**PRACTICAL 1: Quality Control for GWAS**

**Introduction**

The purpose of our practical sessions is to provide guidance on performing a GWAS analysis step-by-step, as well as to introduce you to useful software for undertaking a GWAS. In this first practical, we introduce the quality control steps required when undertaking a GWAS.

**Software and datasets**

A commonly used program for GWAS analysis is called 'PLINK', which is freely available for download. The webpage for PLINK is available at:

<http://pngu.mgh.harvard.edu/~purcell/plink/>

The webpage includes a full description of the commands for running PLINK, as well as input and output file formats etc., and is a very useful point of reference.

PLINK can be installed in various operating systems including Apple Mac, MS-DOC or Linux. In our practicals we will be using the Linux version of PLINK.

For the purpose of running PLINK, input files need to be in a specific format. For undertaking a GWAS, two input files are required: one ‘.ped’ file and one ‘.map’ file. The formats of the two files are as follows:

**The .ped file**

The PED file is a white-space (space or tab) delimited file. It includes one row for each participant within the study, and at least seven columns. The first six columns are mandatory, and are as follows:

|  |  |
| --- | --- |
| Family ID | This column gives a patient identifier for each group of individuals within the same family. In the type of GWAS analyses we are undertaking, we assume all individuals are independent (non-related), so the family ID is unique for each individual and is identical to the ‘Individual ID’ (see below). |
| Individual ID | This column gives a patient identifier for each individual. |
| Paternal ID | In GWAS where we have unrelated individuals (like in our analyses), all entries in this column are set to 0. |
| Maternal ID | In GWAS where we have unrelated individuals (like in our analyses), all entries in this column are set to 0. |
| Sex | Denotes gender of individual (1=male; 2=female; other=unknown) |
| Phenotype | Denotes phenotype (i.e. outcome) for the individual. Phenotype can either be quantitative in which case each individual will have either a numerical value, or ‘-9’ to denote missing (Note: there are commands that can be used to change the code for missing data), or binary in which case each individual will be coded 1=unaffected, 2=affected or -9=missing. |

The remaining columns (column 7 onwards) include genotype data for each of the SNPs genotyped. Each SNP is represented by two columns –one for each of the two alleles held by the individual at the SNP. Typical coding would be A, C, G or T to represent the four possible alleles, or similarly 1, 2, 3 or 4. Coding for missing genotype is ‘0 0’. (Note: for a SNP either none or both of the alleles can be set to missing – it is not possible to mark a SNP as having one allele missing and one not missing).

**The .map file**

The .map file includes information on the SNPs genotyped, and each row represents a SNP. It includes 4 columns, as follows:

|  |  |
| --- | --- |
| Chromosome | The number of the chromosome in which the SNP resides (1-22; 23 (chr X); 24 (chr Y); 25 (pseudo-autosomal region of chr X); 26 (Mitochondrial). |
| rs number | This entry is a number which uniquely identifies the genotyped SNP. An rs number is an accession number used by researchers and databases to refer to specific SNPs. It stands for ‘Reference SNP cluster ID’. |
| Genetic distance | For GWAS analyses like those we are undertaking, this can be set to ‘0’ for all SNPs. |
| Base-pair position | This gives the base-pair position of the SNP along the chromosome, and measurement unit is bp. Values can be any positive integer within the range of typical human chromosome sizes. |

For the purpose of our practical, we will be working on analysing the GWAS dataset ‘genstudy’. The study is a case-control study with 252 diseased (cases) and 252 non-diseased (controls) individuals. All individuals have been genotyped for 198,684 SNPs on chromosome 10 and 22,379 SNPs on chromosome X. The purpose of the study is to try and identify SNPs on chromosome 10 associated with disease. (Note: In a normal GWAS, individuals will typically be genotyped at several hundred thousand SNPs across all chromosomes. Due to time constraints, we have limited our practical to analyse associations with SNPs on chromosome 10 only. Genotypes for SNPs on chromosome X have also, however, been provided for the purpose of running gender checks in the QC process).

A .map and .ped file for ‘genstudy’ (‘genstudy.map’ and ‘genstudy.ped’) have been provided.

* *Type the following command to obtain a list of files within the /tmp/course folder that start with ‘genstudy’:*

**ls -l genstudy\***

You will see a file named ‘genstudy.ped’ and ‘genstudy.map’.

* *Type the following command to view the first row and first seven columns of the ‘genstudy.ped’ file:*

**awk 'FNR == 1 { print $1,$2,$3,$4,$5,$6,$7,$8} ' genstudy.ped**

You will see there are eight columns, corresponding to the first six mandatory columns plus two columns representing the first SNP genotyped. Altogether there are several hundred thousand columns of course, to represent the large number of SNPs genotyped.

* *Type the following command to view the first two rows of the ‘genstudy.map’ file:*

**head -n 2 genstudy.map**

You will see that each row has entries in four columns, corresponding to those described above.

**Checking the input datasets**

Before starting a genotype QC or GWAS association analysis, it is important to check that the input files are in good order, and contain the individuals and SNPs that we think they should. We can do this, focussing on the SNPs on chromosome 10, by typing the following PLINK command:

**plink --file genstudy --chr 10 --noweb**

Note: the ‘-noweb’ option is included to make PLINK run faster – otherwise it connects to the web to check for updates before running the analysis.

* *Run the above PLINK command and read through the output to see how PLINK validates the dataset. Do the number of samples and SNPs agree with what we expected ? Are there any error messages/warning messages ?*

The output is also saved in a log file called ‘plink.log’.

* *Type the command below to take a look at the log file:*

**cat plink.log**

**Binary PED files**

To save space and time, .map and .ped files can be converted to binary format. In doing this, we create three new files as follows:

|  |  |
| --- | --- |
| ‘.bed’ file | This includes all the genotype data from the previous ‘.ped’ file, but in binary format. |
| ‘.fam’ file | This includes the first six columns from the previous ‘.ped’ file. |
| ‘.bim’ file | This is the same as the previous ‘.map’ file, but also includes an additional two columns with the alleles names (which are otherwise lost when converting the genotype data from .ped file to .bed file.) |

* *Create a set of binary format files for the ‘genstudy’ dataset, using the command below:*

**plink --file genstudy --make-bed --out genstudy --noweb**

Using the ‘--out’ option in the command above ensures that the binary files are all called ‘genstudy’, i.e. genstudy.bed, genstudy.fam and genstudy.bim.

If we wish to use the binary format files instead of the .ped and .map files, we just substitute the option --file with --bfile in our PLINK command line. For example, the command:

**plink --file genstudy --chr 10 --noweb**

we previously used to check the files would become:

**plink --bfile genstudy --chr 10 --noweb**

**Sample QC:**

We will now start undertaking genotype QC on the ‘genstudy’ dataset– starting first of all with sample QC, one step at a time.

1. **Checks for missing data**

In PLINK it is possible to obtain a summary of the amount of missing genotype data per sample and per SNP by using the option ‘- -missing’ in the PLINK command line. This option creates two output files:

1. a ‘.imiss’ file which summarises the proportion of missing genotype data per individual
2. a ‘.lmiss’ file which summarises the proportion of missing genotype data per SNP

For example, to obtain this information for the ‘genstudy’ dataset we would use the following command:

**plink --bfile genstudy --missing --out genstudy.CRstats --noweb**

Adding the ‘--out genstudy.CRstats’ option ensures that the two output files are called ‘genstudy.CRstats.imiss’ and ‘genstudy.CRstats.lmiss’ respectively.

The ‘.imiss’ output file includes one row per individual and six columns as follows:

|  |  |
| --- | --- |
| FID | Family ID as per .ped/.fam file. |
| IID | Individual ID as per .ped/.fam file. |
| MISS\_PHENO | Denotes whether the individual has missing phenotype data or not (Y=yes; N=no). |
| N\_MISS | Number of SNPs where genotype is missing for the individual. |
| N\_GENO | Number of SNPs genotyped. |
| F\_MISS | Proportion of missing genotypes for the individual. |

The most important column for sample QC purposes of course is column 6.

The ‘.lmiss’ output file includes one row per SNP, and five columns as follows:

|  |  |
| --- | --- |
| SNP | SNP rs number as per .map/.bim file. |
| CHR | Chromosome number for the SNP as per .map/.bim file. |
| N\_MISS | Number of individuals for which genotype missing at this SNP. |
| N\_GENO | Number of individuals genotyped. |
| F\_MISS | Proportion of missing genotypes for the SNP. |

The most important column for SNP QC purposes of course is column 5.

* *Create the two output files ‘genstudy.CRstats.imiss’ and ‘genstudy.CRstats.lmiss’by running the relevant PLINK command.*
* *Take a look at the first ten rows of your ‘genstudy.CRstats.imiss’ file by typing the command:*

**head -n 10 genstudy.CRstats.imiss**

* *Take a look at the first ten rows of your ‘genstudy.CRstats.lmiss’ file by using a similar command.*

1. **Gender Checks**

PLINK allows us to use genotype data from the X chromosome to determine gender (i.e. based on heterozygosity rates), and compares this to reported gender as per the .ped/.fam file. PLINK then flags any discrepancies i.e. individuals for whom reported gender does not match the gender estimated based on genotype data.

To undertake this comparison, we can use the option ‘- - check-sex’ in PLINK. For example, the following code will undertake the comparison and save results in an output file called ‘genstudy.sexcheck’:

**plink --bfile genstudy --check-sex --out genstudy --noweb**

The output file has six columns, as follows:

|  |  |
| --- | --- |
| FID | Family ID of individual, as per .ped/.fam file. |
| IID | Individual ID, as per .ped/.fam file. |
| PEDSEX | Gender as specified in .ped/.fam file. |
| SNPSEX | Gender as determined by X chromosome genotype data. |
| STATUS | This displays ‘OK’ if no discrepancy and ‘PROBLEM’ if there is a discrepancy between the entry in ‘PEDSEX’ and that in ‘SNPSEX’. |
| F | The actual X-chromosome inbreeding (homozygosity) estimate (Male call made if F>0.8; Female if F<0.2). |

The most important column for the purpose of sample QC is column 5, which indicates whether there is a gender discordance or not.

* *Undertake a gender check for the ‘genstudy’ dataset by using the relevant PLINK command. Take a look at the first 10 rows of the output file by typing the following command:*

**head -n 20 genstudy.sexcheck**

1. **Duplicate/related samples**

PLINK will also calculated identity-by-state (IBS) and identity-by-descent (IBD) statistics for our individuals to help us identify duplicated and/or related samples. Before we ask PLINK to do this, however, we need to generate a pruned subset of SNPs using the option ‘--indep-pairwise’ in PLINK, which is based on pairwise genotypic correlation. An example of a command line for generating a pruned subset of SNPs is as follows:

**plink --bfile genstudy --indep-pairwise 1500 150 0.2 --noweb**

The options 1500, 150 and 0.2 in the command above relate to the following steps, and of course these options can be changed depending on your requirements:

a) consider a window of 50 SNPs,

b) calculate LD between each pair of SNPs in the window, 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.

Running this command line will create two output files, ‘plink.prune.in’ and ‘plink.prune.out’.

plink.prune.in

plink.prune.out

Each is simply a list of SNP IDs – the ‘.in’ file lists all SNPs remaining after pruning; the ‘.out’ file lists all SNPs removed during pruning. Both these files can subsequently be specified as the argument for PLINK’s ‘--extract’ or ‘--exclude’ command respectively to provide a pruned set of SNPs.

* *Create a list of SNPs (‘plink.prune.in’) to include in a pruned version of the genstudy dataset, by using the relevant PLINK command. (Note: running this command will take a while and so you may wish to skip this step during the practical – in which case, a list of SNPs ‘plink.prune.in’ has been provided for you.)*
* *Create a pruned version of the ‘genstudy’ dataset and call it ‘p.genstudy’, by typing the command below*

**plink --bfile genstudy --extract plink.prune.in --make-bed --out p.genstudy --noweb**

We can generate IBS and IBD estimates for all pairs of individuals by using the PLINK option ‘—genome’. The command line for running this is as follows:

**plink --bfile p.genstudy --genome --out genstudy --noweb**

Running this command creates an output file called "genstudy.genome". This includes a row per each unique pair of individuals, and fourteen columns as follows:

|  |  |
| --- | --- |
| FID1 | Family ID for first individual |
| IID1 | Individual ID for first individual |
| FID2 | Family ID for second individual |
| IID2 | Individual ID for second individual |
| RT | Relationship type between the two individuals, given the .ped/.fam file |
| EZ | Expected extent of IBD sharing given .ped/.fam file |
| Z0 | Probability (IBD=0) |
| Z1 | Probability (IBD=1) |
| Z2 | Probability (IBD=2) |
| PI\_HAT | Proportion IBD |
| PHE | Pairwise phenotypic code (1,0,-1=case/case; case/control; control/control |
| DST | IBS distance |
| PPC | IBS binomial test |
| RATIO | Ratio of HETHET : IBS 0 SNPs (expected value is 2) |

The most important column for the purpose of sample QC is column 10 which includes an estimate of the IBD proportion between individuals.

* *Type the command above to run the ‘--genome’ option on the ‘genstudy’ dataset. View the first five rows of the output file by typing the following command:*

**head -n 5 genstudy.genome**

1. **Heterozygosity**

To run heterozygosity checks in PLINK, we can use the option ‘- -het’. The command line for undertaking heterozygosity checks is as follows:

**plink --bfile genstudy --het --out genstudy**

Running this command produces the output file "genstudy.het", which contains one row per individual and six columns, as follows:

|  |  |
| --- | --- |
| FID | Family ID as per .ped/.fam file |
| IID | Individual ID as per .ped/.fam file |
| O(HOM) | Observed number of homozygote genotypes |
| E(HOM) | Expected number of homozygote genotypes |
| N(NM) | Number of non-missing genotypes |
| F | Inbreeding coefficient estimate, F |

The most important column for the purpose of sample QC is column 6 which includes the estimate of the inbreeding coefficient, F.

* *Run a heterozygosity check on the ‘genstudy’ dataset by running the PLINK command above.*
* *Take a look at the first five rows of the output file by typing the following command:*

**head -5 genstudy.het**

**Producing QC plots**

Once we have run the sample QC steps above, we can produce some QC plots in R. We have already written R code to do this, and this code can be run by typing the commands below. (Note: you will need to type in R on your command line first to take you to the R environment. Type q() to escape from that environment after you have finished running the plot commands.

**CR=read.table("genstudy.CRstats.imiss",header=T)**

**pdf("genstudy.CR.pdf")**

**hist(CR$F\_MISS,xlab="Proportion missing per individual",breaks=50)**

**dev.off()**

**HET=read.table("genstudy.het",header=T, as.is=T)**

**H=((HET$N.NM.-HET$O.HOM.)/HET$N.NM.)**

**pdf("genstudy.Het\_vs\_imiss.pdf")**

**dat = data.frame(x=H, y=CR$F\_MISS)**

**plot(dat,log="x",xlim=c(0.001,0.99),ylab="Fraction of missing genotypes",xlab="Mean Heterozygosity")**

**abline(h=0.05,col=2,lty=2)**

**dev.off()**

**PI=read.table("genstudy.genome",header=T)**

**pdf("genstudy.Homo.pdf")**

**hist(PI$PI\_HAT,50)**

**dev.off()**

**CR=read.table("genstudy.CRstats.lmiss",header=T)**

**pdf("genstudy.snp\_missing.pdf")**

**hist(CR$F\_MISS,xlab="Proportion missing per SNP",breaks=50,xlim=c(0,0.2))**

**dev.off()**

**Removing individuals who fail sample QC**

The final steps in our sample QC process are those required to remove individuals who fail any of the QC steps above. These steps are as follows:

1. **Removing due to missingness**

Before removing individuals due to extent of missing genotype data, we first of all need to decide which missingness threshold to apply. There are two approaches we can use to choose a threshold. The first is to review the plot ‘genstudy.CR.pdf’ which provides the distribution of missing genotype data per individual. The second is to investigate how many individuals have missing data above certain thresholds. To use the second approach, we can run the following code (Note that this code is effectively filtering your previously created ‘genstudy.CRstats.imiss ‘ file with reference to a given threshold (in this case 5%).

**awk 'NR>1 && $6 >= 0.05 {print}' genstudy.CRstats.imiss | wc -l**

* *Run the above code to investigate how many individuals are excluded at difference threshold levels (e.g. 1%, 5%, and 10%)*
* *For the purpose of this practical we will assume that we are going to exclude all individuals with more than 5% missing data. Identify these individuals and write their identifiers to the output file ‘genstudy.CRremove’ by using the following code:*

**awk '$6 >= 0.05 {print}' genstudy.CRstats.imiss > genstudy.CRremove**

* *View the list of individuals to be removed due to missing by using the following code:*

**cat genstudy.CRremove**

* *Remove these individuals from the ‘genstudy’ dataset, and create a new dataset, ‘genstudy.CR’ by typing the command below. How many samples have been removed ?*

**plink --bfile genstudy --noweb --remove genstudy.CRremove --make-bed --out genstudy.CR**

1. **Removing due to gender discordance**

Similarly, we can create a list of individuals to remove due to gender discordance by filtering the previously created ‘genstuy.sex.sexcheck’ file to select only those with an entry of ‘PROBLEM’ in column 5. To do this, we can use the following code:

**awk '$5=="PROBLEM"{print}' genstudy.sexcheck>genstudy.sexremove**

* *Type the command above to create a file named ‘genstudy.sexremove’ to include a list of individuals with gender discordance.*
* *View the list of individuals to be removed due to gender discordance by using the following code:*

**cat genstudy.sexremove**

* *Use the command below to remove these individuals from the ‘genstudy.CR’ dataset, and create a new dataset named ‘genstudy.sex’.*

**plink --bfile genstudy.CR --noweb --remove genstudy.sexremove --make-bed --out genstudy.sex**

* *How many individuals are removed at this stage? Is this different to the number of individuals listed in the file ‘genstudy.sexremove’? If so, why do you think this is?*

1. **Remove duplicates and related samples**

We use an IBD threshold of 0.1875 to filter out either duplicated or closely related samples. We can create a list of individuals who appear to be duplicated samples or related by filtering column 10 of our previously created file, ‘genstudy.PI.genome’ with reference to the value 0.1875.To do this, we can use the following code:

**awk '$10 >= 0.1875 {print}' genstudy.genome > genstudy.PIremove**

* *View this list of individuals by using the command:*

**cat genstudy.PIremove**

* *Create a list of duplicated/closely related samples (remove one from each pair). For related pairs or family groups, the usual QC step is to leave one individual in the dataset and drop the other or others, based for example on the one with the least missingness. We can manually create the list of samples to be removed:*

**echo "FID IID**

**idR\_250 idR\_250**

**idR\_251 idR\_251**

**idD\_252 idD\_252**

**idD\_253 idD\_253" > genstudy.PIremove.list**

* *Remove this list of individuals from the ‘genstudy.sex’ dataset to create a new dataset ‘genstudy\_sampleqc’, by using the command below :*

**plink --bfile genstudy.sex --noweb --remove genstudy.PIremove.list --make-bed --out genstudy\_sampleqc**

1. **Heterozygosity**

Look at the "genstudy.Het\_vs\_imiss.pdf" and see if outliers exist. If not, there are no samples to remove. If there were outliers, we would need to exclude these individuals from the ‘genstudy\_sampleqc’ file to create a final file.

After following all the procedures above you now you have a Sample QC'd dataset in binary fomat, which includes the following files:

genstudy\_sampleqc.bed

genstudy\_sampleqc.bim

genstudy\_sampleqc.fam

**SNP QC**

We will now undertake SNP QC, one step at a time.

1. **Checks for missing data**

Similar to what we did for sample QC, we can investigate how many SNPs fail different missingness thresholds by filtering column 5 in ‘genstudy\_sampleqc.CRstats.lmiss’ using different threshold values. We can do this using the following command (e.g. for threshold of 1%):

**plink --bfile genstudy\_sampleqc --missing --out genstudy\_sampleqc.CRstats --noweb**

**awk 'NR>1 && $5 >= 0.01 {print}' genstudy\_sampleqc.CRstats.lmiss | wc -l**

* *Investigate how many SNPs fail various missingness thresholds (e.g. 1%, 5%, and 10%).*

1. **Checks for minor allele frequency (MAF)**

Similarly, we can investigate how many SNPs have MAF<1% by first of all running the ‘--freq’ option in PLINK, using the following command:

**plink --bfile genstudy\_sampleqc --noweb --freq --out genstudy\_sampleqc**

This provides an output file ‘genstudy\_sampleqc.frq’, and the minor allele frequency is in column 5 of the file.

* *Create a ‘genstudy\_sampleqc.frq’ file using the command above, and use the following command to see how many SNPs have MAF<1%. Do the same for a 5% and 10% threshold.*

**awk '$5<0.01' genstudy\_sampleqc.frq | wc -l**

1. **Checks for adherence to Hardy-Weinberg Equilibrium**

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

**plink --bfile genstudy\_sampleqc --noweb --hardy --out genstudy\_sampleqc**

This provides and output file ‘genstudy.hwe’ and the p-value for the test for Hardy-Weindberg Equilibrium is in column 9 of this file.

* *Create a ‘genstudy.hwe’ file using the command above. Use the following command to see how many SNPs have HWE p-value<0.0001 (in controls "UNAFF"). Do the same for a 0.00001 threshold.*

**awk '$3=="UNAFF" && $9<0.0001' genstudy\_sampleqc.hwe | wc -l**

Let’s now use PLINK to apply some SNP QC filters to our already sample-QC’d dataset, *‘genstudy\_sampleqc’*. This is done by applying some simple filters for illustration purposes (in practice one would look at the data much more carefully).

In applying our filtering, we will select only the SNPs for which

* Minor allele frequency (MAF) >0.05 – option “--maf 0.05”
* SNP genotyping rate >95% (i.e. <5% missing genotypes for the SNP, across all individuals) – option “-- geno 0.05”
* Hardy-Weinberg p-value >0.0001in controls – option “-- hwe 0.0001”

To apply this filtering, we can use the following PLINK command:

**plink --bfile genstudy\_sampleqc --noweb --maf 0.05 --geno 0.05 --hwe 0.0001 --make-bed --out genstudy\_qc**

where a fully QC'd fileset is created:

* *Run the command above to filter out SNPs failing our SNP QC criteria.*
* *How many individuals and SNPs are in this final QC’d fileset?*