Effect of using imputed missing data on QTL detection on a wheat GWAS panel

SP Brandariz[1], A González-Reymúndez[1], B Lado[1,2], M Quincke[2], J von Zitzewitz[3], M Castro[2], I Matus[4], A del Pozo[5], L Gutiérrez[1]

Molecular markers are an essential component of plant and animal breeding programs. One inexpensive way of obtaining molecular markers is through Next-Generation Sequencing (NGS). Genotyping-by-sequencing (GBS) is one of the NGS techniques which have been successfully used for complex genomes like wheat. A particularity of GBS is that it generates a lot of missing information which is generally imputed. Imputation is required for Genomic Prediction studies and several studies demonstrate its value. However, the effectiveness of missing data imputation for Genome-wide association (GWAS) studies has not been demonstrated. Data imputation for GWAS where one marker at a time is being studied could potentially create biased estimates. The aim of this study was to compare the effects of using either missing or imputed data for Quantitative Trait Loci (QTL) detection in a wheat GWAS pannel. A set of 384 advanced lines of wheat was included in this study consisting of 186 genotypes from INIA (Instituto Nacional de Investigación Agropecuaria) - Uruguay, 55 genotypes from INIA - Chile and 143 genotypes from CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo). SNPs were obtained using the Tassel-GBS Pipeline. We excluded SNPs with more than 50 % missing data and SNPs with a minor allele frequency (MAF) more extreme than 10%. Sequence database available from the SyntheticxOpata map (synop) was used to construct the maps, obtaining a final data set with more than 18K SNPs. Missing data was handled in three different ways to create the SNP datasets used for QTL detection: 1) data not-imputed, 2) data imputed by the realized relationship matrix method multivariate normal expectation maximization (MVN-EM), and 3) data imputed by the mean. A number of QTL (either 25 or 50) with different heritability (0.2 and 0.7) were simulated on top of each dataset. The following mixed model was used to recover QTL: ,
where y: phenotypic vector, X: SNPs matrix, β: unknown vector of allele effects to be estimated, Q: matrix that relates each measurement to population origin, v: populations vector, Z: kinship matrix, u: vector of random background polygenic effects, and e: residual error. We used a liberal 0.01 significance level. The power to detect QTL was estimated for each dataset and differences among medians of QTL detection power among imputed datasets were studied using the Friedman test and non-parametric contrasts. For this purpose, methods of imputations were defined as treatments and simulation scenarios as blocks. The QTL detection power with the MVN-EM matrix was lower than with the mean imputed matrix or the no imputed matrix. No differences in QTL detection power were found between the mean imputed matrix or the no imputed matrix. Based on our results, imputing does not seem to improve QTL detection power.

u

[1]Departamento de Estadística, Biometría y Cómputos. Facultad de Agronomía, Universidad de la República. Garzón 780, Montevideo 12900, Uruguay,
[2]Programa Nacional de Investigación Cultivos de Secano, Instituto Nacional de investigación Agropecuaria, Est. Exp. La Estanzuela, Colonia 70000, Uruguay,
[3]Secobra Saatzucht GmbH Feldkirchen 3, 85368 Moosburg, Germany,
[4]Instituto de Investigaciones Agropecuarias, Centro Regional de Investigación Quilamapu, Casilla 426, Chillán, Chile,
[5]Facultad de Ciencias Agrarias, Universidad de Talca, Casilla 747, Talca, Chile.
E-mail: agugonrey@gmail.com brandarizsofia@gmail.com