# Introduction to PLINK

GBIO0009

Kridsadakorn Chaichoompu

University of Liege

# PLINK: Why PLINK?

- PLINK is a whole genome association analysis software, and it is FREE!

  http://pngu.mgh.harvard.edu/~purcell/plink/
- PLINK has a well-documented manual to explain all features
- PLINK is available for Linux, Mac OS, and MS-DOS
- PLINK has 2 versions, the stable version (1.07) and the beta version (1.9)
  - PLINK 1.9 works much faster than 1.07
  - PLINK 1.9 has many new features
- gPLINK is the other version of PLINK that provides graphical user interface. Please be aware that using PLINK for a while genome analysis usually takes a long time, it is better to use a command-line version
- Recommend to use PLINK 1.07

# PLINK: Let's get started

- To download PLINK:

http://pngu.mgh.harvard.edu/~purcell/plink/dist/plink-1.07-i686.zip

- In *plink-1.07-xxx.zip*, there is an example set of input files which is a good point to explore
  - test.map contains the marker information
  - test.ped contains genotype data and sample information
- Check what are inside the example files!

  plink --file test

# Example data

- Download the example data from the course website
  - TSI_JPT_chr20_case_control.bed
  - TSI_JPT_chr20_case_control.bim
  - TSI_JPT_chr20_case_control.fam
  - TSI_JPT_chr20_pheno_header.txt
  - TSI_JPT_chr20_pheno.txt

# PLINK: File Formats

PLINK mainly supports 3 types of formats

- Standard text format (PED and MAP) Note that all files must have the same name, otherwise we need to clearly indicate by using *--ped* and *–map*

  plink --file test

- Binary format (BED, BIM, and FAM)

  plink --bfile test

- Transposed text format (TPED, and TFAM) Note that all files must have the same name, otherwise we need to clearly indicate by using *--tped* and *--tfam*

  plink --tfile test

# Format conversion

- To convert or to indicate output as text format (PED and MAP

  plink --file test --recode --out test_ped
- To convert or to indicate output as Binary format (BED, BIM, and FAM)

  plink --file test --make-bed --out test_bin
- To convert or to indicate output as Transposed text format (TPED, and TFAM)

  plink --file test --transpose --recode --out test_tp
- Alternatively, it is possible to recode data as 1/2 encoding

  plink --file test --recode12 --out test_12
- To convert to additive encoding

  plink --file test --recodeAD --out test_12
- It is possible to switch between A,T,G,C encoding to 1,2,3,4 encoding by using --allele1234 or --alleleACGT vice versa

# Alternate phenotype files

To specify an alternate phenotype for analysis, i.e. other than the one in the *.ped file (or, if using a binary fileset, the *.fam file), use the --pheno option:

```
plink --file mydata --pheno pheno.txt
```

where pheno.txt is a file that contains 3 columns (one row per individual):

Family ID

Individual ID

Phenotype

The original PED file must still contain a phenotype in column 6, unless the --no-pheno flag is given.

The order of the alternate phenotype file need not be the same as for the original file. If the phenotype file contains more than one phenotype, then use the --mpheno N option to specify the Nth phenotype is the one to be used:

```
plink --file mydata --pheno pheno2.txt --mpheno 4
```

where pheno2.txt contains 5 different phenotypes, this command will use the 4th for analysis (phenotype D):

Family ID

Individual ID

Phenotype A

Phenotype B

Phenotype C

Phenotype D

Phenotype E

If your file is coded 0/1 to represent unaffected/affected, then use the --1 flag:

```
plink --file mydata --1
```

# Data manipulation: SNPs (1/3)

To get a set of SNPs, you can specify a single SNP and, optionally, also ask for all SNPs in the surrounding region, with the --window option:

plink --bfile mydata --snp rs652423 --window 20

which extracts only SNPs within +/- 20kb of rs652423 based on multiple SNPs and ranges (--snps)

The --snps command will accept a comma-delimited list of SNPs, including ranges based on physical position. For example,

plink --bfile mydata --snps rs273744-rs89883,rs12345-rs67890,rs999,rs222

Based on physical position (--from-kb, etc)

plink --bfile mydata --chr 2 --from-kb 5000 --to-kb 10000

to select all SNPs within this 5000kb region on chromosome 2.

# Data manipulation: SNPs (2/3)

To merge more than two standard and/or binary filesets, it is often more convenient to specify a single file that contains a list of PED/MAP

For example, consider we had 4 PED/MAP filesets (labelled fA.* through fD.*) and 4 binary filesets, labelled fE.* through fH.*).

Then using the command:

plink --file fA --merge-list allfiles.txt --make-bed --out mynewdata

# Data manipulation: SNPs (3/3)

To exclude some sets of SNPs

plink --file data --exclude mysnps.txt

where the file mysnps.txt is, as for the --extract command, just a list of SNPs, one per line.

# Data manipulation: individuals (1/3)

To get a set of individuals

```
plink --file data --keep mylist.txt
```

where the file mylist.txt is, as for the --remove command, just a list of Family ID / Individual ID pairs, one set per line, i.e. one person per line. (fields can occur after the 2nd column but they will be ignored -- i.e. you could use a FAM file as the parameter of the --keep command, or have comments in the file. For example

```
F101   1
F1001  2_B
F3033  1_A  Drop this individual because of consent issues
F4442  22
```

# Data manipulation: individuals (2/3)

To exclude a set of individuals

plink --file data --remove mylist.txt

where the file mylist.txt is, as for the --keep command, just a list of Family ID / Individual ID pairs, one set per line, i.e. one person per line (although, as for --keep, fields after the 2nd column are allowed but they will be ignored).

# Data manipulation: individuals (3/3)

Filter some individuals

```
plink --file data --filter myfile.raw 1 --freq
```

implies a file myfile.raw exists which has a similar format to phenotype and cluster files: that is, the first two columns are family and individual IDs; the third column is expected to be a numeric value (although the file can have more than 3 columns), and only individuals who have a value of 1 for this would be included in any subsequent analysis or file generation procedure. e.g. if myfile.raw were

```
F1 I1 2
F2 I1 7
F3 I1 1
F3 I2 1
F3 I3 3
```

Because filtering on cases or controls, or on sex, or on position within the family, will be common operations, there are some shortcut options that can be used instead of --filter. These are:

```
--filter-cases
--filter-controls
--filter-males
--filter-females
--filter-founders
--filter-nonfounders
```

# Quality control processes

- Missing genotype
- Hardy-Weinberg Equilibrium
- Minor Allele frequency
- Linkage disequilibrium pruning
- Mendel errors

# Missing genotype

To generate a list genotyping/missingness rate statistics:

plink --file data --missing

This option creates two files:

plink.imiss

plink.lmiss

which detail missingness by individual and by SNP (locus), respectively. For individuals, the format is:

| | |
|---|---|
| FID | Family ID |
| IID | Individual ID |
| MISS_PHENO | Missing phenotype? (Y/N) |
| N_MISS | Number of missing SNPs |
| N_GENO | Number of non-obligatory missing genotypes |
| F_MISS | Proportion of missing SNPs |

For each SNP, the format is:

| | |
|---|---|
| SNP | SNP identifier |
| CHR | Chromosome number |
| N_MISS | Number of individuals missing this SNP |
| N_GENO | Number of non-obligatory missing genotypes |
| F_MISS | Proportion of sample missing for this SNP |

# Clustering based on missing genotypes

Systematic batch effects that induce missingness in parts of the sample will induce correlation between the patterns of missing data that different individuals display. One approach to detecting correlation in these patterns, that might possibly idenity such biases, is to cluster individuals based on their identity-by-missingness (IBM).

plink --file data --cluster-missing

which creates the files:

plink.matrix.missing

plink.cluster3.missing

which have similar formats to the corresponding IBS clustering files.

# Missing rate per person

The initial step in all data analysis is to exclude individuals with too much missing genotype data. This option is set as follows:

```
plink --file mydata --mind 0.1
```

which means exclude with more than 10% missing genotypes. A line in the terminal output will appear, indicating how many individuals were removed due to low genotyping. If any individuals were removed, a file called

```
plink.irem
```

will be created, listing the Family and Individual IDs of these removed individuals. Any subsequent analysis also specifeid on the same command line will be performed without these individuals.

# Missing rate per SNP

Subsequent analyses can be set to automatically exclude SNPs on the basis of missing genotype rate, with the --geno option: the default is to include all SNPS (i.e. --geno 1).

To include only SNPs with a 90% genotyping rate (10% missing) use

```
plink --file mydata --geno 0.1
```

As with the --maf option, these counts are calculated after removing individuals with high missing genotype rates.

# Hardy-Weinberg Equilibrium (1/2)

To generate a list of genotype counts and Hardy-Weinberg test statistics for each SNP, use the option:

plink --file data --hardy

which creates a file:

plink.hwe

This file has the following format

| | |
|---|---|
| SNP | SNP identifier |
| TEST | Code indicating sample |
| A1 | Minor allele code |
| A2 | Major allele code |
| GENO | Genotype counts: 11/12/22 |
| O(HET) | Observed heterozygosity |
| E(HET) | Expected heterozygosity |
| P | H-W p-value |

# Hardy-Weinberg Equilibrium (2/2)

To exclude markers that failure the Hardy-Weinberg test at a specified significance threshold, use the option:

```
plink --file mydata --hwe 0.001
```

By default this filter uses an exact test. The standard asymptotic (1 df genotypic chi-squared test) can be requested with the --hwe2 option instead of --hwe.

The following output will appear in the console window and in plink.log, detailing how many SNPs failed the Hardy-Weinberg test, for the sample as a whole, and (when PLINK has detected a disease phenotype) for cases and controls separately:

```
Writing Hardy-Weinberg tests (founders-only) to [ plink.hwe ]
30 markers failed HWE test ( p <= 0.05 ) and have been excluded
    34 markers failed HWE test in cases
    30 markers failed HWE test in controls
```

This test will only be based on founders (if family-based data are being analysed) unless the --nonfounders option is also specified.

# Allele frequency

To generate a list of minor allele frequencies (MAF) for each SNP, based on all founders in the sample:

```
plink --file data --freq
```

will create a file:

```
plink.frq
```

with five columns:

| | |
|---|---|
| CHR | Chromosome |
| SNP | SNP identifier |
| A1 | Allele 1 code (minor allele) |
| A2 | Allele 2 code (major allele) |
| MAF | Minor allele frequency |
| NCHROBS | Non-missing allele count |

# Minor Allele frequency

Once individuals with too much missing genotype data have been excluded, subsequent analyses can be set to automatically exclude SNPs on the basis of MAF (minor allele frequency):

```
plink --file mydata --maf 0.05
```

means only include SNPs with MAF >= 0.05. The default value is 0.01. This quantity is based only on founders (i.e. individuals for whom the paternal and maternal individual codes and both 0).

This option is appropriately counts alleles for X and Y chromosome SNPs.

# Linkage disequilibrium pruning (1/2)

Sometimes it is useful to generate a pruned subset of SNPs that are in approximate linkage equilibrium with each other. This can be achieved via two commands: --indep which prunes based on the variance inflation factor (VIF), which recursively removes SNPs within a sliding window; second, --indep-pairwise which is similar, except it is based only on pairwise genotypic correlation.

The VIF pruning routine is performed:

plink --file data --indep 50 5 2

will create files

plink.prune.in

plink.prune.out

Each is a simlpe list of SNP IDs; both these files can subsequently be specified as the argument for a --extract or --exclude command.

The parameters for --indep are: window size in SNPs (e.g. 50), the number of SNPs to shift the window at each step (e.g. 5), the VIF threshold. The VIF is 1/(1-R^2) where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously. That is, this considers the correlations between SNPs but also between linear combinations of SNPs.

# Linkage disequilibrium pruning (2/2)

The second procedure is performed:

plink --file data --indep-pairwise 50 5 0.5

This generates the same output files as the first option; the only difference is that a simple pairwise threshold is used. The first two parameters (50 and 5) are the same as above (window size and step); the third parameter represents the r^2 threshold.

To give a concrete example: the command above that specifies 50 5 0.5 would a) consider a window of 50 SNPs, b) calculate LD between each pair of SNPs in the window, b) remove one of a pair of SNPs if the LD is greater than 0.5, c) shift the window 5 SNPs forward and repeat the procedure.

To make a new, pruned file, then use something like (in this example, we also convert the standard PED fileset to a binary one):

plink --file data --extract plink.prune.in --make-bed --out pruneddata

# Mendel errors

To generate a list of Mendel errors for SNPs and families, use the option:

plink --file data --mendel

which will create files:

plink.mendel

plink.imendel

plink.fmendel

plink.lmendel

The *.mendel file contains all Mendel errors (i.e. one line per error); the *.imendel file contains a summary of per-individual error rates; the *.fmendel file contains a summary of per-family error rates; the *.lmendel file contains a summary of per-SNP error rates.

The *.mendel file has the following columns:

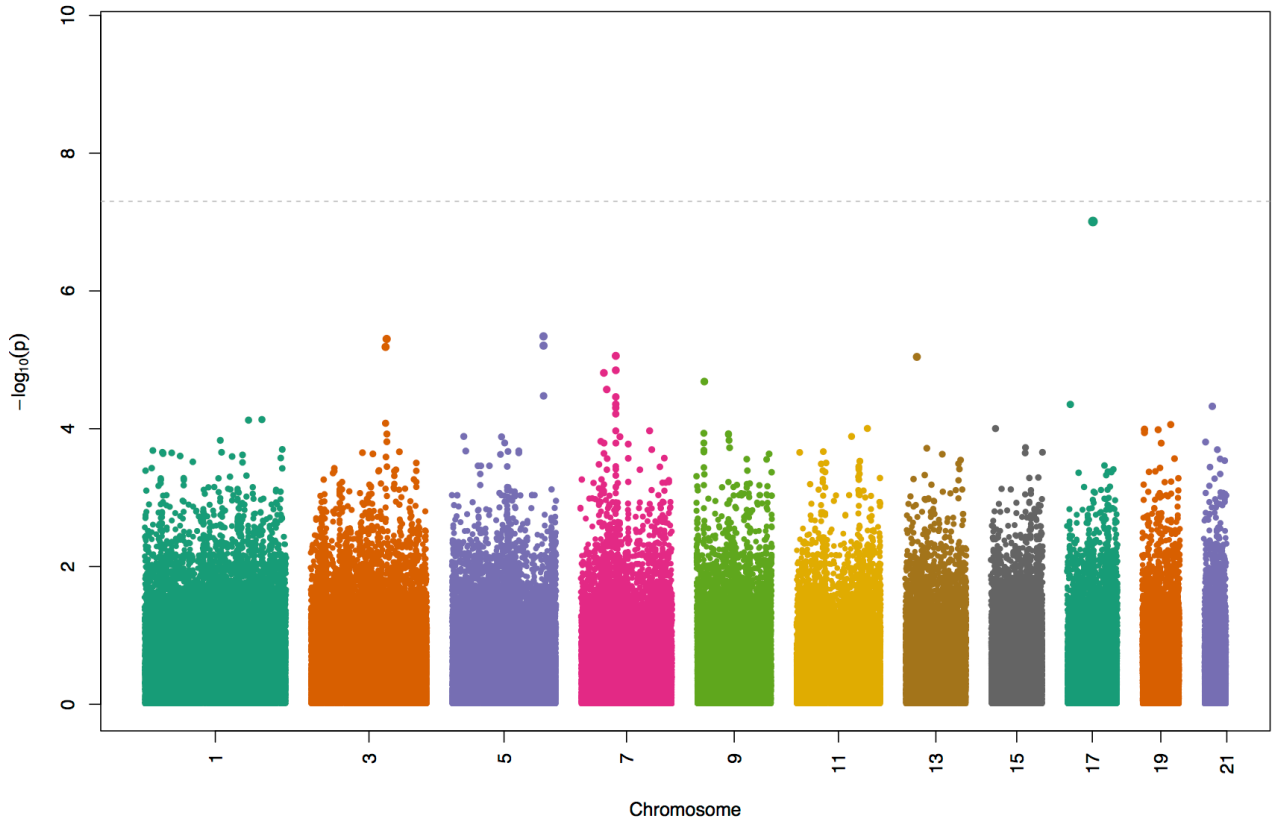
| | |
|---|---|
| FID | Family ID |
| KID | Child individual ID |
| CHR | Chromosome |
| SNP | SNP ID |
| CODE | A numerical code indicating the type of error (see below) |
| ERROR | Description of the actual error |

# Association Analysis

- Case/control
- Fisher's exact
- Full model
- Quantitative trait
- Linear and logistic models
- Multiple-test correction

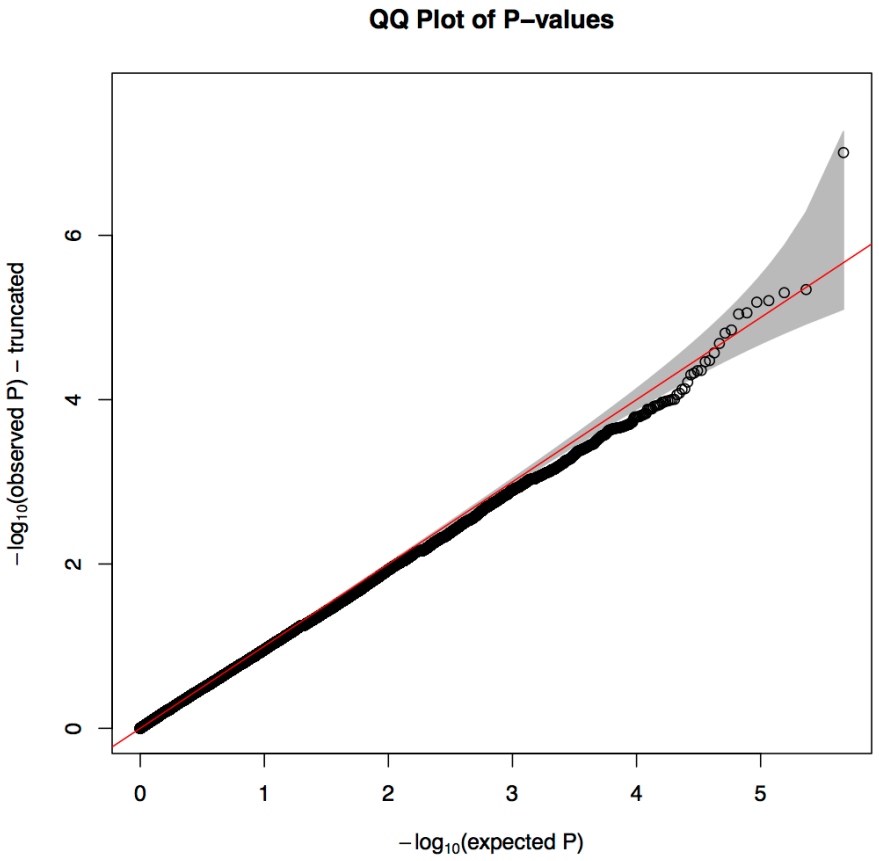
# Manhattan plot using GWASTools

```
manhattanPlot(assoc$P,chromosome=assoc$CHR)
```



# QQ plot using GWASTools

```
qqPlot(pval=assoc$P,truncate=TRUE, main="QQ
Plot of P-values")
```



# Basic case/control association test

To perform a standard case/control association analysis, use the option:

plink --file mydata --assoc

which generates a file

plink.assoc

which contains the fields:

| | |
|---|---|
| CHR | Chromosome |
| SNP | SNP ID |
| BP | Physical position (base-pair) |
| A1 | Minor allele name (based on whole sample) |
| F_A | Frequency of this allele in cases |
| F_U | Frequency of this allele in controls |
| A2 | Major allele name |
| CHISQ | Basic allelic test chi-square (1df) |
| P | Asymptotic p-value for this test |
| OR | Estimated odds ratio (for A1, i.e. A2 is reference) |

# Fisher's Exact test (allelic association)

To perform a standard case/control association analysis using Fisher's exact test to generate significance, use the option:

plink --file mydata --fisher

which generates a file

plink.fisher

which contains the fields:

| | |
|---|---|
| CHR | Chromosome |
| SNP | SNP ID |
| BP | Physical position (base-pair) |
| A1 | Minor allele name (based on whole sample) |
| F_A | Frequency of this allele in cases |
| F_U | Frequency of this allele in controls |
| A2 | Major allele name |
| P | Exact p-value for this test |
| OR | Estimated odds ratio (for A1) |

As described below, if --fisher is specified with --model as well, PLINK will perform genotypic tests using Fisher's exact test.

# Alternate / full model association tests

It is possible to perform tests of association between a disease and a variant other than the basic allelic test (which compares frequencies of alleles in cases versus controls), by using the --model option. The tests offered here are (in addition to the basic allelic test):

- Cochran-Armitage trend test
- Genotypic (2 df) test
- Dominant gene action (1df) test
- Recessive gene action (1df) test

The genotypic test provides a general test of association in the 2-by-3 table of disease-by-genotype. The dominant and recessive models are tests for the minor allele (which is the minor allele can be found in the output of either the --assoc or the --freq commands. That is, if D is the minor allele (and d is the major allele):

| | | | | |
|---|---|---|---|---|
| Allelic: | D | versus | d | |
| Dominant: | (DD, Dd) | versus | dd | |
| Recessive: | DD | versus | (Dd, dd) | |
| Genotypic: | DD | versus | Dd | versus dd |

As mentioned above, these tests are generated with option:

plink --file mydata --model

which generates a file

plink.model

which contains the following fields:

| | |
|---|---|
| CHR | Chromosome number |
| SNP | SNP identifier |
| TEST | Type of test |
| AFF | Genotypes/alleles in cases |
| UNAFF | Genotypes/alleles in controls |
| CHISQ | Chi-squated statistic |
| DF | Degrees of freedom for test |
| P | Asymptotic p-value |

# Quantitative trait association

Quantitative traits can be tested for association also, using either asymptotic or empirical significance values. If the phenotype (column 6 of the PED file or the phenotype as specified with the --pheno option) is quantitative, then PLINK will automatically treat the analysis as a quantitative trait analysis.

plink --file mydata --assoc

will generate the file

plink.qassoc

with fields as follows:

| | |
|---|---|
| CHR | Chromosome number |
| SNP | SNP identifier |
| BP | Physical position (base-pair) |
| NMISS | Number of non-missing genotypes |
| BETA | Regression coefficient |
| SE | Standard error |
| R2 | Regression r-squared |
| T | Wald test (based on t-distribtion) |
| P | Wald test asymptotic p-value |

If permutations were also requested, then an extra file, either

plink.assoc.perm or plink.assoc.mperm

will be generated, depending on whether adaptive or max(T) permutation was used (see the next section for more details). The empirical p-values are based on the Wald statistic.

# Linear and logistic models

These two features allow for multiple covariates when testing for both quantitative trait and disease trait SNP association, and for interactions with those covariates. The covariates can either be continuous or binary (i.e. for categorical covariates, you must first make a set of binary dummy variables). In this section we consider:

- Basic uasge
- Covariate and interactions
- Flexibly specifying the precise model
- Flexibly specifying joint tests

Basic usage

For quantitative traits, use

plink --bfile mydata --linear

For disease traits, specify logistic regression with

plink --bfile mydaya –logistic

These commands will either generate the output file

plink.assoc.linear or plink.assoc.logistic

depending on the phenotype/command used. The basic format is:

| | |
|---|---|
| CHR | Chromosome |
| SNP | SNP identifier |
| BP | Physical position (base-pair) |
| A1 | Tested allele (minor allele by default) |
| TEST | Code for the test (see below) |
| NMISS | Number of non-missing individuals included in analysis |
| BETA/OR | Regression coefficient (--linear) or odds ratio (--logistic) |
| STAT | Coefficient t-statistic |
| P | Asymptotic p-value for t-statistic |

# Adjustment for multiple testing

To generate a file of adjusted significance values that correct for all tests performed and other metrics, use the option:

plink --file mydata --assoc --adjust

which generates the file

plink.adjust

which contains the fields

| | |
|---|---|
| CHR | Chromosome number |
| SNP | SNP identifer |
| UNADJ | Unadjusted p-value |
| GC | Genomic-control corrected p-values |
| BONF | Bonferroni single-step adjusted p-values |
| HOLM | Holm (1979) step-down adjusted p-values |
| SIDAK_SS | Sidak single-step adjusted p-values |
| SIDAK_SD | Sidak step-down adjusted p-values |
| FDR_BH | Benjamini & Hochberg (1995) step-up FDR control |
| FDR_BY | Benjamini & Yekutieli (2001) step-up FDR control |

This file is sorted by significance value rather than genomic location, the most significant results being at the top.