Lewis University

Meta-Analysis of the Genes Ubiquitously

Associated with Human Uterine Leiomyoma Development

in Healthy Humans Using

the Gene Expression Omnibus Data

by

Janis L. Corona

Romeoville, IL

July 2019

Abstract

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

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

Table of contents

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

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

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

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

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

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

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

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

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

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

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

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

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

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

r statistical software for statistical analysis……………… 9

Deriving the data sets

Samples simulated in the Population

machine learning algorithms used

chapter 3 – results…………………………………………………………………. 13

chapter 4 – conclusions………………………………………………………….. 34

chapter 5 – literature cited…………………………………………………… 36

appendix………………………………………………………………………………... 42

List of tables

Table 1. Table of Six Ubiquitous Genes in Majority of Chromosome Expressed……..… 16

Table 2. Simulated Means of Top 10 plus 6 from Ten Thousand Samplings per Gene…. 20

Table 3. Top 10 Plus Six Genes with the Full Gene Name..……………………………… 25

Table 4. Table of the Predicted Outcomes for LDA, RF, and GBM Algorithms……….. 28

Table 5. Table of the Combined Model…………………………...…………………….. 30

Table 6. Various Data Set Results from Machine Learning Algorithms………………… 33

List of figures

Figure 1. Heatmap of Six Ubiquitous Genes in Common in All Samples………………… 12

Figure 2. Heatmap of Six Ubiquitous Genes in UL Samples Only………… ……………. 13

Figure 3. Heatmap of Six Ubiquitous Genes in Non-UL Samples Only...………………... 14

Figure 4. Reverse Strand of Cytoband 11p15.5 Genes Expressed More in UL Near *BET1L* 18

Figure 5. Forward Strand of Cytoband 12q14.3 Genes Expressed Less in UL Near *HMGA2* 19

Figure 6. Reverse Strand of Cytoband 17q25.3 Genes Expressed Less in UL Near *CCDC57* and *FASN* 20

Figure 7. Forward Strand of Cytoband 22q13.1 Genes Expressed More in UL Near *CYTH4* and *TNRC6B* 21

Figure 8. Histogram of TOP 10 PLUS 6 Simulated Means of 10K Samplings per Gene………….. 26

List of abbreviations

BMI Body Mass Index

DE Differential Expression

GBM Generalized Boosted Regression Models

GEO Gene Expression Omnibus Online Data Repository

GWAS Genome Wide Association Studies

HGNC HUGO Gene Nomenclature

LDA Latent Dirichlet Allocation

MAF Minor Allele Frequency

RF Random Forest method in the caret package

RF2 Random Forest function in the randomForest R package

SNP Single Nucleotide Polymorphism

UL Uterine Leiomyoma

Introduction

Description of UL

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

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

UL Described in Populations

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

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

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

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

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

Significant Genes for UL

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

Cytoband Location of UL Gene

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

Chromosome 11

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

Chromosome 12

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

Chromosome 17

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

Chromosome 22

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

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

Methods

GEO Data of UL and Non-UL Samples

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

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 fields needed to combine all the samples together, see the Appendix items 1 through 8. The script that merged all of these data sets to make a universal set of all genes in common is in the Appendix as item 9 ‘All\_analysis.R’ using R. An extension was added to each sample column name as ‘ul’ if the sample was UL to keep an order of samples by UL and non-UL fields in the data before merging all sets together by NCBI gene ID labeled ‘GENE’ and keeping the ‘CYTOBAND’ field for creating a subset of data by cytoband location of the six genes associated with UL risk. Fields that weren’t necessary to merge by were excluded but kept in a separate file of meta data to use later as needed. This file is in the Appendix as item 10 listed as “GSE\_array\_meta.csv.” The data set of all genes and samples that excludes most meta data information is in the Appendix as item 11 listed as ‘mrg5.csv’ and it is 1.1 Gb in size. This data set has 1,954,853 genes with many duplicate gene entries from the merge process and 123 column fields that include the 121 samples with the extended ‘ul’ name attached to the UL samples and two fields 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 fields as above. This data set is in the Appendix as item 12 listed as “DE\_means\_Per\_Gene\_Chr.csv.”

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

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

Stable ID,’ ‘Strand,’ ‘Chromosome/Scaffold name,’ ‘Gene Start (bp),’. ‘Gene end (bp),’ and ‘Gene Name.’ When done the results were exported as csv format for ‘all’ entries and saved as “mart\_export.txt” with a file size of 14.1 MB. This file is in the Appendix as listed item number 17. The last data set was then merged with the “ub\_genes\_ensembl.csv.” data set by ENSEMBL transcript ID after making minor modifications to the imported “mart\_export.txt” data set. The modifications made were to drop unnecessary fields and modify the strand values by changing the ‘-1’ to ‘-‘ and the ‘1’ to ‘+’ to use in Gviz. A field 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 fields with names shortened to "chromosome," "start," "end," "width," "strand,” "gene," "transcript," and "symbol." This data set is in the Appendix as item 18 listed as “ub\_genes\_ensembl\_gviz.csv.”

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

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

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

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

Samples Simulated in the Population

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

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

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

The “MemberMagnitude\_130\_142.csv.” listed as item 22 in the Appendix was then used to generate more data sets based on this subset of genes on the same cytoband location as the six UL risk genes. One data set is a subset of the overall top 16 genes out of the 130 genes that have the highest magnitude of change. This data set, “most\_DE\_ml\_ready\_130.csv,” did not add in the six genes ubiquitous to UL risk studies and can be found in the Appendix as item 26. To create item 26 of the Appendix, the data set it was derived from as item 22 in the Appendix removed the fields 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 field was added as the first field called, type, that would attach the type of each sample ID as either UL or non-UL. This was easy since the first 51 were already non-UL and the last 70 were UL with an extension to the ID that also showed it as a UL sample. This was so that this data set could be used in the following machine learning algortihms to see how accurate the results use these genes as gene targets in predicting a sample as UL or non-UL.

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

Using this same data set that created the first three data sets, “MemberMagnitude\_130\_142.csv,” dplyr was used to add a fold change field 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 fields 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 field was added to label each of the samples as either UL or non-UL so that the type field would be the field with which to predict accuracy in determining a sample as UL or not using the genes as variables. That data set is item 28 in the Appendix listed as “'FOLD16\_ml\_ready.csv.”

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

Additional data sets were made from all genes in common using the “universe\_12173.csv” data set in the Appendix as item 30, made from the data set as item 12 in the Appendix. The means of UL and non-UL were added to each row, then the difference between the two groups, then the magnitude as the absolute value of the difference, then the fold change as the absolute value of the ratio of the UL mean to the non-UL mean. One data set listed as item 31 in the Appendix as “most\_universe\_fold.csv” was made from that data set by adding a fold change field 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. Fields other than the sample ID fields 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 field was added to attach what type of sample each observational sample is as either a UL or non-UL sample. This made each data set ready to be used in the machine learning algorithms to predict the outcome as the type based on the regressions on the genes as variables for each row sample. If the accuracy from the models scored well, this could be an indicator that some genes out of all the genes in common having the most change in UL compared to non-UL are gene targets for evaluating if those genes are related to UL pathogenesis.

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

Finally, another data set made from the item 12 data set in the Appendix, “DE\_means\_Per\_Gene\_Chr.csv,” was a data set that used the same magnitude of differential expression between UL and non-UL samples. This data set took the bottom 16 or 16 least expressed or inhibited genes in UL compared to non-UL samples by magnitude of change between the UL and non-UL means for each gene. This data set is item 33 in the Appendix and listed as “least\_universe\_DE.csv.” The same fields other than the sample ID fields 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 fields. Then a field was added as the first field 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 field is the outcome field each model was regressed or clustered against to produce an outcome of either UL or non-UL based on the type field.

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 field on all the genes. The MASS package is used with caret for the support functions and generalized linear models, poisson, binomial, and ‘modern applied statistics with S’ (Ripley, et al., 2019). The randomForest package uses its own built in algorithm for random forest classification using the e1071 package that stands for ‘Miscellaneous Functions of the Probability and Statistics Group’ (Breiman, et al.,2018; Meyer, et al., 2015). The RF2 is the second version of the random forest algorithm that used the randomForest package of R instead of the random forest method of the caret package. The tuning parameters for this RF2 algorithm by default sample with replacement on 500 trees in classifying data based on the training set (Breiman, et al., 2018). The RF2 algorithm settings for the purposes of training on the data sets of this research set the method to ‘class’ in the default settings of the randomForest function in this RF2 algorithm. The caret package is the classification and regression training in R that supplies the LDA, RF, rpart, GLM, KNN, and GBM algorithms as methods in its ‘train’ function (Kuhn, et al, 2019).

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

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

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

Results

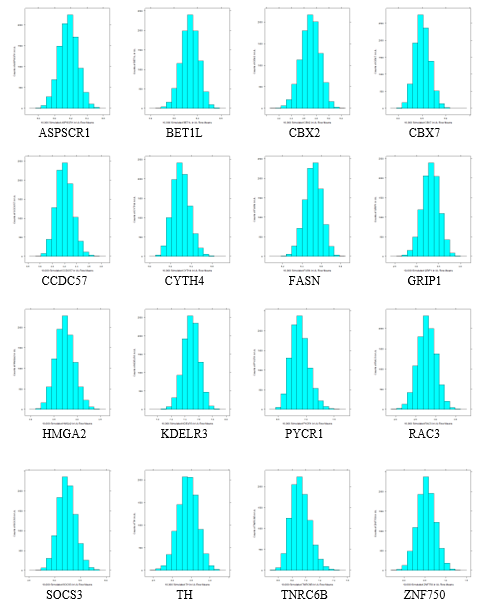
The

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

|  |  |  |
| --- | --- | --- |
| **gene** | **GENE\_NAME** | **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 |

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Top 10 + 6 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 |



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

A screenshot of a map

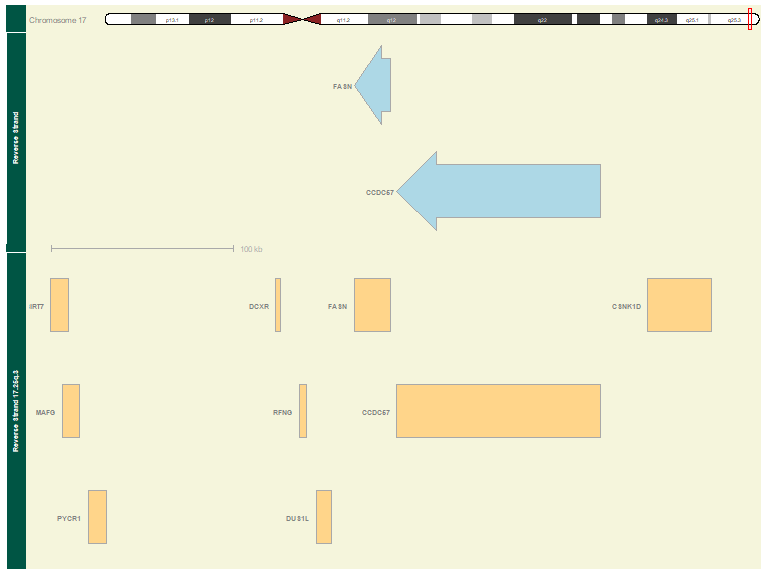
Description automatically generated

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

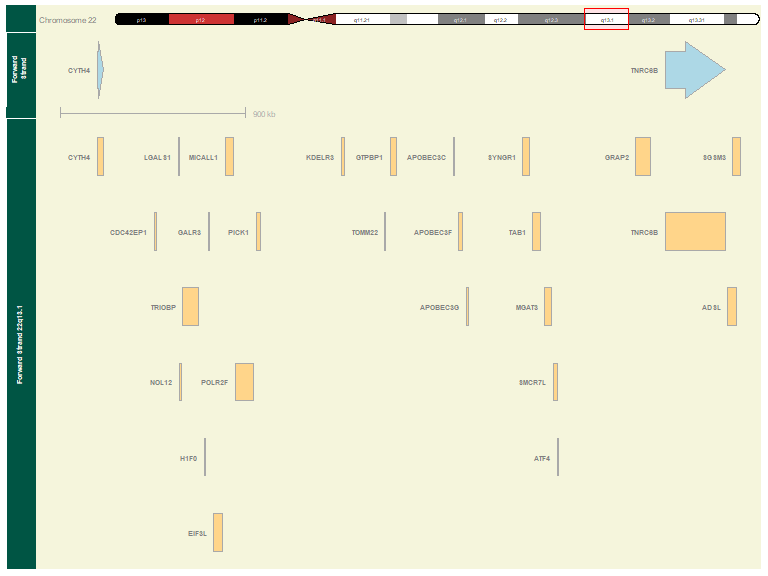
A close up of a map

Description automatically generated

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



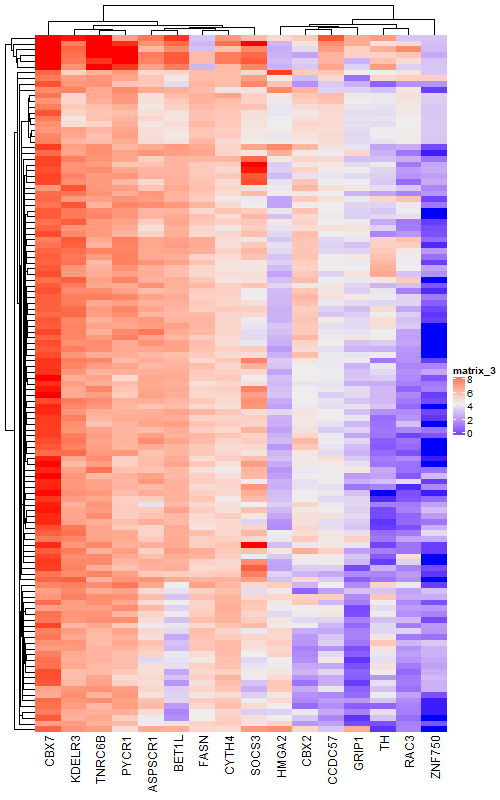
**Figure** **4**. Gviz Map of Reverse Strand of Cytoband 17q25.3 Genes Expressed Less in UL.



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

**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 |



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

A screenshot of a cell phone

Description automatically generated

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

A close up of a map

Description automatically generated

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

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

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

**Table 11:** Machine Learning Results on All Data Sets

Conclusions

.

literature cited

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

1. GPL96. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL96](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL96). This is one of three GEO platforms that was combined with the microarray samples from the five GEO microarray series listed above as items 1 through 5. This platform identified the probe IDs of GSE593, GSE2724, and GSE23112. Only the ‘ID’ field was used to merge with the other four data sets and then most of the fields from GPL6480 were used. There are 22,283 genes and 16 fields of additional information with some directly quoted from the table excel file. These fields are identical to the GPL570 platform because they are both the Affymetrix Human Genome U133 Array, but GPL570 is the ‘Plus 2’ version.
   1. ID : this is the ID field to merge with GSE593, GSE2724, and GSE23112 GEO series
   2. GB\_ACC: Factor. This is the gene bank accession number for each gene
   3. SPOT\_ID: Factor. This is either ‘control’ or ‘NA’
   4. Species.Scientific.Name: Factor. This is equal to ‘Homo sapiens’ for all
   5. Annotation.Date: Factor. The date the data platform IDs annotated, all equal ‘Oct 6, 2014’
   6. Sequence.Type: Factor with three values of ‘Exemplar Sequence,’ ‘Control Sequence,’ or ‘Consensus Sequence’
   7. Sequence.Source: Factor with one level of ‘Affymetrix Proprietary Database GenBank.’ Described as ‘the database from which the sequence used to develop this probe set was taken’
   8. Target.Description: Factor with 21,362 levels describing each gene
   9. Representative.Public.ID: Factor. The accession number of a representative sequence.
   10. Gene.Title: Factor. The title of the gene represented by the probe set.
   11. Gene.Symbol: UniGene gene symbol
   12. ENTREZ\_GENE\_ID: Factor. ENTREZ gene database UID
   13. RefSeq.Transcript.ID: Factor. References to multiple sequences in RefSeq
   14. Gene.Ontology.Biological.Process: Factor. ‘Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence’
   15. Gene.Ontology.Cellular.Component: Factor. ‘Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
   16. Gene.Ontology.Molecular.Function: ‘Gene Ontology Consortium Molecular Function derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
2. GPL570. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL570](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570). This is one of three GEO platforms that was combined with the microarray samples from the five GEO microarray series listed above as items 1 through 5. This platform identified the probe IDs of GSE13319. There are 54,675 genes and 16 fields that are identical to GPL96, because this is the Affymetrix Human Genome U133 Plus 2.0 Array and GPL96 is the Affymetrix Human Genome U133 Array an earlier version.
   1. ID : this is the ID field to merge with GSE593, GSE2724, and GSE23112 GEO series
   2. GB\_ACC: Factor. This is the gene bank accession number for each gene
   3. SPOT\_ID: Factor. This is either ‘control’ or ‘NA’
   4. Species.Scientific.Name: Factor. This is equal to ‘Homo sapiens’ for all
   5. Annotation.Date: Factor. The date the data platform IDs annotated, all equal ‘Oct 6, 2014’
   6. Sequence.Type: Factor with three values of ‘Exemplar Sequence,’ ‘Control Sequence,’ or ‘Consensus Sequence’
   7. Sequence.Source: Factor with one level of ‘Affymetrix Proprietary Database GenBank.’ Described as ‘the database from which the sequence used to develop this probe set was taken’
   8. Target.Description: Factor with 21,362 levels describing each gene
   9. Representative.Public.ID: Factor. The accession number of a representative sequence.
   10. Gene.Title: Factor. The title of the gene represented by the probe set.
   11. Gene.Symbol: UniGene gene symbol
   12. ENTREZ\_GENE\_ID: Factor. ENTREZ gene database UID
   13. RefSeq.Transcript.ID: Factor. References to multiple sequences in RefSeq
   14. Gene.Ontology.Biological.Process: Factor. ‘Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence’
   15. Gene.Ontology.Cellular.Component: Factor. ‘Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
   16. Gene.Ontology.Molecular.Function: ‘Gene Ontology Consortium Molecular Function derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence”.’
3. GPL6480. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GPL6480](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6480). This is one of three GEO platforms that was combined with the microarray samples from the five GEO microarray series listed above as items 1 through 5. This platform identified the probe IDs of GSE68295. There are 41,108 genes and 17 identifying fields in this platform. The fields in this data set are all factors. The following are the listed fields used to merge all other GSE series and GPL platforms to while keeping only the needed fields from this table. The field IDs are labeled as how they are described in the downloaded SOFT text file.
   1. ID : Agilent feature number
   2. SPOT\_ID : Spot identifier
   3. CONTROL\_TYPE : Control type
   4. REFSEQ : RefSeq Accession number
   5. GB\_ACC : GenBank Accession number
   6. GENE : Entrez Gene ID
   7. GENE\_SYMBOL : Gene Symbol
   8. GENE\_NAME : Gene Name
   9. UNIGENE\_ID : UnigeneID
   10. ENSEMBL\_ID : EnsemblID
   11. TIGR\_ID : TIGRID
   12. ACCESSION\_STRING : Accession String
   13. CHROMOSOMAL\_LOCATION : Chromosomal Location
   14. CYTOBAND : Cytoband
   15. DESCRIPTION : Description
   16. GO\_ID : GoIDs
   17. SEQUENCE : Sequence
4. GSE593. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE593](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE593). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL96 GEO platform listed as item 7. This data set shares the same Probe ID as GSE23112 and GSE2724 because they all share GPL96. This data contributed five UL and five non-UL samples to the 121 total samples. There are 22,283 genes in this raw data.
5. GSE2724. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE2724](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2724). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL96 GEO platform listed as item 7. There are 7 UL and 11 non-UL samples as headers with one probe ID field the same as GSE593 and GSE23112. There are 22,283 genes in this raw data.
6. GSE68295. Retrieved March 2019 from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE68295 This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL6480 GEO platform listed as item 8. This data set added 3 UL and 3 non-UL samples to the total 121 samples, but also was needed for the attached information the platform to this data set contained. The various recognized gene names, chromosome, cytoband information, and other meta fields was useful for the analysis. This raw data set had 41,078 genes in it.
7. GSE13319. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE13319](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13319). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL570 GEO platform listed as item 6. This data set used only the human samples from a combined set of human and rat UL. In total 50 UL samples and 27 non-UL samples were added to the 121 total samples. This data had 54,675 genes.
8. GSE23112. Retrieved March 2019 from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc:GSE23112](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23112). This is one of five microarray gene expression data sets from GEO that was merged with its corresponding platform and the other four series of samples and two other platforms to find the universe of all genes in common among the UL and non-UL samples. The platform for this series is the GPL96 GEO platform listed as item 7. This data contributed five UL and five non-UL samples to the 121 samples total. With the same Probe ID field as GSE2724 and GSE593. There are 22,283 genes in this raw data set.
9. All\_analysis.R. Accessible from <https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research/blob/master/All_anlaysis2.R> . This is the R script for all data tables and images produced on the raw data of items 1 through 8 of the Appendix. The version this script used is version 3.6. The packages used are listed in the script but commented out. The packages installed into R to run the script in some sections are: ‘dplyr’,’rpart’,’caret’,’MASS’,’e1071’,’randomForest’,’ggplot2’,’lattice’,’heatmaply’,’plotly’,’Gviz’, ‘ComplexHeatmap’,’GenomicRanges’, and ’UsingR’
10. GSE\_array\_meta.csv.
11. mrg5.csv
12. DE\_means\_Per\_Gene\_Chr.csv
13. chr\_loci\_top\_genes.csv
14. ub\_genes\_gviz.csv
15. ensembl\_generated\_id.csv
16. ub\_genes\_ensembl.csv
17. mart\_export.txt
18. ub\_genes\_ensembl\_gviz.csv
19. All-ggplot2-type-sample-derived.csv
20. DE\_data\_unordered.csv
21. MemberGviz\_130\_141.csv
22. MemberMagnitude\_130\_142.csv
23. 'TOP16\_ml\_ready.csv'
24. ubiq\_and\_top10\_samples\_only.csv
25. Stats16.csv
26. most\_DE\_ml\_ready\_130.csv
27. least\_DE16\_ml\_ready\_130.csv
28. FOLD16\_ml\_ready.csv
29. majority\_ml\_ready\_10\_total.csv
30. universe\_12173.csv
31. most\_universe\_fold.csv
32. most\_universe\_DE.csv
33. least\_universe\_DE.csv
34. “Results\_predictions\_DE16\_8\_algorithms\_used.csv”
35. “Results\_predictions\_Least\_DE16\_8\_algorithms\_used.csv”
36. “Results\_predictions\_FOLD16\_8\_algorithms\_used.csv”
37. “Results\_predictions\_majority10\_8\_algorithms\_used.csv”
38. ML results top 16 universe Fold, “ ”
39. ML results most universe DE, “ “
40. ML results least universe DE, “ “