Traits in AlphaSimR

Chris Gaynor

2023-03-27

This vignette describes AlphaSimR's biological model for traits. The biological model is responsible for converting an individual's genotype into a genetic value. The genetic value is used to create a phenotype. AlphaSimR's biological model is primarily based on classic models used in quantitative genetics. Thus, users that have taken an introductory quantitative genetics course should already be familiar with most elements of AlphaSimR's biological model. This vignette will assume the reader is such a person and thus focus on aspects of AlphaSimR's biological model that may not be obvious or don't follow classic models. The vignette will begin by describing a novel classification system used in AlphaSimR to name different types of traits. Then it will present the full biological model along with a detailed description for each component in the model.

Traits in AlphaSimR are classified according to the biological effects they model using the **ADEG** framework. Under this framework, each trait is assigned a name consisting of one or more letters. The letters come from the name **ADEG**, whose letters correspond to biological effects. The biological effects are: **A**dditive, **D**ominance, **E**pistatic and **G**enotype-by-environment. For example, a trait with only additive effects is called an **A** trait and a trait with both additive and dominance effects is called an **AD** trait. The following traits are modeled in AlphaSimR: **A**, **AD**, **AE**, **AG**, **ADE**, **ADG**, **AEG**, and **ADEG**.

The most complex trait in AlphaSimR is an **ADEG** trait, because it includes all biological effects. This trait is represented using the equation below.

$$GV(x, w) = \mu + A(x) + D(x) + E(x) + G(x, w) \tag{1}$$

The left-hand side of equation (1) consists of the following terms: GV, which represents an individual's genetic value; x, which represents a vector of QTL genotype dosages; and w, which represents an environmental covariate. The right-hand side of the equation consists of an intercept (μ) and four functions. The intercept is used to obtain the user specified trait mean in a founder population. The functions model each of the biological effects. All other trait in the **ADEG** framework can be modeled with equation (1) by simply removing any functions for biological effects that those traits lack.

The discussion of the **ADEG** framework will continue below with an explanation for each function in equation (1) after first describing the concept of genotype dosage scaling. Genotype dosage scaling is discussed first, because it is applied within each of the aforementioned functions.

Genotype Dosage Scaling

AlphaSimR defines an individual's raw genotype dosage as the number of copies of the "1" allele at a locus. Since all loci in AlphaSimR are biallelic with alleles "0" and "1", this definition for raw genotype dosage fully explains an individual's genotype. The number of copies of the "0" allele is just the individual's ploidy level minus its raw genotype dosage. This is usually a rather irrelevant detail, because most user will want to model diploid organism whose ploidy level is always two. However, AlphaSimR can also be used to model a wide range of autopolyploid organisms and it even allows for mixing ploidy levels within a simulation. This means that the allowable range for raw genotype dosage is not a fixed value and that software must account for this fact.

AlphaSimR accounts for different levels of ploidy by scaling an organism's genotype dosage in accordance with its ploidy level. The primary motivation for using scaled dosages is to unify user inputs. The use of scaled dosages can also make it easier to compare simulations with different levels of ploidy. However, the user must use care when making such comparisons, because the underlying assumptions used by this model may not be valid.

The use of scaled genotype dosages implies that an individual's genetic value depends on the relative ratio of alleles and that it is independent of the organism's ploidy level. For some traits in some organisms there is evidence to support this assumption or at least not reject it, but there is also plenty of evidence clearly rejecting this assumption in other traits (Gallais 2003). Thus, the direct comparisons that are possible in AlphaSimR are not always biologically relevant. However, the vast majority of users won't be running a simulation that depends on this assumption, so the information presented below is solely for the purpose of understanding the functions presented in the following sections.

There are two types of scaled genotype dosages in AlphaSimR: additive and dominance. An explanation for both is given below. This is followed by a table providing an example of dosage scaling in both diploid and autotetraploid organisms.

The scaled additive genotype dosage (x_A) is shown in equation (2). This equation linearly scales relative dosage to set the values for opposing homozygotes to -1 and 1.

$$x_A = \left(x - \tfrac{ploidy}{2}\right)\left(\tfrac{2}{ploidy}\right) \tag{2}$$

The scaled dominance genotype dosage (x_D) is shown in equation (3). This equation uses non-linear scaling to fit the value of the opposing homozygotes to 0 and middlemost heterozygote to 1. The middlemost heterozygote is the genotype with an equal ratio of "0" and "1" alleles. For a diploid organism, the scaled dominance genotype dosage matches the classic parameterization for dominance. For autopolyploid organisms, the scaled dominance genotype dosage is consistent with digenic dominance (discussed later).

$$x_D = x\left(ploidy - x\right)\left(\tfrac{2}{ploidy}\right)^2 \tag{3}$$

Table 1 provides an example of raw and scaled genotype dosages for a diploid and an autotetraploid organism. The diploid and tetraploid columns represent the raw genotype dosages for the respective organisms. The additive and dominance columns represent the corresponding scaled genotype dosages. This table shows how the 0, 1 and 2 genotypes of a diploid organism are treated as being equivalent to the 0, 2 and 4 genotypes of an autotetraploid organism.

Table 1: Raw and scaled genotype dosages.

Diploid	Tetraploid	Additive	Dominance
0	0	-1	0
	1	$-1/2$	3/4
1	2	0	1
	3	1/2	3/4
2	4	1	0

Additive Effects

$$A(x) = \sum ax_A \tag{4}$$

The function for additive effects is given above in equation (4). The right-hand side is a summation over all QTL for the product of the additive effect (a) and the scaled additive dosage (x_A). This equation is

equivalent to parameterizations in classic quantitative trait models. The only unique aspect of additive effects in AlphaSimR is the sampling of those effects.

Additive effects are sampled in two stages. The first stage involves sampling initial values and is similar to methods used by other stochastic simulation software programs. It is the second stage that is unique. This stage involves scaling the magnitude of the initial values to achieve a desired genetic variance, that being either total or additive genetic variance.

The first stage of sampling additive effects is setting initial values for the effects. Initial values are sampled from either a standard normal distribution or a gamma distribution. When using a gamma distribution, the user specifies the shape parameter and the scale parameter is set to 1. The deviates sampled from a gamma distribution are randomly assigned either a positive or negative sign. Random assignment of this sign results in an expected distribution that is symmetric distribution with mean 0.

The second stage is scaling the magnitude of the effects to achieve a user specified genetic variance. The user specified genetic variance can be either total or additive. The scaling procedure involves not just the additive effect, but also dominance and epistatic effects if the trait includes those effects as well. The procedure works by first calculating the variance in the founder population using the initially sampled effects and then calculating a scaling constant that is applied to all effects to achieve the desired variance in the founder population. The scaling constant equals the square-root of the target genetic variance (total or additive) divided by the square-root of the initial genetic variance (total or additive). Note that the founder population is the population the user uses to initialize the simulation parameters object.

Scaling the magnitude of effects allows AlphaSimR to build simulations with trait values matching real-world counterparts. For example, a user simulating grain yield in a plant species is likely to have estimates for the means and variances in tons per hectare. The user can use those to create a simulation with the same values. The primary benefit of matching these values is to make interpretation of simulation results more intuitive. The significance of the results will be more apparent to the user because they can work in a scale that is familiar to them. It also allows the user to more easily identify potential flaws in the simulation when values move outside the range of biologically acceptable values.

Dominance Effects

$$D(x) = \sum dx_D \tag{5}$$

The function for dominance effects is given above in equation (5). The right-hand side of the equation is a summation over all QTL for the product of the dominance effect (d) and the scaled dominance dosage (x_D). This equation is equivalent to the parameterization of dominance in classic quantitative trait models for diploid organisms. As with the additive effects, the method for sampling dominance effects in AlphaSimR requires special attention.

Dominance effects are calculated in AlphaSimR using the concept of dominance degrees. The formula for this calculation is given below in equation (6). The equation shows that the dominance effect (d) at a locus is the dominance degree (δ) at that locus times the absolute value of its additive effect (a).

$$d = \delta\,|a| \tag{6}$$

The rationale behind using dominance degrees is their intuitive biological interpretation in diploid organisms. For example, a dominance degree of 0 represents no dominance and an additive model. A dominance degree of 1 corresponds to complete dominance. Dominance degrees between 0 and 1 correspond to partial dominance, and values above 1 correspond to over-dominance.

Dominance effects, as with additive effects, are sampled in two stages. The first stage is sampling of initial values and the second stage is scaling the magnitude of those values. The sampling of initial effects involves two user supplied parameters, the mean and variance for a normal distribution used to sample dominance degrees. The dominance degrees are then used in conjunction with the additive effects to calculate dominance

effects. The scaling procedure is then performed as described above in the additive effects section to calculate the scaling constant. The scaling constant is then directly applied to the dominance effects. Note that this scaling changes the value of the dominance effect, but does not change the value of the dominance degree. This means that the specification of the dominance degree distribution is independent of the requested genetic variance.

Dominance effects become more complicated when discussing polyploid organisms. This is because polyploids have additional heterozygous genotypes. For example, an autotetraploid organism has two homozygous genotypes (0 and 4) and three heterozygous genotypes (1, 2 and 3). Each additional heterozygous genotype requires an additional dominance parameter to obtain a fully parameterized model. Several different parameterizations have been used for polyploids and a full discussion of these parameterizations is outside the scope of this document. Interested readers should refer to Gallais's textbook for more details (2003).

AlphaSimR does not include additional term to its dominance model for polyploids, so it does not make use of a fully parameterized model. Instead, AlphaSimR uses a digenic dominance model for all ploidy levels due to its use of scaled dominance genotype dosages. This model was chosen because it provides consistency in user interface and internal coding regardless of ploidy level and it provides as a reasonable approximation for partial dominance in highly polygenic traits.

An unfortunate side effect of the dominance model in polyploids is that the previously described interpretation of dominance degrees breaks. Consider a single QTL with an additive effect of 1 and a dominance degree of 1. The genotypes and genetic values for the example for both a diploid and an autotetraploid organism are given below in Table 2. For the diploid organism, the heterozygous genotype is equivalent to the best homozygote, so a dominance degree of 1 indicates complete dominance. However, the autotetraploid organism has three heterozygous genotypes. The value of the middlemost heterozygous genotype is equivalent to the best homozygote, but it is not the heterozygous genotype with the highest value. The heterozygous genotype with the highest value is better than the best homozygote, so a dominance degree of 1 actually represents over-dominance in an autotetraploid.

Table 2: Example genetic values ($a = 1$, $d = 1$).

Diploid Dosage	Tetraploid Dosage	Genetic Value
0	0	-1
	1	$1/4$
1	2	1
	3	$5/4$
2	4	1

Epistatic Effects

$$E(x) = \sum ex_{A_1}x_{A_2} \tag{7}$$

The function for epistatic effects is given above in equation (7). The summation in the right-hand side is over pairs of QTLs, because AlphaSimR uses a simplified model of epistasis that restricts epistatic interactions to pairs of loci. Each QTL must be present in one and only one pair, so the number of pairs is equal to half the number of QTL. The remaining elements in the right-hand side of equation (7) are epistatic effects (e), the scaled additive dosage for the first locus in a pair (x_{A_1}) and the scaled additive dosage for the second locus in a pair (x_{A_2}).

The epistatic model in equation (7) is somewhat constrained. The model only allows epistasis between pairs of loci and it only models additive-by-additive epistasis. The motivation behind using this constrained model is to maintain computational tractability when using very large numbers of QTL. As with the dominance model, this model is considered a reasonable approximation for a more complicated reality.

The sampling of epistatic effects is similar to the sampling of additive effects with one additional user specified parameter for the relative ratio of the additive-by-additive epistatic variance. The relative ratio refers to the ratio between additive-by-additive epistatic variance and additive variance. This parameter sets the variance of epistatic effects. Specifically, the variance of the effects is set to a value expected to achieve the desired ratio in a random mating population whose QTL are in linkage equilibrium and have an allele frequency of 0.5. Note that most populations, including simulated populations, will not meet these assumptions, so the observed relative ratio between additive and additive-by-additive epistatic variance is unlikely to match the requested ratio. The observed additive-by-additive epistatic variance will typically be smaller, because its value is maximized at allele frequency 0.5. Note that unselected bi-parental populations derived from inbred parents are expected to have allele frequencies of 0.5 for segregating alleles. Thus, these populations are ideal for estimating a reasonable value for the relative ratio of additive-by-additive epistatic variance.

Genotype-by-Environment Effects

$$G(x, w) = wb(x) \tag{8}$$

$$b(x) = \mu_G + \sum gx_A \tag{9}$$

The function for genotype-by-environment effects is given above in equations (8) and (9). The right-hand side of the equation (8) contains two parts: an environmental covariate (w) and a genotype specific slope ($b(x)$). The formula for the genotype specific slope is shown in equation (9). The right-hand side of this equation includes an intercept value (μ_G) and a summation over all QTL for the product of a genotype-by-environment effect (g) and the scaled additive dosage (x_A).

The following paragraphs explain how the above equations model genotype-by-environment interactions. Initially, the explanation will cover a case where a population has genotype-by-environment interaction variance, but no environmental variance. An explanation for how the parameters of these equations are changed to model both genotype-by-environment interaction variance and environmental variance is given in the next section.

To begin, a description of how the variables in the above equations are sampled is needed. This is because it is much easier to understand how the formulas come together to model genotype-by-environment interactions when it is understood what variables represent.

The environmental covariate (w) in equation (8) represents an environmental component of the genotype-by-environment interaction. The value of the environmental covariate is randomly sampled from a standard normal distribution. By definition, this means that the average value of the environmental covariate is zero and its variance is one. The average value of the environmental covariate is considered to be the target environment. Thus, the value for equation (8) in the target environment is always zero.

The genotype specific slope in equation (9) represents the genetic component of the genotype-by-environment interaction. The astute reader will notice that this equation is very similar to the equation for an additive trait. Indeed, the genotype specific slope is really just an additive trait. The term μ_G serves a similar role to μ in equation (1) and the summation on the right-hand side of equation (9) is similar to the function for additive effects in equation (4). Also like an additive trait, effects for the genotype specific slope are scaled to achieve a specific mean and variance in the founder population. In this case, the mean is set to zero and the variance is set to the user specified genotype-by-environment interaction variance.

To understand how the environmental covariate and the genotype specific slope model genotype-by-environment interactions it helps to review the properties of these variables. The environmental covariate is just a random variable with a mean of zero and a variance of one. The genotype specific slope is also random variable, with regards to the founder population, that has a mean of zero and a variance equal to the genotype-by-environment variance. This means that equation (8) is just the product of two random variables. These random variables are independent, so the formula for the variance of this product is given in the equation below.

$$Var(wb) = E[w]^2 Var(b) + Var(w)E[b]^2 + Var(w)Var(b) \tag{10}$$

Equation (10) above gives the variance for the product of w and b in equation (8). The term $Var()$ indicates the variance of a variable and the term $E[]$ indicates the expectation (mean) of a variable. In the founder population, the expectations for w and b are zero, so the first two terms in the right-hand side of the equation drop out. Equation (10) reduces to the product of the two variances. The variance of w is one and the variance of b equals the genotype-by-environment interaction variance, so the variance in equation (8) equals the genotype-by-environment interaction variance and is equivalent to the variance in the genotype specific slopes. It must be repeated that the above description is limited to the founder population. This is an important point, because the mean and variance for genotype specific slope can and will be different in other populations since it is under genetic control.

Adding Environmental Variance

The genotype-by-environment effects model described above is altered to model a founder population with a non-zero environmental variance. All of the above equations remain relevant under the altered model and only the sampling distributions of the effects in the equations are changed. Specifically, the variance of environmental covariate is set to the environmental variance (σ^2_E) and the genotype specific slope is set to a different mean and variance. The mean of the genotype specific slope is set to one and the variance is set to the genotype-by-environment interaction variance divided by the environmental variance ($\frac{\sigma^2_{GE}}{\sigma^2_E}$).

The logic behind these modifications becomes clear with an examination of equation (10). The right-hand side of equation (10) contains three terms. The first term equals zero, because the environmental covariate has a mean of zero (*i.e.* $E[w] = 0$). The second term equals the environmental variance, because the mean for genotype specific slope equals one and the variance of the environmental covariate equals the environmental variance (*i.e.* $E[b] = 1$ and $Var(w) = \sigma^2_E$). Finally, the value of the last term equals the genotype-by-environment interaction variance (*i.e.* $Var(w)Var(b) = \sigma^2_E \frac{\sigma^2_{GE}}{\sigma^2_E} = \sigma^2_{GE}$).

An examination of equation (10) also shows how both the environmental and genotype-by-environment interaction variances are populations specific parameters that are under genetic control. For example, a population with a higher average genotype specific slope will also have a higher value for the second term in equation (10). This population thus have a higher environmental variance even though the environment itself has not changed. Likewise, the amount of genotype-by-environment interaction variance depends on the variance for the genotype specific slope.

Relationship to Finlay-Wilkinson Regression

AlphaSimR's model for genotype-by-environment effects is effectively a biological model for Finlay-Wilkinson regression. Finlay-Wilkinson regression is a classic technique for analyzing genotype-by-environment interactions (Finlay and Wilkinson 1963). In its simplest form, an individual's genotype-by-environment interaction is reduced to an intercept and a slope. The slope in Finlay-Wilkinson regression is roughly equivalent to genotype specific slope in equation (9). The intercept in Finlay-Wilkinson regression is roughly equivalent to the genetic value of an individual in the target environment.

Initially, the similarity between these two models was not intentional. The model used in AlphaSimR originates with the model used by Gaynor *et al.* (2017). This model treated each QTL's additive effect as random variable whose value depended on the value of an environmental covariate. That model is roughly equivalent to AlphaSimR's model when the founder population has zero environmental variance. The difference between the models is that AlphaSimR scales effects to exactly achieve a desired variance and the Gaynor *et al.* model randomly sampled effects so that the expectation of the sampled effects equaled the desired variance. It was only realized later that a slight change to the model, as described above, could introduce a specified amount of environmental variance and that those modifications would result in a model with properties matching those measured by Finlay-Wilkinson regression.

User Interaction

AlphaSimR users do not directly observe nor set the value of the environmental covariate. Instead, they indirectly set its value by providing a randomly sampled p-value for each environment. AlphaSimR then uses this p-value to calculate the appropriate value of w for any level of environmental variance. Users should note that p-values follow a uniform distribution over the range of zero to one, so it is recommended that users randomly generate p-values using R's built-in `runif` function. A p-value of 0.5 corresponds to w equaling zero and is equivalent to the target environment. If the user does not supply a p-value, AlphaSimR will sample one at random.

Genetic values and genetic variances reported by AlphaSimR are always for the target environment. Since the target environment corresponds to w equaling zero, equation (8) is effectively ignored in these calculations. Only the values for phenotypes and phenotypic variance ever reflect a contribution from genotype-by-environment interaction.

References

Finlay, K. W., and G. N. Wilkinson. 1963. "The Analysis of Adaptation in a Plant-Breeding Programme." *Australian Journal of Agricultural Research* 14 (6): 742–54. https://doi.org/10.1071/ar9630742.

Gallais, A. 2003. *Quantitative Genetics and Breeding Methods in Autopolyploid Plants*. Paris: INRA.

Gaynor, R. Chris, Gregor Gorjanc, Alison R. Bentley, Eric S. Ober, Phil Howell, Robert Jackson, Ian J. Mackay, and John M. Hickey. 2017. "A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines." *Crop Science* 57 (5): 2372–86. https://doi.org/10.2135/cropsci2016.09.0742.