PLINK WGAS Practical Exercise; March 2009
(http://pngu.mgh.harvard.edu/purcell/plink/)

Shaun Purcell
(shaun@pngu.mgh.harvard.edu)

OVERVIEW OF GOALS

The first part of this tutorial can be approached in one of two ways: either using the graphical tools *gPLINK* and *HaploView* before using *PLINK* itself, or alternatively, using command-line *PLINK* straight away, in place of *gPLINK* and *Haploview*. In the first part of the tutorial, we cover the following:

- Examine the file formats for input data
- Generate basic descriptive statistics for whole genome SNP data
- Perform basic quality control filtering
- Perform a basic association analysis
- Incorporate population in a stratified whole-genome association analysis

In the second part of the tutorial, using *PLINK* as a command-line tool, we will:

- Further examine a putative "hit" SNP/locus, for example considering the genetic model and evidence of between-population heterogeneity and interaction with sex
- Empirically detect population substructure in the sample

Finally, we will continue to explore this dataset using *PLINK* and *HaploView*:

- Incorporate new data on extra SNPs typed in the region
- Examining linkage disequilibrium patterns in the associated region
- Perform haplotype analysis of the region
- Perform conditional haplotypic tests

EXAMPLE DATASETS AND SOFTWARE

This tutorial is written with MS-DOS users in mind. Users of Linux/Unix and Mac OS should have no problems running this either (in fact, using one of these alternate operating systems is advised).

This practical uses the following software:

PLINK	Command-line genetic analysis toolset
gPLINK	(Optional) Graphical interface for basic *PLINK* functions
Haploview	(Optional) Graphical tool for viewing *PLINK* results and SNP analysis

The data used in this exercise are from the 90 Asian HapMap individuals (Han Chinese from Beijing and Japanese from Tokyo). From the actual HapMap SNP data, ~250,000 SNPs have been extracted, which are the autosomal SNPs on half of the Affymetrix 500K SNP Array product. Along with a simulated disease phenotype, these data are in the files

`wgas1.ped`	Genotype data for 228,694 SNPs on 90 individuals
`wgas1.map`	Map file for these SNPs

In addition, a small subset of SNPs (N=29) genotyped on the same individuals represent a "follow-up genotyping" exercise, focused on a single locus; these will be used later in the practical.

`extra.ped`	Genotype data for 29 SNPs on the same individuals
`extra.map`	Map file for these SNPs

Finally, the true population membership (Chinese or Japanese) is encoded in a file

```
pop.cov        Population membership (coded 1=CH /2=JP)
```

Open a DOS prompt (Start Menu → Run → Type "cmd" → Hit Return) and set the working directory to the previously used gPLINK folder. For example, if it was on the D: drive, type

```
D:                  { return }
cd example          { return }
```

Check you are in the correct folder by typing

```
plink --file extra
```

which should start *PLINK* and generate some output describing the extra PED/MAP fileset. If you get an error message, you are in the wrong directory. On some computers, you might need to type

```
./plink --file extra
```

CARDINAL RULES & CAVEATS

When using *PLINK* there are a few key points to remember.

- Always consult the LOG file (console output)
- *PLINK* has no memory
  - each run loads data anew, previous filters lost
- Exact syntax and spelling is **very important**
  - "minus minus" …
- Not every option can be combined with every other option
  - For example, basic haplotype tests cannot take covariates
  - *PLINK* doesn't always warn you
  - LOG file often shows what has happened (or not)
- Consult the web documentation (http://pngu.mgh.harvard.edu/purcell/plink/)
  - regularly

USING PLINK AND VIEWING OUTPUT ON THE COMMAND LINE

Note: for all PLINK commands, despite the way I've formatted the commands in this document, all the options must be typed on a single line, i.e. only hit Return after typing all options; put spaces between all options, e.g. the command below is typed as a single line:

```
plink --bfile wgas3 --recode --snp rs11204005 --out tophit   { Return }
```

When the output files are relatively small, they can be viewed on the console, by typing, for example, either the "more" or "type" DOS commands:

```
more plink.lmiss
```

or

```
type plink.assoc
```

or, on Unix of Mac OS computers,

```
less plink.assoc
```

PART 1: BASIC HANDLING AND ANALYSIS OF GWAS SNP DATA

We assume that genotypes have been called for all SNPs previously and a "PED" format file is the starting point for analysis. This format is also used by Merlin, HaploView, etc. The first step is to transform it into a more compact binary fileset, to speed up subsequent analysis.

Purpose	Create a compact binary dataset from the raw data	
Command	`plink --file wgas1` `--make-bed` `--out wgas2`	
Input	`wgas1.ped` `wgas1.map`	*Initial whole genome SNP fileset*
Output	`wgas2.bed` `wgas2.bim` `wgas2.fam`	*Binary PED file for whole genome SNP data* *Corresponding marker information* *Corresponding individual information*
Notes	*This command creates a new fileset containing the same data, that is easier to work with. The file* `wgas2.log` *(which is also printed to the console) contains a lot of useful information about the data.*	

The LOG file contains a lot of useful information: errors and warning messages will be displayed here also. Having performed the above command, we will attempt to read the data back in, as a sanity check that things worked, and look at the LOG file in more detail:

Purpose	Basic validation of this new fileset	
Command	`plink --bfile wgas2` `--out validate`	
Input	`wgas2.bed` `wgas2.bim` `wgas2.fam`	*Initial whole genome SNP binary fileset*
Output	`validate.log`	*LOG file with information*
Notes	*We use* `--bfile` *and not* `--file` *to load a binary fileset.*	

The same information that is displayed in the console is also saved in the file `validate.log`. Here it is:

```
Writing this text to log file [ validate.log ]
Analysis started: Mon Dec  1 13:57:26 2008

Options in effect:
        --bfile wgas2
        --out validate

Reading map (extended format) from [ wgas2.bim ]
228694 markers to be included from [ wgas2.bim ]
Reading pedigree information from [ wgas2.fam ]
90 individuals read from [ wgas2.fam ]
90 individuals with nonmissing phenotypes
```

```
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
49 cases, 41 controls and 0 missing
45 males, 45 females, and 0 of unspecified sex
Reading genotype bitfile from [ wgas2.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 228694 SNPs
90 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 0.993346
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 228694 SNPs
After filtering, 49 cases, 41 controls and 0 missing
After filtering, 45 males, 45 females, and 0 of unspecified sex

Analysis finished: Mon Dec  1 13:57:31 2008
```

The key points to note are:

- PLINK always makes a note of when the analysis was started and what the output is saved as.
- It then lists the commands as they were entered on the command line.
- It lists how many SNPs and individuals were read in, and some other basic information, including the number individuals with phenotype data, counts of cases and controls (if appropriate), and of males and females, genotyping rate, etc.
- In this instance, the QC filters are set to include all SNPs and all individuals, so the lines about "0 SNPs failed…" can be ignored

What we are interested in here is that the dataset contains 90 individuals and 228,694 SNPs and was successfully load into PLINK.

GENERATING BASIC SUMMARY STATISTICS

Of the numerous summary statistics that can be generated for SNPs or individuals, here we choose to look at two: calculating allele frequencies and performing Hardy-Weinberg equilibrium tests for all SNPs.

Purpose	Obtain allele frequencies for all SNPs
Command	`plink --bfile wgas2` `--freq` `--out freq1`
Input	`wgas2.bed` *Initial whole genome SNP binary fileset* `wgas2.bim` `wgas2.fam`
Output	`freq1.frq` *Allele frequency information for each SNP*
Notes	*The output file will be large. Counts instead of frequencies can be obtained by also adding* `--counts`. *Frequencies can be stratified by two or more groups, e.g. by adding* `--within pop.cov`

The output in the file `freq1.frq` contains as many rows as there are SNPs and has the following format:

```
 CHR         SNP   A1  A2        MAF  NCHROBS
   1   rs3094315    G   A     0.1236      178
   1   rs6672353    A   G   0.005618      178
   1   rs4040617    G   A     0.1167      180
```

```
   1   rs2905036    0   T          0       180
   1   rs4245756    0   C          0       180
   1   rs4075116    C   T    0.05556       180
  …
```

meaning that allele `A1` has frequency `MAF`.

Hint: If useful text processing tools such as grep or awk are available on your computer (e.g. Linux and Mac users), you can easily extract out required information, such as a list of all SNPs on chromosome 8 with MAF of 40% or above:

```
awk ' $1 == 8 && $5 >= 0.4 { print $2 } ' freq1.frq > mylist.snps
```

Aside: this particular example of post-processing output could also be performed within PLINK:

```
plink --bfile wgas2  --chr 8  --maf 0.4  --write-snplist  --out mylist
```

Otherwise, you could use Excel to view this file, or a stats package such as R. It is white-spaced delimited, plain-text, with a regular number of fields on each line (6 in this case). Alternatively, you can use *Haploview* to load any PLINK results file, as it offers a simple table-viewer. See the parallel *gPLINK* / *Haploview* tutorial for details.

Here we test for deviation from Hardy-Weinberg equilibrium:

Purpose	Perform Hardy-Weinberg tests for all SNPs
Command	`plink --bfile wgas2` `--hardy` `--out hwe1`
Input	`wgas2.bed` *Initial whole genome SNP binary fileset* `wgas2.bim` `wgas2.fam`
Output	`hwe1.hwe` *LOG file with information*
Notes	*This command calculates a test for departure from HWE for cases, controls and also cases and controls combined, using an exact test.*

The output file contains three rows per SNP (for case/control data), listing all individuals, then cases only, then controls only. A non-significant p-value indicates there is no evidence for deviation from HWE (although for less frequent alleles in small samples, this will not necessarily be a very powerful test). The other fields are described in the PLINK web documentation.

```
 CHR          SNP     TEST   A1  A2       GENO   O(HET)   E(HET)          P
   1   rs3094315      ALL    G   A   0/22/67   0.2472   0.2166     0.3476
   1   rs3094315      AFF    G   A   0/15/33   0.3125   0.2637     0.5771
   1   rs3094315    UNAFF    G   A    0/7/34   0.1707   0.1562          1
```

MAKING A FILTERED, "QC+" DATSET

Here we apply a set of simple quality control filters, for illustration only. In practice, one would want to look at the data much more carefully, and consider other factors that are beyond the scope of this basic tutorial.

Specifically, we will remove individuals who have a genotyping rate of less than 95% (i.e. more than 5% missing data, `--mind`). We will then remove SNPs that have less than a 95% genotyping rate (i.e. more than 5% missing data, `--geno`). We will also remove SNPs that have a minor allele frequency (MAF, `--maf`) of less than 1%, or that fail the HW test with $p < 0.001$ (`--hwe`). These per-SNP metrics are calculated after first removing individuals with below-threshold genotype rate.

We make a new dataset as before (with `--make-bed`), but now adding options to impose the above filters.

Purpose	Create a "QC+" SNP fileset
Command	`plink --bfile wgas2` `--maf 0.01` `--geno 0.05` `--mind 0.05` `--hwe 1e-3` `--make-bed` `--out wgas3`
Input	`wgas2.bed` *Initial whole genome SNP binary fileset* `wgas2.bim` `wgas2.fam`
Output	`wgas3.bed` *"QC+" whole genome SNP binary fileset* `wgas3.bim` `wgas3.fam` `wgas3.irem` *Excluded individuals*
Notes	*This command creates a new binary fileset, after filtering SNPs and individuals for 95% genotyping rate, a minor allele frequency of at least 1% and a HWE test p-value (in controls) of p>0.001.*

As noted in the LOG file, the new dataset contains 89 individuals and 179,493 SNPs. The excluded individual is noted in the file `wgas3.irem`. This dataset (`wgas3.*`) will form the basis for subsequent association analyses.

SINGLE SNP ASSOCIATION ANALYSIS

To perform a basic allelic test of association for single SNPs with disease state, run the following:

Purpose	Basic single SNP association analysis
Command	`plink --bfile wgas3` `--assoc` `--adjust` `--out assoc1`
Input	`wgas3.bed` *QC+ whole genome SNP binary fileset* `wgas3.bim` `wgas3.fam`
Output	`assoc1.assoc` *Basic allelic single SNP association tests* `assoc1.assoc.adjusted` *Adjusted p-values (multiple testing corrections)*
Notes	*The main output file contains, among other things, the case and controls allele frequencies, a p-value and the odds ratio for each SNP. To obtain only SNPs with significant p-values below some threshold, add the command, e.g.* `--pfilter 1e-4`

Adding the `--adjust` flag also makes PLINK report the genomic control inflation factor (lambda) in the LOG file. You can plot the p-values (or better, the minus $\log_{10}$ of the p-values) using Haploview, or, if it is installed on your computer, the stats package R:

```
d <- read.table("assoc1.assoc",header=T)
plot( -log10( d$P ) , col = d$CHR )
```

The "adjusted" file contains the single SNP results, with various other p-values that represent various simple adjustments for multiple testing. Unlike the main `.assoc` output file (which is sorted by genomic location), the `.adjusted` file is sorted by most to least significant SNP. Therefore, looking at the top of this file is a quick way to find the most associated SNPs.

The basic file has this format: chromosome, SNP, base-position, minor allele (A1), case A1 frequency, control A1 frequency, alternate allele (A2), association chi-squared statistic, p-value, odds ratio.

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs3094315	792429	G	0.1489	0.08537	A	1.684	0.1944	1.875
1	rs4040617	819185	G	0.1354	0.08537	A	1.111	0.2919	1.678
1	rs4075116	1043552	C	0.04167	0.07317	T	0.8278	0.3629	0.5507
1	rs9442385	1137258	T	0.3723	0.4268	G	0.5428	0.4613	0.7966
1	rs11260562	1205233	A	0.02174	0.03659	G	0.3424	0.5585	0.5852
1	rs6685064	1251215	C	0.3854	0.439	T	0.5253	0.4686	0.8013
1	rs3766180	1563420	T	0.1771	0.09756	C	2.317	0.128	1.991
…									

To obtain a more manageable file with just the most associated SNPs, e.g. p<1e-5, add the flag:

```
--pfilter 1e-5
```

The analysis is identical, except only highly associated SNPs are listed in the `.assoc` and `.adjusted` files.

The genomic control lambda from the previous analysis is quite high for such a small dataset (~1.26). This implies there are a greater number of associated SNPs than we'd expect by chance (at the p<0.5 level). Given that our sample comprises both Chinese and Japanese individuals, one concern might be that population stratification between these two groups is biasing the association statistics. We can perform a analysis that conditions on these two groups:

Purpose	Stratified single SNP association analysis	
Command	`plink --bfile wgas3` `--mh` `--within pop.cov` `--adjust` `--out cmh1`	
Input	`wgas3.bed` `wgas3.bim` `wgas3.fam`	*QC+ whole genome SNP binary fileset*
	`pop.cov`	*"Cluster file" containing population code*
Output	`cmh1.cmh`	*Stratified allelic single SNP association tests*
	`cmh1.cmh.adjusted`	*Adjusted p-values (multiple testing corrections)*
Notes	*This test performs a Cochran-Mantel-Haenzsel test of a common odds ratio across the K=2 strata (Japanese and Chinese, in this case).*	

What is the new genomic control lambda? What is the new best SNP/region? Does it survive strict (Bonferroni) correction for multiple testing? What is the stratified association result for the best SNP in the previous analysis, that did not correct for potential differences between Chinese and Japanese groups?

SUMMARISING ASSOCIATION STATISTICS

Rather than reporting long lists of associated SNPs, many of which will be in linkage disequilibrium (LD), it is sometimes convenient to summarise the output of association tests as groups of SNPs in LD, or "clumps". We can also supply a list of the genomic co-ordinates for regions – in this case representing genes – in a file, to report clumps of highly associated SNPs and also the genes they are near:

Purpose	Clumping of previous CMH association results	
Command	`plink --bfile wgas3` `--clump cmh1.cmh` `--clump-range glist-hg18.txt` `--out clumps1`	
Input	`wgas3.bed` `wgas3.bim` `wgas3.fam`	*Initial whole genome SNP binary fileset*
	`cmh1.cmh`	*Stratified allelic single SNP association tests*
	`glist-hg18.txt`	*RefSeq known gene co-ordinates (hg18)*
Output	`clumps1.clumped`	*Most highly associated regions*
	`clumps1.clumped.ranges`	*Genes near associated regions*
Notes	*The dataset (`wgas3`) is only used to calculate LD between SNPs. It is possible to clump more than one result file simultaneously (they need not all come from the same dataset also).*	

END OF PART 1

In summary, we have performed basic QC on a GWAS dataset, resulting in 179,493 SNPs and 89 individuals. Population stratification was raised as a potential issue; conditioning on known population membership seems to help. Stratified analysis identified a SNP/region of interest on chromosome 8 (rs11204005).

PART 2: FOLLOW-UP ANALYSIS OF ASSOCIATED REGION

At this point, we assume you have followed the initial *gPLINK* tutorial and have created the binary QC+ fileset `wgas3`. The previous analyses showed that the SNP rs11204005 was the most highly associated when using the stratified CMH test.

```
more cmh1.cmh.adjusted
```

which should display (some of the text deleted here for space)

```
 CHR          SNP      UNADJ         GC       BONF   ...
   8  rs11204005 3.432e-007 4.171e-007     0.0616
   8   rs2460338 2.277e-006 2.696e-006     0.4088
  13   rs4943327  1.28e-005 1.479e-005          1
  13   rs4941815  1.28e-005 1.479e-005          1
  13   rs9531117 1.386e-005 1.599e-005          1
   5    rs839220 1.949e-005  2.24e-005          1
   5    rs373386 3.033e-005 3.463e-005          1
   5    rs444800 3.424e-005 3.903e-005          1
   5    rs454540 3.424e-005 3.903e-005          1
 ...
```

The goal of the next few steps is to extract the data for rs11204005 and perform a series of more detailed analyses on this single SNP.

Purpose	Extract data for single SNP rs11204005
Command	`plink --bfile wgas3` `--recode` `--snp rs11204005` `--out tophit`
Input	`wgas3.bed` *QC+ whole genome SNP binary fileset* `wgas3.bim` `wgas3.fam`
Output	`tophit.ped` *Standard PED file for this single SNP* `tophit.map` *Corresponding marker information*
Notes	*We are converting back from the binary format to standard text format. The* `--snp` *command is a filter, just extracting data for this one SNP.*

For this single SNP, we shall next examine the genotyping rate and, second, the Hardy-Weinberg test statistic.

Purpose	Examine genotyping rate for rs11204005
Command	`plink --file tophit` `--missing`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map`
Output	`plink.lmiss` *Missing rate per locus (SNP)* `plink.imiss` *Missing rate per individual*
Notes	Note use of `--file` instead of `--bfile` as `tophit` is in standard PED format. Also note that we do not always need to specify a unique output name when using PLINK directly, so all output files start `plink.`*`ext`* by default

Purpose	Examine Hardy-Weinberg equilibrium *P*-value for rs11204005
Command	`plink --file tophit` `--hardy`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information*
Output	`plink.hwe` *Hardy-Weinberg statistic and genotype counts*
Notes	For case/control datasets, tests given for all individual, as well as for cases and controls separately

Make a note of the genotyping rate and HWE *P*-value (in controls). What do they tell you?

TEST OF POPULATION-SPECIFIC EFFECTS FOR TOP SNP

Next, we shall examine whether this association varies between the two populations. When using the Cochran-Mantel-Haenszel test, we can request an additional Breslow-Day test for heterogeneous odds ratios between strata. Following this, we will use two alternate approaches that use different statistical methods to answer the same question (i.e is the effect different between Chinese and Japanese individuals?)

Purpose	Repeat stratified CMH test for association with disease for rs11204005 with Breslow-Day test for heterogeneity
Command	`plink --file tophit` `--mh` `--within pop.cov` `--bd`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information* `pop.cov` *File indicating Chinese (1) or Japanese (2)*
Output	`plink.cmh` *Cochran-Mantel-Haenszel statistic and odds ratio, including Breslow-Day test for heterogeneous odds ratios*
Notes	Unlike the CMH, which is appropriate for many sparse strata, the Breslow-Day test assumes a large sample N within each strata

Purpose	Repeat stratified test for association with disease for rs11204005 using a different approach (partitioning effects into total, between and within strata)
Command	`plink --file tophit` `--homog` `--within pop.cov`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information* `pop.cov` *File indicating Chinese (1) or Japanese (2)*
Output	`plink.homog` *Test for homogeneity*
Notes	*This command gives strata-specific odds ratios and p-values, unlike the CMH. The CMH is probably a better more general test for stratified data however.*

Next, we will repeat these basic analyses but using instead the framework of logistic regression analysis, that can incorporate multiple covariates, both continuous and binary. In the first instance, we will enter the Japanese/Chinese group membership as a binary covariate; second, we can explicitly test for an interaction with the SNP effect, to provide yet another way of addressing potential between-group heterogeneity in effect.

Purpose	Repeat test for association with disease for rs11204005 using a different approach (logistic regression, including population as a covariate)
Command	`plink --file tophit` `  --logistic` `  --covar pop.cov`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information* `pop.cov` *Indicates Chinese (1) or Japanese (2)*
Output	`plink.assoc.logistic` *Logistic regression results*
Notes	*The equivalent test for quantitative traits is obtained with* `--linear`

Purpose	Explicitly test for between-population heterogeneity using logistic regression allowing for an interaction effect
Command	`plink --file tophit` `  --logistic` `  --covar pop.cov` `  --interaction`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information* `pop.cov` *Indicates Chinese (1) or Japanese (2)*
Output	`plink.assoc.logistic` *Logistic regression results*
Notes	*One can also have >1 covariate and interaction terms*

The above analyses suggest that the association is equally present in both populations (make a note of what the precise results are that suggest this). Next, we can ask the more basic question of whether allele frequency (not the odds ratio for association) differs between the two groups. This involves using the population label as the *phenotype* of an association test rather than as a *covariate*. Because we have conveniently coded group membership as "1" and "2", we can directly treat it as a phenotype (e.g. "2"= Japanese = "affected").

Purpose	Explicitly test whether allele frequency for rs11204005 differs between populations
Command	`plink --file tophit` `  --assoc` `  --pheno pop.cov`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information* `pop.cov` *Indicates Chinese (1) or Japanese (2)*
Output	`plink.assoc` *Association (with population) results*
Notes	*Here we specify population as the phenotype, not a covariate*

Purpose	Explicitly test whether allele frequency for rs11204005 differs between populations, allowing for association with disease
Command	`plink --file tophit` `  --logistic` `  --pheno pop.cov` `  --covar tophit.ped` `  --covar-number 4`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information* `pop.cov` *Indicates Chinese (1) or Japanese (2)*
Output	`plink.assoc.logistic` *Association (with population) results*
Notes	*We treat the PED file as a covariate file, extracting just the phenotype (i.e. the 4th column after family ID and individual ID)*

These results would suggest that the frequency does indeed differ (again, make a note of exactly why this is).

EXAMINING GENOTYPIC MODELS

For simplicity in this Practical, we will ignore the effect of population for subsequent exercises.
This would not be advised with real data, as in this case, we in fact know that both allele frequency and disease rate differ between populations. It would therefore normally be important to perform analysis within-population or to include population as a covariate.

The previous association statistics were all based on allelic models (that each extra copy of the risk allele increases risk equally). We can also ask whether specific *genotype* configurations (heterozygotes versus homozygotes) have specific risk profiles.

Purpose	Test genotypic models for rs11204005
Command	`plink --file tophit` `--model` `--cell 1`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information*
Output	`plink.model` *Genotypic association tests*
Notes	*The* `--cell` *command sets the minimum cell size for which to perform genotypic tests (i.e. otherwise PLINK would skip this marker with cells < 5 observations, which is the default value).*

Purpose	Test genotypic models for rs11204005 using logistic regression
Command	`plink --file tophit` `--logistic` `--genotypic`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information*
Output	`plink.assoc.logistic` *Genotypic association tests*
Notes	*Covariates can also be included with this approach; dominant and recessive models can be explicitly requested with the options* `--dom` *and* `--rec`

Purpose	Test genotypic models for rs11204005 using logistic regression with an alternate genotypic coding
Command	`plink --file tophit` `--logistic` `--genotypic --hethom`
Input	`tophit.ped` *Standard PED file for single SNP* `tophit.map` *Corresponding marker information*
Output	`plink.assoc.logistic` *Genotypic association tests*
Notes	*Instead of tests of additive and dominance components, the* `--hethom` *option presents explicit tests for the heterozygote and homozygote effects*

These analyses suggest that the effect is an allele-dosage one, rather than showing dominant or recessive non-additivity. Make a note of the exact results that support this conclusion.

SEX-SPECIFIC EFFECTS

Next, in the same manner as we tested for between-population heterogeneity, we can ask whether the effect varies between males and females. We do this first by performing sex-specific analyses; second, by including sex as a covariate in a logistic regression model.

Purpose	Test for association specific in males	
Command	`plink --file tophit` `--filter-males` `--logistic`	
Input	`tophit.ped`	*Standard PED file for single SNP*
	`tophit.map`	*Corresponding marker information*
Output	`plink.assoc.logistic`	*Genotypic association tests*
Notes	*Can also include other covariates here (population, etc)*	

Purpose	Test for association specific in females	
Command	`plink --file tophit` `--filter-females` `--logistic`	
Input	`tophit.ped`	*Standard PED file for single SNP*
	`tophit.map`	*Corresponding marker information*
Output	`plink.assoc.logistic`	*Genotypic association tests*
Notes		

Purpose	Test for different effects in males versus females	
Command	`plink --file tophit` `--logistic` `--sex` `--interaction`	
Input	`tophit.ped`	*Standard PED file for single SNP*
	`tophit.map`	*Corresponding marker information*
Output	`plink.assoc.logistic`	*Genotypic association tests*
Notes	*The* `--sex` *command adds sex as a covariate (0=males, 1=females)*	

These results suggest no sex differences in the nature of the association. Again, make a note of the exact supporting statistical evidence for this. What are the odds ratios in males and females?

EMPIRICAL ASSESSMENT OF POPULATION STRATIFICATION

In the initial Practical session, we used the known population labels of Chinese versus Japanese. In many studies, we might not have this direct information, or the potential differences in ancestry can be subtle (for example, individuals of US individuals of predominantly Northern European descent versus US individuals of predominantly Southern European descent).

In this set of exercises, we will use the SNP data to empirically investigate the ancestry of the samples and to assign individuals to groups. First, we see that we can largely recapture the Chinese/Japanese distinction, although there are some outlying individuals. In addition, we also generate a multi-dimensional scaling (MDS) plot that can be used to visualize the results.

These analyses should be performed on a set of SNPs that are approximately in linkage equilibrium: we achieve this by using *PLINK's* command to remove highly correlated, nearby SNPs.

Purpose	Create a LD pruned set of markers (first step)
Command	`plink --bfile wgas3` `--indep-pairwise 50 10 0.2` `--out prune1`
Input	`wgas3.bed` *QC+ whole genome SNP binary fileset* `wgas3.bim` `wgas3.fam`
Output	`prune1.prune.in` *List of SNPs included after pruning* `prune1.prune.out` *List of SNPs excluded after pruning*
Notes	*This option does not actually remove any SNPs, it just creates two lists of SNPs, which we use below. This removes any SNP that has r-squared > 0.2 with another SNP within a 50-SNP window; this window is shifted across the chromosome 10 SNPs at a time.*

We next calculate identity-by-state (IBS) allelic similarity between of all possible pairs of all 89 QC+ individuals, and store this information in a file.

Purpose	Calculate genome-wide IBS sharing based on pruned marker list
Command	`plink --bfile wgas3` `--extract prune1.prune.in` `--genome` `--out ibs1`
Input	`wgas3.bed` *QC+ whole genome SNP binary fileset* `wgas3.bim` `wgas3.fam`
Output	`ibs1.genome` *IBS sharing data (1 row per pair of individuals)*
Notes	*Equivalently, one could* `--exclude prune1.prune.out`

Finally, using the pairwise IBS information in `ibs1.genome`, we perform stratification analysis:

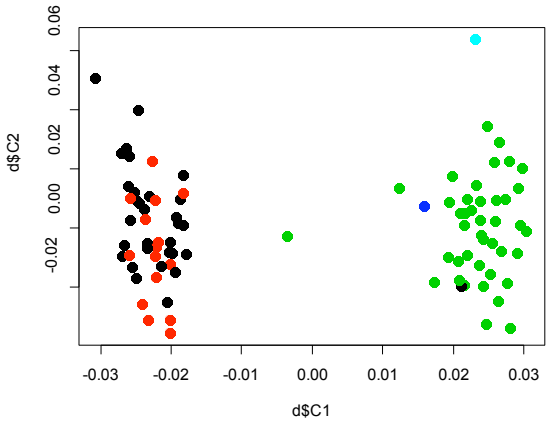
Purpose	Cluster individuals into homogeneous groups and perform a multidimensional scaling analysis
Command	`plink --bfile wgas3` `--read-genome ibs1.genome` `--cluster` `--ppc 1e-3` `--cc` `--mds-plot 2` `--out strat1`
Input	`wgas3.bed` *QC+ whole genome SNP binary fileset* `wgas3.bim` `wgas3.fam` `ibs1.genome` *Pre-calculated pairwise IBS values*
Output	`strat1.cluster2` *Assignment to cluster for each individual* `strat1.mds` *First 2 MDS components for each individual*
Notes	*Constraints on clustering are the PPC test* (`--ppc 1e-3`) *and to ensure that each cluster contains at least one case and one control* (`--cc`)

For more details on the clustering procedure, please refer to the PLINK manuscript (AJHG, 2007). How many clusters are in this solution? To visualize the cluster solution, you can use R. Start R, and set the current working folder/directory to the one your data are in (From File/Change dir… menu option). Then type

```
p <- read.table("strat1.mds", header=T)
plot( d$C1 , d$C2 , pch = 20 , cex = 2 , col = d$SOL + 1 )
```

which should generate a plot of the first two MDS components, with individuals coloured according to the cluster assignment based on the SNP data. The two main cluster represent Chinese (left) and Japanese (right) individuals.

You could also load the MDS file into HaploView and plot it using that software package.



MERGING IN NEW GENOTYPE DATA

The files `extra.ped` and `extra.map` contain new SNP data on the same set of individuals. These are SNPs taken from the region around rs11204005, the best SNP in the previous WGAS analysis. We first examine these SNPs by themselves, and then merge them into the SNPs in that region from the original WGAS dataset.

Purpose	Examine the new SNPs, testing for association stratified by population	
Command	`plink --file extra` `--mh --within pop.cov` `--out strat2`	
Input	`extra.ped`	*New followup SNP genotyping*
	`extra.map`	
	`pop.cov`	*Population label*
Output	`strat2.cmh`	*CMH results for new genotypes*
Notes		

As evident in the result file `strat2.cmh`, there are some very strongly associated SNPs in this new set, in particular rs7835221 (with a *P*-value = 3×10^{-14}). We next merge this new data with the old.

Purpose	Focus on region of association in WGAS data, and merge in new genotype data, creating a new fileset	
Command	`plink --bfile wgas3` `--snp rs11204005 --window 100` `--merge extra.ped extra.map` `--make-bed` `--out followup`	
Input	`wgas3.bed`	*QC+ binary fileset*
	`wgas3.bim`	
	`wgas3.fam`	
	`extra.ped`	*New genotype data (same individuals)*
	`extra.map`	
Output	`followup.bed`	*Merged fileset for region around top hit*
	`followup.bim`	
	`followup.fam`	
Notes	*The* `--snp` *and* `--window` *commands extract a particular region from* `wgas3` *first, and then merge in the new genotype data in* `extra.ped`	

We can check that the associations remain the same after merging these two filesets:

Purpose	Re-run association to check integrity of file
Command	`plink --bfile followup` `  --mh --within pop.cov` `  --out followup-cmh`
Input	`followup.bed` *Merged binary fileset for best region* `followup.bim` `followup.fam`
Output	`followup-cmh.cmh` *CMH for top region in merged dataset*
Notes	*Now focusing on the top region, using* `--adjust` *is no longer appropriate*

EXPLORING LINKAGE DISEQUILIBRIUM AND HAPLOTYPES

In our new dataset, "`followup`" (a binary fileset), we can use *PLINK's* LD-clumping procedure a set of SNPs that are all correlated with the same association signal (above an r-squared of 0.1).

Purpose	LD-based results clumping
Command	`plink --bfile followup` `  --clump followup-cmh.cmh` `  --clump-verbose` `  --clump-r2 0.1` `  --clump-annotate OR,A1`
Input	`followup.bed` *Merged binary fileset for best region* `followup.bim` `followup.fam` `followup-cmh.cmh` *Results file for the clump procedure.*
Output	`plink.clumped` *CMH for top region in merged dataset*
Notes	*Clumping options can also clump results from >1 results file. The OR and A1 in the command refer to fields in the file* `followup-cmh.cmh` *that are to be included in the* `plink.clumped` *report*

This analysis indicates four other SNPs that are associated and in LD with the primary SNP rs7835221. These five SNPs will form the focus of haplotype analysis below.

rs2460915 rs7835221 rs2460911 rs11204005 rs2460338

We can also use HaploView to explore the LD in this region visually.

Purpose	Create an output file for HaploView of this region
Command	`plink --bfile followup` `  --recodeHV` `  --out hv1`
Input	`followup.bed` *Merged binary fileset for best region* `followup.bim` `followup.fam`
Output	`hv1.ped` *Haploview-friendly version of follow-up fileset* `hv1.info`
Notes	*Load as standard "Linkage" format file in Haploview, not a "PLINK" file*

Load this newly created dataset (`hv1.ped` and `hv1.info`) in HaploView to examine the LD around this best SNP.

For the final exercises, we will extract just these five SNPs in another dataset (purely for convenience).

Purpose	For convenience, focus on the 5 clumped SNPs for further analysis (and so create a new dataset containing just these)
Command	`plink --bfile followup` `  --snps rs2460915,rs7835221,rs2460911,rs11204005,rs2460338` `  --make-bed` `  --out followup2`
Input	`followup.bed` *Merged binary fileset for best region* `followup.bim` `followup.fam`
Output	`followup2.bed` *Binary fileset of 5 SNPs in LD in top region* `followup2.bim` `followup2.fam`
Notes	*Note that* `--snps` *(versus* `--snp`*) can take a comma-delimited list of SNPs*

As an aside, the pairwise LD (r-squared) can also be calculated using *PLINK*. By default, only SNP pairs with high LD are shown in the output file.

Purpose	Report pairwise LD (r-squared) for SNPs in this region
Command	`plink --bfile followup2` `  --r2`
Input	`followup2.bed` *Merged binary fileset for best region* `followup2.bim` `followup2.fam`
Output	`plink.ld` *List of r-squared LD values (above threshold)*
Notes	*Add the* `--matrix` *option to get a 5×5 matrix of r-squared statistics*

HAPLOTYPIC ANALYSIS

We will use *PLINK*'s haplotype phasing algorithm to test for haplotypic association between these five SNPs and disease. The haplotype command used here ("`--chap`") is explicitly designed to focus on small regions such as this, rather than automated, genome-wide haplotype-based scans.

Purpose	Omnibus and haplotype-specific association tests for this region
Command	`plink --bfile followup2` `  --chap` `  --hap-snps rs2460915-rs2460338` `  --each-versus-others`
Input	`followup2.bed` *Merged binary fileset for best region (5 SNPs)* `followup2.bim` `followup2.fam`
Output	`plink.chap` *Haplotype test results*
Notes	*The* `--chap` *command requires that* `--hap-snps` *is explicitly specified; ranges can be used, as in the above example (i.e .implies all 5 SNPs)*

Here we see the omnibus test is significant; some haplotype-specific tests are also highly significant.

We can go further and use this framework to try to provide evidence consistent with an effect of a variant being solely due to *indirect association* with another variant, or whether it perhaps has its own independent

effect (and so might be causal). For example, here we ask whether the first SNP `rs2460915`, has any effect independent of the haplotypic background formed by all five of these SNPs.

Purpose	Tests of independent effect for each SNP
Command	`plink --bfile followup2` `--chap` `--hap-snps rs2460915-rs2460338` `--independent-effects rs2460915`
Input	`followup2.bed` *Merged binary fileset for best region (5 SNPs)* `followup2.bim` `followup2.fam`
Output	`plink.chap` *Conditional haplotype test results*
Notes	*Asks whether* rs2450915 *has any effect independent of haplotypic background; repeat this for the other 4 SNPs in this dataset*

We see that although `rs2460915` has a highly significant standard test statistic, it is not significant independent of the other SNPs. Repeat this exercise for the other four SNPs. What do you conclude?

Looked at another way, we can ask whether a particular SNP by itself can explain the entire omnibus association result we observed, as below.

Purpose	"Sole-variant" tests for each SNP
Command	`plink --bfile followup2` `--chap` `--hap-snps rs2460915-rs2460338` `--control rs2460915`
Input	`followup2.bed` *Merged binary fileset for best region (5 SNPs)* `followup2.bim` `followup2.fam`
Output	`plink.chap` *Conditional haplotype test results*
Notes	*Asks whether there is any omnibus association after controlling for* rs2450915 *(i.e. similar to asking whether the other four SNPs jointly have any independent effect). This test can also control for haplotypes, e.g.* `--control TGTAG`

Repeat this for all 5 SNPs. What do you conclude?

This table represents three types of test (basic single SNP, and these two conditional haplotypic tests) for these five SNPs. They are consistent with rs7835221 (the most highly associated SNP) being the sole cause of all association in this region. This doesn't mean it is necessarily the true, causal variant, of course, as there may be other unmeasured variants that have a more direct association. Nonetheless, it does tell us that the other four SNPs do not contribute any association information beyond rs7835221 alone. So, for example, all other things being equal, it would not necessarily be worth genotyping them all, as well as rs7835221, in any replication sample.

	***P*-values**		
SNP	**Single SNP**	**Independent SNP effect**	**Omnibus test controlling for SNP**
rs2460915	0.002	0.32	2×10^{-8}
rs7835221	**2×10^{-15}**	**1×10^{-6}**	**0.83**
rs2460911	0.0004	0.19	5×10^{-8}
rs11204005	8×10^{-6}	0.66	7×10^{-6}
rs2460338	0.001	0.70	1×10^{-7}

IN SUMMARY

- We performed whole genome association analysis
  - summary statistics and QC
  - stratified and stratification analyses
  - detailed follow-up tests and genotyping
  - conditional and unconditional association analysis

- The best SNP to emerge from the WGAS scan was rs11204005, that we found
  - showed no missing data or HW biases
  - was consistent with an allelic, dosage effect
  - had a common (47%) A allele with a strong protective effect (~0.09 odds ratio)
  - alternatively, a 53% G allele with strong risk effect (~11 odds ratio)
  - had similar effects (but not frequencies) in Japanese and Chinese subpopulations

- We went on to find a new single SNP rs7835221, not in the original scan, that is highly significant (*P*-value = 2×10^{-15}) and that, based on haplotype analysis, appears to explain the multiple associated SNPs in that region, including rs11204005.

- The SNP rs7835221 was indeed the only simulated true disease variant.

- Life is not always this simple