MA5106 – Practical Report 2
GWAS Analysis

Tyler Medina
ID 13101308
7 November, 2017

Introduction

In this practical, a GWAS analysis will be run on a sample of 2000 males and 2000 females, for which 306,102 SNPs were sequenced. The PLINK toolset will be used for both QC and analysis of the data, followed by visualization with Haploview.

Setup

The data used for this practical exists as a group of four files with the name “gwas”, followed by a file extension:

```
[nextgen2015@node032 originals]$ ls
gwas.bed  gwas.bim  gwas.covar  gwas.fam
```

The files are formatted as binary files as opposed to plain text files, which facilitates faster processing and smaller file sizes. These files include:

- Raw genotype data for each snp in *gwas.bed*
- SNP mapping in *gwas.bim*
- Individual pedigree information in *gwas.fam*
- Covariate information in *gwas.covar*

To begin analysing this data, PLINK is loaded as a module on the High-Powered Computing Cluster:

```
[nextgen2015@node002 low_hwe]$ module load plink/1.07
```

This will allow PLINK to be used throughout the session from any folder. In addition to logging information about each individual call in a “.log” file, PLINK uses various options to produce different file types containing information and analyses of the data.

In all commands with PLINK, the options “--noweb” and --”bfile” will be used:

- --noweb: Instructs PLINK to skip the initial web check, for expediency
- --bfile: Followed by a file name without the file extension, this instructs PLINK to use the specified group of binary files. For example, “--bfile gwas” would be used to instruct PLINK to use gwas.fam, gwas.bed, and gwas.bim.

Summary Statistics

The initial step in this GWAS analysis is to run summary statistics on the raw data for missingness, allele frequency, and Hardy-Weinberg Equilibrium testing.

Missingness:

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile gwas --missing --out missing_gwas
```

Here, the gwas dataset is checked for missing information with "--missing", which produces two separate files with the extensions ".lmiss" for missing loci information, and ".imiss" for missing individual information. SNPs that have not been well-genotyped across individuals in the study and individuals with high amounts of missing SNP information should be excluded from analysis, as this data can lead to results with low reproducibility.

The following excerpt of the output from the "missing_gwas.lmiss" file shows:

- CHR: chromosome number of an SNP
- SNP: the SNP name
- N_MISS: the number of individuals that are missing genotype information for the SNP
- N_GENO: the total number of individuals
- F_MISS: the frequency at which the SNP is missing.

High missingness scores indicate SNPs that should be removed from the study.

CHR	SNP	N_MISS	N_GENO	F_MISS
1	rs3934834	182	4000	0.0455
1	rs3737728	18	4000	0.0045
1	rs6687776	85	4000	0.02125
1	rs9651273	57	4000	0.01425
1	rs4970405	17	4000	0.00425
1	rs12726255	10	4000	0.0025

The output from the "missing_gwas.imiss" file shows:

- FID: family ID of an individual
- IID: the individual's ID
- MISS_PHENO: an indication of whether the individual's phenotype is missing
- N_MISS: the number of missing SNPs in the individual's genotyping
- N_GENO: the total number of SNPs in the study
- F_MISS: the frequency of how many SNP's the individual is missing

High F_MISS values indicate individuals who are missing more information about their SNPs, and should then be removed.

FID	IID	MISS_PHENO	N_MISS	N_GENO	F_MISS
0	A2001	N	5168	306102	0.01688
1	A2002	N	5175	306102	0.01691
2	A2003	N	5150	306102	0.01682
3	A2004	N	5076	306102	0.01658
4	A2005	N	5137	306102	0.01678
5	A2006	N	5236	306102	0.01711

From the --missing log file, the raw data shows 98.3323% genotyping, indicating high quality sequencing data.

```
Total genotyping rate in remaining individuals is 0.983323
```

Allele Frequency:

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile gwas --freq --out freq_gwas
```

The allele frequency at an SNP in the study can be tabulated using the "--freq" option, generating a ".frq" file that displays the following:

- CHR: Chromosome number of an SNP
- SNP: SNP name
- A1: Minor allele code or nucleotide
- A2: Major allele code or nucleotide
- MAF: Minor allele frequency
- NCHROBS: The number of times the SNPs is genotyped in the study (non-missing)

CHR	SNP	A1	A2	MAF	NCHROBS
1	rs3934834	T	C	0.09966	7636
1	rs3737728	A	G	0.3386	7964
1	rs6687776	T	C	0.06117	7830
1	rs9651273	A	G	0.4065	7886
1	rs4970405	G	A	0.03038	7966
1	rs12726255	G	A	0.03421	7980

For GWAS studies, the large multiple-testing correction factors involved severely reduce statistical power for rare alleles. Because of this, SNPs that show a very low MAF may be excluded from the study to increase power.

Hardy-Weinberg Equilibrium

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile gwas --hardy --out hardy_gwas
```

The "--hardy" option checks each allele for Hardy-Weinberg Equilibrium. Alleles found to be significantly out of equilibrium may indicate problems with sampling, genotyping errors, or population stratification. However, SNPs that are found to be out of equilibrium are often found to be truly associated as well. For this reason, care must be taken when deciding whether to exclude SNPs that fall significantly outside of Hardy-Weinberg Equilibrium.

The following output of the resultant ".hwe" file shows:

- CHR: Chromosome number
- SNP: SNP name
- TEST: Tested group, indicating whether the allele was compared against all individuals, case only, or control only
- A1: Minor allele code or nucleotide
- A2: Major allele code or nucleotide
- GENO: Genotype totals
- O(HET): Observed heterozygote frequency
- E(HET): Expected heterozygote frequency
- P: P-value

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)	P
1	rs3934834	ALL	T	C	46/669/3103	0.1752	0.1795	0.1486
1	rs3934834	AFF	T	C	23/348/1582	0.1782	0.1814	0.4528
1	rs3934834	UNAFF	T	C	23/321/1521	0.1721	0.1774	0.1919
1	rs3737728	ALL	A	G	428/1841/1713	0.4623	0.4479	0.04379
1	rs3737728	AFF	A	G	206/950/842	0.4755	0.4493	0.009647
1	rs3737728	UNAFF	A	G	222/891/871	0.4491	0.4465	0.8406

Quality Control of the Data

After seeing the summary data, the sample data as a whole can now "cleaned" of data that may distort results due to poor data quality. Four inclusion criteria will be used in the following code to filter out less-useful data and produce a new set of binary files without them.

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile gwas --maf 0.05 --mind 0.05
--hwe 0.001 --geno 0.05 --make-bed --out pruned_data
```

Options used:

- --maf 0.05
  - Filters out any SNPs with a minor allele frequency less than 5%
  - 5% was chosen to reduce any rare alleles and increase test power
- --mind 0.05
  - Filters out any individuals missing more than 5% of their genotyping data
  - 5% was only chosen as a demonstration, as all individuals had high genotyping with only 1.5% – 1.7% missing. As such, any threshold for missingness would remove either no individuals, or far too many individuals.
- --hwe 0.001
  - Filters out SNPs that score below 0.1% significance in the Hardy-Weinberg Equilibrium
  - The default value of 0.001 was chosen to filter out only the most extreme SNPs in disequilibrium and avoid discarding SNPs that may have true association.
- --geno 0.05
  - Filters out any SNPs missing in more than 5% of individuals
  - 5% was chosen to reduce SNPs that did not have enough information
- --make-bed
  - Creates a new set of binary files
- --out pruned_data
  - Names the new binary files

The result is as follows:

```
0 of 4000 individuals removed for low genotyping ( MIND > 0.05 )
314 markers to be excluded based on HWE test ( p <= 0.001 )
        287 markers failed HWE test in cases
        314 markers failed HWE test in controls
Total genotyping rate in remaining individuals is 0.983323
5552 SNPs failed missingness test ( GENO > 0.05 )
10010 SNPs failed frequency test ( MAF < 0.05 )
After frequency and genotyping pruning, there are 290407 SNPs
After filtering, 2000 cases, 2000 controls and 0 missing
After filtering, 2000 males, 2000 females, and 0 of unspecified sex
```

After filtering, 100% of individuals and approximately 95% of SNPs remain in the study, further indicating high quality data. Genotyping only increased by ~0.1%, due to the already very high rate in the raw data. Despite the high data retention, this may indicate that thresholds for GENO and MAF, which removed a total of ~15,000 SNPs, could be lowered to increase the amount of data still in the study. Because of the size of the study, it is possible that rare alleles left in could still have potential power.

All further analysis will now be performed on the new set of remaining data in the binary files "pruned_data".

Specific Data

While summary statistics can also be generated for the filtered data, specific information can also be retrieved from the data set by using the "grep" command in Linux to search summary statistic files of the filtered data:

- The amount of SNPs individual A2038 is missing: 4716 SNPs
    - This data can be pulled from the individual missingness file, ".imiss", using grep. Out of 306,102 SNPs, individual A2038 is missing genotype information for 4,716 of them, or 1.624%.

```
[nextgen2015@node002 low_hwe]$ grep 'A2038' missing_prunes.imiss
  37   A2038          N     4716   290407  0.01624
```

- The amount of individuals missing information for SNP rs2493272: 111 individuals
    - From the other file generated from the missingness summary, ".lmiss", the number of individuals missing genotyping of a single SNP can be retrieved using "grep" again, showing that 111 individuals are missing genotypes at SNP rs2493272.

```
[nextgen2015@node002 low_hwe]$ grep 'rs2493272' missing_prunes.lmiss
   1   rs2493272      111     4000  0.02775
```

- Total missingness in the filtered data: 100% - 98.4117% = 1.5883%
    - The total missingness refers to the overall amount of missing genotype information, and can be found by reading the log file from the creation of the new filtered binary files. This line indicates that, after removing the data below our defined thresholds, the remaining data is 98.4117% genotyped.

```
Total genotyping rate in remaining individuals is 0.984117
```

- Minor allele frequency and identity of SNP rs4970357: "C" at 5.028%
    - Frequency information can be found in the ".frq" file by using "grep". From this, the minor allele at SNP rs4970357 is C, which has a minor allele frequency of 5.028%.

```
[nextgen2015@node002 low_hwe]$ grep 'rs4970357' freaqy_prunes.frq
   1   rs4970357    C    A      0.05028     7856
```

Genetic Model Testing

The next step in analysis will be genetic model association. In a basic association test, the frequency at which an allele is found in case individual is compared to the frequency at which it is found in control individuals. In this way, an allele is checked for association with a phenotype. However this only takes into consideration *presence* of the allele, and does not take into account factors such as dominance. If an allele is recessive, it may be found associated with the phenotype less often due to the fact that control individuals that carry one copy of the allele will not show the phenotype.

In the genetic model association, "--model" is used to generate a ".model" file that tests each minor allele under five different genetic models.

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile pruned_data --model --out supermodel_prunes
```

In the output, we see the same notation for chromosome, SNP, and alleles. The model file also shows the following:

- TEST: the genetic model being tested. These include:
  - GENO: Tests homozygous minor allele vs. heterozygotes vs. homozygous major allele
  - TREND: Cochran-Armitage trend test, which is similar to the basic association but does not assume Hardy-Weinberg Equilibrium
  - ALLELIC: Basic test for minor allele vs. major allele, as in the basic association test described above
  - DOM: Tests homozygous minor allele plus heterozygotes vs. homozygous major allele
  - REC: Tests homozygous minor allele vs. homozygous major allele plus heterozygotes
- AFF: The counts of affected individuals with the genotypes or alleles being tested
- UNAFF: The counts of unaffected individuals with the genotypes or alleles being tested
- CHISQ: The chi-squared statistic for the two distributions being tested
- DF: Degrees of freedom
- P: The p-value of the chi-squared statistic with specified degrees of freedom

CHR	SNP	A1	A2	TEST	AFF	UNAFF	CHISQ	DF	P
1	rs3934834	T	C	GENO	23/348/1582	23/321/1521	0.2607	2	0.8778
1	rs3934834	T	C	TREND	394/3512	367/3363	0.1277	1	0.7209
1	rs3934834	T	C	ALLELIC	394/3512	367/3363	0.1307	1	0.7177
1	rs3934834	T	C	DOM	371/1582	344/1521	0.1906	1	0.6625
1	rs3934834	T	C	REC	23/1930	23/1842	0.02475	1	0.875
1	rs3737728	A	G	GENO	206/950/842	222/891/871	2.931	2	0.231
1	rs3737728	A	G	TREND	1362/2634	1335/2633	0.1778	1	0.6733
1	rs3737728	A	G	ALLELIC	1362/2634	1335/2633	0.172	1	0.6783
1	rs3737728	A	G	DOM	1156/842	1113/871	1.257	1	0.2623

In the output excerpt below, the model data for SNP rs9651273 has been pulled from the .model file using "grep" again:

```
[nextgen2015@node002 low_hwe]$ grep 'rs9651273' supermodel_prunes.model
   1    rs9651273    A    G     GENO   322/1013/650    305/939/714        6.085    2      0.04773
   1    rs9651273    A    G    TREND      1657/2313      1549/2367        3.995    1      0.04563
   1    rs9651273    A    G  ALLELIC      1657/2313      1549/2367        3.892    1      0.04853
   1    rs9651273    A    G      DOM       1335/650       1244/714        6.029    1      0.01407
   1    rs9651273    A    G      REC       322/1663       305/1653       0.3062    1        0.58
```

From the five models tested for this SNP, all but the recessive model show significance, the dominant model showing the lowest p-value at 0.01407. This would indicate that SNP rs9651273 would be most significantly associated with the phenotype if the minor "A" allele follows a dominance model over the "G" allele.

Association Testing

As mentioned in the genetic model analysis, the basic association test compares the frequency of the minor allele between case and control individuals. In this way, an association between the allele and a phenotype can be determined. The "--assoc" option is used to generate a ".assoc" file.

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile pruned_data --assoc --out assoc_prunes
```

Besides basic information about the SNPs, the ".assoc" file shows the following information:

- F_A: The frequency of the minor allele in affected individuals
- F_U: The frequency of the allele in unaffected individuals
- CHISQ: The chi-squared statistic in the comparison of the two distributions
- P: The p-value of the chi-squared test
- OR: The odds ratio of the chi-squared test

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs3934834	995669	T	0.1009	0.09839	C	0.1307	0.7177	1.028
1	rs3737728	1011278	A	0.3408	0.3364	G	0.172	0.6783	1.02
1	rs6687776	1020428	T	0.06023	0.06212	C	0.1221	0.7268	0.9676
1	rs9651273	1021403	A	0.4174	0.3956	G	3.892	0.04853	1.095
1	rs2298217	1054842	T	0.07387	0.07089	C	0.2607	0.6097	1.045
1	rs4970357	1066927	C	0.05051	0.05005	A	0.00868	0.9258	1.01
1	rs4970362	1084601	A	0.3554	0.3467	G	0.6674	0.4139	1.039
1	rs4970420	1096336	A	0.1606	0.1813	G	5.725	0.01673	0.8643
1	rs1320565	1109721	T	0.08401	0.07372	C	2.895	0.08884	1.152

Even from the first few lines of the association test, we already see two SNPs with significant association. However, these are unadjusted p-values, which have not been corrected for multiple testing. To correct these, the "--adjust" option is added to create a ".assoc.adjusted" file.

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile pruned_data --assoc --adjust --out assoc_prunes
```

This file contains adjusted p-values using a variety of methods, including genomic control, Bonferroni, Holm, Sidak, Benjamini-Hochberg, and Benjamin-Yekutieli, in addition to the unadjusted p-values . This file is automatically sorted by significance.

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
3	rs6802898	2.327e-20	3.442e-20	6.758e-15	6.758e-15	INF	INF	6.758e-15	8.891e-14
10	rs7901695	6.563e-12	8.161e-12	1.906e-06	1.906e-06	1.906e-06	1.906e-06	9.53e-07	1.254e-05
16	rs8050136	1.006e-08	1.172e-08	0.002921	0.00292	0.002916	0.002916	0.0007381	0.009711
16	rs3751812	1.017e-08	1.185e-08	0.002952	0.002952	0.002948	0.002948	0.0007381	0.009711
10	rs7904519	2.478e-08	2.865e-08	0.007197	0.007197	0.007171	0.007171	0.00135	0.01776
3	rs7615580	2.789e-08	3.221e-08	0.008099	0.008099	0.008066	0.008066	0.00135	0.01776
10	rs7903146	3.889e-08	4.478e-08	0.01129	0.01129	0.01123	0.01123	0.001362	0.01791
3	rs6768587	3.966e-08	4.566e-08	0.01152	0.01152	0.01145	0.01145	0.001362	0.01791
3	rs2028760	4.22e-08	4.855e-08	0.01225	0.01225	0.01218	0.01218	0.001362	0.01791
3	rs12635120	3.563e-07	4.024e-07	0.1035	0.1035	0.0983	0.0983	0.01035	0.1361
10	rs12255372	1.212e-06	1.355e-06	0.3521	0.3521	0.2968	0.2968	0.03201	0.4211
5	rs11947998	1.516e-06	1.69e-06	0.4402	0.4402	0.3561	0.3561	0.03654	0.4807
3	rs307560	1.636e-06	1.823e-06	0.475	0.475	0.3781	0.3781	0.03654	0.4807
3	rs6798713	2.373e-06	2.635e-06	0.6891	0.689	0.498	0.4979	0.04922	0.6475
5	rs4976806	5.738e-06	6.324e-06	1	1	0.8111	0.811	0.1111	1
20	rs6023759	7.48e-06	8.225e-06	1	1	0.8861	0.8861	0.1358	1
10	rs10885409	8.883e-06	9.754e-06	1	1	0.9242	0.9242	0.1518	1
5	rs10038486	1.051e-05	1.152e-05	1	1	0.9527	0.9527	0.1662	1
8	rs4404875	1.087e-05	1.192e-05	1	1	0.9575	0.9574	0.1662	1
2	rs6736007	1.521e-05	1.662e-05	1	1	0.9879	0.9879	0.2208	1
3	rs9816982	2.488e-05	2.708e-05	1	1	0.9993	0.9993	0.3441	1
2	rs2683687	3.108e-05	3.376e-05	1	1	0.9999	0.9999	0.4103	1
6	rs10498910	3.526e-05	3.826e-05	1	1	1	1	0.4449	1
1	rs10492967	3.677e-05	3.988e-05	1	1	1	1	0.4449	1

Obviously, the corrected p-values show a much lower amount of significance, with only 9 SNPs showing significance across all corrected and unadjusted values, and the rest quickly approaching the max p-value of 1. The log file produced also lists the genomic inflation factor and mean chi-squared statistic, both of which are used to estimate possible errors due to population structure or stratification. However, both of these values are nearly 1, indicating that there is little evidence of such effects.

```
Genomic inflation factor (based on median chi-squared) is 1.00914
Mean chi-squared statistic is 1.01652
```

Logistic Regression

Finally, the association data can be used to generate a logistic regression:

```
[nextgen2015@node002 low_hwe]$ plink --noweb --bfile pruned_data --logistic --sex
 --covar gwas.covar --covar-name AGE --out log_prunes
```

Several options are used in this command:

- --logistic: Runs the logistic regression on the data
- --sex: Instructs PLINK to include individuals' gender as a covariate
- --covar gwas.covar: Instructs PLINK to include a list of covariates from a file named "gwas.covar"
- --covar-name AGE: Instructs PLINK to use only the covariates in the file under the column AGE

The covariate file includes the family ID, individual ID, and age of each individual:

```
FID IID AGE
0 A2001 38
1 A2002 30
2 A2003 22
3 A2004 51
4 A2005 31
5 A2006 62
6 A2007 79
7 A2008 54
8 A2009 74
```

The output file ".assoc.logistic" contains the results from the logistic regression, and includes:

- TEST: the type of t-test, ehich includes here:
  - ADD: additive effect model, which measures gene dosage effects
  - SEX: association with the gender covariate
  - AGE: association with the age covariate
- NMISS: The number of individuals with genotyping at the SNP that were used in the regression
- OR: The odds ratio
- STAT: The t-statistic of the logistic regression coefficients
- P: The p-value for the t-statistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs3934834	995669	T	ADD	3818	1.029	0.3812	0.7031
1	rs3934834	995669	T	SEX	3818	1.012	0.1908	0.8487
1	rs3934834	995669	T	AGE	3818	1.002	1.118	0.2635
1	rs3737728	1011278	A	ADD	3982	1.019	0.3867	0.6989
1	rs3737728	1011278	A	SEX	3982	1.006	0.09899	0.9211
1	rs3737728	1011278	A	AGE	3982	1.002	1.098	0.2721

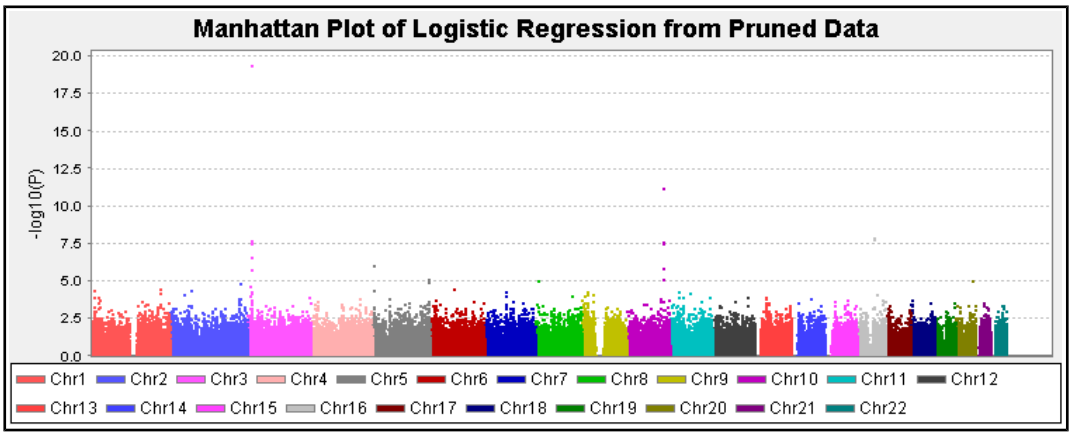
Sorting this list by p-value shows that only the additive effect model, and not the covariates, show significance.

These p-values are once again not corrected. However, an adjusted regression can also be obtained by adding the "--adjust" option.

```
[nextgen2015@node028 low_hwe]$ plink --noweb --bfile pruned_data --adjust --logistic
 --sex --covar gwas.covar --covar-name AGE --out log_adjust_prunes
```

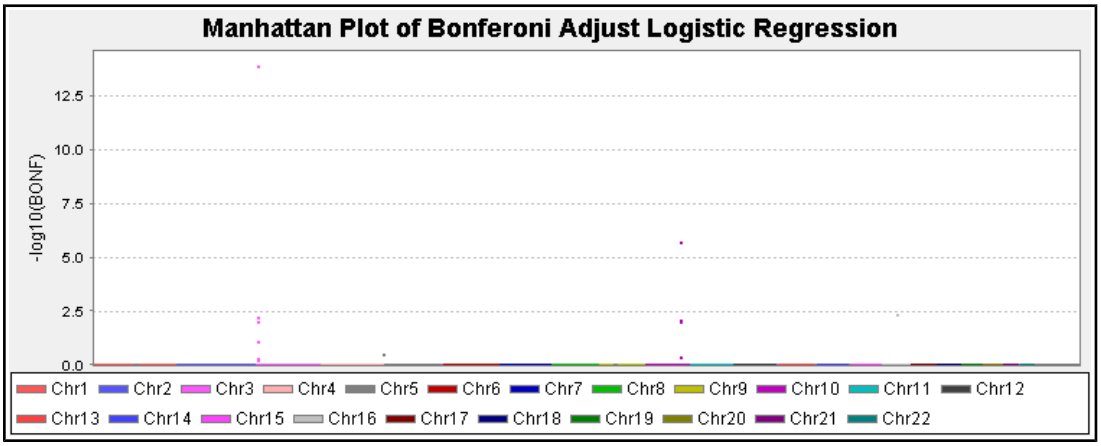
Manhattan Plot

While the logistic regression data is useful, it is much easier to visualize the data via a Manhattan plot. By using Haploview, both the logistic regression and adjusted regression can be plotted by chromosome number and base pair position against log-transformed p-value.



In the above plot, there are clearly at least two areas of significance, with one in chromosome 3 and another in chromosome 10. The clustering of significant SNPs in these two locations indicates that a possible causative variant for the phenotype could be found near these clusters, as these SNPs are highly associated with the affected phenotype.

To eliminate possible noise in the plot, a corrected p-value method can be plotted instead. In the following Manhattan plot, the same significant areas are highly visible due to the use of Bonferroni correction.



While there is less evidence to suggest true association in the SNPs plotted with low p-values in chromosomes 5 and 11, it may be useful to examine these areas further.

Ultimately, the analysis suggests that the region from approximately 10-14Mb on chromosome 3, which includes highly-associated SNP rs6802898, and the region from approximately 113-117Mb on chromosome 10, containing SNP rs7901695, are the most likely candidate areas of interest for locating a causal variant for the phenotype.