**Slide 1:** Title

**Slide 2**: UL are benign tumors of the myometrium of the uterus in females. Females can get more than one UL and can grow in different layers of the uterus, such as the subserosal, submucosal, and intramural layers of the uterus.

Symptoms of UL heavier than normal menstruation, irregular menstruation cycles that last longer than normal or are more frequent than a normal menstrual cycle. Other symptoms include painful cramping, and obvious abdominal growth or enlargement as the UL grows.

**Slide 3:** This is an image of how a UL can grow in different layers of a uterus. The submucosal layer UL projects into the uterus, the subserosal layer UL projects outwards from the uterus, and the intramural UL is in the muscle layer of the uterus.

**Slide 4:** Symptoms of UL heavier than normal menstruation, irregular menstruation cycles that last longer than normal or are more frequent than a normal menstrual cycle. Other symptoms include painful cramping, and obvious abdominal growth or enlargement as the UL grows.

**Slide 5:** Female mammals such as humans and mice get UL.

Risk factors for UL include: having a low age at menarche, being of child-birthing age, being pre-menopausal, being black or European decent, being obese, drinking alcohol, and having a disease such as kidney disease, endometrial cancer in stage III or IV, and having a thyroid dysregulation.

**Slide 6:** Treatment for UL involves an estrogen analogue to inhibit gonadotropin hormone production such as Leuprolide, or an estrogen antagonist like cetrolexin to compete with the estrogen binding sites on a UL. These all stop the UL from growing, but the only treatment to end the symptoms of UL is a myomectomy or a hysterectomy. Removal of the UL (myomectomy) does not keep the UL from coming back.

UL development is still being researched, but there are six genes ubiquitous to the UL risk within populations.

**Slide 7:** TNRC6B: UL risk factor for all above, but also associated with volume or the size of the UL in Chinese and European Americans.

BET1L: UL risk factor for all above, but also associated with the number of fibroids in Chinese and the location of the fibroid as either sub, sera, or intra-mucosal of the uterus in European and European Americans.

CYTH4: UL risk in AA populations.

**Slide 8**: FASN: UL risk in EA, Austr., Icelandic, and other Europeans.

CCDC57: UL risk for EA, Austr. , Icel., other Euro.

HMGA2: same as other two

**Slide 9:** Gene transcription and translation in the cell is shown.

The transcription occurs in the nucleus and the translation occurs outside the cell.

Transcription selects which genes to express or inhibit in the cell due to environmental factors and stress of some sort such as chemical, radiation, diet, time of day, and current health condition or stage of life. Changes in gene expression are mediated by the number of protein copies made through translation (Dr. Sara Powers-Lewis University).

**Slide 10**: Gene expression can be affected by heat, UV light, environmental disturbances, chemical disturbances, and other stresses like diet, disease, stage in life, current health, and time of day.

Gene expression of DNA into mRNA will happen at the transcription phase of gene expression and is one way that expression or inhibition of genes can change.

When outside the cell and ribosomes translate the mRNA into amino acids and proteins at the translational phase of gene expression there can be changes as well. These proteins have a shelf life in the cell and are regulated so that not all of the proteins are being used, but these proteins also mediate the changes in gene expression.

Gene expression of the microarray samples across UL and non-UL samples could indicate some gene targets for treating or further researching UL risk to begin alternative therapies to treat UL symptoms or prevent UL pathogenesis from occuring.

**Slide 11**: This image shows how the transcription in the nucleus of the cell happens due to stress or adaptive changes in the environment. There are 23 chromosomes of which 22 are for genes and the sex chromosome is for duplicating life ultimately.

Referring only to the 22 chromosomes that genes are expressed the chromatin is the tightly coiled DNA that wraps around the nucleosomes of histones so that genes needed for expression or to be inhibited can be activated or blocked by proteins in the cell.

The DNA is packed tightly in the nucleus and stress causes the enzymes and proteins in the cell to inhibit or express the genes along the DNA by weakening the tight chromatin wrapped histone and nucleosomes to bind to the DNA, create mRNA and leave the nucleus to the cytoplasm to be translated by tRNA and rRNA or ribosomes to form proteins

The DNA is then expressed as mRNA in the nucleus and exported into the cytoplasm to be translated into proteins by ribosomes.

**Slide 12**: Gene expression is a result of changes at the transcription or translational process of protein creation and can be a result of stress in the environment.

Current UL risk studies have examined gene expression data and found gene targets to be TNRC6B, BET1L, HMGA2, CYTH4, CCDC57, and FASN as hereditary determinants of UL predisposition and as targets of UL pathogenesis. Not all these genes are significant for UL risk in each population.

Cytoband location of six of the genes has shown that CYTH4 and TNRC6B share the same cytoband location 22q13.1 on the same forward strand with other genes. The genes FASN and CCDC57 also share the same cytoband location on the reverse strand of 17q25.3, while BET1L and HMGA2 have different cytoband locations of 11p15.5 and 12q14.3. Are there other genes along these same cytobands that have an association to UL? Given that chromosomal location is independent of gene expression, are there other genes that reside on these same cytoband locations that also associate with UL pathogenesis?

Can a set of data including these cytoband locations and the six genes ubiquitous to UL risk studies predict a UL sample using machine learning on a sample of UL and non-UL data from the Gene Expression Omnibus of five separate studies?

Slide 13: Five different UL risk studies in the Gene Expression Omnibus (GEO) online gene expression data repository were downloaded and merged together to get a data set of 121 total samples of 70 UL and 51 non-UL gene expression observations from non race identified donors. The UL samples were taken from cells of UL and the non-UL samples were extracted from healthy myometrial uterine cells. Some cells were from the same donor’s uterus after a myomectomy or hysterectomy where the UL cells were extracted and the cells next to that UL in healthy myometrial tissue were extracted.

When combined there were 12, 173 total genes in common between the two types of samples that were filtered to work with only the genes on the same cytoband location as the six gene ubiquitous to UL risk studies.

The rational for filtering for only the four cytoband locations that these six genes reside is because the UL risk studies that confirmed these genes to have significant association to UL in specific populations mentioned in a few analysis that other genes along the same cytoband location had been ignored to test the genes significance for UL in that studies population of Japanese and Saudi American females (Cha, et al., 2011; Bondagji, et al., 2017).

**Slide 14**: Use R, lattice, to visually see any relationships between ten of the genes in the 130-subset having the most change in UL. This subset is of cytoband locations only belonging to the six genes ubiquitous to UL risk studies.

**Slide 15:** This is a pairwise comparison plot of genes that are expressed or inhibited the most in UL samples compared to non-UL samples. Visually this seemed to be a shot in the dark to see patterns or relationships to other genes. None really move with the others, and some like NOL12 and SYNGR1 seem to stay at the same level of expression for each gene it is compared to. The scatters are from the 51 non-UL samples. The UL samples were similar in comparison.

**Slide 16:** The method for grabbing the top 10 differentially expressed genes by magnitude and adding the six genes ubiquitous to UL risk studies was to build a data table that would show how well the six ubiquitous genes compare to the genes along the same cytoband location they reside to see if they can produce good machine learning results on predicting a sample as UL or not. A total of 130 genes reside in the same cytoband locations as the six genes ubiquitous to UL risk studies from a universe of 12,173 genes in common.

**Slide 17:** The R package dplyr was used to create new fields for UL and non-UL means of the subset of 130 genes common to the same cytoband locations as the six ubiquitous UL risk study genes and including those six genes. The difference in the two as Differentially Expressed, and the magnitude of those values were two additional fields added, so that when ordering most to least expressed, those genes having very low expression (showing more inhibition of gene expression in UL) weren’t discarded.

**Slide 18**: The gene ID field was saved as a vector of 130 genes as character values to be the column names of the data table once transposed. The fields other than the sample ID were removed and ordered to be the first 51 non-UL samples and the next 70 samples of UL from the total set and labeled with ‘ul’ at the end by sample type. This table was transposed so that it had 121 samples as row observations and 130 genes as column values after relabeling the header names with the Gene ID names saved first. Then a field was added labeled ‘TYPE’ that indicated whether the sample was a non-UL or UL sample by using a data frame vector to repeat ‘nonUL’ 51 times, then add to it a data frame vector of ‘UL’ repeated 70 times. This produced a machine learning data table that was 121 samples as rows by 131 columns for the 130 genes and 1 TYPE field of UL or non-UL.

Rational for ten: ten was chosen randomly, but the idea in math is related to the squeeze theorem for upper and lower bounds to squeeze out the limits of variable x. In life, it would be called trial and error, to find a subset of a bigger set and shrink it down to better categories. This was the first test run of a subset to gauge the first subset of predictors to include the six genes ubiquitous to UL risk studies.

**Slide 19**: This is a table of the TOP16 genes with their human genome nomenclature symbol and their descriptive names.

**Slide 20:** Bootstrap results, for each UL and non-UL of each gene with standard errors, were added to the data of TOP16 Genes.

This was to see how well these samples could represent the population of genes in UL and non-UL samples.

Since the sample is just that a sample of a much larger population, it is idea to see how well the Central Limit Theorem and the Law of Large Numbers can statistically recognize these 16 genes to be representative of the large population this data doesn’t have by using simulation randomly with replacement 10,000 times on each gene in each set for UL and non-UL samples. There were 51 non-UL samples and 70 UL samples. This data working with tested the ten genes with the most change in gene expression on the same cytoband location as the six genes ubiquitous to UL risk studies. This set was the only set that simulated a population mean for each gene.

Slide 21: This is a table of those 16 genes in the TOP16 after running bootstrap simulations on each gene for the UL and the non-UL means, the difference in means (DE), and the standard error (sd).

**Slide 22**: Histograms to show how symmetric each of the TOP16 genes were in the population using the Bootstrap simulated means in R was done, to lead up to how well these genes would be as predictors of UL

**Slide 23:** Most of the TOP16 (10 most DE+6 ubiquitous genes) are symmetrical, but some are skewed where the median is on the right or left of the mean making one of the tails appear to point in the opposite direction, such as with GRIP1 it is right skewed) having the median on the left of the mean making the tail longer on the right. BET1L is almost perfectly symmetrical and so is CANT1.

**Slide 24**: A visualization of the TOP16 genes by chromosomal location was needed to see how the simulated means of the UL and non-UL means compare, and to see where each of the chromosomal genes is above or below the line in comparison. This was to see if there was an indication of chromosomal location being higher in magnitude of change for UL than non-UL in the population based on this sample.

This would show that there is a good reason to look further into these chromosomes and uncover whether some other genes in the same neighborhood of these genes play a role in UL pathogenesis.

**Slide 25:** This plot was made using ggplot2 in R. It shows the 16 genes of 10 most DE and 6 ubiquitous genes to UL risk studies in the subset of 130 genes living in the same cytoband locations as the six ubiquitous genes, or the TOP16 genes for short. All of chromosome 11 and 22 genes have higher expression means than the non-UL means from the 10,000 simulations and all genes on chromosome 12 have higher non-UL simulated means than UL means. Chromosome 17 shows simulated means mixed for UL and non-UL means.

This result of the simulated means between UL and nonUL samples would point to FSCN2 as a possible gene target for UL pathogenesis because it is expressed more in non-UL than UL for the simulated population means. This could mean it is inhibited in UL and is either a defense the body uses to stop producing as much FSCN2 or is not produced as much because some other gene is blocking its production.

The FASN and CCDC57 genes are also expressed less in non-UL than UL samples, and both belong to the same cytoband of chromosome 17. (q25.3)

The TNRC6B and CYTH4 genes are expressed more in UL than non-UL samples and both belong to the same cytoband of chromosome 22. (q13.1)

**Slide 26**: The gene expression data is continuous, numeric data. So it made sense to use most of these algorithms, and some were used to see how these continuous values would be grouped to predict the outcome of the testing set in determining if the sample is UL or non-UL. The training set of 70% equal to 85 samples to build each model. This same training set was used on every model by setting a seed value in the R software to make random selections in partitioning the same exact samples from each data set with each run of the program, for reproducibility of original results. The testing set was set aside as a partition of all the samples of 30% equal to 36 samples and also remained the same testing set for each algorithm used.

These algorithms are produced in R, mostly with the caret package.

The LDA method is a collapsed Gibbs sampling method called Latent Dirichlet Allocation is used for categorical grouping of text data called Topic Modeling that did score surprisingly well next to these other algorithms.

The RF and RF2 algorithms use a decision tree method to classify samples categorically.

The GBM model uses the Adaboost and gradient boosted regression models that rely on least squares, logistic, poisson, quantile, multi-nomial logistic, and t-distribution loss methods. GBM uses the bootstrapping method for simulation of the genes in the sample and creating a regression on the numeric data inputs on the categorical output

The K Nearest Neighbor uses the centroid of groups of data as unsupervised data points or non-categorized data to keep samples in the group of other samples closest to their characteristic groupings.

The Rpart method uses decision trees by recursive partitioning of the data.

GLM is a generalized linear regression model. The combined model uses the outcomes from all algorithm results as a dataframe to select the most frequent prediction or best prediction.

The data being analyzed is continuous, numeric data. Any of these algorithms would be able to regress numeric values or categorize by clusters or decision trees. There is one outcome variable for the type of the sample, and many predictor variables as genes selected. The random forest and k-nearest neighbor can group by ranges of values and creating subsets within sets to predict an outcome based on the training set build up of the model and the testing set to test that model’s accuracy.

**Slide 27:** Results above are for the TOP16 genes in the subset of 130 genes belonging to the same cytoband locations as the six genes ubiquitous to UL risk studies. The combined prediction model scored the best with an accuracy of 83 per cent and the next best algorithm is the LDA with a score of 74 per cent accuracy in predicting UL or non-UL for the TYPE of sample. The furthest left column is the sample name in GEO. The last column is the true value of the sample as UL labeled ‘UL’ or non-UL labeled ‘nonUL.’ The header is the name of the algorithm used in R. predRF is the caret random forest, predRF2 is the randomForest package, the predlda is the caret Latent Dirichlet Allocation algorithm, the predKNN is the K-nearest neighbor algorithm in caret, the predRPART is the recursive partitioning and regression tree algorithm in the rpart package, the predGLM is the generalized linear model of MASS package, and the CombinedPredictions2 is the caret packages best results of all algorithms used in R.

**Slide 28**: GEO data from five studies on UL risk analyzed and combined to one set of 12,173 genes universally in common. Then, this data set was then used to gather a subset of 130 genes that are in common among the five studies but also only on the same chromosomal locations as the six genes ubiquitous to UL risk in current studies. All data sets had the same fields for means, magnitude, but also fold change added other than the first dataset, TOP16.

The genes in this dataset are the subset of 16 genes that have the greatest magnitude of change in UL compared to non-UL.

**Slide 29:** The data table of the machine learning ready data table with the 16 most expressed/inhibited genes out of the subset of 130 having the most magnitude of change is shown above.

These genes are: "FSCN2" "CARD10" "GRIP1" "CANT1" "IRF7" "ARHGDIA" "NOL12"

"SLC25A10" "SLC38A10" "RNH1" "SYNGR1" "TALDO1" "FN3K" "POLR2F"

"LLPH" "SMCR7L"

**Slide 30:** The results using highest expression/inhibition of subset 130 cytoband genes was 92% combined and best of 92% using KNN, this leads to belief there are some good gene targets in this set that don’t include the 6 genes ubiquitous to UL risk studies. These results also show some other genes in the cytoband location also associate to UL risk in some way.

**Slide 31**: The least DE were used in UL to non-UL to test if the machine learning algorithms were just good on all genes or are worse as expecting for the genes with the least change in UL to non-UL.

**Slide 32:** The 16 least expressed/inhibited genes by magnitude in the subset of 130 genes belonging to the cytobands of the six UL risk genes. The Type field show nonUL because it is not scrolled down to the 121 samples completely.

"PSMD13" "DNAL4" "ATHL1" "CDHR5" "SIGIRR" "MRPL23" "FOXK2"

"RASSF7" "SIRT3" "GRAP2" "AATK" "TH" "TMEM184B" "WDR45L"

"KDELR3" "GCGR"

**Slide 33:** The least DE of subset 130 genes in cytobands scored poorly as expected, with a combined score of 72% and best score of 58% using rpart. Not a good choice to use as UL risk gene targets.

**Slide 34**: The cytoband 130 genes having the most fold change and including the top 6 ubiq genes is used to test UL risk gene targets

**Slide 35:** The 16 genes with the highest fold change in the subset of 130 genes is displayed.

"CANT1" "ARHGDIA" "SLC25A10" "ATF4" "CARD10" "MRPL12" "BET1L" "RNH1" "NOL12" "IRF7" "DRD4" "TNRC6B" "CYTH4" "CCDC57" "FASN" "HMGA2"

**Slide 36**: The fold change results from the cytoband genes of 130 showed a combined score of 78% with the best score of 72% for GLM and LDA. Algorithms. This isn’t good enough to use as predictors for these genes.

**Slide 37**: GEO data from five studies on UL risk analyzed and combined to one set of 12,173 genes universally in common. The same means, magnitude, and additional fold change field was added as all the 130 gene data sets excluding the TOP16 data set. The data was transposed like all other data sets to make them each ready to be analyzed with machine learning algorithms. Where a 70 per cent partition training set to train the data on that model, and the remaining 30 per cent of data is used to test if the sample is UL or not as the testing set using the added TYPE field to the header of genes when transposed with the samples as observations.

**Slide 38:** Those genes having the most fold change in all the genes in common are used as predictors in a data set with the same training/testing ratio of 70/30 from the samples.

**Slide 39:** Data Table above is of the 16 genes in the universe of 12,173 genes with the highest fold change in UL compared to non-UL samples.

"HSPB1" "DSTN" "S100A6" "CNN1" "ACTG2" "VIM" "SPARCL1" "TPM2" "ACTA2" "PCP4" "TAGLN" "DES" "RAMP1" "CYR61" "UBC" "ACTB"

**Slide 40:** The results of those genes having the most fold change in all 12,173 genes in common between UL and non-UL samples showed a combined model score of 92 per cent accuracy and the best model scored 86% for RF, RF2, LDA, KNN, and Rpart. These genes are good gene target to use in UL risk.

**Slide 41**: The genes inhibited/expressed most in all genes excluding top 6 ubiquitous genes is used as a predictive model data set to test how well these predictors score.

**Slide 42**: Data table above is the 16 genes with the most magnitude of change in UL compared to non-UL out of all 12,173 genes.

"HSPB1" "DSTN" "S100A6" "CNN1" "ACTG2" "VIM" "SPARCL1"

"TPM2" "ACTA2" "PCP4" "TAGLN" "DES" "RAMP1" "CYR61" "UBC" "ACTB"

**Slide 43**: The most expressed/inhibited genes in all genes of 12,173 in common comparing UL to non-UL samples showed a combined model score of 86 per cent, and the best model of LDA scored 81%. These are good gene targets to use for the samples.

**Slide 44**: This is the universe of 12,173 genes in common that are expressed the least and made ready for machine learning with the same 70-30 training/testing ratio of the samples tested in each of the machine learning algorithms used on the TOP16 data set.

Slide 45: “USP32P2" "RCVRN" "SYNGR3" "MORC1" "KLK2" "SUV39H1" "LIG4"

"KLHDC4" "GRIK4" "FABP1" "TLX3" "LAMB4" "DNTT" "VN1R1" "LEFTY1"

"C7orf64"

These 16 genes are the least in magnitude of change in the universe of genes comparing UL to non-UL

**Slide 46:** The least expressed in all genes comparing UL to non-UL samples showed 58% accuracy using rpart as the best algorithm, but a combined model score of 75% accuracy. There is not a need to consider any of the least expressed genes as gene targets for UL pathogenesis.

**Slide 47**: The header to the next table is the prediction algorithm used and the first column is the name abbreviated for each type of data set used, the last column is the true value of the sample being UL or not. Anything with 130 is the subset of genes only on the same chromosomes as the six genes ubiquitous to UL risk studies, and the abbreviated names with ‘universe’ in it are the entire set of 12,173 genes in common being evaluated. None of the universal data sets or the data sets with ‘130’ genes include the six genes in question. The TOP16 include the six genes ubiquitous to UL risk studies.

**Slide 48:** This analysis put together all the results of the entire study to show conclusively whether the six genes ubiquitous to UL pathogenesis studies are good predictors of UL risk by gene expression of UL compared to non-UL samples.

Results above are the results from machine learning predictions made on all data sets derived. The combined model of all seven algorithms performed best with an accuracy of 92 per cent on two separate files of data sets named DE16\_most\_130\_results.csv and the universe16\_fold\_results.csv. These two data sets do not include the six genes ubiquitous to UL risk studies. Those genes having the most change in magnitude as observed in UL samples compared to non-UL samples in the DE16\_most\_130\_results is from the subset of 130 genes in common along the same cytoband locations as the six UL risk studies’ genes. The universe16\_fold\_results data set is from all of the genes in the universe of genes these five GEO studies had in common having the most fold change in UL samples compared to non-UL samples. The next best data sets for best accuracy in prediction scores is from the universe16\_DE\_most\_results with 86 per cent accuracy. It is data on the top 16 genes with the highest magnitude of change between UL and non-UL samples in the universe of combined genes between these five GEO studies. Both data sets that chose the 16 genes having the least change in magnitude scored the worst, making it likely none of these genes are suitable to associate with UL risk. The TOP16 of genes that contains five expressed more in UL and five expressed less in UL from the subset of 130 in common by cytoband location both scored 83 per cent accuracy, which is pretty good on a scale of accuracy compared to the least expressed genes.

**Slide 49**: The data from five different UL studies of microarray gene expression data was combined from GEO and used to derive a TOP16 data set to analyze UL risk genes by expression, bootstrap simulations to show how the population looks showed good symmetry, but comparing means didn’t show any strong relationships, nor by chromosome location which was actually the cytoband location of those six genes significant for UL risk.

**Slide 50:** The data from five different UL studies of microarray gene expression data was combined from GEO and used to derive eight different data sets that could be used to determine if the six genes ubiquitous to current research studies can make an accurate prediction of a sample being UL or non-UL. When including these six genes and the next genes that were the top genes expressed in UL by most magnitude of change an accuracy of 83 per cent was achieved with a first run on a 30 per cent partition (36/121) UL and non-UL samples trained on the other 70 per cent partition (85/121).

**Slide 51**: There is no need to rule out the six ubiquitous genes to predicting UL samples and associating with UL pathogenesis based on previous UL risk studies. The genes having the most or least gene expression in UL compared to non-UL measured by fold change or differential expression in magnitude make good predictors in machine learning algorithms.

Since the best accuracy was obtained by a combined model of algorithms that involved unsupervised or non-categorical groupings (KNN) it proves the machine learning algorithms are useful for predicting outcomes in a given data set. Also, the least expressed genes do show little connection to UL risk or onset.

Testing the middle range of genes’ change in gene expression could have been done to see how well those genes closer to a mid-line of least and highest expressed genes could make for predicting UL.

**Slide 52**: Testing the most differentially expressed genes and highest fold change genes in all genes was best at predicting UL and locating gene targets for UL pathogenesis. Defining the logic for how or why these gene expression changes in all data set combinations show UL risk is a further extension to this study.

**Slide 53**: The many ways that gene expression can be altered from transcription to translation mean that those genes expressed less in UL could be associated with UL pathogenesis and how the body regulates UL development by inhibiting those genes’ production. And for the genes expressed more in UL, it could mean that the body needs more of these genes to protect the body from other diseases that could be related to UL, though no current study this research encountered has observed a UL being a predisposition to later cancers or diseases. Current studies have associated diseases to developing UL but not the inverse, and also having thyroid dysfunction associate with having a UL when using CYTH4 as a gene target. Much more work in UL risk gene targeting needs to be done, with more sampling, more ruling out of genes that could be associated with UL risk, more testing for significance of gene targets already established as having UL pathogenesis in some populations but not in few others. Gene expression does play a pivotal role in how our bodies protect themselves from stresses in the environment and UL development has shown that many genes do change in gene expression values in UL samples compared to non-UL samples. There is enough evidence of this in the predictive models that showed up to 92 per cent accuracy in predicting UL. There is no reason to stop screening microarray gene expression samples to find gene targets to UL risk.

**Slide 54**: references start