Project
Transplant kidney rejection
High Dimensional Data Analysis

Jan Alexander* Annabel Vaessens† Steven Wallaert‡

27 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. Data exploration methods show that there is no clear way to map the gene expression levels onto the kidney rejection statuses. In other words, only probabilistic claims can be made. From the 54675 genes, 18081 genes are identified as having a differential expression between the group of rejected and the group of accepted kidneys. Kidney rejection can be predicted sufficiently from the gene expressions with 17 genes.

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
LASSO	Least absolute shrinkage and selection operator
LLE	Locally linear embedding
MDS	Multi dimensional scaling
PCA	Principle component analysis
ROC curve	Receiver operating characteristic curve
sd	Standard deviation
t-SNE	t-distributed stochastic neighbor embedding

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.

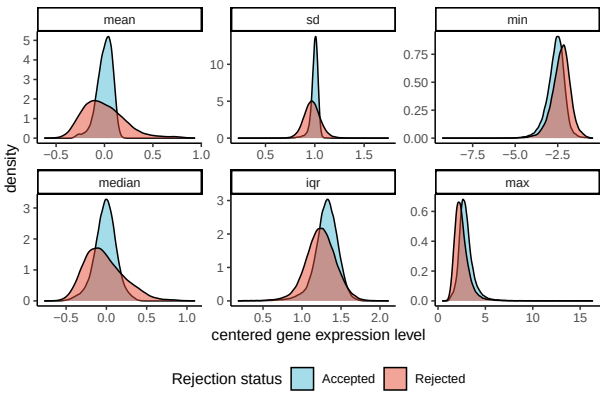

Figure 1: Descriptive statistics across grenes and between groups.

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 (6.1.2 to 6.1.7) 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 reasonable since we only used it as an exploratory tool. A small simulation was performed to verify its potential as such a tool (see appendix 6.1.8). 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: 11719, 15960, 16361, 19568, 29636, 30940, 38692, 45591, 49663, 53849.

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.

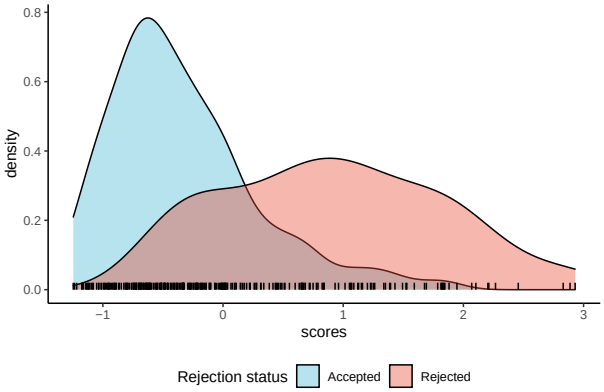

Figure 2: Density plot of linear discriminant scores based on the selected subset of genes.

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 hy-

potheses were tested against alternative hypotheses as follows.

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

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 the adjusted p-values or q-values (see figure 3).

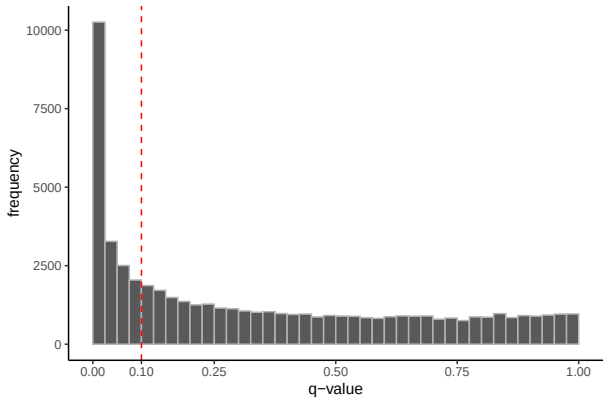

Figure 3: Histograms of adjusted p-values. The dashed line indicates the threshold

The histogram shows a non-uniform distribution. More importantly, it tells there are 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.

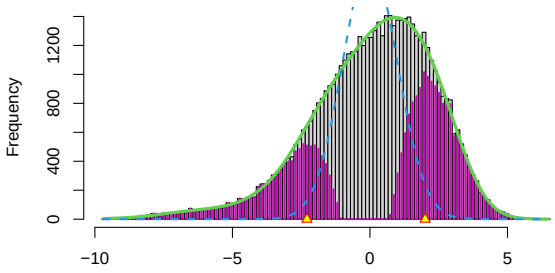

Figure 4: Histogram of normalised test statistics. Approximated density (green line overlay). Theoretical null distribution density (blue dashed line overlay). True discovery likelihood indication (purple overlay).

Fromthe graph in figure 4 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 hypothesis, 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 (70% of the data) and test dataset (30% of the data).

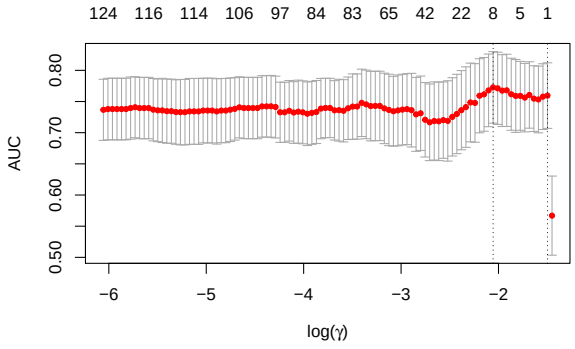

Figure 5: CV plot LASSO regression.

4.1 LASSO regression

In figure 5, one can see that for γ equal to 0.13, the 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 in figure 6.

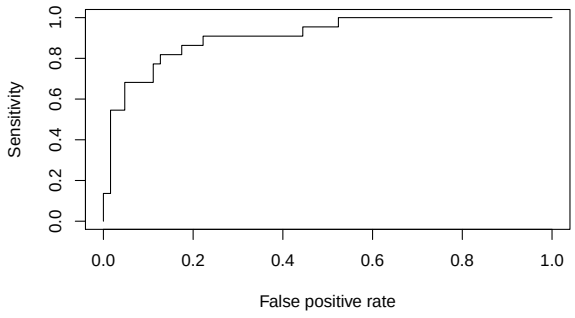

Figure 6: ROC curve LASSO regression model.

This model only uses 8 of the genes: 395, 11719, 13462, 17177, 29062, 29275, 30940, 49663. 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 66.77, the optimal AUC (0.896) is obtained.

The ROC curve, estimated with the cross-validation dataset, is shown in figure 9.

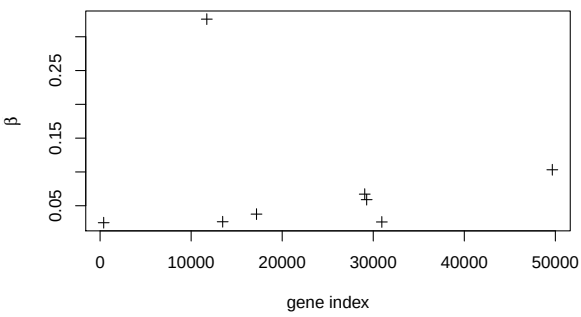

Figure 7: Coefficient plot LASSO regression model.

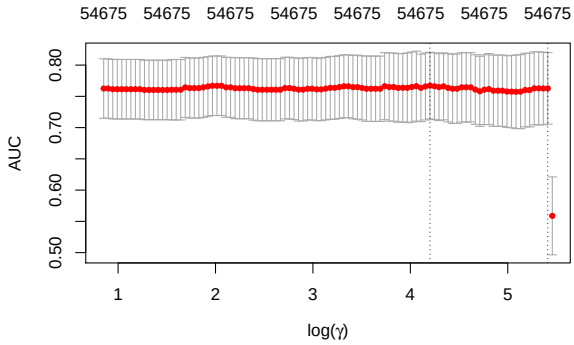

Figure 8: CV plot ridge regression.

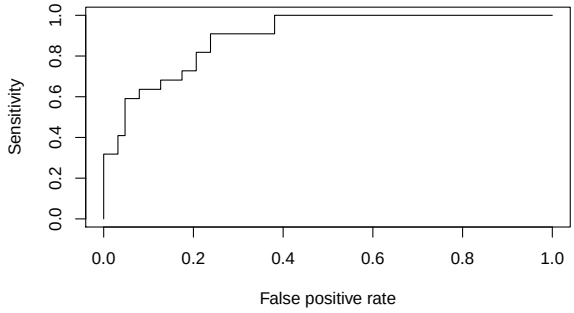

Figure 9: ROC curve ridge regression model.

4.3 Principal component regression

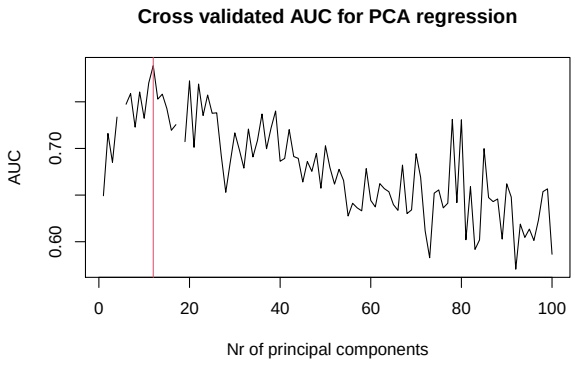

Figure 10: CV plot PCR. The red line shows the number of principal components for which the cross-validete AUC is maximal.

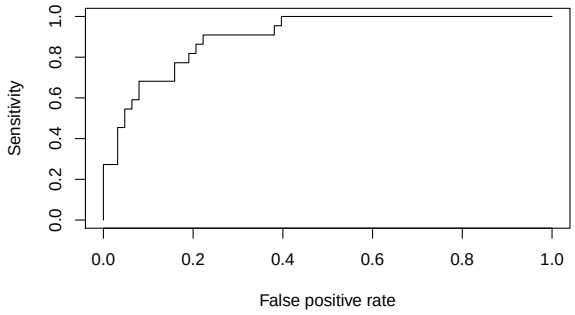

Figure 11: ROC curve PCR model.

The AUC for this model is 0.902.

4.4 Final model evaluation

The model performance in terms of AUC is similar for the 3 models. Since LASSO regression is the simplest model, this is preferred. By chosing cutoff c = 0.34, we can achieve a F1-score of 0.75. This is represented on the following graph. Below, the F1 graph, the confusion matrix for the Lasso model with cutoff c = 0.34 is shown.

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

This confusionmatrix clearly shows a sensitivity of 0.77, and a specificity of 0.87. In short, this predictor seems to strike a balance between both, but the performance is not perfect. If a person tests positive,

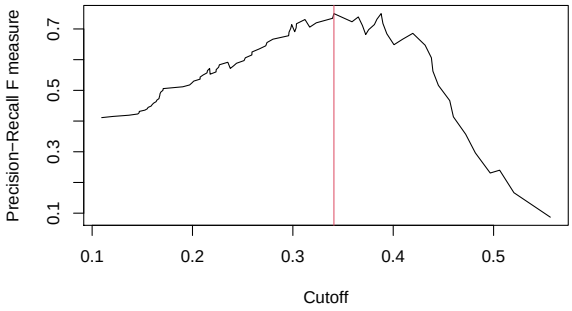

Figure 12: Cutoff value selection based on F1 score. The red line shows the cutoff value for which the F1 is maximal.

	accept pred	reject pred
accept obs	55	8
reject obs	5	17

Table 1: Confusion matrix final model.

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

In the exploratory analysis we found that the 2 groups are different, but not different enough to be able to make clear distinction. Through applyication of sparse LDA a few genes were flagged as potentially interesting for further research. From the 54675 genes in the dataset, 18081 genes are differentially expressed between the two kidney groups, based on multi-scale Welch t-test at an FDR of 10%.
From the 3 modeling approaches used in this study, the LASSO regression gave the best results both in terms of AUC (based on the test data) and in terms of parsimony. Interestingly, from the selected genes by the LASSO model, the genes 11719, 30940, 49663 were also detected by the sparse LDA.

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 described by Roweis and Saul (2000) was performed. From the next 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.

6.1.8 Simulation sparse LDA

In order to have an idea whether sparse LDA split in 3 parts can be used as an exploratory tool we ran a small simulation. We simulated high dimensional data in such a way that also the split parts were high dimensional (n = 33, p = 102). We kept the number of observations and variables as low as possible to make it computationally feasible and still high enough to be able to get some insights from the results. We constructed the data in such a manner that only 2 variables were predictive of the response (note this is 2%, which is different from the real data: 0.2%). The response was constructed in such a way that approximately 27% were successes (allowing for variation over simulation repetitions).

The simulation was repeated 50 times and we looked at following measures: correlation between coefficients (mean correlation =0.79, $.25^{\text{th}}$ and $.75^{\text{th}}$ quantiles (0.72, 0.95), 6 correlations were not computable because one or both of the methods gave only 1 or no coefficients), proportion of simulations in which at least 1 variable was detected by both methods (0.84), proportion of variables detected by the full method that also were detected by the split method (mean proportion = 0.85, $.25^{\text{th}}$ and $.75^{\text{th}}$ quantiles =(0.67, 1, higher is better), and proportion of variables detected by the split method that weren't detected by the full method (mean proportion =0.47, $.25^{\text{th}}$ and $.75^{\text{th}}$ quantiles (0, 0.79, lower is better).

Although this not a formal way of making a comparison, we feel confident in using this approach, albeit only as an exploratory tool. Please note that we don't claim that the method is valid.

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 parameteri-

zation. 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.