SNPscan breeder – a computer program to test genomic tools in breeding programs

Bernd Degen[1]* and Niels A. Müller[1]

[1] Thünen Institute of Forest Genetics, Sieker Landstrasse 2, 22927, Grosshansdorf, Germany

*Correspondence author: Bernd Degen, e-mail: bernd.degen@thuenen.de

Abstract

SNPscan breeder is a software that enables the simulation of breeding programs using simulated individual whole genome data, different genetic architectures of a trait of interest, different mating designs and different selection criteria, i.e. phenotypes, breeding values from progeny tests, marker-assisted selection (MAS) and genomic selection (GS). The impact of breeding population size, mating design, selection intensity, genetic architecture, heritability and selection criteria on genetic gains, kinship, inbreeding and genetic diversity can be evaluated to optimize the breeding program. A special feature is the possibility for post-hoc analysis of different strategies to identify causal SNPs and allele effects within the frame of genome-wide association studies (GWAS). The proportion of true and false positive SNPs and the correlation of estimated and true allelic effects can be measured and the overall impact of their use for MAS on the success of the breeding program can be tested.

Keywords: *gBLUP, genetic gain, genomic selection, GWAS, mating design, marker-assisted selection, stochastic simulation*

Introduction

Since decades, computer simulations are used to help breeders to make decisions on the choice of individuals for the next breeding cycle and on the different options for the breeding strategy (Bellmann and Ahrens 1966, Sun et al. 2011). Although, there are quite a few simulation tools and R-packages available to analyse different breeding strategies such as AlphaSimR (Gaynor et al. 2021), Xsim (Chen et al. 2022) and ADAM-Plant (Liu et al. 2019), there is still a gap for a user friendly windows program that enables simulations of simple genome-wide SNP data, implements user-defined genetic architectures and offers a broad set of selection criteria for forward simulations in breeding programs. Such functionalities are essential to optimise sample designs for the number of individuals and SNPs needed for accurate genomic predictions and marker-assisted selection based on GWAS results. Our program *SNPscan breeder* aims to fill this gap. In an example we use the program to compare the impact of different selection criteria in a simple tree breeding program.

Program description

With *SNPscan breeder* the user defines distribution parameters for the generation of genomes and the genetic architecture of the trait. Then the program generates individual genomes and phenotypes according to these parameters. The options "Mating" and "Selection" offer various alternatives for the mating design and the selection of parents for stochastic simulations of breeding cycles. The genomes of all individuals are stored in separate text files and can be aggregated and exported for further downstream analyses as files in phased HapMap-format (figure 1).

Chromosomes and SNPs

SNPscan breeder assumes diploid sets of chromosomes and cosexual individuals (monoecious or hermaphroditic). The user defines the number of chromosomes, the total number of SNPs, the genome size and the average number of crossing-overs per chromosome. The SNPs are equally distributed over all chromosomes. The chromosomes have equal sizes. E. g., a genome with 20 chromosomes and a total of 2 million SNPs

DOI:10.2478/sg-2023-0013
edited by the Thünen Institute of Forest Genetics

will have 100,000 SNPs per chromosome. The probability for a crossing-over is identical along the chromosomes. Other parameters to be specified by the user are the proportions of bi-allelic, tri-allelic and tetra-allelic SNPs as well as the distribution of frequencies of the common alleles at the SNPs. Re-sequencing data of different tree species, specifically beech (*Fagus sylvatica*), oak (*Quercus robur*) and ash (*Fraxinus excelsior*) are used as default values (Plomion et al. 2016, Sollars et al. 2017, Pfenninger et al. 2021).

Genetic architecture

The user defines the name of a trait, its mean phenotypic value in the founder or wild population, the variance of the phenotypes and the heritability of the trait. Further a parameter controlling the inbreeding depression on the phenotypes as a function of the inbreeding coefficient can be specified. The program uses a linear relationship between inbreeding coefficient and trait value reduction (Durel et al. 1996). The number of causal SNPs is specified and one of three alternative functions for the distribution of the allelic effects selected: a) negative exponential distribution, b) normal distribution, or c) uniform distribution.

Phenotypes

The phenotypes of each individual *i* (p_i) are computed as the sum of the mean phenotype of the population at the start ($\bar{p}$ = parameter of the model), the genetic value (g_i) and an environmental value (e_i) using the following formula:

$$p_i = \bar{p} + g_i + e_i$$

Genetic value (genomic breeding value)

The genetic values of each individual *i* (g_i) are calculated as the sum of all additive effects at each causal locus j (a_{ij}) plus a correction *m* used to centre the mean of all genetic values of the starting population to zero, multiplied with the scale factor *sf*:

$$g_i = \sum_{j=1}^{n} a_{ij} + m \; x \; sf$$

The scale factor (*sf*) is defined as: $sf = \frac{\sqrt{(\sigma_p^2 * h^2)}}{\sigma_{g-values}}$

σ_p^2 = variance of the phenotypes in the population (parameter of the model)

h^2 = heritability (parameter of the model)

$\sigma_{g-values}$ = standard deviation of the genetic values of the population at the start

Environmental values

Environmental values (e_i) are sampled from a normal distribution with a mean of zero and a standard deviation of environmental effects: $N(0,\sigma_{env})$, $\sigma_{env} = \sqrt{(\sigma_p^2 - h^2 * \sigma_p^2)}$

Mating design

Common mating designs of plant breeding programs (Eriksson et al. 2013) have been implemented in *SNPscan breeder*:

- Diallel
- Half-diallel
- Disconnected half-diallel
- Factorial matings (common tester)

Further the user can select "Random mating" and specify the number of top ranked seed contributors (N top females) and pollen donors (N top males).

Selection

The user can specify different options on how to select the parents for the next mating cycles:

"Phenotypes of adults": Adults are ordered according to their phenotypes and the defined proportion of individuals are selected.

"Phenotypes of progenies": For each adult a given number of offspring ("N progenies") are simulated. The mating is random among all adults. The mean phenotype of the offspring is then used to rank the adults (estimate the genomic breeding values; backwards selection).

"GWAS estimates of allele effects": This option is possible if a genome-wide association study (GWAS) has been performed on the simulation results and the estimated allelic effects stored. The program will ask for the according file. During the simulations, the breeding values of the adult individuals are estimated as the sum of the allelic effects at all identified associated SNPs.

"Genomic selection (gBLUP) single generation": *SNPscan breeder* uses the function "kin.blup" integrated in the R-package "rrBLUP" (Endelman 2011, R-Core-Team 2022). The algorithm computes "Predicted Genomic Breeding Values (PGBV)" for all individuals in the actual generation F using a selected proportion of phenotypes as training data and the kinship matrix of all individuals. The user can choose the number of SNPs used for the predictions. Optionally, the causal SNPs can be excluded. During the simulations, *SNPscan breeder* creates a subdirectory "R" in the project folder to store the input files, the R-script and the results.

"Genomic selection (gBLUP) cross generations": Again, the phenotypes of a selected proportion of individuals of generation F1 are used as a training population and the individuals of the generation F are used as the test population. The algorithm computes "Predicted Genomic Breeding Values (PGBV)" for all individuals in generation F using the phenotypes of generation F1 and the kinship-matrix of all individuals in both generations.

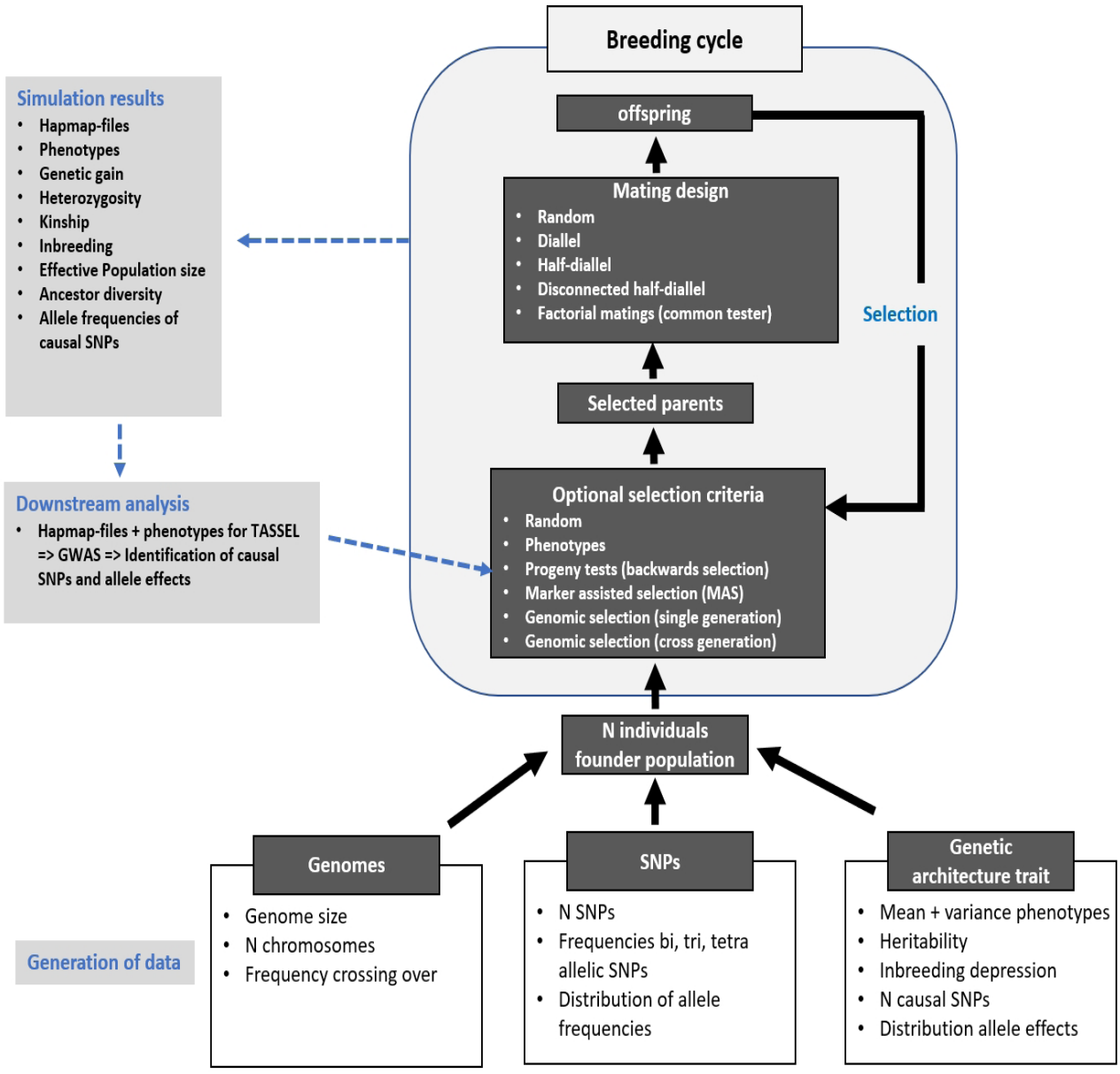

Figure 1
Scheme on the different elements and features of the simulation program *SNPscan breeder*

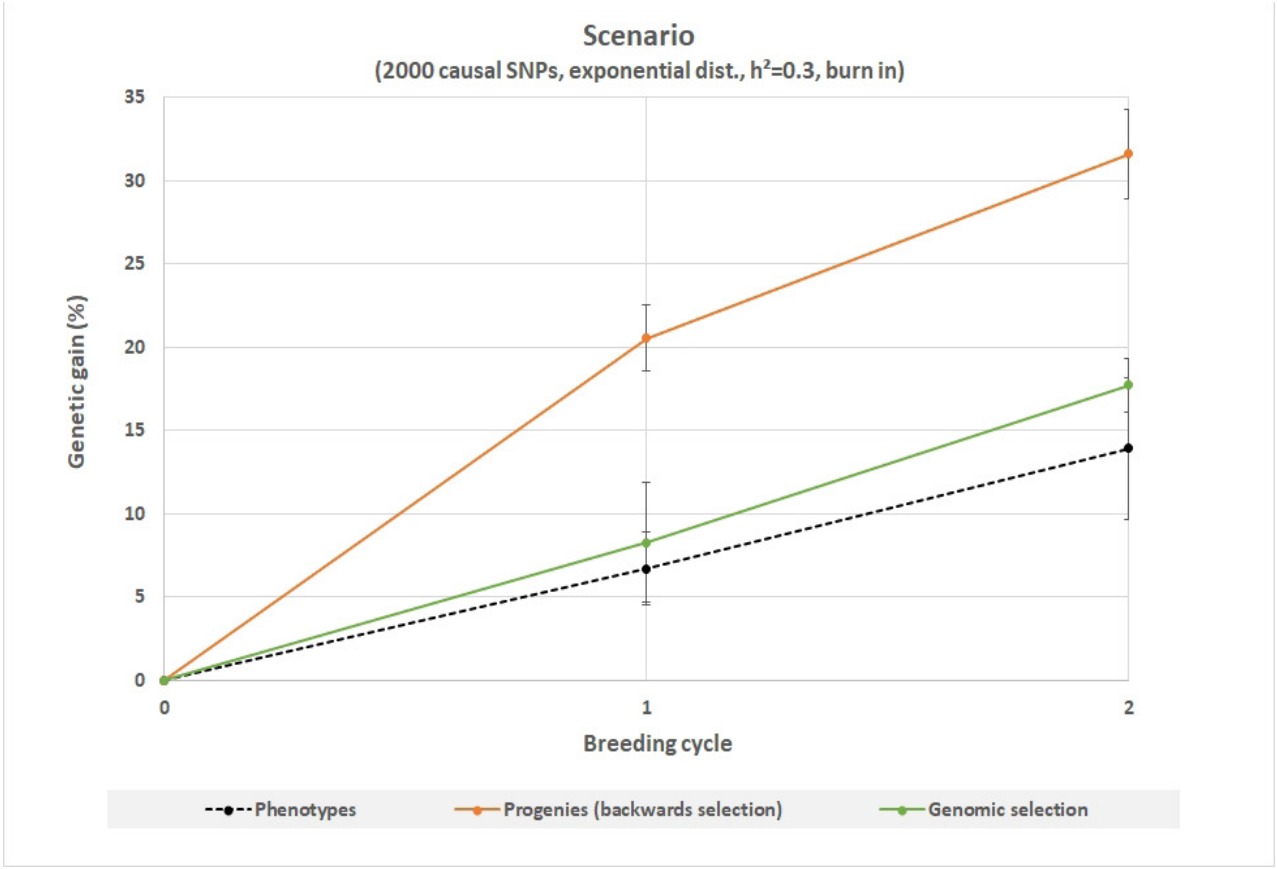

Figure 2
Genetic gains in a simulated simple tree breeding program using different criteria to select the mating partners of the next breeding cycle: phenotypes (Phenotypes), breeding values of progenies-tests (Progenies, backwards selection), and single generation genomic selection with 50 % training and 50 % test data (Genomic selection), the error bars indicate the standard deviation of 10 repetitions.

Genome-wide association studies (GWAS)

SNPscan breeder simulates parents and offspring that can be used for genome-wide association studies (GWAS). For this, simulated genomes of parents or offspring are stored as a HapMap-file and their phenotypes are stored in a text file. There are many different programs and R-scripts available to run GWAS. For the interaction with *SNPscan breeder* we have selected the program Tassel Version 5.0 (Bradbury et al. 2007). The Tassel results on GWAS using the GLM and MLM algorithm can be loaded into *SNPscan breeder* for post-ex analysis. For user defined thresholds of the association probabilities, the proportion of true and false positive associated SNPs are analysed. The estimated allelic effects are compared to the true effects using the Pearson's correlation coefficient. Further correlation coefficients between true and estimated individual breeding values are computed considering the identified SNPs and their allelic effects. Finally, the user can store the allelic effects of all SNPs above the probability threshold as a file used for marker-assisted selection (MAS) in forward simulations with *SNPscan breeder*.

Genetic diversity

Using 100 "dummy" loci with unique alleles for all individuals in the initial founder generation *SNPscan breeder* computes important population genetic parameters:

"Inbreeding F (0-1)": This is the inbreeding coefficient *F* defined as the probability that the two alleles of a homozygous genotype are identical by descent.

"Kinship (0-1)": This is defined as the average probability that alleles at a locus of pairs of individuals are identical by descent.

"Rep.pop.size (*Np*)": The reproductive effective population size is the effective number of parents contributing to the actual generation of offspring weighted by the relative fitness:

$1 \leq N_p = \frac{1}{\sum w_i^2} \leq N$ with (w_i) = proportion of successful male and female gametes of each individual.

"Inbreed. Pop. size (Ne)": The inbreeding effective population size: $1 \leq N_e = \frac{1}{2 \times \Delta F}$, with ΔF = difference of inbreeding coefficient of current and last generation (Falconer and Mackay 1996).

"Ancestor diversity (A_v)": This is the effective number of genetically unrelated ancestors contributing to the actual generation of offspring. During the generation of the parent genomes, each individual is assigned unique alleles at 100 loci *i*. Thus, in the beginning of the simulation the diversity at these "ancestor alleles" is equal to the initial number of parents (N). A_v is computed as the average diversity at all 100 ancestor loci:

$$A_v = \frac{\sum_{i=1}^{100} \frac{1}{\sum_{j=1}^{N} p_{ij}^2}}{100}$$

Example

We simulated a simple tree breeding program and tested the impact of different selection criteria on the genetic gain. In the scenarios it was a hermaphroditic tree species with 10 chromosomes, a genome size of 500 megabase pairs (Mbp) and average crossing over of 1.5 per chromosome. The trait of interest had a heritability of 0.3 and 2000 causal SNPs with exponentially distributed allelic affects according to the default settings of *SNPscan breeder*.

We simulated 1 million SNPs and used five generations of random mating with 10 % of the individuals as seed trees as a burn-in phase to produce the founder generation. This procedure created an average kinship between the individuals of the founder generation of 0.015. Then we simulated two breeding cycles. In each breeding cycles the top 30 individuals were selected as mating partners. With these 30 trees a half-diallel (selfing excluded) was used as mating design. So, we simulated 435 different parent combinations. In each combination 5 seeds were produced, summing up to 2,175 seeds per generation. As selection criteria of the top 30 trees (1.4 % of all individuals) we used: a) the phenotypes, b) the estimated breeding values from progeny tests (backwards selection), and c) single generation genomic selection based on 10,000 SNPs and the gBLUP algorithm (50 % training individuals). Each scenario was repeated 10 times.

After two breeding cycles the highest cumulative genetic gain with about 31 % was realised in the simulated selection based on the backwards selection using progeny tests (figure 2). A good performance was also observed for the single generation genomic selection with an accumulated genetic gain of more than 18 %. The selection based on phenotypes ended with a cumulative genetic gain less than 14 %. The conclusions for practical tree breeding programs should consider the extremely different workloads, costs and time needed for the different selection strategies to generate a certain genetic gain per unit time (Chamberland et al. 2020). The example illustrates the potential of *SNPscan breeder* to optimise tree breeding programs.

Discussion

What makes SNPscan breeder special?

Although there are quite a few other simulation programs and R-packages available to simulate breeding strategies and to study sample strategies, *SNPscan breeder* has special features that make it a useful tool. For example, it is a user-friendly windows application that allows unexperienced users to get started quickly. This offers also the possibility to use the program for teaching purposes. Possible disadvantages of a windows application in terms of computing speed are compensated by the broad use of parallel programming that enables a maximum number of CPU cores at the same time.

The main focus of the program is on the simulation of selection with many different criteria to identify the parents used for the next breeding cycle. To this end, selection by phenotypes, breeding values from progenies tests (backwards selection), genomic selection and MAS based on allelic effects estimated by GWAS are implemented. Import and export options are available to directly interact with the widely used software TASSEL (Bradbury et al. 2007) and the R-package rrBLUP (Endelman 2011). A unique feature is the post-hoc analysis of GWAS performance and the conclusions that can be drawn on the effectiveness of a given sample strategy and used GWAS algorithm to identify SNPs and estimate allelic effects for user-defined genetic architectures.

Compared to other software such as GeneEvolve (Tahmasbi and Keller 2017) *SNPscan breeder* makes strong simplifications of the genome but still maintains important characteristics such as total genome size, number of SNPs, segregation and crossing-over that allow the study of different sample strategies. The software XSim version 2 (Chen et al. 2022) and ADAM-Plant (Liu et al. 2019) are more sophisticated with regard to the simulation of genomic selection in a breeding program and complex crossing schemes but they do not implement MAS based on actual GWAS results.

Outlook

We will focus future work on the integration of real genomic data with *SNPscan breeder* to enable the simulation of more realistic breeding programs such as breeding programs with infusion of unrelated individuals and the sub-division into several breeding populations. For further testing the potential of MAS we will create additional links to other GWAS analysis programs such as GAPIT (Wang and Zhang 2021) for more advanced GWAS algorithms such as BLINK (Huang et al. 2019). Additionally, we will study in more detail the possibilities of MAS based on GWAS with extreme phenotypes, that is selecting only the edges of the phenotypic distribution. The use of genomic selection methods in the field of forest tree breeding is rapidly increasing. It will be interesting to assess the actual impacts on tree growth and quality considering different levels of inbreeding depression. Other aspects to be covered in future developments are breeding programs with multiple

traits and complex selection indices, genotype-environment interactions, pleiotropy and epistatic effects.

Data availability

SNPscan breeder has been programmed with Visual Studio 2019 as a .NET application (framework 4.8) and compiled as 64-bit versions for the operating system Microsoft Windows (Windows 11). The program, the user manual and different videos that explain the program are available on our website: https://www.thuenen.de/en/institutes/forest-genetics/software/SNPscan

Acknowledgement

We are thankful to members of the Centre for Integrated Breeding Research (CiBreed) at the University of Göttingen for helpful discussions on SNPscan breeder. We would like to thank Malte Mader for critical testing of the program and for helpful suggestions of its improvement. We are grateful to two anonymous reviewers for helpful comments and suggestions on a former version of the manuscript.

References

Bellmann K, Ahrens HJ (1966) Modellpopulationen in der Selektionstheorie und einige Ergebnisse aus Simulationsstudien. Der Züchter 36(4):172-185 https://doi.org/10.1007/bf02394156

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss V, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633-2635. https://dx.doi.org/10.1093/bioinformatics/btm308

Chamberland V, Robichaud F, Perron M, Gelinas N, Bousquet J, Beaulieu J (2020) Conventional versus genomic selection for white spruce improvement: a comparison of costs and benefits of plantations on Quebec public lands. Tree Genetics & Genomes 16(1):16. https://dx.doi.org/10.1007/s11295-019-1409-7

Chen CJ, Garrick D, Fernando R, Karaman E, Stricker C, Keehan M, Cheng H (2022) XSim version 2: simulation of modern breeding programs. G3-Genes Genomes Genetics 12(4):9. https://dx.doi.org/10.1093/g3journal/jkac032

Durel CE, Bertin P, Kremer A (1996) Relationship between inbreeding depression and inbreeding coefficient in maritime pine (Pinus pinaster). Theoretical and Applied Genetics 92(3-4):347-356. https://doi.org/10.1007/bf00223678

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4(3):250-255. https://dx.doi.org/10.3835/plantgenome2011.08.0024

Eriksson G, Ekberg I, Clapham D (2013) Genetics applied to forestry: an introduction. Uppsala, Sweden: Department of Plant Biology and Forest Genetics, SLU, ISBN 9157691878

Falconer DS, Mackay TF (1996) Introduction to quantitative genetics. London: Longman London, UK, 464 p, ISBN 0582446791

Gaynor RC, Gorjanc G, Hickey JM (2021) AlphaSimR: an R package for breeding program simulations. G3-Genes Genomes Genetics 11(2):5. https://dx.doi.org/10.1093/g3journal/jkaa017

Huang M, Liu XL, Zhou Y, Summers RM, Zhang ZW (2019) BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 8(2):12. https://dx.doi.org/10.1093/gigascience/giy154

Liu HM, Tessema BB, Jensen J, Cericola F, Andersen JR, Sorensen AC (2019) ADAM-Plant: A Software for Stochastic Simulations of Plant Breeding From Molecular to Phenotypic Level and From Simple Selection to Complex Speed Breeding Programs. Frontiers in Plant Science 9:15. https://dx.doi.org/10.3389/fpls.2018.01926

Pfenninger M, Reuss F, Kiebler A, Schonnenbeck P, Caliendo C, Gerber S, Cocchiararo B, Reuter S, Bluthgen N, Mody K, Mishra B, Balint M, Thines M, Feldmeyer B (2021) Genomic basis for drought resistance in European beech forests threatened by climate change. Elife 10:17. https://dx.doi.org/10.7554/eLife.65532

Plomion C, Aury JM, Amselem J, Alaeitabar T, Barbe V, Belser C, Berges H, Bodenes C, Boudet N, Boury C, Canaguier A, Couloux A, Da Silva C, Duplessis S, Ehrenmann F, Estrada-Mairey B, Fouteau S, Francillonne N, Gaspin C, Guichard C, Klopp C, Labadie K, Lalanne C, Le Clainche I, Leple JC, Le Provost G, Leroy T, Lesur I, Martin F, Mercier J, Michotey C, Murat F, Salin F, Steinbach D, Faivre-Rampant P, Wincker P, Salse J, Quesneville H, Kremer A (2016) Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Molecular Ecology Resources 16(1):254-265. https://dx.doi.org/10.1111/1755-0998.12425

R-Core-Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. [online]. Available from https://www.R-project.org

Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L, Worswick G, Studholme DJ, Zohren J, Salmon DL, Clavijo BJ, Li Y, He ZS, Fellgett A, McKinney LV, Nielsen LR, Douglas GC, Kjaer ED, Downie JA, Boshier D, Lee S, Clark J, Grant M, Bancroft I, Caccamo M, Buggs RJA (2017) Genome sequence and genetic diversity of European ash trees. Nature 541(7636):212-+. https://dx.doi.org/10.1038/nature20786

Sun X, Peng T, Mumm RH (2011) The role and basics of computer simulation in support of critical decisions in plant breeding. Molecular Breeding 28(4):421-436. https://dx.doi.org/10.1007/s11032-011-9630-6

Tahmasbi R, Keller MC (2017) GeneEvolve: a fast and memory efficient forward-time simulator of realistic whole-genome sequence and SNP data. Bioinformatics 33(2):294-296. https://dx.doi.org/10.1093/bioinformatics/btw606

Wang JB, Zhang ZW (2021) GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics Proteomics & Bioinformatics 19(4):629-640. https://dx.doi.org/10.1016/j.gpb.2021.08.005