AARHUS UNIVERSITY

An Evaluation of Software for Performing GWAS Mixed-model Association Analysis

Zhang Leyi

MSc in Bioinformatics, Aarhus University, Denmark

Supervisor: Doug Speed

2023-06

Contents

Abstract

Genome-wide association studies (GWAS) are statistical methods used to identify associations between genes and traits or specific diseases. Two commonly employed methods for GWAS are linear regression (LR) and mixed model analysis (MMA), each with distinct capabilities in preventing false positives, statistical power, and efficiency. This study explores the advantages and limitations of different software tools in assessing various traits under different conditions, including the number of individuals, the number of causal single-nucleotide polymorphisms (SNPs), heritability, heritability model, and quantitative or binary traits. For linear regression, I selected software tools: Plink and LDAK, while for MMA, I evaluated Bolt-lmm, Bolt-lmm-inf, Regenie, and fastGWA. The population data consisted of 66,688 individuals from a multi-ancestry (admixed) population from the UK Biobank. I employed PCA + K-means clustering to stratify the admixed population into five single-ancestry population datasets to compare the performance of the software tools under different population structures. The evaluated criteria included Type I error, statistical power, time, and memory usage. I also discussed the performance of Multi-ancestry Meta-analysis (MAMA) in assessing the multi-ancestry population by combining association test results from multiple single-ancestry groups. My study revealed that fastGWA is not suitable for assessing multi-ancestry populations but can provide feasible statistical power with efficient time usage when evaluating single-ancestry (homogeneous) populations. Bolt-lmm exhibited the slowest speed but performed well in the evaluation of admixed population. Regenie showed limited performance in controlling Type I error and statistical power, making it less favourable. Furthermore, my study demonstrated that MAMA can enhance statistical power by incorporating a large number of individuals from other ethnic groups when the sample size of a specific ethnic group is insufficient, although it may limit the detection ability for the major population. Moreover, MAMA is not an optimal choice in the presence of significant trait heterogeneity. In comparison to MMA software such as Bolt-lmm, MAMA exhibited a more conservative detection ability in admixed populations. Finally, I utilized the results from MMA analyses to construct a polygenic risk score (PRS) model using the Clumping + Threshold (C+T) method, which enables the prediction of complex trait likelihoods for diverse population genotypes. I tested the performance of PRS by Plink with the C+T method on different population structures and software tools. I suggest that the choice of training method should be based on a specific population structure.

Keywords: GWAS, Evaluation, Linear Regression, Mixed Model Analysis, Multi-ancestry Meta-analysis, Polygenic Risk Score