Genome simulation study in GWAS models with heteroscedasticity across management regimens

Eugenia Bortolotto[1], *Cecilia Bruno*[2,3]

[1]bortolotto.eugenia@gmail.com, Doctorado en Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario (UNR), Rosario, Argentina

[2]cebruno@agro.unc.edu.ar, Estadística y Biometría. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba (UNC). Ciudad Universitaria, Córdoba, Argentina

[3]cebruno@agro.unc.edu.ar, Unidad de Fitopatología y Modelización Agrícola (UFYMA - CONICET), Córdoba, Argentina

With the goal of evaluate association mapping (AM) models to detect molecular markers significant in cases where the phenotype is measure in more than one environment and the homoscedastic model cannot assumed, we compared Genome-Wide Association Studies (GWAS) models with different levels of variance. In this study, a dataset based on vegetal parameters was simulated in *xbreed* package in R. There is a database of 240 individuals, 79918 SNP-type (Single-Nucleotide Polymorphisms) molecular markers and 15 QTL. A phenotype trait was also simulated with normal distribution in each environment. Then, to get heteroscedasticity, we simulated three phenotypic repetitions for each genotype at each environment (E1 and E2) by adding a random error term normally distributed with zero mean and three different variances across the environment. First setting with σ^2=0.25 in both environment; second setting 0.25 and 2.25 for σ^2 in E1 and E2, respectively; and the last one setting 0.25 and 12.25 for σ^2. To control for heteroscedasticity, we fitted a Linear Mixed Model (LMM) with Environment (E) as fixed effect and Genotype (G) and G×E interaction (GE) as random effects. In addition, several variance structures were tested in each case and compare with Akaike and Bayesian Information Criteria (AIC and BIC, respectively). Then, the Best Linear Unbiased Predictors (BLUPs) of the model are use in the GWAS. The seven AM models evaluated ranged from simple to complex and included: General Linear Model with Principal Component Analysis, GLM-PCA; Linear Mixed Model with PCA+K (Kinship matrix for family relatedness estimates), LMM-PK; Compressed LMM, CLMM; Enriched CLMM, ECLMM; settlement of LMM under progressively exclusive relationship, SUPER; multiple loci LMM, MLMM; and fixed and random model circulating probability unification, FarmCPU; all of these models were fitted in *GAPIT* package in R. We also compare this model with the LMM considering the markers as covariance matrix with *Sommer* package in R. The correlated multiple testing was done by Li and Ji method that is based in eigenvalues of a correlation matrix. We examine Quantile-Quantile (Q-Q) plots to determining if models control false positives and false negatives. Then compare the number of significant markers identified by Li and Ji method in eight different association models and analyse the proportion of significant SNPs that are less than 5cM from a simulated QTL. The MLMM and ECLMM models were the ones that detected true associations at the different levels of heteroscedasticity. The number of false negative is much higher with the GLM-P and FarmCPU models.

Keywords: variance components; molecular markers; genotype-by-environment interaction, variance heterogeneity, best linear unbiased predictor