

Lecture 3: Introduction to the PLINK Software

Instructors: Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2021

PLINK Overview

- ▶ 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:

<https://www.cog-genomics.org/plink2>

- ▶ PLINK has numerous useful features for managing and analyzing genetic data
- ▶ Data management
 - ▶ Read data in a variety of formats
 - ▶ Recode and reorder files
 - ▶ Merge two or more files
 - ▶ Extracts subsets (SNPs or individuals)
 - ▶ Flip strand of SNPs
 - ▶ Compress data in a binary file format

PLINK Overview

- ▶ Summary statistics for quality control
 - ▶ Allele, genotypes frequencies, HWE tests
 - ▶ Missing genotype rates
 - ▶ Inbreeding, IBS and IBD statistics for individuals and pairs of individuals
 - ▶ non-Mendelian transmission in family data
 - ▶ Sex checks based on X chromosome SNPs
 - ▶ Tests of non-random genotyping failure

PLINK Overview

- ▶ Basic association testing
 - ▶ Case/control
 - ▶ Standard allelic test
 - ▶ Fisher's exact test
 - ▶ Cochran-Armitage trend test
 - ▶ Mantel-Haenszel and Breslow-Day tests for stratified samples
 - ▶ Dominant/recessive and general models
 - ▶ Model comparison tests (e.g. general versus multiplicative)

PLINK Overview

- ▶ Family-based association (TDT, sibship tests)
- ▶ Quantitative traits, association and interaction
- ▶ Association conditional on one or more SNPs
- ▶ Asymptotic and empirical p-values
- ▶ Flexible clustered permutation scheme
- ▶ Analysis of genotype probability data and fractional allele counts (post-imputation)

PLINK Overview

- ▶ Multimarker predictors, haplotypic tests
 - ▶ Suite of flexible, conditional haplotype tests
 - ▶ Case/control and TDT association on the probabilistic haplotype phase
 - ▶ A set of proxy association" methods to study single SNP associations in their local haplotypic context
 - ▶ Imputation heuristic, to test untyped SNPs given a reference panel
- ▶ Copy number variant analysis
 - ▶ Joint SNP and CNV tests for common copy number variants
 - ▶ Filtering and summary procedures for segmental (rare) CNV data
 - ▶ Case/control comparison tests for global CNV properties
 - ▶ Permutation-based association procedure for identifying specific loci

PLINK Overview

- ▶ Gene-based tests of association
- ▶ Screen for epistasis
- ▶ Gene-environment interaction with continuous and dichotomous environments
- ▶ Meta-analysis
 - ▶ Automatically combine several generically-formatted summary files, for millions of SNPs

Input Files

- ▶ Genotype data is a text file
 - ▶ Pedigree file (.ped)
 - ▶ Map file (.map)
- ▶ Genotype data is a compressed binary file
 - ▶ Fam File (.fam)
 - ▶ Bim file (.bim)
 - ▶ Bed file (.bed)

Input Files

- ▶ Pedigree File - the first six columns are mandatory:
 - ▶ Family ID
 - ▶ Individual ID
 - ▶ Paternal ID
 - ▶ Maternal ID
 - ▶ Sex (1=male; 2=female; other=unknown)
 - ▶ Phenotype

Input Files

- ▶ MAP File has 4 columns:
 - ▶ chromosome (1-22, X, Y or 0 if unplaced)
 - ▶ rs# or snp identifier
 - ▶ Genetic distance (morgans)
 - ▶ Base-pair position (bp units)

Creating Binary PLINK files

- ▶ With the PLINK files `myfile.ped` and `myfile.map`, the PLINK command

```
plink --file myfile --make-bed --out myfile
```

generates the following files:

- ▶ `myfile.bed`
- ▶ `myfile.bim`
- ▶ `myfile.fam`

Data Management

- ▶ Inclusion/Exclusion criteria options
 - ▶ `--keep mylist.txt, --remove mylist.txt`
 - ▶ `--extract mysnp.txt, --exclude mysnp.txt`
 - ▶ `--chr 6, --from rs273744 --to rs89883`
- ▶ Other data management options
 - ▶ `--make-bed, --recode, -bmerge`
- ▶ Using files with phenotypes
 - ▶ `--pheno, --all-pheno, --mpheno`

Quality Control (QC)

- ▶ Summary statistics options:
 - ▶ minor allele frequency (MAF): `--freq`
 - ▶ SNP missing rate: `--missing`
 - ▶ Individual missing rate: `--missing`
 - ▶ Hardy-Weinberg: `--hardy`
- ▶ Inclusion/Exclusion criteria
 - ▶ MAF: `--maf`
 - ▶ SNP missing rate: `--geno`
 - ▶ Individual missing rate: `--mind`
 - ▶ Hardy-Weinberg: `--hwe`

Association Analysis with PLINK

- ▶ Basic association testing: `--assoc`, `--qassoc`
- ▶ Stratified analysis: `--within myclusterfile.dat`
- ▶ Covariates: `--covar -- mycovfile.dat`
- ▶ GxE interaction: `--gxe mycovfile.dat`

GWAS of Transferrin

- ▶ PLINK input files:
 - ▶ `Transferrin.bed`
 - ▶ `Transferrin.fam`
 - ▶ `Transferrin.bim`
- ▶ R Script File for Transferrin: Available on course website
 - ▶ `assoc_ex03.R`
- ▶ Plink Script File for Transferrin: Available on course website
 - ▶ `Session_3_PLINK_Code.txt`
- ▶ HELP: Use the PLINK website (very useful!)
pngu.mgh.harvard.edu/~purcell/plink/

Transferrin Data: File Inspection

- ▶ Copy the transferrin PLINK files to a folder
- ▶ Use the R script to inspect files (not the .bed file!)
- ▶ Questions:
 - ▶ How many individuals are there?
 - ▶ How many SNPs are there?
 - ▶ Is the transferrin phenotype approximately normally distributed?

Transferrin Data: QC with PLINK

- ▶ Can estimate allele frequency for all SNPs with PLINK
`plink --bfile Transferrin --freq --out Trans_freq`
- ▶ Calculate SNP and individual missingness with the following option:
`--missing --out Trans_missing`
- ▶ For each SNP, obtain p-values for HWE using the following option:
`--hardy --out Trans_hardy`

Transferrin Data: GWAS with PLINK

- ▶ Run a GWAS analysis of Transferrin with PLINK.
- ▶ Make sure to apply some quality controls
- ▶ Command to apply QC thresholds such as MAF 0.05/ missing 0.01 / HWE 0.001 for the GWAS analysis with PLINK:

```
plink --bfile Transferrin --pheno Tr.pheno --maf 0.05  
--geno 0.01 --hwe 0.001 --assoc --out  
GWAS_T_add
```

Transferrin Data: Analyzing a Subset of SNPs

- ▶ Can easily analyze a subset of SNPs with PLINK
- ▶ The following file contains a list of SNPs that are of interest:

SNP_List.txt

- ▶ Can use the following PLINK command with the "extract" option to perform association testing on a subset of SNPs:

```
plink --bfile Transferrin --pheno Tr.pheno --extract  
SNP_List.txt --assoc --out GWAS_T_add_Subset
```

- ▶ Can use the following command to perform r^2 LD calculations for all possible pairs of SNPs in the subset SNP file

```
plink --bfile Transferrin --extract SNP_List.txt --r2  
--out LD_T_Subset
```

References

- ▶ Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575.