**Module 10 Assignment**

We are using PLINK to analyze sample data, which consists of randomly chosen genotypes (roughly 80,000 autosomal SNPs) from the 89 Asian HapMap subjects. The genotype at one SNP has been used to simulate a phenotype.

##### plink --file hapmap1

The root name of the input files is the only input required by the PLINK software's "--file" option. By default, PLINK expects the root name to correspond to a .ped file containing genotype information for each person and a .map file containing information about the genetic markers used in the study. The current working directory should contain these files. The "--ped" and "--map" options, however, allow you to specify the folders and file names.

Both the PED and MAP files are plain text files; the PED file contains genotype information (one person per row), while the MAP file contains information on the name and location of the markers in the PED file.

**Summarizing the information in the output:**

* The plink.log log file will be saved after the PLINK program has run on the hapmap1 dataset, but you can change it with the --out option, according to the message that appears. The log files should be organized, and each analysis should have its own --out name.
* Then, PLINK read all 89 individuals and just over 80,000 SNPs from the PED file. The phenotype was an affection status variable, and missing values were identified.
* PLINK then removed people and/or SNPs from the data if they did not meet predetermined criteria. In this instance, no individuals were eliminated, but nearly 20,000 SNPs were eliminated due to missingness and frequency. Due to the fact that the Chinese and Japanese samples contained far more rare and monomorphic markers than a typical whole-genome association product, this high percentage of removed SNPs was caused by random HapMap SNPs in these samples.

**Making a binary PED file**

using the following command, create a binary PED file. This data representation is more condensed, which saves storage space and expedites subsequent analysis.

##### plink --file hapmap1 --make-bed --out hapmap1

* The missing rate and allele frequency filters are set to exclude no one when using the --make-bed option by default. However, you can manually specify filters using "--mind", "--geno", and "--maf" options.
* The "--extract"/"--exclude" and "--keep"/"--remove" options are also available during this procedure.
* The "hapmap 1.bed" file, which contains the raw genotype data, "hapmap1.bim," which has two additional columns with the names of each SNP, and "hapmap1.fam," which is a subset of "hapmap1.ped," are all created as a result of selecting this option. .bim and.fam files can be viewed, but.bed files cannot. There should be no manual editing done to any of these files.

For example, you would run the following code if you wanted to create a new file that only contained people with high genotyping (at least 95% complete):

##### plink --file hapmap1 --make-bed --mind 0.05 --out highgeno

which would create the files highgeno.bed, highgeno.bim, and highgeno.fam.

**Working with the binary PED file**

Use the "--bfile" option in place of the "--file" option to specify that the input data is in binary format rather than the default text PED/MAP format. This is the same as repeating the initial command we executed.

##### plink --bfile hapmap1

* The three files hapmap1.bim, hapmap1.fam, and hapmap1.bed were loaded as opposed to the typical two files. As a result, hapmap1.ped and hapmap1.map are not used in this analysis and may be deleted at this time.
* Based on the timestamps at the start and end of the log output, the data is loaded in much more quickly; it took 2 seconds instead of 10 to do so.

**Summary statistics: missing rates**

The --missing option will then be used to create some basic summary statistics on the percentage of missing data in the file:

##### plink --bfile hapmap1 --missing --out miss\_stat

The missingness rate per individual and per SNP are shown, respectively, in the two files produced by the code, miss\_stat.imiss and miss\_stat.lmiss. No individuals were eliminated because of low genotypes, so they accepted individuals with a missingness rate of less than 10%. If no other option was specified, the default output filename would have been "plink". Both files are plain text and can be viewed in any application that can handle large files, including text editors, pagers, spreadsheets, and statistics packages. You can check the missingness rate for each SNP in the miss\_stat.lmiss file.

##### more miss\_stat.lmiss

The number of individuals who are missing (N\_MISS) and the percentage of those who are missing (F\_MISS) are displayed for each SNP.

##### more miss\_stat.imiss

We can see that the genotyping rate is extremely high in this last column, which represents the actual genotyping rate for that person.

It may be advantageous to ask PLINK to analyze the data by chromosome using the --chr option if you are using a spreadsheet program that can only display a finite number of rows (some well-known programs can handle just over 65,000 rows). For chromosome 1, for instance, the following would be the analysis's results:

##### plink --bfile hapmap1 --chr 1 --out res1 --missing

then, for chromosome 2:

##### plink --bfile hapmap1 --chr 2 --out res2 --missing

and so on.

**Summary statistics: allele frequencies**

The following command generates a file called freq\_stat.frq that contains the minor allele frequency and allele codes for each SNP.

##### plink --bfile hapmap1 --freq --out freq\_stat

This frequency analysis can also be stratified by a categorical, cluster variable, just like the missingness analysis. In this instance, we will make use of the pop.phe file, which identifies whether the person comes from the Chinese or the Japanese sample. The three columns in this cluster file each have a row representing a person.

To perform a stratified analysis, use the --within option.

##### plink --bfile hapmap1 --freq --within pop.phe --out freq\_stat

Now, instead of producing freq\_stat.frq, the output will say that freq\_stat.frq.strat was created. If we view this file:

##### more freq\_stat.frq.strat

When stratified by subpopulation, the allele frequency for each SNP is now shown in each row of the output.

Each SNP is represented twice, as can be seen. The CLST column, coded according to the pop.phe file, indicates whether the frequency is from the Chinese or Japanese populations.

Use the --snp option to select this SNP if you were only interested in learning how frequently it occurred in the two populations:

##### plink --bfile hapmap1 --snp rs1891905 --freq --within pop.phe --out snp1\_frq\_stat

would create a file called snp1\_frq\_stat.frq.strat containing just the frequencies that apply to a particular population for this one SNP. The --window kb option, as well as the --from and --to options, each with a different SNP after them, can be used to specify a range of SNPs.

**Basic association analysis**

The fundamental directive is to run a simple association analysis on the disease trait for each individual SNP.

##### plink --bfile hapmap1 --assoc --out as1

which generates an output file as1.assoc which contains the following fields

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

* Chromosome
* SNP identifier
* Code for allele 1 (the minor, rare allele based on the entire sample frequencies)
* The frequency of this variant in cases
* The frequency of this variant in controls
* Code for the other allele
* The chi-squared statistic for this test (1 df)
* The asymptotic significance value for this test
* The odds ratio for this test

To sort the list of association statistics and print the top ten in a Unix/Linux environment, one need only use the available command line tools, as in the following example:

##### sort --key=7 -nr as1.assoc | head

The simulated disease variant rs2222162, which was discovered to be the second most significant SNP on the list, is provided in the output, indicating a significant difference in allele frequencies between cases and controls. With comparable allele frequencies in cases and controls, another SNP on chromosome 13 had slightly better test outcomes. It is unclear whether this outcome is the result of chance or confounding brought on by the sample's population structure. This circumstance demonstrates how it can be challenging to differentiate between true positive results and the best false positive results when conducting numerous tests, especially in a small sample. Nevertheless, it is comforting to know that the simulated disease variant was among the top ten SNPs.

Use the --adjust flag to obtain a sorted list of association findings along with a range of significance levels that are corrected for multiple testing:

##### plink --bfile hapmap1 --assoc --adjust --out as2

In addition to producing the standard as2.assoc output file, this also produces as2.assoc.adjust. Using more, it is simple to examine your strongest connections:

##### more as2.assoc.adjusted

An ordered list of association results appears in the output. The fields are as follows:

* Chromosome
* SNP identifier
* Unadjusted, asymptotic significance value
* Genomic control adjusted significance value. This is based on a simple estimation of the inflation factor based on median chi-square statistics. These values do not control for multiple testing therefore.
* Bonferroni adjusted significance value
* Holm step-down adjusted significance value
* Sidak single-step adjusted significance value
* Sidak step-down adjusted significance value
* Benjamini & Hochberg (1995) step-up FDR control
* Benjamini & Yekutieli (2001) step-up FDR control

After genome-wide correction, we can see that in this specific instance, no single variant is significant at the 0.05 level. The choice of which correction measure to use and how to interpret it is up to the investigator because different correction measures have different properties that are outside the purview of this tutorial.

When the --adjust command is used, the log file logs the median chi-squared statistic (which under the null assumption should be 1) as well as the inflation factor calculated for the genomic control analysis:

Genomic inflation factor (based on median chi-squared) is 1.18739

Mean chi-squared statistic is 1.14813

As both values are greater than 1.00, it would seem that although there isn't a very strong stratification, there may be a hint of a higher false positive rate.

The unadjusted significance values are the default foundation for the adjusted significance values, which account for multiple testing. These adjusted values will instead be based on the genomic-control significance value if the flag --gc is specified along with --adjust.

In this specific case, where we are already familiar with the Chinese/Japanese subpopulations, it might be interesting to look directly at the inflation factor that results from using population membership as the phenotype in a case/control analysis to learn more about the sample. Specifically, when the command is executed with the alternate phenotype option (that is, when the disease phenotype is substituted with the subpopulation membership phenotype found in pop.phe):

##### plink --bfile hapmap1 --pheno pop.phe --assoc --adjust --out as3

When comparing the frequency of Chinese and Japanese people, we observe that there is some departure from the null distribution:

Genomic inflation factor (based on median chi-squared) is 1.72519

Mean chi-squared statistic is 1.58537

In other words, the inflation factor of 1.7 represents the highest inflation factor that could result from a perfect correlation between the disease and the subpopulation that could result from the Chinese/Japanese split in the sample (this, of course, does not take into account any potential within-subpopulation structure that might also increase SNP-disease false positive rates).

**Genotypic and other association models**

The 2-by-3 genotype table, the common allelic test, and tests assuming dominant or recessive action of the minor allele can all be used in conjunction with the "--model" command to compute association statistics. It also permits the Cochran-Armitage trend test to be run instead of the fundamental allelic test. All SNPs can be applied to with ease using the "--model" command, just like with "--assoc." However, in this instance, the command will only be carried out for our target SNP, rs2222162.m.

##### plink --bfile hapmap1 --model --snp rs2222162 --out mod1

When the "--model" command is run, a file with the name "mod1.model" is created. This file has more than one row for each SNP, signifying the various tests that were run on each SNP. The documentation gives a description of the file format. The following tests are run: the basic allelic test, the Cochran-Armitage trend test, the dominant and recessive models, and a genotypic test. All test statistics, with the exception of the genotypic test, which has two degrees of freedom, have a chi-squared distribution with one degree of freedom under the null.

Running the basic model command will not, however, yield results for the genotypic tests in this particular instance because, by default, each cell in the 2-by-3 table must have at least 5 observations, which is not the case in this instance. By specifying the bare minimum number of counts in each cell of the 2-by-3 table before the extended analyses are run, the "--cell" option allows this default to be changed. For illustration, we need to execute the following command in order to conduct genotypic tests for this SNP:

##### plink --bfile hapmap1 --model --cell 0 --snp rs2222162 --out mod2

The genotypic tests will also be computed following the setting of the "--cell" option to 0 to remove the minimum number of counts from each cell. In the output, you can see the genotype counts for both affected and unaffected people. These values are comforting because they coincide with the values produced in the table when the trait was simulated.

All other test statistics are highly significant, as would be predicted given the strength and prevalence of the allelic effect. Though it makes sense given that the data were essentially simulated under an allelic (dosage) model, the allelic test has the most significant p-value.

**Stratification analysis**

The fact that the sample consists of two subpopulations with different disease prevalences—the Chinese and Japanese samples—did not factor into the earlier analyses. One solution to this problem is to group individuals into homogeneous groups using whole genome data. PLINK provides a number of options for carrying out this kind of analysis, but for the sake of illustration, we'll concentrate on a cluster analysis that pairs individuals based on genetic identity. Running this command could take several minutes:

##### plink --bfile hapmap1 --cluster --mc 2 --ppc 0.05 --out str1

The command performs IBS clustering on the dataset with the restrictions that each cluster can only contain two individuals (--mc 2) and that any pairs of individuals with a significance value less than 0.05 for the test to determine whether they belong to the same population are not merged (--cluster). The available SNP data are the foundation for the significance test. This command's objective is to group people according to their genetic similarity into homogeneous groups. The command may take a few minutes to run.

The clustering results are presented in the following file, str1.cluster1, in an accessible format:

##### more str1.cluster1

* Using the PLINK command and the flags --cluster, --mc 2, and --cc, a cluster analysis was carried out to pair up individuals based on genetic identity.
* No more than two people were allowed in each cluster, according to the restriction.
* On the basis of the available SNP data, a threshold was set at a significance value of less than 0.05 to determine whether two individuals belonged to the same population.
* For either being Chinese or Japanese, all but one pair of individuals were concordant, and the lone unpaired individual would not contribute to any further association testing that relied on this cluster solution.
* Instead of JPT257\_1, the Japanese person JPT260\_1 was paired with a Chinese person HCB181\_1, suggesting that the clustering algorithm had limited ability to distinguish people from subpopulations that were very similar to each other or that one of these people might be of mixed ancestry.
* With extremely long stretches of homozygous genotypes and a high inbreeding coefficient, JPT257\_1 was discovered to be somewhat unusual, which likely explains why the clustering algorithm did not consider this individual to be similar to the other Japanese individuals.

**Association analysis, accounting for clusters**

Since the individuals have been matched based on their genetic similarity, the association test can now be run while taking the matching into consideration. To accomplish this, we will use the str1.cluster2 file, which has the exact same data as str1.cluster1 but is in a different format that can be used with the --within option. We will employ the Cochran-Mantel-Haenszel (CMH) association statistic, which examines the relationship between SNPs and diseases while taking the clusters from the cluster file into consideration. The --adjust option will also be used to get a sorted list of the CMH association results.

##### plink --bfile hapmap1 --mh --within str1.cluster2 --adjust --out aac1

We can see that PLINK correctly divided the people into 45 clusters (one of these clusters has a size of 1, the others are pairs), and it then ran the CMH test. Since the distribution of test statistics in the previous analysis was inflated by some substructure, the genomic control inflation factors have now been reduced to almost 1.00.

adjusted results file:

##### more aac1.cmh.adjusted

The "disease" variant, rs2222162, has risen from position two to position one in this list, but even after genome-wide correction, it is still not significant.

In the last case, we asked PLINK to pair the individuals who were most similar. Although there may be fewer or different restrictions on the ultimate solution, we are still capable of performing the clustering. Here, for instance, we do not impose a maximum cluster size; instead, we use the --cc option to request that each cluster contain at least 1 case and 1 control (i.e., to make it informative for association) and --ppc to specify a threshold of 0.01:

##### plink --bfile hapmap1 --cluster --cc --ppc 0.01 --out version 2

The cluster file str1.cluster2 contains the same information as str1.cluster1 but in a different format. Each cluster in str1.cluster2 is on its own line, and each person's ID is followed by their phenotype in parentheses. Individuals from China and Japan are largely divided into different clusters by the clusters. Based on the --ppc constraint, the clustering produces a five-class solution. As soon as any of these five clusters were merged, two people who differ at the 0.01 level would have been merged, which would have caused the clustering to stop. We can use this new clustering scheme to repeat our association analysis. However, actual analysis of real data should not be done by experimenting with various analyses, clusters, etc., until one finds the most significant result. This is merely intended to demonstrate the options available.

##### plink --bfile hapmap1 --mh --within version2.cluster2 --adjust --out aac2

Specifying the number of clusters desired in the final result is a third method for carrying out the stratification analysis. Here, we'll use the --K option to specify two clusters and the omission of the --ppc option will remove the significance test restriction:

##### plink --bfile hapmap1 --cluster --K 2 --out version3

Finally, since the true ancestry of each member of this sample is known, we can always use the following external clustering in the analysis:

##### plink --bfile hapmap1 --mh --within pop.phe --adjust --out aac3

Unsurprisingly, this gives very similar results to the two-class solution derived from cluster analysis.

**In summary,**

* Simple IBS-based clustering can successfully distinguish between various subpopulations, like Chinese and Japanese people in this case.
* By reducing false positive rates and boosting power, population substructure can help association analysis be more precise.
* The sample's characteristics, such as its composition—whether it contains a large number of distinct subpopulations or a complex mixture of people with various ancestries—determine the best clustering strategy to use.
* A pairwise IBS distance matrix and a multidimensional scaling plot can be produced using statistical software like R to visualize the substructure in a sample.

##### plink --bfile hapmap1 --cluster --matrix --out ibd\_view

creating the file ibd\_view.mdist in the process. Afterward, run the following commands in R:

##### m <- as.matrix(read.table("ibd\_view.mdist"))

##### mds <- cmdscale(as.dist(1-m))

##### k <- c( rep("green",45) , rep("blue",44) )

##### plot(mds,pch=20,col=k)

A plot is produced by the output. This plot certainly seems to suggest that at least two quite distinct clusters exist in the sample. One would be better able to choose which method of stratification to use moving forward after viewing this kind of plot.

**Quantitative trait association analysis**

Direct analysis of the quantitative trait, which was based on a straightforward median split, will be done in this section. The majority of the analysis options are still available, with the exception that the --mh approach does not work with quantitative traits. Instead, the --assoc flag is employed, which instantly recognizes whether the phenotype is a case/control status or a quantitative trait and executes the proper analysis, in this case ordinary least squares regression. Instead of using the default phenotype file, we need to specify the quantitative trait file (qt.phe) in PLINK.

##### plink --bfile hapmap1 --assoc --pheno qt.phe --out quant1

The quant1.qassoc file is produced by this analysis. The fields in this file represent:

* Chromosome
* SNP identifier
* Number of non-missing individuals for this analysis
* Regression coefficient
* Standard error of the coefficient
* The regression r-squared (multiple correlation coefficient)
* t-statistic for regression of phenotype on allele count
* Asymptotic significance value for coefficient

A file called quant1.qassoc.adjust would be created if we added the --adjust option:

After genome-wide correction, we can see that the disease variant is still important in this instance. However, unlike our previous tests, these ones do not take into account the sample's clustering. Currently, the estimated genomic control inflation factor is:

Genomic inflation factor (based on median chi-squared) is 1.19824

Mean chi-squared statistic is 1.21478

One strategy is to use permutation instead of stratifying the data or including covariates: In particular, only within clusters can phenotypes be permuted, or label-swapped, between individuals. As between-cluster associations are constant across all permuted datasets, this accounts for any such associations. We request clustered permutations as follows, using the original pairing approach to matching:

##### plink --bfile hapmap1 --assoc --pheno qt.phe --perm --within str1.cluster2 --out quant2

Depending on the speed of the computer, it might take an hour or more to analyze the quantitative trait using the adaptive permutation procedure. With a default maximum of 1 million permutations and 25 SNPs still in the analysis, the number of permutations completed and the number of SNPs still in the analysis will be shown. The outcomes will be kept in a file called quant2.qassoc.perm that can be sorted using the empirical p-value (EMP1, the fourth column). The top of the list shows that the disease variant rs2222162 has an empirical significance value of 1e-6, indicating that no permuted datasets had a statistic for this variant that was higher than this value.

* The --within option and PLINK's permutation method can be used to adjust for the clustering variable.
* The standard chi-squared statistics (STAT) will not reflect the within-cluster analysis when the --within option is used, but the empirical significance values (EMP1) will show that we have controlled for the clustering variable.
* The NP field displays the number of permutations that were performed for each SNP. If some SNPs are obviously not highly significant, PLINK may give up on them after only a few permutations.

Using the --mperm option will control for multiple testing.

##### plink --bfile hapmap1 --assoc --pheno qt.phe --mperm 1000 --within str1.cluster2 --out quant3

The population membership specified as a covariate with the --covar option can be used in conjunction with the --gxe option to determine whether the association with the continuous phenotype varies between the two populations. Instead of analyzing every SNP, only the primary SNP of interest will be the subject of the analysis. The --gxe option will examine the relationship between the genotype and the covariate, in this case the population membership. Both the interaction and main effect p-values for the SNP are displayed in the output that follows. This analysis will assist in determining whether the associations between the SNP and the quantitative trait are different between the two populations.

##### plink --bfile hapmap1 --pheno qt.phe --gxe --covar pop.phe --snp rs2222162 --out quant3

The output will show the creation of a file called quant3.qassoc.gxe that contains information about the number of non-missing people in each category as well as the regression coefficient and standard error. This information is followed by tests to determine whether these two regression coefficients are significantly different (Z\_GXE) and an asymptotic significance value (P\_GXE). Since both populations exhibit a similar effect in this instance (regression coefficients around -2), the test for SNP x population interaction is not significant.

**Extracting a SNP of interest**

Once you've located a SNP, a group of SNPs, or a region of interest, you may want to extract those SNPs into a different, more manageable file that is smaller and less complex. You specifically need to convert from the binary PED file format to a standard PED format so that other applications can analyze the data. Utilizing --recode options allows for this. There are different variations of this option; we'll use --recodeAD, which codes the genotypes in a way that makes them easy to analyze in R or any other statistical software that isn't genetic. To extract only this single SNP, use:

##### plink --bfile hapmap1 --snp rs2222162 --recodeAD --out rec\_snp1

(Instead, use the --to and --from options to choose a region, or --window 100 with --snp to pick a 100 kb region around that SNP, for instance.) This specific recording feature stores genotypes in a file called rec\_snp1.recode.raw as additive (0, 1, 2) and dominance (0, 1, 0) components. In order to quickly perform additional analyses, we can import this file into our statistics program. For example, to replicate the primary analysis as a simple logistic regression using the R package (without taking into account clusters):

##### d <- read.table("rec\_snp1.recode.raw" , header=T)

##### summary(glm(PHENOTYPE-1 ~ rs2222162\_A, data=d, family="binomial"))

which supports the analysis done initially. It goes without saying that tasks like survival analysis or the use of other models not included in PLINK can now be carried out.