Review of:

"*Genome-wide transcriptional analysis of T cell activation reveals differential gene expression associated with psoriasis"* [1]

Sam Brown, Wayne Kunze, Caleb Perry

EECS E6690, Spring 2018

1. Introduction

The paper selected for review investigates the genetic mechanisms that cause activation of T cells, responsible for skin inflammation in sufferers of Psoriasis, a common autoimmune disease.

The paper discusses findings from T cell gene expression data collected from 24 individuals, 17 of which had Psoriasis. The dataset has expression rates for thousands of different genes, all of which are identified with markers. The paper concluded that the two genes SPATS2L and KLF6 are most strongly associated with T cell activation in Psoriasis patients. The dataset itself was interesting due to the much larger number of features compared to samples. This fact plays an important role in much of the analysis conducted throughout this paper.

In addition to attempting to replicate the results of the paper, we chose to examine the data from a several new perspectives. We used supervised learning methods such as decision trees to identify the genes indicating a patient has psoriasis. We also demonstrated how Support Vector Machines (SVM) and regression techniques can be used to train very accurate models. Since the dataset has many more components than data points, feature reduction was performed using Principle Component Analysis (PCA) as well as shrinkage and subset methods. Feature reduction allowed us to reduce the dimensions of the data by several orders of magnitude. Lastly, we used unsupervised learning techniques in which we ignored the response variable to determine if clusters could be found indicating the presence of the disease.

2. Data Set and Paper

The paper examines gene expression profiles from in vitro activated T cells from 17 psoriasis patients and 7 control subjects. The data set developed for this study contains 47,222 transcripts for each sample cataloging the level of gene expression in the activated T-cells for each gene in each individual.

3. Results

The paper, “Genome-wide transcriptional analysis of T cell activation reveals differential gene expression associated with psoriasis”, identifies several genes that significantly up-regulated or down-regulated in patients with psoriasis. These results were found by first pruning the data via removing all genes that were not expressed in at least 3 patients. Then the average expression rate of those who had the disease was computed and compared to the average expression rate of those who didn’t have the gene. This was used to produce tables of relative fold changes. Attempts to reproduce these results were made however results differ from the paper. Table 1 shows the 13 most upregulated gene indicators as a result of this analysis.

Table 1: List of most up-regulated genes

```
"ILMN_2058782" "ILMN_2305112" "ILMN_1701789" "ILMN_2410826" "ILMN_1721113"
“ILMN_1658247" "ILMN_2054297" "ILMN_2184373" "ILMN_1739428" "ILMN_1729749"
“ILMN_1700967" "ILMN_2347798" "ILMN_1670134"
```

The results from Table 1 differ from the results listed in the paper. In the paper the principle gene discussed is SPATS2L, which the paper presented as being upregulated 1.37-fold in patients with psoriasis, however in our analysis, the gene was only upregulated 1.003101-fold. The cause of this discrepancy is unclear but may be due to different pre-processing and scaling of the dataset.

4. New methods used

Decision Trees

We decided to utilize Random Forests to analyze the data. This is a process that can be computationally intensive. We focused on large forests (each forest with ten thousand trees) but we only considered the default number of factors for each tree.

Working with more than forty-seven thousand factors is a challenge. This is larger than RStudio can handle with the randomForest function without crashing. To address this problem, we had to choose a form of factor reduction. We chose to use the importance parameter from randomForest to take a subset of the "more important" factors from subsets of one thousand (or less) factors.

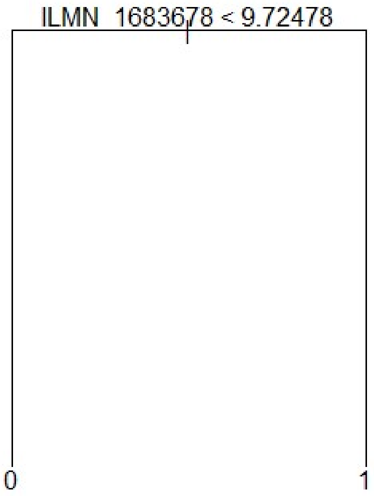


Figure 1 SPATS2L Tree

Once the "more important" factors have been identified, we used randomForest to create trees amongst these factors to identify the "most important" factors across the entire dataset.

We were surprised to note very limited overlap from Random Forests to other methods used by ourselves and the original paper. It's not surprising that ILMN_1683678 (SPATS2L) always makes the top slot (it has 100 percent accuracy in predicting the end result by itself, see Figure 1), but other factors not predicted elsewhere are also ranked very highly in terms of importance for Random Forests (see Figure 2).

Creating individual trees from these "most important" factors has reasonably good accuracy. We only see 100% accuracy in the case of ILMN_1683678, but other trees use two factors for 96% accuracy (for example, see Figure 3 and Figure 4). Even though the 96% accuracy would be maintained with only a single factor, narrowing the field of where potential error would occur could have value in future prediction applications.

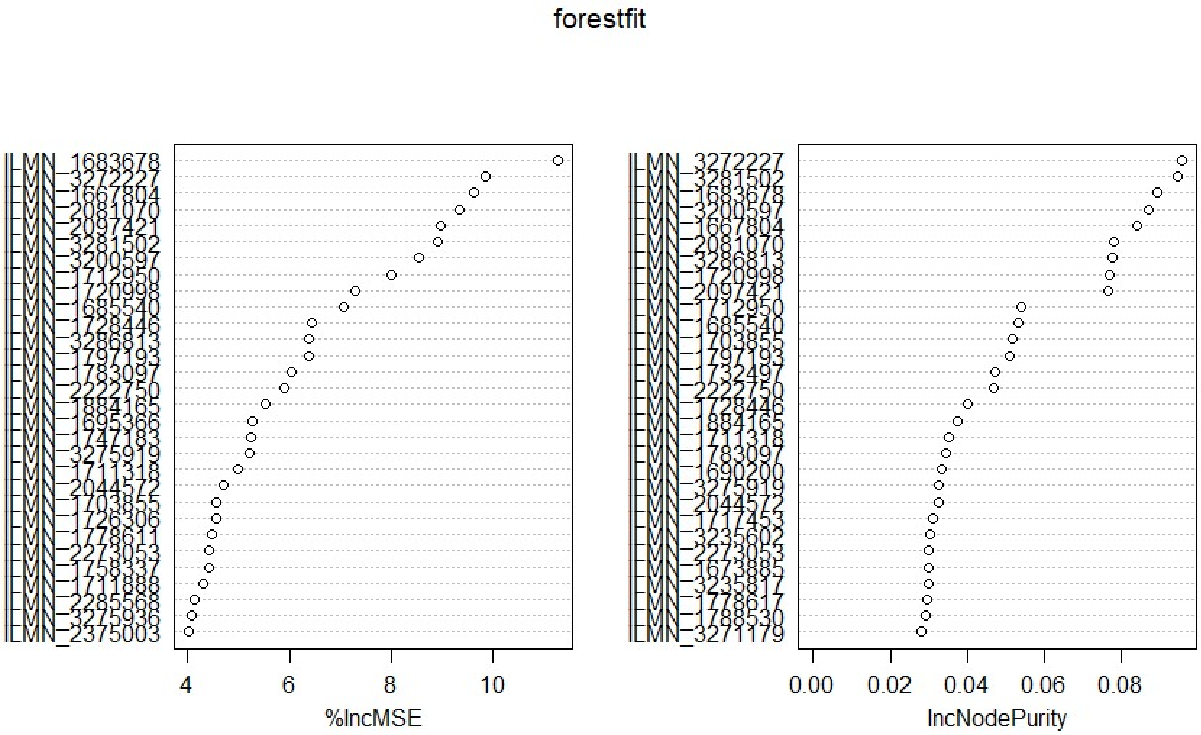


Figure 2 Relative Importance of "Most Important" Factors

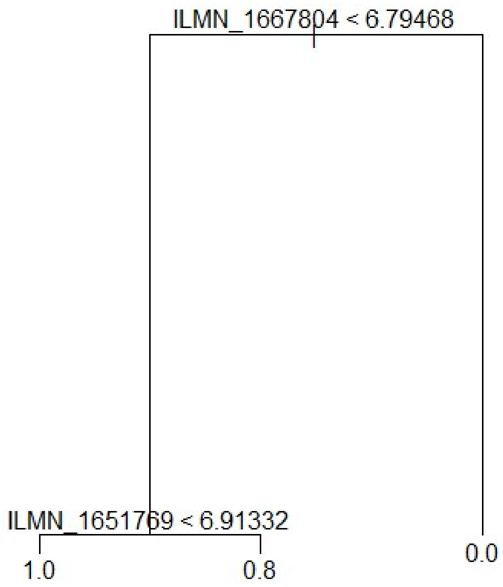


Figure 3 Sample Tree 1

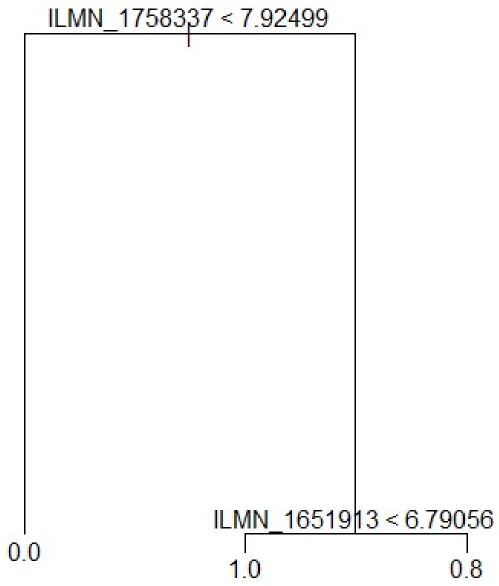


Figure 4 Sample Tree 2

Dimension Reduction

Traditional machine learning methods start with the assumptions that the model features are (mostly) independent and that there are more samples then there are features ($n > p$). In our data set there are 47,222 features (genes) and only 24 samples (patients), p is three orders of magnitude larger than n. In addition, genes express in groups, therefore many of the features are not independent. Attempting to use traditional methods such as least squares linear regression would result in severe overfitting and a poor predictive model on the Psoriasis dataset.

In this situation methods to reduce the dimensionality (discarding of "unnecessary" features) can offer a more reasonably sized model that may still accurately perform classification.

We experimented with three different methods for reducing a high dimensional data set like Psoriasis: Subset selection, Shrinkage, and Principal Component Analysis. The objective was to perform feature reduction and then assess the effectiveness by comparing the reduced dataset to the list of highly expressed genes from the Psoriasis paper. If the method is effective, then a many of the identified genes should still be present in the resulting reduced model.

Subset Selection

Three general approaches to subset selection are commonly used in statistical learning, with speed versus optimality the primary tradeoff. The results of these techniques can be rated using either cross validation or indirect training error estimates, such as RSS or R^2

Best Subset will provide an optimal model (based on either CV or indirect estimate) but at a significant processing cost. To reach the optimal model in the Psoriasis data set would require analysis of the almost inconceivably large 2^{4722} potential models and is therefore not appropriate for our dataset.

Similarly, Backward Stepwise Selection, which would start from with the full, 47,222 feature model, and works back removing the feature that provides the least value to the indirect estimate. Unfortunately, when the model has more features than samples, like ours, the full model will not fit and therefore cannot be used for our set.

That leaves Forward Stepwise Selection, which starts from the null model and adds the most useful feature at each stage. Forward selection requires inspection of on the order of 47222^2 potential models, not trivial but also not intractable. Because it starts with a model of no features it can be used when the full model has more predictors than samples.

Step.r performed forward stepwise reduction on the dataset using a logistic regression model (appropriate for binary classification) for fitting and produced reduced feature models based on the AIC criteria. The objective was to determine if a “good” reduced model would include the genes identified as important to the activation of T-cells in Psoriasis patients from Table 1 of the original paper (reproduced in Table 6 and Table 7 at the end of this paper). The data set was first reduced using the criteria of coefficient of variability, similar to the original paper, to reduce the number of features to ~11,000. This filter allowed R to manage the dataset size without crashing.

Table 2: Forward Stepwise reduced model features

Probe	Gene	Deviance	AIC
ILMN_1683678	SPATS2L	0	4
ILMN_3286813		10.152	14.152
ILMN_3281502		11.93	15.93
ILMN_1735014	KLF6	11.956	15.956
ILMN_1781285	DUSP1	12.519	16.519
ILMN_1778617	TAF9	12.753	16.753
ILMN_2321064	BAX	13.31	17.31
ILMN_3309534		13.765	17.765
ILMN_2143250	FAR1	14.448	18.448
ILMN_3200597		14.904	18.904
ILMN_2119421		15.149	19.149
ILMN_1860954		15.579	19.579
ILMN_2285568		15.642	19.642
ILMN_2246956	BCL2	15.764	19.764
ILMN_1731107	CCDC92	15.922	19.922
ILMN_2397721	GLB1	15.934	19.934

Yellow are up-regulated genes in Psoriasis patients, Blue are down-regulated versus the control

Table 2 illustrates, nine of the important genes are selected in the first seventeen features Forward Stepwise identified as the most useful from a minimal AIC standpoint, including SPATS2L and KLF6—the newly identified genes from the Psoriasis paper. The remaining genes appear in the first 108 features selected by the algorithm, implying that a model reduced from ~47,000 to ~100 features would still include the primary genetic predictors included in the Psoriasis paper.

<u>Shrinkage</u>

The two shrinkage methods attempt to fit the full model through least squares (and attempting to minimize RSS) but with an additive penalty for larger coefficient values.

Ridge regression uses the easier to compute l_2 penalty but the cost of this is a model that shrinks the magnitude of each coefficient but does not remove them. Figure 5 is a plot of the ridge coefficients as a function of the shrinking penalty and demonstrated how the coefficients trend towards 0.

LASSO is a similar algorithm to Ridge but one that forces many coefficients to exactly zero—offering feature selection/reduction like Forward Stepwise. *Lasso.r* performs LASSO and Ridge on a logistic regression model fitted to the dataset across different shrinkage values. LASSO Models were selected manually for feature comparison to the Psoriasis paper list to determine if LASSO selected a similar set.

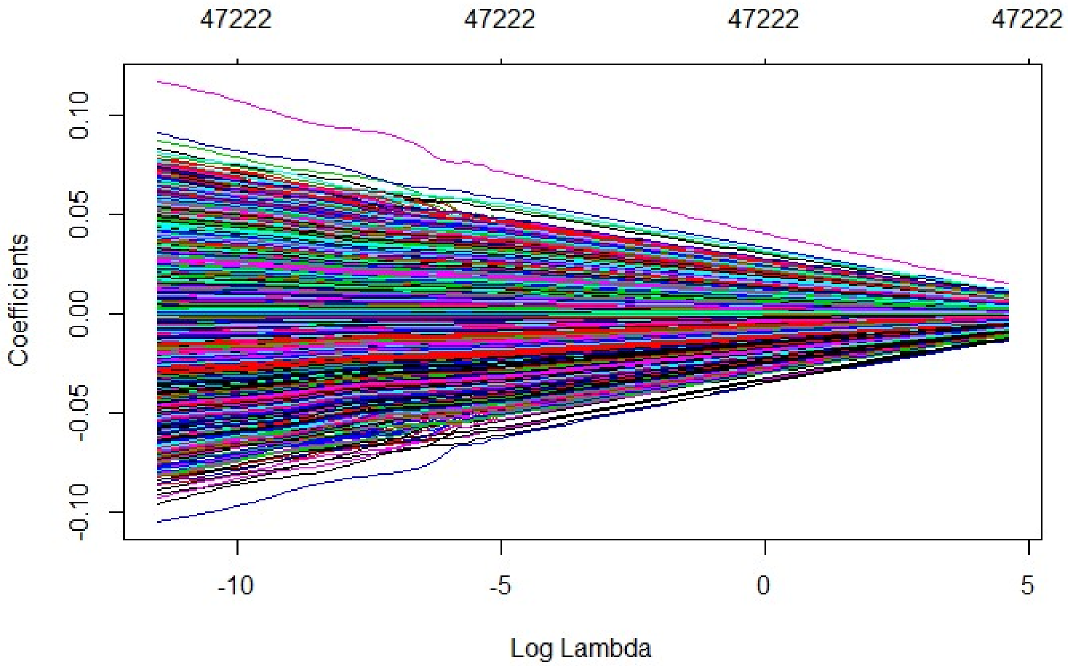


Figure 5: Ridge Regression

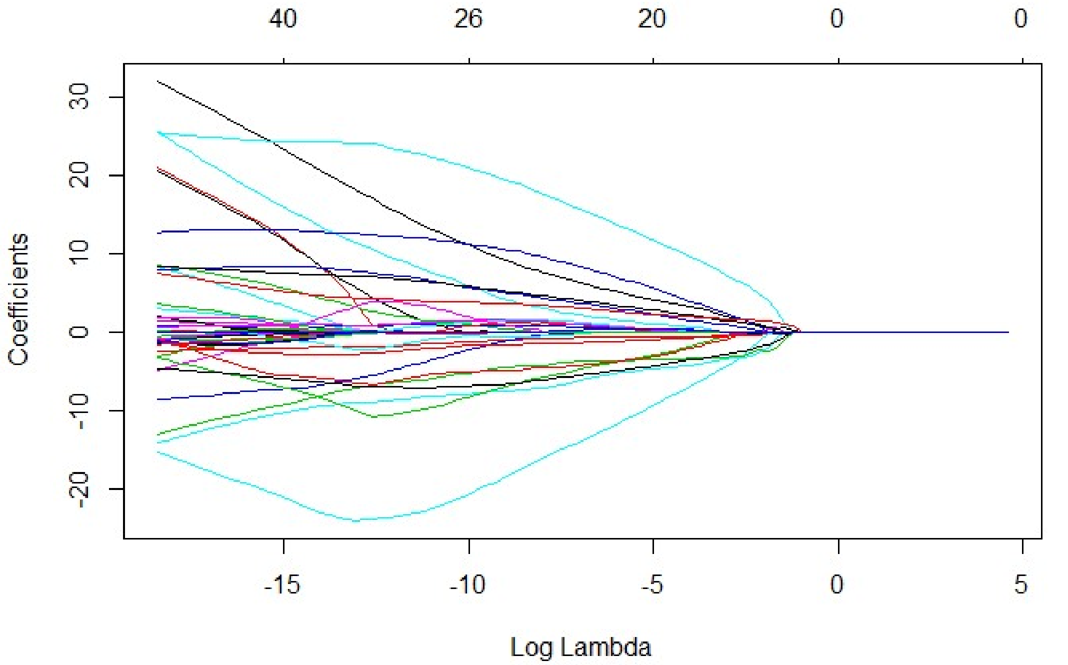


Figure 6: LASSO

Table 3: LASSO coefficients

Probe	Gene	Beta(j)
ILMN_1651913		21.15
ILMN_1652431		-3.07
ILMN_1652784		0.78
ILMN_1654942		3.06
ILMN_1655827		1.40
ILMN_1656052		20.58
ILMN_1656421		0.09
ILMN_1659378		-0.52
ILMN_1661886		-1.10
ILMN_1662807		8.34
ILMN_1663767		-4.82
ILMN_1667804		2.02
ILMN_1670219		0.27
ILMN_1670385		3.67
ILMN_1671004		0.01
ILMN_1673885		25.67
ILMN_1679647		-0.02
ILMN_1683036		-1.34
ILMN_1683678	SPATS2L	-2.43
ILMN_1685540		8.64
ILMN_1688749		-8.63

ILMN_1694742		-14.19
ILMN_1703855		32.11
ILMN_1712913		-3.31
ILMN_1722916		-15.38
ILMN_1735014	KLF6	-0.76

The variables selected by LASSO did include the two newly identified genes, SPATS2L and KLF6, but few of the other significantly up/down regulated genes listed. Our suspicion is the standardization of the expression levels masked some of the more strongly expressed genes by other associated genes. An interesting side note is the relatively moderate penalty required to force all the LASSO coefficients to zero. This is not surprising as the penalty is proportional to the sum of a function of the coefficients and with 47,222 coefficients this summation will grow to a significant penalty quickly.

Principal Component Analysis

The final reduction method explored was Principal Component Analysis. This method differs from the previous in that it doesn't depend on a response to perform feature reduction. Instead it simply selects the "direction" along which the data is most varied. Subsequent components are found in the same manner with the added requirement that the direction of each new component much be orthogonal to the previous ones (in other words uncorrelated). The result is a reduced set of independent features each blended from the original data set. The downside of PCA compared to subset or shrinkage methods is reduced visibility into the original feature set.

PCA.r performs the principal component analysis on the Psoriasis data set. As we are unable to compare the list of genes to that in the original paper the analysis how effectively the PCA based models predict the correct response. LOOCV validation was performed on models using the first 5, 10, 15 and 20 principal components fitted to a logistic regression model with the results displayed in Table 3.

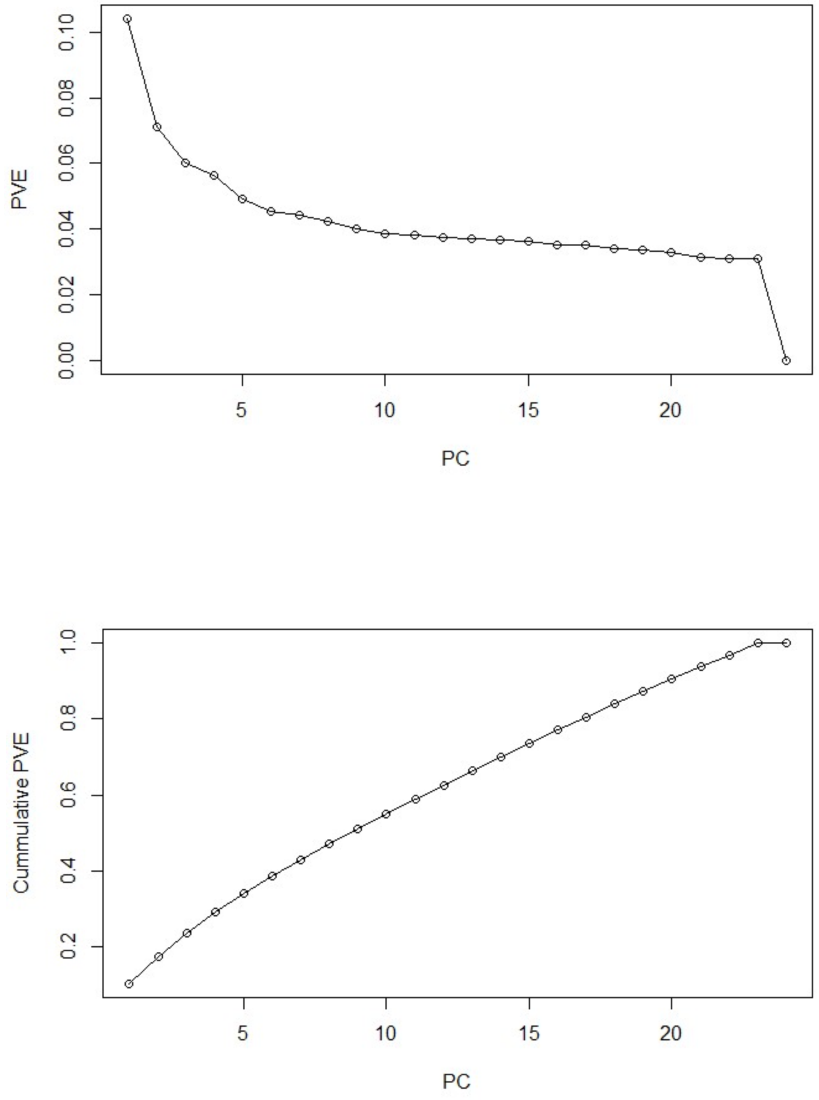


Figure 7: PVE and Cumulative PVE

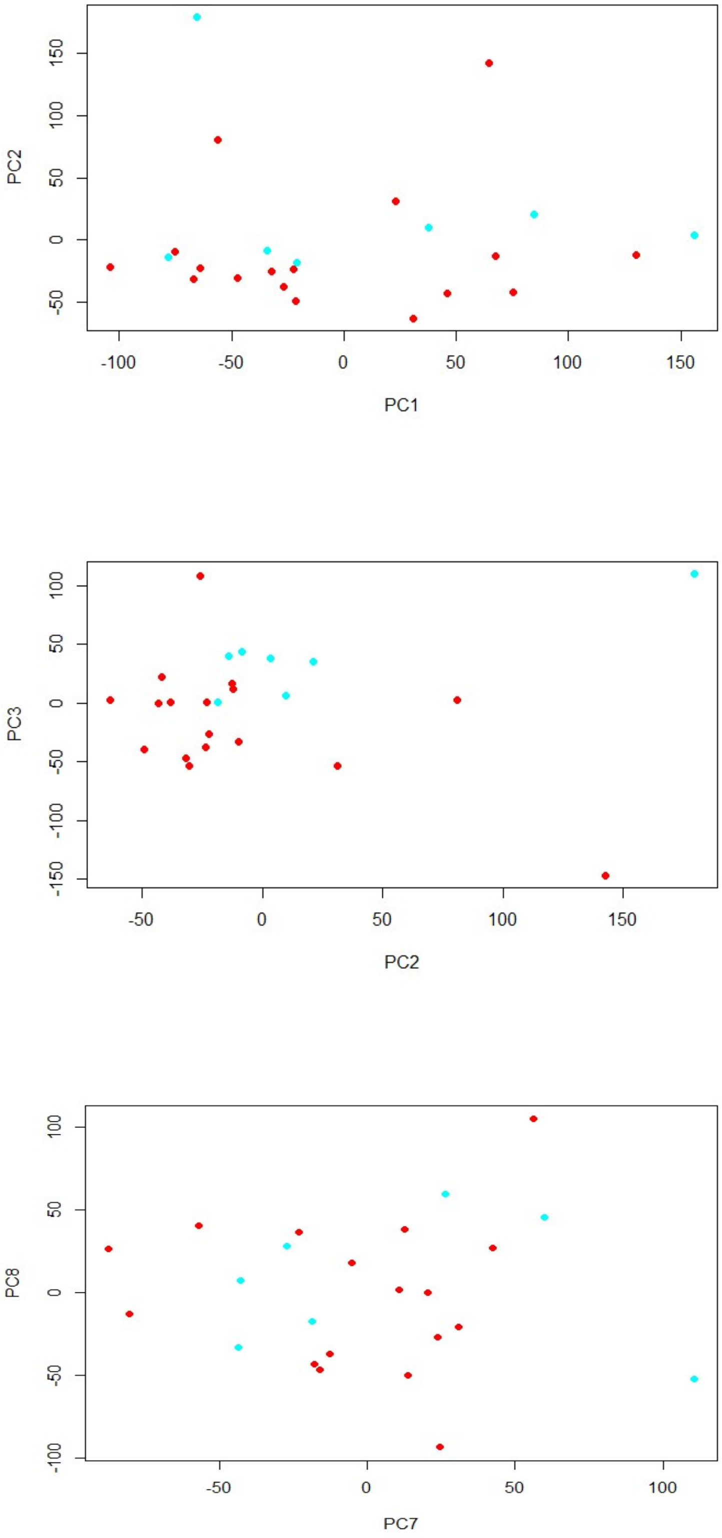


Figure 8: Visualizations of PCs (Red are patients, Blue are control samples)

Table 4: Misclassification Rate for different sized PCA models

		All Samples	LOOCV			
	Actual	10 PCs	5 PCs	10 PCs	15 PCs	20 PCs
GSM1152973	Yes	Yes	Yes	Yes	Yes	Yes
GSM1152974	Yes	Yes	Yes	Yes	No	No
GSM1152975	Yes	Yes	Yes	Yes	Yes	Yes
GSM1152976	Yes	Yes	Yes	Yes	Yes	No
GSM1152977	Yes	Yes	Yes	Yes	Yes	Yes
GSM1152978	Yes	Yes	Yes	Yes	Yes	No
GSM1152979	Yes	Yes	Yes	Yes	No	No
GSM1152980	Yes	Yes	Yes	Yes	Yes	Yes
GSM1152981	Yes	Yes	Yes	Yes	Yes	No
GSM1152982	Yes	Yes	Yes	Yes	Yes	Yes
GSM1152983	Yes	Yes	Yes	Yes	Yes	Yes
GSM1152984	Yes	Yes	Yes	Yes	Yes	Yes
GSM1152985	Yes	Yes	Yes	No	No	Yes
GSM1152986	No	No	No	Yes	Yes	No
GSM1152987	Yes	Yes	No	Yes	No	Yes
GSM1152988	No	No	No	No	No	No
GSM1152989	Yes	Yes	Yes	Yes	Yes	No
GSM1152990	No	No	Yes	No	No	No
GSM1152991	Yes	Yes	Yes	Yes	Yes	No
GSM1152992	No	No	No	No	No	No
GSM1152993	Yes	Yes	No	No	No	Yes
GSM1152994	No	No	Yes	Yes	Yes	No
GSM1152995	No	No	No	Yes	No	No
GSM1152996	No	No	No	Yes	No	Yes
Error		0	4	6	7	8
Error Rate		0.00%	16.67%	25.00%	29.17%	33.33%

Figure 7 illustrates the “Proportion of Variance Explained” by each principal component. The first component only explains ~10% of the variance with 100% of the variance reached with 23 principal components. Figure 8 helps visualize the relationship between PCs, with the PC2/PC3 plot demonstrating good clustering between the two sample groups.

Table 4 highlights an interesting result: as more principal components were added to the model, the ability of the model (using LOOCV validation) to accurately predict the presence of Psoriasis worsened. Examination of the scree plot of PVE in Figure 7 provides a clue to this result. At approximately seven principal components there is a “knee” were the additional amount of variance explained levels out. From this point on the additional principal components add little new information to the model in exchange for greatly increased noise and unwanted flexibility.

SVM

Another statistical learning method that was employed was Support Vector Machines (SVM). Here many SVMs were trained on the data using a grid search with epsilon ranging from 0 to 1 in increments of 0.01, and the cost ranging from 2^2 to 2^9 in powers of 2. Each of these 700 different SVMs was tested using 10-fold cross validation to determine optimal choices for cost and epsilon. The process was highly computationally expensive and took over 8 hours of computation time, however it yielded strong results. The best SVM utilized an epsilon value of 0 and a cost of 4. This SVM had a Root Mean Squared error of 0.0001073526, which is highly accurate. Figure 9 shows the error in each of the 24 patients. The data points were found to be completely separable. Figure 10 shows the results of the training. As can be seen larger values of epsilon produced higher error, with cost being somewhat independent of error. These results are to be expected, with data where the number of components from each sample far exceeds the number of samples. Since the data is perfectly separable, cost therefore has nearly no impact on the SVM.

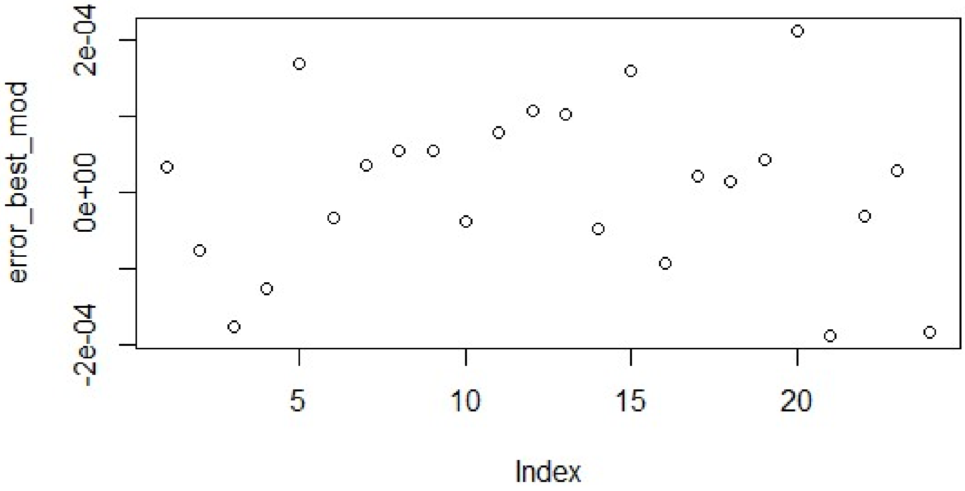


Figure 9: SVM error for each Patient

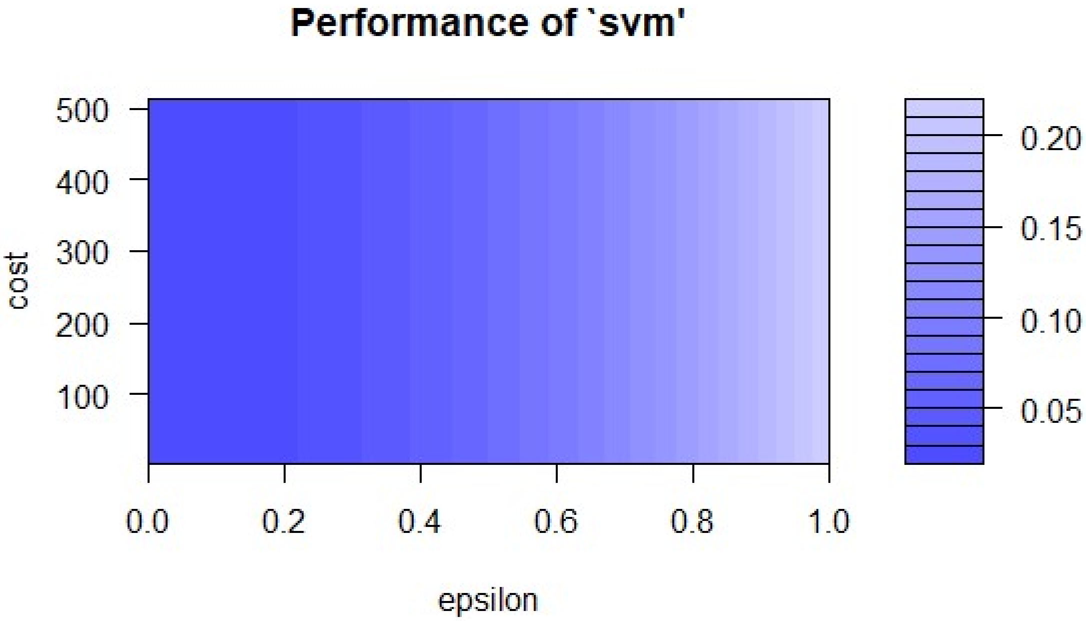


Figure 10: Performance of SVM using Grid search on Parameters

Unsupervised learning methods were also tried on the data. K-means was attempted to determine if clusters would form between the two classes without using the response variable. An important metric for K-means clustering is the distance. Figure 11 shows Euclidean distance between all points while Figure 12 shows the Manhattan distance. From these distance plots, one can see that some patients had more similar gene expression rates than others. K-means clustering was done using the Euclidian distance, with 50 random starts, taking the best of the 50. Attempting to split the data into 2 clusters produced clusters of size 9 and 15; however, the patients in each cluster had a similar rate of the disease compared to the original cluster. PCA was done to make the data viewable and Figure 13 shows the 2 clusters. Expanding the search to look from 2 to 8 clusters, the presence of the disease did not seem to be hugely impactful. Using 6 or more clusters, the patients who did not have the disease appeared to be more highly clustered; however, the disease still did not seem to be a significant factor. Table 5 shows the results of clusters with various sizes. Here patients who did not have the disease are denoted as ‘1’, while those who did are ‘0’.

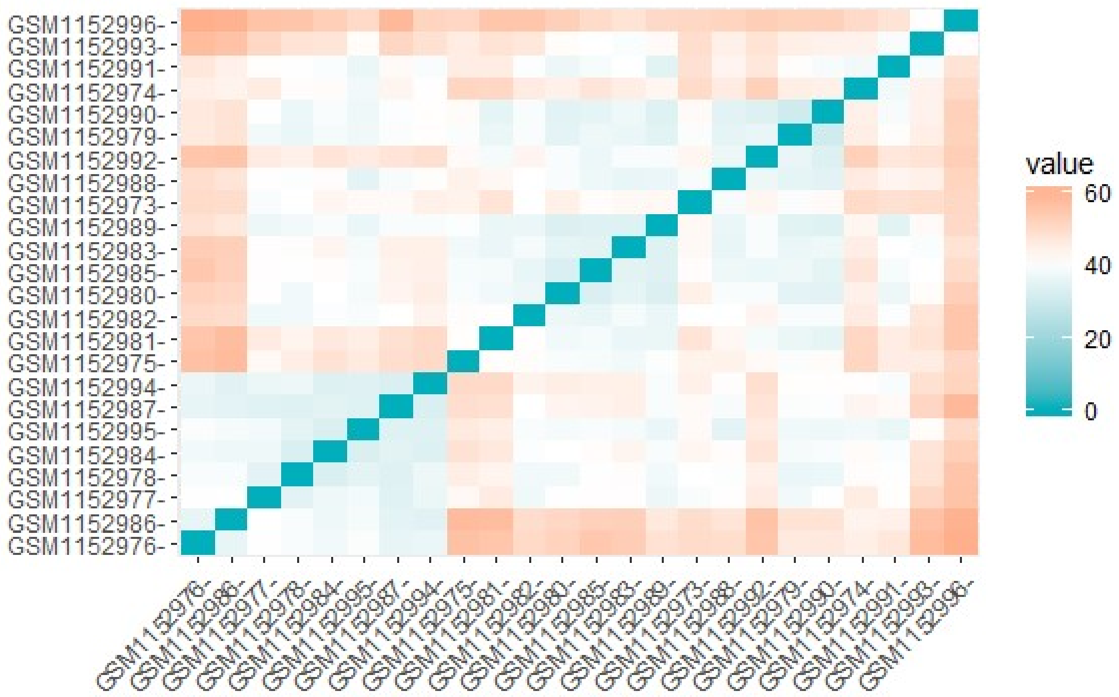


Figure 11: Euclidean Distance

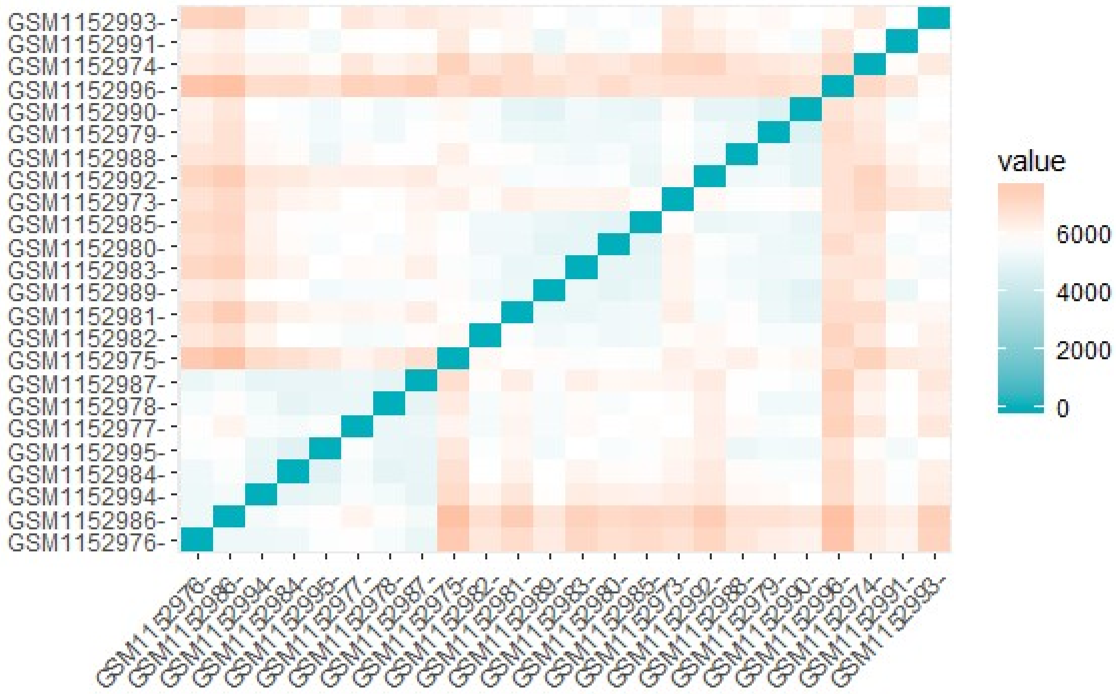


Figure 12: Manhattan Distance

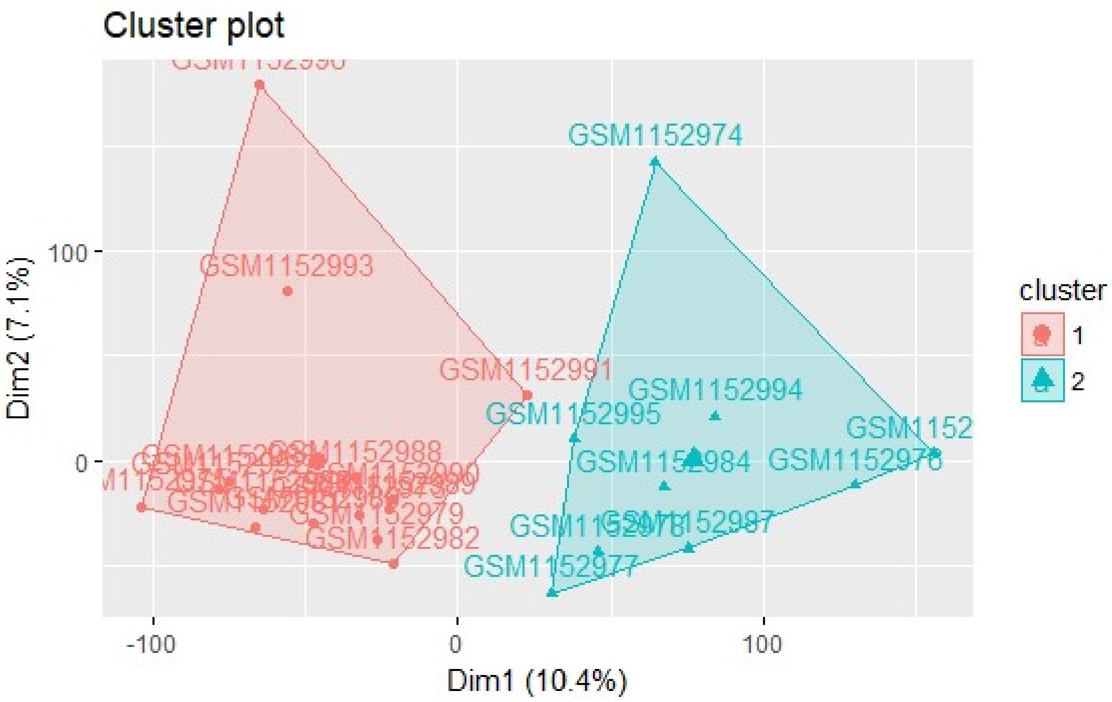
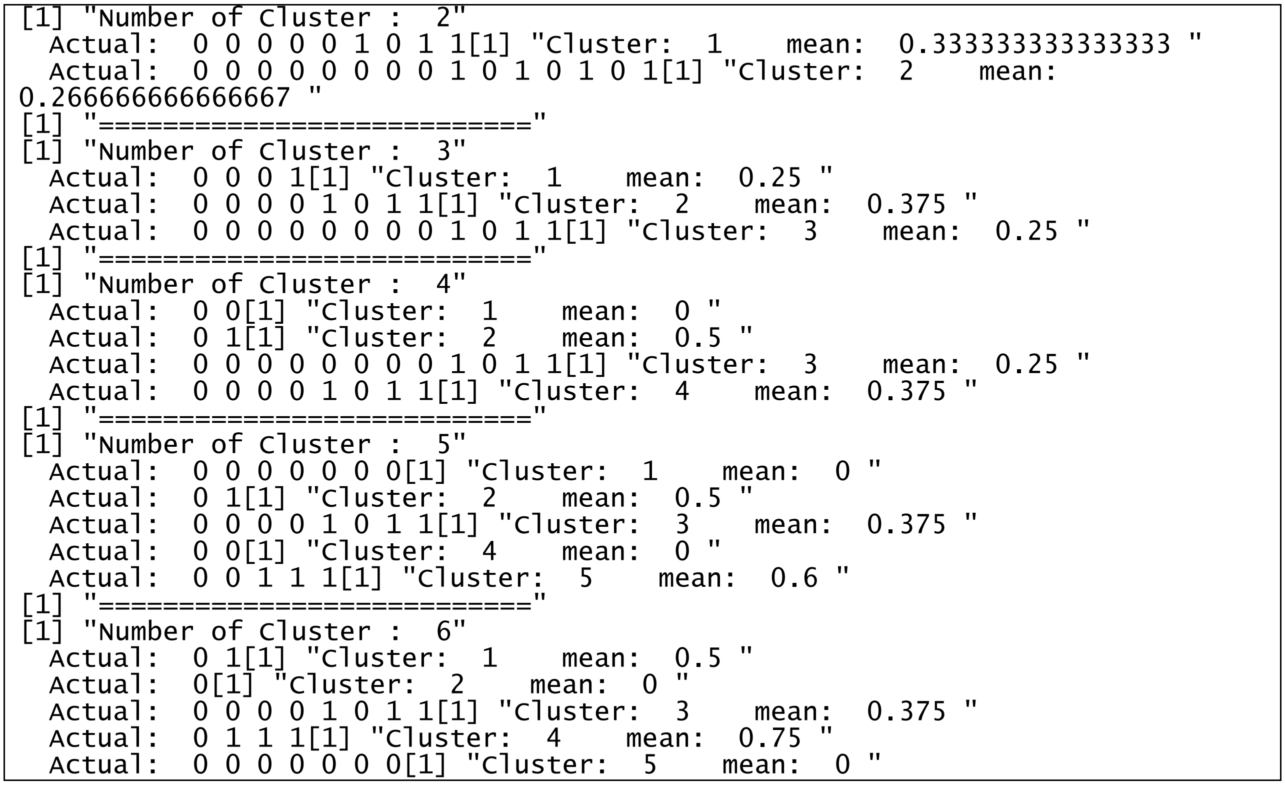
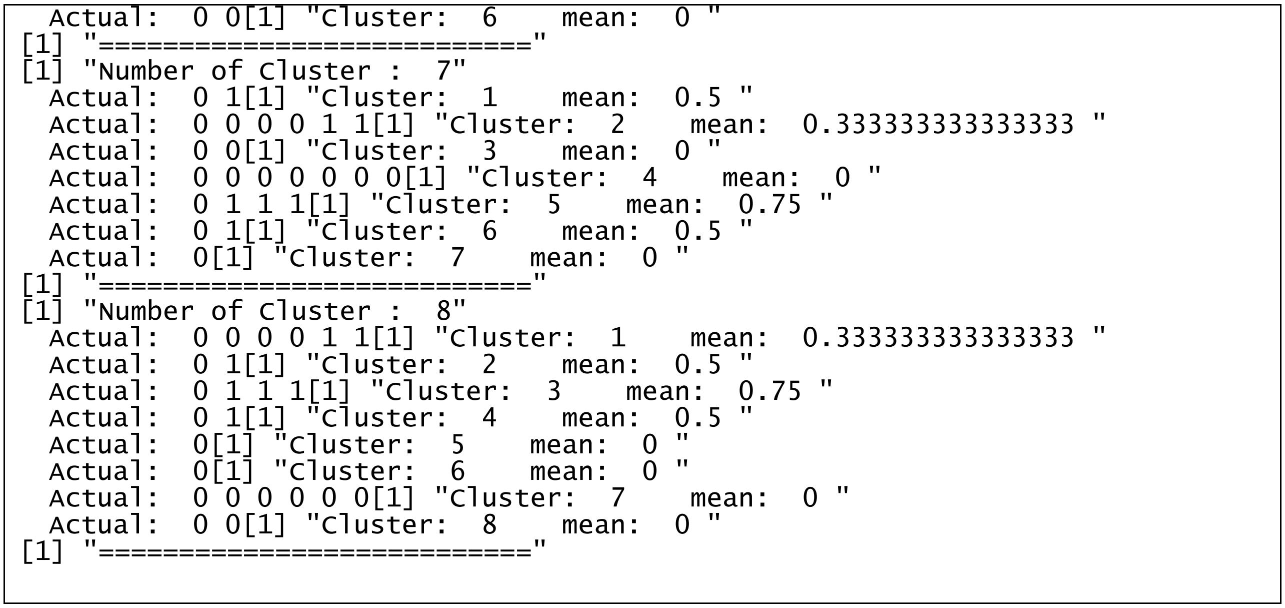


Figure 13: Clusters found using K-means

Table 5: results of Clusters with many different cluster sizes



```
[1] "Number of Cluster :  2"
  Actual:  0 0 0 0 0 1 0 1 1[1] "Cluster:  1    mean:  0.333333333333333 "
  Actual:  0 0 0 0 0 0 0 0 1 0 1 0 1 0 1[1] "Cluster:  2    mean:
0.266666666666667 "
[1] "=========================="
[1] "Number of Cluster :  3"
  Actual:  0 0 0 1[1] "Cluster:  1    mean:  0.25 "
  Actual:  0 0 0 0 1 0 1 1[1] "Cluster:  2    mean:  0.375 "
  Actual:  0 0 0 0 0 0 0 0 1 0 1 1[1] "Cluster:  3    mean:  0.25 "
[1] "=========================="
[1] "Number of Cluster :  4"
  Actual:  0 0[1] "Cluster:  1    mean:  0 "
  Actual:  0 1[1] "Cluster:  2    mean:  0.5 "
  Actual:  0 0 0 0 0 0 0 0 1 0 1 1[1] "Cluster:  3    mean:  0.25 "
  Actual:  0 0 0 0 1 0 1 1[1] "Cluster:  4    mean:  0.375 "
[1] "=========================="
[1] "Number of Cluster :  5"
  Actual:  0 0 0 0 0 0 0[1] "Cluster:  1    mean:  0 "
  Actual:  0 1[1] "Cluster:  2    mean:  0.5 "
  Actual:  0 0 0 0 1 0 1 1[1] "Cluster:  3    mean:  0.375 "
  Actual:  0 0[1] "Cluster:  4    mean:  0 "
  Actual:  0 0 1 1 1[1] "Cluster:  5    mean:  0.6 "
[1] "=========================="
[1] "Number of Cluster :  6"
  Actual:  0 1[1] "Cluster:  1    mean:  0.5 "
  Actual:  0[1] "Cluster:  2    mean:  0 "
  Actual:  0 0 0 0 1 0 1 1[1] "Cluster:  3    mean:  0.375 "
  Actual:  0 1 1 1[1] "Cluster:  4    mean:  0.75 "
  Actual:  0 0 0 0 0 0 0[1] "Cluster:  5    mean:  0 "
```



```
  Actual:  0 0[1] "Cluster:  6    mean:  0 "
[1] "=========================="
[1] "Number of Cluster :  7"
  Actual:  0 1[1] "Cluster:  1    mean:  0.5 "
  Actual:  0 0 0 0 1 1[1] "Cluster:  2    mean:  0.333333333333333 "
  Actual:  0 0[1] "Cluster:  3    mean:  0 "
  Actual:  0 0 0 0 0 0 0[1] "Cluster:  4    mean:  0 "
  Actual:  0 1 1 1[1] "Cluster:  5    mean:  0.75 "
  Actual:  0 1[1] "Cluster:  6    mean:  0.5 "
  Actual:  0[1] "Cluster:  7    mean:  0 "
[1] "=========================="
[1] "Number of Cluster :  8"
  Actual:  0 0 0 0 1 1[1] "Cluster:  1    mean:  0.333333333333333 "
  Actual:  0 1[1] "Cluster:  2    mean:  0.5 "
  Actual:  0 1 1 1[1] "Cluster:  3    mean:  0.75 "
  Actual:  0 1[1] "Cluster:  4    mean:  0.5 "
  Actual:  0[1] "Cluster:  5    mean:  0 "
  Actual:  0[1] "Cluster:  6    mean:  0 "
  Actual:  0 0 0 0 0 0[1] "Cluster:  7    mean:  0 "
  Actual:  0 0[1] "Cluster:  8    mean:  0 "
[1] "=========================="
```

5. Conclusions

In conclusion, while our findings did not exactly match the papers, we were able to examine the data in new ways. Our tree-based approaches revealed that the expression of the SPATS2L gene is enough to classify patients without any more data, which a very important fact supporting the arguments is made in the original paper. Our work reducing the model shows that the model can be massively reduced without significant loss of information. Using SVM's we were able to completely model the data with a very low error rate. Unsupervised learning however underscored the complexity of biology. Even though supervised methods were able to model the data with high accuracy, there are so many more elements influencing gene expression than just psoriasis. A major factor in all this analysis was the problem of a large feature space combined with a small sample set. This property makes drawing conclusions difficult, as there is not much data to back up any point of view (only 24 patients). Beyond the limited quantity of data, the emphasis on patients with psoriasis leads to the data being a poor model for the population. Future work can therefore be done gathering larger datasets for more complete evaluations. These new tactics proved very useful in providing new insights into this dataset.

6. References

Hastie, Tibshirani, Friedman. *The Elements of Statistical Learning*. New York, New York: Springer, 2009. Ebook.

James, Witten, Hastie, Tibshirani. *An Introduction to Statistical Learning with Applications in R*. New York, New York: Springer, 2013. Ebook.

Meltzer, Sean Davis and Paul. "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics* (2007): 1846 -- 1847.

Nuria Palau, Antonio Julia, Carlos Ferrandiz, Lluis Puig, Eduardo Fonsesa, Emilia Fernandez, Maria Lopez-Lasanta, Raul Tortosa, and Sara Marsal. "Genome-wide transcriptional analysis of T cell activation reveals differential gene expression associated with psoriasis." *BMC Genomics* (2013).

7. Source Code: The R code used in this paper was submitted separately

   Function List:

   1. ExamineData.R
   2. Lasso.r
   3. PCA.R
   4. Step.r
   5. Trees.r
   6. Trees2.r
   7. SVM.r
   8. Kmeans.r
   9. Ps-DataSet.r

8. Referenced Tables from Psoriasis paper

Table 6: Significantly Upregulated genes in Psoriasis patients compared to healthy controls

Probe	Gene	Fold Change	P value
ILMN_1683678	*SPATS2L*	1.37	0.0009
ILMN_1735014	*KLF6*	1.32	0.0012
ILMN_1703263	*SP140*	1.38	0.0025
ILMN_2322498	*RORA*	1.31	0.0041
ILMN_2246956	*BCL2*	1.23	0.0062
ILMN_2397721	*GLB1*	1.23	0.0062
ILMN_1781285	*DUSP1*	1.21	0.0071
ILMN_2321064	*BAX*	1.24	0.0096
ILMN_1731107	*CCDC92*	1.33	0.0136
ILMN_1729749	*HERC5*	1.72	0.0267
ILMN_2095660	*TMEM156*	1.29	0.0453
ILMN_1681301	*AIM2*	1.42	0.0464
ILMN_1767470	*SCPEP1*	1.26	0.0469
ILMN_1735979	*BCKDHA*	1.21	0.0639
ILMN_1719543	*MAF*	1.36	0.0658
ILMN_2188333	*CD69*	1.33	0.0677
ILMN_1776723	*PHF11*	1.24	0.0683
ILMN_1721626	*ARID5B*	1.27	0.0788
ILMN_1773742	*DNAJB9*	1.23	0.0805
ILMN_1741003	*ANXA5*	1.27	0.0896
ILMN_1729374	*ETFB*	1.23	0.0921
ILMN_1660368	*TRRAP*	1.20	0.0922
ILMN_2406410	*RHBDD2*	1.39	0.0925
ILMN_2305112	*CTH*	1.57	0.0944
ILMN_2284998	*SP100*	1.27	0.0944

Table 7: significantly Downregulated genes in Psoriasis patients compared to healthy controls

Probe	Gene	Fold Change	P value
ILMN_3286813	*LOC391019*	-1.38	0.0001
ILMN_3281502	*LOC653375*	-1.31	0.0009
ILMN_1778617	*TAF9*	-1.25	0.0009
ILMN_1689294	*LOC85390*	-1.20	0.0120
ILMN_2143250	*FAR1*	-1.20	0.0207
ILMN_1720114	*GMNN*	-1.20	0.0269
ILMN_2135175	*SNORD36A*	-1.27	0.0303
ILMN_1764163	*LOC644330*	-1.20	0.0577
ILMN_1746148	*LRRC33*	-1.22	0.0656
ILMN_1715401	*MT1G*	-1.82	0.0677
ILMN_2299072	*CROP*	-1.20	0.0683
ILMN_2124802	*MT1H*	-1.64	0.0778
ILMN_1655827	*COPS2*	-1.20	0.0800
ILMN_2402936	*LOC440926*	-1.22	0.0827
ILMN_3230435	*LOC729086*	-1.24	0.0886
ILMN_1803799	*LOC649555*	-1.21	0.0921
ILMN_1704873	*TCEB1*	-1.21	0.0970