PENNSTATE

June 18, 2023

From: C. Cratty, J. Lee, A. Thibeault, Graduate Student Consultants

To: E. Green

PROJECT TITLE: Mutation Rate Factors in the Human Genome

EXECUTIVE SUMMARY

Using human genome information provided by the research team at Penn State Huck Institute for Life Sciences, analyses were carried out to understand the relationship between four gene mutation rates and a series of genetic factors and markers. All statistical tests in this study used the standard 5% significance level threshold.

Results revealed that all pairs of mutation rate share a statistically significant positive linear correlation, noting that the microsatellite repeat alterations has the weakest correlation among all pairings. Examining the states revealed that there is a difference in the mutation rates. In particular, states 1 and 6 differ in the insertion, deletion, and nucleotide substitution rate but did not differ in their microsatellite repeat alteration rate. The genomic factors nCGm and LINE showed a significant relationship with the insertion mutation rate but were found to only explain a small portion of the variation in mutation rate as observed.

1.0 - PROJECT DESCRIPTION

This analysis focuses on examining the rates of four types of common genetic mutations as observed in the human genome. The analysis uses information collected and cleaned by a research team headed by Eva Green at Penn State Huck Institute for Life Sciences. The statistical team was tasked to analyze the data for covariation among the mutation rates, the effect of genetic divergence states on mutation rates, and the covariation of mutation rates with a set of genomic factors. Any initial results shown in this study would be grounds for further, more in-depth, research and analysis by the human genome research teams.

1.1 - RESEARCH QUESTIONS

The following questions are addressed in this report:

1. Characterize the linear co-variation among the four mutation rates (if any). Illustrate the co-variations in a two-dimensional plot.
2. Do our four mutation rates differ in these states? Determine whether average mutation rates for state 1 are different from state 6.

3. Link mutation rates and their co-variation to the genomic features and illustrate the relationship between mutation rates and genomic features.

1.2 - STATISTICAL QUESTIONS

Tables A.1, B.1 and C.1 in the appendices show the specific null and alternative hypotheses used in this report to provide answers for questions 1, 2, and 3, respectively.

1.3 - VARIABLES

The response variables of interest for this study are four standardized mutation rates within human DNA. These are: ins.std, del.std, sub.std, and microsat.std. There are three location variables; chr, start, and end. Finally, there are seven genomic factor variables: GC, CpG, nCGm, LINE, SINE, NLp, and state. Variable details can be found in Table 1.

2.0 - EXPLORATORY DATA ANALYSIS (EDA)

We first look at the correlation between the full set of continuous variables. A 2-D graphic in the form of a heat map is shown in Figure 1. We see that the microsatellite repeat alterations rate is the only mutation rate not strongly correlated with the other mutation rates, GC has a strong correlation with both CpG and SINE, and there appears to be little linear correlation between the mutation rates and the genetic features.

We continue the EDA examining the mutation rates by chromosome and by genetic state. These are included in the appendix as Figures 2 through 9. The difference between chromosomes does not appear to be significant, with a large portion of overlap between the plotted quartiles. Again, the microsatellite repeat alteration rate is the least affected and holds steady across all chromosomes. The genetic divergent states show more difference supporting examination between states 1 and 6.

3.0 - STATISTICAL ANALYSIS

Question 1:

In a linear setting, we use the correlation matrix and a scatterplot matrix to evaluate the covariance between all pairs of the 4 mutation rates. Correlation and covariance are both a method to describe the degree to which sets of random variables deviate from their expected value and are related in the following way. See Theory A.10. Since correlation is the standardized form of covariance, we will consider the correlation values throughout the analysis as they are easier to interpret.

Our method of verification is by way of the Student's t-distribution. See Theory A.11. In our case, the null and alternative hypotheses for these tests are listed in Table A.1. Before performing the correlation tests, we address the required assumptions:

- Both variables in consideration follow a normal distribution.
    - The small insertion and microsatellite repeat number alterations passed the Anderson-Darling tests for Normality. Upon further exploration, the deletion and nucleotide substitution mutation rates appear normally distributed enough that we can consider these empirically normal, given the large number of observations. Test results and histograms can be found in Figures A.2 - A.6.
- Neither variable has significant outliers.
    - Each variable was determined to have no extreme outliers. The highest and lowest observations of each mutation rate were confirmed to be within reasonable range and represented less than 1% of the total observations for each mutation rate.

The pair-wise test results are in Table A.2, from output in Figure A.1. From Table A.2, all tests reject the null hypothesis. We can conclude that there is sufficient evidence that all pairs of mutation rates have non-zero correlation at the 0.05 significance level.

We see all pairs are positively correlated. More importantly, we can also see the pairs (insertion, deletion), (insertion, nucleotide substitution), and (deletion, nucleotide substitution) are more strongly positively related, whereas the other pairings are only slightly related, almost to a point of insignificance. We visually confirm our findings in the scatterplot matrix Figure A.2. We see a positive covariation between the pairs mentioned above, as their scatterplots illustrate the positive linear relationship.

To investigate the covariation between the four mutation rates, we PCA. PCA is a dimension reduction machine learning technique that transforms data into a new coordinate system where each principal component describes the characteristics of most variation in the remaining data. When plotted this allows us to visually identify dimensions of the data that are linearly uncorrelated. We use this technique to identify the joint variability between the mutation rates. As the first two components account for approximately 85.24% of the variation in the data, we focus our analysis on the first two principal components. This is confirmed by observing the Scree and Cumulative Contribution Plots in Figure A.7.

We observe the joint covariation, in Figure A.8. The PCA biplot overlays a score plot with a loading plot: input scores of each observation are mapped below and the loading vectors (principal components 1 and 2) are overlaid on top, showing the contribution of each feature to the vector. The mapping visually confirms the insertion, substitution, and deletion mutation rates are linearly correlated while the microsatellite repeat number alterations mutation rate is not strongly correlated with the other rates.

Question 2:

The MANOVA method of analysis is used in this section. MANOVA is a method that allows us to determine if the mutation rates differ between each State. Contrasts are used to determine if specific mutation rates differ between State 1 and 6.

First, we must assess the MANOVA assumptions.

1) The data from each state has a common mean vector and there are no subpopulations within the states.
2) The data from all states have a common variance-covariance matrix as indicated by the result of testing hypothesis 2.5. We calculated a χ^2 test statistic of 57.1 with 50 degrees of freedom, resulting in a p-value of 0.2281. Therefore, we fail to reject H_0 at a 5% level of significance and the matrices are the same.
3) There is an assumption of independence in sampling each mutation rate.
4) By inspection of the squared Mahalanobis distances vs χ^2 quantiles we can see that the there is a relatively straight, consistent line indicating that the data are approximately Multivariate Normal. See Figure B.7 for the plot.

Because of the garnered interest in relationship of regional variations in human genome, tests are performed to understand if the location of the 1 Mb window along the human genome played a role in the mutation rates as well as the interaction between the location and State. The starting position of the 1 Mb window is added as a predictor for location. However, both the starting location and the interaction term between starting location and state are not significant, while State was significant. The hypotheses are noted in tests 2.1 through 2.3 in Table B.1 and the results are shown in Appendix B, Figures B.1 and B.3. As such, the model with State as the only predictor is assessed.

Using only the State, we perform hypothesis test 2.3 in Table B.1 to determine if there is a significant difference in at least one pair of mutation rates for at least one pair of states. We calculate a Wilk's Lambda test statistic of 0.376 and a p-value less than 0.0001. We reject the null hypothesis at a 5% level of significance. There is enough statistical evidence to conclude there are differences in at least one pair of mutation rates between at least one pair of states. The results are shown in Appendix B, Figure B.4.

A follow-on question was asked to determine if there are any differences between mutation rates between States 1 and 6. A contrast between States 1 and 6 is tested. We calculate a Wilk's Lambda test statistic of 0.930 and a p-value less than 0.0001 and reject the null hypothesis at a 5% level of significance. There is enough statistical evidence to conclude that there is at least one mutation rate that is different between States 1 and 6. The results are shown in Appendix B, Figure B.5. To better understand which rates differ, we calculate simultaneous 95% confidence intervals. A confidence interval that contains 0 indicates that

there is not a significant difference between State 1 and State 6. The intervals are calculated using the equation in Figure B.6. The point estimates of the contrasts are shown in Figures B.8 through B.11.

The simultaneous confidence intervals of the State 1 and State 6 differences in each mutation rate are tabulated in Table B.2. From the table, we see that there is a significant difference in small insertion, deletion, and nucleotide substitution rates with State 6 having larger rates than State 1. There is no significant difference in microsatellite repeat number alteration rate between the two states.

Question 3:

For this portion of the analysis, we look at relating the four mutation rates with the six non-state genomic factors. Since the measures are on the same 1Mb of genes, we relate them using CCA. First, we look at the statistically significant dimensionality of the data, with an upper bound of four from the mutation rates. The resulting ANOVA (analysis of variance) table is shown in Table C.2.

The results show that only the first dimension of the data is significant. We now look at the standardized canonical coefficients of the first dimension to see which of the factors and mutation rates have the strongest impact on one another. For CCA, we focus on the absolute value of the coefficients, with the largest value signifying the strongest effect. The factors nGCm and LINE dominate the genomic predictor variables and insertion rate dominates the mutation rate response variables. These values are included below with the full result in Table C.3.

Largest Standardized Canonical Coefficients

NGCM = -0.658 LINE = 0.544 ins.std = 0.600

Since we have reduced the question to a single response variable, we use an OLS regression to confirm the CCA results that the nCGm and LINE are significant predictors of the insertion rate. An OLS using the full predictor set (see Table C.4) supports that of the six genomic factors, only nCGm and LINE are statistically significant predictors. Reducing the OLS to only these two factors (Table C.5) holds the result in significance and provides additional information. We note that while the genetic factors achieve statistical significance, the R^2 value is extremely low at 0.047. Meaning only 4.7% of the variance in insertion rate is explained by these two variables.

4.0 - CONCLUSIONS

4.1 - RECOMMENDATIONS

The pair-wise Pearson Correlation Coefficients and PCA loading vectors show statistically significant relationships between the three mutation rates small insertion, deletion, and nucleotide substitution. There was weak or no statistically significant relationship between the microsatellite repeat number alteration rate and the other mutation rates.

There is a significant difference of at least one pair of mutation rates for at least one pair of States. Furthermore, there is a significant difference between small insertion, deletion, and nucleotide substitution rates between State 1 and 6, with State 6 having the larger rates. There is no significant difference in microsatellite repeat number alteration rate between State 1 and 6.

CCA shows significant relation between the nCGm and LINE genomic factors and the average insertion mutation rate. OLS regression supports this result, but also reveals these predictors do not explain a substantial portion of the variance in the mutation rate.

4.2 - CONSIDERATIONS

Our team had the standardized rates given to us without a hand in that process. Analysis of the original data set may reveal information that is not as readily apparent post-standardization.

While considered empirically normal, our data for the deletion and substitution mutation rates did not pass any of the three statistical tests for normality (see Figure A.4 and A.5). Some results may differ if using more normal data sets or applying a normalizing transform to accommodate. This has applications in all parts of the analysis.

Our analysis is limited to the plausible explanatory genomic features the researchers selected. It is possible that other features may interact with the mutation rates differently that were not considered and may influence the canonical bases-pairing covariance. More research is needed to draw full conclusions.

5.0 - RESOURCES

Thank you for the opportunity to contribute to this research. Please contact the consultants with additional questions or comments.

Appendix (EDA)

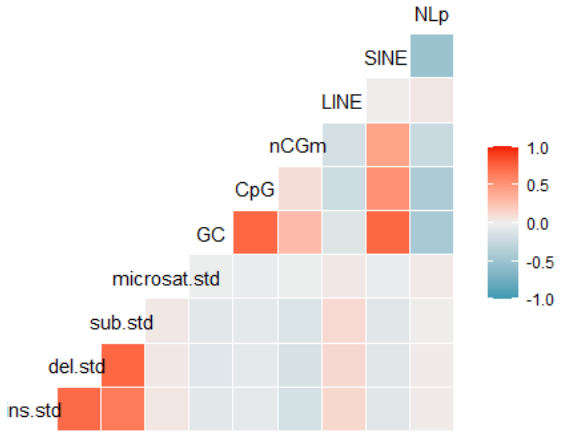

Figure 1: Correlation Heat Map of Variables

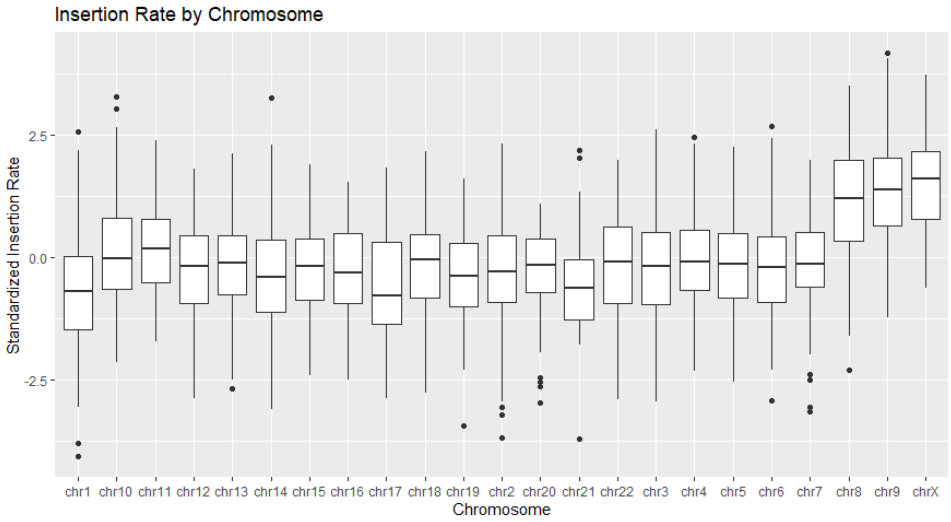

Figure 2: Insertion Rate by Chromosome

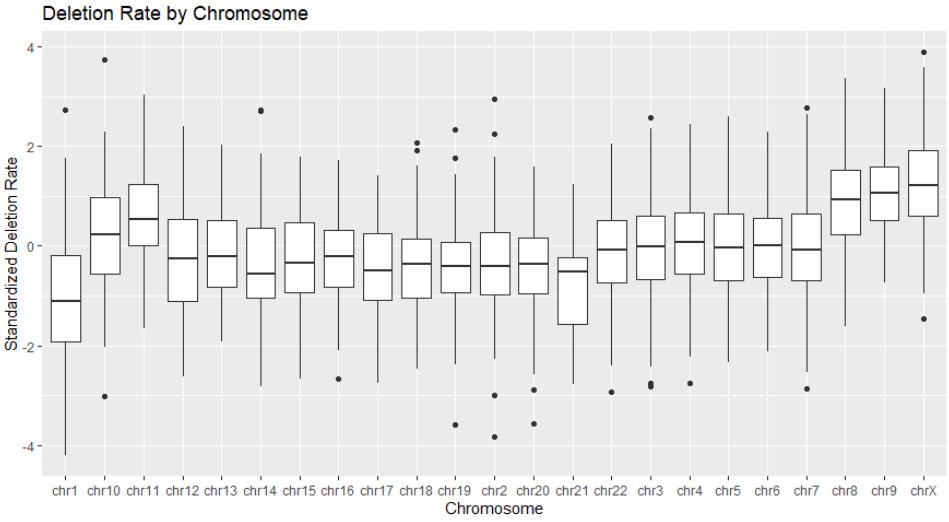

Figure 3: Deletion Rate by Chromosome

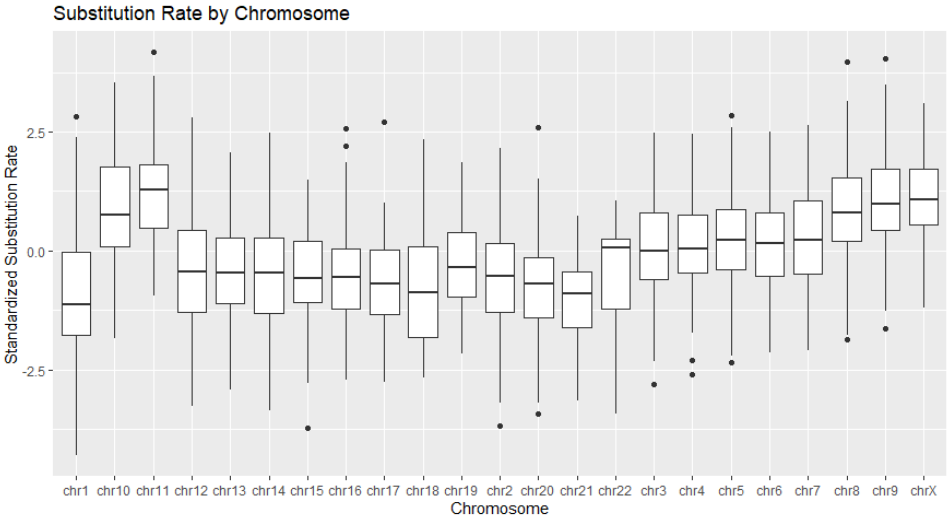

Figure 4: Substitution Rate by Chromosome

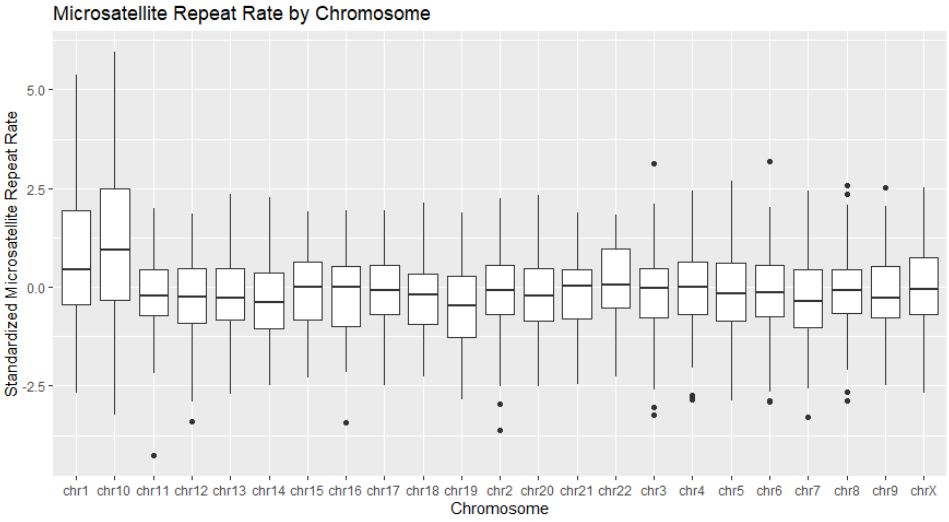

Figure 5: Microsatellite Repeat Alteration Rate by Chromosome

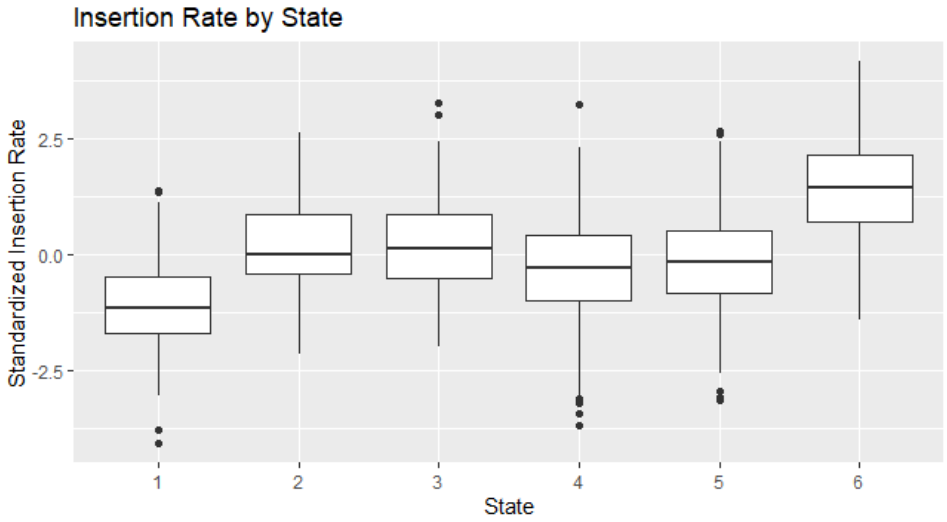

Figure 6: Insertion Rate by Genetic Divergence State

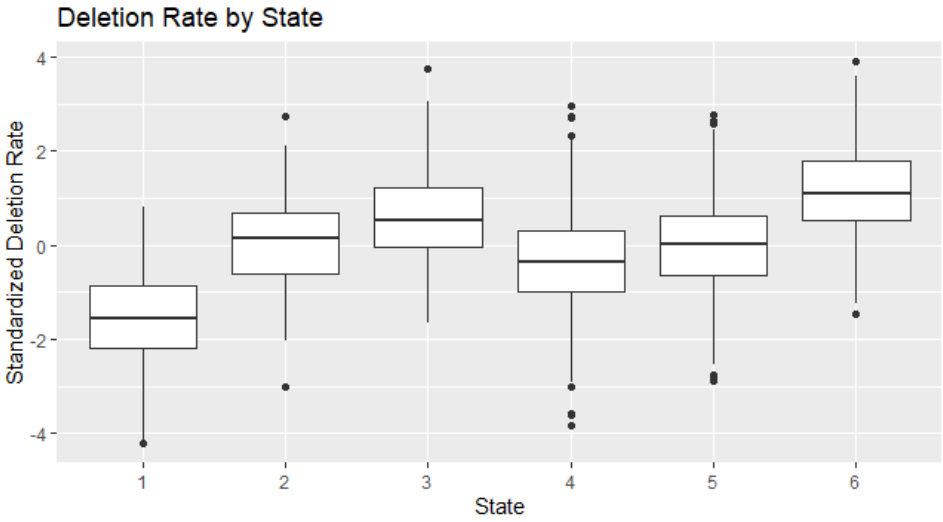

Figure 7: Deletion Rate by Genetic Divergence State

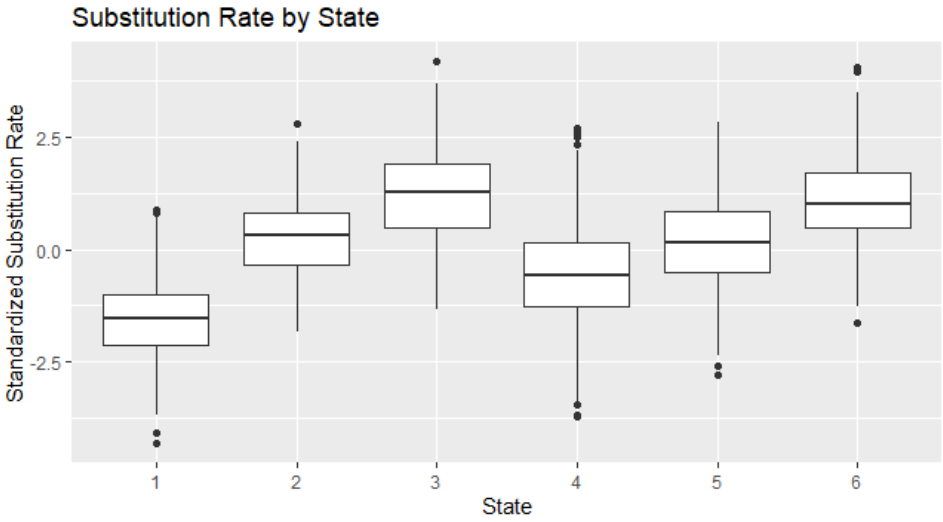

Figure 8: Substitution Rate by Genetic Divergence State

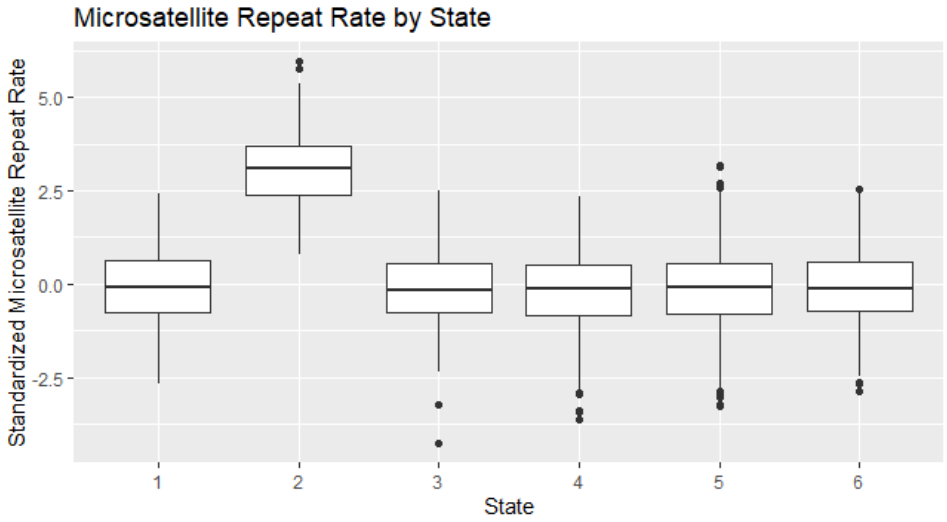

Figure 9: Microsatellite Repeat Alteration Rate by Chromosome

Variable	Description	Data Type
Chr	The Chromosome number to which the 1Mb window belong	categorical
Start	The starting position of the 1Mb window along the human genome	continuous
End	The end position of the 1Mb window along the human genome	continuous
ins.std	Standardized small insertion rates	continuous
del.std	Standardized small deletion rates	continuous
sub.std	Standardized nucleotide substitution rates	continuous
microsat.std	Standardized microsatellite repeat number alterations	continuous
GC	GC content	continuous
CpG	Number of CpG islands	continuous
nCGm	Number of non-CpG methyl-cytosines	continuous
LINE	Number of LINE elements	continuous
SINE	Number of SINE elements	continuous
NLp	Number of nuclear lamina associated regions	continuous
State	State of genetic divergence in the human genome. 1 = state 1, 2 = state 2, etc.	categorical

Table 1: Variable Description and Type

A - Appendix A (Question 1)

Table A.1 Hypothesis Test

Test #	Null Hypotheses (H_0)	Alternative Hypotheses (H_a)
1.1-6	The correlation coefficients for each pair of mutation rates are equal to zero. i.e. $\rho_{X_i,X_j} = 0 \; \forall i,j$	There is at least one pair of mutation rates with a nonzero correlation coefficient.

Table A.2: Pairwise Correlation Testing Results

Mutation Types	Sample correlation	p = Prob > |r|
Insertion, Deletion	0.72884	< 0.0001
Insertion, Nucleotide Substitution	0.64455	< 0.0001
Insertion, Microsatellite Repeat Number Alterations	0.04508	0.0184
Deletion, Nucleotide Substitution	0.74026	< 0.0001
Deletion, Microsatellite Repeat Number Alterations	0.04071	0.0333
Nucleotide Substitution, Microsatellite Repeat Number Alterations	0.04190	0.0284

Pearson Correlation Coefficients, N = 2735
Prob > |r| under H0: Rho=0

	insstd	delstd	substd	microsatstd
insstd	1.00000	0.72884 <.0001	0.64455 <.0001	0.04508 0.0184
delstd	0.72884 <.0001	1.00000	0.74026 <.0001	0.04071 0.0333
substd	0.64455 <.0001	0.74026 <.0001	1.00000	0.04190 0.0284
microsatstd	0.04508 0.0184	0.04071 0.0333	0.04190 0.0284	1.00000

Figure A.1: Hypothesis Test 1.1-6 Results

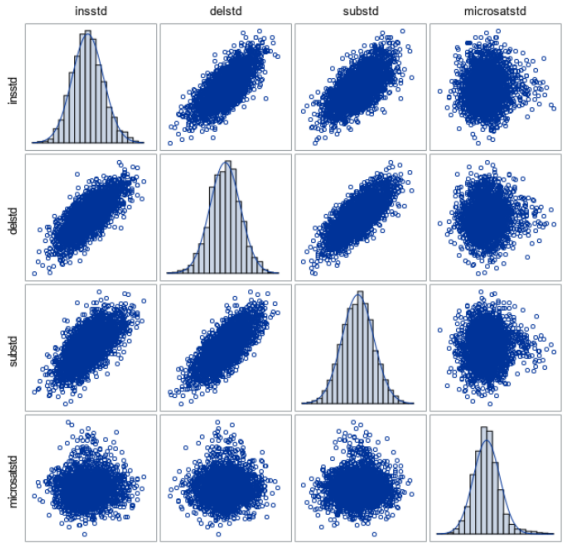

Figure A.2: Scatterplot Matrix with Histograms

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.022909	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.29093	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.947789	Pr > A-Sq	<0.0050

Figure A.3: Normality Test Results for ins.std

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.013464	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.061714	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.462785	Pr > A-Sq	>0.2500

Figure A.4: Normality Test Results for del.std

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.011472	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.031254	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.171112	Pr > A-Sq	>0.2500

Figure A.5: Normality Test Results for sub.std

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.046887	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.764299	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	12.41414	Pr > A-Sq	<0.0050

Figure A.6: Normality Test Results for microstat.std

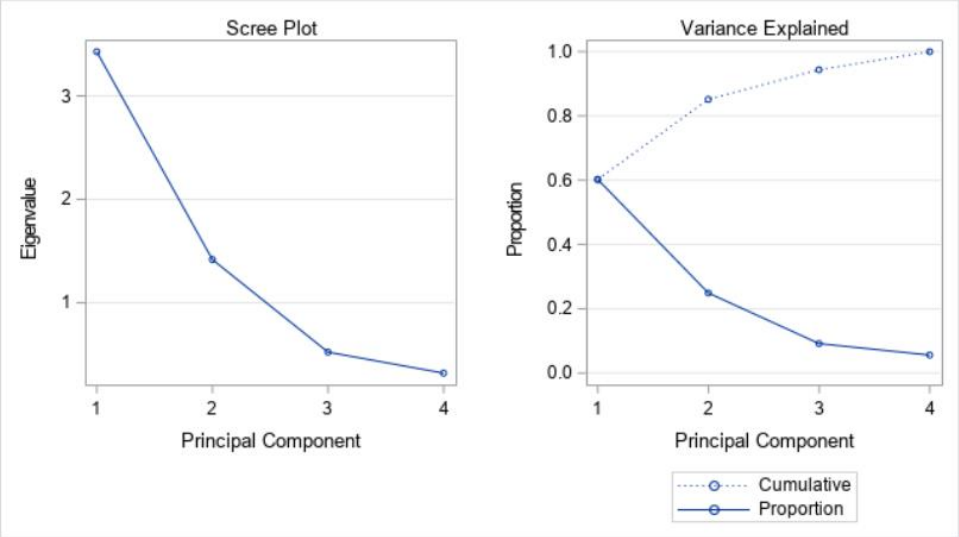

Figure A.7: PCA Scree and Cumulative Contribution Plots

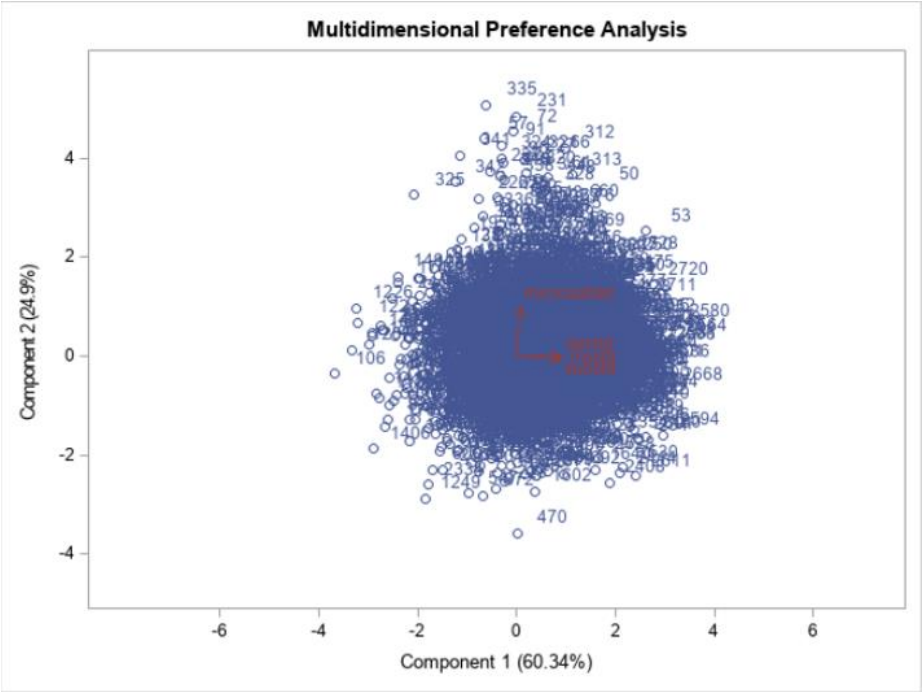

Figure A.8: PCA Biplot

```
options ls=78 NODATE;

data AR1;
   infile "C:\Users\arielle.thibeault\Desktop\STAT 580\Project 1\humanAR_1mbData.csv" firstobs=2 delimiter=',';
   input chr $ start end insstd delstd substd microsatstd state GC CpG nCGm LINE SINE NLp;
   run;

PROC PRINT data = AR1;
RUN;

/* Subsetting our dataset to just the four mutation rates
 * of interest in our analysis.
 */

data AR1_sub;
SET AR1;
DROP chr start end state GC CpG nCGm LINE SINE NLp;
RUN;

/* Some EDA on covariance of the bivariate pairs.
 */
PROC CORR DATA = AR1_sub OUTP = pearson_corr COV;
   VAR insstd delstd substd microsatstd;
RUN;

PROC CORR DATA = AR1_sub OUTP = pearson_corr;
   VAR insstd delstd substd microsatstd;
RUN;

PROC SGSCATTER DATA=AR1_sub;
     MATRIX insstd delstd substd microsatstd /
     DIAGONAL = (histogram normal);
RUN;

/* Anderson Darling test for Normality: two variables pass,
 * the rest are close enough that empirically we will consider
 * them to be normal.
 */
PROC UNIVARIATE data=AR1_sub NORMAL;
VAR insstd delstd substd microsatstd;
HISTOGRAM insstd delstd substd microsatstd;
RUN;
```

```
/* The princomp procedure performs pca on the AR1_sub data.
 * The cov option specifies results are calculated from the covariance
 * matrix, instead of the default correlation matrix.
 */

proc princomp data=AR1_sub cov out=a;
   var insstd delstd substd microsatstd;
   run;

/* The cov procedure is used to calculate pairwise correlations
 * between the first 2 principal components and the original variables.
 */

proc corr data=a;
   var prin1 prin2 insstd delstd substd microsatstd;
   run;

/* The gplot procedure is used to plot the first 2 principal components.
 * axis1 and axis2 options set the plotting window size,
 * and these are then set to vertical and horizontal axes, respectively.
 */

proc gplot data=a;
   axis1 length=5 in;
   axis2 length=5 in;
   plot prin2*prin1 / vaxis=axis1 haxis=axis2;
   run;

/*Biplot*/
proc prinqual data=AR1_sub plots=(MDPref)
              n=2       /* project onto Prin1 and Prin2 */
              mdpref=1; /* use COV scaling */
   transform identity(insstd delstd substd microsatstd);  /* identity transform */
   ods select MDPrefPlot;
run;

proc factor data = AR1_sub cov scree ev method = principal;
var insstd delstd substd microsatstd;
run;
```

Figure A.9: Correlation Testing and PCA Code in SAS

For random variables X and Y, their pair-wise correlation is defined as:

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where:

- $cov(X,Y)$ is the covariance between X and Y,
- σ_X is the standard deviation (or square root of the variance) of X,
- σ_Y is the standard deviation of Y.

Theory A.10: Correlation Definition

Under the null hypothesis, pairs of variables from an uncorrelated bivariate normal distribution follow the Student's t-distribution with n – 2 degrees of freedom.

$$t = \frac{r}{\sigma_r} \sim t_{n-2}$$

Where:

- r is the sample correlation between the two variables,
- σ_r is the standard error associated with the correlation.

Theory A.11: Student's t-distribution Testing

B - Appendix B (Question 2)

Table B.1 Hypothesis Tests

Test #	Null Hypotheses (H_0)	Alternative Hypotheses (H_a)
2.1	There is no difference of the four mutation rates between starting location of the 1 Mb window	There is at least one pair of windows with at least one different mutation rate
2.2	There is no difference of the four mutation rates between the interaction of the starting location and State	There is at least one pair of mutation rates that are different for the interaction between one pair of windows and States
2.3	There is no difference of the four average mutation rates between all states	There is at least one pair of states with at least one different average mutation rate
2.4	There are no differences in mutation rates between States 1 and 6	There is at least one mutation rate that differs from State 1 and 6
2.5	There are homogenous variance-covariance matrices	The variance-covariance matrices are not homogenous

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall start Effect
H = Type III SSCP Matrix for start
E = Error SSCP Matrix

S=1 M=1 N=1359

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.99942101	0.39	4	2720	0.8131
Pillai's Trace	0.00057899	0.39	4	2720	0.8131
Hotelling-Lawley Trace	0.00057933	0.39	4	2720	0.8131
Roy's Greatest Root	0.00057933	0.39	4	2720	0.8131

Figure B.1: Hypothesis Test 2.1 Results

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall start*state Effect
H = Type III SSCP Matrix for start*state
E = Error SSCP Matrix

S=4 M=0 N=1359

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.99392244	0.83	20	9022.2	0.6788
Pillai's Trace	0.00608884	0.83	20	10892	0.6784
Hotelling-Lawley Trace	0.00610338	0.83	20	5976.6	0.6790
Roy's Greatest Root	0.00306265	1.67	5	2723	0.1389

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Figure B.2: Hypothesis Test 2.2 Results

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall state Effect
H = Type III SSCP Matrix for state
E = Error SSCP Matrix

S=4 M=0 N=1359

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.77580460	35.88	20	9022.2	<.0001
Pillai's Trace	0.24152684	35.00	20	10892	<.0001
Hotelling-Lawley Trace	0.26717126	36.32	20	5976.6	<.0001
Roy's Greatest Root	0.14481616	78.87	5	2723	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Figure B.3: Hypothesis Test 2.3 Results (Full Model)

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall state Effect
H = Type III SSCP Matrix for state
E = Error SSCP Matrix

S=4 M=0 N=1362

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.37633076	154.92	20	9042.1	<.0001
Pillai's Trace	0.82461551	141.74	20	10916	<.0001
Hotelling-Lawley Trace	1.17499773	160.09	20	5989.8	<.0001
Roy's Greatest Root	0.57186306	312.12	5	2729	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Figure B.4: Hypothesis Test 2.3 Results (Reduced Model / State Only)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall State 1 - State 6 Effect
H = Contrast SSCP Matrix for State 1 - State 6
E = Error SSCP Matrix

S=1 M=1 N=1362

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.92995729	51.33	4	2726	<.0001
Pillai's Trace	0.07004271	51.33	4	2726	<.0001
Hotelling-Lawley Trace	0.07531820	51.33	4	2726	<.0001
Roy's Greatest Root	0.07531820	51.33	4	2726	<.0001

Figure B.5: Hypothesis Test 2.4

Table B.2: 95% Simultaneous Confidence Intervals for State 1 – State 6

Mutation Type	**95% Confidence Interval**
Small Insertion	(-1.066, -0.535)
Deletion	(-1.433, -0.918)
Nucleotide Substitution	(-1.222, -0.683)
Microsatellite Repeat Number Alteration	(-0.198, 0.342)

$$\hat{\Psi}_j \pm \sqrt{\frac{p(N-g)}{N-g-p+1} F_{p,N-g-p+1}} SE(\hat{\Psi}_j)$$

Where,

$\hat{\psi}_j$ = contrast point estimate of the jth mutation rate

p = number of mutation rates, 4

g = total number of states, 6

N = total number of observations, 2735

$F_{p,\ N\text{-}g\text{-}p+1}$ = F-distribution with p and N-g-p+1 degrees of freedom in the numerator and denominator, respectively

$SE(\hat{\psi}_j)$ = the standard error of the contrast of the jth mutation rate

Figure B.6: Simultaneous Confidence Interval Equation

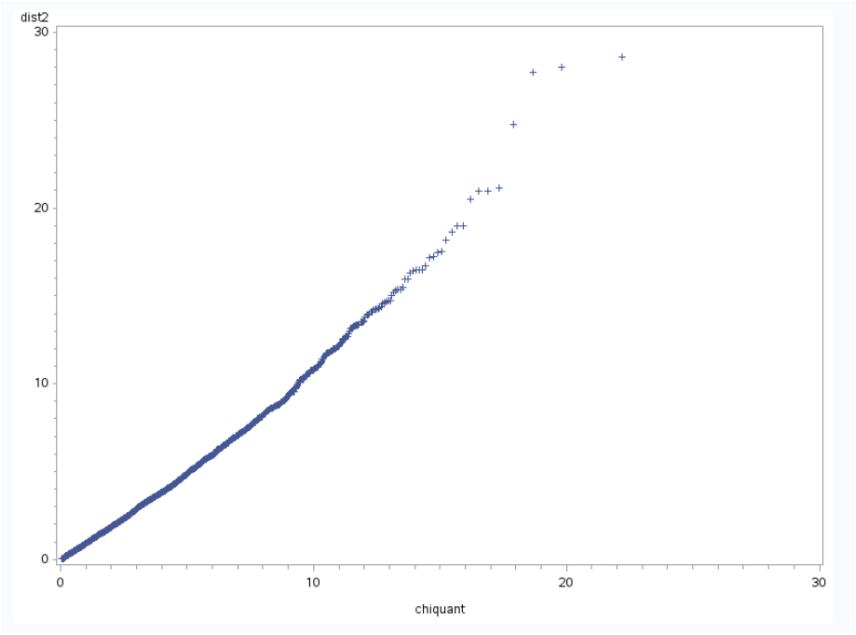

Figure B.7: Mahalanobis Distance^2 vs χ^2 Quantile Plot

Parameter	Estimate	Standard Error	t Value	Pr > |t|
State 1 - State 6	-0.80044145	0.08601374	-9.31	<.0001

Figure B.8: Small Insertion Rate Contrast

Parameter	Estimate	Standard Error	t Value	Pr > |t|
State 1 - State 6	-1.17559333	0.08350965	-14.08	<.0001

Figure B.9: Small Deletion Rate Contrast

Parameter	Estimate	Standard Error	t Value	Pr > |t|
State 1 - State 6	-0.95255066	0.08724750	-10.92	<.0001

Figure B.10: Nucleotide Substitution Rate Contrast

Parameter	Estimate	Standard Error	t Value	Pr > |t|
State 1 - State 6	0.07184957	0.08746638	0.82	0.4115

Figure B.11: Microsatellite Repeat Alteration Rate

```sas
data genome;

    infile "/home/u59993780/sasuser.v94/STAT 580 Project 1/humanAR_1mbData.csv" dlm = ',' firstobs = 2;

    input chr $ start end ins_std del_std sub_std microsat_std state GC cpG nCGm LINE SINE NLp;

run;


proc glm data = genome;

    class state;

    model ins_std del_std sub_std microsat_std = state start state*start;


    manova h = state start state*start / printe printh;

run;
```

Figure B.12: Full Model MANOVA Code in SAS

```sas
data genome;

    infile "/home/u59993780/sasuser.v94/STAT 580 Project 1/humanAR_1mbData.csv" dlm = ',' firstobs = 2;

    input chr $ start end ins_std del_std sub_std microsat_std state GC cpG nCGm LINE SINE NLp;

run;

/* MANOVA */

proc glm data = genome;

    class state;

    model ins_std del_std sub_std microsat_std = state;

    contrast 'State 1 - State 6' state 1 0 0 -1;
    estimate 'State 1 - State 6' state 1 0 0 -1;


    manova h = state / printe printh;

run;


/* Bartlett's test for homogoneous variance-covariance matrices*/
proc discrim data = genome pool = test;

    class state;

    var ins_std del_std sub_std microsat_std;

run;


/*Mahalanobis distance plot for Normality */
proc princomp std out = pcresult;

    var ins_std del_std sub_std microsat_std;

run;

data mahal;
    set pcresult;

    dist2=uss(of prin1-prin4);
run;

proc sort;
    by dist2;
run;

data plotdata;

    set mahal;
    prb = (_n_ -0.5)/2735;
    chiquant = cinv(prb,4);
run;

proc gplot;
```

Figure B.13: State Model MANOVA Code in SAS

C - Appendix C (Question 3)

Table C.1: Hypothesis Tests for Question 3

Test #	Null Hypotheses (H_0)	Alternative Hypotheses (H_a)
3.1	Canonical correlations are zero for dimension n and all following	Canonical correlations are not zero
3.2	The genomic factor is not a significant predictor of insertion rate when other factors are considered	The genomic factor is a significant predictor of insertion rate when other factors are considered
3.3	NGCm and LINE are not significant predictors of the insertion rates	NGCm and LINE are significant predictors of the insertion rates

Dimension	Canonical Corr.	F	Df_1	Df_2	p
1	0.2119	5.98	24	9508	0.000
2	0.0578	1.16	15	7526	0.297
3	0.0436	1.03	8	5454	0.410
4	0.0334	1.02	3	2728	0.384

Table C.2: Statistical Testing for CCA Dimensionality

Table C.3: Standardized Canonical Coefficients for the First Dimension

Genomic Factors		Mutation Rates	
GC	0.031	ins	0.6002
CpG	-0.136	del	0.2654
nCGm	-0.658	sub	0.2310
LINE	0.544	microstat	0.0811
SINE	-0.240		
NLp	-0.235		

Table C.4: OLS of Insertion Mutation Rate by the Six Genomic Factors

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.54e-01   3.88e-01   -0.40    0.691
nCGm        -1.23e-04   1.82e-05   -6.73  2.1e-11 ***
LINE         1.65e-03   3.22e-04    5.11  3.4e-07 ***
GC           5.36e-03   9.60e-03    0.56    0.577
CpG         -2.96e-03   2.60e-03   -1.14    0.255
SINE        -1.38e-04   1.17e-04   -1.18    0.239
NLp         -1.83e-04   1.00e-04   -1.83    0.068 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table C.5: OLS of Insertion Mutation Rate by nCGm and LINE only

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.44e+00   5.06e-01   -4.83  1.4e-06 ***
nCGm         3.15e-04   9.17e-05    3.44  0.00059 ***
LINE         6.24e-03   9.81e-04    6.36  2.4e-10 ***
nCGm:LINE   -8.80e-07   1.80e-07   -4.88  1.1e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.15 on 2731 degrees of freedom
Multiple R-squared:  0.047,     Adjusted R-squared:  0.046
```

```
#Midterm Project for STAT 580
#Charles Cratty
#Human Genome Question #3

#set working drive and load libraries
setwd("D:/STAT 580 - Practicum Part 1/Midterm Project")
library(ggplot2)
library(GGally)#For ggcorr() function
library(CCA)
library(CCP)

#Set output options
options(digits=3)

#Load data
genomeData_Raw = read.csv("humanAR_lmbData.csv", header=TRUE, na.strings="?")
dim(genomeData_Raw)
#Per instruction, no cleaning of the data is expected


names(genomeData_Raw)

genomeData = genomeData_Raw
genomeData$chr = as.factor(genomeData$chr)
genomeData$state = as.factor(genomeData$state)

summary(genomeData)

attach(genomeData)


#Box plot of rates by chromosome
ins_Boxplot = ggplot(data=genomeData, aes(y=ins.std, x=chr))
ins_Boxplot + geom_boxplot() + labs(title="Insertion Rate by Chromosome",
y="Standardized Insertion Rate", x="Chromosome")

del_Boxplot = ggplot(data=genomeData, aes(y=del.std, x=chr))
del_Boxplot + geom_boxplot() + labs(title="Deletion Rate by Chromosome",
y="Standardized Deletion Rate", x="Chromosome")

sub_Boxplot = ggplot(data=genomeData, aes(y=sub.std, x=chr))
sub_Boxplot + geom_boxplot() + labs(title="Substitution Rate by Chromosome",
y="Standardized Substitution Rate", x="Chromosome")

micro_Boxplot = ggplot(data=genomeData, aes(y=microsat.std, x=chr))
micro_Boxplot + geom_boxplot() + labs(title="Microsatellite Repeat Rate by
Chromosome", y="Standardized Microsatellite Repeat Rate", x="Chromosome")


#Box plot of rates by State
ins_Boxplot = ggplot(data=genomeData, aes(y=ins.std, x=state))
ins_Boxplot + geom_boxplot() + labs(title="Insertion Rate by State",
y="Standardized Insertion Rate", x="State")
```

```
del_Boxplot = ggplot(data=genomeData, aes(y=del.std, x=state))
del_Boxplot + geom_boxplot() + labs(title="Deletion Rate by State",
y="Standardized Deletion Rate", x="State")

sub_Boxplot = ggplot(data=genomeData, aes(y=sub.std, x=state))
sub_Boxplot + geom_boxplot() + labs(title="Substitution Rate by State",
y="Standardized Substitution Rate", x="State")

micro_Boxplot = ggplot(data=genomeData, aes(y=microsat.std, x=state))
micro_Boxplot + geom_boxplot() + labs(title="Microsatellite Repeat Rate by
State", y="Standardized Microsatellite Repeat Rate", x="State")

#Remove variables that won't be used going forward
rm(ins_Boxplot, del_Boxplot, sub_Boxplot, micro_Boxplot, genomeData_Raw)

#Since Q3 is interested inrelationships between a set of response variables
and a set of predictor variables, we will use Canonical Correlation Analysis

responses = genomeData[,4:7]
predictors = genomeData[,9:14]

#Create Pairwise plots of each set
ggpairs(responses)
ggpairs(predictors)

#Correlation matrices between and within each set
matcor(predictors, responses)
#Heatmap of the correlation matrix
ggcorr(c(predictors, responses))

#CCA Calculations
cc1 <- cc(predictors, responses)
names(cc1)
#Display the canonical correlations
cc1$cor
#All are less than 0.22

#Raw Canonical Coefficients
cc1$xcoef
#GC has the most impact (on 2 and 3) all other responses have ABS() <=0.01
(1%)
cc1$ycoef


#Compute canonical loadings
cc2 <- comput(predictors, responses, cc1)
names(cc2)
#Display canonical loadings
cc2[3:6]

# tests of canonical dimensions
rho <- cc1$cor
## Define number of observations, number of variables in first set, and number
of variables in the second set.
```

```
n <- dim(predictors)[1]
p <- length(predictors)
q <- length(responses)

## Calculate p-values using the F-approximations of different test statistics:
p.asym(rho, n, p, q, tstat = "Wilks")
p.asym(rho, n, p, q, tstat = "Hotelling")
p.asym(rho, n, p, q, tstat = "Pillai")
#Only the 1 to 4 is statistically significant in these three tests

# standardized psych canonical coefficients diagonal matrix of predictor sd's
s1 <- diag(sqrt(diag(cov(predictors))))
s1 %*% ccl$xcoef

# standardized acad canonical coefficients diagonal matrix of response sd's
s2 <- diag(sqrt(diag(cov(responses))))
s2 %*% ccl$ycoef


#Testing out OLS for the strongest results

#Simple Graphs
ggplot(data = genomeData, aes(x=nCGm, y=ins.std)) + geom_point(size=2)
ggplot(data = genomeData, aes(x=nCGm^2, y=ins.std)) + geom_point(size=2)
ggplot(data = genomeData, aes(x=LINE, y=ins.std)) + geom_point(size=2)
ggplot(data = genomeData, aes(x=LINE^2, y=ins.std)) + geom_point(size=2)
#None of these graphs strongly suggest a linear relationship


full_model <- lm(ins.std ~ nCGm + LINE + GC + CpG + SINE + NLp,
data=genomeData)
summary(full_model)
#Full model supports the CCA in that nCGm and LINE are the only significant
factors

reduced_model <- lm(ins.std ~ nCGm + LINE + nCGm*LINE, data=genomeData)
summary(reduced_model)
#Strong p-values, but very low R^2 value. Strong proof of relationship, but
still a high degree of variability about the regression line.
```

Figure C.6: R Software Code for EDA and Question 3