ID 32

An R script for quality control in genome-wide association studies

Roberto Higa, Embrapa Informática Agropecuária

Fabiana Barrichello, Universidade Federal de São Carlos

Simone Niciura, Embrapa Pecuária Sudeste

Sarah Meirelles, Universidade Federal de Lavras

Luciana Regitano, Embrapa Pecuária Sudeste

Recent advances in massive genotyping technology based on SNP (Single Nucleotide Polymorphisms) markers are pushing animal breeding research into a new era where the entire genome is screened in the search for genes which affect traits of economic interest, called genome wide association studies (GWAS). The first step in setting up a GWAS is to perform a quality control analysis (QC) on the genotyped data in order to filter out samples and SNPs not satisfying a previously defined set of criteria [1]. The most common criteria include sample and SNP call rate, minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE). It is also recommended to analyze the heterozygosity and the presence of population structure and outlier samples. However, a different set of criteria and thresholds may be more appropriate for different datasets. We wrote an R script [2] which implements a number of QC criteria, making them available as functions and allowing users to use those QC criteria which are more appropriate to their dataset. In this work, we describe a set of functions implemented in our R script and illustrate its use in a GWAS of cattle meat quality in a Canchim cattle breed population.