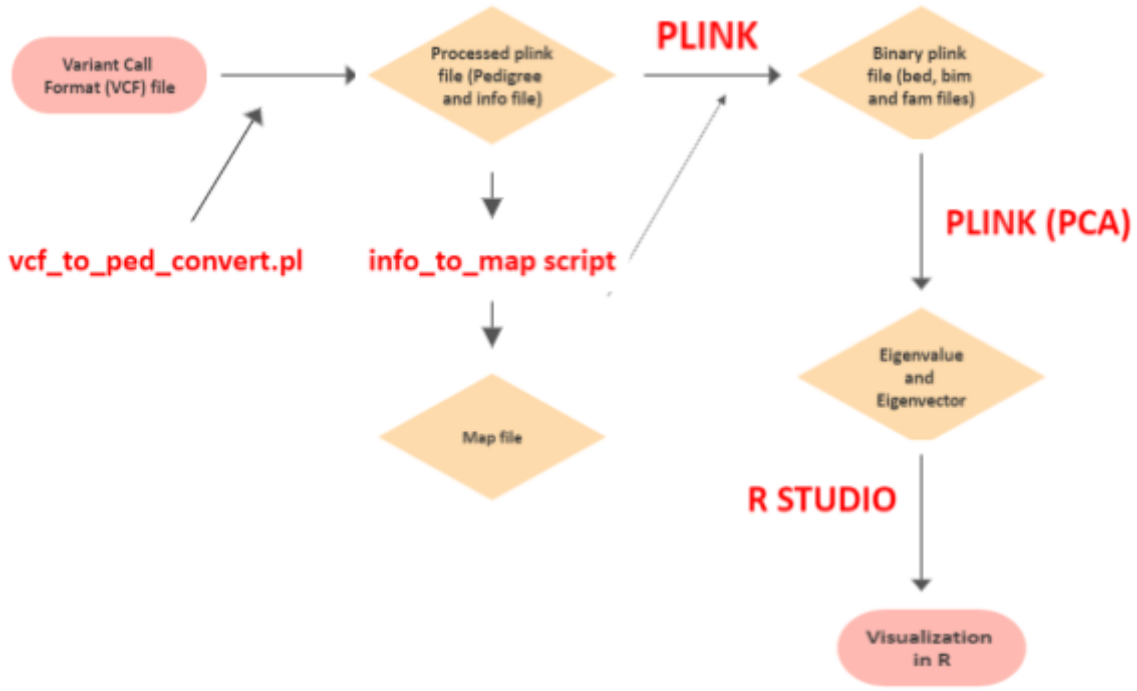
TECHNICAL SPECIFICATION REPORT



METADATA

1. VARIANT CALL FORMAT: VCF is a standardized text file format used for representing SNP, indel, and structural variation calls. Only the vcf for chromosome 1 was used for this project.

2. PED files: A pedigree file is a structured description of the familial relationships between samples. It is the original standard text format for sample pedigree information and genotype calls. The PED file has six fixed columns at the beginning, followed by the SNP information. They are;

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (1=male; 2=female; other=unknown)
6. Phenotype

3. INFO file: An INFO file accompanies a PED file. It is a text file with no header line and one line per variant with two fields:

1. Variant identifier
2. Base-pair coordinate

4. MAP file: The MAP file contains information about every single SNP. Each row corresponds to one SNP in the PED file. The MAP file must have exactly four columns with the following information (the columns should be separated by a space or a tab):

1. Chromosome ID (e.g. Chr1 for Chromosome 1)
2. A unique SNP identifier
3. Genomic distance
4. SNP Position

5. BED file is a text file with no header line and one line per variant with the following fields:

1. Chromosome code
2. Variant identifier
3. Position in morgans or centimorgans
4. Base-pair coordinate

6. BIM file:.bim is an extended variant information file that is accompanied by a BED binary genotype table. It is a text file with the following six fields on one line for each variant and no header line.

1. Chromosome code
2. Variant identifier
3. Position in morgans or centimorgans
4. Base-pair coordinates
5. Allele 1 (corresponding to clear bits in .bed; usually minor)
6. Allele 2 (corresponding to set bits in .bed; usually major)

7. FAM file: a PLINK sample information file that is associated with a BED binary genotype table. It is a text file with no header line and one line per sample with the following six fields:

1. Family ID ('FID')
2. Within-family ID ('IID'; cannot be '0')
3. Within-family ID of father ('0' if father isn't in dataset)
4. Within-family ID of mother ('0' if mother isn't in dataset)
5. Sex code ('1' = male, '2' = female, '0' = unknown)
6. Phenotype value ('1' = control, '2' = case, '-9'/'0'/non-numeric = missing data if case/control)

8. EIGENVEC, EIGENVAL (PRINCIPAL COMPONENTS)

These two files are produced by PCA.

The **Eigenval file** contains one eigenvalue per line.

The **Eigenvec file** is, by default, a space-delimited text file with no header line. The first two columns are the sample's FID/IID, and the rest are principal component weights in the same order as the Eigenval values.

SOFTWARE AND PIPELINES

PLINK

Plink is a free, open-source whole-genome association analysis toolset designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

R Studio

RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It includes a console, a syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging, and managing workspaces.

The VCF to PED converter

This allows the parsing of a vcf file (specification) to create a linkage pedigree file (ped) and a marker information file. There is both an online version of this tool and a perl script (API script)

The info_to_map perl script

This script is designed for the effective conversion of an info file into a map format. This is because the map file format is necessary for generating the binary ped files and not the info files.