GenoSim

Yanlin Liao & Giorgio Tumino

General introduction

Most recent genetic studies in polyploid crops are based on the accurate estimation of marker allelic dosage. However, genotype calling in polyploids is challenging due to its multiple classes of heterozygosity. Simulation studies are helpful to generate a scenario of interest and predict the best genotyping parameter or genotype calling methodologies to be used. Additionally, simulations are ideal to test new methodologies or evaluate competing approaches.

The purpose of GenoSim is to generate simulated data similar to experimental SNP array data or GBS read count data of experimental populations or pedigrees of any ploidy; these simulated data can be used to evaluate or develop genotype calling software for data with specific features. The simulation occurs in two stages. First, actual genotypes are simulated, based on a given population or pedigree and a given or simulated genetic map, according to Mendelian inheritance. Second SNP array data or read counts are simulated, given these simulated genotypes and given parameters affecting data quality. Both the simulated genotypes and the simulated experimental data are saved in data files that can then be used as input for genotype calling or other software.

Operation

GenoSim interface includes three pages: *'Simulation of Genotypes'*, *'SNP array Simulation'*, and *'Sequence Reads Simulation'*. In the *'Simulation of Genotypes'*, you can simulate dosage scores of your desired population, where parameters like population structure, ploidy level, chromosome settings, and etc. need to be specified. The *'SNP array simulation'* and '*Sequence Reads Simulation'* need to use the output from *'Simulation of Genotypes'* (Uploading dataset in same format as the output of *'Simulation of Genotypes'* is also possible). In *'SNP array simulation'* and '*Sequence Reads Simulation'*, parameters that introduce variation in the aspect of its technology need to be chosen. More description regarding to each parameter are explained in the interface.

Simulation of genotypes

We first need to simulate actual genotypes for the desired population or pedigree; this is done on tabsheet Simulation of genotypes. If actual genotypes are already available this step can be skipped; then these genotypes can be taken from a file and uswd directly to simulate SNP array data or sequence readcount data (see next sections).

The simulation of genotypes is performed using the PedigreeSim software and this tabsheets (including its sub-tabsheets) allow you to specify all parameters, including the population

structure (including a specified pedigree if desired), the ploidy (in the case of an F1 population the ploidy of the two parents may be different), the number of replicates of the desired population, a prefix for all filenames generated by the genotype simulation, and parameters for preferential chromosome pairing and quadrivalent formation, the number of chromosomes, their length and centromere position, and the numbers of each type of marker.

SNP array and Sequence reads simulation

After the genotypes have been simulated, they can be used to simulate SNP array data or sequence read counts (see next section).

Go to tabsheet SNP array/Sequence reads simulation. On the left, select the source of the genotypes to use for the simulation: the genotypes simulated in the previous step or a file with genotypes (SNP allele dosages) for each individual and each marker. These genotypes can be used for multiple simulations of SNP array data with different parameters; for each simulation a different file prefix should be entered on the File tabsheet.

In case the genotypes are imported you need to specify what is the ploidy of the population (in case it is an F1 population the parents can have a different ploidy; in that case the ploidy of both parents must be specified, and the parents must be named P1 and P2). This is done on tabsheet Ploidy, which only appears if data are imported.

Output

The output can be located in three different folders: *'Dosages_Simulations'*, *'SNParray_Simulations'*, and *'SequenceReads_Simulations'*. Once simulation is achieved, all related files will be stored in the folder with the name user specified. For users, their simulated datasets can be downloaded from the Download (*.zip) button.

Hope GenoSim can help you achieve your goal. If you would like to cite this paper, please use:

If you encounter any problem, please contact: yanlin.liao@wur.nl OR giorgio.tumino@wur.nl.