GWAS analysis

Statistical Genomics and Bioinformatics workshop

Section 1: Examining the files

Open the EUR folder or move to the folder in the terminal/cmd. Let us examine the files present here.

1. **What files are present in the folder EUR? Can you identify each file?**

Answer: EUR.ped, EUR.map, EUR.bed, EUR.bed, EUR.bim, EUR.fam, EUR.height

Section 2: Data management

2.1 Convert file formats

1. **Convert the PED files to BED files in PLINK**

./plink --file EUR --make-bed --out EUR

Please ensure the path to PLINK and the EUR files are provided correctly. i.e., if you are in the EUR folder, provide the file path to PLINK or vice-verse.

2. **How to convert BED to PED?**

./plink --bfile EUR --recode tab --out EUR

2.2 Sample selection

1. **Create a BED file using the selected samples**

./plink --bfile EUR --keep mysamples.txt --make-bed --out EUR_mysamples

What does the log file say? How many samples are present in EUR_mysamples

2. **Remove samples in removesamples.txt in the BED file**

./plink --bfile EUR --remove removesamples.txt --make-bed --out EUR_removedsamples

How many samples are present in EUR_removedsamples?

2.3 SNP selection

1. **Create a Bed file containing only the specific SNPs present in mysnps.txt**

./plink --bfile EUR --extract mysnps.txt --make-bed --out EUR_mysnps

2. **Create a BED file containing SNPs in chromosome 2 between position 30000000 to 35000000 and chromosome 3 60000000 to 62000000**

First create a myrange.txt file with the follow details:

2	30000000	35000000	Range1
3	60000000	62000000	Range 2

./plink --bfile EUR --extract range myrange.txt --make-bed --out EUR_myrange

3. **Remove the snps in the removesnps.txt in the EUR dataset**

./plink --bfile EUR --exclude removesnps.txt --make-bed --out EUR_removedsnps

4. **Remove the snps in myrange.txt in the EUR dataset**

./plink --bfile EUR --exclude range myrange.txt --make-bed --out EUR_removedrange

5. **Select SNPs on specific chromosomes such as 1-22**

./plink --bfile EUR --chr 1-22 --make-bed --out EUR1-22

./plink --bfile EUR --chr 1 --make-bed --out EURchr1

2.4 Without creating bed-files

Often, we do not have the space to create bed files during every. We create files intermediate files that contain the information

1. **Exclude samples listed in removesamples.txt but do not create a bed file**

./plink --bfile EUR --remove removesamples.txt --make-just-fam --out EUR_removedsamples_famonly

This produces a file called "EUR_removedsamples_famonly.fam". To make bed file, use this file and "keep" flag in Sample selection step 1

2. **Exclude SNPs present in the ranges in reported in myrange.txt. Write out the selected snp list without making bed file**

./plink --bfile EUR --exclude range myrange.txt --write-snplist --out EUR_removedrange_snplist

This creates a file called "EUR_removedrange_snplist.snplist". Use this file and extract flag in SNP selection step 1 to create a bed file.

HINT - ./plink --bfile EUR --extract EUR_removedrange_snplist.snplist --make-bed --out EUR_extractedONLY

Section 3: Quality control

1. **Let us identify and discard samples with low genotyping rate and low quality SNPs including those missing, with maf < 0.01, and hwe < 1e-06**

You can combine multiple flags at once.

./plink --bfile EUR --mind 0.01 --make-just-fam --maf 0.01 --geno 0.01 --hwe 1e-06 --write-snplist --out EUR_QC1

Flags and parameters:

Flag	Parameter	Meaning

--mind	0.01	Removes samples with missing genotype rate greater than the parameter (0.01 or 10%)
--maf	0.01	Discards SNPs with minor allele frequency < 0.01
--geno	0.01	Discards SNPs with missing genotype rate greater than 0.01 or 10%
--hwe	1e-06	Filters out all variants which have Hardy-Weinberg equilibrium exact test p-value below the provided threshold
--make-just-fam	-	Writes out a fam file with the selected samples
--write-snplist	-	Writes out the list of selected SNPs
--out	-	Output file name

2. Prune snps

We will remove SNPs that are highly correlated. Briefly, it uses the first SNP (in genome order) and computes the correlation with the following SNPs (e.g., 199 SNPs). When it finds a large correlation, it removes one SNP from the correlated pair, keeping the one with the largest minor allele frequency (MAF). Using this list of pruned SNPs, we will perform further QC.

./plink --bfile EUR --keep EUR_QC1.fam --extract EUR_QC1.snplist --indep-pairwise 200 50 0.25 --out EUR_QC2

Parameters and flags:

Flag	Parameter	Meaning
--keep	<file>	Keeps the samples present in the file provided as parameter
--extract	<file>	Keeps the SNPs present in the file provided
--indep-pairwise	200 50 0.25	Estimates pairwise correlation between SNPs and remove one from the pair that are highly correlated. 200 is the window size 50 is step size 0.25 is LD or correlation threshold
--hwe	1e-06	Filters out all variants which have Hardy-Weinberg equilibrium exact test p-value below the provided threshold
--out	-	Output file name

This command results in two files: *EUR_QC2.prune.in* and *EUR_QC2.prune.out.* The prune.in file contains the list of selected representative SNPs.

3. Remove samples with high heterozygosity rate.

High heterozygosity is due to sample contamination. To remove those with high heterozygosity rate, we will first estimate the heterozygosity rates for each sample using the flag *--het*.

./plink --bfile EUR --keep EUR_QC1.fam --extract EUR_QC2.prune.in --het --out *EUR_QC3_SelectedSamples*

This creates a file called EUR_QC3.het containing the heterozygosity rate (column F).

Using your preferred data analysis program (R or excel) identify samples with F within 3 standard deviations of sample mean. You can find the list of selected samples in EUR_QC3_SelectedSamples.txt.

4. Identify samples with discrepancy between reported sex and genetically predicted sex

Discrepancy between the reported sex and genetically predicted sex could be due to low sample quality. We use the flag *--check-sex* to identify those samples with discrepancy.

./plink --bfile EUR --keep EUR_QC3_SelectedSamples.het --extract EUR_QC2.prune.in --check-sex --out EUR_QC4

The above command creates a file *EUR_QC4.sexcheck*. The samples with status PROBLEM indicate discrepancy.

How many samples have been reported to have discrepancy in reported sex and genomic sex?

Answer: 4

Now, we need to remove the samples marked as “PROBLEM”. For the purpose of the tutorial, we have listed these samples in file SexdiscrepantSamples as well as created a new file called EUR_QC4_SelectedSamples.txt that do not contain the sex discrepant samples. The EUR_QC4_SelectedSamples.txt was created using EUR_QC3_SelectedSamples.txt

5. Remove related samples

Having related individuals can pose problems in association analysis. Hence, these have to be removed. We use the flag *--rel-cutoff* to estimate the relatedness between samples. The associated parameter is the relatedness measure threshold known as Pi_hat. We ask PLINK to remove those individuals with relatedness greater than 0.125.

./plink --bfile EUR --keep EUR_QC4_SelectedSamples.txt --extract EUR_QC2.prune.in --rel-cutoff 0.125 --out EUR_QC5

The above command creates a file *EUR_QC5.rel.id* that has all related samples removed.

How many samples removed due to relatedness?

Answer: 0

6. Principal components

It is common practice to include genotype-based principal components (PCs) in GWAS. These PCs represent the population structure and sample ancestry. As population structure can induce confounding in GWAS analysis. The flag *--pca 10* a file with first 10 PCs. The output includes two files: .eigenvec containing the first 10 PCs and .eigenvals containing the eigen values.

./plink --bfile EUR --keep EUR_QC5.rel.id --extract EUR_QC2.prune.in --pca 10 --out EUR_PCs

7. **Creating a bed file with the good quality SNPs (identified in step 1) and good quality samples.**

We create a bed file based on the selected samples and SNPs (from section 3 step 1). GWAS generally are conducted on autosomes and hence, we restrict the SNPs to chromosomes 1 to 22.

./plink --bfile EUR --chr 1-22 --keep EUR_QC5.rel.id --extract EUR_QC1.snplist --make-bed --out EUR_QCFinal

How many samples and variants are present in the final EUR QC file?

Answer: 540534 variants and 483 people

Section 4: GWAS analysis

1. Regression analysis with covariates

Given a quantitative phenotype and possibly some covariates (in a --covar file), --linear writes a linear regression report to .assoc.linear. If it is categorical case/control phenotype, --logistic performs logistic regression given the phenotype and some covariates.

./plink --bfile EUR_QCFinal --pheno EUR.height --covar EUR_PCs.eigenvec --linear --out EUR_HeightGWAS

This results in a file called EUR_HeightGWAS.assoc.linear that contains the results of the association analysis. You will notice that the result file also includes association results for covariates. These can be hidden using the flag *hide-covar*.

./plink --bfile EUR_QCFinal --pheno EUR.height --covar EUR_PCs.eigenvec --linear hide-covar --out EUR_HeightGWAS_onlyADD

 # to reduce size of file and select only significant SNPs choose a pfilter

./plink --bfile EUR_QCFinal --pheno EUR.height --covar EUR_PCs.eigenvec --linear hide-covar --pfilter 1e-5 --out EUR_HeightGWAS_onlyADD

Note: If the pheno file has more than 1 phenotype, then if either flag is used with --all-pheno, the type of regression will automatically adapt based on whether the current phenotype is case/control or not.

How many SNPs are genome-wide significantly associated with height?

Answer: 0

How many SNPs have suggestive significance?

Answer: 4

CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
2	rs17672635	105286334	A	ADD	472	0.5106	4.491	8.99E-06
3	rs67163263	17153361	A	ADD	472	-0.7403	-4.676	3.85E-06
8	rs12547998	3506728	G	ADD	472	0.7006	4.636	4.64E-06
20	rs2425873	44963439	G	ADD	472	-0.3086	-4.571	6.26E-06