Chapter 1

Introduction

1.1 Background

Genome Wide Association Studies (GWAS) is widely used to detect disease-associated genetic variants within a population (Uffelmann et al., 2021). The emergence of GWAS studies has benefitted the scientific community by detecting pathogenic variants in a population for complex diseases such as migraines (Kaur et al., 2019), diabetes (Xue et al., 2018), coronary artery disease (Christiansen et al., 2017), and Alzheimer's disease (Bellenguez et al., 2022). Detection of disease-specific variants allows for further analysis of the disease's mechanisms and its interactions with other biological systems. The primary output of GWAS is the detection of single nucleotide polymorphisms (SNPs) with the most statistically significant p-values as computed from association algorithms (White et al., 2019). In this thesis project, the association tools from PLINK and Hail/VariantSpark will be evaluated based on their efficiency and accuracy in calling variants related to Late-onset Alzheimer's GWAS data.

PLINK is an open-source, C/C++ based, command-line program to process and analyze various genomic data. PLINK's five main domains of functions, which are data management, summary statistics, population stratification, association analysis, and identity-by-descent estimation, make it possible for PLINK to perform association analysis in GWAS studies (Purcell et al., 2007). PLINK is capable of running a simple association, linear regression, or logistic regression algorithm to compute p-values of variants. In GWAS, the typical method for identifying genetic variations linked to continuous phenotypes is linear regression, and in epidemiological studies, it is paired with adjustments for covariates to estimate the independent effect of predictor variables (Wang et al., 2018). However, standard linear regression is unoptimized to adjust for heritable covariates in genetic association analysis. Therefore, GWAS of polygenic diseases would be better suited with more advanced association algorithms.

VariantSpark is a random-forest Machine Learning algorithm based in ApacheSpark that is optimized for association studies of complex phenotypes in genomic datasets. VariantSpark is made to run in tandem with the Python library Hail. Compared to the general ApacheSpark library and other association algorithms, VariantSpark is more efficient in resource utilization, memory usage, and is suited for genomic datasets (Bayat et al., 2020). The main appeal of using a random forest algorithm instead of a linear regression algorithm is to detect polygenic gene interactions of complex diseases (Stephan et al., 2015). As such, VariantSpark's random-forest model is expected to more accurately detect the polygenic variants in the Late-onset Alzheimer's data.

Late-onset Alzheimer's disease (LOAD) is the most common form of Alzheimer's disease, occurring after the age of 65 and accounting for 90% of all Alzheimer's cases (Isik, 2010). As a complex and multifactorial disease, multiple genes are suspected to be involved in the occurrence of Alzheimer's, although LOAD tends to happen sporadically without any family history. Previous genome-wide association studies have identified SNPs and genes that lead to increased risk of LOAD, for example the Apolipoprotein E (APOE) gene on chromosome 19q13 (Moreno-Grau et al., 2018). As the prevalence of Alzheimer's disease continues to rise and conventional treatment options prove to hardly be beneficial (Livingston et al., 2020), the need for early detection methods and genetic susceptibility testing is evident. Therefore, GWAS plays a role to detect variants that increase the risk of Alzheimer's disease.

In this thesis project, the association tools from PLINK, Hail, and VariantSpark will be evaluated based on their efficiency and accuracy in calling variants related to Late-onset Alzheimer's GWAS data. As the datasets for GWAS are large and the genetic variations that may affect the phenotypes are often complex, the need to find an efficient and accurate software to analyze these data is justified. Currently there is no research that evaluates the performance of PLINK compared to Hail and VariantSpark for SNP calling accuracy, RAM usage, and time elapsed in the context of conducting GWAS. Therefore, this thesis project is aimed to fill in this gap in research.

1.2 Objective

The main objective of this thesis project is to develop and evaluate the efficiency and accuracy of using PLINK, Hail, and VariantSpark in a GWAS analysis pipeline. The specific objectives of this thesis project are:

- To develop a GWAS analysis pipeline using PLINK and Hail/VariantSpark
- To run association analysis with PLINK using the Simple Association, Linear Regression, and Logistic Regression algorithm
- To run association analysis with Hail using the Linear Regression and Logistic Regression algorithm, and to run the Random Forest algorithm with VariantSpark
- To evaluate the SNP calling accuracy, RAM usage, and time elapsed of the previously mentioned association analysis pipelines

1.3 Hypothesis

The hypothesis of this study is that the random forest algorithm will perform better than algorithms in PLINK and Hail in terms of variant calling accuracy. PLINK and Hail will perform similarly as they provide the same algorithms (logistic and linear regression), but PLINK will have faster processing speed due to its binary file format.