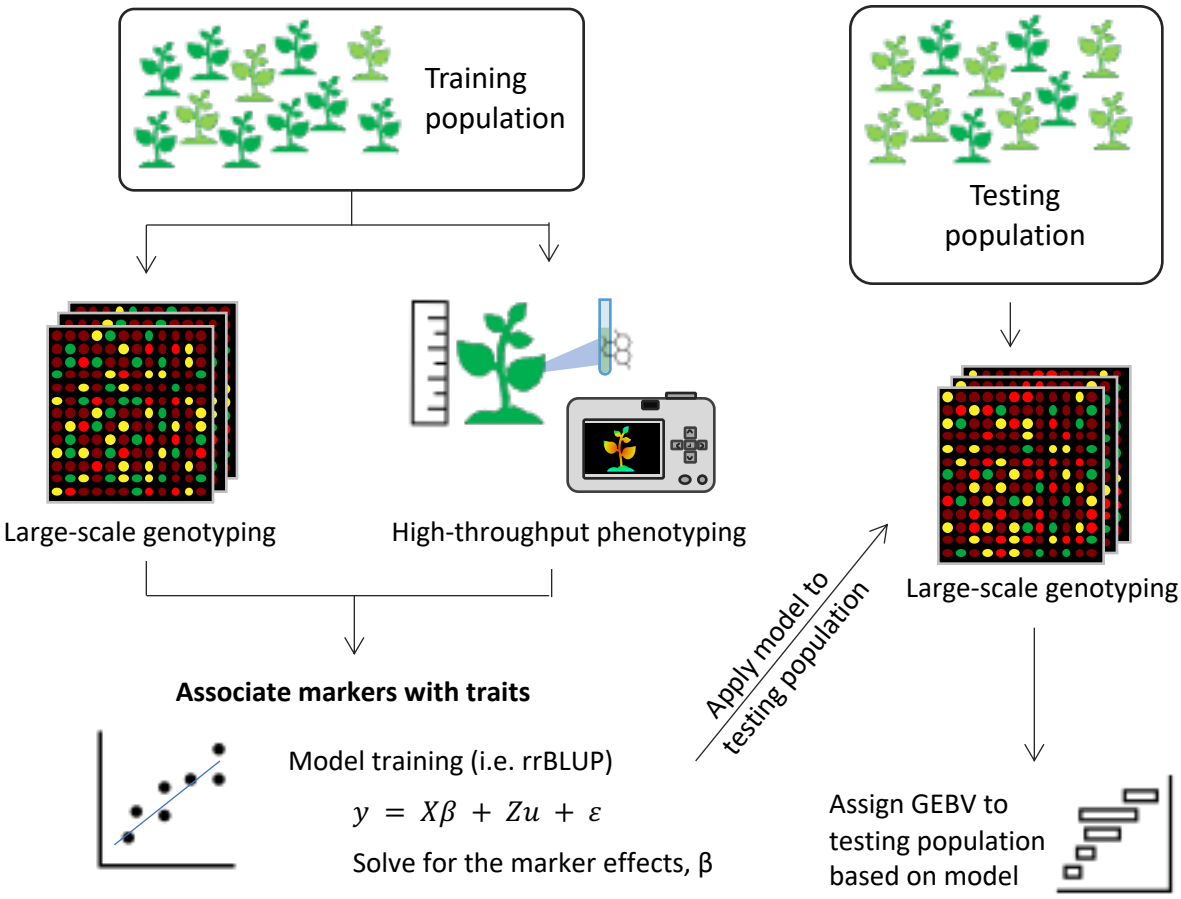

Figure 2.6: Genomic selection scheme

2.2.2 QU-GENE simulation workflow and simulation files

A number of simulations were conducted to compare four different numbers of initial parents, three different traits, three breeding methods, and five breeding strategies. Each simulation consisted of 10 cycles with 50 runs. A summary of the simulation criteria is displayed in Table 2.1. All of the files required by the simulation can be found on the lab GitHub page (McGill University Pulse Breeding and Genetics Laboratory, 2021).

Table 2.1: 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

Figure 2.7 shows the workflow in QU-GENE for the simulation of conventional breeding, as well as the new proposed breeding methods, which require additional steps.

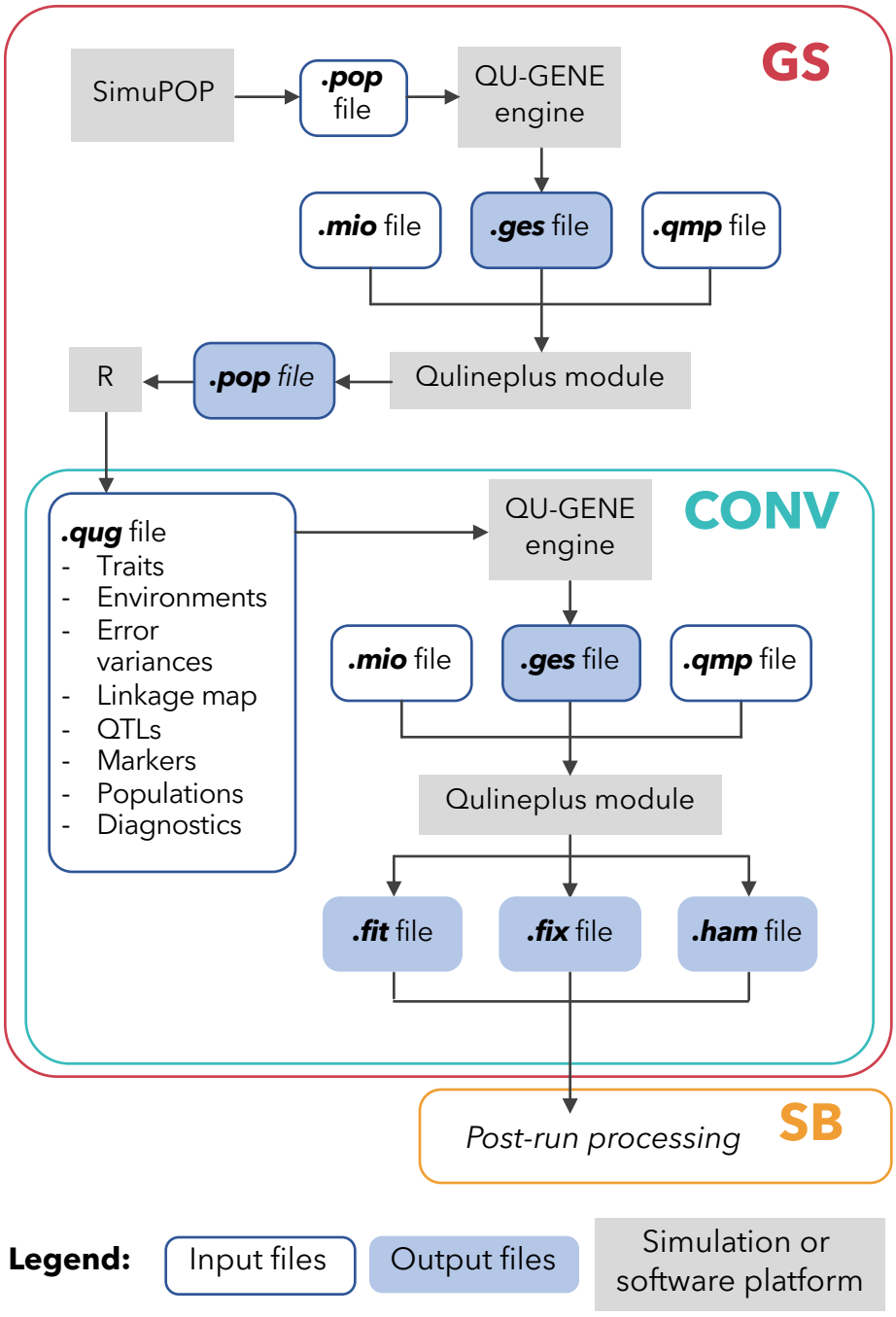

Figure 2.7: QU-GENE simulation workflow for the simulation of genomic selection (GS), conventional methods (CONV), and speed breeding (SB).

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 the narrow sense heritability reported for each trait from the literature. Heritability estimates obtained for each trait in each environment are summarized in Table 2.2. The linkage map, QTL, and markers described in a previous section were included in the ***.qug*** file. The population is automatically generated by QU-GENE. The diagnostic indicated that the file was error free and was able to be run in the QU-GENE engine.

Table 2.2: 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	0.6895	(Nienhuis & Singh, 1988)
	Field	0.92	(Atuahene-Amankwa, Beatie, Michaels, & Falk, 2004)
WM	Nursery	0.33	(Carneiro, Santos, Gonçalves, Antonio, & Souza, 2011)
	Winter nursery	0.65	(Carvalho, Lima, Alves, & Santos, 2013)
	Field	0.78	(Miklas et al., 2001)
SY	Nursery	0.21	(Jeffrey W. White & Singh, 1991)
	Winter nursery	0.29	(Mendes, Botelho, Ramalho, Abreu, & Furtini, 2008)
	Field	0.7	(Kolkman & 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 heritability, it was necessary to calculate these values based on the per plot heritability estimates reported in the literature. The following equation was used to determine the per plant heritability:

$$h^2{}_{per\ plant} = \frac{V_g}{V_g + \frac{V_e}{y} \times n} \quad [2.2]$$

where Vg is the genotypic variance, the phenotypic variance $V_p = \frac{1}{h^2{}_{per\ plot}}$, the error variance $V_e = V_p - V_g$, n is the plot size, and y is the year.

The ***.qmp*** file included information on the breeding 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 breeding strategy such as propagation type, generation advance method, number of replications, plot size, number of testing locations, and how selection was to be done. 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. The generation advance method indicated how the selected plants were harvested. This experiment will use 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 Table 2.3.

Table 2.3: Detailed steps for the breeding strategies specified in the .qmp file

Gen	Mass selection			Bulk breeding			Single seed descent			Pedigree method			Modified pedigree method		
	Harvest method	Family selection		Harvest method	Family selection		Harvest method	Family selection		Harvest method	Family selection		Harvest method	Family selection	
		Among	Within		Among	Within		Among	Within		Among	Within		Among	Within
CB	bulk			bulk			bulk			bulk			bulk		
F_1	superbulk			bulk			bulk			bulk			bulk		
F_2	bulk			bulk			pedigree		Random 1	pedigree	Top 80%	Top 12.5%	pedigree	Top 50%	Top 1
F_3	bulk			bulk			bulk		Random 1	pedigree	Top 50%	Top 2	bulk		
F_4	bulk			bulk			bulk		Random 1	pedigree	Top 50%	Top 2	pedigree	Top 50%	Top 1
F_5	bulk		Top 20%	bulk		Top 10%	bulk		Top 20%	pedigree	Top 50%	Top 2	bulk		
F_6	bulk		Top 20%	bulk	Top 15%		bulk	Top 15%		bulk	Top 20%		bulk	Top 50%	
F_7	bulk		Top 20%	bulk	Top number		bulk	Top number		bulk	Top number		bulk	Top number	
F_8	superbulk		Top 20%	bulk	Top number		bulk	Top number		bulk	Top number		bulk	Top number	

The QuLinePlus module was used to simulate the breeding strategies. It is capable of simulating both self-pollinating and cross-pollinating species, making it quite versatile (Hoyos-Villegas et al., 2019). Output files obtained from the Qu-Gene engine were used as input files for QuLinePlus.

As the simulations required a high level of computing power, they were performed remotely on servers provided by Compute Canada (Digital Research Alliance of Canada, 2020). Access to remote servers required establishing a secure shell via the terminal on MacOS. To browse and manipulate files, the cloud storage browser, Cyberduck was used (iterate GmbH, 2020).

2.2.3 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. A second reason for the use of the consensus linkage map is that QU-GENE does not accept maps with physical distances.

A total of 38 QTL found in the literature were considered for this study, more specifically 11 seed yield QTL, 8 white mold disease incidence QTL, and 19 days to flowering QTL were selected (Table 2.4). Seed yield QTL effect sizes ranged from -36.91 to -197.46. Effect sizes for white mold disease incidence QTL ranged from 3.16 to -7.2. Lastly, QTL effect sizes in days to flowering ranged from 0.68 to -1.21. 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. 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 2*s*, Aa would have effect size 0, and aa would have effect size -2*s*.

Table 2.4: Description of QTLs used in the simulation

Trait	QTL name	Linkage group	Position (cM)	Effect size	Mapping population	Reference
DF	DF41	4	167.11	0.68	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	DF51	5	45.21	0.45	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	DF52	5	56.71	0.49	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	DF53	5	82.21	0.46	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	DF54	5	105.21	0.43	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	DF11a	11	96.51	-0.6	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	DF11b	11	108.71	-0.49	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	EM86	2	21.6	0.57	Bunsi × Newport	(Ender & Kelly, 2005)
	EM78	7	1.1	-0.6	Bunsi × Newport	(Ender & Kelly, 2005)
	EM550	7	13.6	-0.96	Bunsi × Newport	(Ender & Kelly, 2005)
	EM223	7	8.6	-1.21	Bunsi × Newport	(Ender & 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)
	DFmn1	1	16.9	-0.8	AN-37 × P02630	(Hoyos-Villegas, Mkwaila, Cregan, & Kelly, 2015)
	DFmn2	1	105.7	-0.8	AN-37 × P02630	(Hoyos-Villegas et al., 2015)
WM	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 & Kelly, 2005)
	DSI2	2	21	-2.66	Bunsi × Newport	(Ender & Kelly, 2005)
	DSI3	5	27.7	3.16	Bunsi × Newport	(Ender & Kelly, 2005)
	DSI4	7	8.6	-4.17	Bunsi × Newport	(Ender & Kelly, 2005)
	DSI5	7	14.8	-4.01	Bunsi × Newport	(Ender & Kelly, 2005)
	DSI6	8	1.4	2.93	Bunsi × Newport	(Ender & Kelly, 2005)
SY	Yd21	2	151.2	-46.88	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	Yd71	7	35.1	-36.91	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)
	Yd72	7	47.8	-97.3	DOR 364 × BAT 477	(Lucy M Diaz et al., 2017)

syMO14	3	113.7	-153.6	BK004-001 × H68-4	(Sandhu, You, Conner, Balasubramanian, & Hou, 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)
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)

2.2.4 Model for genomic selection

$$y = X\beta + Zu + \varepsilon \quad [2.3]$$

The model used to determine the marker effects in genomic selection is shown in Equation 2.3, where u ~ $N(0, K\sigma^2{}_u)$, 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 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 & 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.

2.2.5 Simulating LD through SimuPOP

By default, QU-GENE will generate populations in Hardy-Weinberg equilibrium with little to no linkage disequilibrium (LD). This is an issue for simulating genomic selection since adequate LD is necessary for markers to be linked to QTL. LD can be formally defined as a non-random association between alleles found at different loci (Flint-Garcia, Thornsberry, & Buckler, 2003).

Two popular methods for estimating LD make use of the parameters D' and r^2. For verifying LD measures, the r^2 parameter was used. To understand how r^2 is calculated, one may consider the following example. 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:

$$D_{AB} = P_{AB} - P_A P_B \quad [2.4]$$

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

$$r^2 = \frac{(D_{AB})^2}{P_A P_B P_a P_b} \quad [2.5]$$

There are a number of 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, the 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 a population with an adequate level of LD, the forward-in-time simulation tool, SimuPOP was used. SimuPOP is implemented in python. Supplemental code can be found on the lab GitHub page (McGill Pulse Breeding and Genetics Lab, 2021). The program can be used to evolve a population over time *in silico*. By allowing a population to undergo natural selection via the simuPOP program, 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 population generated by QU-GENE had essentially no LD (Figure S2.1), while the one generated by simuPOP had substantial LD (Figure S2.2).

2.2.6 Handling simulation output data

QU-GENE produces a number of 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 Equation 2.6, where F is the fitness, TG_h is the highest target genotypic value, and TG_l is the lowest target genotypic value**.**

$$F_{Ad} = \frac{F - TG_l}{TG_h - TG_l} \times 100 \quad [2.6]$$

The adjusted genetic gain can then be calculated as the difference from one cycle to the next, as shown in Equation 2.7, 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.

$$\Delta G_{Ad} = F_{Ad(n)} - F_{Ad(n-1)} \quad [2.7]$$

The ***.fix*** file reports the percentage of fixed favourable and unfavourable alleles after each cycle. sThis 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; C. Wang, Kao, & Hsiao, 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.

2.2.7 Statistical analysis

A multi-way ANOVA was performed based on a mixed model which was defined using the lme4 package in R. 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 **ggbiplot** package in R. All other figures were created using **ggplot2** package in R.

2.2.8 Model

The following equation specifies the general formula for ANOVA:

$$y_{ijkmn} = u + parent_i + framework_{ij} + strategy_{ijk} + cycle_{ijkm} + e_{ijkmn} \quad [2.8]$$

The terms of the model are defined by the following:

y_{ijkmn}: the reported genetic gain variable in the n^{th} run of the m^{th} cycle of the k^{th} strategy of the j^{th} framework of the i^{th} parental population size

i: parental population size; i= 15, 30, 60, 100
j: framework; j=1, 2, 3 which corresponds to conventional breeding, speed breeding, GS
k: strategy; k=1, 2, 3, 4, 5 (corresponds to mass selection, bulk breeding, single seed descent, pedigree method, and the modified pedigree method
m: cycle; m= 1, 2, 3…10
n: a run

u: overall genetic gain variable irrespective of cycle, strategy, framework, and parental population size

$parent_i$: the fixed effect of the i^{th} parental population size on the genetic gain variable in a run

$framework_{ij}$: the fixed effect of the j^{th} framework on the genetic gain variable of a run. The framework is nested within the parental population size

$strategy_{ijk}$: the fixed effect of the k^{th} strategy on the genetic gain variable of a run. The strategy is nested within the framework, which is nested within the parental population size.

$cycle_{ijkmn}$: the fixed effect of the m^{th} cycle on the genetic gain variable of a run. The cycle is nested within the strategy, which is nested within the framework, which is nested within the parental population size.

e_{ijkmn}: the random residual associated with the nth run of the m^{th} cycle of the k^{th} strategy of the j^{th} framework of the i^{th} parental population size.

$e_{ijkmn} \sim N(0, \sigma^2_e)$

The parameters of the model are defined as follows:

u, $parent_i$, $framework_{ij}$, $strategy_{ijk}$, $cycle_{ijkm}$: fixed effects

σ^2_e : random effect

The nested model was compared to an unnested model in terms of goodness of fit, which was dictated by AIC and BIC scores. According to these scores the nested model led to a greater goodness of fit (Tables S2.1, S2.3, S2.5, and S2.7)

2.3 Results

2.3.1 Genetic variance

The breeding strategies and methods 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 (Table S2.2). In general, the relative genetic variance saw a decrease over the five cycles. For days to flowering, as the number of initial parents increased, less relative genetic gain was maintained (Figure 2.8). Similar trends were observed for white mold tolerance (Figure 2.9) and seed yield (Figure 2.10). 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.

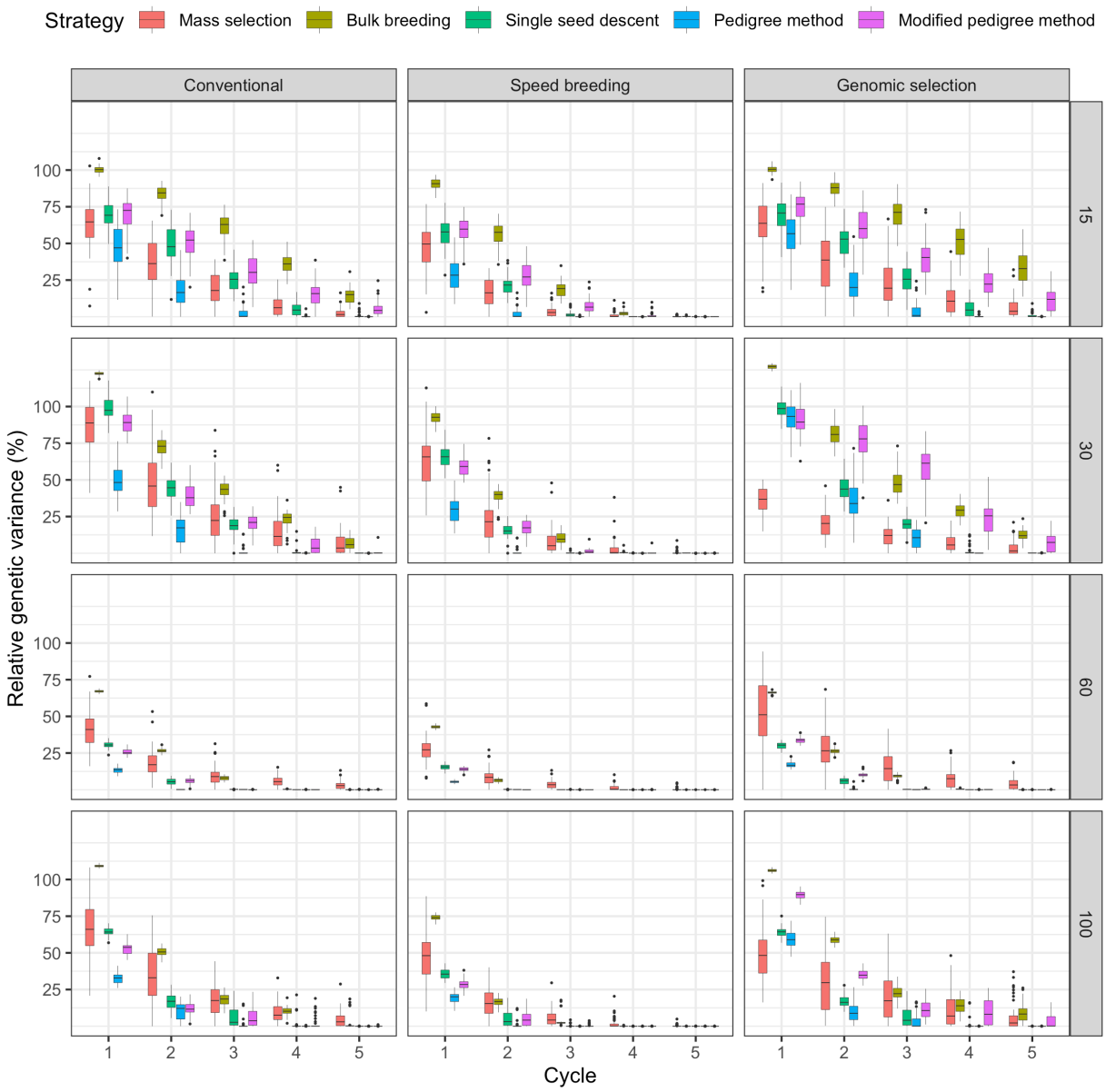

Figure 2.8: Comparison of five breeding strategies in terms of genetic variance over 5 cycles of selection across 50 runs in a closed system. Selection for days to flowering was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies included mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Genetic variance is relative to cycle 0, which is 100%.

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.

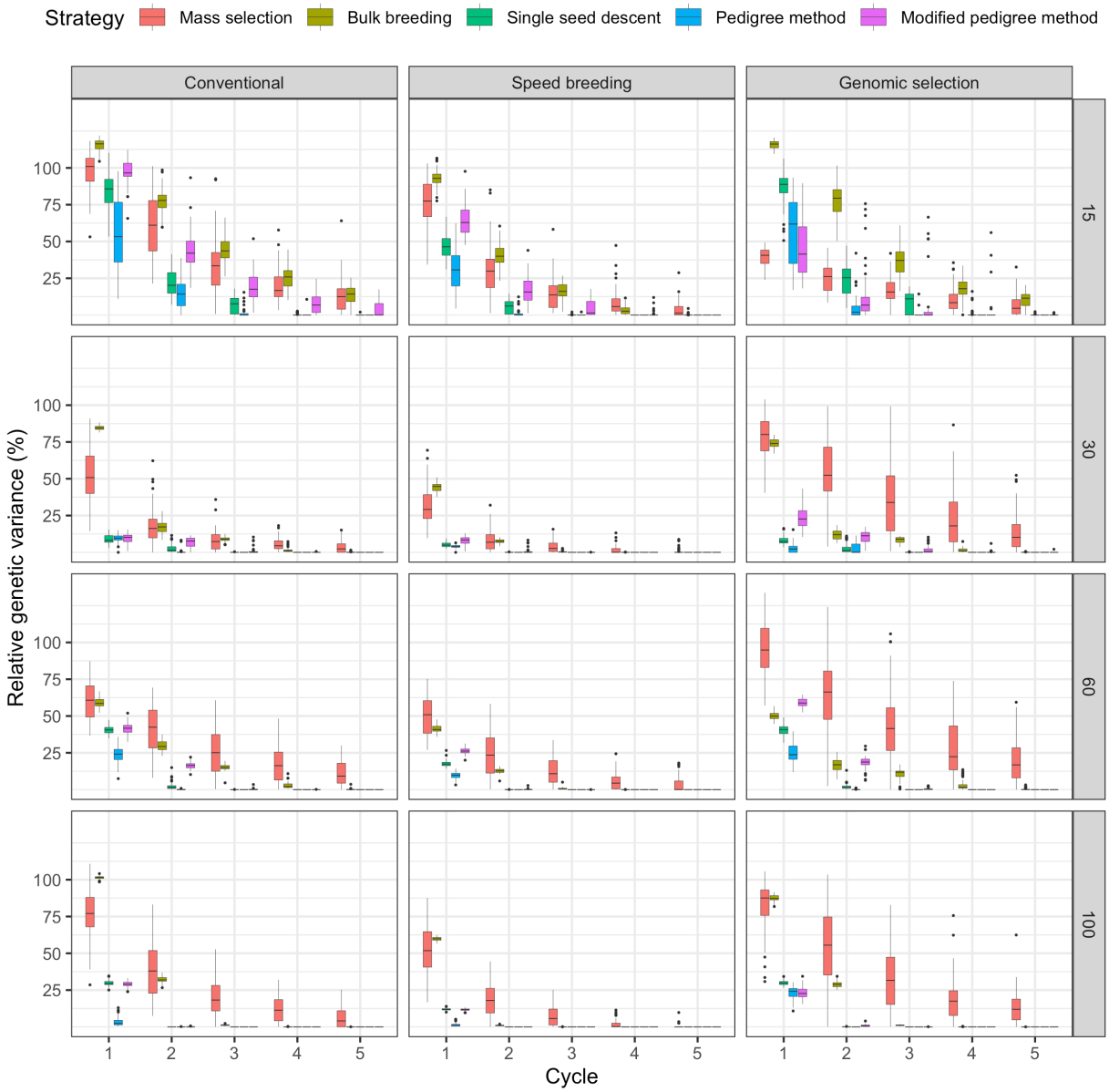

Figure 2.9: Comparison of five breeding strategies in terms of genetic variance over 5 cycles of selection across 50 runs in a closed system. Selection of white mold tolerance was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Genetic variance is relative to cycle 0, which is 100%.

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.

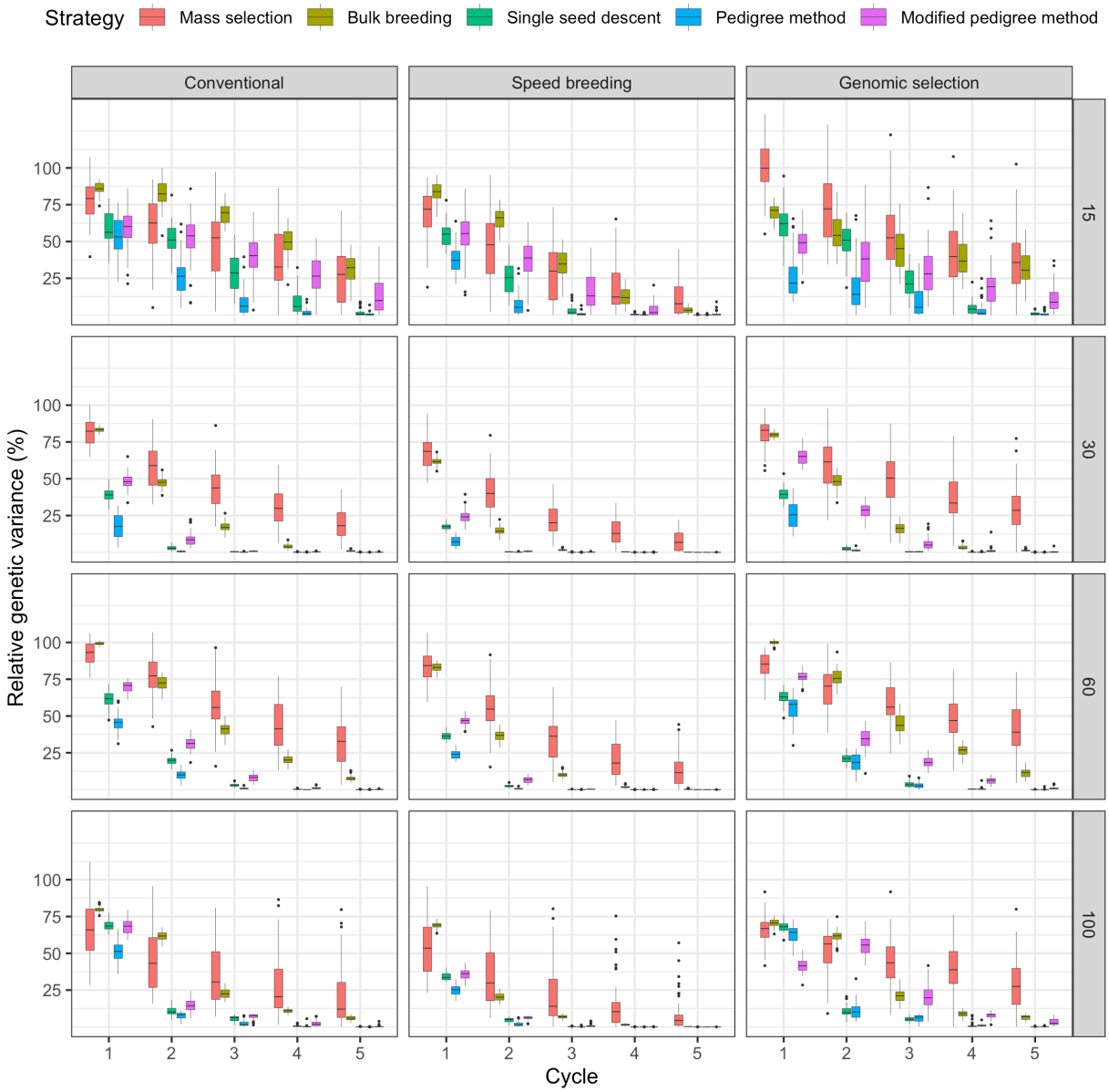

Figure 2.10: Comparison of five breeding strategies in terms of genetic variance over 5 cycles of selection across 50 runs in a closed system. Seed yield selection was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Genetic variance is relative to cycle 0, which is 100%.

2.3.2 Fixation of favourable alleles and Hamming distance

The fixation of favourable alleles was plotted over 10 cycles. The ANOVA demonstrated that the strategy, framework, and parental population size were statistically significant (Table S2.4). Figures 2.11, 2.12, and 2.13 display the plots for the fixation of favourable alleles in days to flowering, white mold tolerance, and seed yield, respectively. For days to flowering, 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.

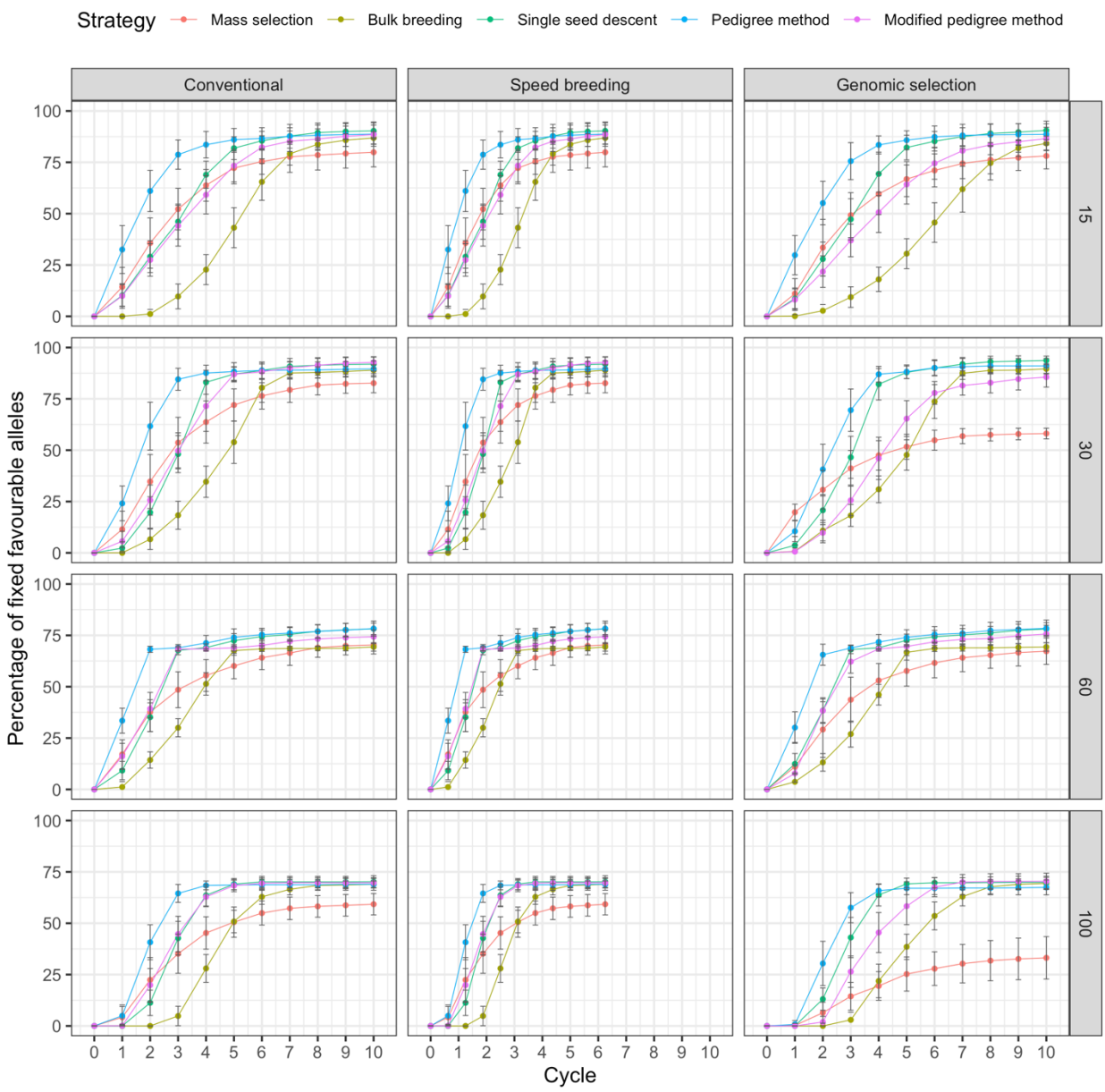

Figure 2.11: Comparison of five breeding strategies in terms of fixation of favourable alleles over 10 cycles of selection across 50 runs in a closed system. Selection for days to flowering was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

For white mold tolerance, multiple scenarios led to 100% of favourable alleles being fixed. 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.

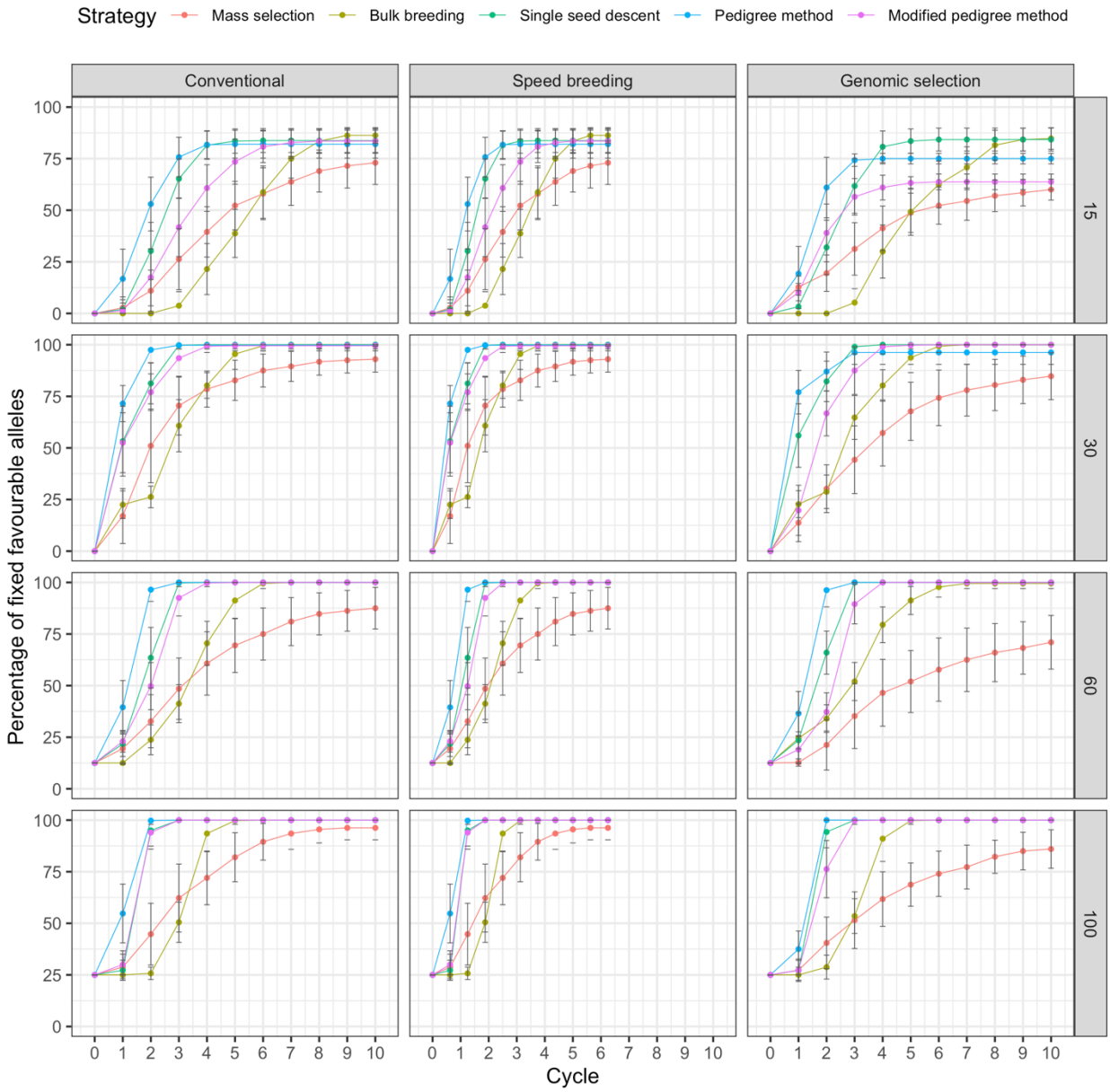

Figure 2.12: Comparison of five breeding strategies in terms of fixation of favourable alleles over 10 cycles of selection across 50 runs in a closed system. Selection for white mold tolerance was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

For seed yield, a parental population size of 15 resulted in the greatest fixation of alleles.

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.

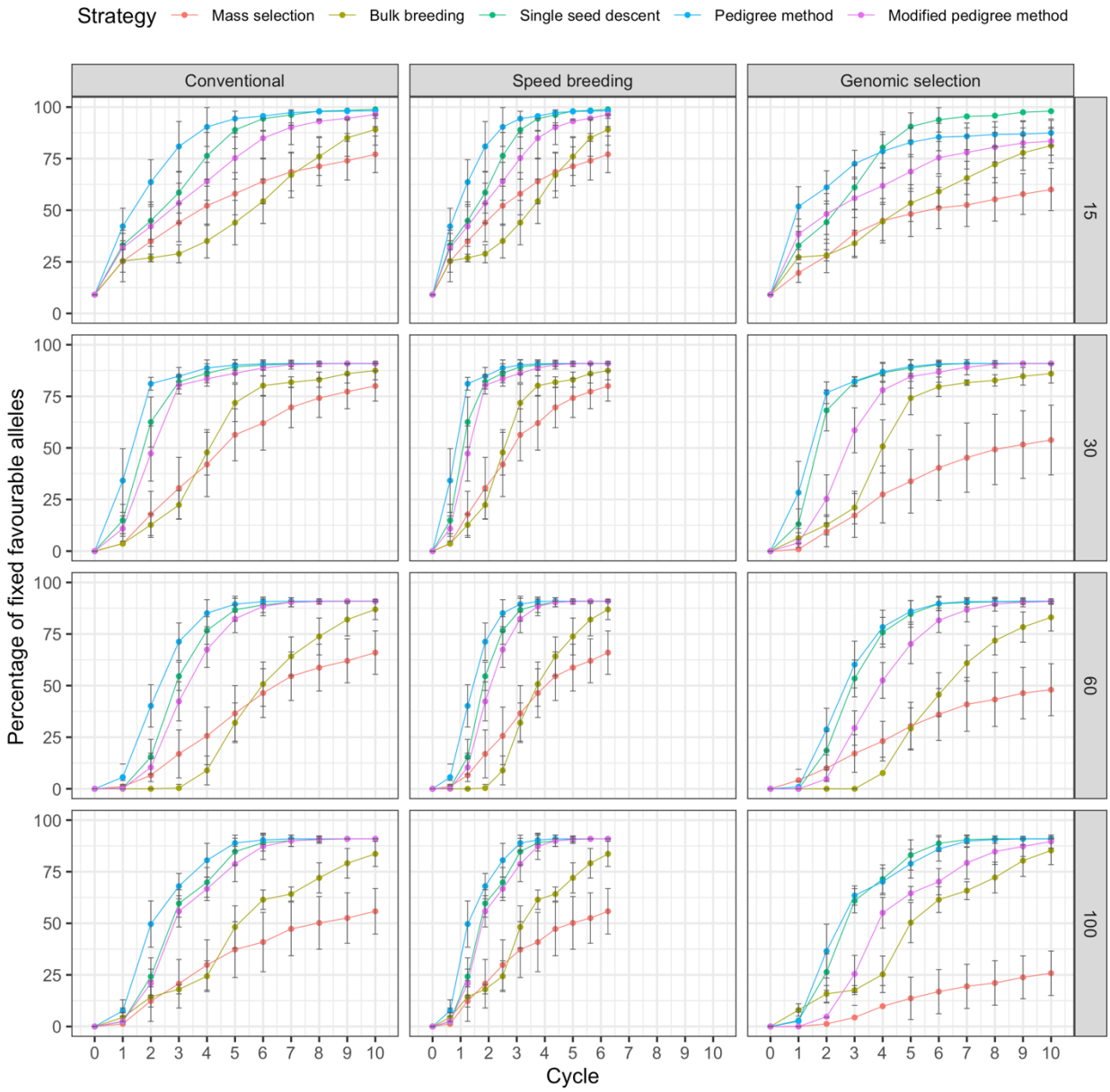

Figure 2.13: Comparison of five breeding strategies in terms of fixation of favourable alleles over 10 cycles of selection averaged across 50 runs in a closed system. Selection for seed yield was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

The plots for average Hamming distance are displayed in Figures 2.14, 2.15, and 2.16. The ANOVA for Hamming distance indicated that the strategy, framework, and parental population size were all statistically significant (Table S2.6). Overall, the Hamming distance had a general decreasing trend which eventually plateaued. For days to flowering, the 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.

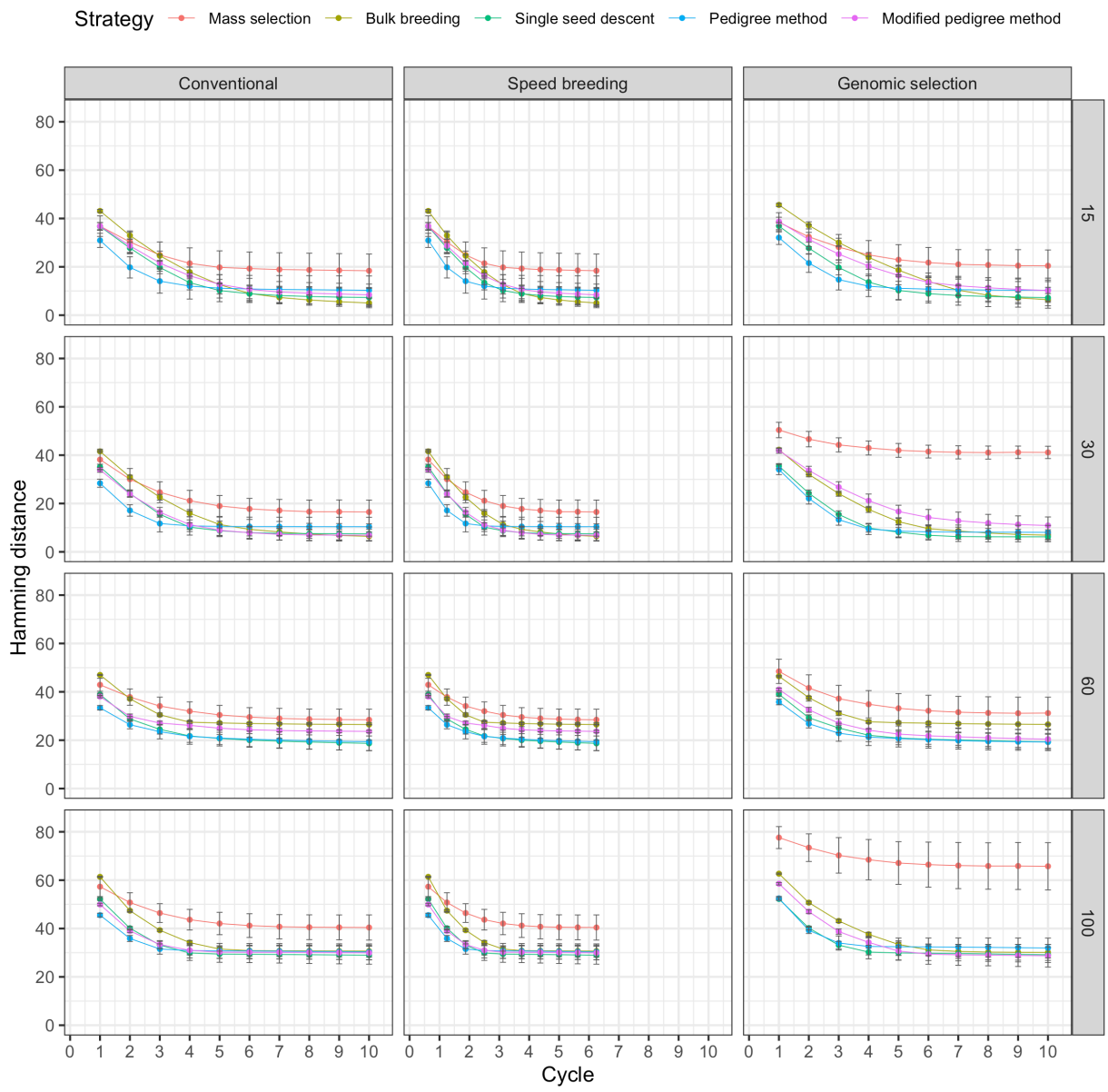

Figure 2.14: Comparison of five breeding strategies in terms of Hamming distance over 10 cycles of selection averaged across 50 runs in a closed system. Selection for days to flowering was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

For white mold tolerance, 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, with the exception of mass selection, resulted in the same Hamming distance after 10 cycles.

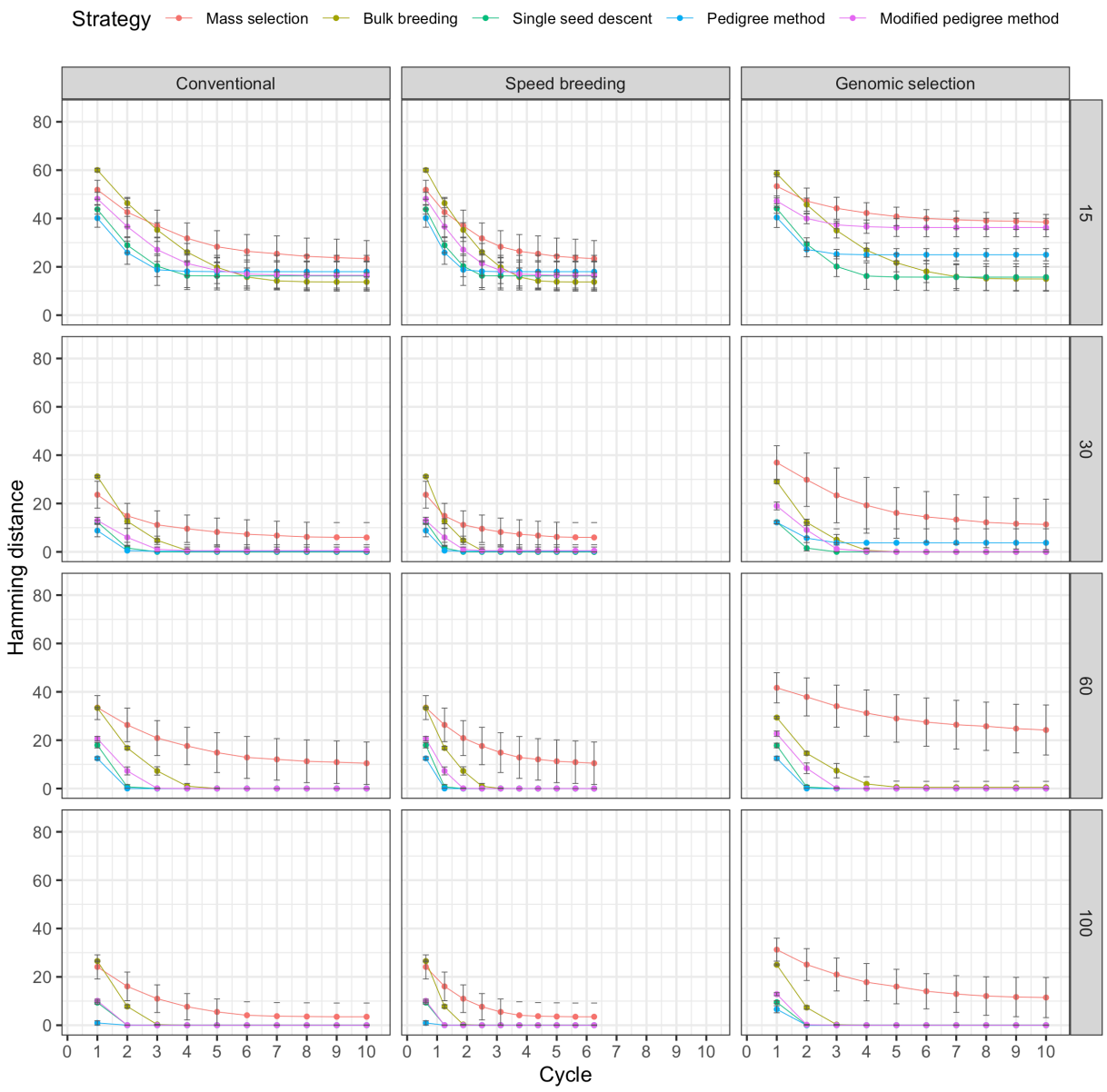

Figure 2.15: Comparison of five breeding strategies in terms of Hamming distance over 10 cycles of selection averaged across 50 runs in a closed system. Selection for white mold tolerance was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

For seed yield, 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.

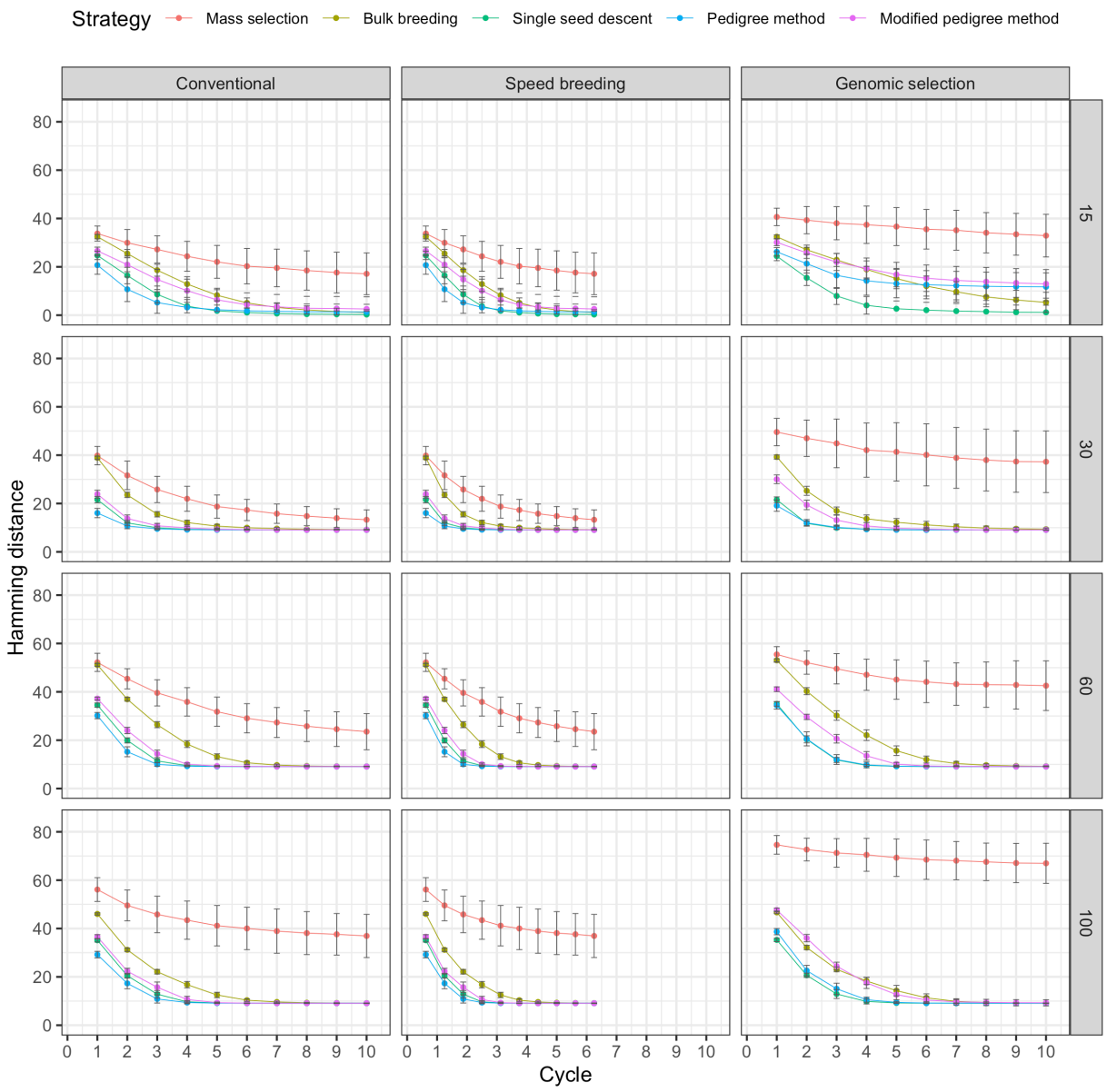

Figure 2.16: Comparison of five breeding strategies in terms of Hamming distance over 10 cycles of selection averaged across 50 runs in a closed system. Seed yield selection was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Error bars indicate standard error.

2.3.3 Genetic gain

The relative genetic gain averaged across runs was determined for each cycle for the various simulation scenarios. The ANOVA revealed that the strategy, framework, and parental population size were all statistically significant (Table S2.8). Figure 2.17 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. There was a general decreasing trend for the average genetic gain, where it eventually plateaued at 0. The cumulative genetic gain was greater in conventional and speed breeding compared to genomic selection for all parental population sizes. Figure 2.18 displays a similar plot for white mold tolerance, while Figure 2.19 shows the plot for seed yield. The parental population size of 100 led to the greatest percent cumulative genetic gain, followed by 30, 15, and 60. For days to flowering, the 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.

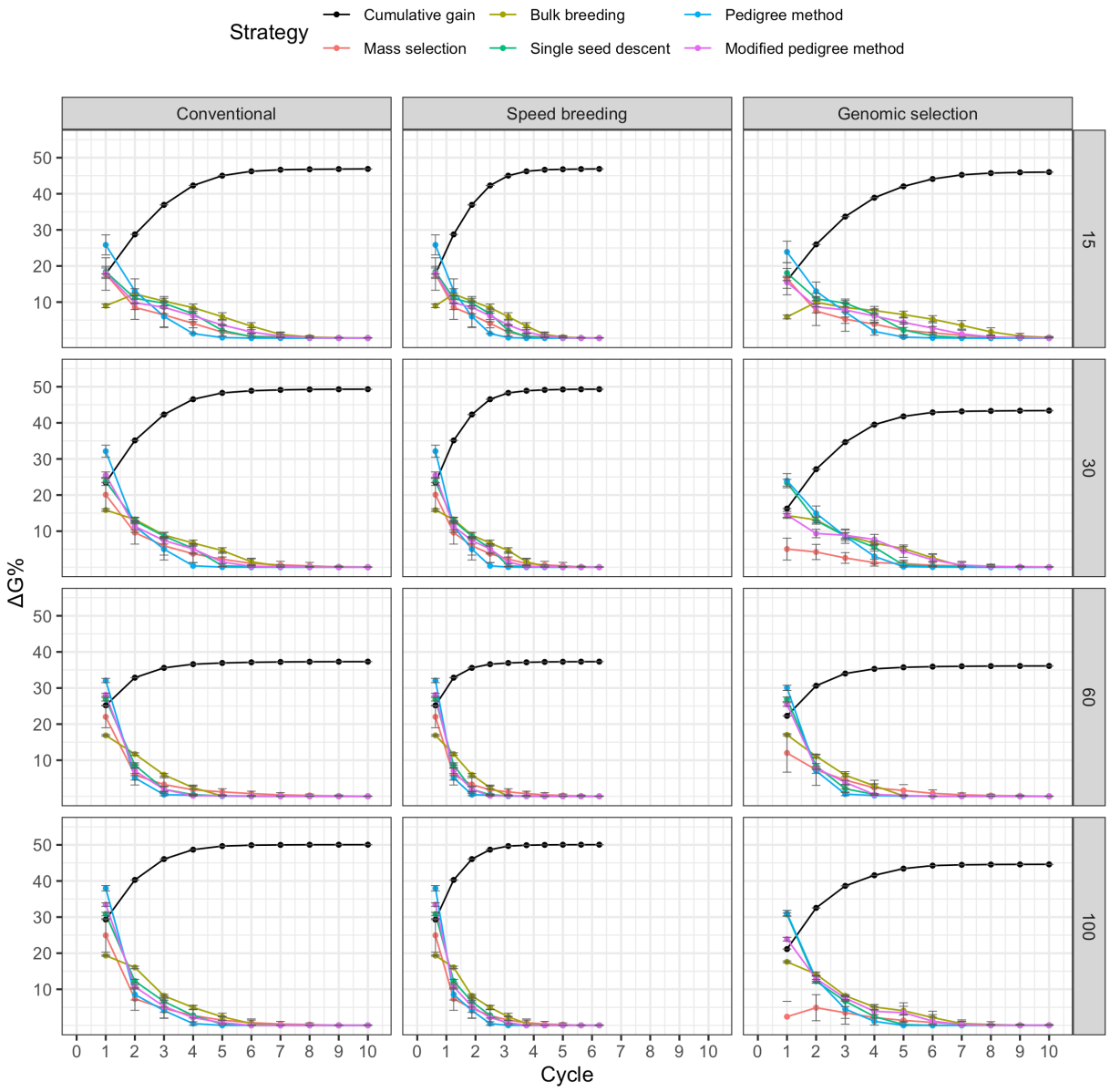

Figure 2.17: Comparison of five breeding strategies in terms of genetic gain over 10 cycles of selection averaged across 50 runs in a closed system. Selection for days to flowering was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at 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. Error bars indicate standard error.

For white mold tolerance, 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.

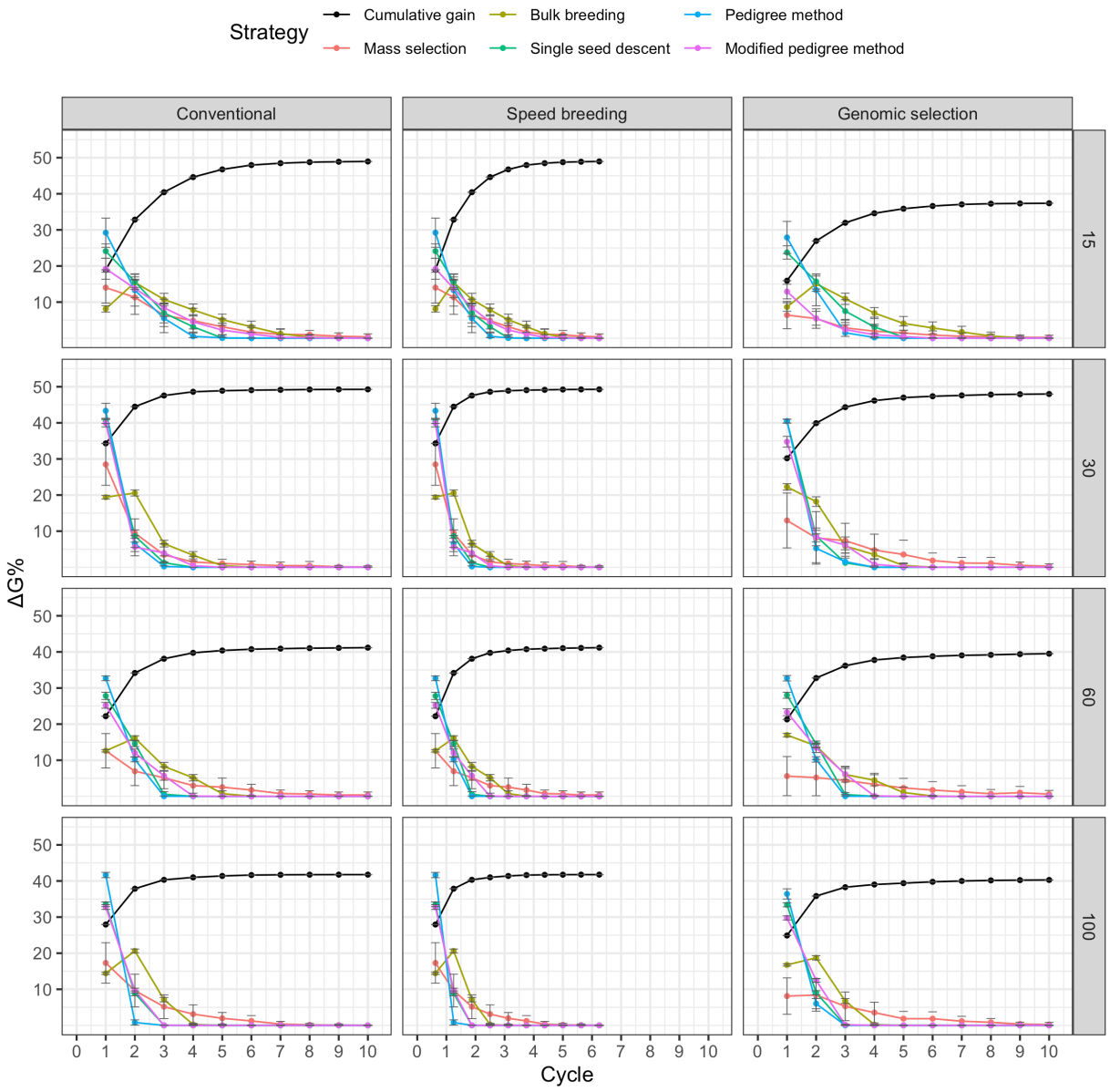

Figure 2.18: Comparison of five breeding strategies in terms of genetic gain over 10 cycles of selection averaged across 50 runs in a closed system. Selection for white mold tolerance was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown 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. Error bars indicate standard error.

For seed yield, 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.

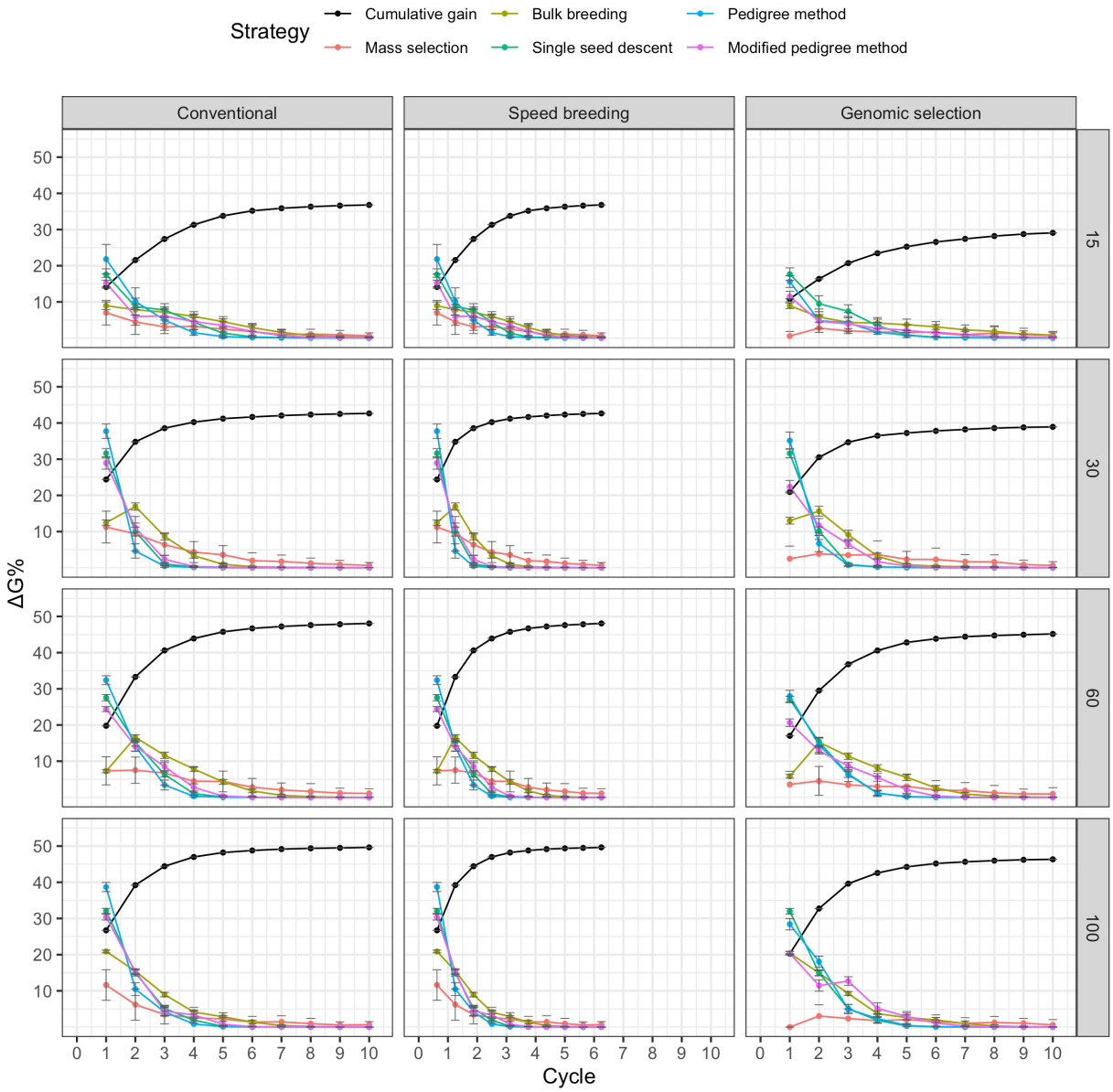

Figure 2.19: Comparison of five breeding strategies in terms of genetic gain over 10 cycles of selection averaged across 50 runs in a closed system. Selection for seed yield was simulated with increasing numbers of initial parents displayed on the right and differing breeding methods shown at the top. Breeding strategies include mass selection, bulk breeding, single seed descent, pedigree method, modified pedigree method. Cumulative genetic gain averaged across strategies indicated in black. Error bars indicate standard error.

The proportion of cumulative genetic gain was determined for each cycle when averaged across all strategies. The proportions were determined for the simulation of days to flowering. For conventional methods, by cycle five, on average the 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 6 cycles. In the simulation for improving white mold tolerance, 83 to 97% of cumulative genetic gain was achieved by cycle 3 for conventional methods. 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 2.20 shows the number of cycles required for 95% cumulative ΔG. On average across all scenarios, it took 3.31 cycles 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.

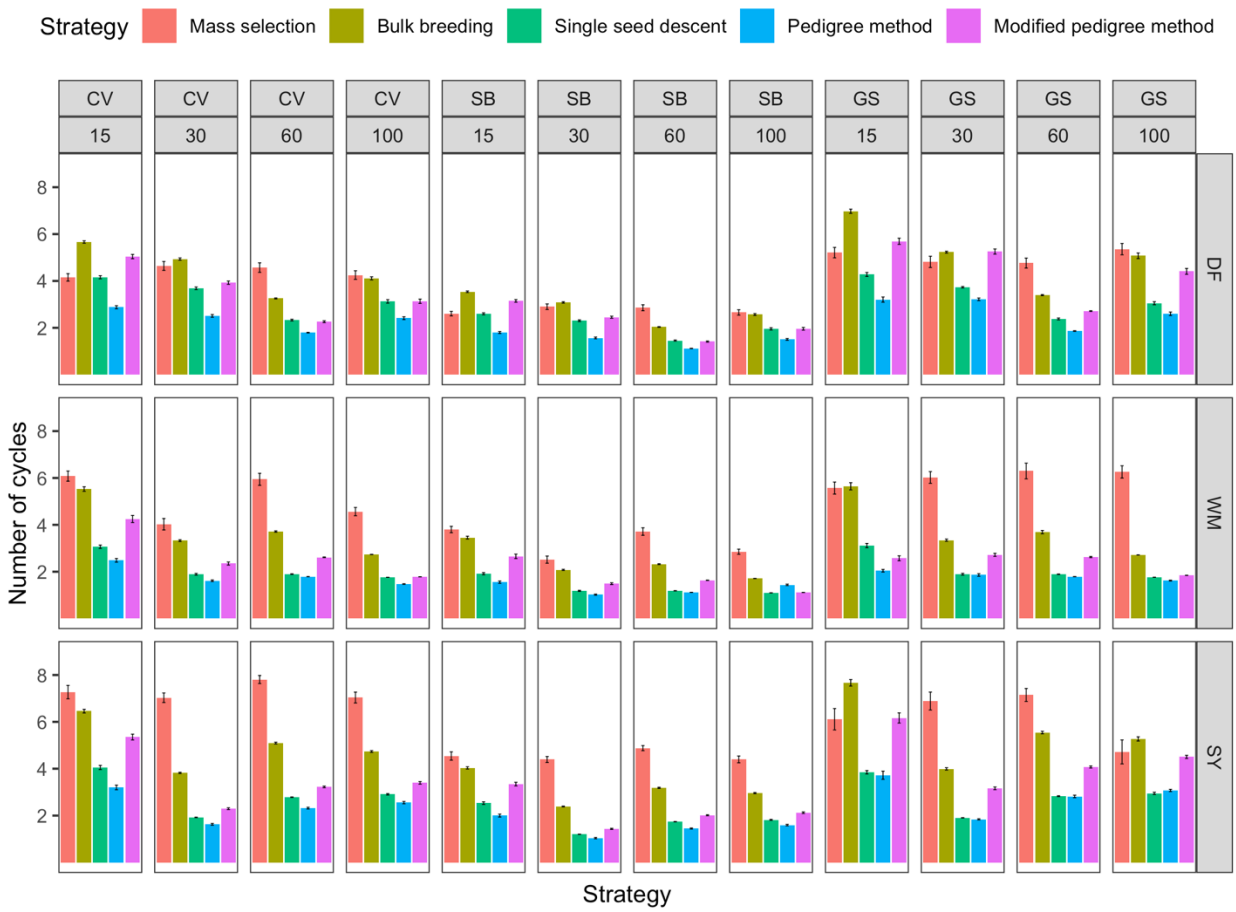

Figure 2.20: Comparison of five breeding strategies in terms of number of cycles until 95% cumulative of genetic gain for 10 cycles averaged over 50 runs in a closed system. Selected traits include days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Increasing numbers of initial parents displayed on the top along with different breeding methods. Breeding methods 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.

The average ΔG per cycle was determined for all scenarios (Figure 2.21). On average across all strategies, 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.

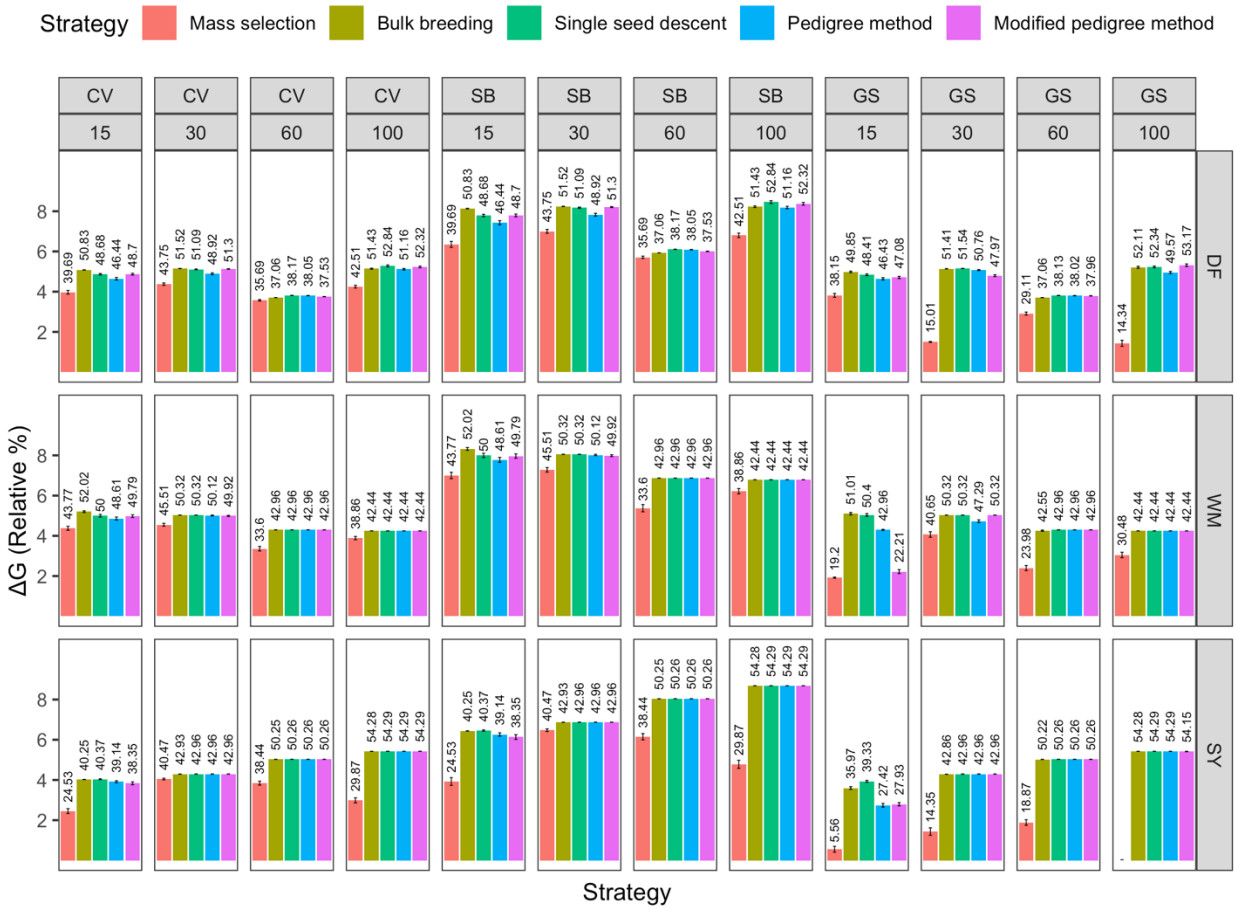

Figure 2.21: Comparison of five breeding strategies in terms of relative genetic gain per cycle across 10 cycles averaged over 50 runs in a closed system. Selected traits include days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Increasing numbers of initial parents displayed on the top along with different breeding methods. Breeding methods 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.

2.3.4 Principal component analysis

Principal component analyses were plotted to show overall patterns in the simulation outputs. Results for conventional selection are shown in Figures 2.22, 2.23, and 2.24, which correspond to the selection of days to flowering, white mold tolerance, and seed yield, respectively. For days to flowering, the first two principal components explained 78.8% of the variance. Under conventional breeding, a notable cluster was formed for parental population size of 100, which separated it from other parental population sizes. There were overlaps observed for the other parental population sizes. A cluster for bulk breeding with a parental population size of 15 formed in the direction of the eigenvector for effective population size. Alongside this was an overlapping cluster consisting of single seed descent and the modified pedigree method, both with a parental population size of 15. The pedigree method, with parental population sizes of 15, 30, and 60, formed a cluster in the direction of the eigenvector for the fixation of favourable alleles. Mass selection, with parental population sizes of 15, 30, and 60, mainly clustered around the center of the PCA plot.

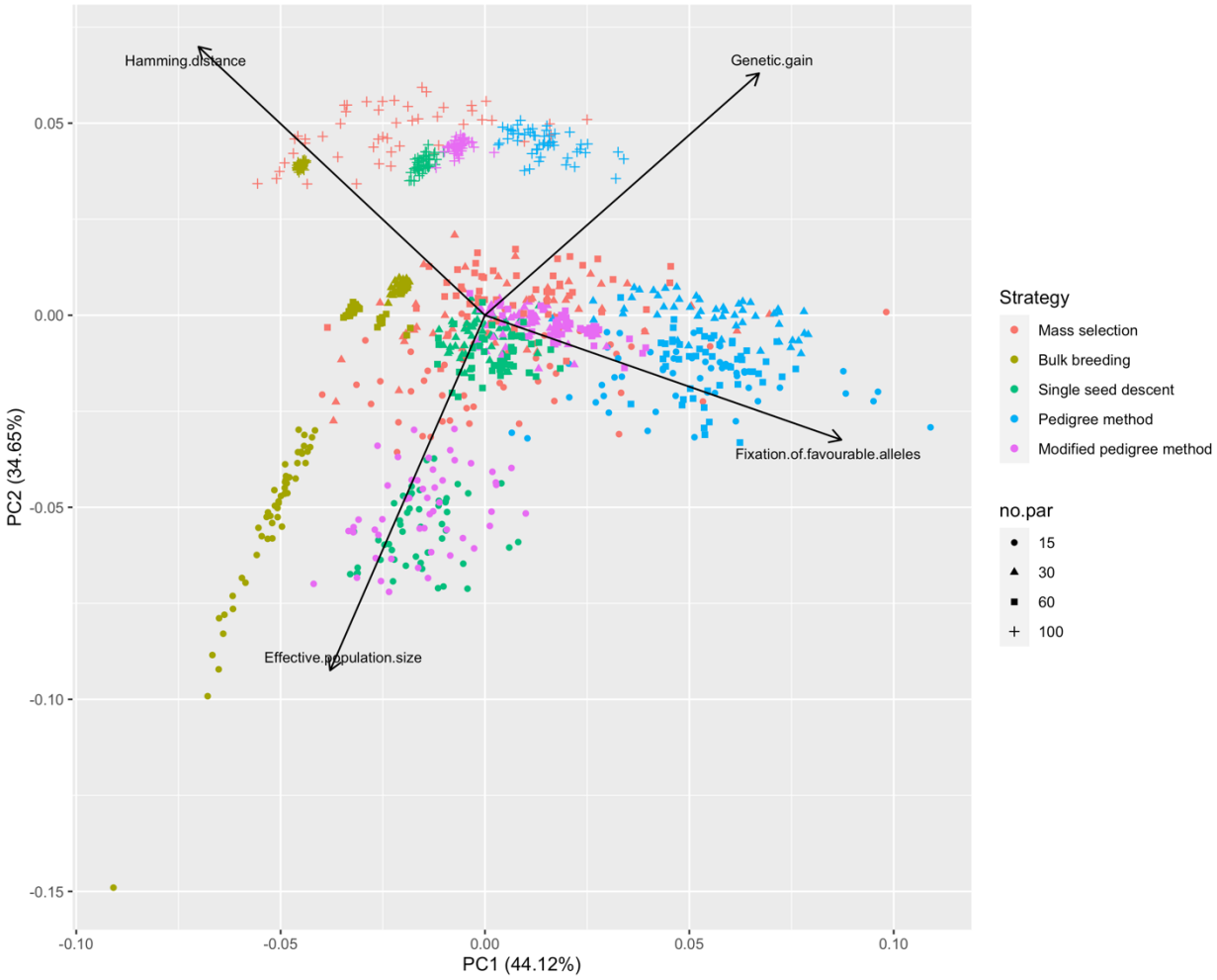

Figure 2.22: Principal component analysis (PCA) plot displaying the variation among five breeding strategies under conventional breeding in terms of genetic gain variables in a closed system. Days to flowering was selected with increasing parental population sizes represented by different shapes. Breeding 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 78.8% of the variance.

For white mold tolerance, the first two principal components explained 89.7% of the variance. The eigenvectors for genetic gain and fixation of favourable alleles point in similar directions. Single seed descent, the pedigree method, and the modified pedigree method with parental population sizes of 30 formed a cluster in the direction of the fixed favourable alleles. The pedigree method with a parental population size of 100 was grouped in the direction of genetic gain. Single seed descent and the modified pedigree with 60 parents formed clusters next to each

other along the eigenvector for effective population size. Another cluster consisting of these two strategies with 100 parents was found to the right. In the direction of the Hamming distance eigenvector is a large cluster with overlaps for all five strategies. The cluster found in the outermost part of the axis for the Hamming distance vector is bulk breeding and mass selection with a parental population size of 15.

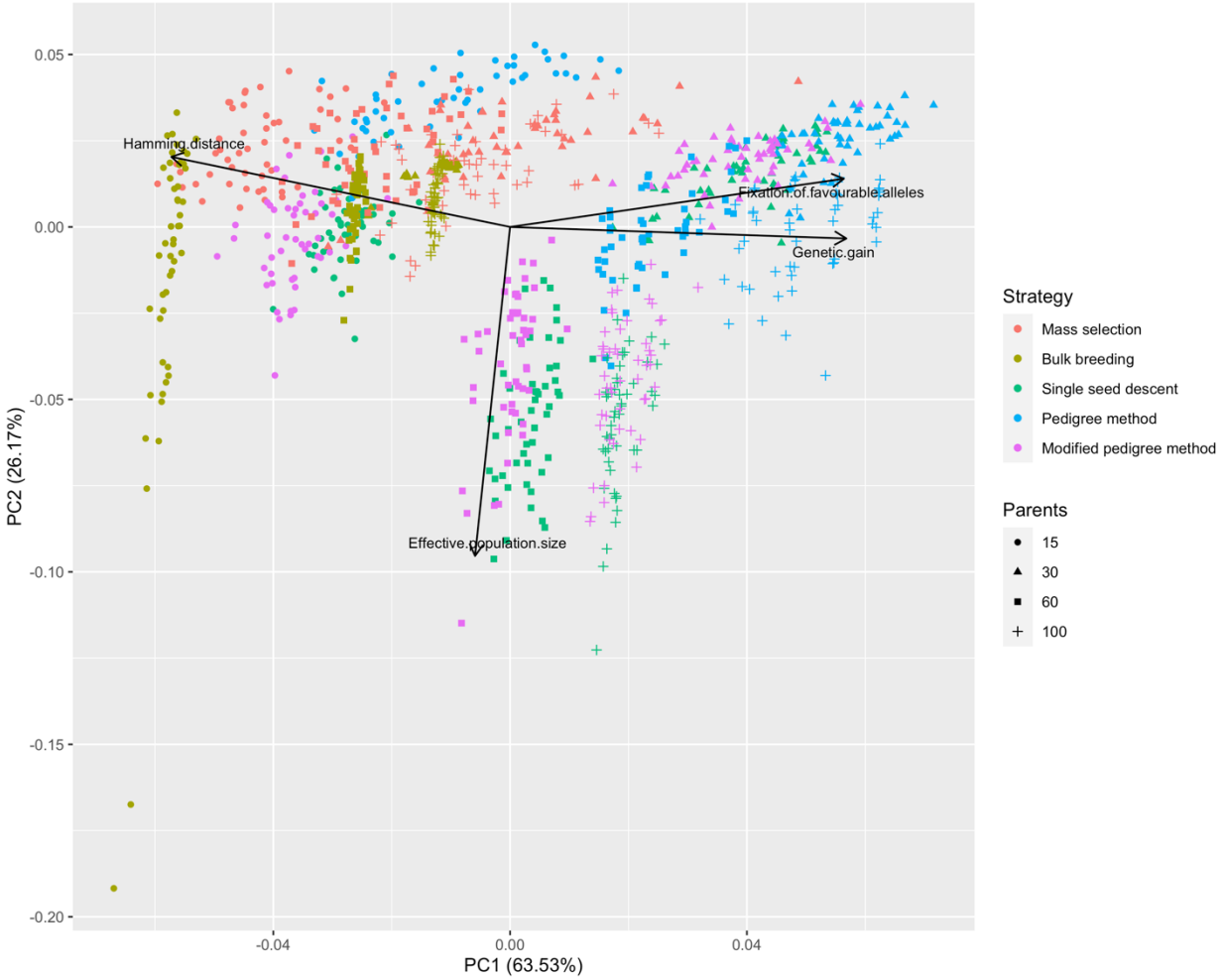

Figure 2.23: Principal component analysis (PCA) plot displaying the variation among five breeding strategies under conventional breeding in terms of genetic gain variables in a closed system. White mold tolerance was selected with increasing parental population sizes represented by different shapes. 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 89.7% of the variance.

The PCA plot for seed yield revealed some regular patterns, with multiple linear-like clusters. The first two principal components accounted for 76.9% of the variance. Mass selection and bulk breeding were distinctly separate from single seed descent, the pedigree method, and the modified pedigree method. Clusters for bulk breeding and mass selection formed in the direction of the Hamming distance and effective population size eigenvectors. Meanwhile, the clusters for single seed descent, the pedigree method, and the modified pedigree method formed along the eigenvectors for genetic gain and the fixation of favourable alleles. Clusters for mass selection and bulk breeding with a parental population size of 15 formed to the left of the fixed favourable alleles eigenvector. The cluster for mass selection with a parental population size of 100 was located in the extreme of the Hamming distance eigenvector. A large linear-like cluster consisting of bulk breeding with 60 parents formed in between the eigenvectors for effective population size and Hamming distance. In general, single seed descent and the modified pedigree method overlapped with each other. Found in the most extreme of the genetic gain eigenvector was the pedigree method with a cluster for the parental population size of 100 and a single point representing a parental population size of 30. The Pedigree method with a parental population size of 15 formed a cluster in the most extreme of the fixed favourable alleles eigenvector.

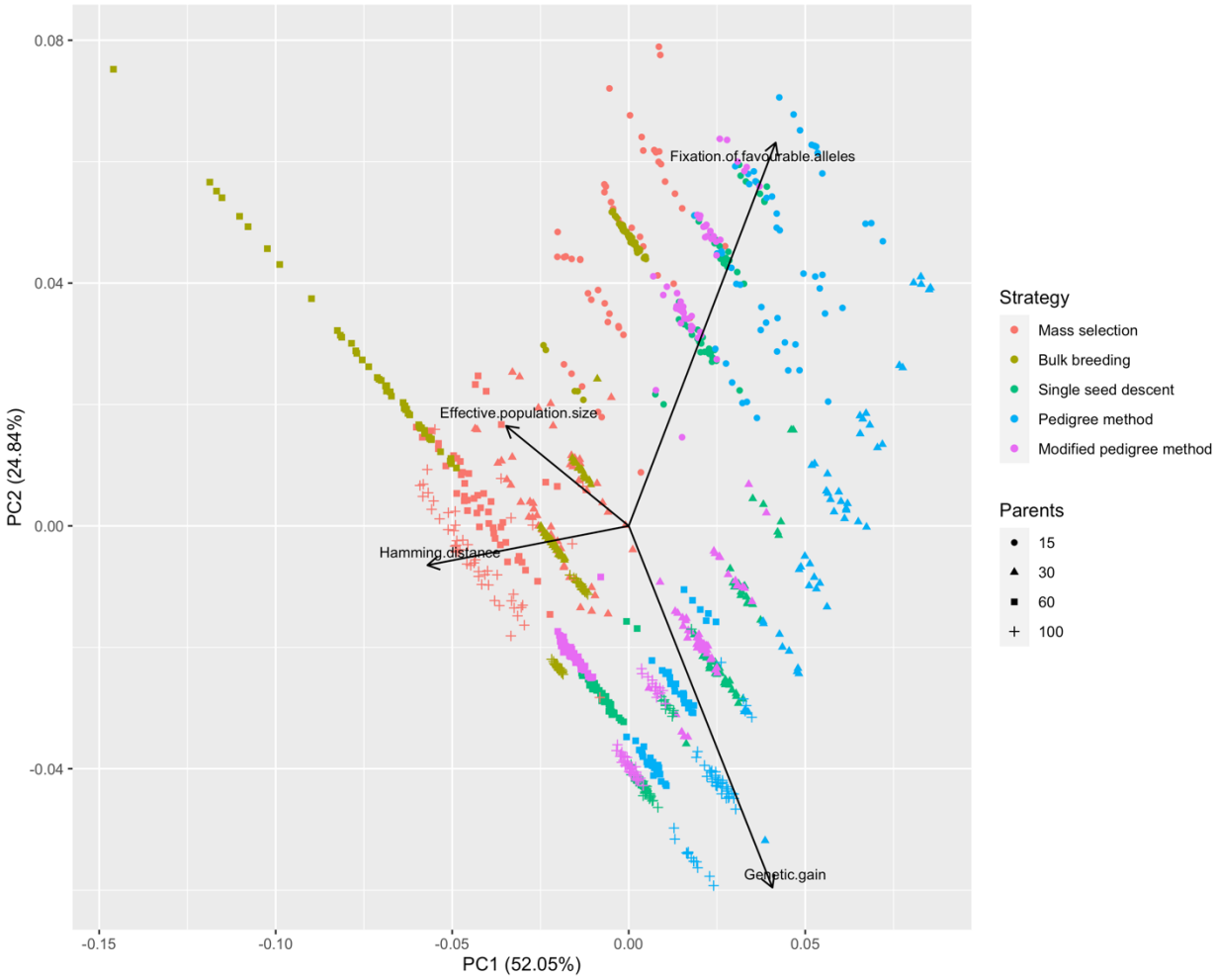

Figure 2.24: Principal component analysis (PCA) plot displaying the variation among five breeding strategies under conventional breeding in terms of genetic gain variables in a closed system. Seed yield was selected with increasing parental population sizes represented by different shapes. Breeding strategies 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 76.9% of the variance.

The PCA plots were also generated with the inclusion of genomic selection and speed breeding. For days to flowering, the first two principal components explained 75.1% of the variance (Figure 2.25). To the right side of the PCA plot between the eigenvectors for genetic gain and Hamming distance, there was a large linear-like cluster representing a parental population size of 100. In the extreme of the eigenvector for Hamming distance, was a cluster 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. In 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, there was a large cluster corresponding to the pedigree method under genomic selection and speed breeding. 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, there was a sparse cluster consisting of multiple strategies including mass selection, the pedigree method, single seed descent, and the modified pedigree method.

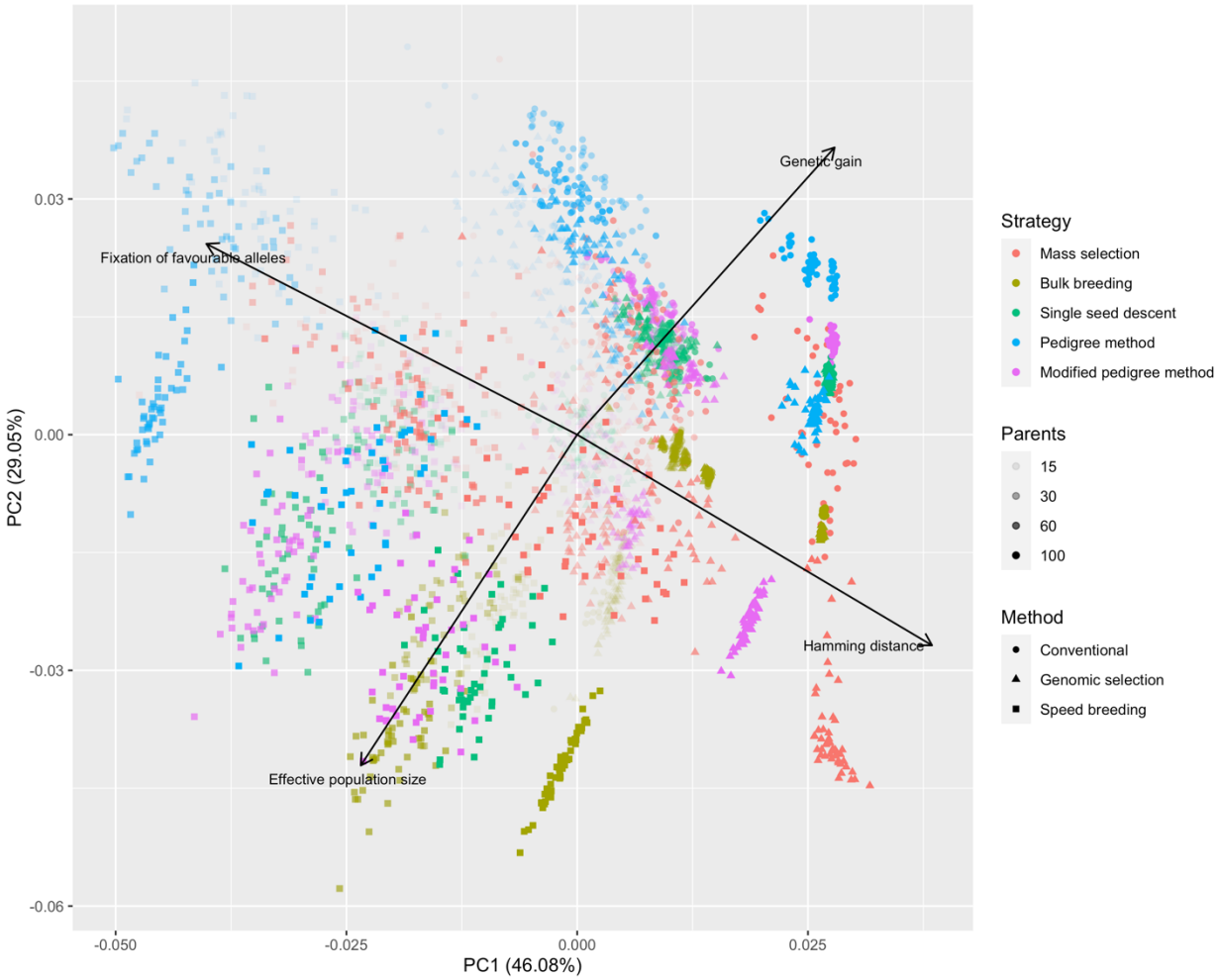

Figure 2.25: Principal component analysis (PCA) plot displaying the variation among five breeding strategies in terms of genetic gain variables in a closed system. Days to flowering was selected with increasing parental population sizes represented by different shapes. Breeding strategies include mass 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.

For white mold tolerance, the first two principal components accounted for 81.8% of the variance (Figure 2.26). 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 linear-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.

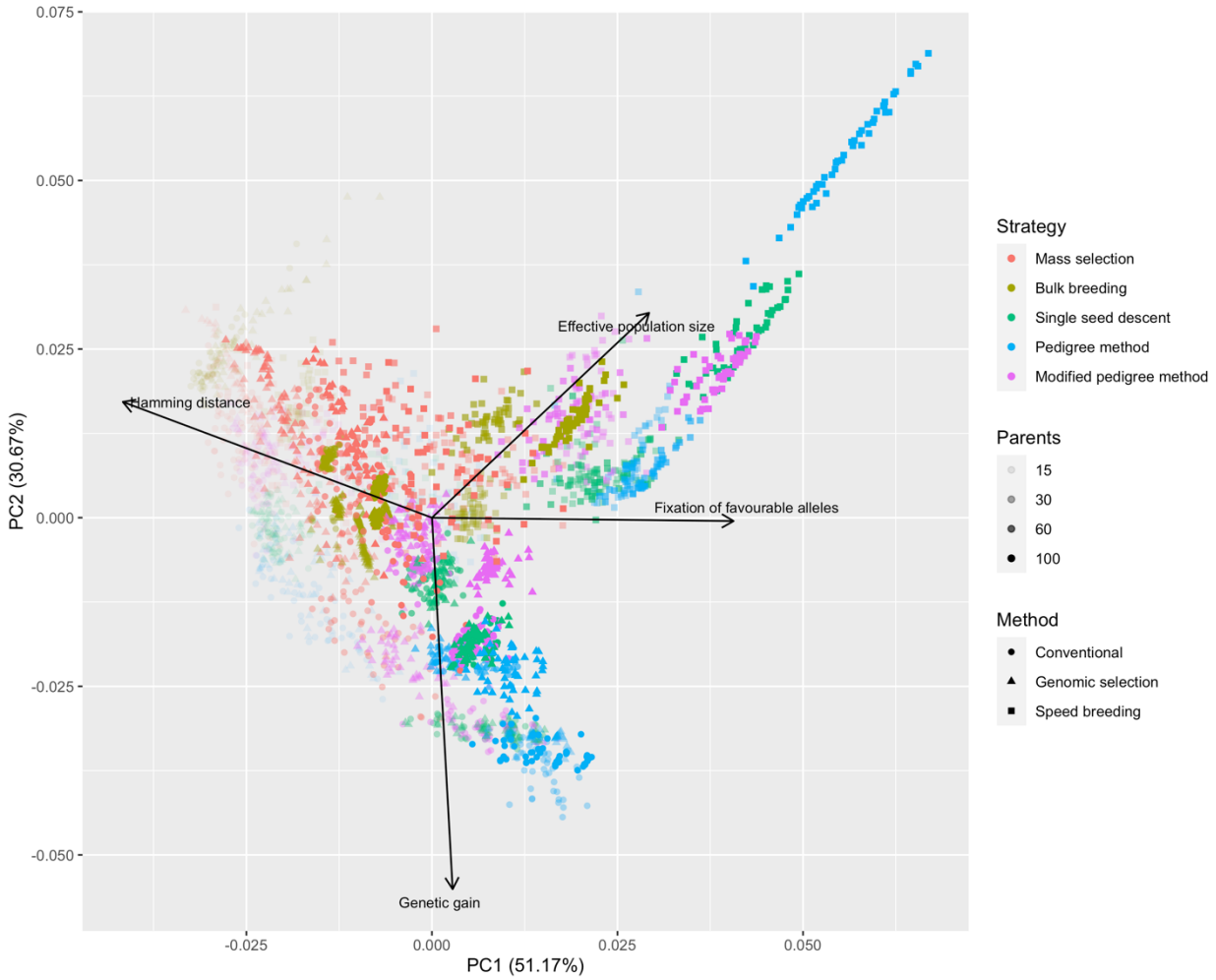

Figure 2.26: Principal component analysis (PCA) plot displaying the variation among five breeding strategies under conventional breeding in terms of genetic gain variables in a closed system. White mold tolerance was selected with increasing parental population sizes represented by different shapes. 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 81.8% of the variance.

For seed yield, the two major principal components explained 72.3% of the variance (Figure 2.27). Overall, there were many linear-like clusters that formed. 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. As one moves towards the center of the plot, there were clusters for bulk breeding that corresponded to speed breeding and genomic selection, as well as multiple points constituting 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.

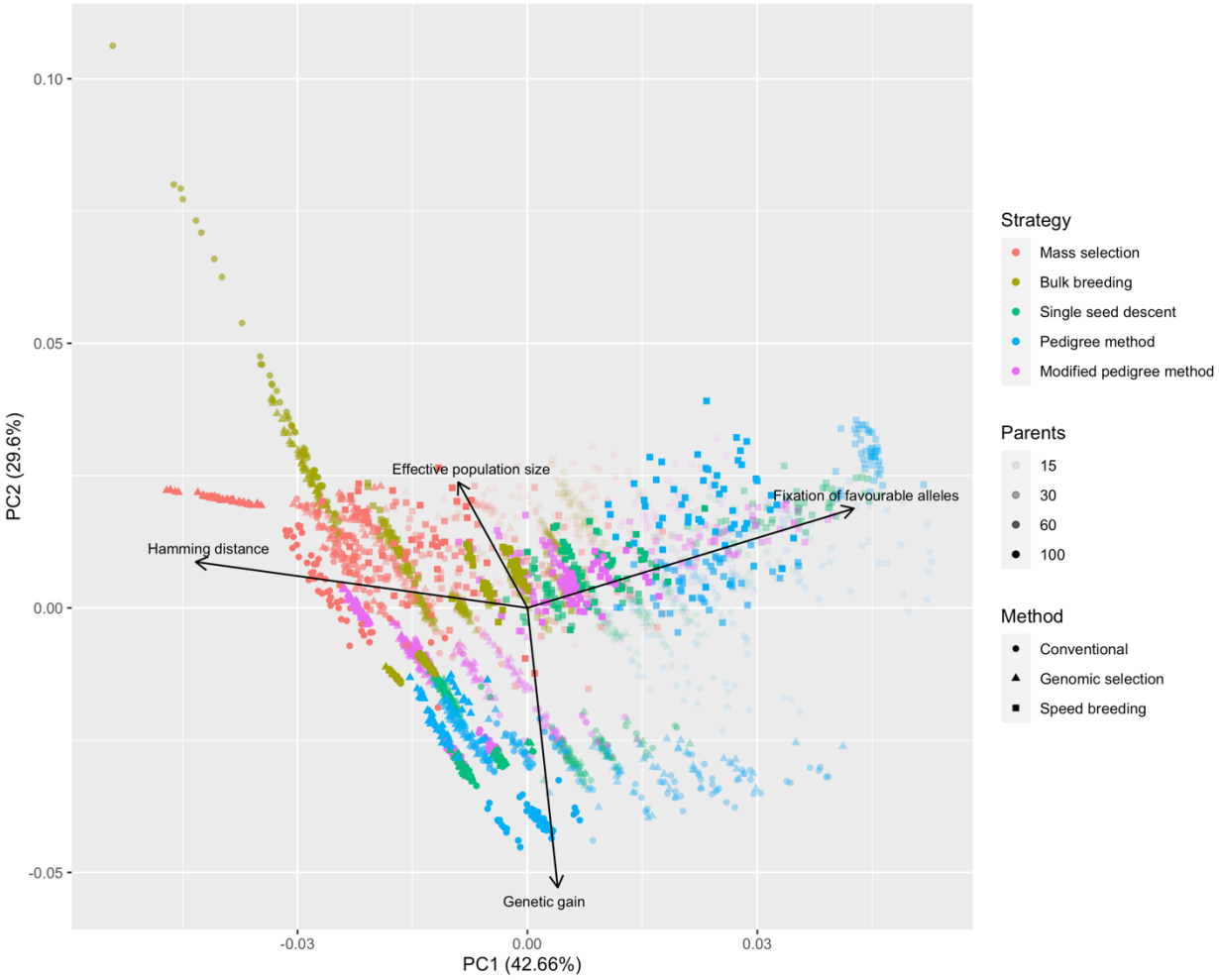

Figure 2.27: Principal component analysis (PCA) plot displaying the variation among five breeding strategies under conventional breeding in terms of genetic gain variables in a closed system. Seed yield was selected with increasing parental population sizes represented by different shapes. Breeding strategies 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 72.3% of the variance.

2.4 Discussion

2.4.1 Comparison of breeding strategies

The breeding strategies performed differently for each breeding scenario simulated and depended on the trait being selected. For days to flowering, the scenario utilizing single seed descent led to the highest genetic gain per cycle. Meanwhile, 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 breeding strategies. However, researchers have investigated the use of different breeding 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. (Agric Res, 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 the 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. For days to flowering, a less complex trait,

2.4.2 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 see if they may be worthwhile to implement 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 methods. 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 in the time and cost savings that could 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. They may also need to consider the expenses tied with establishing a good 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 (Taylor, 2014). The prediction accuracies of genomic selection are discussed in chapter 3.

2.4.3 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. 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. Based on mathematical formulation and simulated data, the use of more crosses with smaller populations was more effective (Bernardo, 2003; Witcombe & Virk, 2001; Yonezawa & Yamagata, 1978). This was based on the assumption 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 cross 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, 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 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.

2.4.4 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; Jiankang 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 Figure 2.21, 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.

2.4.5 Patterns observed from the PCA plots

The 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 resulted in little to no genetic variance early on. Other patterns observed from the PCA plots indicated that single seed descent and the modified pedigree method 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 the two strategies 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, strategies clustering in the direction of the Hamming distance eigenvector are likely to underperform compared to the other strategies. Thus, the Hamming distance eigenvector is a useful indicator of the performance of a strategy, unlike the fixation of favourable alleles, which may be misleading.

2.4.6 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 or a polygenic trait, a large parental population size produced the greatest genetic gain per cycle. For trait involving few QTL, use of a small parental population size is sufficient. In terms of the optimal strategy, single seed descent was the 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 QTLs. 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. In addition, seed yield is a complex trait with many small effect QTLs that are difficult to detect, meaning the QTLs included in this study represent a small sample of the total QTLs that contribute to the trait.

2.5 Supplemental data

Table S2.1: Goodness of fit for the genetic variance model

Trait	**Model†**	**Df**	**AIC**	**BIC**
DF	1	12	122965.6	123057.0
	2	62	119742.5	120214.6
WM	1	12	124321.0	124412.4
	2	62	121197.8	121670.0
SY	1	12	120999.9	121091.3
	2	62	118965.5	119437.7

† Model 1 refers to an unnested model, while model 2 refers to a nested model

Table S2.2: Analysis of variance (ANOVA) for percent genetic variance

Trait	Source	Sum Sq	Mean Sq	NumDF	DenDF	F value	
DF	Parents	55219.92	18406.64	3	14995	108.50	***
	Framework (Parents)	136114.56	17014.32	8	14995	100.29	***
	Strategy (Framework)	1788840.96	37267.52	48	14995	219.68	***
WM	Parents	96172.96	32057.65	3	14995	171.46	***
	Framework (Parents)	409903.96	51237.99	8	14995	274.05	***
	Strategy (Framework)	2000944.47	41686.34	48	14995	222.96	***
SY	Parents	56267.46	18755.82	3	14995	116.43	***
	Framework (Parents)	250154.13	31269.27	8	14995	194.11	***
	Strategy (Framework)	3163532.34	65906.92	48	14995	409.14	***

*** indicates a 0.001 significance level

Table S2.3: Goodness of fit for the fixation of favourable alleles model

Trait	**Model†**	**Df**	**AIC**	**BIC**
DF	1	12	231244.9	231344.6
	2	62	227671.8	228187.0
WM	1	12	240336.5	240436.2
	2	62	238545.2	239060.3
SY	1	12	238983.0	239082.7
	2	62	235970.3	236485.5

† Model 1 refers to an unnested model, while model 2 refers to a nested model

Table S2.4: Analysis of variance (ANOVA) for fixation of favourable alleles

Trait	Source	Sum Sq	Mean Sq	NumDF	DenDF	F value	
DF	Parents	117931.59	39310.53	3	29981	343.03	***
	Framework (Parents)	330146.12	41268.26	8	29985	360.12	***
	Strategy (Framework)	2370853.31	49392.78	48	29981	431.01	***
WM	Parents	282025.86	94008.62	3	29981	570.74	***
	Framework (Parents)	208787.54	26098.44	8	29983	158.45	***
	Strategy (Framework)	3345298.00	69693.71	48	29981	423.12	***
SY	Parents	171666.81	57222.27	3	29981	378.63	***
	Framework (Parents)	440869.60	55108.70	8	29985	364.64	***
	Strategy (Framework)	6895079.20	143647.48	48	29981	950.48	***

*** indicates a 0.001 significance level

Table S2.5: Goodness of fit for the Hamming distance model

Trait	**Model†**	**Df**	**AIC**	**BIC**
DF	1	12	190300.7	190400.4
	2	62	174181.4	174696.6
WM	1	12	194058.8	194158.5
	2	62	186527.6	187042.7
SY	1	12	201445.1	201544.8
	2	62	181481.6	181996.8

† Model 1 refers to an unnested model, while model 2 refers to a nested model

Table S2.6: Analysis of variance (ANOVA) for Hamming distance

Trait	Source	Sum Sq	Mean Sq	NumDF	DenDF	F value	
DF	Parents	165737.3	55245.78	3	29981	2866.33	***
	Framework (Parents)	332251.9	41531.49	8	29983	2154.79	***
	Strategy (Framework)	1140641.9	23763.37	48	29981	1232.92	***
WM	Parents	167665.3	55888.43	3	29981	1920.50	***
	Framework (Parents)	133810.0	16726.25	8	29960	574.77	***
	Strategy (Framework)	1014786.1	21141.38	48	29981	726.48	***
SY	Parents	149621.2	49873.72	3	29981	2028.44	***
	Framework (Parents)	440943.1	55117.89	8	29980	2241.73	***
	Strategy (Framework)	3430251.5	71463.57	48	29981	2906.53	***

*** indicates a 0.001 significance level

Table S2.7: Goodness of fit for the genetic gain model

Trait	**Model†**	**Df**	**AIC**	**BIC**
DF	1	12	153313.5	153413.2
	2	62	152762.4	153277.5
WM	1	12	171961.2	172060.9
	2	62	171772.7	172287.9
SY	1	12	169043.5	169143.2
	2	62	168792.8	169307.9

† Model 1 refers to an unnested model, while model 2 refers to a nested model

Table S2.8: Analysis of variance (ANOVA) of genetic gain

Trait	Source	Sum Sq	Mean Sq	NumDF	DenDF	F value	
DF	Parents	191.90	63.97	3	29981	6.78	***
	Framework (Parents)	4856.43	607.05	8	29985	64.34	***
	Strategy (Framework)	12827.25	267.23	48	29981	28.32	***
WM	Parents	423.86	141.29	3	29981	7.9	***
	Framework (Parents)	2039.48	254.94	8	29984	14.3	***
	Strategy (Framework)	7620.33	158.76	48	29981	8.9	***
SY	Parents	801.16	267.05	3	29981	16.58	***
	Framework (Parents)	2658.14	332.27	8	29982	20.63	***
	Strategy (Framework)	19176.76	399.52	48	29981	24.80	***

*** indicates a 0.001 significance level

Chapter 3 Accuracy of Genomic selection

Abstract

Genomic selection is a technique that predicts the performance of an individual according to genotypes that are predicted to be desirable based on a model. The effectiveness of genomic selection is strongly tied to its prediction accuracy. Previous studies have evaluated the accuracy of genomic selection using simulations. The aim of this study was to evaluate changes in accuracy of genomic selection based on many known QTLs identified in the literature and determine their relationship with true breeding values. Simulation results revealed that correlation-based prediction accuracies (also referred to as realized accuracy) fluctuate depending on trait genetic architecture, breeding strategy and the number of initial parents involved in the breeding program. Generally, maximum accuracies were achieved under a mass selection strategy followed by pedigree single seed descent methods. Model updating benefitted some breeding strategies more than others (e.g., single seed descent vs mass selection). For low heritability traits (i.e., yield), conventional methods provided comparable rates of genetic gain, but genetic gain under genomic selection reached a plateau in a lower number of cycles.

3.1 Introduction

3.1.1 Genomic selection

First described by (Meuwissen, Hayes, & Goddard, 2001), genomic selection (GS) is a technique that can make use of the vast amount of information from genetic markers. With advances DNA technology and declining costs for genotyping, breeders can now gain access to large quantities of genetic information. In particular, genome wide association studies (GWAS) have allowed for the discovery of quantitative trait loci (QTL). These QTL are predicted to contribute to the

phenotype of a trait. In the past, markers that were closely linked to a QTL could be used to select on individuals with a desired allele. Much of the success with from marker assisted selection (MAS) was in traits that were controlled by single genes. Application of MAS to polygenic traits, or traits controlled by many genes, has seen less success. Even with high density markers, there are limitations to the effectiveness of MAS. This is because the linkage phase between a marker and a QTL must be determined each time before its use. GS is a method that relies on estimating breeding values based on model-predicted phenotypic values. The key component in genomic selection is that all markers across the genome are used for prediction. A training population, where individuals are both genotyped and phenotyped, is first used to train a model. Then, the model is applied to a testing population, where individuals have only been genotyped, to predict their phenotypes and assign genomic estimated breeding values (GEBV) to each individual. The advantage to using genomic selection is that it has the potential to save the time and resources that would normally be put towards phenotyping individuals. This is because individuals would only need to be genotyped, so only genotyping costs would need to be considered.

3.1.2 Factors impacting genomic selection accuracy

The main drawback to the use of genomic selection is the accuracy with which the model can predict phenotypes from the genotypes. Genomic selection has been widely used in animal breeding programs. For example, in dairy cattle, one study found annual genetic gain increases of 33 to 77% in three different breeds following the implementation of genomic selection (Doublet et al., 2019). Despite the promising findings for animal breeding, the move towards implementing genomic selection in plants, especially for complex quantitative traits, has been slow. This is likely due to a number of factors that affect the accuracy of genomic selection.

Training population size and trait heritability

A number of studies have found that the training population size greatly impacts the accuracy of genomic selection. A larger training population may increase accuracy by up to 20%. Furthermore, the heritability of a trait can impact the training population size required, especially when the h^2 is less than 0.4. For example, to obtain an accuracy of 0.7, a training population size of 9000 is required for a trait with $h^2 = 0.2$ if the effective population size is 1000. This greatly contrasts a training population size of 3000 when the trait heritability is 0.5 (Lorenz et al., 2011)

Population structure

Accounting for population structure is a key factor for successfully implementing GS. Isidro et al. (2015) demonstrated that stratifying populations can improve the accuracy of GS. Another group of researchers considered the effects of relatedness between individuals when designing a training population. GS accuracy was determined to be highest when individuals in the training population was closely related to individuals in the testing population. Furthermore, in cases where relatedness is low, increasing the diversity of a training population can improve accuracy (Norman, Taylor, Edwards, & Kuchel, 2018).

Genomic selection model

A number of different models are available for predicting marker effects. (Heslot, Yang, Sorrells, & Jannink, 2012) previously compared the effectiveness of 11 GS models. These included random regression best linear unbiased prediction (rrBLUP), Bayesian ridge regression (BRR), and Bayesian Lasso (BL), BayesB, weighted Bayesian shrinkage regression (wBSR), BayesCπ, empirical Bayes (E-Bayes), elastic net, reproducing kernel Hilbert space (RKHS), support vector machine (SVM), random forest (RF), and neural network (NNET). The authors recommended the use of rrBLUP, BL, and wBSR due to their ease of implementation, versatility, and limited

overfitting. They noted that BayesCπ was not an ideal model due to the high computational time. Meanwhile, E-Bayes and NNET both led to overfitting, with E-Bayes also having reduced accuracy and NNET requiring more computational power. Interestingly, RKHS also resulted in overfitting, however the accuracy was not impacted, meaning that while the model picked up more noise, it was able to capture more genetic signal. RF led to promising accuracies, but may require more validation before being established as a GS model ((Heslot et al., 2012). As rrBLUP has been demonstrated to be a reliable model, it was used to determine the marker effects to simulate GS in the study. The model for rrBLUP is shown in Equation 3.1

$$y = X\beta + Zu + \varepsilon \quad [3.1]$$

where y is a list of phenotypes, X is a design matrix for the fixed effects β, Z is a design matrix for the random effects u; where $u \sim N(0, K\sigma^2_u)$, and ε is residual variance.

Model update

A simulation study conducted based on a sorghum breeding program found that updating the genomic selection model every year can increase genetic gains up to 39% (Muleta, Pressoir, & Morris, 2019). Accuracy is greater when the training population contains individuals in the same generation as the selection candidates. In essence, as the number of generations separating the training population and selection candidates increases, the accuracy will decrease. Thus, model updates are required to ensure the genomic selection accuracy is maintained (Heffner, Lorenz, Jannink, & Sorrells, 2010).

3.1.3 Objectives

Although genomic selection has been widely implemented in animal breeding, its use in plant breeding still requires further validation. The objective of this study is to investigate the accuracy

of genomic selection in a simulation study. Five breeding strategies were simulated with the selection of three traits. The following hypotheses were tested:

1. Genomic selection accuracy, measured via the correlation between the breeding value and genomic estimate breeding value, will be greater in traits with a high heritability.
2. The genomic selection accuracy estimated from the simulation will be similar to accuracies predicted from the formula described by Hans D Daetwyler, Ricardo Pong-Wong, Beatriz Villanueva, & John A Woolliams, 2010
3. Updating the model will lead to an increase in GS accuracy.

3.2 Methods

3.2.1 Simulation setup

Simulation parameters in the Unchanged GS model simulation are described in chapter 2. A second simulation, henceforth referred to as the Updated GS model simulation, consisted of updating the genomic selection model after the third cycle of selection. For the Updated GS model simulation, five breeding strategies with a parental population size of 30 were simulated with the selection of three traits, which included days to flowering, white mold tolerance, and seed yield. The Updated GS model simulation consisted of 20 runs, with 6 cycles in total. To simulate GS, a parental population was created using the SimuPOP software and run through QU-GENE to obtain genotypic and phenotypic values for the individual in the population. The mixed.solve function in the rrBLUP package was used to estimate marker effects using the genotypic and phenotypic values obtained from the SimuPOP parental population. The marker effects were then used to make selections for 3 cycles. To simulate updating the model, a random sample of individuals at the third cycle were used to re-train the model. The genotypic and

phenotypic values of these individuals were used to determine marker effects, which were then used to perform selections from cycle 4 to 6.

3.2.2 Expected genomic selection accuracy

(Hans D Daetwyler et al., 2010) described a number of components that impact the accuracy of GS. The authors derived a formula for GS accuracy, as follows:

$$r_{g\hat{g}G} = \sqrt{\frac{N_p h^2}{N_p h^2 + n_G}} \qquad [3.2]$$

Where N_P refers to the number of individuals in a training population, h^2 is the heritability, and n_G is the number of independent loci. Based on the derived formula, the accuracy of GS is influenced by the heritability of the trait, the number individuals in the training population, and the number of loci being considered. This formula, however, does not properly account for situations with a very large number of loci. Based on Equation 3.2, as the number of loci increases, the accuracy will wrongly shift towards 0, since there cannot be an infinite number of independent loci. As LD will result in some of the loci being linked, the number of independent chromosome segments, M_e, should be used in place of n_G (H. D. Daetwyler, R. Pong-Wong, B. Villanueva, & J. A. Woolliams, 2010). By replacing n_G with M_e, one can derive Equation 3.3:

$$r_{g\hat{g}G} = \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}} \qquad [3.3]$$

Where N_P refers to the number of individuals in a training population, h^2 is the heritability, and M_e is the number of independent chromosome segments. The equation to calculate M_e is shown below:

$$M_e = \frac{2N_eL}{log(4N_eL)} \quad [3.4]$$

Where N_e is the effective population size, and L is the genome length in Morgans.

The effective population size (N_e) is an important concept in population genetics. In a closed population with a finite number of individuals, the genetic variation within that population will diminish after several generations. The number of individuals in that population will determine how well the genetic variability can be sustained. Maintaining genetic variation in a population will reduce inbreeding and its negative effects. However, one factor that causes the genetic variation to decline is genetic drift, which is a random occurrence that leads to the fixation of alleles at polymorphic loci. Effective population size is a term that was coined by Sewall Wright to refer to the size of an ideal population, in reference to an actual population, if genetic drift was the only force that was acting on the population (Soulé, 1987). A number of models have been proposed for the estimation of N_e (Caballero & Toro, 2000; Crow & Morton, 1955; Jinliang Wang & Hill, 2000; S. Wright, 1938). Depending on the model used, certain assumptions are made regarding the population under investigation. For plant species in particular, few estimates have been made for N_e. Siol *et al*. (2007) first reported estimates for the highly-selfing, model legume species, *Medicago truncatula*. To estimate N_e, the authors used the variance effective size estimator described by Waples (1989):

$$\widehat{N}_e = \frac{t}{2\left[\hat{F}_c - \left(\frac{1}{2S_0} + \frac{1}{2S_t}\right)\right]} \quad [3.5]$$

where t refers to the number of generations that have elapsed between the two sampled populations, $\hat{F}_c$ refers to the estimator for the standardized variance of gene frequency changes at

a single locus, S_0 and S_t indicate the sample sizes of the population at time 0 and time t, respectively. The estimator Fc can be written as:

$$F_c = \frac{1}{k}\sum_{i=1}^{k}\frac{(x_i - y_i)^2}{(x_i+y_i)/2 - x_i y_i} \qquad [3.6]$$

where k is the number of alleles, x_i is the observed allele frequency at time 0, and y_i is the observed allele frequency at time t. Average Fc estimates for all loci was determined and used to determine the N_e. From there, the genomic selection accuracy was estimated using Equation 3.2.

3.2.3 In silico realized genomic selection accuracy

To estimate the in silico realized GS accuracy, the outputs from the simulation were used. QU-GENE reports the genotypic values obtained from conventional breeding and genomic selection. In QU-GENE, the phenotypic selection used for conventional breeding is based entirely on the QTLs provided. QU-GENE will output genotypic values for each individual and assign phenotypic values drawn from a distribution, which depends on the error variance supplied. In essence, the genotypic value reported by QU-GENE may be considered the true breeding value (TBV), as it assumes that the QTLs are the genes controlling a trait. Since the genomic selection model is trained on the phenotypic values assigned by QU-GENE for a training population, the genomic estimated breeding values (GEBVs) are indirectly based upon the TBV. The ratio between the mean population GEBV and TBV may be used as a rough estimator of genomic selection accuracy.

3.2.4 Principal component analysis

Principal component analyses were conducted to visualize the relationships between the factors that influence both genetic gain and genomic selection accuracy. The family means for 7 different factors that contribute to the genetic gain and genomic selection accuracy in the first

cycle were determined for each of the 20 runs. The 7 factors included genetic gain, fixation of favourable alleles, Hamming distance, genetic variance, effective population size, true breeding value, and genomic estimated breeding value. The fixation of favourable alleles described the average percentage of beneficial alleles that were fixed in the population. Meanwhile, the Hamming distance was used to describe the distance of an individual from an ideal genotype. This distance was determined as the number of base pairs that differ from the optimal genotype. The effective population size was calculated according to equation 5. All calculations were performed using original code written in R and may be located on our lab GitHib page ((McGill University Pulse Breeding and Genetics Laboratory, 2021). Lastly, the R package ggbiplot was used to create the principal component analyses.

3.3 Results

3.3.1 Unchanged GS model

The GS accuracies obtained from the simulations described in chapter 2 are presented here. The simulation conducted in chapter 2 used an unchanged GS model. GS accuracies were estimated in two manners, the first being formula-based, using Equation 3.2, and the second being in silico realized GS accuracies based on correlations between the TBV and the GEBV.

3.3.1.1 Expected formula-based GS accuracy

Using the unchanged GS model, the GS accuracies determined using Equation 3.2 ranged from 0.07 to 0.63 (Figure 3.1). For most breeding scenarios, prediction accuracy decreased over the 10 cycles. The decline was smaller with parental population sizes of 15 and 30. Prediction accuracies were higher with larger parental population sizes. The strategies had similar accuracies and followed similar trends when the parental population size was small. However, with large parental population sizes, mass selection had a much greater prediction accuracy

compared to the other strategies. Furthermore, the accuracy remained relatively high for mass selection. The accuracy was highest under days to flowering, followed by white mold tolerance and then seed yield. For days to flowering under mass selection with 100 parents, the accuracy decreased from 0.63 to 0.47 over 10 cycles. For white mold tolerance under mass selection with 100 parents, the accuracy decreased from 0.46 to 0.39 over 10 cycles. For seed yield under mass selection with 100 parents, the accuracy decreased from 0.43 to 0.29 over 10 cycles. In most breeding scenarios, bulk breeding resulted in the lowest prediction accuracies. For days to flowering with 15 parents, the accuracy in bulk breeding decreased from 0.18 to 0.10 over 10 cycles. For white mold tolerance with 15 parents, the accuracy declined from 0.11 to 0.09 over 10 cycles when bulk breeding was used. For the selection of seed yield with 15 initial parents, the accuracy with bulk breeding decreased from 0.09 to 0.07 over 10 cycles. Heritability had an impact on GS accuracy, where accuracy was highest under days to flowering, followed by white mold tolerance and then seed yield. However, selection strategies had similar accuracies when the parental population size was small, regardless of heritability.

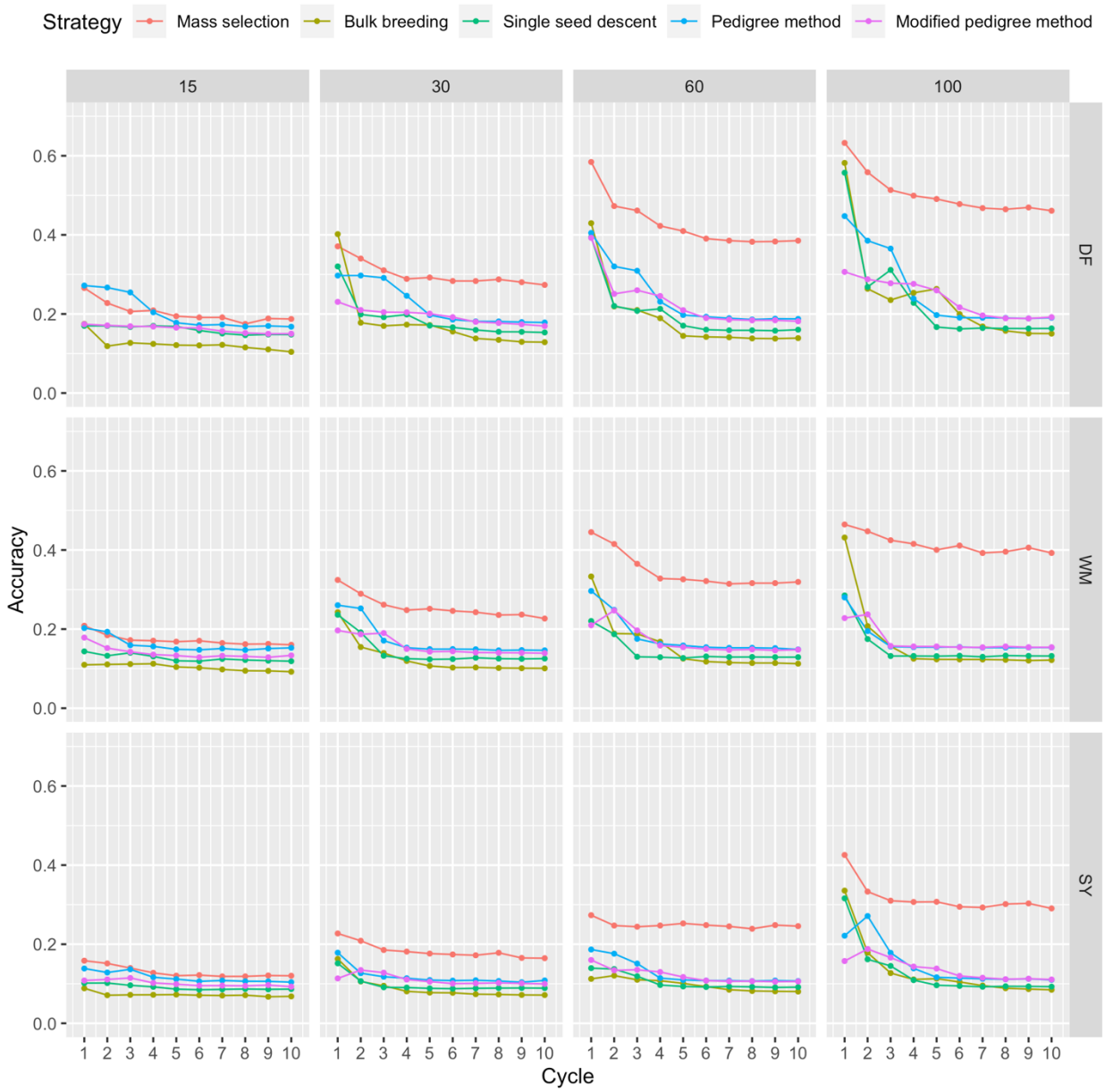

Figure 3.1: Expected genomic selection accuracies predicted from Equation 3.2 using an unchanged GS model. Coloured lines correspond to breeding strategies, which include mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method. Three traits were selected with differing parental population sizes indicated at the top and right-hand side of the panels. Traits included days to flowering (DF), white mold tolerance (WM), seed yield (SY).

3.3.1.2 True breeding values (TBV)

True breeding values were obtained from the QU-GENE output files and plotted over 10 cycles (Figure 3.2). For days to flowering and seed yield, the TBVs increased and eventually plateaued.

The opposite was true for white mold tolerance, where TBVs plummeted before reaching a plateau. There were notable differences between the TBVs when different numbers of parents were used at the beginning of the cycle. For each of the traits, as the number of parents increased, the average TBV for the strategies decreased. For days to flowering, the average TBVs across strategies at the end of the 10^{th} cycle were 20.21, 17.69, 13.95, 9.96 for 15, 30, 60, and 100 parents, respectively. For white mold tolerance after 10 cycles, the average TBVs were -46.17, -62.32, -62.51, -62.55 for 15, 30, 60, and 100 parents. For seed yield, the average TBVs were 2830.95, 2759.40, 2190.16, 2189.79 for 15, 30, 60, and 100 parents. For most breeding scenarios, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method led to similar TBVs. Mass selection resulted in a lower TBV for days to flowering and seed yield, while it led to a higher TBV for white mold tolerance in comparison to the other four strategies.

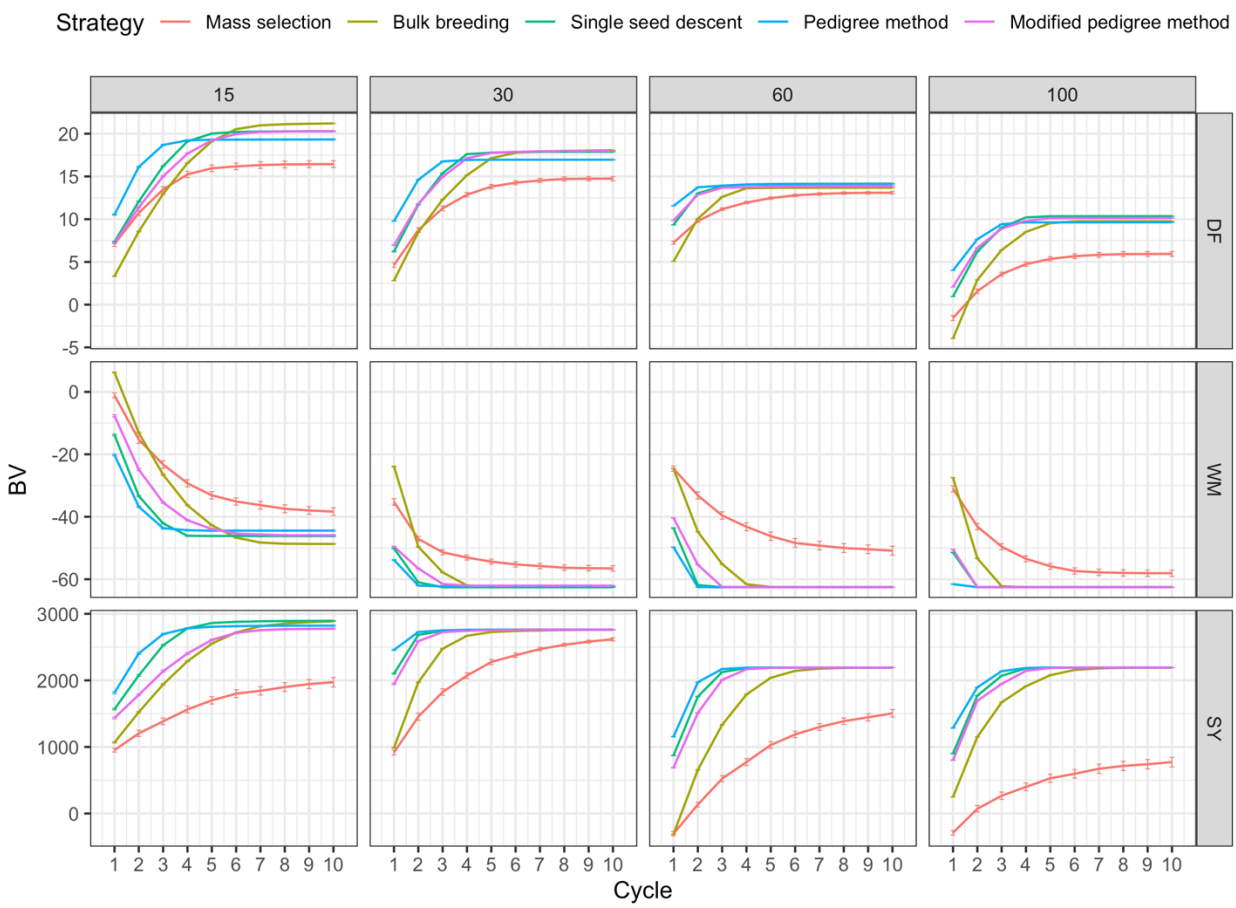

Figure 3.2: True breeding values provided by QU-GENE plotted over 10 cycles for increasing parental population sizes and three traits for an unchanged GS model. Traits include days to flowering (DF), white mold tolerance (WM), seed yield (SY).

3.3.1.3 Genomic estimated breeding values (GEBV)

Genomic estimated breeding values were determined for each cycle for 10 cycles (Figure 3.3). For days to flowering and seed yield, the GEBVs increased rapidly before plateauing. The opposite trend was observed for white mold tolerance. The parental population sizes had an impact on the GEBVs at the end of the breeding program. For days to flowering, the GEBVs averaged across the strategies were 19.90, 17.47, 13.98, and 9.87 for 15, 30, 60, and 100 parents. For white mold tolerance, parental population sizes of 15, 30, 60, and 100 resulted in average GEBVs of -35.26, -61.52, -62.34, and -62.52. Lastly, for seed yield, the average GEBVs were 2420.18, 2745.80, 2185.04, 2183.38 for parental population sizes of 15, 30, 60, and 100.

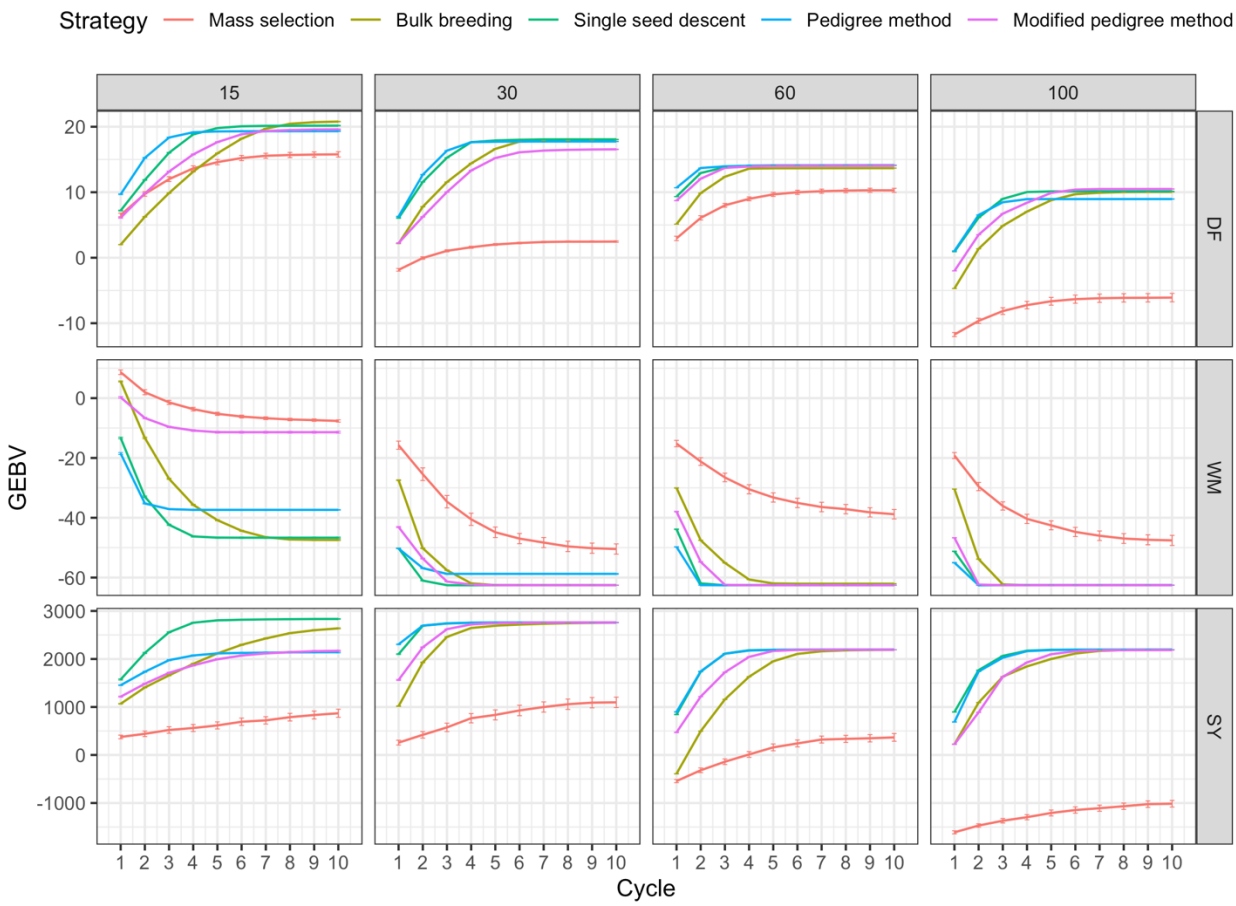

Figure 28 Genomic estimated breeding values plotted over 10 cycles for increasing parental population sizes and three traits for an unchanged model. Traits include days to flowering (DF), white mold tolerance (WM), seed yield (SY).

3.3.1.4 *In silico* Realized GS accuracy

In silico realized GS accuracies were obtained from the correlation between the TBV and the GEBV. They ranged from -0.35 to 0.32 (Figure 3.4). The mean accuracies for each strategy were -0.03, -0.02, 0.02, 0.05, and -0.01, for mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method, respectively. For days to flowering, the highest accuracy (0.31) was observed under the pedigree method with 30 parents, while the lowest accuracy (-0.35) was in bulk breeding with 100 parents. When considering white mold tolerance, single seed descent with 15 parents resulted in the greatest accuracy (0.32). The lowest

accuracy was seen under mass selection with 60 parents (-0.29). Interestingly, in cycle 2, there was an increase in accuracy, after which the accuracy declined rapidly and became negative by cycle 4. For seed yield, both the highest (0.24) and lowest (-0.34) accuracies were observed in mass selection. Notably, the parental population size of 30 led to the highest mean accuracy of 0.01, while the population size of 60 led to the lowest mean accuracy of -0.03. For certain cycles, a correlation could not be obtained. In these cycles, the variance was zero and the correlation was undefined.

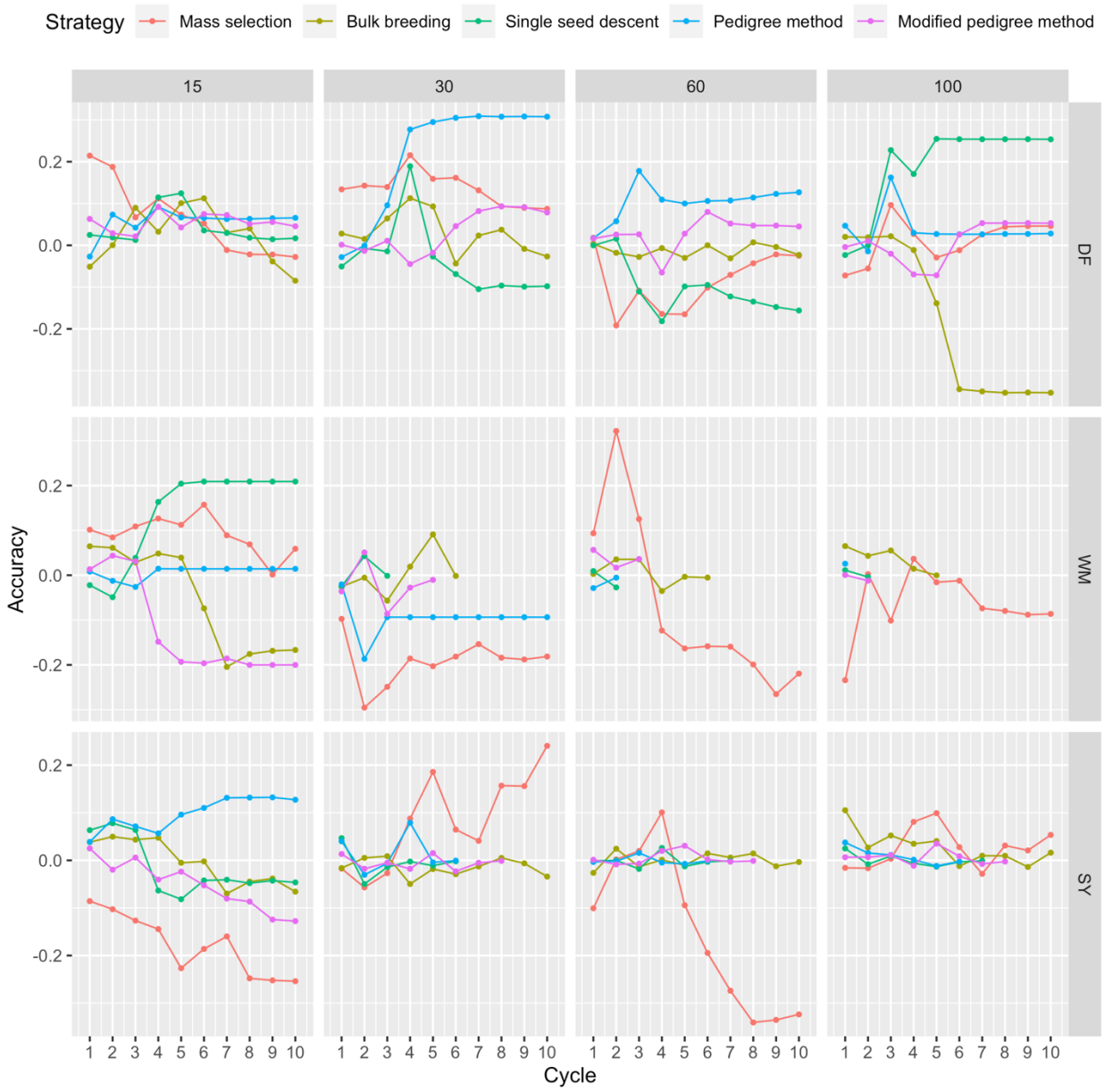

Figure 29: In silico realized genomic selection accuracy estimated from QU-GENE for three selected traits with an unchanged GS model. Traits included days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Accuracies were calculated as the correlation between the true breeding value and the genomic estimated breeding value.

3.3.2 Updated GS Model

The Updated GS model simulation described here in chapter 3 involved updating the GS model.

The simulation was based on a parental population size of 30. The GS model was updated at

cycle 3. GS accuracies were assessed via two measurements, the first being formula-based, where estimated were calculated from Equation 3.2, and the second being in silico realized GS accuracies determined from the correlation between TBV and GEBV.

3.3.2.1 Genetic gain following GS model update

The results from the model update indicated that there was a sharp increase followed by a rapid decline in genetic gain. Model update only seemed to improve genetic gain in one or two cycles immediately after the update, only to return to the rates of genetic gain prior to the update. Conventional breeding was included alongside genomic selection as a comparison for model update. Figure 3.5 shows that updating the GS model resulted in an increase in genetic gain right after cycle 3 for mass selection, the pedigree method, and the modified pedigree method when selecting for days to flowering and seed yield. However, it led to a decrease in genetic gain immediately after cycle 3, followed by an increase after cycle 4, and a decrease after cycle 5 for all strategies when selecting for white mold tolerance. When compared to conventional breeding, genomic selection led to much higher levels of genetic gain for certain strategies in the cycle following the GS model update. For days to flowering, mass selection under genomic selection was 23.8% higher compared to conventional breeding. Meanwhile, the pedigree method and the modified pedigree method were 30.2% and 34.0% higher in genomic selection, respectively. For white mold tolerance, mass selection led to 17.0% greater genetic gain using genomic selection than conventional breeding, while the modified pedigree method under genomic selection resulted in 9.94% higher genetic gain. Finally, for seed yield, mass selection, the pedigree method, and the modified pedigree method resulted in 22.7%, 11.3%, and 18.2% higher genetic gain, respectively using genomic selection compared to conventional breeding. For all other

breeding scenarios, there was little to no difference between genomic selection and conventional breeding in the cycle after the GS model update.

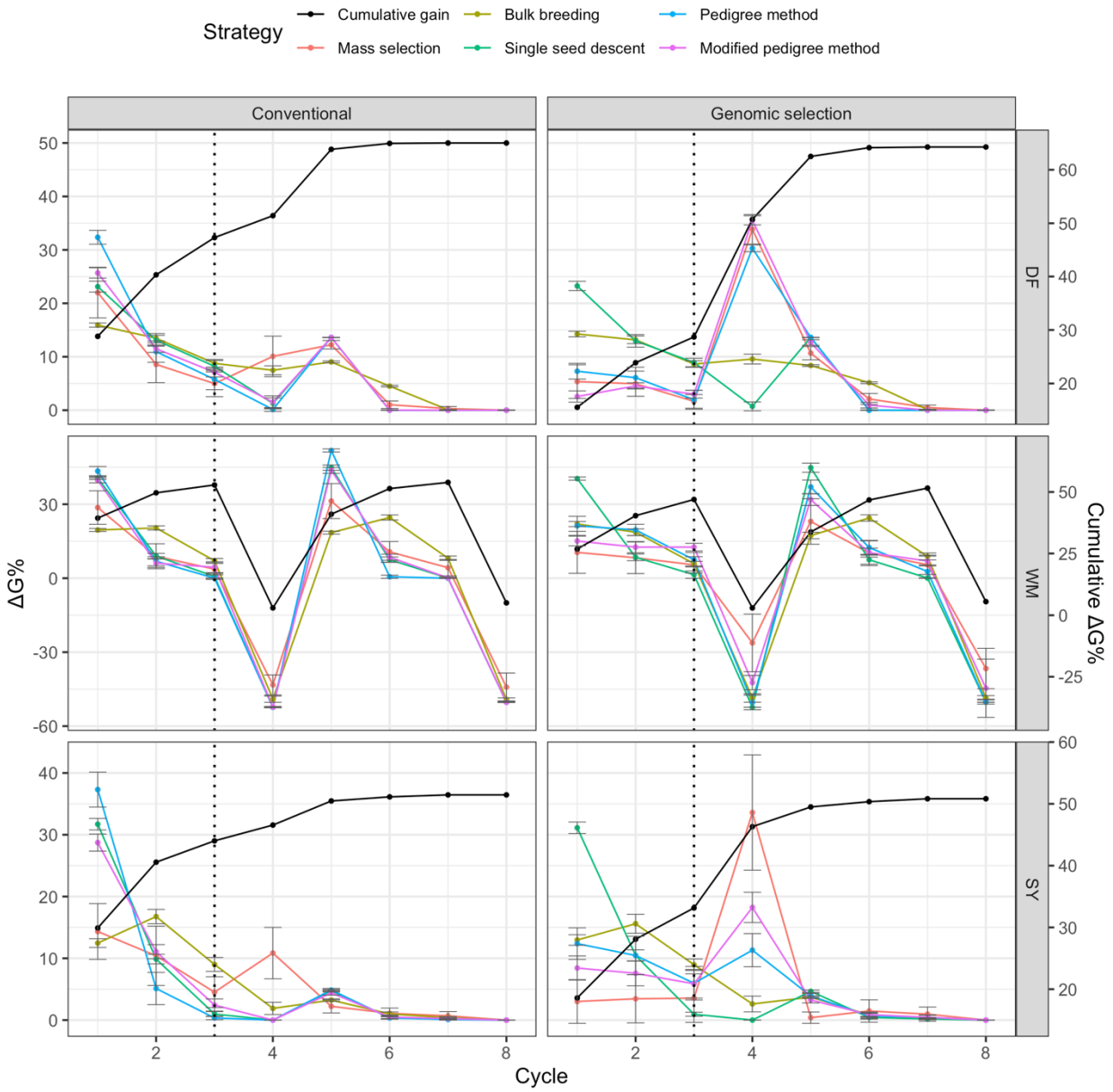

Figure 30 Comparison of five breeding strategies in terms of relative genetic gain following model update. Conventional breeding was included as a control for interruption of the simulation run. Coloured lines correspond to the breeding strategies, mass selection, bulk breeding, single seed descent, pedigree method, and modified pedigree method. Black line indicates the cumulative genetic gain averaged across the five strategies. Simulated traits included days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Dotted line shows when model update took place.

3.3.2.2 Expected formula-based GS accuracy

For the updated GS model, the GS accuracies determined using Equation 3.2 ranged from 0.08 to 0.58 (Figure 3.6). For all breeding scenarios, a general trend was observed where an increase in accuracy occurred after the GS model update at cycle 3, followed by a decline from cycle 4 to 5. The peak accuracy predicted from the days to flowering simulation was 0.58, occurring at cycle 4 with mass selection. For white mold tolerance, the peak accuracy was 0.51, occurring at cycle 4 with the pedigree method. The peak accuracy for seed yield was 0.38 at cycle 4 using the pedigree method. Across the breeding strategies and cycles, the average GS accuracies were 0.28, 0.26, and 0.18 for days to flowering, white mold tolerance, and seed yield, respectively.

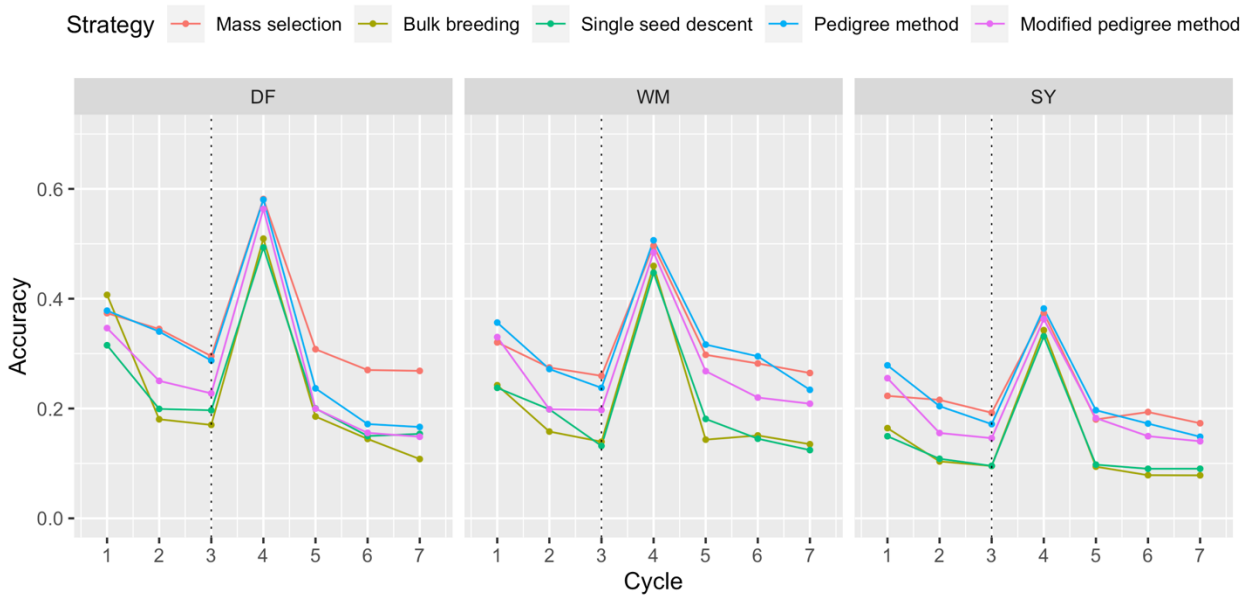

Figure 31 Expected genomic selection accuracies estimated from Equation 3.2 with a GS model update. Prediction accuracies shown over 7 cycles. Coloured lines correspond to strategies, which include mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method.

3.3.2.3 True breeding values (TBV)

After model update, the true breeding values were determined and plotted over 6 cycles (Figure 3.7). At cycle 3, where the update occurred, there was an increase in the TBV for all breeding

scenarios. For days to flowering, the TBV increased by 9.53, 7.05, 6.32, 4.34, and 6.38 from cycle 3 to cycle 4 for mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method respectively. For white mold tolerances, TBVs rose by 15.0, 38.3, 9.44, 0.73, and 10.7 from cycle 3 to cycle 4 for mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method, respectively. Lastly, for seed yield from cycle 3 to cycle 4, mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method had increases in TBVs of 761, 299, 134, 129, and 134, respectively. TBVs appeared to plateau after cycle 4 for days to flowering and seed yield. However, for white mold tolerance, TBVs rapidly declined after cycle 4.

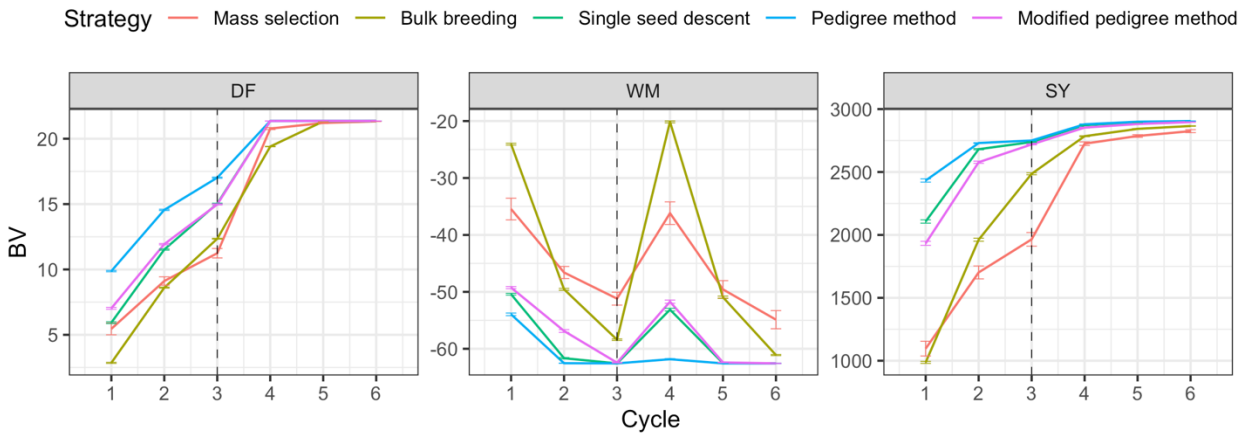

Figure 32 True breeding values plotted over 6 cycles for increasing parental population sizes and three traits with an updated GS model. Traits include days to flowering (DF), white mold tolerance (WM), seed yield (SY). Vertical dotted line indicates the point at which the GS model was updated.

3.3.2.4 Genomic estimated breeding values (GEBV)

Next, genomic estimated breeding values were plotted over 6 cycles (Figure 3.8). For days to flowering, there was a pronounced increase from cycle 3 to cycle 4 for mass selection, the pedigree method, and the modified pedigree method, with increases of 19.0, 18.8, and 20.6, respectively. Smaller increases were observed for the other two strategies. GEBVs increased by

7.66 and 6.02 between cycle 3 and 4 for bulk breeding and single seed descent, respectively. From cycle 3 to cycle 4 for white mold tolerance, GEBVs increased by 4.03, 39.1, 9.43, 16.7, and 13.1 for mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method, respectively. Lastly, for seed yield, all five strategies resulted in an increase in GEBV following model update, with the greatest increase observed in mass selection and the smallest increase in single seed descent. GEBVs increased by 1867, 367, 130, 886, and 1251 for mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method, respectively.

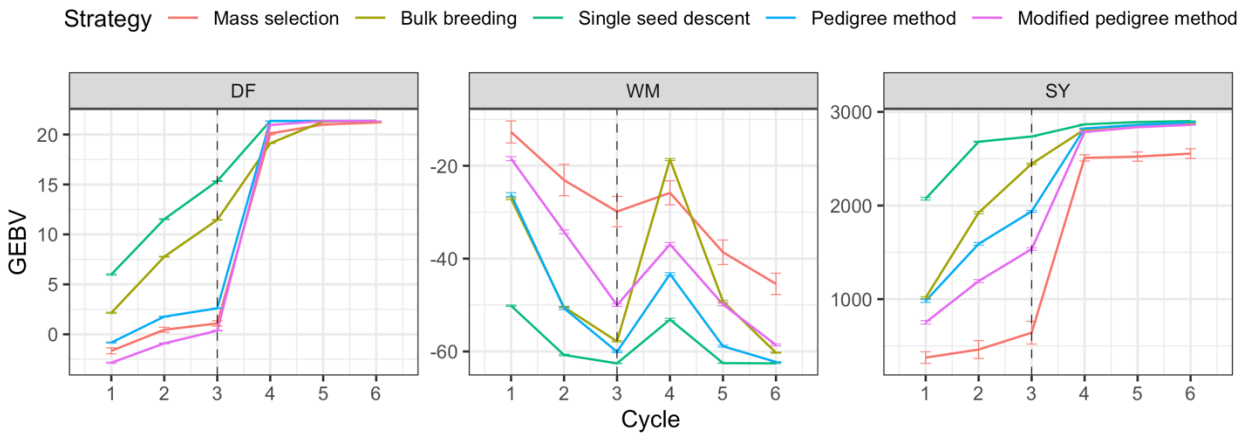

Figure 33 Genomic estimated breeding values plotted over 6 cycles for increasing parental population sizes and three traits with an updated GS model. Traits include days to flowering (DF), white mold tolerance (WM), seed yield (SY). Vertical dotted line corresponds to GS model updated.

3.3.2.5 *In silico* Realized GS accuracy

In silico realized GS accuracies for the updated GS model were obtained and plotted over 6 cycles (Figure 3.9). Once again, the accuracy fluctuated over the different cycles. In particular, mass selection had the greatest variability in accuracy, in some cycles having the highest accuracy, while in others having the lowest accuracies. For days to flowering, following the model update at cycle 4, there was a small improvement in accuracy for single seed descent and

the modified pedigree method, where accuracies increased by 0.08 and 0.04, respectively. The other three strategies saw a decrease in accuracy. From cycle 3 to cycle 4, white mold tolerance GS accuracies declined by 0.09, 0.13, and 0.12 for mass selection, bulk breeding, and the pedigree method, respectively. Lastly, for seed yield, GS accuracies increased by 0.14, 0.03, and 0.02 between cycle 3 and 4 for mass selection, bulk breeding, and single seed descent, respectively. Decreases in GS accuracy after the third cycle were observed for the pedigree method and the modified pedigree method. However, in the last cycle for seed yield, the pedigree method had the greatest accuracy, with a value of 0.06. Across the breeding strategies and cycles, in silico GS accuracies were -0.01, -0.03, and -0.01 for days to flowering, white mold tolerance, and seed yield, respectively.

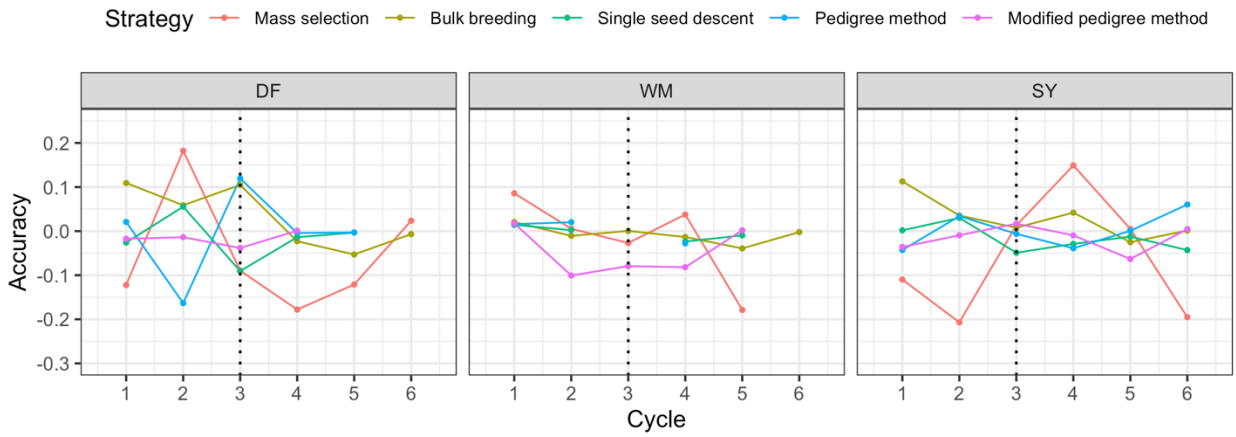

Figure 34: Genomic selection accuracy estimated from QU-GENE following model update. Accuracies were calculated as the correlation between the true breeding value and the genomic estimated breeding value. Simulations began with a parental population size of 30. Traits simulated include days to flowering (DF), white mold tolerance (WM), and seed yield (SY). Coloured lines correspond to the breeding strategies: mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method.

3.3.2.6 Principal component analysis

A principal component analysis was conducted to show the overall result of the simulation with the model update. Figure 3.10 shows the PCA plot for days to flowering, where 80.53% of the

variance is explained by the first two principal components. Notably, the eigenvectors for TBV and GEBV are very close together and point towards similar directions. The eigenvector for Genetic gain and Hamming distance point in similar directions. Towards the right side of the PCA plot, there were two clusters for mass selection that formed on the extreme of the GEBV and TBV eigenvectors. On the opposite side, a cluster containing all five strategies was found in the extremes of both the Hamming distance vector and the genetic gain vector. In the direction of the eigenvector for fixation of favourable alleles, there was a cluster consisting of bulk breeding.

No clusters formed in the extreme of the effective population size eigenvector. Near the center of the plot was a cluster consisting of the pedigree method and single seed descent.

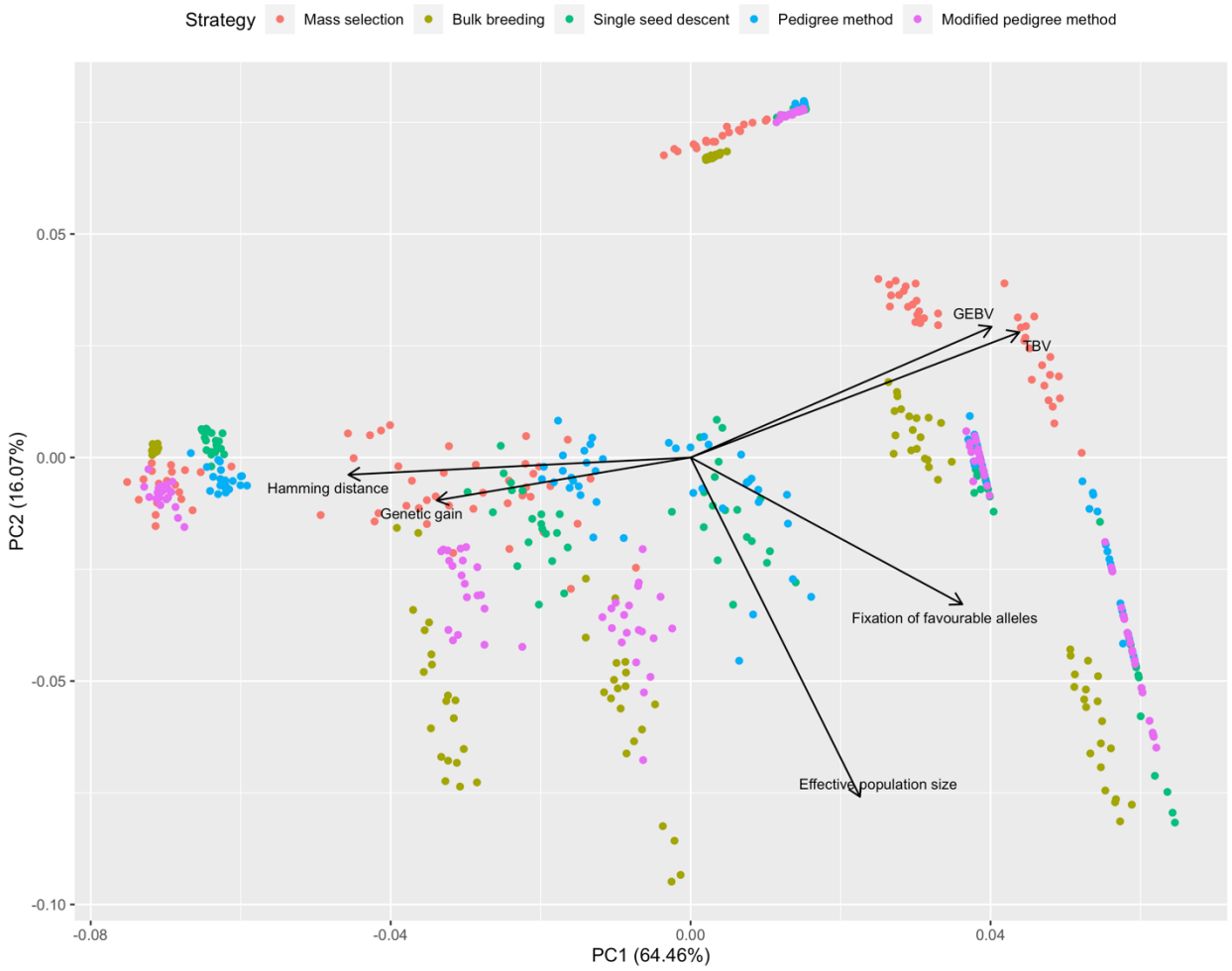

Figure 35: Principal component analysis (PCA) plot displaying the two major principal components accounting for 80.53% of the variance in a simulation with GS model update. Days to flowering was selected for, with colours corresponding to the breeding strategy used. Breeding strategies include mass selection, bulk breeding, single seed descent, the pedigree method, the modified pedigree method.

The first two principal components in the white mold tolerance PCA explained 79.13% of the variance (Figure 3.11). Like days to flowering, the eigenvectors for GEBV and TBV were close to each other. Meanwhile, the eigenvectors for genetic gain, effective population size, and fixation of favourable alleles pointed in similar directions. In the extreme of the Hamming

distance eigenvector, there was a cluster consisting of mass selection. Along the axis of the genetic gain eigenvector, there were some points corresponding to the modified pedigree method and bulk breeding. In the direction of the eigenvector for the fixation of favourable alleles, there was a cluster for bulk breeding.

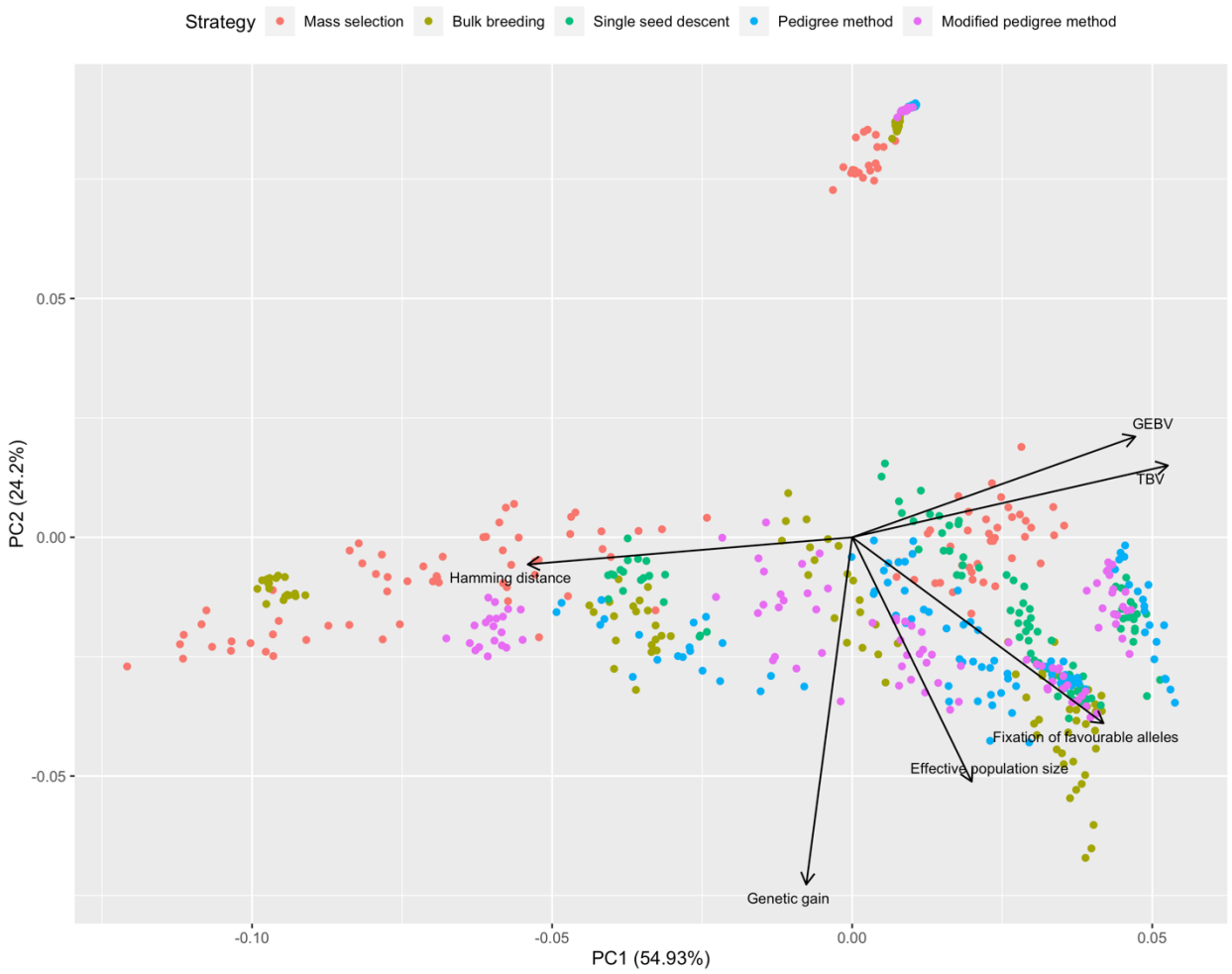

Figure 36: Principal component analysis (PCA) plot with the first two principal components, which explain a total of 79.13% of the variance. White mold tolerance was selected for in a simulation with GS model update. Breeding strategies are indicated by the colour, and include mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method.

For seed yield, the first two principal components described 79.47% of the variance (Figure 3.12). The GEBV and TBV eigenvectors are very close together and point in similar directions.

On the opposite end are the Hamming distance and genetic gain eigenvectors, which are located close together. Towards the extreme of the Hamming distance eigenvector is a cluster made of mass selection and bulk breeding. Between the eigenvectors for fixed favourable alleles and effective population size, there was a cluster corresponding to bulk breeding.

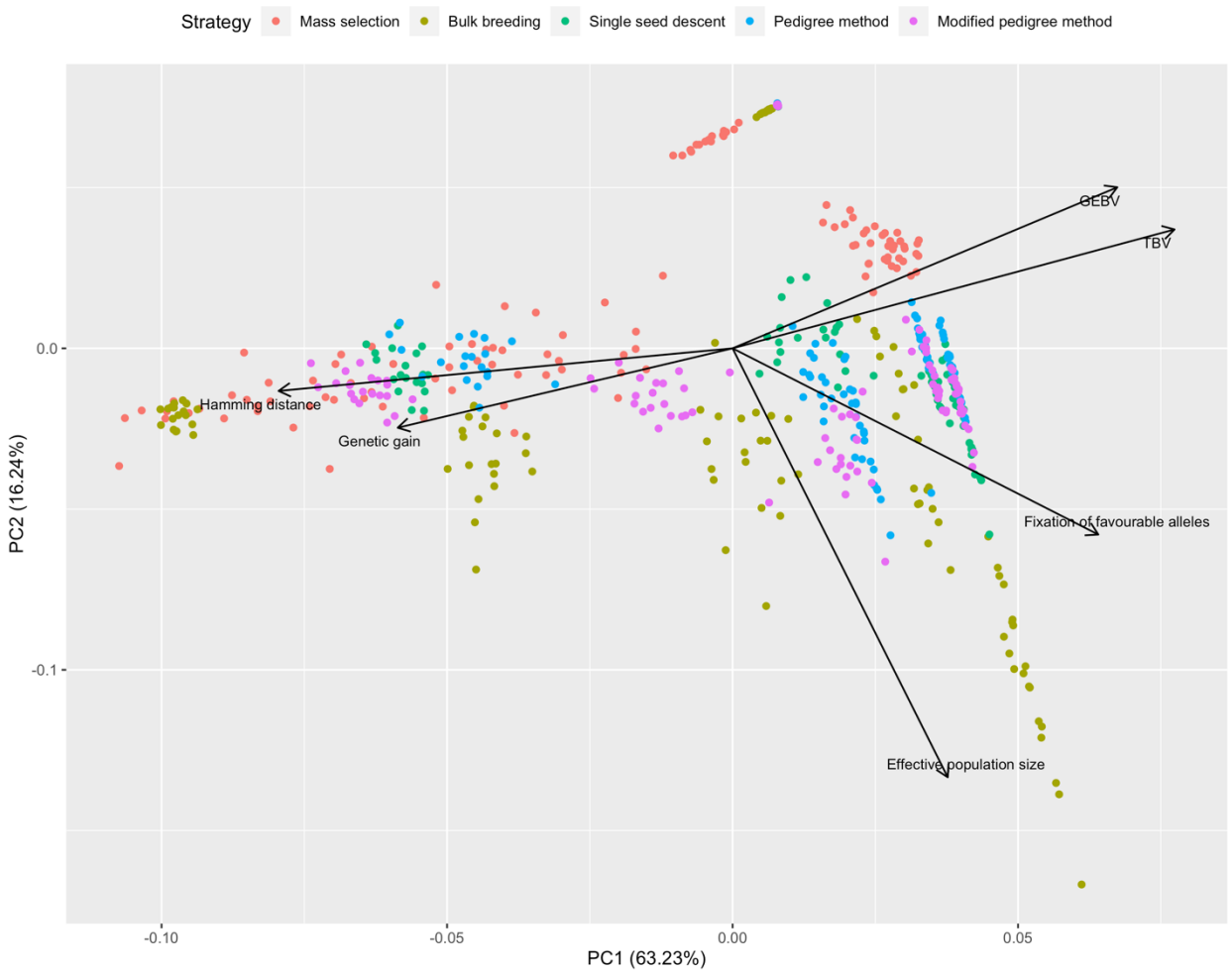

Figure 37: Principal component analysis (PCA) plot consisting of the two principal components that explain the greatest amount of variance, together accounting for 79.47% of the variance. Seed yield was selected for in a simulation with GS model update. Colours refer to the breeding strategy used. Strategies include mass selection, bulk breeding, single seed descent, the pedigree method, and the modified pedigree method.

3.4 Discussion

For the unchanged GS model, the GS accuracy results obtained from Equation 3.2 showed that increased parental population sizes should lead to a higher accuracy. In addition, this increase should be particularly evident in mass selection. The accuracies predicted from Equation 3.2 did not reflect the *in silico* realized GS accuracies. The formula-based accuracies were much higher than the *in silico* realized accuracies. Moreover, the *in silico* realized accuracies greatly fluctuated from one cycle to the next, while the equation-based accuracies saw a gradual decline over six cycles. According to the *in silico* realized GS accuracies, the pedigree method and single seed descent led to the greatest accuracies by the end of the 10 cycles under selection for days to flowering. For white mold tolerance and seed yield, correlations could not be obtained for some of the later cycles due to the correlation being undefined. As previously stated in chapter 2, this may be due to the low number of QTLs included, which consequently increased the efficiency of selection. Within the first few cycles of selection, the variance of genotypic values decreased to zero. The fluctuating accuracies may have been influenced by the simulation setup. The training population consisted of the parents that were used at the beginning of the simulation. Since the simulation involved a closed system, where the progeny at the end of the cycle are used as the parents for the next cycle, a limited number of parents were used due to computational restraints. The small training population size likely reduced the performance of the GS model in its predictive abilities. Interestingly, for certain strategies, in particular single seed descent and the pedigree method, in silico GS accuracies continued to remain high even at the end of the 10^{th} cycle. This was an unexpected result, as theoretically, when a GS model is used to predict on the progeny of a population after several generations, prediction accuracy would be expected to

decline. Thus, future investigations into these strategies must be done to determine the reason that GS accuracies remained high.

For the updated GS model, there was a short-term increase in genetic gain immediately following the model update. Conventional breeding was presented alongside genomic selection to account for the interruption that occurred during the simulation. For mass selection, the pedigree method, and the modified pedigree method across the three traits, genomic selection resulted in higher genetic gain compared to conventional breeding between cycle 3 and 4, while the other strategies saw little to no difference between genomic selection and conventional methods. Thus, for most breeding scenarios, there was an improvement in genetic gain following the model update. The expected formula-based accuracies following GS model update indicated that there would be a spike in accuracy after cycle 3 that would quickly decline once again in the next cycle. However, these formula-based accuracies were not reflected by the in silico realized accuracies. Similar to the unchanged GS model, the in silico accuracies fluctuated from one cycle to the next. Updating the GS model was most beneficial for the single seed descent and the modified pedigree method for days to flowering simulation. It also appeared that mass selection may also benefit from the GS model update, as there was an increase from cycle 4 to 6 after the initial decrease after cycle 3 in days to flowering. For white mold tolerance, use of mass selection and the modified pedigree method may benefit from updating the GS model. Meanwhile, the model update was beneficial for the use of the mass selection and single seed descent in selecting for seed yield. To further investigate how the different genetic gain variables played a role in the simulations overall, PCA plots were generated. Notably, the strategies did not cluster separately from one another. Most of the clusters that formed consisted of more than one strategy. Another key aspect is that the eigenvectors for TBV and GEBV pointed in similar

directions. This would suggest that the correlations would be an effective manner to access the accuracy of GS. For days to flowering and seed yield, the genetic gain eigenvector was closest to the Hamming distance eigenvector and far away from the eigenvector for the fixation of favourable alleles. Thus, results for the GS model update should be taken cautiously. One of the major factors that influence genomic selection accuracy is the training population size. The closed system that was simulated in QU-GENE demanded few parents and many crosses. Since the parents were used for training, the training population size was very small. As a result, the in silico GS accuracies were very low, averaging -0.02 across all breeding scenarios, cycles, and traits. Population structure, which can also impact the accuracy, would not have been a concern in this study because there was no population structure present (results not included). Overall, simulation of GS using this method may require further validation.

3.5 Conclusion

GS has been widely used in animal breeding, however its effectiveness in plant breeding still requires more validation. GS will only be useful if the model can accurately predict the phenotype of a trait from the genotype. Numerous studies have investigated the prediction accuracy in simulations. However, in those studies, QTLs were simulated and were evenly distributed across the genome with effect sizes drawn from a random distribution. This study aimed to assess the accuracy of GS in a simulation that better reflected the real world, in which QTLs effect sizes and positions were based on reported literature. The findings from the study indicate that equation-based estimates of accuracy do not reflect of accuracies obtained from correlations between TBV and GEBV. Nonetheless, according to the correlation-based accuracies, there may be some benefits to using single seed descent or the modified pedigree method.

Chapter 4 General Conclusions

Breeding programs are complex systems. Plant breeders must take into account the time, labour, breeding materials, land, and phenotyping means, in order make the appropriate decisions to enhance genetic gain. By making use of computer simulations, multiple breeding scenarios may be compared at the same time. This study aimed to simulate breeding scenarios that would closely reflect the real world. Rather than simulating QTL positions and effect sizes, real QTLs were identified in the literature and incorporated into the simulation. The findings demonstrated that the chosen strategy, framework, and parental population size significantly contributed to the genetic gain that can be achieved. The optimal breeding scenario leading to the greatest %ΔG differed according to the trait being selected. The genetic architecture of the trait likely contributed to this result. With the versatility of computer simulations, repeating the experiments in the study with the inclusion of more QTLs once they have been discovered, could improve the robustness of the findings presented. Another key finding was that genomic selection either underperformed or performed equally to conventional methods. This led to investigations into the accuracy of genomic selection. It was found that equation-based estimates for accuracy did not correspond to correlation-based estimates for accuracy. Thus, it is imperative to consider the applicability and assumptions of the equation prior using it for evaluating GS accuracy. Overall, the correlation-based estimates were quite low. However, in certain scenarios, the pedigree method and single seed descent outperformed other strategies and maintained accuracy even in later cycles. Finally, updating the GS model resulted in an increase in genetic gain. Model update also improved the accuracy in the pedigree method for days to flowering and seed yield. Therefore, the pedigree method may be beneficial if GS is to be used in a breeding program. Although this study did not factor in the costs, implementing GS can save both time and money.

In scenarios where GS performed equally to conventional methods, it may be worthwhile to use GS to save on time and phenotyping costs.

References

Acquaah, G. (2009). *Principles of plant genetics and breeding*. Wiley.

Adams, M. (1967). Basis of yield component compensation in crop plants with special reference to the field bean, Phaseolus vulgaris L. *Crop Science, 7*(5), 505-510.

Agric Res, J. (2019). Comparison of different breeding methods for developing superior genotypes in soybean. *Agricultural Research Journal, 56*, 628.

AlBallat, I. A., & Al-Araby, A. A. A.-M. (2019). Correlation and path coefficient analysis for seed yield and some of its traits in common bean (Phaseolus vulgaris L.). *Egyptian Journal of Horticulture, 46*(1), 41-51.

Ali, M., Zhang, L., DeLacy, I., Arief, V., Dieters, M., Pfeiffer, W. H., . . . Li, H. (2020). Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. *The Crop Journal, 8*(5), 866-877. https://doi.org/10.1016/j.cj.2020.04.002

Assefa, T., Mahama, A. A., Brown, A. V., Cannon, E. K., Rubyogo, J. C., Rao, I. M., . . . Cannon, S. B. (2019). A review of breeding objectives, genomic resources, and marker-assisted methods in common bean (Phaseolus vulgaris L.). *Molecular Breeding, 39*(2), 20.

Ates, D., Asciogul, T. K., Nemli, S., Erdogmus, S., Esiyok, D., & Tanyolac, M. B. (2018). Association mapping of days to flowering in common bean (Phaseolus vulgaris L.) revealed by DArT markers. *Molecular Breeding, 38*(9), 113. https://doi.org/10.1007/s11032-018-0868-0

Atuahene-Amankwa, G., Beatie, A., Michaels, T. E., & Falk, D. (2004). Cropping system evaluation and selection of common bean genotypes for a maize/bean intercrop. *African Crop Science Journal, 12*(2), 105-113.

Beaver, J. S., & Osorno, J. M. (2009). Achievements and limitations of contemporary common bean breeding using conventional and molecular approaches. *Euphytica, 168*(2), 145-175. https://doi.org/10.1007/s10681-009-9911-x

Bernardo, R. (2003). Parental selection, number of breeding populations, and size of each population in inbred development. *Theoretical and Applied Genetics, 107*(7), 1252-1256. https://doi.org/10.1007/s00122-003-1375-0

Blair, M. W., Iriarte, G., & Beebe, S. (2006). QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean × wild common bean (Phaseolus vulgaris L.) cross. *Theoretical and Applied Genetics, 112*(6), 1149-1163. https://doi.org/10.1007/s00122-006-0217-2

Brim, C. A. (1966). A modified pedigree method of selection in soybeans. *Crop science, 6*(2), 220-220.

Caballero, A., & Toro, M. A. (2000). Interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genetical research, 75*(3), 331-343. https://doi.org/10.1017/S0016672399004449

Carneiro, F. F., Santos, J. B. d., Gonçalves, P. R. C., Antonio, R. P., & Souza, T. P. d. (2011). Genetics of common bean resistance to white mold. *Crop Breeding and Applied Biotechnology, 11*, 165-173.

Carvalho, R. S. B., Lima, I. A., Alves, F. C., & Santos, J. B. d. (2013). Selection of carioca common bean progenies resistant to white mold. *Crop Breeding and Applied Biotechnology, 13*(3), 172-177.

Chacón, S. M., Pickersgill, B., & Debouck, D. G. (2005). Domestication patterns in common bean (Phaseolus vulgaris L.) and the origin of the Mesoamerican and Andean cultivated races. *Theoretical and Applied Genetics, 110*(3), 432-444. https://doi.org/10.1007/s00122-004-1842-2

Crow, J. F., & Morton, N. E. (1955). Measurement of gene frequency drift in small populations. *Evolution, 9*(2), 202-214. https://doi.org/10.2307/2405589

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics, 185*(3), 1021-1031. https://doi.org/10.1534/genetics.110.116855

Diaz, L. M., Ricaurte, J., Cajiao, C., Galeano, C. H., Rao, I., Beebe, S., & Raatz, B. (2017). Phenotypic evaluation and QTL analysis of yield and symbiotic nitrogen fixation in a common bean population grown with two levels of phosphorus supply. *Molecular Breeding, 37*(6), 76.

Diaz, L. M., Ricaurte, J., Tovar, E., Cajiao, C., Terán, H., Grajales, M., . . . Raatz, B. (2018). QTL analyses for tolerance to abiotic stresses in a common bean (Phaseolus vulgaris L.) population. *Plos One, 13*(8), e0202342. https://doi.org/10.1371/journal.pone.0202342

Digital Research Alliance of Canada. (2020) *Compute Canada*. https://ccdb.computecanada.ca/

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.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics, 3*(1), 43-52. https://doi.org/10.1038/nrg703

Doublet, A.-C., Croiseau, P., Fritz, S., Michenet, A., Hozé, C., Danchin-Burge, C., . . . Restoux, G. (2019). The impact of genomic selection on genetic diversity and genetic gain in three French dairy cattle breeds. *Genetics Selection Evolution, 51*(1), 52. https://doi.org/10.1186/s12711-019-0495-1

Ender, M., & Kelly, J. D. (2005). Identification of QTL associated with white mold resistance in common bean. *Crop Science, 45*(6), 2482-2490.

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology, 54*(1), 357-374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: an R package for breeding program simulations. *G3 Genes|Genomes|Genetics, 11*(2). https://doi.org/10.1093/g3journal/jkaa017

Geil, P. B., & Anderson, J. W. (1994). Nutrition and health implications of dry beans: a review. *Journal of the American College of Nutrition, 13*(6), 549-558. https://doi.org/10.1080/07315724.1994.10718446

Gepts, P., Debouck, D., Voysest, O., Dessert, M., Hidalgo, R., Singh, S. P., . . . Graf, W. (1991). *Common Beans: Research for Crop Improvement*: C.A.B. International

Centro Internacional de Agricultura Tropical.

Gepts, P., Osborn, T. C., Rashka, K., & Bliss, F. A. (1986). Phaseolin-protein Variability in Wild Forms and Landraces of the Common Bean(Phaseolus vulgaris): Evidence for Multiple Centers of Domestication. *Economic Botany, 40*(4), 451-468. https://doi.org/10.1007/BF02859659

Gogoi, N., Baruah, K. K., & Meena, R. S. (2018). Grain legumes: impact on soil health and agroecosystem. In *Legumes for Soil Health and Sustainable Management* (pp. 511-539): Springer.
González, A. M., Yuste-Lisbona, F. J., Saburido, S., Bretones, S., De Ron, A. M., Lozano, R., & Santalla, M. (2016). Major Contribution of Flowering Time and Vegetative Growth to Plant Production in Common Bean As Deduced from a Comparative Genetic Mapping. *Frontiers in Plant Science, 7*(1940). https://doi.org/10.3389/fpls.2016.01940
Gordillo, G. A., & Geiger, H. H. (2008). MBP (Version 1.0): A Software Package to Optimize Maize Breeding Procedures Based on Doubled Haploid Lines. *Journal of Heredity, 99*(2), 227-231. https://doi.org/10.1093/jhered/esm103
Haas, J. H., & Bolwyn, B. (1972). Ecology and epidemiology of Sclerotinia wilt of white beans in Ontario. *Canadian Journal of Plant Science, 52*(4), 525-533.
Heffner, E. L., Lorenz, A. J., Jannink, J. L., & Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop science, 50*(5), 1681-1690.
Heslot, N., Yang, H. P., Sorrells, M. E., & Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop science, 52*(1), 146-160.
Hoyos-Villegas, V., Song, Q., Wright, E. M., Beebe, S. E., & Kelly, J. D. (2016). Joint Linkage QTL Mapping for Yield and Agronomic Traits in a Composite Map of Three Common Bean RIL Populations. *Crop Science, 56*(5), 2546-2563. https://doi.org/10.2135/cropsci2016.01.0063
Hoyos-Villegas, V., Mkwaila, W., Cregan, P. B., & Kelly, J. D. (2015). Quantitative trait loci analysis of white mold avoidance in pinto bean. *Crop Science, 55*(5), 2116-2129.
iterate GmbH. (2020) *Cyberduck.* https://cyberduck.io/
Jahufer, M., & Luo, D. (2018). DeltaGen: A comprehensive decision support tool for plant breeders. *Crop Science, 58*(3), 1118-1131.
Kelly, J. D., Kolkman, J. M., & Schneider, K. (1998). Breeding for yield in dry bean (Phaseolus vulgaris L.). *Euphytica, 102*(3), 343-356.
Khosla, G., Gill, B. S., Sharma, P. (2019). Comparison of different breeding methods for developing superior genotypes in soybean. *Agricultural Research Journal, 56*(4), 628-634
Koenig, R., & Gepts, P. (1989). Allozyme diversity in wild Phaseolus vulgaris: further evidence for two major centers of genetic diversity. *Theoretical and Applied Genetics, 78*(6), 809-817. https://doi.org/10.1007/BF00266663
Kolkman, J. M., & Kelly, J. D. (2002). Agronomic Traits Affecting Resistance to White Mold in Common Bean. *Crop Science, 42*(3), 693-699. https://doi.org/10.2135/cropsci2002.6930
Kumar, J., Choudhary, A. K., Solanki, R. K., & Pratap, A. (2011). Towards marker-assisted selection in pulses: a review. *Plant Breeding, 130*(3), 297-313.
Larsen, J., Morneau, E., Zhang, B., Digweed, Q., Page, E. R., Mylnarek, J. J., & Wally, O. S. D. (2019). *Speed Breeding in Dry Beans*. Poster retrieved from
Li, X., Zhu, C., Wang, J., & Yu, J. (2012). Chapter six - Computer Simulation in Plant Breeding. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 116, pp. 219-264): Academic Press.
Liu, H., Tessema, B. B., Jensen, J., Cericola, F., Andersen, J. R., & Sørensen, A. C. (2019). ADAM-Plant: A Software for Stochastic Simulations of Plant Breeding From Molecular to Phenotypic Level and From Simple Selection to Complex Speed Breeding Programs. *Frontiers in Plant Science, 9*(1926). https://doi.org/10.3389/fpls.2018.01926

Lorenz, A. J. (2013). Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. *G3 Genes|Genomes|Genetics, 3*(3), 481-491. doi:10.1534/g3.112.004911

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., . . . Jannink, J.-L. (2011). Chapter Two - Genomic Selection in Plant Breeding: Knowledge and Prospects. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 110, pp. 77-123): Academic Press.

Maurer, H. P., Melchinger, A. E., & Frisch, M. (2004). *Plabsoft: software for simulation and data analysis in plant breeding.* Paper presented at the Genetic variation for plant breeding. Proceedings of the 17th EUCARPIA General Congress, Tulln, Austria, 8-11 September 2004.

Mendes, M. P., Botelho, F. B. S., Ramalho, M. A. P., Abreu, Â., & Furtini, I. V. (2008). Genetic control of the number of days to flowering in common bean. *Embrapa Arroz e Feijão-Artigo em periódico indexado (ALICE)*.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics, 157*(4), 1819.

McGill University Pulse Breeding and Genetics Laboratory. (2021). peas-andlove GitHub Repository. https://github.com/McGillHaricots/peas-andlove/tree/master/Simulation-files

Miklas, P. N., Johnson, W. C., Delorme, R., & Gepts, P. (2001). QTL conditioning physiological resistance and avoidance to white mold in dry bean. *Crop Science, 41*(2), 309-315.

Mukeshimana, G., Butare, L., Cregan, P. B., Blair, M. W., & Kelly, J. D. (2014). Quantitative Trait Loci Associated with Drought Tolerance in Common Bean. *Crop Science, 54*(3), 923-938. https://doi.org/10.2135/cropsci2013.06.0427

Muleta, K. T., Pressoir, G., & Morris, G. P. (2019). Optimizing Genomic Selection for a Sorghum Breeding Program in Haiti: A Simulation Study. *G3 Genes|Genomes|Genetics, 9*(2), 391-401. doi:10.1534/g3.118.200932

Nienhuis, J., & Singh, S. P. (1988). Genetics of Seed Yield and its Components in Common Bean (Phaseolus vulgaris L.) of Middle-American Origin. *Plant Breeding, 101*(2), 155-163. https://doi.org/10.1111/j.1439-0523.1988.tb00281.x

Norman, A., Taylor, J., Edwards, J., & Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3 Genes|Genomes|Genetics, 8*(9), 2889-2899. doi:10.1534/g3.118.200311

Oladzadabbasabadi, A., Mamidi, S., Miklas, P. N., Lee, R., & McClean, P. (2019). Linked candidate genes of different functions for white mold resistance in common bean (Phaseolus vulgaris. L) are identified by QTL-based pooled sequencing.

Osorno, J. M., Vander Wal, A. J., Posch, J., Simons, K., Grafton, K. F., Pasche, J. S., . . . Pastor-Corrales, M. (2021). A new black bean with resistance to bean rust: Registration of 'ND Twilight'. *Journal of Plant Registrations, 15*(1), 28-36.

Peng, B., & Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics,* 21:3686-3687.

Pérez-Vega, E., Pañeda, A., Rodríguez-Suárez, C., Campa, A., Giraldez, R., & Ferreira, J. J. (2010). Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (Phaseolus vulgaris L.). *Theoretical and Applied Genetics, 120*(7), 1367-1380.

Sandhu, K. S., You, F. M., Conner, R. L., Balasubramanian, P. M., & Hou, A. (2018). Genetic analysis and QTL mapping of the seed hardness trait in a black common bean (Phaseolus

vulgaris) recombinant inbred line (RIL) population. *Molecular Breeding, 38*(3), 34. doi:10.1007/s11032-018-0789-y
Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., . . . Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics, 46*(7), 707-713. doi:10.1038/ng.3008
Siddiq, M., & Uebersax, M. A. (2012). *Dry Beans and Pulses: Production, Processing and Nutrition*: Wiley.
Singh, S. P., Lépiz, R., Gutiérrez, J. A., Urrea, C., Molina, A., & Terán, H. (1990). Yield Testing of Early Generation Populations of Common Bean. *Crop Science, 30*(4), cropsci1990.0011183X003000040022x. doi:https://doi.org/10.2135/cropsci1990.0011183X003000040022x
Sinha, N. K., Hui, Y. H., Evranuz, E. Ö., Siddiq, M., & Ahmed, J. (2010). *Handbook of Vegetables and Vegetable Processing*: Wiley.
Soltani, A., Bello, M., Mndolwa, E., Schroder, S., Moghaddam, S. M., Osorno, J. M., . . . McClean, P. E. (2016). Targeted analysis of dry bean growth habit: Interrelationship among architectural, phenological, and yield components. *Crop Science, 56*(6), 3005-3015.
Soulé, M. E. (1987). *Viable Populations for Conservation*: Cambridge University Press.
Stagnari, F., Maggio, A., Galieni, A., & Pisante, M. (2017). Multiple benefits of legumes for agriculture sustainability: an overview. *Chemical and Biological Technologies in Agriculture, 4*(1), 2. doi:10.1186/s40538-016-0085-1
Taylor, J. F. (2014). Implementation and accuracy of genomic selection. *Aquaculture, 420-421*, S8-S14. doi:https://doi.org/10.1016/j.aquaculture.2013.02.017
Tinker, N., & Mather, D. (1993). GREGOR: software for genetic simulation. *Journal of Heredity, 84*(3), 237-237.
UN. (2021). *Our growing population*, 2021. https://www.un.org/en/global-issues/population
Wang, C., Kao, W.-H., & Hsiao, C. K. (2015). Using Hamming Distance as Information for SNP-Sets Clustering and Testing in Disease Association Studies. *PLOS ONE, 10*(8), e0135918. doi:10.1371/journal.pone.0135918
Wang, J., & Hill, W. G. (2000). Marker-Assisted Selection to Increase Effective Population Size by Reducing Mendelian Segregation Variance. *Genetics, 154*(1), 475-489. doi:10.1093/genetics/154.1.475
Wang, J., van Ginkel, M., Podlich, D., Ye, G., Trethowan, R., Pfeiffer, W., . . . Rajaram, S. (2003). Comparison of Two Breeding Strategies by Computer Simulation. *Crop Science, 43*(5), 1764-1773. doi:10.2135/cropsci2003.1764
Westermann, D., & Crothers, S. (1977). Plant population effects on the seed yield components of beans. *Crop Science, 17*, 493-496.
White, J. W., & Laing, D. R. (1989). Photoperiod response of flowering in diverse genotypes of common bean (Phaseolus vulgaris). *Field Crops Research, 22*(2), 113-128. https://doi.org/10.1016/0378-4290(89)90062-2
White, J. W., & Singh, S. P. (1991). Sources and inheritance of earliness in tropically adapted indeterminate common bean. *Euphytica, 55*(1), 15-19. doi:10.1007/BF00022554
Witcombe, J., & Virk, D. (2001). Number of crosses and population size for participatory and classical plant breeding. *Euphytica, 122*(3), 451-462.

Wright, E. M., & Kelly, J. D. (2011). Mapping QTL for seed yield and canning quality following processing of black bean (*Phaseolus vulgaris* L.). *Euphytica, 179*(3), 471-484. doi: 10.1007/s10681-011-0369-2

Wright, S. (1938). Size of population and breeding structure in relation to evolution. *Science, 87*(2263), 430-431. doi:10.1126/science.87.2263.425-a

Yonezawa, K., & Yamagata, H. (1978). On the number and size of cross combinations in a breeding programme of self-fertilizing crops. *Euphytica, 27*(1), 113-116.