AlphaSim: Software for Breeding Program Simulation

Citation for published version:
Faux, A, Gorjanc, G, Gaynor, RC, Battagin, M, Hoj-Edwards, S, Wilson, DL, Hearne, SJ, Gonen, S & Hickey, JM 2016, 'AlphaSim: Software for Breeding Program Simulation', *The Plant Genome*, vol. 9, no. 3. https://doi.org/10.3835/plantgenome2016.02.0013

Digital Object Identifier (DOI):
10.3835/plantgenome2016.02.0013

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
The Plant Genome

Published September 22, 2016

ORIGINAL RESEARCH

AlphaSim: Software for Breeding Program Simulation

Anne-Michelle Faux, Gregor Gorjanc, R. Chris Gaynor, Mara Battagin, Stefan M. Edwards, David L. Wilson, Sarah J. Hearne, Serap Gonen, and John M. Hickey*

Abstract

This paper describes AlphaSim, a software package for simulating plant and animal breeding programs. AlphaSim enables the simulation of multiple aspects of breeding programs with a high degree of flexibility. AlphaSim simulates breeding programs in a series of steps: (i) simulate haplotype sequences and pedigree; (ii) drop haplotypes into the base generation of the pedigree and select single-nucleotide polymorphism (SNP) and quantitative trait nucleotide (QTN); (iii) assign QTN effects, calculate genetic values, and simulate phenotypes; (iv) drop haplotypes into the burn-in generations; and (v) perform selection and simulate new generations. The program is flexible in terms of historical population structure and diversity, recent pedigree structure, trait architecture, and selection strategy. It integrates biotechnologies such as doubled-haploids (DHs) and gene editing and allows the user to simulate multiple traits and multiple environments, specify recombination hot spots and cold spots, specify gene jungles and deserts, perform genomic predictions, and apply optimal contribution selection. AlphaSim also includes restart functionalities, which increase its flexibility by allowing the simulation process to be paused so that the parameters can be changed or to import an externally created pedigree, trial design, or results of an analysis of previously simulated data. By combining the options, a user can simulate simple or complex breeding programs with several generations, variable population structures and variable breeding decisions over time. In conclusion, AlphaSim is a flexible and computationally efficient software package to simulate biotechnology enhanced breeding programs with the aim of performing rapid, low-cost, and objective in silico comparison of breeding technologies.

Published in Plant Genome
Volume 9. doi: 10.3835/plantgenome2016.02.0013

Core Ideas

- AlphaSim allows breeders and researchers to simulate genomic data with specific user criteria.
- AlphaSim is flexible, computationally efficient, and easy to use for a wide range of possible scenarios.
- AlphaSim can also be used in animal breeding, human genetics, and population genetics.

This paper introduces AlphaSim, a software package for simulating breeding programs. AlphaSim combines features from three previous simulation packages, AlphaDrop (Hickey and Gorjanc, 2012), AlphaSimPlant, and AlphaMPSim (Hickey et al., 2014), with new features to form a comprehensive software package capable of simulating a wide range of mating designs, biotechnologies, and selection strategies. This allows a user to perform plant or animal breeding simulations in any species using a wide range of strategies. AlphaSim offers the user a high degree of simulation flexibility making it a useful tool for designing and optimizing new breeding strategies using newly developed technologies.

Simulation has been an effective platform for the evaluation and development of new breeding strategies.

A.M. Faux, G. Gorjanc, R.C. Gaynor, M. Battagin, S.M. Edwards, D.L. Wilson, S. Gonen, and J.M. Hickey, The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian, Scotland, United Kingdom; S.J. Hearne, CIMMYT, Texcoco, Mexico. Received 9 Feb. 2016. Accepted 6 July 2016. *Corresponding author (john.hickey@roslin.ed.ac.uk). Assigned to Associate Editor Jianming Yu.

Abbreviations: DH, doubled-haploid; G × E, genotype × environment; gEBV, genomic-estimated breeding value; MAGIC, multiparent advanced-generation intercross; pEBV, pedigree-estimated breeding value; QTN, quantitative trait nucleotide; RIL, recombinant inbred line; SNP, single-nucleotide polymorphism; TBV, true breeding value; TGV, true genotypic value.

Large-scale field-testing of breeding strategies is either impractical or impossible because of the time and resources needed; simulation offers a comparatively quick and inexpensive alternative. Many software packages for plant breeding simulations are currently available (Sun et al., 2011). These packages have been useful for evaluating existing breeding strategies in actual field-based breeding programs (Wang et al., 2003) and have been used to develop new breeding strategies. For example, the PLABSIM software package aided in the development of an efficient marker-assisted backcross design used to transfer the stripe rust (*Puccinia striiformis* f. sp. *tritici*) resistance gene *Yr15* to the spring wheat (*Triticum aestivum* L.) cultivar Zak (Randhawa et al., 2009). The historical use of simulation and the expanding range of technological options for breeding programs indicates that simulation will continue to play a role and indeed play an increasingly relevant role in future research focusing on design and optimization of plant and animal breeding programs.

New breeding strategies are required to efficiently optimize the implementation of new technologies in breeding programs. Genomic selection (Meuwissen et al., 2001; Bernardo and Yu, 2007) and genome editing (Shan et al., 2014; Jenko et al., 2015) are two such technologies. Genomic selection in particular has been widely promoted as a technology of great value to plant breeding (Bernardo and Yu, 2007; Heffner et al., 2009; Jannink et al., 2010). While it has a large potential to improve plant breeding, implementation of genomic selection requires optimization to maximize return on investment. Simulation is the ideal tool to develop optimal breeding strategies while assessing costs and benefits (e.g., Hickey et al., 2014; Gorjanc et al., 2016). However, to our knowledge, existing software packages lack the ability to simulate breeding programs with genomic selection with sufficient flexibility and computational efficiency.

We designed AlphaSim to fill the need for a software package that is capable of simulating new breeding designs and application of biotechnologies, such as genomic selection and gene editing, in a flexible and computationally efficient manner. This paper describes the simulation method and operation of AlphaSim in varied plant breeding applications with an emphasis on its main features (Fig. 1) and computational performance. Examples of how to use the software are included with measures of computational efficiency. Table 1 gives a list of symbols used throughout this paper.

Materials and Methods

Method

AlphaSim simulates breeding programs in five main steps (Fig. 2):

1. Simulate haplotype sequences and pedigree.
2. Drop haplotypes into the base generation and select SNP and QTN.
3. Assign QTN effects, calculate genetic values, and simulate phenotypes.
4. Drop haplotypes into the burn-in generations.
5. Perform selection and simulate new generations.

For each generation, AlphaSim writes information about the haplotype sequences, SNP and QTN genotypes, and breeding values in output files, which canww be used for further analysis or for running alternative scenarios. This also helps to keep the memory requirements of AlphaSim low. The following five subsections provide details of each step of the method considering the simulation of a single trait. The remainder of this section describes additional features, output, and the data storage system of the software.

First Step: Simulate Haplotype Sequences and Pedigree (Fig. 2, Step 1)

By default, haplotype sequences are simulated through a system call to program MaCS. MaCS is a coalescent simulation program that simulates, for each chromosome successively, a sample of haplotype sequences according to specified ancestral population with, at a minimum, a specified chromosome size, mutation rate, recombination rate, and effective population size. Alternatively, the user can also generate their own haplotype sequences externally and import them into AlphaSim. This external source can be either real sequences or sequences simulated using other methods.

Second Step: Drop Haplotypes into the Base Generation of the Pedigree and Select Single-Nucleotide Polymorphism and Quantitative Trait Nucleotide (Fig. 2, Step 2)

AlphaSim samples haplotypes with replacement from the base set of haplotype sequences and drops them into the first generation of the pedigree. Dropping of haplotypes involves recombination events, which are randomly distributed across the genome ignoring interference. Should the user prefer nonrandom distribution of recombination events, a file can be supplied that specifies the proportion of recombination events in specific regions of the genome, that is, recombination hot spots and cold spots.

After the haplotypes are dropped into the base generation, AlphaSim samples segregating sites to become either SNP markers or QTNs. The SNP markers constitute distinct SNP panels, and the user has control over the number of panels, their density, the minimum and maximum allele frequency of SNP, whether the panels are nested within each other or not, whether these panels include QTN or not, and which panel will be used in selection. The user can also control whether the full sequence and phased data are provided as output.

AlphaSim samples two sets of segregating sites to become biallelic QTN. Both sets include the same user-specified number of QTN, denoted as n_{QTN}. The first set, referred to as unrestricted, is comprised of QTN selected at

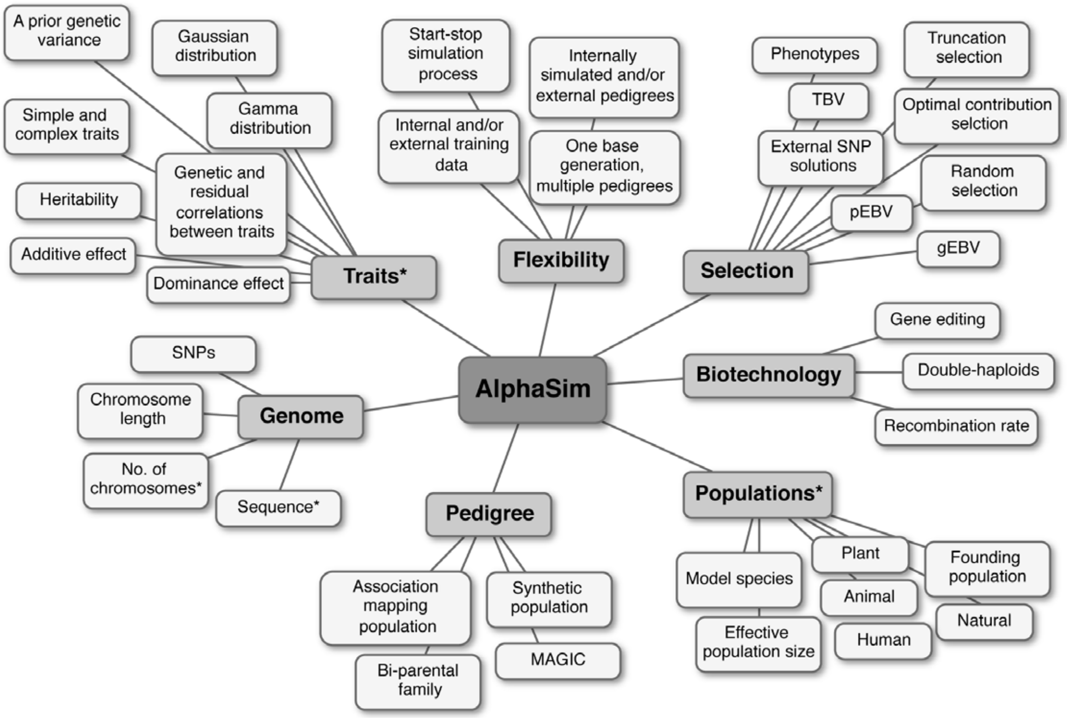


Fig. 1. Some of the AlphaSim parameters that can be specified by the user, of which most can be changed during the course of a simulation. Asterisk denotes parameters that are immutable.

Table 1. List of symbols.

Symbol	Definition†
$\mathbf{a}$	Vector of breeding values
$\hat{\mathbf{a}}$	Vector of estimated breeding values
$\mathbf{A}$	Pedigree or genomic numerator relationship matrix
a_k	Additive effect at QTN k simulated for a given trait
α_k	Average allele substitution effect at QTN k computed for a given trait
$\mathbf{b}$	Vector of SNP effects
$\hat{b}_j$	Estimated effect of SNP j
d_k	Dominance effect at QTN k computed for a given trait
δ_k	Dominance degree at QTN k simulated for a given trait
$\mathbf{e}$	Vector of residual effects
$e_{i,r}$	Residual effect for individual i and trait r
gEBV$_i$	Genomic estimated breeding value of individual i
H^2	Broad-sense heritability
h^2	Narrow-sense heritability
i	Indicates a given individual, varies from 1 to n_{Indiv}
j	Indicates a given SNP, varies from 1 to n_{SNP}
k	Indicates a given QTN, varies from 1 to n_{QTN}
$\mathbf{L}_A$, $\mathbf{L}_E$	Lower triangular matrix obtained from the Cholesky decomposition of $\mathbf{V}_A$ and $\mathbf{V}_E$
$L_{A_{r,s}}$, $L_{E_{r,s}}$	Entry of $\mathbf{L}_A$ and $\mathbf{L}_E$ between traits r and s
λ	Penalty factor applied on the loss of genetic diversity in optimal contribution selection
m_δ	User-specified mean dominance degree
μ	Intercept of the regression models (Eq. [10] and [12])
μ_0	Mean value of the base generation of the pedigree

(cont'd.)

random from across the genome. The second set, referred to as restricted, is comprised of QTN selected at random from across the genome with the restriction that the minor allele frequency must be in a specified range. The restrictions in allele frequency of both SNP markers and QTN allow the user to manage the possibility that QTN have different allele frequencies than SNP. Should the user prefer nonrandom distribution of QTN, a file can be supplied that specifies the proportions of QTN in specific regions of the genome, that is, gene jungles and deserts.

Third Step: Assign Quantitative Trait Nucleotide Effects, Calculate Genetic Values and Simulate Phenotypes (Fig. 2, Step 3)

AlphaSim assigns coded genetic values for additive and dominance effects to the frequency-restricted and unrestricted sets of QTN independently (i.e., a, d, and $-a$; Bernardo, 2010). For additive genetic values, let k indicate a QTN from one of the QTN sets. For each QTN k, AlphaSim randomly samples deviates from a standard Gaussian distribution, defined as a Gaussian distribution

Table 1. Continued.

Symbol	Definition†
n_{Indiv}	No. of individuals
n_{QTN}	Total no. of unrestricted or frequency-restricted QTN in the genome
n_{SNP}	Total no. of SNP in the genome
n_{Traits}	No. of simulated traits
pEBV_i	Pedigree-estimated breeding value of individual *i*
p_k, q_k	Frequencies of the nonzero and zero alleles, respectively, at QTN *k* in the base generation of the pedigree
r, *s*	Indicate two distinct traits: *r* varies from 1 to n_{Traits}, and *s* varies from 1 to *r*
RandDev	Random deviate sampled from a Gaussian or Gamma distribution
σ^2_{a}	A priori additive genetic variance specified by the user for a given trait
$\sigma^2_{\mathrm{a}_0}$	Additive genetic variance computed for a given trait in the base generation
$\sigma^2_{\mathrm{a}_{\mathrm{tr}}}$	Additive genetic variance computed in the training population using the TBV of the training individuals.
σ^2_{b}	Variance of the SNP effects
$\sigma^2_{\mathrm{d}_0}$	Dominance genetic variance computed for a given trait in the base generation
σ^2_{δ}	User-specified variance of the dominance degrees
σ^2_{e}	Residual variance computed for a given trait
$\sigma^2_{\mathrm{g}_0}$	Genotypic variance computed for a given trait in the base generation
TBV_i, TDV_i, TGB_i	True breeding value, true dominance value, and true genotypic value of individual *i* for a trait characterized by a given set of QTN, unrestricted or frequency-restricted, and a given distribution, Gaussian or Gamma
$\mathbf{V_A}$, $\mathbf{V_E}$	Additive genetic and residual correlation matrix, respectively, dimensions $n_{\mathrm{Traits}}n_{\mathrm{Traits}}$
$\mathbf{x}$	Vector providing the contribution of each selection candidate to the next generation
$\mathbf{X}$	Incidence matrix linking phenotypes to **b**
$x_{i,k}$, $x_{i,j}$	Genotype of individual *i* at QTN *k* or SNP *j*, coded as 0, 1, or 2 according to the number of copies of the nonzero allele
$\mathbf{y}$	Vector of phenotype records
$\mathbf{Z}$	Incidence matrix linking phenotypes to **a**

† QTN, quantitative trait nucleotide; SNP, single-nucleotide polymorphism.

with mean zero and unit variance or a Gamma distribution with user-defined shape and scale parameters. If a Gamma distribution is used, the deviates are scaled to unit variance by dividing by the expected standard deviation of the distribution, which is the square root of the product of the shape parameter and the square of the scale parameter. These scaled Gamma deviates are then randomly assigned to have either a positive or negative effect. The final additive genetic value for each QTN locus is obtained as follows:

$$a_k = \mathrm{RandDev}\sqrt{\frac{\sigma^2_{\mathrm{a}}}{n_{\mathrm{QTN}}}} \qquad [1]$$

where RandDev is a random deviate from either the Gaussian distribution or the scaled Gamma distribution, and σ^2_a is the a priori additive genetic variance specified by the user.

Dominance genetic values are assigned to QTN following the methods of Wellmann and Bennewitz (2011, 2012). First, dominance degrees, δ_k, are sampled from a Gaussian distribution with mean m_δ and variance σ^2_δ specified by the user. The user can specify no dominance by setting both m_δ and σ^2_δ to zero. Values for δ_k are obtained as follows:

$$\delta_k = m_\delta + \mathrm{RandDev}\sqrt{\sigma^2_\delta} \qquad [2]$$

where RandDev is a random deviate from a standard Gaussian distribution. The coded genetic values for dominance at each locus is then calculated as follows:

$$d_k = \delta_k\left|a_k\right| \qquad [3]$$

The coded genetic values are used to calculate true genotypic values (TGVs) for individuals and genetic variances in the base generation of the pedigree. For each individual, TGV is calculated by summing its coded genetic value for its genotype across all QTN loci. An initial mean μ_0 is subtracted from that sum to set mean TGV in the base generation to zero:

$$\mathrm{TGV}_i = \sum\nolimits_{k=1}^{n_{\mathrm{QTN}}}\left(-a_k, d_k, a_k\right)\left[x_{i,k}=0, x_{i,k}=1, x_{i,k}=2\right]-\mu_0 \quad [4]$$

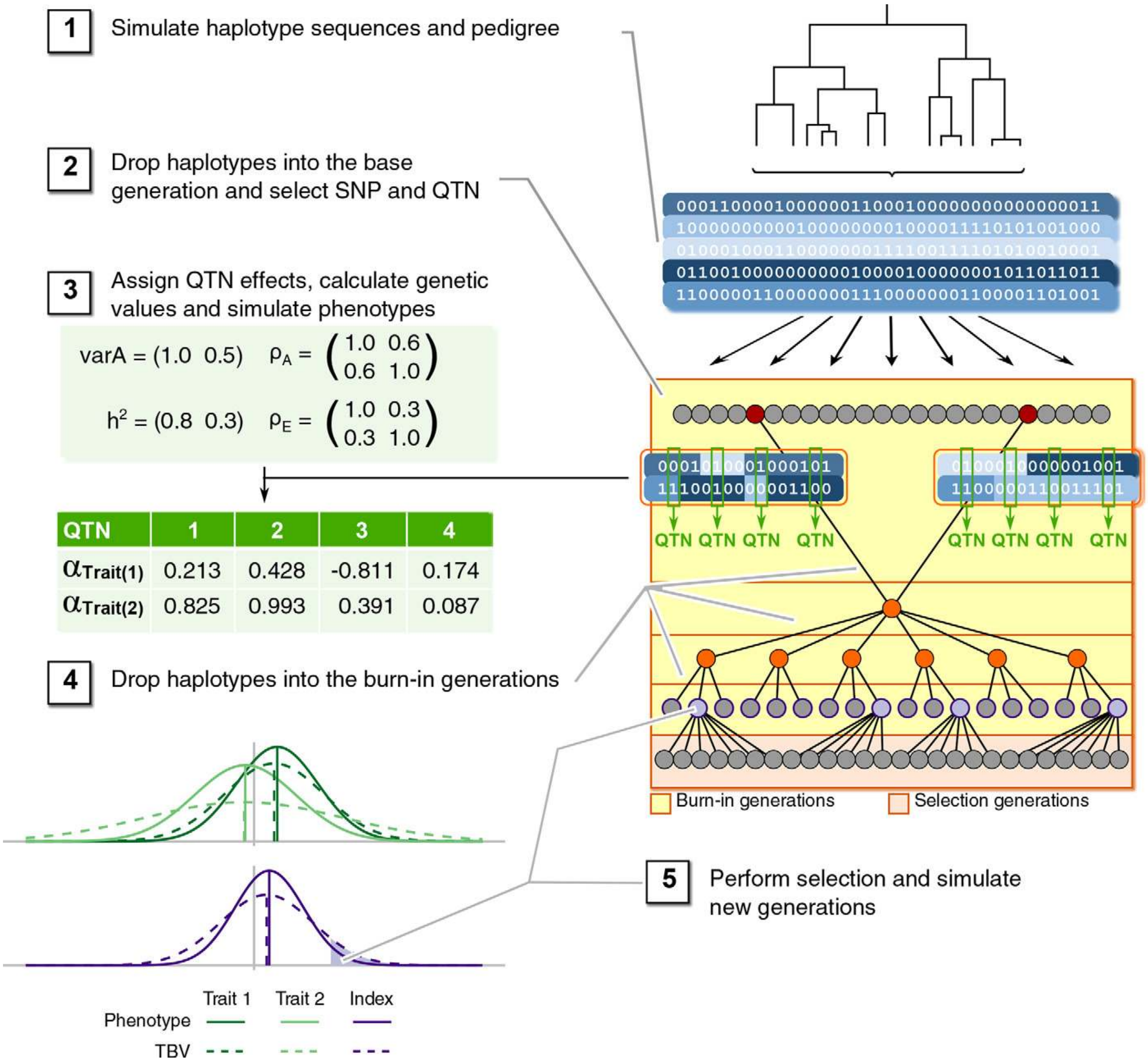
The total genotypic variance in the base generation $\sigma^2_{\mathrm{g}_0}$ is calculated by taking the variance of TGV in the base generation.

Calculating the additive and dominance components of genetic variance requires first calculating the average allele substitution effect α_k for each QTN locus (Bernardo 2010):

$$\alpha_k = a_k + d_k\left(q_k - p_k\right) \qquad [5]$$

where p_k and q_k are the frequencies of the nonzero and zero alleles in the base generation. The average allele substitution effects are then used to calculate true breeding values (TBVs) for each individual in the base generation by summing breeding values at each locus (Bernardo, 2010):

$$\mathrm{TBV}_i = \sum\nolimits_{k=1}^{n_{\mathrm{QTN}}}\left[-2p_k\alpha_k, \left(q_k - p_k\right)\alpha_k, 2q_k\alpha_k\right] \times\left[x_{i,k}=0, x_{i,k}=1, x_{i,k}=2\right] \qquad [6]$$



QTN	1	2	3	4
$\alpha_{\text{Trait(1)}}$	0.213	0.428	-0.811	0.174
$\alpha_{\text{Trait(2)}}$	0.825	0.993	0.391	0.087

Fig. 2. Principle of an AlphaSim simulation illustrated using a pedigree structured in four burn-in generations and one selection generation for two traits characterized by an additive genetic model. (1) Haplotype sequences and an internal pedigree are simulated. (2) Haplotypes are recombined and dropped into the base generation of the pedigree. At this step, single-nucleotide polymorphisms (SNPs) and quantitative trait nucleotides (QTNs) are selected. (3) An effect is assigned to each QTN, and, for each individual of the base generation of the pedigree, genetic values are calculated and phenotypes simulated. (4) Haplotypes of the base generation are recombined and dropped into the burn-in generations of the pedigree successively. Similar to the base generation, genetic values are calculated and phenotypes simulated for each individual of the burn-in generations. (5) A selection generation is simulated according to the selection method and strategy as defined by the user.

The average allele substitution effects are also used to calculate true dominance values (TDV) for each individual in the base generation by summing dominance deviations at each locus (Bernardo 2010):

$$\mathrm{TDV}_i = \sum_{k=1}^{n_{\mathrm{QTN}}} \left(-2p_k^2 d_k, 2p_k q_k d_k, -2q_k^2 d_k\right) \times [x_{i,k} = 0, x_{i,k} = 1, x_{i,k} = 2] \quad [7]$$

The additive genetic variance in the base generation $\sigma^2_{\mathrm{a}_0}$ is calculated by taking the variance of the TBV, and the dominance genetic variance in the base generation $\sigma^2_{\mathrm{d}_0}$ is calculated by taking the variance of the TDV (Bernardo, 2010). Since these calculations, except TGV, depend on allele frequencies, AlphaSim recalculates each in subsequent generations using the generation specific allele frequencies. This means that only TGV and not TBV and TDV should be compared across generations.

Phenotypes are simulated by adding a random residual deviate to each individual's TGV. The residual deviates are sampled from a Gaussian distribution with mean zero and a variance equal to the residual variance σ^2_{e}. The residual variance is calculated so as to obtain the user defined value for trait heritability. The user defines the heritability as either broad-sense heritability, H^2, or narrow-sense heritability, h^2. If the user defines broad-sense heritability, the calculation for residual variance is as follows:

$$\sigma^2_e = \frac{\sigma^2_{\mathrm{g}_0}}{H^2} - \sigma^2_{\mathrm{g}_0} \qquad [8]$$

If narrow-sense heritability is defined, the calculation for residual variance is the following:

$$\sigma^2_{\mathrm{e}} = \frac{\sigma^2_{\mathrm{a}_0}}{h^2} - \sigma^2_{\mathrm{g}_0} \qquad [9]$$

Fourth Step: Drop Haplotypes into the Burn-In Generations (Fig. 2, Step 4)

AlphaSim distinguishes burn-in and selection generations. If the pedigree is internally simulated, burn-in generations are generated by mating randomly selected parents. Internally simulated selection generations are generated by mating parents selected via different selection methods. The generation size and number of parents in each generation can be constant or variable. AlphaSim allows for three distinct types of matings regarding the sex of parents: (i) crosses between male and female individuals, (ii) crosses between bisexual individuals used as male and female parents interchangeably while preventing selfing, and (iii) selfing. Note that an external pedigree can also be imported for any selection generation and combined with the internally simulated pedigrees so that almost any pedigree structure can be defined. If an external pedigree is provided, the user has the option to run the breeding program using only the individuals in the external pedigree or to extend the external pedigree with simulated generations. Extension of the supplied pedigree or internal simulation of the pedigree from the first generation both require the user to specify the number of generations, the size of each generation, the number of parents for each generation, and the mating design to be used in each generation.

Fifth Step: Perform Selection and Simulate New Generations (Fig. 2, Step 5)

Selection in AlphaSim proceeds by selecting individuals in a given generation to become parents of the next generation. Truncation selection is used by default, that is, the best-performing individuals are selected. The number of individuals to be selected can be made constant or variable across generations, and selection can be performed with or without considering gender. AlphaSim enables the selection of individuals based on their TGV, TBV, genomic-estimated breeding values (gEBVs), pedigree-estimated breeding values (pEBVs), or phenotypes; all are obtained using the set of pedigree, SNP, and QTN that characterize the trait under selection as specified by the user.

For computation of both gEBV and pEBV, a training population and a test population are defined. The training population is used to estimate the model parameters, while the test population is used to quantify the accuracy of selection. The test population includes the individuals that will become parents of the next generation. There are several options for constructing the training population: (i) include all individuals in all generations up to the current generation, (ii) include all individuals in all generations up to and including the current generation, (iii) include all individuals in the previous generation only, (iv) as in (ii) but using information from males only, (v) random sampling of a given number of individuals from a range of generations, or (vi) user-specified set of individuals. For each of these options, AlphaSim allows the use of the same training set across different user-specified selection generations. This latter possibility can be used in combination with an externally defined training population to simulate complex selection processes.

To compute gEBV, the phenotypes of the training individuals are regressed onto SNP genotypes in a ridge regression model (Hoerl and Kennard, 1976; Whittaker et al., 2000; Meuwissen et al., 2001):

$$\mathbf{y} = \mu + \mathbf{Xb} + \mathbf{e} \qquad [10]$$

where $\mathbf{y}$ is a vector of phenotype records, μ is the intercept, $\mathbf{b} \sim N\left(0, \mathbf{I}\sigma^2_{\mathrm{b}}\right)$ is a vector of allele substitution effects, $\mathbf{X}$ is the incidence matrix linking phenotypes to $\mathbf{b}$, $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma^2_{\mathrm{e}}\right)$ is a vector of residuals, and σ^2_{e} and σ^2_{b} are respectively variances of residuals and SNP allele substitution effects. The ridge regression is solved through a call to the program AlphaBayes with the variance components set to the simulated values; σ^2_{e} is residual variance and $\sigma^2_{\mathrm{b}} = \frac{\sigma^2_{\mathrm{a}_{\mathrm{tr}}}}{n_{\mathrm{SNP}}}$ where $\sigma^2_{\mathrm{a}_{\mathrm{tr}}}$ is the additive genetic variance in the training population computed using the TBV of the training individuals with training population allele frequencies. Finally, AlphaSim computes gEBV for the selection candidates using their SNP genotypes and the estimated SNP effects $\hat{b}_j$:

$$\mathrm{gEBV}_i = \sum\nolimits_{j=1}^{n_{\mathrm{SNP}}} x_{i,j}\hat{b}_j \qquad [11]$$

Computation of pEBV is based on regressing the phenotypes onto pedigree using the standard mixed model (Henderson, 1984):

$$\mathbf{y} = \mu + \mathbf{Za} + \mathbf{e} \qquad [12]$$

where $\mathbf{y}$ is a vector of phenotype records, μ is the intercept, $\mathbf{a} \sim N\left(0, \mathbf{A}\sigma^2_{\mathrm{a}_{\mathrm{tr}}}\right)$ is a vector of breeding values with $\mathbf{A}$ as the pedigree numerator relationship matrix calculated from an optional number of ancestral generations, $\mathbf{Z}$ is the incidence matrix linking phenotypes to $\mathbf{a}$, and $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma^2_{\mathrm{e}}\right)$ is a vector of residuals. The pedigree regression is solved through a call to the program AlphaBayes with the variance components set to the simulated values.

Additional Features

Multiple Traits

AlphaSim allows the simulation of multiple traits, multiple environments, and genotype × environment (G × E) interactions with the restriction that correlated traits must all be characterized by the same set of QTN (although not all of the QTN in this set need to affect both traits). All the traits are simulated for each individual. When simulating different traits, the user specifies the a priori additive genetic variance and the heritability of each of the traits and the genetic and residual correlations between traits (Fig. 2, Step 3). The a priori additive genetic variances and the additive genetic correlation matrix are used to derive the additive genetic covariance matrix, $\mathbf{V}_{\mathbf{A}}$. After Cholesky decomposition, $\mathbf{V}_{\mathbf{A}} = \mathbf{L}_{\mathbf{A}}\mathbf{L}^{\mathrm{T}}_{\mathbf{A}}$, the Cholesky factor $\mathbf{L}_{\mathbf{A}}$ is used to compute additive genetic effects for each QTN k and trait r:

$$a_{k,r} = \sum\nolimits_{r=1}^{n_{\mathrm{Traits}}} \sum\nolimits_{s=1}^{r} \mathrm{RandDev} \frac{L_{\mathrm{A}_{r,s}}}{\sqrt{n_{\mathrm{QTN}}}} \qquad [13]$$

where r and s are trait indicators, n_{Traits} is the number of traits, and RandDev is a random deviate sampled as in Eq. [1]. The dominance genetic effects of QTN are assumed independent and simulated as in Eq. [2, 3]. The correlated additive genetic effects (Eq. [13]) are further used to compute TGV and allele substitution effects as detailed for the simulation of a single trait (Eq. [4, 5]).

The correlated residual effects are generated for each trait in a process that is similar to that of the correlated additive genetic effects. The residual variances are computed given the specified trait heritabilities and the genetic variances in the base generation of the pedigree. The residual variances are then used jointly with the specified residual correlation matrix to derive the residual covariance matrix $\mathbf{V}_{\mathbf{E}}$. After Cholesky decomposition, $\mathbf{V}_{\mathbf{E}} = \mathbf{L}_{\mathbf{E}}\mathbf{L}^{\mathrm{T}}_{\mathbf{E}}$, the Cholesky factor $\mathbf{L}_{\mathbf{E}}$ is used to compute residual effects $e_{i,r}$ for each individual I and trait r:

$$e_{i,r} = \sum\nolimits_{r=1}^{n_{\mathrm{Traits}}} \sum\nolimits_{s=1}^{r} \mathrm{RandDev} L_{\mathrm{E}_{r,s}} \qquad [14]$$

where RandDev is a random deviate sampled from a standard Gaussian distribution.

When simulating multiple traits and performing selection, a selection index is used to rank individuals. The selection index weights the values of each of the traits as specified by the user. The input values for the selection index can be TBV, gEBV, pEBV, or phenotypes.

Doubled-Haploids

AlphaSim allows the use of DHs. Doubling can be achieved for all individuals included in any given generation as specified by the user. Operationally, AlphaSim simulates DHs by first generating a recombined gamete from the two haplotypes of an individual and then doubling this gamete to produce a diploid individual with identical haplotypes.

Genome Editing

Genome editing is a new technology that has great potential for empowering breeding programs. In recent years, several applications of genome editing have been demonstrated in plant breeding. For example, heritable resistance to powdery mildew has been conferred to bread wheat by simultaneously editing three homeologs (Wang et al., 2014). In maize (*Zea mays* L.), editing technologies were used to modify endogenous loci and add an herbicide tolerance gene at a targeted locus (Shukla et al., 2009). To evaluate the potential of genome editing in breeding programs, genome editing functionality has been added to AlphaSim and demonstrated in an animal breeding application by Jenko et al. (2015). This functionality gives the user the capacity to determine the number of individuals to be edited if these are the top or bottom ranked individuals among the selected and the number of QTN to be edited for each individual. The QTN to be edited are selected in descending order of magnitude of their effect, that is, the QTN with large effect in absolute value are preferentially edited. AlphaSim then performs gene editing such that each edited individual bears the favorable allele in a homozygous state at the edited QTN.

Breeding by Optimal Contribution Selection

In addition to truncation selection, AlphaSim can perform optimal contribution selection, which seeks to find the balance between maximizing the response to selection and minimizing the loss of genetic variance and thereby increases the opportunity for greater response to selection in the long term (Wray and Goddard, 1994; Meuwissen, 1997).

Broadly, optimal contribution selection works by optimizing the contributions of each individual to the next generation by maximizing the genetic mean of the selected individuals while minimizing the genetic relatedness between them, that is, minimizing inbreeding of the next generation. In AlphaSim, optimal contribution selection is performed by calling the program AlphaMate, which maximizes the following objective:

$$\mathbf{x}^{\mathrm{T}}\hat{\mathbf{a}} - \lambda \frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{2} \qquad [15]$$

where $\mathbf{x}$ is a vector of contributions of each individual to the next generation, $\hat{\mathbf{a}}$ is the vector of estimated breeding values of selection candidates, $\mathbf{x}^{\mathrm{T}}\hat{\mathbf{a}}$ is the mean genetic merit passed to the next generation, λ is an unknown penalty on the loss of genetic diversity, **A** is the pedigree or genomic numerator relationship matrix between the selection candidates, and $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}/2$ is an average expected inbreeding in progeny (Wray and Goddard, 1994; Meuwissen, 1997). AlphaMate searches for the value of penalty that gives the user a specified allowed increase in rate of inbreeding, and given that value, solves Eq. [15] for the vector of contributions **x**.

Flexibility

AlphaSim includes three restart functionalities, which make it more flexible than the packages from which it is derived. The first restart functionality enables a simulation process to be stopped after a user-specified generation and to be resumed with some program parameters changed. For example, truncation selection could be used for a number of generations, and then the simulation is stopped and resumed with optimal contribution selection activated or using an alternative genomic selection training population or SNP panel. This feature also enables the simulation of a base population from which different scenarios can be derived or the combination of external and internally simulated pedigrees for both burn-in and selection generations.

The second restart functionality makes AlphaSim flexible in terms of the method used to perform selection. Selection methods or statistical methods that are not implemented in AlphaSim (e.g., marker-assisted selection) can be applied using third-party software to analyze simulated data, select individuals, and mate them, and the externally created pedigree can then be imported into AlphaSim. This functionality thus allows the use of any user-defined pedigree structure in one or more selection generations. For this purpose, AlphaSim provides the user with information about the genotypes, phenotypes, TGV, and TBV of both selection candidates and training individuals as well as the gEBV or pEBV of the selection candidates obtained through a call to the program AlphaBayes.

The third restart functionality enables output from different AlphaSim runs to be merged into a single run. This enables further flexibility and parallel processing. The merge functionality can be used in two ways. The first is to merge information across a range of AlphaSim runs, that is, by run directory merge. The second way is to merge information from specified sets of individuals from a range of runs, that is, by individual merge. This means that the user has the choice of independently performing selection at the end of each run and then combining specific individuals to form a new merged population.

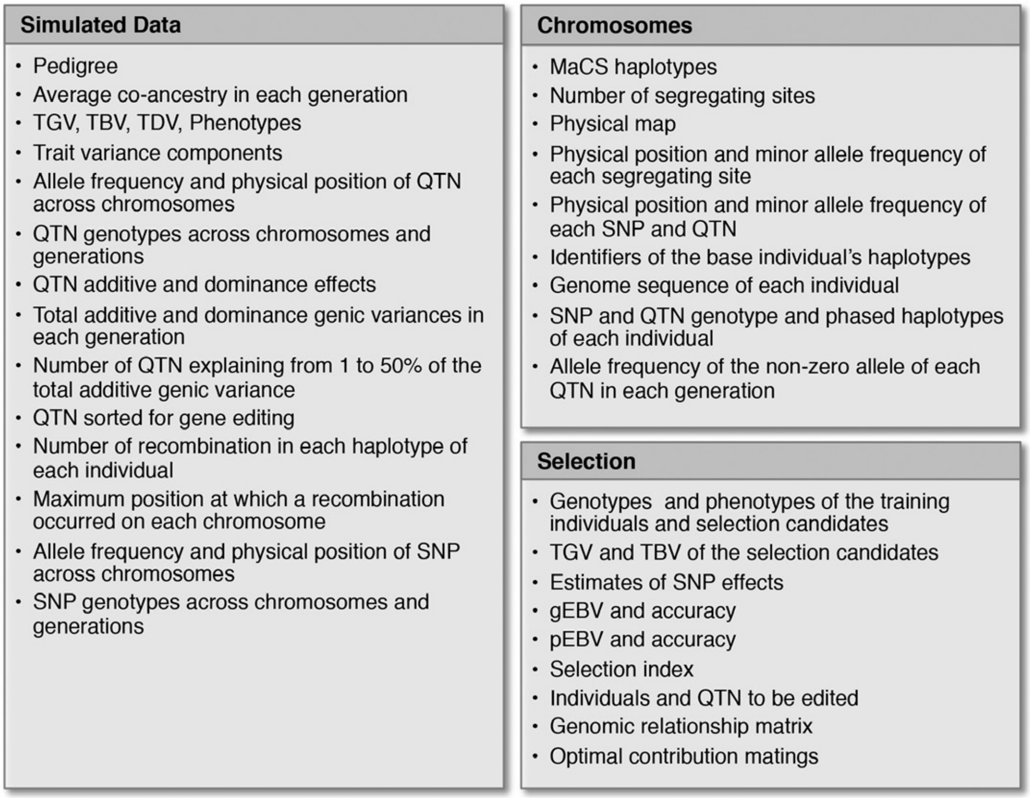
For example, one AlphaSim run can be performed to generate a base population. From this base population, 100 AlphaSim runs can be spawned in which each run would generate a biparental family from two inbred parents. Because these runs all spawn from the same base population, the genetic architecture of traits and other parameters of the founding population is shared between the biparental families. Once these 100 runs are finished, another run of AlphaSim can be performed in which a subset of the individuals from each of the biparental families can be selected and merged into a single population, forming a selected set of lines that can serve as parents of a new set of biparental families. This process of splitting and merging can be repeated several times in many different ways.

Output and Data Storage

The output files of AlphaSim are organized in three directories: Chromosomes, Selection, and SimulatedData (Fig. 3). The Chromosomes directory stores detailed information about the segregating sites, SNP panels, and QTN as well as the phased haplotypes and genotypes of the simulated individuals for each chromosome and for each generation. The Selection directory stores the information required to perform selection for each selection cycle: the TGV of the selection candidates when selection is based on TGV, the TBV when selection is based on TBV, their phenotypes when selection is based on phenotypes, and the SNP genotypes of both the training individuals and selection candidates and the phenotypes of the training individuals when selection is based on gEBV. It also stores the input and output files of AlphaBayes when selection is based on estimated breeding values and the input and output files of AlphaMate when optimal contribution selection is used. The SimulatedData directory stores results of the simulation process. This directory includes the pedigree; the gender of each individual; the simulated TGV, TBV, and phenotypes; the allele frequency and physical position of each SNP and QTN; the simulated QTN effects; the SNP and QTN genotypes; and the trait variance components.

AlphaSim has an efficient system of data storage that makes the simulation of whole-chromosome haplotype sequences in very large pedigrees computationally feasible. This system includes, among other aspects, the representation of strings of zeros and ones in segments of genome as long integers, meaning that more sequence information can be stored in a given segment of memory. Also, the user can define a rate at which the genome is reduced in its representation, which specifically means that only a portion of the segregating sites in the base haplotypes are used in the subsequent part of the simulation. This option allows a reduction in the computational time and memory requirements for the simulation while maintaining all or most of its properties depending on the aims of the simulation. Additionally, standard file zipping procedures are used to compress the larger files.

AlphaSim makes extensive use of the hard disk to store files, which allows the required virtual memory to be managed. The files are stored in the Chromosomes directory and account for the largest part of disk space that is used by the simulation process. To release this disk space, the user has the option to discard the files stored in the Chromosomes directory once the



Simulated Data

- Pedigree
- Average co-ancestry in each generation
- TGV, TBV, TDV, Phenotypes
- Trait variance components
- Allele frequency and physical position of QTN across chromosomes
- QTN genotypes across chromosomes and generations
- QTN additive and dominance effects
- Total additive and dominance genic variances in each generation
- Number of QTN explaining from 1 to 50% of the total additive genic variance
- QTN sorted for gene editing
- Number of recombination in each haplotype of each individual
- Maximum position at which a recombination occurred on each chromosome
- Allele frequency and physical position of SNP across chromosomes
- SNP genotypes across chromosomes and generations

Chromosomes

- MaCS haplotypes
- Number of segregating sites
- Physical map
- Physical position and minor allele frequency of each segregating site
- Physical position and minor allele frequency of each SNP and QTN
- Identifiers of the base individual's haplotypes
- Genome sequence of each individual
- SNP and QTN genotype and phased haplotypes of each individual
- Allele frequency of the non-zero allele of each QTN in each generation

Selection

- Genotypes and phenotypes of the training individuals and selection candidates
- TGV and TBV of the selection candidates
- Estimates of SNP effects
- gEBV and accuracy
- pEBV and accuracy
- Selection index
- Individuals and QTN to be edited
- Genomic relationship matrix
- Optimal contribution matings

Fig. 3. Output of AlphaSim by directory.

simulating process has ended. Finally, the user can use the flexibility of AlphaSim to further reduce memory and storage requirements by breaking large simulations into manageable blocks (by generation or biparental family, etc.) using the restart functionality.

Results

In this section we provide four examples of plant breeding programs simulated using AlphaSim and illustrate the computational requirements of the software. We have demonstrated some examples of the animal breeding applications elsewhere (e.g., Hickey et al., 2011; Gorjanc et al., 2015a,b; Jenko et al., 2015).

Example 1: Genomic Best Linear Unbiased Prediction selection and Genotype × Environment Interactions in Biparental Families

We simulated a pedigree comprising five biparental families in which recombinant inbred lines (RILs) were selected and evaluated in contrasting environments using an experimental design with several replicates. Each biparental family was derived from a cross between two DH lines. Selfing each of the five F_1 individuals was simulated to result in four F_2 individuals per family, that is, 20 F_2 individuals in total. The F_2 individuals were then selfed through a single-seed descent process for eight generations to simulate 20 F_{10} RILs. Five F_{10} RILs were selected based on their gEBV to generate a new generation. In total, the pedigree included 12 burn-in generations and one selection generation (Fig. 4). Because performing genomic selection requires the presence of a population of individuals for which both phenotypes and SNP genotypes are available to train the prediction Eq. [10], a base generation including a large number of individuals (e.g., 1000 individuals) was simulated. As illustrated in Fig. 4, the number of individuals and the number of parents to be selected for each generation was specified according to the applied mating design. AlphaSim then simulated the pedigree so the plants in a given generation were equally distributed among the matings.

Genotype × environment interactions were simulated using the multiple traits capability of AlphaSim with each trait representing a distinct environment. The heritability was set to 0.8 and 0.2, respectively, for the Environment 1 and 2. The a priori additive genetic variance was set to 1.0 in both environments, the genetic correlation between the environments was set to 0.8, and the residual correlation was set to zero. Dominance effects were assumed to be null. This setting simulated a correlated G + G × E value for each individual in each of the two environments. These G + G × E values were a sum of the main genotypic effect and interaction with the environment. Adding independent residuals gave rise to phenotypic values in each of the two environments.

Genomic selection was conducted in F_{10} by using the default truncation selection method. The five best-performing F_{10} RILs were selected based on their gEBV. For this purpose, the training population was comprised of the 1000 individuals from the first generation of the

Generation	Number of Individuals	Number of parents				Create double haploid
				bisexual		
		male	female	crossed	selfed	
1 (Founders)	1000	10	0	0	0	No
2 (Parents)	10	5	5	0	0	Yes
3 (F$_1$)	5	0	0	0	5	No
4 (F$_2$)	20	0	0	0	20	No
5 (F$_3$)	20	0	0	0	20	No
6 (F$_4$)	20	0	0	0	20	No
7 (F$_5$)	20	0	0	0	20	No
8 (F$_6$)	20	0	0	0	20	No
9 (F$_7$)	20	0	0	0	20	No
10 (F$_8$)	20	0	0	0	20	No
11 (F$_9$)	20	0	0	0	20	No
12 (F$_{10}$)	20	0	0	0	5	No
13 (F$_{11}$)	15	-	-	-	-	No

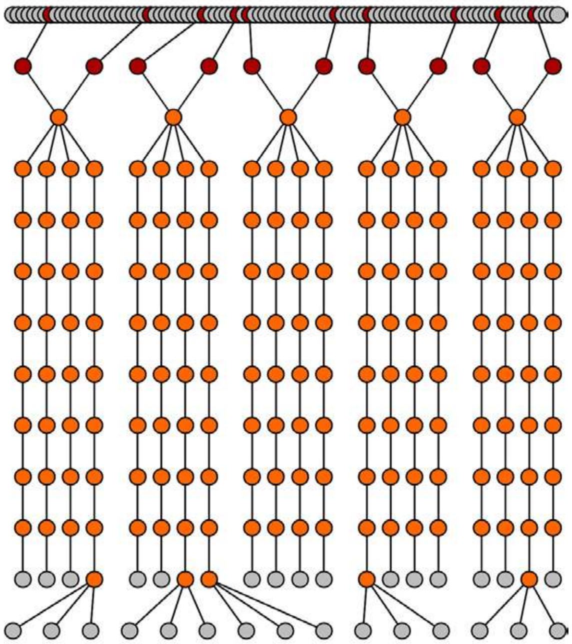


Fig. 4. Simulation of the plant breeding pedigree in Example 1. The pedigree includes 12 burn-in generations (from founders to F$_{10}$) and one selection generation (F$_{11}$). Ten randomly selected founders are used to generate 10 double haploids (DHs). These DHs are crossed to simulate five unique F$_1$ individuals. Selfing the F$_1$ individuals results in 20 F$_2$ individuals (i.e., four per one F$_1$). The F$_2$ genotypes are selfed through single-seed descent for eight generations to generate 20 F$_{10}$ or recombinant inbred lines (RILs). Five RILs are then selected and selfed to create three F$_{11}$ each.

pedigree. The SNP effects were estimated in each environment independently using the genotypes of the training individuals and their phenotypes, which were simulated in each environment. The gEBVs were then computed in each environment independently before being integrated into selection indices using the provided index weights. Here, the same importance was given to each environment by setting the index weights to 0.5. Finally, selection was performed in F$_{10}$ based on the genomic-estimated selection indices.

The five selected F$_{10}$ plants were selfed to generate F$_{11}$ seeds. These latter were tested in the two distinct environments with three replicates so that three phenotypic values were simulated for each RIL in each environment. The five simulated RILs showed contrasting performance in the two distinct environments (Fig. 5).

Example 2: User-Defined Selection in Biparental Families

A pedigree including five burn-in generations and one generation derived from selection was generated with a structure similar to Example 1 (Fig. 6). The best-performing F$_3$ individual in each of the five biparental families was selected and selfed to create three F$_4$ individuals. The selection of one single F$_3$ individual in each family was achieved using the restart functionalities of AlphaSim. Specifically, the simulating process was stopped after the creation of the gEBV for the F$_3$ individuals, enabling selection decisions to be made outside the program. The pedigree of Generation 6, that is, the pedigree of the F$_4$ individuals, was externally created and then imported into AlphaSim before resuming the simulation process.

Example 3: Eight-Parent Multiparent Advanced-Generation Intercross Population

Some populations used in plant breeding have a pedigree structure that includes a very specific crossing scheme. For this example, we used the pedigree of an eight-parent multiparent advanced-generation intercross (MAGIC) population, whose power for the dissection of the genetics of traits has been demonstrated (Mackay et al., 2014). The pedigree included a total of 561 individuals: the eight parental varieties, the 28 possible F$_1$ individuals derived from crossing two parents (excluding reciprocal crosses), the 210 possible F$_2$ individuals derived from crossing two unrelated F$_1$ parents, and the 315 F$_3$ individuals derived from crossing two unrelated F$_2$ parents (Mackay et al., 2014). Because of the specificity of the crossing structure, the pedigree of the eight-parent MAGIC population was constructed externally and then imported into AlphaSim. Since externally imported pedigrees and internally simulated pedigrees are compatible in AlphaSim,

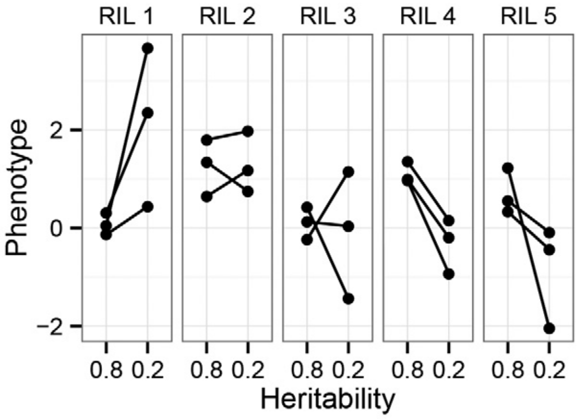


Fig. 5. Results of genotype × environment interaction simulated in Example 1. Five recombinant inbred lines (RILs) tested in three replicates in two contrasting environments with heritabilities of 0.8 and 0.2.

additional generations can be integrated into the MAGIC pedigree. This feature could be used to derive RILs from the 315 F_3 individuals as demonstrated in Example 1.

Example 4: Plant Breeding Programs

A further plant breeding capability of AlphaSim is demonstrated with a simulation of the development of (pseudo) F_4 RILs using single-seed descent combined with recurrent selection on F_2 plants. The simulation included three scenarios that differed from each other by the number of cycles of recurrent selection: 0, 2, or 4 (Fig. 7). The scenarios begin with a common pair of initial parents simulated using a single run of AlphaSim. Crossing these parents generated the F_1 population, which was then selfed to generate F_2 plants. The output from this latter run was copied to three distinct new locations, in which each scenario was run using the flexibility option. Recurrent selection consisted of selecting the two best performing F_2 individuals based on their gEBV and crossing them to generate new F_2 plants (Fig. 7). Because the simulation did not include a training population for genomic selection, we took gEBV to be a phenotype with a heritability of 0.6. After completing all cycles of recurrent selection, F_4 RILs were developed using single-seed descent. The resulting F_4 RILs were used to compare the performance of the three scenarios (Table 2).

Computational Requirement

AlphaSim was benchmarked using the simulation of two distinct scenarios that were each run twice with or without requesting the full genome sequence to be written out (Table 3). The two scenarios differed from each other by the number of segregating sites in the genome, the numbers of SNP and QTN, the size of the pedigree, and the size of the genomic selection training population, all larger in Scenario 2 than in Scenario 1 (Table 3). The genome was comprised of 10 chromosomes, each 1 Morgan in length. In Scenario 1, MaCS used parameters relating to the historical effective population size, mutation rate, and recombination rate, resulting in an average of 71,190 segregating sites across the genome, while in Scenario 2, there was an average of 163,590 segregating sites across the genome. Totals of 5000 and 20,000 SNP, and 2500 and 10,000 QTN, respectively, were sampled from the segregating sites of Scenarios 1 and 2. Two traits were simulated with heritability and variance–covariance components as described in Example 1. The structures of the pedigrees were as described in Fig. 4. In Scenario 1, the pedigree included 1210 individuals distributed along the pedigree as shown in Fig. 4. In Scenario 2, the pedigree included 234,500 individuals; 50,000, 2000, and 1000 individuals in Generations 1, 2, and 3, respectively; 20,000 in Generations 4 to 12; and 1500 in Generation 13. Genomic selection was performed in Generation 12 using a training population of 1000 and 30,000 individuals, respectively, that were sampled from the first generation of Scenarios 1 and 2.

Computations were performed on a Linux server. Scenario 1 was run using one CPU core with 2 GB of RAM available from a dual Intel Westmere E5620 2.4-GHz quad-core processor. Scenario 2 was run using 12 CPU cores with 5 GB of RAM available from dual Intel Westmere E5645 2.4-GHz six-core processors. For Scenario 1, running time was 1 min 34 s and increased to 3 min 39 s when the full sequence information was written to disk (Table 3). For Scenario 2, the running time was 4 h 9 min 54 s and increased to 19 h 6 min 11 s when the full sequence was written to disk.

Discussion

AlphaSim is a new software package for simulating breeding program designs that use sequence data, pedigrees, genotypes, and phenotypes. Different mating systems enable simulation of plant or animal populations. AlphaSim extends the scope of the currently available plant breeding simulation packages because of its wide flexibility, enabling the design of almost any pedigree structure and the application of many selection methods, in particular genomic selection and genome editing as demonstrated in the above examples and other previously published work (Clark et al., 2012; Daetwyler et al., 2013; Hickey et al., 2014, 2015; Gorjanc et al., 2015a).

AlphaSim can be used for the simulation of small datasets in a very short time. Simulating large pedigrees with large genome sequence significantly increases the running time, particularly when writing the full sequence data to disk (Table 3). However, when simulating distinct scenarios characterized by the same SNP panels, QTN, and trait information, the simulation time can be significantly reduced using the restart functionality of the software, that is, by deriving each scenario from a common base generation. For example, we have successfully used this approach to simulate a wheat breeding program with genomic selection spanning 41 yr (overlapping generations) with 1.7 million unique genotypes per year or a pig (*Sus scrofa domesticus*) breeding program with genomic selection spanning 30 yr (overlapping

Generation	Number of Individuals	Number of parents				Create double haploid
				bisexual		
		male	female	crossed	selfed	
1 (Founders)	1000	10	0	0	0	No
2 (Parents)	10	5	5	0	0	Yes
3 (F_1)	5	0	0	0	5	No
4 (F_2)	20	0	0	0	20	No
5 (F_3)	20	0	0	0	5	No
6 (F_4)	15	-	-	-	-	No

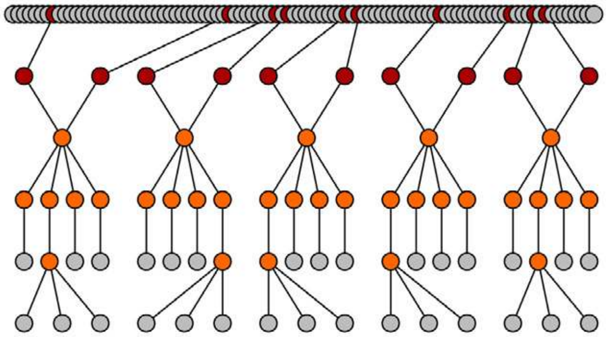


Fig. 6. Simulation of the plant breeding pedigree in Example 2. The pedigree includes five burn-in generations (from founders to F_3) and one selection generation. Ten randomly selected founders are used to generate 10 double haploids (DHs). These DHs are crossed to simulate five unique F_1 genotypes. Selfing the F_1 individuals result in 20 F_2 individuals, four per F_1. The F_2 individuals are selfed through single-seed descent for one generation to simulate 20 F_3, and the best performing F_3 in each of the five biparental families is selected and selfed to create three F_4 individuals.

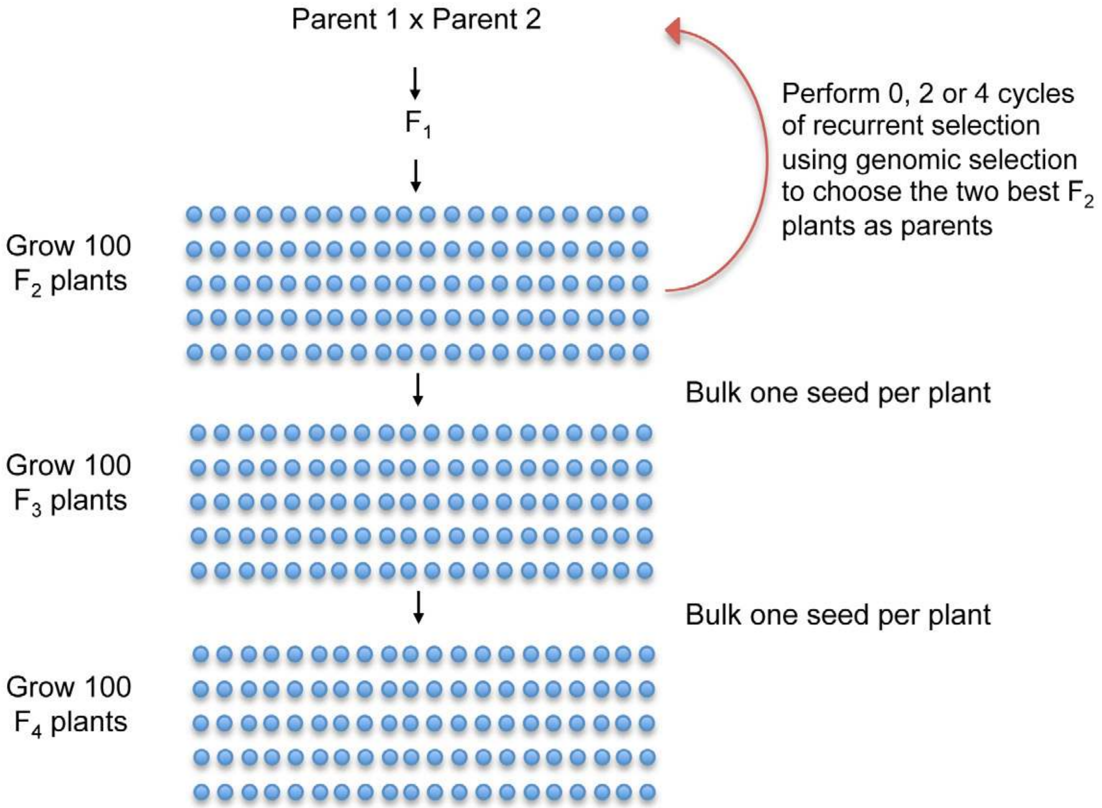


Fig. 7. Simulation of the development of F_4 derived recombinant inbred lines (RILs) using single-seed descent combined with recurrent selection on F_2 plants. The simulation included three scenarios differing from each other by the number of cycles of recurrent selection, which was 0, 2, or 4. The scenarios begin with a common pair of parents, which were crossed to generate the F_1 plants. Selfing the F_1 generated F_2 plants. Recurrent selection consisted of selecting the two best performing F_2 individuals based on their genomic estimated breeding values and crossing them to generate new F_2 plants. The F_4 RILs were then developed using single-seed descent.

generations) with 35,000 unique genotypes per year (R.C. Gaynor and J.M. Hickey, unpublished data, 2016).

In conclusion, we make three points: (i) AlphaSim allows breeders and researchers to simulate genomic data controlled by very specific user criteria, to evaluate the power of diverse breeding programs, and to optimize their requirements in terms of sequencing, genotyping, and phenotyping resources; (ii) AlphaSim is flexible, computationally efficient, and easy to use for a wide range of possible scenarios; and (iii) AlphaSim was designed to

Table 2. Benchmarking of AlphaSim through the simulation of three distinct plant breeding programs.

	Base	Scenario 1	Scenario 2	Scenario 3
User features				
Number of chromosomes	7			
Number of segregating sites	30,356			
Start–stop generation	1–2	3–5	3–9	3–13
Number of individuals	52	300	504	708
Results				
Genetic variance F_4 stage	–	0.333	0.110	0.002
Mean gEBV† F_4 stage	–	1.315	2.182	2.584
Computational feature				
Running Time	0 m 6 s	0 m 4 s	0 m 7 s	0 m 10 s

† gEBV, genomic estimated breeding value.

Table 3. Benchmarking of AlphaSim through the simulation of two distinct scenarios run with or without requesting the full genome sequence to be written out.

	Scenario 1		Scenario 2	
User features				
Write out the full genome sequence	No	Yes	No	Yes
Number of segregating sites	70,140	72,240	162,780	164,400
Number of SNP†	5000		20,000	
Number of QTN‡	2500		10,000	
Pedigree size (no. of individuals)	1210		234,500	
Size of the training population	1000		30,000	
Computational feature				
Running Time	1 m 34 s	3 m 39 s	4 h 9 m 54 s	19 h 6 m 11 s

† SNP, single-nucleotide polymorphism.

‡ QTN, quantitative trait nucleotide.

simulate plant breeding programs; however, it can be used in many other fields of genetics, including animal breeding, human genetics, and population genetics.

Finally, we plan to continue to add new features to AlphaSim to increase its functionality and to respond to the requirements of simulation technology and breeding program designs that will emerge in the coming years.

Availability

AlphaSim is available from http://www.alphagenes.roslin.ed.ac.uk/alphasuite/alphasim/. Material available includes the compiled programs for 64-bit Linux, Mac OSX, and Windows together with a user manual. This material also includes a 47-page user manual and a 51-page set of simple examples with step-by-step instructions that is aimed at prospective users who have no experience with Linux or Mac OSX.

Acknowledgments

The development of AlphaSim has been funded under multiple projects. The authors would like to acknowledge specific contribution from The Australian Research Council, Aviagen LTD, Genus PLC, Pfizer Inc., The Sheep CRC, CIMMYT, Advanta Semillas, the Seeds of Discovery Project supported by SAGARPA (La Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación), Mexico under the MasAgro (Sustainable Modernization of Traditional Agriculture) initiative, the ISPG funding from the BBSRC to The Roslin Institute, and the GplusE project funded by the BBSRC.

References

Bernardo, R. 2010. Breeding for quantitative traits in plants. 2nd ed. Stemma Press, Woodsbury, MN.

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082. doi:10.2135/cropsci2006.11.0690

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4. doi:10.1186/1297-9686-44-4

Daetwyler, H.D, M.P.L. Calus, R. Pong-Wong, G. de Los Campos, and J.M. Hickey. 2013. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365. doi:10.1534/genetics.112.147983

Gorjanc, G., P. Bijma, and J.M. Hickey. 2015a. Reliability of pedigree-based and genomic evaluations in selected populations. Genet. Sel. Evol. 47:65. doi:10.1186/s12711-015-0145-1

Gorjanc, G., M.A. Cleveland, R.D. Houston, and J.M. Hickey. 2015b. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. 47:12. doi:10.1186/s12711-015-0102-z

Gorjanc, G., J. Jenko, S.J. Hearne, and J.M. Hickey. 2016. Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17:30. doi:10.1186/s12864-015-2345-z

Heffner, E.L., M.E. Sorrells, and J.L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512

Henderson, C.R. 1984. Applications of linear models in animal breeding. University of Guelph, ON, Canada.

Hickey, J.M., and G. Gorjanc. 2012. Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3: Genes, Genomes, Genet. 2:425–427. doi:10.1534/g3.111.001297

Hickey, J.M., G. Gorjanc, B.E. Huang, and S. Hearne. 2014. AlphaMPSim: Flexible simulation of multi-parent crosses. bioinformatics. 30:2686–2688. doi:10.1093/bioinformatics/btu206

Hickey, J.M., G. Gorjanc, R.K. Varshney, and C. Nettelblad. 2015. Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden Markov Model. Crop Sci. 55:1934–1946. doi:10.2135/cropsci2014.09.0648.

Hickey, J.M., B.P. Kinghorn, B. Tier, J.F. Wilson, N. Dunstan, and J.H.J. van der Werf. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. 43:12. doi:10.1186/1297-9686-43-12

Hoerl, A.E., and R.W. Kennard. 1976. Ridge regression iterative estimation of the biasing parameter. Commun. Stat. Theory Methods 5:77–88. doi:10.1080/03610927608827333

Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomics 9:166–177. doi:10.1093/bfgp/elq001

Jenko, J., G. Gorjanc, M.A. Cleveland, R.K. Varshney, C.B. Whitelaw, J.A. Woolliams, and J.M. Hickey. 2015. Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. Genet. Sel. Evol. 47:55. doi:10.1186/s12711-015-0135-3

Mackay, I.J., P. Bansept-Basler, T. Barber, A.R. Bentley, J. Cockram, N. Gosman, et al. 2014. An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: Creation, properties, and validation. G3: Genes, Genomes, Genet. 4:1603–1610. doi:10.1534/g3.114.012963.

Meuwissen, T.H.E. 1997. Maximizing the response of selection with a predefined rate of inbreeding. J. Animal Sci. 75:934–940.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Randhawa, H.S., J.S. Mutti, K. Kidwell, C.F. Morris, X. Chen, and K.S. Gill. 2009. Rapid and targeted introgression of genes into popular wheat cultivars using marker-assisted background selection. PLoS One 4:e5752. doi:10.1371/journal.pone.0005752

Shan, Q., Y. Wang, J. Li, and C. Gao. 2014. Genome editing in rice and wheat using the CRISPR/Cas system. Nat. Protoc. 9:2395–2410. doi:10.1038/nprot.2014.157

Shukla, V.K., Y. Doyon, J.C. Miller, R.C. DeKelver, E.A. Moehle, S.E. Worden, et al. 2009. Precise genome modification in the crop species *Zea mays* using zinc-finger nucleases. Nature 459:437–441. doi:10.1038/nature07992

Sun, X., T. Peng, and R.H. Mumm. 2011. The role and basics of computer simulation in support of critical decisions in plant breeding. Mol. Breed. 28:421–436. doi:10.1007/s11032-011-9630-6

Wang, J., M. van Ginkel, D. Podlich, G. Ye, R. Trethowan, W. Pfeiffer, I.H. DeLacy, M. Cooper, and S. Rajaram. 2003. Comparison of two breeding strategies by computer simulation. Crop Sci. 43:1764–1773. doi:10.2135/cropsci2003.1764

Wang, Y., X. Cheng, Q. Shan, Y. Zhang, J. Liu, C. Gao, and J.L. Qiu. 2014. Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat. Biotechnol. 32:947–951. doi:10.1038/nbt.2969

Wellmann, R., and J. Bennewitz. 2011. The contribution of dominance to the understanding of quantitative genetic variation. Genet. Res. 93:139–154. doi:10.1017/S0016672310000649

Wellmann, R., and J. Bennewitz. 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet. Res. 94:21–37. doi:10.1017/S0016672312000018

Whittaker, J.C., R. Thompson, and M.C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. 75:249–252. doi:10.1017/S0016672399004462

Wray, N.R., and M.E. Goddard. 1994. Increasing long-term response to selection. Genet. Sel. Evol. 26:431. doi:10.1186/1297-9686-26-5-431