


# Geno-Diver: A combined coalescence and forward-in-time simulator for populations undergoing selection for complex traits

J. T. Howard<sup>1</sup>  | F. Tiezzi<sup>1</sup> | J. E. Pryce<sup>2,3</sup> | C. Maltecca<sup>1</sup>

<sup>1</sup>Department of Animal Science, North Carolina State University, Raleigh, NC, USA

<sup>2</sup>Department of Economic Development, Jobs, Transport and Resources and Dairy Futures Cooperative Research Centre, Bundoora, Vic., Australia

<sup>3</sup>La Trobe University, Bundoora, Vic., Australia

## Correspondence

Jeremy Howard, Department of Animal Science, North Carolina State University, Raleigh, NC, USA.  
Email: jthoward@ncsu.edu

## Summary

Geno-Diver is a combined coalescence and forward-in-time simulator designed to simulate complex traits with a quantitative and/or fitness component and implement multiple selection and mating strategies utilizing pedigree or genomic information. The simulation is carried out in two steps. The first step generates whole-genome sequence data for founder individuals. A variety of trait architectures can be generated for quantitative and fitness traits along with their covariance. The second step generates new individuals forward-in-time based on a variety of selection and mating scenarios. Genetic values are predicted for individuals utilizing pedigree or genomic information. Relationship matrices and their associated inverses are generated using computationally efficient routines. We benchmarked Geno-Diver with a previous simulation program and described how to simulate a traditional quantitative trait along with a quantitative and fitness trait. A user manual with examples, source code in C++11 and executable versions of Geno-Diver for Linux are freely available at <https://github.com/jeremyhoward/Geno-Diver>.

## KEYWORDS

Animal breeding, fitness, quantitative genetics, simulation

## 1 | INTRODUCTION

The use of data simulation to generate genome and genetic architectures and test novel selection or mating methods/strategies has traditionally been a fundamental aspect of animal and plant breeding. Currently a variety of disciplines including conservation (McMahon, Teeling, & Höglund, 2014), animal and plant breeding (De los Campos, Hickey, Pong-Wong, Daetwyler, & Calus, 2013) and human genetics (Yang et al., 2010) are making use of an increasingly large amount of genomic information. Within animal breeding programs, the use of molecular markers to predict the genetic merit of individuals has become a routine practice (Jonas & de Koning, 2015). This has resulted in a significant increase in the number of genotyped individuals within a herd/population. Simulation has often been employed to compare a wide range of hypotheses relating to genetics and/or the genetic management of populations

at a low cost (Daetwyler, Calus, Pong-Wong, de Los Campos, & Hickey, 2013). The use of simulated data is particularly useful in determining the impact of current selection and management practices across time, which is often not possible using real data due to time and cost requirements. Simulation is also a useful tool to optimize the construction of marker panels in terms of SNP uniformity across the genome, the impact of the inclusion of preselected candidate causative mutations and the proportion of individuals to genotype in a population for a given marker density. Most previous research in this area has focused on how to best use genomic information to accurately predict the genetic merit of an individual (Henryon, Berg, & Sorensen, 2014). Conversely the ability to use this information to efficiently manage agricultural populations at the genomic level, both for preserving genetic diversity and lessening inbreeding depression, has received comparatively less attention (Henryon et al., 2014). Previous research has

highlighted that genomic information maintains greater diversity compared to traditionally utilized pedigree information in the context of minimizing parental relationships (Howard, Tiezzi, Huang, Gray, & Maltecca, 2016; Rodríguez-Ramilo, García-Cortés, & de Cara, 2016). Furthermore, in the context of optimal contribution selection, genomic information allows for specific regions to be constrained and in some cases (i.e., large full-sib families) results in increased genetic gain compared to pedigree information (Clark, Kinghorn, Hickey, & van der Werf, 2013; Gómez-Romano, Villanueva, Fernández, Woolliams, & Pong-Wong, 2016).

A number of simulation programs have been developed that mimic livestock breeding populations, but they primarily focus on testing strategies where the genetic architecture is based solely on one or multiple quantitative traits (Faux et al., 2016; Pérez-Enciso & Legarra, 2016; Sargolzaei & Schenkel, 2009). Currently there is an absence of self-contained software that can simulate complex traits involving both quantitative and fitness components along with the ability to generate complex pedigrees common to livestock breeding programs. Consequently, determining how various selection and management practices impact the fitness and the overall genetic variability of a population undergoing selection for a quantitative trait remains challenging. Furthermore, methodologies to identify lethal mutations utilizing genomic information have been successfully implemented (VanRaden, Olson, Null, & Hutchison, 2011). Yet optimal mating procedures to minimize the frequency of a large number of lethal, and more importantly, sublethal mutations across generations have not been fully implemented. As the popularity of genotyping individuals increases, the possibility of utilizing genomic information from multiple sources to manage the genome of a population will also increase. Methods that make effective use of information from multiple sources including performance, genomic diversity and inbreeding load at the selection and/or mating step are an increasing need. Here, we propose “Geno-Diver,” a combined coalescence and forward-in-time simulator designed to simulate complex traits involving quantitative and/or fitness components and implement multiple selection and mating strategies utilizing pedigree or genomic information. The software will be outlined in two sections: (i) a general overview of the software and its associated parameters and, (ii) a brief comparison with a previously developed simulation program (Sargolzaei & Schenkel, 2009) and two simulation scenarios that describe the parameters that were utilized.

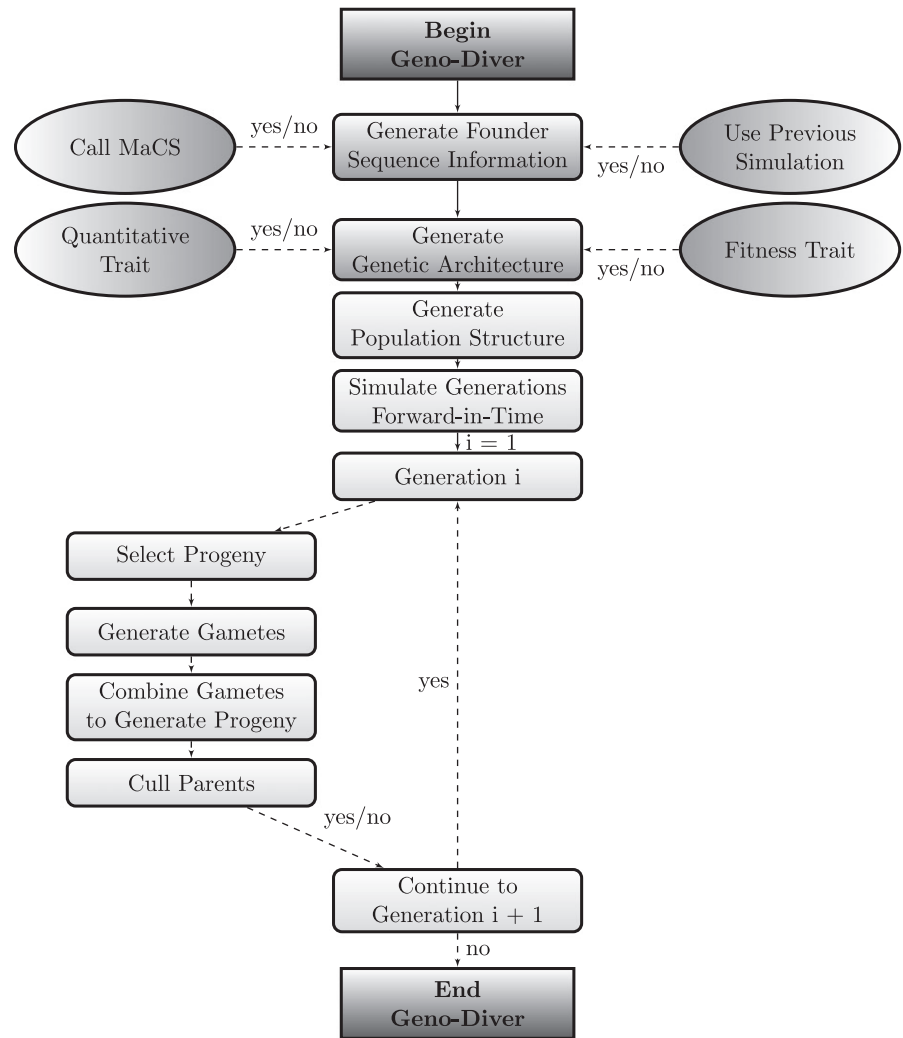
## 1.1 | Software overview

A schematic overview of Geno-Diver simulation strategy is outlined in Figure 1. The simulation can be split into two

steps: (i) generate the historical population using coalescence based methods and, (ii) generate new individuals for a given selection and mating scenario across multiple generations using a forward-in-time strategy. The program calls the Markovian Coalescence Simulator (MaCS; Chen, Marjoram, & Wall, 2009) to set up the historical population. The MaCS software makes it possible to generate a wide range of population scenarios in terms of the size and structure of the ancestral population across time. Once the historical population is generated, the software builds the genetic architecture which can include quantitative and/or fitness traits. Following the first step, new individuals are generated using a forward-in-time approach. A wide range of recent population structures, selection procedures and mating plans can be utilized to closely mimic livestock breeding populations.

### 1.1.1 | Step 1: Generating historical population and genetic architecture

The software is initiated by reading a file that specifies the parameters used in the simulation. To simplify the program initialization, only a small portion of the parameters are required to be explicitly set, while the remainder are set to default values. Once the parameters are read in the program internally calls MaCS, which simulates a sample of haplotypes with sequence information for a user-specified number of chromosomes. We have chosen to employ a coalescence simulator (specifically MaCS) in this step due to the flexibility of the approach in generating haplotype sequences for a wide range of population scenarios in terms of the size and structure of the ancestral population. Furthermore, the coalescent approach is a computationally more parsimonious approach that scales better with the use of sequence data. Within Geno-Diver, the historical population is simulated assuming that all mutations are neutral with respect to either quantitative or fitness traits. The assumption of mutations being neutral is a common approach to initialize the founder population in simulation programs to generate a genome with the desired level of linkage disequilibrium (LD; Faux et al., 2016; Pérez-Enciso & Legarra, 2016). In Geno-Diver, the user can specify a custom population history and structure following MaCS protocol. Additionally, five default scenarios can be directly generated. The scenarios mimic different population histories and result in the founder population having varying amounts of LD between genetic markers across the genome. We have chosen the default scenarios to resemble options specified by AlphaSim (Faux et al., 2016), as they represent LD patterns typical of agricultural species. After haplotype sequences are generated for each chromosome, founder individuals are generated by randomly sampling sequence haplotypes without replacement from the complete set.



**FIGURE 1** Overview of simulation program

Given the founder population, SNP derived from sequence information are utilized to generate a SNP panel with a user-specified density. The marker panel is generated by assigning SNP randomly across the genome. The user can alter the number of markers within a chromosome, as well as a minimum minor allele frequency (MAF) threshold for a SNP to be included in the marker panel across all chromosomes. The QTL are generated by randomly assigning a QTL position across the genome. The user can specify the minimum (maximum) allele frequency for the quantitative (fitness) trait. Therefore, the user can generate a SNP chip with varying amounts of common versus rare variants, thus allowing for the possibility of a QTL having a lower MAF than markers on the SNP chip. Within the software, a quantitative trait is one that impacts the phenotype of an individual and is assumed to follow an additive model. Alternatively, a fitness trait is one that impacts the ability of an individual to survive to breeding age and is assumed to follow a multiplicative model. For the quantitative trait, both additive and dominance effects can be specified for a QTL; although at the current time, epistasis cannot be modelled. Each additive effect is

generated from a gamma distribution with a user defined shape and scale parameter. The effect generated from the gamma distribution has an equal chance of being positive or negative. The dominance effects are calculated following the methods of Wellmann and Bennewitz (2012). First, the degree of dominance is sampled from a normal distribution and then the associated dominance effect is calculated as the absolute value of the additive effect times the degree of dominance. The user can alter the parameters that specify the distribution that generates additive and dominance effects. The additive and dominance variance for the quantitative trait is scaled to achieve a user-specified narrow ( $h^2$ ) and broad ( $H^2$ ) sense heritability. Once additive and dominance effects are determined, the phenotype for an individual<sub>*i*</sub> ( $y_i$ ) is generated as:

$$y_i = \mu + \sum_q^{nQTL} (\gamma_i a_q + \delta_i d_q) + e_i,$$

where  $\mu$  is the general mean,  $nQTL$  is the number of QTL,  $\gamma$  is the genotype (i.e., 0 for the homozygote; 2 for the alternative homozygote; 1 for the heterozygote) for individual<sub>*i*</sub> at QTL<sub>*q*</sub>,  $a$  is the additive substitution effect for QTL<sub>*q*</sub>,

$\delta$  is the dominance genotype (i.e., 1 for heterozygote; 0 for either homozygote) for individual<sub>*i*</sub> at QTL<sub>*q*</sub>,  $d$  is the dominance effect for QTL<sub>*q*</sub> and  $e_i$  is a normal residual ( $e \sim N(0, (1-H^2))$ ).

For each fitness trait locus (FTL), the traditional relative fitness parameterization is employed (Falconer & Mackay, 1996). Under the relative fitness parameterization, the favourable homozygote genotype has a fitness value of 1 while the heterozygote and homozygotes have fitness values relative to the favoured genotype. Relative fitness is parameterized by two coefficients: the selection coefficient ( $s$ ) and the dominance coefficient ( $h$ ). The  $s$  measures how much worse the unfit allele is, compared to the fittest allele. The  $h$  measures the degree of dominance that the heterozygote displays in terms of the reduced fitness compared to the unfit homozygote. The unfavourable homozygote genotype has a value of  $1-s$  and the heterozygote genotype has a fitness value of  $1-hs$ . Lethal (i.e., high probability of not making it to breeding age) and/or sublethal (i.e., high probability of making it to breeding age) FTL can be generated with a varying number belonging to each class. The FTL are split into lethal and sublethal to provide the user with a more straightforward approach of characterizing the consequences of a given scenario on lethal versus sublethal FTL across time. Furthermore, the user can revert back to a single fitness category by setting one of the categories to zero. The fitness of an individual ( $v_i$ ) is defined here as the ability to survive to breeding age is calculated using a multiplicative model, as follows:

$$v_i = \prod_q^{n_{FTL}} v_q$$

where  $n_{FTL}$  is the number of FTL and  $v$  is the relative fitness value for individual<sub>*i*</sub> at FTL<sub>*q*</sub>. For each progeny, a random number is obtained from a standard uniform distribution  $U(0,1)$  and compared with  $v_i$  to determine whether the individual survived to breeding. The individual will survive to breeding age if the random number is less than  $v_i$  and will not survive if the random number is greater than or equal to  $v_i$ . Therefore, individuals with a fitness value close to one have a higher probability of surviving to breeding age compared to individuals with a fitness value close to zero. As a consequence, the number of surviving offspring may vary among families, although each family generates the user-specified number of offspring. The method utilized here to calculate the fitness of an individual and whether a gamete survived to breeding age is the same employed by Wang and Hill (1999). The selection coefficient and degree of dominance are simulated from a gamma and normal distribution, respectively, with the possibility for the user to alter the parameters that specify the distributions. In the paper by Wang and Hill (1999), an exponential distribution was employed to

generate fitness effects and effects were scaled to obtain a user-defined mean population fitness. This approach was not utilized in the current study to obtain covariances between quantitative and fitness traits in a more straightforward manner.

Covariance between the two traits is generated based on the trivariate reduction method, which is a common method to generate covariance between two distributions that are non-normal (Sarabia & Gómez-Déniz, 2008). For example, to generate a given correlation ( $\rho$ ) between a Gamma<sub>1</sub> ( $shape_1, scale_1$ ) and Gamma<sub>2</sub> ( $shape_2, scale_2$ ), the following steps are taken:

1. Generate  $Y_1 \sim \text{gamma}(shape_1 - \rho\sqrt{shape_1shape_2}, 1)$ .
2. Generate  $Y_2 \sim \text{gamma}(shape_2 - \rho\sqrt{shape_1shape_2}, 1)$ .
3. Generate  $Y_3 \sim \text{gamma}(\rho\sqrt{shape_1shape_2}, 1)$ .
4. Generate value for gamma1:  $scale_1(Y_1 + Y_3)$ .
5. Generate value for gamma2:  $scale_2(Y_2 + Y_3)$ .

Rank correlation is employed in the program as the sampled distributions are not normal. The trivariate reduction method only allows the correlation to be positive and bounded between 0 and  $\frac{\min(shape_1, shape_2)}{\sqrt{shape_1, shape_2}}$ . Consequently, based on the positive correlation, selecting high values for the quantitative trait will result in the two traits being antagonistic and changing the selection direction for the quantitative trait will alter the interpretation. Future versions should allow for greater flexibility in the available sampling distributions as well as in the ability to simulate more than 1 trait within quantitative and fitness traits. The default parameterization for each effect is outlined in the software manual.

### 1.1.2 | Step 2: simulate individuals forward-in-time

The second step of the simulation generates new individuals across generations according to a specified selection and mating scenario. New mutations can be generated within this step and each mutation follows the infinite-site model (Kimura, 1969), which assumes that a new mutation always results in a new polymorphism instead of occurring at a site where a polymorphism already exists. The number of mutation events is drawn from a Poisson distribution with mean equal to the nucleotide length times the mutation rate. Furthermore, quantitative trait mutation effects were multiplied by the scaling factor that was utilized in step 1 to achieve the given narrow and broad sense heritability in the founder population. This was performed to keep the effect of new mutations on a similar scale as mutation events that were generated in the founder generation. Mutations that impact fitness were not multiplied by any scaling factor. The number of recombination events is

sampled from a Poisson distribution with a mean set at 1 Morgan. Options are available to alter the frequency of occurring at a given genomic location to investigate the effect of recombination rate on genomic diversity and inbreeding load.

Selection and culling can be implemented using different criteria and can be based on the phenotype, true genetic value or the estimated breeding value (EBV) of an individual. The EBV is an estimate of the additive genetic value of an individual and is obtained using best linear unbiased prediction via an animal model based either on pedigree or genomic relationship matrices. Genomic-based relationships and their inverse are calculated with efficient algorithms that are scalable to large marker panels (i.e., in excess of 500,000 SNP; Aguilar, Misztal, Legarra, & Tsuruta, 2011). The mating design part of the simulation can be based on random mating, avoidance of animals above a certain relationship, or can be optimized by minimizing inbreeding using either pedigree or genomic-based relationship matrices. Optimization of inbreeding is carried out using the simulated annealing method (Kirkpatrick, Gelatt, & Vecchi, 1983). Complex population structures can be generated by manipulating the differential contribution of gametes to the next generation and the minimum number of siblings selected within a family. Many data files and summary statistics for each generation are generated, and a description of the files is outlined in Figure 2. In order to provide the user with a snapshot overview for a given simulation scenario, multiple summary statistics are produced and include the phenotypic and genetic performance, genetic diversity, inbreeding load, LD-decay and QTL and/or FTL frequency for each generation. A full description of the summary statistics and data files generated is provided in the user manual (<https://github.com/jeremyhoward/Geno-Diver>).

## 1.2 | Implementation and applications

The program is written in C++11 and is accompanied with executable files for the Linux platform. Geno-Diver requires the Eigen and Intel MKL libraries and is multi-threaded to make the simulation computationally efficient. For example, computing time after sequences were simulated for a 50 K SNP panel on 550 individuals (50 sires & 500 dams) with 1 progeny per mating pair undergoing EBV selection based on pedigree or genomic selection for 10 generations was 7.52 and 12.48 min, respectively. The computations were performed using a Dell Precision T3500 with two Intel Xeon X5482 3.20 GHz processors and 24 GB of RAM utilizing four threads. To limit the space requirements for sequence and output files, multiple options can be utilized by the program, including saving genotype information only for a user-specified generation number(s) or using sequence information from a previous scenario. In the current version of the software, the number of generations allowed for a quantitative trait under selection is tailored to a medium time frame (~30 generations). Future versions should allow a longer horizon. It should be noted that if the genetic architecture only involves a fitness trait, the number of generations that can be run is much greater, although the users should limit the amount of generations that are saved. A comprehensive user manual that outlines how to run the program and a description of each parameter is available at <https://github.com/jeremyhoward/Geno-Diver>.

### 1.2.1 | Benchmark with another simulation program

We have compared results from Geno-Diver to a previously developed simulation program. We have produced a

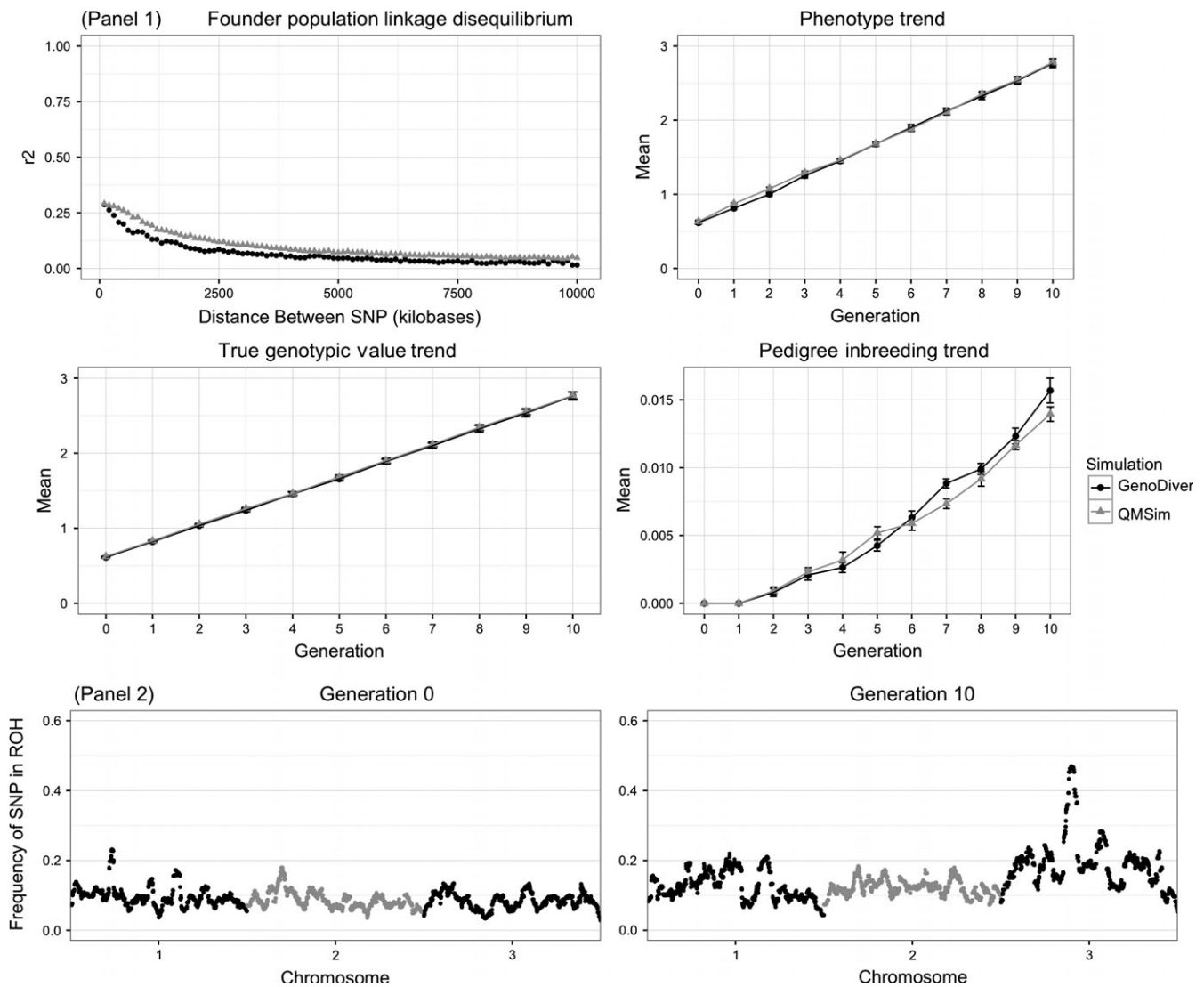
**FIGURE 2** Output files generated from Geno-Diver

Simulated Data Files	Summary Statistics Files
<ul style="list-style-type: none"> <li>• Log file with description of simulation.</li> <li>• Information on animals that died due to fitness.</li> <li>• Marker map.</li> <li>• Marker, QTL and FTL genotypes for each individual.</li> <li>• QTL and FTL information.</li> <li>• Master file with a large amount of information on each individual.</li> <li>• Pedigree File.</li> <li>• Marker Genotype File.</li> </ul>	<ul style="list-style-type: none"> <li>• Summary statistics by generation on number of QTL and FTL segregating.</li> <li>• Summary of Inbreeding metrics by generation based on genome-wide inbreeding levels and fitness-related metrics.</li> <li>• Summary of phenotypic and genetic performance by generation.</li> <li>• Marker LD decay by generation.</li> <li>• QTL and FTL frequency across generations</li> </ul>

comparable scenario using both the QMSim simulation program (Sargolzaei & Schenkel, 2009) and Geno-Diver. For Geno-Diver, the founder population was generated using the “Ne100\_Scen1” historical population parameter scenario. To make the two founder populations similar, the historical population in QMSim was generated for 2001 generations with a constant size of 2,000 individuals for 1,000 generations followed by a gradual decrease in population size from 2,000 to 100 across 1,000 generations. The last generation (i.e., 2001) was increased to 1,400 individuals to provide enough individuals for the first generation of the recent population. Across both simulation programs, the population consisted of 300 females and 50 males. The replacement rate for females and males was 0.1 and 0.5, respectively. New progeny was selected and parents culled based on their true breeding value and then mated at random. Selection was conducted for 10 generations. The

number of QTL across three chromosomes was set at 150 and was generated based on an additive model (i.e., dominance variance was 0). The narrow-sense heritability was 0.35. Each chromosome was assumed to be 150 Mb and contained 4,000 markers. Within each simulation program, the scenario was replicated 10 times. The map for QMSim is based on cM; therefore, the cM map positions were converted to nucleotide position by multiplying them by 1 Mb. Within each simulation, 10 replicates were generated.

The founder population LD-decay for one of the replicates and average phenotype, true breeding value and pedigree inbreeding trends across replicates based on either QMSim or Geno-Diver are outlined in Panel 1 of Figure 3. Across both simulation programs, the LD-decay in the founder population was similar, such that there are high levels of short range LD and the LD-decays as the distance



**FIGURE 3** Comparison of Geno-Diver results across generations with a previously developed simulation program (Panel 1) and change in autozygosity across the genome from the founder generation to generation 10 (Panel 2)

between the markers increases. Thus, the use of coalescence-based methods or randomly mating for many generations based on forward-in-time methods to initialize the founder population (i.e., QMSim) gives rise to similar results. Furthermore, the mean phenotype, true breeding value and pedigree inbreeding value across generations were similar across the simulation programs. Variation in the autozygosity frequency across the genome due to multiple factors including genetic drift and selection is an important aspect of genomes in real livestock populations. To determine the autozygosity levels and the change across generations, the frequency of a SNP occurring in a contiguous run of homozygosity (ROH) with a length of at least 5 Mb was calculated as outlined by Howard et al. (2016) for one of the replicates. As illustrated in Panel 2 of Figure 3, the frequency of a SNP being in an ROH is heterogeneous across the genome and increased from the unselected founder generation.

### 1.2.2 | Applications of geno-diver

Geno-Diver can be used to gain an understanding of a diverse array of topics, a few of which are outlined in the next two examples. Both examples involve two active areas of research and include understanding the effectiveness of different models to predict the phenotype or the impact of the FTL strength of selection on its frequency across generations. More emphasis was placed on how the simulation scenarios were set up, and the parameter values that were utilized to fully describe how complex scenarios can be generated in Geno-Diver.

### 1.2.3 | Impact of including dominance to predict the phenotype

The advent of genomic selection has led to a renewed interest in including dominance effects in genetic evaluations to more accurately predict the phenotype of an individual. Thus, a simulation scenario was developed to determine the impact of predicting the phenotype with or without the inclusion of dominance effects across four common Bayesian models. The Geno-Diver simulation program was utilized to generate the training and validation populations. The founder population was generated using two scenarios with varying degrees of short range LD and included the “Ne1000” and “Ne70” historical population parameter scenario. For each scenario, the genome had a total of five chromosomes that were 150 Mb long containing 4,000 and 100 markers and QTL, respectively. The narrow and broad sense heritabilities were scaled to a value of 0.35 and 0.40, respectively. The trait was simulated to display directional dominance, and the majority of dominance effects displayed partial dominance as compared to

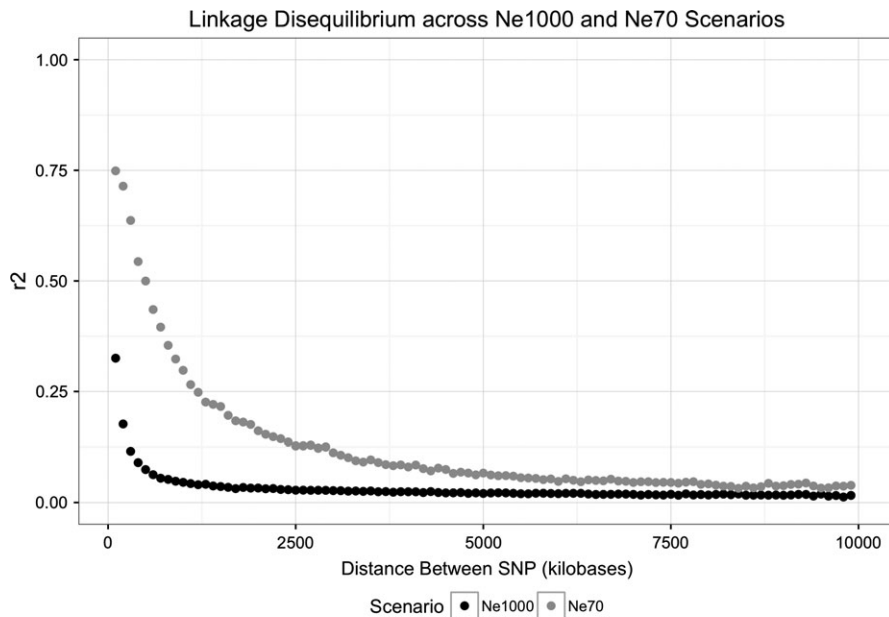
over-dominance. After the founder population and genetic architecture of the trait were generated, a selection scenario mimicking a livestock population was implemented for 10 generations. Within each generation, the population consisted of 50 males and 600 females. The replacement rate for both female and male parents was 20%. The EBV were generated from an animal model based on pedigree information.

Progeny with a high EBV was selected to serve as parents for the next generation. Animals were mated at random, and one progeny was produced for each mating pair. Individuals from generation 7 to 9 were the training population and animals from generation 10 were the validation population. Within each LD scenario, 10 replicates were generated. The Bayesian models investigated include Bayesian Ridge Regression (BRR), Bayesian Lasso (BL; Park & Casella, 2008), BayesB (Meuwissen, Hayes, & Goddard, 2001) and BayesC (Habier, Fernando, Kizilkaya, & Garriick, 2011). The marker effects across the four models were estimated using the “BGLR” package in R (Pérez & de los Campos, 2014). For the additive effect of a SNP, the genotypes were coded as 0 for the homozygote, 2 for the other homozygote and 1 for the heterozygote. For the dominance effect of a SNP, the genotypes were coded as 1 for the heterozygote and 0 for either homozygote. A total of 55,000 iterations were run with the first 5,000 discarded as burn-in and a thinning rate of 5. For each individual, the estimated breeding value (EBV; i.e., only additive) and genotypic value (EGV; i.e., additive + dominance) were generated by multiplying the additive and when applicable, the dominance genotype for an individual by the estimated marker effect and summing across all markers.

The correlations between the EBV and EGV and phenotype across the four Bayesian models are presented in Figure 4. In general, the inclusion of dominance effects across the 4 models resulted in a small improvement in the ability to predict the phenotype. Furthermore, greater improvement was seen in the scenario with increased amounts of LD (i.e., “Ne70” scenario), which agrees with previous results across multiple traits and species. Previous reports have found the inclusion of dominance effects (i.e., EGV) to predict the phenotype of an individual results in little improvement in comparison with relying solely on the additive genotypic value (i.e., EBV) of an individual (Lopes, Bastiaansen, Janss, Knol, & Bovenhuis, 2016; Sun, VanRaden, O’Connell, Weigel, & Gianola, 2013).

### 1.2.4 | Effectiveness of purging sublethal and lethal fitness trait loci

The ability of a population to purge deleterious alleles over time is dependent on the degree by which they reduce the fitness relative to the most-fit genotype. Such purging will



Correlation with phenotype across models and scenarios

Model	Ne1000		Ne70	
	EBV (SE)	EGV (SE)	EBV (SE)	EGV (SE)
BRR	0.413 (0.013)	0.414 (0.014)	0.459 (0.025)	0.482 (0.024)
BL	0.415 (0.013)	0.418 (0.014)	0.462 (0.025)	0.484 (0.024)
BayesB	0.425 (0.014)	0.433 (0.015)	0.467 (0.026)	0.496 (0.022)
BayesC	0.414 (0.013)	0.416 (0.014)	0.461 (0.026)	0.484 (0.024)

**FIGURE 4** Impact of including dominance effects on predicting the phenotype across four Bayesian models

be most efficient against lethal mutations (i.e., high selection coefficient). However, this is unlikely to be as effective against the mildly deleterious mutations. Thus, when considering the genetic background of fitness, the effects of rare recessive deleterious (or even lethal) genes are most striking. Slightly detrimental mutations, however, have a much larger fixation probability and collectively lead to a substantial reduction in fitness (Meuwissen & Woolliams, 1994). Therefore, two simulation scenarios were developed to understand the frequency and number of segregating lethal and sublethal FTL across generations within a small population. To understand the impact of genetic drift on the number of FTL across time, one scenario had the same maximum allele frequency (0.06) in the founder population across both lethal and sublethal FTL. The other scenario had lethal FTL at a lower maximum allele frequency (0.03) than sublethal FTL (0.08) in the founder population.

Across both scenarios, the founder population was generated using the “Ne100\_Scen2” historical population parameter scenario. The genome had a total of three chromosomes that were each 150 Mb long containing 3,000 and 150, markers and QTL, respectively. The narrow and broad sense heritability was scaled to a value of 0.35 and 0.40, respectively. The lethal (sublethal) selection coefficient was generated from a gamma with a shape and scale parameter of 3.0 (0.2) and 0.1 (0.2), respectively. Furthermore, the lethal (sublethal) degree of dominance was

generated from a normal distribution with a mean of 0.02 (0.30) and standard deviation of 0.05 (0.1). Within each lethal and sublethal category, 100 FTL was generated within each chromosome and placed randomly across the genome. Following these parameterizations, the average ( $\pm SD$ ) selection coefficient for the lethal and sublethal FTL across both scenarios were 0.30 ( $\pm 0.01$ ) and 0.04 ( $\pm 0.01$ ), respectively. The average ( $\pm SD$ ) degree of dominance for the lethal and sublethal FTL across both scenarios were 0.05 ( $\pm 0.01$ ) and 0.30 ( $\pm 0.01$ ), respectively. Previous results have found that the heterozygote genotype for FTL with a large selection coefficient (i.e., lethal or nearly lethal) has its fitness reduced by 2%, while sublethal mutations tend to be much closer to semi-dominance, with respect to their fitness effects (Charlesworth, 2012). After the founder population and genetic architecture of the trait were generated, a total of 50 generations were simulated. Within each generation, the population consisted of 50 males and 250 females. The replacement rate for female and male parents was 20%. New progeny was selected, and parents were culled based on their phenotype. Parents were mated at random, and 1 offspring was produced per mating. Within each scenario, a total of 20 replicates were generated.

The mean number of segregating lethal and sublethal FTL and the difference in the mean frequency of lethal and sublethal FTL from the founder mean frequency across

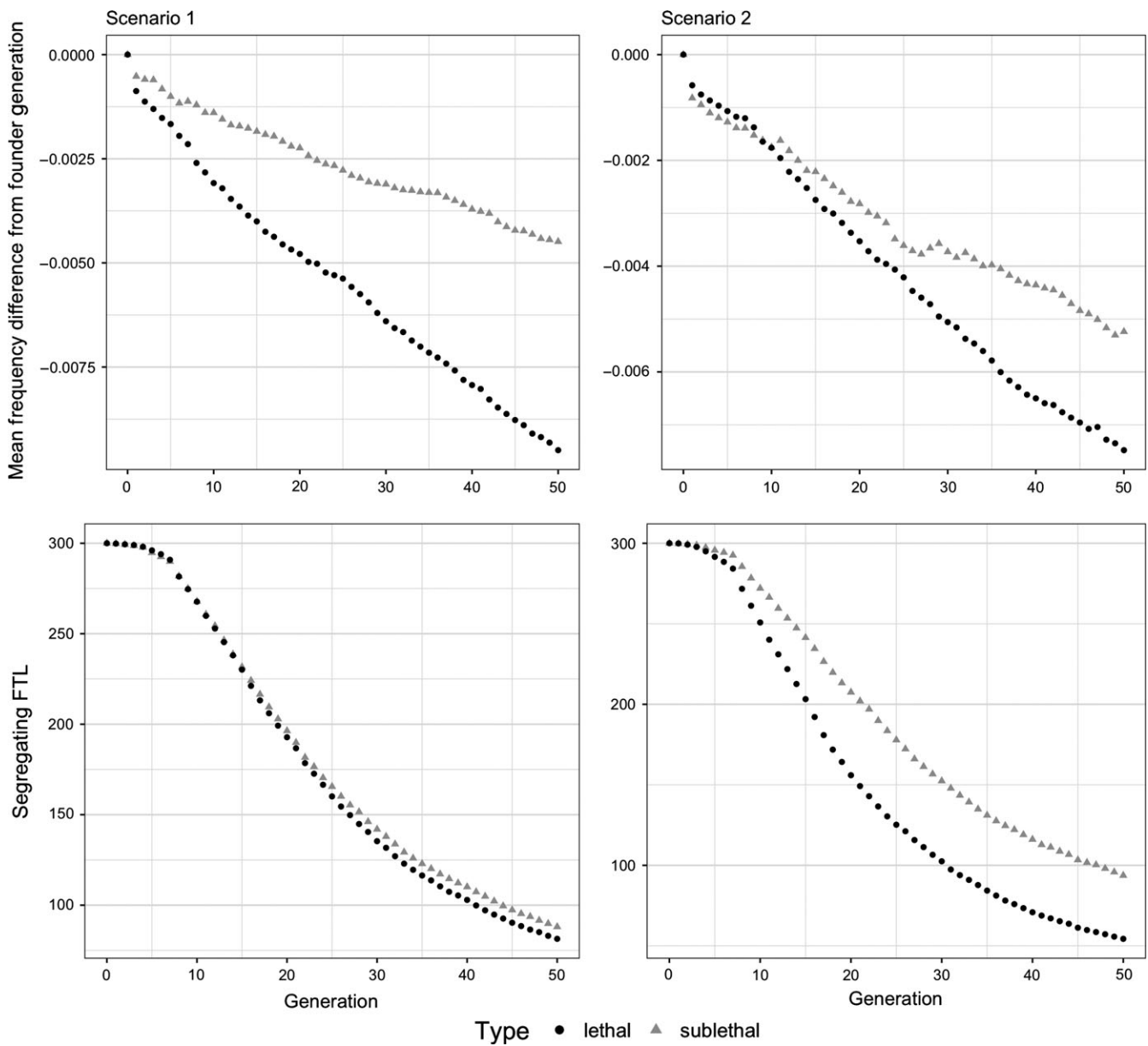


generations is outlined in Figure 5. Under the scenario of equal maximum frequency, the effect of drift is equal across both lethal and sublethal FTL and the number of segregating FTL is reduced more severely for lethal than for sublethal FTL. Furthermore, the change in mean frequency of the segregating FTL is reduced to a greater degree for the lethal than sublethal FTL. It is expected that at equilibrium lethal or nearly lethal FTL should be at a lower frequency than sublethal FTL, and therefore, the effect of drift is greater for lethal compared to sublethal FTL. Thus, the second scenario is more than likely closer to reality and as outlined in Figure 4 where trends are similar to scenario 1, but a greater number of lethal FTL are

lost compared to sublethal FTL due to a greater degree of drift occurring for the first category.

## 2 | DISCUSSION

We have presented a simulation software capable of efficiently simulating complex traits involving quantitative and/or fitness components and implementing multiple selection and mating strategies utilizing pedigree or genomic information. The program combines coalescence and forward-in-time methods and was designed to be flexible and simple to implement. Previous simulations programs



**FIGURE 5** Impact of purging lethal versus sublethal fitness trait loci across generations for two scenarios<sup>1</sup>. <sup>1</sup>Scenario 1 refers to lethal and sublethal fitness trait loci have the same maximum allele frequency of 0.6 in the founder population; Scenario 2 refers to lethal fitness trait loci being at a lower maximum allele frequency (0.02) in comparison with sublethal fitness trait loci (0.08)

have assumed the genetic architecture is a quantitative trait with only additive (Sargolzaei & Schenkel, 2009) or additive and dominance effects (Faux et al., 2016; Pérez-Enciso & Legarra, 2016). Alternatively, simulation programs have been developed that allow for fitness characters to be simulated and include FREGENE (Chadeau-Hyam et al., 2008) and SLiM (Messer, 2013). To the best of our knowledge, Geno-Diver is the only software that can simulate both trait categories along with a covariance between them. Geno-Diver can provide a platform to facilitate fundamental research on how to maximize the fitness of livestock populations while increasing yield and efficiency traits. Currently, Geno-Diver has been primarily developed to understand the impact of fitness mutations in a livestock breeding program. As a result, other methods such as optimum contribution, gene editing and the use of advanced reproductive technologies have not been fully introduced into the software, although we plan to introduce these and other options with later versions.

Methods to manage population diversity have relied for the most part on genome-wide estimates of inbreeding, either based on pedigree or genomic information. These metrics disregard the fact that genetic diversity and inbreeding depression are heterogeneous across the genome. The heterogeneity of genetic diversity across the genome has been recently discussed by Jiménez-Mena, Hospital, and Bataillon (2016) and Howard et al. (2016). Jiménez-Mena et al. (2016) found the effective population size ( $N_e$ ) to vary considerably across the genome ( $N_e$ : 40–250) in a Danish Holstein population, implying the accumulation of inbreeding is heterogeneous across the genome. Heterogeneous accumulation of inbreeding results in certain regions of the genome being inbred at a faster rate than others. Alternative metrics to manage livestock populations at the genomic level and their impact on the fitness and genetic value of the population will be increasingly evaluated in the future. Geno-Diver can facilitate in measuring and understanding the implications alternative genomic measures in evaluating or comparing genetic gain and overall fitness across generations.

## REFERENCES

- Aguilar, I., Misztal, I., Legarra, A., & Tsuruta, S. (2011). Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal Breeding and Genetics*, *128*, 422–428. <https://doi.org/10.1111/j.1439-0388.2010.00912.x>
- Chadeau-Hyam, M., Hoggart, C. J., O'Reilly, P. F., Whittaker, J. C., de Iorio, M., & Balding, D. J. (2008). Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, *9*, 1–11. <https://doi.org/10.1186/1471-2105-9-364>
- Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked sites. *Genetics*, *190*, 5–22. <https://doi.org/10.1534/genetics.111.134288>
- Chen, G. K., Marjoram, P., & Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Research*, *19*, 136–142. <https://doi.org/10.1101/gr.083634.108>
- Clark, S. A., Kinghorn, B. P., Hickey, J. M., & van der Werf, J. H. J. (2013). The effect of genomic information on optimal contribution selection in livestock breeding programs. *Genetics Selection Evolution*, *45*, 44. <https://doi.org/10.1186/1297-9686-45-44>
- Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de Los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*, *193*, 347–365. <https://doi.org/10.1534/genetics.112.147983>
- De los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-Genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*, 327–345. <https://doi.org/10.1534/genetics.112.143313>
- Falconer, D.S., & Mackay, T.F.C. (1996) *Introduction to quantitative genetics*, 4th ed. New York, NY: Longman Scientific and Technical.
- Faux, A. M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., ... Hickey, J. M. (2016). AlphaSim: Software for breeding program simulation. *Plant Genome*, *9*(3), 1–14. <https://doi.org/10.3835/plantgenome2016.02.0013>
- Gómez-Romano, F., Villanueva, B., Fernández, J., Woolliams, J. A., & Pong-Wong, R. (2016). The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. *Genetics Selection Evolution*, *48*, 2. <https://doi.org/10.1186/s12711-015-0172-y>
- Habier, D., Fernando, R., Kizilkaya, K., & Garrick, D. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*, 186. <https://doi.org/10.1186/1471-2105-12-186>
- Henryon, M., Berg, P., & Sorensen, A. C. (2014). Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. *Livestock Science*, *166*, 38–47.
- Howard, J. T., Tiezzi, F., Huang, Y., Gray, K. A., & Maltecca, C. (2016). Characterization and management of long runs of homozygosity in parental nucleus lines and their associated crossbred progeny. *Genetics Selection Evolution*, *48*, 91. <https://doi.org/10.1186/s12711-016-0269-y>
- Jiménez-Mena, B., Hospital, F., & Bataillon, T. (2016). Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. *Conservation Genetics Resources*, *8*, 35–41. <https://doi.org/10.1007/s12686-015-0508-5>
- Jonas, E., & de Koning, D. J. (2015). Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Frontiers in Genetics*, *6*, 49. <https://doi.org/10.3389/fgene.2015.00049>
- Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, *61*, 893–903.
- Kirkpatrick, S., Gelatt, C. D. Jr, & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 672–680. <https://doi.org/10.1126/science.220.4598.671>
- Lopes, M. S., Bastiaansen, J. W. M., Janss, L., Knol, E. F., & Bovenhuis, H. (2016). Genomic prediction of growth in pigs based on a model including additive and dominance effects. *Journal of Animal Breeding and Genetics*, *133*, 180–186. <https://doi.org/10.1111/jbg.12195>

- McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? *Evolutionary Applications*, *7*, 999–1007. <https://doi.org/10.1111/eva.12193>
- Messer, P. W. (2013). SLiM: Simulating evolution with selection and linkage. *Genetics*, *194*, 1037–1039. <https://doi.org/10.1534/genetics.113.152181>
- Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829.
- Meuwissen, T. H. E., & Woolliams, J. A. (1994). Effective sizes of livestock populations to prevent a decline in fitness. *TAG. Theoretical and Applied Genetics*, *89*, 1019–1026. <https://doi.org/10.1007/BF00224533>
- Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of American Statistical Association*, *103*, 681–686.
- Pérez, P., & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*, 483–495. <https://doi.org/10.1534/genetics.114.164442>
- Pérez-Enciso, M., & Legarra, A. (2016). A combined coalescence gene-dropping tool for evaluating genomic selection in complex scenarios (ms2gs). *Journal of Animal Breeding and Genetics*, *133*, 85–91. <https://doi.org/10.1111/jbg.12200>
- Rodríguez-Ramilo, S. T., García-Cortés, L. A., & de Cara, M. Á. R. (2016). Artificial selection with traditional or genomic relationships: Consequences in coancestry and genetic diversity. *Frontiers in Genetics*, *6*, 127. <https://doi.org/10.3389/fgene.2015.00127>
- Sarabia, J. M., & Gómez-Déniz, E. (2008). Construction of multivariate distributions: A review of some recent results (with discussion). *SORT*, *32*, 3–36. <https://doi.org/10.1016/j.insmathco.2008.09.002>
- Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*, 680–681. <https://doi.org/10.1093/bioinformatics/btp045>
- Sun, C., VanRaden, P. M., O'Connell, J. R., Weigel, K. A., & Gianola, D. (2013). Mating programs including genomic relationships and dominance effects. *Journal of Dairy Science*, *96*, 8014–8023. <https://doi.org/10.3168/jds.2013-6969>
- VanRaden, P. M., Olson, K. M., Null, D. J., & Hutchison, J. L. (2011). Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *Journal of Dairy Science*, *94*, 6153–6161. <https://doi.org/10.3168/jds.2011-4624>
- Wang, J., & Hill, W. G. (1999). Effect of selection against deleterious mutations on the decline in heterozygosity at neutral loci in closely inbreeding populations. *Genetics*, *153*, 1475–1489.
- Wellmann, R., & Bennewitz, J. (2012). Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genetical Research*, *94*, 21–37. <https://doi.org/10.1017/S0016672312000018>
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*, 565–569. <https://doi.org/10.1038/ng.608>

**How to cite this article:** Howard JT, Tiezzi F, Pryce JE, Maltecca C. Geno-Diver: A combined coalescence and forward-in-time simulator for populations undergoing selection for complex traits. *J Anim Breed Genet*. 2017;134:553–563. <https://doi.org/10.1111/jbg.12277>