# Revised Proposal, Prospectus, and Annotated Bibliography_Janis

*by* Janis Corona

**Submission date:** 07-Jun-2019 12:07PM (UTC-0500)
**Submission ID:** 1141128460
**File name:** Week1_Proposal_Prospectus_AnnotatedBibliography.docx (50.39K)
**Word count:** 8902
**Character count:** 45924

Janis Corona

BIOL 59000: Data Science Project for Life Sciences

Week 1 Assignment: research proposal, prospectus, and annotated bibliography

June 3, 2019

# Meta-Analysis of the Ubiquitous Genotypes Associated with Human Uterine Leiomyoma Development in Healthy Human Tissue

In this research project, the top genetic markers for heterogenetic risk in developing uterine leiomyomas (UL) will be analyzed in data made available for gene expression using the Gene Expression Omnibus (GEO) online data repository. There are many genome wide association studies (GWAS) on the many single nucleotide polymorphisms (SNP) associated with UL, the studies have been exclusive to subsets of races by only 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, Hartmann, & Edwards, 2013; Liu et al., 2018; Hellwege et al., 2017; Rafnar et al., 2018; Cha et al., 2011; Aissani, Zhang, & Weiner, 2015). The distinct difference between this study and previous studies is using a subset of non-race specific gene expression samples instead of solely samples from one specific race to find or not find those same ubiquitous genotypes to be associated with UL development.

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 are estrogen responsive and that discontinuing hormone therapy can make UL symptoms and tumor growth recur (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).

The most ubiquitous SNP genotypes highlighted in these GWAS population specific studies are the rs2280543 genotype belonging to the Bet1 Golgi Vesicular Membrane Trafficking Protein like gene called BET1L and the rs12484776 genotype belonging to 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 genotypes have been 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 genotype rs2280543 is not associated with UL. However, two other separate studies by Eggert et al. (2012) and Edwards et al. (2013) found that the BET1L genotype rs2280543 is associated with UL risk for white women and European Americans.

A study on European Americans by Edwards et al. (2013) found that rs2280543 of 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). This same phenotype of BET1L genotype rs2280543 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, additional SNPs for fatty acid synthase (FASN) and rs4247357 of 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 target found to be an exclusive heterogenetic risk of UL in African American populations is the rs739187 SNP of 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. It 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 gene SNPs from the GWAS samples were gathered from most expressed genes along a region of one of those chromosomes, ande further analyzed to determine which genes had significantly high gene expression of SNPs in UL cases (Aissani et al., 2015; Eggert et al., 2012; Bondagji et al., 2017; Edward et al., 2013).

Currently, significant SNP 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 SNP located genotypes along chromosomes 11, 12, 17, 22, and 23 that have regions associated with developing UL among varying populations. Thus, it makes sense to further study these genotypes associated with UL except for the MED12 gene 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).

The design of this research will be to examine the gene expression data collected from the GEO data repository of six independent studies involving healthy human uterine myometrial tissue and human UL tissue (Crabtree, Jelinsky, Harris, & Choe et al, 2009; Hoffman, Milliken, Gregg, Davis, & Gregg, 2004; Miyata et al., 2015; Quade, Mutter, & Morton, 2004; Vanharanta et al., 2006; Zavadil et al, 2010). The R biostatistics software Bioconductor will be used for machine learning on the GEO data to

find significant SNPs associated with UL among this study's UL and non-UL GEO gene samples found in chromosomes 11, 12, 17, and 22 and compare to the ubiquitous top UL risk genotypes of CCDC57, BET1L, TNRC6B, FASN, HGMA2, or CYTH4.

Many of the methods similar to Zhang et al. (2012) and Eggert et al. (2012) will be used for developing visualizations for analyzing the data that could include: a differential expression MA-plot for highly up or down regulated genes between groups of UL and healthy samples, box plots per gene, a QQ plot (with top SNPs being far from the diagonal), PCA analysis of variable genes, link analysis of top genes to their closest genes to elucidate the functional inhibition that may or may not occur, the highest expressed up or down regulated genes between the healthy and UL samples, and other analytics as needed using only the packages in R and Bioconductor. The analysis work will focus on finding significant genes associated with UL development in healthy people using the GEO multi-racial pool of 132 microarray samples of 77 UL and 55 non-UL samples.

Machine learning will then be used to develop a network of linked genes that impact other genes along the chromosomes associated with UL development. Using a training set of 70% of the 132 samples randomly selected and a test set of the other 30% will give a training set of 92 samples and a testing set of 40 samples. If the model built produces accuracy of a certain threshold like 90%, then it will be assumed a good measure of which genotypes are associated with UL. Models could be built in R for log linear regression, random forest, k-nearest neighbor, and other regression models like Bayesian to test accuracy in predicting UL risk genes. If the accuracy compare across all models is below 50% , then the original chromosomal markers need to be identified for the hundreds of genes left out when originally searching for SNPs associated with UL as some studies suggested (Bondagji, et al., 2017; Aissani, et al., 2015).

# Prospectus for Gene Expression Analysis of the Ubiquitous Genotypes Associated with Uterine Leiomyoma Development in Healthy Human Tissue

I. **Introduction** This section introduces uterine leiomyomas (UL) and the race specific genotypes associated with pathogenesis of UL. The focus of this research is to contribute to the study of UL development in otherwise healthy human tissue using GEO data across all races. Other studies on UL development used race specific studies of UL development to find genotypes associated with UL risk within each race. The distinct difference between this study and previous studies is using non-race specific gene expression samples instead of solely samples from one specific race to find or not find those same ubiquitous genotypes to be associated with UL development. The software used to analyze the data sets is R and Bioconductor using RStudio to determine if these top UL risk genes are found to be sufficient in determining UL risk in any female patient the chromosomal regions where these top genes reside.

A. The ubiquitous genotypes associated with the pathogenesis of UL in women from current research publications on UL in race specific populations includes Japan, White European Americans or Icelandic populations, African Americans, Australia, Hans Chinese, and Saudi Arabia (Cha et al., 2011; Edwards, Hartmann, & Edwards, 2013; Hellwege et al., 2017; Aissani, Zhang, & Weiner, 2015; Liu et al., 2018; Bondagji et al., 2017; Rafnar et al., 2018).

   a. These ethnicity specific studies and other UL studies unrelated to ethnicity found certain genotypes in RNA samples of UL and non-UL RNA to either indicate a risk for UL, indicate the size of the UL, or indicate risk of having more than one UL in one patient (Cha et al., 2011; Edwards et al., 2013;

Hellwege et al., 2017; Aissani et al., 2015; Liu et al., 2018; Bondagji et al., 2017; Rafnar et al., 2018; Eggert et al., 2012; Hodge et al., 2012).

b. The contribution to UL research this project will obtain is to discover if these same ubiquitous genotypes are found in six multiple microarray data sets in the Gene Expression Omnibus (GEO) online data repository of UL gene expression data (Crabtree et al., 2009; Hoffman, Milliken, Gregg, Davis, & Gregg, 2004; Miyata et al., 2015; Quade, Mutter, & Morton, 2004; Vanharante et al., 2006; Zavadil et al., 2010).

B. The software for analyzing the UL data after combining the data into one data set of microarray data.

a. R using RStudio (R, 2019)

b. Bioconductor for R (Bioconductor, 2019).

II. **Background**. This section describes the Ubiquitous Genotypes Currently Associated with Uterine Fibroids and explains what they are and the impact they have on lives.

A. **Description of uterine leiomyomas (UL)**. This sub-section describes who UL affect, the known risk factors for UL, and what known SNP or genotypes are significantly associated with UL.

a. **UL described in populations**. This sub-sub section describes the population demographic risks, symptoms of UL, asymptomatic patients' impact on reproducible research, and the most chosen treatment for UL being hysterectomy (Hellwege et al., 2017; Eggert, et al., 2012; Rafnar et al., 2018; Bondagji et al., 2018; Liu et al., 2018; Hodge et al., 2012; Cha et al., 2011).

b. **Significant genotypes for UL**. This sub-sub-section lists the SNPs associated with UL development in women. These genotypes associated with UL are:

BET1L, TNRC6B, HGMA2, FASN, CCDC57, and CYTH4 (Hellwege et al., 2017; Eggert et al., 2012; Rafnar et al., 2018; Bondagji et al., 2018; Liu et al., 2018; Hodge et al., 2012; Cha et al., 2011).

B. **Chromosomal location of UL associated SNPs**. This sub-section lists and describes each chromosome associated with regions of SNPs that house the most common SNPs significantly associated with UL.

a. **Chromosome 11**. This sub-sub-section describes the BET1L genotype SNP that is described as having various 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). This genotype, 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).

b. **Chromosome 12**. This sub-sub-section describes the location of HGMA2 and how it is associated with high expression in UL samples (Hodge at al., 2012). Another study stated HGMA2 to be a factor in tumorigenesis from studies done in 1988 researching HGMA2 and tumor formation (Aissani et al., 2015).

c. **Chromosome 17**. This sub-sub-section describes the two genes with SNPs on chromosome 17 named CCDC57 and FASN that are associated with UL in Europeans (Eggert et al., 2012; Aissani et al., 2015). Another study tested these two gene SNPs and found no significance in UL for Saudi Arabian populations (Bondagji et al., 2017).

d. **Chromosome 22**. This sub-sub-section describes the two genes that are found on Chromosome 22 to be significant in UL in specific races. For the first gene SNP called 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, one of the other genes on Chromosome 22 found to be significant in UL for African Americans will be described in this sub-sub-section as well as TNRC6B (Hellwege et al., 2017).

III. **Materials and Methods**. This section will describe the material sources used and the type of software used to analyze the data. The ubiquitous SNPs found most differentially expressed between non-UL and UL samples of gene expression data will be used to build a model (D. Zhang, Sun, Ma, Dai, & W. Zhang, 2012; Eggert et al., 2012). Link analysis will be generated linking influential genes used in their SNP regions of the chromosomes and the separate population studies found to be significant in UL association (Hellwege et al., 2017; Eggert et al., 2012; Rafnar et al., 2018; Bondagji et al., 2018; Liu et al., 2018; Cha et al., 2011; Aissani et al., 2015; Hodge et al., 2012; Edwards et al., 2013).

A. **Gene Expression Omnibus (GEO) data of UL and healthy uterine samples**. This sub-section explains the use of GEO to access six microarray data sets that list their gene expression data on GEO for sharing (Crabtree et al, 2009; Hoffman et al., 2004; Miyata et al., 2015; Quade et al., 2004; Vanharanta et al., 2006; Zavadil et al, 2010). These 'in situ oligonucleotide' or microarray platforms were derived from five separate studies and one study not published (Hoffman et al., 2004; Vanharanta et al., 2006; Zavadil et al., 2010; Crabtree et al., 2009; Miyata et al., 2015; Quade et al., 2004).

a. The combined data sets will combine into one big data table of 55 non-UL and 77 UL observations for analyzing using R.

b. This final big data table will be analyzed with differential gene expression between non-UL and UL samples to find the relation of the ubiquitous genotypes mapped next to the other runner up genotypes not further explored in the population studies.

B. **R Statistical software for statistical analysis**. This sub-section describes how R software will be used in this research.

a. R to transform the six independent microarray data sets into the same scale as one data set GSE68295 used the log 2 scale.

b. R will then be used to combine the microarray data by common probe IDs using the GEO platform data that each independent study was derived from (Edgar, Domrachev, & Lash, 2019).

i. Four of the array data sets have the same probe IDs but two of the array data sets have different IDs that can be aligned to the four other sets by using the platform array data containing the alternate gene names.

ii. The gene symbol field of all six data sets aligned to their respective platform is the item that will be merged on.

iii. R will then extract only the first listed gene symbol from each data frame, omit the NAs and duplicate fields, keep only the field meta fields of alternate symbol names on one data frame to merge the other five data frames to, then merge each data frame together by gene symbol.

iv. R will then extract only those observations that the cytoband location of each gene is on one of the ubiquitous gene loci so that a data set of only those genes and loci of genes in the current studies on UL are used to compare gene expression data for UL risk in healthy (non-cancerous UL) patients.

c. The final data table produced with R will have 1857 observations and 149 fields including 17 meta data fields, UL samples, and non-UL samples. An interactive heatmap showing gene expression data with the 'heatmaply' package in R will be provided in this section demonstrating findings on UL associated risk genes of current studies (Galili, O'Callaghan, Sidi, & Benjamin, 2019).

C. **Bioconductor for biostatistics in R**. This sub-section describes use of Bioconductor and R packages Gviz, DESeq, and ggbio for analyzing gene expression data with M-A plots, karyogram plots, circular plots, and others as needed for analyzing the GEO data (Bioconductor, 2019; Anders, & Huber, 2019; Hahne, 2019; Yin, 2019).

IV. **Results**. This section will display the results and provide a detailed summary of each result and what it means.

A. **Analyzing the biostatistical data from R**. This sub-section will explain the meaning of the scatter points or box plots developed using the R statistical packages. Any plots, tables, or other visualizations for analyzing the data before making predictions on the population from the GEO data set of samples will be displayed in this section.

B. **Machine Learning with Bioconductor**. This sub-section will display the visualizations of the tables and/or graphical plots created in Bioconductor and explain the meaning of each visualization as it is displayed.

IV. **Discussion**. This section will synthesize all the information from the results gathered and bring it back to answering the research question of which ubiquitous genotypes are associated with UL in healthy females. The conclusion will be within the end of this section and will interpret the results. There is an open question for further analysis of the next set of five or six subsequent top genotypes having a risk for UL development in healthy patients. Right now, the idea is that UL development is currently unique to each race and personalized treatment for UL in each race is needed (Hodge et al., 2012; Hellwege et al., 2017; Liu et al., 2018; Edwards et al., 2013).

V. **Supplementary Material**. This section will provide the database of samples used from GEO as one table of UL and non-UL microarray samples. In this section, there will also be a link to view the original GEO files used, the R coding script file that created the combined table of all six microarray UL and non-UL samples, and the analysis script used to create the visualizations in R and Bioconductor.

VI. **References**. This section will display all of the references used in this research that include data resources, research articles, and software used. If any other types of data are used, such as supplemental articles or review articles published in peer review journals, then they will also be listed in this section.

## References

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

This research study does a meta-analysis between the NIEHS-UFA National Institute of Environmental Health Study, The Right from The Start (RFTS) and BioVu (Vanderbuilt University biorepository of DNA extracted from discarded blood of de-identified medical patients) study on European Americans, the studies on Japanese, Chinese, Africans, and other ethnic studies published to analyze the significant gene markers in all the SNP genome wide association studies associated with uterine fibroids. This study's findings show that SNPs in TNRC6B labeled rs139909 and rs138089 in the NIEHS-UFA study to be significantly associated with UL in European Americans. Also, this study found that genotype rs12484776 is significant in the RTFS study for UL tumor size and UL risk. The reproducing of the African American population study for UL risk didn't show this gene target of SNPs to be significant as the African American study originally produced the same results. The BET1L UL risk was reproduced in the Han Chinese, Japanese, and European American populations but not the African American population. This article also suggested THDS7B and the ECM or extracellular matrix components are factors in UL risk and UL size. The study discussed that there could be misclassification in the control group of UL in the African American population based on the self-reported method in the Black Women's Health Study (BWHS). If that misclassification was true, then there could be a new significant finding in the TNRC6B and BET1L genotypes being associated with tumorigenesis in UL among African American populations as they have shown for Chinese, Japanese, and European populations.

Anders, S. & Huber, W. (2019). Differential expression of RNA-Seq data at the gene level -the DESeq package. Retrieved June 3, 2019, from https://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf

This resource is for information on the functions the R package DESeq gives users in manipulating data for analysis. The modules inside this package will be used for statistical purposes for machine learning on a training and testing partition of the data table of all microarray UL and non-UL samples. It produces a dispersion plot, differential gene expression data tables, heat maps, MA-plots, and true positive/ false negative tables for statistical analysis in classification of UL or non-UL samples.

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

This is an open source and free software to analyze genomic and other biological data using biostatistics. It operates with R software version 3.5.0 or higher as this version of Bioconductor is version 3.8. There are more than 1600 biostatistics packages available using this software for R. R is also an open source free software for statistical analysis and machine learning. R is easily able to be updated and when compared to Python it is not as package version dependent, nor platform dependent like Python2 and Python3 platforms. R is easier to upload with less time spent downloading and configuring packages that need to be rolled back or updated. The coding of R is similar to python and other coding languages. The packages in Bioconductor that will be used for this research report are ggbio and DESeq. This software is built to use the data repository that GEO provides.

Bondagji, N., Morad, F., Al-Nefaei, A., et al., (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

This study used Saudi Arabian women with UL and without UL to study the putative genes shown in Japanese and European populations to be a risk factor for UL. The results showed that the SNP in TNRC6B as rs12484776 is a UL risk factor in Arab populations and that the SNPs rs2280543 in BET1L, as well as rs7913069 of LOC102724351 and rs1056836 of CYP1B1 also increase the risk of UL in Arab populations. The study reports on the current findings on UL such as the gene MED12 is the only gene directly shown to have a causal relationship to UL. Also, TNRC6B and BET1L are risk factors for UL in Japanese populations and in developing UL in European populations, and that FASN and CCDC57 are risk factors for UL in Europeans and Australian populations. This study also examined the location of the UL inside the uterus as being either in the uterine wall (intramural), below the mucosal layer of the uterus (subserous), or below the endometrial layer of the uterus (submucosal). Most of this population had more than one UL (79%) compared to only one UL (21%). The study admits to not being large enough given its roughly 100 samples of UL and non-UL case and control groups for applying the findings to the entire Arab population. Results also did not analyze whether the rs12464776 SNP of TNRC6B is a biomarker for the number of UL a patient has, the uterine location of each UL a patient has within the layers of the uterus, or the size of the UL in each patient due to perceived statistical errors in running those tests. The study also reports that although TNRC6B was found to be a risk factor in UL in Arab populations it is a possibility that the missing gene expression data and linked gene region could be the real pathogenesis of UL because this study only examined the five genes shown in previous UL studies to be risk factors of UL in Japanese and European populations. And these five SNPs were screened for frequencies while ignoring the

millions of other genes that showed variations in humans so the results could be omitting other genes that might be risk factors of UL in Arab populations.

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).

This is the original study that preceded the other race specific studies on Europeans, European Americans, Australians, Chinese, Saudi Arabian, and African Americans. It is a Japanese population study of UL risk that showed BET1L, TNRC6B, OBFC1, and SLK significantly associated with UL risk in Japanese females.

Crabtree, J.S., Jelinsky, S.A., Harris, H.A., Choe, S.E., Cotreau, M.M., Kimberland, M.L., ... Walker, C.L. (2009). Comparison of human and rat uterine leiomyomata: identification of a dysregulated mammalian target of rapamycin pathway. *Cancer Research*, 69(15), 6171-8. PMID:19622772

This database was published August 20, 2009 and is of rat and human uterine samples. The data samples used for this research were from the UL and non-UL human samples found at the GEO online data repository at https://www.ncbi.nlm.nih.gov/geo/ by entering the GEO Access ID of GSE13319. The platform this particular data sequence uses is the platform GPL570. There are 23 healthy human myometrial tissue samples (non-UL) and 23 human UL tissue samples. The human tissue samples are from the Affymetrix Human Genome U133 Plus 2.0 Array. GPL570 platform was made public on November 7, 2003. All the gene expression arrays of the human samples in this data will be added to the human samples from five other GEO microarray UL and non-UL samples.

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

This is the Gene Expression Online database repository (GEO) for finding UL and non-UL gene expression data donated by research scientists. This website was used for all the gene expression microarray data used in this research. 33

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

This research used the Right from the Start (RFTS) study and BioVu (Vanderbuilt University's biorepository of de-identified DNA extracted from discarded blood samples of patients) data. This research used DNA data specifically at the BET1L, TNRC6B, and SLK gene targets discovered in a Japanese population study on UL to analyze the European American population's risk of UL, size of UL, and uterine layer location of UL. It found that there wasn't significance in the SLK gene target due to minor allele frequency (MAF) below a specific threshold value of 0.01. However, it did confirm the significance of BET1L in uterine location of UL in layers of the uterus as either subserous (below the mucosal layer of the uterus) or intramural (in the uterine muscle). Confirmation that an associated risk of UL with both genotypes of BET1L rs2280543 and TNRC6B rs12484776 in European populations was one conclusion of this study. Another conclusion of this study was that TNRC6B genotype rs12484776 is significant as a biomarker for the size of UL in European populations.

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

This study used white female sisters having medically diagnosed UL and other family members totaling 385 pairs of sisters from 261 families with 1,103 individuals to search for gene targets in UL that have cytogenetic abnormalities. Two other studies involving sisters, twins, and moms was also used in this study by adding the data to this study's analysis of gene targets and abnormalities in UL. The gene Fatty acid Synthase (FASN) was found in prostate, breast, colon cancers, and also found in UL. Research from this study showed that impairing this gene's receptors can slow down colon and breast cancer. FASN is highly expressed in hormone sensitive cells, found to be regulated in transcriptional and post-transcriptional levels. Primary transcription factor is sterol-regulating-element-binding-transcription-factor 1 (SREBP-1) activated downstream of growth hormone and hormone receptors. Inhibiting FASN has led to cancer cell line apoptosis, tumor growth ceasing, and has shown little or minimal effects on surrounding normal cells. There is a connection between FASN and the P13K/Akt signaling pathway which is a commonly dysregulated pathway in human cancers. The major and minor allele rs4247357 of gene coiled-coil domain containing 57 or CCDC57 was genotyped from the healthy and UL tissues of 20 UL tumors from 12 women of which half had the major allele and the other half had the minor allele. In order to satisfy a genome-wide association significance level the P value of 0.05 is not good enough and having even a P value of 10^-4 or .0005 is not significant enough to mark a SNP as a gene target for UL risk. After analyzing the CCDC57 alleles and using linkage analysis, a 35 Mb (million base pair chromosome length) genomic region was found containing hundreds of genes that possibly pose a risk for UL. Candidate SNPs found on chromosome 17 in this region are FASN, CCDC57, and SLC16A3. The rs4247357 SNP of CCDC57

was found to be at a genome-wide association significance level for UL risk in white patients using the Finding Genes for Fibroids (FGFF) study, the Women's Genome Health Study (WGHS) and the Australian cohort study.

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

This is the information on using the R package called heatmaply for creating heatmaps of the differentially expressed gene data that is obtained from a collection of the six microarray gene expression data sets found in GEO. This package, 'heatmaply,' is not a Bioconductor package, but is an R package.

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

This is an R package that doesn't rely on Bioconductor and uses UCSC or ENSEMBL data on genes to build tracks as maps of gene tracks and gene expression data mapped to the chromosome and next to neighboring genes in the same chromosomal track. This research uses GEO gene expression data that has multiple names as meta field to the samples that can switch between names as needed to build the gene tracks for visualization and inference purposes that add to this research. The meta fields in the GEO data include the National Center for Biotechnology Information (NCBI) Reference sequence ID, the University of Santa Cruz (UCSC) ID, the Human Gene Nomenclature (HGNC) ID, the HGNC approved symbol, the ENSEMBL ID, and chromosome location information from which the ENSEMBL ID would be used to build the tracks with this R package.

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

The research in this study focused exclusively on African American females in the 23andMe.com database and the GWAS database. This study found an increase in UL risk in the SNP rs739187 of cytohesin 4 (CYTH4). Results showed a lower predicted gene expression in the thyroid tissue was significant in determining UL risk in African Americans. The study recognizes thyroid problems such as overt hypothyroidism, thyroid nodules, and thyroid cancer are associated with UL, and that genes HMGA2 and PLAG1 are associated with UL and strongly correlated with thyroid tumors. None of the BET1L, TNRC6B, or SLK genotype SNPs for UL risk in the Japanese population were found to be a risk for UL in African American populations. Also, women not at risk of developing UL (age beyond menopause) and self-reporting of UL were part of the control group which may have affected the results due to UL misclassification. UL cases are reported only when symptomatic or clinically obvious, which according to Aissani et al. (2015) is in 33% of patients who have UL. This measure of symptomatic UL increases to 50% for Liu et al. (2018) and decreases to 20-25% symptomatic UL for Eggert et al. (2012). Clearly, given those measures for UL patients that are seeking help, as many as 80% of other UL populations do not either know they have a UL or are not reporting they have a UL. Therefore, it could be likely a misclassification of UL occurred in this self-reporting control group for UL. This case study implies heterogenetic risk factors for UL by presenting evidence that the genotypes of TNRC6B do not associate with risk of UL in African American populations. But TNRC6B has been shown in Chinese, Japanese, Saudi, European, and other white races as being associated with risk of UL. At the same time, this study provides evidence of a genotype rs739187 of CYTH4 that is exclusively associated with risk of UL in African American populations only. This study also

confirms the decrease in angio-tension for renal homeostasis (AGT) in thyroid tissue and high ALDH2 (involved in alcohol metabolism) as risk factors for UL in African American populations.

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

This study used nine female patients having more than one UL or multiple UL (MUL) to do a paired comparison study of the recurrent chromosome abnormalities of t(12;14) and non-t(12;14) genomic regions of UL. The study confirmed the putative gene HGMA2 of the high mobility group being over expressed in UL and as the most significant altered gene. Using unsupervised PCA the variation between patients was greater than the t(12;14) UL and non-t(12;14) UL. Also, patient variation was a greater classifier than del(7q) UL and t(12;14) UL. This suggests personalized UL treatment because of the heterogeneity in patient demographics having UL. The patients were of different racial origin in one cohort which differs from the other UL studies looking for genome wide associations in each race. And all patients had more than one UL each. There are 7.5% of UL having translocation differences in the 12q14-15 region of the genome, specifically at t(12;14)(q14-15;q23-24). The samples were weighted according to the percent of cells in the sample which controlled patient demographic variability. The results of this study created a gene expression profile for the subgroup t(12;14) of UL. This study made it clear that genetic heterogeneities exist in the pathogenesis of UL and the first step to treating UL is to identify the genetic profiles for UL. The way the UL gene is listed is different from other studies that list the chromosome location and the base pair location along that region showing

high expression in the genes between control and patient groups. The other studies used a GEO type of identification or one recognized by the NCBI like ENSEMBL or UCSC genomic data repositories. This study used fluorescence in situ hybridization and mentioned HGMA2 as being on chromosome 12. The study stated that HGMA2 is responsible for allowing transcription factors to reach their target genes and restricted mostly to proliferative embryonic tissue that include the derivatives of the mesenchymal tissue. This includes the myometrium from which UL are found to grow. HGMA2 has been found expressed in the myometrium in vivo (in a live person).

Hoffman, P.J., Milliken, D.B., Gregg, L.C., Davis, R.R., & Gregg, J.P. (2004). Molecular characterization of uterine fibroids and its implication for underlying mechanisms of pathogenesis. *Fertility and Sterility*,82(3), 639-49. PMID:15374708

This resource is data collected from the GEO archive located at: https://www.ncbi.nlm.nih.gov/geo/. The data was published to GEO October 17, 2003. The GEO platform ID is GPL96 and the GEO Access ID is GSE593. In this data there are five healthy myometrial samples and five UL samples used to compare gene expression using the samples with a fold change greater than 1.5 or less than -1.5. The results that were significant were those that had a P value less than or equal to 0.05. The human gene samples of healthy myometrial and UL gene samples will be added to the database of the other five GEO data sets of healthy human and UL human gene samples. In total there will be 132 human samples for comparing genes differentially expressed the most between 55 healthy myometrial samples and 77 UL samples.

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

The female case and control groups for UL in this study are not related and all from the Han Chinese population. The genes BET1L and TNRC6B were found to be risk factors of UL in Han Chinese women. SNPs associated with these two genes also showed significance in the number of UL and the size of the UL in each patient. This study found the SNP for TNRC6B called rs12484776 with genotypes GG, GA, and AA (most common genotype is AA) is significantly associated with the size of the UL. It also found that the SNP rs2280543 associated with BET1 has two genotypes TT and CT (most are CC) that are significantly associated with the number of ULs one patient has. Neither of these gene targets are associated with gene expression in the uterus using eQTL-expression quantitative trait loci from gtexportal.org/home or the gtex database that uses healthy tissue samples for gene expression mapping. This study also compares itself to the Japanese, Saudi Arabian, and European population studies on UL from 2011-2017.

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*. PMID: 27057136

The GEO repository of data located at https://www.ncbi.nlm.nih.gov/geo/ was accessed for the data of healthy myometrial and uterine leiomyoma (UL) gene data for this resource's data. This specific study also includes cancerous leiomyomas called leiomyosarcomas (LMS) that will not be included. The GEO Access ID is GSE68295 using GEO platform GPL6480. This data was

published to GEO on February 9, 2017. In the GSE68295 data series, the samples are separately identified as UL, uterine leiomyosarcoma (cancerous UL), or healthy myometrial tissue. This is how all the data is separated in the other five GEO resources used. In each GEO series prepended with 'GSE' and followed by the ID number of that study, there will be a list of separate data sets that identify what type of data it is. These data sets begin with 'GSM' and are followed by their data set ID number. In this study, the GSM IDs are clearly explained as to what data it is. For instance, GSM1667147 is titled, 'uterine leiomyoma tissue from case 4,' and GSM1667145 is titled, 'uterine normal myometrium tissue from case 2.' There are three healthy and three UL samples totaling six samples to add to the other five GEO data sets on healthy myometrium gene samples and UL gene samples. The six gene samples of LMS will not be added to the data needed for research on gene expression of most differentially expressed genes between normal and UL tissue. When all of the six GEO resources used are extracted for human normal myometrium and human UL gene samples is completed, the database for studying UL gene expression will total 132 gene samples of 55 non-UL and 77 UL samples.

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

37

The above data is from the GEO repository with link provided above. The data was published January 1, 2004 with no listed research article associated with it judging from the GEO citation information missing. The GEO Access ID is GSE764 using GEO platform GPL80. This data does not have an attached study that was published. The data for this GEO series has seven UL gene samples and four healthy non-UL gene samples all from humans. This data is similar to another data series in comparing tissue from not only UL and non-UL tissue samples, but also using

cancerous LMS. The data sets all clearly identify which data sets beginning with 'GSM' and followed by the data set ID number are either 'Myo' for healthy myometrial gene samples, 'Leio' for UL gene samples, 'LMS' for leiomyosarcoma, or 'exULMS' for extra cancerous uterine leiomyosarcoma. Only the data sets that begin with 'Myo' or 'Leio' will be used. From this GEO database there will be gene samples from four healthy myometrial and seven uterine leiomyomas. This data contributes to the accumulation of data on healthy non-UL and UL gene samples for researching high differentially expressed genes between normal and UL human tissue. There will be a total of 55 healthy or normal myometrial gene samples and 77 UL gene samples after all six of the GEO data studies are collected and combined into one data set.

R (2019). CRAN: Comprehensive R Archive Network. R, version 3.5.2, for Windows 64-bit Operating System. Retrieved March 3, 2019 from: https://cran.cnr.berkeley.edu/

This source is the R statistical programming software that will be used to install Bioconductor, and access statistical and machine learning package libraries as needed to conduct research on gene expression data from the GEO online data repository. This version is for the Windows 64-bit operating system and is 79 megabytes in size. Bioconductor has to be installed from within R once installed. The basic R packages used will be Gviz and heatmaply. The R Bioconductor packages that will be used are the ggbio and DESeq packages.

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

The research done in this article involved a meta-analysis of two GWAS studies of UL using Icelandic and English European females. The patients with UL are the case group and the volunteers without UL are the control group. There are two separate studies in this research. One study is on genes expressed in cancers and other benign tumors that are also expressed in UL. The other study is on the putative loci regions associated with hormone related diseases and the changes in those loci in UL. This research elucidated the relationship that hormones have on UL growth and made explicit the common genes being expressed between UL, cancer and benign tumors elsewhere in the body. The information about hormonal therapy to treat symptoms of estrogen responsive leiomyomas before hysterectomy can cause the symptoms to recur was found in this article. The TNRC6B and the BET1L genes were found to also be associated with UL in this study. But the BET1L gene found in Japanese populations and the CYTH4 gene found in African American populations were not found to be associated with UL in this study on European women. This study on Europeans confirmed one of the endometrial cancer genes associated with UL is r10917151 of CDC42/WNT4. This study found reason to exclude the other seven genes previous GWAS studies found to be associated with UL and cancer. This study also shows in a table of polygenic risk scores that there is a significant association of ULs in patients with thyroid cancer (R-squared = 21% and P value: 3.0*10^-5), endometriosis Stage III and IV (R-squared = 11% and P value =4.1*10^-3), and kidney cancer (R-squared = 10% and P value = 2.43*10^-3). This research on Europeans, shows a connection between thyroid disfunctions and thyroid cancer that the research done on African American populations also found to be associated with UL risk.

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. PMID:16319128

This reference is used for the data collected from the GEO online data repository at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi The GEO Access ID is GSE2724 using GEO platform GPL96. This data was made publicly available December 20, 2005. In this data series, there are seven UL samples and 11 non-UL healthy gene samples that will be added to the data of non-UL and UL gene samples for studying gene expression between the two groups excluding race as a factor in otherwise healthy patients. This original data series has labeled the UL samples to include an 'm' for fibroid tissue, and an 'n' for normal myometrial tissue. The 'm' is actually for mutated fumarate hydrase (FH) as this study was examining the mutated FH in UL to find a connection of FH expression in UL. When this data is added to the data from the other GEO data on non-UL and UL gene samples there will be a total of 132 gene samples comparing 55 normal uterine tissue to 77 UL tissue.

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

This is an R and Bioconductor package that will allow colorful and detailed visualizations and analysis of the six combined GEO microarray data of UL and non-UL genes. The plotting visualizations involve circular plotting, gene track plotting, nucleotide mapping, and more. The gene IDs used in this example are 'hg19' a GEO platform version for human genome 19 of ENSEMBL as Bioconductor is built to use GEO data.

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

This article uses data mining to extract gene expression data from the microarrays of the ArrayExpress gene expression data repository and the GEO database. Both are available online. In total this research gathered 45 samples of four different phases of fertility in women during the implantation window of fertility. This study and the study by Hodge et al. (2012) are two studies that perform gene expression analysis independent of race for UL donor tissue origin and of having only a UL. Almost 200 potential biomarkers were discovered and tested to predict the likelihood of a woman capable of becoming pregnant to get pregnant. The experimental study of how the samples were collected that were uploaded into GEO and ArrayExpress was detailed and bioinformatic statistics was conducted using a software called GeneSifter. This software was used to perform differential gene expression analysis and validate for fold-change of at least 10 so that the most highly changed genes could be compared. Principal Component Analysis (PCA) was done to evaluate the multivariate data complexities as well as hierarchical clustering after PCA on the data files to conditionally categorize the genes. Afterwards, a software program called Ingenuity was used to compare the fold-change values of genes at least equal to two for the genes that are potential biomarkers along the canonical pathways the Ingenuity Pathway Alignment library produced. This was verified by Fisher Exact Test for P values of at most 0.05. Network analysis was then used on the top 35 genes having at least a fold-change value of 10 from the GeneSifter analysis done earlier to visualize the network of connections each gene has to other genes using the algorithm generated by Ingenuity software. To validate the data, genes that were selected for validation were either significantly different from the same gene in another array, had a very low P value, or a high fold-change in the

previous analysis. Statistical analysis of the genes being validated was performed using the regular expression algorithm, the Least Significant Difference (LSD), and one-way analysis of variance. Those genes resulting in a P value less than or equal to .05 were labeled statistically significant. The chromosomes 4, 9, and 14 were found to be the chromosomes with the most differentially expressed genes, and these chromosomes are not in any of the current studies reviewed for UL risk exclusively by race. The other studies having ubiquitous genotypes associated with UL risk are found on Chromosomes 22, 11, 17, and 10 (Hellwege et al., 2017; Eggert et al., 2012; Rafnar et al., 2018; Bondagji et al., 2018; Liu et al., 2018; Cha et al., 2011; Aissani et al., 2015; Hodge et al., 2012; Edwards et al., 2013).

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

This is a GEO data series of information published July 24, 2010. GEO can be accessed at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi and using the GEO Access ID of GSE23112 in the search box will give a list of the data sets in this study. The platform used in GEO is GPL96. The study examines gene expression qualities of micro RNA (miRNA) and messenger RNA (mRNA) in tumorigenesis and tumor regulation of UL. In total this data contributes 10 samples of five healthy non-UL gene samples and five healthy UL gene samples. After the addition of these data sets to the other five GEO data sets on normal and UL gene samples, there will be a database of 55 healthy uterine myometrial gene samples and 77 UL gene samples.

# Revised Proposal, Prospectus, and Annotated Bibliography_Janis

ORIGINALITY REPORT

| 80% | 12% | 12% | 80% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to Lewis University<br>Student Paper | 77% |
| 2 | Submitted to Western Governors University<br>Student Paper | 2% |
| 3 | espace.library.uq.edu.au<br>Internet Source | <1% |
| 4 | sydney.edu.au<br>Internet Source | <1% |
| 5 | ddd.uab.cat<br>Internet Source | <1% |
| 6 | hal.archives-ouvertes.fr<br>Internet Source | <1% |
| 7 | www.ingentaconnect.com<br>Internet Source | <1% |
| 8 | support.bioconductor.org<br>Internet Source | <1% |
| 9 | bioinfo.na.iac.cnr.it | |

| | | |
|---|---|---|
| | Internet Source | <1% |
| 10 | helda.helsinki.fi<br>Internet Source | <1% |
| 11 | www.illumina.com.cn<br>Internet Source | <1% |
| 12 | elifesciences.org<br>Internet Source | <1% |
| 13 | www.nature.com<br>Internet Source | <1% |
| 14 | academic.oup.com<br>Internet Source | <1% |
| 15 | Virgil S. Bideau, Angela T. Alleyne. "A preliminary study of fatty acid synthase gene and the risk of uterine leiomyoma in an Afro-Caribbean female population", Meta Gene, 2019<br>Publication | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | Off | | |

# Revised Proposal, Prospectus, and Annotated Bibliography_Janis

GRADEMARK REPORT

FINAL GRADE

12/15

GENERAL COMMENTS

## Instructor

You have defined data that you will pull together, and it is clear that you propose to combine several smaller studies into one bigger analysis. However, at this time, there are some parts that remain unclear.

1.) What is your overall goal for this project? In part, the ambiguity seems to come from a lack of clarity between SNP and gene expression. Please make sure to clear up this aspect so there is less ambiguity about what you are aiming to achieve.

2.) Aside from having numerical expression data, what other information will be present in your data set? How will this other information be important for analysis?

3.) Machine learning is vague -- what do you hope to accomplish? How will you define training versus test sets?

PAGE 1

### Comment 1

Check the structure of this sentence -- the phrase before and after the comma are not properly joined together.

### Comment 2

Please clarify here. "Hormone therapy" is typically used to describe the giving of a hormone. As far as I understand (please correct me if I am wrong), UL tumors are promoted by the presence of estrogen; however, the giving of certain hormone agonists (which work by decreasing the amount of active estrogen) can reduce UL. I think the confusion in my reading this is the result of an over-simplification in your description -- moving forward, be as specific about the mechanism as possible so that the responsiveness of the tumor is clear.

PAGE 2

## Comment 3

FYI about comments to follow: In the paragraphs that follow, you summarize two different classifications of data observations. In some places, you are conveying information about SNPs (which means that there are differences in the DNA code associated with the gene), while in other places you discuss information about changes not in the DNA code of the gene, but rather alterations in how much the gene is expressed. I am attempting to make sure that I understand your proposed research study -- so to help me with organizing the data, I am going to label the data as one of these two categories...and then I will evaluate later in the proposal.

## Comment 4

SNP

## Comment 5

SNP

## Comment 6

SNP



## Comment 7

SNP

## Comment 8

SNP

## Comment 9

Gene Expression

## Comment 10

SNP

## Comment 11

I'm guessing this is SNP?

## Comment 12

Gene Expression

## Comment 13

Gene Expression

PAGE 4

## Comment 14

Gene Expression

## Comment 15

Implied SNPs

## Comment 16

SNPs

## Comment 17

SNPs

## Comment 18

SNP

## Comment 19

Gene Expression

PAGE 5

## Comment 20

SNPs

## Comment 21

My concern about what you have provided here is that you seem to interchange the concepts of SNP and gene expression. SNPs are instances where there is a single nucleotide change in the DNA code for a given gene. There can be many consequences of having a SNP incorporated -- sometimes this can lead to a mutated protein that is generated by the gene which may alter the function or work that gene is able to accomplish (it may work "better" or "worse" relative to the wild-type ("normal") sequence. Sometimes the SNP could lead to a truncation, which means the protein is not generated (or cannot perform its task), or could lead to a protein that is not regulated normally, so there are different dosages of protein present. The only way to detect a SNP is by having sequencing data.

Another concept is gene expression, which is often measured by the amount of mRNA that is present in the cell at a given time -- this can be assessed using a microarray (which seems to be the type of data you propose working with). Sometimes, SNPs may have a consequence of having altered expression, but that is not always the case. There can be other reasons that

gene expression is changed, without a SNP being present. Thus, you cannot make the statement that SNPs and gene expression changes are exactly the same.

The methods that you provide give some information about what you propose to do with looking at differential gene expression. What is unclear is how this is connected to the background information you provided about SNPs. As a result, the purpose/objective of your work is not clear. Are you proposing to look at genes where there are known SNPs to see if there is also difference in gene expression? Are you looking for changes in gene expression without any specific link to SNPs? In your report during week 2, please clarify the intention of your project and make a clear distinction between the importance of considering SNPs versus gene expression.

PAGE 6

## Comment 22

In the introduction section, you will need to make sure the research objective for this project is clear.

## Comment 23

It is important to make a clear distinction between genotype (which allele version(s) of a gene are present) and the level of gene expression.

PAGE 7

## Comment 24

To determine a genotype, it is necessary to have sequencing data, which is different from microarray expression data.

PAGE 8

PAGE 9

## Comment 25

Section C?

To this point, you have not provided the reader with background information about why we want to know more about gene expression. Ultimately, the functionality in a cell (normal or diseased) is not just about the coded information in the DNA, but also how much of the corresponding gene product is made. What is known about differences in gene expression in different UL cases?

## Comment 26

What are the specific identifiers for these data sets? I had a hard time looking up the sets to check what they contain without this information.

PAGE 10

PAGE 11

## Comment 27

What kind of data will be present in this large data set? What is included in the "meta data" category? Will you only be comparing numerical attributes? Does your dataset also include categorical information beyond UL vs. non-UL (such as ethnic/ancestral group, age, etc.) . Please make sure to describe moving forward.

## Comment 28

How will you select data for training versus testing?

## Comment 29

What type of models do you anticipate generating?

PAGE 12

## Comment 30

You will also need to consider limitations of your work.

PAGE 13

PAGE 14

PAGE 15

PAGE 16

## Comment 31

You'll need to make sure you are exclusively studying human, as combining different species can be difficult to interpret.

## Comment 32

Please link directly. Instructions for how to do so can be found here: https://www.ncbi.nlm.nih.gov/geo/info/linking.html

PAGE 17

## Comment 33

You will need to specifically designate which individual data sets you have selected and will use to make your master set.

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

### Comment 34

Please link directly. Instructions for how to do so can be found here: https://www.ncbi.nlm.nih.gov/geo/info/linking.html

PAGE 23

### Comment 35

Please eliminate the use of PMID...but if you would like to add a digital tag, report the doi.

### Comment 36

Please link directly. Instructions for how to do so can be found here: https://www.ncbi.nlm.nih.gov/geo/info/linking.html

PAGE 24

### Comment 37

Please link directly. Instructions for how to do so can be found here: https://www.ncbi.nlm.nih.gov/geo/info/linking.html

PAGE 25

PAGE 26

PAGE 27

### Comment 38

Please link directly. Instructions for how to do so can be found here: https://www.ncbi.nlm.nih.gov/geo/info/linking.html

PAGE 28

PAGE 29

### Comment 39

Please link directly. Instructions for how to do so can be found here: https://www.ncbi.nlm.nih.gov/geo/info/linking.html

## PROGRESSION (34%) 4 / 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 |

## PURPOSE (33%) 3 / 5

| | |
|---|---|
| 5 (5) | Main purpose clearly evident throughout, with all aspects clearly connected, and demonstration of extensions that will be made from the existing literature |
| 4 (4) | Above average expression of the main purpose, ideas well defined and connected with above average demonstration of knowledge, extensions that will be made. |
| 3 (3) | Adequate expression of main purpose, connection of ideas, and allusion to extensions that will be made |
| 2 (2) | Expression of the main purpose and included concepts inconsistent and somewhat disconnected; extensions that will be made stated, but not clearly tied to publication record |
| 1 (1) | Inadequate expression of main purpose; ideas unclear and disjointed; not clear how new contribution will be made |

## BIBLIOGRAPHY (33%) 5 / 5

| | |
|---|---|
| 5 (5) | Exemplary report of at least 10 references (mostly primary sources, supplemented by additional resources) with full citation (APA format) and summary of how each will support the project |
| 4 (4) | Report of at least 10 references (even number of primary and secondary sources) with full citation (APA format) and summary |
| 3 (3) | Report of 10 sources, but not in correct format or incomplete summary |
| 2 (2) | Less than 10 sources given, but correct in format, summary |

| | |
|---|---|
| 1 (1) | Insufficient number of sources, inadequate information provided |