GWAS Tutorial with PLINK and Haploview

This tutorial is derived from the PLINK tutorial at http://pngu.mgh.harvard.edu/~purcell/plink/tutorial.shtml, simplified and adapted to use gPlink (a graphical version of PLINK which is integrated with the haplotype viewer Haploview)

What is a Genome Wide Association Study?

"GWAS depend on the fact that the 3 billion base pairs of genetic letters in humans are 99.9 percent identical in every person. So, small changes are significant are reasonably rare. If you can find them, then you can look for associations between the changes and disease risk.
Researchers believe there are some 10 million common SNPs in the human genome. Scanning the genomes of large numbers of patients for such a large number of variants would be prohibitively expensive. Fortunately, a major shortcut has been discovered that reduces the workload about 30-fold. When the International HapMap Project was completed in October 2005, the researchers demonstrated that the 10 million variants cluster into local neighborhoods, called haplotypes, and that they can be accurately sampled by as few as 300,000 carefully chosen SNPs. New technological systems allow these SNPs to be systematically studied in high-throughput facilities that dramatically lower the cost."
Source: http://www.genome.gov/17516714

VIDEO: Disease Susceptibility - Gene-disease Association Studies (2min)

Complex traits GWAS browser

Genome-Wide Associations (GWA) Karyogram

What happens in a GWAS?

Essentially a large case/control study: (Blue text = bioinformatics activity)

Requires large groups of individuals, some with a particular phenotype/disease (cases),

the rest without (controls). Try and match for 'confounding variables' like race, sex, ethnicity. If not, stratify the dataset so like groups are compared.

1. Genotype each individual (that is, assay a large number of single nucleotide polymorphisms (SNPs) spread throughout the genome. Large number = 10,000-1,000,000)

2. Quality control of genotype data. E.g. remove individuals/SNPs with a high percentage of missing data.

3. Look for statistical associations between the genotypes at each locus and phenotype status to identify areas that are linked to disease susceptibility. Use SNP by SNP statistical test - chi-squared or similar

4. Fine mapping of association signal by directed genotyping of additional SNPs in relevant areas. Fine mapping of Linkage Disequilibrium in relevant areas. Empirical derivation of haplotypes (strings of SNPS on the same chromosome)

5. Replication on another large independent cohort of cases and controls. Genotype only nominated SNPs (much cheaper). Replicate results using association tests.

6. Biological Validation: Identify risk-enhancement variant, examine functional consequences of variant, determine mechanism of risk enhancement

---

GWAS tools

There are a number of good free GWAS tools available. One of the most widely used is the PLINK toolset, developed by the Broad Institute.

PLINK

*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. The focus of **PLINK** is purely on analysis of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data).*

What does PLINK do?

1. **Data management**
2. **Summary statistics for quality control**
3. Population stratification detection
    - Complete linkage hierarchical clustering
    - In the population, there are different groups characterised by common genotypes
4. **Basic association testing**
    - **Case/control**
        - **Standard allelic test**
        - **Fisher's exact test**
        - **Cochran-Armitage trend test**
        - **Dominant/recessive and general models**
5. Multimarker predictors, haplotypic tests
6. Copy number variant analysis

Haploview

Haploview is an application for haplotype and statistical genomics analysis and visualisation.

What does Haploview do?

- Linkage Disequilibrium & haplotype block analysis
- single SNP and haplotype association tests
- **visualization and plotting of PLINK whole genome association results including advanced filtering options**

gPLINK

What does gPLINK do?
gPLINK = a common interface for data analysis (via PLINK) + visualisation (via Haploview)

Background to the tutorial

The tutorial is based on analysis of a artificial GWAS dataset representing the genotypes of 83534 SNPs from each of 89 individuals, in which 44 of the individuals are the 'cases', exhibiting a disease phenotype and the remaining 45 of the individuals are the 'controls' with no disease phenotype. The actual disease is not important, we're just looking for an association between genotype and phenotype in the data.

The GWAS data is stored in 2 text files:

genotypes.ped holds the phenotype and genotype data. It's formatted as 89 lines of 167077 columns:

```
Column1 = FamilyID
Column2 = IndividualID
Column3 = PaternalID
Column4 = MaternalID
Column5 = Sex
Column6 = Phenotype (represented as a 1, 2 or 0 [1=unaffected, 2=affected, 0=missing])
Column7+8 = genotype pair at SNP1 (represented as a pair of bases, one column for each allele)
Column9+10 = genotype pair at SNP2
.
.
Column167073+167074 = genotype pair at SNP83534
```

Each genotype is represented as a pair of bases, one for each allele. Example:

```
3 SNP genotypes: G G  C T  A T
```

Columns 2,3,4 and 5 are not relevant to this tutorial. They are relevant for linkage studies and not required for basic association testing. We will only consider autosomal chromosomes (1-22) for simplicity. PLINK can analyse the X chromosome, but it does so in a simplistic way which may not be optimal.

ped file format is described more fully here

genotypes.map represents the chromosomal positions of each SNP that has been

genotyped in the following format:

```
Column1 = Chromosome#
Column2 = SNPIdentifier
Column3 = Genetic Distance (in morgans)
Column4 = Physical base-pair position (bp units)
```

Columns 3 and 4 are not required for basic association testing.

map file format is described more fully here

Examine the data files

Create a directory 'gwas_ws' (for the examples below, this directory has been created in the C: drive on a Windows machine 'C:\gwas_ws', but the tutorial will work on any computer for which a PLINK binary is available)
Download and extract the workshop to this directory
Open the gwas_ws directory; you should see 4 files (genotypes.map, genotypes.ped, genotypes_5snps.map, genotypes_5snps.ped) and one subdirectory 'PLINK'

genotypes.ped is too large to easily view in a text editor - even a single line is ~170000 characters. Instead we'll examine a smaller example set, containing the first 5 SNPs from each entry in **genotypes.ped**.

Open up the file **genotypes_5snps.map** file in a text editor, or in an spreadsheet. Note that there are 5 SNPs listed according to the file format described above. All of these SNPs are on chromosome 2 (the first column is chromosome number). Open up **genotypes.map** for comparison, and note that the 83534 SNPs are spread across all 22 autosomal chromosomes.

Now open **genotypes_5snps.ped** in a text editor or a spreadsheet. You should see lines in this format:

```
Sample_181 1 0 0 1 1  G G  T T  T T  A G  C C
Sample_182 1 0 0 1 1  G G  T T  T T  A G  A C
Sample_183 1 0 0 1 2  G G  T T  T T  G G  C C
Sample_184 1 0 0 1 1  G G  T T  T T  G G  A C
.
.
Sample_269 1 0 0 1 2  G G  T T  T T  G G  A C
```

Note there are 89 lines in the file, one for each sample (i.e one per individual in the study).

If you're comfortable with spreadsheets, you might like to open the file in a spreadsheet and count the number of affected and unaffected individuals in column 6 (remembering that roughly equal numbers will provide the best statistical power). You may need to right click on the file and select the program you want to open the file with (e.g. Microsoft Excel). The Excel function COUNTIF() will be useful.

If you're game, open up the full **genotypes.ped** file in a text editor or spreadsheet and repeat this task. You'll notice that some of the genotypes are '0 0', meaning that that SNP genotype is missing for that sample. This is common in genotyping, and can occur due to poor sample quality or technical problems with the assay.

Start gPLINK and configure it

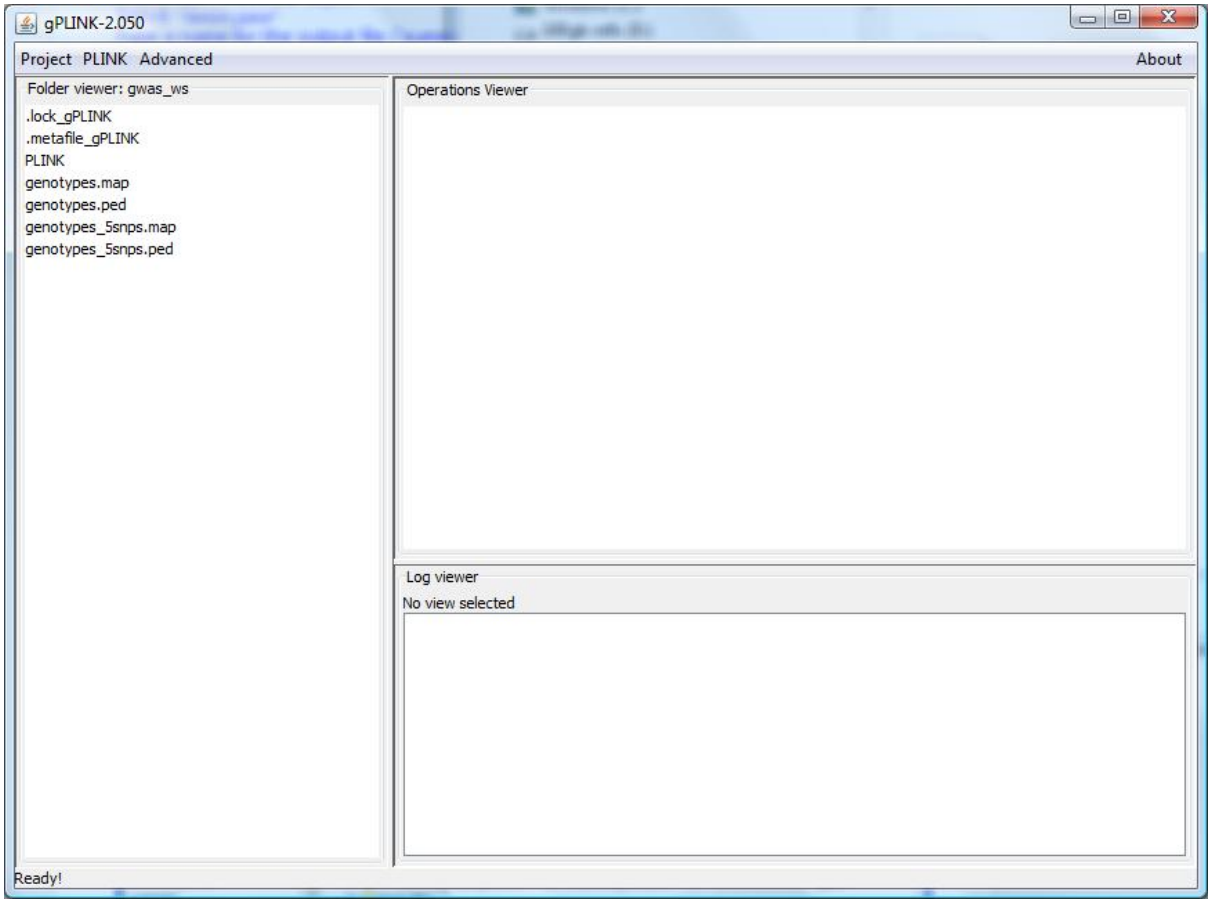
Change to the gwas_ws directory created above
Change to the PLINK directory

If your computer is set up for it, you might just be able to double click on the gPLINK file and it will start. If this doesn't work:

Start a command prompt and go to the PLINK directory
Start the gPLINK application with the following command: `java -jar gPLINK.jar`

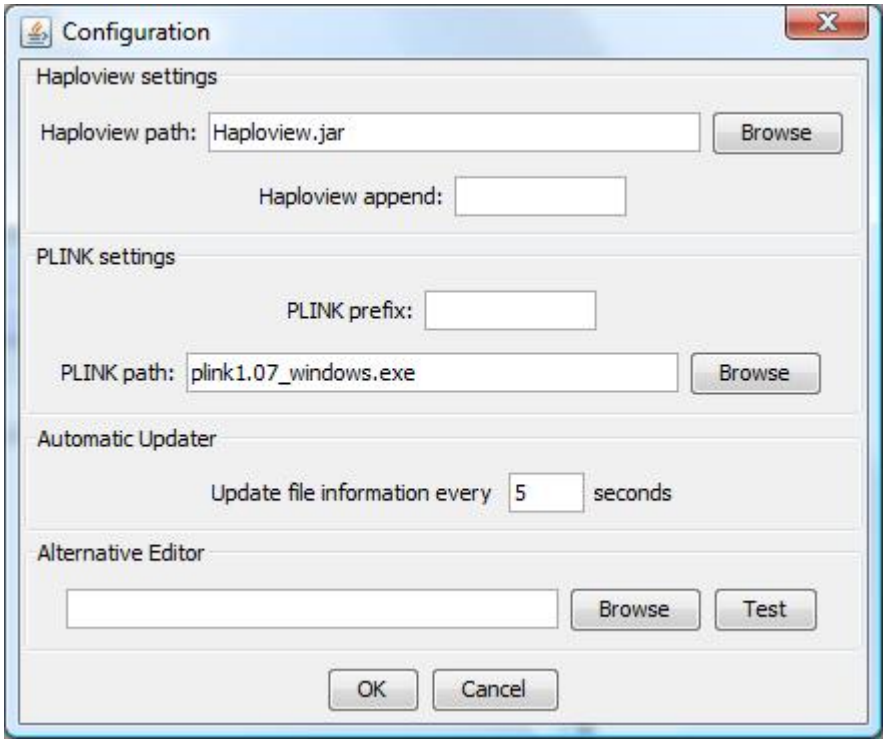
The gPLINK window will open.

Select `Project>Open and` browse to your `gwas_ws` directory
Do NOT click the '`SSH link...`' option
Click OK.



Then, configure gPLINK so it can find the PLINK and Haploview applications:

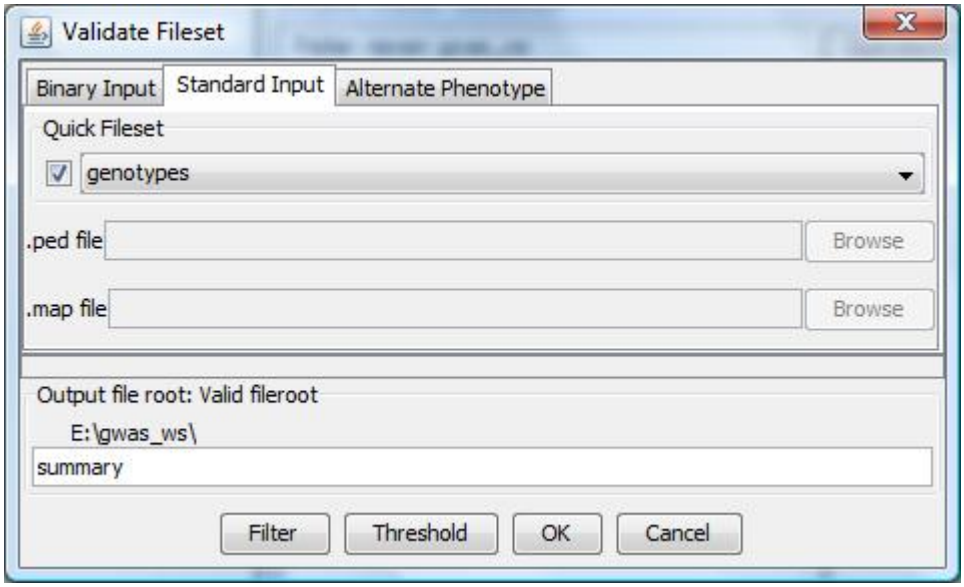
Select `Project>Configure` and browse to the correct application locations (both Haploview and PLINK are in the `gwas_ws\PLINK` folder). **You'll need to select the version of PLINK that's appropriate for your computer (Windows, Mac, Linux).**
Click OK.



Validate and generate summary statistics on the GWAS dataset

Before we do any analysis, we will need to check that the dataset is valid (conforms to the specifications of the PLINK file format)

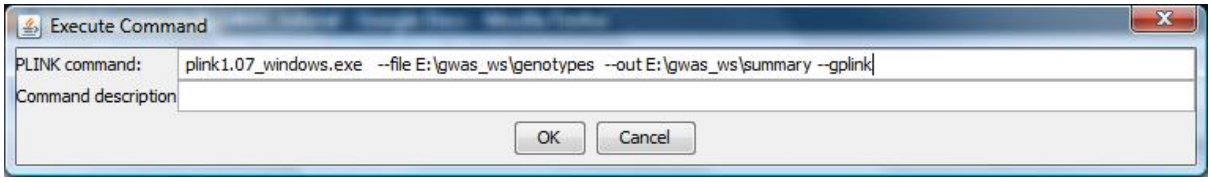
`Select PLINK>Summary Statistics>Validate Fileset`
Select tab `'Standard Input'`
Select `'Genotypes'`
Type a name for the output file ('summary')



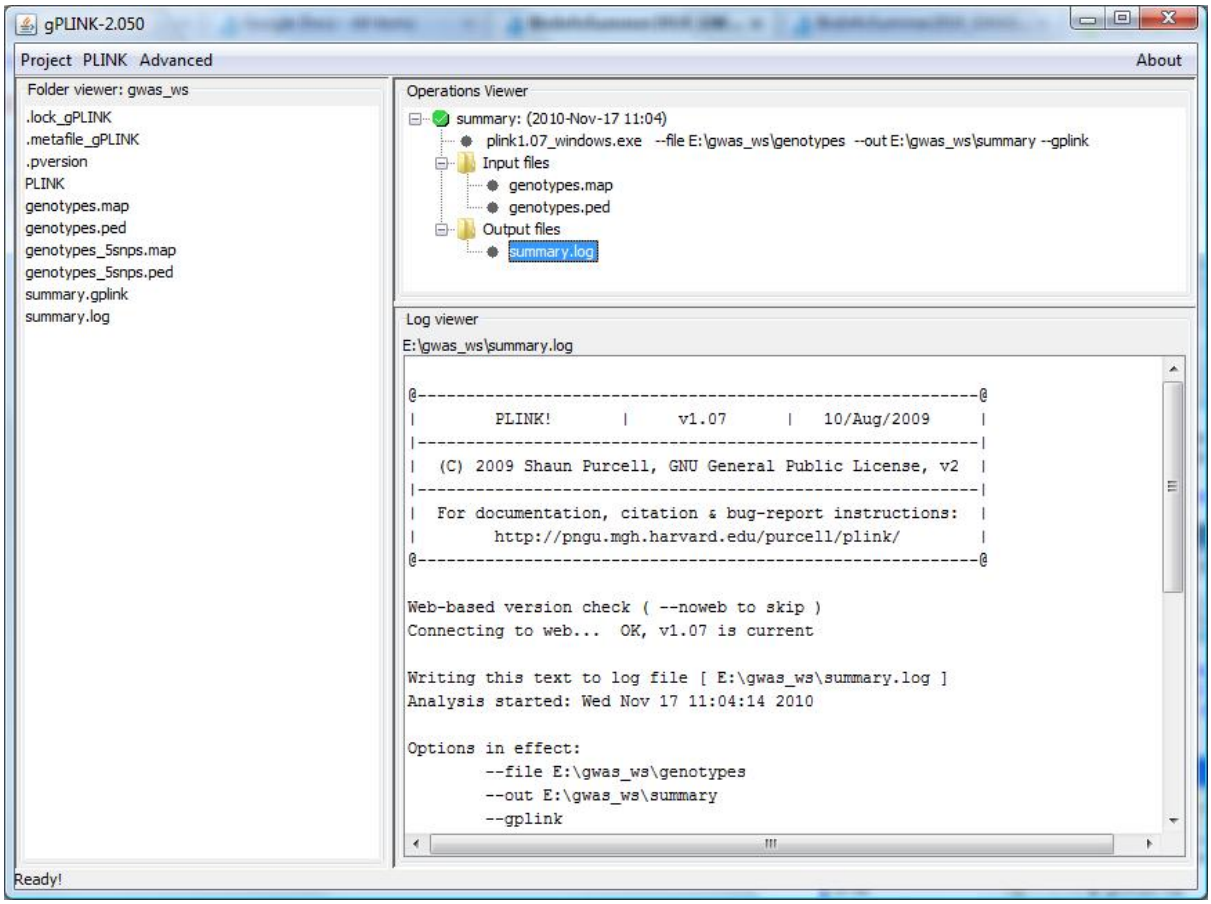
The following window appears, showing the PLINK command that is generated. gPLINK is

just a graphical interface to PLINK that uses menus and buttons to create PLINK commands.

You can give your command a name, such as 'summarise' below, for later reference.



The command will run and the results will be written to a file in c:\gwas_ws as whatever filename you specified above (in this case 'c:\gwas_ws\summary'). gPLINK will identify the output files and list all relevant information about the command you constructed and ran, as below:



Now we will examine the results.

Every PLINK command generates a log file which provides useful information about the data set and the command that was just executed. This file will have the extension '.log'. Most PLINK commands (but not this one) will generate additional files containing results. These have different file extensions specific to each command.

Open the '**summary.log**' file in gPLINK under the 'Ouput files' folder. You should see various information about the **genotypes.ped** dataset.

```
@----------------------------------------------------------@
|        PLINK!       |     v1.07      |   10/Aug/2009     |
|----------------------------------------------------------|
|  (C) 2009 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-based version check ( --noweb to skip )
Connecting to web...  OK, v1.07 is current

Writing this text to log file [ C:\gwas_ws\summary.log ]
Analysis started: Sun Mar 14 12:10:34 2010

Options in effect:
    --file C:\gwas_ws\genotypes
    --out C:\gwas_ws\summary
    --gplink

83534 (of 83534) markers to be included from [ C:\gwas_ws\genotypes.map ]
89 individuals read from [ C:\gwas_ws\genotypes.ped ]
89 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
44 cases, 45 controls and 0 missing
89 males, 0 females, and 0 of unspecified sex
Before frequency and genotyping pruning, there are 83534 SNPs
89 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 0.99441
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 83534 SNPs
After filtering, 44 cases, 45 controls and 0 missing
After filtering, 89 males, 0 females, and 0 of unspecified sex

Analysis finished: Sun Mar 14 12:10:45 2010
```

Summary statistics: Missing rates

```
0 of the 89 individuals removed for low genotyping ( MIND > 1 )
```

What does this mean?
The initial step in all data analysis is to exclude individuals with too much missing genotype data. MIND is the Maximum INDividual missingness rate - i.e. SNP data missing for an individual. Intially it is set to 1, so we only exclude samples with 100% of SNP data missing.

Shortly we will use a more stringent test for our data, specifying MIND > 0.1. This will remove any samples that have more than 10% of SNP genotypes missing

```
0 SNPs failed missingness test ( GENO > 1 )
```

What does this mean?
The second analysis step is to remove individual SNPs which are missing from too many samples - that is, check by column rather than by row, and if a SNP is missing values in >10% of samples. Initially GENO is set to >1, so no SNP columns are removed.

Shortly we will apply more stringent criteria, such that GENO > 0.1. In this case, 0.1*89 = 8.9 samples, meaning that if a SNP is missing in 9 more more samples, that SNP will be removed from the dataset.

Summary statistics: Allele frequencies

```
0 SNPs failed frequency test ( MAF < 0 )
```

What does this mean?
MAF is the Minor Allele Frequency. It can be used to exclude SNPs which are not informative because they show little variation in the sample set being analysed. For instance, if a SNP shows variation in only 1 of the 89 individuals, it is not useful statistically and should be removed.
Initially, MAF is set to < 0, meaning that no SNPs are excluded.

Now we will run a more stringent allele frequency test specifying MAF < 0.01 is grounds for exclusion.

```
Select PLINK>Summary Statistics>Missingness
```
Select the **genotypes** dataset as before. This time, click 'Threshold', and set the missingness and allele frequency thresholds as described above (that is, set MIND > 0.1, GENO > 0.1 and MAF < 0.01). Make sure you check the appropriate boxes. Check the 'Maximum individual missingness rate' and make sure it's set to 0.1. Then run the command and check the output.

Note that 858 SNPs have been removed by this criteria.
Also note 16993 SNPs failed the frequency test (MAF < 0.01)

```
0 of 89 individuals removed for low genotyping ( MIND > 0.1 )
Total genotyping rate in remaining individuals is 0.99441
858 SNPs failed missingness test ( GENO > 0.1 )
16993 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 65804 SNPs
After filtering, 44 cases, 45 controls and 0 missing
After filtering, 89 males, 0 females, and 0 of unspecified sex
```

What does this mean?
858 SNPs were missing more than 10% of their genotypes across the sample set, so should be excluded from the subsequent analysis. We do this by applying the thresholds in any subsequent steps.
16993 SNPs had a Minor Allele Frequency of < 1%, so should also be excluded.

Note that these thresholds are somewhat arbitrary, but seem to be a good empirical starting point.

Basic Association Analysis

Now we will try a basic association analysis - that is, we will identify the alleles that are most closely associated with the phenotype. We will set thresholds to exclude SNPs with high missingness and low minor allele frequencies, as above.

PLINK>Association>Allelic Association Tests
Select the dataset, keep all the default values for parameters, set the thresholds as in the previous section, choose a meaningful output filename (e.g. 'genotypes_allelictest') and run the command. It will take a few seconds.

Look at the output file (not the log file) in a text editor (you can do this from inside gPLINK by right clicking on a file and selecting 'Open in default viewer')

The output file should look like this:

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P
1	rs6681049	789870	C	0.1591	0.2667	T	3.067	0.07991
1	rs4074137	1016570	A	0.07955	0.07778	C	0.001919	0.9651
1	rs1891905	1090080	A	0.4091	0.4	G	0.01527	0.9017
1	rs9729550	1125105	A	0.1705	0.08889	C	2.631	0.1048
1	rs3813196	1159244	A	0.03409	0.02222	G	0.2296	0.6318
1	rs12044597	1698661	A	0.5	0.4889	G	0.02198	0.8822
1	rs10907185	1723079	A	0.3068	0.2667	G	0.3509	0.5536
1	rs11260616	1751865	A	0.2326	0.2	T	0.2754	0.5998
1	rs745910	1819342	A	0.1395	0.1932	G	0.9013	0.3424
1	rs262688	2103425	A	0.2045	0.1667	C	0.4227	...

Each row is a single SNP association result. The fields are:

- Chromosome
- SNP identifier
- Location of the SNP on the chromosome (called physical distance and measured in bp)
- Code for allele 1 (the more rare or 'minor' allele based on the entire sample frequencies)
- The frequency of this variant in affected individuals (cases)
- The frequency of this variant in unaffected individuals (control)
- Code for allele 2 (the more common or 'major' allele)
- The chi-squared statistic for this test (1 df)
- The asymptotic p-value for this test
- The odds ratio for this test

P-value

The important field we are looking for is the P-value, which effectively is a measure of how strongly associated the values for that SNP are with the phenotype values across the entire sample set. So, we want to sort the results by P value.

Open the output file in a spreadsheet and sort by P-value (Data > Sort in Excel). Which SNPs are most strongly associated with the phenotype?

Other informative fields are F_A and F_U, for Frequency_Affected (cases) and Frequency_Unaffected (controls) - clearly, SNP alleles associated with the phenotype will be seen more frequently in cases than controls, or vice versa. As we will see in the next section, allele frequency is actually the basis of the P-value.

PLINK can of course sort the SNPs for us, and we will do this in the next sections when we do a more advanced association test.

Understanding the statistics

The idea behind association analysis is to look through each SNP one by one, testing to see if there is a difference in the frequency of alleles seen with cases vs controls. If this difference is statistically significant, then that allele can be said to be associated with the phenotype.

The method of statistically testing the frequencies is through a chi-squared test or similar. This essentially summarises the difference between the observed allelic frequencies at a SNP and the expected allelic frequencies (if there were no difference between cases and controls, this should just reflect the allele frequencies of this SNP in the population).

The simplest type of frequency analysis is called an 'allelic test'.

Basic Allelic Test for association
The unit of this test is the allele rather than the individual (remember each individual contributes two alleles). Each of the genotypes aa, aA, and AA are divided into pairs of alleles a and a, a and A, or A and A. The associations of the phenotype with these individual alleles are then tested.
An advantage of this test is that the number of observations has been doubled.

Essentially, this test asks: **is any single allele associated with this phenotype?**
It ignores any interaction or relationship between alleles at a locus.

This is perhaps the easiest test to understand. Let's manually run through an allelic test at a single locus. Say we had a group of 100 individuals, split into 50 affected (cases) and 50 unaffected according to some phenotype. For a particular SNP, we assume that alleles A and a are equally common in the population (50:50 distribution). If there is no association between the allele present and the phenotype we would expect this distribution of allele counts:

The **expected** contingency table would be:

	allele a	allele A
Case (aff)	50 [$E_{a,aff}$]	50 [$E_{A,aff}$]
Control (unaff)	50 [$E_{a,unaff}$]	50 ($E_{A,unaff}$]

When we genotype the group, we might observe a different distribution of allelic frequencies for this SNP. The **observed** contingency table might be:

	allele a	allele A
Case (aff)	25 [$O_{a,aff}$]	75 [$O_{A,aff}$]
Control (unaff)	75 [$O_{a,unaff}$]	25 [$O_{A,unaff}$]

To determine if these two distributions were significantly different, we perform a chi-

squared independence test.

What does this mean?

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

X^2 = the test statistic that asymptotically approaches a χ^2 distribution
O_i = an observed frequency
E_i = an expected (theoretical) frequency
n = the number of possible outcomes of each event

*The purpose of the chi-squared test is to get a single numerical value that represents the difference between what we would expect if a SNP was **not** associated with a phenotype - i.e. similar frequencies of occurance of the SNP in both cases and controls - and what we actually observe in the data we are analysing. As most genes and, by implication, most SNPs **won't** be associated with the phenotype, most of the time the difference will be small, and so will the value of the chi-squared statistic. If there's a large chi-squared value, then this means that a particular SNP is seen more commonly in samples with the phenotype than in samples without the phenotype - and so the SNP is **associated** with the phenotype.*

In the data above, we can intuitively see that there is a difference between the **observed** and **expected** contingency tables - allele A is much more common in cases than controls. We quantify this difference using chi-squared:

n=4 (there are 4 possible combinations of allele and outcome), so:
$\chi^2 = (O_{a,aff}-E_{a,aff})^2/E_{a,aff}+(O_{a,unaff}-E_{a,unaff})^2/E_{a,unaff}+(O_{A,aff}-E_{A,aff})^2/E_{A,aff}+(O_{A,unaff}-E_{A,unaff})^2/E_{A,unaff}$
$= (25-50)^2/50+(75-50)^2/50+(75-50)^2/50+(25-50)^2/50$
$= 625/50+625/50+625/50+625/50$
$= 12.5 + 12.5 +12.5 +12.5$
$= 50$

We then need to determine the significance of this value. To do this we calculate the P-value of the chi-squared result according to the degrees of freedom of the test. A 2x2 contingency table has 1 degree of freedom.

For these data, **P-value < 0.0001**, which means that the probability of obtaining these frequencies by chance is very low (<0.01%), and therefore the significance of the association is high.

What does this mean?
The P-value is a measure of the significance of the chi-squared test: essentially, an interpretation on what the chi-squared statistic means in terms of probability. Formally, a P-value is the probability of observing a value of the chi-squared statistic as large or larger than the one observed, by chance, if the null hypothesis is true. Here our null hypothesis is that there is truly no underlying difference in allele frequences between cases and controls.

If the P-value is very small, then that means that the observed case:control split of values for the SNP are very different to the expectation (that is, the expected values if the SNP and the outcome were independent). The P-value is very unlikely to occur by chance if the underlying population of cases and controls have the same allele

frequency, so we conclude that cases and controls do have different allele frequencies. The lower the P-value, the more unlikely our results are to have occurred under the null hypothesis.

The P-value resulting from a chi-squared test depends both on the chi-squared statistic and on the degrees of freedom (df) of the comparison. The latter is equal to (number of rows in contingency table - 1) x (number of columns -1), so it is (2-1) x (2-1) = 1 for a 2x2 table. For a given value of degrees of freedom, increasing the value of the chi-squared statistic lead to smaller p-values. The higher the degrees of freedom, the higher a chi-squared value we need for the same significance.

Any P-value less than 0.05 is considered statistically significant for a single test.

This approach does not require equal frequencies between the two alleles or between cases and controls. It will also work for different expected frequencies e.g. for a lower Minor Allele Frequency (MAF), and for which the split of case:controls is not 50:50
Let's assume that allele frequency is ~40:60 a:A (observed from the data), and the case control split is 30:70.
The contingency tables are then:

	Observed	Observed	Expected	Expected	
	a	A	a	A	*Totals*
Case (30)	**15**	**45**	**24** (80*60/200)	**36** (120*60/200)	*60*
Control (70)	**65**	**75**	**56** (80*140/200)	**84** (120*140/200)	*140*
Totals	*80*	*120*	*80*	*120*	*200*

In this case chi-squared
$= (15-24)^2/24+(65-56)^2/56+(45-36)^2/36+(75-84)^2/84$
$= 8.036$, and P-value = 0.0046

Try verifying these test values using this online chi-squared calculator. Calculate the P-value using the Chi-squared without Yates correction option, and a 2-tailed P-value. Alternatively, verify the p-value in Excel using the CHIDIST() function.

Calculating chi-squared for GWAS data

Go through the process of generating some P-values for the first SNP in the smaller **genotypes_5snps.ped** file. You'll need to count up the alleles for each of the 89 samples, counting separately for the cases (phenotype=2) and controls (phenotype=1) (summing the counts across each of the two possible values for the SNP), then enter them into the online chi-squared calculator. It's probably easiest to count the alleles by importing the data into a spreadsheet and using the COUNTIF() function.

Example: Counting allele frequencies for the **last** SNP in **genotypes_5snps.ped** for the first 12 samples (remember column 6 is the phenotype, and each SNP has two columns, one for each diploid allele)

```
Sample_181 1 0 0 1 1 G G T T T T A G C C
Sample_182 1 0 0 1 1 G G T T T T A G A C
Sample_183 1 0 0 1 2 G G T T T T G G C C
Sample_184 1 0 0 1 1 G G T T T T G G A C
Sample_185 1 0 0 1 1 G G T T T T A G A C
Sample_186 1 0 0 1 1 G G T T T T G G A C
Sample_187 1 0 0 1 1 G G T T T T A G C C
Sample_188 1 0 0 1 1 G G T T T T A A C C
Sample_189 1 0 0 1 1 G G T T T T A G A A
Sample_190 1 0 0 1 1 G G T T T T A G A A
Sample_191 1 0 0 1 2 G G T T T T G G C C
Sample_192 1 0 0 1 1 G G T T T T G G A C
```

Allele 'A': **Occurs 9 times controls, 0 times in cases**
Allele 'C': **Occurs 11 times in controls, 4 times in cases**

24 alleles total, as expected for 12 samples.

Check the P-value and chi-squared values against those from the output file you generated in the previous section. You can look up the name of the SNP you checked in the **genotypes_5snps.map** file.

Then find the two most significant SNPs from your prior assocation testing and build contingency tables for them. To make it a little easier, one of these SNP genotypes is included at the end of the **genotypes_5snps.ped** file.

Intuitively, you can see that the numbers for these SNPs look very different between case and controls. The chi-squared test is just a way to measure that difference, and the P-value is a way to objectively judge how statistically significant that chi-squared value is.

NOTE: There are other ways to calculate significance values! This is just one. The Association Analysis in PLINK reference describes the different tests in more detail.

Different association tests, different assumptions

Allelic tests are the simplest, but not necessarily the best test for association in humans, as they ignore the overall genotype of the two chromosomes. Genotypic tests form another category of test. The units for these tests are the genotype, rather than the allele.

Genotypic Tests

Basic genotypic test
We tabulate each genotype (homozygote aa, heterozygote Aa and homozygote AA) against case/control status. This means that instead of the 2x2 contingency table seen previously, we have a table with 2 rows and 3 columns, so the test has (2-1) x (3-1) = 2 degrees of freedom. The basic test makes no assumptions about the genetic model, but the extra degree of freedom means that a larger chi-squared value is needed to obtain the same p-value.

Additive Model
This model assumes that having two copies of the minor allele* (AA genotype) has twice the effect of having a single copy of the minor allele (Aa genotype). In other words, the more copies of the minor allele you have, the greater the effect, with heterozygotes having phenotypes lying between those of the two homozygotes. This test has 1 df. It is also known as the 'Cochran-Armitage test for trend'.

Dominant Model
This model assumes that an effect on phenotype is only seen if you have at least one copy of the minor allele. We categorise individuals into two groups based on whether they have at least one minor allele A (either Aa or AA) or no copies of the minor allele (aa). So we have a 2 x 2 table and our test has 1 df.

Recessive Model
This model assumes that an effect on phenotype is only seen if you have two copies of the minor allele. We categorise individuals into two groups based on whether this is the case (i.e. AA vs Aa or aa). Again, we have a 2 x 2 table and our test has 1 df.

*note the nomenclature here: the 'major' allele for a SNP is the one seen more frequently in the population.

Experience tells us that if we're not sure of the genetic model behind the genotype-phenotype association, then the 'Additive' genotypic test a good compromise. The basic test makes the fewest assumptions but is less powerful due to the extra degree of freedom.

Redo the association analysis using an additive genotypic test:

`PLINK>Association>Genotypic C/C association tests`

Choose the 'Permute trend test' [PLINK's name for the Cochran-Armitage additive model]
Check the 'adjusted p-values' box. This sorts the association results, and also includes p-values that are adjusted for multiple testing using a range of different methods.

What does this mean?
We are testing 83,534 SNPs here, not just one. The PLINK command performs the allelic test and the four genotypic tests, so we are performing 5 x 83,534 = 417,670 tests in total!
If we set the significance threshold to p=0.05, then each test has a 5% chance of falsely concluding that a difference does not exist. As the number of tests increases, so does the probability of observing at least one false positive test result. This is called 'the problem of multiple testing'. We need to account for a higher probability of seeing spurious associations just because we are performing so many tests. This can be done by either decreasing the significance threshold, or adjusting the p-values. PLINK takes the latter approach.

There are different methods to adjust p-values for multiple testing. PLINK presents the results of 7 different methods of adjustment. We won't go through the details, but note that the Bonferroni method assumes that each test is independent, and is too strict when this assumption is not true. SNPs that are located close to each other tend to have correlated values, and in addition we would expect different tests on the same SNP to be correlated.

Call your output file something sensible like: '**genotypes_trendtest**'

Run the command.

This should generate three files, the log file and two results files: genotypes_trendtest.model and genotypes_trendtest.model.trend.adjusted (the sorted list adjusted for multiple testing)

First, have a look at genotypes_trendtest.model.

genotypes.ped

CHR	SNP	A1	A2	TEST	AFF	UNAFF	CHISQ	DF
1	rs6681049	C	T	GENO	3/8/33	5/14/26	NA	NA
1	rs6681049	C	T	TREND	14/74	24/66	2.426	1
1	rs6681049	C	T	ALLELIC	14/74	24/66	3.067	1
1	rs6681049	C	T	DOM	11/33	19/26	NA	NA
1	rs6681049	C	T	REC	3/41	5/40	NA	NA
1	rs4074137	A	C	GENO	1/5/38	0/7/38	NA	NA
1	rs4074137	A	C	TREND	7/81	7/83	0.001794	1
1	rs4074137	A	C	ALLELIC	7/81	7/83	...	

genotypes_chr12.ped

CHR	SNP	A1	A2	TEST	AFF	UNAFF	CHISQ	DF
1	rs6681049	C	T	GENO	3/8/33	5/14/26	NA	NA
1	rs6681049	C	T	TREND	14/74	24/66	2.426	1
1	rs6681049	C	T	ALLELIC	14/74	24/66	3.067	1
1	rs6681049	C	T	DOM	11/33	19/26	NA	NA
1	rs6681049	C	T	REC	3/41	5/40	NA	NA
1	rs4074137	A	C	GENO	1/5/38	0/7/38	NA	NA
1	rs4074137	A	C	TREND	7/81	7/83	0.001794	1
1	rs4074137	A	C	ALLELIC	7/81	7/83	0.001919	1
1	rs4074137	A	C	DOM	6/38	7/38	NA	NA
1	rs4074137	A	C	REC	1/43	0/45	NA	NA
1	rs1891905	A	G	GENO	8/20/16	10/16/19	0.9127	2

This lists the results of the different tests done on the data, including genotype, additive, allelic, dominant and recessive. Of particular note are the frequencies of AFFected (cases) vs UNAFFected (controls) for each test. The TEST columns explains which test was used (GENO= basic genotypic, TREND=additive test, DOM=dominant test and REC=recessive test).

Find the results for the SNP 'rs2222162', which has a low p-value. This is the SNP you analysed in the previous section with a manual chi-squared test (the last SNP in genotypes_5snps.ped). You should see the allele frequencies match those that you calculated, and you should have the same chi-squared value.

CHR	SNP	A1	A2	TEST	AFF	UNAFF	CHISQ	DF	
2	rs2222162	A	C	GENO	3/19/22	17/22/6	NA	NA	
2	rs2222162	A	C	TREND	25/63	56/34	19.15	1	1
2	**rs2222162**	**A**	**C**	**ALLELIC**	**25/63**	**56/34**	**20.51**	**1**	**5**
2	rs2222162	A	C	DOM	22/22	39/6	NA	NA	
2	rs2222162	A	C	REC	3/41	17/28	NA	NA	

Open the other file '**genotypes_trendtest.model.trend.adjusted**' in a text editor. Note that the SNPs are now sorted by UNADJusted P-value, but looking at the other significance metrics, it seems like the ranking would remain the same. The columns after UNADJ show p-values adjusted for multiple testing using the 7 different methods. We can see that p-values values in these 7 columns are much larger than the unadjusted p-values.

We will focus on the FDR_BH column, which shows p-values adjusted using the Benjamini and Hochberg False Discovery Rate (FDR) method. If an adjusted p-value <0.05, then we know that the probability of the result under the null hypothesis is less than 5%, just as it is when we perform a single test and get an unadjusted p-value <0.05. In this case FDR_BH=8.41e-006, which is much lower than 0.05, so there is enough evidence to conclude that allele frequencies differ between cases and controls. If the FDR_BH value was higher than 0.05, we can't conclude that an association doesn't exist for that SNP - it still might - but the evidence is not strong enough to rule out a false positive result. In this case, we are confident that there is a real assocation. Interestingly, there look to be other SNPs on the same chromosome that are associated as well.

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	
12	rs11830226	1.278e-010	3.326e-009	8.41e-006	8.41e-006	8.41e-006	8.41e-006	8.4
12	rs11068090	3.185e-010	7.225e-009	2.096e-005	2.096e-005	2.096e-005	2.096e-005	1.04
12	rs2008370	1.195e-009	2.223e-008	7.867e-005	7.867e-005	7.867e-005	7.866e-005	2.0
12	rs12302542	1.222e-009	2.265e-008	8.04e-005	8.04e-005	8.04e-005	8.039e-005	2.0
12	rs7310400	2.53e-009	4.204e-008	0.0001665	0.0001665	0.0001665	0.0001664	3.32
12	rs1895624	4.489e-009	6.845e-008	0.0002954	0.0002954	0.0002954	0.0002954	4.92
12	rs1872650	7.017e-009	1.001e-007	0.0004617	0.0004617	0.0004616	0.0004616	6.59

Viewing and interpreting the analysis results

Haploview is a viewer specifically for statistical genomics studies. One of its functions is to plot the results of an association analysis across the genome so that they can be visualised.

Select the **genotypes_trendtest.model.trend.adjusted** (or whatever you called it) file and right click; then select 'Open with Haploview'. Choose the appropriate .map file (**genotypes.map**), which lists the chromosomal positions of the SNPs listed in the model file.

Haploview 4.2 -- genotypes_chr12_trendtest.model.trend.adjusted

File Display Analysis Help Key

PLINK

CHROM	MARKER	POSITION	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
12	rs11830226	114486655	1.278E-10	3.326E-9	8.41E-6	8.41E-6	8.41E-6	8.41E-6	8.41E-6	9.816E-5
12	rs11068090	115497733	3.185E-10	7.225E-9	2.096E-5	2.096E-5	2.096E-5	2.096E-5	1.048E-5	1.223E-4
12	rs2008370	114814120	1.195E-9	2.223E-8	7.867E-5	7.867E-5	7.867E-5	7.866E-5	2.01E-5	2.346E-4
12	rs12302542	114910478	1.222E-9	2.265E-8	8.04E-5	8.04E-5	8.04E-5	8.039E-5	2.01E-5	2.346E-4
12	rs7310400	115088109	2.53E-9	4.204E-8	1.665E-4	1.665E-4	1.665E-4	1.664E-4	3.329E-5	3.886E-4
12	rs1895624	115007688	4.489E-9	6.845E-8	2.954E-4	2.954E-4	2.954E-4	2.954E-4	4.924E-5	5.747E-4
12	rs1872650	114533586	7.017E-9	1.001E-7	4.617E-4	4.617E-4	4.616E-4	4.616E-4	6.596E-5	7.699E-4
12	rs35369	114080930	4.586E-8	4.938E-7	0.003018	0.003017	0.003013	0.003013	3.353E-4	0.003913
12	rs11829788	115698898	4.586E-8	4.938E-7	0.003018	0.003017	0.003013	0.003013	3.353E-4	0.003913
12	rs2172953	115662495	5.483E-8	5.749E-7	0.003608	0.003607	0.003601	0.003601	3.608E-4	0.004211
12	rs1424612	114737328	8.002E-8	7.929E-7	0.005266	0.005265	0.005252	0.005251	4.388E-4	0.005122
12	rs3858652	114697065	8.002E-8	7.929E-7	0.005266	0.005265	0.005252	0.005251	4.388E-4	0.005122
12	rs10744863	114636812	2.133E-7	1.826E-6	0.01403	0.01403	0.01394	0.01393	9.424E-4	0.011
12	rs10507269	115044710	2.133E-7	1.826E-6	0.01403	0.01403	0.01394	0.01393	9.424E-4	0.011
12	rs1895625	114966049	2.148E-7	1.837E-6	0.01414	0.01413	0.01404	0.01403	9.424E-4	0.011
12	rs1466852	116076226	3.504E-7	2.786E-6	0.02306	0.02305	0.02279	0.02279	0.001404	0.01639
12	rs7957270	115604900	3.84E-7	3.012E-6	0.02527	0.02526	0.02495	0.02495	0.001404	0.01639
12	rs1874894	114370689	3.84E-7	3.012E-6	0.02527	0.02526	0.02495	0.02495	0.001404	0.01639
12	rs10774890	115741081	4.742E-7	3.604E-6	0.0312	0.03119	0.03072	0.03071	0.001642	0.01917
12	rs3907612	115165677	5.565E-7	4.131E-6	0.03662	0.03661	0.03596	0.03595	0.001831	0.02137
12	rs3858646	114060076	1.044E-6	7.058E-6	0.0687	0.06868	0.06639	0.06637	0.003271	0.03818
12	rs4767332	114416655	1.638E-6	1.036E-5	0.1078	0.1078	0.1022	0.1022	0.004583	0.05349
12	rs10850682	115547107	1.639E-6	1.036E-5	0.1079	0.1078	0.1022	0.1022	0.004583	0.05349
12	rs3803088	115634433	1.718E-6	1.078E-5	0.113	0.113	0.1069	0.1068	0.004583	0.05349
12	rs1007791	115407141	1.741E-6	1.091E-5	0.1146	0.1145	0.1082	0.1082	0.004583	0.05349
12	rs2454400	114311894	1.825E-6	1.136E-5	0.1201	0.1201	0.1132	0.1131	0.004619	0.05392
12	rs11067863	114871593	2.04E-6	1.249E-5	0.1342	0.1342	0.1256	0.1256	0.004971	0.05802
12	rs2063943	116519039	2.729E-6	1.6E-5	0.1796	0.1795	0.1644	0.1643	0.006414	0.07486

Viewing 65804 results

Chr: Start kb: End kb: Filter: Filter

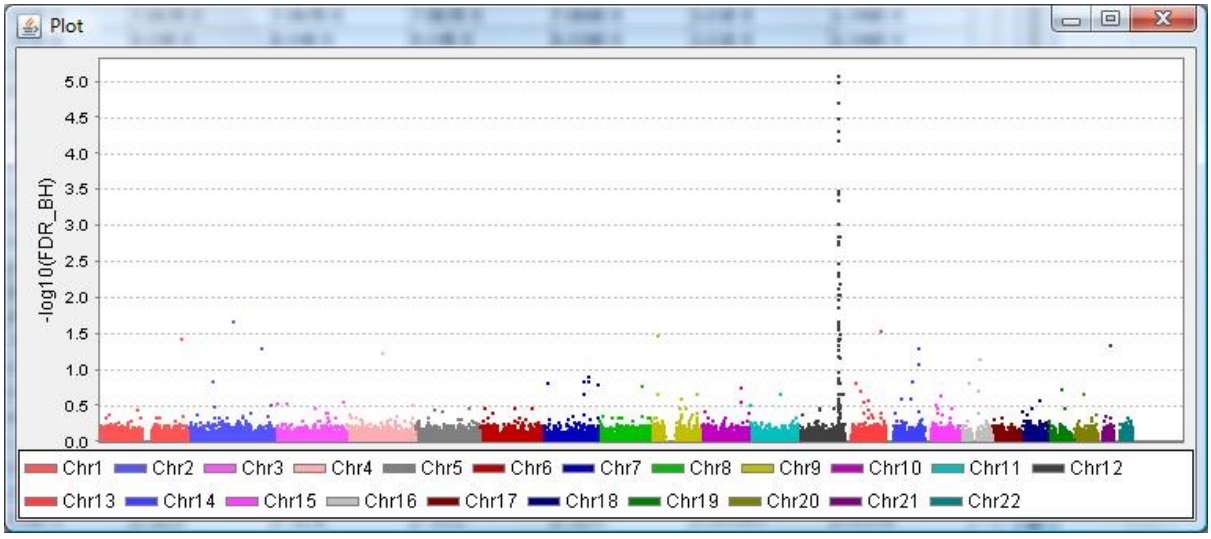
Specify Marker: Prune Table Remove Column: Remove

Load Additional Results Combine P-Values Plot Reset View Active Filters

Go to Selected Region

You can now sort by any of the columns. Note that all the significant SNPs are on chromosome 12 by most criteria.

More interestingly, Haploview will allow plotting of the association data against genomic position. Select 'Plot', and in the Plot dialogue box, select FDR_BH as the Y-value. Select '-log10' for the scale, give the plot a name, and click OK. You should see the following chromosome:significance plot:

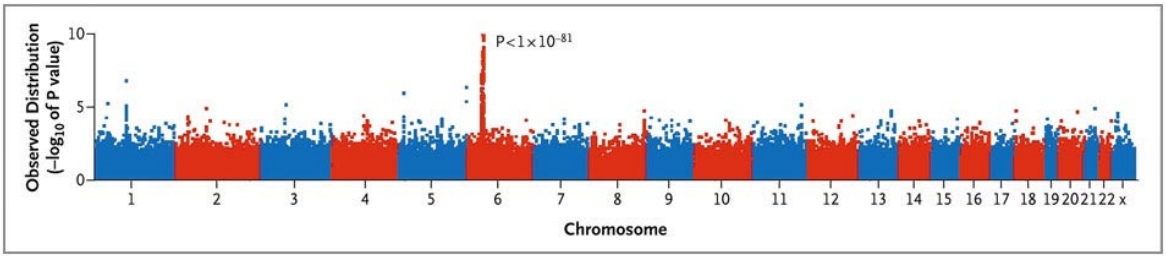


The highest blue dot represents the most significant SNP 'rs11830226' on chromosome 12

If there were a true association, we would expect some of its neighbouring SNPs to also be associated with the phenotype, as in the population we would expect some of those SNP alleles to be coinherited with the rs11830226 allele. This is indeed what we see, with a number of SNPs on chromosome 12 also having significant p-values.

You might now look at the genes in the area of the most significant SNPs to see if there are candidates for explaining the genotype-phenotype relationship, using the chromosomal position information. Note that most of the significant SNPs occur between 114000000 and 116000000 on chromosome 12. So, for instance, you might go to the UCSC human genome browser and look at this area to see what's there.

The data you used in this tutorial was not real, but the analysis you have done is the same as you would for real data, as shown by this example plot from a Multiple Sclerosis GWAS:



Source: International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PI, Gabriel SB, Mirel DB, Ivinson AJ, Pericak-Vance MA, Gregory SG, Rioux JD, McCauley JL, Haines JL, Barcellos LF, Cree B, Oksenberg JR, Hauser SL. ***Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study*** *N Engl J Med. 2007 Aug 30;357(9):851-62*

More advanced analyses

Now you are familiar with simple association analyses, you might try the more advanced PLINK tutorial on which this tutorial was based. This tutorial addresses issues such as confounding factors in the data due to population stratification, and quantitative trait association analysis.

You might also be interested in the Haploview tutorial, although this focuses mostly on haplotype analysis.