Lactose Tolerance Exploration

Research Motivation

The data collected was extracted from opensnp.org which is public to anyone. The data was extracted and was preprocessed in our legacy code that is able to accomplish the following:
1. Preprocessing the genomic data - This step converts user genotypes at each SNP to a mutation count.
2. Building the model - This step uses the preprocessed data to build a model to predict phenotype.
3. Using the model - This step uses the model from the previous step to predict phenotype for additional users.

By using elastic net, this model had the highest accuracy with 92% and it provide us with the top candidates (Rsid) for the phenotype that is lactose tolerance we are currently researching for a specific gene and ethnic group.

Context and SNPs of Interest

- **Lactose intolerance** is a condition in which people have digestive symptoms—such as bloating, diarrhea, and gas—after eating or drinking milk or milk products.
- **Lactose** is a sugar found in milk and milk products. It breaks down into two simpler forms of sugar: glucose and galactose by an enzyme called **lactase** which produced in the small intestines.
- The **LCT gene** provides instructions for making the enzyme lactase.
- For some people, the production of the lactase enzyme stops (caused by gradually decreasing activity (expression) of the LCT gene after infancy) when they become an adult, driven by a genetic variation near the LCT gene.
- The ability to produce lactase in adulthood depends on SNPs within the **MCM6 gene.**
- MCM6 lies upstream of the LCT gene and a specific DNA sequence within the MCM6 gene called a regulatory element helps control the activity (expression) of the LCT gene.
- The percentage of the population with genetic variations differs quite a bit among people with different backgrounds.
- The percentage of the population with lactose intolerance differs based on the backgrounds. Adult Caucasian populations are likely to produce lactase. While the majority of Asian populations do not produce lactase as an adult.

Figure 1: Lactose Tolerance Facts

SNP of interest

MCM6 Variant	Allele	Populations	Lactase Persistence
rs4988235 & rs182549	C	The ancestral "Wild type" Common in Asian and African ancestry	CC Likely to be lactose intolerant in adulthood
	T	Common in European Caucasian ancestry populations	CT or TT Likely to digest dairy in adulthood.

References:

1. https://ghr.nlm.nih.gov/condition/lactose-intolerance
2. https://ghr.nlm.nih.gov/gene/MCM6#conditions
3. https://www.geneticlifehacks.com/lactose-intolerance/
4. https://www.toolboxgenomics.com/blog/snp-highlight-lactose-intolerance-mcm6/
5. https://www.snpedia.com/index.php/Lactose_intolerance

Figure 2: MCM6 Variant Table

Top 20 coefficients (SNPs) legacy code capture using Elastic Net

	A	B
1	intercept: -4.021494625991822	
2		
3	main effects:	
4	feature	coefficient
5	gene_105374595_LOC105374595_rs6545107	-77.167356
6	gene_9378_NRXN1_rs988179	-73.094943
7	gene_256987_SERINC5_rs7712447	-69.483344
8	gene_105370576_LOC105370576_rs4903435	65.1401132
9	gene_55843_ARHGAP15_rs10178148	-64.251604
10	gene_9223_MAGI1_rs4396824	-64.12448
11	gene_84952_CGNL1_rs1995990	-62.455192
12	gene_1608_DGKG_rs1558910	-62.447148
13	gene_2045_EPHA7_rs3799807	59.5539707
14	gene_638_BIK_rs4988406	59.3999856
15	gene_4088_SMAD3_rs17293443	59.0636884
16	gene_55084_SOBP_rs9386654	-58.435392
17	gene_221938_MMD2_rs10229311	-57.850042
18	gene_105377864_LOC105377864_rs12209650	57.7589239
19	gene_2131_EXT1_rs11562695	57.6630297
20	gene_4175_MCM6_rs182549	57.1057324
21	gene_107984687_LOC107984687_5732_PTGER2_rs1254601	57.0868164
22	gene_4175_MCM6_rs4988235	57.0340293
23	gene_105371531_LOC105371531_rs9914374	54.9289899
24	gene_5583_PRKCH_rs1088682	-54.89943
25	gene_107984719_LOC107984719_145814_PGPEP1L_rs2715423	53.9565203

Figure 3: Top 20 Coefficients

Data distribution

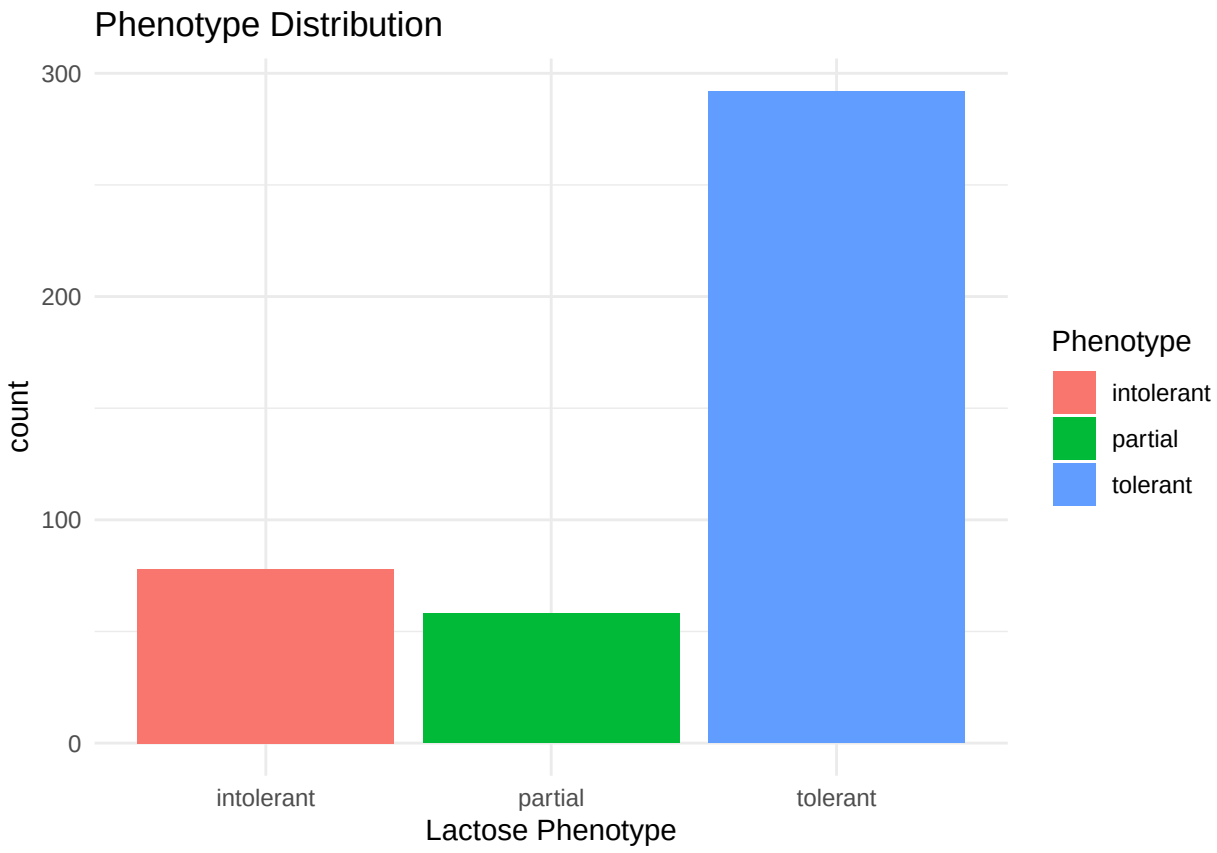
```
dim(original_data)
```

```
## [1] 428 972
```

```
table(original_data$pheno)
```

```
##
## intolerant    partial   tolerant
##         78         58        292
```

```
p<-ggplot(original_data, aes(x=pheno, fill=pheno)) +
  geom_bar(stat="count")+theme_minimal() +
  ggtitle("Phenotype Distribution")
p <- p + labs(fill = "Phenotype") +  xlab("Lactose Phenotype")
p
```



EDA and Data Preprocessing For Top 20 Rsid

Steps applied were the following:
i) Drop rows that has users' phenotype as partial.
ii) Threshold to drop columns with more than 50% of NA and drop phenotype with partial label.
iii) Imputation by mode.

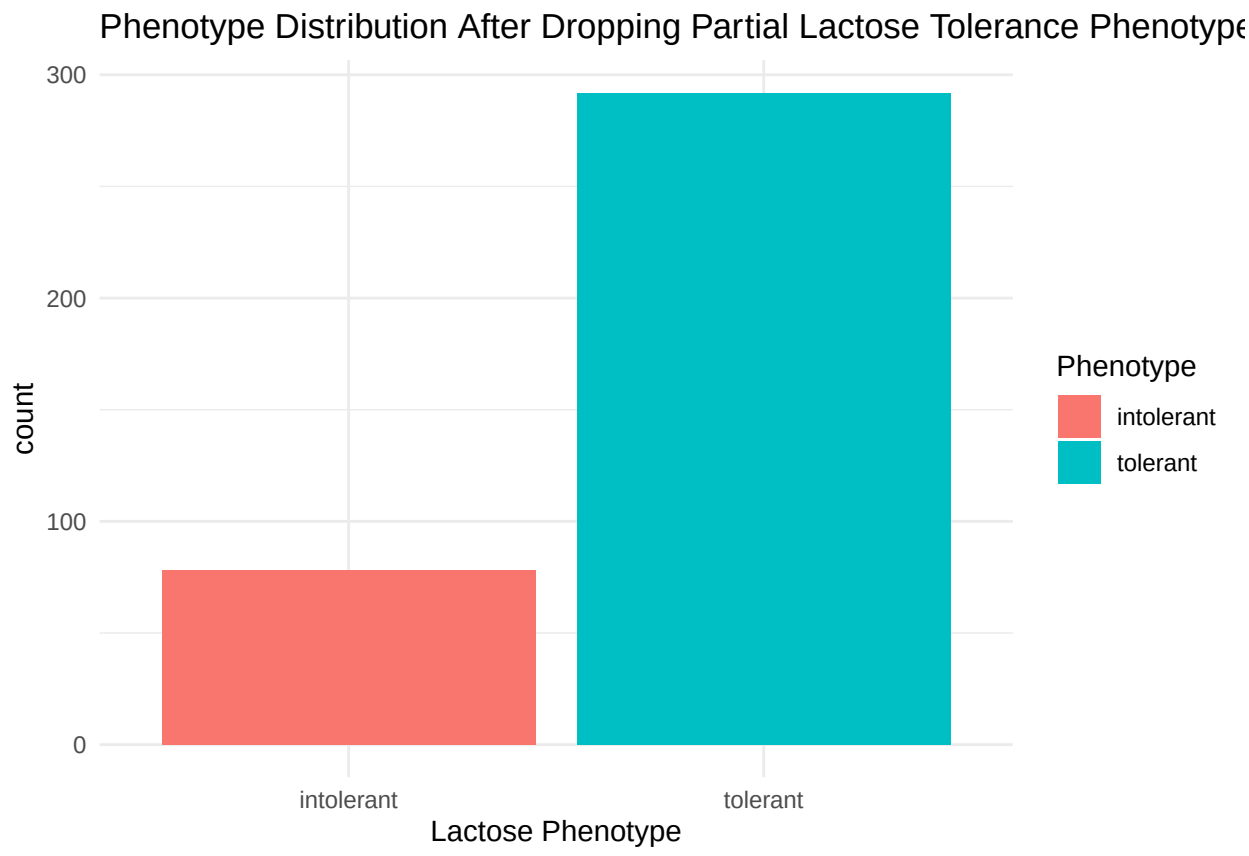
```
dim(original_data_50thres)
```

```
## [1] 370 966
```

```
table(original_data_50thres$pheno)
```

```
##
## intolerant    partial   tolerant
##         78          0        292
```

```
p<-ggplot(original_data_50thres, aes(x=pheno, fill=pheno)) +
  geom_bar(stat="count")+theme_minimal() +
  ggtitle("Phenotype Distribution After Dropping Partial Lactose Tolerance Phenotype")
p <- p + labs(fill = "Phenotype") +  xlab("Lactose Phenotype")
p
```



Subset of Top 20 Rsid Dimensions

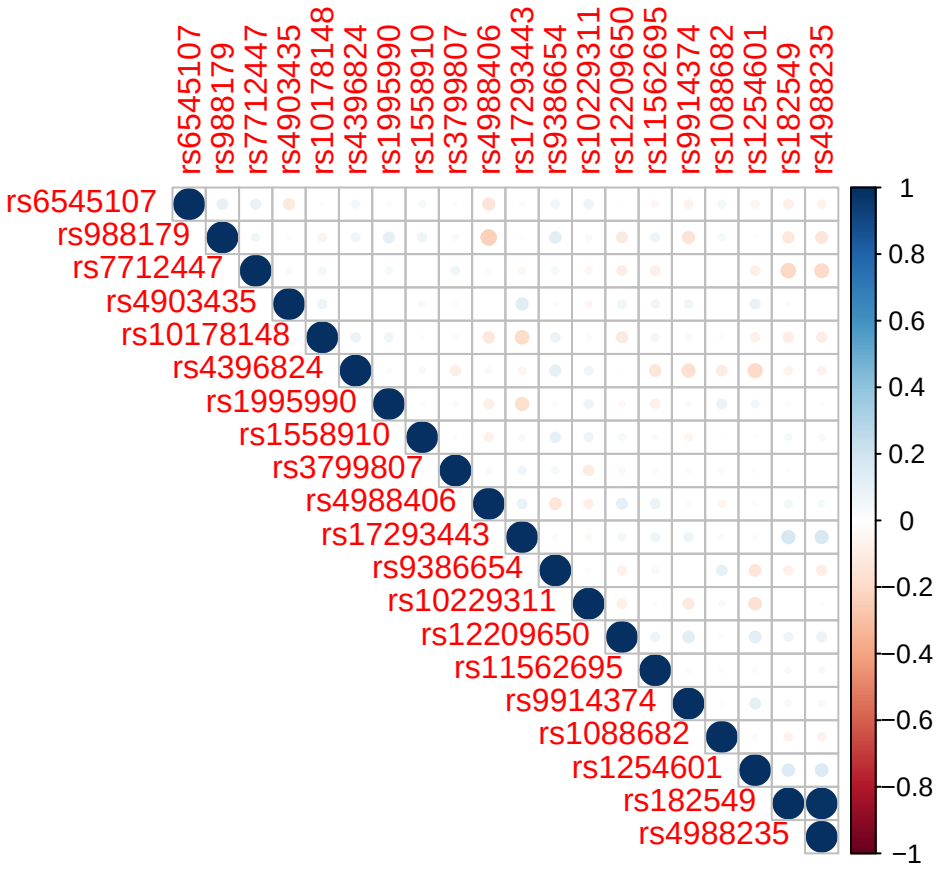
```
top_20 <- original_data_50thres[,c("rs6545107","rs988179", "rs7712447", "rs4903435","rs10178148",
          "rs4396824", "rs1995990", "rs1558910",
          "rs3799807", "rs4988406", "rs17293443", "rs9386654", "rs10229311",
          "rs12209650", "rs11562695", "rs9914374", "rs1088682", "rs1254601",
          "rs182549",  "rs4988235","pheno")]

dim(top_20)
```

```
## [1] 370  21
```

Correlation Plot

Below displayed we can observed that rs182549 and rs4988235 are highly correlated with a 0.985596271.

```
corrplot(cor(top_20[1:20], use = "pairwise.complete.obs"), type = "upper")
```



Correlation Info of rs182549 with the other Rsid

```
corr_data <- cor(top_20[1:20], use = "pairwise.complete.obs")
corr_data[,19]
```

```
##    rs6545107     rs988179    rs7712447    rs4903435   rs10178148    rs4396824
## -0.081303508 -0.121346964 -0.206785739 -0.016835273 -0.097105807 -0.066456333
##    rs1995990    rs1558910    rs3799807    rs4988406   rs17293443    rs9386654
## -0.020740134  0.042082190  0.011596778  0.057124243  0.170618259 -0.077917630
##   rs10229311   rs12209650   rs11562695    rs9914374    rs1088682    rs1254601
## -0.001603606  0.071089776  0.032269597  0.037566108 -0.064854586  0.145349545
##     rs182549    rs4988235
##  1.000000000  0.985596271
```

Correlation Info of rs4988235 with the other Rsid

```
corr_data[,20]
```

```
##    rs6545107     rs988179    rs7712447    rs4903435   rs10178148    rs4396824
## -0.063989849 -0.136480290 -0.192133758 -0.001564568 -0.097668925 -0.066406694
##    rs1995990    rs1558910    rs3799807    rs4988406   rs17293443    rs9386654
## -0.018451137  0.037359784  0.008506523  0.042423308  0.163235102 -0.095281865
##   rs10229311   rs12209650   rs11562695    rs9914374    rs1088682    rs1254601
##  0.013351234  0.087700331  0.034135577  0.033352795 -0.068802558  0.158221920
##     rs182549    rs4988235
##  0.985596271  1.000000000
```

PCA Analysis with Using Varimax Rotation

Factor rotations make the expression of a particular subspace simpler. Subspaces are smaller vector spaces within a R^n vector space. The orthogonal basis is rotated to align with the coordinate system. By making use of varimax rotation the orthogonal rotation produce that these factors are not correlated.Varimax rotation seeks to maximize the sum of the variance of the squared loadings, where these location means correlations between variables and factors.

PCA with Varimax Rotation, Rotatated Components = 2

```
pca_varimax <- principal(top_20[1:20], nfactors = 2, rotate = "varimax")
rotation2 <- data.frame(cbind(pca_varimax$score, pheno=top_20[,"pheno"]))
pca_varimax$loadings
```

```
##
## Loadings:
##            RC1    RC2
## rs6545107  -0.129  0.270
## rs988179   -0.135  0.543
## rs7712447  -0.250  0.324
## rs4903435         -0.383
## rs10178148 -0.277
## rs4396824  -0.165 -0.121
## rs1995990          0.346
## rs1558910          0.447
## rs3799807         -0.352
## rs4988406   0.199 -0.255
## rs17293443  0.286 -0.219
## rs9386654  -0.198
## rs10229311         0.108
## rs12209650  0.190 -0.295
## rs11562695        -0.330
## rs9914374         -0.383
## rs1088682          0.359
## rs1254601   0.283 -0.101
## rs182549    0.941
## rs4988235   0.941
##
##                  RC1   RC2
## SS loadings    2.266 1.707
## Proportion Var 0.113 0.085
## Cumulative Var 0.113 0.199
```

```
pca_varimax$Vaccounted
```

```
##                             RC1        RC2
## SS loadings           2.2661758 1.70681356
## Proportion Var        0.1133088 0.08534068
## Cumulative Var        0.1133088 0.19864947
## Proportion Explained  0.5703956 0.42960436
## Cumulative Proportion 0.5703956 1.00000000
```

Principal Component Regression (PCR) with 2 Rotated Components

```
linModel <- glm(pheno ~ ., data = rotation2, family = "binomial")
summary(linModel)
```

```
##
## Call:
## glm(formula = pheno ~ ., family = "binomial", data = rotation2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.62646  -0.50093  -0.25644  -0.08251   2.90396
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1968     0.2291  -9.590  < 2e-16 ***
## RC1          -1.3658     0.1976  -6.911 4.83e-12 ***
## RC2           1.5683     0.2141   7.326 2.37e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 381.12  on 369  degrees of freedom
## Residual deviance: 244.19  on 367  degrees of freedom
## AIC: 250.19
##
## Number of Fisher Scoring iterations: 6
```

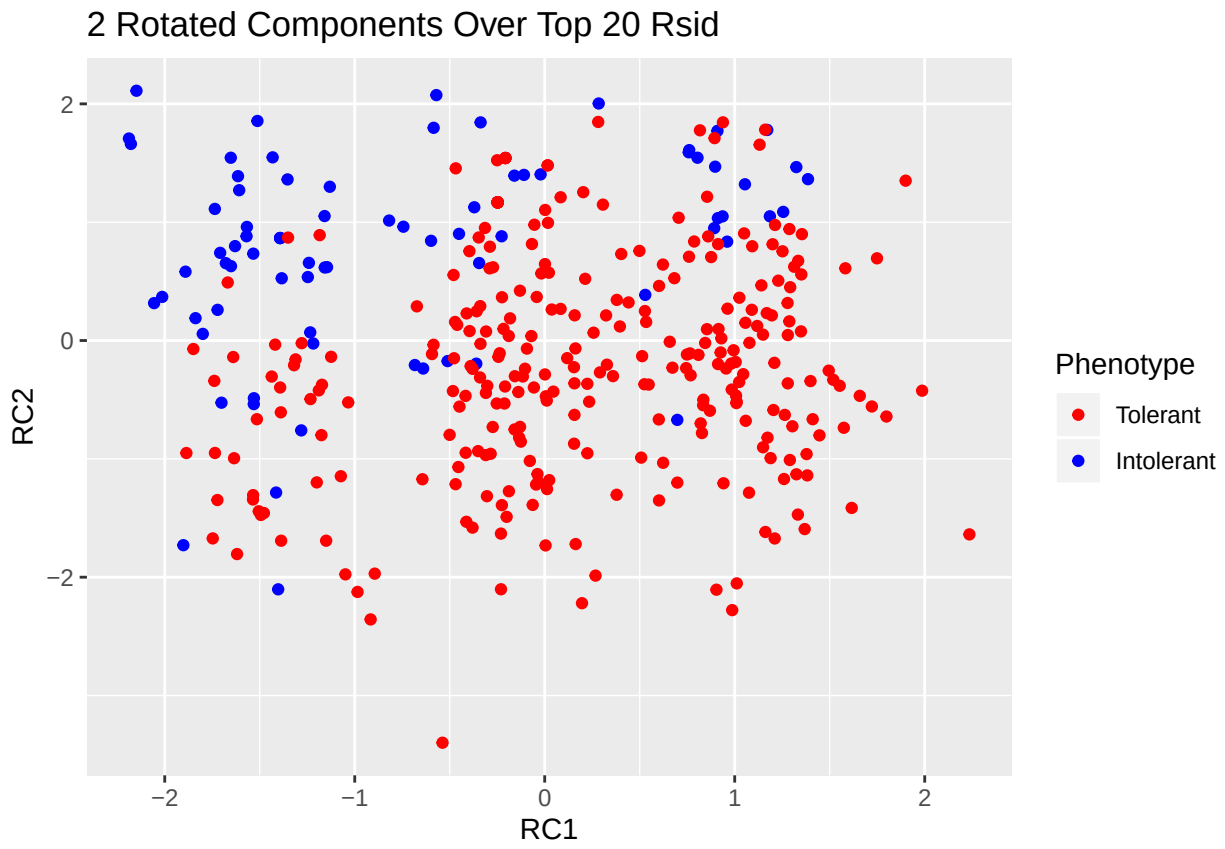
```
pR2(linModel)[4]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.3592752
```

PCA with 2 Rotated Components Plot

```
ggplot(rotation2, aes(x = RC1, y = RC2, label = as.factor(pheno))) +
  geom_point(aes(colour = as.factor(pheno)), show.legend = TRUE) +
  scale_color_manual(name="Phenotype",
                     labels=c("Tolerant","Intolerant"),
                     values=c("red","blue")) +
   ggtitle("2 Rotated Components Over Top 20 Rsid")
```



PCA with Varimax Rotation, Rotatated Components = 3

```
pca_varimax <- principal(top_20[1:20], nfactors = 3, rotate = "varimax")
rotation3 <- data.frame(cbind(pca_varimax$score, pheno=top_20[,"pheno"]))
pca_varimax$loadings
```

```
##
## Loadings:
##             RC1    RC2    RC3
## rs6545107           0.236  0.225
## rs988179   -0.106  0.543
## rs7712447  -0.332  0.389 -0.175
## rs4903435         -0.390
## rs10178148 -0.139         0.420
## rs4396824         -0.226  0.523
## rs1995990          0.352
## rs1558910   0.109  0.442
## rs3799807         -0.340
## rs4988406         -0.189 -0.393
## rs17293443  0.242 -0.209 -0.168
## rs9386654                  0.454
## rs10229311  0.139         0.423
## rs12209650  0.118 -0.266 -0.226
## rs11562695        -0.374  0.158
## rs9914374         -0.329 -0.303
## rs1088682          0.388
## rs1254601   0.122        -0.482
## rs182549    0.960        -0.119
## rs4988235   0.958        -0.126
##
##                  RC1   RC2   RC3
## SS loadings    2.127 1.695 1.554
## Proportion Var 0.106 0.085 0.078
## Cumulative Var 0.106 0.191 0.269
```

```
pca_varimax$Vaccounted
```

```
##                             RC1        RC2        RC3
## SS loadings           2.1270460 1.69516685 1.55382291
## Proportion Var        0.1063523 0.08475834 0.07769115
## Cumulative Var        0.1063523 0.19111064 0.26880179
## Proportion Explained  0.3956532 0.31531912 0.28902764
## Cumulative Proportion 0.3956532 0.71097236 1.00000000
```

PCR with 3 Rotated Components

```
linModel <- glm(pheno ~ ., data = rotation3, family = "binomial")
summary(linModel)
```

```
##
## Call:
## glm(formula = pheno ~ ., family = "binomial", data = rotation3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.06666  -0.51230  -0.23027  -0.06094   2.92887
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3231     0.2440  -9.519  < 2e-16 ***
## RC1          -1.0657     0.1892  -5.631 1.79e-08 ***
## RC2           1.5336     0.2178   7.042 1.90e-12 ***
## RC3           1.5064     0.2339   6.440 1.20e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 381.12  on 369  degrees of freedom
## Residual deviance: 229.09  on 366  degrees of freedom
## AIC: 237.09
##
## Number of Fisher Scoring iterations: 6
```

```
pR2(linModel)[4]
```

```
## fitting null model for pseudo-r2

##  McFadden
## 0.3988967
```

PCA with 3 Rotated Components Plot

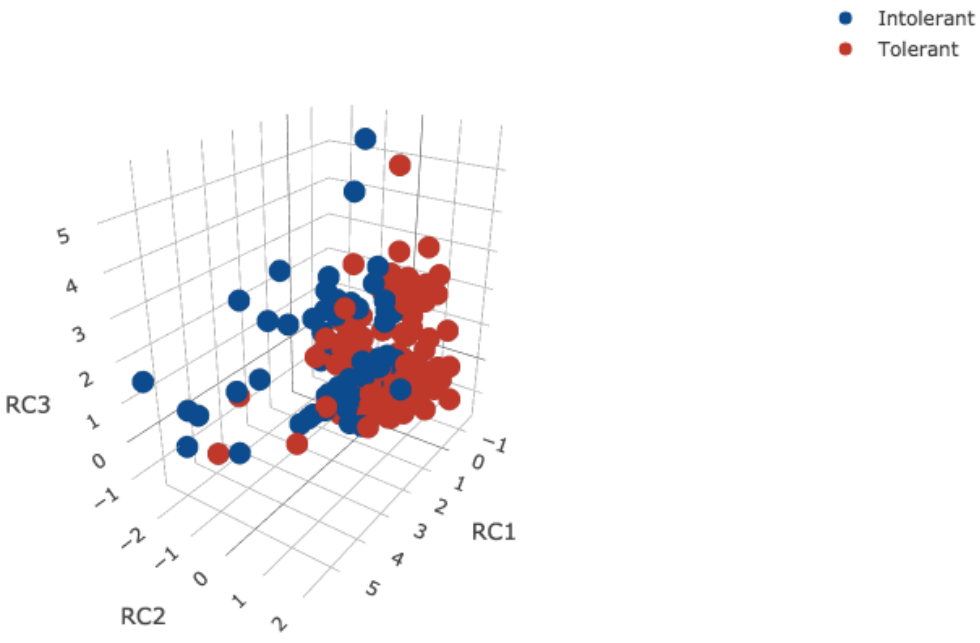


Figure 4: 3D Viz of Top 20 Coefficients (1)

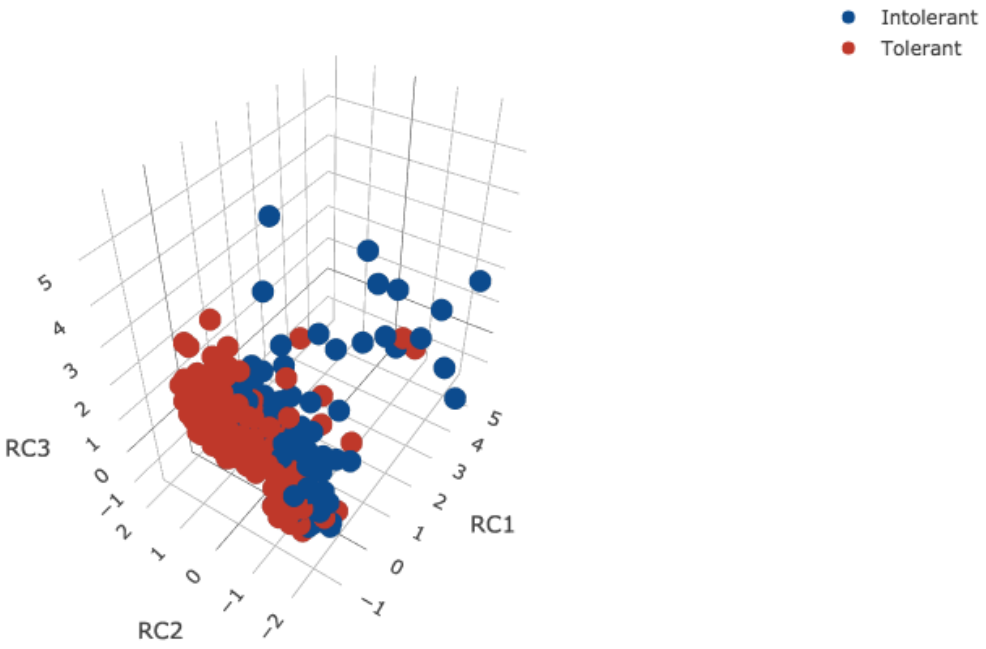


Figure 5: 3D Viz of Top 20 Coefficients (2)

EDA and Data Preprocessing Over Full Data

Steps applied were the following:
i) Drop rows that has users' phenotype as partial.
ii) Threshold to drop columns with more than 60% of NA and drop pheno with partial label.
iii) Imputation by mode.
iv) Elimininate columns that are highly correlated (threshold > 0.99).

```
original_data <- read.csv('/Users/student1/Desktop/lact_exploration/three_label_with_selected_features_
table(original_data$pheno)
```

```
##
## intolerant    partial   tolerant
##         78         58        292
```

```
print(dim(original_data))
```

```
## [1] 428 972
```

Dimensions after data prerpocessing

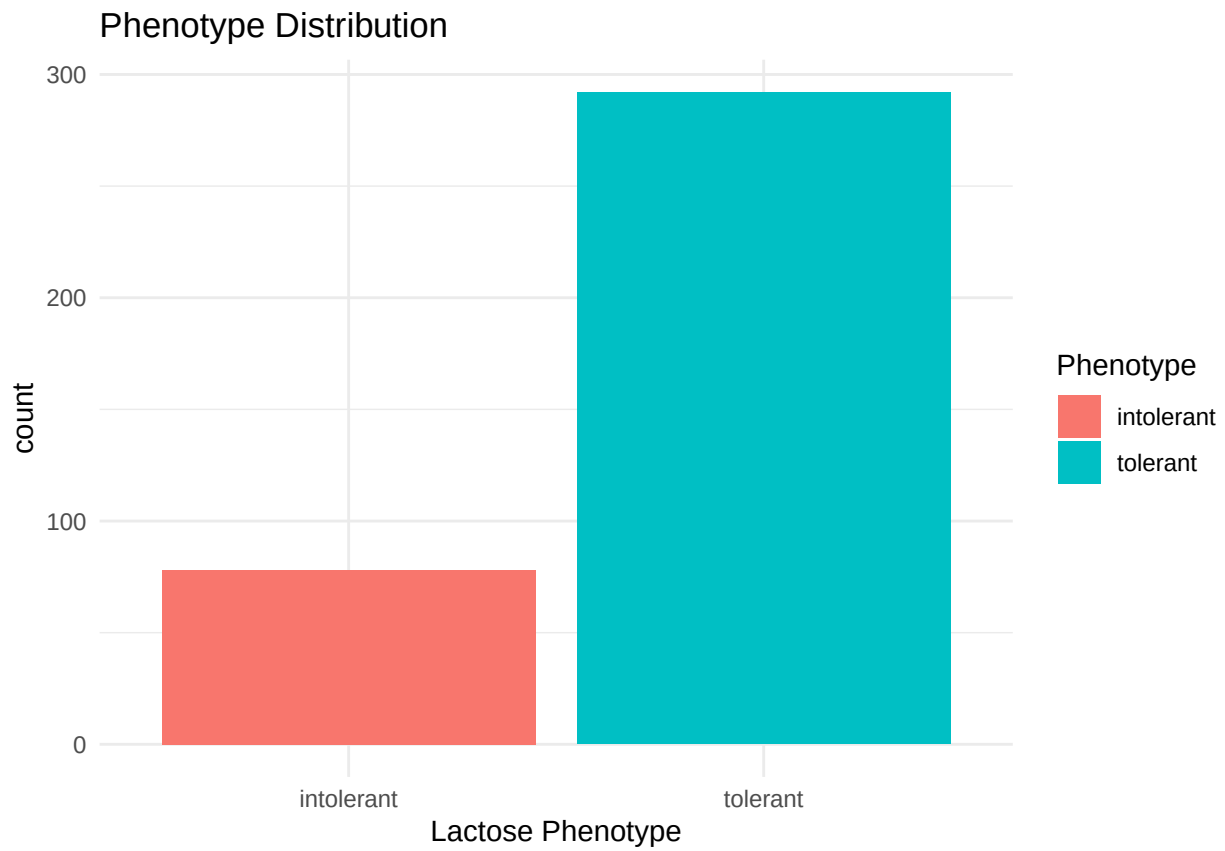
```
## Elimininate columns that are highly correlated

corr <- cor(original_data_60thres[,2:length(original_data_60thres)-1])
highCorr <- findCorrelation(corr, cutoff = .99, names = T)
clean_data <- original_data_60thres[, !names(original_data_60thres) %in% highCorr]
dim(clean_data)
```

```
## [1] 370 357
```

```
table(clean_data$pheno)
```

```
##
## intolerant    partial   tolerant
##         78          0        292
```

```
temp <- clean_data[, 2:length(clean_data)]
p<-ggplot(temp, aes(x=pheno, fill=pheno)) +
  geom_bar(stat="count")+theme_minimal() +
  ggtitle("Phenotype Distribution")
p <- p + labs(fill = "Phenotype") +  xlab("Lactose Phenotype")
p
```



PCA with Varimax Rotation, Rotatated Components = 2

Displaying only MCM6 gene that is of interest for this research

```
pca_varimax <- principal(clean_data[,2:length(clean_data)-1], nfactors = 2, rotate = "varimax")
rotation2 <- data.frame(cbind(pca_varimax$score, pheno=clean_data[,"pheno"]))
# pca_varimax$loadings
pca_varimax$loadings[c(59,345),]
```

```
##                  RC1       RC2
## rs182549  -0.3841644 0.3220302
## rs4988235 -0.3903394 0.3084034
```

```
pca_varimax$Vaccounted
```

```
##                              RC1         RC2
## SS loadings           35.19691966 10.68137693
## Proportion Var         0.09886775  0.03000387
## Cumulative Var         0.09886775  0.12887162
## Proportion Explained   0.76718018  0.23281982
## Cumulative Proportion  0.76718018  1.00000000
```

PCR with 2 Rotated Components

```
linModel <- glm(pheno ~ RC1 + RC2 , data = rotation2, family = "binomial")
summary(linModel)
```

```
##
## Call:
## glm(formula = pheno ~ RC1 + RC2, family = "binomial", data = rotation2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5818  -0.5534  -0.3276  -0.1919   2.3457
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6769     0.1799  -9.320  < 2e-16 ***
## RC1           1.5951     0.2702   5.903 3.58e-09 ***
## RC2          -1.5074     0.2110  -7.145 9.00e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 381.12  on 369  degrees of freedom
## Residual deviance: 252.89  on 367  degrees of freedom
## AIC: 258.89
##
## Number of Fisher Scoring iterations: 6
```

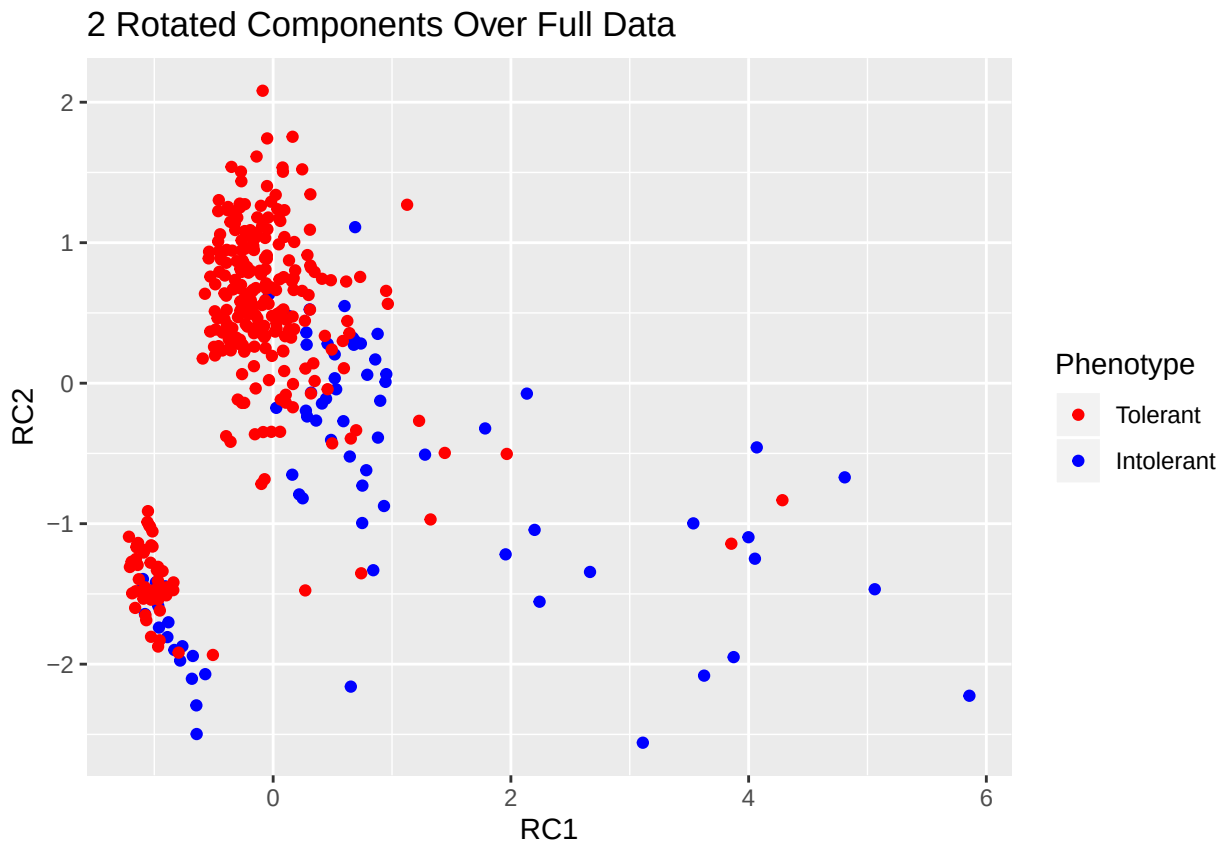
```
pR2(linModel)[4]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.3364617
```

PCA with 2 Rotated Components Plot

```
ggplot(rotation2, aes(x = RC1, y = RC2, label = as.factor(pheno))) +
  geom_point(aes(colour = as.factor(pheno)), show.legend = TRUE) +
  scale_color_manual(name="Phenotype",
                     labels=c("Tolerant","Intolerant"),
                     values=c("red","blue")) +
   ggtitle("2 Rotated Components Over Full Data")
```



PCA with Varimax Rotation, Rotatated Components = 3

Displaying only MCM6 gene that is of interest for this research

```
pca_varimax <- principal(clean_data[,2:length(clean_data)-1], nfactors = 3, rotate = "varimax")
rotation3 <- data.frame(cbind(pca_varimax$score, pheno=clean_data[,"pheno"]))
pca_varimax$loadings[c(59,345),]
```

```
##                   RC1         RC3       RC2
## rs182549  -0.4045487 -0.04937727 0.3387450
## rs4988235 -0.4077086 -0.05696198 0.3233873
```

```
pca_varimax$Vaccounted
```

```
##                              RC1         RC3         RC2
## SS loadings           31.40889507 12.40002799 10.46942356
## Proportion Var         0.08822723  0.03483154  0.02940849
## Cumulative Var         0.08822723  0.12305877  0.15246727
## Proportion Explained   0.57866345  0.22845257  0.19288398
## Cumulative Proportion  0.57866345  0.80711602  1.00000000
```

PCR with 3 Rotated Components

```
linModel <- glm(pheno ~ ., data = rotation3, family = "binomial")
summary(linModel)
```

```
##
## Call:
## glm(formula = pheno ~ ., family = "binomial", data = rotation3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9214  -0.5332  -0.3215  -0.1735   2.3306
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6901     0.1832  -9.227  < 2e-16 ***
## RC1           1.6347     0.2761   5.920 3.22e-09 ***
## RC3           0.6592     0.1682   3.920 8.87e-05 ***
## RC2          -1.5472     0.2148  -7.201 5.96e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 381.12  on 369  degrees of freedom
## Residual deviance: 248.80  on 366  degrees of freedom
## AIC: 256.8
##
## Number of Fisher Scoring iterations: 6
```

```
pR2(linModel)[4]
```

```
## fitting null model for pseudo-r2

##  McFadden
## 0.3471833
```

PCA with 3 Rotated Components Plot

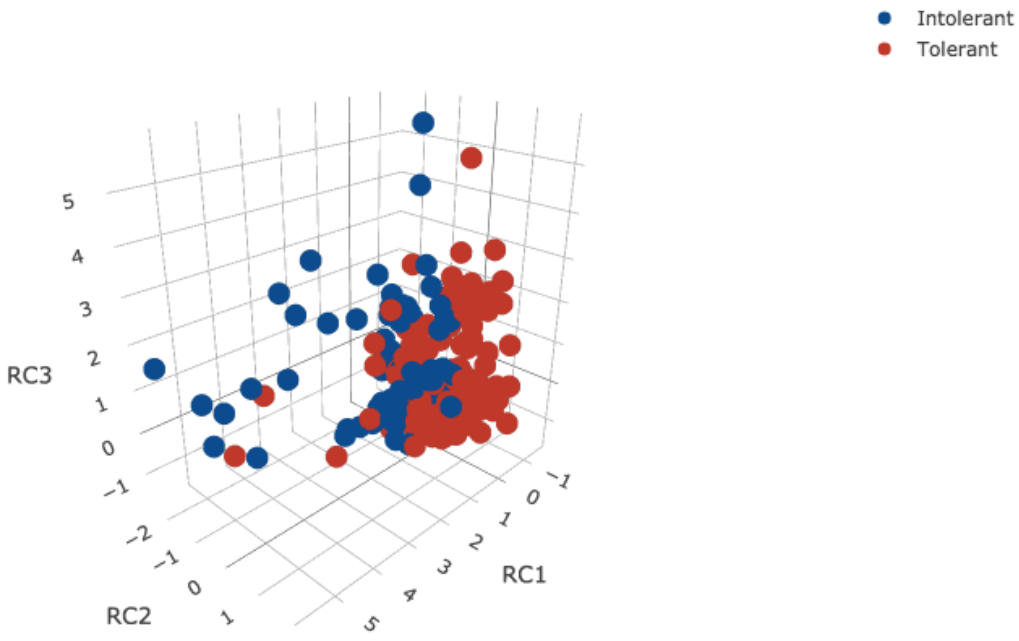


Figure 6: 3D Viz of All Rsid (1)

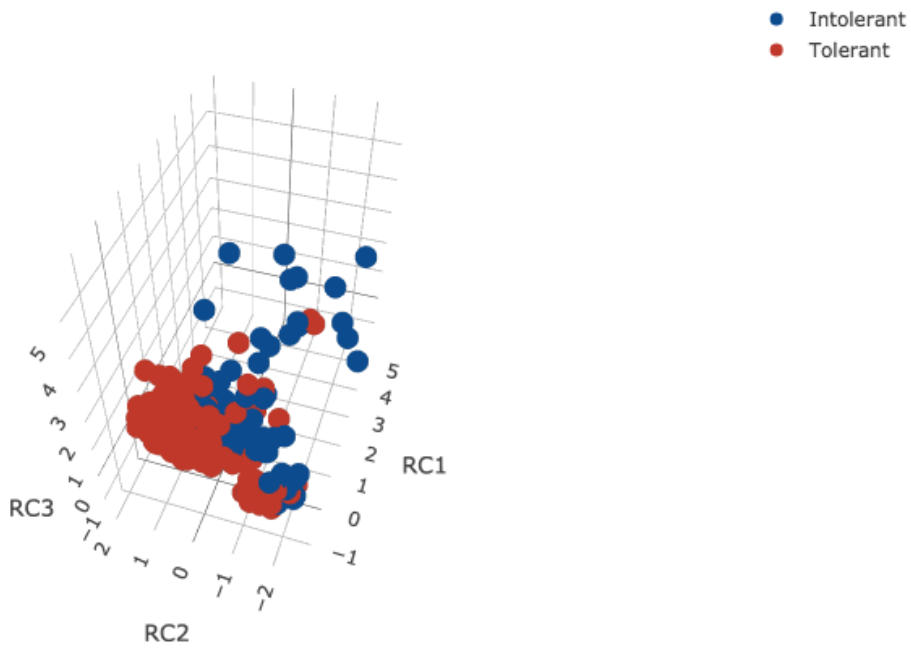


Figure 7: 3D Viz of All Rsid (2)