Janis Corona

BIOL 59000: Data Science Project for Life Sciences

Week 5 Rough Draft Prospectus

July 3, 2019

Meta-Analysis of the Genes Ubiquitously Associated with Human Uterine Leiomyoma Development in Healthy Humans Using the Gene Expression Omnibus Data

**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 Gene Expression Omnibus (GEO) data across all races. Other studies on UL development used race specific studies of UL development to find genes having genotypes associated with UL risk within each race. The distinct difference between this study and previous studies is using non-race specific gene expression microarray samples instead of solely samples from one specific race to find or not find those same ubiquitous genes 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 significant in determining UL risk in any female patient in the chromosomal regions where these top genes reside.

1. The ubiquitous genes having an association 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).
   1. These ethnicity specific studies and other UL studies unrelated to ethnicity found certain genes in RNA samples of UL and non-UL 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).
   2. The contribution to UL research this project will obtain is to discover if these same ubiquitous genes are found in five 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).
2. The software for analyzing the UL data after combining the data into one data set of microarray data.
   1. R using RStudio (R, 2019)
   2. Bioconductor for R (Bioconductor, 2019).
3. **Background**. This section describes the ubiquitous genes currently associated with Uterine Leiomyomas and explains what they are and the impact they have on lives.
4. **Description of uterine leiomyomas (UL)**. This sub-section describes who UL affect, the known risk factors for UL, and what known genes are significantly associated with UL.
   1. **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).
   2. **Significant genes for UL**. This sub-sub-section lists the genes associated with UL development in women. These genes 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).
5. **Chromosomal location of UL associated genes**. This sub-section lists and describes each chromosome having genes significantly associated with UL.
   1. **Chromosome 11**. This sub-sub-section describes the BET1L gene 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). 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).
   2. **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).
   3. **Chromosome 17**. This sub-sub-section describes the two genes 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 genes and found no significance in UL for Saudi Arabian populations (Bondagji et al., 2017).
   4. **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 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 considered significant for UL risk in African Americans, will be described in this sub-sub-section as well as the other UL risk gene on chromosome 22 called TNRC6B (Hellwege et al., 2017).
6. **Materials and Methods**. This section will describe the material sources used and the type of software used to analyze the data (D. Zhang, Sun, Ma, Dai, & W. Zhang, 2012; Eggert et al., 2012). The ubiquitous genes found most differentially expressed between non-UL and UL samples of gene expression data will be used to build a model to test on UL risk prediction. The entire samples will be partitioned into 70 per cent for training a model and 30 per cent for testing that model. The model will be based on a known algorithm such as linear, random forest, k-nearest neighbor, and others to predict whether any sample in the testing partition is a UL sample or a non-UL. Using a Bioconductor package called Gviz, produce a layout to visualize the UL risk genes on the chromosomes they reside next to the neighborhood of genes that also show up in the GEO data on that chromosome (Hahne, 2019).
7. **GEO data of UL and non-UL samples**. This sub-section explains the use of GEO to access five 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; Vanharanta et al., 2006; Zavadil et al, 2010). These ‘in situ oligonucleotide’ or microarray platforms were derived from five separate studies (Hoffman et al., 2004; Vanharanta et al., 2006; Zavadil et al., 2010; Crabtree et al., 2009; Miyata et al., 2015).
   1. The combined data sets will combine into one big data table of 51 non-UL and 70 UL observations for analyzing using R.
   2. 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 genes mapped next to the other runner up genes not further explored in the population studies.
8. **R Statistical software for statistical analysis**. This sub-section describes how R software will be used in this research.
   1. R to transform the five independent microarray data sets into the same scale as one data set GSE68295 used the log 2 scale.
   2. 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).
      1. Four of the array data sets have the same probe IDs but one of the array data sets has different IDs that can be aligned to the four other sets by using the platform array data containing the alternate gene names.
      2. The gene symbol field of all five data sets aligned to their respective platform is the item that will be merged on.
      3. 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.
      4. R will then extract only those observations that the cytoband location of each ubiquitous gene is on 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.
   3. The final data table produced with R for analysis and prediction model training will have 183 gene observations and 121 fields including 1 ID field, 70 UL samples, and 51 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 compared to the top genes found in the GEO data for UL risk in healthy females (Galili, O'Callaghan, Sidi, & Benjamin, 2019).
9. **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 with visualizations (Bioconductor, 2019; Anders, & Huber, 2019; Hahne, 2019; Yin, 2019).
10. **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 genes are associated with UL risk 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 genes posing 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). Limitations of this study will be discussed in this section, such as not exploring the actual SNPs of the top genes associated with UL risk, not having a large number of UL and non-UL samples to compare, and having to exclude some samples in GSE764 because four of the six ubiquitous genes were not in that data set (Quade, Mutter, & Morton, 2004).

1. **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 five microarray UL and non-UL samples, and the analysis script used to create the visualizations in R and Bioconductor.
2. **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

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., 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

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (March 2018). 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. (November 2015). 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.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.

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

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. (January 2019). 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.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.

Karatzoglou, A., Smola, A., Hornik, K., Maniscalco, M., & Teo, C. (August 2018). 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. (April 2019). 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. (January 2008). 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. (June 2019). 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. (April 2019). 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>

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. (June 2019). 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