R AS THE CORE TOOL FOR DATA ANALYSIS AND VISUALIZATION IN A BREEDING PROGRAM: FROM FIELD AND GENOTYPIC EVALUATION TO INDIVIDUAL SELECTION

key word: experimental design, high-throughput genotyping, association analyses

Abstract

Plant and animal breeding is a process of controlled reproduction and selection to improve characteristics of human interest. The main objective of breeding is to combine favorable characteristics from two parents and pass them on to the next generation. Breeding programs can be found in various fields, including agriculture, where crops are improved for increased yield, nutritional quality and resistance to biotic and abiotic stresses. Livestock breeding aims to select animals with desirable traits such as meat quality, milk production, or health properties.

In particular, plant breeding programs involve careful selection of traits to be improved, selection of crosses, evaluation of offspring, and multiplication of selected plants. Multiple statistical and visualization analysis should be carried out in each stage, and R is a powerful statistical and data analyses software to integrate all of these analyses within the same framework.

Reliable phenotypic evaluation requires well-designed experiments and flexible models for data analysis that include, in addition to design factors, relevant additional information. Typically, data come from field trials, in which many lines are evaluated over a period of time, generating spatial and/or temporal correlations between observations. There are several packages in R for designing more complex trials such as p-rep, alpha-lattice, augmented designs (‘agricolae’, ‘ExpDes’, ‘pRepDesigns’), as well

as for describing or incorporating the recorded environmental information into the analysis model (‘nlme’, ‘lme4’, ‘emmeans’, ‘gstat’, ‘sp’, ‘spplot’, ‘INLA’, ‘INLAspacetime’, among others). Moreover, the ‘Digger’ package offers sophisticated data analysis and optimization tools, greatly enhancing the capacity of breeding programs optimization. It encompasses various criteria of optimality, allowing breeders to fine-tune their selection strategies based on specific objectives and priorities. The ‘ggplot2’ package is essential for data visualization, spatial data analysis, and the study of genotype by environment interaction. Besides, in case of including environmental covariates and needing to perform data imputation, the Apsim package could be used, in order to better explain the behavior of different species in the environment.

Genetic evaluation of individuals in a breeding program requires high-throughput genotyping technology to obtain high-density nucleotide variants (SNP markers) from the DNA of each individual. This information is used in breeding to study genetic diversity and predict phenotypic values of plants that have not been field evaluated. High-throughput genotyping data could be thousands to millions of SNPs on hundreds of individuals, being extremely difficult to manipulate this kind of data with standard computer programs. R is a powerful software to filter, visualize, impute and modify marker genotypic matrices using principal functions of the base package. Presence of diversity between individuals from a breeding program is necessary for improving desirable characteristics through parental crossing. To study genetic diversity there are many packages in R (‘adegenet’,‘hierfstat’, ‘pegas’, ‘factoextra’, ‘princomp’, among others that are included in the base package).

Finally, the selection of lines through the combination of genotypic and phenotypic information implies the identification of genomic regions responsible for desirable characteristics, and prediction of not evaluated individuals performance. To perform these analyses combining genotypic and phenotypic information some packages are used in R (‘GAPIT’, ‘rrBLUP’, ‘GWAS poly’, ‘ggplot2’, ‘sommer’,’BGLR’).

In summary, leveraging experimental design, high-throughput genotyping, and association analysis in breeding programs empowers breeders to make informed decisions, enhance genetic diversity, and improve desirable traits for sustainable agriculture and livestock production. R software, with its extensive range of packages, provides a comprehensive platform for conducting these analyses and accelerating the progress of breeding programs.