Rough Draft-pdf version-Janis Corona

by Janis Corona

Submission date: 07-Jul-2019 10:18AM (UTC-0500)
Submission ID: 1149798079
File name: RoughDraft.pdf (533.46K)
Word count: 6949
Character count: 35034

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

10

LIST OF TABLES

11

12

LIST OF FIGURES

16

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
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

17

18

INTRODUCTION

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, Hartmann, & Edwards, 2013; Liu et al., 2018; Hellwege et al., 2017; Rafnar et al., 2018; Cha et al., 2011; Aissani, Zhang, & Weiner, 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

19

Sub-Verb Agree

Verb tense

Verb tense

Fragment

Verb tense

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).

Description of UL

Many 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 et al. (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 control groups used in this study (Aissani et al., 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, *SCDC57* 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).

mosome 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, Wang, & Wiener, 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

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).

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*. 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 ENSEMBL 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 (H[illegible]n, 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 130 genes unique and common to each of the five GEO microarray series of UL and non-UL tissue samples was generated and 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.

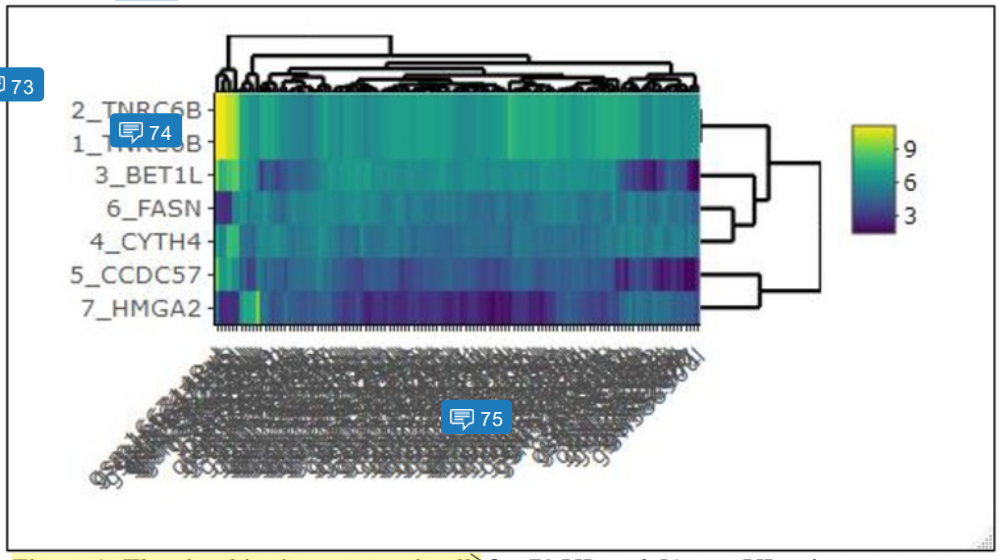

Figure 1: The six ubiquitous genes in all for 70 UL and 51 non-UL microarray gene expression samples are displayed in a heatmap produced in R using the heatmaply package.

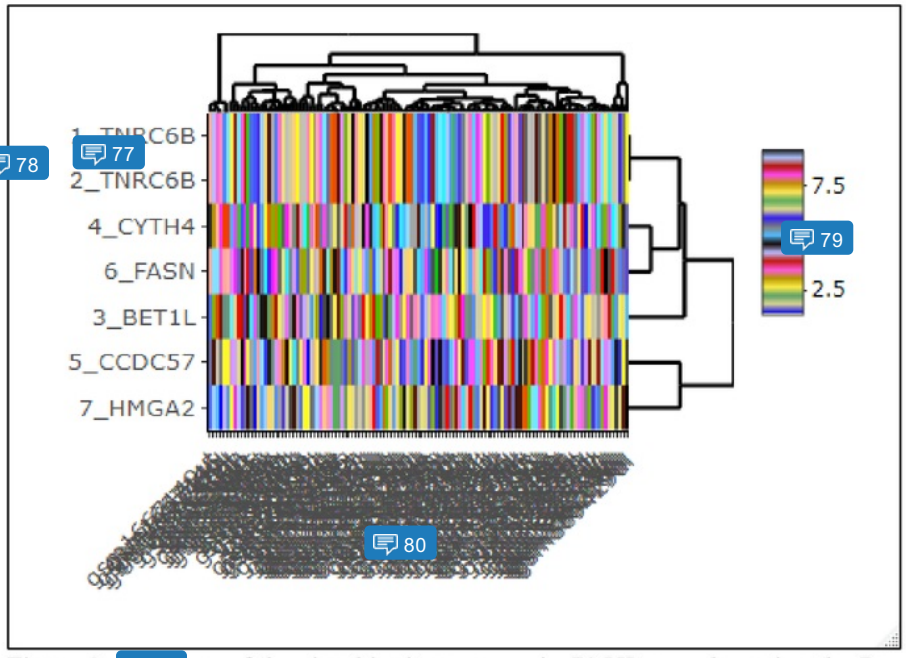

Figure 2: Heatmap of the six ubiquitous genes in 70 UL samples using the R package heatmaply.

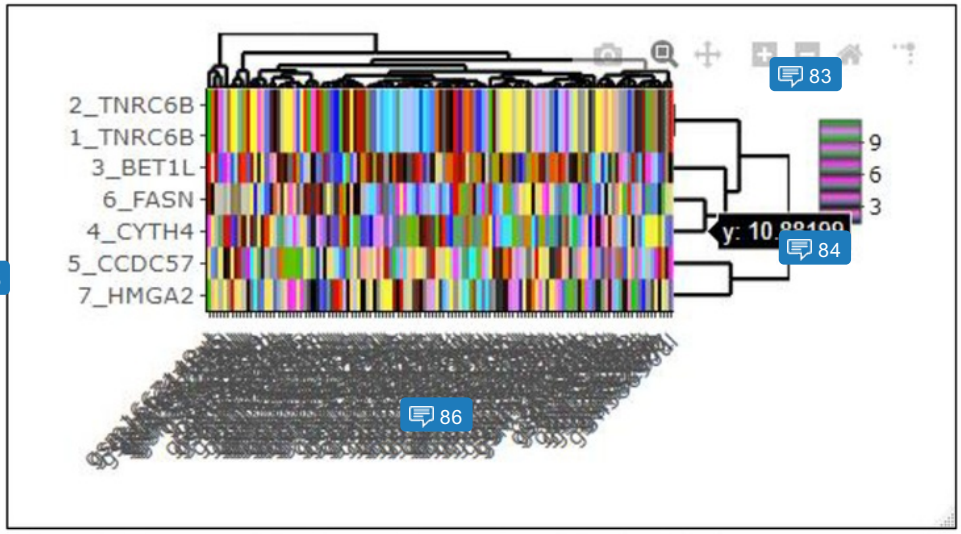

Figure 3. Heatmap of the 51 non-UL samples of the six ubiquitous genes using the R heatmaply package.

The heatmaps in Figure 1, Figure 2, and Figure 3 didn't show any useful information to pursue based on the quick plotted hea[89] A table made to calculate the mean of each of the 130 genes among all 121 samples for UL and non-UL separately[90] 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[91] 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[92] There were 70 genes expressed more, exactly 60 genes expressed less in UL, and one gene, *KDELR3*, expressed the[93] Gviz was used to map out these genes on their[94] chromosomal bands, but not with much useful info and some overlap.[95] Then, each chromosome of the four 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. 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, according to Table 1.[96][WC][97]

Table 1. This table shows the six ubiquitous genes and if it is expressed more as 'Up' or less as 'Down' and if it is part of the majority of genes on that chromosome that is expressed more or less in UL.

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 using the Gviz R package.

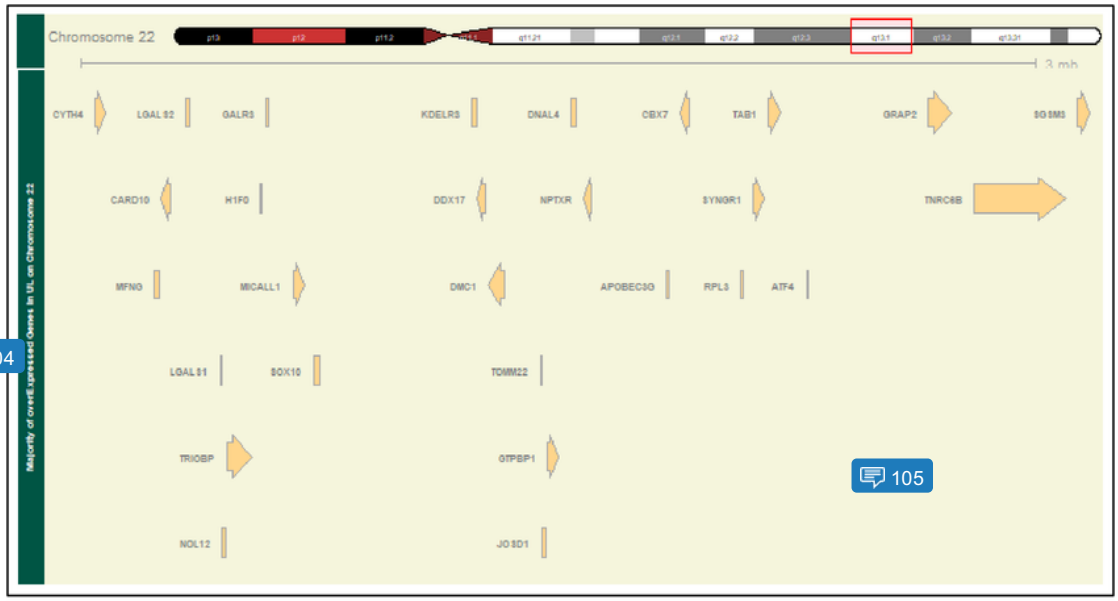

Figure 4. Chromosome 22 Gviz map of the genes that are expressed more in UL, this shows cytoband, direction, and chromosome. *TNRC6B* and *CYTH4* are both part of the majority of genes expressed more in UL.

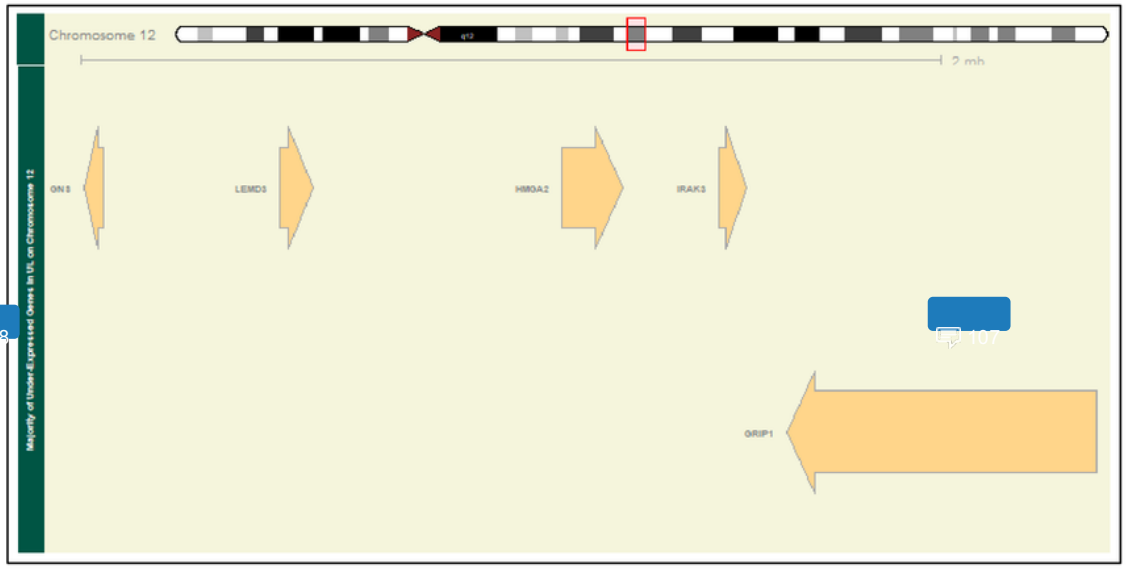

Figure 5. Chromosome 12 Gviz map of the genes that are a part of the majority of genes expressed less in UL. *HMGA2* is a part of this group majority.

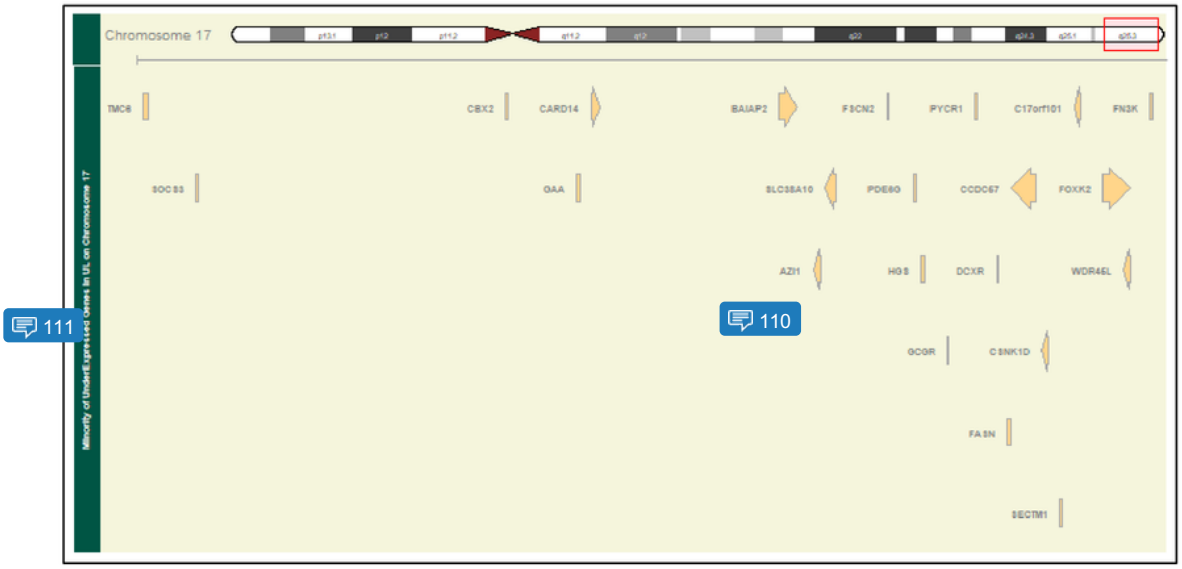

Figure 6. Minority of genes expressed less in UL on chromosome 17 using the R package Gviz. *CCDC57* and *FASN* are part of this minority of genes expressed less on chromosome 17 when the majority of genes in UL are expressed more on chromosome 17.

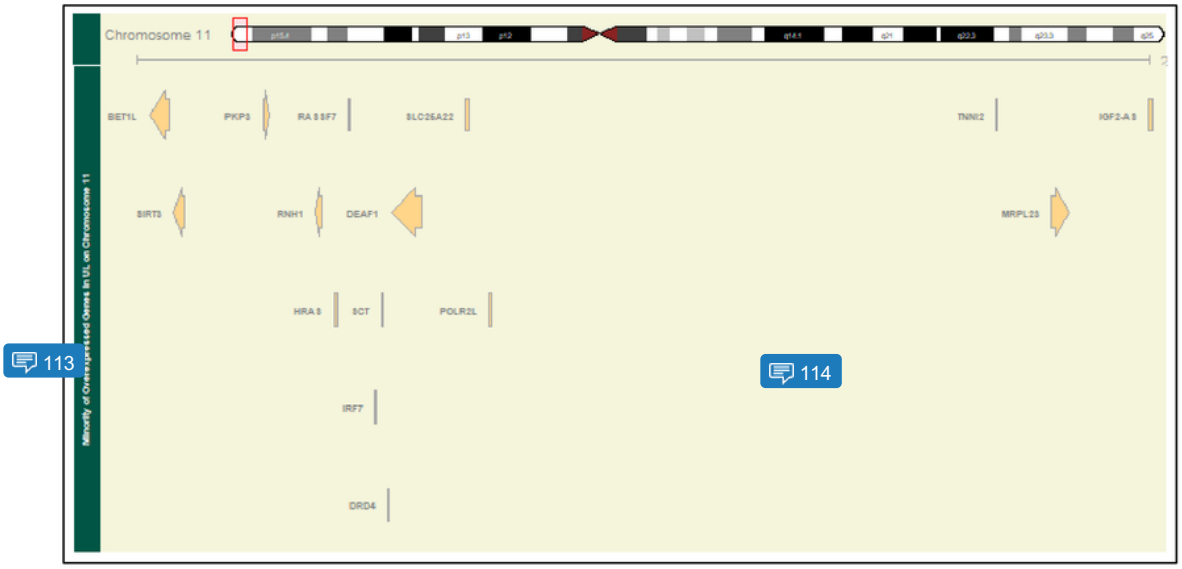

Figure 7. Minority of genes expressed more in UL on chromosome 11 using the R package Gviz. *BET1L* is part of this group of genes expressed more when the majority is expressed less on chromosome 11.

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.

Table 2. [redacted] 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.

	simulatedMean10k	simulatedSD10K	leftTail2.5	rightTail97.25	ulStatus	gene
1	[redacted]3693	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	[redacted]3228	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. 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 mean generated in Table 2 for UL means.

Table 3. The TOP16 genes and their descriptive names. These are the 10 most expressed or under expressed genes in UL compared to non-UL samples per gene plus the six genes ubiquitous to UL risk studies.

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

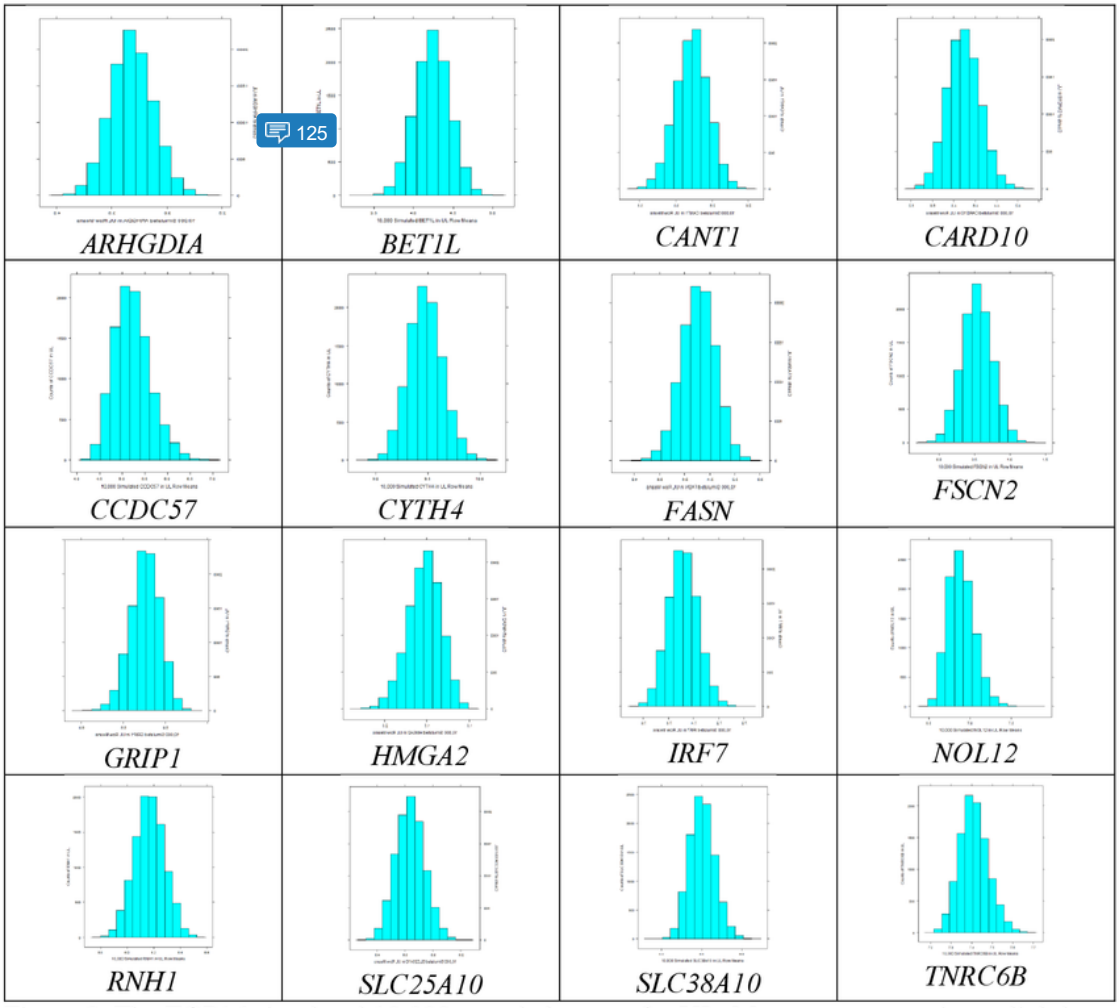

Figure 8. The TOP16 simulated 10,000 sampled means for UL. Most are symmetric or slightly skewed with the median value to the right or left which direction the skew points.

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 126 LDA and RF models performed the best. The results for the predictions are in Table 4. 127 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. In the combined model using the 'gam' method produced accuracy of 80 per cent as can be seen in Figure 9.

Table 4. The predicted outcomes for each model algorithm on the TOP16 genes. Row by row comparisons can be seen. The true value of the testing set is under the 'type' field. The highest scoring algorithm for this data table was the LDA model with 74 per cent.

	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

```
confusionMatrix(CombinedPredictions, testingSet$TYPE)
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
```

Figure 9. The combined model of all three previous algorithms scored 80% accuracy in predicting the sample as either UL or non-UL.

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. 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.

141

LITERATURE CITED

1
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

12
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. 10 Retrieved July 2019 from https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

2
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).

142

4
Chang, J. (2015, November). Collapsed gibbs sampling methods for topic models. 'lda,' version: 1.4.2. Retrieved July 2019 from 22 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=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

149

Del.

Rough Draft-pdf version-Janis Corona

ORIGINALITY REPORT

37%	16%	14%	36%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Lewis University Student Paper	30%
2	helda.helsinki.fi Internet Source	1%
3	journals.plos.org Internet Source	<1%
4	Submitted to College of Eastern Utah Student Paper	<1%
5	d-nb.info Internet Source	<1%
6	Coyle, Jessica R., and Allen H. Hurlbert. "Environmental optimality, not heterogeneity, drives regional and local species richness in lichen epiphytes : Regional and local lichen species richness", Global Ecology and Biogeography, 2016. Publication	<1%
7	Submitted to University of Sheffield Student Paper	<1%

8	Submitted to Australian Catholic University Student Paper	<1%
9	espace.library.uq.edu.au Internet Source	<1%
10	www.tandfonline.com Internet Source	<1%
11	library.wur.nl Internet Source	<1%
12	res.mdpi.com Internet Source	<1%
13	scholarworks.gsu.edu Internet Source	<1%
14	Submitted to National College of Ireland Student Paper	<1%
15	support.bioconductor.org Internet Source	<1%
16	Jing Shen. "Genome-wide DNA methylation profiles in hepatocellular carcinoma", Hepatology, 2012 Publication	<1%
17	Submitted to University of Sunderland Student Paper	<1%
18	ddd.uab.cat Internet Source	<1%

19	Submitted to University of Pretoria Student Paper	<1%
20	E. W. M. McDermott. "Prognostic variables in patients with gastrointestinal carcinoid tumours", British Journal of Surgery, 07/1994 Publication	<1%
21	www.11ssslisbon.pt Internet Source	<1%
22	www.onepetro.org Internet Source	<1%
23	Submitted to University of Dundee Student Paper	<1%
24	bmcmedresmethodol.biomedcentral.com Internet Source	<1%
25	journals.sagepub.com Internet Source	<1%
26	Submitted to Massey University Student Paper	<1%
27	www.dtic.mil Internet Source	<1%
28	link.springer.com Internet Source	<1%
29	drum.lib.umd.edu Internet Source	<1%

30	Submitted to Universitaet Hamburg Student Paper	<1%
31	Liddicoat, C, D Maschmedt, D Kidd, and R Searle. "Modelling soil carbon stocks using legacy site data, in the Mid North region of South Australia", GlobalSoilMap, 2014. Publication	<1%
32	Große-Stoltenberg, André, Christine Hellmann, Christiane Werner, Jens Oldeland, and Jan Thiele. "Evaluation of Continuous VNIR-SWIR Spectra versus Narrowband Hyperspectral Indices to Discriminate the Invasive Acacia longifolia within a Mediterranean Dune Ecosystem", Remote Sensing, 2016. Publication	<1%
33	Submitted to National University of Singapore Student Paper	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Rough Draft-pdf version-Janis Corona

GRADEMARK REPORT

FINAL GRADE

29/45

GENERAL COMMENTS

Instructor

Janis - I have left extensive comments throughout your document to try to point you in the right direction with revisions. Some of my ability to guide you was limited, though, as the document did not provide access to the data. With holes in the explanation given in the methods section, I made my best effort to understand what you were working with, but I made assumptions along the way. Within the Results section, there are analyses that were performed but it is not clear why you decided to complete that work. As I've attempted to address in other email correspondence, it is important to make sure that the work you do with the data yields outcomes that are interpretable in the biological context. For instance, why it may be interesting that there are certain genes that are localized on the same chromosome, why does that spatial location matter? Isn't it just as likely that key changes in expression (key to changing a key into the UL category) could all be located on different chromosomes? What is the utility in mapping the location of genes? How do these analyses reflect what other researchers have previously performed? You will need to make some substantial revisions to your work -- even if you stay with the analyses done so far (and do not complete any others to address the biology more thoroughly), you will need to provide documentation to solidify why you are performing these analyses (why are they in fact appropriate)?

I'm happy to speak with you -- sometimes a phone call can help clear things up. Please let me know if you would like to take me up on this offer.

Some essential items to work on within your revision:

- Make sure that, for any sentence about your research, that verb/verb phrase should be written in past tense. I've attempted to mark instances of verbs that need to be revised, but check that there are not other instances I have skipped flagging.

Check organization. Especially within the Introduction section, you want to provide information more generally, and then add on more specific details. That way, readers with limited content knowledge will have an easier time understanding your writing. In contrast, if you begin with more specific concepts without presenting defining elements, readers not well-versed in the subject will struggle (until, perhaps, they get to the more general information, and then they will need to double back and re-read). This means it will be important to give background information before you pose your research question.

- At times, similar/related information tends to be split into different sentences distributed throughout a section of writing. Always make sure that related concepts are grouped together (if not within the same sentence, then in adjacent sentences) so that the reader can learn all related concepts in clusters, rather than feeling like concepts "bounce" back and forth.

- When writing about information you've been thinking about for quite awhile, a common pitfall is to assume a certain level of pre-existing knowledge and so the information that makes it onto the page is only a partial representation of the knowledge you want to reference. With the context you already know, it might be possible to have the writing make sense (because you are filling in the gaps without realizing it), but that doesn't work so well for readers that are not already intimately aware with the topic. Thus, it is extremely important to fully convey all necessary and relevant information in your writing, without assuming that concepts are implied or that your reader will know what you're conveying and be able to fill in the gaps.

- Throughout your Introduction section, you have identified some genes of interest, but have not consistently explained by they are deemed interesting. It seems that some have been identified because there are gene variants (either different alleles or SNPs/haplotypes) that are associated with disease. In other cases, there are references to changes in the level of expression of these genes. These two concepts are not interchangeable -- although there may be examples where one gene variant leads to disease because it is also changed in expression, there are also cases where a gene may lead to different cellular function because of a change in coding (but not because of a change in

expression) OR because of a change in expression (but without any change in sequence).

- Within the Methods section, it is essential that you start with the original data, describe the layout and components included within that set, and then step by step explain the manipulations you performed to derive your working set. Right now, the information provided is too vague and could not be used to accurately replicate your work.

- In the Results section, each figure should have an included narration about what the reader should learn from the information as presented. I have included a comment along with the "figure 1" highlight on page 17 in the document -- use this as a model to write your own explanation for this figure, and also create similar text explanation for each of your other figures.

- Each analysis within the Results section must be fully explained, and outcomes clearly articulated. It is important that the analyses you completed have biological interpretations. Thus, for each of your outcomes, you need to make clear how they are able to address your overall research objective.

- I am concerned about your Conclusion. First, make sure that you re-state your research objective. You will then need to systematically review the outcomes presented in the results section, thoroughly placing your observations in the context of the literature. It is ok if some of the analyses did not perform as well as desired, but you need to always justify why that analysis was an appropriate attempt to address the biological objective. There are several instances currently where you have performed some work with your data, but it is not clear why. What is the biological interpretation in all instances. Then, as you think about limitations in your analysis, make sure that they are in terms of the work you set out to complete. You should identify (based on the type of data you worked with) reasons why analyzing gene expression data may not give the outcomes expected -- it is not appropriate to point to not doing another type of analysis (such as looking at SNPs, as that was not the focus of your project). Ultimately, what are your recommendations for moving forward?

PAGE 1

Comment 1

Please remove the running head from your document.

PAGE 2

Comment 2

The top margin of this page should be set to 1 inch.

QM

Verb tense

Additional Comment

Remember, in all sentences written about your research/work/study, you should write in past tense.

Comment 3

Check word choice here -- you are pulling together data from five studies, right? As written, this could be interpreted as data from only five individuals (whereas the data set you created contains many more individuals, giving you more power in your analysis).

Comment 4

This is as compared to non-UL tissues, correct? You need to make clear exactly what you have done -- as written, this suggests that you only looked at samples from UL but not with any comparison to non-UL.

QM

Verb tense

QM

Verb tense

Comment 5

10

QM

Del.

Delete

Additional Comment

This is a concept/term that you coined? Keywords are typically phrases or words that are used in high repetition across may sources.

Comment 6

Please use lowercase for the roman numeral numbering of these leading pages.

PAGE 3

Comment 7

The top margin on this page should be set to 1 inch.

Comment 8

Make sure that this is correctly iv (rather than 1V) -- all characters should be lowercase for these roman numerals.

Comment 9

Currently, no appendix is included. Please plan on using an appendix to share access to your newly created data sets.

PAGE 4

Comment 10

The top margin for this page should be set to 1 inch.

Comment 11

Please consider editing this title, as it is not clear what you are intending to convey here. It is expected that each chromosome has some genetic information expressed, so there should not be a differential of expressed versus non-expressed chromosomes.

Comment 12

Tables should be numbered based on the page order of their appearance. Thus, table 1 should not appear after table 2 (as this numbering would suggest).

PAGE 5

Comment 13

The top margin of this page should be set to 1 inch.

Comment 14

Each figure should be placed on its own individual page -- therefore, figure 2 and 3 should not both be listed as appearing on page 12. On the previous page, you have also indicated that a table is located on page 12. This will all need to be updated in the final paper.

Comment 15

In your introduction section, you gave background information on the chromosomes in the order of their numbers (11, 12, 17, 22). What is the rationale for presenting the order as listed here (22, 12, 17, 11)? If there is a rationale for providing the data in this order (say there are more important genetic influencers on chromosome 22 relative to all others), then it might make sense to preserve this order (and also considering reordering how you provide information in the introduction section). If there is no particular reason for this ordering of data, then please consider reorganizing into numerical order.

PAGE 6

Comment 16

There should be a 1 inch top margin on this page.

Comment 17

Be mindful of how you use this abbreviation -- in the abstract, you define the plural (uterine leiomyomas) as UL. Either you should make the abbreviation plural for this term (ULs) or be aware that each time you use UL, you are actually referring to tumors (plural). Thus -- you should not write "a UL" as you are using a singular article with a plural term.

PAGE 7

Comment 18

Please note the expected format for page numbers:

- On the first page of a chapter (such as here), the number should be placed bottom center on the page.

- On second and subsequent pages within a chapter, please place the number in the top right corner.

Comment 19

You likely should not assume that all readers will understand what this means -- I suggest that you include an explanation as part of your Introduction.

QM

Sub-Verb Agree

Subject-verb agreement

QM

Verb tense

QM

Verb tense

QM

Fragment

Fragment

QM

Verb tense

PAGE 8

QM

Verb tense

QM

Verb tense

Comment 20

Based on what? Expression level patterns? You should clearly identify the focus of the analysis here.

QM

Del.

Delete

Additional Comment

See note on abbreviation page about singular/plural for this term.

Comment 21

Check your organization in this paragraph. You first list some risk factors, then give some treatment information, then provide more information about risk. It would be stronger to reorganize the most similar content together in sequential sentences, and perhaps subdivide into two different paragraphs. For example, you might want to have the sentence with "age at menarche" followed by the sentence at the end of this page including "thyroid dysregulation" as well as the genetic content at the top of the next page together as one paragraph. You could then have a second paragraph to provide information relative to estrogen -- the first sentence could be the one that starts with "overweight females" then the one starting with "Because estrogen" and then follow up with the "treatment involving an estrogen antagonist".

PAGE 9

Comment 22

What about BET1L is associated with these tumor locations? Is the observation that there is a particular allele or SNP variants for this gene? Or is the association having to do with how much of this gene is expressed? You need to clarify what about this gene is interesting, as presumably the gene is present in all human genomes -- the variable part amongst humans is not made clear here.

Comment 23

Be consistent with how you report gene names -- in the previous paragraph you only gave the shortened form, and in the next paragraph you gave both the full length name and the shortened form. You could go with either approach, but whichever you decide on should be used for introducing ALL genes and not different from one case to another.

QM

Italicize

Additional Comment

The full length gene name should also be italicized.

QM

Italicize

Additional Comment The full length gene name should also be italicized.

Comment 24

What kind of difference? Since you mentioned GWAS, I am guessing that it is SNP/allele differences? Be clear.

Comment 25

Here, do you mean these same three genes? Or are there specific genotypes for these genes? So far, you have not included genotype information.

Comment 26

I do not think that enough context about diagnosis process, or perhaps the identification of impacted individuals used in the study, has been provided for this statement to fully make sense. Remember, your reader does not have intimate knowledge of all of the literature you reference here -- it is very important to make sure that your summary is inclusive of key content.

Comment 27

Reported to whom?

QM

Italicize

Comment 28

At abnormally low levels? Or maybe "under-expressed" would be appropriate to state?

Comment 29

Is this dependent on SNPs? Make sure to clearly identify the thing about this gene that is different in the populations you identified.

Comment 30

How so? Due to expression? Genetic differences (SNPs or alleles)?

Comment 31

What does this mean?

Comment 32

This information is not readily available to all readers. Can you convey the same general concept using less jargon? Make sure that the sentence also is obviously related to the previous sentence in the paragraph -- without a lot of specialized content knowledge to make an assumed connection, not all readers will understand the point that you are making here.

Comment 33

Please make a more clear description of how this was performed. I am assuming that the samples were of endometrium? After the cells were harvested, what type of assessment was performed? Was the focus on mRNA levels (gene expression)? If variation in gene expression were assessed, by microarray? RNA-seq? What samples were compared to look at differential expression?

Comment 34

What do you mean by "most significant"? It is difficult to know (if in fact four different phases were compared) what comparisons were made, and what level of change was determined as significant.

Comment 35

Which GWAS are you referencing? Which populations was this true for?

Comment 36

The focus was only on genes more highly expressed in UL than non-UL (but not samples downregulated in UL compared to non-UL)?

Comment 37

Once you have given the full length and shortened name of the gene once, you do not need to provide both again. Include these on your list of abbreviations.

Comment 38

What about the gene is associated? Expression or SNP?

Comment 39

You are missing an "on" in this phrase.

Comment 40

I suggest that you get in the habit of organizing these by increasing number of the associated chromosome -- that will make comparison of lists easier.

Comment 41

Reorder by value.

PAGE 12

Comment 42

You have previously mentioned LD, but have not explained the significance as a genetic principle. It seems that you will go on to devote a significant amount of your work to considering placement of genes close to each other on a chromosome. Therefore, it is important for you to explain here, as part of your Introduction why this is a factor worth studying. What is the implication (especially for inheritance)?

Comment 43

It is important to avoid repetition of information -- there is a lot of content about the key genes within the paragraphs on population differences that is repeated here in the paragraphs about the genes, divided by chromosome location. If organizationally you decide to provide information about a gene in these two places, then the content given in each location should really be distinct. If, however, you feel that the information provided in this section on chromosome location is largely redundant, you may need to revise your organization strategy such that all information about a gene is unified in one location.

Comment 44

Unless there is a ranked order that identifies BET1L as the most important gene (and genes on subsequent chromosomes as a decreasing order of significance), please reorder your descriptions of chromosomes numerically. Otherwise, your choice to report about chr11 first implies more importance than actually exists.

Comment 45

What about the gene is associated with UL?

PAGE 13

Comment 46

Begin paragraphs with tab indentation.

Comment 47

Begin paragraphs with tab indentation.

QM

Del.

Delete

Comment 48

Begin paragraphs with tab indentation.

Comment 49

Please revise the ordering within your Introduction section, such that you begin with more general information and then become more specific -- that is, you should provide background information first so that the reader has an idea of what was already known before you started your project, and then at the end of the chapter, you should make clear your research objective(s) and how you were able to address that. Most of the needed content is present in this chapter, it just needs to be reordered.

Comment 50

The first level subheading should be underlined.

Comment 51

Please provide a direction to exactly where you accessed these data sets, including the names as appropriate.

Comment 52

Please explain your process for bringing together all of this information -- were the sets imported into a uniform space within GEO before the set was exported, or were they exported individually and then combined? As currently provided, it is not possible to definitely replicate the work you performed.

Comment 53

It is important to maintain a chronological explanation. You focus on R usage in the next section. Rather than explaining in two places, in this section, give information that is only relevant to how the data were accessed in GEO (locations to find), any work done within GEO, and then how the files were exported. That should then end this section, and the next step in the procedure picked up in the next section.

Comment 54

This gets back to the need to an explanation of how the sets were all assembled together…how were these managed such that the IDs were uniform with the others?

QM

Del.

Delete

Additional Comment

Make sure to focus on what was done -- with any study, there will usually be other options that could have been used but trying to be fully exhaustive of these options is not possible.

Comment 55

Format

Comment 56

This is the total number of genes. What other description of the data set can be provided? How many samples were included? What other categories of information (such as UL or non-UL, age, race) were included?

Comment 57

As you describe the data to start, make clear what information was present in total (without focusing on specific genes). If you subsequently then filter the data to keep some genes and remove others, you can then detail that in the next step.

Comment 58

This is an aspect of why it is important to first describe what information is present in the data set. How was it possible to retain genes on certain chromosomes? Did you perform a filter focused on these specific genes somehow, or did you elected to retain genes on Chr. 11, 12, 17, and 22? If that's the process you used, then describe that way (otherwise you need to describe how you were able to associate other genes specifically with these six). You can remind the reader of the location of the six genes of interest within the Results section.

Comment 59

Why were there some duplicates?

Comment 60

Do you mean Ensembl?

Also, please be more clear about the process that was used here -- what build of Ensembl was used? What information from the GEO set was used to coordinate with Ensembl, and what information was then built into the new data set?

Comment 61

Why was this the case?

Comment 62

Because you did not explain the organization of the original data, it is not clear how this information was initially presented, and then subsequently available due to your work. Please explain.

Comment 63

How was this accomplished? Doesn't this require that each UL sample was matched to a non-UL sample? You need to first describe the status of the data set to begin with, and then fully explain what manipulation was performed.

Comment 64

This seems to be the point at which you completed the generation of your working data set. At this point, please name the data set and then make reference to how access to it is provided via the appendix.

Comment 65

After the creation of the working data set, this seems to be where you began analysis to learn some information from the data. At the very least, a new paragraph should begin here. Alternatively, you may want to create subheadings that describe the steps of the process, rather than the software used. For instance, you could have one heading for "Establishment of the Working Data Set" and a second for "Determination of Differential Gene Expression" and later "Analysis of Location by Chromosomes"

QM

Fragment

Fragment

Comment 66

Previously you made mention of the top 10, but you have not clarified the TOP16 here in the methods. Were these included in a secondary (restricted) data set, or were you always working in the single larger set, focusing only on the TOP16? Please explain.

Comment 67

Check formatting here -- it is not clear why the line of text ends here and picks up at the top of the next page.

Comment 68

Formatting.

Comment 69

Was this all performed using the single large master data set, or was a specialized working data

set created? Please define, and if a secondary working set was created, make available via the appendix.

QM

Awk.

Awkward:
The expression or construction is cumbersome or difficult to read. Consider rewriting.

PAGE 17

Comment 70

Please clarify this statement. Remember, all humans should have the same fundamental genes (even though expression might be varied from one cell to another). Here, it would be good to first explain what information you began with. For example:

"Data was gathered from five independent GEO microarray data sets, each with expression levels normalized to facilitate comparisons. In total, there were X individuals representing UL and Y individuals representing non-UL. From the complete list of Z genes, the list was reduced to 130 genes, based on [insert criterion here -- I believe this was based on chromosome location?]. For each gene, the expression value was known in triplicate [is this true? based on your statement I believe you made earlier, this might be the case, or there might be a single expression value for each -- make sure this is clear]. Analysis was then performed to determine whether gene expression data could be used to stratify samples into UL and non-UL categories [am I correct that this was your intention?]."

QM

Del.

Delete
Additional Comment

The details should be present in the Methods section, but here in the Results you can rely on stating the type of analysis that was performed, rather than the software.

Comment 71

Each figure should have at least a robust paragraph included in the text to describe what is conveyed through the data panel. At the start of this paragraph, you made mention of 130 genes in total, but in figure 1, there is a focus on only six genes. You will need to make clear the process by which your analysis became focused on only these six. I am not sure that I fully understand what is shown in this first figure, but I will make an attempt at writing an appropriate example paragraph:

“The data set was streamlined to generate a heatmap in which expression levels of the six previously established genes associated with UL (TNRC6B, BET1L, FASN, CYTH4, CCDC57, HMGA2) were used to cluster different patient samples based on expression levels. In total, 70 UL and 51 non-UL samples were compared and are presented in Figure 1. [I’m not sure what the color gradient means, so I am going to make something up here – you will need to input the correct interpretation]. Genes with expression that is more highly expressed in UL than in non-UL are represented in yellow, whereas genes that are expressed less in UL than in non-UL are represented in dark blue. Each mark on the X-axis represents an individual patient. From this analysis, patients are grouped into five larger sub-categories [I made this assessment based on the brackets on the top of the chart] but do not cleanly segregate samples into categories of

UL and non-UL [this may not actually be true, but since I cannot read the labels at the bottom, I do not know]."

Comment 72

Make sure to provide descriptive text for this figure, similar to the model given for figure 1.

Comment 73

Each data panel must be placed on its own unique page in the document. Since figure 1 is first referenced in the text here on page 17, then the figure should appear on page 18. Don't forget, the accompanying text within the results chapter must be fully descriptive of what is included in the figure (it should not be reduced to "figure 1 shows X"). That way, there should be plenty of writing that fills out the remainder of the page. If there is still room on the page of text and a second figure is referenced, then figure 2 would be placed on the next new page in the document (in this case, on page 19).

Comment 74

Please make sure that the resolution within a figure is high -- this looks blurry.

Comment 75

All font included within a figure should be readable -- here, the labels for vertical items is so highly overlapping that it is not possible to understand the information provided. Either reduce the number of items used in the heatmap (to a level where each label can be clearly read) or remove the labels here and present those data in another format that can more easily be determined (such as by providing a table).

Comment 76

Here, this should read: "Figure 1. The same title as what was given on the list of figures page. Following the title, you may then provide additional information that aids the reader in interpreting the figure." Within the legend, it would be important to make clear the number of samples in each group. What is shown on the X axis? What is shown on the Y-axis? What do the different colors represent?

Comment 77

What is the significance of the numbers given here? What are these items? Why does it appear that the top two rows re repeats?

Comment 78

Each figure should be placed on its own unique page within the document.

Comment 79

It will be very important to explain these colors as you compose the associated paragraph of text.

Comment 80

See the comment given for figure 1.

Comment 81

Make sure that the figure title is exactly the same as you have reported in the list of figures at the start of the document.

Comment 82

As part of the legend, please describe what the different colors indicate. It is not clear the significance of the color scheme across figures.

Comment 83

As you capture the image to use for your figure, make sure that these types of "hovering" labels are not included.

Comment 84

As you capture the image to use for your figure, make sure that these types of "hovering" labels are not included.

Comment 85

Each figure should be placed on its own unique page in the document.

Comment 86

See previous comments about this section (non-readable information should be removed).

Comment 87

Improve the title of the figure, and make sure that it is consistent with the title you provide on your list of figures.

Comment 88

Within the legend, make sure to include information that explains how to interpret the different colors displayed.

Comment 89

Here, I think you have proposed an alternate approach. Am I correct in my understanding? I think you took the value for, say, BET1L from all UL samples and averaged it, and then the BET1L from all non-UL, and then this sample approach was completed for the remaining 129 genes? This may be a valid way to re-attempt the analysis, but it would be stronger if you can cite that another research study has used the same approach.

Comment 90

Rather than writing this about making a table, explain the process.

Comment 91

Before you move on to the differential expression approach, if you take the averages of the 130 genes and create a heatmap to compare the UL versus non-UL, what kind of pattern emerges? Can you see clear cohorts of gene clusters?

Comment 92

This sentence is very complex and difficult to interpret. Please revise -- it might be helpful to use several, less complex sentences rather than a single long one.

Comment 93

expressed less in UL for that gene.

Comment 94

How can you present these data with a visualization?

Comment 95

Please check the values here -- if there are 130 genes in total, the values here cannot be correct.

Comment 96

What was your rationale for this step? It is not clear why DE should be associated with chromosome location.

QM

WC

Word choice error:
Sometimes choosing the correct word to express exactly what you have to say is very difficult to do. Word choice errors can be the result of not paying attention to the word or trying too hard to come up with a fancier word when a simple one is appropriate. A thesaurus can be a handy tool when you're trying to find a word that's similar to, but more accurate than, the one you're looking up. However, it can often introduce more problems if you use a word thinking it

has exactly the same meaning.

Comment 97

What is the rationale for this categorization? I do not understand the reasoning behind this stratification. Remember, analyses need to have a biological relevance in order to be meaningful.
Also, the way that this information is presented in Table 3 is difficult to interpret. Assuming that you can substantiate why this analysis makes sense, you will need to revise how the information is presented, so that your results are easily interpreted (not obscured).

Comment 98

Please edit the table title, and make sure that it is consistent with the title provided on the list of tables.

Comment 99

I haven't read your paper for content yet (at this point I'm looking at format) -- it seems that you have merged two different ideas into a single table, and that means that some information is duplicated (something that should always be avoided). I suggest that you create one table to report information that is unique to specific genes, and a second table to report information about the cohort of genes based on chromosome number.

Comment 100

Please update the format of this table so that it is not just squares with lines to denote boxes -- please select a table format feature (such as within Word) to make the presentation of data look less stark.

Comment 101

Make sure that you are consistent with formatting.

Comment 102

This is can be biologically interesting, as it may explain why these cells behave abnormally. Have you performed an analysis to determine whether these genes have in common functionality? For instance, if you input these genes into DAVID, are there common pathways represented? Can you extrapolate why this change in expression may be linked to UL?

Comment 103

Each figure will need its own text to explain. Remember, though, that you will need to consider why, biologically, it is worthwhile to know this information.

Comment 104

Make sure that each figure is placed on its own independent page in your document.

Comment 105

Within the figure, make sure that the text given is large enough and clear enough so that it can be clearly read.

As presented, it is not clear:

- Why the left part of the chromosome has black/red cytoband shading, while the right part of the chromosome is white or grey.

- Why some characters have left or right facing arrows, while others have no arrows.

- Where along the length of the chromosome these genes fall (I am guessing that they all fall within the section that is highlighted with the red box?).

- Why different genes are placed on the lines (top to bottom) that they are.

Comment 106

Here, the title needs to be fully descriptive of the information that is shown. For example, the title could be:

"Distribution of genes more highly expressed in UL along chromosome 22." The legend should then explain or identify features needed to interpret the meaning of the figure.

Comment 107

Make sure that it is easy to read all text given in this figure.

What is the rationale for distributing the genes on two different rows?

What is the connection between the schematic of the chromosome that is given at the top and the location of the genes, as listed at the bottom?

Comment 108

Make sure that each figure is placed on its own independent page in your document.

Comment 109

Make sure that the figure title is clear and is consistent with the information included on the list of figures.

Comment 110

See questions posed for the previous two figures. Make sure that the included information is

easy for the reader to interpret.

Comment 111

Each figure should be placed on its own unique location within the document.

Comment 112

You seem to be presenting similar information across different figures, each for a different chromosome. I suggest that you create one common way to title this type of figure, and then use that consistently across all figures, only changing the part unique to that figure (such as what chromosome is represented).

Comment 113

Make sure that each figure is placed on its own independent page in your document.

Comment 114

See previous figures for comments on how to better clarify what information is shown.

Comment 115

Make sure that the figure title clearly conveys what is displayed in that figure.

PAGE 23

Comment 116

Make sure that your terminology is used consistently. Make a definition wherever you first use the term/abbreviation and then use that consistently.

Comment 117

You will need to explain the utility of performing this analysis. Remember, there needs to be a biological relevance to all of the analyses you perform, and the interpretation will be most valuable when connected to a citation that demonstrates the validity of this approach.

PAGE 24

Comment 118

The information you have provided after the table number gives the reader info about how you arrived with these results, but that is different from having a descriptive title that summarizes what is contained within the table. Please improve.

Comment 119

I would reconsider the order of your columns here. Typically, identifier information is given on

I would reconsider the order of your columns here. Typically, identifier information is given on the left,

Comment 120

Please format all tables within your document. They should not be screen captures of information within a spreadsheet.

Comment 121

Make sure to generate the table as one, rather than capturing two images and splicing together the pictures.

PAGE 25

Comment 122

Have you compared the simulation distribution to the distributions of the values actually represented in the initial data set? Is the simulation representative of what has been experimentally collected? How is this analysis relevant for better understanding gene expression in UL versus non-UL samples?

Comment 123

All data tables need to be placed on their own independent page.

Comment 124

Please improve format.

The title also implies that the reader should know information about which genes are up in UL or down in UL compared to non-UL, but that information is not included. As it currently stands, is it important to provide full length gene names for each?

PAGE 26

Comment 125

As currently presented, it is not possible to read the axis labels.

Also, what is the significance of having some y axis labels on the right and others on the left? You cannot assume that your reader knows what this means -- it must be explained. You might want to include in the figure legend.

PAGE 27

Comment 126

Please describe the analyses thoroughly and, algorithm by algorithm, report the results.

Comment 127

Each analysis should be explained independently. Of the TOP16 genes, can you determine which are the most important for determining UL versus non-UL?

PAGE 28

Comment 128

Looking at this table on its own, it is not clear what each of the columns represent. Also, it is not clear what each row represents.

The legends makes reference to scores, but no scores are provided here. This table is very difficult to interpret on its own.

Please improve the format of this table.

PAGE 29

Comment 129

Is it necessary to provide this information as a "figure," or could the information be presented in another way? In other articles that you've referenced as part of your research, is this the format by which this information is typically presented in a formal presentation?

Comment 130

All figures should be placed on their own independent page.

Comment 131

Please describe each independently, making clear the strength of that analysis relative to other approaches.

Comment 132

This needs to be better defined.

PAGE 30

Comment 133

What was your rationale for performing this analysis? You should have a justification, tied to the publication record, for your decisions.

PAGE 31

Comment 134

At the start of this chapter, it would be stronger if you began by first re-stating the objective(s)

of this study. By refreshing the reader's memory, it makes it easier to assess whether the goals were actually met or not.

Comment 135

This section needs to extensively integrate your outcomes with the literature. Do your results make sense? How are the outcomes perhaps not surprising? Surprising?

Comment 136

Within this section, you need to discuss all of the results, not just the last portion of your analysis. Please consider your initial heat maps, explain why they were not helpful and so forth. Based on those outcomes, how did that part of the project integrate with the later parts of your analysis?

Comment 137

Expand upon this further...why did you perform the analysis the way you did? What other needs might there be in order to complete the analysis in a better way?

Comment 138

Did you conclusively demonstrate this in your work? Ultimately, what about your research improved upon what was already known?

Comment 139

How so? What other information might you need? Why might these genes, in particular, be important for changing cellular function? What is already known about the cellular role for these gene products?

Comment 140

But your work focused on gene expression. Why would SNPs be necessary? You need to make clear in your writing the distinction between performing an analysis for expression level changes versus genotype differences. Limitations then need to explain what were issues in your project, given that you elected to study expression changes. The objective of your project was not, as I understand it, to evaluate genetic (DNA level) differences -- thus, this should not be the limitation you focus on.

Comment 141

The top margin of this page should be set to 1 inch.

Comment 142

Page number(s)?

PAGE 33

Comment 143

Page number(s)?

Comment 144

This reference seems to be incomplete. Are you referencing an up-to-date data entry point, or are you referencing an article that was published? I can find this https://www.ncbi.nlm.nih.gov/pubmed/11752295

but the year is different.

Comment 145

Page numbers?

QM

Del.

Delete

PAGE 34

Comment 146

To be consistent with some of your other references, this should also be an active link.

PAGE 35

Comment 147

This should be *Sarcoma* 2015, 412068.
See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707342/

Comment 148

Be consistent with your formatting: use volume number, page number

PAGE 36

PAGE 37

Comment 149

Make sure that this source is reported correctly, including the e-page number. If you want to include the doi, please incorporate that, rather than the PMID. See the citation section here:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012362

QM

Del.

Delete

RUBRIC: ROUGH DRAFT RUBRIC 3.22 / 5

ABSTRACT (11%) 5 / 5

5 (5)	Concisely and clearly covers all key components of the document: rationale, objective(s), methods, results, conclusions and implications
4 (4)	Concisely and clearly covers all but one key component OR clearly covers all key components but could be a little more concise
3 (3)	Covers most key components but could be conveyed more clearly and/or concisely
2 (2)	Many key components are missing; those stated are unclear and/or are not stated concisely
1 (1)	Abstract is missing or, if present, provides no relevant information

INTRODUCTION (11%) 4 / 5

5 (5)	Clearly, concisely and logically presents all key components: relevant and correctly cited background information, rationale, objectives, approach
4 (4)	Concisely and clearly covers all but one key component (with exception of rationale) OR clearly covers all key components but could be a little more concise
3 (3)	Covers most key components but could be done much more logically, clearly, and/or concisely
2 (2)	Many key components are very weak or missing; those stated are unclear and/or are not stated concisely. Weak/missing components make it difficult to follow the rest of the paper
1 (1)	Introduction provides little to no relevant information

METHODS (11%) 2 / 5

5 (5)	Concisely, clearly and chronologically describes procedure used so that a knowledgeable reader could replicate; methods appropriate for the project
4 (4)	Concisely, clearly and chronologically describes procedure used so that reader could replicate most with the exception of a few minor details; methods used are appropriate. Minor problems with organization OR some irrelevant information
3 (3)	Procedure is presented such that a reader could replicate only after learning a few more key details OR methods used are reasonably appropriate for project, though a more straightforward approach might have been taken
2	Procedure is presented such that a reader could replicate BUT methods are largely

(2)	inappropriate; OR procedure is presented such that a reader could replicate only after learning several more key details
1 (1)	So little information is presented that reader could not possibly replicate OR methods are entirely inappropriate

RESULTS (11%) 2 / 5

5 (5)	Contains concise, well-organized narrative text and tables/figures that highlight key trends/patterns/output produced through applied methodology; refrains from providing interpretation of outcomes. Tables/figures have appropriate labels with legends (can stand on their own)
4 (4)	Has presented both a concise, narrative text and informative tables/figures without interpretation but has made a few minor omissions or has other relatively small problems
3 (3)	Has presented findings with a reasonably good narrative text, has informative tables/figures, but has 2-3 problems such as: relevant data mixed with unnecessary information; information provided in tables/figures but not written about in the text; tables/figures are not adequately labeled/described such that they could stand on their own; interpretation briefly made
2 (2)	Has 3-5 problems such as: narrative text and tables/figures are minimal and mostly uninformative; some relevant data given, but mixed with irrelevant information; major concepts are obscured in tables/figures, not explicitly noted in the text; interpretations and conclusions given
1 (1)	Major problems that leave the reader uninformed; lack of narrative text; tables/figures contain unclear and/or irrelevant information

CONCLUSION (11%) 2 / 5

5 (5)	Clearly, concisely and logically presents all key components of research; evaluates outcomes in terms of the objective of the study; compares outcomes with relevant findings in literature; evaluates experimental design, evaluates reliability of data; states implications of results, suggests next investigation steps; ends with final conclusion
4 (4)	Concisely, clearly and logically covers all but 1-2 key components OR clearly covers all key components but could be more concise OR has discussed outcomes of project but also included a laundry list of experimental problems without discussing their impact on conclusions
3 (3)	Covers most key components but could be done much more logically, clearly and/or concisely
2 (2)	Many key components are very weak or missing; those stated are unclear and/or are not concise

1 (1)	Most key components are missing or very weakly done

PROGRESSION (12%) 3 / 5

5 (5)	Excellent progression of the content with logical synthesis of evidence
4 (4)	Content displays logical progression with appropriate synthesis of evidence
3 (3)	Content adequately structured with some synthesis of evidence
2 (2)	Content partially organized, evidence disjointed
1 (1)	Content not well organized, evidence lacking

SYNTHESIS (11%) 3 / 5

5 (5)	Exemplary use of the literature to support the project; literature used exemplifies current state of knowledge and is used critically
4 (4)	Literature used supports the project and is accurately portrayed, but contains some unnecessary information and/or is not a complete representation of the literature
3 (3)	Literature used appropriate to the project topic and adequate comprehension of material displayed, but conceptual detail is lacking; adequate but limited review of literature.
2 (2)	Literature used appropriate to the project topic but evidence of comprehension lacking; inadequate review of the literature
1 (1)	Literature used not always appropriate to the project topic; no evidence of comprehension; insufficient review of the literature

FORMAT (11%) 3 / 5

5 (5)	All sections are included in the correct order; tables/figures are correctly formatted; citations are correctly provided; margins, page numbers and sections/sub-sections are provided correctly; appropriate use of person and tense; other misc. stylistic components as designated in the Format Guide
4 (4)	All major sections of the paper are correctly provided, with 1-2 minor errors in stylistic formatting
3 (3)	All major sections of the paper are correctly provided, with 3-4 errors in stylistic formatting

2 (2)	Most major sections of the paper are provided, but are not presented in the appropriate format
1 (1)	Missing components of the paper; lack of regard for formatting requirements

GRAMMAR (11%) 5 / 5

5 (5)	Less than 3 spelling, grammatical or mechanical errors
4 (4)	No more than 5 spelling, grammatical and/or mechanical errors
3 (3)	Fewer than 8 spelling, grammatical and/or mechanical errors
2 (2)	Less than 10 spelling, grammatical and/or mechanical errors
1 (1)	More than 10 spelling, grammatical and/or mechanical errors