Project
Transplant kidney rejection
High Dimensional Data Analysis

Jan Alexander* Annabel Vaessens† Steven Wallaert‡

24 05 2020

Executive summary

This research examines whether some genes are responsible for a patient's likelihood of rejecting a kidney after transplantation, for the Gene Expression Omninibus (GEO) dataset. This dataset consists of gene expression levels of 54675 genes from 282 patients. The variability of the gene expression is ... From the 54675 genes, 18080 genes are identified as having a differential expression between the group of rejected and the group of accepted kidneys. The list of these genes can be found in ... (insert appendix) Kidney rejection can be predicted sufficiently from the gene expressions with 17 genes. The most inmportant genes in predicting rejection are in... (insert appendix). The prediction model (insert which model is best) perfoms the best and (say something about the performance)

Contents

*jan.alexander@ugent.be
†annabel.vaessens@vub.be
‡steven.wallaert@ugent.be

1 Abbreviations

abbreviation	meaning
AUC	Area under the curve
IQR	Interquartile range
LDA	Linear discriminant analysis
LLE	Locally linear embedding
PCA	Principle component analysis
sd	Standard deviation

2 Exploratory Analysis

In this section general descriptive statistics are given and multiple methods for high dimensional data exploration are used.

2.1 Basic descriptive summary

In this study 54675 gene expression levels of 282 samples were analysed. In total, 76 or 27% of the transplanted kidneys were rejected.

Several descriptive statistics (mean, sd, median, iqr, min, and max) were calculated for every gene and kidney rejection status combination. This resulted in 2 (accepted vs. rejected) distributions of every statistic across genes. Note that these statistics were only calculated to perform a visual inspection.

The resulted plot is presented in figure 1. From this figure we can see there are, at least on this level, differences between the two groups. Most notable are the mean and median expression levels which tend to be closer to the overall mean (across groups) expression levels in the *accepted* group and more varying

in the *rejected* group. There seems to be more variablity in the measures of dispersion in the *rejected* group. Finally there are minimal differences between the min/max expression level distributions, perhaps suggesting that gene expression levels in the *rejected* group are slightly less extreme than in the *accepted* group.

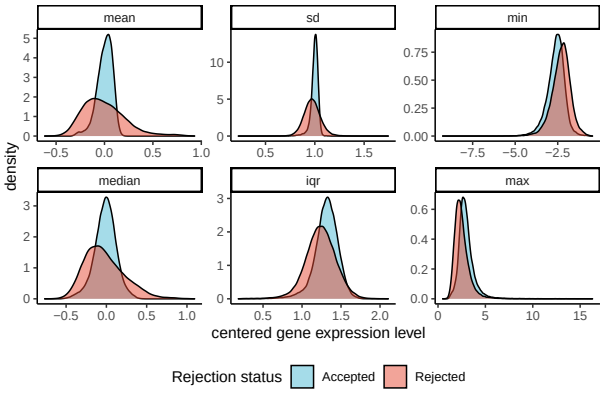


Figure 1: Descriptive statistics across grenes and between groups.

2.2 Advanced exploratory analyses

Multiple methods for exploration and visualisation of high dimensional data were applied (sparse PCA, MDS, sparse LDA, LLE, ISOMAP, Sammon Mapping, Diffusion maps, and t-SNE), yet without clear results. Because the sparse LDA gave the best results we discuss the results here and refer to the appendices **??** to **??** for the results of the other techniques.

The sparse LDA was performed to find potential candidate genes for future investigation. Due to computational constraints (our system ran out of memory) we needed to split the data set in 3 parts (each part consisting of 282 observations on 18225 genes). We considered this approach to be valid since we only used it as an exploratory tool. In total 116 genes (or 0.2%) had non-zero loadings.

Because this is still a substantial amount, only the genes with loadings in absolute value larger than two standard deviations were further considered ($|v_i| > 2sd(v)$, with $i = \{1, ..., 116\}$, where v_i is the ith loading). This resulted in a list of 10 genes. The list of genes can be found in appendix **??**.

By using these gene's loadings we calculated the scores of the linear discriminant for every sample and used these to construct the following graph.

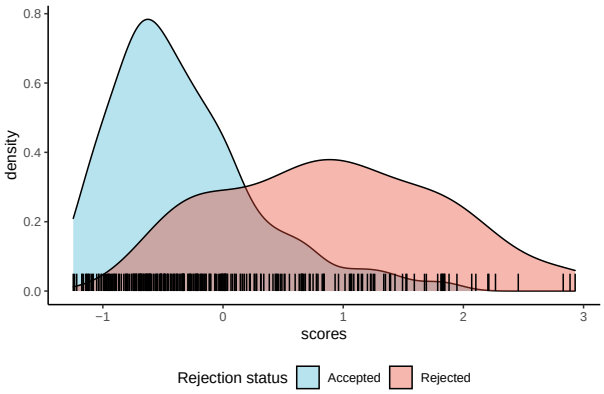


Figure 2: Density plot of linear discriminant scores based on the selected subset of genes.

From this graph we can see that to a degree a distinction can be made, albeit with a substantial overlap.

2.3 Conclusions Exploratory Analysis

Although differences between groups could be found within the data, no articulate distinction between the rejection status groups could be made with any of the used methods. This finding suggests there is relevant information at the genetic level w.r.t. transplant kidney rejection, but more factors need to be taken into account in order to arrive at a better understanding.

The main directions of variability in the gene expression dataset do not coincide with the separation between the rejection status groups. Nevertheless, certain genes were identified as potentially closely related to the differentiation between the two groups using sparse LDA.

3 Differentially expressed genes between kidney rejection groups

In order to find out which genes are differentially expressed between rejection status groups null hypotheses were tested against alternative hypotheses as follows.

$$\left.\begin{array}{l} H_{0,i} : \mu_{rejected,i} = \mu_{accepted,i} \\ H_{a,i} : \mu_{rejected,i} \neq \mu_{accepted,i} \end{array}\right\} \quad with i = \{1, \ldots, 54675\}$$

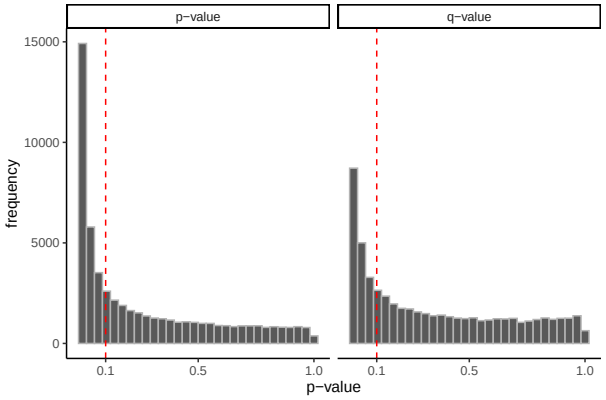
In these hypotheses $\mu_{rejected,i}$ and $\mu_{accepted,i}$ are the population means of the gene expression level of the

ith gene in the rejected and accepted kidney rejection group respectively. Before testing, a visual inspection of 30 QQ plots, from 15 randomly drawn variables, was done to assess whether the variables follow a normal distribution for both groups separately. These QQ plots showed that some genes were normally distributed, but also that many genes were not.

Nevertheless, two-sided Welch t-tests were performed on the uncentered data to determine whether the two groups can be differentiated based on the gene expression level for every gene. The choice for the Welch t-test is motivated by the presence of unequal variances between the two groups, even though not all genes were normally distributed. We included a small random subset 6 QQ plots in the appendix so the reader, if she/he wishes, can have a rough idea of the divergence from normality.

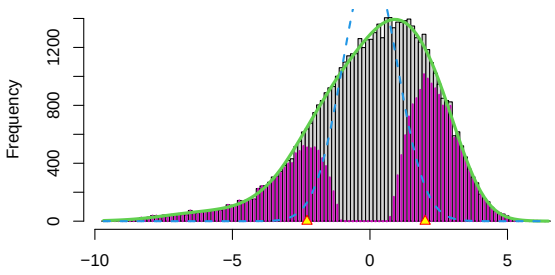
To address the multiple testing problem at this large scale (54675 simultaneous tests), the FDR is controlled at 0.10 through application of the method of Benjamini and Hochberg (1995).

To summarise the results we constructed a histogram of both the regular and adjusted p-values or q-values.



Both histograms reveal non-uniform distributions. More importantly, they have many small values, indicating that for many genes the null hypothesis was rejected. Based on the q-values, there were 18081 rejected null hypotheses. As such we conclude that the mean gene expression differs between the accepted and rejected kidney groups for those 18081 genes. As the FDR is controlled at 10%, it is expected to have around 1808 false discoveries.

Next, the normalised test statistics (z-scores) are plotted and compared to the local false discovery rate.

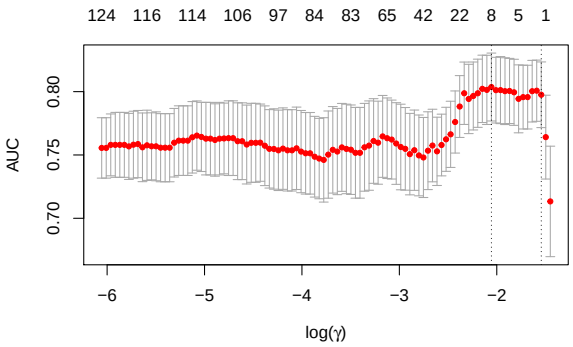


From the figure **??** can be concluded that a small lfdr can be obtained for small and large z-scores (a lfdr smaller than 0.2 for z-scores smaller than -2 and larger than 2). For these z-scores, it is more likely that when rejecting the null hypotheses, a true discovery is made.

4 Prediction of kidney transplant rejection

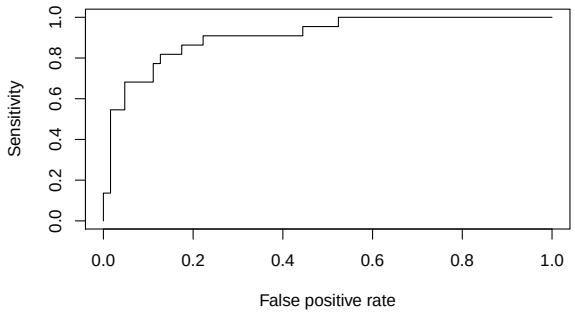
The objective of this final part is to construct a classifier for kidney acceptance or rejection based on the measured gene expressions. Three approaches are compared: lasso, ridge regression and principle component regression. These approaches will be compared based on the AUC. The final modelling approach is chosen as the one that shows the largest cross-validated AUC. The cutoff is chosen based on the F1-score. The dataset is split into a training and test dataset.

4.1 Lasso regression

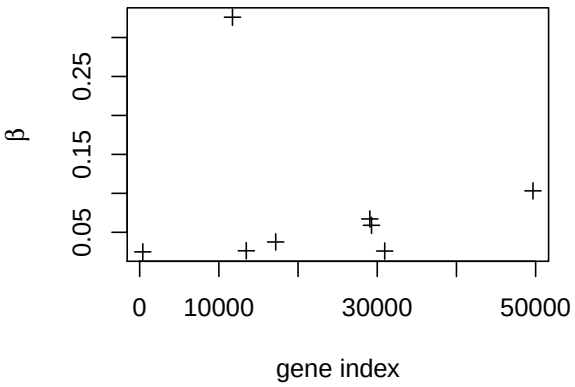


In the figure above, one can see that for γ equal to 0.1280675 , the area under the curve (AUC) is maximal (0.909) for the train dataset based on a 10-fold cross-validation over the train dataset.

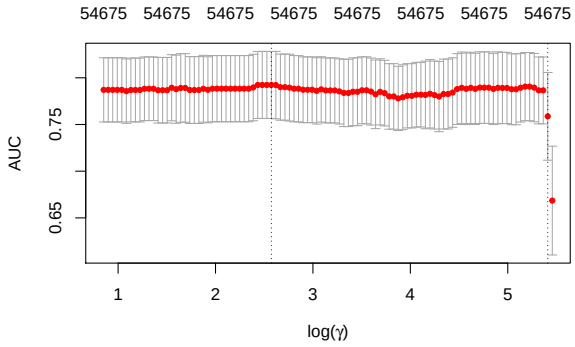
The ROC curve, estimated with the cross-validation dataset, is shown below:



This model only uses 8 of the genes: 396, 11720, 13463, 17178, 29063, 29276, 30941, 49664. Compared to the interesting genes found with the Sparse LDA method: This is a considerable dimensional reduction. This is illustrated below. This figure shows the loadings of the 8 selected values.

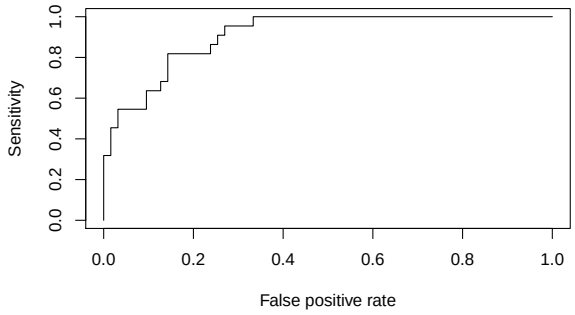


4.2 Ridge regression

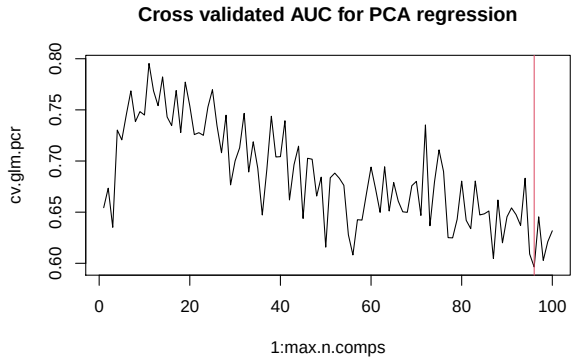


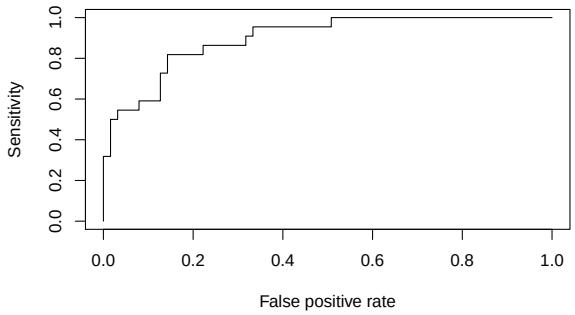
Likewise as for the Lasso regression, for γ equal to 13.1081041, the optimal AUC (0.911) is obtained.

The ROC curve, estimated with the cross-validation dataset, is shown below:



4.3 Principal component regression



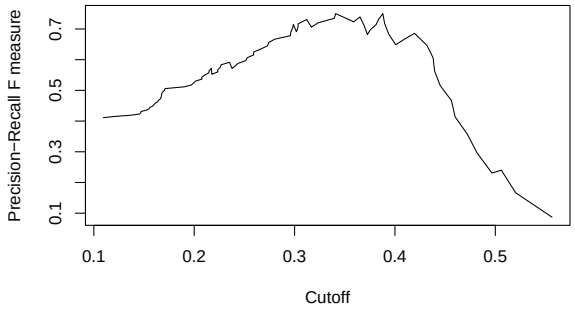


AUC is 0.899.

4.4 Final model evaluation

The model performance in terms of AUC is extremely similar for the 3 models. Since LASSO regression is the simplest model, this is preferred. By chosing cutoff c = 0.23, we can achieve a F1-score of over 0.7. This is represented on the following graph. Below, the F1 graph, the confusion matrix for the PCR model with cutoff c = 0.21 is shown.

The confusion matrix shows no false negatives at all, but the number of false positives is high.



	accept predicted	reject predicted
accept observed	35	28
reject observed	1	21

This confustionmatrix clearly shows a sensitivity of 0.4285714, and a specificity of 0.9722222. In short, this predictor seems to strike a balance between both, but the performance is not perfect. If a person tests positive, there is still a considerable chance the kidney will not be rejected.

This behaviour could be *explained* (or at least understood a little better) when looking back at the exploratory analysis. Here it was already clear that the gene expressions of the patients with rejected kidneys overlap with those of the patients with accepted kidneys. Both are not perfectly separable.

5 Conclusions

From the 54675 genes in the dataset, 18080 genes are differentially expressed between the two kidney groups, based on multi-scale Welch t-test at an FDR of 10%.

6 Appendices

6.1 Exploration methods for high dimensional data

6.1.1 Sparse principle components analysis

Unfortunately, naive sparce principle component analysis cannot be used to make a distinction between the accepted and rejected kidneys.

6.1.2 Multi-dimensional scaling:

In the biplots of the three first dimensions of the svd (**??**), no distinction can be made between rejected and accepted kidneys.

In the scree plot **??** it can be seen that the two first dimensions account for only 25% of the total variance in the dataset and the first three dimensions for 29%. To account for 80% of the total variance, 120 dimensions are needed.

6.1.3 LLE (locally linear embedding)

Locally linear embedding described by Roweis and Saul (2000) was performed. From figures **??** can be seen that no distinction between the accepted and rejected kidneys can be made.

6.1.4 ISOMAP

ISOMAP presented by Tenenbaum, Silva and Langford in 2000 is performed. The parameter k is varied manually so that the maps are optimal. From figure **??** can be seen that with ISOMAP, it is also not possible to make a distinction between the group of accepted and rejected kidneys.

6.1.5 Sammon mapping

Sammon mapping presented by Sammon (1969). The result is in figures **??**: no distinction can be made between the two groups.

6.1.6 Diffusion maps

Diffusion mapping was presented by Nadler et al (2006) and LAfon and Lee (2006). From figure **??**, no distinction between the two groups can be made with diffusion maps.

6.1.7 t-SNE

t-stochastic neighbor embedding is presented by Van Den Maaten and Hilton (2008). The resulting plot can be seen in figure **??** and indicates again that a simple distinction between the two groups cannot be made. Yet, there seems to be roughly two groups that differ in heterogeneity: one largely heterogeneous group and one group that is less heterogeneous, though far from homogeneous.

6.2 QQ-plots

7 References

Benjamini Y and Hochberg Y, 1995. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. Journal of the Royal Statistical Society. Series B: Methodological 57, 289-300.

Lafon S and Lee AB, 2006. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, 1393–1403.

Nadler B, Lafon S, Coifman RR, and Kevrekidis IG, 2006. Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems. Applied and Computational Harmonic Analysis: Special Issue on Diffusion Maps and Wavelets, 21, 113–127.

Roweis ST and Saul LK, 2000. Nonlinear dimensionality reduction by locally linear embedding. Science, 290, 2323-2326.

Sammon JW, 1969. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 18, 401–409.

Tenenbaum JB, De Silva V and Langford JC, 2000. A global geometric framework for nonlinear dimensionality reduction. Science, 290, 2319-2323.

Van Der Maaten L and Hilton G, 2008. Visualising data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605.