




ORIGINAL ARTICLE

Special Section: Computational Design of Changing Cropping Systems

Plant breeding simulations with AlphaSimR

Jon Bančič^{1,†}  | Philip Greenspoon^{1,†}  | R. Chris Gaynor^{1,2}  | Gregor Gorjanc¹ 

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, Midlothian, UK

²Bayer Crop Science, Chesterfield, Missouri, USA

Correspondence

Jon Bančič, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, EH25 9RG, Midlothian, UK. Email: jbancic@ed.ac.uk

[†]Jon Bančič and Philip Greenspoon contributed equally to this work.

Assigned to Associate Editor Collins Kimbeng.

Funding information

Bayer Crop Science; BASF; Data-Driven Innovation - Edinburgh and South East Scotland City Region Deal; Lantmännens Forskningsstiftelse; Biotechnology and Biological Sciences Research Council, Grant/Award Numbers: BB/L020467/1, BB/R002061/1, BB/R019940/1, BBS/E/D/30002275, BBS/E/RL/230001A, BBS/E/RL/230001C; The University of Edinburgh; Marie Skłodowska-Curie Action; Limagrain

Abstract

Plant breeding plays a crucial role in the development of high-performing crop varieties that meet the demands of society. Emerging breeding techniques offer the potential to improve the precision and efficiency of plant breeding programs; however, their optimal implementation requires refinement of existing breeding programs or the design of new ones. Stochastic simulations are a cost-effective solution for testing and optimizing new breeding strategies. The aim of this paper is to provide an introduction to stochastic simulation with software AlphaSimR for plant breeding students, researchers, and experienced breeders. We present an overview of how to use the software and provide an introductory AlphaSimR vignette as well as complete AlphaSimR scripts of breeding programs for self-pollinated, clonal, and hybrid crops, including relevant breeding techniques, such as backcrossing, speed breeding, genomic selection, index selection, and others. Our objective is to provide a foundation for understanding and utilizing simulation software, enabling readers to adapt the provided scripts for their own use or even develop completely new plant breeding programs. By incorporating simulation software into plant breeding education and practice, the next generation of plant breeders will have a valuable tool in their quest to provide sustainable and nutritious food sources for a growing population.

1 | INTRODUCTION

Stochastic simulation is a cost-effective tool for refining breeding programs to increase their rate of genetic gain through the development, validation, and optimization of

breeding program modifications. Stochastic simulation also serves as a valuable communication tool and an educational platform for learning quantitative genetics and plant breeding principles. In this contribution, we describe the process of building a plant breeding program simulation and provide complete AlphaSimR scripts of different breeding programs and techniques, suitable for training, research, or even practical implementation.

Plant breeding is the genetic improvement of plants to meet the nutritional, cultural, economic, and technological

Abbreviations: AYT, advanced yield trial; DH, doubled haploid; EYT, elite yield trial; G×E, genotype by environment; GCA, general combining ability; GS, genomic selection; GWAS, genome-wide association study; PYT, preliminary yield trial; QTL, quantitative trait locus; SNP, single nucleotide polymorphism; SSD, single seed descent.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Crop Science* published by Wiley Periodicals LLC on behalf of *Crop Science* Society of America.

requirements of humans. Since its inception, plant breeding has gone through a significant transformation. The first form of selective breeding by humans started about 10,000 years ago with selection on traits such as non-shattering, seed/fruit appearance, and plant stature (Kingsbury, 2011). This domestication of plants lasted for millennia, transmitting knowledge and improved plant materials through generations, and developed a set of plant species that are still widely cultivated today (Meyer et al., 2012). From the beginning of the twentieth century, advances in our understanding of inheritance and genetics, along with new technologies and improved management practices, have enabled plant breeding to become more science-driven and effective (Bradshaw, 2017). This is reflected in maize and wheat yields nearly tripling and rice and soybean crop yields doubling between 1961 and 2021 (Roser, 2023). As the world continues to face global challenges such as climate change and population growth, ongoing improvements in plant breeding will be essential to develop resilient, high-yield crops that can meet the demands of the future.

A typical plant breeding pipeline involves steps consistent across most crop species (Figure 1). The pipeline begins with defining the product profile by identifying a list of “must-have” and “value-added” traits, the target growing area, and assembling germplasm as a source of new genetic variation (Cobb et al., 2019). Subsequently, the breeding program operates through continuous cycles of: (i) crossing parents to create a new segregating (recombining) population, (ii) evaluation and selection across multiple breeding stages, locations, and years to identify superior individuals, and (iii) recycling superior individuals as parents for the next breeding cycle. The final phase of the breeding pipeline usually takes a few years more and involves registering and multiplying the seed of a new variety before its commercial release (Atlin & Econopoulou, 2022; Covarrubias-Pazaran et al., 2022). The variation among programs arises from factors such as the crop species considered; their biology (e.g., age of sexual maturity); their reproductive system (e.g., self-pollinated, cross-pollinated, or clonally propagated); the importance of dominance genetic variation, which drives both heterosis and inbreeding depression; and the resources available to fund and support the breeding program (Schnell, 1982). The choice of release variety type (e.g., line, clone, hybrid, open-pollinated variety) also contributes further to program differences, with variety types differing in their desired level of genetic uniformity and heterozygosity.

Despite significant advancements in plant breeding, annual genetic gains in most staple crops fall short of the recommended targets stipulated by the Food and Agriculture Organization of the United Nations (FAO). On average, annual gains in staple crops are estimated to be around 1%, while FAO guidelines stipulate a minimum of 2% to meet the ever-growing global food demand (FAO, 2017; Ray et al.,

Core Ideas

- This research highlights the value of stochastic simulation in plant breeding research and practice.
- The key steps of building a plant breeding simulation with AlphaSimR are described.
- AlphaSimR example scripts for different plant breeding programs and techniques are provided.

2013). Several factors contribute to these limited annual genetic gains in plant breeding. First, breeders are unable to develop new stable varieties that perform well in increasingly challenging climate conditions in time. Second, breeders need to develop varieties that are not only higher yielding but also resource efficient, stress tolerant, and nutritious (Zhang & Cai, 2011). Breeding of multiple traits simultaneously reduces yield due to trade-offs and requires maintenance of genetic diversity that sometimes involves lengthy cycles to introgress donor material into elite material (Salgotra & Chauhan, 2023). Third, long breeding cycles, spanning 7–12 years in many cereal crops and up to 20 years in fruit and tree species, present a significant challenge for the timely development and release of improved varieties (Bernardo, 2014). To deal with these challenges, breeders must efficiently manage their program resources and continuously refine their programs to maximize the rate of genetic gain per unit of time and investment.

Many new technologies and methods offer the potential to transform plant breeding programs to improve their efficiency and increase genetic gain (Varshney et al., 2021). These innovations can be used to target different parameters of the breeder's equation (Cobb et al., 2019; Lush, 1937). For example, genomics, CRISPR-Cas technology, and gene editing can be used in pre-breeding programs to generate new genetic variation through introgression, targeted allele manipulation, and targeted recombination (Allier et al., 2020; Breider et al., 2022; Bernardo, 2017; Gorjanc et al., 2016; Jenko et al., 2015; Johnsson et al., 2019; Yin et al., 2017). Advanced statistical models and, more recently, machine learning algorithms that integrate genomics (Bernardo & Yu, 2007; Bernardo, 2009; Bančič et al., 2023; Gorjanc et al., 2015; Meuwissen et al., 2001), environomics (Costa-Neto et al., 2021; Jarquín et al., 2014; Tolhurst et al., 2022), phenomics (Araus & Cairns, 2014; Cobb et al., 2013; Xu, 2016), and other -omics (Chao et al., 2023; Christensen et al., 2021; Zhao et al., 2022) can be used to improve selection accuracy and increase selection intensity. The combination of speed breeding and genomic selection can shorten the length of a breeding cycle within recurrent selection programs (Gaynor et al., 2017; Watson et al., 2018). However, applying some of these innovations in breeding programs is constrained by uncertainties about

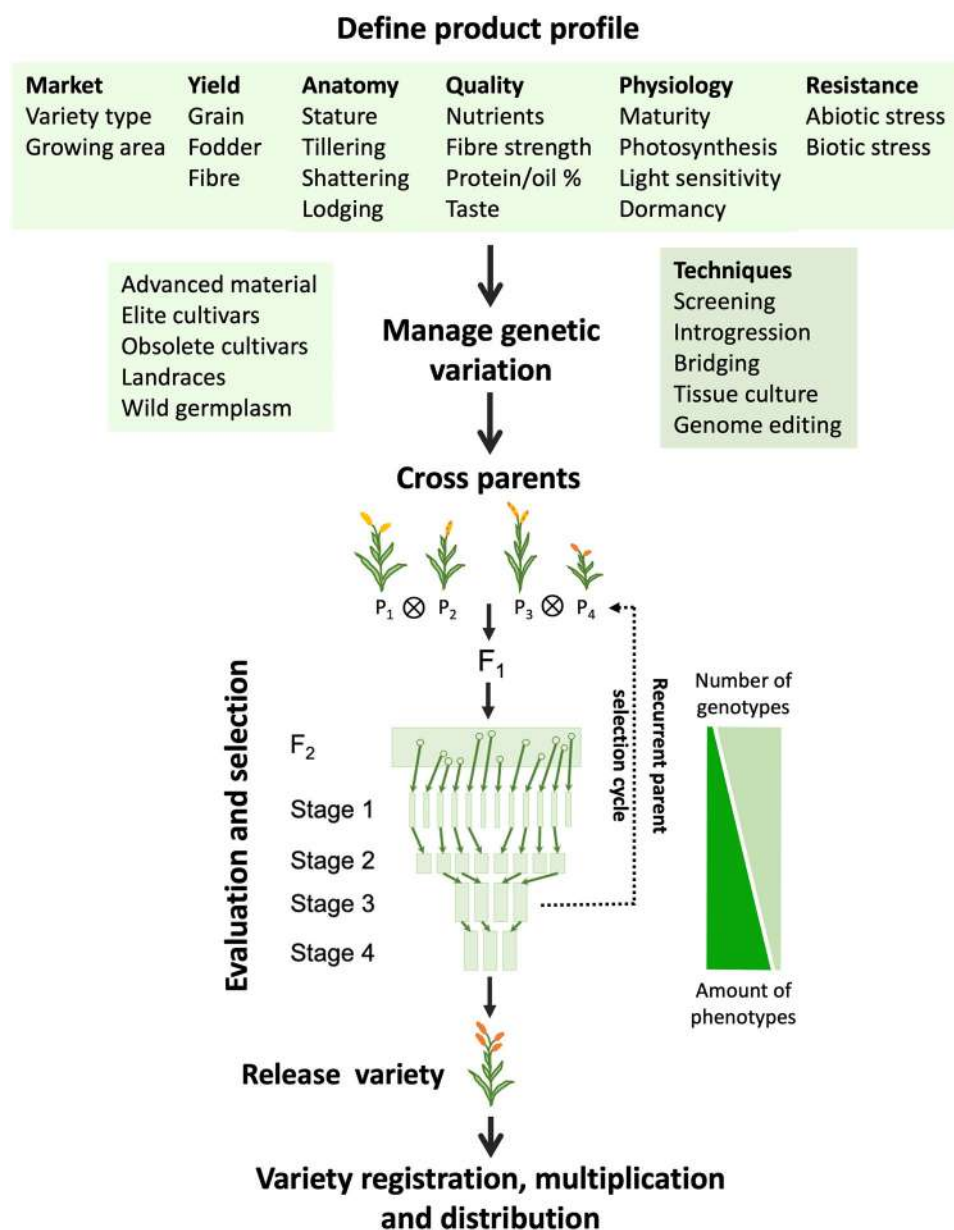


FIGURE 1 Common steps in a plant breeding program pipeline.

their potential impact on long-term genetic trends, logistical feasibility, and cost implications.

Stochastic simulation is a cost-effective tool for testing new ideas and theory, as well as guiding the integration of new technologies and methods to refine real breeding programs. Unlike simpler short-term modeling approaches using deterministic equations such as breeder's equation (e.g., Atlin & Econopouly, 2022), stochastic simulation, incorporating the stochasticity arising during founder initialization, non-random mating, inheritance, and artificial selection, enables modeling the entire complexity of breeding programs to assess expected long-term genetic trends as well as variation of potential outcomes (Li et al., 2012). This approach is particularly useful for the development, valida-

tion, and optimization of targeted program modifications. Hence, stochastic simulation can help breeders make decisions, mitigating the risk of resource wastage before the real-world empirical testing and implementation of these modifications. The power of stochastic simulation, bolstered by modern computing, has motivated the release of several simulation software, available as stand-alone programs or as packages for the R, Python, or Julia environment. Prominent examples include AlphaSim(R) (Faux et al., 2016; Gaynor et al., 2021), AdamPlant (Liu et al., 2019), ChromaX (Younis et al., 2023), MoBPS (Pook et al., 2020), PyBrOpS (Shrote & Thompson, 2023), QU-GENE (Podlich & Cooper, 1998), and XSim (Chen et al., 2022). These software provide flexibility by accommodating features such as additive, dominance,

and epistatic genetic effects; non-random mating and complex family structures; selection on a single or multiple (correlated) traits controlled by different numbers of quantitative trait loci (QTLs); and to some extent, genotype by environment interaction. The R package AlphaSimR has become a particularly popular choice for researchers due to its flexibility and user-friendliness and is therefore the focus of this paper.

The flexibility of stochastic simulation to answer diverse research and practical questions is reflected in the wide range of published literature. Simulation has been applied to study the implementation of genomic selection in breeding programs (Gaynor et al., 2017; Labroo et al., 2023), propose new breeding systems (Bančič et al., 2021), optimize mating strategies (Allier et al., 2020; Bernardo & Yu, 2007), compare statistical models (Bančič et al., 2024; De Jong et al., 2023; Gezan et al., 2010), evaluate genotyping and imputation strategies (Gorjanc et al., 2017, 2017), and test strategies for introducing and managing genetic variation (Allier et al., 2020; Bernardo, 2009; Gorjanc et al., 2018). As a reference, Table 1 provides a list of simulation studies relevant to plant breeding that used AlphaSimR software. The wide range of studies highlights the usefulness and importance of gaining proficiency in simulation for students, researchers, and practitioners. Despite many simulation studies, perspective papers on how to approach such studies (Covarrubias-Pazaran et al., 2022; Simianer et al., 2021), and online course material (<https://www.edx.org/course/breeding-programme-modelling-with-alphasimr>), there is still a shortage (and demand) for literature guiding the practical implementation and deployment of breeding simulations.

This paper serves as an introductory guide to stochastic simulation of plant breeding programs using the R package AlphaSimR (Gaynor et al., 2021). Our target readers are plant breeding students and experienced breeders or researchers interested in integrating simulations into their toolkit. Through a detailed walk-through of an AlphaSimR wheat breeding program simulation, we share a systematic approach to building a plant breeding program simulation. Additionally, we provide scripts for various breeding programs (self-pollinated, clonal, and hybrid plants) and breeding techniques and features (backcrossing, speed breeding, genomic selection, index selection, etc.) commonly used in plant breeding to highlight AlphaSimR's flexible nature and plasticity of R scripting. These resources are intended as educational and reference materials that offer a starting point for designing bespoke plant breeding programs.

2 | MATERIAL AND METHODS

This section is divided into three parts to provide an introduction to stochastic simulation of breeding programs using the R package AlphaSimR. The first part outlines the key steps for

building a plant breeding simulation. The second part summarizes examples of breeding programs for self-pollinated, clonal, and hybrid crop species, along with different strategies for each breeding program. The third part describes examples of common breeding techniques and features that can be implemented within a breeding program or used independently for hypothesis testing. All examples are supported with scripts and are available on GitHub repository https://github.com/HighlanderLab/jbancic_alphasimr_plants. Throughout, we maintain a simple narrative, avoiding mathematical descriptions of the simulation steps involving quantitative genetics and statistics, and refer readers to previous publications for these details (Faux et al., 2016; Gaynor et al., 2017, 2021).

We recommend that readers first review Section 2 to familiarize themselves with the paper's content and then proceed to Section 3, which provides a detailed description of a wheat breeding program simulation along with key results. This simulation description is supplemented by a walk-through R Markdown vignette (see files `LineBreeding.{Rmd,html}` in the GitHub repository).

2.1 | Key simulation steps

The process of building a breeding program simulation can be split into seven steps:

1. Outlining the breeding program,
2. Specifying global parameters,
3. Simulating genomes and founders,
4. Filling the breeding pipeline,
5. Running the burn-in phase,
6. Running the future phase with competing scenarios, and
7. Replication and examining the results.

Detailed explanations of these steps are in Section 3 along with a practical example of a wheat breeding program and the corresponding walk-through R Markdown vignette in the GitHub repository (https://github.com/HighlanderLab/jbancic_alphasimr_plants). Although the steps may seem linear, they typically involve continuous revisions and iterations, which all require considerable time to complete. We used AlphaSimR version 1.5.3 and expect future versions to retain backward compatibility with the code presented.

2.2 | Common breeding programs

This subsection introduces three distinct plant breeding programs for self-pollinated, clonal, and hybrid crop species. Each breeding program is briefly introduced and supported with a corresponding AlphaSimR script (Table A.1). The

TABLE 1 Simulation studies relevant to plant breeding using AlphaSimR.

Study	Description	Code
Gaynor et al. (2017)	Develops a two-part genomic selection strategy in a wheat line program	no
Cowling et al. (2020)	Tests a strategy for the formation of heterotic pools	no
Powell et al. (2020)	Tests a two-part genomic selection strategy in a maize hybrid program	no
Bančič et al. (2021)	Develops strategies for intercrop breeding with genomic selection	yes
Batista et al. (2021)	Compares long-term independent culling and index selection	no
Borges da Silva et al. (2021)	Compares performance of spatial correction models with simulated data	yes
Gonen et al. (2021)	Examines performance of a new imputation software with simulated data	no
Lara et al. (2021)	Analyzes genetic variance over time and genome in a breeding program	yes
Yang et al. (2021)	Examines the usefulness of molecular markers in DUS evaluation system	yes
Aono et al. (2022)	Examines a machine learning approach for genomic prediction	no
Breider et al. (2022)	Proposes a strategy for introgression of exotic germplasm into an elite program	yes
Covarrubias-Pazaran et al. (2023)	Examines a genetic complementation method for creation of heterotic groups	no
Huang et al. (2022)	Investigates factors influencing the rate of genetic gain in sugar kelp program	yes
Mancin et al. (2022)	Demonstrates restricted index selection with multiple traits	yes
Pocrnic et al. (2022)	Optimizes APY method for large scale genomic analyses	yes
Powell et al. (2022)	Integrates gene-to-phenotype biological model for bud outgrowth in simulation	yes
Sabadin et al. (2022)	Develops genomic selection strategies for single-seed decent rice program	yes
Yang et al. (2022)	Develops a method for modeling allele frequency change over years	yes
De Jong et al. (2023)	Compares selection on GCA with different statistical models	yes
Epstein et al. (2023)	Examines the impact of engineering the crossover/recombination landscape	yes
Fritsche-Neto et al. (2024)	Compares different index selection approaches	yes
Jannink et al. (2023)	Uses Bayesian optimization for resource allocation in two-part clonal program	yes
Krause et al. (2023)	Compares linear mixed models to estimate realized genetic gains in a soybean program	no
Labroo et al. (2023)	Develops hybrid breeding strategies for clonal programs	yes
Lubanga et al. (2023)	Develops genomic selection strategies for a tea program	yes
Lanzl et al. (2023)	Examines the impact of mating designs on estimation of genetic variance	yes
Oliveira et al. (2023)	Analyzes drivers of genetic change in a breeding program	yes
Platten and Fritsche-Neto (2023)	Compares three methods of QTL introgression into elite material	yes
Werner et al. (2023)	Develops genomic selection strategies for a strawberry clonal program	no
Fritsche-Neto et al. (2024)	Examines factors influencing the genetic gain in a rice hybrid program	no
Azevedo et al. (2024)	Examines factors influencing genomic prediction accuracy for visual score traits	no
Bakare et al. (2024)	Evaluates narrow-sense and broad-sense selection strategy in a cassava program	yes
Bančič et al. (2024)	Develops framework for simulating genotype by environment interaction	yes
Peixoto et al. (2024)	Examines the impact of high-throughput phenotyping and genomic selection in a hybrid program	yes
Endelman (2024)	Compares long-term effectiveness of a new mate allocation method with OCS	yes

Abbreviations: APY, algorithm for proven and young; DUS, distinctness uniformity and stability; GCA, general combining ability; OCS, optimal contribution selection; QTL, quantitative trait locus.

AlphaSimR scripts for each breeding program adhere to the above-mentioned steps and report trends in genetic mean, genetic variance, and selection accuracy.

2.2.1 | Breeding for self-pollinated species

Breeding programs for self-pollinated species include those of wheat, rice, barley, oat, and soybean. Briefly, a breed-

ing cycle begins by crossing inbred parental lines to produce an F_2 segregating population. This is followed by several rounds of growing, selection on highly heritable traits, and self-pollination of individual plants. Once a desired level of homozygosity is reached, lines are evaluated for other traits in trials across multiple environments and years. Line breeding programs leverage additive and epistatic genetic variation, and the release variety is a fixed inbred/pure line, which can be multiplied through repeated selfing and bulking.

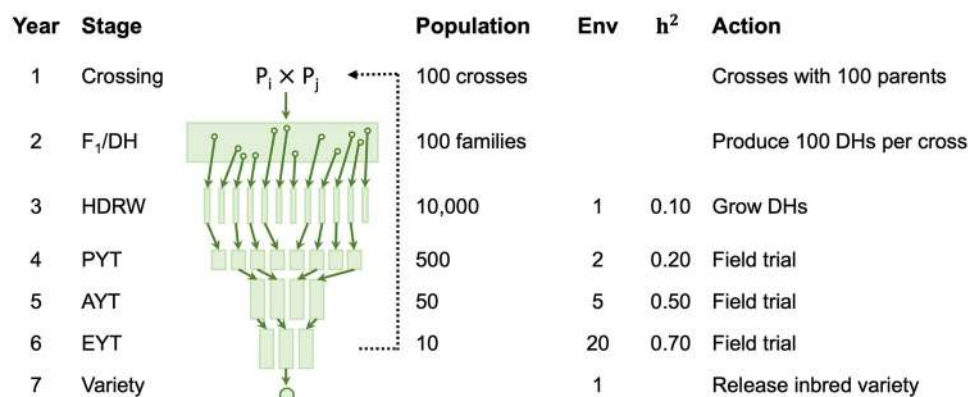


FIGURE 2 Key features of a hypothetical wheat line breeding program using double haploid technology following Gaynor et al. (2017). Presented are key breeding stages with different populations (AYT, advanced yield trial; DH, doubled haploid; EYT, elite yield trial; HDRW, headrow; PYT, preliminary yield trial), each involving different numbers of individuals, numbers of environments (Env), heritability of yield, and action taken. The black dotted line indicates the recurrent selection of new parents with the best phenotypic performance.

Different traditional breeding methods can be used to derive inbred lines, such as pedigree, bulk, and single seed descent (SSD) methods (Allard, 1999; Acquaah, 2009; Fehr et al., 1987; Rutkoski et al., 2022). These methods typically require several years of selfing to derive fixed inbred lines. Modern breeding programs use doubled haploid (DH) technology to obtain true homozygous lines in less than 2 years after parental crosses, thereby substantially shortening the breeding cycle. In contrast to traditional methods which rely on F_2 segregating populations, DH lines are directly generated from F_1 plants using either inducer line (e.g., wheat, corn) or anther/microspore culture (e.g., oilseed rape) to produce haploids, followed by chromosome doubling induction via a chemical agent.

AlphaSimR scripts (Table A.1, Programs 1– 4) demonstrate the simulation of a wheat breeding program using the mass selection method, SSD method, pedigree method, and DH technology. The mass selection method involves recurrent selection of best individual plants or families to increase the frequency of desirable alleles. The pedigree method, a slower and more intense method, involves making specific parental crosses to produce F_2 families for successive family and within-family selection until plants are sufficiently inbred and ready for multi-environment trials. Parent–progeny information is recorded for every candidate line to assist the retention of the variation among lines and produce new crosses. The SSD method accelerates inbreeding by selecting a single seed per plant each generation, maintaining genetic diversity by tracing each individual back to a unique F_2 individual. Each selfing round, seeds can be bulked or kept separate per F_2 plant depending on the strategy. Once sufficiently inbred, plants advance to multi-environment trials for further line development. A program shown in Figure 2 creates DH lines with line development and multi-environment trials commencing immediately after their creation. These

scripts showcase AlphaSimR's capability to model various breeding approaches.

In addition to different selfing methods with phenotypic selection, we provide AlphaSimR scripts for two breeding scenarios that implement genomic selection with DH technology (Table A.1, Programs 5 and 6). The first uses a genomic selection strategy within the context of a conventional wheat breeding program with concurrent population improvement and product development, while the second uses genomic selection to separate the components of population improvement and product development into the two-part strategy of Gaynor et al. (2017). For faster execution on the personal computer, both genomic selection scenarios are implemented in our scripts with a small training population with only 2 years of data, which may not produce optimal results.

2.2.2 | Breeding for clonally propagated species

Breeding programs for clonally propagated species include those of tea, strawberries, citrus trees, cassava, and banana (Allard, 1999; Acquaah, 2009; Fehr et al., 1987). Briefly, a breeding cycle begins by crossing heterozygous parents through sexual reproduction to produce a segregating F_1 population. Each cross produces unique true F_1 seed, which immediately holds the potential to develop into a new variety. The F_1 seedlings undergo preliminary screening in an unreplicated trial to identify promising candidate plants. Selected F_1 plants are then propagated through clonal or vegetative means (such as cuttings, corms, or tubers) for further evaluation in trials across multiple environments and years. Clonal breeding programs leverage the additive and non-additive genetic variation, and the release variety is an outbred individual, which can be multiplied through repeated vegetative propagation.

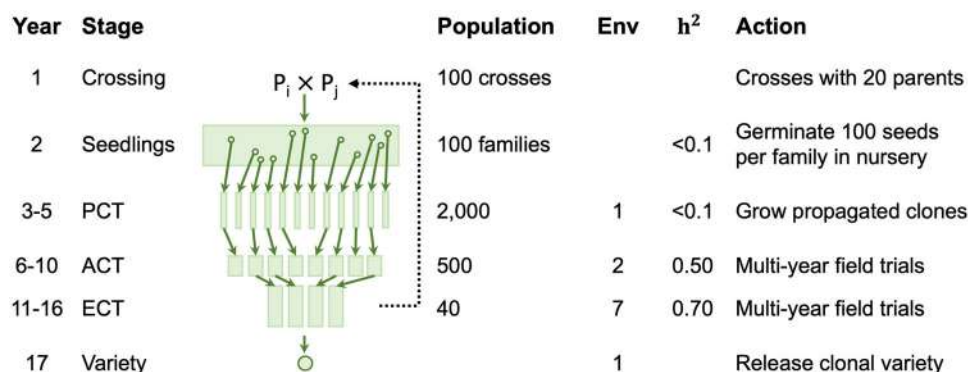


FIGURE 3 Key features of a hypothetical tea clonal breeding program following Lubanga et al. (2023). Presented are key breeding stages with different populations (ACT, advanced clonal trial; ECT, elite clonal trial; PCT, preliminary clonal trial), each involving different numbers of individuals, numbers of environments (Env), heritability, and action taken. The black dotted line indicates the recurrent selection of new parents with the best phenotypic performance.

AlphaSimR scripts (Table A.1, Programs 7–9) demonstrate simulations of a hypothetical tea breeding program that uses phenotypic selection as shown in Figure 3 and two program variations using either pedigree- or genomic-based selection strategy as described in Lubanga et al., 2023. Both scenarios use limited training population sizes for faster execution on personal computers.

2.2.3 | Breeding for hybrid cultivars

Breeding for hybrid cultivars is used in self-pollinated and out-crossing species, including maize, sorghum, wheat, rice, sunflower, and various vegetables (Allard, 1999; Acquaah, 2009; Fehr et al., 1987). Briefly, the process of developing hybrid cultivars uses reciprocal recurrent selection, which consists of three main steps: First, inbred parental lines within two or more genetically distinct (heterotic) groups are crossed to produce segregating populations. Second, selected inbred lines from one heterotic group are crossed with tester lines from the opposite group to produce testcross F_1 hybrids. These hybrids are evaluated in a few trials to identify inbred lines with the highest general combining ability (GCA), measured as the average performance of a line across various testcross combinations. Third, inbred lines with the highest GCA from each heterotic group are intercrossed to produce F_1 hybrids. These hybrids are evaluated across multiple environments and years to identify those with the highest specific combining ability (SCA), measured as the average performance of a specific hybrid combination. The F_1 hybrids typically have a large proportion of heterozygous loci that exhibit some level of dominance, resulting in heterosis (i.e., F_1 progeny surpass the performance of their inbred parents). Hybrid programs leverage both additive and dominance variation, and the release variety is an outbred F_1 hybrid, which can be multiplied through repeated crossing of its inbred parents.

AlphaSimR scripts (Table A.1, Programs 10–12) demonstrate the simulation of a hypothetical maize hybrid breeding program with phenotypic selection as shown in Figure 4 and two scenarios with genomic selection scenarios. The first scenario applies genomic selection in the context of a traditional hybrid breeding program with concurrent population improvement and product development, and the second scenario applies genomic selection to separate the components of population improvement and product development into the two-part strategy of Gaynor et al. (2017) and Powell et al. (2020). Scripts for both scenarios use limited training population sizes for faster execution on personal computers, which may not produce optimal results. The R scripts also highlight features of AlphaSimR such as the specification of heterotic groups, performing testcrosses and hybrid crosses, as well as calculating, predicting, and selecting on GCA.

2.3 | Common breeding techniques and features

In this subsection, we describe several common plant breeding program techniques and features, presented as independent AlphaSimR scripts (Table A.2) that can be readily integrated into breeding simulations.

Example 1: Mating plans: Planned crosses serve multiple purposes in plant breeding, including the generation of new genetic variation, estimation of genetic variances, formation of heterotic groups, and mapping of major QTLs. AlphaSimR script (Table A.2, Feature 1) demonstrates several mating designs, including biparental crosses, testcrosses, half- and full-diallel crosses within both a single and two populations. AlphaSimR script (Table A.2, Feature 2) for GWAS simulation is also provided.

Example 2: Setting heritability: There are several ways that heritability can be set in AlphaSimR. AlphaSimR script

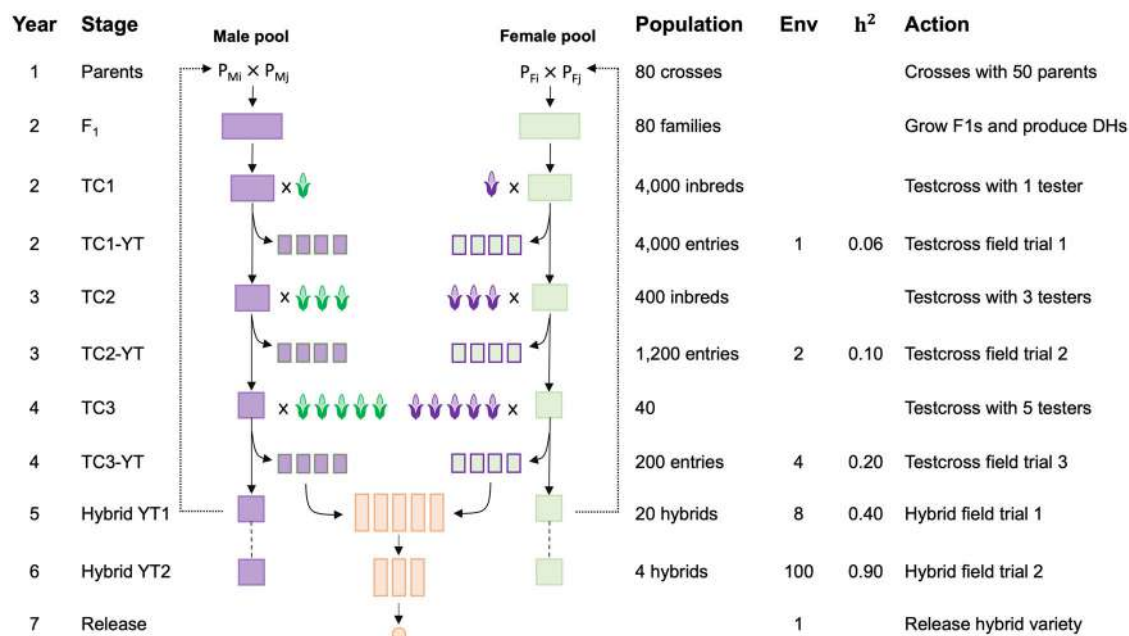


FIGURE 4 Key features of a hypothetical maize hybrid breeding program following Bernardo (2009). Presented are key breeding stages (TC, testcross; TC-YT, testcross yield trial; Hybrid YT, hybrid yield trial), each involving different numbers of individuals in a population, numbers of environments (Env), heritability, and action taken. The black dotted line indicates the reciprocal recurrent selection of new parents with the best general combining ability.

(Table A.2, Feature 3) demonstrates setting narrow-sense heritability, broad-sense heritability, and error variance with the number of replications.

Example 3: Multi-trait selection with selection index: This approach is commonly used for the simultaneous selection of multiple traits, each with an assigned (economic) weight. AlphaSimR script (Table A.2, Feature 4) demonstrates how to simulate uncorrelated or correlated traits and how to perform index selection. AlphaSimR offers a miscellaneous slot to store additional information about the population such as selection index values for each individual to be used during selection (see Feature 5 in Table A.2).

Example 4: Genotype by environment (G×E) interaction: Individual genotypes respond differently across environments, complicating selection outcomes. AlphaSimR offers two approaches for modeling G×E: one driven by a single latent environmental covariate and the other using multiple correlated traits, interpreted as the same trait across different environments. AlphaSimR script (Table A.2, Feature 6) demonstrates how both approaches work.

Example 5: Genomic models: Different models can be fit to genomic data. Models can include only additive random effects or can include both additive and dominance random effects. Models can also incorporate population-specific random effects (i.e., GCA and SCA), applicable to hybrid breeding. AlphaSimR script (Table A.2, Feature 7) demonstrates the fitting of different models using internal and external model functionality.

Example 6: Trait introgression: This technique facilitates the transfer of a new trait, such as a major QTL allele or a transgene, from a donor parent into the superior genetic background of the recurrent parent (Bernardo, 2009). Typically, breeders aim to recover the genome of the recurrent parent while introducing the new allele of interest into the resulting line. This process can be achieved through phenotypic selection but can be expedited with marker-assisted selection. AlphaSimR script (Table A.2, Feature 8) demonstrates the use of backcrossing with marker-assisted selection to introgress a desired trait, controlled by a single locus, into the background of a recurrent parent.

Example 7: Genome editing: This technique allows for precise modification of plant genomes by targeting specific locus (e.g., editing QTLs with CRISPR-Cas) to introduce or modify specific traits at a faster rate than a traditional trait introgression. AlphaSimR script (Table A.2, Feature 9) demonstrates the editing of a single locus in the entire population.

Example 8: Speed breeding; This technique shortens the breeding cycle by manipulating plants under controlled greenhouse conditions to accelerate their growth and development. For example, speed breeding allows up to six generations of wheat plants to be grown annually under ideal conditions, compared to typical two generations in outdoor conditions (Watson et al., 2018). AlphaSimR script (Table A.2, Feature 10) demonstrates a simulation of four generations per year. Speed breeding can also be applied to accelerate selfing and for rapid cycling with DH technology in two-part breeding

programs, as we show in the implementation of a two-part wheat breeding program (Table A.1, Program 6).

These examples highlight just a small number of potential applications in plant breeding. For more potential applications and research ideas, users are encouraged to review the literature in Table 1.

3 | RESULTS

In this section, we demonstrate the key simulation steps outlined in Section 2.1. We use the wheat breeding program with DH technology as a case study to compare phenotypic selection and genomic selection. Each simulation step is accompanied by a description, highlighting aspects particular to the wheat breeding program example. We suggest readers consult the information provided below concurrently with the R Markdown vignette (files `LineBreeding.{Rmd,html}`). The vignette replicates steps 3.2–3.5 over a single year of the breeding program, including some of the text for completeness. The complete multi-year simulations of the wheat breeding program are available as AlphaSimR scripts.

3.1 | Outlining the breeding program

Before initiating a breeding simulation, it is essential to clearly define the research question and determine the desired complexity of the simulation. Once this is established, we start by mapping out the key stages and actions of the breeding program. This step requires gathering biological, logistical, statistical, and agronomic details from breeding program managers and specialists. The information gathered in this step is the foundation for the subsequent translation of a breeding program into code. What is below includes a non-exclusive list of things to consider:

- *Breeding objectives*: identifying key traits for improvement.
- *Stage-specific features*: determining the traits evaluated, the number of genotypes, and field trials at each stage.
- *Genomic features*: gathering information on the crop's ploidy, genome, evolution and genetic architecture of key traits.
- *Breeding population estimates*: gathering information on genetic diversity, population structure, levels of inbreeding and heterosis, genetic means and variances of key traits, their heritability estimates at each breeding stage, and correlations with other traits.
- *Program specifics*: understanding the amount and type of data generated for key traits, statistical models used, and the target growing area.
- *Crop's biology*: understanding reproductive biology, life cycle, and identifying the crop's biological restrictions,

such as mating incompatibility and seed multiplication rate limits.

- *Logistical and cost limitations*: understanding the annual limitations and maximum capacity of parental crosses, line production, and field trials in a breeding program.

Establishing an overall budget for the baseline breeding program involves gathering data on costs per unit, including expenses associated with parent crossing, DH production, nursery maintenance, genotyping, and phenotyping. With a budget in hand, it is possible to ensure fair comparisons of different strategies within common cost constraints.

The aim of our wheat breeding example is to compare a phenotypic program to an alternative program implementing genomic selection. We present the key features of our phenotypic wheat breeding program, including the number of individuals at each stage, heritability of a hypothetical grain yield trait, and selection actions in Figure 2. Additionally, in Table 2, we outline approximate costs for both phenotypic and genomic selection breeding programs, using approximate unit costs for demonstration purposes. To constrain the costs of the genomic program to those of the phenotypic program (\$503,500), we reduced the number of DH individuals produced per parental cross from 100 to 89 and skipped the headrow stage. Other strategies for cost reduction, such as decreasing the extent of field testing, can be considered when a new costly strategy is introduced. However, these strategies need to be logistically viable and made in consultation with the breeding program manager. We stress that changing the costs of actions can have a profound impact on the choice of the optimal strategy.

3.2 | Specifying global parameters

To translate a breeding program into AlphaSimR code, we begin by defining simulation parameters based on the information gathered in the previous step. We specify values such as the number of crosses, progeny per cross, and individuals evaluated at different stages and genetic details including variances, the number of QTLs, degree of dominance, and demographic history. In Table 3, we list parameters, definitions, and the values used for simulating the wheat breeding program example.

Careful selection of global parameter values is crucial to ensure that the resulting properties of a simulated population, such as population structure, level of heterosis or inbreeding, and genetic variation, somewhat reflect those of a real breeding program. Parameter selection may rely on estimates from prior analyses of empirical data or observed long-term trends. Deciding on parameter values may require simulating the founder parents and breeding program multiple times until the realized population properties resemble expectations. We

TABLE 2 Cost comparison between the phenotypic and the cost-constrained genomic selection wheat breeding program.

Action	Cost (\$)	Env	Phenotypic		Genomic	
			# Units	Cost (\$)	# Units	Cost (\$)
Cross	30/cross		100	3000	100	3000
Grow F1s	30/plant		100	3000	100	3000
Make DHs	30/plant		10,000	300,000	8900	267,000
Genotype	15/plant				8900	<i>133,400</i>
HDRW	10/plot	1	10,000	100,000		
PYT	20/plot	5	500	50,000	500	50,000
AYT	50/plot	15	50	37,500	50	37,500
EYT	50/plot	20	10	10,000	10	10,000
			Total	503,500	Total	504,500

Note: Text in italics indicates the cost incurred by genotyping in the genomic selection program, offset by the smaller number of DH individuals and skipping the HDRW stage.

Abbreviations: AYT, advanced yield trial; DH, doubled haploid; EYT, elite yield trial; HDRW, headrow; PYT, preliminary yield trial.

stress that users must be cautious to avoid biasing parameters to favor a specific scenario in the study.

3.3 | Simulating genomes and founders

In the next step, we initialize a founder population and specify trait-related and other features. Many crop species have experienced significant and repeated bottlenecks during domestication and selective breeding, resulting in distinct population structure and linkage disequilibrium patterns (Meyer et al., 2012). To account for this, AlphaSimR embeds MaCS software (Chen et al., 2009) to generate founder genomes through a backward-in-time (coalescent) simulation. The software creates genealogical trees based on recombination and demographic parameters and drops mutations onto the trees to produce whole-chromosome haplotypes that form genomes and genotypes of founder individuals. These founder individuals then serve as the initial parents in the breeding program simulation.

In AlphaSimR, we can create founder genomes via three approaches: (i) selecting from pre-defined species' demographic histories, (ii) creating custom species histories, or (iii) importing externally obtained haplotypes (Table A.2, Features 11 and 12). For quick testing, we can sample founder haplotypes randomly using a Bernoulli distribution, which results in no linkage disequilibrium and allele frequencies that follow a normal distribution. Alternatively, the user can specify a "generic" species to rapidly generate founders with a somewhat realistic demography for an agricultural species. To create a custom species demography, we need to specify parameters such as the number of chromosomes, loci, mutation and recombination rates, and demographic history in terms of changes in effective population size over time. It is important to note that simulating realistic demographic

histories is challenging, both due to the lack of knowledge and excessive run times of complex demographic simulations. We have observed that different methods of simulating genomes and their parameters can influence breeding program outcomes, though they generally do not impact relative performance of different scenarios. Finally, we can import externally obtained haplotypes, either simulated with other software or derived from observed genotypes, which must be phased into haplotypes. Although some users might assume that importing their genomic data is sufficient to replicate observed phenotypic variation in simulations, this is generally not the case because phased genotypes for SNP markers usually lack QTL information, let alone their effects on key traits.

Additionally, AlphaSimR enables simulations with haploid, diploid, and polyploid species. Some functions even work with variable ploidy within a simulation. For allopolyploid species, we recommend specifying their polyploid genome as diploid with more chromosome pairs exhibiting disomic inheritance. For autopolyploid species, we recommend specifying their polyploid genome as the number of chromosome groups exhibiting polysomic inheritance. While AlphaSimR's forward-in-time simulation adequately accommodates variable ploidy, the MaCS call for backward-in-time simulation of autopolyploid genomes requires appropriate modifications (Arnold et al., 2012; Labroo et al., 2023).

After creating the founder population, we can add trait-related and other features. We can add single or multiple (correlated or uncorrelated) traits and specify the distribution of QTL effects (normal or gamma), means, and variance-covariance matrices for QTL effects (additive, dominance, and epistatic). It is also possible to import a trait with manually specified QTLs and their effects (see Feature 10 in Table A.2). Furthermore, we have the ability to incorporate sex differentiation (if relevant for species), employ one or more SNP

TABLE 3 Global parameters and definitions with values used for simulating the wheat breeding program.

Parameter	Definition	Value
nReps	Number of simulation replications	10
nBurnin	Number of years in the burn-in phase	20
nFuture	Number of years in future phase	20
nQTL	Number of QTLs per chromosome	20
nSnp	Number of SNPs per chromosome	400*
initMeanG	Initial population mean genetic value for yield trait	0
initVarG	Initial population genetic variance for yield trait	1
initVarGE	Initial GxE interaction variance for yield trait	2
varE	Yield trial error variance for yield trait	4
nParents	Number of parents to start a breeding cycle	50
newParents	Number of new parents each breeding cycle	50
nCrosses	Number of crosses among parents to start a breeding cycle	100
nDH	Number of DH individuals produced per cross	100 (89)
famMax	Maximum number of DH individuals per cross to enter PYT	10
nPYT	Number of entries in PYT	500
nAYT	Number of entries in AYT	50
nEYT	Number of entries in EYT	10
repHDRW	Effective replication in HDRW	4/9
repPYT	Effective replication in PYT	1
repAYT	Effective replication in AYT	4
repEYT	Effective replication in EYT	8
startTP	Year to start collecting training records for GS	18*

Note: Values in italics represent the reduced number of DH individuals to offset the cost of implementing genomic selection and * indicates parameters added for the genomic selection program.

Abbreviations: AYT, advanced yield trial; DH, doubled haploid; EYT, elite yield trial; GS, genomic selection; HDRW, headrow; QTL, quantitative trait locus; PYT, preliminary yield trial; SNP, single nucleotide polymorphism.

arrays (including QTL or not), account for sex-specific recombination, adjust for the frequency of quadrivalent pairing in autopolyploids, and apply the parameters of the gamma-sprinkling recombination model (Falque et al., 2009). We can also enable tracking of pedigree, founder haplotypes, recombination, and segregation events.

In our wheat breeding program example, we utilize a pre-defined wheat species demographic history and consider a single grain yield trait for demonstration purposes. In Table 4, we summarize the founder genomes and trait parameters used in the simulation.

3.4 | Filling the breeding pipeline

Once the breeding program stages are defined and the founder population is simulated, we continue by filling each stage with a distinct population or cohort. This step is necessary to mimic the simultaneous running of different breeding cycles in real breeding programs. In practical terms, this means that a breeding program consists of multiple cohorts, each originating from the same parental population, that progress through

different stages of the breeding pipeline simultaneously. For each cohort, we carry out the filling process by crossing the founder parents and progressing the new populations of genotypes through the stages of our breeding program, saving a different cohort for each breeding stage.

Our wheat breeding program example has six stages, meaning we need six distinct cohorts, and the filling process is illustrated in Figure 5. The first cohort (orange) progresses through all stages and is saved in the sixth (final) stage (elite yield trial, EYT), making it the oldest cohort. The second cohort (yellow) progresses to the fifth stage (advanced yield trial, AYT), making it the second oldest cohort. This process repeats six times until all stages are populated with distinct cohorts (see year 0), with the last being the most recent (dark gray). Despite using the constant founder parents, unique cohorts arise due to different crossings, randomness of genetic inheritance (drift, recombination, and mutation), and selections at each stage. As the burn-in phase commences in year 1 (described in the next subsection), all cohorts, excluding the oldest which has already completed all stages, will simultaneously progress diagonally down the pipeline each year. A new cohort (cohort 7) is also generated in the earliest stage,

TABLE 4 Simulation features of the wheat breeding program. Presented are steps taken in the burn-in and the future phase pertaining to the genome, founders, and trait.

Phase	Action	Feature
Burn-in	Specifying founder genomes	100,000 generations of evolution
		50 inbred founders
		10 chromosome pairs
		1.43 Morgans per chromosome
		8×10^8 base pairs per chromosome
	Specifying trait features	2×10^{-9} mutation rate
		Grain yield
		1000 QTLs per chromosome
	Simulating recent breeding	Normally distributed QTL effects
		For other values, see Table 3
Future	Simulating future breeding	20 years of breeding
		Doubled haploid lines
		Phenotypic selection
		Track mean, variance, and selection accuracy
		20 years of breeding
		Test genomic selection
		Constrained and unconstrained costs
		4K SNP array
		5 years of training records for genomic selection
		Ridge regression BLUP for genomic selection

Abbreviations: BLUP, best linear unbiased prediction; QTLs, quantitative trait loci; SNP, single nucleotide polymorphism.

representing the choosing of parents to start a new breeding cycle. Once the burn-in phase begins, the distinct cohorts that were created during the filling process will either stay distinct (as in our mass selection example, Table A.1, Program 1) or, more commonly, will mix (e.g., in the other programs in Table A.1). The mixing will take place during the formation of parents for a new breeding cycle, during which parents chosen from the EYT stage of the previous breeding cycle (curved upward arrow) are combined with parents that have been used to begin previous breeding cycles (right-pointing arrow).

3.5 | Running the burn-in phase

Before comparing different competing scenarios, we first simulate a burn-in phase to provide a common starting point from which scenarios can be compared. This phase is necessary for two reasons: (i) to represent a historical breeding period preceding the implementation of a new strategy; and (ii) to achieve realistic evolution of linkage-disequilibrium across the genome under selection so that genetic structure and diversity mimic those of a real breeding program. During the burn-in phase, we also initiate the collection of data, which will be analyzed once the simulation ends. To streamline this process, we set up empty data frames that serve as

containers for various variables. These data frames typically include identifier columns (e.g., scenario name, replication number, breeding stage, and simulation year), specific simulation input parameters, and the variables tracked throughout the simulation. AlphaSimR provides many built-in functions for calculating genetic parameter values that align with classical genetics theory, including genetic and phenotypic means and variances for each trait (under both random and non-random mating), as well as partitioning total genetic variance into genic, additive, dominance, additive-by-additive components, along with their covariances. The collected data enable results to be reported on specific stages or across all breeding stages over the duration of the simulation. It also assists in identifying any errors in the simulation code. In some cases, we might wish to save entire populations throughout the simulation (e.g., for a temporal analysis study), which can be achieved by storing them in lists.

In our wheat breeding program example, the burn-in phase lasts 20 years and we use a phenotypic selection program as the baseline (Figure 2 and Table 4). Following the process shown in Figure 5, each new burn-in year involves: (i) selecting the best individuals from the previous year's EYT stage as parents and crossing them to create a new cohort, and (ii) progressing all cohorts one stage ahead. We initiate the collection of the genetic mean, the genetic variance, and the selection

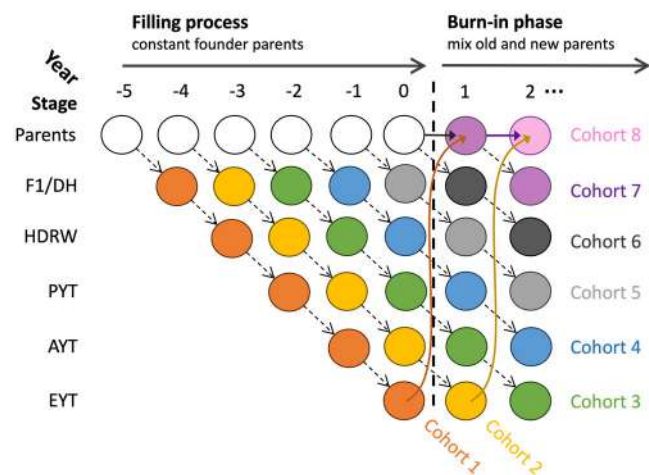


FIGURE 5 An illustration of filling the breeding pipeline and the beginning of the burn-in phase in the wheat breeding program. *Filling process*: initiates cohorts of the breeding program using constant founder parents over six iterations. In the first iteration, cohort 1 (orange) passes to the sixth stage (elite yield trial, EYT). In the second iteration, cohort 2 (yellow) passes to the fifth stage (advanced yield trial, AYT), and so on. After the last iteration, each stage contains a population of a different cohort. *Burn-in phase*: In the breeding cycles after filling, each cohort progresses diagonally from left to right to the next stage of the breeding program. The cohorts are transitioning through the breeding program simultaneously, and a new cohort is generated each year (e.g., cohorts 7 and 8) due to the creation of new parents. DH, doubled haploid; HDRW, headrow; PYT, preliminary yield trial.

accuracy at the DH stage. We chose the DH stage because it is the earliest common stage among all competing scenarios in the future phase and contains the largest number of individuals upon which to calculate genetic parameter values. Additionally, we initiate the collection of historical records for training a genomic prediction model that will be used in the future phase when we introduce genomic selection.

3.6 | Running the future phase with competing scenarios

Following the burn-in phase, we proceed to assess various breeding scenarios, each introducing novel features. The choice and properties of these scenarios depend on the objectives of our research question. For example, the future phase could help determine the most effective crossing strategy for maximizing genetic variation with available parents, identifying the optimal number and recycling of parents for maximizing genetic gain, optimizing the number of observational trials at a particular stage, or studying the effect of training population composition on prediction accuracy (see Table 1). In our wheat breeding example, the future phase lasts 20 years in addition to the burn-in phase, and we use

it to compare a phenotypic program to an alternative program implementing genomic selection (Table 4).

Evaluating a new feature in the future phase may require the use of external software. If the software is available as an R package, we can directly incorporate it into the AlphaSimR script. However, for software not available as an R package, we need to pause the simulation each time external software is required, export the dataset in a format compatible with the external software, conduct the analysis, import the results back into R, and then resume the simulation. This process can be automated using R's functionality to interact with the operation system (e.g., using `system()` R function). Several functions already exist in AlphaSimR that can assist with tasks such as extracting SNP or QTL genotype or haplotype matrices and phenotype data from a target population or preparing export data (e.g., for export to PLINK).

In Figure 6, we present the implementation of genomic selection in our example to: (i) reduce the program's breeding length by advancing individuals directly from the DH stage to the preliminary yield trial (PYT) stage, (ii) shorten the program's breeding cycle by selecting parents from the earliest stage (DH) rather than the EYT stage using genomic estimated breeding values (Figure 2), and (iii) improve selection accuracy in the DH stage with genomic prediction. This represents just one potential implementation of genomic selection, and typically, several different scenarios would be tested to identify the optimal one. We use an internal AlphaSimR ridge regression function to train a genomic prediction model and assign estimated breeding values to individuals in a population. Alternatively, a different software could be used for prediction as demonstrated in Feature 7 (see Table A.2). The collection of descriptive and genetic parameter values continues into the future phase for each scenario to allow comparisons to be made. Note that the selection accuracy is measured as the correlation between estimated (genomic or phenotypic) and true genetic values and is recorded before advancing individuals from HDRW or DH to the PYT stage for phenotypic and genomic selection, respectively.

3.7 | Replication and examining the results

In the final step, we summarize and examine the replicated outcomes from different competing scenarios. Due to simulation stochasticity, we need to replicate the entire simulation multiple times. Replication serves at least two purposes: (i) to capture the key sources of variation and their impact on the breeding program and (ii) to compute and compare summary statistics of tracked parameters (genetic mean, genetic variance, accuracy, etc.) across multiple scenarios. For statistical analysis, outcomes of the simulations may be modeled with a mixed effects model with replicate as a random effect and parameters of interest or scenarios as fixed effects. The

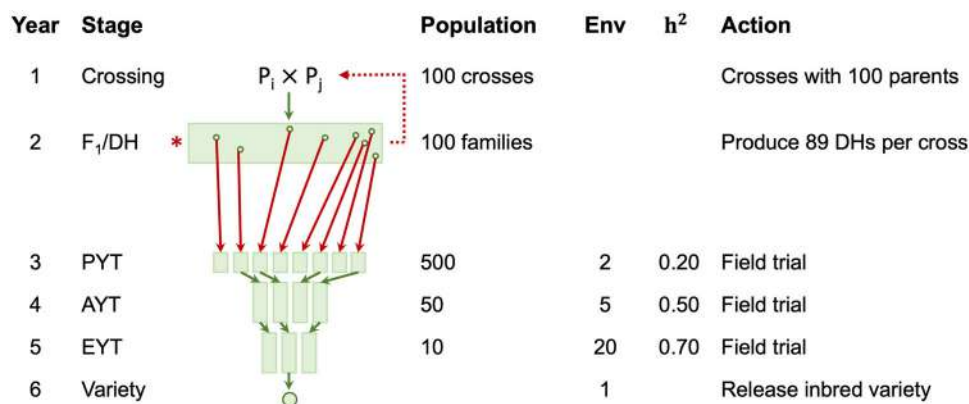


FIGURE 6 Key features of a hypothetical wheat breeding program with doubled haploid technology implementing genomic selection. Presented are key breeding stages with different populations (AYT, advanced yield trial; DH, doubled haploid; EYT, elite yield trial; PYT, preliminary yield trial), each involving different numbers of plant genotypes in a population, environments (Env), heritability of yield, and action taken. The actions in red highlight the changes enabled by genomic selection and are described in Section 3.6. The red dotted line indicates the rapid recurrent genomic selection of new parents with the highest genomic estimated breeding values.

optimal number of simulation replications is determined by the objectives of the study, the required level of precision, the complexity of the simulation, and available computational resources. We recommend that users conduct as many replications as feasible to achieve stable distributions of tracked statistics and their means. To overcome long compute times, especially for complex simulations, we suggest users parallelize multiple replicates using a computer cluster. To do so, we first run one burn-in phase for each replicate and save a snapshot of the R environment, including states of variables and populations, at the end of the burn-in. Next, we start parallel, independent R sessions for the future phase of each scenario, importing our burn-in snapshots to serve as a common starting point (within a replicate), with results saved and collated afterward. Doing so, each burn-in replicate is used by several scenarios at the same time, and the elapsed time of running all scenarios is significantly reduced. Finally, we stress the importance of thoroughly examining the results by discussing them with colleagues and providing explanations for unexpected trends, which may lead to multiple iterations of the simulation experiment before final conclusions can be made.

In Figure 7, we show the trends of genetic mean, genetic variance, and selection accuracy in our wheat breeding program example over 40 years comparing phenotypic selection and genomic selection. Trends are presented as an average across 100 simulation replicates (thick lines) in both the burn-in phase and the future phase and for genetic gain also as individual simulation replicates (thin lines). Our results indicate a clear advantage of genomic selection over phenotypic selection at the expense of more rapid loss in genetic variance. Consistent with previous studies (referenced in Table 1), the positive difference in the genetic mean is due to the shortened length of the breeding program and breed-

ing cycle and increased selection accuracy in breeding stages where genomic selection was applied. Variability around the mean line and the boxplot showing the simulation outcomes in the final year highlight the highly stochastic nature of simulations and the importance of replication. Furthermore, the figure also compares the outcome of an unconstrained genomic selection program with a total cost of \$553,500 to test the achievable gain at the program's maximum capacity. The results suggest that an additional investment of \$49,000 to genotype 2000 more individuals increases genetic gain only slightly compared to the cost-constrained genomic selection scenario. We note that to shorten simulation running times, all genomic selection scenarios restrict the size of training populations to only 2 years' worth of data, which may have influenced the outcomes since a larger training population can increase selection accuracy. For further insights and discussions, readers are encouraged to consult the literature cited in Table 1.

4 | DISCUSSION

Stochastic simulation has great potential to help improve plant breeding programs to meet the rising food demands amidst global population growth and climate change. Previous research has highlighted the value of simulation in optimizing breeding programs, implementing new technologies, answering research and practical questions, and designing novel breeding approaches (Table 1). However, there remains a lack of literature providing practical guidance on implementing and deploying such simulations. This paper aims to fill this gap by providing an introduction to the use of stochastic simulation in plant breeding using the R package AlphaSimR (Gaynor et al., 2021). We illustrate the process of building

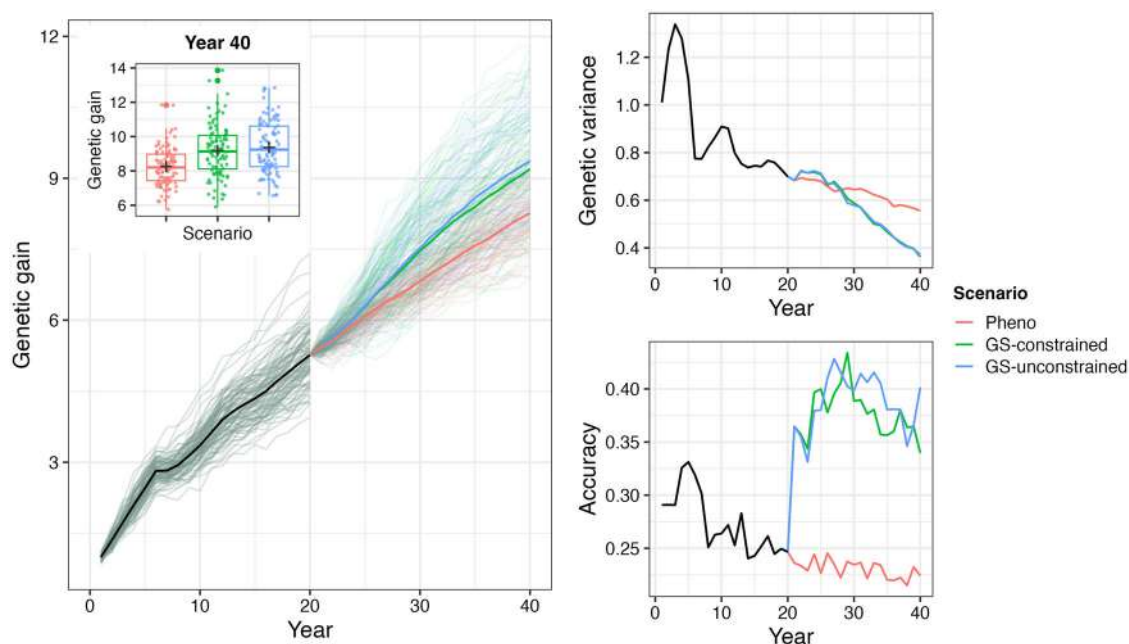


FIGURE 7 Genetic trends of the wheat breeding program using phenotypic selection (Pheno) or genomic selection (GS) with either constrained or unconstrained costs over 40 years. Presented are genetic mean (left), genetic variance (top-right), and selection accuracy (bottom-right) at the doubled haploid stage. Black lines represent the burn-in phase and colored lines represent different scenarios of the future phase. Thin lines indicate gain in individual replicates, and thick lines indicate the average gain across all 100 simulation replicates. Boxplots represent the distribution of outcomes in the final simulation year. For clarity, we omit individual replicates for genetic variance and selection accuracy.

a plant breeding program simulation through a detailed walk-through of a wheat breeding program example. Additionally, we provide AlphaSimR scripts for various breeding programs (self-pollinated, clonal, and hybrid crops) and techniques (backcrossing, speed breeding, genomic selection, index selection, etc.) to highlight AlphaSimR's flexibility coupled with the plasticity of R scripting. These resources are intended as educational and reference materials that offer a starting point for designing bespoke plant breeding programs or studies of them.

Simulations, like any theoretical model, rest upon a set of assumptions typically made for simplification. While these assumptions can never fully capture the complexities of real-world systems, the aim is to approximate reality closely enough to provide meaningful insights. If an assumption is seriously flawed, it may lead to results that are unsuitable for the study system. For example, if genotype-by-environment interactions are assumed to be absent for the sake of simplicity, but are in fact significant, then simulation outcomes may differ importantly from reality. Other assumptions are made due to the current lack of knowledge. For example, the positions and numbers of QTLs in the simulated genomes do not match their true locations in the genomes of the studied crop. This caveat means that the genomes we model are not an accurate *in silico* reproduction of real genomes, and the results we obtain (e.g., from testing genomic prediction) can be largely regarded as a qualitative guide. These limitations highlight the

need to view simulation results as guiding insights that inform further research and hypothesis testing, rather than providing exact predictions.

In light of the limitations facing the modeling of breeding programs, there are several areas where development is ongoing or needed to make simulation stronger. Here, we highlight six areas of development:

- *Genotype by environment (G×E) interaction* presents an important challenge in plant breeding. The existing methods provided by AlphaSimR do not adequately capture the complexity of G×E interaction. A scalable and reproducible framework for simulating G×E interaction of any complexity and structure is being developed (see Bančič et al., 2024). This will enable the construction of realistic multi-environment field trials within breeding simulation to allow for comparison of different statistical models relevant to plant breeding, evaluation of different selection strategies and experimental designs, and obtaining more realistic genetic trends.
- *Optimizing large parameter space* is one of the inherent challenges of stochastic simulation studies. The number of potential parameter combinations, coupled with the time-consuming and computationally demanding nature of simulations often limits users to explore only a small subset of possibilities. Recently, Bayesian optimization-type approaches have been proposed to manage the exploration

of a large set of possible parameters more efficiently (Diot & Iwata, 2023; Hassanpour et al., 2023; Jannink et al., 2023).

- *Genome simulation* is often done approximately because crop species evolution and domestication are only partially known. With rapid advances in genomics, the characterization of genomes and demographies of multiple species and populations is becoming available (Gower et al., 2022; Lauterbur et al., 2023). Community-curated population genetics models for various species, facilitated by software such as stdpopsim and msprime (Baumdicker et al., 2022; Lauterbur et al., 2023), are becoming available and will enhance user convenience and reproducibility. This will allow a more accurate and efficient simulation of the founder genomes (Haller & Messer, 2019).
- *Ancestral recombination graphs* (ARGs) can be used to represent the genealogical history of haplotypes, which enables genotypic data to be stored more efficiently. Current efforts for implementing ARGs in AlphaSimR will enable simulations with whole-genomes for larger populations that are typical of commercial breeding programs as well as import simulated founder genomes with the complete mutation and recombination history from the software mentioned above.
- *Multi-population selection* is important for some crop species, such as forages, where breeding programs perform group selection rather than individual selection. We are expanding AlphaSimR functionality to work with entries that are themselves populations, as has already been done in an insect breeding simulator (Obšteter et al., 2023).
- *Model-specified phenotype simulation* is required to specify custom sources of variation. Currently, AlphaSimR limits the user in the number of fixed terms and random terms with pre-defined distributions that can be added to the simulation of a phenotypic value. While there are existing attempts to add additional sources of variation, for example, plot errors (Werner et al., 2024), a more general model-based approach is needed for simulating phenotypes. This approach would allow users to specify custom fixed and random effects such as location, genotype by location, genotype by location by year, and various linear and non-linear functions for different covariates.

As these developments continue, plant breeding simulations will become an increasingly powerful tool for accelerating the genetic improvement of crops. We encourage readers to actively contribute to this effort by participating in the free online edX course, Breeding Programme Modelling with AlphaSimR (<https://www.edx.org/course/breeding-programme-modelling-with-alphasimr>), consulting the technical documentation (<https://cran.r-project.org/web/packages/AlphaSimR/AlphaSimR.pdf>), participating in user discus-

sions and issues on GitHub (<https://github.com/gaynorrr/AlphaSimR>), and extending the existing literature outlined in Table 1. Ultimately, every breeding program would benefit from having a corresponding virtual/digital twin as a testing ground for new ideas before incorporating them into practice. By integrating the use of simulation software into plant breeding education and practice, future plant breeders will be equipped with a useful tool.

AUTHOR CONTRIBUTIONS

Jon Bančič: Data curation; formal analysis; investigation; methodology; resources; validation; visualization; writing—original draft; writing—review and editing. **Philip Greenspoon:** Data curation; formal analysis; investigation; methodology; resources; validation; visualization; writing—original draft; writing—review and editing. **R. Chris Gaynor:** Data curation; investigation; methodology; software; supervision; writing—review and editing. **Gregor Gorjanc:** Conceptualization; data curation; formal analysis; funding acquisition; project administration; resources; supervision; validation; writing—review and editing.

ACKNOWLEDGMENTS

The authors acknowledge funding from BBSRC (grants BBS/E/D/30002275, BBS/E/RL/230001A, BBS/E/RL/230001C, BB/L020467/1, BB/R019940/1, and BB/R002061/1), Bayer Crop Science, BASF, Limagrain, Lantmännen, Data-Driven Innovation—Edinburgh and South East Scotland City Region Deal, Marie Skłodowska-Curie Action, and The University of Edinburgh. For the purpose of open access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Several sources have funded the development of AlphaSimR software and its use in plant breeding and genetics research, consulting, and teaching.

CONFLICT OF INTEREST STATEMENT


The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

AlphaSimR scripts for all the breeding programs and techniques presented in this manuscript (Tables A.1 and A.2) are available at the GitHub repository: https://github.com/HighlanderLab/jbancic_alphasimr_plants.

ORCID

Jon Bančič  <https://orcid.org/0000-0001-7077-7163>

Philip Greenspoon  <https://orcid.org/0000-0001-6284-7248>

R. Chris Gaynor  <https://orcid.org/0000-0003-0558-6656>

Gregor Gorjanc  <https://orcid.org/0000-0001-8008-2787>

REFERENCES

- Acquaah, G. (2009). *Principles of plant genetics and breeding*. John Wiley & Sons.
- Allard, R. W. (1999). *Principles of plant breeding*. John Wiley & Sons.
- Allier, A., Teyssèdre, S., Lehermeier, C., Moreau, L., & Charcosset, A. (2020). Optimized breeding strategies to harness genetic resources with different performance levels. *BMC Genomics*, 21, 1–16. <https://doi.org/10.1186/s12864-020-6756-0>
- Aono, A. H., Ferreira, R. C. U., Moraes, A. d. C. L., frontiers, L. A. d. C., Pimenta, R. J. G., Costa, E. A., Pinto, L. R., Landell, M. G. d. A., Santos, M. F., Jank, L., Barrios, S. C. L., do Valle, C. B., Chiari, L., Garcia, A. A. F., Kuroshu, R. M., Lorena, A. C., Gorjanc, G., & de Souza, A. P. (2022). A joint learning approach for genomic prediction in polyploid grasses. *Scientific Reports*, 12(1), 12499. <https://doi.org/10.1038/s41598-022-16417-7>
- Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, 19(1), 52–61. <https://doi.org/10.1016/j.tplants.2013.09.008>
- Arnold, B., Bomblies, K., & Wakeley, J. (2012). Extending coalescent theory to autotetraploids. *Genetics*, 192(1), 195–204. <https://doi.org/10.1534/genetics.112.140582>
- Atlin, G. N., & Econopoulou, B. F. (2022). Simple deterministic modeling can guide the design of breeding pipelines for self-pollinated crops. *Crop Science*, 62(2), 661–678. <https://doi.org/10.1002/csc2.20684>
- Azevedo, C. F., Ferrão, L. F. V., Benevenuto, J., de Resende, M. D. V., Nascimento, M., Nascimento, A. C. C., & Munoz, P. R. (2024). Using visual scores for genomic prediction of complex traits in breeding programs. *Theoretical and Applied Genetics*, 137(1), 9. <https://doi.org/10.1007/s00122-023-04512-w>
- Bakare, M. A., Kayondo, S. I., Kulakow, P., Rabbi, I. Y., & Jannink, J.-L. (2024). Evaluating breeding for broad versus narrow adaptation for cassava in Nigeria using stochastic simulation. *Crop Science*, 64(2), 603–616. <https://doi.org/10.1002/csc2.21170>
- Bančič, J., Gorjanc, G., & Tolhurst, D. (2024). A framework for simulating genotype by environment interaction using multiplicative models. *Research Square*. <https://doi.org/10.21203/rs.3.rs-3855188/v1>
- Bančič, J., Ovenden, B., Gorjanc, G., & Tolhurst, D. (2023). Genomic selection for genotype performance and stability using information on multiple traits and multiple environments. *Theoretical and Applied Genetics*, 136(5), 104. <https://doi.org/10.1007/s00122-023-04305-1>
- Bančič, J., Werner, C. R., Gaynor, R. C., Gorjanc, G., Odeny, D. A., Ojulung, H. F., Dawson, I. K., Hoad, S. P., & Hickey, J. M. (2021). Modeling illustrates that genomic selection provides new opportunities for intercrop breeding. *Frontiers in Plant Science*, 12, 605172. <https://doi.org/10.3389/fpls.2021.605172>
- Batista, L. G., Gaynor, R. C., Margarido, G. R., Byrne, T., Amer, P., Gorjanc, G., & Hickey, J. M. (2021). Long-term comparison between index selection and optimal independent culling in plant breeding programs with genomic prediction. *PLoS ONE*, 16(5), e0235554. <https://doi.org/10.1371/journal.pone.0235554>
- Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, C. E., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretschmar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., ... Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220, iyab229. <https://doi.org/10.1093/genetics/iyab229>
- Bernardo, R. (2009). Genomewide selection for rapid introgression of exotic germplasm in maize. *Crop Science*, 49(2), 419–425. <https://doi.org/10.2135/cropsci2008.08.0452>
- Bernardo, R. (2017). Prospective targeted recombination and genetic gains for quantitative traits in maize. *The Plant Genome*, 10(2), plantgenome2016–11. <https://doi.org/10.3835/plantgenome2016.11.0118>
- Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3), 1082–1090. <https://doi.org/10.2135/cropsci2006.11.0690>
- Bernardo, R. N. (2014). *Essentials of plant breeding*. Stemma Press.
- Borges da Silva, E. D., Xavier, A., & Faria, M. V. (2021). Joint modeling of genetics and field variation in plant breeding trials using relationship and different spatial methods: A simulation study of accuracy and bias. *Agronomy*, 11(7), 1397. <https://doi.org/10.3390/agronomy11071397>
- Bradshaw, J. E. (2017). Plant breeding: past, present and future. *Euphytica*, 213, 1–12. <https://doi.org/10.1007/s10681-016-1815-y>
- Breider, I., Gaynor, R. C., Gorjanc, G., Thorn, S., Pandey, M. K., Varshney, R. K., & Hickey, J. M. (2022). A multi-part strategy for introgression of exotic germplasm into elite plant breeding programs using genomic selection. *Research Square*. <https://doi.org/10.21203/rs.3.rs-1246254/v1>
- Chao, H., Zhang, S., Hu, Y., Ni, Q., Xin, S., Zhao, L., Ivanisenko, V. A., Orlov, Y. L., & Chen, M. (2023). Integrating omics databases for enhanced crop breeding. *Journal of Integrative Bioinformatics*, 20(4), 20230012. <https://doi.org/10.1515/jib-2023-0012>
- Chen, C. J., Garrick, D., Fernando, R., Karaman, E., Stricker, C., Keehan, M., & Cheng, H. (2022). XSim version 2: Simulation of modern breeding programs. *G3 Genes|Genomes|Genetics*, 12(4), jkac032. <https://doi.org/10.1093/g3journal/jkac032>
- Chen, G. K., Marjoram, P., & Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Research*, 19(1), 136–142. <https://doi.org/10.1101/gr.083634.108>
- Christensen, O. F., Börner, V., Varona, L., & Legarra, A. (2021). Genetic evaluation including intermediate omics features. *Genetics*, 219(2), iyab130. <https://doi.org/10.1093/genetics/iyab130>
- Cobb, J. N., DeClerck, G., Greenberg, A., Clark, R., & McCouch, S. (2013). Next-generation phenotyping: Requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics*, 126, 867–887. <https://doi.org/10.1007/s00122-013-2066-0>
- Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., & Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: Lessons from the breeder's equation. *Theoretical and Applied Genetics*, 132, 627–645. <https://doi.org/10.1007/s00122-019-03317-0>
- Costa-Neto, G., Fritsche-Neto, R., & Crossa, J. (2021). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity*, 126(1), 92–106. <https://doi.org/10.1038/s41437-020-00353-1>
- Covarrubias-Pazarán, G., Gebeyehu, Z., Gemenet, D., Werner, C., Labroo, M., Sirak, S., Coaldrake, P., Rabbi, I., Kayondo, S. I., Parkes, E., Kanju, E., Mbanjo, E. G. N., Agbona, A., Kulakow, P., Quinn, M., & Debaene, J. (2022). Breeding schemes: What are they, how to formalize them, and how to improve them? *Frontiers in Plant Science*, 12. <https://doi.org/10.3389/fpls.2021.791859>
- Covarrubias-Pazarán, G., Werner, C., & Gemenet, D. (2023). Reciprocal recurrent selection based on genetic complementation: An efficient way to build heterosis in diploids due to directional dominance. *Crop Science*. <https://doi.org/10.1002/csc2.21018>

- Cowling, W. A., Gaynor, R. C., Antolín, R., Gorjanc, G., Edwards, S. M., Powell, O., & Hickey, J. M. (2020). In silico simulation of future hybrid performance to evaluate heterotic pool formation in a self-pollinating crop. *Scientific Reports*, 10(1), 4037. <https://doi.org/10.1038/s41598-020-61031-0>
- De Jong, G., Powell, O., Gorjanc, G., Hickey, J. M., & Gaynor, R. C. (2023). Comparison of genomic prediction models for general combining ability in early stages of hybrid breeding programs. *Crop Science*. <https://doi.org/10.1002/csc2.21105>
- Diot, J., & Iwata, H. (2023). Bayesian optimisation for breeding schemes. *Frontiers in Plant Science*, 13, 1050198. <https://doi.org/10.3389/fpls.2022.1050198>
- Endelman, J. B. (2024). *Genomic prediction of heterosis, inbreeding control, and mate allocation in outbred diploid and tetraploid populations*. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2024.07.01.601581>
- Epstein, R., Sajai, N., Zekowski, M., Zhou, A., Robbins, K., & Pawlowski, W. (2023). Exploring impact of recombination landscapes on breeding outcomes. *Proceedings of the National Academy of Sciences USA*, 120(14), e2205785119. <https://doi.org/10.1073/pnas.2205785119>
- Falque, M., Anderson, L. K., Stack, S. M., Gauthier, F., & Martin, O. C. (2009). Two types of meiotic crossovers coexist in maize. *The Plant Cell*, 21(12), 3915–3925. <https://doi.org/10.1105/tpc.109.071514>
- FAO. (2017). *The future of food and agriculture: Trends and challenges*. FAO. http://www.fao.org/publications%0Ahttp://www.fao.org/3/a-i6583e.pdf%0Ahttp://siteresources.worldbank.org/INTARD/825826-1111044795683/20424536/Ag_ed_Africa.pdf%0Awww.fao.org/cfs%0Ahttp://www.jstor.org/stable/4356839%0Ahttps://ediss.uni-goettingen.de/bitstream/han
- Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., Hearne, S. J., Gonen, S., & Hickey, J. M. (2016). AlphaSim: Software for breeding program simulation. *The Plant Genome*, 9(3), plantgenome2016.02.0013. <https://doi.org/10.3835/plantgenome2016.02.0013>
- Fehr, W. R. (1987). *Principles of cultivar development: Theory and technique* (Vol. 1). Macmillan Publishing Company.
- Fritsche-Neto, R., Ali, J., De Asis, E. J., Allahgholipour, M., & Labroo, M. R. (2024). Improving hybrid rice breeding programs via stochastic simulations: Number of parents, number of hybrids, tester update, and genomic prediction of hybrid performance. *Theoretical and Applied Genetics*, 137(1), 3. <https://doi.org/10.1007/s00122-023-04508-6>
- Gaynor, C. R., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: An R package for breeding program simulations. *G3: Genes, Genomes, Genetics*, 11(2), jkaa017. <https://doi.org/10.1093/g3journal/jkaa017>
- Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., & Hickey, J. M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, 57(5), 2372–2386. <https://doi.org/10.2135/cropsci2016.09.0742>
- Gezan, S. A., White, T. L., & Huber, D. A. (2010). Accounting for spatial variability in breeding trials: A simulation study. *Agronomy Journal*, 102(6), 1562–1571. <https://doi.org/10.2134/agronj2010.0196>
- Gonen, S., Wimmer, V., Gaynor, R. C., Byrne, E., Gorjanc, G., & Hickey, J. M. (2021). Phasing and imputation of single nucleotide polymorphism data of missing parents of biparental plant populations. *Crop Science*, 61(4), 2243–2253. <https://doi.org/10.1002/csc2.20409>
- Gorjanc, G., Battagin, M., Dumasy, J.-F., Antolín, R., Gaynor, R. C., & Hickey, J. M. (2017). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Science*, 57(1), 216–228. <https://doi.org/10.2135/cropsci2016.06.0526>
- Gorjanc, G., Cleveland, M. A., Houston, R. D., & Hickey, J. M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genetics Selection Evolution*, 47, 1–14. <https://doi.org/10.1186/s12711-015-0102-z>
- Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R. C., Antolín, R., & Hickey, J. M. (2017). Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Science*, 57(3), 1404–1420. <https://doi.org/10.2135/cropsci2016.08.0675>
- Gorjanc, G., Gaynor, R. C., & Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, 131, 1953–1966. <https://doi.org/10.1007/s00122-018-3125-3>
- Gorjanc, G., Jenko, J., Hearne, S. J., & Hickey, J. M. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics*, 17(1), 1–15. <https://doi.org/10.1186/s12864-015-2345-z>
- Gower, G., Ragsdale, A. P., Bisschop, G., Gutenkunst, R. N., Hartfield, M., Noskova, E., Schiffels, S., Struck, T. J., Kelleher, J., & Thornton, K. R. (2022). Demes: A standard format for demographic models. *Genetics*, 222, iyac131. <https://doi.org/10.1093/genetics/iyac131>
- Haller, B. C., & Messer, P. W. (2019). Evolutionary modeling in SLiM 3 for beginners. *Molecular Biology and Evolution*, 36(5), 1101–1109. <https://doi.org/10.1093/molbev/msy237>
- Hassanpour, A., Geibel, J., Simianer, H., & Pook, T. (2023). Optimization of breeding program design through stochastic simulation with kernel regression. *G3 Genes/Genomes/Genetics*, 13(12), jkad217. <https://doi.org/10.1093/g3journal/jkad217>
- Huang, M., Robbins, K. R., Li, Y., Umanzor, S., Marty-Rivera, M., Bailey, D., Yarish, C., Lindell, S., & Jannink, J.-L. (2022). Simulation of sugar kelp (*Saccharina latissima*) breeding guided by practices to accelerate genetic gains. *G3*, 12(3), jkac003. <https://doi.org/10.1093/g3journal/jkac003>
- Jannink, J.-L., Astudillo, R., & Frazier, P. (2023). Insight into a two-part plant breeding scheme through Bayesian optimization of budget allocations. *Crop Science*. <https://doi.org/10.1002/csc2.21124>
- Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127, 595–607. <https://doi.org/10.1007/s00122-013-2243-1>
- Jenko, J., Gorjanc, G., Cleveland, M. A., Varshney, R. K., Whitelaw, C. B. A., Woolliams, J. A., & Hickey, J. M. (2015). Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. *Genetics Selection Evolution*, 47(1), 1–14. <https://doi.org/10.1186/s12711-015-0135-3>
- Johnsson, M., Gaynor, R. C., Jenko, J., Gorjanc, G., De Koning, D. J., & Hickey, J. M. (2019). Removal of alleles by genome editing (RAGE) against deleterious load. *Genetics Selection Evolution*, 51(1), 1–18. <https://doi.org/10.1186/s12711-019-0456-8>
- Kingsbury, N. (2011). *Hybrid: The history and science of plant breeding*. University of Chicago Press.
- Krause, M. D., Piepho, H.-P., Dias, K. O., Singh, A. K., & Beavis, W. D. (2023). Models to estimate genetic gain of soybean seed yield from annual multi-environment field trials. *Theoretical and Applied Genetics*, 136, 252. <https://doi.org/10.1007/s00122-023-04470-3>

- Labroo, M. R., Endelman, J. B., Gemenet, D. C., Werner, C. R., Gaynor, R. C., & Covarrubias-Pazarán, G. E. (2023). Clonal diploid and autopolyploid breeding strategies to harness heterosis: Insights from stochastic simulation. *Theoretical and Applied Genetics*, 136(7), 147. <https://doi.org/10.1007/s00122-023-04377-z>
- Lanzl, T., Melchinger, A. E., & Schön, C.-C. (2023). Influence of the mating design on the additive genetic variance in plant breeding populations. *Theoretical and Applied Genetics*, 136(11), 236. <https://doi.org/10.1007/s00122-023-04447-2>
- Lara, L. A. d. C., Pocrnic, I., Oliveira, T. d. P., Gaynor, R. C., & Gorjanc, G. (2021). Temporal and genomic analysis of additive genetic variance in breeding programmes. *Heredity*, 128, 21–32. <https://doi.org/10.1038/s41437-021-00485-y>
- Lauterbur, M. E., Cavassim, M. I. A., Gladstein, A. L., Gower, G., Pope, N. S., Tsambos, G., Adrion, J., Belsare, S., Biddanda, A., Caudill, V., Cury, J., Echevarria, I., Haller, B. C., Hasan, A. R., Huang, X., Iasi, L. N. M., Noskova, E., Obšteter, J., Pavinato, V. A. C., ... Gronau, I. (2023). Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *eLife*, 12, RP84874. <https://doi.org/10.7554/eLife.84874.1>
- Li, X., Zhu, C., Wang, J., & Yu, J. (2012). Computer simulation in plant breeding. *Advances in Agronomy*, 116, 219–264. <https://doi.org/10.1016/B978-0-12-394277-7.00006-3>
- 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(January), 1–15. <https://doi.org/10.3389/fpls.2018.01926>
- Lubanga, N., Massawe, F., Mayes, S., Gorjanc, G., & Bancic, J. (2023). Genomic selection strategies to increase genetic gain in tea breeding programs. *The Plant Genome*, 16, e20282. <https://doi.org/10.1002/tpg2.20282>
- Lush, J. L. (1937). *Animal breeding plans*. Collegiate Press, Incorporated.
- Mancin, E., Mantovani, R., Tuliozi, B., & Sartori, C. (2022). Economic weights for restriction of selection index as optimal strategy for combining multiple traits. *Journal of Dairy Science*, 105(12), 9751–9762. <https://doi.org/10.3168/jds.2022-22085>
- Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), 1819–1829. <https://doi.org/10.1093/genetics/157.4.1819>
- Meyer, R. S., DuVal, A. E., & Jensen, H. R. (2012). Patterns and processes in crop domestication: An historical review and quantitative analysis of 203 global food crops. *New Phytologist*, 196(1), 29–48. <https://doi.org/10.1111/j.1469-8137.2012.04253.x>
- Obšteter, J., Strachan, L. K., Bubnič, J., Prešern, J., & Gorjanc, G. (2023). SIMPLYBee: R package for simulating honeybee populations and breeding programs. *Genetics Selection Evolution*, 55, 31. <https://doi.org/10.1186/s12711-023-00798-y>
- Oliveira, T. P., Obšteter, J., Pocrnic, I., Heslot, N., & Gorjanc, G. (2023). A method for partitioning trends in genetic mean and variance to understand breeding practices. *Genetics Selection Evolution*, 55, 36. <https://doi.org/10.1186/s12711-023-00804-3>
- Peixoto, M. A., Coelho, I. F., Leach, K. A., Bhering, L. L., & Resende Jr., M. F. R. (2024). Simulation-based decision-making and implementation of tools in hybrid crop breeding pipelines. *Crop Science*, 64(1), 110–125. <https://doi.org/10.1002/csc2.21139>
- Platten, J. D., & Fritsche-Neto, R. (2023). Optimizing quantitative trait loci introgression in elite rice germplasms: Comparing methods and population sizes to develop new recipients via stochastic simulations. *Plant Breeding*. <https://doi.org/10.1111/pbr.13118>
- Pocrnic, I., Lindgren, F., Tollhurst, D., Herring, W. O., & Gorjanc, G. (2022). Optimisation of the core subset for the APY approximation of genomic relationships. *Genetics Selection Evolution*, 54, 76. <https://doi.org/10.1186/s12711-022-00767-x>
- Podlich, D. W., & Cooper, M. (1998). QU-GENE: A simulation platform for quantitative analysis of genetic models. *Bioinformatics*, 14(7), 632–653. <https://doi.org/10.1093/bioinformatics/14.7.632>
- Pook, T., Schlather, M., & Simianer, H. (2020). MoBPS—Modular Breeding Program Simulator. *G3 Genes/Genomes/Genetics*, 10(6), 1915–1918. <https://doi.org/10.1534/g3.120.401193>
- Powell, O., Gaynor, R. C., Gorjanc, G., Werner, C. R., & Hickey, J. M. (2020). A two-part strategy using genomic selection in hybrid crop breeding programs. *bioRxiv*. <https://doi.org/10.1101/2020.05.24.113258>
- Powell, O. M., Barbier, F., Voss-Fels, K. P., Beveridge, C., & Cooper, M. (2022). Investigations into the emergent properties of gene-to-phenotype networks across cycles of selection: A case study of shoot branching in plants. *in silico Plants*, 4(1), diac006. <https://doi.org/10.1093/insilicoplants/diac006>
- Ray, D. K., Mueller, N. D., West, P. C., & Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLoS ONE*, 8(6), e66428. <https://doi.org/10.1371/journal.pone.0066428>
- Roser, M. (2023). Crop yields, world, 1961 to 2021. OurWorldIn-Data.org. <https://ourworldindata.org/grapher/key-crop-yields>
- Rutkoski, J. E., Krause, M. R., & Sorrells, M. E. (2022). Breeding methods: Line development. In *Wheat improvement: Food security in a changing climate* (pp. 69–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-90673-3_5
- Sabadin, F., DoVale, J. C., Platten, J. D., & Fritsche-Neto, R. (2022). Optimizing self-pollinated crop breeding employing genomic selection: From schemes to updating training sets. *Frontiers in Plant Science*, 13, 3770. <https://doi.org/10.3389/fpls.2022.935885>
- Salgotra, R. K., & Chauhan, B. S. (2023). Genetic diversity, conservation, and utilization of plant genetic resources. *Genes*, 14(1), 174. <https://doi.org/10.3390/genes14010174>
- Schnell, F. W. (1982). A synoptic study of the methods and categories of plant breeding. *Z. Pflanzenzüchtung*, 89, 1–18.
- Shrote, R. Z., & Thompson, A. M. (2023). PyBrOpS: A Python package for breeding program simulation and optimization for multi-objective breeding. *bioRxiv*. <https://doi.org/10.1101/2023.02.10.528043>
- Simianer, H., Büttgen, L., Ganesan, A., Ha, N. T., & Pook, T. (2021). A unifying concept of animal breeding programmes. *Journal of Animal Breeding and Genetics*, 138(2), 137–150. <https://doi.org/10.1111/jbg.12534>
- Tollhurst, D. J., Gaynor, R. C., Gardunia, B., Hickey, J. M., & Gorjanc, G. (2022). Genomic selection using random regressions on known and latent environmental covariates. *Theoretical and Applied Genetics*, 135(10), 3393–3415. <https://doi.org/10.1007/s00122-022-04186-w>
- Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., & Sorrells, M. E. (2021). Designing future crops: Genomics-assisted breeding comes of age. *Trends in Plant Science*, 26(6), 631–649. <https://doi.org/10.1016/j.tplants.2021.03.010>
- Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M.-D., Asyraf Md Hatta, M., Hinchliffe, A., Steed, A., Reynolds,

- D., & Adamski, N. M. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature Plants*, 4(1), 23–29. <https://doi.org/10.1038/s41477-017-0083-8>
- Werner, C., Garment, D., & Tolhurst, D. (2024). Fieldsimr: An R package for simulating plot data in multi-environment field trials. *Frontiers in Plant Science*, 15, 1330574. <https://doi.org/10.3389/fpls.2024.1330574>
- Werner, C. R., Gaynor, R. C., Sargent, D. J., Lillo, A., Gorjanc, G., & Hickey, J. M. (2023). Genomic selection strategies for clonally propagated crops. *Theoretical and Applied Genetics*, 136, 74. <https://doi.org/10.1007/s00122-023-04300-6>
- Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theoretical and Applied Genetics*, 129, 653–673. <https://doi.org/10.1007/s00122-016-2691-5>
- Yang, C. J., Ladejobi, O., Mott, R., Powell, W., & Mackay, I. (2022). Analysis of historical selection in winter wheat. *Theoretical and Applied Genetics*, 135(9), 3005–3023. <https://doi.org/10.1007/s00122-022-04163-3>
- Yang, C. J., Russell, J., Ramsay, L., Thomas, W., Powell, W., & Mackay, I. (2021). Overcoming barriers to the registration of new plant varieties under the DUS system. *Communications Biology*, 4(1), 302. <https://doi.org/10.1038/s42003-021-01840-9>
- Yin, K., Gao, C., & Qiu, J.-L. (2017). Progress and prospects in plant genome editing. *Nature Plants*, 3(8), 1–6. <https://doi.org/10.1038/nplants.2017.107>
- Younis, O. G., Turchetta, M., Ariza Suarez, D., Yates, S., Studer, B., Athanasiadis, I. N., Krause, A., Buhmann, J. M., & Corinzia, L. (2023). Chromax: A fast and scalable breeding program simulator. *Bioinformatics*, 39(12), btad691. <https://doi.org/10.1093/bioinformatics/btad691>
- Zhang, X., & Cai, X. (2011). Climate change impacts on global agricultural land availability. *Environmental Research Letters*, 6(1), 014014. <https://doi.org/10.1088/1748-9326/6/1/014014>
- Zhao, T., Zeng, J., & Cheng, H. (2022). Extend mixed models to multilayer neural networks for genomic prediction including intermediate omics data. *Genetics*, 221(1), iyac034. <https://doi.org/10.1093/genetics/iyac034>

How to cite this article: Bančič, J., Greenspoon, P., Gaynor, R. C., & Gorjanc, G. (2025). Plant breeding simulations with AlphaSimR. *Crop Science*, 65, e21312. <https://doi.org/10.1002/csc2.21312>

APPENDIX A: INVENTORY OF ALPHASIMR CODE

This appendix contains an inventory of the AlphaSimR scripts we provide for a variety of breeding programs and techniques—available at the GitHub repository https://github.com/HighlanderLab/jbancic_alphasimr_plants. In Tables A.1 and A.2, we indicate the folder and file path for each breeding program and technique.

TABLE A.1 List of breeding programs and methods presented in this paper and their file paths.

Program	Description	Directory path
	Line breeding	01_LineBreeding/
	- Phenotypic selection	01_LineBreeding/01_PhenotypicSelection
1	- Mass selection	01_LineBreeding/01_PhenotypicSelection/01_MassSelection/
2	- Single seed selection	01_LineBreeding/01_PhenotypicSelection/02_SingleSeedDescent/
3	- Pedigree selection	01_LineBreeding/01_PhenotypicSelection/03_PedigreeSelection/
4	- Doubled-haploid	01_LineBreeding/01_PhenotypicSelection/04_DoubledHaploid/
5	- Genomic selection	01_LineBreeding/02_GenomicSelection/
6	- Two-part genomic selection	01_LineBreeding/03_TwoPartGS/
	Clonal breeding	02_ClonalBreeding/
7	- Phenotypic selection	02_ClonalBreeding/01_PhenotypicSelection/
8	- Pedigree selection	02_ClonalBreeding/02_PedigreeSelection/
9	- Genomic selection	02_ClonalBreeding/03_GenomicSelection/
	Hybrid breeding	03_HybridBreeding/
10	- Phenotypic selection	03_HybridBreeding/01_PhenotypicSelection/
11	- Genomic selection	03_HybridBreeding/02_GenomicSelection/
12	- Two-part genomic selection	03_HybridBreeding/03_TwoPartGS/

TABLE A.2 List of breeding techniques presented in this paper and their file paths.

Feature	Description	File path
1	Mating plans	04_Features/matingPlans.R
2	GWAS	04_Features/simulateGWAS.R
3	Set heritability	04_Features/setHeritability.R
4	Multiple traits	04_Features/multipleTraits.R
5	Miscellaneous slot	04_Features/miscellaneousSlot.R
6	G×E interaction	04_Features/simulateGxE.R
7	Genomic models	04_Features/genomicModels.R
8	Trait introgression	04_Features/traitIntrogression.R
9	Genome editing	04_Features/genomeEditing.R
10	Speed breeding	04_Features/speedBreeding.R
11	Import external haplotypes	04_Features/importExternalHaplo.R
12	Specify demography	04_Features/specifyDemography.R

Abbreviation: GWAS, genome-wide association study.