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 examines five microarray gene expression samples of uterine leiomyomas (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 is UL or not. The algorithms used were Latent Dirichlet Allocation (LDA), random forest (RF), general boosted regression models (GBM), k-nearest neighbors (KNN) and principal component analysis (PCA). The LDA model and random forest models could accurately predict whether or not a sample is UL or non-UL with 88-91% accuracy using the six genes ubiquitous to the current research on UL risk and the ten 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, top 16 genes, 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

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
GWAS	Genome Wide Association Studies
HGNC	HUGO Gene Nomenclature
LD	Linkage Disequilibrium
LDA	Latent Dirichlet Allocation
MAF	Minor Allele Frequency
PCA	Principal Component Analysis
RF	Random Forest
SNP	Single Nucleotide Polymorphism
TOP16	Top 10 Most Expressed Genes in Magnitude in UL, Plus the Six Genes Ubiquitous to Current Research on UL Risk
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 oestrogens 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 SNPs of 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 *BET1L* associated with what part of 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 (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 genotypes 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 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 Bet1 Golgi Vesicular Membrane Trafficking Protein like gene called *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 SNPs shown in separate

population specific studies to associate to the number of UL one patient has (rs2280543, *BET1L*) and the size of the UL one person has (rs12484776, *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.

Chromosomal 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 is used for LD analysis in some of the current

literature to find genes with a high LD and significant association to UL risk (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 was analyzed in data made available for gene expression using GEO. There are many genome wide association studies (GWAS) on the few genes having certain genotypes associated with UL, after evaluating the single nucleotide polymorphisms (SNP) in those genotypes (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 are combined by genes that are in common, and then filtered for those genes that are along the same chromosomal bands indicated in some of the studies of UL risk for possibly having an association for UL risk (Bondagi, et al., 2017; Cha, et al., 2011) . This study is limited by the data not providing all the genotypes of the six genes ubiquitous to UL risk in the current UL risk studies from the data on the gene expression values in microarray samples made available through GEO. As only one study, GSE68295, of the five

GEO studies combined has the sequence information for each gene's genotype (Miyata et al., 2017; Vanharanta, et al., 2006; Hoffman, Milliken, Gregg, Davis, & Gregg, 2004; Zavadil, et al., 2010; Crabtree, et al., 2009). Thus, the same type of data science methods employed by the studies on UL risk are not able to be used, such as filtering for SNPs by either of having minor allele frequency (MAF), fold change, or linkage disequilibrium (LD) greater than a set threshold (Eggert, et al., 2012; Hodge, et al., 2012). Because of this limitation, 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 will be analyzed (TOP16). Data science methods will be used to determine a model based on an algorithm in RStudio and Bioconductor software that is best to predict if a sample is a UL or non-UL sample (R, 2019; Bioconductor, 2019).

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. GEO had two other UL risk studies available to analyze but were excluded as they were oligonucleotide beads and not the microarray 'in situ oligonucleotide' gene expression data type.

R Statistical Software for Statistical Analysis.

The R software was used to combine the GEO independent studies into one larger data set of genes in common among all the studies, but that also have the six genes ubiquitous to the current UL risk population studies. This larger data set of 12,173 genes was then filtered in R to only use the genes along the same chromosomal locations as those six genes *TNRC6B, BET1L, FASN, CYTH4, CCDC57,* and *HMGA2* (the raw data sets can be linked to by visiting the Appendix section, 'Data Sets, Files, and Images in this Research Study'). That data set gave a

table of 183 genes but was then filtered to only have 130 unique genes. The base statistical functions in R were primarily used to combine by merging the GEO data sets and applying the first listed item in a field when many observations had multiple entries. More work was done to add ENSEMBLE data fields to the data to use with the Gviz Bioconductor package in R for visualizing the gene locations. The R package, dplyr, was used to add fields that describe the means of each gene in the samples of UL and non-UL separately, then add a field of changes in expression by means of each type of sample and use as a way to group the genes by differential expression in up and down expression in UL compared to non-UL samples, and for determining if the genes were part of the majority of genes expressed more or less in each chromosome (Francois, Lionel, & Muller, 2019). 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). The R packages, ggplot2, heatmaply, and lattice were used along with R base package to create plots that could describe the data visually to look for patterns between the genes, samples, or stats of the samples (Wickham, 2019; Galili, O'Callaghan, Sidi, & Benjamini, 2019; Sarker, 2018). Bootstrap simulations using the 'UsingR' r package using 10,000 simulations with replacement for each TOP16 gene (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). This was generated per TOP16 gene based on a generalization of the Central Limit Theorem and the Law of Large Numbers which state that a sample of a larger population will converge to the true population mean when sampling with replacement is done a large amount of times. One simulated population mean for UL and one for non-UL converged from 10,000 samplings per TOP16 gene of the combined 121 GEO samples. The predictive algorithms of PCA, LDA, RF, KNN, and

GBM were used on this dataset of TOP16 genes using kernlab, caret, gbm, lda, randomForest, e1071, and MASS r packages (Karatzoglou, Smola, Hornik, Maniscalco, & Teo, 2018; 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).

Bioconductor for Biostatistics in R.

One of the packages from Bioconductor, Gviz, was used to map out the chromosomes 11, 12, 17, and 22 separately with the TOP16 genes (Hahn, F., 2019; Bioconductor, 2019). Gviz was able to place the ENSEMBL gene ID onto a map of each of the four chromosomes each of the TOP16 genes reside, as well as the cytoband location on the gene, a strand direction of forward or reverse with an arrow pointing left or right respectively, and the start and end width of each gene by size of each arrow in that chromosome cytoband location (ENSEMBL, 2019). This was useful to separately see a group of the genes in common among the studies, next seeing which are more up or down expressed in UL compared to non-UL, then observing which are part of the majority of genes up or down regulated, and finally which genes are one of the TOP16 genes.

RESULTS

The data table of 12,173 genes unique and common to each of the five GEO microarray series of UL and non-UL tissue samples was generated in R from a much larger combined data table of all five tables that had many duplicate entries created when merging the five series of GEO samples. That much larger data set was 1,954,853 genes long with many duplicate genes and entries because it attached every different sequence from one data series to the other four data series that did not have sequence information. This file and all files are in the Appendix section of this document with a digital link to the files and R script. These duplicates were removed and means with counts of each gene were produced in the data table of 12,173 genes. Then, exploratory data analysis was used on the data table to see if some type of relationship between the UL and non-UL samples could be observed visually with plotting by lattice, ggplot2, and heatmaply R software graphical plotting packages. Figure 1 shows the heatmap of the ubiquitous genes and samples as they are, grouping the most similar clusters together. Figure 2 shows the heatmap of the ubiquitous genes and UL samples. Figure 3 shows the heatmap of the ubiquitous genes in non-UL samples. The heatmaps in Figure 1, Figure 2, and Figure 3 didn't show any useful information to pursue based on the quick plotted heatmap.

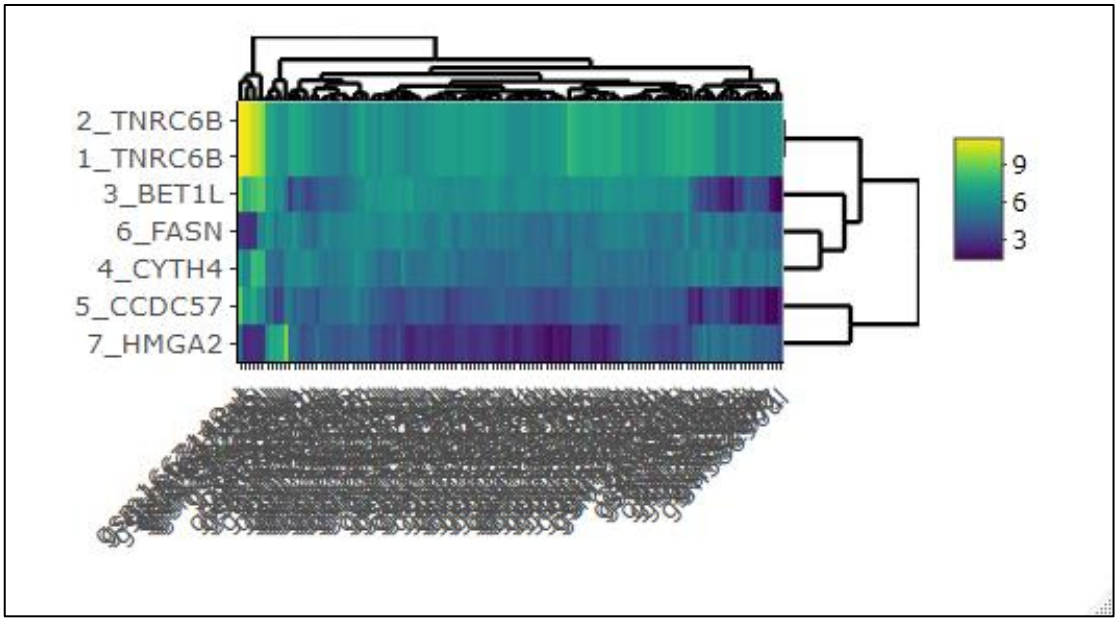

Figure 1: Heatmap of Six Ubiquitous Genes in Common in All Samples

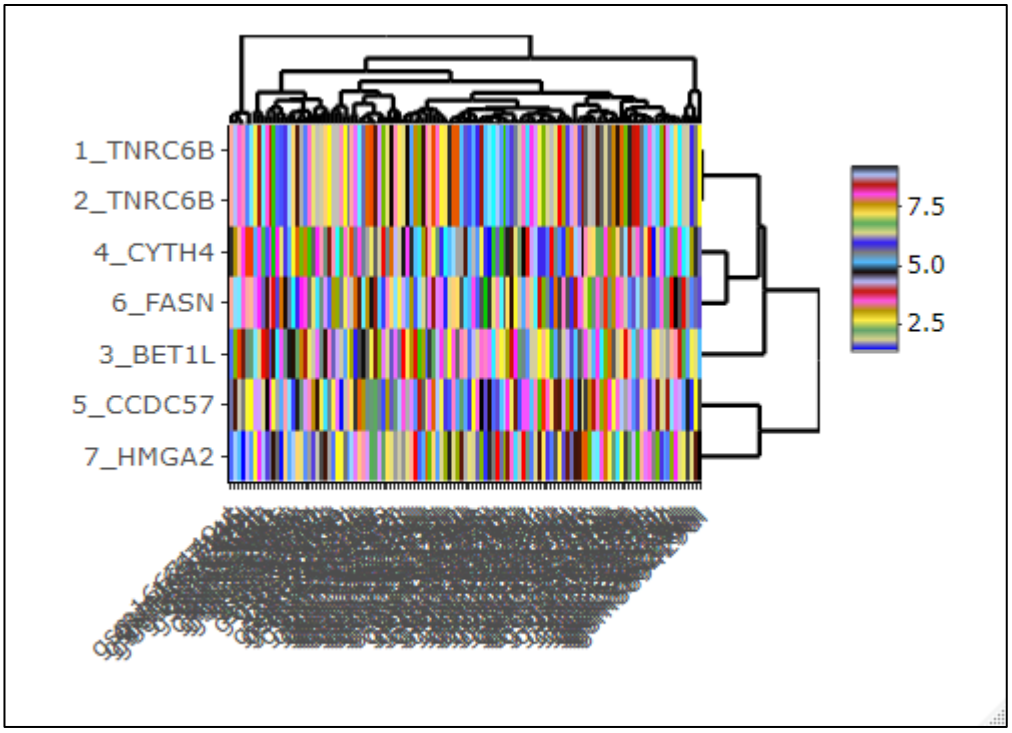

Figure 2: Heatmap of Six Ubiquitous Genes in UL Samples Only

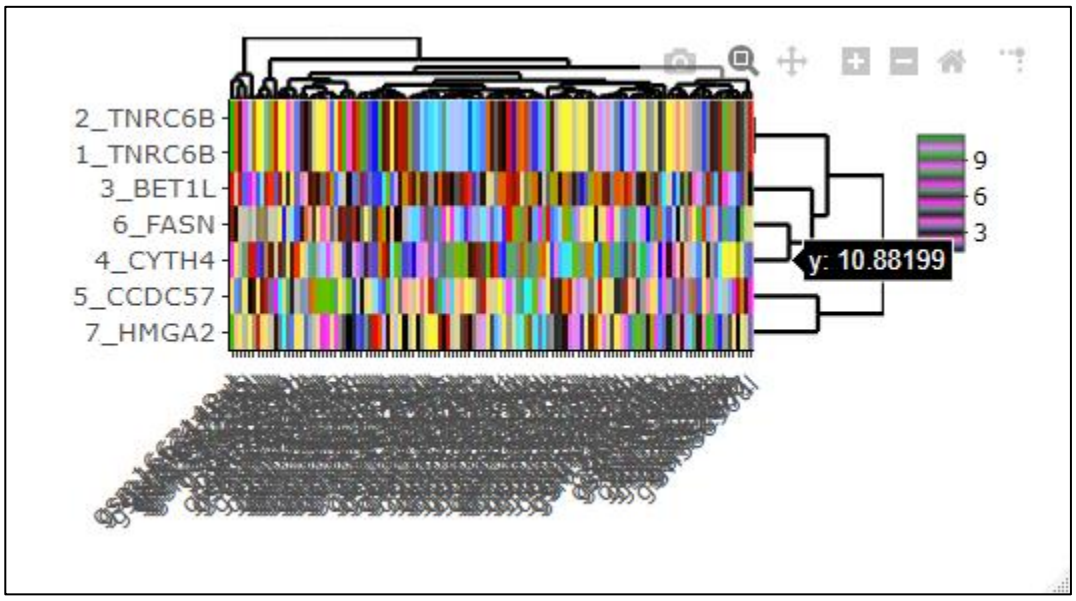

Figure 3. Heatmap of Six Ubiquitous Genes in Non-UL Samples Only

A table made to calculate the mean of each of the 130 genes among all 121 samples for UL and non-UL separately and the difference between the mean of the UL and non-UL samples as the mean of each gene in the UL subset minus the mean of the same gene in the non-UL subset to get the differential expression (DE) between the non-UL and UL samples. This data was further divided as a part of a decision tree algorithm by whether the DE value is positive for expressed more in UL for that gene or negative for expressed less in that gene for UL. There were 70 genes expressed more, exactly 60 genes expressed less in UL, and one gene, *KDELR3,* expressed the same. Gviz was used to map out these genes on their chromosomal bands, to show the neighborhood of genes along each of the 130 genes in common among these five data series. Then, each chromosome of the four that the six ubiquitous genes reside were further divided into groups of those that are expressed as a part of the majority of genes expressed more in UL for each chromosome and those that are not part of the majority of genes expressed more in UL for each chromosome. This was done to see if there was some sort of pattern that could be shown from those that differ in behavior in UL samples compared to non-UL samples.

Table 1 shows the six ubiquitous gene values for each chromosome and what genes are not part of the group of genes expressed in each chromosome. This could be an indicator for those genes that do have changes in UL, but not following the changes that the other genes follow as in negative correlation between those genes that change in UL but opposite most genes that change. There were 53 genes out of 130 that were not part of the 77 genes that did change most in UL for each chromosome the six ubiquitous genes are located. *HMGA2* is expressed less while *TNRC6B* and *CYTH4* are expressed more as most genes expressed in UL, and *BET1L* is expressed more while *FASN* and *CCDC57* are expressed less as the minority of genes expressed in UL as Table 1 shows.

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

Genes	Chromosomes	Type	All	Up	Down	Majority
BET1L	11	Up	33	15	18	FALSE
TNRC6B	22	Up	43	28	15	TRUE
CYTH4	22	Up	43	28	15	TRUE
CCDC57	17	Down	48	26	22	FALSE
FASN	17	Down	48	26	22	FALSE
HMGA2	12	Down	6	1	5	TRUE

The 130 genes common to all these UL and non-UL samples were than complemented with a magnitude field which sorted in order of most change (up or down) the genes to pick the top 10 most expressed genes in UL. In Figures 4 through Figure 7, some chromosomal mapping was done to show where these top 10 genes and the six ubiquitous genes live on each of the four chromosomes with the genes that are ubiquitous to the current UL risk meta-analysis studies by using the Gviz R package. One of the current meta-analysis UL risk studies have mentioned how other genes along the loci of the genes ubiquitous to UL risk studies were found but ignored due to low MAF or being unobservable at the GWAS level of significance (Aissani, 2015). Another study by Bondagji, et al. (2017) said these loci showed a strong linkage to genes associated with UL predisposition or risk because the regression score for accuracy was higher than 80 per cent for LD scores when looking at peaks in the chromosomes where *CCDC57*, *FASN*, and *BET1L* reside.

The idea of mapping the cytoband location for these genes came from these studies mentioning other genes being ignored but having strong associations with UL risk by showing strong linkage to genes associated with UL risk in some populations. The data set that can be used for this plot with the 130 genes in common, the Gviz fields needed to run the R script, and other fields for analysis such as the UL and non-UL means, the differential expression between UL and non-UL by means, the magnitude of change between UL and non-UL samples, and the fold change of the ratio of UL means to non-UL means for each of the 130 genes is in the Appendix under the ‘Data Set of All Added Fields for 130 Common Genes’ section.

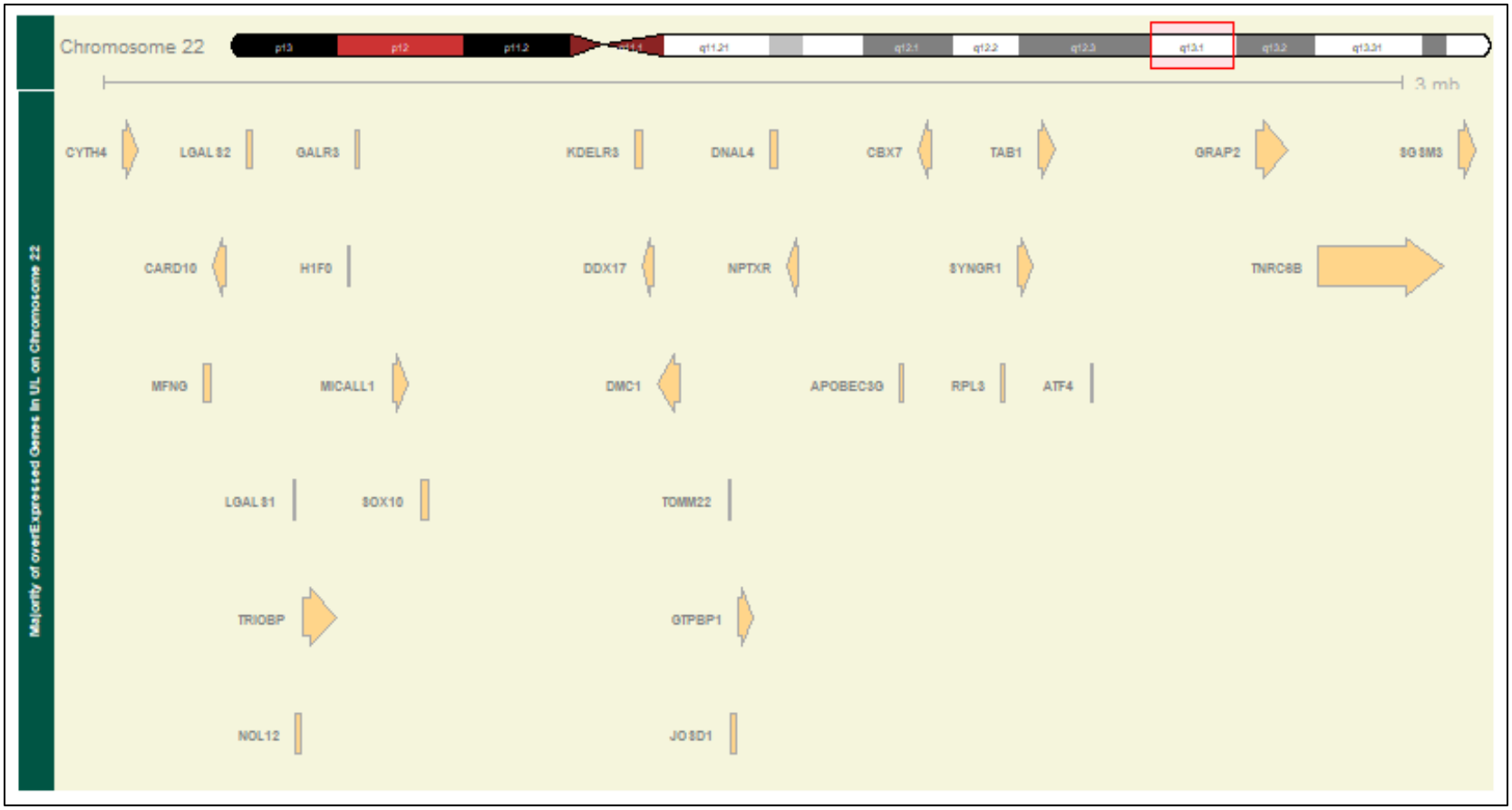

Figure 4. Gviz Map of Chromosome 22 Majority of Genes Expressed More in UL

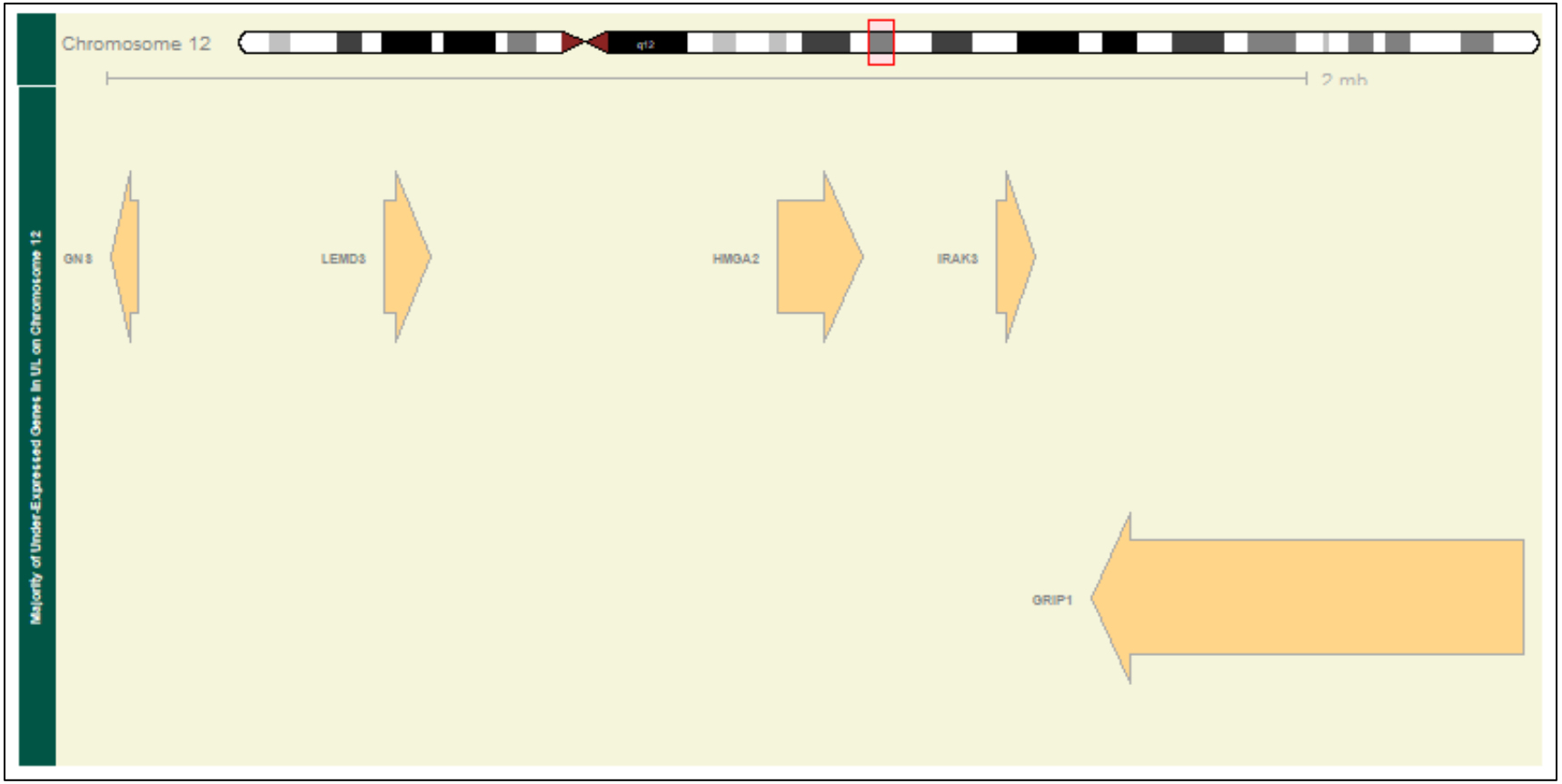

Figure 5. Gviz Map of Chromosome 12 Majority of Genes Expressed Less in UL

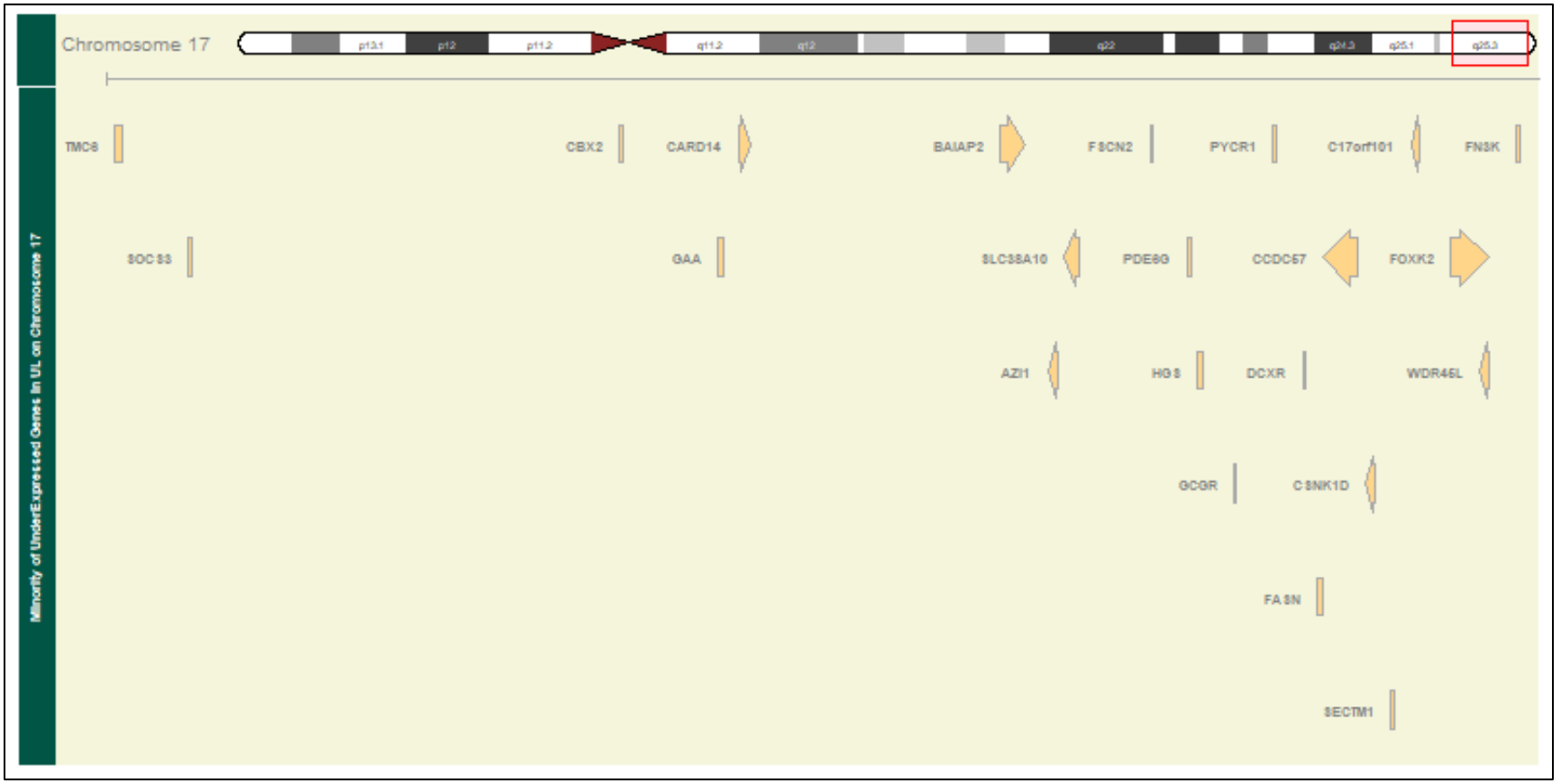

Figure 6. Gviz Map of Chromosome 17 Minority of Genes Expressed Less in UL.

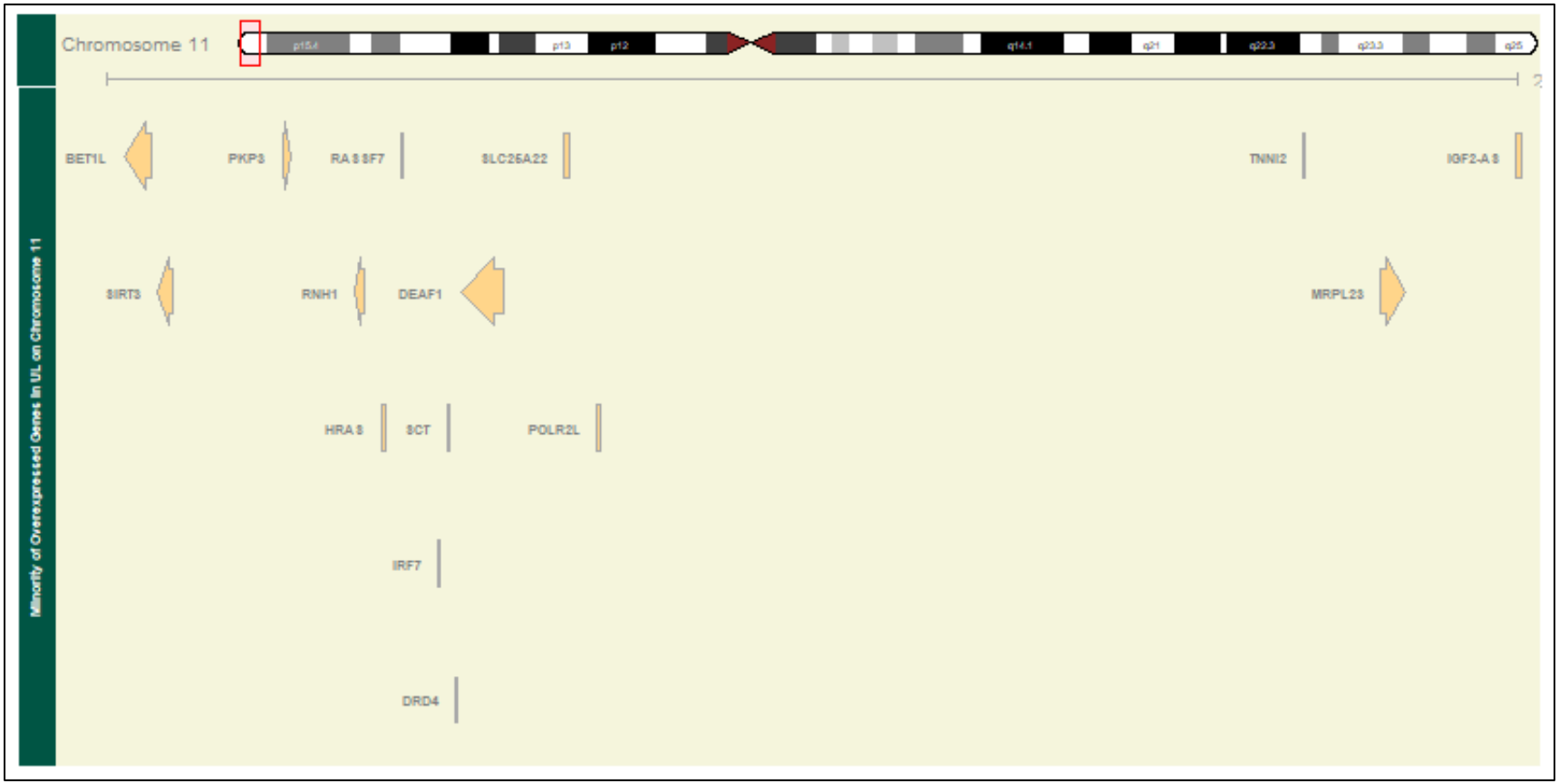

Figure 7. Gviz Map of Chromosome 11 Minority of Genes Expressed More in UL

When a magnitude field was added to the table of all 130 genes, this allowed the top 10 genes having the most differential expression of change in either up or down gene expression in UL. From this table of 10 most expressed or under expressed genes in UL compared to non-UL samples, the six ubiquitous genes were added which has been abbreviated earlier in the methods section as TOP16 but isn't the top 16 genes expressed. The bootstrap results of 10,000 simulated samplings of each gene resulted in Table 2. This table shows the TOP16 genes and how they are comparable in magnitude when expressed across UL samples. There are 32 observations because there are mean and standard deviations for each gene as 'UL' or as 'nonUL' in the 'ulStatus' field. The 16 genes with the simulated means of 10,000 samples of each gene drawn with replacement from a pool of 70 UL and 51 non-UL samples. These are the TOP16 genes. The 'simulatedMean10K' field is the mean of each gene from 10,000 simulated samplings with replacement, the 'simulatedSD10' field is the standard error of those genes' means, the 'leftTail2.5' field is the left side of the two-tailed 95% confidence interval, the 'rightTail97.25' is the right side of the two-tailed 95% confidence interval, the 'ulStatus' field is whether or not the gene simulated results are from the UL or non-UL sample labeled 'UL' and 'nonUL' respectively, and the 'gene' field is the gene that is either the top 10 genes with the largest magnitude of change in UL compared to non-UL samples, or one of the six genes ubiquitous to current UL risk meta-analysis studies.

Table 2. Simulated Means of TOP16 from Ten Thousand Samplings per Gene

	simulatedMean10k	simulatedSD10K	leftTail2.5	rightTail97.25	ulStatus	gene
1	0.5323693	0.21577234	0.1152786	0.9488729	UL	FSCN2
2	4.7027510	0.15190607	4.3912714	4.9797143	UL	CARD10
3	5.2372015	0.17186176	4.9135245	5.5778431	UL	GRIP1
4	2.7801885	0.22669439	2.3370539	3.2145098	UL	CANT1
5	7.4854601	0.11669587	7.2535245	7.7039216	UL	IRF7
6	3.3206441	0.19550963	2.9243088	3.6905936	UL	ARHGDIA
7	6.8912070	0.15764669	6.6229412	7.2241230	UL	NOL12
8	3.4063468	0.25704965	2.9019167	3.8825544	UL	SLC25A10
9	8.5049034	0.15824961	8.2066667	8.8207843	UL	SLC38A10
10	6.1686359	0.11400604	5.9421569	6.3876471	UL	RNH1
11	4.2316854	0.22174494	3.7966618	4.6476471	UL	BET1L
12	1.5327457	0.19534709	1.1780392	1.9308005	UL	HMGA2
13	9.4750098	0.16203686	9.1735245	9.7956917	UL	CYTH4
14	5.2188861	0.39960207	4.5362745	6.0803975	UL	CCDC57
15	3.4883228	0.13786381	3.2276422	3.7551034	UL	FASN
16	7.4099588	0.07051176	7.2831373	7.5531426	UL	TNRC6B
17	1.3533233	0.24191907	0.8768431	1.8096348	nonUL	FSCN2
18	3.9073855	0.16439952	3.5756765	4.2133387	nonUL	CARD10
19	6.0091253	0.25920759	5.5241176	6.5301961	nonUL	GRIP1
20	2.0392911	0.19571653	1.6656863	2.4201961	nonUL	CANT1
21	6.7638797	0.10061099	6.5637157	6.9517701	nonUL	IRF7
22	2.6410369	0.22302368	2.2078333	3.0711819	nonUL	ARHGDIA
23	6.2069839	0.14204988	5.9523480	6.5003975	nonUL	NOL12
24	2.7575306	0.25160843	2.2613627	3.2311819	nonUL	SLC25A10
25	9.1558256	0.18822490	8.8093922	9.5333387	nonUL	SLC38A10
26	5.5463611	0.12016566	5.3166618	5.7807843	nonUL	RNH1
27	3.7283169	0.19786562	3.3278431	4.0986328	nonUL	BET1L
28	2.0137804	0.25947227	1.5429167	2.5450980	nonUL	HMGA2
29	9.1321560	0.20108231	8.7580294	9.5264760	nonUL	CYTH4
30	5.5418017	0.44041553	4.7762696	6.4670750	nonUL	CCDC57
31	3.5689284	0.14433342	3.2886225	3.8509912	nonUL	FASN
32	7.3534082	0.05615986	7.2441176	7.4605936	nonUL	TNRC6B

After creating the simulated means, histograms of the 16 genes were made to see how symmetric the simulated sampling looks as a predictor for the larger population mean of each gene in UL and non-UL samples. From this most were slightly skewed, but some genes had almost perfect symmetry indicating that these values are close to true representations of gene expression values in UL compared to non-UL samples in a much larger population than 121 samples. Table 3 has the TOP16 genes and their descriptive name. Figure 8 shows these 16 genes from a glance to see how a Gaussian curve would fit over the simulated means generated in Table 2. It is clear from the symmetrical layouts of each gene that these can be considered good candidates for predictors in determining if a sample is UL or not.

Table 3. TOP16 with the Full Gene Name

	genes	GENE_NAME
1	ARHGDIA	Rho GDP dissociation inhibitor (GDI) alpha
2	BET1L	blocked early in transport 1 homolog (S. cerevisiae)-like
3	CANT1	calcium activated nucleotidase 1
4	CARD10	caspase recruitment domain family, member 10
5	CCDC57	coiled-coil domain containing 57
6	CYTH4	cytohesin 4
7	FASN	fatty acid synthase
8	FSCN2	fascin homolog 2, actin-bundling protein, retinal (Strongylocentrotus purpuratus)
9	GRIP1	glutamate receptor interacting protein 1
10	HMGA2	high mobility group AT-hook 2
11	IRF7	interferon regulatory factor 7
12	NOL12	nucleolar protein 12
13	RNH1	ribonuclease/angiogenin inhibitor 1
14	SLC25A10	solute carrier family 25 (mitochondrial carrier; dicarboxylate transporter), member 10
15	SLC38A10	solute carrier family 38, member 10
16	TNRC6B	trinucleotide repeat containing 6B

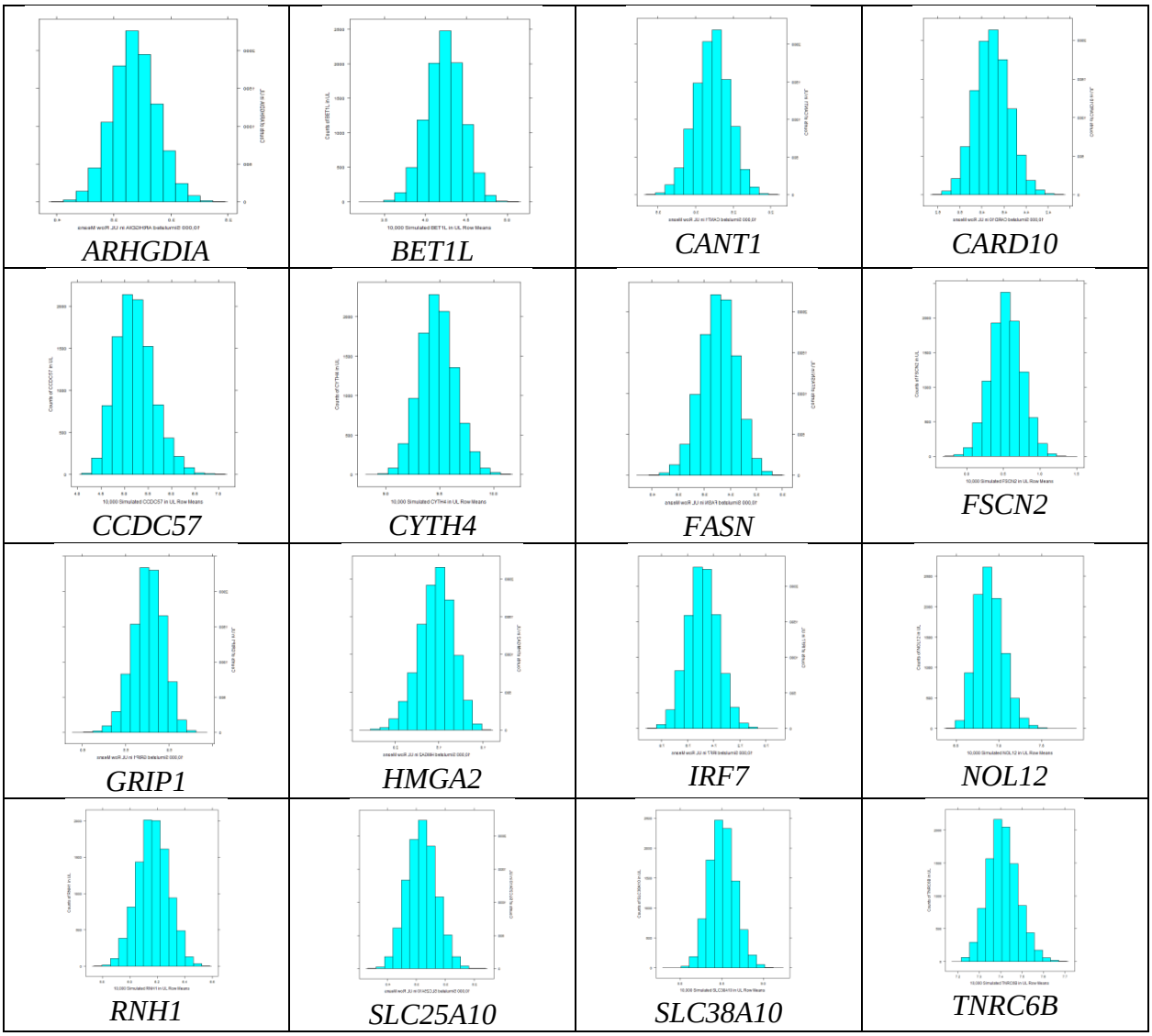

Figure 8. Histogram of TOP16 Simulated Means of 10K Samplings per Gene

The results from the machine learning algorithms on the TOP16 genes produced good results of 88 per cent to 91 per cent accuracy when a model was built using 70% or 85 samples of UL and non-UL to train the algorithm on testing the remaining 30% or 35 samples of UL and non-UL. The LDA and RF models performed the best. The results for the predictions are in Table 4. The ‘type’ field is the actual value the sample in the testing set should be, and the other three fields ‘predRF,’ ‘predGBM,’ and ‘predlda’ are for the RF, GBM, and LDA algorithms which scored per cents of 71,71, and 74 respectively.

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

	predRF	predGbm	predlda	type
1	UL	UL	nonUL	nonUL
2	nonUL	nonUL	nonUL	nonUL
3	nonUL	nonUL	nonUL	nonUL
4	nonUL	nonUL	nonUL	nonUL
5	nonUL	nonUL	nonUL	nonUL
6	nonUL	nonUL	nonUL	nonUL
7	nonUL	nonUL	nonUL	nonUL
8	nonUL	nonUL	nonUL	nonUL
9	nonUL	nonUL	nonUL	nonUL
10	UL	UL	nonUL	nonUL
11	UL	UL	nonUL	nonUL
12	UL	nonUL	nonUL	nonUL
13	UL	UL	nonUL	nonUL
14	nonUL	nonUL	nonUL	nonUL
15	nonUL	nonUL	nonUL	nonUL
16	nonUL	nonUL	nonUL	UL
17	UL	UL	UL	UL
18	UL	UL	UL	UL
19	UL	UL	UL	UL
20	UL	UL	UL	UL
21	UL	UL	UL	UL
22	nonUL	UL	UL	UL
23	UL	UL	UL	UL
24	UL	UL	UL	UL
25	UL	UL	UL	UL
26	UL	nonUL	nonUL	UL
27	UL	nonUL	nonUL	UL
28	UL	UL	UL	UL
29	nonUL	UL	nonUL	UL
30	nonUL	nonUL	UL	UL
31	nonUL	nonUL	nonUL	UL
32	UL	UL	nonUL	UL
33	nonUL	nonUL	nonUL	UL
34	UL	nonUL	nonUL	UL
35	UL	UL	nonUL	UL

Using the same model algorithms of LDA, RF, and GBM, a combined model of the three is compared using the gam method. Table 5 shows the combined model using the ‘gam’ method producing accuracy of 80 per cent.

Table5. Table of the Combined Model Confusion Matrix Results

```
Confusion Matrix and Statistics

          Reference
Prediction nonUL UL
     nonUL    14  6
     UL        1 14

               Accuracy : 0.8
                 95% CI : (0.6306, 0.9156)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : 0.003999

                  Kappa : 0.608

 Mcnemar's Test P-Value : 0.130570

            Sensitivity : 0.9333
            Specificity : 0.7000
         Pos Pred Value : 0.7000
         Neg Pred Value : 0.9333
             Prevalence : 0.4286
         Detection Rate : 0.4000
   Detection Prevalence : 0.5714
      Balanced Accuracy : 0.8167

       'Positive' Class : nonUL
```

Using the package for random forests outside of the R caret package method, the random forest package in R performed worse than all three algorithms just tested. PCA and Random Forest scored 64-65 per cent, while KNN was not able to be used to test this data with. When this model was used with exact predictors on a second training set the accuracy for the two top performers was 91 per cent for LDA and 88 per cent for RF.

This model was then tested on the actual top 16 highest under or over expressed genes in UL compared to non-UL and scored similarly to above but not as well. The LDA model scored 74 per cent compared to 91 per cent, and the RF model scored 86 per cent compared to 88 per cent.

To see if the models would do better with the lowest magnitude of change genes, the table was filtered to take the last 16 genes of the table sorted by magnitude in UL change in means from non-UL means of each gene, and tested the same as the above algorithms. The LDA and RF methods of the caret package in R gave accuracy results of 41 percent for LDA and 47 per cent for RF.

CONCLUSIONS

The findings were able to predict with up to 91% accuracy using the Latent Dirichlet Allocation (LDA) algorithm, and 88% using the Random Forest (RF) algorithm on the top 10 most expressed genes and the six genes ubiquitous to the current research studies on UL risk. Interestingly, when comparing the 16 least expressed genes these algorithms scored around 40% accuracy, but when run a second time on the least expressed genes using different testing samples but the same training samples the accuracy went up to 66% for LDA and 75% for RF. This could be because the sample from the training set were mixed in with the testing set for better accuracy the second run using the first run of predictor values, instead of a second run of predictor values.

This shows that the six genes ubiquitous to the current literature on UL risk and the topmost under or over expressed genes in UL compared to non-UL samples can be used to associate risk of UL by examining microarray expression levels of these genes. When looking at Figure 8 of the histograms of simulated means for TOP16, some genes had better symmetry than others, like *IRF7, FSCN2, RNH1, GRIP1, BET1L, CCDC57, and TNRC6B*. These genes could be explored more to see if they are correlated and have been explored more in the Additional Questions Answered section of the Appendix using six of the TOP16 genes having the lowest standard errors in the 10,000-bootstrap sampling with replacement simulations. The genes that are in the majority group for *ARGHDIA, TNRC6B,* and *GRIP* show that the first are expressed more while the last is expressed less in UL. The genes expressed in the non-majority are *RNH1* with more expression in UL and less expression in UL from *IRF7*, *FSCN2*, and *CCDC57*. This

could be explored further. Limitations for this study, are that the SNPs were not included in all combined data sets from GEO to run analytics based on fold change and MAF threshold values as the studies have.

Some additional analysis was done after this work that could answer some additional questions. The fold change of the ratio of UL gene expression to non-UL gene expression was a new field added to the data that was used to see if these same algorithms on the TOP16 genes would give better or similar results to using the 16 genes with the highest fold change from non-UL gene expression to UL gene expression, without considering the chromosomal location of the six genes ubiquitous to current UL risk meta-analysis studies. The results excluded all six and all TOP16 genes that were used in this research and showed better prediction accuracy. Additional questions answered can be shown in the link to the R-markdown pdf file in the Appendix section.

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.

Karatzoglou, A., Smola, A., Hornik, K., Maniscalco, M., & Teo, C. (2018, August). Kernel-Based machine learning lab. 'kernlab,' version: 0.9-27. Retrieved July 2019 from https://cran.r-project.org/web/packages/kernlab/kernlab.pdf

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

Data Sets, Files, and Images in this Research Study

Data tables and files can be obtained from https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research. The R script and location of the original files the script used to create the csv files of data tables and plots are in this folder too. The five GEO series and corresponding GEO platforms located in this folder were used to run the script Text_2_CSV_R_script_manipulations_analysis.R in RStudio to get all the tables used for analysis. The GEO files can also be obtained from GEO at the following links using the SOFT file download options in the GEO site:

- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68295
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13319
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE593
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23112
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2724
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL96
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6480

The file of all the merged data that was filtered for unique genes, on all chromosomes and the added means per gene is DE_means_Per_Gene_Chr.csv. The separate meta information the GEO platform GPL6480 provided is the file GSE_array_meta.csv.

All header column names beginning with ‘GSM’ are the GEO series sample, and if the field begins with ‘GSM’ and ends with ‘UL’ it is a UL sample. Otherwise, the ‘GSM’ column is a non-UL sample. This ‘UL’ extension to the column field name was added to the sample name in the R script, Text_2_CSV_R_script_manipulations_analysis.R. The all_common_12173_130_fold_magnitude.csv file has the universe of 12,173 common and unique genes between all five GEO studies and 129 columns with the row names the HGNC

gene name and the column names of 121 samples as columns of GSM IDs and ending in UL if the sample is a UL sample, and nine fields that are:

- GENE: this is a field for the NCBI gene ID located as column 1 in the table
- CYTOBAND:this is the cytoband location along the human genome where as an example, ‘hs|12p13.31’ is ‘hs’ for ‘homo-sapien’ or human genome, ‘12’ is the chromosome of this observation, and ‘p13.31’ is the chromosome band the gene is located in the chromosome.
- GENE_SYMBOL: this is HGNC approved gene symbol. All data sets are filtered and combined using this field. More information is retrievable at https://www.genenames.org.
- Counts: this field is what generated the unique genes to shrink down the 1,954,853 observational set of duplicates to unique values per gene. The values for each gene under each sample is the mean from all the counts of that gene per sample. This is because when merging the GPL6480 data set included many sequence values for each gene and the other four data sets didn’t have a sequence field. This resulted in many observations repeated to fill in the sequence fields for those genes in other sets when combined. The mean of each sample per gene is the same value. For example, if you have a ‘Counts’ value of ‘3’ for sample GSM167144 the value is 196.4667 for GENE_SYMBOL of ABCA1 with a GENE name of ’19,’ this means there were 3 duplicates in the larger set that the mean value of 196.4667 times 3 entries is divided by 3 entries. No information is lost.

UL_mean: The sample means for UL samples per gene across 70 samples of UL.

nonUL_mean: The sample means for non-UL samples per gene across 51 samples of non-UL.

DE: The difference in means from the UL mean per gene to the non-UL mean per gene

Magnitude: The absolute value of the change seen in up or down values of gene expression when comparing each gene in UL to non-UL sample mean values.

foldchange: The ratio of the UL_mean to nonUL_mean per gene. A value of ‘2’ means the gene had gene expression values that doubled in UL samples compared to non-UL samples.

Data Set of All Added Field for 130 Common Genes

The data set to get the meta field, fields for using Gviz, the 130 genes in common between all five GEO data series of 121 UL and non-UL samples and along the cytoband locations as the six ubiquitous genes, the fold change, means, difference in means, and magnitude fields are all in this data set. The data set can be retrieved from the same folder as all the other data on this research under the name, fold_magnitude_member_gviz_130_143.csv. The link, https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research/blob/master/fold_magnitude_member_gviz_130_143.csv, will have this data set for download all the time.

The following is the field identifiers for this data set, the GSM fields are samples of UL or non-UL and if it ends with 'ul' it is a UL sample. There are 22 meta fields or non-numeric fields to describe the genes of the samples:

1. genes: this is the HGNC gene ID
2. chromosome: this is the chromosome the gene is located
3. type: this is if this gene is 'up' (more) or 'down' (less) expressed in UL compared to non-UL samples
4. all: this tells how many genes including this gene are on this chromosome from the 130
5. up: this tells how many of those gene counts in 'all' are 'up' or expressed more in UL
6. down: this tells how many of those gene counts in 'all' are 'down' or expressed less
7. majority: this tells if this gene is part of the majority of genes expressed more or less on that chromosome
8. start: this is the ENSEMBL value for what million base pair start site this gene starts on that chromosome
9. end: this is the ENSEMBL value for what million base pair end site this gene ends on that chromosome
10. width: this is the ENSEMBL width from start to end of this gene on that chromosome
11. strand: this will be '+' if on forward strand and point right, and '-' on reverse strand pointing left
12. gene: this is the ENSEMBL gene ID
13. transcript: this is the ENSEMBL transcript ID for this gene
14. GENE: this is the NCBI gene ID
15. GENE_NAME: this is the full name of the HGNC ID
16. CYTOBAND: this is the cytoband location for this gene

17. DESCRIPTION: this describes the gene with its HGNC name and if it is a mRNA, transc ript variant, protein, and other gene functions
18. nonUL_Mean: this is a calculated field made by taking the row means of each gene belonging to the non-UL samples for this gene
19. UL_Mean: this is a calculated field made by taking the row means of each gene belonging to the subset of UL samples for this gene
20. Difference_UL_minus_non_means: this is the difference between the UL_Mean and the nonUL_Mean for each gene across all samples
21. Magnitude: this is the absolute value of the Difference_UL_minus_non_means or magnitude of change each gene had in the UL_Mean minus the nonUL_Mean
22. foldchange: this is the fold change that each gene had in the ratio of UL_Mean to nonUL_Mean

There are 130 genes, 51 non-UL samples, 70 UL samples, and fields for added additional information. When running predictive analytics, exclude the non-numeric fields once you filter for which ever subset observing. The two data sets ready for machine learning and used in this research study involved the top 10 genes having the highest magnitude of change plus the six genes ubiquitous to current UL risk meta-analysis studies and separately as the top 10 genes having the highest fold change in the ratio of UL means to non-UL means plus those same six ubiquitous genes in UL risk studies. The first file is named TOP16_ml_ready.csv and the second file is named FOLD16_ml_ready.csv. Both files are transposes of the larger data frame with row names as the genes and the header as only samples and one categorical field ‘TYPE’ that is used for predicting whether the testing set of 30 per cent of the 121 samples left out during training is a UL or non-UL sample. The script to run the code and create all data sets used from the original files is in this same file folder https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research/blob/master/ as All_analysis.R.

Additional Questions Answered

The pdf file of the R markdown version of the R code with results showing some additional questions answered after this research can be obtained from https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research/blob/master/SomeQuestionsAnswered.pdf . Here are the questions asked by the researcher after this research was concluded:

1. How well do the lda, rf, gbm, combined models, and rf2 predict ul or non-UL on the entire 12k genes in common and having the highest magnitude of change in UL compared to non-UL samples?
2. What about using the best indicator fields from the bootstrap simulations showing the most symmetrical genes among the TOP16 of genes only on cytobands of the six ubiquitous genes, and those six genes? How well do these models accurately predict UL samples? Use the genes with the lowest sd.
3. What about those genes part of the minority of genes that are up/down expressed? Are these good indicators of UL or non-UL?
4. What about those genes part of the majority of genes that are up/down expressed? Are those good indicators of UL or non-UL?

The tables produced rely on the table in the github link from which to run analysis on. This is the table that was made from all of the chromosomes, unique genes, and dropped duplicates from the five GEO series of 121 samples of 70 UL and 51 non-UL tissue samples. This table is located as a csv file, DE_means_Per_Gene_Chr.csv, in same folder as the R-markdown document that answers these questions. The field values are in the Appendix section, 'Data Sets, Files, and Images in this Research Study.' The R-markdown script to run in RStudio

is SomeQuestionsAnswered.Rmd and printed with results as a pdf file as SomeQuestionsAnswered.pdf from https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research.