Abstract

Genome Wide Association Studies (GWAS) aims to find genomic variants that correlates with the occurrence of certain traits within a population. Various tools and algorithms can be utilized to conduct GWAS, among them are PLINK, Hail, and VariantSpark. In this study, a GWAS analysis pipeline was developed using PLINK, Hail, and VariantSpark to assess their performance in conduction GWAS association analysis. The Late-onset Alzheimer's (LOAD) GWAS dataset from a previous study by Webster et al. (2009) was used on these softwares. PLINK ran Simple Association, Linear Regression, and Logistic Regression algorithms, while Hail ran Linear regression and Logistic regression. The novel Random Forest algorithm in VariantSpark was ran and compared with the previous algorithms. The softwares was evaluated based on SNP calling accuracy, RAM usage, and processing time. Overall, PLINK's algorithms performed the best in terms of SNP calling accuracy, RAM usage, and processing time. Hail was comparable with PLINK in terms of RAM and processing time, but PLINK was more accurate in detecting variants. VariantSpark used a lot more RAM and processing time for the same performance as PLINK's Simple Association algorithm. In conclusion, PLINK and Hail are great tools to conduct GWAS Association Testing. More testing with different datasets is needed for more accurate assessments of these tools.

Keywords: GWAS, Late-onset Alzheimer's (LOAD), PLINK, Hail, VariantSpark