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 were currently 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), generalized 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. Combined model results show that using the top genes and the six UL risk genes in the same cytobands as the six UL risk genes scored 83 per cent accuracy, but the top 16 genes in most fold change in all 12,173 genes scored 100 per cent in the combined model.

Keywords: uterine leiomyomas, uterine fibroids, latent dirichlet allocation, 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, geo, gene expression data

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

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
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 were 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 were 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 were 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 were 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 were *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 were 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 were excluded because they were only found to be associated with UL in patients who have endometrial cancer, and this research focus was 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 was on chromosome 11 and it was described as having significant associations with UL such as which uterine layer a UL was originating from or how many UL were 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 was on chromosome 12 along cytoband 12q14.3 and it was 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 were 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 were found on Chromosome 22 to be significant in UL were along cytoband 22q13.1. For the first gene *TNRC6B*, it was 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 was 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 were 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 was combined by genes that were in common, and then filtered for those genes that were along the same 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 were 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 were 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 were 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 was 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 were 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 was 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 was 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 was in the Appendix as item 11 listed as 'mrg5.csv' and it was 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 was 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 was 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 were 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 was 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 was 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 was 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 was 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 was 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 were 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 was 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 was listed as item 22 in the Appendix as “MemberMagnitude_130_142.csv.” The top ten genes that had the most magnitude of change was made into a subset and the six genes ubiquitous to UL risk were added. This data set was 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 were genes, the first column was the gene column, and the other 121 columns were 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 was 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 was “ubiq_and_top10_samples_only.csv” and it was 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 was 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 was 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 algorithms 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 was 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 was 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 was 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 were 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 was 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 was 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 were gene targets for evaluating if those genes were related to UL pathogenesis.

Another data set made from the same data set of item 12 in the Appendix was 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 was 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 was 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 was 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 was 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 was 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 was 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 was 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 was 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 were 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 was 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 was 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 was used in four other subsets to predict the result on a left out subset. The five results were 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 were closest to. This was repeated while recalculating the centroid of each cluster as more neighbors were 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 was 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 was 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 was a generalized linear regression model (Kuhn, et al., 2019). The gene expression data was continuous numeric data, so this seemed logical to use. The GBM algorithm was 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 was a generalized boosted regression model that bootstrap aggregates the samples similar to the AdaBoost and gradient boosting algorithms do as it was based on those algorithms.

These same algorithms were used to test variations of the data of genes that were 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 were in this section. There were many results for the many methods previously described. The first result was 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 was located, and the cytoband that each gene was located. The strand was forward if the value was '+' and reverse if the value was '-' for location in the cytoband region for each gene. Looking at the table both *CCDC57* and *FASN* were both on the reverse strand of cytoband 17.q25.3 of chromosome 17, and *CYTH4* and *TNRC6B* were both on the forward strand of cytoband 22q13.1. The gene *BET1L was* on the reverse strand of cytoband 11p15.5 of chromosome 11, and *HMGA2* was on the forward strand of cytoband 12q14.43. The other genes were 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* were on the same reverse strand in the same cytoband as *FASN* and *CCDC57.* Also, the gene, *TH,* was on the same reverse strand as *BET1L.* The gene *KDELR3* was 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 was 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 were from the subset of 130 genes found on the same cytobands of those six UL risk genes. The ‘Genes’ column was the gene symbol, the ‘Non-UL Mean’ was the mean of each gene simulated from 10,000 samplings as was the ‘UL-Mean’ column but from the UL samples. The ‘Non-UL Std Dev’ and ‘UL Std Dev’ columns were for the standard deviations of those simulated means for non-UL and UL samples respectively. The ‘Simulated Magnitude Changed’ column was 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 was 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 was 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*, was almost perfectly symmetrical. There was 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.

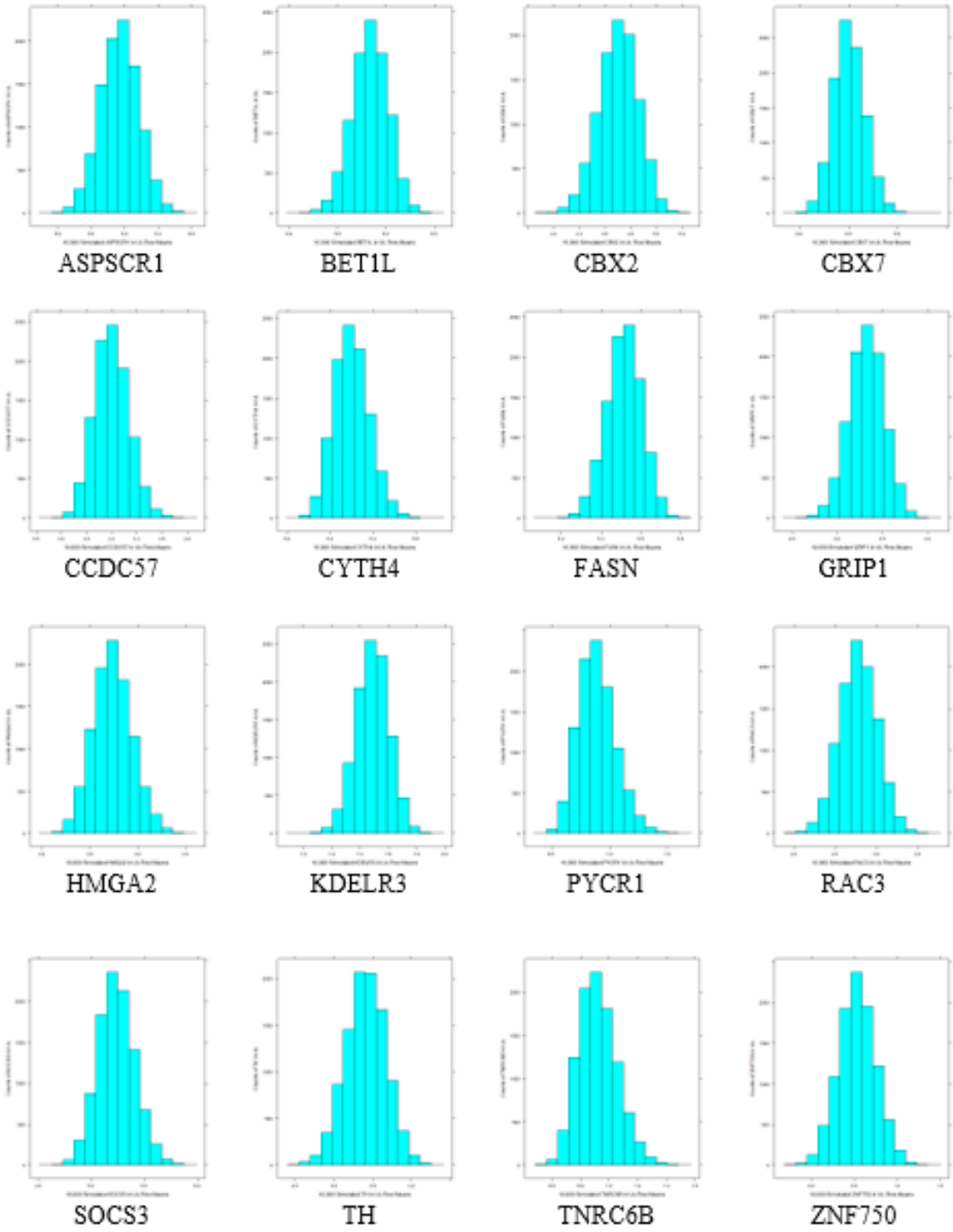

Figure 1: Histograms of UL Simulated Means for Top 10 Plus 6 Genes

The next result was 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 was only one other gene, *SIRT,* down-stream of BET1L and it was 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 was the reverse strand that *BET1L* was located. This image was in Figure 2 that follows.

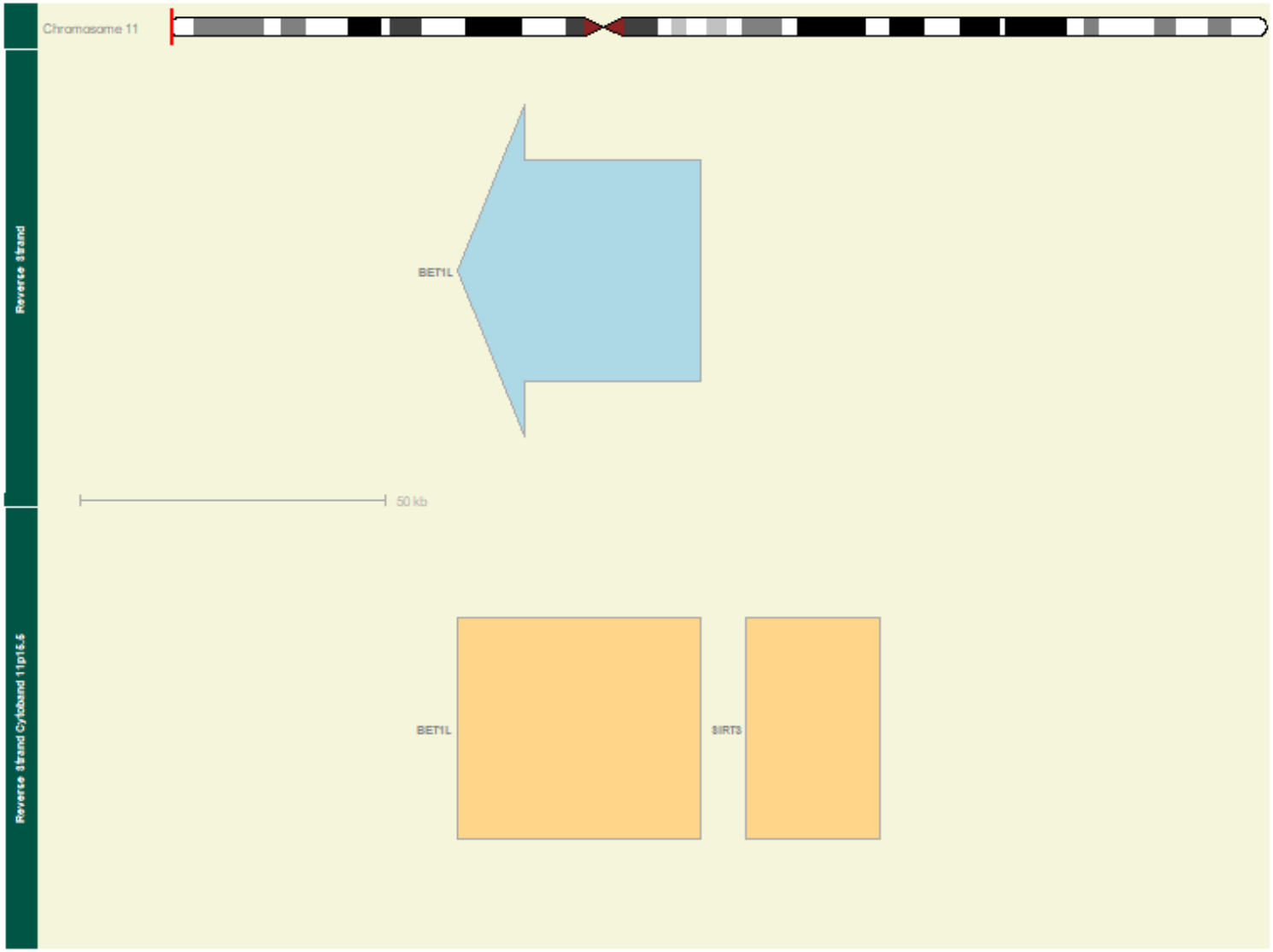

Figure 2: Reverse Strand of Cytoband 11p15.5 Genes Expressed More in UL Near *BET1L*

The next result was 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 were *LEMD3* down stream of HMGA2 and *IRAK3* up stream of *HMGA2.* The arrow in the top half of the image was pointing right to indicate the forward strand and highlights the UL risk gene, HMGA2. This image was shown in Figure 3.

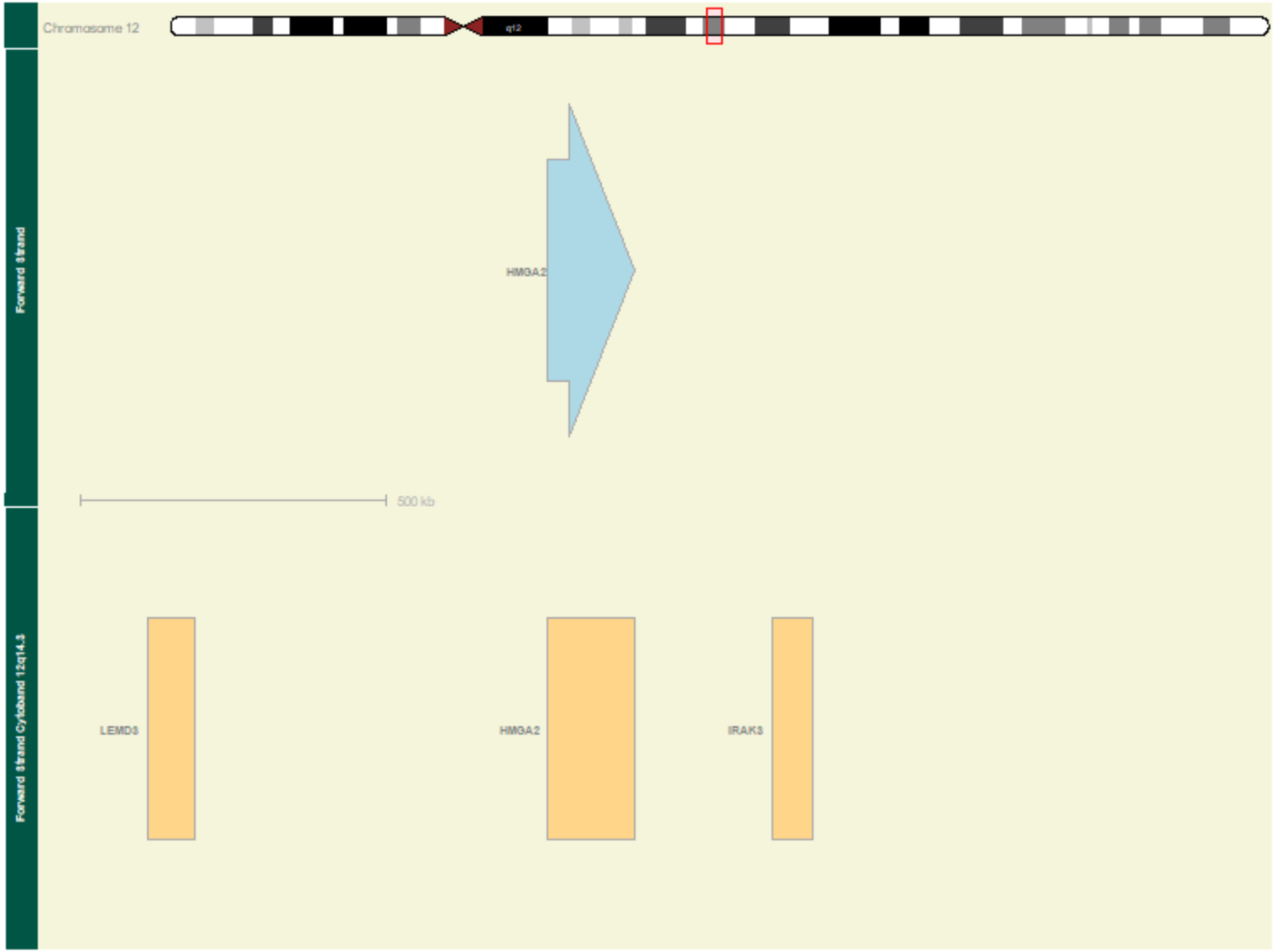

Figure 3. Forward Strand of Cytoband 12q14.3 Genes Expressed Less in UL Near *HMGA2*

The following result was for those genes along the same cytoband as *CCDC57* and *FASN* that were in the same group of genes on the reverse strand that were expressed less in UL compared to non-UL samples. The only ‘top 10 plus 6 UL risk genes’ that was on this same strand of cytoband 17q25.3 was *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 was the reverse strand as indicated by the arrow pointing left. This image was shown in Figure 4.

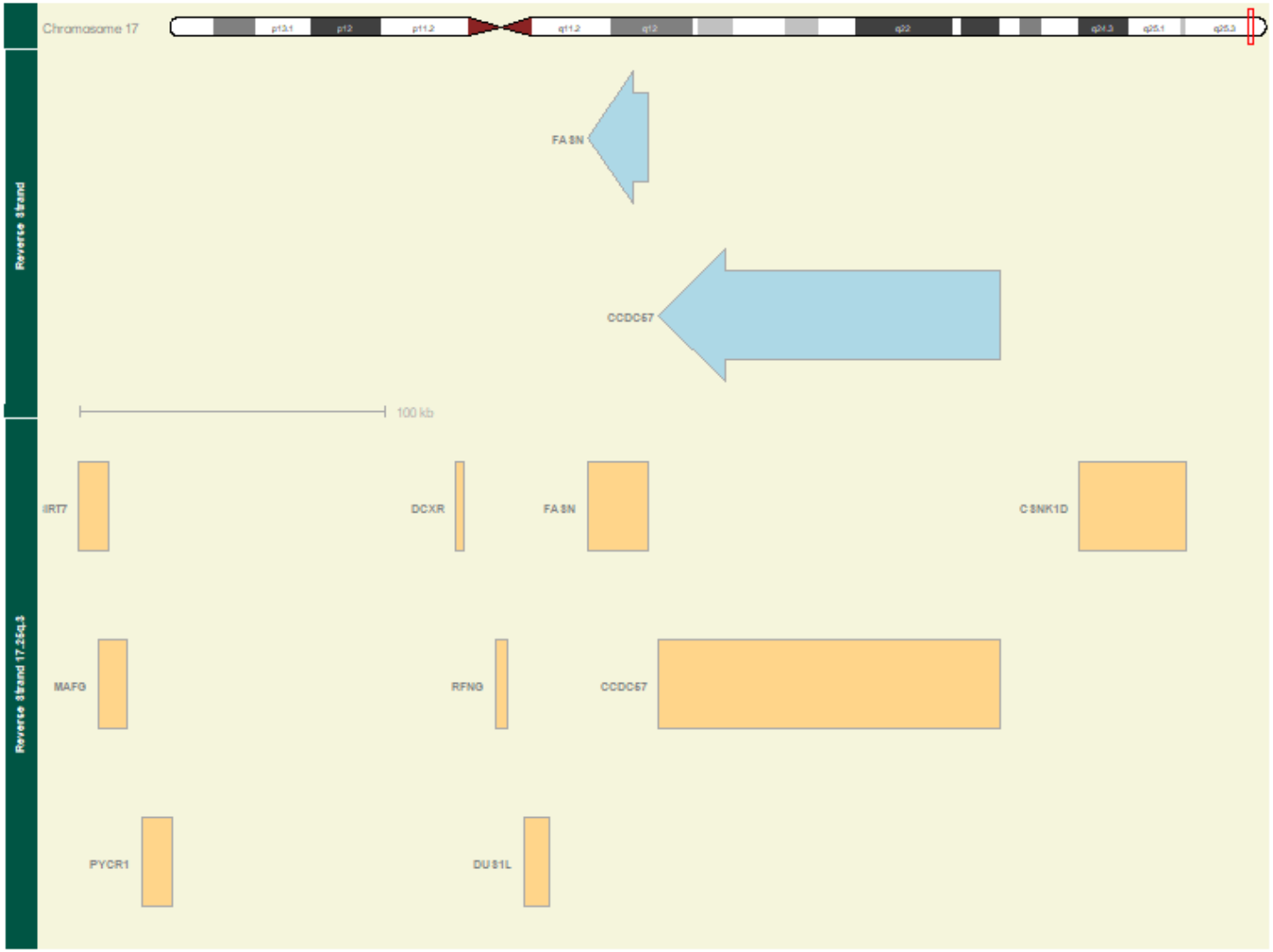

Figure 4. Gviz Map of Reverse Strand of Cytoband 17q25.3 Genes Expressed Less in UL.

The next result was 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 were 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* was almost half the distance between these two UL risk genes. This image was shown in Figure 5.

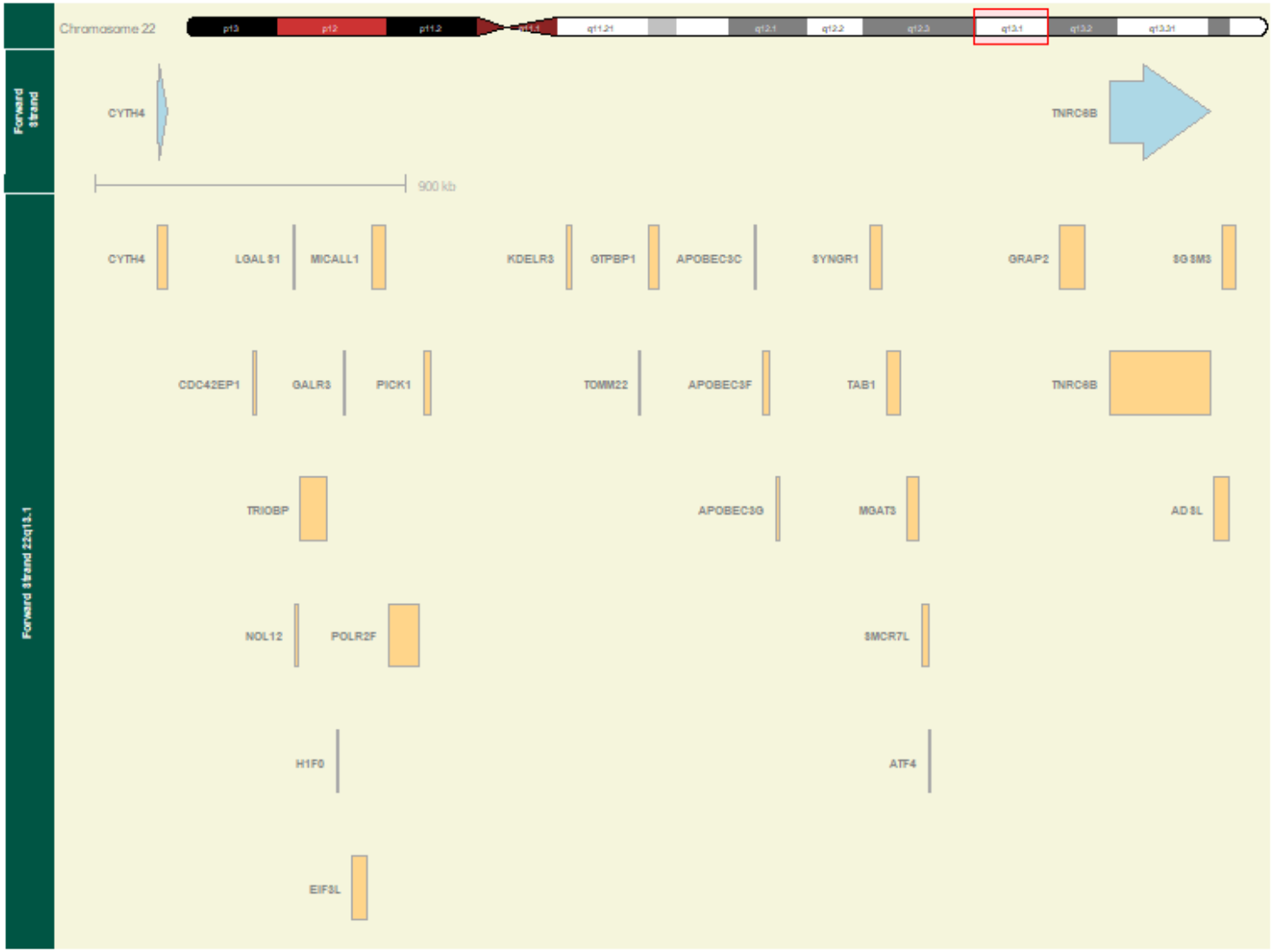

Figure 5. Forward Strand of Cytoband 22q13.1 Majority of Genes Expressed More in UL

The next result was 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 was 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 was the gene symbol of each gene. The ‘type’ column indicates if the gene was up or down regulated in UL compared to non-UL. The all field shows how many genes in all were in that cytoband location in the subset of 130 genes. The ‘up’ column indicates how many of the ‘all’ column were up regulated. The ‘down’ column indicates how many in the ‘all’ column were down regulated. The majority column indicates if that gene was in the majority of genes up or down regulated on that cytoband. The strand column indicates if that gene was on the forward or reverse strand indicated with ‘+’ or ‘-‘ respectively. The cytoband column indicates what cytoband the gene was 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 were 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 were *GRIP1* and *KDLR3*. The gene *KDLR3* was also in the same forward strand of cytoband 22q13.1 of *CYTH4* and *TNRC6B*. This table was 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 was 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 were 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 were those six UL risk genes. The scale was 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, were *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 were* the two genes that had the most simulated magnitude of change in UL compared to non-UL samples, so it makes sense that it was in this heatmap showing large changes in expression values between cools and hots on this scale. Three of the genes were already UL risk genes, *BET1L, HMGA2,* 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 was in Figure 6.

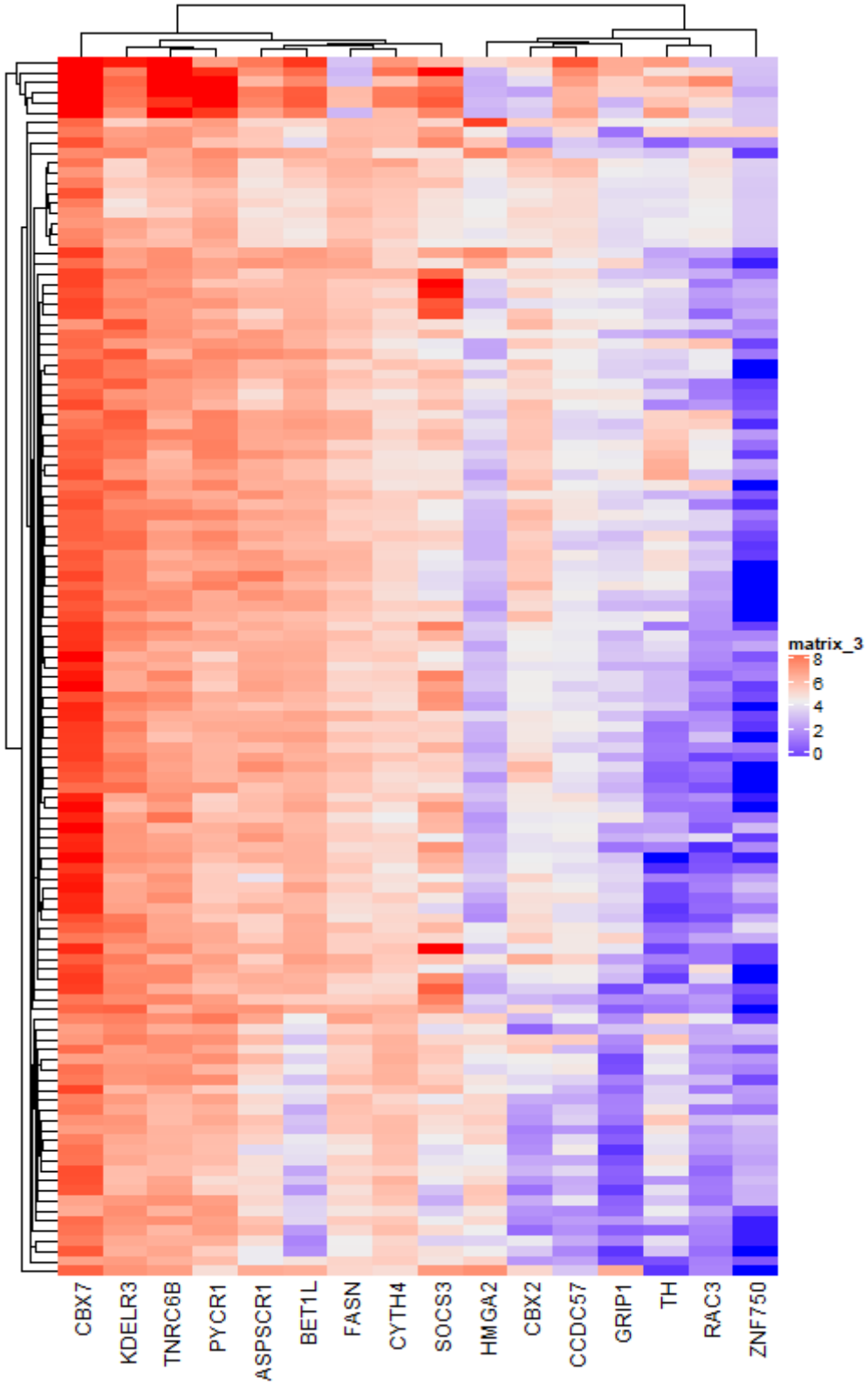

Figure 6: Heatmap of Top 10 Plus Six Genes in All Samples

The next results were 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 was 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 was in Figure 7.

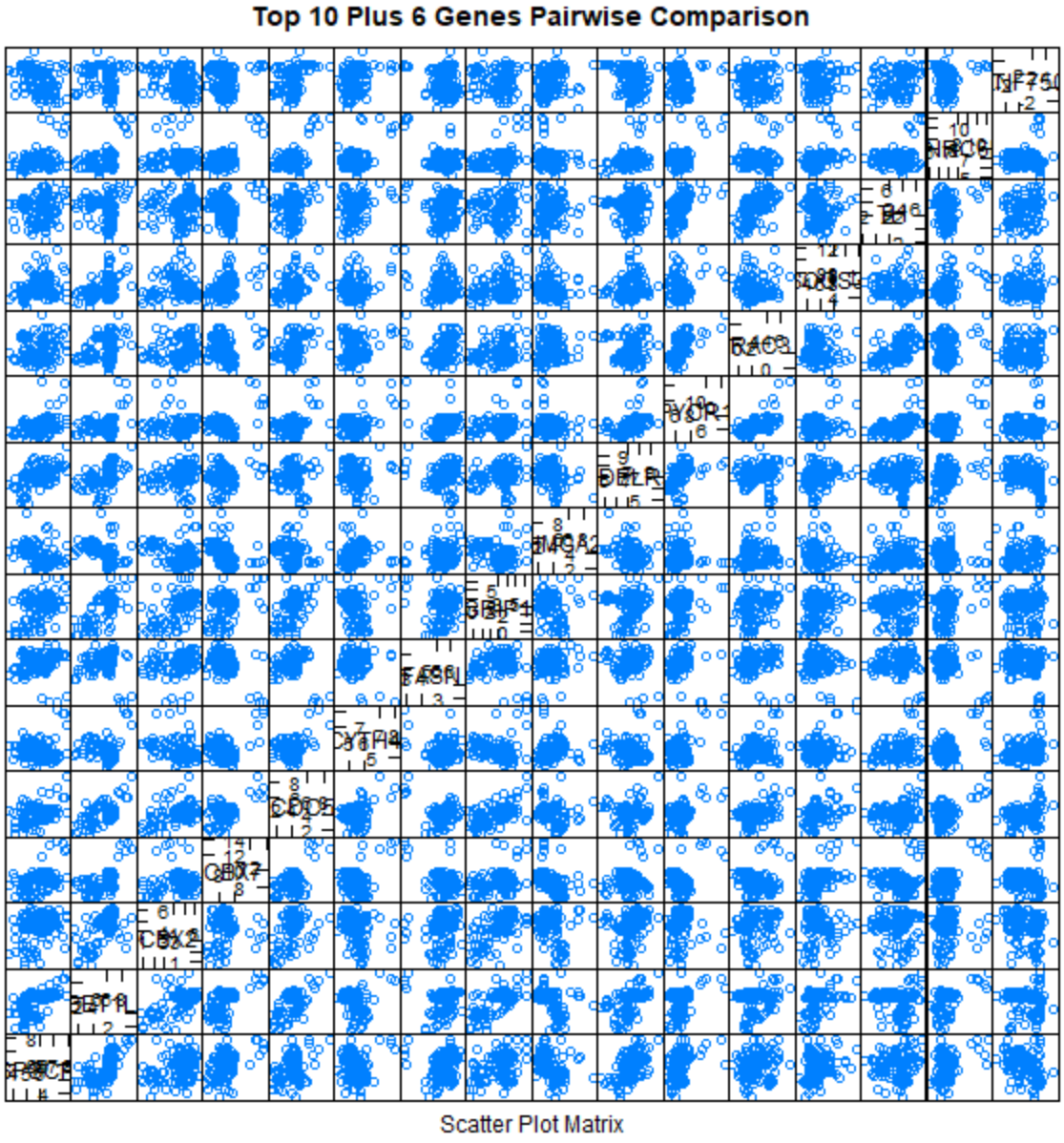

Figure 7: Pairwise Comparison of All Top 10 Plus 6 Genes**.**

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 was a factor aesthetic used to distinguish what cytoband each of these genes belongs to. Three sets of scatters were close to each other but not the same exact expression values. The bottom expression values of the first set of close genes were *GRIP1* and *TH*. The second close group was in the middle as genes *CYTH4* and *SOCS3*. The last close scatter genes were higher in expression values for genes *TNRC6B* and *PYCR1*. Any genes below the red line were under expressed in UL compared to non-UL samples, and anything above the red line were over expressed in UL compared to non-UL samples. This ggplot2 image in in Figure 8.

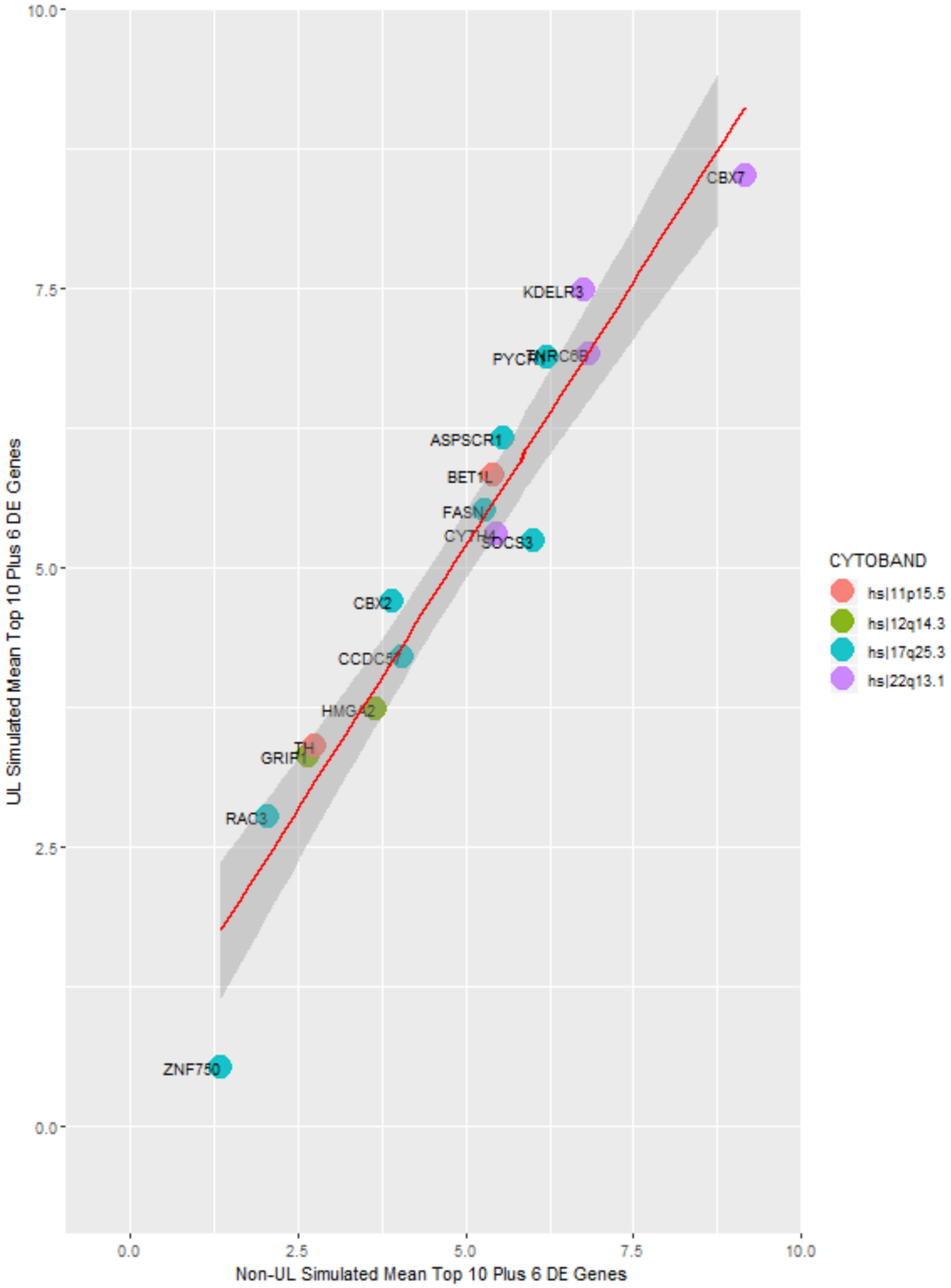

Figure 8: Comparison of Simulated Means for Non-UL and UL Top 10 Plus Six Genes

The next result shows the results of taking the top 16 in magnitude of change out of all the genes in the subset of 130 genes. The six genes ubiquitous to UL risk studies was not in this set of genes, and were replaced with *C1QTNF1, CARD10, GRAP2, MICALL1, SLC38A10,* and *EIF4A3.* Most of these top 16 genes with the most change in UL compared to non-UL samples in the same cytobands as the six UL risk genes were found on cytoband 22q13.1 and 17q25.3. The 'Gene' column was the gene symbol. The 'HGNC Gene Name' column was the Hugo Nomenclature descriptive gene name. The 'Strand' field was the strand the gene was located as indicated with a '+' for forward strand and a '-' for the reverse strand. The 'cytoband' field was one of the four cytobands these genes belong to in the subset of genes belonging to the same cytoband regions as the six UL risk genes. The details of this table were in Table 4.

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

The next results were the results of the subset of 130 genes pulled from the same cytobands of the six UL risk genes having the 16 genes with least magnitude of change in UL compared to non-UL samples. This data set has an entirely new set of genes that don't include any of the six genes ubiquitous to UL risk studies. One observation was that the gene seen earlier in the same group of genes expressed more on the 11p15.5 reverse strand was in this set of genes with the least amount of change in UL compared to non-UL exclusive only to the same cytoband regions of the six UL risk genes. That gene was *SIRT* and this data was shown in Table 5.

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

The next results show the data set created of the same subset of 130 genes on the same cytobands as the six UL risk genes but with the top 10 genes having the most magnitude of fold change as the ratio of UL means per gene to non-UL means per gene. Additionally, the six UL risk genes were included to make this set have 16 genes in total for possible gene target to UL pathogenesis. Most of the top 10 genes having most magnitude of change in this same subset were not found in this subset. Except for *GRIP1, KDELR3, CBX2, TH, PYCR1,* and *RAC3.* The gene with the most magnitude of change in the previous top 10 plus 6 UL risk genes set, *ZNF750,* was not in this subset. The six UL risk genes were added to this subset, so they were in this set, but not for having the most fold change. This could mean that the six gene above and the five new genes of *APOBEC3F, ASCL2, APSDR1, FSCN2,* and *NPTX1* were possible gene targets for UL pathogenesis. These genes were shown in Table 6, with four columns of 'Gene,' 'HGNC Gene Name,' 'Strand,' and 'cytoband.' The strand was '+' if on the forward strand and '-' if on the reverse strand. The Gene was the gene symbol. The HGNC Gene Name was the descriptive gene name, and the 'cytoband' was which cytoband of the six UL risk cytobands the gene belongs to.

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

The next result was the table of the 10 most magnitude of change in the subset that belong to the majority group of up or down regulated genes along the cytobands the UL risk genes reside. These top 10 genes were shown earlier and were the top 5 genes in up regulation as seen in UL sample means compared to non-UL sample means, and the top 5 genes in down regulation. The following table has four columns of ‘Gene’ for the gene symbol, ‘HGNC Gene Name’ for the descriptive gene name, ‘Strand’ for the forward or reverse strand the gene resides, and ‘cytoband’ for the cytoband the gene resides. The forward strand was indicated with ‘+’ and the reverse was indicated with ‘-‘ in that column. Two of the possible gene targets outside the already known UL risk gene targets were also in the previous data sets of top expressed genes in magnitude of difference and fold change in UL compared to non-UL samples. These genes were *GRIP1* and *KDELR3*. This information was shown in Table 7.

Table 7: Majority of 10 Most Differentially Expressed Genes 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

The next result was taken from the data set of the universe of genes in common in all chromosomes. This data had a total of 12,173 unique genes in common before sub-setting into those 130 unique genes only on the same cytobands as the six UL risk genes. In this data set those 16 genes that had the most magnitude of fold change as the ratio of UL means per gene to non-UL means per gene were selected. None of the six UL risk genes made this set, and neither did any of the genes from the subset of 130 genes pulled from those six UL risk genes' cytobands. Many of these genes were spread throughout the chromosomes, and none of the same cytobands of the six UL risk genes made this set. Two genes were from the female sex chromosome X, *CAPN6* and *PLP1*. Two other genes were on the same cytoband of 1q32.1 but different strands for *PPFIA4* and *CHI3LI* on the forward and reverse strands respectively. Cytobands of 11 (11q14.3, and 11p14.1)made this list, but not the same cytoband as the 130 gene subset of the six UL risk genes' cytobands on 11p15.5. These genes were shown in Table 8 with their gene symbol under the 'Gene' column, the 'HGNC Gene Name' descriptive name, the 'Strand' column for forward as '+' and reverse strand as '-', and the 'Cytoband' as the cytoband the gene was located. The first *DCX* group on genenames.org was used for the *DCX* gene in the data set as *DDC1*, because it didn't have an HGNC name. This could be one of the other *DCX* genes in that group found on different cytobands. The same with *FOHL1*. All other genes

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
DDC1	doublecortin domain containing 1 (DCX group ensemble.org)	-	hs|11p14.1
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

The next result was the table o the genes in the data set that were from the 12,173 unique genes in all. This data set was of those 16 genes having the most magnitude of differential expression between UL and non-UL means per gene. These were identical to the genes in the previous data set of the 16 genes with the most magnitude of change in all.

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
DDC1	doublecortin domain containing 1 (DCX group ensemble.org)	-	hs|11p14.1
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

The next result, was also from the universe of all 12,173 unique genes in common between these separate UL risk studies obtained from GEO. This data set was of the least expressed 16 genes in all the genes. None of the six UL risk genes or most expressed genes were in this data set. None of the same subset of 130 genes' cytobands were in this set, but an X chromosome gene was in this set and a few genes on chromosomes 11 and 17 that were the same chromosomes as the six UL risk genes. This could mean these genes could be used as tissue mediators or genes that were in this uterine tissue operating to maintain the uterus functions normally, regardless of UL or non-UL condition. This was the last data set made to test the machine learning algorithms on accuracy in predicting the testing set samples as either UL or non-UL. This table of least expressed genes in all 12,173 genes was shown in Table 10. Same columns and values as previous tables. The 'Gene' column was the gene symbol of each gene. The 'HGNC Gene Name' was the full name the gene symbol abbreviates. The 'Strand' column was the strand the gene was located as '+' if on the forward strand and '-' if on the reverse strand. The 'Cytoband' column was the cytoband the gene was located.

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

The next result was a table of how the machine learning results compare for each of the 36 samples in the identical set used as the testing set for each data set in each algorithm to record accuracy in predictions. There was a list of row names as the Sample IDs of each sample in this testing set, appended with ‘ul’ on all of the UL samples. There was also a ‘Type’ field to identify each of the samples as UL or non-UL as ‘nonUL’ and used as the outcome to regress the other genes in that data set on and create a predicted outcome by each algorithm as UL or not. The accuracy in prediction was recorded below each column of the algorithm used. The other columns were the algorithms used as the random forest caret package method as ‘RF,’ and the ‘RF2’ uses the randomForest package. The ‘LDA’ uses Latent Dirichlet allocation, and the ‘GBM’ uses Generalized Boosted Regression Models. The ‘KNN’ uses the k-nearest neighbor algorithm, and the ‘RPART’ column uses the regressive partitioning and regression trees algorithm. The ‘GLM’ column uses the generalized linear regression models algorithm. The ‘Combined’ column uses the best outcome from the previous seven algorithms in a data frame. The combined score was 86 per cent for the top 10 plus 6 UL risk genes data set from the subset of 130 genes on the same cytobands as the six UL risk genes. The best algorithm used was tied with another algorithm. Those two algorithms were the LDA and the KNN algorithms which both scored 77 per cent. You can see the results in Table 11.

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**

The last result was a result of how well all of the data sets did in each of the algorithms side by side to compare. The columns were the same columns as Table 11 for each algorithm identified in each column, the samples in the testing set were the same for each data set and algorithm used, and the Type column identifies each samples as UL or non-UL with a score of 100 per cent because those were the true values of each testing set sample.

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

The data sets that used the majority in the subset of 130 genes and the least magnitude of differential expression in the universe of all 12,173 genes and the subset of 130 genes scored the worst with a combined prediction of 75 per cent, 69 per cent, and 75 per cent respectively. The majority of genes that were the 10 most changed in up and down regulation scored better than the least magnitude of change in the universe. But it scored no better than the 16 genes with the least magnitude of change in the same subset of genes the majority was pulled from. So, the majority and least expressed genes data sets can all be excluded from containing any gene targets for UL pathogenesis.

The data set that scored the best with a combined prediction of 100 per cent was the data set with 16 genes out of the 12,173 genes that had the highest magnitude of fold change. The two algorithms that predicted the type of the sample the best was the randomForest R package and LDA algorithms with each scoring 94 per cent. The next data set from the 12,173 genes in all with the 16 most magnitude of differential expression scored 94 per cent combined. The best algorithm used on that data set was the randomForest R package algorithm. The data set with the top genes excluding the 6 UL risk genes in the subset of 130 for those 16 genes having the highest magnitude of change in differential expression in UL and non-UL means scored 92 percent. The top 10 plus 6 UL risk genes data set in the 130-subset scored 83 per cent, better than the three worst data sets in predicting UL with those genes.

These results from Table 12 indicate that the most expressed genes in magnitude of change in UL compared to non-UL samples make great predictors. But these data sets also might not hold UL risk gene targets as the algorithms that were used to predict the type of sample used unsupervised machine learning algorithms of random forest, k-nearest neighbors and rpart regression tree training.

CONCLUSIONS

The results from the data sets that included the six UL risk genes in the same cytoband regions of those six UL risk genes did score moderately well on predicting UL at 83 per cent. This could indicate there were some genes that could be gene targets for UL pathogenesis that weren't examined for significance when determining the TNRC6B and CYTH4 genes had UL risk associated with them on cytoband 22q13.1.

The best performing data sets developed to use in the machine learning algorithms to predict if the sample was a UL or not were from those most expressed in magnitude of change in fold change or differential expression between UL means per gene and non-UL means per gene.

The least performing data sets were those that were developed from the least magnitude of change and the majority data sets. The majority data set of genes was the five best of each up and down regulated genes that were expressed more or less in UL when compared to non-UL samples. And the two least magnitude of differential expression of genes were each of 16 least changed genes in the subset of 130 and the entire set of 12,173 genes. The majority did score as well as the least changed genes in all and better than the least changed in the subset of genes, but not well enough to beat the score of the top 10 plus six UL risk genes data set from the subset of 130 genes.

A limitation of this study was that only subsets of genes were chosen to look for UL risk gene targets based on gene expression data. The best genes as a subset were selected by having the most change in UL compared to non-UL samples and some from being genes in the same cytobands of those six UL risk genes. The entire set of 12,173 genes in common were not ran in

any of the machine learning algorithms because the file was too wide to run and might have stopped the program with 12,173 variables regressed on 1 added 'Type' field. When using R to calculate row means on each gene of the 1,954,853 total genes containing duplicates from the merge of all five data sets, the process took 45 minutes to shrink down to a data set of genes that still had to have the NAs removed. This would have shown if the algorithms were good on predicting any data outcomes and not necessarily finding gene targets to UL pathogenesis.

Another limitation to this study was that those previous gene targets that showed the most expression in the 130 gene subsets could possibly point to themselves or neighboring genes as being UL gene targets. As those genes were on the same cytobands as the six UL risk genes. The genes that could be tested to see whether they were gene targets by using a data set made up of only those genes having the most change in UL compared to non-UL were *KDELR3, ZNF750, TH, PYCR1, SOCS3,GRIP1,* and *CBX2.*

Moving forward with additional UL risk gene targeting using gene expression data, it would be interesting to observe smaller subsets of genes having the highest change in fold change in all 12,173 genes to find if these genes do have a connection to UL.

Transcription selects which genes to express or inhibit in the cell due to environmental factors and stress of some sort such as chemical, radiation, diet, time of day, and current health condition or stage of life. Changes in gene expression were mediated by the number of protein copies made through translation. Knowing how these genes might play a role in the cycle of UL development would be a big step in treating UL or preventing it.

Currently, it was still unknown how UL form but that they can be hereditary and linked to certain genes that have associated UL risk significantly proven in certain population studies on UL risk (Edwards, 2013; Eggert, et al., 2012; Liu et al., 2018; Hellwege et al., 2017; Rafnar et al., 2018; Cha et al., 2011; Aissani, 2015).

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**

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. This was 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 were 22,283 genes and 16 columns of additional information with some directly quoted from the table excel file. These columns were identical to the GPL570 platform because they were both the Affymetrix Human Genome U133 Array, but GPL570 was the 'Plus 2' version.

    1. ID : this was the ID column to merge with GSE593, GSE2724, and GSE23112 GEO series
    2. GB_ACC: Factor. This was the gene bank accession number for each gene
    3. SPOT_ID: Factor. This was either 'control' or 'NA'
    4. Species.Scientific.Name: Factor. This was 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.

8. 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'
9. Target.Description: Factor with 21,362 levels describing each geneRepresentative.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. This was 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 were 54,675 genes and 16 columns that were identical to GPL96, because this was the Affymetrix Human Genome U133 Plus 2.0 Array and GPL96 was the Affymetrix Human Genome U133 Array an earlier version. These were the fields:.ID : this was the ID column to merge with GSE593, GSE2724, and GSE23112 GEO series

1. GB_ACC: Factor. This was the gene bank accession number for each gene
2. SPOT_ID: Factor. This was either ‘control’ or ‘NA’
3. Species.Scientific.Name: Factor. This was equal to ‘Homo sapiens’ for all
4. Annotation.Date: Factor. The date the data platform IDs annotated, all equal ‘Oct 6, 2014’
5. Sequence.Type: Factor with three values of ‘Exemplar Sequence,’ ‘Control Sequence,’ or ‘Consensus Sequence’
6. 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’
7. Target.Description: Factor with 21,362 levels describing each gene
8. Representative.Public.ID: Factor. The accession number of a representative sequence.
9. Gene.Title: Factor. The title of the gene represented by the probe set.
10. Gene.Symbol: UniGene gene symbol
11. ENTREZ_GENE_ID: Factor. ENTREZ gene database UID
12. RefSeq.Transcript.ID: Factor. References to multiple sequences in RefSe

13. Gene.Ontology.Biological.Process: Factor. ‘Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence’
14. Gene.Ontology.Cellular.Component: Factor. ‘Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
15. 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. This was 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 were 41,108 genes and 17 identifying columns in this platform. The columns in this data set were all factors. The following were 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 were labeled as how they were 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 numbe

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. This was 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 was 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 were 22,283 genes in this raw data as the 22,283 rows. There were 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. This was 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 was the GPL96 GEO platform listed as item 7. There were 7 UL and 11 non-UL samples as headers with one probe ID column the same as GSE593 and GSE23112. There were 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 was 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 was 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. This was 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 was 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 were 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.

23. GSM336222: UL

24. GSM336223: UL

25. GSM336224: UL

26. GSM336225: UL

27. GSM336226: UL

28. GSM336227: UL

29. GSM336228: UL

30. GSM336229: UL

31. GSM336230: UL

32. GSM336231: UL

33. GSM336232: UL

34. GSM336233: UL

35. GSM336234: UL

36. GSM336235: UL

37. GSM336236: UL

38. GSM336237: UL

39. GSM336238: UL

40. GSM336239: UL

41. GSM336240: UL

42. GSM336241: UL

43. GSM336242: UL

44. GSM336243: UL

45.

46. GSM336244: UL

47. GSM336245: UL

48. GSM336246: UL

49. GSM336247: UL

50. GSM336248: UL

51. GSM336249: UL

52. GSM336250: UL

53. GSM336251: UL

54. GSM336252: non-UL

55. GSM336253: non-UL

56. GSM336254: non-UL

57. GSM336255: non-UL

58. GSM336256: non-UL

59. GSM336257: non-UL

60. GSM336258: non-UL

61. GSM336259: non-UL

62. GSM336260: non-UL

63. GSM336261: non-UL

64. GSM336262: non-UL

65. GSM336263: non-UL

66. GSM336264: non-UL

67. GSM336265: non-UL

68. GSM336266: non-UL
69. GSM336267: non-UL
70. GSM336268: non-UL
71. GSM336269: non-UL
72. GSM336270: non-UL
73. GSM336271: non-UL
74. GSM336272: non-UL
75. GSM336273: non-UL
76. GSM336274: non-UL
77. GSM336275: non-UL
78. GSM336276: non-UL
79. GSM336277: non-UL
80. GSM336278: non-UL

8.) GSE23112. Retrieved March 2019 from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE23112. This was 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 was 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 were 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 was 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 was version 3.6. The packages used were 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 was the same exact columns as GPL6480 renamed to know it was 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 were 17 columns identical to item 3 in this Appendix. This file was retrievable from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL6480 and it was 25.6 mb in size with 41,078 rows of genes by 17 columns of genes information. This file was too big to be placed in the github file folder by 1 mb as the limit was 25 mb for files stored to the folder of these files in the data repository of Github.com . The following columns were 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 was 1.1 Gb in size, too large for the github repository. It was the data of all five series merged together, including duplicate entries and missing values. This was 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 were not one of the 121 samples were the 'GENE' and 'CYTOBAND' columns from the GSE_array_meta.csv data table. The 'GENE' column was the ENTREZ gene ID and the 'CYTOBAND' column was the cytoband location of the gene in each chromosome. The columns were not listed in order of sample type like later data sets, so there was 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 were 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 was the same as the mrg5.csv file above but there were 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 was 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 were identical to the columns above in item 12, but instead 12,173 genes as rows there were 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 was 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 were the last 70 columns, the first six columns were the meta data, and the columns after the first six columns were 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 was 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 was 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:

1. Gene.stable.ID. This column has a prepend of “ENSG” followed by 11 numeric values, this was not needed so much as the next column to merge with the meta data from GPL6480 to get additional meta data on each gene

2. Transcript.stable.ID. This was 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.

16.) ub_genes_ensembl.csv. This data set was 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 were duplicate gene entries, but all genes were only those genes found on the same cytoband regions of the six UL risk genes. Those cytoband regions were 11p15.5, 12q14.3, 17q25.3, and 22q13.1. The first seven columns were meta data, columns 8 through 58 were non-UL samples, and columns 59 through 128 were the UL samples. The 128 columns are:

1. transcript. The ensemble.org gene transcript ID
2. ensemble. This was the ensemble gene stable ID
3. symbol. This was the HGNC gene symbol
4. chromosome. This was the chromosome each gene belongs to
5. start. This was the start of each gene in base pairs along its cytoband
6. end. This was the end of each gene in base pairs on its cytoband

7. width. This was the width of each gene from start to end, including the start nucleic acid in base pairs and along each cytoband
8. gsm1667144. This was a UL sample, all IDs 8 through 58 were non-UL samples. The UL samples end in ‘ul’ and were columns 59 through 128
9. gsm1667145
10. gsm1667146
11. gsm336252
12. gsm336253
13. gsm336254
14. gsm336255
15. gsm336256
16. gsm336257
17. gsm336258
18. gsm336259
19. gsm336260
20. gsm336261
21. gsm336262
22. gsm336263
23. gsm336264
24. gsm336265
25. gsm336266

26. gsm336267
27. gsm336268
28. gsm336269
29. gsm336270
30. gsm336271
31. gsm336272
32. gsm336273
33. gsm336274
34. gsm336275
35. gsm336276
36. gsm336277
37. gsm336278
38. gsm52661
39. gsm52662
40. gsm52663
41. gsm52664
42. gsm52665
43. gsm52666
44. gsm52667
45. gsm52668
46. gsm52669
47. gsm52670
48. gsm52671

49. gsm9098
50. gsm9099
51. gsm9100
52. gsm9101
53. gsm9102
54. gsm569424
55. gsm569425
56. gsm569426
57. gsm569427
58. gsm569428
59. gsm1667147ul. This was the start of the UL columns and all end in ‘ul’ to identify the samples as being UL derived. Columns 59 through 128 were UL samples, and columns 8 through 58 were non-UL samples.
60. gsm1667148ul
61. gsm1667149ul
62. gsm336202ul
63. gsm336203ul
64. gsm336204ul
65. gsm336205ul
66. gsm336206ul
67. gsm336207ul
68. gsm336208ul

69. gsm336209ul
70. gsm336210ul
71. gsm336211ul
72. gsm336212ul
73. gsm336213ul
74. gsm336214ul
75. gsm336215ul
76. gsm336216ul
77. gsm336217ul
78. gsm336218ul
79. gsm336219ul
80. gsm336220ul
81. gsm336221ul
82. gsm336222ul
83. gsm336223ul
84. gsm336224ul
85. gsm336225ul
86. gsm336226ul
87. gsm336227ul
88. gsm336228ul
89. gsm336229ul
90. gsm336230ul
91. gsm336231ul

92. gsm336232ul
93. gsm336233ul
94. gsm336234ul
95. gsm336235ul
96. gsm336236ul
97. gsm336237ul
98. gsm336238ul
99. gsm336239ul
100. gsm336240ul
101. gsm336241ul
102. gsm336242ul
103. gsm336243ul
104. gsm336244ul
105. gsm336245ul
106. gsm336246ul
107. gsm336247ul
108. gsm336248ul
109. gsm336249ul
110. gsm336250ul
111. gsm336251ul
112. gsm38689ul
113. gsm38690ul
114. gsm38691ul

115. gsm38692ul
116. gsm38693ul
117. gsm38694ul
118. gsm38695ul
119. gsm9093ul
120. gsm9094ul
121. gsm9095ul
122. gsm9096ul
123. gsm9097ul
124. gsm569429ul
125. gsm569430ul
126. gsm569431ul
127. gsm569432ul
128. gsm569433ul

17.) 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 was 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:

1. Gene.stable.ID. This was the ENSEMBL gene stable ID
2. Transcript.stable.ID. This was the ENSEMBL transcript ID
3. Strand. This was the column for what direction of the cytoband the gene was 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.
4. [4] Gene.end..bp. This was the end of each gene on the cytoband in base pairs
5. Gene.start..bp. This was the start of each gene in base pairs on the cytoband
6. Chromosome.scaffold.name. This was the chromosome gene was located
7. Gene.name. This was the HGNC gene ID

18.) ub_genes_ensembl_gviz.csv. This was a data set with 149 genes as rows and 129 columns of meta data and samples by sample ID. The genes were duplicated for some and were 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 columns were meta columns for each gene, the next 51are the non-UL samples, and the next 70 columns were the UL samples identified with ‘ul’ appended to the end of the Sample ID. The columns are:

1. chromosome. The chromosome the gene was found

2. start. The start of the gene in bp along the cytoband
3. end. The end of the gene in bp along the cytoband
4. width. The length of the gene in base pairs from start to end
5. strand. The strand the gene was located on the cytoband as either the forward (‘+’) or reverse (‘-‘) strand
6. gene. The ENSEMBL gene stable ID.
7. transcript. The ENSEMBL gene transcript ID
8. symbol. The HGNC gene name.
9. gsm1667144. The first of 51 non-UL samples
10. gsm1667145
11. gsm1667146
12. gsm336252
13. gsm336253
14. gsm336254
15. gsm336255
16. gsm336256
17. gsm336257
18. gsm336258
19. gsm336259
20. gsm336260
21. gsm336261
22. gsm336262
23. gsm336263

24. gsm336264
25. gsm336265
26. gsm336266
27. gsm336267
28. gsm336268
29. gsm336269
30. gsm336270
31. gsm336271
32. gsm336272
33. gsm336273
34. gsm336274
35. gsm336275
36. gsm336276
37. gsm336277
38. gsm336278
39. gsm52661
40. gsm52662
41. gsm52663
42. gsm52664
43. gsm52665
44. gsm52666
45. gsm52667
46. gsm52668

47. gsm52669
48. gsm52670
49. gsm52671
50. gsm9098
51. gsm9099
52. gsm9100
53. gsm9101
54. gsm9102
55. gsm569424
56. gsm569425
57. gsm569426
58. gsm569427
59. gsm569428
60. gsm1667147ul. The first of 70 UL samples
61. gsm1667148ul
62. gsm1667149ul
63. gsm336202ul
64. gsm336203ul
65. gsm336204ul
66. gsm336205ul
67. gsm336206ul
68. gsm336207ul
69. gsm336208ul

70. gsm336209ul
71. gsm336210ul
72. gsm336211ul
73. gsm336212ul
74. gsm336213ul
75. gsm336214ul
76. gsm336215ul
77. gsm336216ul
78. gsm336217ul
79. gsm336218ul
80. gsm336219ul
81. gsm336220ul
82. gsm336221ul
83. gsm336222ul
84. gsm336223ul
85. gsm336224ul
86. gsm336225ul
87. gsm336226ul
88. gsm336227ul
89. gsm336228ul
90. gsm336229ul
91. gsm336230ul
92. gsm336231ul

93. gsm336232ul
94. gsm336233ul
95. gsm336234ul
96. gsm336235ul
97. gsm336236ul
98. gsm336237ul
99. gsm336238ul
100. gsm336239ul
101. gsm336240ul
102. gsm336241ul
103. gsm336242ul
104. gsm336243ul
105. gsm336244ul
106. gsm336245ul
107. gsm336246ul
108. gsm336247ul
109. gsm336248ul
110. gsm336249ul
111. gsm336250ul
112. gsm336251ul
113. gsm38689ul
114. gsm38690ul
115. gsm38691ul

116. gsm38692ul
117. gsm38693ul
118. gsm38694ul
119. gsm38695ul
120. gsm9093ul
121. gsm9094ul
122. gsm9095ul
123. gsm9096ul
124. gsm9097ul
125. gsm569429ul
126. gsm569430ul
127. gsm569431ul
128. gsm569432ul
129. gsm569433ul

19.) 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 was used to plot with ggplot 2. The samples were the row names and the genes were 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 was 121 rows and 132 columns in size. There were no gene duplicates because they were removed earlier. The columns of genes as the gene symbol for each gene and two meta columns are:

1. UL_nonUL. The value was ‘nonUL’ if not a UL sample and ‘UL’ if it was a UL sample

2. samples. This was the column specifying which GEO series the sample was from of the five GEO series used and appended with '_UL' if it was a UL. The factor values are: GSE68295, GSE13319, GSE2724, GSE593, GSE23112, GSE68295_UL, GSE13319_UL, GSE2724_UL, GSE593_UL, and GSE23112_UL
3. AATK. The first of 130 genes labeled with the gene symbol of each gene
4. ADSL
5. APOBEC3C
6. APOBEC3F
7. APOBEC3G
8. ARHGDIA
9. ASCL2
10. ASPSCR1
11. ATF4
12. ATHL1
13. AZI1
14. BAHCC1
15. BAIAP2
16. BET1L
17. BIRC5
18. C11orf21

19. C17orf101
20. C1QTNF1
21. CANT1
22. CARD10
23. CARD14
24. CBX2
25. CBX7
26. CCDC57
27. CD7
28. CD81
29. CDC42EP1
30. CDHR5
31. CEND1
32. CHMP6
33. CSNK1D
34. CSNK1E
35. CTSD
36. CYTH1
37. CYTH4
38. DCXR
39. DDX17
40. DEAF1
41. DMC1

42. DNAH17
43. DNAL4
44. DRD4
45. DUS1L
46. EIF3L
47. EIF4A3
48. ENGASE
49. EPS8L2
50. FASN
51. FN3K
52. FN3KRP
53. FOXK2
54. FSCN2
55. GAA
56. GALR3
57. GCGR
58. GNS
59. GRAP2
60. GRIP1
61. GTPBP1
62. H1F0
63. HGS
64. HMGA2

65. HRAS
66. IFITM3
67. IGF2.AS
68. INS
69. IRAK3
70. IRF7
71. JOSD1
72. KDELR3
73. LEMD3
74. LGALS1
75. LGALS2
76. LLPH
77. MAFG
78. MFNG
79. MGAT3
80. MICALL1
81. MKL1
82. MRPL12
83. MRPL23
84. NOL12
85. NPLOC4
86. NPTX1
87. NPTXR

88. PDE6G
89. PICK1
90. PKP3
91. PLA2G6
92. PNPLA2
93. POLR2F
94. POLR2L
95. PSMD13
96. PYCR1
97. RAB40B
98. RAC2
99. RAC3
100. RASSF7
101. RFNG
102. RNH1
103. RPL3
104. RPLP2
105. SCT
106. SECTM1
107. SGSM3
108. SIGIRR
109. SIRT3
110. SIRT7

111. SLC16A8
112. SLC25A10
113. SLC25A22
114. SLC38A10
115. SMCR7L
116. SOCS3
117. SOX10
118. SYNGR1
119. TAB1
120. TALDO1
121. TBCD
122. TH
123. TMC6
124. TMEM184B
125. TMEM80
126. TNNI2
127. TNRC6B
128. TOMM22
129. TRIOBP
130. TSSC4
131. WDR45L
132. ZNF750

20.) DE_data_unordered.csv. This was 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 were row names that were the 130 genes. This file can be retrieved at https://www.dropbox.com/s/q9oqlquuyu2xz8f/DE_data_unordered.csv?dl=0 . The column names were listed as they were in this data set, with the mean values as the last columns This was 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:

1. gsm1667144. Beginning of the non-UL samples
2. gsm1667145
3. gsm1667146
4. gsm336252
5. gsm336253
6. gsm336254
7. gsm336255
8. gsm336256
9. gsm336257
10. gsm336258
11. gsm336259
12. gsm336260
13. gsm336261
14. gsm336262

15. gsm336263
16. gsm336264
17. gsm336265
18. gsm336266
19. gsm336267
20. gsm336268
21. gsm336269
22. gsm336270
23. gsm336271
24. gsm336272
25. gsm336273
26. gsm336274
27. gsm336275
28. gsm336276
29. gsm336277
30. gsm336278
31. gsm52661
32. gsm52662
33. gsm52663
34. gsm52664
35. gsm52665
36. gsm52666
37. gsm52667

38. gsm52668
39. gsm52669
40. gsm52670
41. gsm52671
42. gsm9098
43. gsm9099
44. gsm9100
45. gsm9101
46. gsm9102
47. gsm569424
48. gsm569425
49. gsm569426
50. gsm569427
51. gsm569428. Last of the non-UL samples
52. gsm1667147ul. Beginning of the UL samples
53. gsm1667148ul
54. gsm1667149ul
55. gsm336202ul
56. gsm336203ul
57. gsm336204ul
58. gsm336205ul
59. gsm336206ul
60. gsm336207ul

61. gsm336208ul
62. gsm336209ul
63. gsm336210ul
64. gsm336211ul
65. gsm336212ul
66. gsm336213ul
67. gsm336214ul
68. gsm336215ul
69. gsm336216ul
70. gsm336217ul
71. gsm336218ul
72. gsm336219ul
73. gsm336220ul
74. gsm336221ul
75. gsm336222ul
76. gsm336223ul
77. gsm336224ul
78. gsm336225ul
79. gsm336226ul
80. gsm336227ul
81. gsm336228ul
82. gsm336229ul
83. gsm336230ul

84. gsm336231ul
85. gsm336232ul
86. gsm336233ul
87. gsm336234ul
88. gsm336235ul
89. gsm336236ul
90. gsm336237ul
91. gsm336238ul
92. gsm336239ul
93. gsm336240ul
94. gsm336241ul
95. gsm336242ul
96. gsm336243ul
97. gsm336244ul
98. gsm336245ul
99. gsm336246ul
100. gsm336247ul
101. gsm336248ul
102. gsm336249ul
103. gsm336250ul
104. gsm336251ul
105. gsm38689ul
106. gsm38690ul

107. gsm38691ul
108. gsm38692ul
109. gsm38693ul
110. gsm38694ul
111. gsm38695ul
112. gsm9093ul
113. gsm9094ul
114. gsm9095ul
115. gsm9096ul
116. gsm9097ul
117. gsm569429ul
118. gsm569430ul
119. gsm569431ul
120. gsm569432ul
121. gsm569433ul. The last UL samples listed
122. nonUL_Mean. This was the non-UL mean for each gene
123. UL_Mean. This was the UL mean for each gene
124. Difference_UL_minus_non_means. This was the difference in the UL mean and the non-UL mean.

21.) MemberGviz_130_141.csv. This data set was 130 rows as unique genes and 141 columns of meta data at the beginning and all 121 samples at the end. This data set was only of the genes that were found along the same cytobands as the six UL risk genes. Like most other

files the samples of UL were 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 was of the 141 columns in order from left to right:

1. Genes. The gene symbol
2. Chromosome. The chromosome the gene was in of either chr11, chr12, chr17, or chr22
3. type. If the gene was up or down regulated in UL compared to non-UL
4. all. How many genes in all along that cytoband there are
5. up. How many of the genes in the same cytoband as this gene were up regulated in UL compared to non-UL
6. down. How many of the genes in the same cytoband as this gene were down regulated in UL compared to non-UL
7. majority. If this gene was in the majority as 'TRUE,' 'Equal,' or not as 'FALSE' of genes that were 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.
8. start. The start of each gene in base pairs on its cytoband
9. end. The end of each gene in base pairs on its cytoband
10. width. The length of each gene from start to end in base pairs
11. strand. The forward ('+') or reverse ('-') strand of each gene's location in the cytoband

12. gene. The ENSEMBL gene stable ID
13. transcript. The ENSEMBL transcript ID
14. GENE. The Unicode gene ID
15. GENE_NAME. The HUGO Nomenclature full name of each gene
16. CYTOBAND. The cytoband of each gene
17. DESCRIPTION. What the gene does in the cell
18. nonUL_Mean. The non-UL means of each gene
19. UL_Mean. The UL means of each gene
20. Difference_UL_minus_non_means. The difference in UL minus non-UL in means per gene
21. gsm1667144. The start of the 51 non-UL samples
22. gsm1667145
23. gsm1667146
24. gsm336252
25. gsm336253
26. gsm336254
27. gsm336255
28. gsm336256
29. gsm336257
30. gsm336258
31. gsm336259
32. gsm336260

33. gsm336261
34. gsm336262
35. gsm336263
36. gsm336264
37. gsm336265
38. gsm336266
39. gsm336267
40. gsm336268
41. gsm336269
42. gsm336270
43. gsm336271
44. gsm336272
45. gsm336273
46. gsm336274
47. gsm336275
48. gsm336276
49. gsm336277
50. gsm336278
51. gsm52661
52. gsm52662
53. gsm52663
54. gsm52664
55. gsm52665

56. gsm52666
57. gsm52667
58. gsm52668
59. gsm52669
60. gsm52670
61. gsm52671
62. gsm9098
63. gsm9099
64. gsm9100
65. gsm9101
66. gsm9102
67. gsm569424
68. gsm569425
69. gsm569426
70. gsm569427
71. gsm569428. End of the non-UL samples
72. gsm1667147ul. Start of the UL samples
73. gsm1667148ul
74. gsm1667149ul
75. gsm336202ul
76. gsm336203ul
77. gsm336204ul
78. gsm336205ul

79. gsm336206ul
80. gsm336207ul
81. gsm336208ul
82. gsm336209ul
83. gsm336210ul
84. gsm336211ul
85. gsm336212ul
86. gsm336213ul
87. gsm336214ul
88. gsm336215ul
89. gsm336216ul
90. gsm336217ul
91. gsm336218ul
92. gsm336219ul
93. gsm336220ul
94. gsm336221ul
95. gsm336222ul
96. gsm336223ul
97. gsm336224ul
98. gsm336225ul
99. gsm336226ul
100. gsm336227ul
101. gsm336228ul

102. gsm336229ul
103. gsm336230ul
104. gsm336231ul
105. gsm336232ul
106. gsm336233ul
107. gsm336234ul
108. gsm336235ul
109. gsm336236ul
110. gsm336237ul
111. gsm336238ul
112. gsm336239ul
113. gsm336240ul
114. gsm336241ul
115. gsm336242ul
116. gsm336243ul
117. gsm336244ul
118. gsm336245ul
119. gsm336246ul
120. gsm336247ul
121. gsm336248ul
122. gsm336249ul
123. gsm336250ul
124. gsm336251ul

125. gsm38689ul
126. gsm38690ul
127. gsm38691ul
128. gsm38692ul
129. gsm38693ul
130. gsm38694ul
131. gsm38695ul
132. gsm9093ul
133. gsm9094ul
134. gsm9095ul
135. gsm9096ul
136. gsm9097ul
137. gsm569429ul
138. gsm569430ul
139. gsm569431ul
140. gsm569432ul
141. gsm569433ul. End of the UL samples

22.) 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 were 130 rows of unique genes and 142 columns of meta columns and 121 samples. The samples were all at the end of the columns and meta at the beginning of the columns from left to right. The list of the columns is:

1. Genes. The gene symbol
2. Chromosome. The chromosome the gene was in of either chr11, chr12, chr17, or chr22
3. type. If the gene was up or down regulated in UL compared to non-UL
4. all. How many genes in all along that cytoband there are
5. up. How many of the genes in the same cytoband as this gene were up regulated in UL compared to non-UL
6. down. How many of the genes in the same cytoband as this gene were down regulated in UL compared to non-UL
7. majority. If this gene was in the majority as 'TRUE,' 'Equal,' or not as 'FALSE' of genes that were 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.
8. start. The start of each gene in base pairs on its cytoband
9. end. The end of each gene in base pairs on its cytoband
10. width. The length of each gene from start to end in base pairs
11. strand. The forward ('+') or reverse ('-') strand of each gene's location in the cytoband
12. gene. The ENSEMBL gene stable ID
13. transcript. The ENSEMBL transcript ID

14. GENE. The Unicode gene ID
15. GENE_NAME. The HUGO Nomenclature full name of each gene
16. CYTOBAND. The cytoband of each gene
17. DESCRIPTION. What the gene does in the cell
18. nonUL_Mean. The non-UL means of each gene
19. UL_Mean. The UL means of each gene
20. Difference_UL_minus_non_means. The difference in UL minus non-UL in means per gene
21. 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.
22. gsm1667144. The start of the 51 non-UL samples
23. gsm1667145
24. gsm1667146
25. gsm336252
26. gsm336253
27. gsm336254
28. gsm336255
29. gsm336256
30. gsm336257
31. gsm336258

32. gsm336259
33. gsm336260
34. gsm336261
35. gsm336262
36. gsm336263
37. gsm336264
38. gsm336265
39. gsm336266
40. gsm336267
41. gsm336268
42. gsm336269
43. gsm336270
44. gsm336271
45. gsm336272
46. gsm336273
47. gsm336274
48. gsm336275
49. gsm336276
50. gsm336277
51. gsm336278
52. gsm52661
53. gsm52662
54. gsm52663

55. gsm52664
56. gsm52665
57. gsm52666
58. gsm52667
59. gsm52668
60. gsm52669
61. gsm52670
62. gsm52671
63. gsm9098
64. gsm9099
65. gsm9100
66. gsm9101
67. gsm9102
68. gsm569424
69. gsm569425
70. gsm569426
71. gsm569427
72. gsm569428. End of the non-UL samples
73. gsm1667147ul. Start of the UL samples
74. gsm1667148ul
75. gsm1667149ul
76. gsm336202ul
77. gsm336203ul

78. gsm336204ul
79. gsm336205ul
80. gsm336206ul
81. gsm336207ul
82. gsm336208ul
83. gsm336209ul
84. gsm336210ul
85. gsm336211ul
86. gsm336212ul
87. gsm336213ul
88. gsm336214ul
89. gsm336215ul
90. gsm336216ul
91. gsm336217ul
92. gsm336218ul
93. gsm336219ul
94. gsm336220ul
95. gsm336221ul
96. gsm336222ul
97. gsm336223ul
98. gsm336224ul
99. gsm336225ul
100. gsm336226ul

101. gsm336227ul
102. gsm336228ul
103. gsm336229ul
104. gsm336230ul
105. gsm336231ul
106. gsm336232ul
107. gsm336233ul
108. gsm336234ul
109. gsm336235ul
110. gsm336236ul
111. gsm336237ul
112. gsm336238ul
113. gsm336239ul
114. gsm336240ul
115. gsm336241ul
116. gsm336242ul
117. gsm336243ul
118. gsm336244ul
119. gsm336245ul
120. gsm336246ul
121. gsm336247ul
122. gsm336248ul
123. gsm336249ul

124. gsm336250ul
125. gsm336251ul
126. gsm38689ul
127. gsm38690ul
128. gsm38691ul
129. gsm38692ul
130. gsm38693ul
131. gsm38694ul
132. gsm38695ul
133. gsm9093ul
134. gsm9094ul
135. gsm9095ul
136. gsm9096ul
137. gsm9097ul
138. gsm569429ul
139. gsm569430ul
140. gsm569431ul
141. gsm569432ul
142. gsm569433ul. End of the UL samples

23.) TOP16_ml_ready.csv. This data set was 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 was the list of columns in this column. The first column was the type of sample the observation was as UL or non-UL and the next 16 were 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:

1. TYPE. The type each row sample was as either ‘UL’ or ‘nonUL’
2. ASPSCR1. The first of the top 10 plus 6 genes most differential expression in UL compared to non-UL samples.
3. BET1L
4. CBX2
5. CBX7
6. CCDC57
7. CYTH4
8. FASN
9. GRIP1
10. HMGA2
11. KDELR3
12. PYCR1
13. RAC3
14. SOCS3
15. TH
16. TNRC6B
17. ZNF750

24.) ubiq_and_top10_samples_only.csv. This data set was 16 rows by 122 columns big. It was 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 was a list of the 122 columns in this data set:

1. genes. The 16 genes of the top 10 plus 6 UL risk genes using magnitude of change.
2. gsm1667144. The start of the 51 non-UL samples
3. gsm1667145
4. gsm1667146
5. gsm336252
6. gsm336253
7. gsm336254
8. gsm336255
9. gsm336256
10. gsm336257
11. gsm336258
12. gsm336259
13. gsm336260
14. gsm336261
15. gsm336262
16. gsm336263

17. gsm336264
18. gsm336265
19. gsm336266
20. gsm336267
21. gsm336268
22. gsm336269
23. gsm336270
24. gsm336271
25. gsm336272
26. gsm336273
27. gsm336274
28. gsm336275
29. gsm336276
30. gsm336277
31. gsm336278
32. gsm52661
33. gsm52662
34. gsm52663
35. gsm52664
36. gsm52665
37. gsm52666
38. gsm52667
39. gsm52668

40. gsm52669
41. gsm52670
42. gsm52671
43. gsm9098
44. gsm9099
45. gsm9100
46. gsm9101
47. gsm9102
48. gsm569424
49. gsm569425
50. gsm569426
51. gsm569427
52. gsm569428. End of the non-UL samples
53. gsm1667147ul. Start of the UL samples
54. gsm1667148ul
55. gsm1667149ul
56. gsm336202ul
57. gsm336203ul
58. gsm336204ul
59. gsm336205ul
60. gsm336206ul
61. gsm336207ul
62. gsm336208ul

63. gsm336209ul

64. gsm336210ul

65. gsm336211ul

66. gsm336212ul

67. gsm336213ul

68. gsm336214ul

69. gsm336215ul

70. gsm336216ul

71. gsm336217ul

72. gsm336218ul

73. gsm336219ul

74. gsm336220ul

75. gsm336221ul

76. gsm336222ul

77. gsm336223ul

78. gsm336224ul

79. gsm336225ul

80. gsm336226ul

81. gsm336227ul

82. gsm336228ul

83. gsm336229ul

84. gsm336230ul

85. gsm336231ul

86. gsm336232ul
87. gsm336233ul
88. gsm336234ul
89. gsm336235ul
90. gsm336236ul
91. gsm336237ul
92. gsm336238ul
93. gsm336239ul
94. gsm336240ul
95. gsm336241ul
96. gsm336242ul
97. gsm336243ul
98. gsm336244ul
99. gsm336245ul
100. gsm336246ul
101. gsm336247ul
102. gsm336248ul
103. gsm336249ul
104. gsm336250ul
105. gsm336251ul
106. gsm38689ul
107. gsm38690ul
108. gsm38691ul

109. gsm38692ul
110. gsm38693ul
111. gsm38694ul
112. gsm38695ul
113. gsm9093ul
114. gsm9094ul
115. gsm9095ul
116. gsm9096ul
117. gsm9097ul
118. gsm569429ul
119. gsm569430ul
120. gsm569431ul
121. gsm569432ul
122. gsm569433ul. End of the UL samples

25.) Stats16.csv. This file can be retrieved at https://www.dropbox.com/s/k90cchkjkcb86x0/Stats16.csv?dl=0 .This was 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 were also six columns of those results for either the UL or non-UL gene. This was the name of the columns:

1. simulatedMean10k. This was the simulated mean of the bootstrap results for each gene

2. simulatedSD10K. This was the simulated standard deviation of each gene
3. leftTail2.5. This was 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
4. rightTail97.25. This was the right tail of the 95 per cent confidence interval for the simulated means of each gene in UL and non-UL
5. ulStatus. This column separates the two types of gene into its UL result or non-UL (nonUL) result
6. Gene. This was the gene symbol for each gene

26.) 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 was 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 was the most differentially expressed 16 genes overall that were 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 was 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:

1. TYPE. This gives the type of sample each of the 121 rows are.
2. ZNF750
3. CBX2

4. SOCS3
5. RAC3
6. KDELR3
7. GRIP1
8. PYCR1
9. TH
10. CBX7
11. ASPSCR1
12. MICALL1
13. C1QTNF1
14. SLC38A10
15. CARD10
16. GRAP2
17. EIF4A3

27.) 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 was 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 was 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 was 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:

1. TYPE. This was 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.
2. 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.
3. TRIOBP
4. DDX17
5. RPL3
6. RASSF7
7. CD7
8. GAA
9. IFITM3
10. PICK1
11. TBCD
12. SIRT3
13. SLC16A8
14. AZI1
15. BAIAP2
16. PLA2G6
17. SIRT7

28.) 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 was 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 were 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:

1. TYPE
2. RAC3
3. GRIP1
4. TH
5. ASCL2
6. CBX2
7. FSCN2
8. NPTX1
9. ASPSCR1
10. PYCR1
11. KDELR3
12. APOBEC3F
13. TNRC6B
14. CYTH4
15. CCDC57
16. FASN
17. HMGA2

29.) 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 was of a TYPE column identifying each sample as UL or non-UL and the other ten genes as the variables that were in the majority group. These majority of genes were 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:

1. TYPE. Identifies each row of samples as UL or non-UL (nonUL)
2. EPS8L2. First of the ten majority genes
3. TNNI2
4. SCT
5. INS
6. RPLP2
7. KDELR3
8. GRIP1
9. MICALL1
10. ADSL
11. MGAT3. Last of the ten majority genes

30.) universe_12173.csv. This was the data set of all unique genes in common between all five GEO series. There were 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 was the column variables in this data set with the samples of UL identified by an appended 'UL' to the end of its sample ID:

1. nonUL_Mean. This was the non-UL means of each of the unique 12, 173 genes
2. UL_Mean. This was the UL means of each of the 12, 173 unique genes
3. DE. This column was the difference in up or down change in UL means of each gene compared to non-UL means of each gene.
4. Magnitude. This was the absolute value or magnitude of change each gene had in means for UL samples minus non-UL sample means for each gene
5. foldchange. This was the amount of fold change each gene had as a ratio of UL mean to non-UL mean per each 12,173 genes
6. GSM1667144. This was the beginning of the 121 mixed samples by sample ID as variables. This doesn't have 'UL'

7. appended to the end so it was a non-UL sample. This applies to all of the following sample IDs.
8. GSM1667145
9. GSM1667146
10. GSM1667147UL. This was 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
11. GSM1667148UL
12. GSM1667149UL
13. GSM336202UL
14. GSM336203UL
15. GSM336204UL
16. GSM336205UL
17. GSM336206UL
18. GSM336207UL
19. GSM336208UL
20. GSM336209UL
21. GSM336210UL
22. GSM336211UL
23. GSM336212UL
24. GSM336213UL
25. GSM336214UL
26. GSM336215UL

27. GSM336216UL

28. GSM336217UL

29. GSM336218UL

30. GSM336219UL

31. GSM336220UL

32. GSM336221UL

33. GSM336222UL

34. GSM336223UL

35. GSM336224UL

36. GSM336225UL

37. GSM336226UL

38. GSM336227UL

39. GSM336228UL

40. GSM336229UL

41. GSM336230UL

42. GSM336231UL

43. GSM336232UL

44. GSM336233UL

45. GSM336234UL

46. GSM336235UL

47. GSM336236UL

48. GSM336237UL

49. GSM336238UL

50. GSM336239UL

51. GSM336240UL

52. GSM336241UL

53. GSM336242UL

54. GSM336243UL

55. GSM336244UL

56. GSM336245UL

57. GSM336246UL

58. GSM336247UL

59. GSM336248UL

60. GSM336249UL

61. GSM336250UL

62. GSM336251UL

63. GSM336252

64. GSM336253

65. GSM336254

66. GSM336255

67. GSM336256

68. GSM336257

69. GSM336258

70. GSM336259

71. GSM336260

72. GSM336261

73. GSM336262

74. GSM336263

75. GSM336264

76. GSM336265

77. GSM336266

78. GSM336267

79. GSM336268

80. GSM336269

81. GSM336270

82. GSM336271

83. GSM336272

84. GSM336273

85. GSM336274

86. GSM336275

87. GSM336276

88. GSM336277

89. GSM336278

90. GSM38689UL

91. GSM38690UL

92. GSM38691UL

93. GSM38692UL

94. GSM38693UL

95. GSM38694UL

96. GSM38695UL
97. GSM52661
98. GSM52662
99. GSM52663
100. GSM52664
101. GSM52665
102. GSM52666
103. GSM52667
104. GSM52668
105. GSM52669
106. GSM52670
107. GSM52671
108. GSM9093UL
109. GSM9094UL
110. GSM9095UL
111. GSM9096UL
112. GSM9097UL
113. GSM9098
114. GSM9099
115. GSM9100
116. GSM9101
117. GSM9102
118. GSM569424

119. GSM569425
120. GSM569426
121. GSM569427
122. GSM569428
123. GSM569429UL
124. GSM569430UL
125. GSM569431UL
126. GSM569432UL
127. GSM569433UL

31.) most_universe_fold.csv. This data set was 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 was 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 was 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:

1. TYPE. This column identifies each sample as UL or non-UL (nonUL)
2. FOLH1B. This was the first gene of the top 16 genes having the most fold change in absolute value in all genes
3. STMN2
4. TNN

5. AKR1B10
6. DCX
7. CAPN6
8. KIAA1199
9. PLP1
10. PRL
11. IL17B
12. PPFIA4
13. GRP
14. CARTPT
15. GRIA2
16. CHI3L1
17. DLK1

32.) most_universe_DE.csv. This file can be retrieved at https://www.dropbox.com/s/8sg3ysidosfhzlb/most_universe_DE.csv?dl=0 . This was a data set of the 16 most differentially expressed genes in magnitude for all 12,173 genes. It was 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 was a list of all the variables as columns:

1. TYPE. The type of each of the 121 samples as either UL or non-UL (nonUL)
2. HSPB1. The first of the top 16 genes having the most magnitude of change in all 12, 173 genes

3. DSTN S
4. 100A6
5. CNN1
6. ACTG2
7. VIM
8. SPARCL1
9. TPM2
10. ACTA2
11. PCP4
12. TAGLN
13. DES
14. RAMP1
15. CYR61
16. UBC
17. ACTB

33.) least_universe_DE.csv. This data set was 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 was 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 was of the 17 columns in this data set:

1. TYPE. The type of each sample as UL or non-UL (nonUL)

2. 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.
3. RCVRN
4. SYNGR3
5. MORC1
6. USP32P2
7. FABP1
8. GRIK4
9. LIG4
10. SUV39H1
11. TLX3
12. KLHDC4
13. DNTT
14. GRM8
15. INSM1
16. POU3F2
17. SOX11

34.) 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 was 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 was of the 9 columns:

1. RF. This was the caret package random forest method of machine learning algorithm results.
2. RF2. This was the randomForest and partner e1071 package of the random forest machine learning algorithm results
3. LDA. This was the latent Dirichlet allocation machine learning method of the caret package results.
4. GBM . This was the global boosted regression models machine learning method results of the caret package and gbm package working together.
5. KNN. This was the k-nearest neighbor method results of the caret package for knn machine learning.
6. RPART. This was the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
7. GLM . This was the generalized linear model machine learning algorithm results of the glm method of the caret package
8. Combined. This was 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.

9. Type. This was the true value type of each of the 36 samples in the testing set these algorithms used.

35.) Results_predictions_Least_DE16_8_algorithms_used.csv. There were 37 rows and 9 columns in this data set. The last row was 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 was 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 were 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 was a list of these columns:

1. RF. This was the caret package random forest method of machine learning algorithm results.
2. RF2. This was the randomForest and partner e1071 package of the random forest machine learning algorithm results
3. LDA. This was the latent Dirichlet allocation machine learning method of the caret package results.
4. GBM . This was the global boosted regression models machine learning method results of the caret package and gbm package working together.

5. KNN. This was the k-nearest neighbor method results of the caret package for knn machine learning.
6. RPART. This was the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
7. GLM . This was the generalized linear model machine learning algorithm results of the glm method of the caret package
8. Combined. This was 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.
9. Type. This was the true value type of each of the 36 samples in the testing set these algorithms used.

36.) Results_predictions_FOLD16_8_algorithms_used.csv. There were 37 rows and 9 columns in this data set that used the subset of 130 genes. The last row was 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 was 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 were

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 was the columns in this data set:

1. RF. This was the caret package random forest method of machine learning algorithm results.
2. RF2. This was the randomForest and partner e1071 package of the random forest machine learning algorithm results
3. LDA. This was the latent Dirichlet allocation machine learning method of the caret package results.
4. GBM . This was the global boosted regression models machine learning method results of the caret package and gbm package working together.
5. KNN. This was the k-nearest neighbor method results of the caret package for knn machine learning.
6. RPART. This was the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
7. GLM . This was the generalized linear model machine learning algorithm results of the glm method of the caret package
8. Combined. This was the combined results of all seven machine learning algorithms used that uses the gam method

9. of the caret package and selects the best results from a data frame of all the algorithm results.
10. Type. This was the true value type of each of the 36 samples in the testing set these algorithms used.

37.) Results_predictions_majority10_8_algorithms_used.csv. There were 37 rows and 9 columns in this data set that used the subset of 130 genes. The last row was 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 was 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 were 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 was a list of the columns in this data set:

1. RF. This was the caret package random forest method of machine learning algorithm results.
2. RF2. This was the randomForest and partner e1071 package of the random forest machine learning algorithm results
3. LDA. This was the latent Dirichlet allocation machine learning method of the caret package results.

4. GBM . This was the global boosted regression models machine learning method results of the caret package and gbm package working together.
5. KNN. This was the k-nearest neighbor method results of the caret package for knn machine learning.
6. RPART. This was the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
7. GLM . This was the generalized linear model machine learning algorithm results of the glm method of the caret package
8. Combined. This was 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.
9. Type. This was the true value type of each of the 36 samples in the testing set these algorithms used.

38.) Results_predictions_universe16_fold_8_algorithms_used.csv. There were 37 rows and 9 columns in this data set that used the universe of 12,173 genes. The last row was 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 was 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 were 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 was the columns in this data set:

1. RF. This was the caret package random forest method of machine learning algorithm results.
2. RF2. This was the randomForest and partner e1071 package of the random forest machine learning algorithm results
3. LDA. This was the latent Dirichlet allocation machine learning method of the caret package results.
4. GBM . This was the global boosted regression models machine learning method results of the caret package and gbm package working together.
5. KNN. This was the k-nearest neighbor method results of the caret package for knn machine learning.
6. RPART. This was the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
7. GLM . This was the generalized linear model machine learning algorithm results of the glm method of the caret package

8. Combined. This was 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.
9. Type. This was the true value type of each of the 36 samples in the testing set these algorithms used.

39.) Results_predictions_universe16_DE_8_algorithms_used.csv. There were 37 rows and 9 columns in this data set that used the universe of 12,173 genes. The last row was 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 was 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 were 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 was the columns in this data set:

1. RF. This was the caret package random forest method of machine learning algorithm results.
2. RF2. This was the randomForest and partner e1071 package of the random forest machine learning algorithm results

3. LDA. This was the latent Dirichlet allocation machine learning method of the caret package results.

4. GBM . This was the global boosted regression models machine learning method results of the caret package and gbm package working together.
5. KNN. This was the k-nearest neighbor method results of the caret package for knn machine learning.
6. RPART. This was the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
7. GLM . This was the generalized linear model machine learning algorithm results of the glm method of the caret package
8. Combined. This was 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.
9. Type. This was the true value type of each of the 36 samples in the testing set these algorithms used.

40.) Results_predictions_universe16_DE_least_8_algorithms_used.csv. There were 37 rows and 9 columns in this data set that used the universe of 12,173 genes. The last row was 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 was 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 were 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 was the columns in this data set:

1. RF. This was the caret package random forest method of machine learning algorithm results.
2. RF2. This was the randomForest and partner e1071 package of the random forest machine learning algorithm results
3. LDA. This was the latent Dirichlet allocation machine learning method of the caret package results.
4. GBM . This was the global boosted regression models machine learning method results of the caret package and gbm package working together.
5. KNN. This was the k-nearest neighbor method results of the caret package for knn machine learning.
6. RPART. This was the recursive partitioning and regression tree modeling machine learning algorithm results using the rpart package.
7. GLM . This was the generalized linear model machine learning algorithm results of the glm method of the caret package

8. Combined. This was 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.
9. Type. This was the true value type of each of the 36 samples in the testing set these algorithms used.