LONGITUDINAL ANALYSIS OF GENETIC RISK FACTORS OF CARDIOVASCULAR DISEASE

by

Victoria D. Causer

B.S Applied Math, Biological Science, Chemistry, University of Pittsburgh- Greensburg, 2010

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Victoria D. Causer

It was defended on

July 28, 2014

and approved by

Thesis Advisor: Lisa Weissfeld PhD, Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Joyce Chang PhD, Professor, Biostatistics, Medicine, Clinical and Translational Science, Graduate School of Public Health, School of Medicine University of Pittsburgh

Committee Member: Indrani Halder PhD, Assistant Professor, Medicine, School of Medicine, University of Pittsburgh

LONGITUDINAL ANALYSIS OF GENETIC RISK FACTORS OF CARDIOVASCULAR DISEASE

Victoria D. Causer, M.S.

University of Pittsburgh, 2014

ABSTRACT

Cardiovascular disease (CVD) continues to be the leading cause of death in the United States. For this reason, CVD and CVD risk factor prevention poses high public health significance, due to its prevalence and financial burden on society. CVD and risk factors prevalence as well as genetic structures are known to be different in White and African American populations. This indicates that genetic differences could be responsible for differences in disease prevalence. The purpose of this study was to examine if genotype was a significant predictor of CVD risk factor measurements over time using mixed modeling and trajectory group analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 CVD RISK FACTORS AND STATISTICS

Cardiovascular disease (CVD) continues to be the leading cause of death for men and women in the United States, consisting of nearly every 1 in every 4 deaths (Murphy, 2013). Coronary heart disease, the most common type, contributes to nearly $108.9 billion in yearly costs to the United States (Heidenreich, 2011). For this reason, CVD and CVD risk factor prevention poses high public health significance. Americans with high levels of the metabolic syndrome components are at high risk for CVD. Metabolic syndrome is defined by the NIH as a group of risk factors that raises your risk for heart disease and other health problems (NIH). These Metabolic Syndrome risk factors include high blood pressure, high cholesterol, high glucose, high triglyceride, and high BMI. The CDC recognizes High blood pressure, high LDL cholesterol, and smoking as key risk factors for heart disease, with about 49% of the population having at least one of the key factors (CDC, 2011). CVD and risk factors prevalence were also found to be different in White and Black populations. For example, CVD age-adjusted death rates are 33% higher for blacks than for the overall population and the black population also has a higher prevalence of high blood pressure. This has been previously contributed to lack of health insurance and limited access to quality health care (AHA, 2014). More recently, these differences have also been found to be partially genetic (Halder et al., 2008, 2012).

1.2 GENETICS AND PREVIOUS STUDIES

Black and white populations are also known to have different gene structures. This is seem both within our dataset but also validated in publically available databases. With this, it is possible that genetic differences could also be partially responsible for differences in disease and risk factor prevalence. Although blood pressure is thought to be considered highly heritable, only about 0.9% of phenotypic variance has been identified. This is thought to be due to specific interactions with inflammation, blood coagulation, cellular adhesion molecules, and lipid metabolism (El Shamieh, 2012). Multiple loci on chromosome 1 have been previously reported to be linked to blood pressure phenotypes, with fine locus mapping tracing the variation to ATP1B1, RGS5 and SELE genes. Polymorphisms in the SELE gene found to be associated with both SBP and DBP (Faruque, 2011). Although prior analysis has been performed to find association between inflammation genes and blood pressure measurements, to our knowledge, nothing has been done to look at the longitudinal effects of genotype on blood pressure.

1.3 DATA DESCRIPTION

For this study, data was collected longitudinally as a part of the University of Pittsburgh HeartSCORE Study. HeartSCORE was instituted in 2003 with hopes to explain racial and socioeconomic disparities in cardiovascular risk. Study subjects included 771 Whites and 464 African Americans, between 45-74 years in age, and represented all Framingham risk strata. The study population was made up of approximately 64% females. The Illumina iSelect IBC Chip was used to genotype approximately 49k single nucleotide polymorphisms (SNPs) covering

approximately 2100 genes for each individual enrolled in the study. A subset containing 400 single nucleotide polymorphisms (SNPs), in 8 inflammation and 8 serotonin pathway genes, were examined specifically. These included both known functional variants and tag SNPs with no known functions. Metabolic risk factor measurements, systolic blood pressure (SBP), diastolic blood pressure (DBP), (high-density lipoprotein) HDL and (low-density lipoprotein) LDL Cholesterol, Triglycerides, and (body mass index) BMI measurements, as well as psychological risk factors, such as anxiety and depression (CES-D), were collected yearly. Risk factors of CVD were then examined longitudinally using group based trajectory modeling was used using the TRAJ procedure, developed by Jones, Nagin, and Roeder of CMU. All analysis was performed in SAS v. 9.3 and plink v. 1.07.

1.4 OBJECTIVE

The objective of this thesis was to examine if relationships exist between SNPs, located within the Selectin-E (SELE) gene, and longitudinal measurements blood pressure.

2.0 METHODOLOGY

2.1 EXPLORATORY/UNIVARIATE ANALYSIS

Analysis was performed separately by race. CES-D score, cholesterol levels, HDL levels, and triglyceride levels were all log-transformed to help with normalization prior to analysis. Univariate analysis was performed using ANOVAs to compare means at baseline for each outcome variable (diastolic BP, systolic BP, BMI, ln(CESD score), ln(cholesterol), ln(HDL), and ln(triglyceride levels) by genotype for all SNPs (280 inflammation and 129 serotonin SNPs). ANOVAs allow for the generalization of the t-test to more than two groups. The null hypothesis for the ANOVA is that the group means do not differ between groups. This test is based on the F statistics which is defined by:

$$F = \frac{\text{Between group variance}}{\text{Within group Variance}}$$

Since the grouping variable variance is in the numerator, the larger the between group variance, the larger the F-statistics indicating a greater likelihood that differences in mean are not due to chance (Agresti, 2007).

2.2 LINEAR REGRESSION

Linear regression is often used when examining a continuous outcome with a number of independent predictors. Specific assumptions must meet in order for linear regression to be a valid method for data analysis. These include that there is an underlying linear relationship between the dependent (outcome) and independent (predictors) variables, errors resulting from model fitting must be uncorrelated, independent, and normally distributed, and that equal variances are present (homoscedasticity) (Chapman, 2001).

Linear regression models can be represented in the following general matrix form:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

Where **y** is an n x 1 vector of the outcome variable; **X** is an n x p matrix of the p predictor variables; **β** is a p x 1 column vector of the regression coefficients, and **ε** is the error (Chapman, 2001).

There are various model selection methods which allow for the determination of the inclusion of covariates; backwards, forwards, and stepwise. The backwards method involves the initial fitting of the full model with removal of variables sequentially, based on a cutoff p-value selected a priori, until all variables remain significant. The forwards method involves the addition of variables one at a time sequentially into the model. If the variable is significant it remains in the model, it the variable fails to reach the significant p-value it is removed. Lastly, the stepwise method allows the addition and elimination of variables in the model, dropping or adding variables at the various steps (Chapman, 2001).

For this project, covariates were preselected based on literature and backward selection methodology. Age, sex, socio-economic status measurement composed of education and income, anti-hypertensive medication usage, time, and genotype (predictor of interest) were included in each model as well as. BMI was also included as a covariate for blood pressure variables.

2.3 LOGISTIC REGRESSION

Logistic regression is often used when examining a binary outcome with a number of independent predictors. The general form of logistic regression model is

$$logit(P(y)) = ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \boldsymbol{X\beta}$$

Where

$$p(X) = \frac{e^{\boldsymbol{X\beta}}}{1+e^{\boldsymbol{X\beta}}} = \frac{1}{1+e^{-(\boldsymbol{X\beta})}}$$

P(X) is the probability of "1", in the binary outcome, X is the list of predictors, β**'s** are parameters estimated. The logit link function for logistic regression models log(odds); therefore, odds can be computed for both the "0" and "1" group in order to compute an odds ratio for group comparisons.

2.4 LINK FUNCTIONS

Link functions and families allow for the extension of mixed models for other types of outcome variables such as count, binary, and continuous outcomes by linking beta parameters estimated to real parameters (Agresti, 2007). The link function restricts the range of values of a particular estimate. Without the link, the estimates of the regression function can range from ($-\infty$, ∞) which may not necessary fit the underlying distributions (Bates, 2010). For example, binary outcome variables in logistic regression can only take on values of 0 or 1.

Table 1. Commonly Used Link Functions

Link	Formula	Domain
Identity	μ	$(-\infty,\infty)$
Log	$\log \mu$	$(0, \infty)$
Inverse	$1/ \mu$	$(0, \infty)$
Sq. root	$\sqrt{\mu}$	$(0, \infty)$
Logit	$\text{Log } (\mu/(1- \mu))$	[0,1]
Log-log	$\text{Log}(-\log(1- \mu))$	[0,1]
Power	μ^k	$(0, \infty)$

2.5 LONGITUDINAL MIXED MODELS

Longitudinal studies are defined as a study which the participant is followed and data is collected over an extended period of time. These studies are very beneficial because they allow for the observation of incident events, allow for a prospective look at exposures, allow for the observation of individual changes in outcomes, and allow for the separation of time effects. Potential weaknesses of this study design are that it is strongly affected by patient follow-up.

This method also allows for the analysis of correlated data, as well as the introduction of more complex modeling due to the influence of time-varying covariates (Brown, 2006).

Previously, alternative methodology would have been used for the analysis of longitudinal data such as analyzing mean response over time, analyzing the data at each time point, and analyzing data at all-time points with fixed subject effects. These alternative methods do have some disadvantages over mixed modeling. The first two approaches do not allow for the observation of differences over time. The fixed subject effect models, although should be equivalent to random subject effect model, may result in some estimation errors if there are too many subjects and very few time points (Brown, 2006).

Mixed modeling was performed using the PROC MIXED procedure with a REPEATED statement to control for the repeated measures found in the data as well as a RANDOM statement to allow for the differences in intercepts seen on the spaghetti plots for each individual. Spaghetti plots were used as a visual tool allowing for individual trajectories to be seen over time. For this reason, spaghetti plots are generated to show the time course profile for each outcome by genotype group and race in addition to mean trajectory plots (SAS Institute Inc, 2002).

Mixed models for modeling longitudinal data allows for two components: a fixed and a random component. The fixed components are interpreted as they would be in normal linear regression. The random component adjustments and corrects for correlations present in the data, due to the repeated measurements and lack of independence of the observations.

The general form of a mixed model is given by:

$$\mathbf{y=X\beta+Z\gamma+\varepsilon}$$

Where **y** is a n x 1 vector of the outcome variable; **X** is a n x p matrix of the p predictor variables; **β** is a p x 1 column vector of the fixed-effects regression coefficients; **Z** is the n x q design matrix for the q random effects; **γ** is a q x 1 vector of the random effects and is also considered the random complement to the fixed **β**; and **ε** is a n x 1 column vector of the residuals (Brown, 2006).

The variance of Y is defined by:

$$\mathbf{V=ZGZ' +R}$$

Where R is the variance component resulting from the fixed effects, var=σ^2 I, G is the variance component attributed to the random part of the model, and Z is the design matrix. If the variance of the random component is zero, G=0, the resulting model is equivalent to regular linear regression (Brown, 2006).

2.6 COVARIANCE STRUCTURES

Covariance structures were determined by running models with the various structures (compound symmetry, AR(1), toeplitz, and unstructured) and comparing AIC, with the model containing the lowest AIC selected as the best model. Inclusion of a covariance structure accounts for the lack of independence within the sample population, due to the repeated measurements. Common covariance structures include: compound symmetry (CS), Auto-regressive 1 (AR-1), Toeplitz, and unstructured (Brown, 2006).

CS example:

$$\begin{pmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{pmatrix}$$

AR(1) example:

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

Toeplitz example:

$$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_4 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_5 & \sigma_4 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$$

Unstructured example:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} & \sigma_{35} & \sigma_{36} \\ \sigma_{41} & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 & \sigma_{56} \\ \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_6^2 \end{pmatrix}$$

2.7 MODEL DIAGNOSTICS

Model diagnostics were outputted using the VCIRY option in the PROC MIXED SAS procedure. This adds scaled marginal residual to output data sets and assesses model fit by examining departures from normality. Once adequate fit was observed, a SAS macro was created to examine the SNPs in the selected serotonin and inflammation pathway genes as predictors for the risk measurements longitudinally for all of the outcome variables. Once single locus results were obtained haplotypes will be generated using Haploview Tagger v.4.2 and PHASE v. 2.1 software and the analysis was repeating using haplotypes as predictor for the longitudinal risk factors (Barrett, 2005).

3.0 RESULTS AND CONCLUSIONS

Table 2: Baseline Population Characteristics

Variable	Gender	Medication Use	Age	DBP	BMI	SES	AF
	Mean (SD) or N (%)						
Category	Male	Yes					
Blacks	135 (29.1%)	244 (52.6%)	58.01 (7.34)	82.91 (9.95)	32.37 (6.59)	-0.042 (0.87)	0.68 (0.16)
Whites	294 (38.1%)	251 (32.6%)	59.60 (7.34)	79.24 (9.96)	28.61 (5.26)	-0.042 (0.84)	
p-value	0.001	<0.001	<0.001	<0.001	<0.001	0.921	

T-test p-values are shown for continuous variables and X^2 p-values are shown for categorical variables comparing the Black and White populations.

Baseline characteristics differ between Black and White populations with gender proportions, medication usge, age, blood pressure, and BMI. With Blacks overall having higher proportions of females, medication usage, DBP, and BMI but overall being slightly younger. There was not a statistically significant difference in the two populations in SES. Percent African ancestry (AF) was only examined in the Black population with an average AF of approximately 68%.

Table 3: Number of observations by Time point

Variable	Time Point (years)	Whites (N)	Blacks (N)
DBP	Baseline	771	463
	1	734	404
	2	692	383
	3	672	372
	4	664	354

Table 3 shows adequate sample size and follow-up to analyze the data longitudinally. The White population is larger, starting with 771 individuals, with approximately 14% loss of follow-up between baseline and year 4 measurements. The Black population, starting with 463 individuals, had about 24% loss of follow-up between baseline and year 4 measurements.

Table 4: Minor Allele Frequencies of SELE SNPs tested

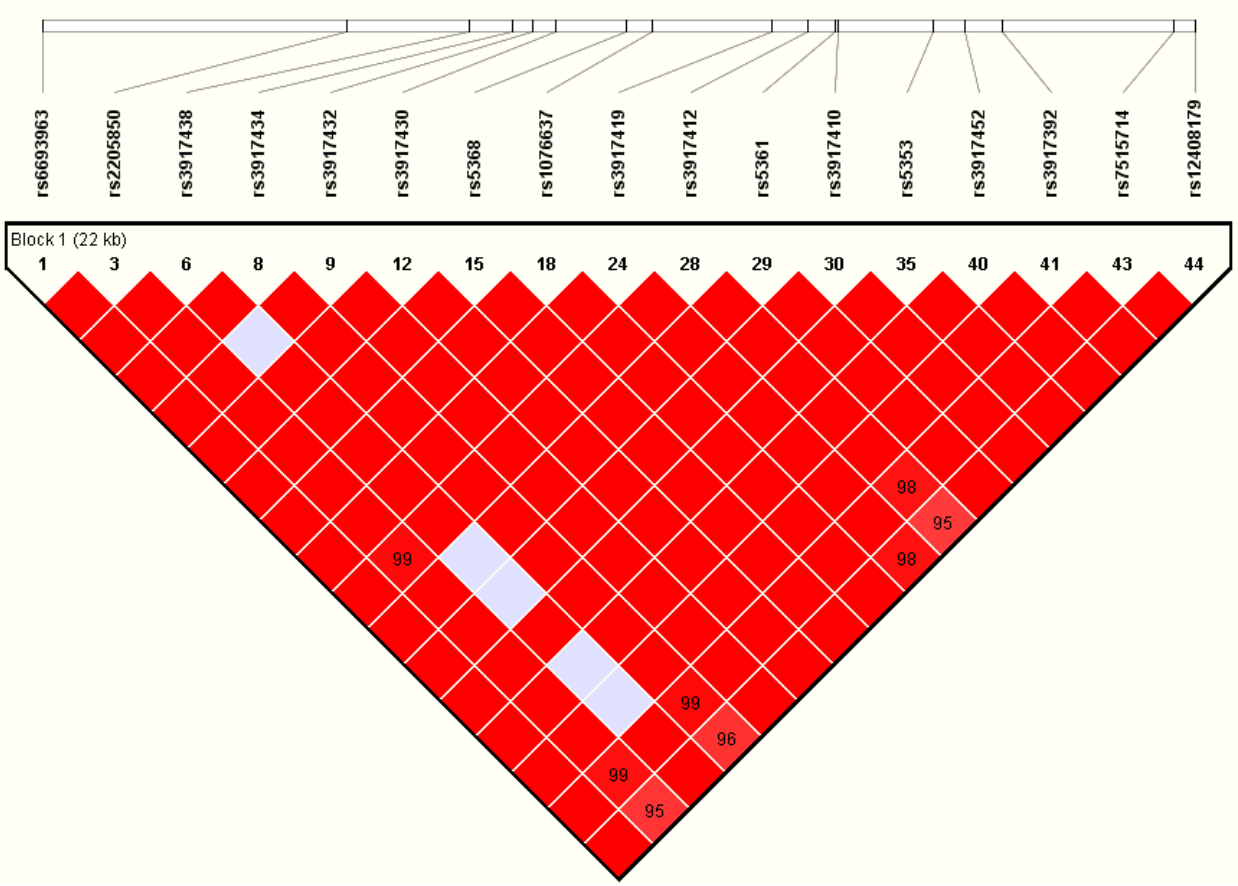
Name	White Alleles (Major: Minor)	White MAF	White HWE p-value	Black Alleles (Major: Minor)	Black MAF	Black HWE p-value
rs6693963	C:G	0.148	0.613	C:G	0.112	0.572
rs2205850	A:G	0.290	0.150	A:G	0.094	0.007
rs3917439	C:C	0.000	1.000	C:A	0.053	0.525
rs3917438	G:A	0.053	0.726	G:A	0.013	1.000
rs3917437	G:A	0.001	1.000	G:A	0.058	0.391
rs3917434	A:G	0.293	0.247	A:G	0.135	1.000
rs3917432	T:A	0.096	1.000	T:A	0.026	0.526
rs3917430	C:G	0.155	0.564	C:G	0.261	0.329
rs5368	G:A	0.108	0.883	G:A	0.084	0.284
rs1076637	G:A	0.159	0.751	G:A	0.302	0.336
rs3917419	G:A	0.419	0.322	G:A	0.282	0.215
rs3917415	G:C	0.001	1.000	G:C	0.094	0.755
rs3917413	A:G	0.498	0.242	G:A	0.326	0.983
rs3917412	G:A	0.250	0.606	G:A	0.069	0.681
rs5361	A:C	0.111	0.759	A:C	0.033	1.000
rs3917410	A:G	0.111	0.759	A:G	0.033	1.000
rs727909	G:A	0.170	0.004	G:A	0.352	0.992
rs5353	A:G	0.287	0.311	A:G	0.420	0.280
rs3917397	A:G	0.002	1.000	A:G	0.058	0.389
rs3917452	C:A	0.111	0.759	C:A	0.033	1.000
rs3917392	A:G	0.109	0.958	A:G	0.033	1.000
rs10919229	A:T	0.050	1.000	A:T	0.209	1.000
rs7515714	G:A	0.288	0.378	G:A	0.233	0.532
rs12408179	A:G	0.150	0.706	A:G	0.088	0.210

Minor allele frequencies and HWE- p-values were calculated as part of inclusion criteria for SNP's to be tested. If a SNPs MAF or HWE p-value was less than 0.05, it was excluded from the analysis. Four White and 6 Black SNPs were excluded based on MAF and 1 SNP for each race was also excluded based on HWE.

Linkage disequilibrium plots were then visualized in HAPLOVIEW. Essentially, a heat map is outputted with D=1 indicated in RED with other percentages clearly marked. Grey indicates very low LD. From this we can see that the genetic structures differ between the two races which is also validated in publically available databases such as HapMap.

The LD plot for Blacks, pictured above, indicates that variation in SELE can be broken into one large block containing SNPs rs6693963-rs7515714 and rs12408179. This indicates that the majority of variation in this gene can be summarized by selecting one SNP from rs6693963-rs7515714 as well as rs12408179.

Figure 1: SELE Linkage disequilibrium plot for Black population



The LD plot for Whites, pictured above, indicates that variation in SELE can be summarized by one large haplotype block and that all SNPs tested were in very high LD.

Figure 2: SELE Linkage Disequilibrium Plot in Whites

Single SNP Analysis Schematic and Example:

All SNPs which met the HWE and MAF requirement were analyzed in the same manner. For this reason, one example, rs5368, will be looked at in detail. This SNP was selected because it met inclusion criteria for both whites and blacks. It is also a coding SNP within SELE which is known to causes an amino acid substitution. This SNP has not been looked at for associations with BP but has previously looked at its association with Multiple Sclerosis (Fenoglio, 2009) and SELE levels (Wu, 2012).

Table 5: SNP Sample Characteristics

SNP	Race	Allele Counts	MAF	χ^2 p-value
rs5368	Whites	AA (8) AG(151) GG(612)	0.108	0.551
	Blacks	AA (1) AG(76) GG(387)	0.084	

χ^2 p-value shows there is not a significant difference in allele frequencies between the two races.

3.1 WHITE MIXED MODEL ANALYSIS RESULTS

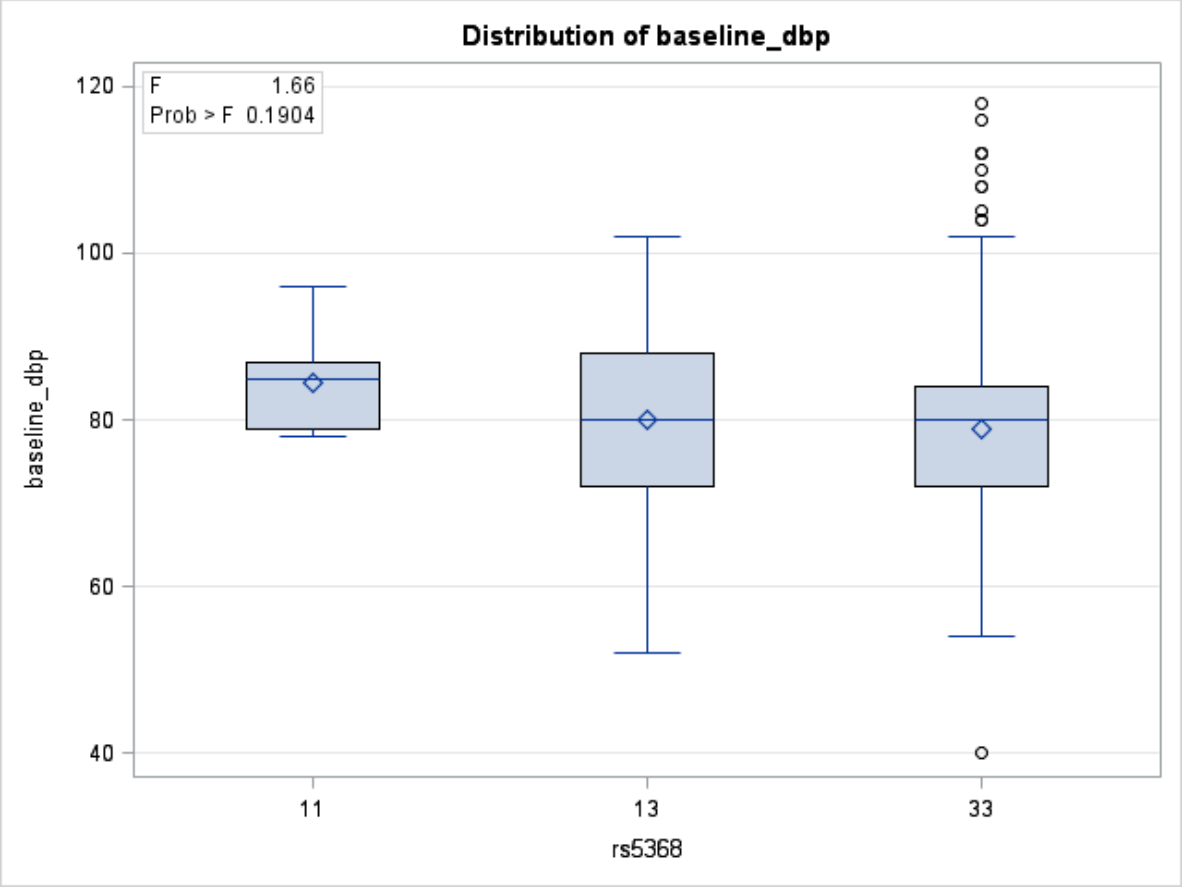


Figure 3: ANOVA results comparing baseline DBP by rs5368 genotype for Whites

ANOVA results indicate there is a difference in mean DBP at baseline between rs5368 genotype groups (F=5.45, p=0.004).

Individual trajectories were then visualized for each population by genotype. Spaghetti plots allow for the identification of trends across time as well as the visualization of individual variations within the data.

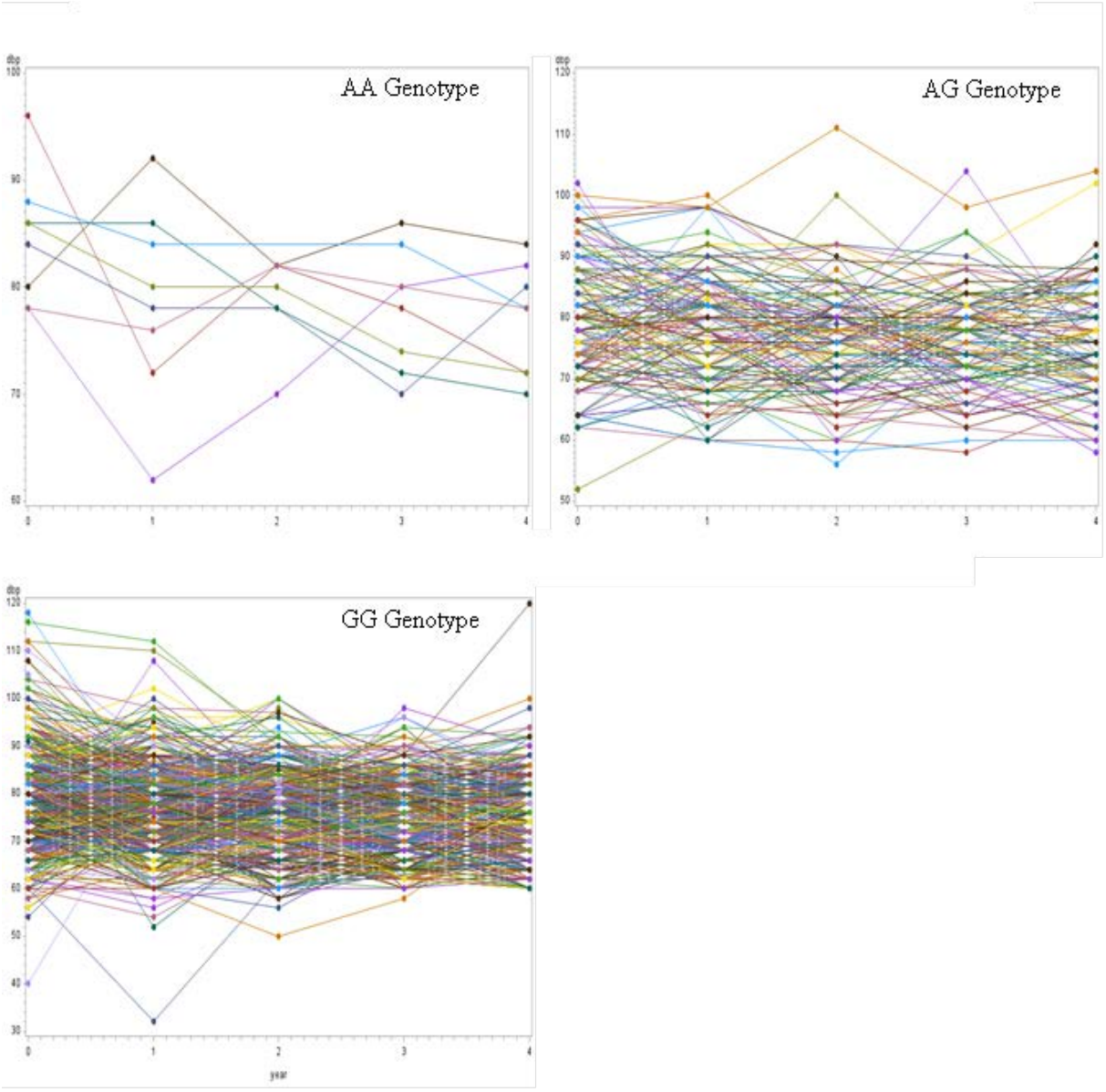


Figure 4: White Spaghetti Plots for DBP over time by rs5368 genotype

Variation in spaghetti plots were then summarized by plotting mean trajectories of each genotype group.

The mean trajectory plot for whites (Figure) shows differences in slope as well as intercepts for the three genotype groups, indicating that it is appropriate to allow for a random intercept as well as test for a possible interaction between genotype and the time variable (year).

Figure 5: White Spaghetti Plots for DBP over time by rs5368 genotype

Mixed univariate analysis was then performed in SAS using the PROC MIXED procedure with a REPEATED statement to control for the repeated measures found in the data as well as a RANDOM statement to allow for the differences in intercepts seen on the spaghetti plots for each individual. Rs5368 and time were included as predictors (Tables 8 and 13,

respectively). Using the models only containing the SNP and year, covariance structure were determined by re-running models with the various structures (compound symmetry, AR(1), toeplitz, and unstructured) and comparing AIC, with the model containing the lowest AIC selected as the best model.

Table 6: AIC for various Univariate models with different covariance structures in whites

	Covariance Structure	AIC	Order
Model 1	CS	24284.8	4
Model 2	AR(1)	24277.5	2
Model 3	Toeplitz	24282.0	3
Model 4	Unstructured	24155.9	1*

For whites, an unstructured correlation structure preformed the best using AIC selection criteria (Table 6).

Table 7: Univariate Model for DBP and rs5368 in whites

Effect	**rs5368**	**Estimate**	**S.E.**	**DF**	**P-value**	**Type III test of Fixed Effects**
Intercept		76.34	0.25	760	<.0001	
rs5368	**11**	2.55	2.20	724	0.247	0.152
rs5368	**13**	0.93	0.58	770	0.106	
rs5368	**33**	0	.	.	.	

Since rs5368 was found to have a priori p-value < 0.20 univariately (p= 0.152), models were adjusted for covariates. Age, sex, socio-economic status, medication usage, year, baseline levels, and time varying BMI were included in each model. Backwards selection was used until all variables remained significant in the model.

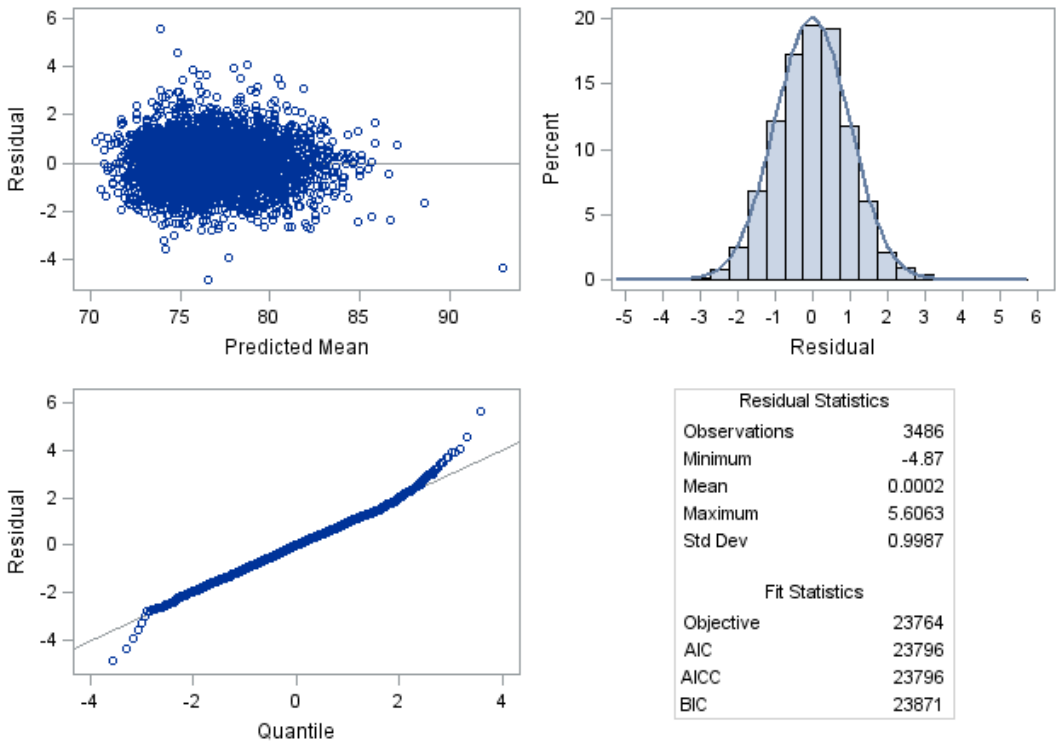
Table 8: Full Model for DBP and rs5368 in whites

Effect	rs5368	year	Estimate	S.E.	DF	P-value	Type III test of Fixed Effects
Intercept			70.25	2.73	657	<.0001	
age			0.01	0.04	566	0.8598	0.8598
sex			-1.55	0.52	572	0.0032	0.0032
zses			-0.48	0.32	618	0.1342	0.1342
bmi			0.37	0.04	885	<.0001	<.0001
Antihypertensive medication			1.28	0.55	578	0.0197	0.0197
year		**1**	-1.37	0.50	557	0.0067	
year		**2**	-3.06	0.49	557	<.0001	
year		**3**	-3.63	0.49	545	<.0001	0.0339
year		**4**	-3.04	0.50	535	<.0001	
year		**0**	0.00	.	.	.	
rs5368	**11**		5.68	4.46	567	0.2034	
rs5368	**13**		0.36	1.02	569	0.7278	0.4087
rs5368	**33**		0.00	.	.	.	
rs5368*year	**11**	**1**	-5.06	4.73	543	0.2855	
rs5368*year	**11**	**2**	-3.33	4.60	533	0.4692	
rs5368*year	**11**	**3**	-3.74	4.51	516	0.4075	
rs5368*year	**11**	**4**	-3.79	4.60	502	0.41	
rs5368*year	**11**	**0**	0.00	.	.	.	
rs5368*year	**13**	**1**	0.46	1.11	565	0.681	
rs5368*year	**13**	**2**	-0.34	1.08	562	0.7517	
rs5368*year	**13**	**3**	0.64	1.07	556	0.5479	0.9413
rs5368*year	**13**	**4**	0.58	1.09	546	0.5952	
rs5368*year	**13**	**0**	0.00				
rs5368*year	**33**	**1**	0.00				
rs5368*year	**33**	**2**	0.00				
rs5368*year	**33**	**3**	0.00				
rs5368*year	**33**	**4**	0.00				
rs5368*year	**33**	**0**	0.00				

The interaction term and zses were sequentially removed from the model. Age was not significant but was forced into the model due to its clinical significance.

Table 9: Final model for DBP and rs5368 in whites after backward selection

Effect	rs5368	year	Estimate	S.E.	DF	P-value	Type III test of Fixed Effects
Intercept			68.90	2.27	879	<.0001	0.872
age			0.00	0.03	768	0.872	0.025
sex			-0.99	0.44	766	0.025	<.0001
BMI			0.40	0.04	1130	<.0001	<.0001
year		**1**	-1.93	0.38	751	<.0001	<.0001
year		**2**	-2.96	0.37	746	<.0001	
year		**3**	-3.48	0.37	738	<.0001	
year		**4**	-3.36	0.38	722	<.0001	
year		**0**	0.00	.	.	.	
rs5368	**11**		3.04	2.05	720	0.140	0.124
rs5368	**13**		0.80	0.54	772	0.137	
rs5368	**33**		0.00	.	.		



The final model in whites contained age, gender, BMI, and year. The SNP was not statistically significant based on a $p<0.05$; however, model diagnostics indicate appropriate fit. This can be seen visually on the qq-plot and distribution of errors which appear to be random without any noticeable patterning.

Figure 6: Model Diagnostics for Final Model rs5368 in whites

Mixed modeling allowed for the examination of the effect of genotype longitudinally over time. Analysis was then also performed to examine the effect of genotype on group membership of clustering within the data. Trajectory analysis was then performed using the PROC TRAJ procedure in SAS. Different combinations of groups and ordered models were tested; however, the two groupings with linear relationships were favored over the other grouping combinations (based on BIC).

3.2 WHITE TRAJECTORY ANALYSIS RESULTS

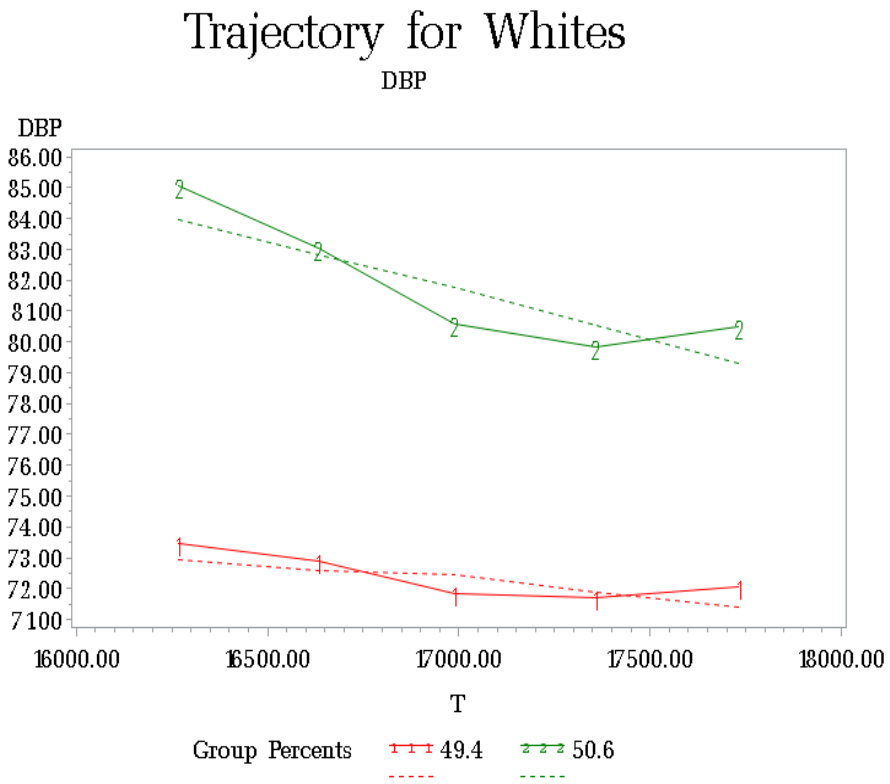


Figure 7: White DBP Group Membership

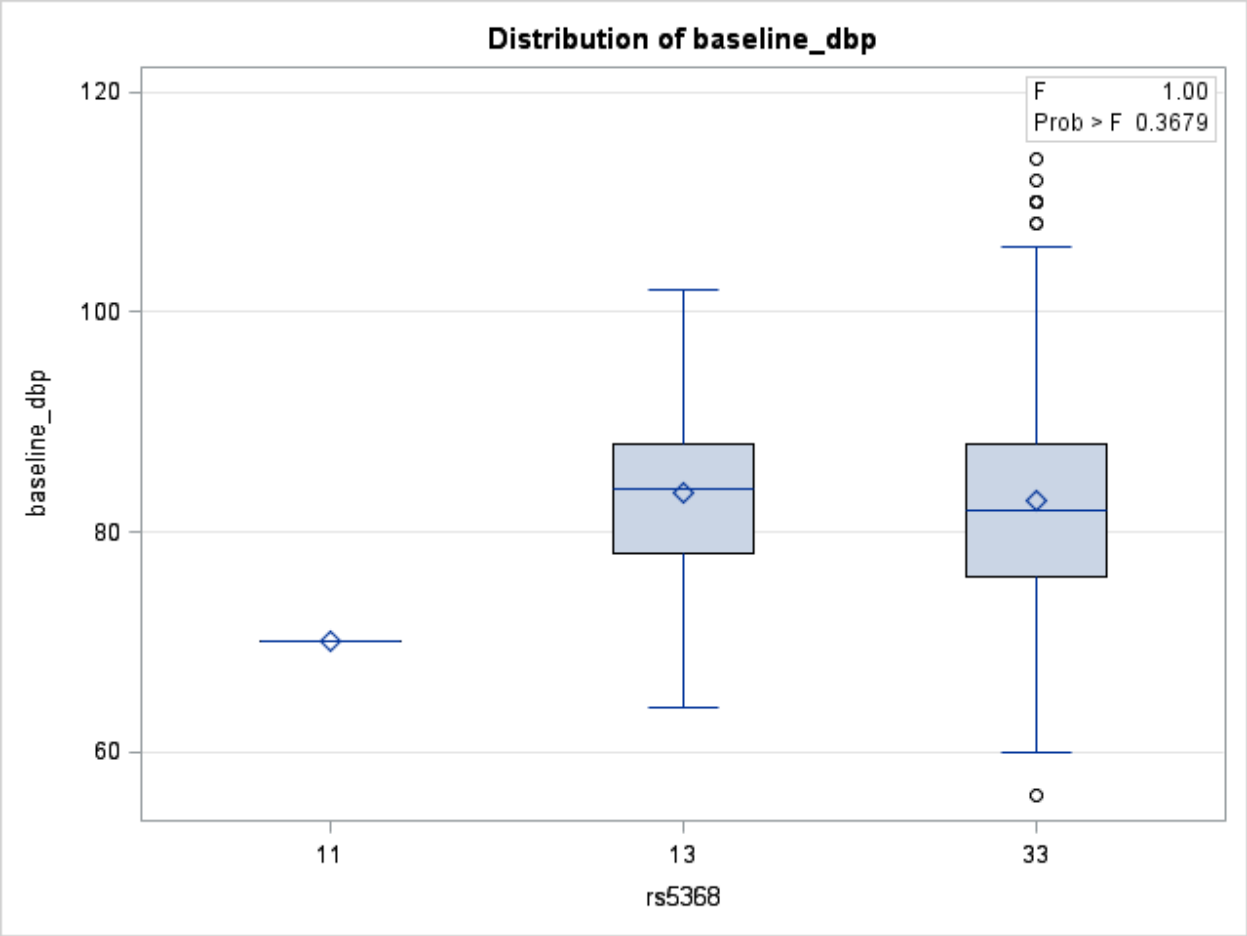
The two groupings were then recoded into indicator (0=low, 1=high DBP group) and logistic regression was performed in Plink v1.07.

Table 10: Univariate Logistic Results using DBP group membership as outcomes Whites

Race	SNP	BP	Beta	OR	Allele	STAT	P-value
Whites	rs5368	167963570	0.504	1.655	A	2.573	0.01008

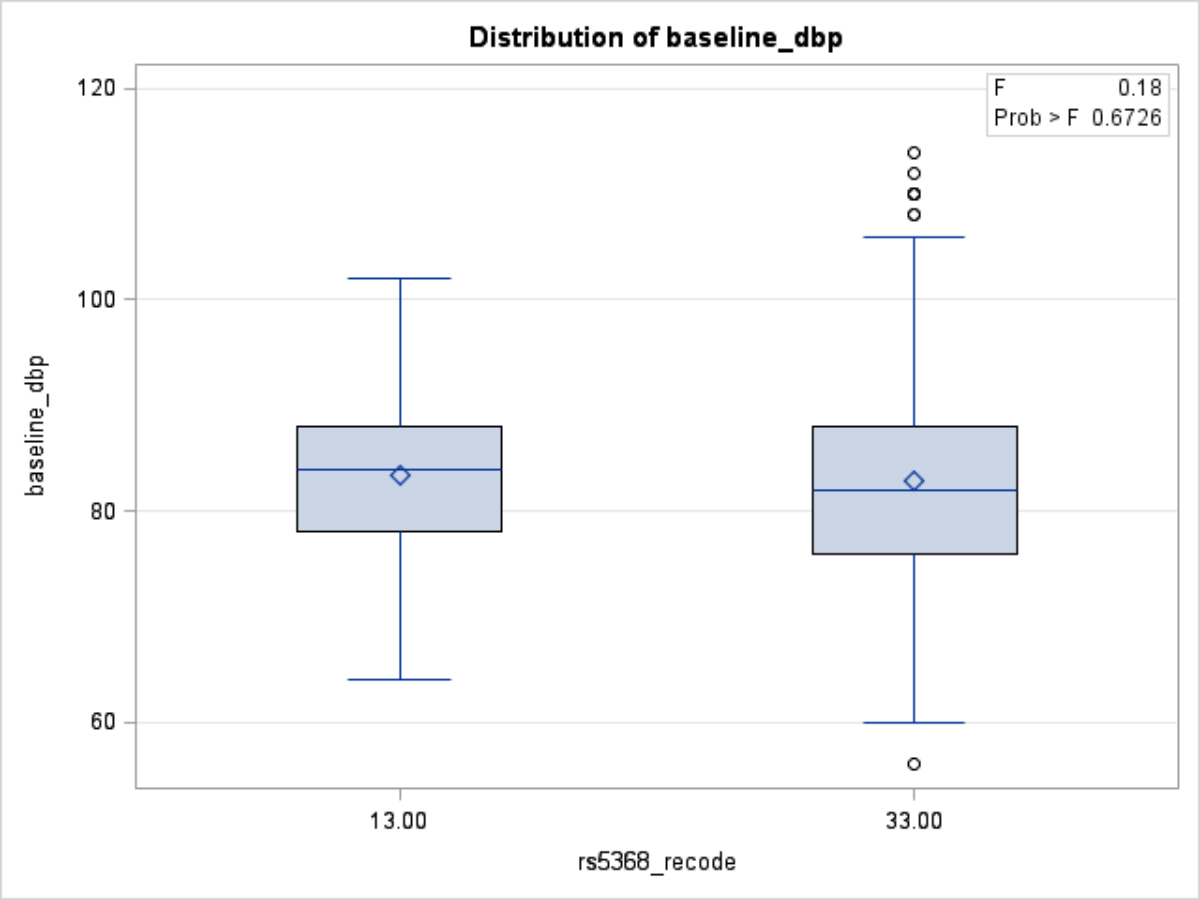
Plink outputs the effect of each extra minor allele. In the white trajectory analysis OR=1.655 indicating that odds of being in the "high" DBP category is ~66% higher for those individuals with each additional A allele.

3.3 BLACK MIXED MODEL ANALYSIS RESULTS



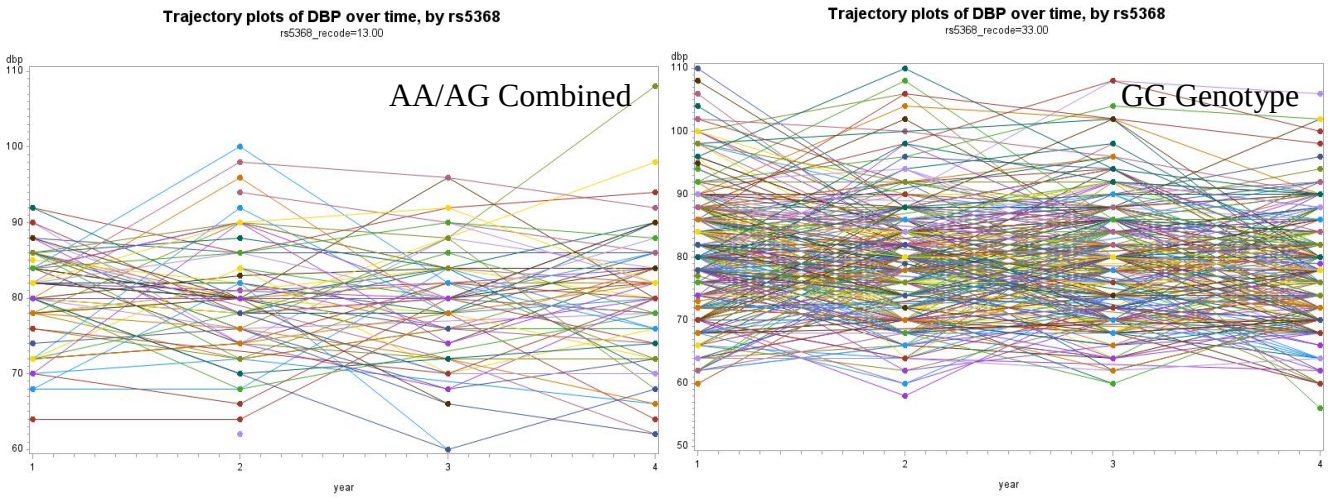
From Figure 7 above, we can see that only one individual has the minor/minor allele genotype. This would cause potential problems with the estimation and evaluation of statistical models due to the small sample size. For this reason, for the remaining analysis the 11 (AA) and 13 (AG) groups will be combined.

Figure 8: ANOVA results comparing baseline DBP by rs5368 genotype for Blacks



ANOVA results indicate that there was not a statically significant difference in mean DBP at baseline for Blacks (F=0.18, P=0.673).

Figure 9: ANOVA results comparing baseline DBP by rs5368 genotype for Blacks with AA and AG groups combined



Similarly in Blacks, spaghetti plots show variation in individual trajectories of DBP over time.

Figure 10: Individual trajectory plots for black population using combined SNP groups

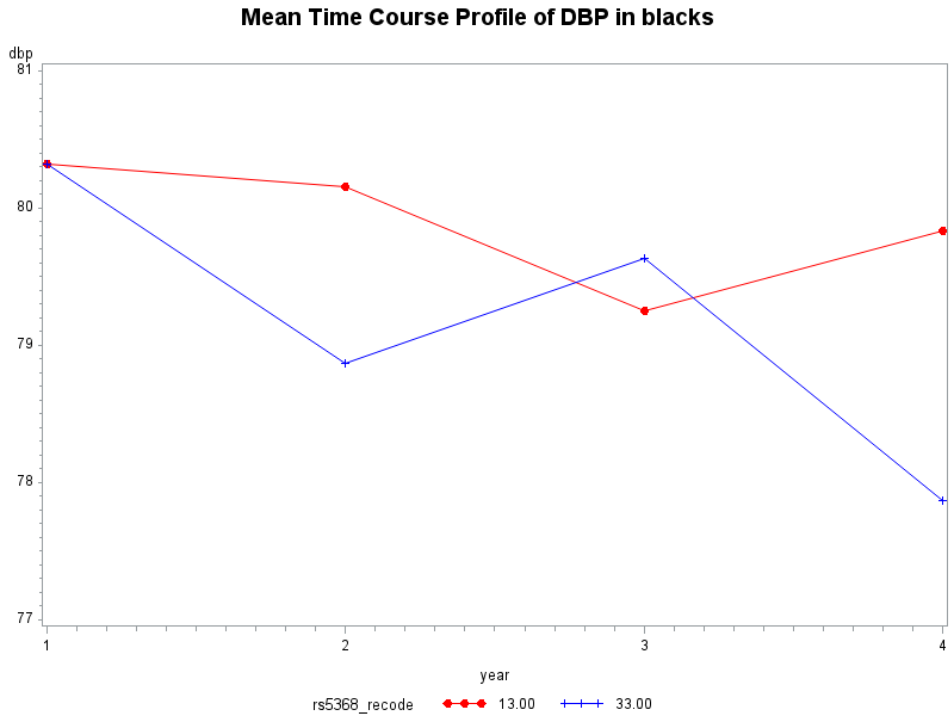


Figure 11: Mean Trajectories and best fit for DBP over time by rs5368 combined genotype in Blacks

Mean time course profiles indicate a different slope and intercept between genotype groups. This means that mixed modeling would be appropriate which a REPEATED statement is used to control for the repeated measures found in the data as well as a RANDOM statement to allow for the differences in intercepts seen on the spaghetti plots for each individual.

Table 11: AIC for various Univariate models with different covariance structures in blacks

	Covariance Structure	AIC	Order
Model 1	CS	13711.3	4
Model 2	AR(1)	13704.6	2
Model 3	Toeplitz	13708.4	3
Model 4	Unstructured	13673.8	1*

AIC tables indicate that an unstructured correlation structure would provide a best fit for the correlations in the data (AIC=13673.8).

Table 12: Univariate Model for DBP and rs5368 in blacks

Effect	**rs5368 recode**	**Estimate**	**S.E.**	**P-value|**
Intercept		79.256	0.345	<.0001
rs5368 recode	11 and 13 combined	0.503	0.841	0.550
rs5368 recode	33	0.000		

Black univariate results indicate a there was not a statistically significant association between rs5368 genotype recoding and DBP (p=0.550).

Although the p-value is greater than the a-priori p-value of p=0.20, models were still adjusted for covariates (age, gender, zses, medication use, and bmi) in this example. Backward model selection was then used to systematically eliminate variables from the model, based on p=0.05, until all variables remained significant.

Table 13: Full model for DBP and rs5368 in blacks

Effect	rs5368	year	Med	Estimate	S.E.	P-value	TIII p-value
Intercept				75.268	4.281	<.0001	
age				-0.033	0.052	0.529	0.529
sex				-2.516	0.807	0.002	0.002
Anti-hyp			Yes	0.410	0.750	0.585	0.585
Anti-hyp			No	0.000	.	.	
BMI				0.266	0.057	<.0001	<.0001
ZSES				-0.309	0.407	0.448	0.448
AF				0.399	2.326	0.864	0.864
year		1		2.686	0.627	<.0001	0.011
year		2		0.229	0.618	0.712	
year		3		1.008	0.552	0.069	
year		4		0.000	.	.	
rs5368	11 and 13 combined			1.875	1.308	0.153	0.383
rs5368	33			0.000	.	.	
rs5368*year	11 and 13 combined	1		-1.658	1.521	0.277	0.690
rs5368*year	11 and 13 combined	2		-1.504	1.483	0.312	
rs5368*year	11 and 13 combined	3		-1.001	1.313	0.447	
rs5368 *year	11 and 13 combined	4		0.000	.	.	

Sex, bmi, antihypertensive medication usage, and year remained after the selection process. Age was forced into the model due to its clinical significance resulting in the final model.

Table 14: Final model for DBP and rs5368 in Blacks after backward selection

Effect	rs5368 recode	year	Estimate	S.E.	P-value	Type III p-value
Intercept			77.296	3.211	<.0001	
age			-0.054	0.041	0.194	0.194
sex			-2.715	0.673	<.0001	<.0001
bmi			0.269	0.047	<.0001	<.0001
year		**1**	2.014	0.483	<.0001	0.0002
year		**2**	0.932	0.466	0.046	
year		**3**	1.369	0.427	0.002	
year		**4**	0.000	.	.	
rs5368 recode	**11 and 13 Combined**		0.522	0.810	0.520	0.520
rs5368 recode	**33**		0.000	.		

The final model in Blacks indicate that rs5368 was not a statistically significant (p=0.520) predictor of DBP.

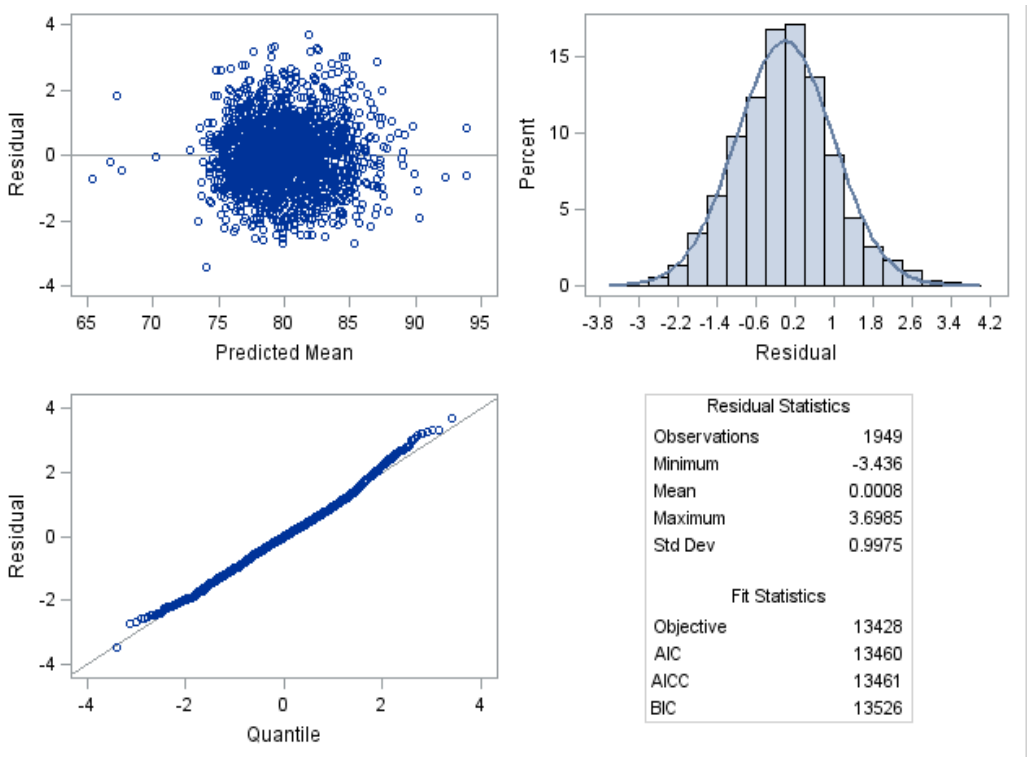


Figure 12: Model Diagnostics for Final Model in blacks

Model diagnosis indicated appropriate fit of the model. This can be visually observed by the plots of the residuals and qq-plot, which show normally distributed residuals with no visual patterning. Trajectory analysis was then performed using the PROC TRAJ procedure in SAS. Different combinations of groups and ordered models were tested; however, the two groupings with linear relationships were favored over the other grouping combinations (based on BIC).

3.4 BLACK TRAJECTORY ANALYSIS RESULTS

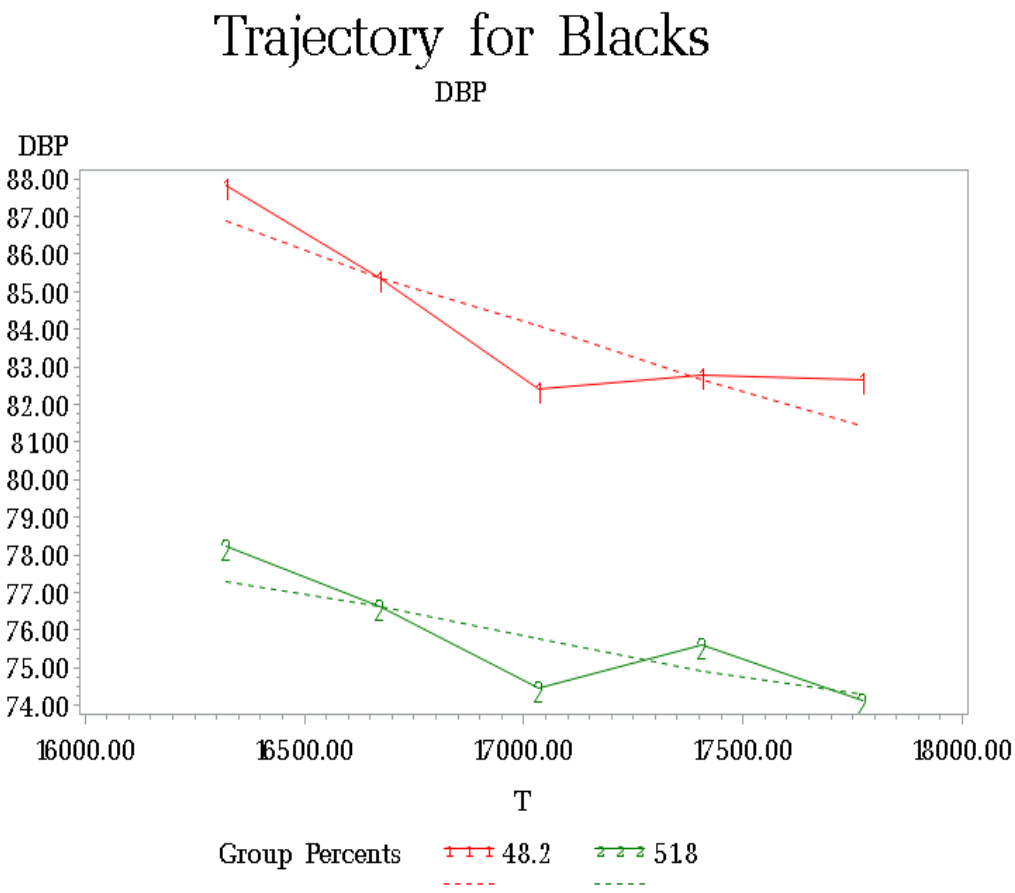


Figure 13: Group Trajectory Plot Blacks

The two groupings were then recoded into indicator (0=low, 1=high DBP group) and logistic regression was performed using plink v.1.07.

Table 15: Univariate Logistic Results using DBP group membership as outcomes Blacks

SNP	BP	Allele	Beta	OR	STAT	P-value
rs5368 (Combined)	167963570	A	-0.403	1.018	-1.355	0.883

Logistic regression results indicate that the minor allele for rs5368 does not significantly effect odds of group membership (p=0.883).

4.0 DISCUSSION

After pruning of SNPs in SELE, based on MAF=0.05 and HWE=0.05, 19 snps were examined in whites, and 18 snps were examined in blacks for longitudinal grouping associations with DBP. In whites, 6 of the 19 snps tested showed significant associations with DBP group membership: rs3917430, rs1076637, rs5353, rs5368, rs3917413, and rs3917432 with p-values of 0.002, 0.002, 0.002, 0.01, 0.022, and 0.048, respectively. In blacks, only one SNP, rs3917412, was statistically significantly associated with group membership p=0.022. In future analysis it may be beneficial to combine genotype groups for snps which had very low minor/minor allele parings to increase power and test whether the presence or absence of the allele is influencing group membership. Haplotype analysis may also be helpful to try to identify if the variation in SELE, as a whole, may be responsible for CVD risk disparities as opposed to one SNP in the gene region. Multiple testing could also be accounted for by applying some sort of correction (FDR, Bonferoni, etc.) to decrease the type II error.

APPENDIX A: White SELE Logistic Regression Results

Table 16: White SELE Logistic Regression Results for Remaining SNPs

SNP	BP	Allele	N	BETA	OR	STAT	P-value
rs6693963	167952069	3	576	-0.153	0.858	-0.887	0.375
rs2205850	167958063	3	576	-0.176	0.838	-1.318	0.187
rs3917438	167960474	1	576	0.118	1.125	0.461	0.645
rs3917434	167961319	3	576	-0.158	0.854	-1.184	0.237
rs3917432	167961734	1	575	0.405	1.499	1.977	0.048
rs3917430	167962186	3	576	0.536	1.710	3.169	0.002
rs5368	167963570	1	576	0.504	1.655	2.573	0.010
rs1076637	167964068	1	576	0.533	1.704	3.159	0.002
rs3917419	167966443	1	576	-0.218	0.804	-1.804	0.071
rs3917413	167966909	3	191	0.507	1.660	2.293	0.022
rs3917412	167967126	1	576	-0.179	0.836	-1.296	0.195
rs5361	167967684	2	576	0.004	1.004	0.021	0.983
rs3917410	167967732	3	576	0.004	1.004	0.021	0.983
rs5353	167969598	3	576	0.424	1.528	3.144	0.002
rs3917452	167970241	1	576	0.004	1.004	0.021	0.983
rs3917392	167970959	3	549	-0.073	0.930	-0.380	0.704
rs7515714	167974353	1	567	-0.174	0.840	-1.299	0.194
rs12408179	167974751	3	576	-0.123	0.884	-0.716	0.474
rs725974	168879875	2	574	0.136	1.146	0.895	0.371

APPENDIX B: Black SELE Logistic Regression Results

Table 17: Black SELE Logistic Regression Results for Remaining SNPs

SNP	BP	Allele	N	BETA	OR	STAT	P-value
rs6693963	167952069	G	320	0.127	1.135	0.497	0.620
rs3917439	167960345	A	320	0.092	1.097	0.257	0.797
rs3917437	167960645	A	319	0.577	1.780	1.683	0.092
rs3917434	167961319	G	320	0.380	1.462	1.623	0.105
rs3917430	167962186	G	320	0.060	1.061	0.349	0.727
rs5368 (Combined)	167963570	A	320	-0.403	1.018	-1.355	0.883
rs1076637	167964068	A	320	0.019	1.020	0.118	0.906
rs3917419	167966443	A	320	-0.031	0.970	-0.163	0.870
rs3917415	167966762	C	320	-0.489	0.613	-1.697	0.090
rs3917413	167966909	A	134	0.154	1.167	0.608	0.543
rs3917412	167967126	A	320	0.760	2.139	2.287	0.022
rs727909	167968965	A	262	0.059	1.061	0.324	0.746
rs5353	167969598	G	320	-0.103	0.902	-0.664	0.507
rs3917397	167969808	G	320	0.570	1.769	1.665	0.096
rs10919229	167971751	T	320	0.289	1.335	1.490	0.136
rs7515714	167974353	A	318	-0.039	0.962	-0.206	0.837
rs12408179	167974751	G	320	0.339	1.404	1.136	0.256
rs725974	168879875	C	319	0.082	1.085	0.422	0.673

BIBLIOGRAPHY

Agresti A. et al. (2007). An Introduction to Categorical Data Analysis, 2nd Edition. Hoboken, NJ: Wiley.

American Heart Association. (2014). FACTS Bridging the Gap CVD Health Disparities.Retrieved July 18, 2014, from http://www.heart.org/idc/groups/heart-public/@wcm/@hcm/@ml/documents/downloadable/ucm_429240.pdf

Barrett JC., et al. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics.[PubMed ID: 15297300]

Brown, H., & Prescott, R. (2006). Applied mixed models in medicine (2nd ed.). Chichester, England: John Wiley.

Bates, D. (2010). Generalized linear models. . Retrieved from http://www.stat.wisc.edu/courses/st849-bates/lectures/GLMH.pdf

CDC. (2011) Million Hearts: strategies to reduce the prevalence of leading cardiovascular disease risk factors. United States. MMWR2011;60(36):1248–51.

Chapman R., Hall D. (2001). A First Course in Linear Model Theory, CRC.

El Shamieh, S., et al. (2012) Functional Epistatic Interaction between rs6046G>A in F7 and rs5355C>T in SELE Modifies Systolic Blood Pressure Levels. PLoS ONE, 7, e40777

Faruque, M. (2011). Association of ATP1B1, RGS5 and SELE polymorphisms with hypertension and blood pressure in African–Americans. Journal of Hypertension, 1906-1912.

Fenoglio, C. et. al. (2009). Candidate gene analysis of selectin cluster in patients with multiple sclerosis. *Journal of Neurology*, 832-833.

Heidenreich PA, et al. (2011) Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011; 123:933-44.

Murphy SL, et al. (2013). Deaths: Final data for 2010. Natl Vital Stat Rep. 61(4). http://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_04.pdf

Nagin, D. S. et al. (2007). Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them. Sociological Methods & Research, 542-571.

National Institutes of Health (NIH). (n.d.). *U.S National Library of Medicine*. Retrieved April 21, 2014, from http://www.nih.gov/

Purcell S, et.al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

Roeder, K., et al. (2001). A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. Sociological Methods & Research, 374-393.

SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2002-2004.

Wu, S, et. al. (2012). Association of SELE genotypes/haplotypes with sE-selectin levels in Taiwanese individuals: interactive effect of MMP9 level. *BMC Medical Genetics*, 115.