Gene Expression Signatures of Endometriosis

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

Introduction

Endometriosis is a painful gynecological disorder that affects nearly 10% of women worldwide [1]. It is an estrogen-dependent condition characterized by the displacement of the endometrial tissue that causes growths and lesions in the abdomen and pelvic cavity, leading to chronic pain and inflammation. Currently, it can only be fully diagnosed at surgery, therefore adding on average a latency period of 11 years before it is diagnosed. Recent studies have suggested that abnormalities in the regulation of specific genes [2] 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.

Methods

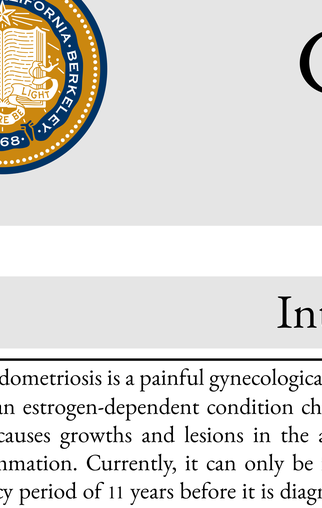

Figure 1. Workflow for DE gene selection.

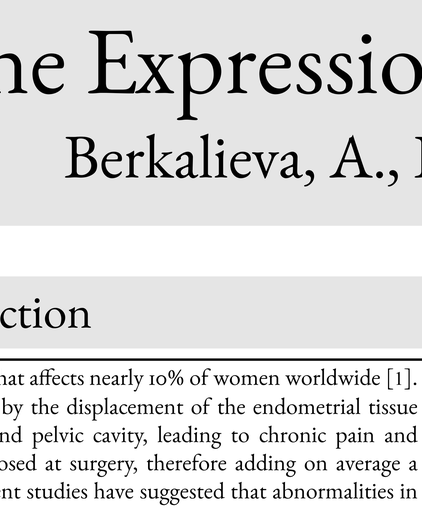

Figure 2. Machine learning pipeline.

The presence of noise due to non-specific binding is a significant problem in microarray data analysis. To take this into account, we used GCRMA to create an expression matrix from the probe level Affymetrix data. The raw intensity values are background corrected, log2 transformed and then quantile normalized.

We created 80/20 and 90/10 datasets for the DE genes selected by 2-fold mean difference method and Mann-Whitney test respectively to fit KNN, Random Forests (RF), and SVM models. The final models using cross-validated parameters were applied to test sets.

Results

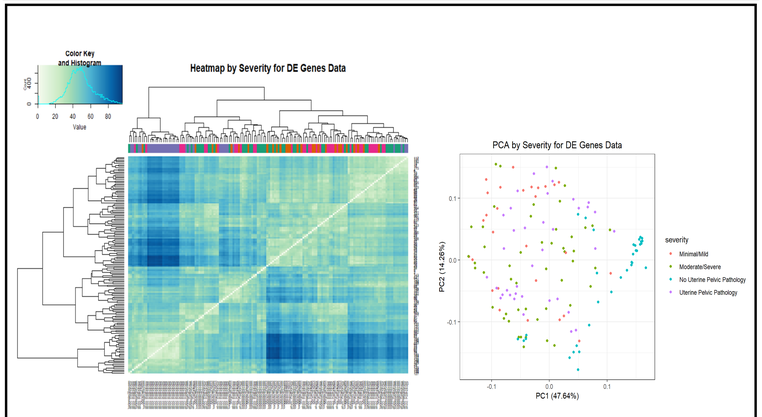

Figure 3. Random selection of Mann-Whitney DE genes show no-UPP (uterine pelvic pathology) samples separated from others.

Figure 4. PCA also shows clear grouping of the no-UPP samples from the other categories of endometriosis.

Table 1. Optimized parameters for classification models.

	Parameters			
	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	ntree = 250	ntree = 250	ntree = 550
	mtry = 17	mtry = 20	mtry = 65	mtry = 11
SVM	cost = 0.01	cost = 0.01	cost = 0.01	cost = 0.01

Table 2. Test prediction accuracies for 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	0.7	0.867	0.8	0.633
RF	0.767	0.867	0.867	0.533
SVM	0.833	0.867	0.933	0.767

Discussion

- Normalization results indicated some skew, therefore we proceeded with the results of GCRMA with caution.Mann-Whitney Test
  - Selected 7,375 DE genes
  - 55.3% of these genes were used in classifiers by Tamaresis
- 2-fold mean difference method
  - Selected 382 DE genes
  - 54.19% of these genes were used in classifiers by Tamaresis
- We attempted to reduce overfitting by (1) using cross-validation for KNN and SVM and (2) Random Forests
- The classifications fitted by Tamaresis had higher accuracies than our models (>93%), but our constructed models used much less data (no menstrual cycle data, less genes) and still achieved a relatively high accuracy (>70%)

Conclusion

- 374/382 genes selected as DE by the 2-fold mean difference method are also considered DE by the Mann-Whitney Test
- SVM with a cost of C=0.01 outperformed KNN and RF for correctly classifying the presence or absence of endometriosis
- All three methods performed moderately well (>70%) on all validation sets

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. x, 155(12), 4986-4999.

2. 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.