Gene Expression Signatures of Endometriosis

Berkalieva, A., Espejo, E., Trikannad, N., and Zhu, H.

Department of Statistics and Division of Biostatistics

University of California, Berkeley

December 6, 2018

Abstract

Endometriosis is an estrogen-dependent condition that affects nearly 10% women of reproductive age and is characterized by chronic pain and inflammation. Currently, it can only be fully diagnosed at surgery and has an average latency period of 11 years before it is diagnosed [1]. Profiling the deviations in gene expressions from normal endometrium to endometriosis is of high value towards understanding the disease and identifying diagnostic and therapeutic targets. We utilized a publicly available microarray data containing archived endometrial samples from women with different stages of endometriosis to derive gene expression signatures of the disease. We use the results of our differential analysis to construct a classification models to predict endometriosis for a given sample.

1. Introduction

Endometriosis is a painful gynecological disorder that affects nearly 10% of women worldwide [1]. It is an estrogen-dependent condition characterized by the ectopic growth of endometrium i.e., the endometrium, which is normally found as a tissue in the inner lining of the uterus, grows and functions outside of the uterus. This displacement of the endometrial tissue causes growths and lesions in the abdomen and pelvic cavity, leading to chronic pain and inflammation that is disruptive to a woman's physical and social well being.

Currently, it can only be fully diagnosed at surgery, therefore adding on average a latency period of 11 years before it is diagnosed. The lack of research in this domain and scarce

knowledge of the physiological underpinnings, that cause and aggravate endometriosis, are major roadblocks in the development of novel diagnostics and therapeutics for this disease.

Recent studies have suggested that abnormalities in the regulation of specific genes [3] are involved in the development of endometriosis and exploring these anomalies is of high value in understanding the disease and identifying diagnostic and therapeutic targets. Our goal is to perform differential gene expression analysis between samples with normal endometrium and endometriosis using microarray data and use those results to build classification models that predict the presence of the disorder.

Tamaresis, et al. [1] developed menstrual cycle phase dependent decision tree models using differentially expressed genes based on t-tests. Bakhtiarizadeh, et al. [2] similarly chose differentially expressed genes. The selection of differentially expressed genes in this manner assumes the normality of gene expressions.

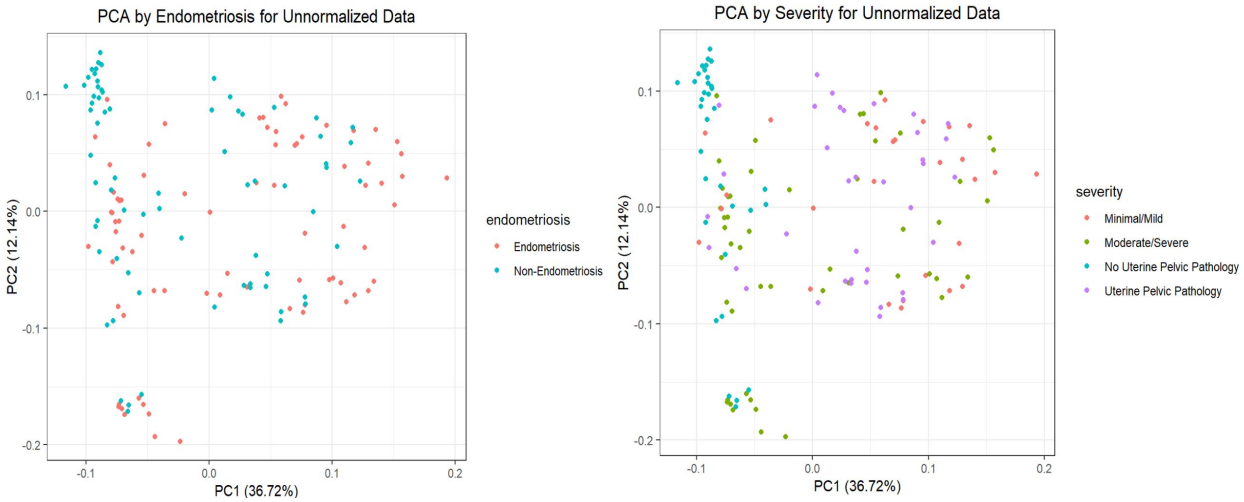
Skew in the endometrial gene expressions violate this assumption. Instead of t-testing for selection of DE genes, we propose to use a well-known method of 2-fold group mean difference [4] and a Mann-Whitney U test to select DE genes. With the two collections of genes selected by these two methods, we aim to develop two sets of machine learning models that accurately classify patients as with endometriosis (E) or no endometriosis (no E) regardless of menstrual cycle state comparably to the classifiers in Tamaresis et al.

By building and comparing three classification models, we will assess which models are most accurate as well as if the selected core genes are useful predictors in outcome of endometriosis. Tamaresis et al. [1] built Classification and Regression Trees (CART) to classify samples into five pairwise outcomes, using menstrual phase cycle and 21,734 probes as their predictor variables. With so many predictor variables applied to CART, overfitting to the data may be an issue. Furthermore, none of the five pairwise classifications included endometriosis versus no endometriosis. As a result, we wanted to see if utilizing other methods used in high-dimensional genomics contexts might perform better at predicting presence of endometriosis without the risk of overfitting. By using K Nearest Neighbors, Random Forests, and Support Vector Machines, we used gene expression levels to predict clinical outcome of endometriosis.

2. Data

Tamaresis et al. [1] analyzed large scale gene expression using microarray data available from UCSF archives measured by the Giudice Lab which contain measurements on 148 subjects and 54,665 probe measurements. The 148 patients were labeled as either with endometriosis or not. Of the 77 patients that had endometriosis, 28 patients were classified as having minimal to mild symptoms and 49 were classified as having moderate to severe symptoms. The remaining patients did not have endometriosis and were classified as whether they had uterine pelvic pathology (*n=37*) or not (*n=34*). Bakhtiarizadeh, et al. [2] studied a subset (*n=105*) of the same array data to examine weighted gene co-expression networks. We accessed the original data from Tamaresis, et al. [1] on NCBI Genbank with accession number GSE51981 and downloaded in CEL format for analysis.

The principal-component-analysis (PCA) plot for the top two principal components on binary (endometriosis vs. no endometriosis) is shown in Fig. 1, while the PCA plot for four categories (Minimal/mild, moderate/severe, no uterine pelvic pathology (N-UPP) and uterine pelvic pathology (UPP)) is shown in Fig. 2. Corresponding dendrograms and heatmaps for the data are attached in the appendix, in Supplemental Figures 7, 8, 9, and 10, respectively.



Figures 1 & 2. PCA with Two Genotype Tags for Unnormalized Data (**Fig 1, left**) and PCA by Severity for Unnormalized Data (**Fig 2,right**)

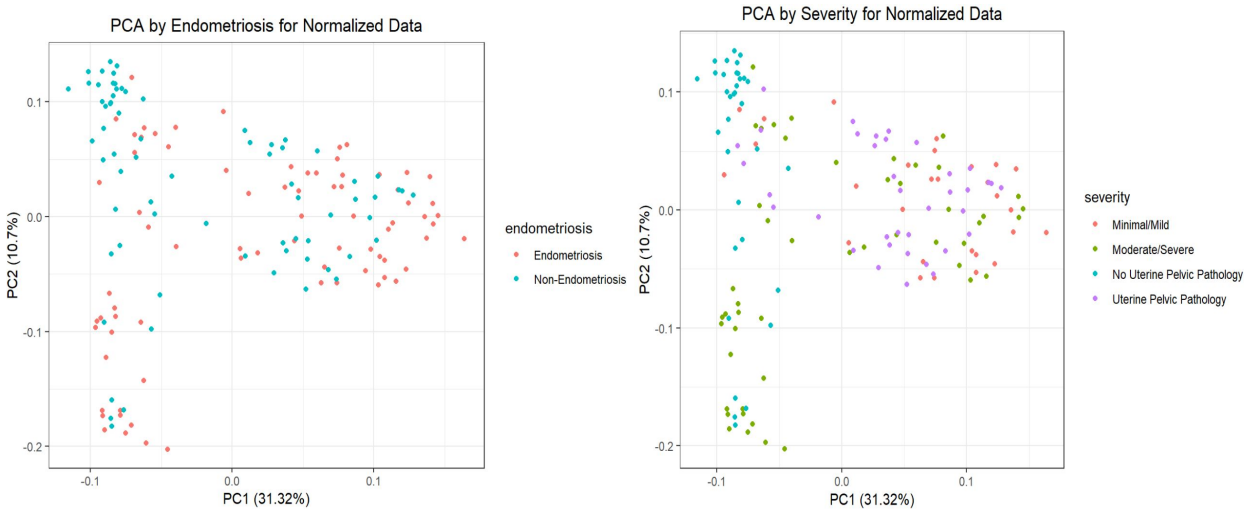
Applying PCA on the unnormalized gene expression data, we do not see a clear distinction between the categories of samples, except for a clear grouping of the N-UPP samples at the top left of the plot.

3. Methods

3.1 Normalization

Our method of choice for normalization is called the GCRMA or the Gene Chip Robust Microarray Averaging method. It is a package from Bioconductor that is used in the normalization of microarray data and is based on the method of RMA (Robust Microarray Averaging). We used GCRMA to create an expression matrix from the the probe level Affymetrix data. The raw intensity values are background corrected, log2 transformed, and quantile normalized. The presence of noise due to non-specific binding is a considerable problem in the analysis of microarray data. We preferred GCRMA over other methods of normalization because, (1) like RMA, it adjusts for background intensities including optical noise and non-specific binding and (2), it uses specific probe sequence information to estimate probe affinity to non-specific binding (NSB) and get accurate expression measures.

PCA plots for the binary data (endometriosis vs. no endometriosis) and the four categories (Minimal/mild, moderate/severe, no uterine pelvic pathology (N-UPP) and uterine pelvic pathology (UPP)) after normalization are presented in Figures 3 and 4, respectively.



Figures 3 & 4. PCA by Endometriosis for Normalized Data (**Fig. 3, left**) and PCA by Four Categories for Normalized Data (**Fig. 4, right**)

We do not see any improvements in the grouping of the samples after normalization but we continue to see the same distinction for the N-UPP samples, as seen in the unnormalized data. The corresponding dendrograms and heatmaps for normalized data are provided in Supplemental Figures 1, 2, 3 and 4.

3.2 Preprocessing

Low variance probes were defined as probes with measurement ranges that did not fall outside of 3 standard deviations of the mean or those that had a standard deviation of zero; these were removed from the dataset, reducing our dataset size from 54,675 probes to 15,430 probes. The remaining probes were translated into gene symbols using the Affymetrix library *hgu133plus2*. Genes with multiple representations were averaged into single column representations, allowing each gene to appear only once for each sample, thus, further reducing the data set to 14,822 columns of probes and genes. For uniformity, we will refer to the columns

of our dataset as "genes" although they contain both genes and probes (that were not matched to any genes).

3.3 Selecting differentially expressed genes

For comparison, we selected differentially expressed genes using (1) the nonparametric Mann-Whitney U test and (2) 2-fold mean difference between E and No E samples.

Mann-Whitney Test. Gene expressions for our data contained some skew (Supplemental Figure 5), thus violating (although not harshly) assumptions necessary to conduct t-tests for the selection of differentially expressed genes. Instead of a t-test, similar to the study done by Vengatesan et al [12], we favored a Mann-Whitney test that does not require the assumption of normality. We tested the following hypotheses.

H_o : Gene expression is independent of endometriosis.

H_1 : Gene expression is not independent of endometriosis.

After calculating p-values under the null hypothesis that a given gene expression is not independent of endometriosis and correcting for False Discovery Rates, 7,375 genes with a p-value of less than 0.05 were selected as differentially expressed and are presented in Supplemental Table 1.

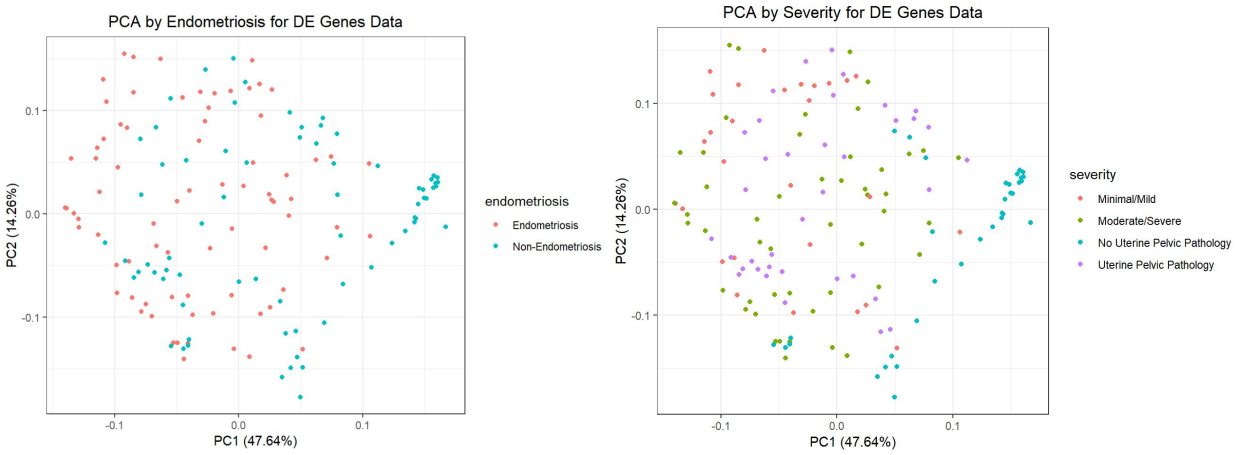
2-fold Mean Difference. DE genes were also selected using two-fold mean difference between E and No E subjects. For each category, the mean gene expression was calculated. The difference of mean expressions between the two categories was taken.

$$M = \bar{x}_E - \bar{x}_{No\,E} = log_2\,(E/No\,E)$$

$$A = \bar{x}$$

The MA plot (Supplemental Figure 6) showed variance smaller toward more extreme values and lacked curves, therefore we have further reason to believe our normalization procedure worked well on our data. Under this criterion, 382 DE genes were selected for values of M that exceeded 1, implying a mean two-fold increase between E and No E for the selected gene/probe. The genes from this method are available in Supplemental Table 2. The two generated sets of genes were then used to create classifiers for E v. No E.

PCA plots for the selected Mann-Whitney DE genes are presented in Figures 5 and 6. These plots failed to show clear patterns between the endometriosis categories. Therefore, we opted to classify our samples for E v. No E using machine learning methods. The corresponding dendrograms and heatmaps for normalized data are provided in Supplemental Figures 11, 12, 13 and 14.



Figures 5 and 6. PCA plots for DE genes data for a random sample (**Fig. 5, left** for two classes for endometriosis and **Fig. 6, right** for four categories of samples) of the Mann-Whitney DE genes do not show clear patterns.

3.4 Classification models

Three classification models were built (on both sets of DE gene data) to assess how accurately our two sets of DE genes can predict the presence or absence of endometriosis. In particular, k-nearest neighbors (KNN), Random Forests, and support vector machine (SVM) were applied and compared to one another. Cross-validation (CV) was used to build the KNN and SVM models, while Random Forests used out-of-bagging and hold-out validation to determine optimal number of variables and number of trees, respectively. All model accuracies were then assessed using hold-out validation on a separate validation set of the original data (Supplemental Figure 15). Due to the small sample size, we built two of each model: one model used 90/10 hold-out validation (test set = 15 samples) while the other used 80/20 hold-out

validation (test set = 30 samples) in order to increase the size of the validation set. As a result, two models for KNN, RF, and SVM were built and assessed. Finally, all models were built using both DE gene datasets (selected by the Mann-Whitney test and 2-fold mean difference method) to compare prediction accuracy between both sets of DE genes.

K-Nearest Neighbors. KNN is a simple and popular model that has been used in previous clinical outcome, gene-expression classification problems, including in predicting outcomes of breast cancer [4] and success of hepatitis treatment [5]. Prediction accuracy in this previous work has reached rates at or above 80%. KNN uses a distance metric and majority vote to classify an observation based on the dominant class of its k neighbors. In addition to its simplicity and widespread use, we selected KNN because as a nonparametric model, it does not require distributional assumptions of the data. k-fold cross-validation was applied to a training set in order to build the optimal model and select the value of k, which was then assessed on a separate test set. In particular, two KNN models were built, using a different training vs. test set divisions. The first model utilized a test set comprising 10% of the original data and 5-fold CV on the training set to build the model, while the second model relied on 20% of the original data as the test set and 4-fold CV on the training set to build the model.

Random Forests. Random Forests were the second classification model applied to the data. Similar to KNN, Random Forests require no assumptions about the data distribution and have become more frequently used in the context of high-throughput -omics platforms [6]. In a report by Lee et al. [7], RF outperformed the remaining machine learning methods they had applied in predicting early-stage ovarian cancer in the analysis of mass-spectrometry data. Additional attractive features of the model within the context of high-dimensional genomics data include that Random Forests avoid overfitting, particularly when the number of features outnumbers the number of samples (*large p, small n*) [8] and the model returns variable importance, a feature that can provide further information about the top differentially expressed genes [9]. A Random Forest is a collection of 'n' unpruned Classification and Regression Trees (CART), which are then all averaged to obtain a new prediction. RF uses bootstrapping and out-of-bag (OOB) sampling to avoid overfitting and obtain an an optimal number of variables (*mtry*) to use at each decision split. As a result, cross-validation is not needed to tune the *mtry*

parameter. However, the bootstrapping/out-of-bagging does not tune for number of trees (*ntree*), and this feature must be selected before building the model. In order to obtain an optimal number of trees, we utilized hold-out validation by finding the optimal *ntree* on a training set, and applied this to a separate validation set to obtain an overall accuracy. Once again, two hold-out validation methods were used (80/20 and 90/10) and thus, two RF models were built on each set of DE gene data.

Support Vector Machines. SVM is a popular method that has recently been expanding its use in genomics classification, particularly in the context of cancer outcome [10]. Advantages to SVM include that it is more powerful than other machine learning methods in recognizing subtle patterns in data [10], which may bode well for gene expression values. Additionally, similar to RF, SVM is effective in high-dimensional spaces [11]. Support vector machines create an optimal decision boundary/hyperplane between observations of binary classes. Linear SVM was applied to the data, and the width of the decision margin (cost) was tuned using the same cross-validation method as applied to KNN.

4. Results

4.1 Normalization

Figure 7 shows the effect of normalization for a random sample of n = 20 for all 54,675 probes.

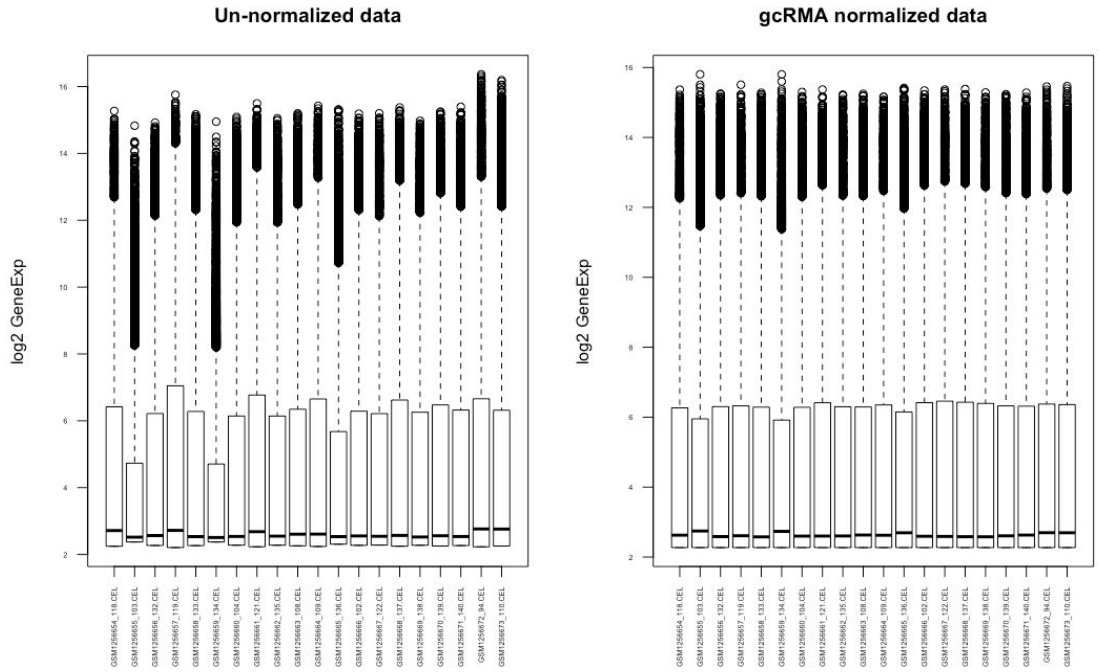


Figure 7. Boxplot showing effect of normalization after GCRMA

4.2 Differential expression results.

Top genes using Mann-Whitney are listed in Supplemental Table 1, while top genes using 2-fold mean difference are listed in Supplemental Table 2. The two methods yielded similarly selected DE genes. Of the 382 genes selected using the 2-fold mean difference, 374 of them were also selected by the Mann Whitney test.

4.3 Classification Results.

Optimal parameters for the three models are presented in Supplemental Table 3. Accuracies for the three models on the DE genes obtained using 2-fold mean differences are presented in Supplemental Table 4. The top five important variables (genes) generated by the Random Forests models are presented in Supplemental Table 5. For both the 90/10 and 80/10 hold-out validation methods, all accuracies are larger than 70%. In the 90/10 validation method, all three methods performed in the same manner on the test data; given a small test set of n = 15

samples, it is not surprising that similar error rates were obtained. When comparing methods on the 80/10 set, SVM outperforms KNN and RF.

Accuracies for the three models on the DE genes obtained using the Mann-Whitney method are presented in Supplemental Table 4, with corresponding optimal parameters presented in Supplemental Table 3. Similarly, SVM outperforms the remaining two models. Accuracies across all three models are lower for the 90/10 validation set while they are higher for the 80/20 set. Comparing across both datasets (2-fold mean difference and Mann-Whitney), SVM consistently outperforms the other two models.

We compared our prediction accuracies to those obtained by Tamaresis et al. The classification trees built by Tamaresis et al. [1] obtained validation set accuracies spanning 93% - 100% when using 21,734 predictors, including the stage of the menstrual phase in addition to probe expression levels. However, full comparisons cannot be made because Tamaresis et al. did not include E vs. no E as one of their pairwise classifications; in addition, they used a different model (CART) and different predictor variables.

5. Discussion

The results of our normalization indicate some some skew (as explained by an overall low median value and large number of outliers). We have, therefore, used the results of gcRMA with caution that we are introducing some bias into our data. In further analyses, we have tried to use methods that do not assume normality, wherever possible.

Selected genes from the Mann-Whitney test were compared with the genes used by Tamaresis et al. to classify E v. No E UPP+ and E v. No E UPP-. Tamaresis used a combined total of 12,057 genes to classify their samples for disease presence. Our Mann-Whitney test identified 7,375 genes (38% less genes), of which at least 55.35% were included in the Tamaresis classifiers. 382 genes (96.8% less genes) were selected by the 2-fold mean increase method, and 54.19% were included in the Tamaresis classifiers. The contrast between the genes selected by our methods was quite large considering that 12,057 genes were selected by Tamaresis et al. and we included much less where we would expect much more of our selected

genes to be in common with Tamaresis. Our selection of DE genes, despite the inconsistency with Tamaresis, was able to create moderately accurate classifiers for E and No E samples.

The classification models yielded prediction accuracies as high as those in previous literature. While the accuracies were higher on the 80/20 hold-out method using the Mann-Whitney method, it is possible that overfitting had occurred because the models relying on the Mann-Whitney method data used 7,375 predictor genes, as compared to the 382 predictors from the 2-fold mean difference method data. While we attempted to reduce the overfitting of the classification models by using cross-validation, it is best to apply the models to new data in the future in order to further circumvent such overfitting in the prediction accuracies. Additionally, while the CART models built by Tamaresis et al. yielded up to 100% prediction accuracies, they risk overfitting due to the use of single decision trees instead of Random Forests and the inclusion of 21,734 predictor variables. In contrast, our models still returned relatively high accuracies, while relying on less information.

Future classification models can mimic Tamaresis et al.'s decision to include menstrual cycle phase as a predictor. However, even without the menstrual phase as a predictor, models solely using core genes as predictors may provide just as much information in predicting clinical outcome. Additionally, future work and models should consider the source of data collection. It is possible that the sample from UCSF is not representative of all patients, and may introduce bias with respect to location and health disparities. Finally, we performed our analysis under the assumption that the experimental information was collected accurately and consistently. If the experimental data was not, in fact, collected accurately, this may have implications on our statistical results.

6. Conclusions

The two methods of selection for differentially expressed genes yielded similar results. 374 out of 382 of the DE genes (97%) selected by the 2-fold mean difference method were also considered DE by the Mann-Whitney test. Of the classifiers built from these two datasets, SVM

with a cost of c = 0.01 outperformed KNN and RF for correctly classifying the presence or absence of endometriosis achieving a highest accuracy of 93.33% and a lowest accuracy of 76.7% using the 20% and 10% validation sets with Mann-Whitney DE genes. However, all three methods performed moderately well (>70%) on all validation sets.

7. References

1. Tamaresis, J. S., Irwin, J. C., Goldfien, G. A., Rabban, J. T., Burney, R. O., Nezhat, C., ... & Giudice, L. C. (2014). Molecular classification of endometriosis and disease stage using high-dimensional genomic data. Endocrinology, 155(12), 4986-4999.

2. Bakhtiarizadeh, M. R., Hosseinpour, B., Shahhoseini, M., Korte, A., & Gifani, P. (2018). Weighted gene co-expression network analysis of endometriosis and identification of functional modules associated with its main hallmarks. Frontiers in genetics, 9.

3. Tsudo, T., Harada, T., Iwabe, T., Tanikawa, M., Nagano, Y., Ito, M., ... & Terakawa, N. (2000).
Altered gene expression and secretion of interleukin-6 in stromal cells derived from endometriotic tissues. Fertility and sterility, 73(2), 205-211.

4. Parry, R. M., Jones, W., Stokes, T. H., Phan, J. H., Moffitt, R. A., Fang, H., Shi, L., Oberthuer, A., Fischer, A., Tong, W., Wang, M.D. (2010). k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. Pharmacogenomics, 10(4), 292-309.

5. Sarasin-Filipowicz, M., Oakeley, E. J., Duong, F. H. T., Christen, V., Terracciano, L., Filipowicz, W., Heim, M. H. (2008). Interferon signaling and treatment outcome in chronic hepatitis C. Proc Natl Acad Sci. 105(19), 7034-7039.

6. Chen, X., Ishwaran, H. (2012). Random Forests for Genomic Data Analysis. Genomics, 99(6), 323-329.

7. Wu, B., Abbot, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics, 19(13), 1636-1643.

8. Okun, O., Prissalu, H. (2007). Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues. IbPRIA, 4478, 484-490.

9. Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics, 99(6), 323-329.

10. Huang, S., Cai, N., Pacheco, P. P., Narandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics-Proteomics, 15(1), 41-51.

11. Byvatov, E., & Schneider, G. (2003). Support vector machine applications in bioinformatics. Applied bioinformatics, 2(2), 67-77.

12. Vengatesan, K., Mahajan, S. B., Sanjeevikumar, P., Mangrule, R., & Kala, V. (2018). Performance Analysis of Gene Expression Data Using Mann–Whitney U Test. In Advances in Systems, Control and Automation (pp. 701-709). Springer, Singapore.

Supplemental Tables and Figures

Figures

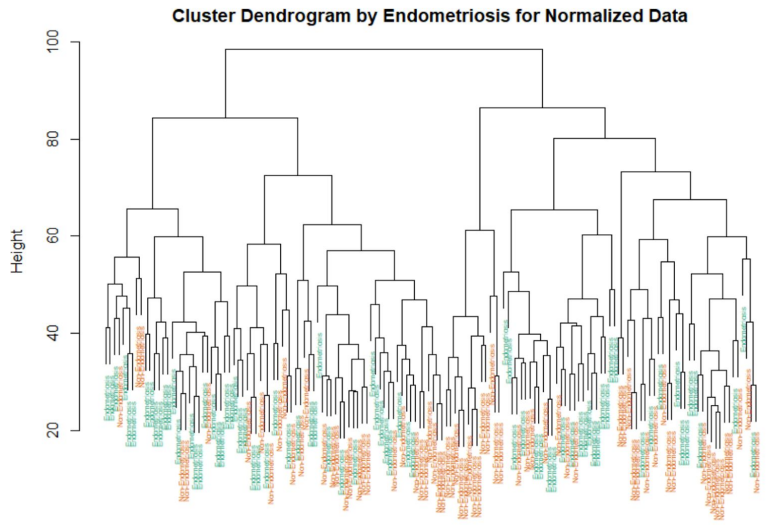


Figure 1. Cluster Dendrogram by Endometriosis for Normalized Data (We only see clusters for Non-Endometriosis)

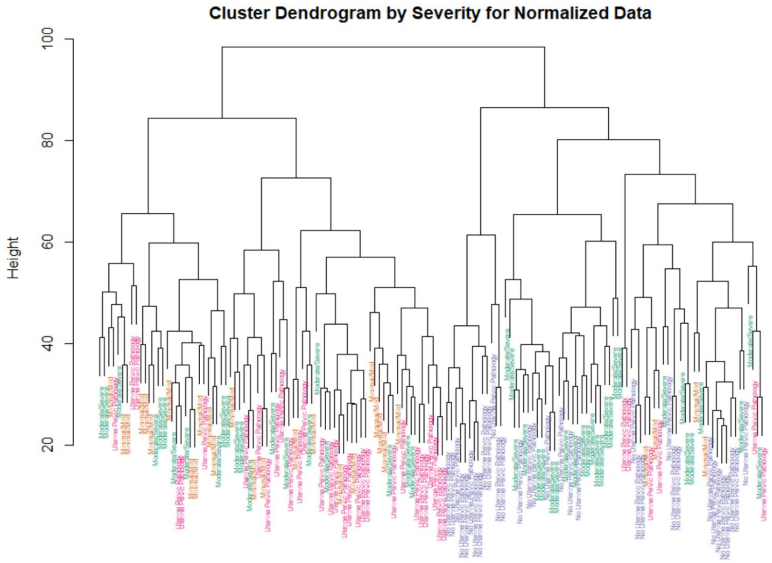


Figure 2. Cluster Dendrogram by Severity for Normalized Data (We only see clusters for Non-Endometriosis)

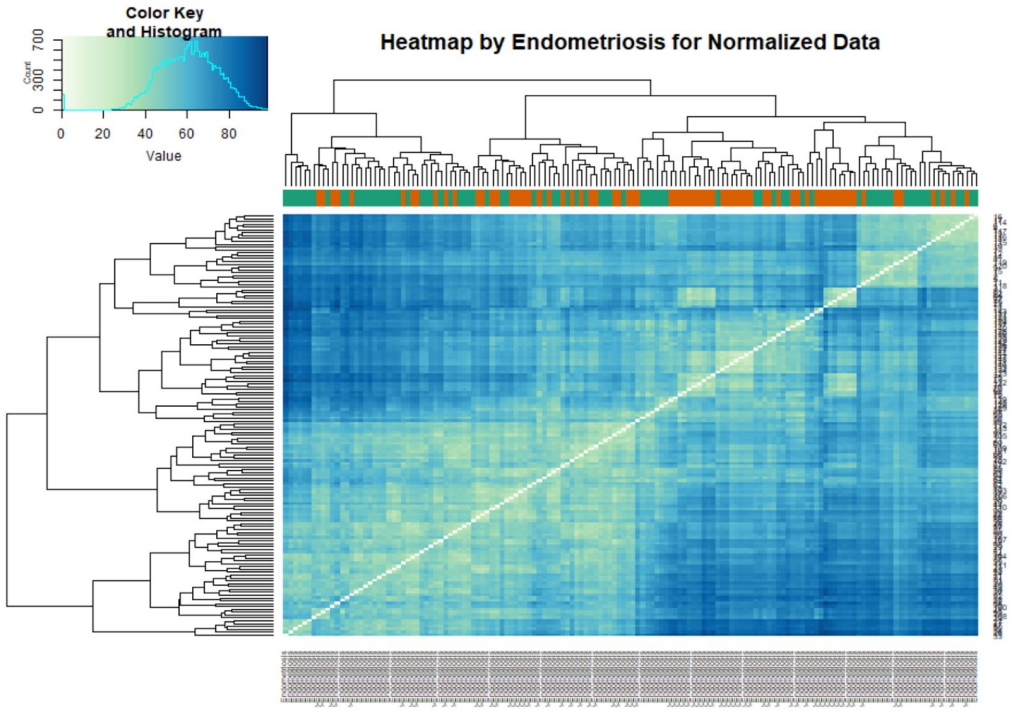


Figure 3. Heatmap by Endometriosis for Normalized Data

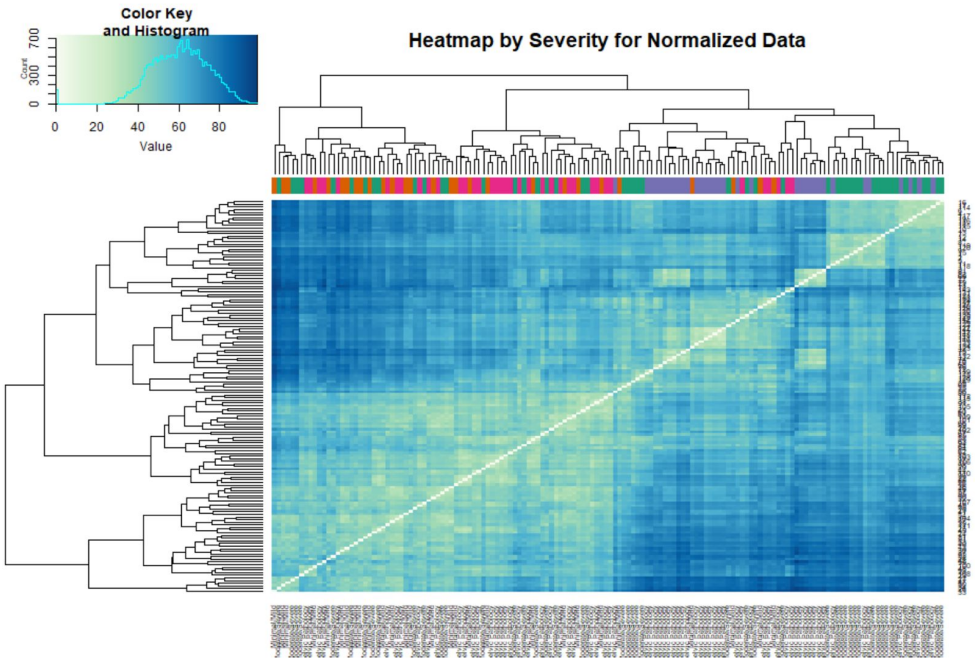


Figure 4. Heatmap by Severity for Normalized Data

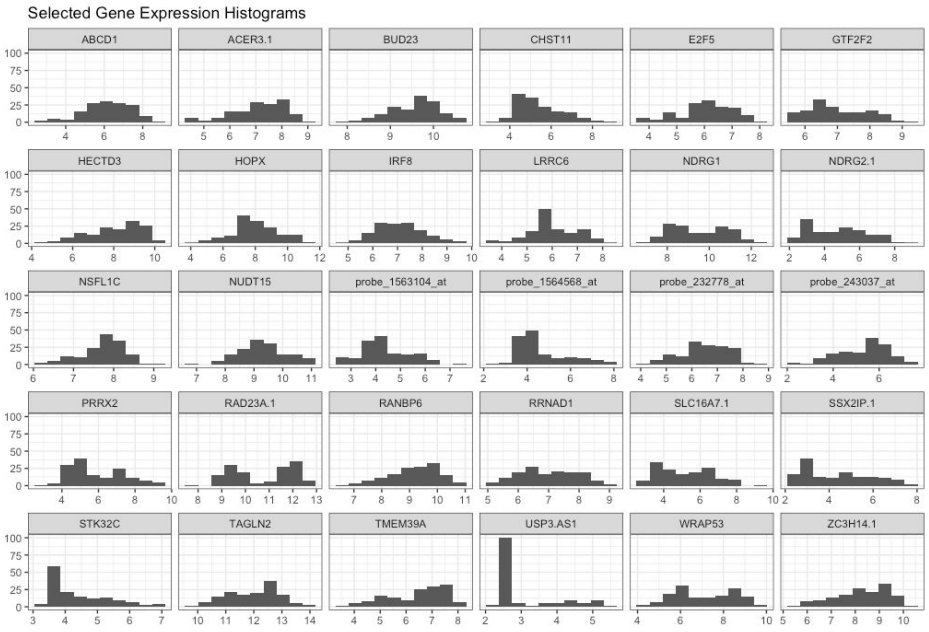


Figure 5. Randomly selected gene expression histograms.

Randomly selected histograms are shown to examine normality. Some plots show near symmetry. Others are bimodal or skewed. USP3.AS1 showed a strong peak near 2.5.

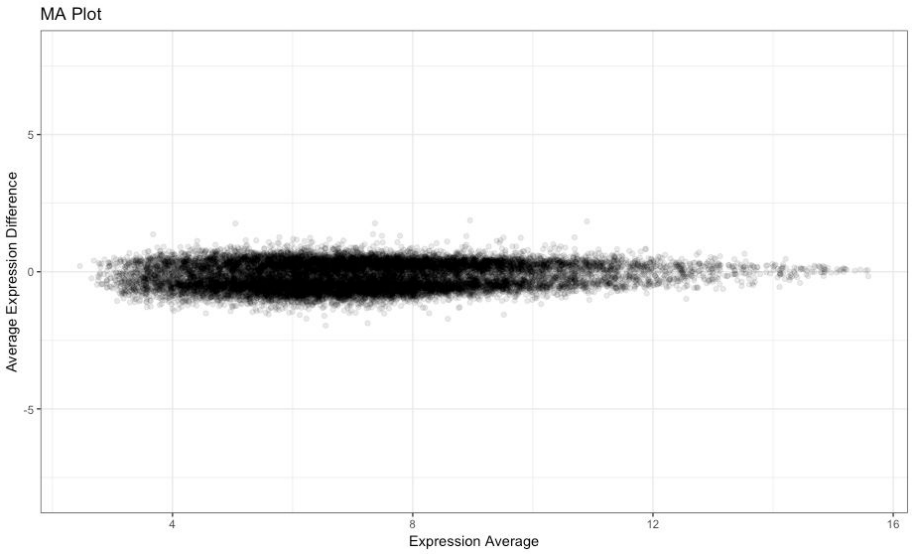


Figure 6. MA plotThe average gene expression was plotted against the mean difference of gene expressions per group. There is less variance with increasing average. We also see that the data are centered well around 0.

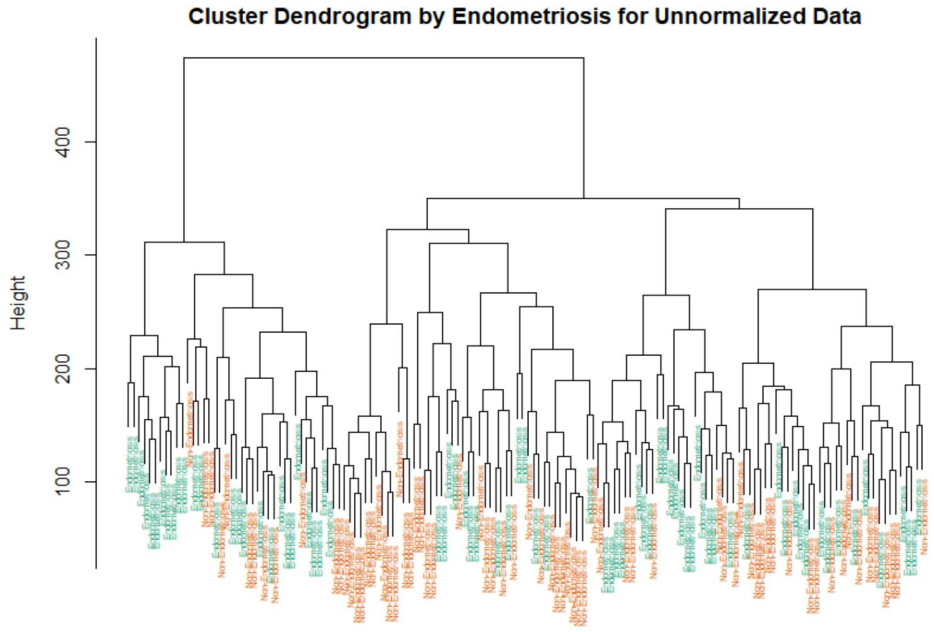


Figure 7. Cluster Dendrogram by Endometriosis for Unnormalized Data

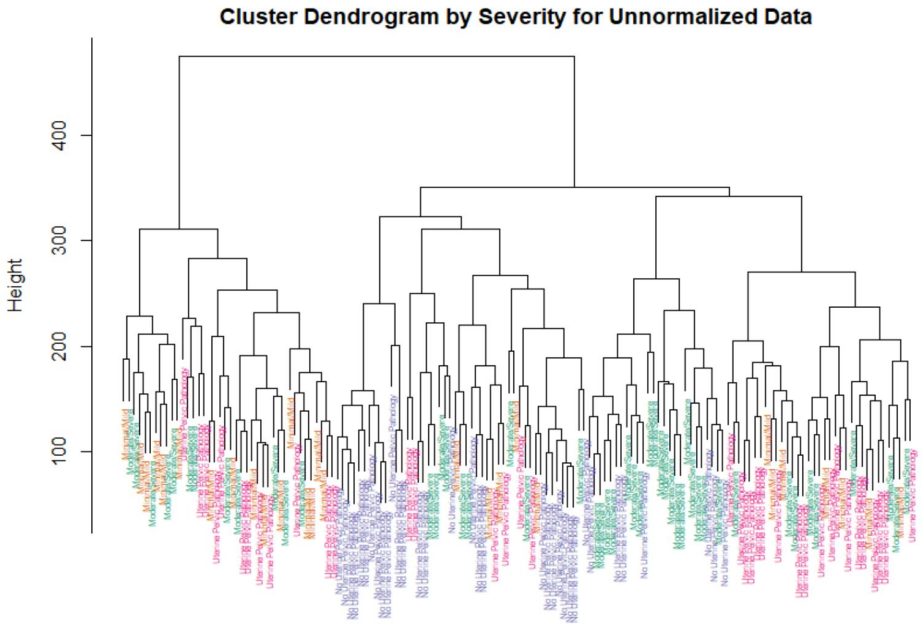


Figure 8. Cluster Dendrogram by Severity for Unnormalized Data

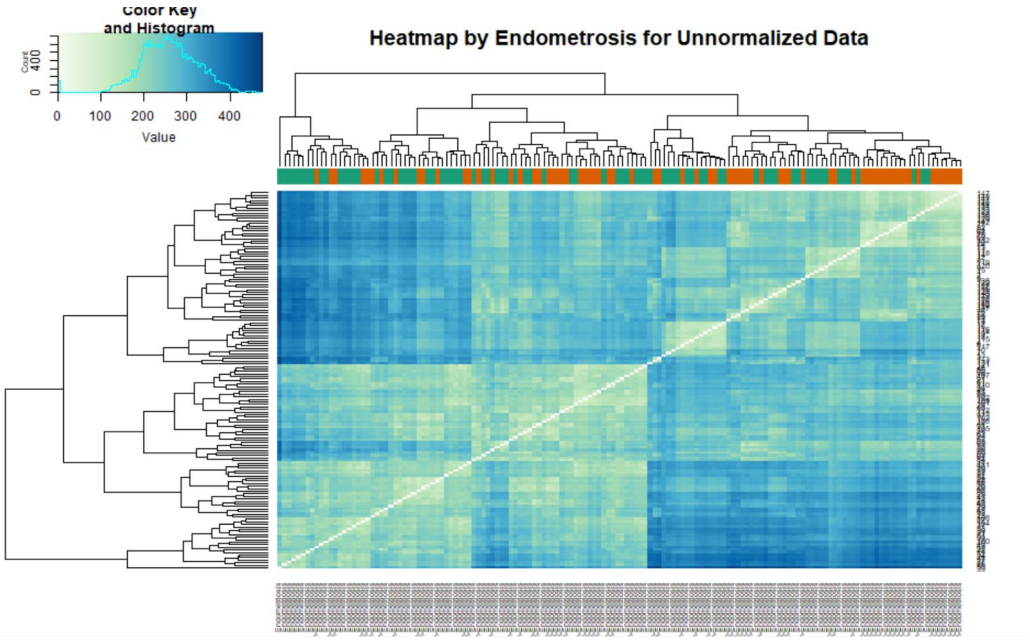


Figure 9. Heatmap by Endometrosis for Unnormalized Data

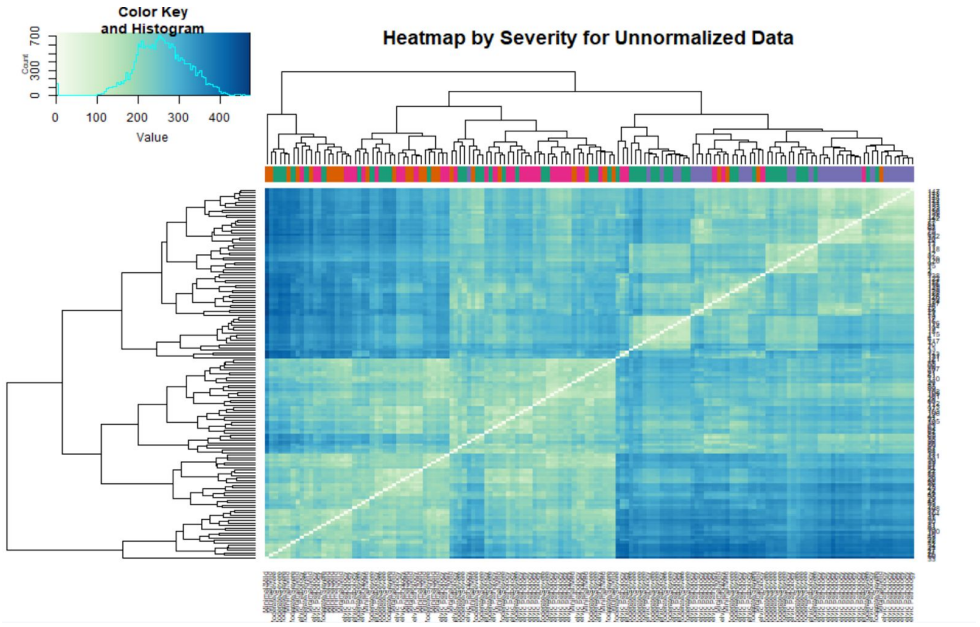


Figure 10. Heatmap by Severity for Unnormalized Data

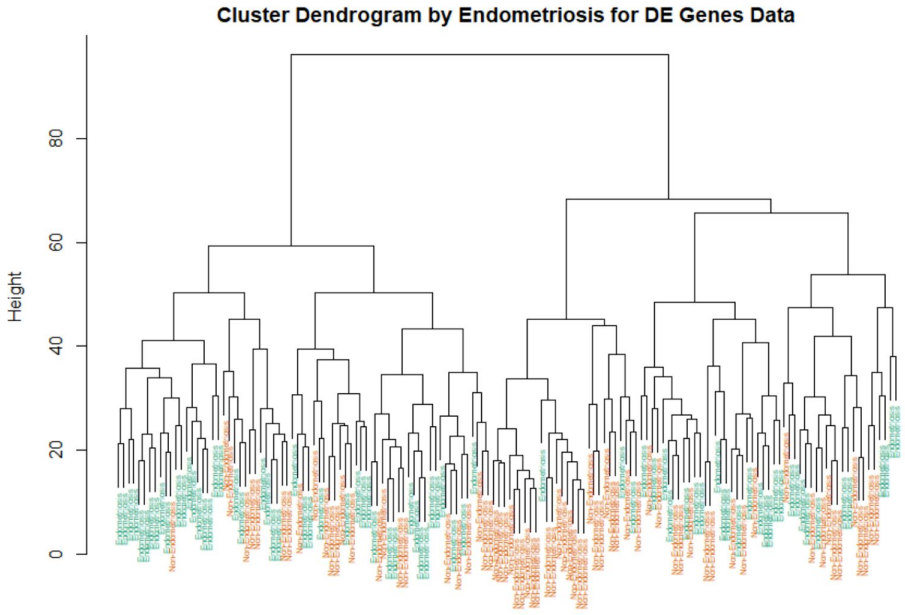


Figure 11. Cluster Dendrogram by Endometriosis for DE Genes Data

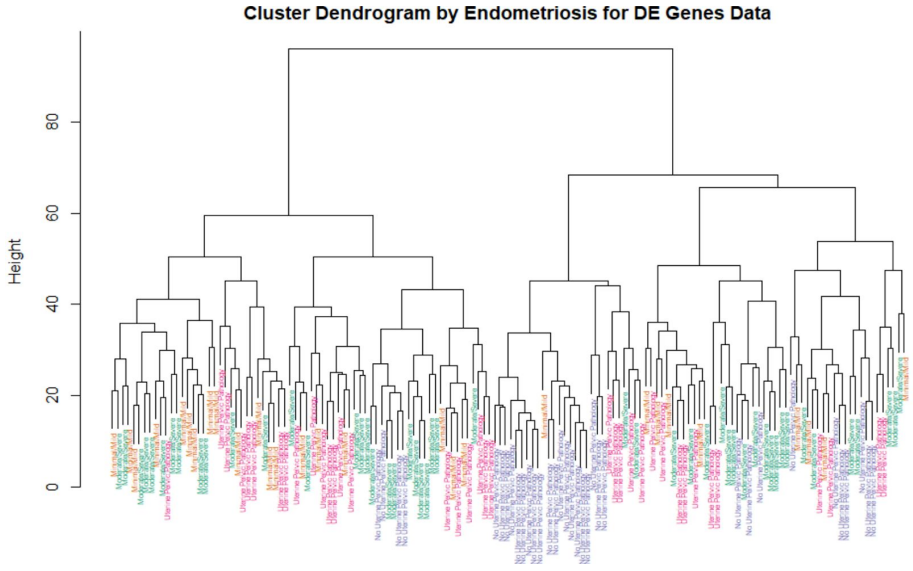


Figure 12. Cluster Dendrogram by Endometriosis for DE Genes Data

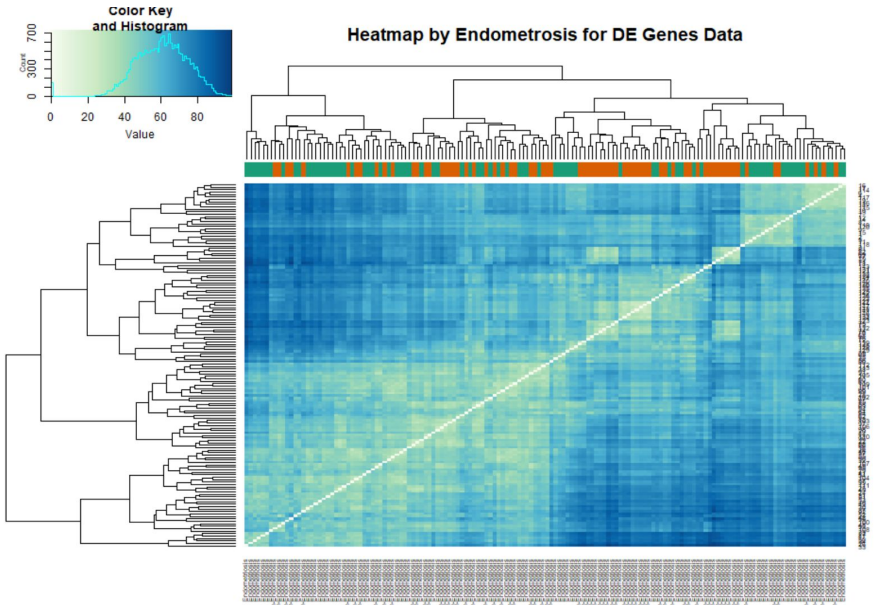


Figure 13. Heatmap by Endometriosis for DE Genes Data

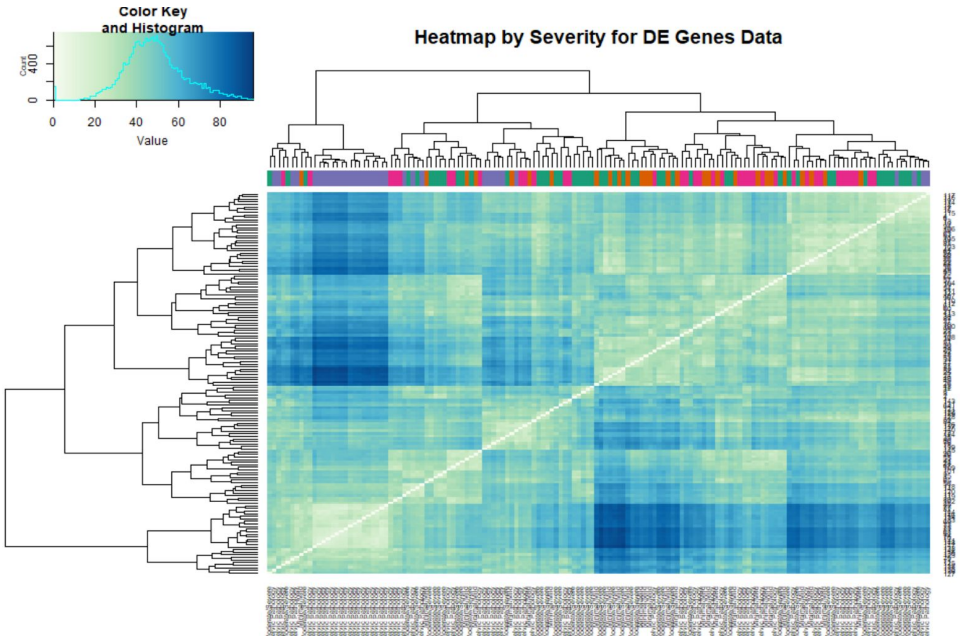


Figure 14. Heatmap by Severity for DE Genes Data

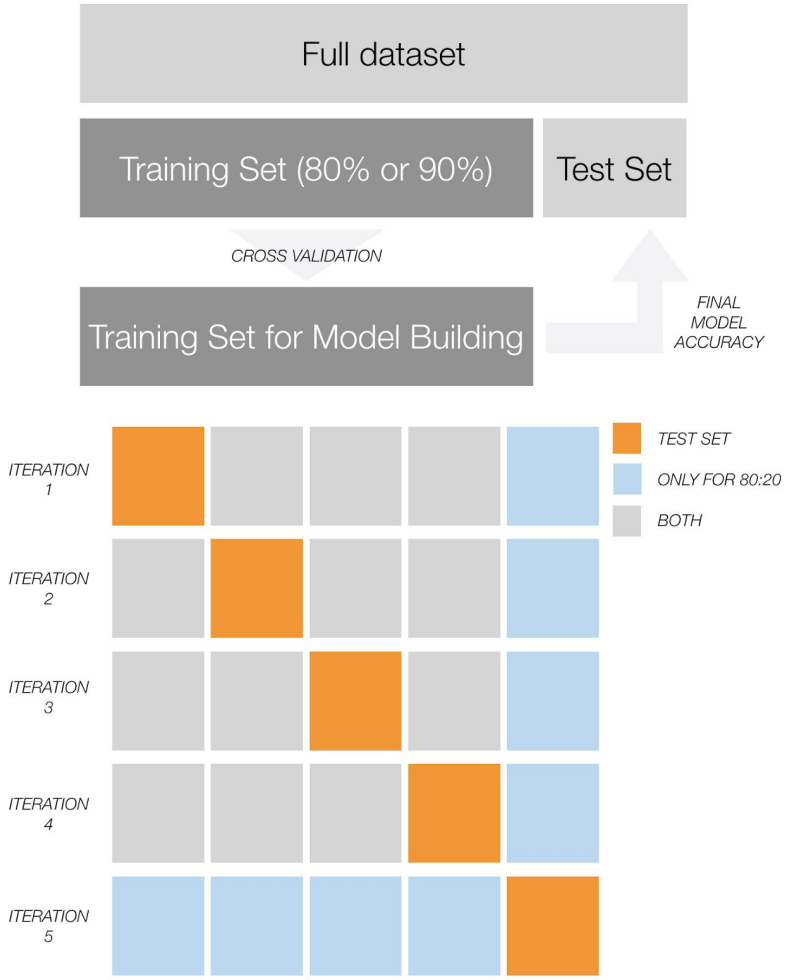


Figure 15. Classification model and assessment validation scheme

Tables

Table 1. Differentially expressed genes via Mann-Whitney Test

DCAF16	ABHD17B
HAUS2.1	MAP4
CDC27	MYLIP.1
ZBTB5	230580_at (Probe)
EMSY.1	IL13RA1.3
ADAT1	LOX.1

These are ordered by p-value significance and non-exhaustive. These are 12 of 7,375 DE genes.

Table 2. Differentially expressed genes via 2-fold mean difference

DIO2	CPM.2
PCSK5.1	OLFM4
FOS	CADM1.1
EGR1.2	NMT2
FOSB	PCSK5
SCGB3A1	ANK2

These are ordered by highest mean expression difference. These are 12 of 382 DE genes

Table 3. Optimal parameters associated with classification models

	Parameters[1]			
	2-fold mean difference DE genes		**Mann-Whitney DE genes**	
Method	**80/20 hold-out validation**	**90/10 hold-out validation**	**80/20 hold-out validation**	**90/10 hold-out validation**
KNN	k = 51	k = 44	k = 13	k = 11
RF	ntree = 400 mtry = 17	ntree = 250 mtry = 20	ntree = 250 mtry = 65	ntree = 550 mtry = 11
SVM	cost = 0.01	cost = 0.01	cost = 0.01	cost = 0.01

1. Parameters were tuned using k-fold cross-validation (k = 4 for 80/20 and k = 5 for 90/10) on the hold-out training set

Table 4. Test accuracies associated with classification models

	Test accuracy			
	2-fold mean difference DE genes		**Mann-Whitney DE genes**	
Method	**80/20 hold-out validation**	**90/10 hold-out validation**	**80/20 hold-out validation**	**90/10 hold-out validation**
KNN	0.700	0.867	0.800	0.633
RF	0.767	0.867	0.867	0.533
SVM	0.833	0.867	0.933	0.767

Table 5. Top 5 Important Variables (DE Genes) Returned by Random Forests

	2-fold mean difference DE genes		Mann-Whitney DE genes	
Rank[1] of Variable	**80/20 hold-out validation**	**90/10 hold-out validation**	**80/20 hold-out validation**	**90/10 hold-out validation**
1	MAP4	DCAF7.2	ZBED6	ZBED6
2	CDC27	MAP4	244470_at	HAX1
3	DCAF16	HECTD1	HECTD1	1556818_at (Probe)
4	HAUS2.1	HAUS2.1	C2CD3	PMS2P5
5	EMSY.2	CDC27	242787_at	ZNF709

1. Rank was based off of 'mean decrease accuracy'