Lewis University

Meta-Analysis of the Genes Ubiquitously

Associated with Human Uterine Leiomyoma Development

in Healthy Humans Using

the Gene Expression Omnibus Data

by

Janis L. Corona

Romeoville, IL

July 2019

Abstract

This study examined five microarray gene expression samples of uterine leiomyomas (UL) and of non-UL in healthy females obtained from the Gene Expression Omnibus (GEO) online data repository for gene expression data. The genes in common between the five studies were combined and examined to see which genes were the most differentially expressed up or down in UL samples compared to non-UL samples in otherwise healthy females. Six genes that are ubiquitous to the association with UL risk in females were compared next to the top 10 most expressed genes in UL to test whether a machine learning model could predict with great accuracy if a sample was UL or not. The algorithms used were Latent Dirichlet Allocation (LDA), random forest (RF), general boosted regression models (GBM), k-nearest neighbors (KNN), generalized linear regression models (GLM), a second version of random forest for classification and regression (RF2), recursive partitioning and regression trees (rpart), and a combined model of best results from all of those algorithms were used. The LDA model and random forest models could accurately predict whether or not a sample was UL or non-UL with 88-91per cent accuracy using the six genes ubiquitous to the current research on UL risk and the 10 genes among the five separate studies having the highest magnitude of change in UL samples compared to non-UL samples.

*Keywords*: uterine leiomyomas, uterine fibroids, latent dirichlet allocation, six ubiquitous UL genes, bet1 golgi vesicular membrane trafficking protein like, trinucleotide repeat containing adaptor 6b, cytohesin 4, fatty acid synthase, high mobility group at-hook 2, coiled-coil domain containing 57

Table of contents

list of tables…………………………………………………………………………… iv

list of figures………………………………………………………………………….. v

list of abbreviations……………………………………………………………….. vi

chapter 1 – introduction………………………………………………………….. 1

Description of ul……………………………………………………………... 1

UL Described in Populations……………………………………… 2

significant genes for ul…………………………………………… 4

cytoband location of ul gene…………………………………………. 5

chromosome 11………………………………………………………... 6

chromosome 12………………………………………………………... 6

chromosome 17………………………………………………………... 6

chromosome 22……………………………………………………….. 7

chapter 2 – methods………………………………………………………………… 9

geo data of ul and non-ul samples………………………………….. 9

r statistical software for statistical analysis……………… 12

Deriving the data sets…………………………………………… 12

Samples simulated in the Population……………………… 16

machine learning algorithms used………………………... 22

chapter 3 – results…………………………………………………………………. 25

chapter 4 – conclusions………………………………………………………….. 66

chapter 5 – literature cited…………………………………………………… 69

appendix………………………………………………………………………………... 75

List of tables

Table 1. The top 10 plus six UL risk genes 26

Table 2: Bootstrap Simulated Results for Top 10 plus 6 Genes 28

Table 3: Member Majorities of Five Most Changed Up or Down 40

Table 4: Top 16 Genes Differentially Expressed in Subset 48

Table 5: Bottom 16 Genes Differentially Expressed in Subset 50

Table 6: Top 16 Fold Change in Subset 52

Table 7: Majority of 10 Most Differentially Expressed Gene Up and Down 54

Table 8: Top 16 Genes in Fold Change from All 56

Table 9: Top 16 Genes Differentially Expressed in All 58

Table 10: Least Expressed 16 Genes in All 60

Table 11: Machine Learning Results on Top 10 Plus 6 62

Table 12: Machine Learning Results on All Data Sets 64

List of figures

Figure 1: Histograms of UL Simulated Means for Top 10 Plus 6 Genes 30

Figure 2: Reverse Strand of Cytoband 11p15.5 Genes Expressed More in UL Near BET1L 32

Figure 3. Forward Strand of Cytoband 12q14.3 Genes Expressed Less in UL Near HMGA2 34

Figure 4. Gviz Map of Reverse Strand of Cytoband 17q25.3 Genes Expressed Less in UL 36

Figure 5. Forward Strand of Cytoband 22q13.1 Majority of Genes Expressed More in UL 38

Figure 6: Heatmap of Top 10 Plus Six Genes in All Samples 42

Figure 7: Pairwise Comparison of All Top 10 Plus 6 Genes 44

Figure 8: Comparison of Simulated Means for Non-UL and UL Top 10 Plus Six Genes 46

List of abbreviations

BMI Body Mass Index

DE Differential Expression

GBM Generalized Boosted Regression Models

GEO Gene Expression Omnibus Online Data Repository

GLM Generalized Linear Regression Models

GWAS Genome Wide Association Studies

HGNC HUGO Gene Nomenclature

LDA Latent Dirichlet Allocation

RF Random Forest method in the caret package

RF2 Random Forest function in the randomForest R package

UL Uterine Leiomyoma

Introduction

Description of UL

Many uterine leiomyoma (UL) research studies define UL as benign tumors in the uterine myometrium or similarly as benign growths in the smooth muscle tissue of the myometrium (Eggert et al., 2012; Bondagji et al., 2018). Some of the known risk factors for developing a UL are age at menarche, alcohol consumption, child birthing age, family history of UL, race, and obesity (Hellwege et al., 2017; Eggert et al., 2012; Rafnar et al., 2018). It is also known that UL treatment involving an estrogen analogue such as Leuprolide will place the body in a hypogonadal state and in some cases decrease the size of a UL but can also cause bone density loss (Dvorská1, Braný, Danková, Halašová, & Višňovský, 2017). Treatment involving an estrogen antagonist such as cetrolexin acetate have been proven to shrink the size of a UL by competing with progesterone, glucocorticoids, and androgens for estrogen receptor binding sites on the UL (Dvorska et al., 2018). Overweight females are more likely to have a UL by 20 per cent for every 10 kg over the normal body mass index (BMI), because a UL has more estrogen binding sites and androgens turn into estrogens in adipose tissue (Dvorska et al., 2017). Because estrogen has an impact on the size of a UL, it is considered estrogen dependent (Rafnar et al., 2018). There is a risk of developing a UL if the UL patient also has thyroid dysregulation, kidney cancer, stage III or higher endometrial cancer, or endometrial cancer with the genotype rs10917151 of the *CDC42*/*WNT4* gene (Rafnar et al., 2018). It is also known that *MED12* is the only gene to have a causal relationship in having a UL (Bandagji et al, 2017). The knowledge of how UL develop is still unknown and many GWAS studies have sought to find gene targets

along highly up or down regulated genes in differential gene expression studies between normal uterine tissue and UL tissue (Eggert et al., 2012; Hodge et al., 2012).

UL Described in Populations

A study on European Americans by Edwards, Hartmann, and Edwards (2013) found that *Bet1 Golgi Vesicular Membrane Trafficking Protein like* *BET1L* associated with where inside the uterus a UL formed in European American populations, such as in the uterine wall (intramural), under the endometrium (submucosal), or under the mucosal layer of the uterus (subserous). *BET1L* is also found to be significant in the Han Chinese population for the number of UL one female can have (Liu et al., 2018).

In a particular study on white races of Australian and European origin, *fatty acid synthase* (*FASN*) and *coiled-coil domain containing 57* gene (*CCDC57*) have been found to have a genome-wide significance for UL in white populations while not showing significance in Arab populations (Eggert et al., 2012; Bondagji et al., 2017).

There is insignificant evidence to include these same genes as biomarkers for UL in the African American females possibly due to misclassification of fibroid by the self-reporting of UL in control groups used in this study (Aissani, Wang, & Wiener, 2015; Hellwege et al., 2017). Because UL diagnosis is only reported if symptomatic and most cases of UL are asymptomatic as only 20-33% of patients with UL show symptoms such as pain in the pelvis and heavy bleeding (Bondagji et al., 2017; Eggert et al., 2012). The gene found to be an exclusive heterogenetic risk of UL in African American populations is *cytohesin-4* (*CYTH4*); when *CYTH4* is expressed low in thyroid tissue there is a risk for developing UL for African American females (Hellwege et al., 2017).

There is also a study by Eggert et al. (2012) on white females, sisters, and other family members from European and Australian data who have UL. In this study there was a genome wide significance level of risk for UL with *CCDC57*. The study also found that *FASN* plays a role in risk of UL in white females. When excluding studies on heterogeneity of UL, Hodge et al. (2012) found that the putative gene *HGMA2* of the *high mobility group* *AT-Hook 2* on chromosome 12 is over expressed in UL and is the most significant altered gene. This same study also suggested that due to the most variation in clustering around patient demographics than clustering of t (12;14) and non-t (12;14), that there is reason to believe that race plays a role in risk for UL development.

Another study that excluded race as a determinant in gene expression analysis of UL is the study by Zhang, Sun, Ma, Dai, & Zhang (2012). In this study on differential gene expression, the four phases of menstruation were analyzed. This was to see when the best time for implantation of a fertilized ova to produce an embryo would occur. This study was not race specific to the uterus samples gathered at different stages of the gene sample extraction. High variation of genes expressed was measured to find the most significant ones. The chromosomes of the genes most expressed were identified as chromosomes 4, 9, and 14. Many of the top genes from the GWAS samples were gathered from most expressed genes along a region of one of those chromosomes, and further analyzed to determine which genes had significantly high gene expression in UL cases (Aissani et al., 2015; Eggert et al., 2012; Bondagji et al., 2017; Edward et al., 2013).

Significant Genes for UL

The most ubiquitous genes highlighted in these GWAS population specific studies are the *BET1L* and the *trinucleotide repeat containing 6B* gene called *TNRC6B* (Edwards et al., 2013; Rafnar et al, 2018; Liu et al, 2018, Bondagji et al., 2017). These genes have been shown in separate population specific studies to associate to the number of UL one patient has (*BET1L*) and the size of the UL one person has (*TNRC6B*) in European American, Japanese, and Han Chinese populations (Edwards et al., 2013; Liu et al, 2018). Saudi Arabian populations found that *TNRC6B* only poses a risk of developing a UL (Bondagji et al., 2017). Two studies by separate researchers Rafnar et al. (2018) (UL in Europeans from the United Kingdom and Iceland) and Aissani, B. et al. (2015) (UL in European Americans) found that *BET1L* is not associated with UL. However, two other separate studies by Eggert et al. (2012) and Edwards et al. (2013) found that the *BET1L* gene is associated with UL risk for white women and European Americans.

Cytoband Location of UL Gene

Currently, significant genes found associated with UL among all of the population studies researched are *BET1L* on chromosome 11, *TNRC6B* on chromosome 22, *FASN* on chromosome 17, *CYTH4* chromosome 22, *CCDC57* on chromosome 17, *HGMA2* on chromosome 12, and *MED12* on chromosome X or 23 (Aissani et al., 2015; Eggert et al., 2012; Bondagji et al., 2017; Edward et al., 2013; Hodge et al., 2012; Hellwege et al., 2017; Liu et al., 2018; Rafnar et al., 2018). Zhang et al. (2012) found chromosomes 4, 14, and 9 to be in healthy uterine tissue capable of impregnation; these chromosomes are not from the UL risk gene chromosomes found along chromosomes 11, 12, 17, 22, and 23 in current population studies. Thus, it makes sense to further study these genes associated with UL except for the *MED12* gene on chromosome 23 that has already been proven causal to UL (Bondagji et al., 2017). The *CDC4* and *WNT4* genes are excluded because they are only found to be associated with UL in patients who have endometrial cancer, and this research focus is on UL development in healthy people (Rafnar et al., 2018). The cytoband location or locus of a chromosome was discovered to hold information on other UL risk gene targets some of the current UL risk studies testing significant association to UL risk for some of the six UL risk genes (Eggert et al., 2012; Aissani et al., 2015).

Chromosome 11

*BET1L* gene is on chromosome 11 and it is described as having significant associations with UL such as which uterine layer a UL is originating from or how many UL are in one uterus making the UL patient have multiple UL (Cha et al., 2011, Liu et al., 2018; Edwards, Hartmann, & Edwards, 2013; Rafnar et al., 2018). *BET1L* was tested for significance in association with UL in studies on other race demographics and determined insignificant in certain races (Bondagji et al., 2017; Aissani, et al., 2015; Rafnar et al., 2018). This chromosome along cytoband location 11p15.5 has two other genes *RIC8A* and *SIRT3* mentioned in two of the current UL risk studies in the same neighborhood of *BET1L* (Cha, et al., 2011; Bondagji, et al., 2017).

Chromosome 12

HGMA2 is on chromosome 12 along cytoband 12q14.3 and it is considered to have high expression levels in UL samples (Hodge at al., 2012). One other study stated HGMA2 to be a factor in tumorigenesis from studies done in 1988 that researched HGMA2 and tumor formation (Aissani et al., 2015).

Chromosome 17

Two genes on chromosome 17 along cytoband 17q25.3 named CCDC57 and FASN are significantly associated with UL in Europeans (Eggert et al., 2012; Aissani et al., 2015). Eggert’s study (2012) used the LD analysis of all chromosomes and found that one specific locus 17q25.3 of houses a handful of genes that also pose some significance, but not a GWAS significance to UL risk. Another study tested these two genes and found no significance in UL for Saudi Arabian populations (Bondagji et al., 2017).

Chromosome 22

Two genes that are found on Chromosome 22 to be significant in UL are along cytoband 22q13.1. For the first gene *TNRC6B*, it is found to be significant in Chinese, Japanese, Europeans, European Americans, and Saudi Arabians (Cha et al., 2011, Rafnar et al., 2018; Liu et al., 2018; Edwards et al., 2013; Aissani et al., 2015; Bondagji et al., 2017). *TNRC6B* was not found to be significant in African Americans (Hellwege et al., 2017). *CYTH4*, the second gene along cytoband 22q13.1 on Chromosome 22 is considered significant for UL risk in African Americans (Hellwege et al., 2017).

In this research, the top genes for heterogenetic risk in developing UL were analyzed in data made available for gene expression using GEO. There are many genome wide association studies (GWAS) on the few genes having certain genes associated with UL (Edwards, et al., 2013; Liu et al., 2018; Hellwege et al., 2017; Rafnar et al., 2018; Cha et al., 2011; Aissani, 2015). These studies have been exclusive to analyzing heterogenous differences between races of European Americans, Japanese, Chinese, African Americans, Australians, White females from Australia or the United Kingdom, and Saudi Arabian females (Edwards, 2013; Liu et al., 2018; Hellwege et al., 2017; Rafnar et al., 2018; Cha et al., 2011; Aissani, 2015). In this study, a subset of non-race specific gene expression microarray samples were combined by genes that were in common, and then filtered for those genes that were along the same chromosome cytoband locations as the six genes ubiquitous to UL risk studies. This was a measure used for analysis because some other genes around the same cytoband location as a few of these six UL risk genes were not ruled out or tested to determine if these other genes might also be gene targets for UL pathogenesis (Bondagi, et al., 2017; Cha, et al., 2011) Initially, only those few genes *TNRC6B*, *BET1L*, *CYTH4,* *FASN*, *HMGA2*, and *CCDC57* ubiquitous to the current UL risk studies and the top 10 genes with the largest magnitude of change between UL and non-UL samples were analyzed with R and Bioconductor and labeled the ‘Top 10 Plus 6’ data set (R, 2019; Bioconductor, 2019). Data science methods were used to determine a model based on seven algorithms in RStudio and Bioconductor software that was best to predict if a sample was a UL or non-UL sample. This was done by making two partitions of the 121 samples using the caret package of R into one of a training set consisting of 70 per cent or 85 samples. The other partition held the remaining 30 per cent or 36 samples as the testing set to test the accuracy in prediction of each model built with the training set using each of the seven machine learning algorithms (R, 2019). Then, seven more data sets were built to test each of the seven chosen machine learning algorithms on to decide the best genes that could be used as gene targets for UL pathogenesis. This was to determine if the gene expression data for the six UL risk genes are good gene targets for UL risk in non-race specific samples of UL and non-UL, but also test to see if the genes near these six genes might also have some missed gene targets for UL pathogenesis. The larger sets were to determine if there are even better gene targets in a mixed non-race specified sample of UL and non-UL gene expression data and to compare results from each data set and combined results of all seven algorithms on what possible gene targets to UL pathogenesis could be.

Methods

GEO Data of UL and Non-UL Samples

The gene expression microarray data collected from the GEO data repository of five independent studies involving healthy human uterine myometrial tissue and human UL tissue were included because they all had the six genes *TNRC6B, BET1L, FASN, HMGA2, CCDC57,* and *CYTH4* ubiquitous to the current UL risk studies (Miyata et al., 2017; Vanharanta, et al., 2006; Hoffman, et al., 2004; Zavadil, et al., 2010; Crabtree, et al., 2009). These data sets came with different probe IDs that were able to be merged together with additional meta columns using the GEO platform from which the GEO samples were a part of. The data from these five separate studies are microarray data that has been normalized to be on the same scale except for the study by Miyata, et al. (2017), which was inverse log2 transformed in R software to be scaled the same as the other four studies.

The process of merging the sets together was to first read in each csv file for GSE23112, GSE593, GSE13319, GSE2724, and GSE68295 with the R ‘read.csv2’ function. Some of the arguments in the read.csv2 function used were ‘comment.ignore = !’ for identifying comment tags in each file to ignore and the ‘skip =’ argument set to the number of commented lines to ignore. The other arguments to the read.csv2 function for ‘sep = ‘,’’ and ‘na.strings=c(‘’,’NA’) allowed the delimiter for csv file to be read in as comma separated and labeled missing values as empty or ‘NA’ so that these could be removed later in the script. Then in each data set, those columns that corresponded to the UL samples according to the information in the commented tags were appended with ‘ul’ to the end of those IDs to identify which samples were UL and

which weren’t. The GSE68295 file was inverse log 2 transformed to make it the same scale as the other values that were not log 2 scaled. This was done by removing the meta columns and taking only those values having numeric values and using the base math of R for ‘2^’ to that matrix version of the data frame GSE68295. Also, those samples not UL or non-UL in GSE68295 were not included by removing them by creating a separate data frame of GSE68295 that removed the sarcoma UL samples. Because this study is on healthy human UL only, the sarcoma samples were not included.

Then each of the platforms GPL96, GPL570, and GPL6480 were read in with the ‘delimit2’ function of R with arguments that indicated ‘sep=’\t’’ and ‘comment.char=’#’’ to indicate which lines of the file are commented information to ignore aside from the data frame in each of these text files. The meta columns of each of these platforms were examined and it was determined the best column to merge all data sets by was the ‘ID’ column for the GSE593, GSE2724, and GSE23112 data sets with GPL96 by ‘ID\_REF’ column using the ‘merge’ function in R. The GSE13319 was merged with GPL570, and the GSE68295 was merged with GPL6480 in the same manner.

After the series were each joined with the meta columns from their platforms, the ‘ENTREZ\_GENE\_ID’ column belonging to the merged platforms with the series of GSE593, GSE13319, GSE2724, and GSE23112 were edited to take the first listed element of that variable as there were multiple entries. This was a series of steps that involved the ‘strsplit’ function with the arguments ‘[///]’ and ‘as.character’ function of the column, and the ‘lapply’ function using the arguments ‘[‘ and ‘1’ to indicate splitting the column by the first listed. Then, these four series were each merged together to keep only the genes in common using the ‘merge’ function defaults and the ‘ENTREZ\_GENE\_ID’ column just modified to one entry per gene. After those

four series were merged into one data frame, that data set was then merged the same way with GSE68295 that was merged with its platform using the ‘GENE’ column of the GSE68295 data set and the ‘ENTREZ\_GENE\_ID’ column of the last data set of all four other series. This created a universal set of genes only in common between the five separate UL risk studies obtained from GEO. The following sub-section explains how to obtain these files.

R Statistical Software for Statistical Analysis

Deriving the Data Sets

The R software was used to combine the GEO independent studies into one large data set of genes in common among all the studies, but that also had the six genes ubiquitous to the current UL risk population studies. For these five data sets and the three platforms that added the columns needed to combine all the samples together, see the Appendix items 1 through 8. The script that merged all of these data sets to make a universal set of all genes in common is in the Appendix as item 9 ‘All\_analysis.R’ using R. An extension was added to each sample column name as ‘ul’ if the sample was UL to keep an order of samples by UL and non-UL columns in the data before merging all sets together by NCBI gene ID labeled ‘GENE’ and keeping the ‘CYTOBAND’ column for creating a subset of data by cytoband location of the six genes associated with UL risk. Columns that weren’t necessary to merge by were excluded but kept in a separate file of meta data to use later as needed. This file is in the Appendix as item 10 listed as “GSE\_array\_meta.csv.” The data set of all genes and samples that excludes most meta data information is in the Appendix as item 11 listed as ‘mrg5.csv’ and it is 1.1 Gb in size. This data set has 1,954,853 genes with many duplicate gene entries from the merge process and 123 column columns that include the 121 samples with the extended ‘ul’ name attached to the UL samples and two columns for the gene and cytoband location of that gene. Using the R package dplyr, this very large set was modified to include only unique gene values per sample by grouping by gene and taking the mean of each gene for each sample (Francois, Lionel, & Muller, 2019). This created a data set that had unique genes in common among the five series without any duplicates. In total this data set had 12,173 genes and the same 123 columns as above. This data set is in the Appendix as item 12 listed as “DE\_means\_Per\_Gene\_Chr.csv.”

This data set was then filtered in R to only include those genes along the same chromosome cytoband locations as those six genes *TNRC6B, BET1L, FASN, CYTH4, CCDC57,* and *HMGA2.* That smaller, filtered data set gave a table of 183 genes with some duplicates. This data set is in the Appendix as listed item 13 named “chr\_loci\_top\_genes.csv.”

From the last data set, modifications were made with R to use the Bioconductor package, Gviz. This was done to look at the strands of six genes ubiquitous to UL risk and the genes in the neighborhood of each gene to see if there are genes close enough to the six UL risk genes on the same strand that could be targets for UL pathogenesis (Hahn, F., 2019; Bioconductor, 2019). After this, the meta data set was modified to split the chromosome column into three columns of the chromosome for each gene as ‘chromosome,’ the start in base pairs of each gene as ‘start,’ and the end of each gene in base pairs as ‘end,’ then a column was added that gave the gene width called ‘width.’ This file is in the Appendix as item 14 titled, “ub\_genes\_gviz.csv.” This file was compared to the actual meta information per gene from the ENSEMBL website using the BioMart tab. The website ensemble.org was visited, then the BioMart tab was clicked, then the ‘New’ option was selected, followed by choosing the ‘Ensembl 96’ database, then selecting the ‘Human Genes (GRCh38.12)’ option for that database. The columns for ‘transcript’ were copied and saved as a csv text file labeled ‘ensembl\_generated\_id.csv’ and listed as item 15 in the Appendix under the same name. To this data set the transcript name was merged with the ‘ub\_genes\_gviz.csv’ data set and listed as item 16 in the Appendix as “ub\_genes\_ensembl.csv.” Then in ENSEMBL at ensemble.org, the ‘BioMart’ tab was selected again, then ‘Ensembl Genes 96,’ followed by ‘Human genes (GRCh38.p12’),’ followed by selecting ‘Structures,’ then by selecting ‘Gene Stable ID,’ and checkbox selecting each of the following items: ‘Transcript

Stable ID,’ ‘Strand,’ ‘Chromosome/Scaffold name,’ ‘Gene Start (bp),’. ‘Gene end (bp),’ and ‘Gene Name.’ When done the results were exported as csv format for ‘all’ entries and saved as “mart\_export.txt” with a file size of 14.1 MB. This file is in the Appendix as listed item number 17. The last data set was then merged with the “ub\_genes\_ensembl.csv.” data set by ENSEMBL transcript ID after making minor modifications to the imported “mart\_export.txt” data set. The modifications made were to drop unnecessary columns and modify the strand values by changing the ‘-1’ to ‘-‘ and the ‘1’ to ‘+’ to use in Gviz. A column for width (length of gene in base pairs) was also calculated as the absolute value of the ‘end’ minus the ‘start’ plus one to include the start number. This data set now had 149 genes that included duplicate genes and 129 columns consisting of 121 samples and eight meta columns with names shortened to "chromosome," "start," "end," "width," "strand,” "gene," "transcript," and "symbol." This data set is in the Appendix as item 18 listed as “ub\_genes\_ensembl\_gviz.csv.”

To this data set, some modifications were done so that duplicates were removed using the dplyr package, then transposing that data set to make the sample header into the genes and the rows as the GEO sample columns. Next, a column for the GEO sample each sample was obtained was added as a header column next to the gene columns. This data set had 121 samples as row observations labeled in each row as the GEO sample it was, and 132 header columns. The header columns included the 130 unique genes along the four cytobands of the six UL risk genes and two meta columns. The two meta columns were of the GEO series origin called ‘samples’ and a column called ‘UL\_nonUL’ that identified each row as a UL or non-UL sample . This data set is in the Appendix as item 19 listed as “All-ggplot2-type-sample-derived.csv.”

Then dplyr was used to create a column that determined the top 10 expressed genes by magnitude of most or least expressed in UL when compared to non-UL samples (Francois, et al., 2019). This data set removed the ‘samples’ and ‘UL\_nonUL’ columns of the last data set and added three new columns for each gene as the UL means, the non-UL means, and the difference in expression of the UL means minus the non-UL means. This data set is listed as item 20 in the Appendix as “DE\_data\_unordered.csv .” Then the set was divided into subsets of those having a majority or minority of gene expression along the cytoband location as the six ubiquitous genes. This was to show how the gene expression values look differently in UL compared to non-UL samples, and show if those genes associated with UL risk are in the group of genes that mostly change in more expression (‘up’) or less expression (‘down’) in UL compared to non-UL samples. This data set is in the Appendix as listed item 21 labeled “MemberGviz\_130\_141.csv.”

From this data set a magnitude column was added that took the absolute value of the difference in expression of the means. This was done so that when ordering from most to least in difference in expression values between UL and non-UL samples, those genes having more changes in inhibition of gene expression weren’t ignored. This data set is listed as item 22 in the Appendix as “MemberMagnitude\_130\_142.csv.” The top ten genes that had the most magnitude of change were made into a subset and the six genes ubiquitous to UL risk were added. This data set is now referred to as the data set of top 10 plus six genes ubiquitous to UL risk. That more manipulations were done to for making it ready for the machine learning algorithms that follow.

The last data set was made machine learning ready by making it a data set of samples only. Where rows are genes, the first column is the gene column, and the other 121 columns are the GEO sample IDs. This was then grouped into two subsets of UL and non-UL, and each transposed into a data set called “TOP16\_ml\_ready.csv.”. These data sets were then used to test bootstrap simulations on each gene in the top 10 plus six gene set to see how well they represent the population at large with 10,000 samplings with replacement on each of these 16 genes for every sample in each subset of UL or non-UL.

Samples Simulated in the Population

Bootstrap simulations were made with the ‘UsingR’ R package that built 10,000 simulations with replacement for each of the top 10 plus 6 genes (Maindonald, 2008). Then histograms of those 16 genes were made using ggplot2 to see how symmetrical each gene in the population would fit the Gaussian bell curve (Wickham, 2019). As this study had 121 samples to base the entire population of humanity upon, it was necessary to use the Law of Large Numbers to discriminate whether these genes could represent the population well and ultimately add credibility to the legitimacy in the subsequent methods and results. The Law of Large Numbers in statistics and probability theory state that a sample of a larger population will converge to the true population mean when random sampling with replacement is done a large amount of times or trials but also while averaging over those trials. One simulated population mean for UL and one for non-UL converged from 10,000 samplings for each of the top 10 plus six genes of the combined 121 GEO samples.

The file for the top 10 genes and six ubiquitous genes used for bootstrap simulations of these 121 samples to fit the population at large is “ubiq\_and\_top10\_samples\_only.csv” and it is in the Appendix as item 24. The file it used was the “MemberMagnitude\_130\_142.csv” listed in the Appendix as item 22. The file that has the results of the bootstrap simulations on these top 10 plus six UL risk genes belonging to the same cytoband location of those 6 genes is in the Appendix as item 25 as “Stats16.csv.”

The R packages, ggplot2, heatmaply, and lattice were used along with R base package to plot the simulated means between the UL and non-UL samples of those top 10 plus six UL risk genes for exploratory data analysis of the results visually (Wickham, 2019; Galili, O'Callaghan, Sidi, & Benjamini, 2019; Sarker, 2018). The R package, ggplot2, was used to visually show how the simulated means of the non-UL samples of those top 10 plus six UL risk genes measure up to the simulated means of the UL samples of those same genes.

The “MemberMagnitude\_130\_142.csv.” listed as item 22 in the Appendix was then used to generate more data sets based on this subset of genes on the same cytoband location as the six UL risk genes. One data set is a subset of the overall top 16 genes out of the 130 genes that have the highest magnitude of change. This data set, “most\_DE\_ml\_ready\_130.csv,” did not add in the six genes ubiquitous to UL risk studies and can be found in the Appendix as item 26. To create item 26 of the Appendix, the data set it was derived from as item 22 in the Appendix removed the columns other than the sample IDs after filtering only for those 16 genes having the most change in magnitude in UL compared to non-UL samples.

Then the data was transposed so that sample IDs became 121 observational rows, and the 130 genes became 130 variables as columns. Another column column was added as the first column called, type, that would attach the type of each sample ID as either UL or non-UL. This was easy since the first 51 were already non-UL and the last 70 were UL with an extension to the ID that also showed it as a UL sample. This was so that this data set could be used in the following machine learning algortihms to see how accurate the results use these genes as gene targets in predicting a sample as UL or non-UL.

Another data set was made from the same data set, “MemberMagnitude\_130\_142.csv,” that is item 22 in the Appendix. From this data set took, the 16 least expressed genes in magnitude of differential expression were extracted to see how well the algorithms that predict UL or non-UL do on the genes having the least expression in the same cytobands as the six UL risk genes. The same manipulations were done to the data set, “MemberMagnitude\_130\_142.csv,” after extracting only those 16 genes that had the lowest magnitude of change in UL compared to non-UL samples by mean for each gene. Predicting those genes that have minimal change in UL compared to non-UL would be based on the added ‘type’ column that would have an outcome of either UL or non-UL. This data set is item 27 in the Appendix and listed as “least\_DE16\_ml\_ready\_130.csv.”

Using this same data set that created the first three data sets, “MemberMagnitude\_130\_142.csv,” dplyr was used to add a fold change column to this data that used the ratio of the UL mean for each gene over the non-UL mean for each gene. A fold change equal to two means the gene doubled in UL samples compared to non-UL samples. This additional data set took the ten genes with the highest magnitude of fold change in UL compared to non-UL samples and added in the six UL risk genes.

Then the same manipulations were done that removed all columns other than the sample ID columns and then transposed the data so that genes were now columns of variables and rows were observations of sample IDs. When done with above steps, a type column was added to label each of the samples as either UL or non-UL so that the type column would be the column with which to predict accuracy in determining a sample as UL or not using the genes as variables. That data set is item 28 in the Appendix listed as “'FOLD16\_ml\_ready.csv.”

The last data set made using the data of genes only on the cytoband locations of the six UL risk genes, “MemberMagnitude\_130\_142.csv,” extracted the top five genes expressed most and the top five genes expressed least in the majority group of genes expressed along the six UL risk genes’ cytoband addresses. The same manipulations were made to get this data set into a machine learning ready format. Those manipulations involved removing the columns other than the sample IDs after gathering the 10 columns needed, then transposing the data so that the sample IDs became observational rows and the columns became 130 genes as variables. Then a ‘type’ column was added so that each of the 121 sample IDs would be labeled as either UL or non-UL. This would be the outcome variable to base accuracy in prediction of the machine learning algorithms using the gene variables to predict the sample as either UL or non-UL. If the accuracy of any of all the algorithms was good, then this could mean there are some genes that reside in the same cytoband location as the six UL risk genes that might hold further evidence to UL pathogenesis. This data set is item 29 in the Appendix and listed as “majority\_ml\_ready\_10\_total.csv.”

Additional data sets were made from all genes in common using the “universe\_12173.csv” data set in the Appendix as item 30, made from the data set as item 12 in the Appendix. The means of UL and non-UL were added to each row, then the difference between the two groups, then the magnitude as the absolute value of the difference, then the fold change as the absolute value of the ratio of the UL mean to the non-UL mean. One data set listed as item 31 in the Appendix as “most\_universe\_fold.csv” was made from that data set by adding a fold change column of the ratio to UL means over non-UL means per gene. Then the top 16 genes having the highest fold change in magnitude were selected. Columns other than the sample ID columns were removed after collecting the top 16 genes with the most fold change in absolute value in UL compared to non-UL samples. Then the data was transposed so that genes became columns and sample IDs became rows listed as first 51 non-UL and next 70 samples the UL samples. Then a type column was added to attach what type of sample each observational sample is as either a UL or non-UL sample. This made each data set ready to be used in the machine learning algorithms to predict the outcome as the type based on the regressions on the genes as variables for each row sample. If the accuracy from the models scored well, this could be an indicator that some genes out of all the genes in common having the most change in UL compared to non-UL are gene targets for evaluating if those genes are related to UL pathogenesis.

Another data set made from the same data set of item 12 in the Appendix is the “most\_universe\_DE.csv” data set that was made by adding a magnitude of differential expression column. Then taking the 16 most expressed genes by magnitude of change in UL compared to non-UL samples. Columns other than the sample ID columns were removed after collecting the top 16 genes of magnitude of change in UL compared to non-UL. Then the data was transposed so that genes became columns and samples became rows listed as first 51 non-UL and next 70 samples the UL samples. Then a type column was added to attach what type of sample each observational sample is as either a UL or non-UL sample.This made each data set ready to be used in the machine learning algorithms to predict the outcome as the type based on the regressions on the genes as variables for each row sample. This data set is listed as item 32 in the Appendix. Gene targets for UL pathogenesis could be found if these genes in this data set of all genes produced results from the machine learning algorithms that indicated great accuracy in predicting UL or non-UL as the type of sample.

Finally, another data set made from the item 12 data set in the Appendix, “DE\_means\_Per\_Gene\_Chr.csv,” was a data set that used the same magnitude of differential expression between UL and non-UL samples. This data set took the bottom 16 or 16 least expressed or inhibited genes in UL compared to non-UL samples by magnitude of change between the UL and non-UL means for each gene. This data set is item 33 in the Appendix and listed as “least\_universe\_DE.csv.” The same columns other than the sample ID columns were removed once the 16 genes having the lowest gene expression changes in UL compared to non-UL were selected. The data was then transposed so that the sample IDs became observational rows, and the genes became header or variable columns. Then a column was added as the first column that labeled each of the row samples as UL or non-UL. This was done so that this data set could be machine learning ready to run into the predictive analytics R functions to see how well these 16 genes make in determining gene targets for UL pathogenesis based on how accurate the models predict each sample as being a UL or not. The type column is the outcome column each model was regressed or clustered against to produce an outcome of either UL or non-UL based on the type column.

Machine Learning Algorithms Used

The seven predictive algorithms of LDA, RF, rpart, GLM, KNN, GBM, and RF2 were used on this dataset of top 10 plus six genes using caret, gbm, lda, randomForest, e1071, and MASS r packages (Kuhn, Wing, Weston, Williams, Keefer, Engelhardt, & Hunt, 2019; Greenwell, Boehmke, & Cunningham, 2019; Chang, 2015; Breiman, Cutler, Liaw, & Wiener, 2018; Meyer, Dimitriadou, Hornik, Weingessel, Leisch, Chang, & Lin, 2015; Ripley, Venables, Bates, Hornik, Gebhardt, & Firth, 2019). All these algorithms were trained on a 70 per cent partition of the top 10 plus 6 genes data set equal to 85 samples of the 121 total samples. Then they were tested on the remaining 30 per cent or 36 samples for accuracy in predicting whether a sample was UL or non-UL based on regressing the type column on all the genes. The MASS package is used with caret for the support functions and generalized linear models, poisson, binomial, and ‘modern applied statistics with S’ (Ripley, et al., 2019). The randomForest package uses its own built in algorithm for random forest classification using the e1071 package that stands for ‘Miscellaneous Functions of the Probability and Statistics Group’ (Breiman, et al.,2018; Meyer, et al., 2015). The RF2 is the second version of the random forest algorithm that used the randomForest package of R instead of the random forest method of the caret package. The tuning parameters for this RF2 algorithm by default sample with replacement on 500 trees in classifying data based on the training set (Breiman, et al., 2018). The RF2 algorithm settings for the purposes of training on the data sets of this research set the method to ‘class’ in the default settings of the randomForest function in this RF2 algorithm. The caret package is the classification and regression training in R that supplies the LDA, RF, rpart, GLM, KNN, and GBM algorithms as methods in its ‘train’ function (Kuhn, et al, 2019).

The LDA algorithm is a method used in the caret package. LDA uses the collapsed Gibbs sampling model for topic modeling renamed latent Dirichlet allocation and typically used to categorize text by topic and not normally used for numeric data as the gene expression values are continuous numeric data types. LDA works by using approximated sequencing of observations gathered from a multivariate or joint probability distributions or at least two variables using the Markov Chain Monte Carlo algorithm (Chang, 2015). The RF algorithm is the random forest method in the caret package of R (Kuhn, et al., 2019). This method tunes the number of trees to decide in categorizing data so that accurate results can be predicted from this classification model built on a training set of data. For the the methods used here, the RF method was trained using cross validation with a value equal to five. This means that the training set was divided into five subsets where one set is left out so that the other four sets predict the result on the left out subset. This was repeated for each set so that each subset left out is used in four other subsets to predict the result on a left out subset. The five results are averaged out to get an estimate for the best result for each gene sample value for predicting the sample to be UL or non-UL. The KNN algorithm of the caret package uses a set ‘K’ number of clusters to group the nearest neighbors or genes that fit the threshold of values this algorithm puts for each cluster (Kuhn, et al., 2019). It takes the centroid of each cluster then groups the neighboring clusters into the groups whose centroids the neighbors are closest to. This is repeated while recalculating the centroid of each cluster as more neighbors are added. The setting for the KNN method in caret that were used for each data set with a pre-process of ‘center’ and ‘scale’ with a tune length of 10 and a training method set to ‘cv’ for cross validation.

The rpart method of the caret package is used in combination with the rpart package and were used with R settings having a tune length of 9 and default settings for rpart to predict using recursive partitioning and regression trees (Therneau, et al., 2019; Kuhn, et al., 2019). The GLM method is from the caret package and was used to run predictive analytics using the default settings in R and caret for the ‘glm’ method (Kuhn, et al., 2019). The glm is a generalized linear regression model (Kuhn, et al., 2019). The gene expression data is continuous numeric data, so this seemed logical to use. The GBM algorithm is also in the caret package and used for predictive analytics on the continuous gene expression data. The only adjustment made to the default settings was to set the verbose parameter to false. This package is a generalized boosted regression model that bootstrap aggregates the samples similar to the AdaBoost and gradient boosting algorithms do as it is based on those algorithms.

These same algorithms were used to test variations of the data of genes that are universally in common between all five GEO series of samples and variations in those genes in the subset of genes universally in common and only on the same chromosomes as the six genes ubiquitous to current UL risk studies. Those data sets can be found in the Appendix as items 23 and items 25-32. The reasoning behind the variations in data sets of predictors in UL samples for the algorithms, was to discover any better predictors out of using those with the most fold change in all, those with the most change in magnitude in all, compare to those with the least fold change in all and the least magnitude of change in all, and to also compare those genes along the cytobands of interest shifting in change with UL or against the majority of genes changing in UL when compared to non-UL samples.

Results

The results from merging all data series on UL risk microarray studies that had the six UL risk genes in them are in this section. There are many results for the many methods previously described. The first result is those top 10 genes having the most magnitude of change in difference between UL means per gene and non-UL means per gene with the added six UL risk genes. The following table, ‘Table 1: The top 10 plus six UL risk genes,’ shows the gene symbol of each of those top 10 plus 6 UL risk genes, the Hugo Nomenclature descriptive name, the strand that each gene is located, and the cytoband that each gene is located. The strand is forward if the value is ‘+’ and reverse if the value is ‘-‘ for location in the cytoband region for each gene. Looking at the table both *CCDC57* and *FASN* are both on the reverse strand of cytoband 17.q25.3 of chromosome 17, and *CYTH4* and *TNRC6B* are both on the forward strand of cytoband 22q13.1. The gene *BET1L* is on the reverse strand of cytoband 11p15.5 of chromosome 11, and *HMGA2* is on the forward strand of cytoband 12q14.43. The other genes are the top 10 highest magnitude of change in UL compared to non-UL in those same cytoband regions. Of those top 10 genes, *PYCR1, SOCS3,* and *ZNF750* are on the same reverse strand in the same cytoband as *FASN* and *CCDC57.* Also, the gene, *TH*, is on the same reverse strand as *BET1L.* The gene *KDELR3* is on the same forward strand of the same cytoband as *CYTH4 and TNRC6B*, and no other gene in these top 10 share the same strand and cytoband as *HMGA2.* These other genes could be gene targets for UL pathogenesis.

**Table 1**. The top 10 plus six UL risk genes

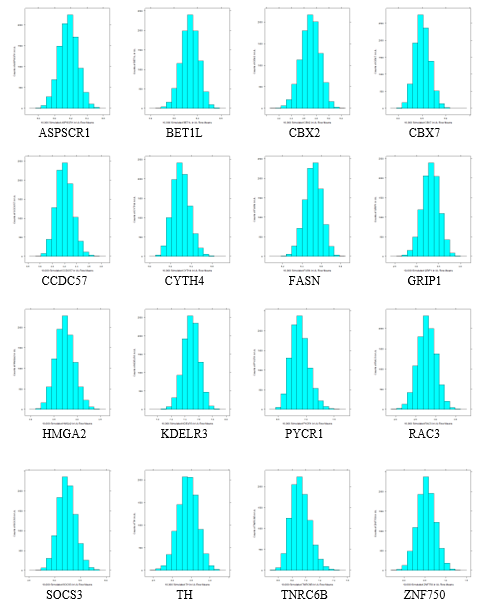
|  |  |  |  |
| --- | --- | --- | --- |
| **Genes** | **HGNC Gene Name** | Strand | **Cytoband** |
| *ASPSCR1* | alveolar soft part sarcoma chromosome region, candidate 1 | + | hs|17q25.3 |
| *BET1L* | blocked early in transport 1 homolog (S. cerevisiae)-like | - | hs|11p15.5 |
| *CBX2* | chromobox homolog 2 | + | hs|17q25.3 |
| *CBX7* | chromobox homolog 7 | - | hs|22q13.1 |
| *CCDC57* | coiled-coil domain containing 57 | - | hs|17q25.3 |
| *CYTH4* | cytohesin 4 | + | hs|22q13.1 |
| *FASN* | fatty acid synthase | - | hs|17q25.3 |
| *GRIP1* | glutamate receptor interacting protein 1 | - | hs|12q14.3 |
| *HMGA2* | high mobility group AT-hook 2 | + | hs|12q14.3 |
| *KDELR3* | KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 | + | hs|22q13.1 |
| *PYCR1* | pyrroline-5-carboxylate reductase 1 | - | hs|17q25.3 |
| *RAC3* | ras-related C3 botulinum toxin substrate 3 (rho family, small GTP binding protein Rac3) | + | hs|17q25.3 |
| *SOCS3* | suppressor of cytokine signaling 3 | - | hs|17q25.3 |
| *TH* | tyrosine hydroxylase | - | hs|11p15.5 |
| *TNRC6B* | trinucleotide repeat containing 6B | + | hs|22q13.1 |
| *ZNF750* | zinc finger protein 750 | - | hs|17q25.3 |

The next result is the bootstrap simulation results for the mean, standard deviation, and magnitude of the change using the difference in means of the UL and non-UL samples. These top 10 plus six UL risk genes are from the subset of 130 genes found on the same cytobands of those six UL risk genes. The ‘Genes’ column is the gene symbol, the ‘Non-UL Mean’ is the mean of each gene simulated from 10,000 samplings as is the ‘UL-Mean’ column but from the UL samples. The ‘Non-UL Std Dev’ and ‘UL Std Dev’ columns are for the standard deviations of those simulated means for non-UL and UL samples respectively. The ‘Simulated Magnitude Changed’ column is the magnitude of change for each gene simulated as the absolute value of the difference of UL means per gene minus the non-UL means per gene. Two of the genes known to be UL risk genes, *HMGA2* and *TNRC6B* have very low magnitude of change values for simulated population means in UL compared to non-UL samples of 0.09 and 0.07 respectively. The gene with the highest change is on the same strand as two of the UL risk genes, , *FASN* and *CCDC57,* on 17q25.3 with a value of 0.82 in magnitude of change in UL compared to non-UL samples. The next highest magnitude of change is 0.80 also belonging to a gene in cytoband 17q25.3, but on the forward strand for gene *CBX2.* These results can be viewed in Table 2.

**Table 2**: Bootstrap Simulated Results for Top 10 plus 6 Genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Genes** | **Non-UL**  **Mean** | **Non-UL**  **Std Dev** | **UL\_**  **Mean** | **UL\_**  **Std Dev** | **Simulated**  **Magnitude**  **Changed** |
| *ASPSCR1* | 5.55 | 0.12 | 6.17 | 0.11 | 0.62 |
| *BET1L* | 5.4 | 0.22 | 5.83 | 0.19 | 0.44 |
| *CBX2* | 3.91 | 0.16 | 4.7 | 0.15 | 0.8 |
| *CBX7* | 9.16 | 0.19 | 8.5 | 0.16 | 0.65 |
| *CCDC57* | 4.05 | 0.16 | 4.2 | 0.14 | 0.15 |
| *CYTH4* | 5.47 | 0.1 | 5.3 | 0.08 | 0.17 |
| *FASN* | 5.29 | 0.09 | 5.52 | 0.09 | 0.23 |
| *GRIP1* | 2.64 | 0.22 | 3.32 | 0.2 | 0.68 |
| *HMGA2* | 3.64 | 0.13 | 3.74 | 0.2 | 0.09 |
| *KDELR3* | 6.76 | 0.1 | 7.49 | 0.12 | 0.72 |
| *PYCR1* | 6.21 | 0.14 | 6.89 | 0.16 | 0.68 |
| *RAC3* | 2.04 | 0.2 | 2.78 | 0.23 | 0.74 |
| *SOCS3* | 6.01 | 0.26 | 5.24 | 0.17 | 0.77 |
| *TH* | 2.76 | 0.25 | 3.41 | 0.26 | 0.65 |
| *TNRC6B* | 6.85 | 0.14 | 6.92 | 0.13 | 0.07 |
| *ZNF750* | 1.35 | 0.24 | 0.53 | 0.22 | 0.82 |

The results from the histograms of each of these simulated means in the UL samples for the top 10 plus six UL risk genes in the 130 sub-set of genes in the same cytobands as those six UL risk genes show mostly good approximations to the population from this sample of 70 UL patients. The gene that had the most change in UL compared to non-UL, *ZNF750*, is almost perfectly symmetrical. There is good enough reason to continue with using these 121 samples as good approximations to the population based on the symmetry in the samples shown in Figure 1.



**Figure 1**: Histograms of UL Simulated Means for Top 10 Plus 6 Genes

The next result is of those genes in the same group of genes as *BET1L* expressed along cytoband 11p15.5 having more up regulation in UL compared to non-UL. There is only one other gene, *SIRT,* down-stream of BET1L and it is not one of the top 10 genes with the most magnitude of change. The arrow points left in the top half of the image to indicate this is the reverse strand that *BET1L* is located. This image is in Figure 2 that follows.

A screenshot of a map

Description automatically generated

**Figure 2**: Reverse Strand of Cytoband 11p15.5 Genes Expressed More in UL Near *BET1L*

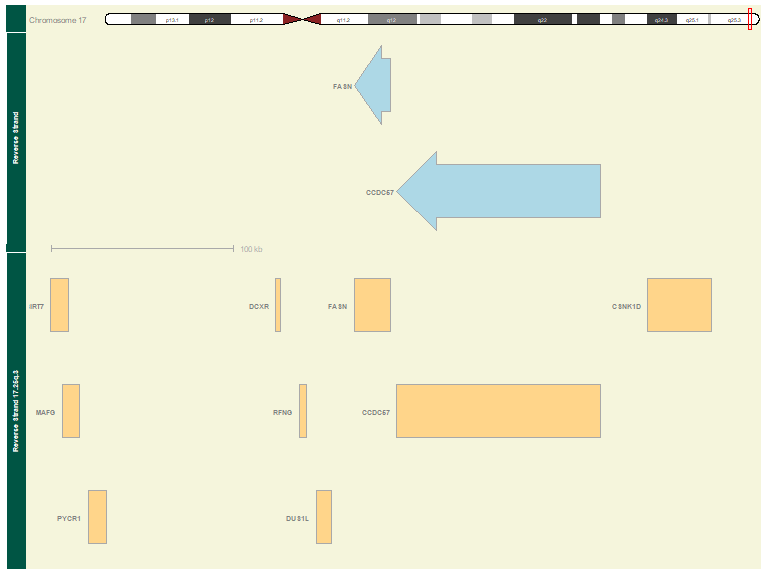
The next result is of the genes in the same group of under or down expression in UL compared to non-UL samples along the forward strand of 12q14.3 of *HMGA2.* The other two genes in this same group of down expression are *LEMD3* down stream of HMGA2 and *IRAK3* up stream of *HMGA2.* The arrow in the top half of the image is pointing right to indicate the forward strand and highlights the UL risk gene, HMGA2. This image is shown in Figure 3.

A close up of a map

Description automatically generated

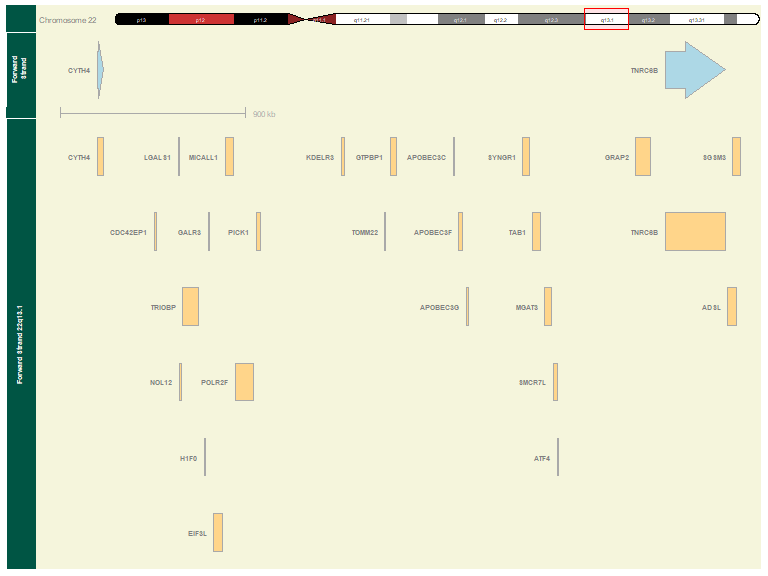
**Figure 3**. Forward Strand of Cytoband 12q14.3 Genes Expressed Less in UL Near *HMGA2*

The following result is for those genes along the same cytoband as *CCDC57* and *FASN* that are in the same group of genes on the reverse strand that are expressed less in UL compared to non-UL samples. The only ‘top 10 plus 6 UL risk genes’ that is on this same strand of cytoband 17q25.3 is *PYCR1* in the lower left corner of the image up stream of *CCDC57* and *FASN*. The arrows highlight the two UL risk genes *CCDC57* and *FASN* in the top half of the image to show this is the reverse strand as indicated by the arrow pointing left. This image is shown in Figure 4.



**Figure** **4**. Gviz Map of Reverse Strand of Cytoband 17q25.3 Genes Expressed Less in UL.

The next result is the result of the group of genes expressed more in UL compared to non-UL samples along cytoband 22q13.1 of the two UL risk genes, *TNRC6B* and *CYTH4.* There are many genes between these two UL risk genes along the forward strand as indicated by the top half of the image showing arrows pointing right. One of the ‘top 10 plus 6 genes,’ *KDELR3* is almost half the distance between these two UL risk genes. This image is shown in Figure 5.



**Figure 5**. Forward Strand of Cytoband 22q13.1 Majority of Genes Expressed More in UL

The next result is a table of the 130 genes along the same cytobands as the six UL risk genes but also in the majority group of genes showing up or down expression the most in UL compared to non-UL samples for cytobands 11p15.5, 12q14.3, and 22q13.1. The cytoband region of 17q25.3 is not in any group because there were an equal amount of genes expressed more and less in UL compared to non-UL samples. This grouping allowed for the top five genes in the majority group being expressed the most in UL and the top five genes being inhibited or down regulated the most in UL compared to non-UL.

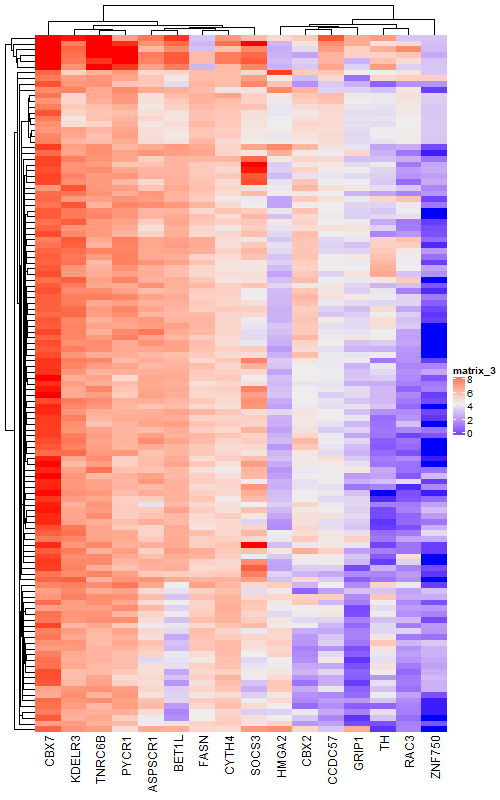
The ‘genes’ column is the gene symbol of each gene. The ‘type’ column indicates if the gene is up or down regulated in UL compared to non-UL. The all field shows how many genes in all are in that cytoband location in the subset of 130 genes. The ‘up’ column indicates how many of the ‘all’ column are up regulated. The ‘down’ column indicates how many in the ‘all’ column are down regulated. The majority column indicates if that gene is in the majority of genes up or down regulated on that cytoband. The strand column indicates if that gene is on the forward or reverse strand indicated with ‘+’ or ‘-‘ respectively. The cytoband column indicates what cytoband the gene is on. The ‘diff\_expr’ column indicates what the difference from UL minus non-UL means per gene is.

None of the six UL risk genes are in this set of genes in the majority of top 5 up and top 5 down regulated in UL compared to non-UL genes. But two of the top 10 most expressed genes by magnitude of change over the entire subset of 130 genes are, and those genes are *GRIP1* and *KDLR3*. The gene *KDLR3* is also in the same forward strand of cytoband 22q13.1 of *CYTH4* and *TNRC6B*. This table is shown in Table 3.

**Table 3**: Member Majorities of Five Most Changed Up or Down

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **genes** | **type** | **all** | **up** | **down** | **majority** | **strand** | **cytoband** | **diff\_expr** |
| EPS8L2 | down | 33 | 16 | 17 | TRUE | + | hs|11p15.5 | -0.5 |
| TNNI2 | down | 33 | 16 | 17 | TRUE | + | hs|11p15.5 | -0.48 |
| SCT | down | 33 | 16 | 17 | TRUE | - | hs|11p15.5 | -0.36 |
| INS | down | 33 | 16 | 17 | TRUE | - | hs|11p15.5 | -0.32 |
| RPLP2 | down | 33 | 16 | 17 | TRUE | + | hs|11p15.5 | -0.32 |
| KDELR3 | up | 43 | 25 | 18 | TRUE | + | hs|22q13.1 | 0.72 |
| GRIP1 | up | 6 | 5 | 1 | TRUE | - | hs|12q14.3 | 0.68 |
| MICALL1 | up | 43 | 25 | 18 | TRUE | + | hs|22q13.1 | 0.58 |
| ADSL | up | 43 | 25 | 18 | TRUE | + | hs|22q13.1 | 0.43 |
| MGAT3 | up | 43 | 25 | 18 | TRUE | + | hs|22q13.1 | 0.42 |

The following result is the result of the Heatmaply package in R to produce a heatmap of all 121 samples for the top 10 plus 6 UL risk genes data set. Those top 10 genes are the genes having the most change in magnitude in UL compared to non-UL samples in the subset of 130 genes found in the cytobands of the 6 UL risk genes. The other six genes are those six UL risk genes. The scale is the default scale of hot reds being the highest expression values and cool violets being the lowest expression values. Most of the genes to the left stay in the hot zone of gene expression changing slightly within the red zone, while the others to the left stay mostly in the cool violet zones for lowest gene expression values. The genes that could be gene target for UL pathogenesis based on this heatmap of genes that change values in the cool and hot zones, are *BET1L, SOCS3, HMGA2, CBX2,CCDC57, GRIP1, TH, RAC3,* and *ZNF750.* These start in the middle of the heatmap and follow towards the right through to the end of the right side of the heatmap. *ZNF750* and *CBX2* are the two genes that had the most simulated magnitude of change in UL compared to non-UL samples, so it makes sense that it is in this heatmap showing large changes in expression values between cools and hots on this scale. Three of the genes are already UL risk genes, *BET1L, HMGA2,* and *and CCDC57*. The other genes, *SOCS3, GRIP1, TH,* and *RAC3* could possibly be gene targets as well as *CBX2* and *ZNF750* for UL pathogenesis. This image is in Figure 6.



**Figure 6:** Heatmap of Top 10 Plus Six Genes in All Samples

The next results are of the lattice R package showing a pairwise comparison of the true sample values of all top 10 plus six UL risk genes. The image shows the splots of samples arranged according to each gene’s expression value in all 121 samples of the UL and non-UL samples. If any of these genes moved the same they would be on the 45 degree line, but none of them do. There were no visual UL risk gene relationships between the genes to display. This is not the same as a quantile-quantile plot of the expected to observed values. If it were, then any scatter outside the 45 degree line would indicate gene targets. This result showed that lattice pairwise comparison of genes to each other in all samples adds no real additional information. This image is in Figure 7.

A screenshot of a cell phone

Description automatically generated

**Figure 7**: Pairwise Comparison of All Top 10 Plus 6 Genes**.**

The next result shows a plot made with the ggplot2 package in R to show the simulated means for UL and non-UL samples for each of the top 10 plus six UL risk genes. The cytobands of each gene is a factor aesthetic used to distinguish what cytoband each of these genes belongs to. Three sets of scatters are close to each other but not the same exact expression values. The bottom expression values of the first set of close genes are *GRIP1* and *TH*. The second close group is in the middle as genes *CYTH4* and *SOCS3*. The last close scatter genes are higher in expression values for genes *TNRC6B* and *PYCR1*. Any genes below the red line are under expressed in UL compared to non-UL samples, and anything above the red line are over expressed in UL compared to non-UL samples. This ggplot2 image in in Figure 8.

A close up of a map

Description automatically generated

**Figure 8:** Comparison of Simulated Means for Non-UL and UL Top 10 Plus Six Genes

**Table 4:** Top 16 Genes Differentially Expressed in Subset

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene** | **HGNC Gene Name** | **Strand** | **Cytoband** |
| ASPSCR1 | alveolar soft part sarcoma chromosome region, candidate 1 | + | hs|17q25.3 |
| C1QTNF1 | C1q and tumor necrosis factor related protein 1 | + | hs|17q25.3 |
| CARD10 | caspase recruitment domain family, member 10 | - | hs|22q13.1 |
| CBX2 | chromobox homolog 2 | + | hs|17q25.3 |
| CBX7 | chromobox homolog 7 | - | hs|22q13.1 |
| EIF4A3 | eukaryotic translation initiation factor 4A3 | - | hs|17q25.3 |
| GRAP2 | GRB2-related adaptor protein 2 | + | hs|22q13.1 |
| GRIP1 | glutamate receptor interacting protein 1 | - | hs|12q14.3 |
| KDELR3 | KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 | + | hs|22q13.1 |
| MICALL1 | MICAL-like 1 | + | hs|22q13.1 |
| PYCR1 | pyrroline-5-carboxylate reductase 1 | - | hs|17q25.3 |
| RAC3 | ras-related C3 botulinum toxin substrate 3 (rho family, small GTP binding protein Rac3) | + | hs|17q25.3 |
| SLC38A10 | solute carrier family 38, member 10 | - | hs|17q25.3 |
| SOCS3 | suppressor of cytokine signaling 3 | - | hs|17q25.3 |
| TH | tyrosine hydroxylase | - | hs|11p15.5 |
| ZNF750 | zinc finger protein 750 | - | hs|17q25.3 |

**Table 5:** Bottom 16 Genes Differentially Expressed in Subset

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene** | **HGNC gene name** | **Strand** | **Cytoband** |
| AZI1 | 5-azacytidine induced 1 | - | hs|17q25.3 |
| BAIAP2 | BAI1-associated protein 2 | + | hs|17q25.3 |
| CD7 | CD7 molecule | - | hs|17q25.3 |
| DCXR | dicarbonyl/L-xylulose reductase | - | hs|17q25.3 |
| DDX17 | DEAD (Asp-Glu-Ala-Asp) box polypeptide 17 | - | hs|22q13.1 |
| GAA | glucosidase, alpha; acid | + | hs|17q25.3 |
| IFITM3 | interferon induced transmembrane protein 3 | - | hs|11p15.5 |
| PICK1 | protein interacting with PRKCA 1 | + | hs|22q13.1 |
| PLA2G6 | phospholipase A2, group VI (cytosolic, calcium-independent) | - | hs|22q13.1 |
| RASSF7 | Ras association (RalGDS/AF-6) domain family (N-terminal) member 7 | + | hs|11p15.5 |
| RPL3 | ribosomal protein L3 | - | hs|22q13.1 |
| SIRT3 | sirtuin 3 | - | hs|11p15.5 |
| SIRT7 | sirtuin 7 | - | hs|17q25.3 |
| SLC16A8 | solute carrier family 16, member 8 (monocarboxylic acid transporter 3) | - | hs|22q13.1 |
| TBCD | tubulin folding cofactor D | + | hs|17q25.3 |
| TRIOBP | TRIO and F-actin binding protein | + | hs|22q13.1 |

**Table 6:** Top 16 Fold Change in Subset

|  |  |  |  |
| --- | --- | --- | --- |
| **Genes** | **HGNC Gene Name** | **Strand** | **Cytoband** |
| APOBEC3F | apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3F | + | hs|22q13.1 |
| ASCL2 | achaete-scute complex homolog 2 (Drosophila) | - | hs|11p15.5 |
| ASPSCR1 | alveolar soft part sarcoma chromosome region, candidate 1 | + | hs|17q25.3 |
| CBX2 | chromobox homolog 2 | + | hs|17q25.3 |
| CCDC57 | coiled-coil domain containing 57 | - | hs|17q25.3 |
| CYTH4 | cytohesin 4 | + | hs|22q13.1 |
| FASN | fatty acid synthase | - | hs|17q25.3 |
| FSCN2 | fascin homolog 2, actin-bundling protein, retinal (Strongylocentrotus purpuratus) | + | hs|17q25.3 |
| GRIP1 | glutamate receptor interacting protein 1 | - | hs|12q14.3 |
| HMGA2 | high mobility group AT-hook 2 | + | hs|12q14.3 |
| KDELR3 | KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 | + | hs|22q13.1 |
| NPTX1 | neuronal pentraxin I | - | hs|17q25.3 |
| PYCR1 | pyrroline-5-carboxylate reductase 1 | - | hs|17q25.3 |
| RAC3 | ras-related C3 botulinum toxin substrate 3 (rho family, small GTP binding protein Rac3) | + | hs|17q25.3 |
| TH | tyrosine hydroxylase | - | hs|11p15.5 |
| TNRC6B | trinucleotide repeat containing 6B | + | hs|22q13.1 |

**Table 7:** Majority of 10 Most Differentially Expressed Gene Up and Down

|  |  |  |  |
| --- | --- | --- | --- |
| **Genes** | **HGNC Gene Name** | **Strand** | **Cytoband** |
| ADSL | adenylosuccinate lyase | + | hs|22q13.1 |
| EPS8L2 | EPS8-like 2 | + | hs|11p15.5 |
| GRIP1 | glutamate receptor interacting protein 1 | - | hs|12q14.3 |
| INS | insulin | - | hs|11p15.5 |
| KDELR3 | KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 | + | hs|22q13.1 |
| MGAT3 | mannosyl (beta-1,4-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase | + | hs|22q13.1 |
| MICALL1 | MICAL-like 1 | + | hs|22q13.1 |
| RPLP2 | ribosomal protein, large, P2 | + | hs|11p15.5 |
| SCT | secretin | - | hs|11p15.5 |
| TNNI2 | troponin I type 2 (skeletal, fast) | + | hs|11p15.5 |

**Table 8:** Top 16 Genes in Fold Change from All

|  |  |  |  |
| --- | --- | --- | --- |
| **Genes** | **HGNC Gene Name** | **Strand** | **Cytoband** |
| TNN | tenascin N | + | hs|1q25.1 |
| GRP | gastrin-releasing peptide | + | hs|18q21.32 |
| PPFIA4 | protein tyrosine phosphatase, receptor type, f polypeptide (PTPRF), interacting protein (liprin), alpha 4 | + | hs|1q32.1 |
| GRIA2 | glutamate receptor, ionotropic, AMPA 2 | + | hs|4q32.1 |
| CARTPT | CART prepropeptide | + | hs|5q13.2 |
| PRL | prolactin | - | hs|6p22.3 |
| DCXR | dicarbonyl/L-xylulose reductase | - | hs|17q25.3 |
| CAPN6 | calpain 6 | - | hs|Xq23 |
| DLK1 | delta-like 1 homolog (Drosophila) | + | hs|14q32.2 |
| AKR1B10 | aldo-keto reductase family 1, member B10 (aldose reductase) | + | hs|7q33 |
| KIAA1199 | KIAA1199 | + | hs|15q25.1 |
| CHI3L1 | chitinase 3-like 1 (cartilage glycoprotein-39) | - | hs|1q32.1 |
| IL17B | interleukin 17B | - | hs|5q32 |
| FOLH1B | folate hydrolase 1B | + | 11q14.3 |
| PLP1 | proteolipid protein 1 | + | Xq22.2 |
| STMN2 | stathmin-like 2 | + | hs|8q21.13 |

**Table 9:** Top 16 Genes Differentially Expressed in All

|  |  |  |  |
| --- | --- | --- | --- |
| **Genes** | **HGNC Gene Name** | **Strand** | **Cytoband** |
| TNN | tenascin N | + | hs|1q25.1 |
| GRP | gastrin-releasing peptide | + | hs|18q21.32 |
| PPFIA4 | protein tyrosine phosphatase, receptor type, f polypeptide (PTPRF), interacting protein (liprin), alpha 4 | + | hs|1q32.1 |
| GRIA2 | glutamate receptor, ionotropic, AMPA 2 | + | hs|4q32.1 |
| CARTPT | CART prepropeptide | + | hs|5q13.2 |
| PRL | prolactin | - | hs|6p22.3 |
| DCXR | dicarbonyl/L-xylulose reductase | - | hs|17q25.3 |
| CAPN6 | calpain 6 | - | hs|Xq23 |
| DLK1 | delta-like 1 homolog (Drosophila) | + | hs|14q32.2 |
| AKR1B10 | aldo-keto reductase family 1, member B10 (aldose reductase) | + | hs|7q33 |
| KIAA1199 | KIAA1199 | + | hs|15q25.1 |
| CHI3L1 | chitinase 3-like 1 (cartilage glycoprotein-39) | - | hs|1q32.1 |
| IL17B | interleukin 17B | - | hs|5q32 |
| FOLH1B | folate hydrolase 1B | + | 11q14.3 |
| PLP1 | proteolipid protein 1 | + | Xq22.2 |
| STMN2 | stathmin-like 2 | + | hs|8q21.13 |

**Table 10**: Least Expressed 16 Genes in All

|  |  |  |  |
| --- | --- | --- | --- |
| **Genes** | **HGNC Gene Name** | **Strand** | **Cytoband** |
| FABP1 | fatty acid binding protein 1, liver | - | hs|2p11.2 |
| GRIK4 | glutamate receptor, ionotropic, kainate 4 | + | hs|11q23.3 |
| GRM8 | glutamate receptor, metabotropic 8 | - | hs|7q31.33 |
| INSM1 | insulinoma-associated 1 | + | hs|20p11.23 |
| KLHDC4 | kelch domain containing 4 | - | hs|16q24.2 |
| KLK2 | kallikrein-related peptidase 2 | + | hs|19q13.33 |
| LIG4 | ligase IV, DNA, ATP-dependent | - | hs|13q33.3 |
| MORC1 | MORC family CW-type zinc finger 1 | - | hs|3q13.13 |
| POU3F2 | POU class 3 homeobox 2 | + | hs|6q16.1 |
| SOX11 | SRY (sex determining region Y)-box 11 | + | hs|2p25.2 |
| USP32P2 | ubiquitin specific peptidase 32 pseudogene 2 | - | hs|17p11.2 |
| DNTT | DNA nucleotidylexotransferase | + | hs|10q24.1 |
| RCVRN | recoverin | - | hs|17p13.1 |
| SUV39H1 | suppresor of variegation 3-9 homolog1 | + | hs|Xp11.23 |
| SYNGR3 | synaptogyrin 3 | + | hs|16p13.3 |
| TLX3 | T cell leukemia homeobox 3 | + | hs|5q35.1 |

**Table 11**: Machine Learning Results on Top 10 Plus 6

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SampleID** | **RF** | **RF2** | **LDA** | **GBM** | **KNN** | **RPART** | **GLM** | **Combined** | **Type** |
| gsm1667145 | nonUL | UL | nonUL | UL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm336254 | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL |
| gsm336258 | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL |
| gsm336260 | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL |
| gsm336270 | UL | UL | nonUL | nonUL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm336273 | nonUL | nonUL | nonUL | nonUL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm336276 | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL |
| gsm52662 | nonUL | nonUL | nonUL | nonUL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm52663 | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL |
| gsm52665 | UL | UL | nonUL | UL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm52667 | UL | UL | UL | UL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm52669 | nonUL | nonUL | nonUL | nonUL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm9099 | UL | UL | nonUL | UL | UL | UL | nonUL | UL | nonUL |
| gsm569425 | nonUL | nonUL | nonUL | nonUL | nonUL | UL | nonUL | nonUL | nonUL |
| gsm569427 | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL |
| gsm336202ul | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | UL |
| gsm336208ul | UL | UL | UL | UL | UL | UL | UL | UL | UL |
| gsm336209ul | UL | UL | UL | UL | UL | UL | UL | UL | UL |
| gsm336214ul | UL | UL | UL | UL | UL | UL | UL | UL | UL |
| gsm336215ul | UL | UL | UL | UL | UL | UL | UL | UL | UL |
| gsm336218ul | UL | UL | UL | UL | UL | UL | UL | UL | UL |
| gsm336220ul | nonUL | nonUL | UL | UL | nonUL | nonUL | nonUL | UL | UL |
| gsm336229ul | UL | UL | UL | UL | UL | nonUL | UL | UL | UL |
| gsm336232ul | UL | UL | UL | UL | UL | nonUL | UL | UL | UL |
| gsm336234ul | UL | UL | UL | UL | nonUL | UL | UL | UL | UL |
| gsm336238ul | UL | UL | nonUL | UL | nonUL | UL | nonUL | nonUL | UL |
| gsm336239ul | UL | UL | nonUL | nonUL | UL | UL | nonUL | UL | UL |
| gsm336240ul | UL | UL | UL | UL | UL | UL | UL | UL | UL |
| gsm336241ul | nonUL | nonUL | UL | nonUL | nonUL | nonUL | UL | UL | UL |
| gsm336245ul | nonUL | nonUL | UL | nonUL | UL | nonUL | UL | UL | UL |
| gsm336248ul | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | nonUL | UL |
| gsm38689ul | nonUL | nonUL | nonUL | UL | UL | UL | nonUL | UL | UL |
| gsm38692ul | nonUL | nonUL | nonUL | nonUL | nonUL | UL | nonUL | nonUL | UL |
| gsm9094ul | UL | UL | UL | nonUL | UL | UL | nonUL | UL | UL |
| gsm569429ul | UL | UL | nonUL | UL | UL | UL | nonUL | UL | UL |
| **results** | **0.69** | **0.66** | **0.77** | **0.69** | **0.77** | **0.54** | **0.74** | **0.86** | **100** |

**Table 12:** Machine Learning Results on All Data Sets

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Sets** | **RF** | **RF2** | **LDA** | **GBM** | **KNN** | **RPART** | **GLM** | **Combined** | **Type** |
| TOP 10 Plus 6 DE 130 | 0.69 | 0.66 | 0.77 | 0.69 | 0.77 | 0.6 | 0.74 | 0.83 | 100 |
| Top 16 DE 130 | 0.61 | 0.67 | 0.78 | 0.67 | 0.89 | 0.61 | 0.72 | 0.92 | 100 |
| Bottom 16 DE 130 | 0.53 | 0.58 | 0.42 | 0.42 | 0.53 | 0.58 | 0.33 | 0.75 | 100 |
| Top 10 Plus 6 Fold 130 | 0.69 | 0.69 | 0.61 | 0.69 | 0.72 | 0.58 | 0.58 | 0.78 | 100 |
| Majority 10 | 0.72 | 0.75 | 0.75 | 0.72 | 0.75 | 0.64 | 0.72 | 0.75 | 100 |
| Universe Top 16 Fold | 0.92 | 0.94 | 0.94 | 0.89 | 0.89 | 0.78 | 0.78 | 1 | 100 |
| Universe Top 16 DE | 0.89 | 0.94 | 0.89 | 0.89 | 0.86 | 0.86 | 0.81 | 0.94 | 100 |
| Universe Bottom 16 DE | 0.47 | 0.42 | 0.36 | 0.5 | 0.5 | 0.5 | 0.42 | 0.69 | 100 |

Conclusions

.

literature cited

Aissani, B., Zhang, K., and Wiener, H. (2015). Evaluation of GWAS candidate susceptibility loci for uterine leiomyoma in the multi-ethnic NIEHS uterine fibroid study. *Frontiers in Genetics*, 6, 241. DOI:10.3389/fgene.2015.00241

Bioconductor, version 3.8, (2019). Bioconductor: Open Source Software for Bioinformatics. Retrieved March 3, 2019 from <https://www.bioconductor.org/install/>

Bondagji, N., Morad, F., Al-Nefaei, A., Khan, I., Elango, R., Abdullah, L., ..., Shaik, N. (2017). Replication of GWAS loci revealed the moderate effect of TNRC6B locus on susceptibility of Saudi women to develop uterine leiomyomas. *Journal of Obstetrics and Gynaecology*, 43(2):330-338. DOI:10.1111/jog.13217

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2018, March). Breiman and cutler's random forests for classiﬁcation and regression. ‘randomForest,’ version: 4.6-14. Retrieved July 2019 from <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>

Cha, P, Takahashi, A., Hosono, N., Low, S., Kamatani, N., Kubo, M., & Nakamura, Y. (2011). A genome-wide association study identifies three loci associated with susceptibility to uterine fibroids. *Nature Genetics*, 43(5).

Chang, J. (2015, November). Collapsed gibbs sampling methods for topic models. ‘lda,’ version: 1.4.2. Retrieved July 2019 from <https://cran.r-project.org/web/packages/lda/lda.pdf>

Crabtree, J., Jelinsky, S., Harris, H., Choe, S., Cotreau, M., Kimberland, M., … Walker, C. (2009). Comparison of human and rat uterine leiomyomata: identification of a dysregulated mammalian target of rapamycin pathway. *Cancer Research*, 69(15), 6171-8.

Dvorská1, D., Braný, D., Danková, Z., Halašová, E., & Višňovský, J. (2017). Molecular and clinical treatment of uterine leiomyomas. *Tumor Biology*, 39(6). DOI: 10.1177/1010428317710226.

Edgar, R., Domrachev, M., & Lash, A. (2019). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.

Edwards, T., Hartmann, K., & Edwards, D. (2013). Variants in BET1L and TNRC6B associate with increasing fibroid volume and fibroid type among European Americans. *Human Genetics,* 132(12). DOI:10.1007/s00439-013-1340-1

Eggert, S., Huyck, K., Somasundaram, P., Kavalla, R., Stewart, E., Lu, A., … Morton, C. (2012). Genome-wide linkage and association analyses implicate FASN in predisposition to uterine leiomyomata. *American Journal of Human Genetics*, 91(4), 621–628. DOI: 10.1016/j.ajhg.2012.08.009

ENSEMBL, (2019). Human genes (GRCh38.12) from ensembl genes 97. Retrieved from <http://uswest.ensembl.org/biomart/martview/7cbd4e5eb92adf75e973b6e01e016a03>

Francois, R., Lionel, H., & Muller, K. (2019, July). A grammar of data manipulation, ‘dlyr’ R package, version: 0.8.3. Retrieved July 5, 2019 from <https://cran.r-project.org/web/packages/dplyr/dplyr.pdf>

Galili, T., O'Callaghan, A., Sidi, J., & Benjamin, Y. (2019). Package 'heatmaply.' Retrieved June 3, 2019, from <https://cran.r-project.org/web/packages/heatmaply/heatmaply.pdf>

Greenwell, B., Boehmke, B., and Cunningham, J. (2019, January). Generalized boosted regression models (‘gbm,’ version: 2.1.5). Retrieved July 2019 from <https://cran.r-project.org/web/packages/gbm/gbm.pdf>

Hahne, F. (2019). The Gviz user guide. Retrieved June 3, 2019, from https://manualzz.com/doc/4237818/the-gviz-user-guide.

Hellwege, J. N., Jeff, J. M., Wise, L. A., Gallagher, C. S., Wellons, M., Hartmann, K. E., … Velez Edwards, D. R. (2017). A multi-stage genome-wide association study of uterine fibroids in African Americans. *Human Genetics*, 136(10), 1363–1373. DOI:10.1007/s00439-017-1836-1

Hodge, J.C., Kim, T., Dreyfuss, J.M., Somasundaram, P., Christacos, N.C., Rouselle, M., … Morton, C.C. (2012). Expression profiling of uterine leiomyomata cytogenetic subgroups reveals distinct signatures in matched myometrium: transcriptional profiling of the t (12;14) and evidence in support of predisposing genetic heterogeneity. *Human Molecular Genetics*, 21, 102312–2329. DOI:10.1093/hmg/dds051

Hoffman, P, Milliken, D, Gregg, L., Davis, R., & Gregg, J. (2004). Molecular characterization of uterine fibroids and its implication for underlying mechanisms of pathogenesis. *Fertility and Sterility*,82(3), 639-49.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Hunt, T. (2019, April). Classiﬁcation and regression training. ‘caret,’ version: 6.0-84. Retrieved July 2019 from <https://cran.r-project.org/web/packages/caret/caret.pdf>

Liu, B., Wang, T., Jiang, J., Li, M., Ma, W., Wu, H., & Zhou, Q. (2018). Association of BET1L and TNRC6B with uterine leiomyoma risk and its relevant clinical features in Han Chinese population. *Scientific Reports*, 8,7401. DOI:10.1038/s41598-018-25792-z

Maindonald, J. (2008, January). Using r for data analysis and graphics: introduction, code and commentary. Retrieved July 2019 from <https://cran.r-project.org/doc/contrib/usingR.pdf>

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., Lin, C. (2019, June). Misc functions of the department of statistics, probability theory group (Formerly: E1071), tU wien (‘e1071,’ version: 1.7-2). Retrieved July 2019 from <https://cran.r-project.org/web/packages/e1071/e1071.pdf>

Miyata, T., Sonoda, K., Tomikawa, J., Tayama, C., Okamura, K., Maehara, K., … Nakabayashi, K. (2015). Genomic, Epigenomic, and Transcriptomic Profiling towards Identifying Omics Features and Specific Biomarkers That Distinguish Uterine Leiomyosarcoma and Leiomyoma at Molecular Levels. *Sarcoma* *2015*.

Quade, B.J., Mutter, G.L., & Morton, C.C. (2004). Comparison of Gene Expression in Uterine Smooth Muscle Tumors. Gene Expression Omnibus. GEO Accession ID: GSE764. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE**764**](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE764)

R (2019). CRAN: Comprehensive R Archive Network. R, version 3.6.0, for Windows 64-bit Operating System. Retrieved March 3, 2019 from <https://cran.cnr.berkeley.edu/>

Rafnar, T., Gunnarsson, B., Stefansson, O.A., Sulem, P., Ingason, A., Frigge, M.L., … Stefansson, K. (2018). Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-related traits. *Nature Communications*, 9:3636. DOI:10.1038/s41467-018-05428-6

Ripley, B., Venables, B., Bates, D., Hornik, K., Gebhardt, A., Firth, D. (2019, April). Support functions and datasets for venables and ripley's MASS (‘MASS,’ version 7.3-51.4). Retrieved July 2019 from <https://cran.r-project.org/web/packages/MASS/MASS.pdf>

Sarker, D., (2018, November). Trellis graphics for r, ‘lattice’ r package, version: 0.20-38. Retrieved July 5, 2019 from <https://cran.r-project.org/web/packages/lattice/lattice.pdf>

Therneau, T., Atkinson, B., & Ripley, B. (April 2019). Recursive partitioning and regression trees. ‘rpart,’ version: 4.1-15. Retrieved July 2019 from <https://cran.r-project.org/web/packages/rpart/rpart.pdf>

Vanharanta, S., Pollard, P.J., Lehtonen, H.J., Laiho, P., Sjoberg, J., Leminen, A., … Aaltonen, L.A. (2006). Distinct expression profile in fumarate-hydratase-deficient uterine fibroids. *Human Molecular Genetics*, 15(1), 97-103.

Wickham, H. (2019, June). Create elegant data visualisations using the grammar of graphics. ‘ggplot2,’ version 3.2.0. Retrieved July 2019 from <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>

Yin, T., Dianne Cook, D., & Lawrence, M. (2012): Ggbio: An R package for extending the grammar of graphics for genomic data Genome Biology 13: R77. Retrieved June 3, 2019, from <http://www.bioconductor.org/packages/release/bioc/vignettes/ggbio/inst/doc/ggbio.pdf>

Zhang, D., Sun, C., Ma, C., Dai, H., & Zhang, W. (2012). Data mining of spatial–temporal expression of genes in the human endometrium during the window of implantation. *Reproductive Sciences*, 19(10), 1085-98. DOI:10.1177/1933719112442248

Zavadil, J., Ye, H., Liu, Z., Wu, J., Lee, P., Hernando, E., … Wei, J.J. (2010). Profiling and functional analyses of microRNAs and their target gene products in human uterine leiomyomas. PLoS One, 5(8). PMID: 20808773

Appendix

1. GPL96. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL96](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL96). This is one of three GEO platforms that was combined with the microarray samples from the five GEO microarray series listed above as items 1 through 5. This platform identified the probe IDs of GSE593, GSE2724, and GSE23112. Only the ‘ID’ column was used to merge with the other four data sets and then most of the columns from GPL6480 were used. There are 22,283 genes and 16 columns of additional information with some directly quoted from the table excel file. These columns are identical to the GPL570 platform because they are both the Affymetrix Human Genome U133 Array, but GPL570 is the ‘Plus 2’ version.
   * 1. ID : this is the ID column to merge with GSE593, GSE2724, and GSE23112 GEO series
     2. GB\_ACC: Factor. This is the gene bank accession number for each gene
     3. SPOT\_ID: Factor. This is either ‘control’ or ‘NA’
     4. Species.Scientific.Name: Factor. This is equal to ‘Homo sapiens’ for all
     5. Annotation.Date: Factor. The date the data platform IDs annotated, all equal ‘Oct 6, 2014’
     6. Sequence.Type: Factor with three values of ‘Exemplar Sequence,’ ‘Control Sequence,’ or ‘Consensus Sequence’
     7. Sequence.Source: Factor with one level of ‘Affymetrix Proprietary Database GenBank.’ Described as ‘the database from which the sequence used to develop this probe set was taken’
     8. Target.Description: Factor with 21,362 levels describing each gene
     9. Representative.Public.ID: Factor. The accession number of a representative sequence.
     10. Gene.Title: Factor. The title of the gene represented by the probe set.
     11. Gene.Symbol: UniGene gene symbol
     12. ENTREZ\_GENE\_ID: Factor. ENTREZ gene database UID
     13. RefSeq.Transcript.ID: Factor. References to multiple sequences in RefSeq
     14. Gene.Ontology.Biological.Process: Factor. ‘Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence’
     15. Gene.Ontology.Cellular.Component: Factor. ‘Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
     16. Gene.Ontology.Molecular.Function: ‘Gene Ontology Consortium Molecular Function derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
2. GPL570. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL570](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570). This is one of three GEO platforms that was combined with the microarray samples from the five GEO microarray series listed above as items 1 through 5. This platform identified the probe IDs of GSE13319. There are 54,675 genes and 16 columns that are identical to GPL96, because this is the Affymetrix Human Genome U133 Plus 2.0 Array and GPL96 is the Affymetrix Human Genome U133 Array an earlier version.
   * 1. ID : this is the ID column to merge with GSE593, GSE2724, and GSE23112 GEO series
     2. GB\_ACC: Factor. This is the gene bank accession number for each gene
     3. SPOT\_ID: Factor. This is either ‘control’ or ‘NA’
     4. Species.Scientific.Name: Factor. This is equal to ‘Homo sapiens’ for all
     5. Annotation.Date: Factor. The date the data platform IDs annotated, all equal ‘Oct 6, 2014’
     6. Sequence.Type: Factor with three values of ‘Exemplar Sequence,’ ‘Control Sequence,’ or ‘Consensus Sequence’
     7. Sequence.Source: Factor with one level of ‘Affymetrix Proprietary Database GenBank.’ Described as ‘the database from which the sequence used to develop this probe set was taken’
     8. Target.Description: Factor with 21,362 levels describing each gene
     9. Representative.Public.ID: Factor. The accession number of a representative sequence.
     10. Gene.Title: Factor. The title of the gene represented by the probe set.
     11. Gene.Symbol: UniGene gene symbol
     12. ENTREZ\_GENE\_ID: Factor. ENTREZ gene database UID
     13. RefSeq.Transcript.ID: Factor. References to multiple sequences in RefSeq
     14. Gene.Ontology.Biological.Process: Factor. ‘Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence’
     15. Gene.Ontology.Cellular.Component: Factor. ‘Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
     16. Gene.Ontology.Molecular.Function: ‘Gene Ontology Consortium Molecular Function derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
3. GPL6480. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL6480](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6480). This is one of three GEO platforms that was combined with the microarray samples from the five GEO microarray series listed above as items 1 through 5. This platform identified the probe IDs of GSE68295. There are 41,108 genes and 17 identifying columns in this platform. The columns in this data set are all factors. The following are the listed columns used to merge all other GSE series and GPL platforms to while keeping only the needed columns from this table. The column IDs are labeled as how they are described in the downloaded SOFT text file.
   * + 1. ID : Agilent feature number
       2. SPOT\_ID : Spot identifier
       3. CONTROL\_TYPE : Control type
       4. REFSEQ : RefSeq Accession number
       5. GB\_ACC : GenBank Accession number
       6. GENE : Entrez Gene ID
       7. GENE\_SYMBOL : Gene Symbol
       8. GENE\_NAME : Gene Name
       9. UNIGENE\_ID : UnigeneID
       10. ENSEMBL\_ID : EnsemblID
       11. TIGR\_ID : TIGRID
       12. ACCESSION\_STRING : Accession String
       13. CHROMOSOMAL\_LOCATION : Chromosomal Location
       14. CYTOBAND : Cytoband
       15. DESCRIPTION : Description
       16. GO\_ID : GoIDs
       17. SEQUENCE : Sequence
4. GSE593. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE593](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE593). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL96 GEO platform listed as item 7. This data set shares the same Probe ID as GSE23112 and GSE2724 because they all share GPL96. This data contributed five UL and five non-UL samples to the 121 total samples. There are 22,283 genes in this raw data as the 22,283 rows. There are 11 columns used from this file as:
   * + 1. ID\_REF: The microarray Affymetrix ID
       2. GSM9093: UL
       3. GSM9094: UL
       4. GSM9095: UL
       5. GSM9096: UL
       6. GSM9097: UL
       7. GSM9098: non-UL
       8. GSM9099: non-UL
       9. GSM9100: non-UL
       10. GSM9101: non-UL
       11. GSM9102: non-UL
5. GSE2724. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE2724](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2724). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL96 GEO platform listed as item 7. There are 7 UL and 11 non-UL samples as headers with one probe ID column the same as GSE593 and GSE23112. There are 22,283 genes in this raw data as rows and 19 columns as:
   * + 1. ID\_REF: The Affymetrix microarray probe ID
       2. GSM38689: UL
       3. GSM38690: UL
       4. GSM38691: UL
       5. GSM38692: UL
       6. GSM38693: UL
       7. GSM38694: UL
       8. GSM38695: UL
       9. GSM52661: non-UL
       10. GSM52662: non-UL
       11. GSM52663: non-UL
       12. GSM52664: non-UL
       13. GSM52665: non-UL
       14. GSM52666: non-UL
       15. GSM52667: non-UL
       16. GSM52668: non-UL
       17. GSM52669: non-UL
       18. GSM52670: non-UL
       19. GSM52671: non-UL
6. GSE68295. Retrieved March 2019 from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE68295>. This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL6480 GEO platform listed as item 8. This data set added 3 UL and 3 non-UL samples to the total 121 samples, but also was needed for the attached information the platform to this data set contained. The various recognized gene names, chromosome, cytoband information, and other meta columns was useful for the analysis. This raw data set had 41,078 genes as rows and 13 columns of UL, non-UL, and sarcoma UL samples. Only the three UL and three non-UL samples were used in this research:
   * + 1. ID\_REF: Affymetrix Probe ID
       2. GSM1667144: non-UL
       3. GSM1667145: non-UL
       4. GSM1667146: non-UL
       5. GSM1667147: UL
       6. GSM1667148: UL
       7. GSM1667149: UL
       8. GSM1667150: UL sarcoma, not added to this research
       9. GSM1667151: UL sarcoma, not added to this research
       10. GSM1667152: UL sarcoma, not added to this research
       11. GSM1667153: UL sarcoma, not added to this research
       12. GSM1667154: UL sarcoma, not added to this research
       13. GSM1667155: UL sarcoma, not added to this research
7. GSE13319. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE13319](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13319). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL570 GEO platform listed as item 6. This data set used only the human samples from a combined set of human and rat UL. In total 50 UL samples and 27 non-UL samples were added to the 121 total samples. This data had 54,675 genes. The original data set included mouse samples, but for the purposes of this study on human females, those rat samples were excluded. In this file, there are 54,675 rows as genes and 78 columns as:
   1. ID\_REF: Affymetrix probe ID
   2. GSM336202: UL
   3. GSM336203: UL
   4. GSM336204: UL
   5. GSM336205L: UL
   6. GSM336206: UL
   7. GSM336207: UL
   8. GSM336208: UL
   9. GSM336209: UL
   10. GSM336210: UL
   11. GSM336211: UL
   12. GSM336212: UL
   13. GSM336213: UL
   14. GSM336214: UL
   15. GSM336215: UL
   16. GSM336216: UL
   17. GSM336217: UL
   18. GSM336218: UL
   19. GSM336219: UL
   20. GSM336220: UL
   21. GSM336221: UL
   22. GSM336222: UL
   23. GSM336223: UL
   24. GSM336224: UL
   25. GSM336225: UL
   26. GSM336226: UL
   27. GSM336227: UL
   28. GSM336228: UL
   29. GSM336229: UL
   30. GSM336230: UL
   31. GSM336231: UL
   32. GSM336232: UL
   33. GSM336233: UL
   34. GSM336234: UL
   35. GSM336235: UL
   36. GSM336236: UL
   37. GSM336237: UL
   38. GSM336238: UL
   39. GSM336239: UL
   40. GSM336240: UL
   41. GSM336241: UL
   42. GSM336242: UL
   43. GSM336243: UL
   44. GSM336244: UL
   45. GSM336245: UL
   46. GSM336246: UL
   47. GSM336247: UL
   48. GSM336248: UL
   49. GSM336249: UL
   50. GSM336250: UL
   51. GSM336251: UL
   52. GSM336252: non-UL
   53. GSM336253: non-UL
   54. GSM336254: non-UL
   55. GSM336255: non-UL
   56. GSM336256: non-UL
   57. GSM336257: non-UL
   58. GSM336258: non-UL
   59. GSM336259: non-UL
   60. GSM336260: non-UL
   61. GSM336261: non-UL
   62. GSM336262: non-UL
   63. GSM336263: non-UL
   64. GSM336264: non-UL
   65. GSM336265: non-UL
   66. GSM336266: non-UL
   67. GSM336267: non-UL
   68. GSM336268: non-UL
   69. GSM336269: non-UL
   70. GSM336270: non-UL
   71. GSM336271: non-UL
   72. GSM336272: non-UL
   73. GSM336273: non-UL
   74. GSM336274: non-UL
   75. GSM336275: non-UL
   76. GSM336276: non-UL
   77. GSM336277: non-UL
   78. GSM336278: non-UL
8. GSE23112. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE23112](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23112). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL96 GEO platform listed as item 7. This data contributed five UL and five non-UL samples to the 121 samples total. With the same Probe ID column as GSE2724 and GSE593. There are 22,283 genes in this raw data set as rows and 11 columns as:
   * + 1. ID\_REF: probe ID for each gene in the microarray sample of this data set
       2. GSM569424: non-UL
       3. GSM569425: non-UL
       4. GSM569426: non-UL
       5. GSM569427: non-UL
       6. GSM569428: non-UL
       7. GSM569429: UL
       8. GSM569430: UL
       9. GSM569431: UL
       10. GSM569432: UL
       11. GSM569433: UL
9. All\_analysis.R. Accessible from https://www.dropbox.com/s/b8a9fjy8wcfptd4/All\_analysis\_2.R?dl=0 . This is the R script for all data tables and images produced on the raw data of items 1 through 8 of the Appendix. The version this script used is version 3.6. The packages used are listed in the script but commented out. The packages installed into R to run the script in some sections are: ‘dplyr’,’rpart’,’caret’,’MASS’,’e1071’,’randomForest’,’ggplot2’,’lattice’,’heatmaply’,’plotly’,’Gviz’, ‘ComplexHeatmap’,’GenomicRanges’, and ’UsingR’ To search for the specific data table made or plot made, select the magnifying glass in the toolbar in RStudio (a GUI for R) and type in the csv file name or plot name. Then backtrack to the steps used since the last file read in to see the steps used to create it.
10. GSE\_array\_meta.csv. This is the same exact columns as GPL6480 renamed to know it is all the meta information to the samples the five GEO series studies have in common for this research on UL and non-UL gene expression data. There are 17 columns identical to item 3 in this Appendix. This file is retrievable from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL6480](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6480) and it is 25.6 mb in size with 41,078 rows of genes by 17 columns of genes information. This file is too big to be placed in the github file folder by 1 mb as the limit is 25 mb for files stored to the folder of these files in the data repository of Github.com . The following columns are identical to those in item 3 of this Appendix:
    * + 1. ID : Agilent feature number
        2. SPOT\_ID : Spot identifier
        3. CONTROL\_TYPE : Control type
        4. REFSEQ : RefSeq Accession number
        5. GB\_ACC : GenBank Accession number
        6. GENE : Entrez Gene ID
        7. GENE\_SYMBOL : Gene Symbol
        8. GENE\_NAME : Gene Name
        9. UNIGENE\_ID : UnigeneID
        10. ENSEMBL\_ID : EnsemblID
        11. TIGR\_ID : TIGRID
        12. ACCESSION\_STRING : Accession String
        13. CHROMOSOMAL\_LOCATION : Chromosomal Location
        14. CYTOBAND : Cytoband
        15. DESCRIPTION : Description
        16. GO\_ID : GoIDs
        17. SEQUENCE : Sequence
11. mrg5.csv. This file is 1.1 Gb in size, too large for the github repository. It is the data of all five series merged together, including duplicate entries and missing values. This is the file before it was cleaned by removing duplicates and missing values. It has 1,954,853 rows of genes, and 123 columns that include 121 samples of UL or non-UL after labeling the UL samples with an extension, ‘UL,’ to the end of the corresponding UL sample ID. The two columns that are not one of the 121 samples are the ‘GENE’ and ‘CYTOBAND’ columns from the GSE\_array\_meta.csv data table. The ‘GENE’ column is the ENTREZ gene ID and the ‘CYTOBAND’ column is the cytoband location of the gene in each chromosome. The columns are not listed in order of sample type like later data sets, so there is a mix between UL and non-UL samples in the organization of the columns. This file was uploaded to dropbox and made shareable at <https://www.dropbox.com/s/bwkiq1h3872u2j2/mrg5.csv?dl=0> . The following are the columns or variables in this file:
    1. GENE: The Entrez gene ID
    2. CYTOBAND: The cytoband location of each gene
    3. GSM1667144: non-UL
    4. GSM1667145: non-UL
    5. GSM1667146: non-UL
    6. GSM1667147UL: UL
    7. GSM1667148UL: UL
    8. GSM1667149UL: UL
    9. GSM336202UL: UL
    10. GSM336203UL: UL
    11. GSM336204UL: UL
    12. GSM336205UL : UL
    13. GSM336206UL: UL
    14. GSM336207UL: UL
    15. GSM336208UL: UL
    16. GSM336209UL: UL
    17. GSM336210UL: UL
    18. GSM336211UL: UL
    19. GSM336212UL: UL
    20. GSM336213UL: UL
    21. GSM336214UL: UL
    22. GSM336215UL: UL
    23. GSM336216UL: UL
    24. GSM336217UL: UL
    25. GSM336218UL: UL
    26. GSM336219UL: UL
    27. GSM336220UL: UL
    28. GSM336221UL: UL
    29. GSM336222UL: UL
    30. GSM336223UL: UL
    31. GSM336224UL: UL
    32. GSM336225UL: UL
    33. GSM336226UL: UL
    34. GSM336227UL: UL
    35. GSM336228UL: UL
    36. GSM336229UL: UL
    37. GSM336230UL: UL
    38. GSM336231UL: UL
    39. GSM336232UL: UL
    40. GSM336233UL: UL
    41. GSM336234UL: UL
    42. GSM336235UL: UL
    43. GSM336236UL: UL
    44. GSM336237UL: UL
    45. GSM336238UL: UL
    46. GSM336239UL: UL
    47. GSM336240UL: UL
    48. GSM336241UL: UL
    49. GSM336242UL: UL
    50. GSM336243UL: UL
    51. GSM336244UL: UL
    52. GSM336245UL: UL
    53. GSM336246UL: UL
    54. GSM336247UL: UL
    55. GSM336248UL: UL
    56. GSM336249UL: UL
    57. GSM336250UL: UL
    58. GSM336251UL: UL
    59. GSM336252: non-UL
    60. GSM336253: non-UL
    61. GSM336254: non-UL
    62. GSM336255: non-UL
    63. GSM336256: non-UL
    64. GSM336257: non-UL
    65. GSM336258: non-UL
    66. GSM336259: non-UL
    67. GSM336260: non-UL
    68. GSM336261: non-UL
    69. GSM336262: non-UL
    70. GSM336263: non-UL
    71. GSM336264: non-UL
    72. GSM336265: non-UL
    73. GSM336266: non-UL
    74. GSM336267: non-UL
    75. GSM336268: non-UL
    76. GSM336269: non-UL
    77. GSM336270: non-UL
    78. GSM336271: non-UL
    79. GSM336272: non-UL
    80. GSM336273: non-UL
    81. GSM336274: non-UL
    82. GSM336275: non-UL
    83. GSM336276: non-UL
    84. GSM336278: non-UL
    85. GSM38689UL: UL
    86. GSM38690UL: UL
    87. GSM38691UL: UL
    88. GSM38692UL: UL
    89. GSM38693UL: UL
    90. GSM38694UL: UL
    91. GSM38695UL: UL
    92. GSM52661: non-UL
    93. GSM52662: non-UL
    94. GSM52663: non-UL
    95. GSM52664 : non-UL
    96. GSM52665: non-UL
    97. GSM52666: non-UL
    98. GSM52667: non-UL
    99. GSM52668: non-UL
    100. GSM52669: non-UL
    101. GSM52670: non-UL
    102. GSM52671: non-UL
    103. GSM9093UL: UL
    104. GSM9094UL: UL
    105. GSM9095UL: UL
    106. GSM9096UL: UL
    107. GSM9097UL : UL
    108. GSM9098: non-UL
    109. GSM9099: non-UL
    110. GSM9100: non-UL
    111. GSM9101: non-UL
    112. GSM9102: non-UL
    113. GSM569424: non-UL
    114. GSM569425: non-UL
    115. GSM569426: non-UL
    116. GSM569427: non-UL
    117. GSM569428: non-UL
    118. GSM569429UL: UL
    119. GSM569430UL: UL
    120. GSM569431UL: UL
    121. GSM569432UL: UL
    122. GSM569433UL: UL
12. DE\_means\_Per\_Gene\_Chr.csv: This file is the same as the mrg5.csv file above but there are added columns for the number of gene observations the row means were taken to remove duplicate gene names, and a HGNC column for the gene symbol of each gene modified from the mrg5.csv data set. This file can be obtained at: <https://www.dropbox.com/s/x08jm2isb0o4j2z/DE_means_Per_Gene_Chr.csv?dl=0> . It has 12,173 rows of unique genes and 125 columns of 121 samples labeled as ‘UL’ at the end if the sample is a UL sample, and four meta columns:
    1. GENE: Entrez gene ID
    2. CYTOBAND: cytoband location of each gene
    3. GENE\_SYMBOL: the HGNC gene symbol of each name
    4. Counts: the number of times this gene was listed in the larger mrg5.csv file, that the row means for each gene was made to produce this more compact data set
    5. GSM1667144
    6. GSM1667145
    7. GSM1667146
    8. GSM1667147UL
    9. GSM1667148UL
    10. GSM1667149UL
    11. GSM336202UL
    12. GSM336203UL
    13. GSM336204UL
    14. GSM336205UL
    15. GSM336206UL
    16. GSM336207UL
    17. GSM336208UL
    18. GSM336209UL
    19. GSM336210UL
    20. GSM336211UL
    21. GSM336212UL
    22. GSM336213UL
    23. GSM336214UL
    24. GSM336215UL
    25. GSM336216UL
    26. GSM336217UL
    27. GSM336218UL
    28. GSM336219UL
    29. GSM336220UL
    30. GSM336221UL
    31. GSM336222UL
    32. GSM336223UL
    33. GSM336224UL
    34. GSM336225UL
    35. GSM336226UL
    36. GSM336227UL
    37. GSM336228UL
    38. GSM336229UL
    39. GSM336230UL
    40. GSM336231UL
    41. GSM336232UL
    42. GSM336233UL
    43. GSM336234UL
    44. GSM336235UL
    45. GSM336236UL
    46. GSM336237UL
    47. GSM336238UL
    48. GSM336239UL
    49. GSM336240UL
    50. GSM336241UL
    51. GSM336242UL
    52. GSM336243UL
    53. GSM336244UL
    54. GSM336245UL
    55. GSM336246UL
    56. GSM336247UL
    57. GSM336248UL
    58. GSM336249UL
    59. GSM336250UL
    60. GSM336251UL
    61. GSM336252
    62. GSM336253
    63. GSM336254
    64. GSM336255
    65. GSM336256
    66. GSM336257
    67. GSM336258
    68. GSM336259
    69. GSM336260
    70. GSM336261
    71. GSM336262
    72. GSM336263
    73. GSM336264
    74. GSM336265
    75. GSM336266
    76. GSM336267
    77. GSM336268
    78. GSM336269
    79. GSM336270
    80. GSM336271
    81. GSM336272
    82. GSM336273
    83. GSM336274
    84. GSM336275
    85. GSM336276
    86. GSM336277
    87. GSM336278
    88. GSM38689UL
    89. GSM38690UL
    90. GSM38691UL
    91. GSM38692UL
    92. GSM38693UL
    93. GSM38694UL
    94. GSM38695UL
    95. GSM52661
    96. GSM52662
    97. GSM52663
    98. GSM52664
    99. GSM52665
    100. GSM52666
    101. GSM52667
    102. GSM52668
    103. GSM52669
    104. GSM52670
    105. GSM52671
    106. GSM9093UL
    107. GSM9094UL
    108. GSM9095UL
    109. GSM9096UL
    110. GSM9097UL
    111. GSM9098
    112. GSM9099
    113. GSM9100
    114. GSM9101
    115. GSM9102
    116. GSM569424
    117. GSM569425
    118. GSM569426
    119. GSM569427
    120. GSM569428
    121. GSM569429UL
    122. GSM569430UL
    123. GSM569431UL
    124. GSM569432UL
    125. GSM569433UL
13. chr\_loci\_top\_genes.csv. This file was made from the item 12 above by creating a subset of that data set to only include the four cytoband regions the six UL risk genes reside. The columns are identical to the columns above in item 12, but instead 12,173 genes as rows there are now 183 genes as rows and the same 125 columns as above. Some genes do still have duplicate entries even though the data set of item 12 above used the row means per gene to make the mrg5.csv file more compact. This item 13 data set can be obtained at <https://www.dropbox.com/s/z9oqwn73k17xxe6/chr_loci_top_genes.csv?dl=0> .
    1. GENE: Entrez gene ID
    2. CYTOBAND: cytoband location of each gene
    3. GENE\_SYMBOL: the HGNC gene symbol of each name
    4. Counts: the number of times this gene was listed in the larger mrg5.csv file, that the row means for each gene was made to produce this more compact data set
    5. GSM1667144
    6. GSM1667145
    7. GSM1667146
    8. GSM1667147UL
    9. GSM1667148UL
    10. GSM1667149UL
    11. GSM336202UL
    12. GSM336203UL
    13. GSM336204UL
    14. GSM336205UL
    15. GSM336206UL
    16. GSM336207UL
    17. GSM336208UL
    18. GSM336209UL
    19. GSM336210UL
    20. GSM336211UL
    21. GSM336212UL
    22. GSM336213UL
    23. GSM336214UL
    24. GSM336215UL
    25. GSM336216UL
    26. GSM336217UL
    27. GSM336218UL
    28. GSM336219UL
    29. GSM336220UL
    30. GSM336221UL
    31. GSM336222UL
    32. GSM336223UL
    33. GSM336224UL
    34. GSM336225UL
    35. GSM336226UL
    36. GSM336227UL
    37. GSM336228UL
    38. GSM336229UL
    39. GSM336230UL
    40. GSM336231UL
    41. GSM336232UL
    42. GSM336233UL
    43. GSM336234UL
    44. GSM336235UL
    45. GSM336236UL
    46. GSM336237UL
    47. GSM336238UL
    48. GSM336239UL
    49. GSM336240UL
    50. GSM336241UL
    51. GSM336242UL
    52. GSM336243UL
    53. GSM336244UL
    54. GSM336245UL
    55. GSM336246UL
    56. GSM336247UL
    57. GSM336248UL
    58. GSM336249UL
    59. GSM336250UL
    60. GSM336251UL
    61. GSM336252
    62. GSM336253
    63. GSM336254
    64. GSM336255
    65. GSM336256
    66. GSM336257
    67. GSM336258
    68. GSM336259
    69. GSM336260
    70. GSM336261
    71. GSM336262
    72. GSM336263
    73. GSM336264
    74. GSM336265
    75. GSM336266
    76. GSM336267
    77. GSM336268
    78. GSM336269
    79. GSM336270
    80. GSM336271
    81. GSM336272
    82. GSM336273
    83. GSM336274
    84. GSM336275
    85. GSM336276
    86. GSM336277
    87. GSM336278
    88. GSM38689UL
    89. GSM38690UL
    90. GSM38691UL
    91. GSM38692UL
    92. GSM38693UL
    93. GSM38694UL
    94. GSM38695UL
    95. GSM52661
    96. GSM52662
    97. GSM52663
    98. GSM52664
    99. GSM52665
    100. GSM52666
    101. GSM52667
    102. GSM52668
    103. GSM52669
    104. GSM52670
    105. GSM52671
    106. GSM9093UL
    107. GSM9094UL
    108. GSM9095UL
    109. GSM9096UL
    110. GSM9097UL
    111. GSM9098
    112. GSM9099
    113. GSM9100
    114. GSM9101
    115. GSM9102
    116. GSM569424
    117. GSM569425
    118. GSM569426
    119. GSM569427
    120. GSM569428
    121. GSM569429UL
    122. GSM569430UL
    123. GSM569431UL
    124. GSM569432UL
    125. GSM569433UL
14. ub\_genes\_gviz.csv. This file is a data set with 173 genes and 127 columns of 121 samples and six meta columns of gene information from ensemble using the merge of the item 13 data set and the next item set “ensemble\_generated\_id.csv.’ The columns or columns were organized so that the UL samples are the last 70 columns, the first six columns are the meta data, and the columns after the first six columns are the non-UL samples. The header columns were also all changed to lowercase values. This file can be retrieved at https://www.dropbox.com/s/aclwb7f4julqk37/ub\_genes\_gviz.csv?dl=0. The following list is of the columns in this data set:
    1. symbol
    2. transcript
    3. chromosome
    4. start
    5. end
    6. width
    7. gsm1667144
    8. gsm1667145
    9. gsm1667146
    10. gsm336252
    11. gsm336253
    12. gsm336254
    13. gsm336255
    14. gsm336256
    15. gsm336257
    16. gsm336258
    17. gsm336259
    18. gsm336260
    19. gsm336261
    20. gsm336262
    21. gsm336263
    22. gsm336264
    23. gsm336265
    24. gsm336266
    25. gsm336267
    26. gsm336268
    27. gsm336269
    28. gsm336270
    29. gsm336271
    30. gsm336272
    31. gsm336273
    32. gsm336274
    33. gsm336275
    34. gsm336276
    35. gsm336277
    36. gsm336278
    37. gsm52661
    38. gsm52662
    39. gsm52663
    40. gsm52664
    41. gsm52665
    42. gsm52666
    43. gsm52667
    44. gsm52668
    45. gsm52669
    46. gsm52670
    47. gsm52671
    48. gsm9098
    49. gsm9099
    50. gsm9100
    51. gsm9101
    52. gsm9102
    53. gsm569424
    54. gsm569425
    55. gsm569426
    56. gsm569427
    57. gsm569428
    58. gsm1667147ul
    59. gsm1667148ul
    60. gsm1667149ul
    61. gsm336202ul
    62. gsm336203ul
    63. gsm336204ul
    64. gsm336205ul
    65. gsm336206ul
    66. gsm336207ul
    67. gsm336208ul
    68. gsm336209ul
    69. gsm336210ul
    70. gsm336211ul
    71. gsm336212ul
    72. gsm336213ul
    73. gsm336214ul
    74. gsm336215ul
    75. gsm336216ul
    76. gsm336217ul
    77. gsm336218ul
    78. gsm336219ul
    79. gsm336220ul
    80. gsm336221ul
    81. gsm336222ul
    82. gsm336223ul
    83. gsm336224ul
    84. gsm336225ul
    85. gsm336226ul
    86. gsm336227ul
    87. gsm336228ul
    88. gsm336229ul
    89. gsm336230ul
    90. gsm336231ul
    91. gsm336232ul
    92. gsm336233ul
    93. gsm336234ul
    94. gsm336235ul
    95. gsm336236ul
    96. gsm336237ul
    97. gsm336238ul
    98. gsm336239ul
    99. gsm336240ul
    100. gsm336241ul
    101. gsm336242ul
    102. gsm336243ul
    103. gsm336244ul
    104. gsm336245ul
    105. gsm336246ul
    106. gsm336247ul
    107. gsm336248ul
    108. gsm336249ul
    109. gsm336250ul
    110. gsm336251ul
    111. gsm38689ul
    112. gsm38690ul
    113. gsm38691ul
    114. gsm38692ul
    115. gsm38693ul
    116. gsm38694ul
    117. gsm38695ul
    118. gsm9093ul
    119. gsm9094ul
    120. gsm9095ul
    121. gsm9096ul
    122. gsm9097ul
    123. gsm569429ul
    124. gsm569430ul
    125. gsm569431ul
    126. gsm569432ul
    127. gsm569433ul
15. ensembl\_generated\_id.csv. This is an ensemble.org file with 2 columns and 229,428 rows retrieved from ensemble.org in the BioMart tab, using the transcript and stable ID selections of the Ensembl 96 platfrom and the Human Genes (GRCh38.12) data base. The content can be retrieved from <https://www.dropbox.com/s/t8jvbf3kipv3h83/ensembl_generated_id.csv?dl=0> . The two columns are:
16. Gene.stable.ID. This column has a prepend of “ENSG” followed by 11 numeric values, this is not needed so much as the next column to merge with the meta data from GPL6480 to get additional meta data on each gene
17. Transcript.stable.ID. This is the column that begins with “ENST” for each entry followed by 11 numeric values. It was used to merge with meta data needed for Gviz and add strand direction of each gene in the chromosome, base pair width, start in base pairs of each gene, and end of each gene in base pairs along the chromosome each gene resides.
18. ub\_genes\_ensembl.csv. This data set is 149 rows and 128 columns of samples and meta data from the merge of the ub\_genes\_gviz.csv data set and the ensembl\_generated\_id.csv data set. This file can be retrieved at <https://www.dropbox.com/s/znk2hiktv88qxm6/ub_genes_ensembl.csv?dl=0> . In this file, there are duplicate gene entries but all genes are only those genes found on the same cytoband regions of the six UL risk genes. Those cytoband regions are 11p15.5, 12q14.3, 17q25.3, and 22q13.1. The first seven columns are meta data, columns 8 through 58 are non-UL samples, and columns 59 through 128 are the UL samples. The 128 columns are:
19. transcript. The ensemble.org gene transcript ID
20. ensemble. This is the ensemble gene stable ID
21. symbol. This is the HGNC gene symbol
22. chromosome. This is the chromosome each gene belongs to
23. start. This is the start of each gene in base pairs along its cytoband
24. end. This is the end of each gene in base pairs on its cytoband
25. width. This is the width of each gene from start to end, including the start nucleic acid in base pairs and along each cytoband
26. gsm1667144. This is a UL sample, all IDs 8 through 58 are non-UL samples. The UL samples end in ‘ul’ and are columns 59 through 128
27. gsm1667145
28. gsm1667146
29. gsm336252
30. gsm336253
31. gsm336254
32. gsm336255
33. gsm336256
34. gsm336257
35. gsm336258
36. gsm336259
37. gsm336260
38. gsm336261
39. gsm336262
40. gsm336263
41. gsm336264
42. gsm336265
43. gsm336266
44. gsm336267
45. gsm336268
46. gsm336269
47. gsm336270
48. gsm336271
49. gsm336272
50. gsm336273
51. gsm336274
52. gsm336275
53. gsm336276
54. gsm336277
55. gsm336278
56. gsm52661
57. gsm52662
58. gsm52663
59. gsm52664
60. gsm52665
61. gsm52666
62. gsm52667
63. gsm52668
64. gsm52669
65. gsm52670
66. gsm52671
67. gsm9098
68. gsm9099
69. gsm9100
70. gsm9101
71. gsm9102
72. gsm569424
73. gsm569425
74. gsm569426
75. gsm569427
76. gsm569428
77. gsm1667147ul. This is the start of the UL columns and all end in ‘ul’ to identify the samples as being UL derived. Columns 59 through 128 are UL samples, and columns 8 through 58 are non-UL samples.
78. gsm1667148ul
79. gsm1667149ul
80. gsm336202ul
81. gsm336203ul
82. gsm336204ul
83. gsm336205ul
84. gsm336206ul
85. gsm336207ul
86. gsm336208ul
87. gsm336209ul
88. gsm336210ul
89. gsm336211ul
90. gsm336212ul
91. gsm336213ul
92. gsm336214ul
93. gsm336215ul
94. gsm336216ul
95. gsm336217ul
96. gsm336218ul
97. gsm336219ul
98. gsm336220ul
99. gsm336221ul
100. gsm336222ul
101. gsm336223ul
102. gsm336224ul
103. gsm336225ul
104. gsm336226ul
105. gsm336227ul
106. gsm336228ul
107. gsm336229ul
108. gsm336230ul
109. gsm336231ul
110. gsm336232ul
111. gsm336233ul
112. gsm336234ul
113. gsm336235ul
114. gsm336236ul
115. gsm336237ul
116. gsm336238ul
117. gsm336239ul
118. gsm336240ul
119. gsm336241ul
120. gsm336242ul
121. gsm336243ul
122. gsm336244ul
123. gsm336245ul
124. gsm336246ul
125. gsm336247ul
126. gsm336248ul
127. gsm336249ul
128. gsm336250ul
129. gsm336251ul
130. gsm38689ul
131. gsm38690ul
132. gsm38691ul
133. gsm38692ul
134. gsm38693ul
135. gsm38694ul
136. gsm38695ul
137. gsm9093ul
138. gsm9094ul
139. gsm9095ul
140. gsm9096ul
141. gsm9097ul
142. gsm569429ul
143. gsm569430ul
144. gsm569431ul
145. gsm569432ul
146. gsm569433ul
147. mart\_export.txt. Retrieved from the ensemble.org website in the BioMart tab using the Ensembl Genes 96 -> Human genes (GRCh38.p12) -> select Structures -> Gene Stable ID, Transcript Stable ID, Strand, Chromosome/Scaffold name, Gene Start (bp). Gene end (bp), and Gene Name, then exporting ‘Results’ as csv format. This is a 14.1 Mb size file with 229,248 rows of genes and 7 columns of the columns above. It was used to add strand direction of each gene. Some genes in the top in all the genes aren’t listed due to being renamed later in other NCBI gene name websites. This file can be found at: <https://www.dropbox.com/s/j8zc8aw0w5lnhgq/mart_export.txt?dl=0> .The column variables in this file are:
148. Gene.stable.ID. This is the ENSEMBL gene stable ID
149. Transcript.stable.ID. This is the ENSEMBL transcript ID
150. Strand. This is the column for what direction of the cytoband the gene is found as forward indicated with ‘1’ or reverse strand indicated as ‘-1’. These values were changed in the next data set to match the Gviz package factor values for strand of ‘+’ for forward and ‘-‘ for reverse strand using the gsub function.
151. [4] Gene.end..bp. This is the end of each gene on the cytoband in base pairs
152. Gene.start..bp. This is the start of each gene in base pairs on the cytoband
153. Chromosome.scaffold.name. This is the chromosome gene is located
154. Gene.name. This is the HGNC gene ID
155. ub\_genes\_ensembl\_gviz.csv. This is a data set with 149 genes as rows and 129 columns of meta data and samples by sample ID. The genes are duplicated for some and are only those on the same cytobands as the six UL risk genes. This file can be obtained from <https://www.dropbox.com/s/pdk2ttucc0zgsdl/ub_genes_ensembl_gviz.csv?dl=0> . The first 7 columnss are meta columns for each gene, the next 51are the non-UL samples, and the next 70 columns are the UL samples identified with ‘ul’ appended to the end of the Sample ID. The columns are:
156. chromosome. The chromosome the gene is found
157. start. The start of the gene in bp along the cytoband
158. end. The end of the gene in bp along the cytoband
159. width. The length of the gene in base pairs from start to end
160. strand. The strand the gene is located on the cytoband as either the forward (‘+’) or reverse (‘-‘) strand
161. gene. The ENSEMBL gene stable ID.
162. transcript. The ENSEMBL gene transcript ID
163. symbol. The HGNC gene name.
164. gsm1667144. The first of 51 non-UL samples
165. gsm1667145
166. gsm1667146
167. gsm336252
168. gsm336253
169. gsm336254
170. gsm336255
171. gsm336256
172. gsm336257
173. gsm336258
174. gsm336259
175. gsm336260
176. gsm336261
177. gsm336262
178. gsm336263
179. gsm336264
180. gsm336265
181. gsm336266
182. gsm336267
183. gsm336268
184. gsm336269
185. gsm336270
186. gsm336271
187. gsm336272
188. gsm336273
189. gsm336274
190. gsm336275
191. gsm336276
192. gsm336277
193. gsm336278
194. gsm52661
195. gsm52662
196. gsm52663
197. gsm52664
198. gsm52665
199. gsm52666
200. gsm52667
201. gsm52668
202. gsm52669
203. gsm52670
204. gsm52671
205. gsm9098
206. gsm9099
207. gsm9100
208. gsm9101
209. gsm9102
210. gsm569424
211. gsm569425
212. gsm569426
213. gsm569427
214. gsm569428
215. gsm1667147ul. The first of 70 UL samples
216. gsm1667148ul
217. gsm1667149ul
218. gsm336202ul
219. gsm336203ul
220. gsm336204ul
221. gsm336205ul
222. gsm336206ul
223. gsm336207ul
224. gsm336208ul
225. gsm336209ul
226. gsm336210ul
227. gsm336211ul
228. gsm336212ul
229. gsm336213ul
230. gsm336214ul
231. gsm336215ul
232. gsm336216ul
233. gsm336217ul
234. gsm336218ul
235. gsm336219ul
236. gsm336220ul
237. gsm336221ul
238. gsm336222ul
239. gsm336223ul
240. gsm336224ul
241. gsm336225ul
242. gsm336226ul
243. gsm336227ul
244. gsm336228ul
245. gsm336229ul
246. gsm336230ul
247. gsm336231ul
248. gsm336232ul
249. gsm336233ul
250. gsm336234ul
251. gsm336235ul
252. gsm336236ul
253. gsm336237ul
254. gsm336238ul
255. gsm336239ul
256. gsm336240ul
257. gsm336241ul
258. gsm336242ul
259. gsm336243ul
260. gsm336244ul
261. gsm336245ul
262. gsm336246ul
263. gsm336247ul
264. gsm336248ul
265. gsm336249ul
266. gsm336250ul
267. gsm336251ul
268. gsm38689ul
269. gsm38690ul
270. gsm38691ul
271. gsm38692ul
272. gsm38693ul
273. gsm38694ul
274. gsm38695ul
275. gsm9093ul
276. gsm9094ul
277. gsm9095ul
278. gsm9096ul
279. gsm9097ul
280. gsm569429ul
281. gsm569430ul
282. gsm569431ul
283. gsm569432ul
284. gsm569433ul
285. All-ggplot2-type-sample-derived.csv. This file can be retrieved from <https://www.dropbox.com/s/s2xsishg608c6g2/All-ggplot2-type-sample-derived.csv?dl=0> . This data set is used to plot with ggplot 2. The samples are the row names and the genes are the columns with two other columns for the Type of gene as a row observation and a sample column for the GEO series the sample was derived. This data set is 121 rows and 132 columns in size. There are no gene duplicates because they were removed earlier. The columns of genes as the gene symbol for each gene and two meta columns are:
286. UL\_nonUL. The value is ‘nonUL’ if not a UL sample and ‘UL’ if it is a UL sample
287. samples. This is the column specifying which GEO series the sample was from of the five GEO series used and appended with ‘\_UL’ if it is a UL. The factor values are: GSE68295, GSE13319, GSE2724, GSE593, GSE23112, GSE68295\_UL, GSE13319\_UL, GSE2724\_UL, GSE593\_UL, and GSE23112\_UL
288. AATK. The first of 130 genes labeled with the gene symbol of each gene
289. ADSL
290. APOBEC3C
291. APOBEC3F
292. APOBEC3G
293. ARHGDIA
294. ASCL2
295. ASPSCR1
296. ATF4
297. ATHL1
298. AZI1
299. BAHCC1
300. BAIAP2
301. BET1L
302. BIRC5
303. C11orf21
304. C17orf101
305. C1QTNF1
306. CANT1
307. CARD10
308. CARD14
309. CBX2
310. CBX7
311. CCDC57
312. CD7
313. CD81
314. CDC42EP1
315. CDHR5
316. CEND1
317. CHMP6
318. CSNK1D
319. CSNK1E
320. CTSD
321. CYTH1
322. CYTH4
323. DCXR
324. DDX17
325. DEAF1
326. DMC1
327. DNAH17
328. DNAL4
329. DRD4
330. DUS1L
331. EIF3L
332. EIF4A3
333. ENGASE
334. EPS8L2
335. FASN
336. FN3K
337. FN3KRP
338. FOXK2
339. FSCN2
340. GAA
341. GALR3
342. GCGR
343. GNS
344. GRAP2
345. GRIP1
346. GTPBP1
347. H1F0
348. HGS
349. HMGA2
350. HRAS
351. IFITM3
352. IGF2.AS
353. INS
354. IRAK3
355. IRF7
356. JOSD1
357. KDELR3
358. LEMD3
359. LGALS1
360. LGALS2
361. LLPH
362. MAFG
363. MFNG
364. MGAT3
365. MICALL1
366. MKL1
367. MRPL12
368. MRPL23
369. NOL12
370. NPLOC4
371. NPTX1
372. NPTXR
373. PDE6G
374. PICK1
375. PKP3
376. PLA2G6
377. PNPLA2
378. POLR2F
379. POLR2L
380. PSMD13
381. PYCR1
382. RAB40B
383. RAC2
384. RAC3
385. RASSF7
386. RFNG
387. RNH1
388. RPL3
389. RPLP2
390. SCT
391. SECTM1
392. SGSM3
393. SIGIRR
394. SIRT3
395. SIRT7
396. SLC16A8
397. SLC25A10
398. SLC25A22
399. SLC38A10
400. SMCR7L
401. SOCS3
402. SOX10
403. SYNGR1
404. TAB1
405. TALDO1
406. TBCD
407. TH
408. TMC6
409. TMEM184B
410. TMEM80
411. TNNI2
412. TNRC6B
413. TOMM22
414. TRIOBP
415. TSSC4
416. WDR45L
417. ZNF750
418. DE\_data\_unordered.csv. This is a data set with 130 rows of genes and 124 columns of all 121 samples, UL and non-UL mean values per gene, and the difference in mean values between UL and non-UL for each gene as three additional columns. There are row names that are the 130 genes. This file can be retrieved at <https://www.dropbox.com/s/q9oqlquuyu2xz8f/DE_data_unordered.csv?dl=0> . The column names are listed as they are in this data set, with the mean values as the last columns This is a list of the columns in this data set with the first 51 columns the non-UL samples and the next columns the UL samples indicated with an appended ‘ul’ to the end of the row name:
419. gsm1667144. Beginning of the non-UL samples
420. gsm1667145
421. gsm1667146
422. gsm336252
423. gsm336253
424. gsm336254
425. gsm336255
426. gsm336256
427. gsm336257
428. gsm336258
429. gsm336259
430. gsm336260
431. gsm336261
432. gsm336262
433. gsm336263
434. gsm336264
435. gsm336265
436. gsm336266
437. gsm336267
438. gsm336268
439. gsm336269
440. gsm336270
441. gsm336271
442. gsm336272
443. gsm336273
444. gsm336274
445. gsm336275
446. gsm336276
447. gsm336277
448. gsm336278
449. gsm52661
450. gsm52662
451. gsm52663
452. gsm52664
453. gsm52665
454. gsm52666
455. gsm52667
456. gsm52668
457. gsm52669
458. gsm52670
459. gsm52671
460. gsm9098
461. gsm9099
462. gsm9100
463. gsm9101
464. gsm9102
465. gsm569424
466. gsm569425
467. gsm569426
468. gsm569427
469. gsm569428. Last of the non-UL samples
470. gsm1667147ul. Beginning of the UL samples
471. gsm1667148ul
472. gsm1667149ul
473. gsm336202ul
474. gsm336203ul
475. gsm336204ul
476. gsm336205ul
477. gsm336206ul
478. gsm336207ul
479. gsm336208ul
480. gsm336209ul
481. gsm336210ul
482. gsm336211ul
483. gsm336212ul
484. gsm336213ul
485. gsm336214ul
486. gsm336215ul
487. gsm336216ul
488. gsm336217ul
489. gsm336218ul
490. gsm336219ul
491. gsm336220ul
492. gsm336221ul
493. gsm336222ul
494. gsm336223ul
495. gsm336224ul
496. gsm336225ul
497. gsm336226ul
498. gsm336227ul
499. gsm336228ul
500. gsm336229ul
501. gsm336230ul
502. gsm336231ul
503. gsm336232ul
504. gsm336233ul
505. gsm336234ul
506. gsm336235ul
507. gsm336236ul
508. gsm336237ul
509. gsm336238ul
510. gsm336239ul
511. gsm336240ul
512. gsm336241ul
513. gsm336242ul
514. gsm336243ul
515. gsm336244ul
516. gsm336245ul
517. gsm336246ul
518. gsm336247ul
519. gsm336248ul
520. gsm336249ul
521. gsm336250ul
522. gsm336251ul
523. gsm38689ul
524. gsm38690ul
525. gsm38691ul
526. gsm38692ul
527. gsm38693ul
528. gsm38694ul
529. gsm38695ul
530. gsm9093ul
531. gsm9094ul
532. gsm9095ul
533. gsm9096ul
534. gsm9097ul
535. gsm569429ul
536. gsm569430ul
537. gsm569431ul
538. gsm569432ul
539. gsm569433ul. The last UL samples listed
540. nonUL\_Mean. This is the non-UL mean for each gene
541. UL\_Mean. This is the UL mean for each gene
542. Difference\_UL\_minus\_non\_means. This is the difference in the UL mean and the non-UL mean.
543. MemberGviz\_130\_141.csv. This data set is 130 rows as unique genes and 141 columns of meta data at the beginning and all 121 samples at the end. This data set is only of the genes that are found along the same cytobands as the six UL risk genes. Like most other files the samples of UL are at the end and identified with an appended ‘ul’ to its sample ID name. This file can be retrieved at <https://www.dropbox.com/s/4uzs7zboc7y4ra2/MemberGviz_130_141.csv?dl=0> .The following list is of the 141 columns in order from left to right:
544. Genes. The gene symbol
545. Chromosome. The chromosome the gene is in of either chr11, chr12, chr17, or chr22
546. type. If the gene is up or down regulated in UL compared to non-UL
547. all. How many genes in all along that cytoband
548. up. How many of the genes in the same cytoband as this gene are up regulted in UL compared to non-UL
549. down. How many of the genes in the same cytoband as this gene are down regulated in UL compared to non-UL
550. majority. If this gene is in the majority as ‘TRUE,’ ‘Equal,’ or not as ‘FALSE’ of genes that are up or down regulated in UL in that cytoband as the majority of genes changed in UL. Some cytobands had an equal number of down and up regulated genes, so the majority was equal.
551. start. The start of each gene in base pairs on its cytoband
552. end. The end of each gene in base pairs on its cytoband
553. width. The length of each gene from start to end in base pairs
554. strand. The forward (‘+’) or reverse (‘-‘) strand of each gene’s location in the cytoband
555. gene. The ENSEMBL gene stable ID
556. transcript. The ENSEMBL transcript ID
557. GENE. The Unicode gene ID
558. GENE\_NAME. The HUGO Nomenclature full name of each gene
559. CYTOBAND. The cytoband of each gene
560. DESCRIPTION. What the gene does in the cell
561. nonUL\_Mean. The non-UL mean of each gene
562. UL\_Mean. The UL mean of each gene
563. Difference\_UL\_minus\_non\_means. The difference in UL minus non-UL in means per gene
564. gsm1667144. The start of the 51 non-UL samples
565. gsm1667145
566. gsm1667146
567. gsm336252
568. gsm336253
569. gsm336254
570. gsm336255
571. gsm336256
572. gsm336257
573. gsm336258
574. gsm336259
575. gsm336260
576. gsm336261
577. gsm336262
578. gsm336263
579. gsm336264
580. gsm336265
581. gsm336266
582. gsm336267
583. gsm336268
584. gsm336269
585. gsm336270
586. gsm336271
587. gsm336272
588. gsm336273
589. gsm336274
590. gsm336275
591. gsm336276
592. gsm336277
593. gsm336278
594. gsm52661
595. gsm52662
596. gsm52663
597. gsm52664
598. gsm52665
599. gsm52666
600. gsm52667
601. gsm52668
602. gsm52669
603. gsm52670
604. gsm52671
605. gsm9098
606. gsm9099
607. gsm9100
608. gsm9101
609. gsm9102
610. gsm569424
611. gsm569425
612. gsm569426
613. gsm569427
614. gsm569428. End of the non-UL samples
615. gsm1667147ul. Start of the UL samples
616. gsm1667148ul
617. gsm1667149ul
618. gsm336202ul
619. gsm336203ul
620. gsm336204ul
621. gsm336205ul
622. gsm336206ul
623. gsm336207ul
624. gsm336208ul
625. gsm336209ul
626. gsm336210ul
627. gsm336211ul
628. gsm336212ul
629. gsm336213ul
630. gsm336214ul
631. gsm336215ul
632. gsm336216ul
633. gsm336217ul
634. gsm336218ul
635. gsm336219ul
636. gsm336220ul
637. gsm336221ul
638. gsm336222ul
639. gsm336223ul
640. gsm336224ul
641. gsm336225ul
642. gsm336226ul
643. gsm336227ul
644. gsm336228ul
645. gsm336229ul
646. gsm336230ul
647. gsm336231ul
648. gsm336232ul
649. gsm336233ul
650. gsm336234ul
651. gsm336235ul
652. gsm336236ul
653. gsm336237ul
654. gsm336238ul
655. gsm336239ul
656. gsm336240ul
657. gsm336241ul
658. gsm336242ul
659. gsm336243ul
660. gsm336244ul
661. gsm336245ul
662. gsm336246ul
663. gsm336247ul
664. gsm336248ul
665. gsm336249ul
666. gsm336250ul
667. gsm336251ul
668. gsm38689ul
669. gsm38690ul
670. gsm38691ul
671. gsm38692ul
672. gsm38693ul
673. gsm38694ul
674. gsm38695ul
675. gsm9093ul
676. gsm9094ul
677. gsm9095ul
678. gsm9096ul
679. gsm9097ul
680. gsm569429ul
681. gsm569430ul
682. gsm569431ul
683. gsm569432ul
684. gsm569433ul. End of the UL samples
685. MemberMagnitude\_130\_142.csv. The same data set as above (item 21) but with an added magnitude column. This file can be retrieved at <https://www.dropbox.com/s/b46jl38676879oz/MemberMagnitude_130_142.csv?dl=0> . There are 130 rows of unique genes and 142 columns of meta columns and 121 samples. The samples are all at the end of the columns and meta at the beginning of the columns from left to right. The list of the columns are:
686. Genes. The gene symbol
687. Chromosome. The chromosome the gene is in of either chr11, chr12, chr17, or chr22
688. type. If the gene is up or down regulated in UL compared to non-UL
689. all. How many genes in all along that cytoband
690. up. How many of the genes in the same cytoband as this gene are up regulted in UL compared to non-UL
691. down. How many of the genes in the same cytoband as this gene are down regulated in UL compared to non-UL
692. majority. If this gene is in the majority as ‘TRUE,’ ‘Equal,’ or not as ‘FALSE’ of genes that are up or down regulated in UL in that cytoband as the majority of genes changed in UL. Some cytobands had an equal number of down and up regulated genes, so the majority was equal.
693. start. The start of each gene in base pairs on its cytoband
694. end. The end of each gene in base pairs on its cytoband
695. width. The length of each gene from start to end in base pairs
696. strand. The forward (‘+’) or reverse (‘-‘) strand of each gene’s location in the cytoband
697. gene. The ENSEMBL gene stable ID
698. transcript. The ENSEMBL transcript ID
699. GENE. The Unicode gene ID
700. GENE\_NAME. The HUGO Nomenclature full name of each gene
701. CYTOBAND. The cytoband of each gene
702. DESCRIPTION. What the gene does in the cell
703. nonUL\_Mean. The non-UL mean of each gene
704. UL\_Mean. The UL mean of each gene
705. Difference\_UL\_minus\_non\_means. The difference in UL minus non-UL in means per gene
706. Magnitude. The added column to the previous data set, MemberGviz\_130\_141.csv, that gives the magnitude of the difference in change of UL\_Mean and nonUL\_Mean columns.
707. gsm1667144. The start of the 51 non-UL samples
708. gsm1667145
709. gsm1667146
710. gsm336252
711. gsm336253
712. gsm336254
713. gsm336255
714. gsm336256
715. gsm336257
716. gsm336258
717. gsm336259
718. gsm336260
719. gsm336261
720. gsm336262
721. gsm336263
722. gsm336264
723. gsm336265
724. gsm336266
725. gsm336267
726. gsm336268
727. gsm336269
728. gsm336270
729. gsm336271
730. gsm336272
731. gsm336273
732. gsm336274
733. gsm336275
734. gsm336276
735. gsm336277
736. gsm336278
737. gsm52661
738. gsm52662
739. gsm52663
740. gsm52664
741. gsm52665
742. gsm52666
743. gsm52667
744. gsm52668
745. gsm52669
746. gsm52670
747. gsm52671
748. gsm9098
749. gsm9099
750. gsm9100
751. gsm9101
752. gsm9102
753. gsm569424
754. gsm569425
755. gsm569426
756. gsm569427
757. gsm569428. End of the non-UL samples
758. gsm1667147ul. Start of the UL samples
759. gsm1667148ul
760. gsm1667149ul
761. gsm336202ul
762. gsm336203ul
763. gsm336204ul
764. gsm336205ul
765. gsm336206ul
766. gsm336207ul
767. gsm336208ul
768. gsm336209ul
769. gsm336210ul
770. gsm336211ul
771. gsm336212ul
772. gsm336213ul
773. gsm336214ul
774. gsm336215ul
775. gsm336216ul
776. gsm336217ul
777. gsm336218ul
778. gsm336219ul
779. gsm336220ul
780. gsm336221ul
781. gsm336222ul
782. gsm336223ul
783. gsm336224ul
784. gsm336225ul
785. gsm336226ul
786. gsm336227ul
787. gsm336228ul
788. gsm336229ul
789. gsm336230ul
790. gsm336231ul
791. gsm336232ul
792. gsm336233ul
793. gsm336234ul
794. gsm336235ul
795. gsm336236ul
796. gsm336237ul
797. gsm336238ul
798. gsm336239ul
799. gsm336240ul
800. gsm336241ul
801. gsm336242ul
802. gsm336243ul
803. gsm336244ul
804. gsm336245ul
805. gsm336246ul
806. gsm336247ul
807. gsm336248ul
808. gsm336249ul
809. gsm336250ul
810. gsm336251ul
811. gsm38689ul
812. gsm38690ul
813. gsm38691ul
814. gsm38692ul
815. gsm38693ul
816. gsm38694ul
817. gsm38695ul
818. gsm9093ul
819. gsm9094ul
820. gsm9095ul
821. gsm9096ul
822. gsm9097ul
823. gsm569429ul
824. gsm569430ul
825. gsm569431ul
826. gsm569432ul
827. gsm569433ul. End of the UL samples
828. TOP16\_ml\_ready.csv. This data set is 121 rows of samples as the row names and 17 columns of the top 10 plus 6 genes in magnitude of difference in means between UL and non-UL samples for each gene and type of sample. This file can be retrieved from <https://www.dropbox.com/s/pknr9d0zumn3iit/TOP16_ml_ready.csv?dl=0> .The following list is the list of columns in this column. The first column is the type of sample the observation is as UL or non-UL and the next 16 are the top 10 plus 6 gene symbols of those genes in the subset of genes only in the same cytoband regions of the 6 UL risk genes:
829. TYPE. The type each row sample is as either ‘UL’ or ‘nonUL’
830. ASPSCR1. The first of the top 10 plus 6 genes most differential expression in UL compared to non-UL samples.
831. BET1L
832. CBX2
833. CBX7
834. CCDC57
835. CYTH4
836. FASN
837. GRIP1
838. HMGA2
839. KDELR3
840. PYCR1
841. RAC3
842. SOCS3
843. TH
844. TNRC6B
845. ZNF750
846. ubiq\_and\_top10\_samples\_only.csv. This data set is 16 rows by 122 columns big. It is a set of the top 10 plus 6 genes as their listed gene symbols under the first column, ‘gene,’ and the remaining columns of 51 non-UL samples and 70 UL samples in that order. This file can be retrieved at <https://www.dropbox.com/s/nwjfzesot66bgmx/ubiq_and_top10_samples_only.csv?dl=0>. The following list is a list of the 122 columns in this data set:
847. genes. The 16 genes of the top 10 plus 6 UL risk genes using magnitude of change.
848. gsm1667144. The start of the 51 non-UL samples
849. gsm1667145
850. gsm1667146
851. gsm336252
852. gsm336253
853. gsm336254
854. gsm336255
855. gsm336256
856. gsm336257
857. gsm336258
858. gsm336259
859. gsm336260
860. gsm336261
861. gsm336262
862. gsm336263
863. gsm336264
864. gsm336265
865. gsm336266
866. gsm336267
867. gsm336268
868. gsm336269
869. gsm336270
870. gsm336271
871. gsm336272
872. gsm336273
873. gsm336274
874. gsm336275
875. gsm336276
876. gsm336277
877. gsm336278
878. gsm52661
879. gsm52662
880. gsm52663
881. gsm52664
882. gsm52665
883. gsm52666
884. gsm52667
885. gsm52668
886. gsm52669
887. gsm52670
888. gsm52671
889. gsm9098
890. gsm9099
891. gsm9100
892. gsm9101
893. gsm9102
894. gsm569424
895. gsm569425
896. gsm569426
897. gsm569427
898. gsm569428. End of the non-UL samples
899. gsm1667147ul. Start of the UL samples
900. gsm1667148ul
901. gsm1667149ul
902. gsm336202ul
903. gsm336203ul
904. gsm336204ul
905. gsm336205ul
906. gsm336206ul
907. gsm336207ul
908. gsm336208ul
909. gsm336209ul
910. gsm336210ul
911. gsm336211ul
912. gsm336212ul
913. gsm336213ul
914. gsm336214ul
915. gsm336215ul
916. gsm336216ul
917. gsm336217ul
918. gsm336218ul
919. gsm336219ul
920. gsm336220ul
921. gsm336221ul
922. gsm336222ul
923. gsm336223ul
924. gsm336224ul
925. gsm336225ul
926. gsm336226ul
927. gsm336227ul
928. gsm336228ul
929. gsm336229ul
930. gsm336230ul
931. gsm336231ul
932. gsm336232ul
933. gsm336233ul
934. gsm336234ul
935. gsm336235ul
936. gsm336236ul
937. gsm336237ul
938. gsm336238ul
939. gsm336239ul
940. gsm336240ul
941. gsm336241ul
942. gsm336242ul
943. gsm336243ul
944. gsm336244ul
945. gsm336245ul
946. gsm336246ul
947. gsm336247ul
948. gsm336248ul
949. gsm336249ul
950. gsm336250ul
951. gsm336251ul
952. gsm38689ul
953. gsm38690ul
954. gsm38691ul
955. gsm38692ul
956. gsm38693ul
957. gsm38694ul
958. gsm38695ul
959. gsm9093ul
960. gsm9094ul
961. gsm9095ul
962. gsm9096ul
963. gsm9097ul
964. gsm569429ul
965. gsm569430ul
966. gsm569431ul
967. gsm569432ul
968. gsm569433ul. End of the UL samples
969. Stats16.csv. This file can be retrieved at <https://www.dropbox.com/s/k90cchkjkcb86x0/Stats16.csv?dl=0> .This is a data set with 32 rows of 16 UL and 16 non-UL bootstrap simulation results for each of the top 10 in magnitude of change in UL compared to non-UL samples plus 6 genes ubiquitous to UL risk. There are also six columns of those results for either the UL or non-UL gene. This is the name of the columns:
970. simulatedMean10k. This is the simulated mean of the bootstrap results for each gene
971. simulatedSD10K. This is the simulated standard deviation of each gene
972. leftTail2.5. This is the left tail of a 95 per cent confidence interval on the bootstrap simulated means of each gene in the UL and non-UL samples
973. rightTail97.25. This is the right tail of the 95 per cent confidence interval for the simulated means of each gene in UL and non-UL
974. ulStatus. This column separates the two types of gene into it’s UL result or non-UL (nonUL) result
975. Gene. This is the gene symbol for each gene
976. most\_DE\_ml\_ready\_130.csv. This file can be retrieved at <https://www.dropbox.com/s/kyrdupp2vhpbz1b/most_DE_ml_ready_130.csv?dl=0> . This data set is 121 rows with the row names of all 121 samples and 17 columns of 16 genes and one column to identify what type of samples the observational row is. It is the most differentially expressed 16 genes overall that are identified by their gene symbol in the subset of 130 genes only belonging to the same cytoband region as the six UL risk genes. This data set is ready for machine learning to be used to determine the TYPE variable based on the other 16 genes as variables. These variables listed as the columns are:
977. TYPE. This gives the type of sample each of the 121 rows are.
978. ZNF750
979. CBX2
980. SOCS3
981. RAC3
982. KDELR3
983. GRIP1
984. PYCR1
985. TH
986. CBX7
987. ASPSCR1
988. MICALL1
989. C1QTNF1
990. SLC38A10
991. CARD10
992. GRAP2
993. EIF4A3
994. least\_DE16\_ml\_ready\_130.csv. This file can be retrieved at <https://www.dropbox.com/s/mjva7aer6jxhrau/least_DE16_ml_ready_130.csv?dl=0>. This data set is 121 rows with the row names of all 121 samples and 17 columns of 16 genes least differentially expressed and one column to identify what type of samples the observational row is. It is the least 16 genes identified by their gene symbol in the subset of 130 genes only belonging to the same cytoband region as the six UL risk genes. This data set is ready for machine learning to be used to determine the TYPE variable based on the other 16 genes as variables. These variables listed as the columns are:
995. TYPE. This is the column variable that identifies each row sample as ‘UL’ or ‘nonUL.’ Machine learning results use this as the outcome to predict based on the 16 gene variables.
996. DCXR. The first least differentially expressed gene out of the least 16 overall in the subset of 130 genes all in the same cytoband regions as the six UL risk genes.
997. TRIOBP
998. DDX17
999. RPL3
1000. RASSF7
1001. CD7
1002. GAA
1003. IFITM3
1004. PICK1
1005. TBCD
1006. SIRT3
1007. SLC16A8
1008. AZI1
1009. BAIAP2
1010. PLA2G6
1011. SIRT7
1012. FOLD16\_ml\_ready.csv. This file can be retrieved at <https://www.dropbox.com/s/s1m09s5zytijgfm/FOLD16_ml_ready.csv?dl=0> .This data set is machine learning ready and has the top 10 genes with the most fold change in the subset of 130 and the six genes ubiquitous to UL risk studies. The columns are the variables of the 16 genes by gene symbol and a column variable TYPE that identifies each of the 121 rows of samples as UL or non-UL. These columns are:
1013. TYPE
1014. RAC3
1015. GRIP1
1016. TH
1017. ASCL2
1018. CBX2
1019. FSCN2
1020. NPTX1
1021. ASPSCR1
1022. PYCR1
1023. KDELR3
1024. APOBEC3F
1025. TNRC6B
1026. CYTH4
1027. CCDC57
1028. FASN
1029. HMGA2
1030. majority\_ml\_ready\_10\_total.csv. This file can be retrieved at <https://www.dropbox.com/s/ch3xry57mrnrp3l/majority_ml_ready_10_total.csv?dl=0> . This data set used the subset of 130 genes belonging to the same cytoband regions as the six UL risk genes. It has 121 rows by row name of each of the 121 samples, and it has 11 variables as columns. One is of a TYPE column identifying each sample as UL or non-UL and the other ten genes as the variables that are in the majority group. These majority of genes are those having the most magnitude of change in UL as the five most up regulated and the five most down regulated in the subset of 130. Some genes were equally divided by the number of genes that showed more up or down regulation in UL compared to non-UL and were not included. The list of those variables as columns are:
1031. TYPE. Identifies each row of samples as UL or non-UL (nonUL)
1032. EPS8L2. First of the ten majority genes
1033. TNNI2
1034. SCT
1035. INS
1036. RPLP2
1037. KDELR3
1038. GRIP1
1039. MICALL1
1040. ADSL
1041. MGAT3. Last of the ten majority genes
1042. universe\_12173.csv. This is the data set of all unique genes in common between all five GEO series. There are 12, 173 rows of unique genes and 126 columns of four meta generated columns used to subset and derive top genes from. All 121 samples of UL and non-UL in the five combined series. This file can be retrieved at <https://www.dropbox.com/s/2u569db2l7m7uhv/universe_12173.csv?dl=0> . The following list is the column variables in this data set with the samples of UL identified by an appended ‘UL’ to the end of it’s sample ID:
1043. nonUL\_Mean. This is the non-UL means of each of the unique 12, 173 genes
1044. UL\_Mean. This is the UL means of each of the 12, 173 unique genes
1045. DE. This column is the difference in up or down change in UL means of each gene compared to non-UL means of each gene.
1046. Magnitude. This is the absolute value or magnitude of change each gene had in means for UL samples minus non-UL sample means for each gene
1047. foldchange. This is the amount of fold change each gene had as a ratio of UL mean to non-UL mean per each 12,173 genes
1048. GSM1667144. This is the beginning of the 121 mixed samples by sample ID as variables. This doesn’t have ‘UL’ appended to the end so it is a non-UL sample. This applies to all of the following sample IDs.
1049. GSM1667145
1050. GSM1667146
1051. GSM1667147UL. This is the first UL sample as indicated by the appended ‘UL’ to the end of the sample ID. This applies to all the following sample IDs
1052. GSM1667148UL
1053. GSM1667149UL
1054. GSM336202UL
1055. GSM336203UL
1056. GSM336204UL
1057. GSM336205UL
1058. GSM336206UL
1059. GSM336207UL
1060. GSM336208UL
1061. GSM336209UL
1062. GSM336210UL
1063. GSM336211UL
1064. GSM336212UL
1065. GSM336213UL
1066. GSM336214UL
1067. GSM336215UL
1068. GSM336216UL
1069. GSM336217UL
1070. GSM336218UL
1071. GSM336219UL
1072. GSM336220UL
1073. GSM336221UL
1074. GSM336222UL
1075. GSM336223UL
1076. GSM336224UL
1077. GSM336225UL
1078. GSM336226UL
1079. GSM336227UL
1080. GSM336228UL
1081. GSM336229UL
1082. GSM336230UL
1083. GSM336231UL
1084. GSM336232UL
1085. GSM336233UL
1086. GSM336234UL
1087. GSM336235UL
1088. GSM336236UL
1089. GSM336237UL
1090. GSM336238UL
1091. GSM336239UL
1092. GSM336240UL
1093. GSM336241UL
1094. GSM336242UL
1095. GSM336243UL
1096. GSM336244UL
1097. GSM336245UL
1098. GSM336246UL
1099. GSM336247UL
1100. GSM336248UL
1101. GSM336249UL
1102. GSM336250UL
1103. GSM336251UL
1104. GSM336252
1105. GSM336253
1106. GSM336254
1107. GSM336255
1108. GSM336256
1109. GSM336257
1110. GSM336258
1111. GSM336259
1112. GSM336260
1113. GSM336261
1114. GSM336262
1115. GSM336263
1116. GSM336264
1117. GSM336265
1118. GSM336266
1119. GSM336267
1120. GSM336268
1121. GSM336269
1122. GSM336270
1123. GSM336271
1124. GSM336272
1125. GSM336273
1126. GSM336274
1127. GSM336275
1128. GSM336276
1129. GSM336277
1130. GSM336278
1131. GSM38689UL
1132. GSM38690UL
1133. GSM38691UL
1134. GSM38692UL
1135. GSM38693UL
1136. GSM38694UL
1137. GSM38695UL
1138. GSM52661
1139. GSM52662
1140. GSM52663
1141. GSM52664
1142. GSM52665
1143. GSM52666
1144. GSM52667
1145. GSM52668
1146. GSM52669
1147. GSM52670
1148. GSM52671
1149. GSM9093UL
1150. GSM9094UL
1151. GSM9095UL
1152. GSM9096UL
1153. GSM9097UL
1154. GSM9098
1155. GSM9099
1156. GSM9100
1157. GSM9101
1158. GSM9102
1159. GSM569424
1160. GSM569425
1161. GSM569426
1162. GSM569427
1163. GSM569428
1164. GSM569429UL
1165. GSM569430UL
1166. GSM569431UL
1167. GSM569432UL
1168. GSM569433UL
1169. most\_universe\_fold.csv. This data set is 121 rows by row name of each of the 121 samples and 17 columns of the type of sample and the 16 genes having the most fold change out of all 12,173 genes in common between the five GEO series. This data set is ready to be used in the machine learning algorithms and can be retrieved at <https://www.dropbox.com/s/np2sirc7vr8bgni/most_universe_fold.csv?dl=0> . The following is a list of the top 16 genes having the most fold change in absolute value in UL compared to non-UL samples, and a TYPE column to identify each sample as UL or non-UL:
1170. TYPE. This column identifies each sample as UL or non-UL (nonUL)
1171. FOLH1B. This is the first gene of the top 16 genes having the most fold change in absolute value in all genes
1172. STMN2
1173. TNN
1174. AKR1B10
1175. DCX
1176. CAPN6
1177. KIAA1199
1178. PLP1
1179. PRL
1180. IL17B
1181. PPFIA4
1182. GRP
1183. CARTPT
1184. GRIA2
1185. CHI3L1
1186. DLK1
1187. most\_universe\_DE.csv. This file can be retrieved at <https://www.dropbox.com/s/8sg3ysidosfhzlb/most_universe_DE.csv?dl=0> . This is a data set of the 16 most differentially expressed genes in magnitude for all 12,173 genes. It is 121 rows by row name of each of the 121 samples and 17 columns of the TYPE column of each sample as UL or non-UL and the 16 genes. The following is a list of all the variables as columns:
1188. TYPE. The type of each of the 121 samples as either UL or non-UL (nonUL)
1189. HSPB1. The first of the top 16 genes having the most magnitude of change in all 12, 173 genes
1190. DSTN S
1191. 100A6
1192. CNN1
1193. ACTG2
1194. VIM
1195. SPARCL1
1196. TPM2
1197. ACTA2
1198. PCP4
1199. TAGLN
1200. DES
1201. RAMP1
1202. CYR61
1203. UBC
1204. ACTB
1205. least\_universe\_DE.csv. This data set is 121 rows of sample IDs by 17 columns of the 16 least differentially expressed genes in magnitude of all genes and a TYPE column to identify each sample as UL or non-UL. It is the data set used for machine learning algorithms for predicting the results of those genes having the least amount of change in UL compared to non-UL samples. This file can be retrieved at https://www.dropbox.com/s/nugc7bnifmdgn1o/least\_universe\_DE.csv?dl=0 . The following list is of the 17 columns in this data set:
1206. TYPE. The type of each sample as UL or non-UL (nonUL)
1207. KLK2. The first of 16 genes in all 12, 173 genes that has the least magnitude of change in difference between UL mean per gene and non-UL mean per gene.
1208. RCVRN
1209. SYNGR3
1210. MORC1
1211. USP32P2
1212. FABP1
1213. GRIK4
1214. LIG4
1215. SUV39H1
1216. TLX3
1217. KLHDC4
1218. DNTT
1219. GRM8
1220. INSM1
1221. POU3F2
1222. SOX11
1223. Results\_predictions\_DE16\_8\_algorithms\_used.csv. This has 37 rows of predictions on the testing set of 36 samples and a row of results for each algorithm on the subset of 130 genes. It also had 9 columns of each of the algorithms used and a TYPE column that is the true type of each sample. This file can be retrieved at <https://www.dropbox.com/s/3wybopxupmscf8s/Results_predictions_DE16_8_algorithms_used.csv?dl=0> . The following list is of the 9 columns:
1224. RF. This is the caret package random forest method of machine learning algorithm results.
1225. RF2. This is the randomForest and partner e1071 package of the random forest machine learning algorithm results
1226. LDA. This is the latent Dirichlet allocation machine learning method of the caret package results.
1227. GBM . This is the global boosted regression models machine learning method results of the caret package and gbm package working together.
1228. KNN. This is the k-nearest neighbor method results of the caret package for knn machine learning.
1229. RPART. This is the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
1230. GLM . This is the generalized linear model machine learning algorithm results of the glm method of the caret package
1231. Combined. This is the combined results of all seven machine learning algorithms used that uses the gam method of the caret package and selects the best results from a data frame of all the algorithm results.
1232. Type. This is the true value type of each of the 36 samples in the testing set these algorithms used.
1233. Results\_predictions\_Least\_DE16\_8\_algorithms\_used.csv. There are 37 rows and 9 columns in this data set. The last row is the numeric results on predictions made for each of the algorithms and the true type of each sample. This data set of the results on the testing set of 36 samples for each of the 8 algorithms used in machine learning used the subset of 130 genes and the 16 least differentially expressed in magnitude genes. There is a TYPE column that shows a side by side comparison to the eight different machine learning algorithm results for the predicted type of each sample. The rows are the row names of the samples in the testing set. The same training and testing set were used for each algorithm. This file can be retrieved at <https://www.dropbox.com/s/v64glm217y6mhr5/Results_predictions_Least_DE16_8_algorithms_used.csv?dl=0> . The following is a list of these columns:
1234. RF. This is the caret package random forest method of machine learning algorithm results.
1235. RF2. This is the randomForest and partner e1071 package of the random forest machine learning algorithm results
1236. LDA. This is the latent Dirichlet allocation machine learning method of the caret package results.
1237. GBM . This is the global boosted regression models machine learning method results of the caret package and gbm package working together.
1238. KNN. This is the k-nearest neighbor method results of the caret package for knn machine learning.
1239. RPART. This is the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
1240. GLM . This is the generalized linear model machine learning algorithm results of the glm method of the caret package
1241. Combined. This is the combined results of all seven machine learning algorithms used that uses the gam method of the caret package and selects the best results from a data frame of all the algorithm results.
1242. Type. This is the true value type of each of the 36 samples in the testing set these algorithms used.
1243. Results\_predictions\_FOLD16\_8\_algorithms\_used.csv. There are 37 rows and 9 columns in this data set that used the subset of 130 genes. The last row is the numeric results on predictions made for each of the algorithms and the true type of each sample. This data set of the results on the testing set of 36 samples for each of the 8 algorithms used in machine learning used the subset of 130 genes and the 16 genes with the most magnitude of change in fold change of genes as a ratio of UL mean to non-UL mean. There is a TYPE column that shows a side by side comparison to the eight different machine learning algorithm results for the predicted type of each sample. The rows are the row names of the samples in the testing set. The same training and testing set were used for each algorithm. This file can be retrieved at <https://www.dropbox.com/s/lxou086vl2d0ra5/Results_predictions_FOLD16_8_algorithms_used.csv?dl=0> . The following list is the columns in this data set:
1244. RF. This is the caret package random forest method of machine learning algorithm results.
1245. RF2. This is the randomForest and partner e1071 package of the random forest machine learning algorithm results
1246. LDA. This is the latent Dirichlet allocation machine learning method of the caret package results.
1247. GBM . This is the global boosted regression models machine learning method results of the caret package and gbm package working together.
1248. KNN. This is the k-nearest neighbor method results of the caret package for knn machine learning.
1249. RPART. This is the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
1250. GLM . This is the generalized linear model machine learning algorithm results of the glm method of the caret package
1251. Combined. This is the combined results of all seven machine learning algorithms used that uses the gam method of the caret package and selects the best results from a data frame of all the algorithm results.
1252. Type. This is the true value type of each of the 36 samples in the testing set these algorithms used.
1253. Results\_predictions\_majority10\_8\_algorithms\_used.csv. There are 37 rows and 9 columns in this data set that used the subset of 130 genes. The last row is the numeric results on predictions made for each of the algorithms and the true type of each sample. This data set of the results on the testing set of 36 samples for each of the 8 algorithms used in machine learning used the subset of 130 genes and the 10 genes in the majority group as the highest five up regulated and highest five down regulated in magnitude of change in UL compared to non-UL samples There is a TYPE column that shows a side by side comparison to the eight different machine learning algorithm results for the predicted type of each sample. The rows are the row names of the samples in the testing set. This file can be retrieved at <https://www.dropbox.com/s/iejyel6l24ixwdu/Results_predictions_majority10_8_algorithms_used.csv?dl=0> . The following list is a list of the columns in this data set:
1254. RF. This is the caret package random forest method of machine learning algorithm results.
1255. RF2. This is the randomForest and partner e1071 package of the random forest machine learning algorithm results
1256. LDA. This is the latent Dirichlet allocation machine learning method of the caret package results.
1257. GBM . This is the global boosted regression models machine learning method results of the caret package and gbm package working together.
1258. KNN. This is the k-nearest neighbor method results of the caret package for knn machine learning.
1259. RPART. This is the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
1260. GLM . This is the generalized linear model machine learning algorithm results of the glm method of the caret package
1261. Combined. This is the combined results of all seven machine learning algorithms used that uses the gam method of the caret package and selects the best results from a data frame of all the algorithm results.
1262. Type. This is the true value type of each of the 36 samples in the testing set these algorithms used.
1263. Results\_predictions\_universe16\_fold\_8\_algorithms\_used.csv. There are 37 rows and 9 columns in this data set that used the universe of 12,173 genes. The last row is the numeric results on predictions made for each of the algorithms and the true type of each sample. This data set of the results on the testing set of 36 samples for each of the 8 algorithms used in machine learning and the 16 genes with the most magnitude of change in fold change of genes as a ratio of UL mean to non-UL mean. There is a TYPE column that shows a side by side comparison to the eight different machine learning algorithm results for the predicted type of each sample. The rows are the row names of the samples in the testing set. This file can be retrieved at <https://www.dropbox.com/s/j9fgfi92cwpbc79/Results_predictions_universe16_fold_8_algorithms_used.csv?dl=0> . The following list is the columns in this data set:
1264. RF. This is the caret package random forest method of machine learning algorithm results.
1265. RF2. This is the randomForest and partner e1071 package of the random forest machine learning algorithm results
1266. LDA. This is the latent Dirichlet allocation machine learning method of the caret package results.
1267. GBM . This is the global boosted regression models machine learning method results of the caret package and gbm package working together.
1268. KNN. This is the k-nearest neighbor method results of the caret package for knn machine learning.
1269. RPART. This is the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
1270. GLM . This is the generalized linear model machine learning algorithm results of the glm method of the caret package
1271. Combined. This is the combined results of all seven machine learning algorithms used that uses the gam method of the caret package and selects the best results from a data frame of all the algorithm results.
1272. Type. This is the true value type of each of the 36 samples in the testing set these algorithms used.
1273. Results\_predictions\_universe16\_DE\_8\_algorithms\_used.csv. There are 37 rows and 9 columns in this data set that used the universe of 12,173 genes. The last row is the numeric results on predictions made for each of the algorithms and the true type of each sample. This data set of the results on the testing set of 36 samples for each of the 8 algorithms used in machine learning and the 16 genes with the most magnitude of change in difference in expression between UL mean to non-UL mean. There is a TYPE column that shows a side by side comparison to the eight different machine learning algorithm results for the predicted type of each sample. The rows are the row names of the samples in the testing set. This file can be retrieved at <https://www.dropbox.com/s/3kjpp22j1mekzr8/Results_predictions_universe16_DE_8_algorithms_used.csv?dl=0> . The following list is the columns in this data set:
1274. RF. This is the caret package random forest method of machine learning algorithm results.
1275. RF2. This is the randomForest and partner e1071 package of the random forest machine learning algorithm results
1276. LDA. This is the latent Dirichlet allocation machine learning method of the caret package results.
1277. GBM . This is the global boosted regression models machine learning method results of the caret package and gbm package working together.
1278. KNN. This is the k-nearest neighbor method results of the caret package for knn machine learning.
1279. RPART. This is the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
1280. GLM . This is the generalized linear model machine learning algorithm results of the glm method of the caret package
1281. Combined. This is the combined results of all seven machine learning algorithms used that uses the gam method of the caret package and selects the best results from a data frame of all the algorithm results.
1282. Type. This is the true value type of each of the 36 samples in the testing set these algorithms used.
1283. Results\_predictions\_universe16\_DE\_least\_8\_algorithms\_used.csv. There are 37 rows and 9 columns in this data set that used the universe of 12,173 genes. The last row is the numeric results on predictions made for each of the algorithms and the true type of each sample. This data set of the results on the testing set of 36 samples for each of the 8 algorithms used in machine learning and the 16 genes with the least magnitude of change in difference in expression between UL mean to non-UL mean. There is a TYPE column that shows a side by side comparison to the eight different machine learning algorithm results for the predicted type of each sample. The rows are the row names of the samples in the testing set. This file can be retrieved at https://www.dropbox.com/s/zcj6al3y3y058tr/Results\_predictions\_universe16\_DE\_least\_8\_algorithms\_used.csv?dl=0. The following list is the columns in this data set:
1284. RF. This is the caret package random forest method of machine learning algorithm results.
1285. RF2. This is the randomForest and partner e1071 package of the random forest machine learning algorithm results
1286. LDA. This is the latent Dirichlet allocation machine learning method of the caret package results.
1287. GBM . This is the global boosted regression models machine learning method results of the caret package and gbm package working together.
1288. KNN. This is the k-nearest neighbor method results of the caret package for knn machine learning.
1289. RPART. This is the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
1290. GLM . This is the generalized linear model machine learning algorithm results of the glm method of the caret package
1291. Combined. This is the combined results of all seven machine learning algorithms used that uses the gam method of the caret package and selects the best results from a data frame of all the algorithm results.
1292. Type. This is the true value type of each of the 36 samples in the testing set these algorithms used.