EXAMPLE

Note: This is a great report, but your report can be considerably briefer than this very thorough example!

Bio 227 - Gene Mapping Final Project

Introduction

Asthma is one of the most common chronic diseases of childhood; family-based and twin studies demonstrate that the disorder arises from complex interaction of genetic and environmental factors[3]. The Childhood Asthma Management Program (CAMP) study was a large-scale, multicenter trial that followed 1041 children with asthma.[2] In 2009, Himes *et al.* performed a genome-wide association study of asthma using CAMP participants as cases and controls from Illumina's database, matched on ancestry principal components.[3] Here, we use the genetic data generated by this study to identify through our own alleles that are either causally associated with asthma or associated with asthma through linkage disequilibrium. Available genetic information consisted of single nucleotide polymorphisms (SNPs) for four selected autosomal chromosomes.

Quality Control

The original dataset included a total of 1205 unrelated individuals (359 asthma affected cases and 846 matched controls) and 556,505 genotyped SNPs available on four autosomal chromosomes (chromosomes 2, 5, 13 and 20). Information about sex chromosomes were not available. Although according to project description imputation and basic quality controls had been done beforehand, we performed additional statistical quality controls to ensure the validity of the data.

Since all individuals and all SNPs had a genotyping success rate of 100%, filtering excess missingness was not necessary in individuals or markers. We excluded SNPs with minor allele frequency of less than 5% (82,022 SNPs). In addition, we counted SNPs that failed the Hardy-Weinberg test at $P < 1 \times 10^{-5}$ as departure from Hardy-Weinberg Equilibrium and removed them from the dataset (997 SNPs). We could not perform a sex check, as information about sex chromosomes were not available; given the limited amount of genomic data from only 4 chromosomes, we did not attempt to infer relatedness to filter cryptically related subjects.

We examined the data for population substructure by principal component (PC) analysis. Since the original design of the study matched controls on the basis of genetic similarities, we suspected that there would be minimal population stratification. Plink was used to generate the first four eigenvalues of the SNP covariance matrix. We then plotted cases versus controls using R to see if any clustering appeared.

As shown in Figure 1, the cases and controls overlapped well in the space, and no major clustering was observed.

Analytic Strategy

Several models were fit to attempt to determine which SNPs were significantly associated with the outcome. Since we did not have any *a priori* knowledge of the mode of inheritance, we felt that models that test an additive mode of inheritance (the alleles test, trend test, codominant model, and logistic regression) could all be appropriate. Since we had available the covariate "gender", we explored whether it could be a confounder in the analysis. As shown in Table 1 in Appendix 3, cases were more common among males than among females. The Pearson's Chi-squared test of independence (Chi-squared=64.5, p<1e-15) rejected the null hypothesis that asthma status was independent of gender. Consequently, gender was included as a covariate in the final model, and we chose logistic regression (assuming additive mode of inheritance) as our primary model to allow adjustment for this covariate. For comparison, we also performed the alleles test, the recessive test, the codominant test, and the trend test.

While our examination of population substructure by principal component analysis suggested that inclusion of PCs would not affect the results, we ran two additive logistic regression models, with and without PCs, to explore whether such adjustment would affect the results.

We adjusted p-values in order to control for false discovery due to multiple testing. Given the inclusion of ~500,000 SNPs in our sample, a Bonferroni threshold of significance, assuming independence of SNPs, would be about 1×10^{-7}.

Results

After frequency and genotyping pruning, there were 1,205 individuals and 473,488 SNPs left. Our QC filters excluded none of the individuals and 83,017 SNPs from the original dataset. Out of 1,205 individuals, there were 359 asthma cases and 846 controls. There was little evidence of population stratification given the genomic inflation factor of 1.02 in our unadjusted analysis. A quantile-quantile (QQ) plot comparing allelic-association *p*-values from our logistic regression model demonstrated deviation at the tail, suggesting true associations between SNPs and asthma. (See Figure 2 in Appendix 3)

None of the models identified any SNPs with p-values below our *a priori* Bonferroni significance threshold of $p < 1 \times 10^{-7}$. We therefore report our top SNPs meeting a nominal *p*-value threshold of $<1 \times 10^{-6}$. Using logistic regression controlling for gender, we found 11 SNPs on Chromosome 5, and one on Chromosome 13 that passed this threshold, shown on the Manhattan plot (Figure 3) in Appendix 3; nominally

significant SNPs are outlined in the table 2 (As seen in Appendix 3). Clearly, the SNPs of interest on chromosome 5 are close together in terms of physical distance. A linkage disequilibrium (LD) diagram (Appendix 3 - Figure 4) illustrates they are in a region of tight LD.

Compared with our findings for logistic regression, the alleles test, dominant test, and codominant test gave very similar results. The recessive test did not identify any SNPs that met our nominal *P*-value threshold, which is as expected as this test has low power unless the true model is recessive.

Discussion

Using a case-control design, we identified 11 SNPs on two chromosomes associated with childhood asthma. The chromosome 5 SNPs are in a region of tight LD in the gene *PDE4D* (phosphodiesterase, 4D, cAMP-specific), consistent with the findings of Himes *et al*.[3] The phosphodiesterases are a family of proteins involved in intracellular cAMP signaling; the PDE4 subfamily is known to be specifically expressed in airway smooth muscle, immune, and inflammatory cells.[4] Thus, this locus is a good biological candidate for involvement in asthma pathogenesis. The SNP on chromosome 13 did not lie in a known gene. While this may mean a spurious association, we note by examining the regional association plot (as seen in Appendix 3, figure 5 and 6) that there are a cluster of SNPs in tight LD just below our nominal *p*-value threshold. Further studies, perhaps with larger cohorts to increase power, are needed to understand this finding.

A strength of this study is that the cases came from a formal study of childhood asthma and thus had a rigorously defined phenotype. However, this study has several limitations. First, the sample size is relatively small and thus underpowered to detect loci with effect sizes in the range expected for complex disease; indeed, we did not detect any associations that met our *a priori* threshold of significance. Second, the cases and controls come from different populations. Matching was done on the basis of genomic data; however, we do not know if the cases and controls differ on non-genetic factors that may affect the asthma phenotype, such as socioeconomic status or exposure to smoking in the home.

References

1. Ober, C. and Hoffjian, S. Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun.* **7**, 95-100 (2006).
2. Childhood Asthma Management Group. The Childhood Asthma Managment Program (CAMP): design, rationale, methods. *Contrl Clin Trials*. **1,** 91-120 (1999).

3. Himes, B.E. *et al*. Genome-wide association analysis identifies *PDE4D* as an asthma-susceptibility gene. *Am. J. Hum. Genet.* **84**, 581–593 (2009).
4. Aspiotis R, Deschênes D, Dubé D, Girard Y, Huang Z, Laliberté F, Liu S, Papp R, Nicholson DW, Young RN. The discovery and synthesis of highly potent subtype selective phosphodiesterase 4D inhibitors. *Bioorg Med Chem Lett.* **18**, 5502-5 (2010).

Individual Contributions

- Suqin Hou - collaborated in the quality control in PLINK, drew the Manhattan plot by Haploview, wrote up the initial draft of the introduction and quality control sections of the report and edited report.
- Dong Yan - collaborated on checking population substructure by principal component analysis using PLINK and R, plotted cases versus controls to see if any clustering appeared; created analysis plan part presentation.
- Tom Madsen - collaborated with Jose on preliminary PLINK association analysis including logistic regression, wrote R script to generate initial QQ plots and to analyze PLINK results, edited report.
- Will Townes- assessment of gender as a covariate, principal components analysis in quality control, and the logistic regression with principal components as covariates. Wrote up the initial draft of the Results section in the article.
- Jose Malagon Lopez-collaborated in the preliminary analysis running in PLINK the association, model and logistic tests, without covariates and with covariate GENDER. Also collaborated in the initial draft of the Analytic Strategy section.
- Jennifer Todd- collaborated on the quality control, contributed to QC and discussion sections, performed literature search, created regional association plots, and created oral presentation.

Appendix 1- PLINK commands

```
/* Remove the individuals with missing rate > 2% and the individuals with missing phenotype */
plink --bfile data/CAMPdata --prune --mind 0.02 --make-bed --out data/CAMPdata1

/* Remove the SNPs with missing rate > 10%, MAF < 5% and HWE p-value < 0.0001 */
plink --bfile data/CAMPdata1 --geno 0.1 --maf 0.05 --hwe 0.0001 --make-bed --out data/CAMPdata2

/* Generate Assoc files for use in R preliminary analysis */
plink --bfile data/CAMPdata2 --assoc --out data/CAMPdata2
plink --bfile data/CAMPdata2 --model --out data/CAMPdata2
plink --bfile data/CAMPdata2 --logistic --covar data/pheno.txt --covar-name GENDER --adjust --out data/CAMPdata2

/* Generate files for Multidimensional Scaling (aka, Principal Components) */
/* Step 1: prune SNPs to get a subset of independent ones */
plink --bfile data/CAMPdata2 --maf 0.05 --indep-pairwise 50 5 .05 --out data/plink.prune
plink --bfile data/CAMPdata2 --extract data/plink.prune.prune.in --make-bed --out data/CAMPdata2indep
/* Step 2: Generate principal components file after pruning, can be visualized in R */
plink --bfile data/CAMPdata2indep --genome --cluster --mds-plot 4 --out data/plink_

/* Re-run logistic regression with PCs included as covars */
plink --bfile data/CAMPdata2 --logistic --covar data/covars_full.txt --covar-name GENDER,C1,C2,C3,C4 --adjust --out
data/CAMPdata2_pc_covar

/* Make subset of SNPs for analysis in Haploview */
plink --bfile data/CAMPdata2 --extract topsnps.txt --recode --out haploview_input
```

Appendix 2- R code

```
#Import data from --assoc command in plink
assoc_data = read.table('data/CAMPdata2.assoc',header=T)

#Code to generate a crude Q-Q plot
sorted_assoc_p_values = sort(assoc_data$P)
uniform_quantiles = seq(from=1/length(sorted_assoc_p_values),to=1-
1/length(sorted_assoc_p_values),length.out=length(sorted_assoc_p_values))
plot(-log10(uniform_quantiles),-log10(sorted_assoc_p_values),cex=0.1)
abline(a=0,b=1)

#How many SNPs meet various levels of significance?
bonferroni_level = 0.05/length(sorted_assoc_p_values)
bonferroni_level
sum(sorted_assoc_p_values < bonferroni_level)
sum(sorted_assoc_p_values < 10^-7)
sum(sorted_assoc_p_values < 10^-6)

#Extract rows of the data frame corresponding to highly significant SNPs
assoc_signif_SNPs = assoc_data[assoc_data$P < 10^-6,]
```

```
assoc_signif_SNPs

#Create a vector of names of significant SNPs
assoc_signif_SNP_names = as.character(assoc_signif_SNPs$SNP)
assoc_signif_SNP_names

#Import data from --model command in plink (WARNING: Potentially very slow)
model_data = read.table('data/CAMPdata2.model',header=T)

#Trend test analysis
trend_data = model_data[model_data$TEST == "TREND",]
trend_signif_SNPs = trend_data[trend_data$P < 10^-6,]
trend_signif_SNP_names = as.character(trend_signif_SNPs$SNP)
trend_signif_SNP_names
length(trend_signif_SNP_names) #how many highly significant genes?
sum(trend_signif_SNP_names %in% assoc_signif_SNP_names) #how many were also identified by assoc?

#Genotypic model analysis
geno_data = model_data[model_data$TEST == "GENO",]
geno_signif_SNPs = geno_data[!is.na(geno_data$P) & geno_data$P < 10^-6,]
geno_signif_SNP_names = as.character(geno_signif_SNPs$SNP)
geno_signif_SNP_names
length(geno_signif_SNP_names) #how many highly significant genes?
sum(geno_signif_SNP_names %in% assoc_signif_SNP_names)  #how many were also identified by assoc?

#Dominant model analysis
dom_data = model_data[model_data$TEST == "DOM",]
dom_signif_SNPs = dom_data[!is.na(dom_data$P) & dom_data$P < 10^-6,]
dom_signif_SNP_names = as.character(dom_signif_SNPs$SNP)
dom_signif_SNP_names
length(dom_signif_SNP_names) #how many highly significant genes?
sum(dom_signif_SNP_names %in% assoc_signif_SNP_names) #how many were also identified by assoc?

#Recessive model analysis
rec_data = model_data[model_data$TEST == "REC",]
rec_signif_SNPs = rec_data[!is.na(rec_data$P) & rec_data$P < 10^-6,]
rec_signif_SNP_names = as.character(rec_signif_SNPs$SNP)
rec_signif_SNP_names
length(rec_signif_SNP_names) #how many highly significant genes?
sum(rec_signif_SNP_names %in% assoc_signif_SNP_names) #how many were also identified by assoc?

#Logistic regression analysis
logistic_data = read.table('data/CAMPdata2.assoc.logistic.adjusted',header=T)
logistic_signif_SNPs = logistic_data[!is.na(logistic_data$UNADJ) & logistic_data$UNADJ < 10^-6,]
logistic_signif_SNP_names = as.character(logistic_signif_SNPs$SNP)
logistic_signif_SNP_names
length(logistic_signif_SNP_names) #how many highly significant genes?
sum(logistic_signif_SNP_names %in% assoc_signif_SNP_names) #how many were also identified by assoc?

# Script to merge the PCs and the main covariate file
library(plyr)
covar1<-read.table("data/pheno.txt",header=TRUE)
pcs<-read.table("data/plink_.mds",header=TRUE)
covar2<-merge(covar1,pcs,by=c("FID","IID"))
write.table(covar2,file="data/covars_full.txt",row.names=FALSE,quote=FALSE)

#get additional information
bps0<-read.table('data/CAMPdata2.assoc.logistic',header=TRUE)
```

```
logistic_signif_SNPs<-join(bps0,logistic_signif_SNPs,by=c("CHR","SNP"),"right")
logistic_signif_SNPs<-logistic_signif_SNPs[logistic_signif_SNPs$TEST=="ADD",]
write.csv(logistic_signif_SNPs,file="top_SNPs_logistic.csv",row.names=FALSE)

#Logistic regression adjusting for PCs
logistic_data2 = read.table('data/CAMPdata2_pc_covar.assoc.logistic.adjusted',header=T)
logistic_signif_SNPs_2 = logistic_data2[!is.na(logistic_data2$UNADJ) & logistic_data2$UNADJ < 10^-6,]
logistic_signif_SNP_names_2 = as.character(logistic_signif_SNPs_2$SNP)
logistic_signif_SNP_names_2
length(logistic_signif_SNP_names_2) #how many highly significant genes?
sum(logistic_signif_SNP_names_2 %in% assoc_signif_SNP_names) #how many were also identified by assoc?

#comparing analysis with and without PCs as covars
#following SNPs were "significant" @ 10^-6 in both logistic analyses
signif_SNP_both<-intersect(logistic_signif_SNP_names_2,logistic_signif_SNP_names)
logistic_signif_SNPs_2$signif_in_original_logistic<-logistic_signif_SNPs_2$SNP %in% signif_SNP_both
#adjusting for PCs removed none of the original top SNPs
setdiff(logistic_signif_SNP_names,logistic_signif_SNP_names_2)
#get additional information
bps<-read.table('data/CAMPdata2_pc_covar.assoc.logistic',header=TRUE)
logistic_signif_SNPs_2<-join(bps,logistic_signif_SNPs_2,by=c("CHR","SNP"),"right")
logistic_signif_SNPs_2<-logistic_signif_SNPs_2[logistic_signif_SNPs_2$TEST=="ADD",]
#write out summary of logistic model with PC covars
write.csv(logistic_signif_SNPs_2,file="top_SNPs_logistic_pc.csv",row.names=FALSE)
#write out list of top SNPs for Plink to generate .PED file for haploview
write.table(logistic_signif_SNP_names_2,file="topsnps.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)
#write out .info file for haploview
write.table(logistic_signif_SNPs_2[,c("SNP","BP")],file="topsnps.info",row.names=FALSE,col.names=FALSE,quote=FALSE)
```

Appendix 3 - Tables and Figures

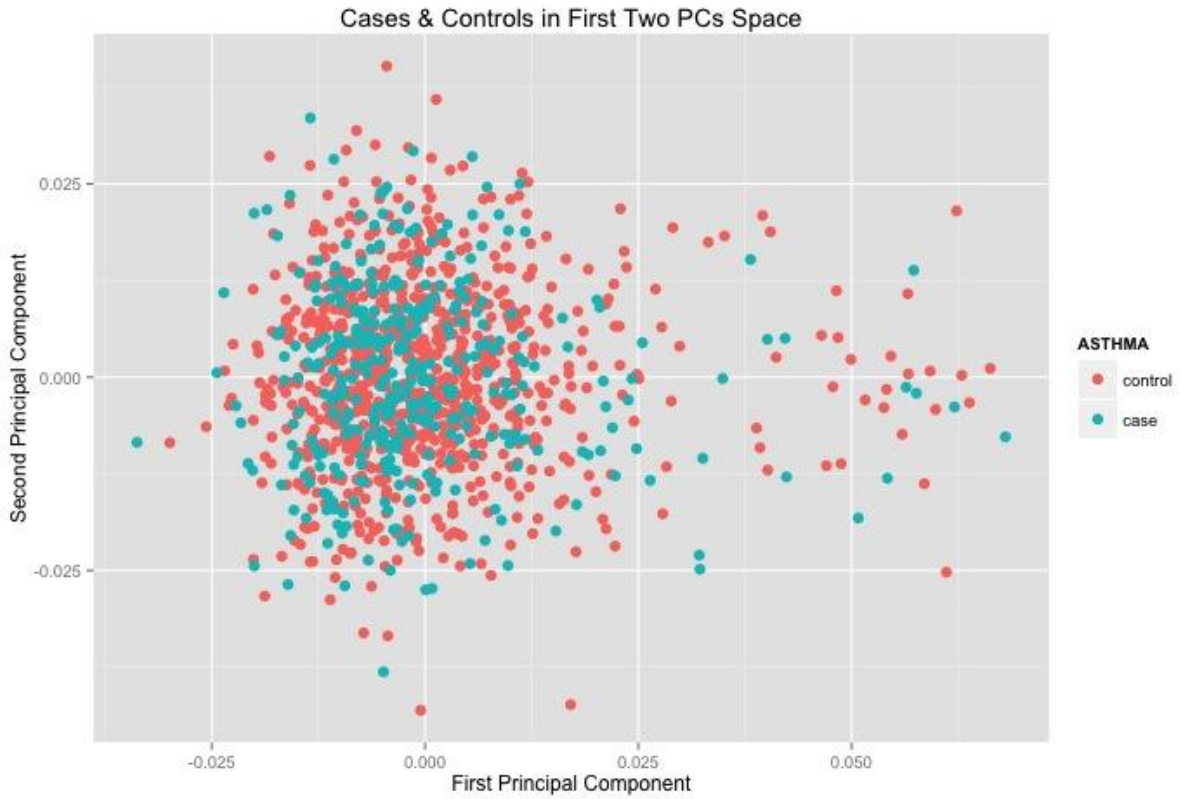

Figure 1. Cases versus controls in the first two principal components space

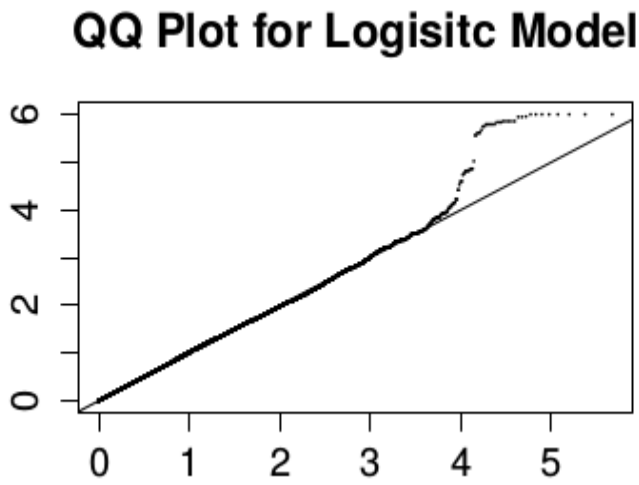

Figure 2. Q-Q plot of allelic-associated p-value for logistic model adjusting for gender

	ASTHMA	

GENDER	control	case
male	26%	18%
female	45%	11%

Table 1. 2 x 2 table of asthma and gender distributions

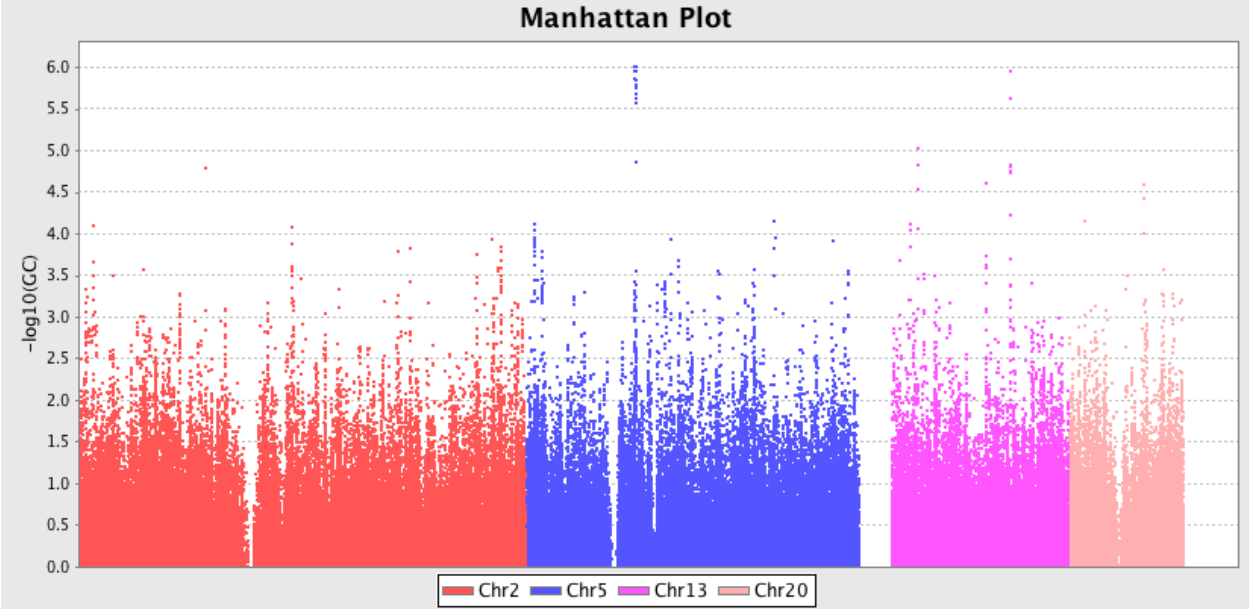

Figure 3. Manhattan Plot

CHR	SNP	location (bp)	A1	OR	P
5	rs7731007	59399588	G	0.5893	7.39E-07
5	rs1508859	59389854	T	0.5893	7.39E-07
5	rs1508864	59388568	C	0.5893	7.39E-07
5	rs2662444	59429941	G	0.5893	7.39E-07
5	rs1100918	59422372	G	0.5893	7.39E-07
5	rs13164971	59414410	C	0.5893	7.39E-07
5	rs10461667	59404215	C	0.5893	7.39E-07
5	rs1588265	59405551	G	0.5893	7.39E-07
5	rs2136203	59418081	C	0.5907	8.22E-07
13	rs9546395	82927920	C	1.626	8.37E-07
5	rs4700355	59368006	G	0.5896	8.38E-07

Table 2. Significant SNPs

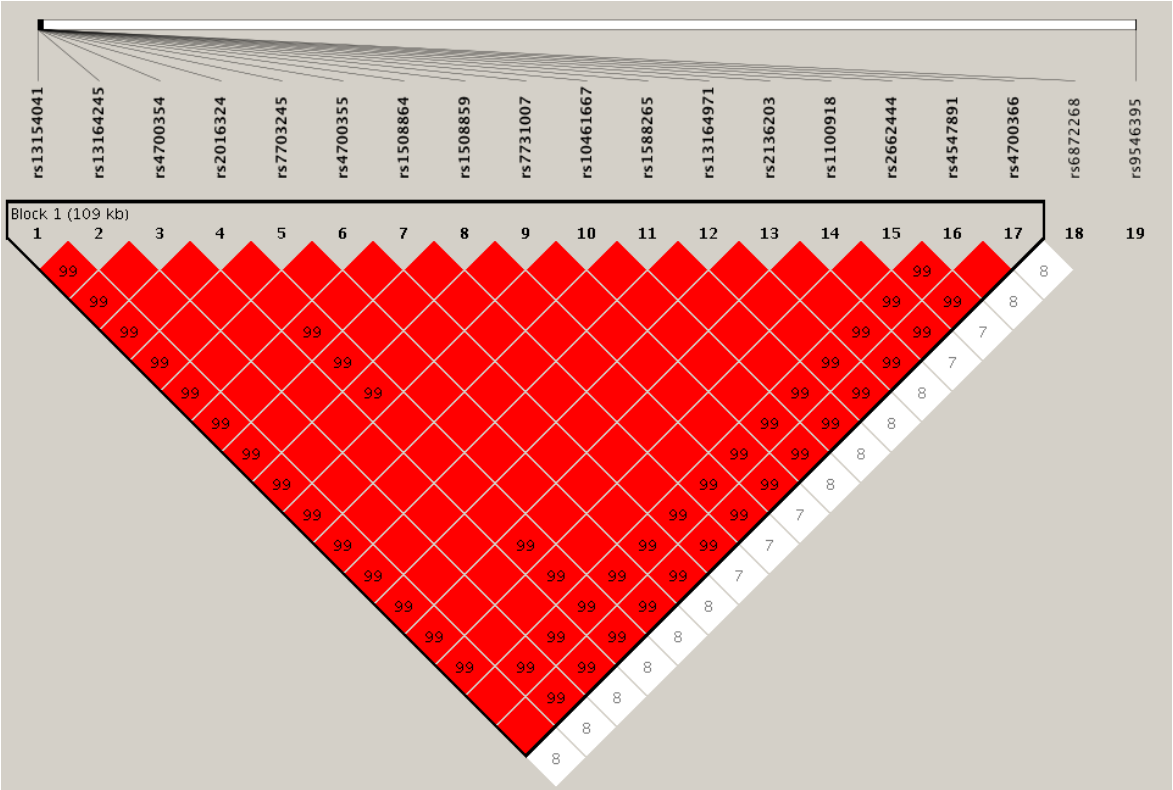

Figure 4. Linkage disequilibrium plot

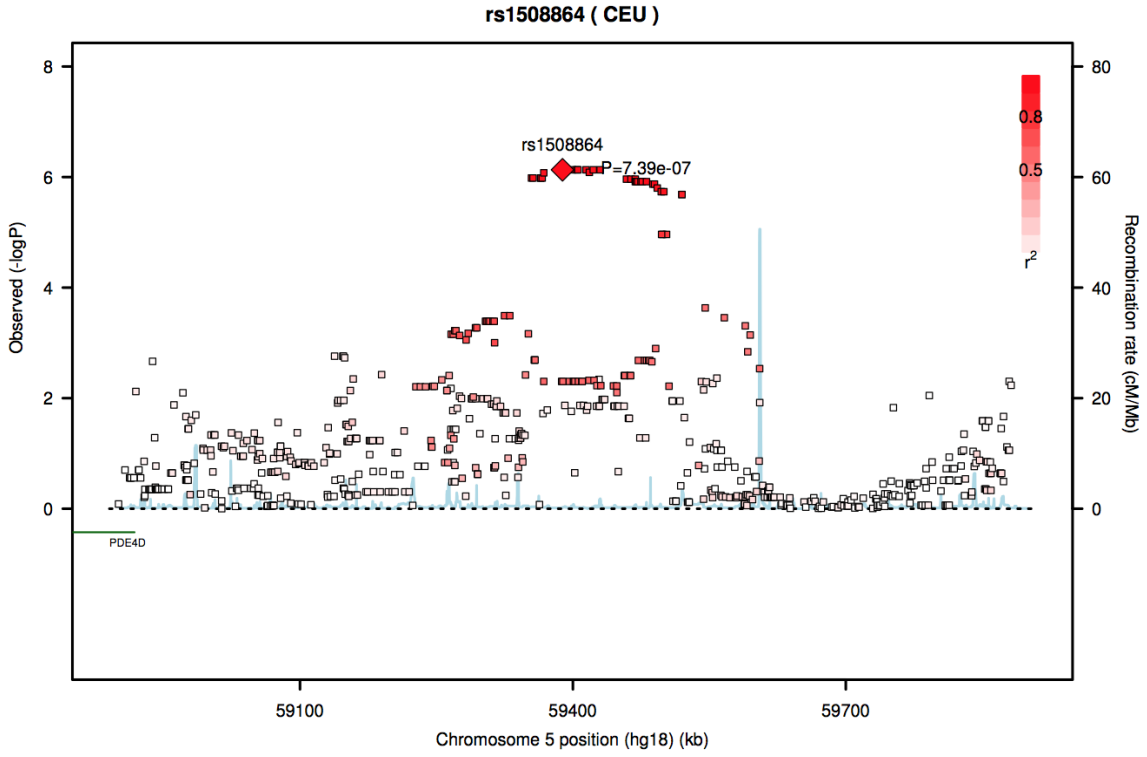

Figure 5. Regional plot of rs1508864

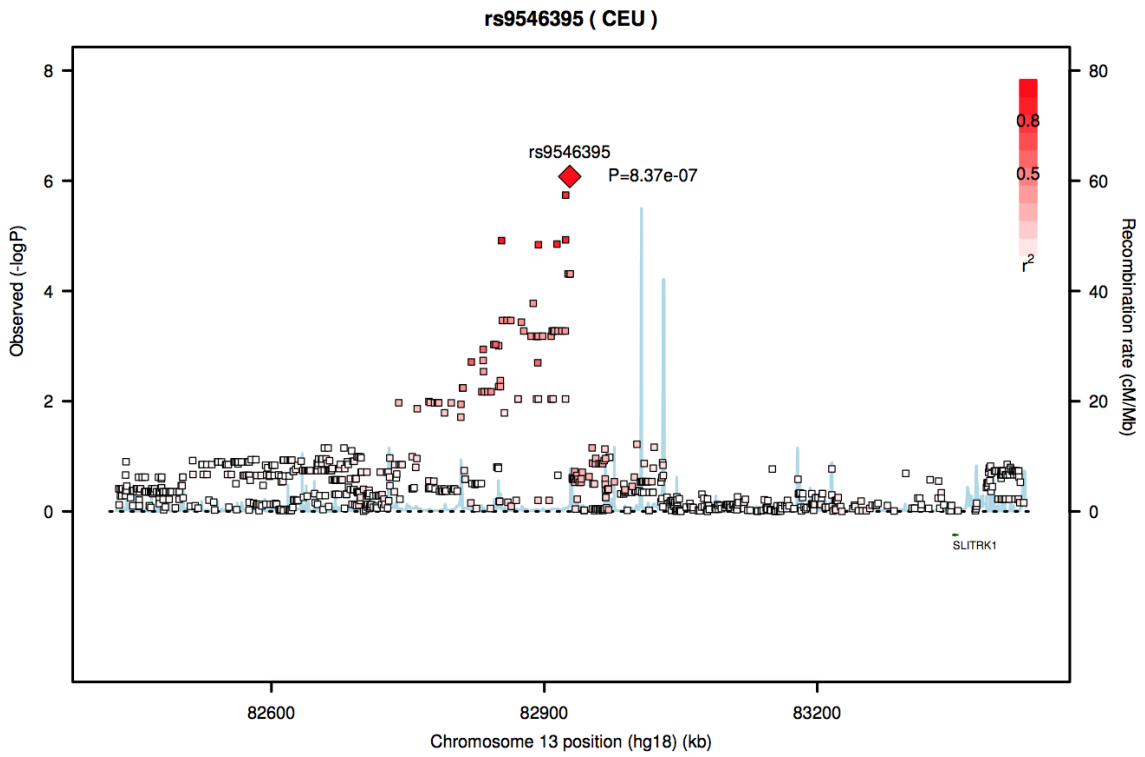

Figure 6. Regional plot of rs9546395