Project 2

Ezgi Karaesmen

November 14, 2016

Simulated Data

Question 1

In order to run `plink` on simulated data, `chrom12.txt` file extension was changed to `.ped` and a `chrom12.map` file was generated from `chrom12_marker_info.txt` file as shown in the Appendix 1.a. PLINK was able to detect 114 markers from a total number of 417 individuals, of which 235 and 182 were cases and controls respectively. Furthermore, 278 founders and 139 non-founders were found. This suggests that 139 affected offspring trios were included in the data.

Question 2

Allele frequencies can be summarized using the `--freq` option, which produces a `.freq` file that contains a data table with 6 columns containing chromosome, SNP identifier, allele 1 code (minor allele), allele 2 code (major allele), minor allele frequency (MAF) and non-missing allele count. A histogram of minor allele frequencies was generated using R as shown in Figure 1. More than 20% of the SNPs have a MAF between 0.10 and 0.15, also more than 15% of the SNPs have a MAF between 0.05 and 0.10.

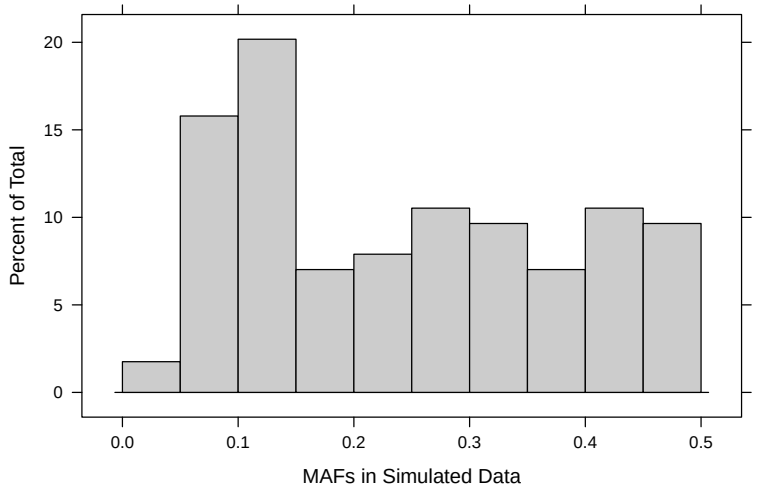

Figure 1: Histogram of minor allele frequencies

Furthermore, summary statistics for the MAFs of all SNPs were generated using R via `summary()` function. The results are shown below:

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.03957 0.11920 0.23740 0.24440 0.36870 0.49820
```

Minimum and maximum MAFs were detected as 0.03957 and 0.49820 respectively. Assuming that rare variants have a MAF less than 0.01, no rare variants could be be detected in the `chrom12` data set as the minimum MAF was determined as 0.03957. However, uncommon SNPs which have a MAF between 0.01 and 0.05 were determined as *SNP51* (MAF=0.03957) and *SNP75* (MAF=0.04496). The frequencies were

determined based only on founders in the sample and Hardy-Weinberg Equilibrium (HWE) is assumed for the population.

Question 3

SNPs were tested for HWE using the default HWE testing option `--hardy` with PLINK, which uses exact statistic to determine the significance with a threshold of `p <= 0.001`, and only considers founders. Additionally, PLINK tests HWE for all, case only and control only founders, which is specified under `TEST` as `ALL`, `AFF` and `UNAFF` in the `.hwe` output file. The summary results provided by PLINK suggested that only 1 marker failed the HWE test in cases. This SNP was determined as *SNP44* and had genotype counts of `28/68/0`. The results from HWE testing for *SNP44* is shown below.

Table 1: Output from the HWE testing for SNP44

CHR	SNP	TEST	A1	A2	GENO	O.HET.	E.HET.	P
12	SNP44	ALL	1	2	44/153/81	0.5504	0.4911	0.05132
12	SNP44	AFF	1	2	28/68/0	0.7083	0.4575	0.00000
12	SNP44	UNAFF	1	2	16/85/81	0.4670	0.4362	0.39800

The fact that only cases were deviating from HWE and the genotype counts were showing opposite trends for affected and unaffected individuals suggests that *SNP44* could be associated with the affection status and therefore it might be expected that it is not in HWE.

Furthermore, the number of SNPs failing the HWE test increases for a threshold of `p < 0.05`, and the results are listed below:

Table 2: List of markers deviating from HWE for a $p < 0.05$ cutoff

CHR	SNP	TEST	A1	A2	GENO	O.HET.	E.HET.	P
12	SNP93	AFF	1	2	6/20/70	0.2083	0.2778	0.0211100
12	SNP85	ALL	2	1	58/161/59	0.5791	0.5000	0.0115900
12	SNP56	UNAFF	2	1	25/102/55	0.5604	0.4864	0.0483800
12	SNP81	ALL	2	1	25/141/112	0.5072	0.4510	0.0460700
12	SNP81	AFF	2	1	4/51/41	0.5312	0.4257	0.0176300
12	SNP44	AFF	1	2	28/68/0	0.7083	0.4575	0.0000000
12	SNP8	ALL	1	2	4/113/161	0.4065	0.3405	0.0007317
12	SNP8	UNAFF	1	2	2/73/107	0.4011	0.3336	0.0064620
12	SNP29	UNAFF	1	2	12/46/124	0.2527	0.3107	0.0158400
12	SNP66	AFF	1	2	5/51/40	0.5312	0.4335	0.0350000
12	SNP91	ALL	1	2	56/116/106	0.4173	0.4838	0.0253700
12	SNP91	AFF	1	2	20/36/40	0.3750	0.4783	0.0345700

Question 4

A TDT test was conducted on all SNPs using the `--tdt` option on PLINK. The output of this option provides a TDT statistic with a 1df chi-squared distribution, including the odds ratio and the p-value for the statistic. TDT simply tests for the difference of transmission from the parents for the alleles of a marker. The output also includes Parental discordance statistic which takes the parents' affection status into account and a combined test statistic for both TDT and parental discordance. These test statistics, however, were ignored for the homework.

For a `p < 0.05`, 11 markers showed association with the affection status. The results for these SNPs are shown in Table 3. However, since 114 markers are tested, it would be appropriate to adjust the p-values for multiple testing. Using the `--adjust` flag in PLINK, an output with corrected p-values is generated. The output file includes

Unadjusted p-values (UNADJ), Genomic-control corrected p-values (GC), Bonferroni single-step adjusted p-values (BONF), Holm (1979) step-down adjusted p-values (HOLM), Sidak single-step adjusted p-values (SIDAK_SS), Sidak step-down adjusted p-values (SIDAK_SD), Benjamini & Hochberg (1995) step-up FDR control (FDR_BH), Benjamini & Yekutieli (2001) step-up FDR control (FDR_BY).

These adjusted p-values are shown in Table 4 for 11 SNPs that have been identified as significant before adjustment. Only 3 SNPs, *SNP43*, *SNP44* and *SNP45* remained significantly associated with the affection for any correction method.

Table 3: TDT results of associated SNPs for p < 0.05 cutoff

CHR	SNP	BP	A1	A2	T	U	OR	CHISQ	P
12	SNP2	1008787	1	2	46	68	0.6765	4.246	0.0393500
12	SNP5	1045764	1	2	10	26	0.3846	7.111	0.0076610
12	SNP43	1047802	2	1	36	79	0.4557	16.080	0.0000608
12	SNP67	1076984	1	2	85	59	1.4410	4.694	0.0302600
12	SNP68	1080202	2	1	57	81	0.7037	4.174	0.0410500
12	SNP44	1089453	1	2	125	28	4.4640	61.500	0.0000000
12	SNP45	1147565	2	1	26	67	0.3881	18.080	0.0000212
12	SNP36	1148589	2	1	64	43	1.4880	4.121	0.0423400
12	SNP98	1202021	2	1	13	26	0.5000	4.333	0.0373700
12	SNP76	1323853	2	1	77	52	1.4810	4.845	0.0277300
12	SNP77	1763582	1	2	43	22	1.9550	6.785	0.0091950

Table 4: TDT p-values for different corrections

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
12	SNP44	0.0000000	0.0000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
12	SNP45	0.0000212	0.0000432	0.002421	0.002399	0.002418	0.002397	0.00121	0.006436
12	SNP43	0.0000608	0.0001147	0.006929	0.006807	0.006905	0.006784	0.00231	0.012280
12	SNP5	0.0076610	0.0103100	0.873300	0.850300	0.583800	0.574100	0.20960	1.000000
12	SNP77	0.0091950	0.0122200	1.000000	1.000000	0.651100	0.638000	0.20960	1.000000
12	SNP76	0.0277300	0.0342300	1.000000	1.000000	0.959500	0.953300	0.43880	1.000000
12	SNP67	0.0302600	0.0371400	1.000000	1.000000	0.969900	0.963800	0.43880	1.000000
12	SNP98	0.0373700	0.0452300	1.000000	1.000000	0.987000	0.983000	0.43880	1.000000
12	SNP2	0.0393500	0.0474700	1.000000	1.000000	0.989700	0.985800	0.43880	1.000000
12	SNP68	0.0410500	0.0493800	1.000000	1.000000	0.991600	0.987700	0.43880	1.000000
12	SNP36	0.0423400	0.0508300	1.000000	1.000000	0.992800	0.988900	0.43880	1.000000

The obtained p-values from TDT test were visualized on a Manhattan plot as shown in Figure 2. Although LD information is unknown, the 3 SNPs that remained significant upon correction are located in close proximity, strongly suggesting that the association is a true-positive for that particular region.

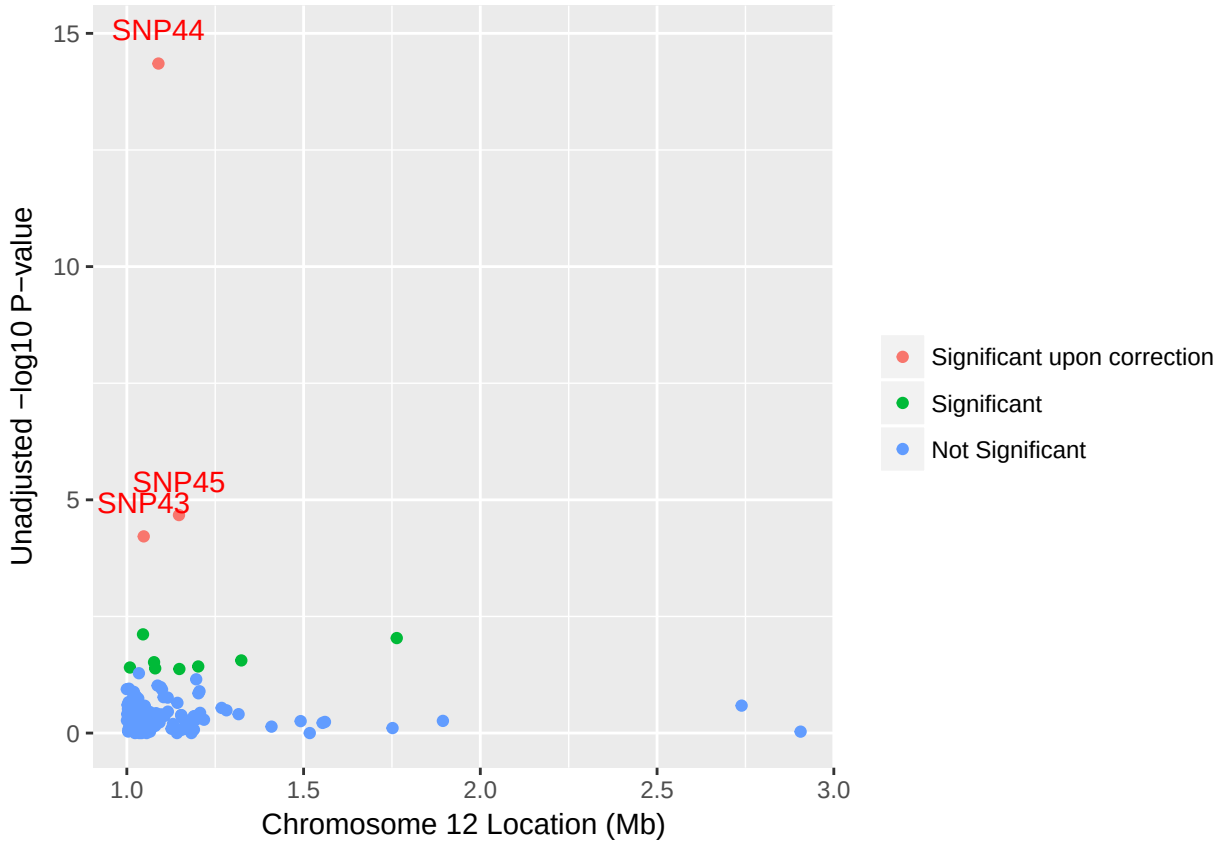

Figure 2: Manhattan plot of TDT results. The x y and axis show the chromosomal location of the marker in Mb and -log10 of the p-values. Colors suggest the significance level of the marker: blue - not significant, green - significant without correction and red - significant upon multi-test correction

NARAC Data

Question 1

In order to conduct analysis with PLINK on NARACA data, `.ped` and `.map` files were generated using the `narac_chr6.txt` file, where dummy genomic locations were assigned for SNPs on chromosome 6. Furthermore dummy family IDs matching the first column of covariate file `cov.txt` were assigned to the `.ped` file and flags `--no-parents`, `--no-sex`, `--1` and `--allow-no-sex` were used to specify missing fields in the data. NARAC data consists of 868 cases and 1194 controls with a total of 2062 individuals. A total of 199 SNPs were included for the analysis.

Question 2

Minor allele frequencies (MAFs) of the SNPs in NARAC data was estimated with PLINK and summary statistics were determined with `R` as shown below:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0211  0.1188  0.2210  0.2262  0.3184  0.4899
```

Minimum MAF was determined as `0.0211` suggesting that none of the SNPs are rare but uncommon SNPs are present. A histogram was generated to observe the MAF distribution of the SNPs in NARAC data as shown in Figure 3. Majority of the SNPs in the NARAC had a MAF between 0.05 and 0.25. Furthermore, 9 uncommon SNPs were detected and the frequency information of these SNPs are shown in Table 5. To estimate the frequency, HWE is assumed for all SNPs.

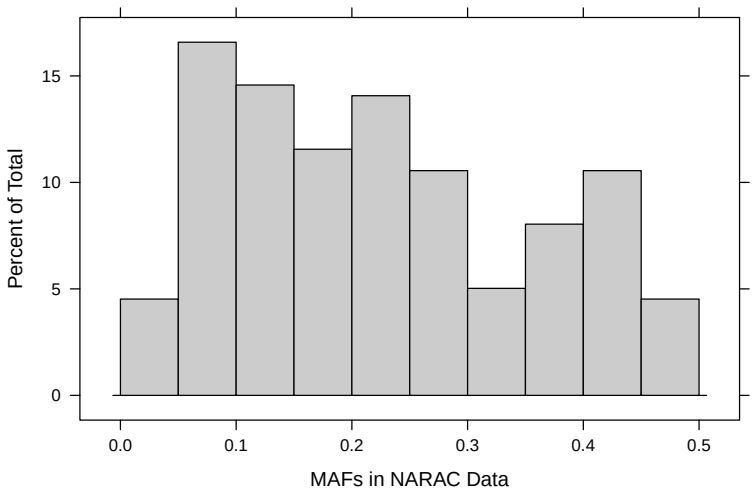

Figure 3: Histogram of minor allele frequencies of NARAC data

Chromosome	SNP	Minor Allele	Major Allele	MAF	Non-missing allele count
6	rs11753872	A	C	0.0212	4110
6	rs11758087	A	G	0.0211	4124
6	rs9258437	A	G	0.0265	4120
6	rs7741100	A	C	0.0433	4044
6	rs2517830	C	A	0.0365	3698
6	rs7760172	G	A	0.0485	4124
6	rs11967812	G	A	0.0334	3836
6	rs12665039	G	A	0.0424	4124

Chromosome	SNP	Minor Allele	Major Allele	MAF	Non-missing allele count
6	rs3757333	A	C	0.0366	4122

Question 3

SNPs in the NARAC data were tested for HWE and overall 10 markers were found to be deviating from HWE (p <= 0.001). PLINK tests HWE for both cases and controls, and provides test results for all individuals as well as for cases and controls separately. Consequently, 8 and 10 SNPs failed the HWE test in PLINK for cases and controls respectively. However, since case population is conditioned based on the affection status that might have an underlying genetic cause and therefore is likely to deviate from the expected HWE, only results from the control population was considered for further investigation.

Table 6: SNPs deviating from HWE in NARAC data

CHR	SNP	TEST	A1	A2	GENO	O.HET.	E.HET.	P
6	rs1610628	UNAFF	C	A	47/227/808	0.2098	0.2527	2.00e-07
6	rs2517930	UNAFF	A	C	129/351/618	0.3197	0.4008	0.00e+00
6	rs2517817	UNAFF	A	G	26/139/1011	0.1182	0.1492	0.00e+00
6	rs1611493	UNAFF	A	C	94/289/782	0.2481	0.3256	0.00e+00
6	rs2517713	UNAFF	C	A	245/404/419	0.3783	0.4867	0.00e+00
6	rs2975042	UNAFF	C	A	214/436/430	0.4037	0.4800	2.00e-07
6	rs2523966	UNAFF	G	A	213/389/408	0.3851	0.4814	0.00e+00
6	rs5009448	UNAFF	A	G	133/362/557	0.3441	0.4188	0.00e+00
6	rs356954	UNAFF	G	A	104/386/630	0.3446	0.3897	1.62e-04
6	rs9261154	UNAFF	A	G	19/116/1024	0.1001	0.1240	1.00e-07

The 10 SNPs that failed HWE testing in the control population are shown in Table 6. No overlap was detected between these SNPs and SNPs that have MAFs less than 0.05. This suggests that HWE deviation is likely due to genotyping error and therefore these SNPs will be excluded from the analyses. To achieve this, list of SNP ID to be excluded was written on the txt file `excl_SNP.txt` and these SNPs were excluded using the `--exclude excl_SNP.txt` option.

Question 4

Because no specific background information was provided for the SNPs, it is unclear what type of model each SNP should be following. Instead of selecting a model randomly to test the association, SNPs can be tested for multiple models and permutation test can be conducted to determine the appropriate p-value for each SNP. Since the given NARAC data has very small number of SNPs (compared to a GWAS), this approach would not be computationally heavy for our case. To test association using multiple models, PLINK's `--model` function can be preferred to also compute the T_{max} p-values `--mperm` option can also be added to the command. This outputs 2 files: `.model` file includes SNP information as well as test statistics and p-values for Cochran-Armitage trend test, Genotypic (2 df) test, Dominant (1df) test, Recessive (1df) test and the basic Allelic test. The second file `narac.model.best.mperm` provides the best results from the max(T) permutation procedure and includes both unadjusted and adjusted p-values from the permutation test. For this exercise, 10^4 iterations were done for the max(T) permutation test. 65 SNPs were determined to be associated based on `p < 0.05` for adjusted max(T) p-values.

A Manhattan plot of the max(T) p-values is shown in Figure 4. Many significant SNPs were observed to have the minimum p-value of `9.999e-05` which has a -log10 value of 4. Empirical p-values for max(T) are calculated as (R+1)/(N+1) where R is the number of times the permuted test is greater than the observed

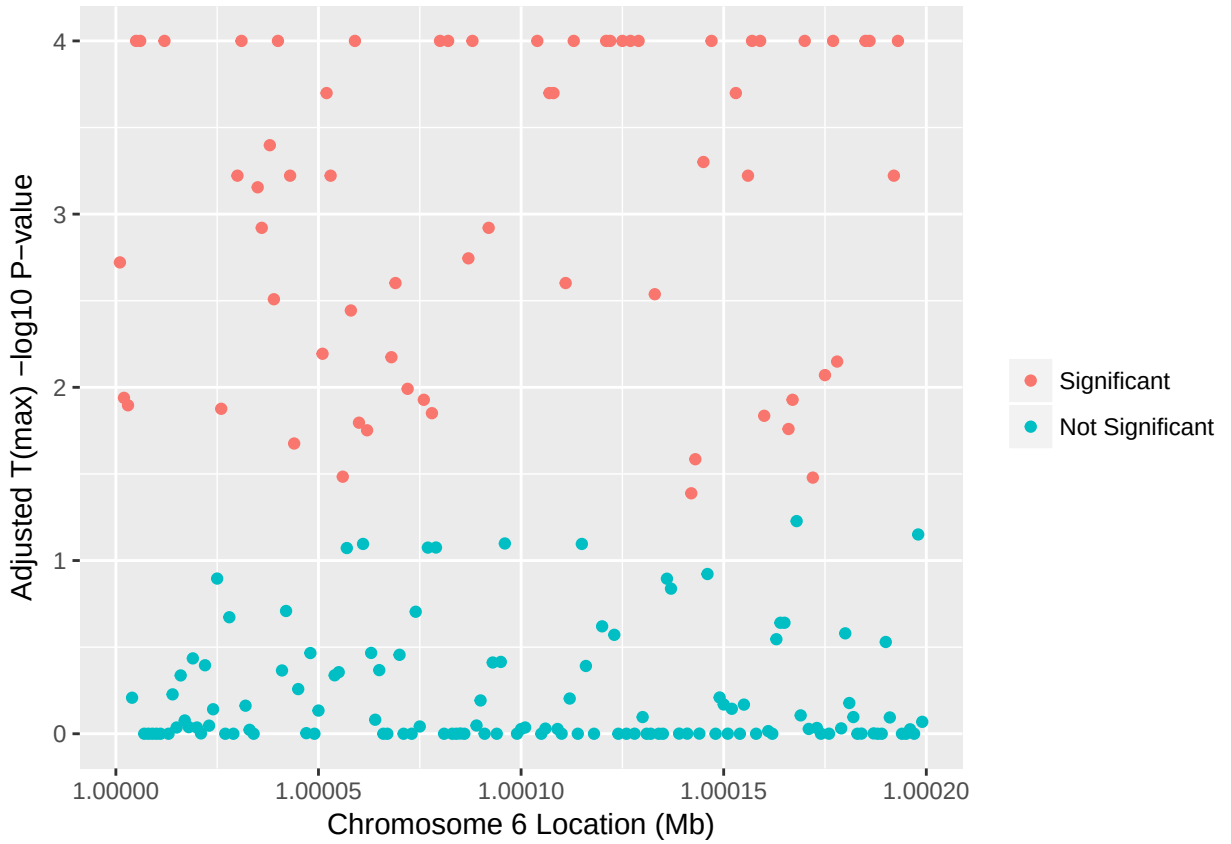

Figure 4: Manhattan plot of T(max) results. The x and y axes show the chromosomal location of the marker in Mb and -log10 of the p-values respectively.

test; N is the number of permutations. Therefore, it is likely that this set `9.999e-05` value is a result of the tests that have a much smaller observed p-value than the permuted p-values, hence R was 0 and N was equal to 10^4, giving an empirical p-value by $(0+1)/(10^4+1) \simeq 9.999e-05$. These results suggest that this genomic region on chromosome 6 is highly associated with the affection status.

Question 5

A logistic regression model was used to investigate the significance of the covariate's effect on the phenotype, also the significance of the SNP considering the covariate, and to detect any possible interaction between the SNP and the covariate. This can be achieved with options `--logistic` which uses an additive model to test the association and `--cov cov.txt` which specifies the file containing the covariate information. Also adding the `--interaction` flag tests for SNP x covariate interaction. These commands generate output file `.assoc.logistic` and the adjusted p-value file `.assoc.logistic.adjusted` (if `--adjust` flag is added). `.assoc.logistic` file includes SNP information, type of the test (`ADD`, `COV1` and `ADDxCOV1`) and the respective test results: odds ratio (`OR`), coefficient t-statistic (`STAT`) and asymptotic p-value for t-statistic (`P`). Additionally, by providing the `--ci 0.95` option, upper (`U95`) and lower (`L95`) limit of the 95% confidence interval of the odds ratio was also computed and include in the output file.

Addition of the `--interaction` flag generated an output file with large number of `NA`s for columns summarizing the test results, namely `OR`, `STAT`, `P`. According to PLINK documentation this is a problem due to multi-collinearity: when the predictor variables are too strongly correlated to each other, the parameter estimates will become unstable. PLINK tries to detect this, and will display NA for the test statistic and p-value for all terms in the model if there is evidence of multi-collinearity. This suggests that there is a strong

correlation between the genotypes and the covariate, and the interaction cannot be tested with PLINK for a large majority of the SNPs. For only 16 SNPs, the `--interaction` option generated results for the tests (`ADD`, `COV1` and `ADDxCOV1`). The `ADDxCOV1` test results for these SNPs are shown in Table 7. None of the SNPs showed significant interaction with the covariate.

Table 7: Interaction testing results for the 16 SNPs

CHR	SNP	BP	A1	TEST	NMISS	OR	SE	L95	U95	STAT	P
6	rs9258170	1000015	C	ADDxCOV1	2062	0.991	0.061	0.880	1.115	-0.157	0.876
6	rs9258186	1000017	G	ADDxCOV1	2049	0.996	0.062	0.882	1.123	-0.072	0.943
6	rs2235383	1000018	G	ADDxCOV1	2061	0.991	0.061	0.880	1.116	-0.151	0.880
6	rs2272874	1000020	G	ADDxCOV1	2062	0.991	0.061	0.880	1.115	-0.157	0.876
6	rs9258218	1000029	A	ADDxCOV1	2029	1.008	0.065	0.887	1.146	0.124	0.901
6	rs9258437	1000050	A	ADDxCOV1	2060	1.105	0.156	0.815	1.499	0.642	0.521
6	rs16896081	1000070	A	ADDxCOV1	2022	1.022	0.101	0.838	1.247	0.218	0.827
6	rs5013093	1000078	A	ADDxCOV1	1982	1.029	0.052	0.930	1.140	0.557	0.578
6	rs2517861	1000079	A	ADDxCOV1	2056	1.013	0.050	0.919	1.117	0.264	0.792
6	rs2517830	1000084	C	ADDxCOV1	1849	0.901	0.097	0.745	1.090	-1.077	0.282
6	rs429511	1000106	G	ADDxCOV1	2044	1.085	0.064	0.958	1.229	1.286	0.198
6	rs6457109	1000109	G	ADDxCOV1	2034	0.918	0.066	0.806	1.046	-1.281	0.200
6	rs16896923	1000144	G	ADDxCOV1	2062	0.991	0.086	0.837	1.173	-0.102	0.918
6	rs1264702	1000179	G	ADDxCOV1	2013	1.088	0.053	0.980	1.207	1.584	0.113
6	rs1264695	1000182	G	ADDxCOV1	2060	1.109	0.053	0.999	1.231	1.936	0.053
6	rs9261407	1000187	A	ADDxCOV1	1928	1.059	0.067	0.929	1.208	0.864	0.388

Since the `--interaction` option did not provide any information for majority of the SNPs, multiple logistic regression was conducted without interaction to explore the data. Covariate was determined to have highly significant effect on the phenotype. The summary of the unadjusted p-value distribution of the covariate as obtained from he logistic regression is shown below:

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
0.000e+00 0.000e+00 0.000e+00 5.316e-99 0.000e+00 5.810e-97
```

The fact that maximum p-value for the covariate is `5.810e-97` already suggests that the covariate has a very significant effect on the phenotype. Therefore, it is important to test the SNP effect while adjusting for the covariate. Consequently, 56 SNPs remained significantly associated with the disease upon multiple regression for a `p < 0.05` for unadjusted p-values. Upon multiple test correction, the number of significant SNPs reduced to 4. Logistic regression results and multiple test correction for these 4 SNPs are shown in Table 8 and Table 9 respectively. As shown in Table 8, rs3807031 is the only SNP with a clinically relevant odds ratio, where all SNPs have very small odds ratios even though their confidence interval does not contain 0. Results of the logistic regression was also visualized as a locus specific Manhattan plot in Figure 5.

Table 8: Logistic regression results for significant SNP effect

CHR	SNP	BP	A1	TEST	NMISS	OR	SE	L95	U95	STAT	P
6	rs3807031	1000159	A	ADD	2050	1.6920	0.1327	1.3040	2.1940	3.962	0.0000744
6	rs1737078	1000035	A	ADD	2055	0.6295	0.1197	0.4978	0.7960	-3.866	0.0001106
6	rs1737076	1000036	G	ADD	2055	0.6333	0.1199	0.5007	0.8011	-3.810	0.0001391
6	rs3131888	1000005	G	ADD	2060	0.6379	0.1209	0.5034	0.8084	-3.720	0.0001993

Table 9: Adjusted p-values of significant SNPs

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
6	rs3807031	0.0000744	0.03756	0.01406	0.01406	0.01397	0.01397	0.008761	0.05100
6	rs1737078	0.0001106	0.04242	0.02091	0.02080	0.02069	0.02059	0.008761	0.05100
6	rs1737076	0.0001391	0.04552	0.02628	0.02601	0.02594	0.02567	0.008761	0.05100
6	rs3131888	0.0001993	0.05086	0.03767	0.03707	0.03697	0.03640	0.009418	0.05483

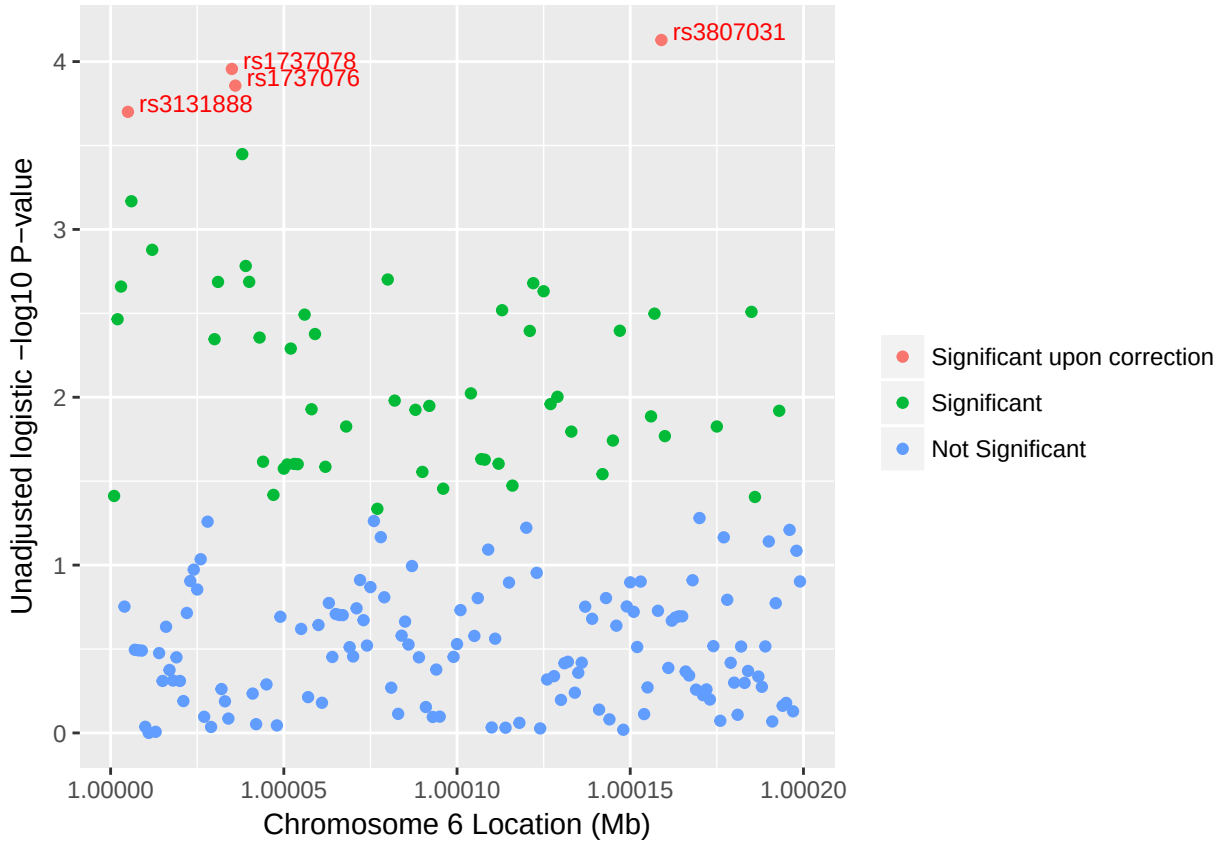

Figure 5: Manhattan plot of multiple logistic regression results. The x and y axes show the chromosomal location of the marker in Mb and -log10 of the p-values respectively.

Compared to association results obtained by the max(T) method, multiple regression had a significant drop in number significant SNPs. This suggests that the given covariate is a strong confounding factor for predicting the phenotype.

Appendix - R codes

1. Simulated Data

1.a Reformatting the .map file

Reformatting with R:

```
setwd("~/Box Sync/Stat Anal of Genetic Data/Project2/Description_Data/")

map.cM <- scan("chrom12_marker_info.txt")

# map distance 1 was added for the first SNP,
# because the chrom12_marker_info.txt file only provide inter-SNP distances
map.cM <- c(1, map.cM)

# map distances were given in cM
# therefore were converted to M
map.M  <- map.cM/100

map.file <- data.frame(CHR = rep(12, length(map.cM)),
                       SNP = paste0("SNP", 1:length(map.cM)),
                       M   = map.M,
                       BP  = map.cM*10^6)
map.file$BP <- format(map.file$BP, scientific = FALSE)
head(map.file, 3)
```

```
##   CHR  SNP          M      BP
## 1  12 SNP1 0.01000000 1000000
## 2  12 SNP2 0.01008787 1008787
## 3  12 SNP3 0.01019971 1019971
```

```
# column and row names were not included in the actual .map file
write.table(map.file, "chrom12.map", sep="\t",col.names = F, row.names = F, quote = F, )
```

1.b Question 1

On terminal:

Command:

```
$ plink --noweb --file chrom12
```

Output:

```
114 (of 114) markers to be included from [ chrom12.map ]
417 individuals read from [ chrom12.ped ]
417 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
235 cases, 182 controls and 0 missing
278 males, 139 females, and 0 of unspecified sex
Before frequency and genotyping pruning, there are 114 SNPs
278 founders and 139 non-founders found
Total genotyping rate in remaining individuals is 1
```

```
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 114 SNPs
After filtering, 235 cases, 182 controls and 0 missing
After filtering, 278 males, 139 females, and 0 of unspecified sex
```

1.c Question 2

On terminal:

Command:

```
$ plink --noweb --file chrom12 --freq --out chrom12_q2
```

Output:

```
114 (of 114) markers to be included from [ chrom12.map ]
417 individuals read from [ chrom12.ped ]
417 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
235 cases, 182 controls and 0 missing
278 males, 139 females, and 0 of unspecified sex
Before frequency and genotyping pruning, there are 114 SNPs
278 founders and 139 non-founders found
```

On R:

```
# histogram
library(lattice)
freq <- read.table("chrom12_q2.frq", header = T)
histogram(freq$MAF, col="gray80", xlab="MAFs in Simulated Data",
          breaks=seq(0, 0.5, by=0.05))
summary(freq$MAF)
```

1.d Question 3

On terminal:

Command:

```
$ plink --noweb --file chrom12 --hardy --out chrom12_hardy
```

Output:

```
Writing Hardy-Weinberg tests (founders-only) to [ chrom12_hardy.hwe ]
0 markers to be excluded based on HWE test ( p <= 0.001 )
    1 markers failed HWE test in cases
    0 markers failed HWE test in controls
```

On R:

```
HWE <- read.table("chrom12_hardy.hwe", header = T)
HWE[ HWE$P < 0.05,]
```

1.d Question 4

On terminal:

Command:

```
$ plink --noweb --file chrom12 --tdt --out chrom12 --adjust
```

Output:

```
139 nuclear families, 0 founder singletons found
139 non-founders with 2 parents in 139 nuclear families
0 non-founders without 2 parents in 0 nuclear families
139 affected offspring trios
90 phenotypically discordant parent pairs found
0 Mendel errors detected in total
```

On R:

```
TDT <- read.table("chrom12.tdt", header = T, stringsAsFactors = F)

TDT.adj <- read.table("chrom12.tdt.adjusted", header = T, stringsAsFactors = F)

# results from TDT
sigTDT <- subset(TDT, P < 0.05, select=1:10)
kable(sigTDT, row.names = F, caption = "TDT results of associated SNPs for p < 0.05 cutoff")

# adj p-values
sig.adj.TDT <- TDT.adj[ TDT.adj$SNP %in% sigTDT$SNP,]
kable(sig.adj.TDT, row.names = F, caption = "TDT p-values for different corrections")

# plotting the manhattan plot
TDT$legend <- cut(TDT$P, c(1, 0.05, 0.05/114, 10^-20) ,
                  labels = c("Significant upon correction" , "Significant",  "Not Significant") )

p <- ggplot(TDT, aes(BP/10^6, -log10(P) ,  label = SNP) )
p <- p + geom_point(aes(colour = legend))
p <- p + labs(list(x = "Chromosome 12 Location (Mb)", y = "-log10 P-value", colour=" "))
p + geom_text(data=subset(TDT, P < 0.05/114), vjust = 0, nudge_y = 0.5, col="red")
```

2. NARAC Data

2.a Reformatting/Generating the .map and .ped files

On R:

```
## narac.ped
narac <- read.table("narac_chr6.txt", header = T, stringsAsFactors = F)
narac[is.na(narac)] <- "0_0"
anyNA(narac)

rsIDs = colnames(narac[ ,-c(1,2)])

temp2 = list()
for(i in rsIDs){
  temp = NULL
  temp <- narac[,i]
  temp = strsplit(temp, split="_")
  temp2[[i]] = do.call(rbind, temp)
}

genotypes <- do.call(cbind, temp2)
narac.ped = data.frame(FID= c(1:nrow(narac)),narac[,1:2], genotypes)
narac.ped[1:10, 1:10]
write.table(narac.ped, "narac.ped", col.names = F, row.names = F, quote = F)

## narac.map

narac.map <- data.frame(CHR = rep(6, length(rsIDs)),
           SNP = rsIDs,
           M   = rep(0, length(rsIDs)),
           BP  = 1:length(rsIDs)+10^6)
write.table(narac.map, "narac.map", col.names = F, row.names = F, quote = F)
```

2.b Question 1

On terminal:

Command:

```
$ plink --noweb --file narac --no-parents --no-sex --1 --allow-no-sex
```

Output:

```
Options in effect:
    --noweb
    --file narac
    --no-parents
    --no-sex
    --1
    --allow-no-sex
    --freq

199 (of 199) markers to be included from [ narac.map ]
Warning, found 2062 individuals with ambiguous sex codes
```

```
Writing list of these individuals to [ plink.nosex ]
2062 individuals read from [ narac.ped ]
2062 individuals with nonmissing phenotypes
Assuming a disease phenotype (0=unaff, 1=aff, other=miss)
868 cases, 1194 controls and 0 missing
0 males, 0 females, and 2062 of unspecified sex
Before frequency and genotyping pruning, there are 199 SNPs
2062 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 0.984328
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 199 SNPs
After filtering, 868 cases, 1194 controls and 0 missing
After filtering, 0 males, 0 females, and 2062 of unspecified sex
```

2.c Question 2

On terminal:

Command:

```
$ plink --noweb --file narac --no-parents --no-sex --1 --allow-no-sex --freq --out narac
```

On R:

```r
narac.freq <- read.table("narac.frq", header = T)

# statistical summary
summary(narac.freq$MAF)

# histogram
histogram(narac.freq$MAF, col="gray80", xlab="MAFs in NARAC Data",
          breaks=seq(0, 0.5, by=0.05))

# subsetting uncommon SNPs
uncomm <- subset(narac.freq, MAF < 0.05)
```

2.d Question 3

Command:

```
$ plink --noweb --file narac --no-parents --no-sex --1 --allow-no-sex --hardy --out narac
```

Output:

```
Writing Hardy-Weinberg tests (founders-only) to [ narac.hwe ]
10 markers to be excluded based on HWE test ( p <= 0.001 )
    8 markers failed HWE test in cases
    10 markers failed HWE test in controls
Total genotyping rate in remaining individuals is 0.984328
```

On R:

```r
narac.freq <- read.table("narac.frq", header = T)

# statistical summary
```

```r
summary(narac.freq$MAF)

# histogram
histogram(narac.freq$MAF, col="gray80", xlab="MAFs in NARAC Data",
          breaks=seq(0, 0.5, by=0.05))

# subsetting uncommon SNPs
uncomm <- subset(narac.freq, MAF < 0.05)
```

2.e Question 4

On Terminal

Command:

```
$ plink --noweb --file narac --exclude excl_SNP.txt --no-parents --no-sex --1
                             --allow-no-sex --model --mperm 10000 --out narac
```

Output:

```
Options in effect:
    --noweb
    --file narac
    --exclude excl_SNP.txt    # excluded SNPs that failed HWE test
    --no-parents
    --no-sex
    --1
    --allow-no-sex
    --model
    --mperm 10000
    --out narac

[...]

Full-model association tests, minimum genotype count: --cell 5
Writing full model association results to [ narac.model ]
Using BEST of ALLELIC, DOM and REC for --model permutation
maxT permutation: 10000 of 10000
```

On R

```r
#read-in the results

tmax <- read.table("narac.model.best.mperm", header = T)
assoc <- read.table("narac.model", header = T)
narac.map <- read.table("narac.map", header = F,
                        col.names = c("CHR", "SNP", "M", "BP"))
```

```r
head(tmax)
```

```
##   CHR       SNP      EMP1      EMP2
## 1   6 rs3117294 2.000e-04 1.900e-03
## 2   6 rs2747453 3.000e-04 1.150e-02
## 3   6 rs2747454 2.000e-04 1.270e-02
## 4   6 rs2747457 1.950e-02 6.194e-01
## 5   6 rs3131888 9.999e-05 9.999e-05
```

```
## 6   6 rs7776082 9.999e-05 9.999e-05
```

```
head(assoc)
```

```
##   CHR       SNP A1 A2    TEST        AFF       UNAFF CHISQ DF         P
## 1   6 rs3117294  C  A    GENO 70/363/432 161/531/502 20.62  2 3.326e-05
## 2   6 rs3117294  C  A   TREND   503/1227    853/1535 19.72  1 8.955e-06
## 3   6 rs3117294  C  A ALLELIC   503/1227    853/1535 20.06  1 7.513e-06
## 4   6 rs3117294  C  A     DOM    433/432     692/502 12.63  1 3.803e-04
## 5   6 rs3117294  C  A     REC     70/795    161/1033 14.64  1 1.302e-04
## 6   6 rs2747453  G  A    GENO 11/206/651  48/330/816 19.39  2 6.151e-05
```

```
# plotting the manhattan plot
tmax_plot <- merge(narac.map, tmax)
tmax_plot$legend <- cut(tmax_plot$EMP2, c(1, 0.05, 0) ,
                  labels = c("Significant" , "Not Significant") )

p <- ggplot(tmax_plot, aes(BP/10^6, -log10(EMP2) ,  label = SNP) )
p <- p + geom_point(aes(colour = legend))
p <- p + labs(list(x = "Chromosome 6 Location (Mb)",
                   y = "Adjusted T(max) -log10 P-value", colour=" "))
p
```

2.e Question 5

On Terminal:

Command 1 - with interaction testing:

```
plink --noweb --file narac --exclude excl_SNP.txt --no-parents --no-sex --1 --allow-no-sex
                           --logistic --interaction --covar cov.txt --out narac_int --adjust --ci 0.95
```

Output:

```
Options in effect:
    --noweb
    --file narac
    --exclude excl_SNP.txt
    --no-parents
    --no-sex
    --1
    --allow-no-sex
    --logistic
    --interaction
    --covar cov.txt
    --out narac_int
    --adjust
    --ci 0.95

[...]

Reading list of SNPs to exclude [ excl_SNP.txt ] ... 10 read
Reading 1 covariates from [ cov.txt ] with nonmissing values for 2062 individuals
Before frequency and genotyping pruning, there are 189 SNPs

[...]
```

```
Converting data to Individual-major format
Writing logistic model association results to [ narac_int.assoc.logistic ]
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is 1
Mean chi-squared statistic is 0.067969
Correcting for 189 tests
Writing multiple-test corrected significance values to [ narac_int.assoc.logistic.adjusted ]
```

Command 2 - without interaction testing:

```
./plink --noweb --file narac --exclude excl_SNP.txt --no-parents --no-sex --1 --allow-no-sex
                            --logistic --covar cov.txt --out narac --adjust --ci 0.95
```

Output:

```
Options in effect:
  --noweb
    --file narac
    --exclude excl_SNP.txt
    --no-parents
    --no-sex
    --1
    --allow-no-sex
    --logistic
    --covar cov.txt
    --out narac
    --adjust
    --ci 0.95

[...]

Reading list of SNPs to exclude [ excl_SNP.txt ] ... 10 read
Reading 1 covariates from [ cov.txt ] with nonmissing values for 2062 individuals
Before frequency and genotyping pruning, there are 189 SNPs

[...]

Writing logistic model association results to [ narac.assoc.logistic ]
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is 3.62919
Mean chi-squared statistic is 3.03185
Correcting for 189 tests
Writing multiple-test corrected significance values to [ narac.assoc.logistic.adjusted ]
```

```
## read-in and subset files with interaction testing
int_log.assc <- read.table("narac_int.assoc.logistic", header = T)
```

```
length(unique(int_log.assc[complete.cases(int_log.assc), "SNP"]))
```

```
## [1] 16
```

```
head(int_log.assc)
```

```
##   CHR       SNP      BP A1     TEST NMISS OR SE L95 U95 STAT  P
## 1   6 rs3117294 1000001  C      ADD  2059 NA NA  NA  NA   NA NA
## 2   6 rs3117294 1000001  C     COV1  2059 NA NA  NA  NA   NA NA
```

```
## 3   6 rs3117294 1000001  C ADDxCOV1  2059 NA NA  NA  NA   NA NA
## 4   6 rs2747453 1000002  G      ADD  2062 NA NA  NA  NA   NA NA
## 5   6 rs2747453 1000002  G     COV1  2062 NA NA  NA  NA   NA NA
## 6   6 rs2747453 1000002  G ADDxCOV1  2062 NA NA  NA  NA   NA NA
```

```
comp_int_log.assc <-int_log.assc[complete.cases(int_log.assc), ]
comp_int_log.assc <- subset(comp_int_log.assc, TEST == "ADDxCOV1")

## Table 7
kable(comp_int_log.assc, row.names = F, digits = 3, caption = "Interaction testing results for the 16 SI



## logistic regression results without interaction
log.assc <- read.table("/narac.assoc.logistic", header = T, stringsAsFactors = F)
adj.assc <- read.table("narac.assoc.logistic.adjusted", header = T, stringsAsFactors = F)
```

```
head(log.assc)
```

```
##   CHR       SNP      BP A1 TEST NMISS     OR      SE    L95    U95   STAT
## 1   6 rs3117294 1000001  C  ADD  2059 0.7753 0.12310 0.6090 0.9869 -2.067
## 2   6 rs3117294 1000001  C COV1  2059 1.8980 0.02904 1.7930 2.0090 22.070
## 3   6 rs2747453 1000002  G  ADD  2062 0.6206 0.16300 0.4508 0.8542 -2.926
## 4   6 rs2747453 1000002  G COV1  2062 1.8950 0.02889 1.7900 2.0050 22.120
## 5   6 rs2747454 1000003  A  ADD  2036 0.6347 0.14840 0.4745 0.8490 -3.063
## 6   6 rs2747454 1000003  A COV1  2036 1.9030 0.02934 1.7960 2.0150 21.920
##            P
## 1  3.873e-02
## 2 6.344e-108
## 3  3.428e-03
## 4 2.210e-108
## 5  2.190e-03
## 6 1.516e-106
```

```
head(adj.assc)
```

```
##   CHR       SNP     UNADJ      GC    BONF    HOLM SIDAK_SS SIDAK_SD
## 1   6 rs3807031 7.441e-05 0.03756 0.01406 0.01406  0.01397  0.01397
## 2   6 rs1737078 1.106e-04 0.04242 0.02091 0.02080  0.02069  0.02059
## 3   6 rs1737076 1.391e-04 0.04552 0.02628 0.02601  0.02594  0.02567
## 4   6 rs3131888 1.993e-04 0.05086 0.03767 0.03707  0.03697  0.03640
## 5   6 rs9258275 3.560e-04 0.06088 0.06729 0.06587  0.06509  0.06376
## 6   6 rs7776082 6.794e-04 0.07450 0.12840 0.12500  0.12050  0.11760
##     FDR_BH  FDR_BY
## 1 0.008761 0.05100
## 2 0.008761 0.05100
## 3 0.008761 0.05100
## 4 0.009418 0.05483
## 5 0.013460 0.07835
## 6 0.021400 0.12460
```

```
sig.assc.adj <- subset(adj.assc, BONF < 0.05)
sig.assc <- log.assc[log.assc$SNP %in% sig.assc.adj$SNP & log.assc$TEST == "ADD",]
sig.assc <- sig.assc[order(sig.assc$P),]



## Tables 6 & 7
```

```r
kable(sig.assc, row.names = F, caption="Logistic regression results for significant SNP effect")
kable(sig.assc.adj, row.names = F, caption="Adjusted p-values of significant SNPs")

# plotting the manhattan plot
log.assc.plot <- log.assc[!log.assc$TEST == "COV1",]
log.assc.plot$legend <- cut(log.assc.plot$P, c(1, 0.05, 0.0002, 10^-20) ,
                  labels = c("Significant upon correction" , "Significant",  "Not Significant") )


p <- ggplot(log.assc.plot, aes(BP/10^6, -log10(P) ,  label = SNP) )
p <- p + geom_point(aes(colour = legend))
p <- p + labs(list(x = "Chromosome 6 Location (Mb)", y = "Unadjusted logistic -log10 P-value", colour="
p + geom_text(data=subset(log.assc.plot, P < 0.0002), col="red", vjust = -0.4)
```