Lewis University

Meta-Analysis of the Genes Ubiquitously

Associated with Human Uterine Leiomyoma Development

in Healthy Humans Using

the Gene Expression Omnibus Data

by

Janis L. Corona

Romeoville, IL

July 2019

Abstract

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

*Keywords*: uterine leiomyomas, uterine fibroids, latent dirichlet allocation, top 16 genes, six ubiquitous UL genes, bet1 golgi vesicular membrane trafficking protein like, trinucleotide repeat containing adaptor 6b, cytohesin 4, fatty acid synthase, high mobility group at-hook 2, coiled-coil domain containing 57

Table of contents

list of tables…………………………………………………………………………… 1v

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

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

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

Description of ul……………………………………………………………... 2

UL Described in Populations……………………………………… 3

significant genes for ul…………………………………………… 5

chromosomal location of ul gene…………………………………… 5

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

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

chromosome 17………………………………………………………... 7

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

chapter 2 – methods………………………………………………………………… 8

geo data of ul and non-ul samples………………………………….. 8

r statistical software for statistical analysis……………… 8

bioconductor for biostatistics in r………………………………... 10

chapter 3 – results…………………………………………………………………. 11

chapter 4 – conclusions………………………………………………………….. 15

chapter 5 – literature cited…………………………………………………… 17

List of tables

Table 1. Data on Genes…………………………………………………………………… 11

Table 2. Data on Majority……………………………..…………………………………. 12

Table 3. Data Table on Genes Having Highest Magnitude of Change…………………... 13

List of figures

Figure 1. Data on Genes…………………………………………………………………… 10

Figure 2. Data on Majority……………………………..…………………………………. 11

Figure 3. Data Table on Genes Having Highest Magnitude of Change…………………... 12

List of abbreviations

BMI Body Mass Index

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

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

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 progesterones, 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 in 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, *CCDC57* on chromosome 17, *HGMA2* on chromosome 12, and *MED12* on chromosome X or 23 (Aissani et al., 2015; Eggert et al., 2012; Bondagji et al., 2017; Edward et al., 2013; Hodge et al., 2012; Hellwege et al., 2017; Liu et al., 2018; Rafnar et al., 2018). Zhang et al. (2012) found chromosomes 4, 14, and 9 to be in healthy uterine tissue

capable of impregnation; these chromosomes are not from the UL risk gene chromosomes found along chromosomes 11, 12, 17, 22, and 23 in current population studies. Thus, it makes sense to further study these genes associated with UL except for the *MED12* gene on chromosome 23 that has already been proven causal to UL (Bondagji et al., 2017). The *CDC4* and *WNT4* genotypes 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).

Chromosome 11

*BET1L* gene is on chromosome 11 and it is described as having significant associations with UL such as which uterine layer a UL is originating from or how many UL are in one uterus making the UL patient have multiple UL (Cha et al., 2011, Liu et al., 2018; Edwards, Hartmann, & Edwards, 2013; Rafnar et al., 2018). *BET1L* was tested for significance in association with UL in studies on other race demographics and determined insignificant in certain races (Bondagji et al., 2017; Aissani, 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

Two genes on chromosome 17 along cytoband 17q25.3 named *CCDC57* and *FASN* are significantly associated with UL in Europeans (Eggert et al., 2012; Aissani et al., 2015). 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 similarly 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 was then filtered in R to only use the genes along the same chromosomal locations as those six genes *TNRC6B, BET1L, FASN, CYTH4, CCDC57,* and *HMGA2*. The base statistical functions in R were primarily used to combine by merging the GEO data sets and applying the first listed item in a field when many observations had multiple entries. More work was done to add ENSEMBLE data fields to the data to use with the Gviz Bioconductor package in R for visualizing the gene locations. The R package, dplyr, was used to add fields that describe the means of each gene in the samples of UL and non-UL separately, then add a field of changes in expression by means of each type of sample and use as a way to group the genes by differential expression in up and down expression in UL compared to non-UL samples, and for determining if the genes were part of the majority of genes expressed more or less in each chromosome (Francois, Lionel, & Muller, 2019). Then dplyr was used to create a field that determined the top 10 expressed genes by magnitude of most or least expressed in UL when compared to non-UL samples (Francois, et al., 2019). The R packages, ggplot2, heatmaply, and lattice were used along with R base package to create plots that could describe the data visually to look for patterns between the genes, samples, or stats of the samples (Wickham, 2019; Galili, O'Callaghan, Sidi, & Benjamini, 2019; Sarker, 2018). Bootstrap simulations using the ‘UsingR’ r package using 10,000 simulations with replacement for each TOP16 gene (Maindonald, 2008). Then histograms of those 16 genes were made using ggplot2 to see how symmetrical each gene in the population would fit the Gaussian bell curve (Wickham, 2019). This was generated per TOP16 gene based on a generalization of the Central Limit Theorem and the Law of Large Numbers which state that a sample of a larger population will converge to the true population mean when sampling with replacement is done a large amount of times. One simulated population mean for UL and one for non-UL converged from 10,000 samplings per TOP16 gene of the combined 121 GEO samples. The predictive algorithms of PCA, LDA, RF, KNN, and GBM were used on this dataset of TOP16 genes using kernlab, caret, gbm, lda, randomForest, e1071, and MASS r packages (Karatzoglou, Smola, Hornik, Maniscalco, & Teo, 2018; Kuhn, Wing, Weston, Williams, Keefer, Engelhardt, & Hunt, 2019; Greenwell, Boehmke, &

Cunningham, 2019; Chang, 2015; Breiman, Cutler, Liaw, & Wiener, 2018; Meyer, Dimitriadou, Hornik, Weingessel, Leisch, Chang, & Lin, 2015; Ripley, Venables, Bates, Hornik, Gebhardt, & Firth, 2019).

Bioconductor for Biostatistics in R.

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

Results

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.

literature cited

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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