Investigation of Gene x Sex Interaction Effects on Idiopathic Pulmonary Fibrosis

Adelle Price

Statistical Methods for Genetic Association Studies

May 20, 2022

Introduction

Idiopathic Interstitial Fibrosis (IPF) is an illness characterized by lung tissue scarring and a progressive lack of oxygen. Median survival rate of IPF is 2-3 years. As of 2013, there were no drugs approved for treatment of IPF in the US; a complete lung transplant was the only option. As of 2015, two drugs were approved for treatment of IPF- but each is only moderately effective.[2] In recent years, several variants have been identified via GWAS to be significantly associated with IPF risk.[1] It is also known that IPF risk is approximately 70% higher for males compared to females.[3] This project investigates whether there are observable variant x sex interactions that significantly modify IPF risk.

Methods

Data

The data used in this project included genotype data from 8059 subjects, 2555 SNPs in total (all on chromosome 11), as well as phenotype data containing the same 8059 subjects' IPF case/control status and sex (male/female). The data went through the following quality control steps before analysis:

- Confirming that reported subject sex corresponds to genetic sex.

- Only subjects with European ancestry included in the data.

- Data filtered to include only variants in Hardy-Weinberg equilibrium in controls ($p > 0.0001$).

- Data filtered to remove SNPs with call rates < 95% and/or different missingness between cases and controls.

- Data filtered to remove all subjects related at a second degree or more.

- Data filtered to include only subjects with < 1% genotype missingness.

- Data filtered to include SNPs with MAF > 3%.

An additive genetic model was assumed for all analyses, and thus the genotypes were coded as 0,1,2 where addition of 1 corresponds to the addition of 1 minor allele in an individual's genotype.

Case/Control Analysis

A separate logistic regression model was used to quantify and assess the significance of interactions between SNP and Sex for each SNP. Each logistic regression model had the following structure:

$$\text{logit}\left[\frac{\text{Case}}{\text{Control}}\right] = \beta_0 + \beta_1 Sex + \beta_2 SNP + \beta_3 (Sex * SNP)$$

An important assumption of the case/control model was that the samples included in the model (individual subjects) were independent and therefore unrelated.

The significance threshold for each logistic regression model was determined to be 1.95e-05 after a Bonferroni Correction for 2555 tests (1 for each SNP).

Case Only Analysis

To set up Case only analysis the data was filtered to only include individuals classified as 'Cases' with regards to IPF. For this method, a separate logistic regression model was built for each SNP, regressing Sex (Male, Female) on SNP:

$$\text{logit}[\text{Sex}] = \beta_0 + \beta_1 * SNP$$

If a SNP and Sex relationship is shown to be statistically significant in the case only model, we conclude that this indicates a significant SNP x Sex interaction with regards to IPF risk. This conclusion hinges on the assumption that the relationship between SNP and Sex is independent in the individuals classified as 'Controls' with regards to IPF.[4]

The significance threshold for each logistic regression model was determined to be 1.95e-05 after a Bonferroni Correction for 2555 tests (1 for each SNP).

Checking Model Assumptions

In the Case/Control method, the assumption that the samples included in the model was addressed in the data quality control portion of the project. The data was filtered to only include unrelated individuals where unrelated was defined as third degree related or less.

In the Case only method, the assumption that the relationship between SNP and Sex is independent in the Controls was assessed by creating a logistic regression model for each of the SNPs- using only the individuals classified as 'Controls' with regards to IPF. Each model had the following structure:

$$\text{logit}[\text{Sex}] = \beta_0 + \beta_1 * SNP$$

If a statistically significant relationship between SNP and Sex was found using this model, our assumptions for the Case Only analysis would not be met and would not be appropriate to use for this project.

Software and Packages

Plink software was used to perform all filter and QC portions of the project.

R software was used to complete all data analysis. Specific R packages/functions used include:

snpStats package for processing genotype data in R,

glm() function for building and analyzing all logistic regression models,

dplyr() package for data manipulation.

Data Exploration

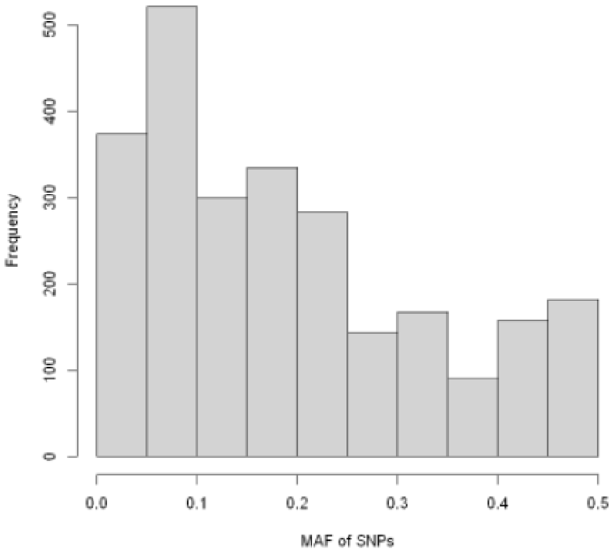

Figure 1:

A histogram showing the distribution of minor allele frequencies in the genotype data. All SNPs included in project had a MAF > 3%.

	Case (N=3617)	Control (N=4442)	Total (N=8059)
Female	1033(29%)	2610(59%)	3643 (45%)
Male	2584(71%)	1832(41%)	4416 (55%)

Figure 2:

A table showing the number of female cases and controls, the number male cases and controls, and the respective proportions of each.

Results

The results from the Case/Control analysis are shown below. The smallest p-value depicted in the figure below, **Figure 3**, is .003.

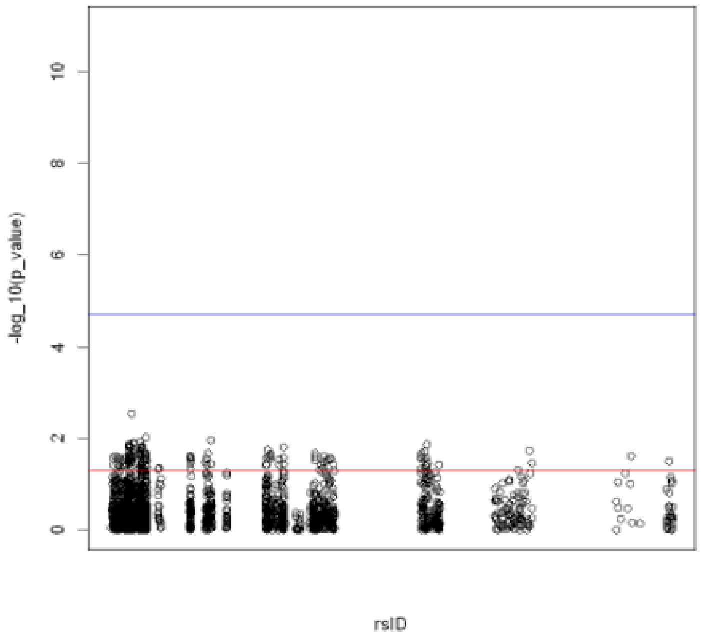

Figure 3:

A Manhattan plot showing the the -log_10(p-value) of the SNP x Sex interaction term from each Case/Control logistic regression model. The x-axis corresponds to the rsID for the SNPs in each SNP x Sex interaction term. The y-axis corresponds the -log_10(p-value).

The blue line corresponds to the threshold of statistical significance (p-value=1.96e-05 after a Bonferroni correction). The red line corresponds to nominal statistical significance(p-value=.05)

The results from the Case only analysis are shown below. The smallest p-value depicted in the figure below, **Figure 4**, is .003.

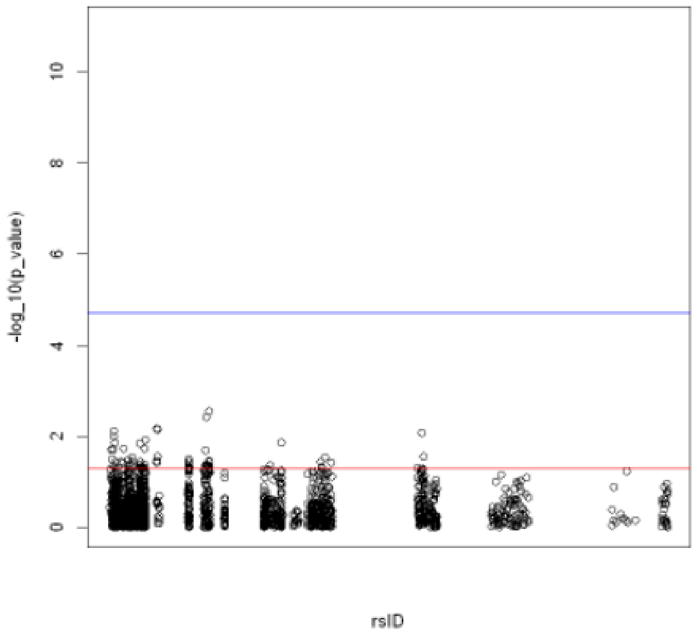

Figure 4:

A Manhattan plot showing the the -log_10(p-value) of the SNP beta coefficient from each Case only logistic regression model. The x-axis corresponds to the rsID for the SNPs in each SNP beta coefficient. The y-axis corresponds the -log_10(p-value).

The blue line corresponds to the threshold of statistical significance (p-value=1.96e-05 after a Bonferroni correction). The red line corresponds to nominal statistical significance(p-value=.05)

The results from the Control only analysis (checking model assumptions of Case only analysis) are shown below. The smallest p-value depicted in the figure below, **Figure 5**, is 0.002.

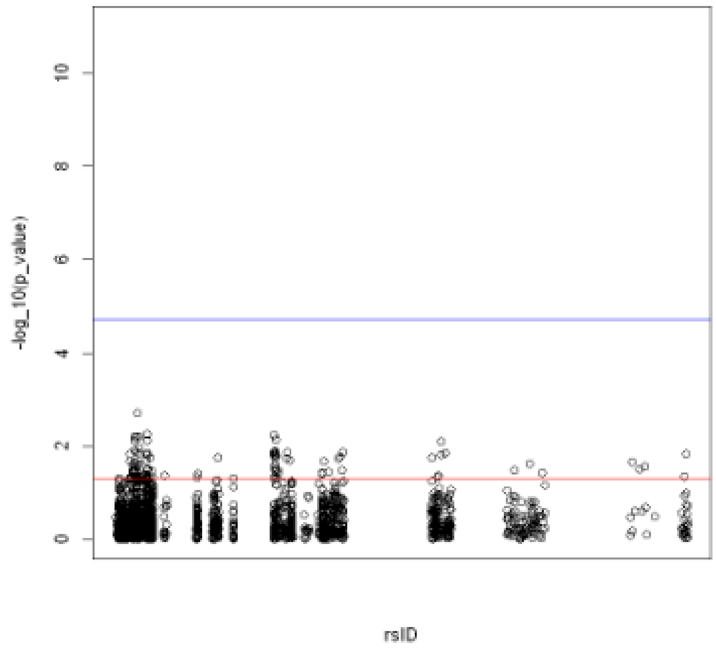

Figure 5:

A Manhattan plot showing the the -log_10(p-value) of the SNP beta coefficient from each Control only logistic regression model. The x-axis corresponds to the rsID for the SNPs in each SNP beta coefficient. The y-axis corresponds the -log_10(p-value).

The blue line corresponds to the threshold of statistical significance (p-value=1.96e-05 after a Bonferroni correction). The red line corresponds to nominal statistical significance(p-value=.05)

	Estimate	95% Confidence Interval	P-value
Intercept	-0.64	(-1.12, -0.18)	.007
rsID7396026	-0.14	(-0.39, 0.10)	.231
rsID7396026:Sex(Male)	0.48	(0.16, 0.79)	.003
Sex (Male)	0.36	(-0.24, 0.97)	.244

Figure 6:

A summary of the Case/Control analysis logistic regression model that includes the topmost nominally significant SNP x Sex interaction.

	Estimate	95% Confidence Interval	P-value
Intercept	0.28	(-0.14, 0.71)	.190
rsID35678986	0.34	(0.12, 0.55)	.003

Figure 7:

A summary of the Case only analysis logistic regression model that includes the topmost nominally significant SNP beta coefficient.

	Estimate	95% Confidence Interval	P-value
Intercept	0.31	(-0.11, 0.73)	.151
rsID34830395	0.32	(0.10, 0.53)	.004

Figure 8:

A summary of the Case only analysis logistic regression model that includes the second topmost nominally significant SNP beta coefficient.

Discussion

As shown in **Figure 3**, the Case/Control analysis did not yield any statistically significant interactions between individual SNPs and Sex (with a significance threshold of 1.56e-05). The analysis did, however, yield 131 nominally significant interactions between individual SNPs and Sex. The topmost significant interaction will be discussed here:

The topmost significant interaction from Case/Control analysis was between SNP rs7396026 and Sex. This interaction had a p-value of approximately .003 and a point estimate of approximately 0.47 (CI 0.16, 0.79). Interpretation of this SNP x Sex interaction with regards to IPF risk is as follows:

The addition of a risk allele at SNP rs7396026 in men corresponds to approximately a 2.31 times increase in the odds of IPF compared the addition of a risk allele at SNP rs7396026 in women.

SNP rs7396026 is located in an intergenic region in chromosome 11. It is possible that this SNP could perhaps play a regulatory role in transcription.[5]

As shown in **Figure 5**, the Control only analysis did not yield any statistically significant relationships between individual SNPs and Sex. This confirmed it was appropriate to proceed with the Case only analysis and interpretation.

As shown in **Figure 4**, the Case only analysis also did not yield any statistically significant relationships between individual SNPs and Sex (with a significance threshold of 1.56e-05). The analysis did, however, yield 65 nominally significant interactions between individual SNPs and Sex. The two topmost significant interaction will be discussed here:

The topmost significant interaction from Case only analysis was between SNP rs35678986 and Sex. This interaction had a p-value of approximately .003 and a point estimate of approximately 0.34 (CI 0.12, 0.55). Interpretation of this SNP x Sex interaction with regards to IPF risk is as follows:

The addition of a risk allele at SNP rs35678986 in men corresponds to approximately a 1.40 times increase in the odds of IPF compared the addition of a risk allele at SNP rs35678986 in women.

SNP rs35678986 is located in the KRTAP5-4 gene (keratin associated protein) on chromosome 11. SNP rs35678986 is known as a 3' UTR variant (UTR is a part of mRNA). Mutations in this region of the chromosome are known to be high consequence for protein structure and function.[5]

The second topmost significant interaction from Case only analysis was between SNP rsID34830395 and Sex. This interaction had a p-value of approximately .004 and a point estimate of approximately 0.32 (CI 0.10, 0.53). Interpretation of this SNP x Sex interaction with regards to IPF risk is as follows:

The addition of a risk allele at SNP rs34830395 in men corresponds to approximately a 1.38 times increase in the odds of IPF compared the addition of a risk allele at SNP rs34830395 in women.

SNP rs34830395 is located in the intron of the MOB2 gene on chromosome 11; MOB2 codes for a kinase activator. SNP34830395 is likely a regulatory of MOB2 transcription.[5]

Conclusion

Based on the results of the various analyses, we conclude there is no strong evidence for interaction between Sex and SNPs with regards to IPF risk. There were, however, 196 nominally significant SNP x Sex interactions observed (p-values <.05). These results are consistent with the SNP x Sex analysis done in Fingerlin et al. which also found no strong evidence for interaction between Sex and SNPs with regards to IPF risk.

Future Investigation

To improve study design for any future studies, it would be prudent to ensure the number of male and female subjects classified as 'Cases" are equivalent as well as the number of male and female subjects classified as 'Controls'.

It is also important to perform investigation of gene x sex interaction in regards to IPF risk for ethnic groups other than European. Various ethnic groups are known to have varying IPF risk and this may give insight into the question of possible gene x sex interaction effects.[6]

References

1. Fingerlin, T., Murphy, E., Zhang, W. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* **45,** 613–620 (2013). https://doi.org/10.1038/ng.2609

2. Kropski JA, Blackwell TS, Loyd JE. The genetic basis of idiopathic pulmonary fibrosis. Eur Respir J. 2015 Jun;45(6):1717-27. doi: 10.1183/09031936.00163814. Epub 2015 Apr 2. PMID: 25837031; PMCID: PMC4849867.

3. Kropski JA, Blackwell TS, Loyd JE. The genetic basis of idiopathic pulmonary fibrosis. Eur Respir J. 2015 Jun;45(6):1717-27. doi: 10.1183/09031936.00163814. Epub 2015 Apr 2. PMID: 25837031; PMCID: PMC4849867.

4. Liu, Chen-yu et al. “Design and analysis issues in gene and environment studies.” *Environmental health : a global access science source* vol. 11 93. 19 Dec. 2012, doi:10.1186/1476-069X-11-93

5. uswest.ensembl.org

6. Swigris JJ, Olson AL, Huie TJ, et al. Ethnic and racial differences in the presence of idiopathic pulmonary fibrosis at death. *Respir Med*. 2012;106(4):588-593. doi:10.1016/j.rmed.2012.01.002

R code Appendix

```r
#Read in phenotype data
phen <- read.csv("C:/Users/adell/Desktop/adelle_price/adelle_price/phenotype.csv", stringsAsFactors=T)
#Make case/control data binary
phen$case.control <- ifelse(phen$phenotype == "case" , 1, 0)
#Read in genotype data as numeric (automatically codes for additive genotype assumption)
suppressPackageStartupMessages(library(snpStats))
bed.fn <- "C:/Users/adell/Desktop/adelle_price/adelle_price/common_11p15.bed"
fam.fn <- "C:/Users/adell/Desktop/adelle_price/adelle_price/common_11p15.fam"
bim.fn <- "C:/Users/adell/Desktop/adelle_price/adelle_price/common_11p15.bim"
gens <- read.plink(bed.fn, bim.fn, fam.fn)
geno <- as(gens$genotypes, "numeric")
#plot MAF of SNP data
Sampstats <- row.summary(gens$genotypes)
SNPstats <- col.summary(gens$genotypes)
hist(SNPstats$MAF, xlab = "MAF of SNPs")
#Make data frame for case/control analysis
full <- cbind.data.frame("ID" = phen$id, "Sex" = phen$SEX, "case.control" = phen$case.control, geno)

#Perform case/control analysis for each SNP
genes_length <- dim(geno)[2]
all_names <- colnames(full[,4:2555])
case_control = data.frame(matrix(, nrow=0, ncol=6))
```

```
for (i in 1:genes_length){
    Object <- glm(formula = case.control ~ full[[as.character(all_names[i])]] * Sex, data = full, family='binomial')
    names(Object$coefficients) = gsub("full[[as.character(all_names[i])]]", paste(all_names[i], "_"), names(Object$coefficients), fixed = TRUE)
    names(Object$coefficients) = gsub("(Intercept)", paste(all_names[i], "_(Intercept)"), names(Object$coefficients), fixed = TRUE)
    names(Object$coefficients) = gsub("SexMale", paste(all_names[i], "_SexMale"), names(Object$coefficients), fixed = TRUE)
    k = summary(Object)$coefficients
    f = names(Object$coefficients)
    z = cbind(k, "names" = rownames(k))
    f <- as.data.frame(f)
    data_all <- as.data.frame(merge(z, f, by.x = "names", by.y = "f", all = TRUE))
    case_control= rbind.data.frame(case_control, data_all)


  }



#Prepare case/control analysis results for Manhattan plot of significance per each interaction coefficient estimate p-value
Significant = case_control
Sig_Int <- Significant[grep("Intercept", Significant$names), ]
Sig_coef <- subset(Significant, !(rownames(Significant) %in% rownames(Sig_Int)))
Sig_coef <- na.omit(Sig_coef)
Sig_Marg <- Sig_coef[grep(":rs", Sig_coef$names), ]
Sig_Marg[,5] <- as.numeric(Sig_Marg[,5])
Sig_Marg$log_pval = -(log10(Sig_Marg[,5]))
```

```
suppressPackageStartupMessages(library(dplyr))

Sig_Marg <- Sig_Marg %>%

  mutate(ID = row_number())

Sig_Marg$names <- gsub("(:).*", "\\1", Sig_Marg$names)

Sig_Marg$names <- gsub(":", "", Sig_Marg$names)

Sig_Marg$rsid <- Sig_Marg$names

Sig_Marg$rsid <- gsub("(_).*", "\\1", Sig_Marg$rsid)

Sig_Marg$rsid <- gsub("_", "", Sig_Marg$rsid)

Sig_Marg$rsid <- gsub("rs", "", Sig_Marg$rsid)

Sig_Marg$rsid <- as.numeric(Sig_Marg$rsid)

Sig_Marg_o <- Sig_Marg[order(Sig_Marg$rsid),]



#Manhattan Plot of -log_10(pvals) of all SNP x Sex interaction terms from Case/Control analysis

plot(Sig_Marg_o$rsid, Sig_Marg_o$log_pval, xaxt = "n", xlab = "rsID", ylab = "-log_10(p_value)", ylim = c(0,11))

abline(h=(-(log10(.05))), col = "red")

abline(h=(-(log10(.05/genes_length))), col = "blue")


#Get SNPxSex interaction with highest significance from Case/Control analysis

Sig_Marg_p <- Sig_Marg[order(Sig_Marg[,5]),]

Sig_Marg_p[1,]
```

```
#Perform case only analysis for each SNP
genes_length <- dim(geno)[2]
full_case = full[full$case.control == 1,]
all_names <- colnames(full_case[,4:2555])
case_only = data.frame(matrix(, nrow=0, ncol=6))
for (i in 1:genes_length){
    Object <- glm(Sex ~ full_case[[as.character(all_names[i])]], data=full_case, family=binomial)
    names(Object$coefficients) = gsub("full_case[[as.character(all_names[i])]]", paste(all_names[i], "_"), names(Object$coefficients), fixed = TRUE)
    names(Object$coefficients) = gsub("(Intercept)", paste(all_names[i], "_(Intercept)"), names(Object$coefficients), fixed = TRUE)
    k = summary(Object)$coefficients
    f = names(Object$coefficients)
    z = cbind(k, "names" = rownames(k))
    f <- as.data.frame(f)
    data_all <- as.data.frame(merge(z, f, by.x = "names", by.y = "f", all = TRUE))
    case_only= rbind.data.frame(case_only, data_all)

   }

#Prepare case only analysis results for Manhattan plot of significance per each interaction coefficient estimate p-value
Significant = case_only
Sig_Int <- Significant[grep("Intercept", Significant$names), ]
Sig_coef <- subset(Significant, !(rownames(Significant) %in% rownames(Sig_Int)))
Sig_Marg_c <- na.omit(Sig_coef)
Sig_Marg_c[,5] <- as.numeric(Sig_Marg_c[,5])
```

```
Sig_Marg_c$log_pval = -(log10(Sig_Marg_c[,5]))
Sig_Marg_c$rsid <- Sig_Marg_c$names
Sig_Marg_c$rsid <- gsub("(_).*", "\\1", Sig_Marg_c$rsid)
Sig_Marg_c$rsid <- gsub("_", "", Sig_Marg_c$rsid)
Sig_Marg_c$rsid <- gsub("rs", "", Sig_Marg_c$rsid)
Sig_Marg_c$rsid <- gsub("chr11:", "", Sig_Marg_c$rsid)
Sig_Marg_c$rsid <- as.numeric(Sig_Marg_c$rsid)
Sig_Marg_o <- Sig_Marg_c[order(Sig_Marg_c$rsid),]

#Manhattan Plot of -log_10(pvals) of all SNP x Sex interaction terms from Case only analysis
plot(Sig_Marg_o$rsid, Sig_Marg_o$log_pval, xaxt = "n", xlab = "rsID", ylab = "-log_10(p_value)", ylim = c(0,11))
abline(h=(-(log10(.05))), col = "red")
abline(h=(-(log10(.05/genes_length))), col = "blue")

#Get two SNPxSex interactions with highest significance from Case only analysis
Sig_Marg_p <- Sig_Marg_c[order(Sig_Marg_c[,5]),]
Sig_Marg_p[1:2,]
```

```
##Perform control only analysis for each SNP
genes_length <- dim(geno)[2]
full_control = full[full$case.control == 0,]
all_names <- colnames(full_control[,4:2555])
control_only = data.frame(matrix(, nrow=0, ncol=6))



for (i in 1:genes_length){
	Object <- glm(Sex ~ full_control[[as.character(all_names[i])]], data=full_control, family=binomial)
	names(Object$coefficients) = gsub("full_control[[as.character(all_names[i])]]", paste(all_names[i], "_"), names(Object$coefficients), fixed = TRUE)
	k = summary(Object)$coefficients
	f = names(Object$coefficients)
	z = cbind(k, "names" = rownames(k))
	f <- as.data.frame(f)
	data_all <- as.data.frame(merge(z, f, by.x = "names", by.y = "f", all = TRUE))
	control_only= rbind.data.frame(control_only, data_all)

      }
#Prepare control only analysis results for Manhattan plot of significance per each interaction coefficient estimate p-value
Significant = control_only
Sig_Int <- Significant[grep("rs", Significant$names), ]
Sig_Marg_con <- Sig_Int[order(Sig_Int[,5]),]
Sig_Marg_con[,5] <- as.numeric(Sig_Marg_con[,5])
Sig_Marg_con$log_pval = -(log10(Sig_Marg_con[,5]))
```

```
Sig_Marg_con$rsid <- Sig_Marg_con$names
Sig_Marg_con$rsid <- gsub("(_).*", "\\1", Sig_Marg_con$rsid)
Sig_Marg_con$rsid <- gsub("_", "", Sig_Marg_con$rsid)
Sig_Marg_con$rsid <- gsub("rs", "", Sig_Marg_con$rsid)
Sig_Marg_con$rsid <- gsub("chr11:", "", Sig_Marg_con$rsid)
Sig_Marg_con$rsid <- as.numeric(Sig_Marg_con$rsid)
Sig_Marg_o <- Sig_Marg_con[order(Sig_Marg_con$rsid),]
plot(Sig_Marg_o$rsid, Sig_Marg_o$log_pval, xaxt = "n", xlab = "rsID", ylab = "-log_10(p_value)", ylim = c(0,11))
abline(h=(-(log10(.05))), col = "red")
abline(h=(-(log10(.05/genes_length))), col = "blue")

#Get the coefficient estimates and confidence intervals for the Case/Control analysis top significance SNPxSex interaction
snp <- full$rs7396026
sex <- full$Sex
case.control <- full$case.control
dat = cbind.data.frame(sex,case.control,snp)
Object <- glm(formula = case.control ~ snp * sex, data = dat, family='binomial')
summary(Object)
confint(Object)
#Get the coefficient estimates and confidence intervals for the Case only analysis top two significance SNPxSex interaction
full <- full[full$case.control == "1",]
snp <- full$rs34830395
sex <- full$Sex
case.control <- full$case.control
dat = cbind.data.frame(sex,case.control,snp)
```

```
Object <- glm(formula = sex~snp, data = dat, family='binomial')

summary(Object)

confint(Object)

snp <- full$rs35678986

sex <- full$Sex

case.control <- full$case.control

dat = cbind.data.frame(sex,case.control,snp)

Object <- glm(formula = sex~snp, data = dat, family='binomial')

summary(Object)

confint(Object)
```