Chapter 8

Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis

Miguel E. Rentería, Adrian Cortes, and Sarah E. Medland

Abstract

Within this chapter we introduce the basic PLINK functions for reading in data, applying quality control, and running association analyses. Three worked examples are provided to illustrate: data management and assessment of population substructure, association analysis of a quantitative trait, and qualitative or case–control association analyses.

Key words Data management, Population stratification, Genetic association, Binomial trait, Quantitative trait, Multidimensional scaling, PLINK

1 Introduction

PLINK is a freely available, widely used open-source toolset for genetic association that allows for the study of large datasets of genotypes and phenotypes [1]. It was initially developed in 2007, when genome-wide association was a very new concept. Prior to this time the most commonly used method for genome level analysis had been linkage analysis that typically used a set of ~400 markers from which identity by descent information was estimated either directly at the marker or, through inference at a 1–2 cM grid, yielding ~3,500 positions at which analysis would be run. With the development of chip-based genome-wide association studies (GWAS) arrays the volume of data and analyses quickly increased by several orders of magnitude and many researchers were daunted by the sheer volume of data. This historical context has shaped the development of PLINK. The ambitious aim of Purcell et al. [1] was to create a single package that could seamlessly integrate data manipulation, quality control, analysis, and annotation. While there are now many programs that can be used for analyses, there remain only a small number of programs for manipulation and quality control of GWAS data and the use of PLINK has become almost ubiquitous for this

Cedric Gondro et al. (eds.), *Genome-Wide Association Studies and Genomic Prediction*, Methods in Molecular Biology, vol. 1019, DOI 10.1007/978-1-62703-447-0_8,

type of analysis. The discipline is currently facing an analogous transition point as exome and genome-wide sequence level data become more affordable, and Purcell and colleagues have responded to this new challenge by developing PLINK-seq [1], which we expect will become as popular as PLINK has been.

The omnibus nature of PLINK makes it difficult to provide a comprehensive overview within a single chapter. Our focus here is on four main functional domains: data management, summary statistics, population stratification and estimation of relatedness, and association analysis. We have developed a series of complementary exercises designed to demonstrate these commonly used PLINK functionalities. Readers are referred to the extensive PLINK website for more information on other procedures (http://pngu.mgh.harvard.edu/~purcell/plink/).

1.1 Getting Started

PLINK is a command line program written in C/C++. Binaries and source code can be downloaded from http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml. For information on how to install it on your computer, please refer to the notes on the download page.

As a command line program, all commands involve typing *plink* (*see* **Note 1**) at the command prompt (e.g., DOS window or Unix terminal) followed by the desired options which specify the data files and methods to be used (all prefaced with two minus signs, i.e.,--). In addition to its functions for genetic association, PLINK provides a simple interface for recoding, reordering, merging, flipping DNA-strand, and extracting subsets of data and a number of other versatile functions.

In addition to the command line version, a number of R interfaces have been developed (*see* http://pngu.mgh.harvard.edu/~purcell/plink/rfunc.shtml). There is also a simple Java-based graphic user interface (GUI), gPLINK which provides access to the more commonly used PLINK commands. To learn more about this package, visit: http://pngu.mgh.harvard.edu/~purcell/plink/gplink.shtml

This chapter covers the basics of data management and manipulation in PLINK through three exercises that illustrate how to check for population stratification and basic genetic association analyses. The materials section provides an introduction to the file formats used in PLINK and provides a step-by-step tutorial for data management.

2 Materials

PLINK can accept data in a number of file formats. The most common are generally variations to the basic ASCII-based *linkage file format* PED/MAP and the binary *PLINK file formats* BED/BIM/FAM.

PED files (e.g., genotypes.ped) are white-space (space or tab) delimited text files that contain phenotype and genotype data and are typically structured as follows (*see* **Notes 2** and **3**):

Column 1	Family ID (FID)
Column 2	Individual ID (IID)
Column 3	Paternal ID (PID)
Column 4	Maternal ID (MID)
Column 5	Sex
Column 6	Phenotype
Column 7...n	Genotype(s)

Depending on the nature of the study and other programs that might have been used to produce the data, there are many options that can be included to allow for variations on this format. For example, in many case-control studies data are stored without family or parental IDs, if you wished to use a file without these columns you could do so by adding the *--no-familyID* and *--no-parentalIDs* options to the command line. Similarly, many genotype files do not contain a phenotype, rather than add a dummy phenotype to the file the user can simply add the *--no-pheno* option to the command line.

There are however, some restrictions on the contents of these variables:

- FIDs, IIDs, PIDs, and MIDs must be alphanumeric and unique for each family/individual.
- By default, sex is coded: 1 = male; 2 = female; other=unknown. If an individual's sex is unknown, any character other than 1 or 2 can also be used.
- A PED cannot contain more than one phenotype, and if a phenotype is included it must be placed in the sixth column. A phenotype can be either binary (i.e., affection status; by default, 1 = unaffected, 2 = affected, 0 = missing) or quantitative. PLINK will automatically detect which type (i.e., If a value other than 0, 1, 2, or the missing genotype code is observed, PLINK assumes the phenotype is a quantitative trait).
- For quantitative traits, the missing phenotype value is, by default, −9, but this can be changed to any integer number by using the *--missing-phenotype* option.

By default, PLINK assumes markers are biallelic. Columns 7 and 8 contain the genotype pair at SNP1; Columns 9 and 10 contain the genotype pair at SNP2; and so on. All SNPs (whether

haploid or not) must have two alleles specified. In the case of missing data, *both* alleles should be missing (i.e., 0) or neither. For haploid chromosomes, i.e., chromosomes X and Y in males or mitochondrial DNA, genotypes should be entered as homozygotes. The default missing genotype character (0) can be changed with the *--missing-genotype* option. Compound genotypes, that is data in which the two alleles are separated by a slash (i.e., A/C) or concatenated (i.e., AC), can be incorporated by adding the *--compound-genotypes*. PLINK can also read genotypic data that has been coded numerically, either with reference to a given allele (i.e., 1 = major and 2 = minor allele) or in alphabetical order (i.e., 1 = A, 2 = C, 3 = G, 4 = T) using the appropriate options.

PLINK can read other variations on the standard PED file, including transposed files (that are used by BEAGLE and IMPUTE), in which the data for SNPs is contained in rows and each column represents an individual. More information about methods for working with these alternative file-formats can be found on the PLINK website (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml).

MAP files (e.g., genotypes.map) are white-spaced (space or tab) delimited text files that contain the chromosomal positions of each SNP that has been genotyped, and typically are structured as follows:

Column 1	Chromosome number
Column 2	SNP identifier (rs# or SNP identifier)
Column 3	Genetic distance (in Morgans)
Column 4	Physical base-pair position (bp units)

As with PED files, there are a number of variations available for map files. For example, genetic distance can be specified in centimorgans with the --cm option. Alternatively, you can use a MAP file that does not contain the genetic distance by adding the --map3 option. If working with human data, autosomes are defined by numbers 1–22. Additionally, the following codes are used to specify other chromosome types: X chr. = 23; Y chr. = 24; pseudo-autosomal region of X (XY) = 25; and MT chromosome=26. Comments can be added to a PED or MAP file by starting the line with a # character. PLINK also supports several nonhuman species. To this end, the following flags can be added: --dog, --horse, --cow, --sheep, --rice, or --mouse.

Each row in the MAP file corresponds to two (or more, if working with multiploid genomes) columns of the PED file, and these need to be in the same order (e.g., the SNP described in row 1 of the MAP file will be assumed to be the one for which genotypes are contained in Columns 7 and 8 of the PED file, and the SNP described in row 2 of the MAP file will be assumed to correspond to the genotypes contained in Columns 9 and 10 of the PED file, etc.).

PLINK has a number of options for reading data. The most common is the –-file function, as in the following example, in which two files (mydata.ped, and mydata.map) will be imported into PLINK:

```
plink --file mydata
```

If the PED and MAP files have different prefixes, they can be specified separately, with the --ped and --map options, as in the following example:

```
plink --ped mydata.ped --map chr1.map
```

To save space and time, it is possible to convert the data to a binary PED file (*.bed). This will store the pedigree/phenotype information in a separate file (*.fam) and create an extended MAP file (*.bim), which contains allelic information that is not stored in the BED file. The prefix of the output files can be specified using the --out option. For example: *plink --file mydata --make-bed --out chr1* would read in the *mydata.ped* and *mydata.map* files and create four files:

chr1.bed	Compressed binary file that contains genotype information
chr1.fam	Contains the first six columns of mydata.ped (FID, IID, PID, MID, sex, and phenotype) if any of these columns are not included in the PED file, they will be set to missing in the FAM file
chr1.bim	Extended MAP file: contains two extra columns = allele names
chr1.log	Contains a running log of the PLINK job which includes basic summary information and a list of the options requested

Often, it is useful to filter out SNPs from datasets based on quality control parameters such as minor allele frequency (MAF) or missingness. This can be achieved using the *--maf* and *--geno* functions, for example:

```
plink --file mydata --make-bed --maf 0.02 --geno 0.1
```

would exclude SNPs with a MAF of 2 % and SNPs with more than 10 % missingness.

To read these file formats (*.bed, *.fam, *.bim) back into PLINK, the --bfile option is used instead of --file, i.e., *plink --bfile mydata*. It is also possible to specify these files separately:

```
plink --bed file1.bed --fam file2.fam --bim file3.bim
```

Data can be converted from *binary* to *linkage* format using the *--recode* function, which will create simple PED and MAP (instead of binary) files. The *--bfile*, *--make-bed*, and *--out* functions are constantly used when subsetting data, performing quality control

procedures and analysis. To limit the amount of memory being used, PLINK does not store data for an "analysis session," rather PLINK uses a sequential approach in which each QC or analysis step is performed as a new job. For example, to generate a clean dataset that has been filtered with the *--maf* and *--geno* options so that it can be used for analysis or additional QC the user needs to add the *--make-bed* and *--out* options to their job.

3 Methods

This section comprises two exercises. In the first exercise we will obtain publicly available data for samples from three different populations (CEU, YRI, and JPT + CHB) of the HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) and introduce some data management features before performing multidimensional scaling (MDS) analysis to quantify population structure of the sample. The second exercise focuses on genetic association, with two examples: copy number variation (CNV) data in the Central European cohort (CEU) of the HapMap Project as a quantitative trait, and association in a simulated case–control dataset.

3.1 Exercise 1: Data Management and Population Stratification

Population stratification, also referred to as population substructure, refers to the presence of systematic differences in allele frequencies between subpopulations within a sample, possibly due to different ancestry, which results from nonrandom mating between groups (e.g., this is often explained by physical separation, as in the case of populations of African and European descent) followed by genetic drift of allele frequencies in each group. The presence of population stratification is a problem for association studies, as it increases type 1 error and leads to spurious results. This is particularly true when both the genotypic and phenotypic data differ between populations.

If the structure of a population is known, or a putative structure is found, there are a number of possible ways to control for this in the association studies and thus ameliorate these population biases. Several methods exist, such as genomic control, structured association, or principal component analysis-based methods. In this exercise, publicly available genotype data from three different populations (Central European [CEU], Yoruba in Ibadan, Nigeria [YRI], and the East Asian combined sample of Japanese in Tokyo and Han Chinese in Beijing [JPT + CHB]) collected by the HapMap project will be downloaded, filtered by chromosome, and merged into a single file to estimate pair-wise identity-by-state (IBS) and MDS values. All scripts needed to perform these analyses and to plot the results are provided on the publisher's website (http://extras.springer.com/).

1. Go to the resources section in PLINK's website (data also available from the book's website):

   http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml

   Download the following files:

Population	File
HapMap2 (rel 23) CEU	hapmap_CEU_r23a_filtered.zip
HapMap2 (rel 23) YRI	hapmap_YRI_r23a_filtered.zip
HapMap2 (rel 23) JPT + CHB	hapmap_JPT_CHB_r23a_filtered.zip

   These files contain filtered data for the founders of the CEU, YRI, and JPT + CHB populations in binary PLINK format (*.bed, *.bim, and *.fam), consisting of genome-wide data for 60 CEU, 60 YRI, and 90 JPT + CHB HapMap samples for over 2.5 million SNPs.

2. After downloading and uncompressing the data files, use the --chr and --make-bed functions to subset the data and create a more manageable dataset (composed of only SNPs in chromosome 1) to work with throughout this exercise:

```
plink --bfile hapmap_CEU_r23a_filtered
--chr 1 --make-bed --out CEU_chr_1

plink --bfile hapmap_YRI_r23a_filtered
--chr 1 --make-bed --out YRI_chr_1

plink --bfile hapmap_JPT_CHB_r23a_filtered
--chr 1 --make-bed --out JPT_CHB_chr_1
```

   The above commands will create binary PED (.bed) files (and their corresponding .bim and .fam files) for the three populations. Table 1 lists other options PLINK offers to filter datasets by SNPs.

   Likewise, PLINK offers a number of functions to filter a dataset by individuals in the sample. These include the options listed in Table 2 below.

3. To merge two different datasets, PLINK offers the --merge function (and its binary version --bmerge). These functions take the names of the files to be merged in the following order: [.ped, .map] or [.bed, .bim, .fam], respectively.

```
plink --file data1 --merge data2.ped data2.map
--recode --out data_merged

plink --bfile data1 --bmerge data2.bed data2.bim
data.fam
--make-bed --out data_merged
```

Table 1
PLINK options to filter by SNPs

Option	Action
--chr VALUE	Only analyze or retain SNPs in chromosome VALUE
--exclude filename	Exclude SNPs in the file, one SNP identifier per line
--include filename	Only analyze or retain SNPs in the file
--exclude filename --range	Exclude regions specified in the file, one region per line. Regions are defined by four fields: chromosome, start, end, and region name. For example, 8 8000000 12000000 R1
--from-kb VALUE/ --to-kb VALUE	Only analyze or retain SNPs between the regions defined. Needs the chromosome options
--thin VALUE	Randomly select SNPs. A value of 0.4 will output only 40 % of the markers

Table 2
PLINK options to filter by individuals

Option	Action
--keep {indlist}	Keep only these individuals
--remove {indlist}	Remove these individuals
--filter-controls	Include only controls
--filter-cases	Include only cases
--filter-males	Include only males
--filter-females	Include only females
--filter-founders	Include only founders
--filter-nonfounders	Include only nonfounders

To merge more than two datasets, use the --merge-list function, which takes a filename as a value. This file must contain the names of the datasets to be merged, one dataset per line. Note that with this function either plain text or binary PLINK files can be used. In our example, the CEU_chr_1 dataset will be merged with the corresponding sets of the YRI and JPT + CHB populations.

The file named **merge_list.txt** contains the following text:

```
JPT_CHB_chr_1.bed JPT_CHB_chr_1.bim
JPT_CHB_chr_1.fam
YRI_chr_1.bed YRI_chr_1.bim YRI_chr_1.fam
```

Table 3
PLINK options when merging two datasets

1	Only keep genotypes that match
2	Only overwrite calls which are missing in original PED file
3	Only overwrite calls which are not missing in new PED file
4	Never overwrite
5	Always overwrite mode
6[a]	Report all mismatching calls (diff mode—do not merge)
7[a]	Report mismatching non-missing calls (diff mode—do not merge)

[a]These options are particularly useful when computing concordance rates between two datasets with overlapping samples and markers

To create a single merged dataset with data of the three populations, use the following command:

```
plink --bfile CEU_chr_1 --merge-list merge_list.
txt
--make-bed --out hapmap_chr_1
```

By default, if an individual is present in two files, SNPs present in one file will be set to missing if not present in one of the others, and any existing genotype data (i.e., in CEU_chr_1.bed) will not be overwritten by data in the second file (JPT_CHB_chr_1.bed). To change this behavior, the --merge-mode {} function provides seven different merging alternatives (*see* Table 3):

4. The following command will drop any SNPs with missingness ≥5 %. These SNPs with low call rates have mostly resulted from the merging process:

```
plink --bfile hapmap_chr_1 --geno .05
--make-bed --out data
```

5. To estimate population stratification, PLINK offers tools to cluster individuals into homogeneous subsets (which is achieved through complete linkage agglomerative clustering based on pair-wise IBS distance) and to perform classical MDS to visualize substructure and provide quantitative indices of population genetic variation that can be used as covariates in subsequent association analysis to control for stratification, instead of using discrete clusters. Generally, the *--mds-plot* option is used in conjunction with the *--cluster* function.

   The following commands will estimate the degree of population stratification within the samples under analysis (*see* **Note 4**):

```
plink --bfile data_chr1 --genome --out genome_chr1
plink --bfile data_chr1 --read-genome geno-
me_chr1.genome
--cluster --mds-plot 4 --silent --out mds
```

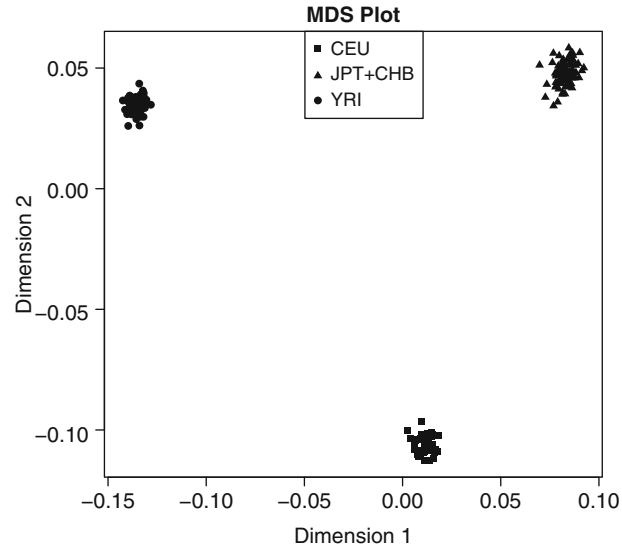

Fig. 1 Multidimensional scaling (MDS) plot of three populations: CEU (Central European), YRI (Yoruba in Ibadan, Nigeria), and JPT + CHB (combined sample of Japanese in Tokyo and Han Chinese in Beijing)

6. Finally, plot the results using R (http://www.r-project.org/) as shown in Fig. 1:

```
d <- read.table("mds.mds",header=T)

cols <- rep("gray",nrow(d))
cols[grepl("^1",d$FID)] <- 'red'
cols[grepl("^NA",d$FID)] <- 'green'
cols[grepl("^Y",d$FID)] <- 'blue'

pchs = rep(19,nrow(d))
pchs[grepl("^1",d$FID)] <- 15
pchs[grepl("^NA",d$FID)] <- 17
pchs[grepl("^Y",d$FID)] <- 19

pdf("Fig1.pdf")

plot(d$C1,d$C2,
   col='black',
   xlab='Dimension 1',
   ylab='Dimension 2',
   main='MDS Plot',
   pch=pchs)

legend('top',
    legend=c("CEU","JPT+CHB","YRI"),
    col='black',
    pch=c(15,17,19))

dev.off()
```

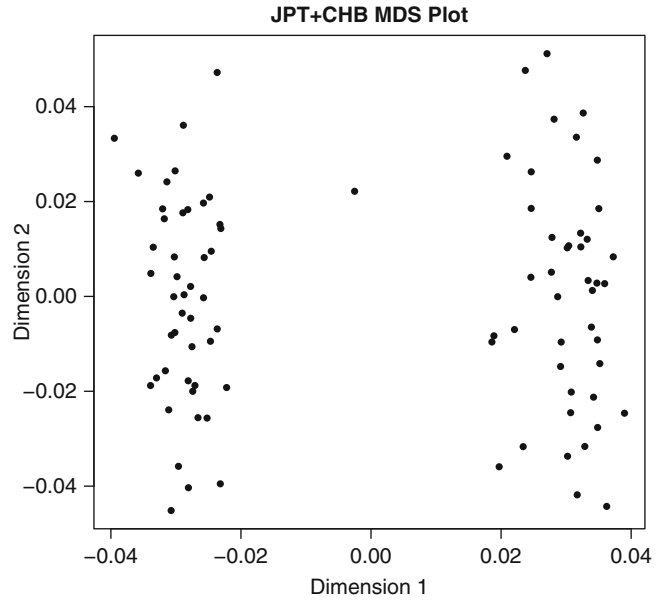

Fig. 2 MDS plot of the combined East Asian (JPT + CHB) populations

As expected, the analyses of these 3 diverse datasets yielded 3 distinct data clusters (on the left). However, in practice, association analyses are generally conducted in rather homogenous populations (e.g., people of European or Asian descent only). For instance, Fig. 2 shows an MDS plot for the JPT + CHB group. Although all individuals were clustered within the same group, the Chinese and Japanese are still distinguishable, and hence MDS coordinates could be used as covariates if controlling for this stratification was needed. The following exercise will provide an example of how to do this.

3.2 Exercise 2: More Data Management Options and Genetic Association

In this section, two genetic association analyses will be conducted. Firstly, we will conduct a quantitative trait GWAS on CNV using publicly available data from the HapMap CEU cohort. This will be followed by a GWAS of a simulated binary affection status (case/control) phenotype.

3.2.1 Quantitative Trait Association

1. CNV data will be used to illustrate how to perform genetic association of a quantitative trait with PLINK. CNV data for the HapMap CEU samples can be found on the HapMap ftp server:

   ftp://ftp.ncbi.nlm.nih.gov/hapmap/cnv_data/hm3_cnv_submission.txt

   The original CNV dataset contains 856 CNV genotypes. For this exercise, we have selected three CNVs (shown in the table below), which are tagged by genotyped SNPs. This data is contained in the **CNV_phenos.txt** file (Table 4).

Table 4
CNV_phenos.txt file example

CNV	Chr	Start pos	End pos
HM3_CNP_35	1	151028547	151035324
HM3_CNP_211	4	34462895	34501120
HM3_CNP_516	9	29084549	29087680

2. Genotype data for this exercise has been prepared and is provided in binary PED format (**CEU_HapMap_GWAS_data.bed**, **CEU_HapMap_GWAS_data.bim** and **CEU_HapMap_GWAS_data.fam**). The original HapMap dataset contains 3,907,239 SNPs for 174 individuals, of which 6 were ethnic outliers and 9 of whom had no CNV data available.

   The script **Exercise_2.sh**, also provided, contains all the commands needed to download and parse the genotype data.

3. The *--pheno* option allows for the specification of alternative (one or more) phenotypes. When using the *--pheno* option, the original PED file must still contain a phenotype Column 6 (even if this is a dummy phenotype, e.g., all missing), unless the *--no-pheno* flag is given. Also, note that the file can contain a header which specifies column name; in this case, the column name can be used with the option *--fillme* to limit the analysis to only this alternative phenotype.

4. To conduct genetic association on all three traits with the *--pheno* option, use the following command:

```
plink --bfile CEU_HapMap_GWAS_data --assoc
--allow-no-sex --pheno CNV_phenos.txt
--all-pheno --missing-phenotype -9 --adjust
--ci 0.95 --out cnv_qtl
```

   Explanations for all PLINK options used in the above command are provided in Table 5:

5. PLINK will generate three pairs of results files (e.g., results for the first CNV [HM3_CNP_35] are written to the files: cnv_qtl.P1.qassoc and cnv_qtl.P1.qassoc.adjusted). The first lines of a *.qassoc output file are shown in Fig. 3 below.

   Table 6 describes the output of each column.

   The **.qassoc.adjusted* files contain adjusted *p*-values with different routines for all SNPs tested and printed in ascending order (most significant SNPs at the top of the file). Figure 4 contains summary results of the six most significant SNPs associated with the HM3_CNP_35 CNV genotype (this corresponds to the first six SNPs in the *adjusted file).

   SNP rs1048535 showed the strongest association with CNV HM3_CNP_35, which is located in chr1: 151028547

Table 5
PLINK commands used for genetic association

Option	Function
--bfile *filename*	Specify *filename*.bed, *filename*.bim, and *filename*.fam files containing binary PED and MAP data
--assoc	Perform association. PLINK will automatically detect whether the phenotype is a binary trait or a continuous trait. If the phenotype contains only 0, 1, or 2 entries, then it is assumed that the trait is binary (0 = missing; 1 = unaffected; 2 = affected). Otherwise, the trait is assumed to be quantitative. In the data provided here, we have added 1 to all entries to prevent PLINK interpreting the trait as binary
--allow-no-sex	Do not include gender as a covariate. If the dataset contains missing gender entries, PLINK will not perform association analysis unless this flag is present
--pheno *filename*	Use the phenotype(s) contained in this file, as opposed to data in the sixth column of the PED/FAM file
--missing-phenotype *value*	The default missing-phenotype value is −9. This option allows for the specification of a different value, which must be an integer
--adjust	Output adjusted *p*-values for multiple testing correction. SNPs are printed in increasing order of significance
--ci *value*	Include 95 % confidence interval of the odds ratio or beta coefficient. This will also force PLINK to print the standard error of the estimate which can then be used for meta-analysis
--out *prefix*	Specify output root filename
--all-pheno	Perform analysis for all phenotypes (all columns) of the file specified with the --pheno option

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
1	rs6650104	554340	158	-0.4968	0.2746	0.02054	-1.809	0.07241
1	rs12565286	711153	154	-0.1962	0.1875	0.00715	-1.046	0.2971
1	rs3094315	742429	157	-0.1063	0.08354	0.08354	-1.273	0.205
1	rs3131972	742584	159	-0.1284	0.08475	0.01441	-1.515	0.1317

Fig. 3 Example of a .qassoc output file

–151035324. The *.qassoc.adjusted* file only contains significance values. To extract information about this SNP from the *.qassoc* file (such as chromosome, location, etc.), a search function can be used in a text editor or directly from a command line. For instance (in Linux):

```
grep rs1048535 cnv_qtl.P1.qassoc
```

Table 6
Output columns

CHR	Chromosome number
SNP	SNP identifier
BP	Chromosomal position (base-pair)
NMISS	Number of non-missing genotypes
BETA	Regression coefficient
SE	Standard error
R^2	Regression *r*-squared
T	Wald test (based on *t*-distribution)
P	Wald test asymptotic *p*-value

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	rs1048535	8.357e-103	1.111e-94	1.092e-96	1.092e-96	INF	INF	1.092e-96	1.601e-95
1	rs7524281	1.163e-79	5.604e-72	1.519e-73	1.519e-73	INF	INF	7.597e-74	1.114e-72
1	rs11586156	3.351e-68	6.634e-61	4.379e-62	4.379e-62	INF	INF	6.256e-63	9.172e-62
1	rs1591077	3.351e-68	6.634e-61	4.379e-62	4.379e-62	INF	INF	6.256e-63	9.172e-62
1	rs10494277	3.351e-68	6.634e-61	4.379e-62	4.379e-62	INF	INF	6.256e-63	9.172e-62
1	rs12239774	3.351e-68	6.634e-61	4.379e-62	4.379e-62	INF	INF	6.256e-63	9.172e-62

Fig. 4 Summary results for the six most significantly associated SNPs with the HM3_CNP_35 CNV genotype (as reported in the .adjusted output file)

will output the following line:

1	rs1048535	151044089	158	−0.957	0.01774	0.9491	−53.95	8.357e-103

6. According to the line shown above, the SNP is located in chromosome 1, position 151044089 (or approximately 9 kb from the HM3_CNP_35 CNV). Genotypes at rs1048535 are strongly correlated ($r^2 = 0.9491$) with genotypes at the CNV and it is highly significant ($p = 8.357\text{e-}103$). Also, strong evidence of association was also observed at rs7524281 (second most significant SNP in the .adjusted file. See above), which is also in chromosome 1 and only about 6 kb away from rs1048535. Therefore, it is reasonable to suspect that there is also high correlation between rs7524281 and rs1048535. PLINK features several functions that allow to easily estimate linkage disequilibrium (LD) values between SNPs. For example, the following command computes the LD of rs1048535 to every SNP in a window of size 500 kb around the SNP. The option *--ld-window-r2* 0 ensures all pair-wise LD calculations are printed to the output file.

```
plink --bfile CEU_HapMap_GWAS_data --r2
--ld-snp rs1048535 --ld-window-kb 5000
--ld-window-r2 0 --out rs1048535_ld
```

CHR_A	BP_A	SNP_A	CHR_B	BP_B	SNP_B	R2
1	151044089	rs1048535	1	151046763	rs7411365	0.0340111
1	151044089	rs1048535	1	151046901	rs1930127	0.87627
1	151044089	rs1048535	1	151047146	rs17670505	0.0297426
1	151044089	rs1048535	1	151049879	rs7524281	0.948987
1	151044089	rs1048535	1	15105184	rs7536191	0.0255985
1	151044089	rs1048535	1	151050348	rs12023196	0.0226745
1	151044089	rs1048535	1	151050879	rs11804609	0.0258134
1	151044089	rs1048535	1	151059829	rs7550676	0.0317649
1	151044089	rs1048535	1	151062595	rs6659798	0.0317649
1	151044089	rs1048535	1	151063073	rs7517755	0.0613721
1	151044089	rs1048535	1	151068040	rs11205114	0.145429
1	151044089	rs1048535	1	151092224	rs12565568	0.0317649
1	151044089	rs1048535	1	151114516	rs12022319	0.0486071

Fig. 5 Screenshot of output file rs1048535_ld.ld, which contains all pair-wise LD calculations

The output file rs1048535_ld.ld (Fig. 5) contains all pair-wise LD calculations. While the first three columns show information (chromosome, position, and identifier) on the first SNP, the next three columns show information on the second SNP and the seventh column contains the LD value, r^2. The two SNPs of our interest, rs7524281 and rs1048535, display high linkage-disequilibrium ($r^2 = 0.948987$). Hence, these two SNPs and the CNV are part of the same haplotype block.

7. To illustrate the case when there exists a need to control for population stratification, we can easily repeat the analysis and include the MDS option (see example in the previous section) as covariates. Since PLINK can take covariate values contained in a file (where the first two columns indicate family and individual IDs and subsequent columns contain the covariates), it is possible to use the MDS file generated with PLINK to produce a covariate file with the following command on the linux terminal:

```
awk {print $1,$2,$4,$5} filename.mds > mds_
covar.txt
```

Finally, we will use a command similar to that described in **step 4**, different only for the addition of the covariate option (*--covar*) and the use of *--linear* instead of *--assoc* (*see* **Notes 5** and **6**).

```
plink --bfile CEU_HapMap_GWAS_data --linear
--allow-no-sex --pheno CNV_phenos.txt
--covar mds_covar.tx
--all-pheno --missing-phenotype –9 --adjust
--ci 0.95 --out cnv_qtl
```

PLINK will generate output files, which are similar to those previously described, except that they contain an additional column "TEST." For every SNP, a line containing the results of the additive test is presented, as well as one additional line per analyzed covariate.

8. In summary, this exercise used a CNV as a quantitative trait to exemplify how to run a genome-wide association in PLINK. We described how to navigate the results files generated by PLINK, how to calculate LD between SNPs and how to include covariates and correct for population stratification in our analyses. In the final exercise, you will learn to simulate data and to perform a case vs. control association analysis.
9. In addition to the results files, PLINK also generates a .log file, which contains information about the analysis and the dataset (i.e., analysis start and finish times, number of markers tested, number of individuals, cases/controls, males/females, parameters, options and filters in effect, etc.). We recommend you have a look at the .log files generated by the different analyses you conduct.

3.2.2 Case vs. Control Association Analysis

1. PLINK offers several tests for binary trait association (also known as case–control association) analyses. There exist repositories of real case–control genotype data (e.g., the database of Genotypes and Phenotypes, dbGAP: http://www.ncbi.nlm.nih.gov/gap), but obtaining access to such datasets commonly requires filling out an application form for data release. Given that PLINK also offers a function for simulating genotype and phenotype data, we will use it to generate a case–control dataset. All scripts and data files you need are provided in the publisher's website (http://extras.springer.com/).
2. We will simulate genotype data for 45,030 SNPs, of which 30 are trait-associated and 45,000 are not. This is indicated in the simulate file **gwas.sim** (shown below), which contains the parameters of the desired genotype data.

   Where:

Column 1	Indicates the number of markers to be generated
Column 2	Specifies an SNP identifier prefix
Column 3, 4	MAF window of SNPs to be generated
Columns 5, 6	Odds ratio of the heterozygous genotype and the homozygous minor genotype. Specifying "mult" implies a multiplicative effect for the homozygote genotype, i.e., ORHOM = ORHET * ORHET

```
20000   nullA   0.00   0.05   1.00   1.00
10000   nullB   0.05   0.10   1.00   1.00
5000    nullC   0.10   0.20   1.00   1.00
10000   nullD   0.20   0.99   1.00   1.00
10      assoc1  0.00   0.05   1.80   Mult
20      assoc2  0.05   0.40   1.20   Mult
```

3. The following command will generate genotypes for 3,000 cases (*--simulate-ncases*) and 3,000 controls (*--simulate-ncontrols*):

```
plink --simulate gwas.sim --make-bed --simulate-
ncases 3000 \
--simulate-ncontrols 3000 --simulate-label POP1
--out simulated_gwas
```

4. As mentioned previously, several association tests have been implemented in PLINK. These include:

Test	Option
Allelic test	--assoc
Allelic test using Fisher's exact test	--fisher
Logistic regression	--logistic
Linear regression	--linear
Full-association	--model
Cochran-Armitage trend test	--model --model-trend
Genotypic (2 df) test	--model --model-gen
Dominant gene action (1 df) test	--model --model-dom
Recessive gene action (1 df) test	--model --model-rec

Allelic tests compare frequencies of alleles in groups of cases and controls. If the *--model* option is specified, PLINK will perform full-model association testing, which includes ALLELIC, TREND, GENO, DOM, and REC for each SNP. The analysis can be restricted to only one of these tests by including an additional option (i.e., *--model --model-dom* will only compute the dominant gene effect model, and *--model --model-rec* will only compute the recessive gene effect model). These options can also be combined with the *--adjusted* option previously described. In GWAS for complex diseases, it is usually assumed that disease alleles have an additive effect and so allelic tests are more widely used. Logistic and linear regression are also commonly used, as they are more flexible than the other tests in that they can account for confounding effects with the use of covariates (e.g., disease onset or gender). Another use of logistic and linear regression is to account for population stratification, as previously shown in the quantitative trait association exercise.

The following command will perform allelic association test on the simulated dataset:

```
plink  --bfile simulated_gwas --allow-no-sex --
assoc
--adjust --ci 0.95 --out simulated_assoc
```

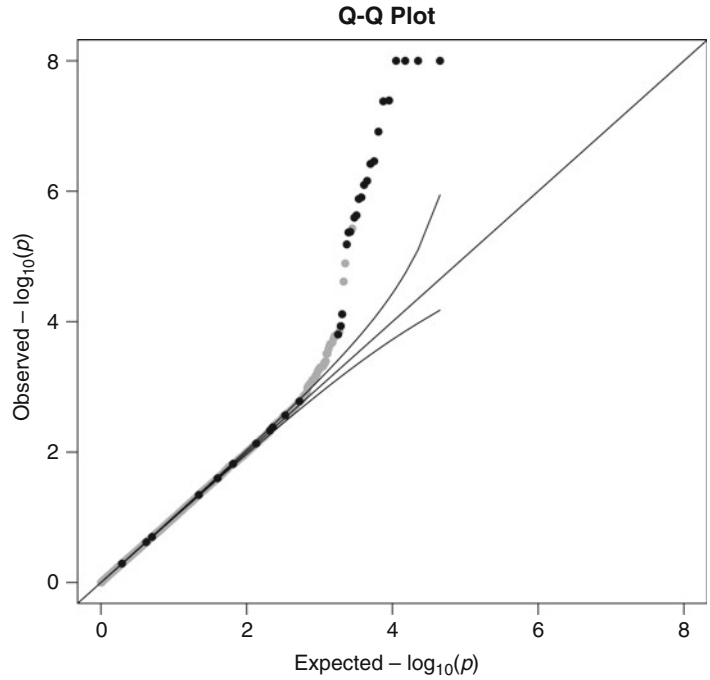

Fig. 6 Quantile-quantile (Q-Q) plot showing expected vs. observed [$-\log_{10}(P)$ values]. Simulated nonassociated SNPs are shown in *gray*, and simulated associated SNPs are shown in *black*

This command is almost identical to that previously used for quantitative trait association. This is because PLINK will automatically assume this is a binary trait when the phenotype column in the PED/FAM file only contain 0, 1, 2 values.

5. The interpretation of whole-genome association studies is deeply facilitated by the use of data visualization tools. For instance, a quantile-quantile plot (Fig. 6), or Q-Q plot, can be used for detecting evidence of systematic bias (be it from unrecognized population structure, genotyping artifacts, etc.). Q-Q plots also show the extent to which the observed distribution of the test statistic follows the expected (null) distribution. The **qq_plot.R** script (provided) will generate a Q-Q plot of the GWAS results obtained from our simulated dataset for chromosome 1:

   In the figure, simulated disease SNPs are colored in red.

6. Similarly, GWAS Manhattan plots display the negative logarithm of the association *p*-value for each SNP (Y-axis) against the genomic coordinates (along the X-axis). Given that the strongest associations have the smallest *p*-values, the $-\log_{10}$ of these *p*-values will have the highest height in the Manhattan plot (Fig. 7). The **manhattan_plot.R** script (also provided) will generate a Manhattan plot of the GWAS results obtained from our simulated dataset for chromosome 1.

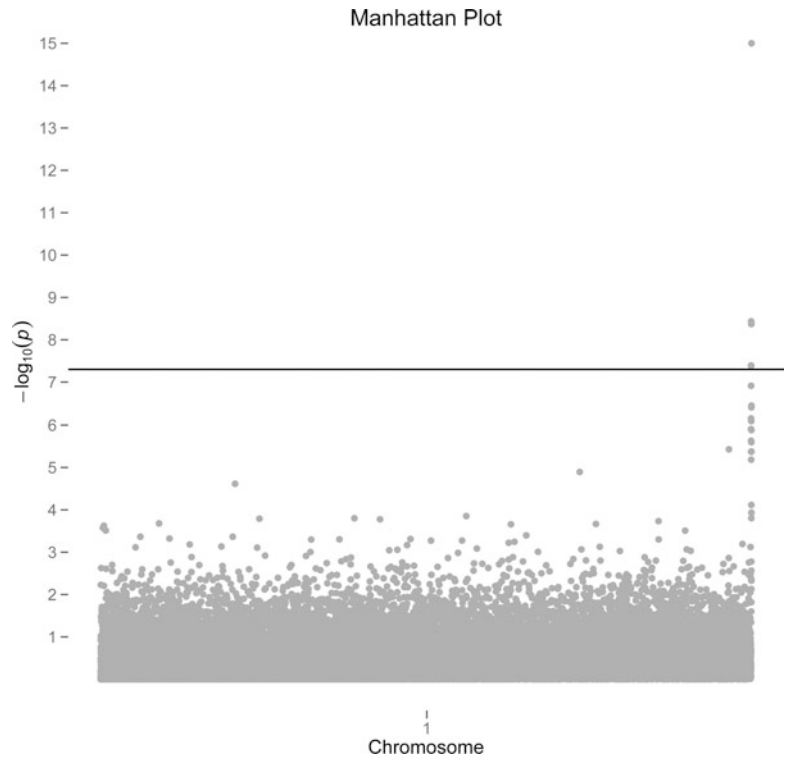

Fig. 7 Manhattan plot of association *p*-values of SNPs in chromosome 1. The *x*-axis shows location and *y*-axis displays the significance of the association ($-\log_{10}(P)$ value)

Note that, in this case, only SNPs from chromosome 1 are plotted. Usually, Manhattan plots also include all other autosomal chromosomes.

On the right hand side of the plot, a column of SNPs with high significance is observed. At least three SNPs display *p*-values above the genome-wide significance threshold. In the case of a true association, we would expect that some of its neighboring SNPs in LD were also associated with the phenotype, since they are expected to be co-inherited in the population.

Although the data used in this exercise are simulated, the same sequence of analyses would be conducted on real data. After finding a significant association, you would typically want to find out more about the genomic context of the SNP. For example, whether there are other SNPs in LD and do they show the expected pattern of *p*-values? Is your SNP, or one in high LD within a transcribed region or a functional element, such as a methylation site? There exist a number of online tools and databases that can help in the annotation and further characterization of GWAS findings.

The Catalog of Published GWAS (http://www.genome.gov/gwastudies/) is a searchable and downloadable database of publications reporting SNP-trait associations. These publications are

identified through periodic PubMed searches, NIH-distributed compilations of news and media reports, and occasional comparisons with other GWAS literature databases. The catalogue is searchable by disease/trait, chromosomal region, gene, or SNP. This is a good starting point if one wants to find out whether a gene or genomic region has previously been associated with the same or other traits of interest by a whole-genome association study.

Locus Zoom (http://csg.sph.umich.edu/locuszoom/) is a web-based plotting tool for generating regional plots of association results in their genomic context with publication-ready quality. This enables a quick visual inspection of the strength of association evidence, the position of the associated SNPs relative to genes in the region, and the extent of the association signal and LD.

SNAP (http://www.broadinstitute.org/mpg/snap/) is an online tool that allows for the rapid retrieval of proxy SNPs based on LD, physical distance, and/or membership in commercial genotyping arrays. Given an input of one or more query SNPs, SNAP can report pair-wise LD estimates and generate LD-plots derived from the genomic data from both the International HapMap Project and the 1K Genomes Project.

3.3 Conclusions

In this chapter we have provided a brief practical tutorial on the analysis of GWAS data with PLINK. Given the space limitations, it is not possible to cover all the functionalities of PLINK within a single book chapter. However, the developers of PLINK have provided comprehensive documentation in an online manual, which contains detailed information about all PLINK options and functions:

http://pngu.mgh.harvard.edu/~purcell/plink/pdf.shtml.

4 Notes

It is important to consider the following, while using PLINK:

1. When PLINK starts it will attempt to contact the web, to check whether there is a more up-to-date version available or not. After checking, PLINK writes a file called *.pversion* to the working directory and uses this cached information for the rest of the day. This option can be disabled with the *--noweb* option on the command line. When using PLINK on a machine with no, or a very slow, web connection, it may be desirable to turn this feature off, as no PLINK updates are being made.
2. In family-based analyses, PLINK can only correct for one relationship type. This means that PLINK cannot be used for the analysis of extended families or twin cohorts.

3. In a PED file, quantitative traits with decimal points must be coded with a period/full-stop character and not with a comma, i.e., 2.394 not 2,394.
4. The *--silent* option suppresses output to console window.
5. Covariates can only be used with the linear and logistic commands. However, including the *--covariate* command with other association commands will not yield an error message.
6. When using covariate files, individuals with missing data are ignored by default.

Reference

1. Purcell S et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81 (3):559–575