SNPscan breeder

1.0

User manual

Bernd Degen & Niels Müller

Thuenen Institute of Forest Genetics
Sieker Landstrasse 2
22927 Grosshansdorf
Germany
e-mail: bernd.degen@thuenen.de

Last updated: 13th of July 2023

How to cite the program?

Degen, B. & Müller N. (2023) *SNPscan breeder* – a computer program to test genomic tools in breeding programs. Silvae Genetica (72), 126-131

Contents

1 Introduction

Advances in DNA sequencing techniques allow the sequencing of whole genomes of hundreds if not thousands of individuals. These data combined with precise and high-throughput phenotyping enable the identification of single nucleotide polymorphisms (SNPs) underlying traits with complex architectures. Merging quantitative genetics and genomics is getting a reality. This offers enormous opportunities, especially for breeding. The simulation program *SNPscan breeder* has been created to estimate, optimise and implement these opportunities.

SNPscan breeder was developed to generate simple simulations of whole-genome SNP data for many individuals in populations in order to:

- simulate phenotypes based on specific genetic architectures and a given level of heritability (number of SNPs with positive and negative effect alleles, distribution of SNPs in the genome)
    - develop and test sample strategies for GWAS
    - predict phenotypes and validate them with empirical data
- test and optimise different breeding strategies using various crossing schemes and based on pedigree data, genomic prediction and GWAS results

2 Download

SNPscan breeder has been programmed with Visual Studio 2019 as a vb.net application (.NET framework 4.8) and compiled as 64-bit versions for the operating system Microsoft Windows (Windows 11 and former versions). The program and different videos that explain the use of the program and the User's manual are available on our webpage:

https://www.thuenen.de/en/institutes/forest-genetics/software/SNPscan

3 Installation

Execute the file "SnpScan breeder.exe". The program uses parallel programming techniques to speed up the simulations. Thus, different CPU kernels are used at the same time. Make sure that you have enough hard disc space available for the simulations. The file of a single individual with 1 million SNPs has a size of ca. 4.9 MB.

4 Structure of the program

The program has four different menus (see figure 4.1):

- "File" => New Project, Open Project, Save Project, Save parents as Hapmap-file, Save offspring as Hapmap-file, End
- "Generate Data" => Chromosomes + Alleles, Genetic Architecture of Trait, Parents, Offspring
- “Breeding” => simulate different breeding strategies by setting a mating design (sub menu “Mating” and defining the criteria for recurrent selection (sub menu “Selection”)
- “Analyse” => analyse GWAS result of simulated individuals
- "About" => general information about the program

The structure represents the order of the user's activities. First you have to create a new project or load an existing project. Each project has a project name and a file directory. The project directory is subdivided in the “main” directory and directories for each simulated generation (“F1”, “F2” etc.). The main directory includes the following files:

- “Chromosome_allele_file.txt” => name of the locus, chromosome number, SNP location on the chromosome, relative frequencies of the four possible alleles: A,C,G,T
- “FileArchitecture.csv” => name of the causal SNP, chromosome location, allele name, effect direction (+ or -), and effect size, R^2 max (maximal amount of phenotypic variation explained by the SNP assuming allele frequencies of 0.5 in the population) and expected relative frequency of the allele with an effect <> 0 in the parent generation. Note that always only one allele has an effect <> 0.
- “LogFile.txt” => project name, file structure, selected parameters. This file needs to be opened if you want to load an existing project.
- “QG.txt” => file with the quantitative genetic data: name of each individual, parents, genetic value and phenotype.
- “Parents_Causal_Loci.txt” => This file stores the genotypes of the parents at the causal SNPs responsible for the trait

- “AlleleFreq_Causal_loci.txt” => changes of the allele frequencies of causal SNPs for parents and each offspring generation
- “P_Ancestor_File.txt” => stores the ancestor information of the parent population. In the beginning 100 loci are created for each initial individual with unique numbers coding for the alleles. During the simulation the ancestor information is used to compute inbreeding, kinship and ancestor based effective population sizes.

The program creates one subdirectory for the “Parents” and then for each following generation: “F1”, “F2” etc. The genomic information of each individual is stored as a single ASCII-file in the respective sub-directory. In case of multiple repetitions of the forward selection, only the files of the last repetition are saved.

Once the project is created the user needs to generate data on the overall genome configuration (chromosomes and alleles) and the genetic architecture of the trait. As a next step, the founder population of parents is generated. Once the parents have been simulated the user can run a breeding strategy by defining the mating scheme (system of crosses among parents) and setting selection criteria for recurrent selection. During this simulation different types of selection (with regard to the direction and intensity) can be simulated. Further, the data of the parents, phenotypes and changes in allele frequencies of the causal SNPs are stored. At the end of a forward simulation the user can create genome files of all parents and all offspring in a particular generation as a hapmap-file. The hapmap-files and the file with the phenotypes can then be loaded in GWAS programs such as TASSEL (Bradbury, et al. 2007).

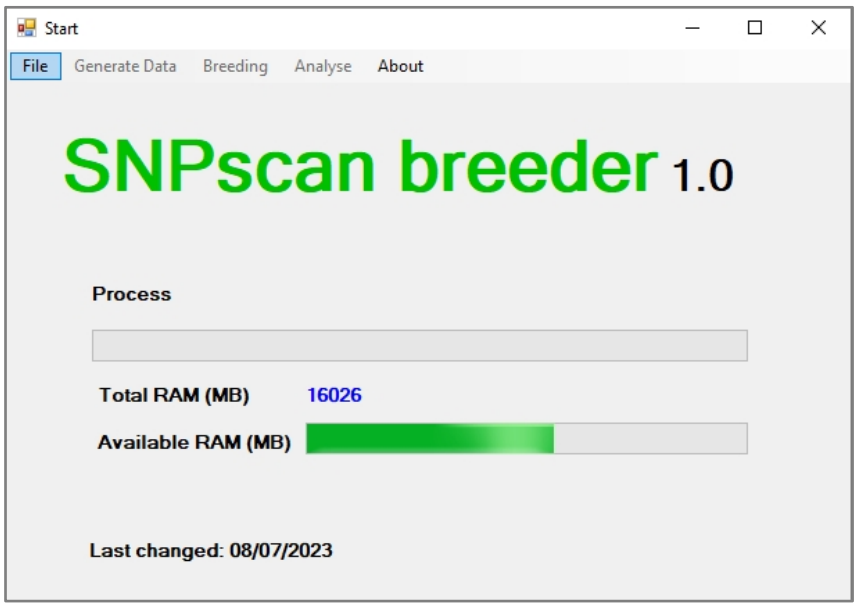


Figure 4.1: Start screen of the program *SNPscan breeder*

5 Menu “File”

See figure 5.1. First, a new project needs to be created (“New Project”) or an existing project opened (“Open Project”). To create a new project, you need to select or create a file directory for your new project and give the project a name. All files linked to the project are stored in this directory. In order to open a project, you need to go to the directory of the project, select the sub-directory “main” and click on the file “LogFile.txt”.

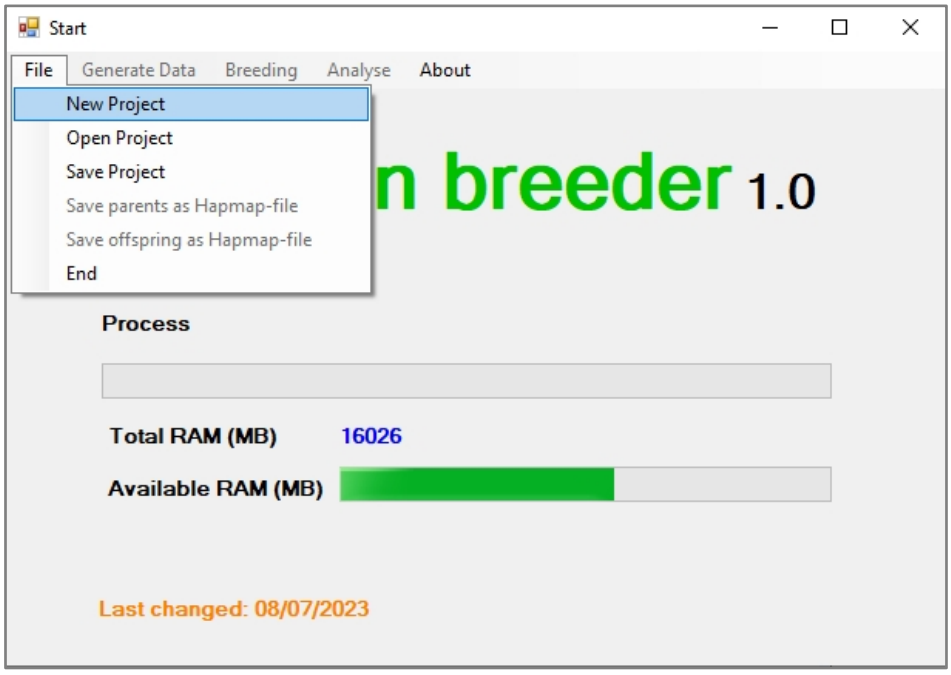


Figure 5.1: The menu “File”

Once the genomes of the parents and offspring have been created they can be saved as single files in the hapmap-format for further data analysis (“Save parents as Hapmap-file” or “Save offspring as Hapmap-file”).

6 Menu "Generate data"

In the menu “Generate Data” (figure 6.1.1) the user can design genomes with a given number of chromosomes and SNPs (“Chromosomes + Alleles”), define the genetic architecture of a trait (“Genetic Architecture of Trait”) and generate the genomes of potential parents (“Parents”).

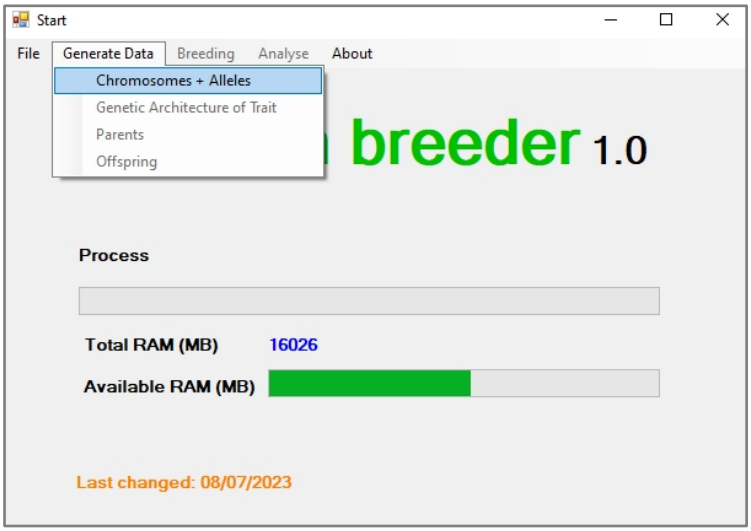


Figure 6.1.1: The menu “Generate Data”

6.1 Chromosomes and alleles

SNPscan breeder assumes diploid sets of chromosomes and the production of male and female gametes by all individuals (monoecy, hermaphroditism). Selfing is possible. The user defines the number of chromosomes, genome size, total number of SNPs and the average number of crossing-overs per chromosome (figure 6.1.2). The SNPs are evenly distributed over all chromosomes. E. g., a genome with 20 chromosomes and a total of 2 million SNPs will have 100,000 SNPs per chromosome. *SNPscan breeder* assumes a simplified recombination landscape with equal probability for crossing-overs along the chromosomes.

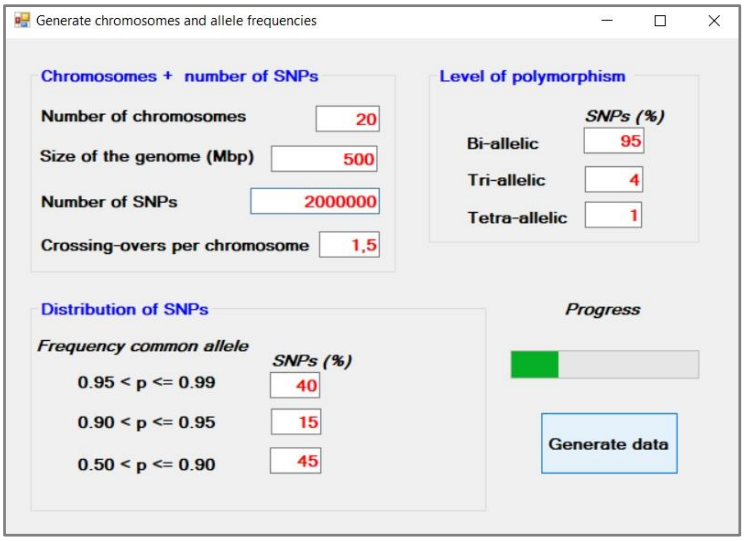


Figure 6.1. 2: Window to “Generate chromosomes and allele frequencies”

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 the frequencies of the common alleles in the genome (figure 6.1). The default values are based on population re-sequencing data of beech (*Fagus sylvatica*), oak (*Quercus robur*) and ash (*Fraxinus excelsior*). For details on these genomes see Sollars, et al. (2017), Pfenninger, et al. (2021) and Plomion, et al. (2016).

6.2 Genetic architecture

The user defines the name of a trait, its mean phenotypic value in the founder or wild population, the variance of the phenotype, the heritability of the trait and the percentage of inbreeding depression (figure 6.2.1). The percentage of inbreeding depression assumes a linear relationship between the inbreeding coefficient and the reduction of the phenotype value (Durel, et al. 1996). The user further specifies the number of causal SNPs and selects one of three alternative functions for the distribution of the allelic effects: a) negative exponential distribution, b) normal distribution, or c) equal distribution. The mean of all allelic effects in the three distributions is 0 and in all cases 50 % of the allelic effects are positive and 50 % of the effects are negative. The negative exponential distribution is controlled by the exponent as a single parameter. Values close to 2 result in many alleles with small effects and only a few alleles with larger effect. With a decreasing exponent the distribution gets flatter with a more even distribution of effect sizes and more loci with larger effects (figure 6.2.2). For the normal distribution the user specifies the standard deviation as parameter. Larger values of the standard deviation will result in a flatter distribution with smaller and larger effects more evenly distributed (figure 6.2.3). Small standard deviations imply many small effects and only few large-effect alleles.

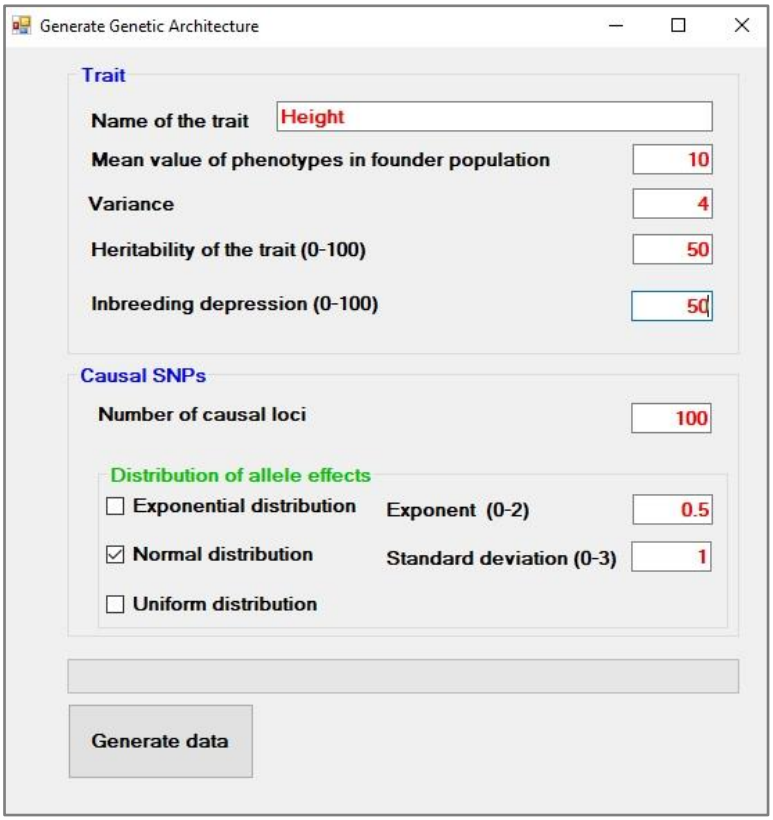


Figure 6.2 1: Window to define the genetic architecture of a trait

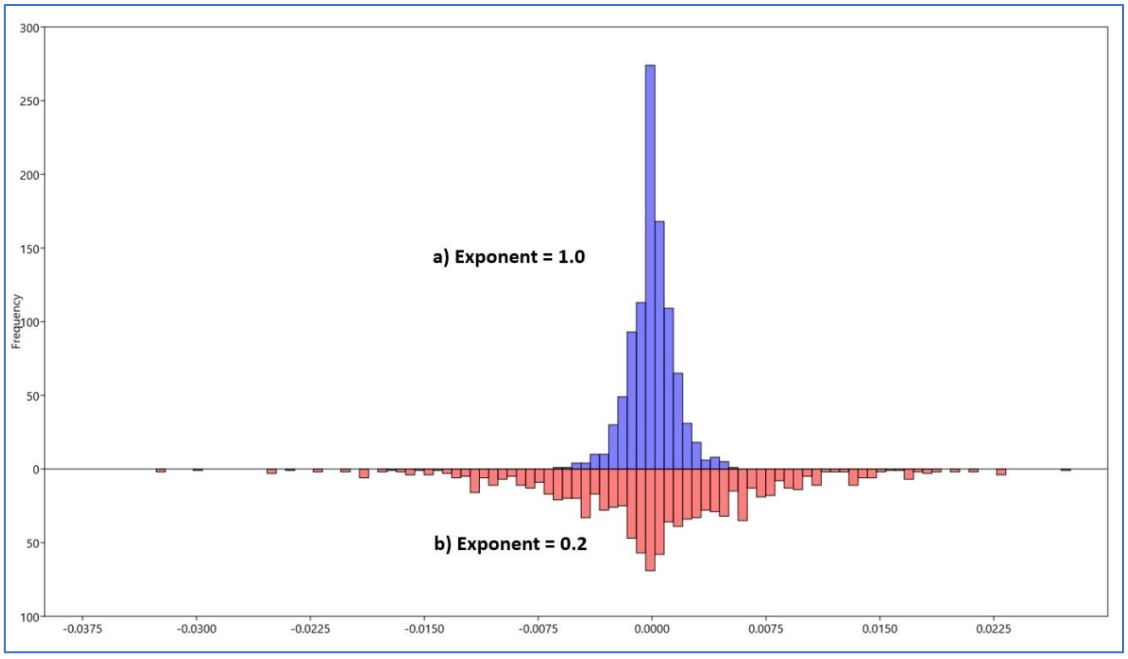


Figure 6.2.2: Examples for the allelic effects for 1000 causal SNPs generated with the negative exponential distribution: a) with an exponent of 1 (blue) and b) with an exponent of 0.2 (red).

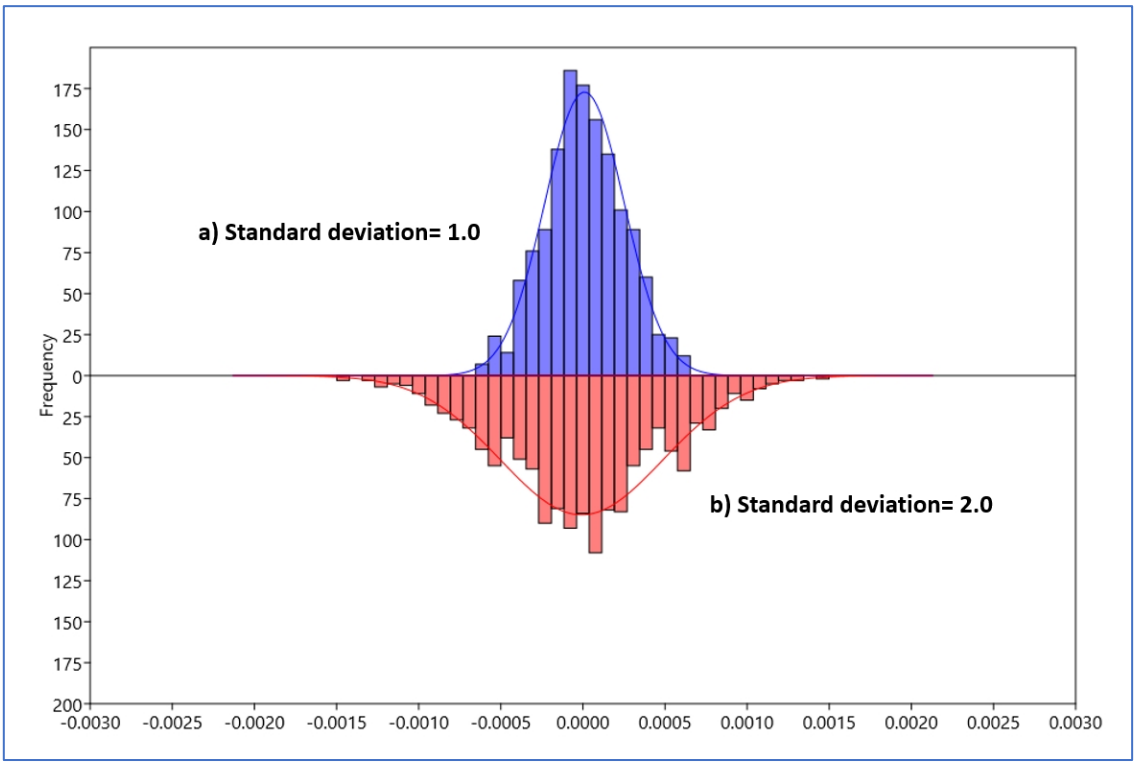


Figure 6.2.3: Examples for the allelic effects for 1000 causal SNPs generated with the normal distribution: a) with standard deviation 1 (blue) and b) with standard deviation of 2 (red).

Phenotypes

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

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

Genetic value (genomic breeding value)

The genetic value of individual *i* (g_i) is computed as the sum of all additive effects at each causal locus j (a_{ij}) + a correction *m* to centralise the mean of all genetic values of the initial population to 0 multiplied with the scale factor *sf*:

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

The scale factor (*sf*) is defined as:

$$sf = \frac{\sqrt{(\sigma_p^2 * h)}}{\sigma_{g-values}}$$

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

h = heritability (0-1, parameter of the model)

$\sigma_{g-values}$ = standard deviation of the genetic values of the population in the beginning

Environmental value

Environmental values (e_i) are sampled from a normal distribution with a mean of 0 and the standard deviation of environmental effects => N(0,σ_{env}).

$$\sigma_{env} = \sqrt{(\sigma_p^2 - h * \sigma_p^2)}$$

6.3 Parents

In order to generate parents, the user needs to specify the number of parents (figure 6.3.1).

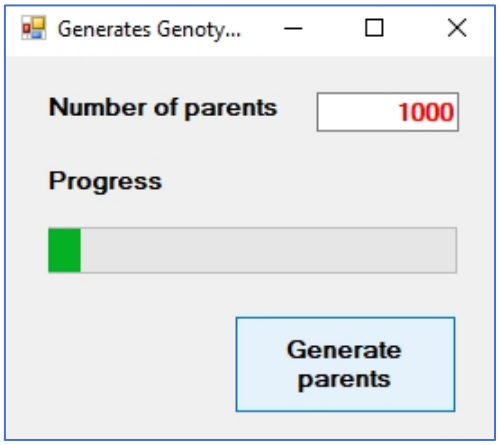


Figure 6.3.1: Window to generate the parents

7 Breeding

After the parents have been created the user can select the menu “Breeding” setup the breeding scenario by first selecting the mating design and then defining the selection criteria (figure 7.1.1).

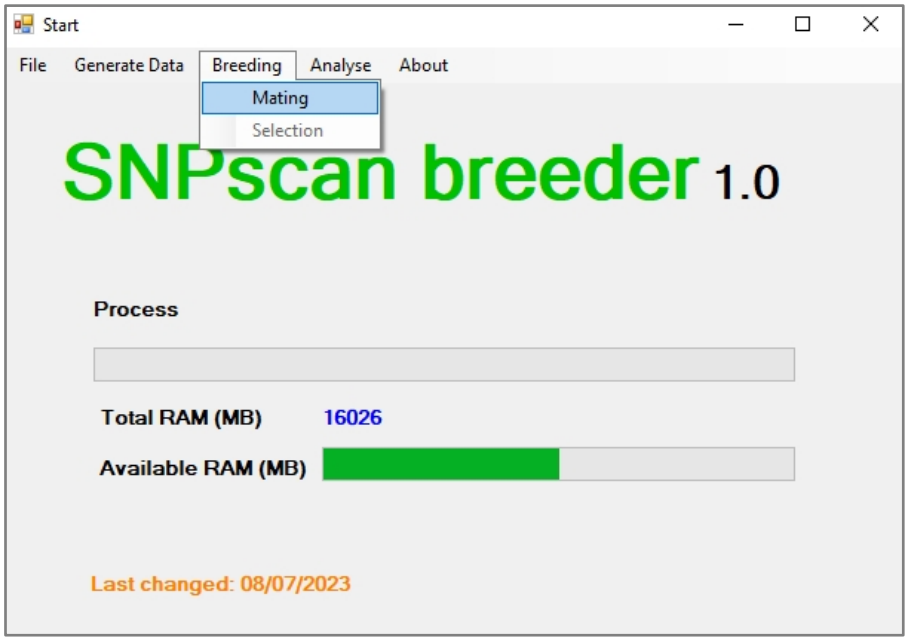


Figure 7.1.1: Window of the menu “Breeding”

7.1 Mating

All common mating designs of plant breeding programs (Eriksson, et al. 2013) have been implemented in *SNPscan breeder* (figure 7.1.2):

- 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). After clicking “OK” the total number of possible combinations and the number of seeds simulated are computed.

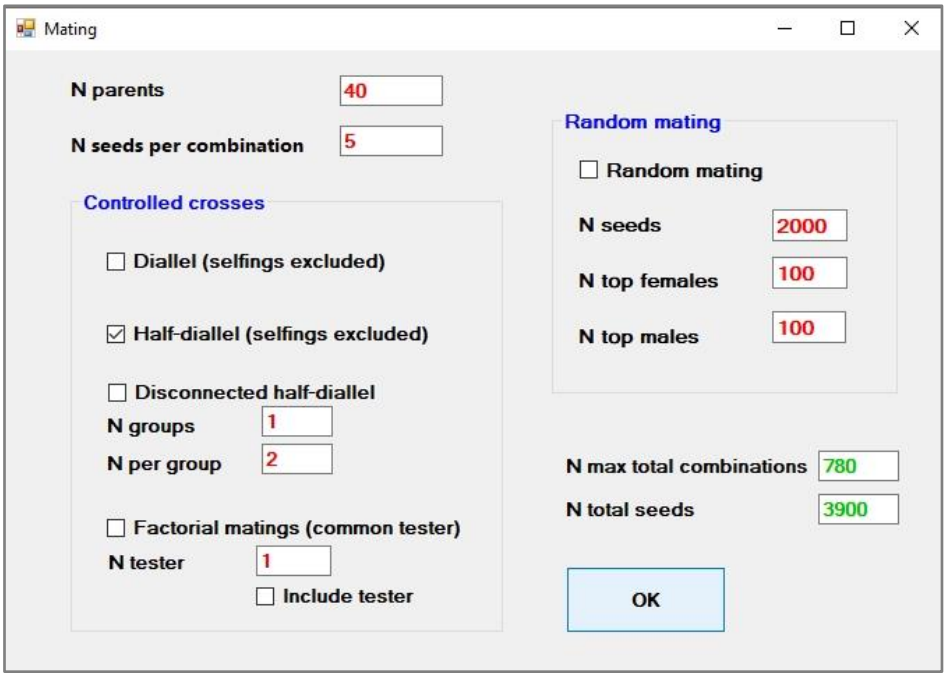


Figure 7.1.2: Window of the menu “Mating”

7.2 Selection

After the mating design has been setup the user can specify the selection criteria by clicking on “Selection” in the menu “Breeding” (figure 7.2.1). The number of simulated generations (non-overlapping generations) can be specified by the user (“Number of simulated generations”). Each series of generations can be repeated as defined (“Number of repetitions).

As indicated in the panel “Criteria for selection of individuals” the user can choose among seven different options on how to select the parents for the next mating cycle:

- “Phenotypes of adults” => adults are ordered according to their phenotypes
- “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 breeding values).
- “Gene markers of adults” => The program computes for each adult the genetic value of causal SNPs. The proportion of causal SNPs involved in this is defined by the selected “Heritability of causal SNPs”. If this value is equal to the total heritability of the trait, all causal SNPs are used. If the value is smaller only a proportion of causal SNPs are included. The selection of causal SNPs is done from the SNPs with the

largest effect on the phenotypic variation down to the smallest ones until the defined heritability is reached. Note: This routine takes the true effects of the alleles as they have been generated by the program.

- “GWAS estimates of allele effects” => This option is possible if you have done a genome-wide association study (GWAS) on the simulation results and stored the estimated allele effects in a file (see chapter 8 “Analyse – GWAS”). The program will ask you to open the file with the estimated allele effects. During the simulations the breeding values of the adult individuals are estimated as the sum of the allele effects at all identified causal SNPs.
- “Genomic selection (gBLUP) single generation” => The program uses the function “kin.blup” integrated in the R-package “rrBLUP” (Endelman 2011). The algorithm computes “Predicted Genomic Breeding Values (PGBV)“ for all individuals in the actual generation F with help of a selected proportion of phenotypes (“% training data set (1-100)”) and the kinship-matrix of all individuals. The user can select the number of SNPs used for the predictions. As an option the causal SNPs can be excluded from these randomly selected SNPs. Note that this function requires the installation of R and the package “rrBLUP” on the computer. During the simulations *SNPscan breeder* creates a subdirectory “R” in the project folder in order to store the input files, r-script for the calculation and the results. *SNPscan breeder* automatically tries to find the file “Rscript.exe”. In case this is not successful, the user needs to enter the path to the program manually. During the simulation, *SNPscan breeder* then repeatedly updates the input files, runs the R-script and reads the results on the PGBVs.
- “Genomic selection (gBLUP) cross generations” => As above but the phenotypes of the selected proportion of individuals of generation F_{-1} are used as training population and the individuals of the generation F are the test population. The algorithm computes “Predicted Genomic Breeding Values (PGBV)“ for all individuals in generation F with help of phenotypes of generation F_{-1} and the kinship-matrix of all individuals in both generations.
- “Random” => the male and female parents get randomly selected. This option is useful to generate as a “burn-in process” a certain level of kinship in the initial population for further simulations => at the end of the simulations the individuals can be saved as parents by clicking on the bottom “Save as parents”.

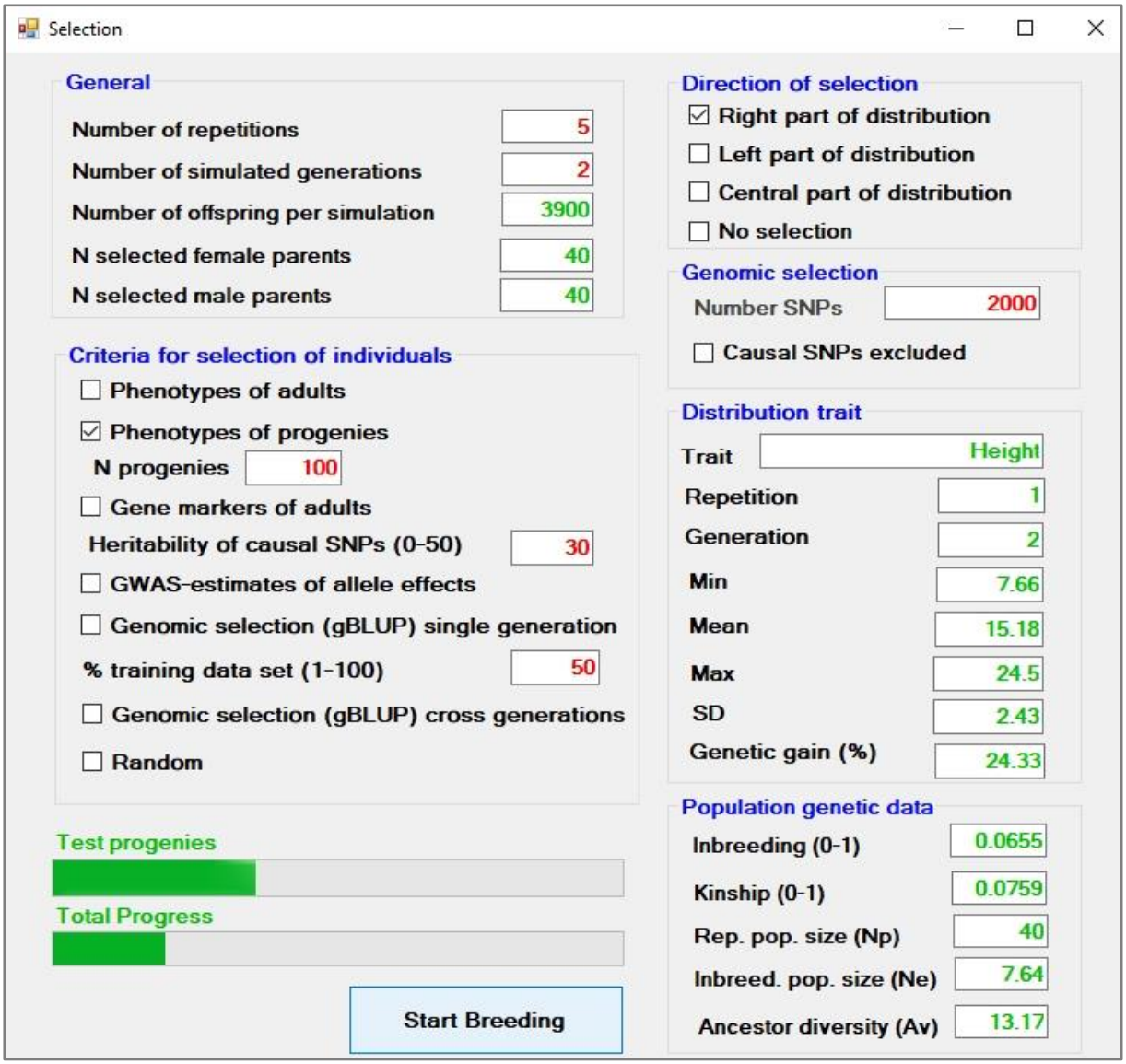


Figure 7.2.1 : Window of the menu “Selection”

The user needs to define the “Direction of selection”. There are four options:

- “Right part of distribution” => the individuals are ordered from the smallest to the largest value of the selection criteria and the x% individuals with the largest values are selected.
- “Left part of distribution” => the x% individuals with the smallest values are selected.
- “Central part of distribution” => (100 – x%)/2 of the individuals with the smallest values and (100 – x%)/2 of the individuals with the biggest values get excluded from the mating. All other individuals participate in the mating
- “No selection” => all individuals are included in the mating.

The fields in the panel “Distribution trait” give information on the current simulated generation:

- Generation => “Parent” or “F1”, “F2” etc.
- Min => minimum phenotypic value of all individuals
- Mean => arithmetic mean of phenotype
- Max => maximum phenotypic value of all individuals
- SD => standard deviation of the phenotype
- Genetic gain (%) => relative difference of the arithmetic means of the phenotype in the current and former generation

In addition, some important population genetic parameters are computed and shown in the panel “Population genetic data”:

- “Inbreeding (0-1)” => average probability that two equal alleles of a homozygote genotype are identical by decent
- “Kinship (0-1)” => average probability that homologous alleles of pairs of individuals are identical by decent
- “Rep.pop.size (*Np*)” => reproductive effective population size. Effective number of parents contributing to the actual generation of offspring weighted by the relative fitness. Proportion of successful male and female gametes of each individual (w_i):

$$1 \leq N_p = \frac{1}{\sum w_i^2} \leq N$$

- “Inbreed. Pop. size (Ne)” => inbreeding effective population size: $1 \leq N_e = \frac{1}{2 \times \delta \mathrm{P}} \leq N$, with $\delta \mathrm{P}$ = difference of average inbreeding of current and last generation, inbreeding = probability of two alleles of an individual to be equal by decent (Falconer and Mackay 1996)
- “Ancestor diversity (A_v)” => Effective number of genetically unrelated ancestors contributing to the actual generation of offspring. During the generation of the parent genomes, each individual got unique alleles at 100 loci *i*. Thus, in the beginning of the simulation the initial diversity at these “ancestor alleles” was equal the initial number of parents (N). A_v is computed as the average diversity at all 100 ancestor loci:

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

- Accuracy PGBV => the accuracy of the predicted genomic breeding values (PGBV). Pearson’s correlation coefficient among the Genomic Breeding Values (GBV) and PGBV in the test population. This column is only provided, if “Genomic selection

(gBLUP)" has been selected. In the first generation no training population is available. Here the phenotypes of the adults are used for selection and no PGBVs are computed leading to "NaN" for the accuracy.

These values are stored in the file "Selection_Records.txt" in the main project directory (table 7. 2).

Repetition	Generation	Min	Mean	Max	SD_Phenotypes	Genetic Gain %	SD Genetic Values	N p	N e	Inbreeding	Kinship	Ancestor_Diversity	Accuracy PGBV
1	1	2,69	11,56	19,06	2,46	14,99	0,028	94,39	100	0,01	0,01	94,39	0,57
1	2	5,79	13,38	21,87	2,35	15,68	0,024	94,62	25,64	0,039	0,036	27,6	0,72
2	1	5,02	11,16	19,38	2,31	10,95	0,026	95,17	83,33	0,012	0,01	95,17	0,57
2	2	6,43	12,94	19,25	2,25	15,99	0,023	94,92	31,25	0,032	0,03	33,04	0,69

Table 7.2: Statistics of the phenotypes in each generation stored in the file "Selection_Records.txt"

8 Analyse GWAS

You can use *SNPscan breeder* to analyse results of genome-wide association studies (GWAS) that have been done with simulated data (figure 8.1.1). The purpose of this analysis might be:

- to elaborate the power of sample designs to detect causal SNPs in GWAS
- to compare the performance of different methods and algorithms of GWAS
- to estimate the impact of different model parameters (e.g. level of heritability, number of causal SNPs, minor allele frequencies, distribution of allele effects, level of kinship) on the performance of GWAS
- to create files with identified causal SNPs and estimated allele effects for later use in forward simulations to compute breeding values (see chapter 7 “Forward selection”).

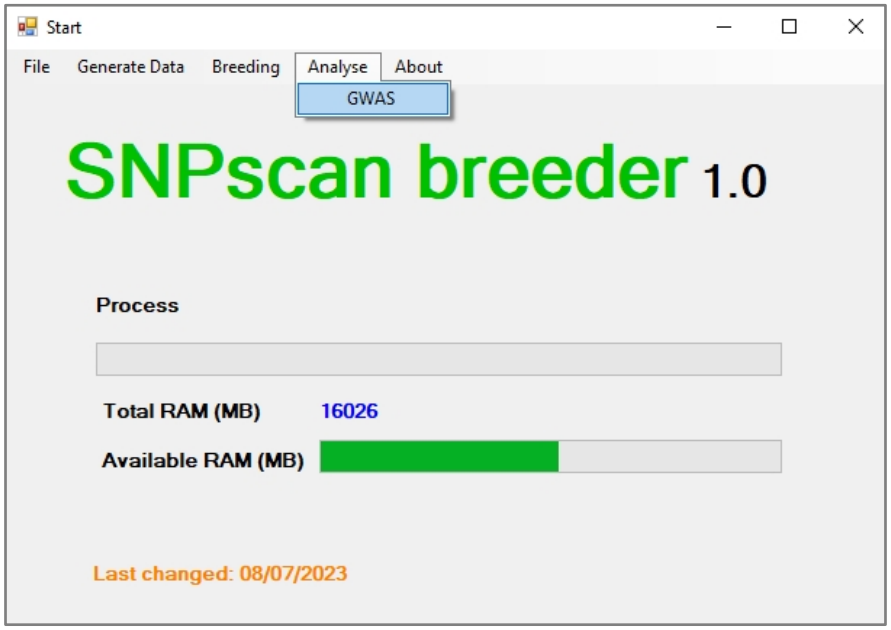


Figure 8.1.1: Window to start the Analysis of GWAS results

SNPscan breeder simulates populations of parents and offspring that can be used for genome-wide association studies (GWAS). For this, you need the parent and at least one offspring generation and store the simulated genomes as hapmap-files. Further, you need data on the phenotypes. They are stored in the sub-directory “main” of your simulation project in the file “QG.txt”. There are many different software and R-scripts available to run a GWAS. For the interaction with *SNPscan breeder* we have selected the program Tassel Version 5.0 (Bradbury, et al. 2007). The format of the hapmap and phenotype files are in the right format so that Tassel can open them and do a GWAS. In Tassel you can use different methods of association analysis. For details see the Tassel web page: https://tassel.bitbucket.io/

So far, we have included into *SNPscan breeder* the possibility to use the results of the General Linear Model (GLM) and the Mixed Linear Model (MLM). In Tassel the output of the GWAS is stored in two files: one with the statistics of each SNP and one with the estimated allele effects. To analyse GWAS results in *SNPscan breeder* you need to open these two files (figure 8.2.1).

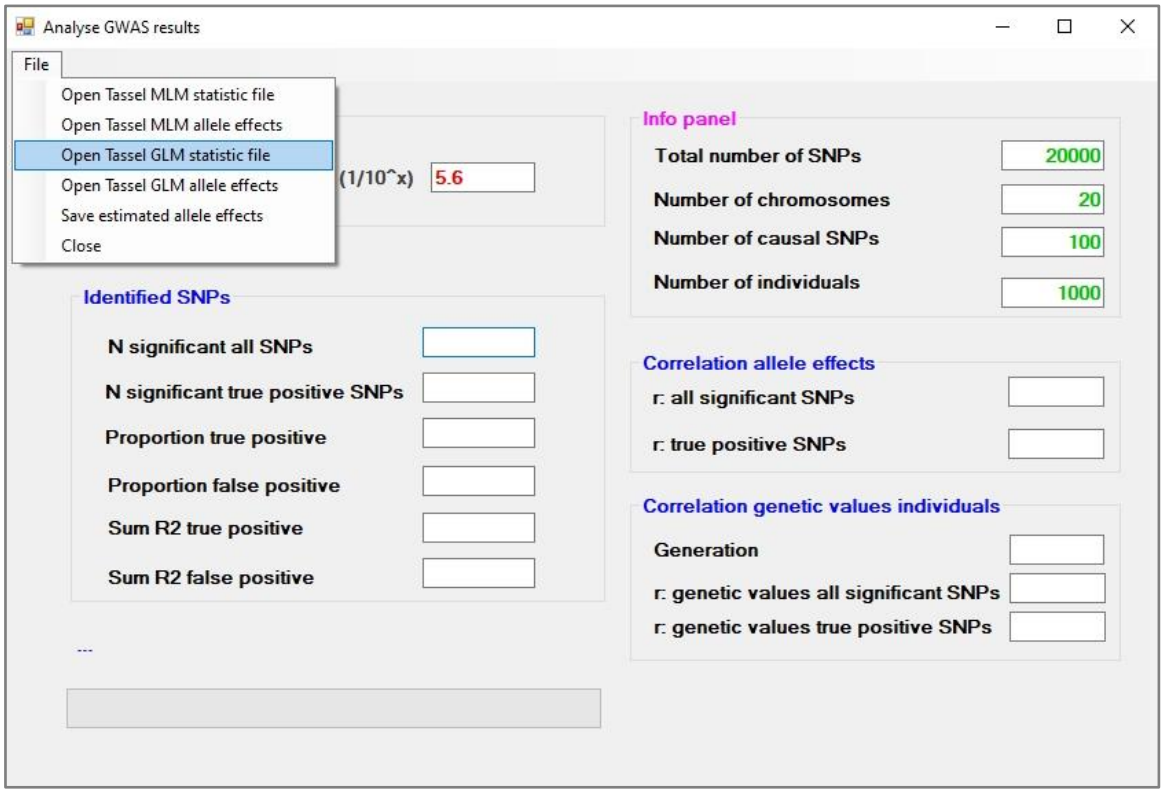


Figure 8.2.1: Window to open the GWAS results computed with Tassel

Once the two files have been opened you can click the bottom “Compare SNPs” (figure 8.3a). This starts a routine that compares the “true” causal SNPs of your simulation with the identified SNPs of the GWAS. The parameter “Threshold prob additive” controls the identification of SNPs. This is the threshold of -log10 p-values computed for SNPs with additive effect.
The standard value (α_b) in *SNPscan breeder* uses an α value of 0.05 after Bonferroni correction based on the total number of tests (m), which is the total number of SNPs of the simulated genome ($\alpha_b = \frac{\alpha}{m}$). In the example shown in figure 8.3a the threshold is a -log10 p-value of 5.6 for 20,000 simulated SNPs. The box “Info panel” summarises important settings of the simulations.

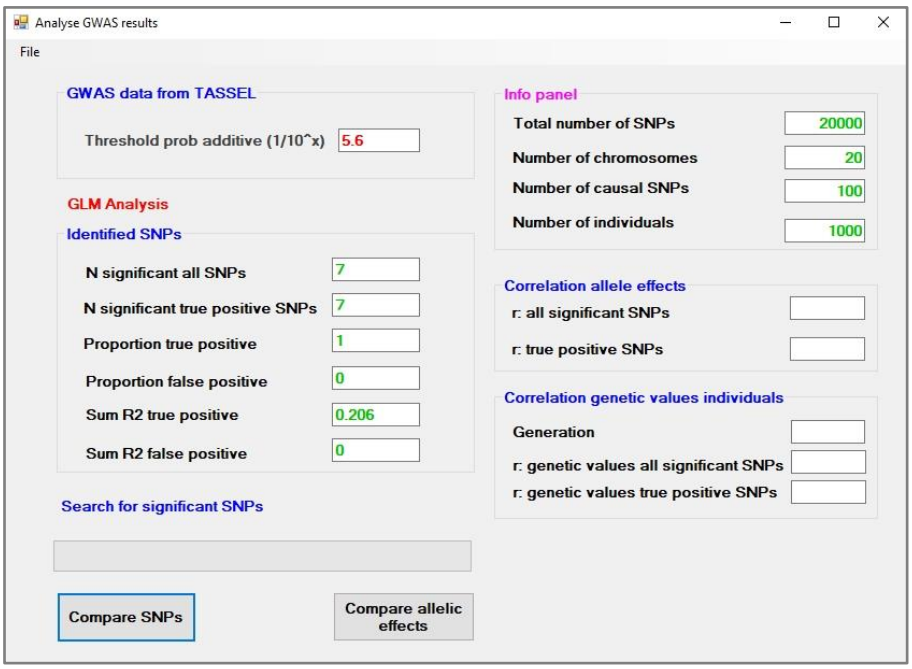
The results of the comparison between true simulated genetic architecture and the estimates of the GWAS are given in the panel “Identified SNPs”:

- “N significant all SNPs” => this is the total number of SNPs in the GWAS that are above the threshold of probability
- “N significant SNPs true positive” => the total number of identified causal SNPs
- “Proportion true positive” => the relative frequency of causal SNPs among all identified SNPs
- “Proportion false positive” => the relative frequency of identified SNPs that are not among the causal SNPs
- “Sum R^2 true positive” => GWAS results of the proportion of phenotypic variation that is explained by all “true positive” SNPs together
- “Sum R^2 false positive” => GWAS results of the proportion of phenotypic variation that is explained by the “false positive” SNPs. Note that this value can be higher than 1 because these SNPs are per definition “false”

The user can change the threshold probability and check the impact on true and false positive SNPs (figure 8.3b).

By clicking on the bottom “compare allele affects” the correlation between the allele effects of the GWAS identified SNPs and the allele effects of the true SNPs of the simulations are computed. Further, the correlation between the genetic values (breeding values) of the individuals of a selected generation are calculated for all GWAS identified SNPs and all true SNPs coding for the trait (figure 8.4). Both correlations are given once for all identified SNPs and once only using the true positive SNPs. The level of correlation indicates how good the estimated allele effects could be used as predictors in breeding programs. The identified SNPs and their effects can be stored as files for the later use in the “Forward simulations” (figure 8.5).

a



b

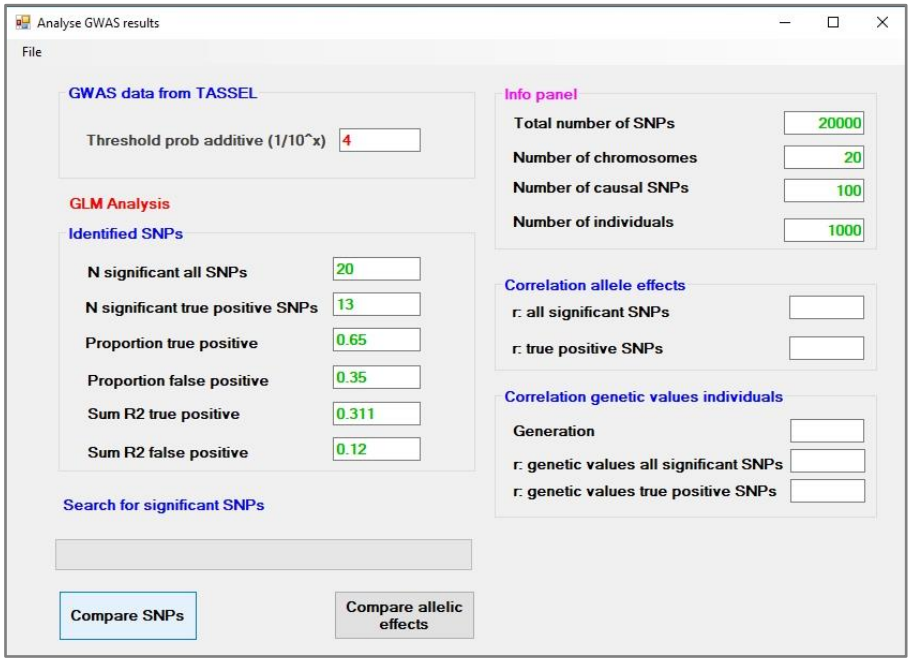


Figure 8.3: Comparison of simulated true causal SNPs and identified SNPs in the GWAS a) using the -log10 p-value of 5.6 (with Bonferroni correction for α = 0.05), b) using a -log10 p-value of 4

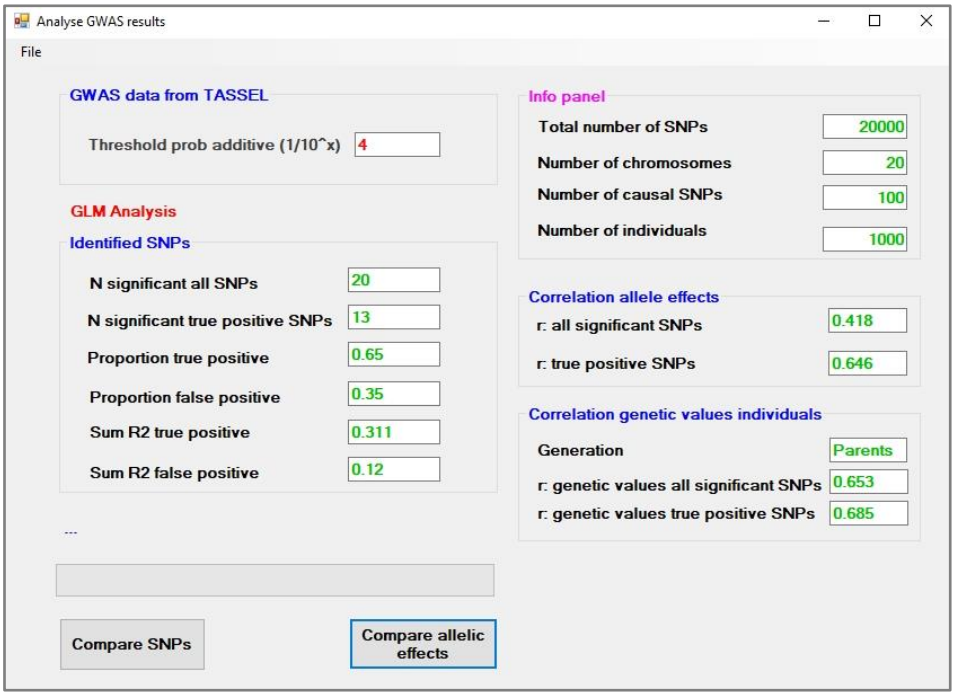


Figure 8.4: Window with computed correlations of allele effects and individual genetic values at the GWAS-identified SNPs and all causal SNPs

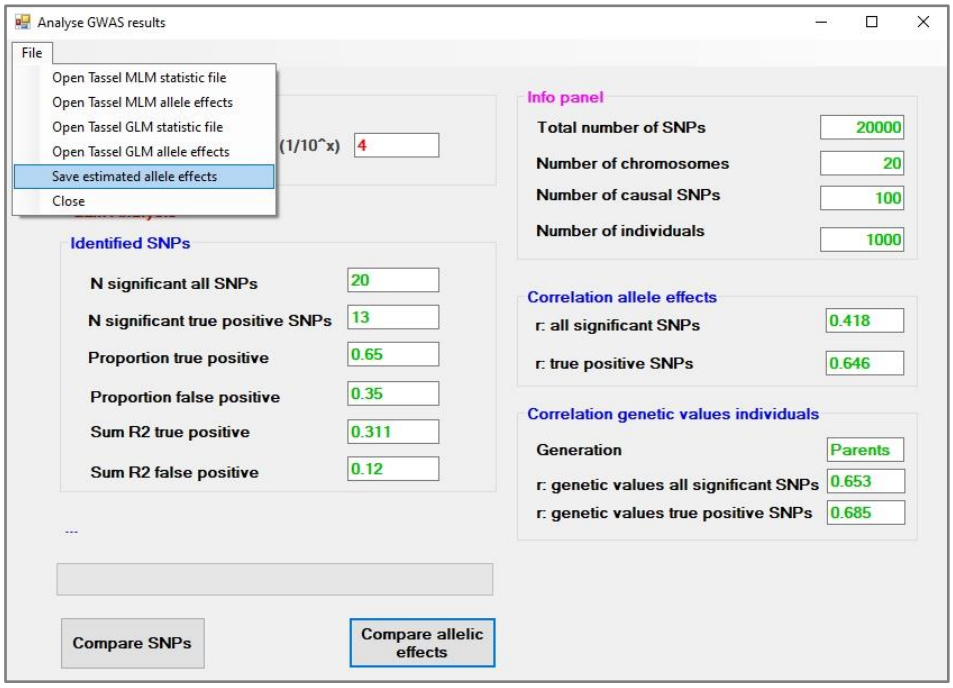


Figure 8.5: Window to save the estimated allele effects as a file for later use in a “Forward simulation”

9 Acknowledgements

We would like to thank Malte Mader for critical testing of the program and for helpful suggestions of its improvement.

10 Literature

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. 2007 TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23** (19), 2633-2635.

Sollars, E.S.A., Harper, A.L., Kelly, L.J., Sambles, C.M., Ramirez-Gonzalez, R.H., Swarbreck, D. *et al.* 2017 Genome sequence and genetic diversity of European ash trees. *Nature*, **541** (7636), 212-+.

Pfenninger, M., Reuss, F., Kiebler, A., Schonnenbeck, P., Caliendo, C., Gerber, S. *et al.* 2021 Genomic basis for drought resistance in European beech forests threatened by climate change. *eLife*, **10**, 17.

Plomion, C., Aury, J.M., Amselem, J., Alaeitabar, T., Barbe, V., Belser, C. *et al.* 2016 Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Mol. Ecol. Resour.*, **16** (1), 254-265.

Durel, C.E., Bertin, P. and Kremer, A. 1996 Relationship between inbreeding depression and inbreeding coefficient in maritime pine (Pinus pinaster). *Theoretical and Applied Genetics*, **92** (3-4), 347-356.

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

Endelman, J.B. 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, **4** (3), 250-255.

Falconer, D.S. and Mackay, T.F. 1996 *Introduction to quantitative genetics*. Longman London, UK: London, 464 p.