Introduction to GWAS studies with PLINK 1.9

A quick introduction on how to perform a GWAS study using variant data at the population-level

From VCF to GWAS

In this part of the tutorial we discuss and provide practical examples on how to perform a Genome-Wide Association study (GWAS) from variants (a single VCF file), produced following steps similar to the example we went over in day 1:

1. Quality control and trimming of raw FASTQ data
2. Mapping to a reference genome
3. SAM/BAM file pre-processing
4. Variant calling (e.g., with FreeBayes)
5. VCF file pre-processing (e.g., filtering and merging with bcftools)
6. VCF file annotation (e.g., with SnpEff and SnpSift)

In this tutorial

- QC procedures and statistical analyses will be illustrated using the open-source whole-genome association analysis toolset PLINK (Purcell et al., 2007)
- The GWAS analysis here is performed using PLINK version 1.9 (Chang et al., 2015) and R, and follows the published tutorial of Marees et al. 2018 (https://github.com/MareesAT/GWA_tutorial/) with minor changes
- Simulated dataset (N=207) with a binary outcome measure from publicly available data of the International HapMap Project
- Detailed documentation for PLINK 1.9 in: https://www.cog-genomics.org/plink/
- Similar software tools:
  - TASSEL5
  - GEMMA

PLINK - Data format

PLINK can either read text-format files or binary files.

- Text PLINK data consist of two files:
  1. A file with information on the individuals and their genotypes (*.ped)
  2. A file with information on the genetic markers (*.map)
- Binary PLINK data consist of three files:
  1. A binary file that contains individual identifiers (IDs) and genotypes (*.bed)
  2. A text file containing information on the individuals (*.fam)
  3. A text file containing information on the genetic markers (*.bim)

PLINK - Data format

*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

Overview of various commonly used PLINK files. Adopted from Marees et al. 2018.

PLINK - Basic commands

- Format and name of input files:
  - --file: For text files
  - --bfile: For binary files
- Output files:
  - --out: Name of the output files
  - --make-bed: Create binary (PLINK format) files as output
- Association analysis:
  - --assoc: X^2 test of association
  - --logistic: logistic regression analysis, optionally with covariates
  - --linear: linear regression analysis, optionally with covariates

PLINK - Quality Control

Overview of seven QC steps that should be conducted prior to genetic association analysis. Adopted from Marees et al. 2018.

Step	Command	Function	Thresholds and explanation
1. Missingness of SNPs and individuals	--geno	Excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed.	We recommend to first filter SNPs and individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02).
	--mind	Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.	Note, SNP filtering should be performed before individual filtering.

PLINK - Quality Control

Overview of seven QC steps that should be conducted prior to genetic association analysis. Adopted from Marees et al. 2018.

Step	Command	Function	Thresholds and explanation
2. Sex discrepancy	--check-sex	Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.	Can indicate sample mix-ups. If many subjects have this discrepancy, the data should be checked carefully. Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.

PLINK - Quality Control

Overview of seven QC steps that should be conducted prior to genetic association analysis. Adopted from Marees et al. 2018.

Step	Command	Function	Thresholds and explanation
3: Minor allele frequency (MAF)	--maf	Includes only SNPs above the set MAF threshold.	SNPs with a low MAF are rare, therefore power is lacking for detecting SNP-phenotype associations. These SNPs are also more prone to genotyping errors. The MAF threshold should depend on your sample size, larger samples can use lower MAF thresholds. Respectively, for large (N = 100.000) vs. moderate samples (N = 10000), 0.01 and 0.05 are commonly used as MAF threshold.

PLINK - Quality Control

Overview of seven QC steps that should be conducted prior to genetic association analysis. Adopted from Marees et al. 2018.

Step	Command	Function	Thresholds and explanation
5: Heterozygosity	-	Excludes individuals with high or low heterozygosity rates.	Deviations can indicate sample contamination, inbreeding. We suggest removing individuals who deviate ±3 SD from the samples' heterozygosity rate mean.

PLINK - Quality Control

Overview of seven QC steps that should be conducted prior to genetic association analysis. Adopted from Marees et al. 2018.

Step	Command	Function	Thresholds and explanation
6: Relatedness	--genome	Calculates identity by descent (IBD) of all sample pairs.	Use independent SNPs (pruning) for this analysis and limit it to autosomal chromosomes only.
	--min	Sets threshold and creates a list of individuals with relatedness above the chosen threshold. Meaning that subjects who are related at, for example, pi-hat >0.2 (i.e., second degree relatives) can be detected.	Cryptic relatedness can interfere with the association analysis. If you have a family-based sample (e.g., parent-offspring), you do not need to remove related pairs but the statistical analysis should take family relatedness into account. However, for a population based sample we suggest to use a pi-hat threshold of 0.2.

PLINK - Quality Control

Overview of seven QC steps that should be conducted prior to genetic association analysis. Adopted from Marees et al. 2018.

Step	Command	Function	Thresholds and explanation
7: Population stratification	--genome	Calculates identity by descent (IBD) of all sample pairs.	Use independent SNPs (pruning) for this analysis and limit it to autosomal chromosomes only.
	--cluster --mds-plot k	Produces a k-dimensional representation of any substructure in the data, based on IBS.	K is the number of dimensions, which needs to be defined (typically 10). This is an important step of the QC that consists of multiple proceedings but for reasons of completeness we briefly refer to this step in the table. This step will be described in more detail in section “controlling for population stratification.”

PLINK – Controlling for population stratification

- Population stratification as a major source of systematic bias in GWAS studies
- Using multidimensional scaling (MDS) approach with PLINK to correct for population stratification
- MDS calculates the genome-wide average proportion of alleles shared between any pair of individuals within the sample to generate quantitative indices (components) of the genetic variation for each individual.
- The individual component scores can be plotted to explore whether there are groups of individuals that are genetically more similar to each other than expected.

PLINK – Controlling for population stratification

- Investigating for which individuals the generated component scores deviate from the sample-target population (ethnic traits)
- Plotting of the scores of the sample under investigation and a population of known ethnic structure (e.g., 1000 genome data)
- Obtain ethnic information for samples and determine possible ethnic outliers
- Remove outliers, perform MDS anew and use the main components as covariates in the association tests to correct for any remaining population stratification

PLINK – Association test

- Depending on the expected genetic model of the trait or disease of interest and the nature of the phenotypic trait studied, the appropriate statistical test should be selected.
- Binary outcome measure:
  - --assoc: option in PLINK performs a X^2 test of association that does not allow the inclusion of covariates
  - --logistic: a logistic regression analysis will be performed which allows the inclusion of covariates
  - The --logistic option is more flexible than the --assoc option, yet it comes at the price of increased computational time.
- Quantitative outcome measure:
  - --assoc: this option will automatically perform an asymptotic version of the usual Student's t-test to compare two means. This option does not allow the use of covariates.
  - --linear: a linear regression analysis with each individual SNP as a predictor.
  - Similar to the --logistic option, the --linear option enables the use of covariates and is somewhat slower than the --assoc option.

PLINK - Correction for multiple testing

- --adjust:
  - Generates output in which the unadjusted p-value is displayed, along with p-values corrected with various multiple testing correction methods (Bonferroni, FDR, etc.)
  - These values tend to be overly conservative when significant Linkage Disequilibrium (LD) is present, with more useful p-values available from e.g., appropriate permutation tests
- --mperm:
  - The outcome measure labels are randomly permuted multiple times which effectively removes any true association between the outcome measure and the genotype
  - An empirical distribution of the test-statistic and the p-values under the null hypothesis of no association are obtained through multiple testing
  - The original test statistic or p-value obtained from the observed data is subsequently compared to the empirical distribution of p-values to determine an empirically adjusted p-value