Project
Transplant kidney rejection
Analysis of High Dimensional Data

Jan Alexander* Annabel Vaessens† Steven Wallaert‡

31 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 (not without error) from the gene expressions with only as few as 8 genes. These genes are 1552807_a_at, 202270_at, 204014_at, 207735_at, 219777_at, 219990_at, 221658_s_at, and 240413_at. From these, **202270_at, 221658_s_at, 240413_at** are also proposed by the exploratory analysis as potentially predictive for kidney rejection status.

Contents

Abbreviations

Abbreviation	Meaning
AUC	Area under the (ROC) curve
CV	Cross validation
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

1 Exploratory Analysis

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

1.1 Basic descriptive summary

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

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

planted 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 results are summarised in figure 1. From this figure we can see there are differences between the two groups, at least on this level. Most notable are the mean and median expression levels which tend to be closer to the overall mean 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 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, though these differences are rather minimal.

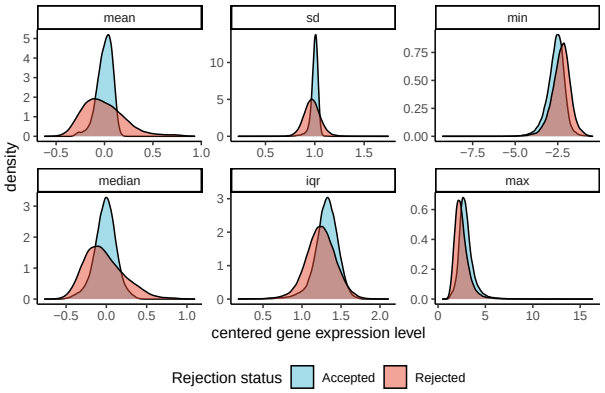


Figure 1: Descriptive statistics across grenes and between groups.

1.2 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 (sections 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 limitations (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 section 6.1.8 in the appendices). 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 $i = \{1, ..., 116\}$, sd_v the standard deviation of the loadings, and v_i the ith loading). This resulted in a list of 10 genes: 202270_at, 206513_at, 206914_at, 210163_at, 220351_at, 221658_s_at, 229437_at, 236341_at, 240413_at, 244598_at.

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.

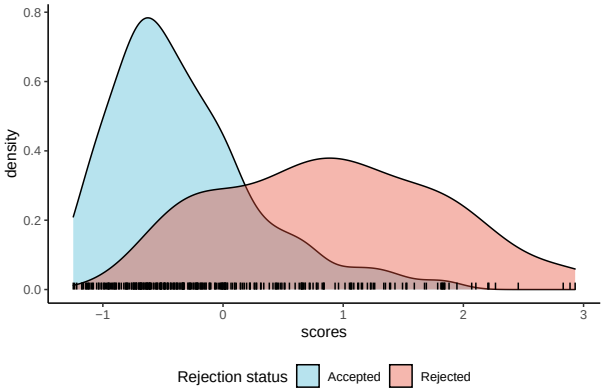


Figure 2: Density plot of linear discriminant scores based on the selected subset of genes.

From this graph we can see that to a degree a distinction can be made, albeit not without a substantial overlap.

1.3 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 should 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.

2 Testing for differential expression

In order to find out which genes are differentially expressed between rejection status groups the following null hypotheses were tested against the following alternative hypotheses:

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

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 some 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 of 6 QQ plots in the appendix (section 6.2) so that the reader, if she/he wishes, can have a rough idea of the divergence from normality (or the absence thereof).

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).

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.

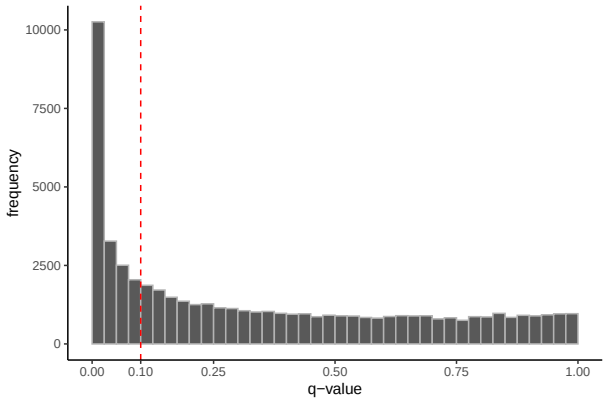


Figure 3: Histograms of adjusted p-values. The dashed line indicates the threshold

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

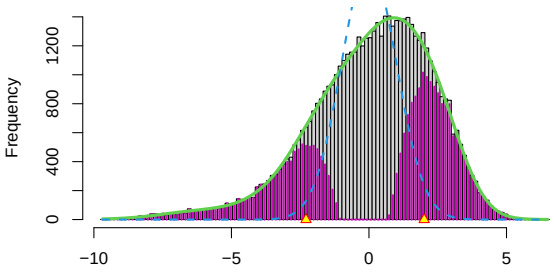


Figure 4: Histogram of normalised test statistics. Approximated density (green line overlay). Theoretical null distribution density (blue dashed line overlay). True discovery likelihood indication (purple overlay).

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 -2.29 and larger than 2.01). For these z-scores, it is more likely that when rejecting the null hypothesis, a true discovery is made.

An online lookup table of all differentially expressed genes, together with their q-values and local false discovery rate can be consulted at https://users.ugent.be/~swallaer/aohd/.

3 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 regression, ridge regression and principle component regression. Each of these approaches will yield a 'best' model, based on the cross validate AUC, using a dedicated training part of the data set (random subset of 70% of the data). Of these 3 'best' models, the model with the largest AUC, calculated using an independent test part of the data (random subset of 30% of the data)—that is, data the models haven't seen yet—, will then be selected as the final model.

3.1 LASSO regression

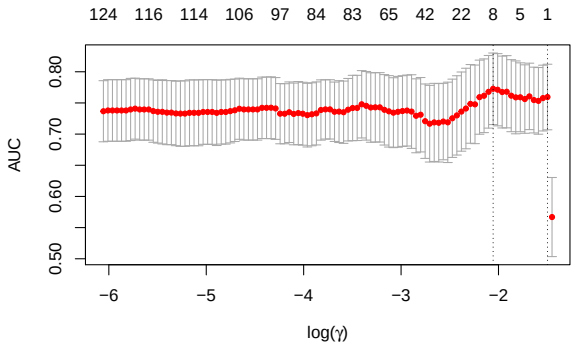


Figure 5: CV plot LASSO regression.

In figure 5, one can see that for γ equal to 0.13, the cross validated AUC is maximal for the train dataset based on a 10-fold cross-validation over the train dataset.
The ROC curve, estimated with the test dataset, is shown in figure 6. The corresponding AUC equals 0.909.

This model only uses 8 of the genes: 1552807_a_at, 202270_at, 204014_at, 207735_at, 219777_at, 219990_at, 221658_s_at, 240413_at. This is a considerable dimensional reduction. Table 1 shows the coefficients of the selected genes.

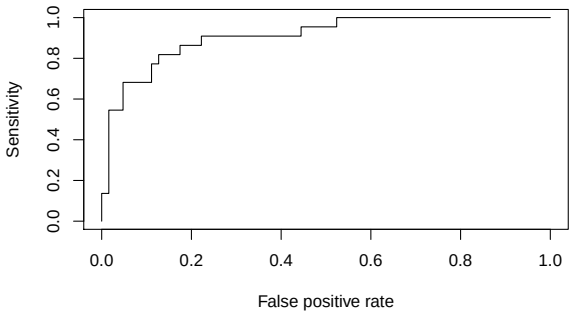


Figure 6: ROC curve LASSO regression model.

Genes	Parameter value
intercept	-1.028
1552807-a-at	0.025
202270-at	0.326
204014-at	0.026
207735-at	0.038
219777-at	0.067
219990-at	0.059
221658-s-at	0.026
240413-at	0.103

Table 1: Coefficients LASSO model

3.2 Ridge regression

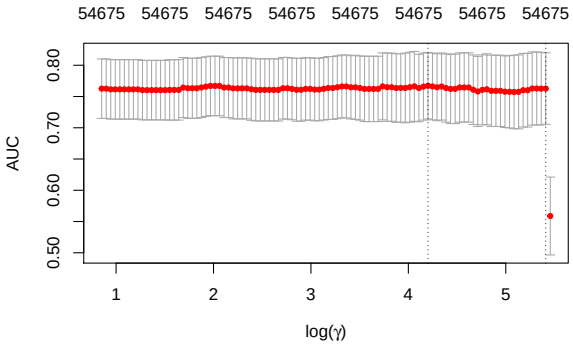


Figure 7: CV plot ridge regression.

For the ridge regression a γ equal to 66.77 delivered an optimal cross validated AUC (also using 10-fold cross validation on the training set).

The ROC curve, estimated with the test dataset, is

shown in figure 8. The corresponding AUC equals 0.896.

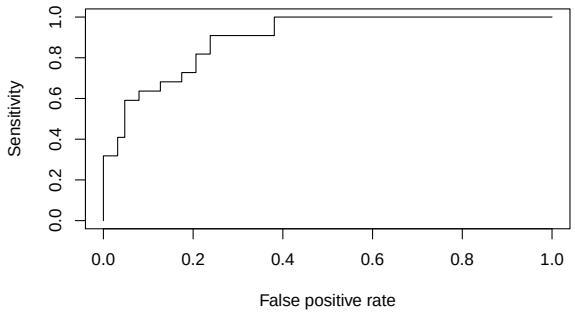


Figure 8: ROC curve ridge regression model.

3.3 Principal component regression

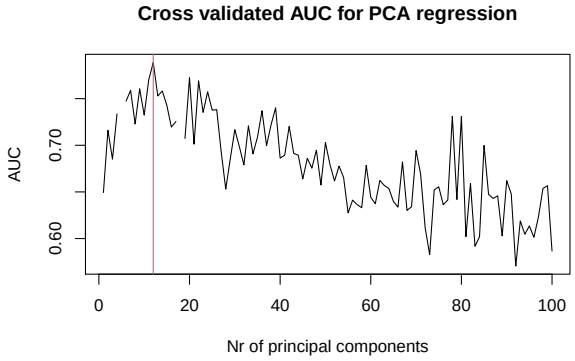


Figure 9: CV plot PCR. The red line shows the number of principal components for which the cross-validete AUC is maximal.

For the principal component regression, the optimal number of principal components was chosen based on the cross validated AUC (using the training dataset). Figure 9 shows the AUC vs. number of components plot (using only 1 to 100 components for computational reasons). The selected model had 12 components. Figure 10 shows the ROC curve for this model, based on its performance on the test dataset. The corresponding AUC for this model is 0.902.

3.4 Final model evaluation

Table 2 summarises the performances and complexities of the three selected models.

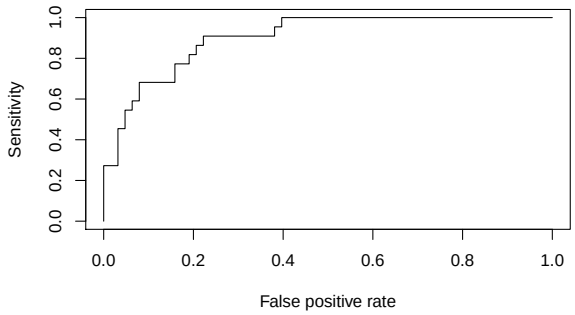


Figure 10: ROC curve PCR model.

Model	AUC	Number.of.beta.parameters
LASSO	0.909	9
Ridge	0.896	54676
PCR	0.902	13

Table 2: Summary of the 3 modelling approaches.

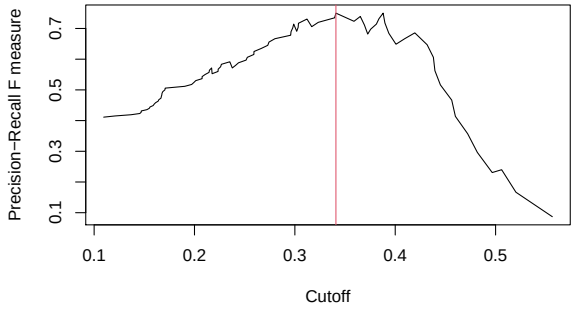


Figure 11: Cutoff value selection. The red line indicates the maximum F1 score and corresponding cutoff value.

The model performance in terms of AUC is similar for the 3 models. Since LASSO regression is the simplest model, this model is preferred. By choosing cutoff c = 0.34, we can achieve a maximal F1-score of 0.75. This is represented in figure 11. The confusion matrix for the LASSO model, based on the test dataset, with cutoff c = 0.34 is shown in table 3.

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

Table 3: Confusion matrix final model.

This confusionmatrix clearly shows a sensitivity of 0.77, and a specificity of 0.87. In short, this prediction model 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 result could be *explained* (or at least understood a little better) by looking back at the exploratory analysis. There 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.

4 Conclusions

In the exploratory analysis we found that the 2 groups are different, but not fully separable. Through application of a procedure based on the sparse LDA (adapted for running on our memory-limited hardware) 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%.
The 3 modeling approaches used in this study resulted in models that performed very similar in terms of AUC. The LASSO regression model was selected based on its spareseness. The final model performed well in terms of estimated sensitivity (0.77) and specificity (0.87). Interesting to note is that, from the selected genes by the LASSO model, the genes 202270_at, 221658_s_at, 240413_at were also detected by the sparse LDA procedure.

5 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.

6 Appendix

6.1 Exploration methods for high dimensional data

6.1.1 Sparse principle components analysis

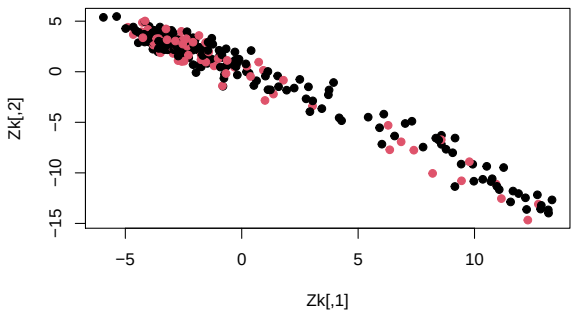


Figure 12: sparse PCA

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:

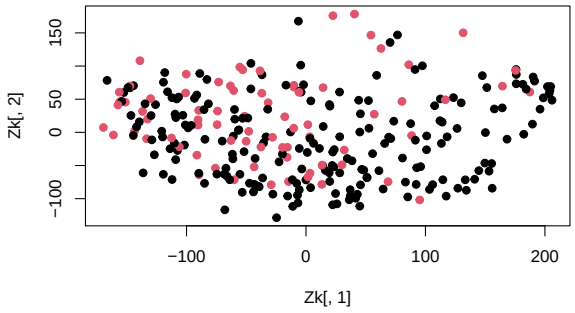


Figure 13: Biplot (not showing the vectors) MDS, dimensions 1 and 2.

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

From the scree plot in figure 15 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

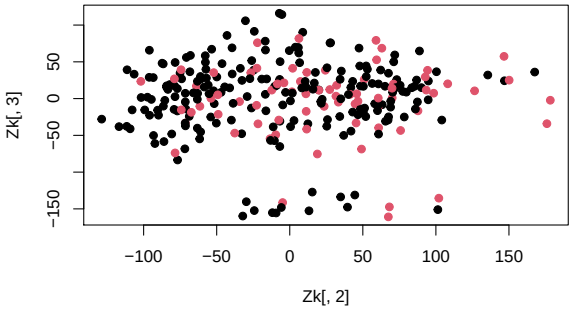


Figure 14: Biplot (not showing the vectors) MDS, dimensions 2 and 3.

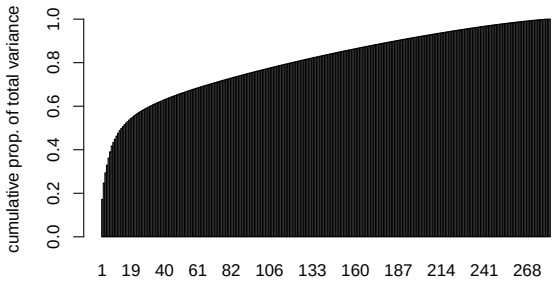


Figure 15: Scree plot MDS.

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.

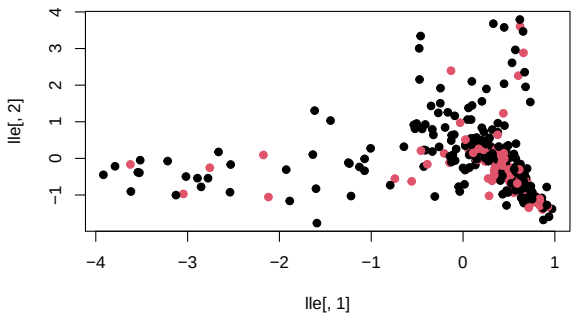


Figure 16: LLE, dimensions 1 and 2.

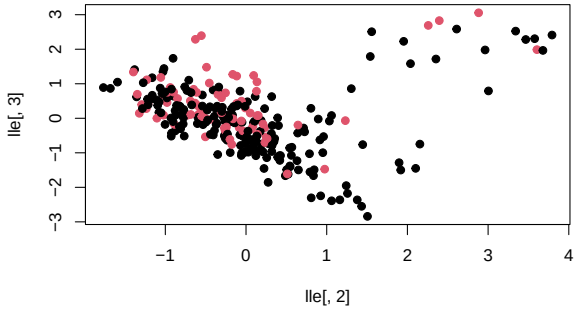


Figure 17: LLE, dimensions 2 and 3.

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 18 can be seen that with ISOMAP, it is also not possible to make a distinction between the group of accepted and rejected kidneys.

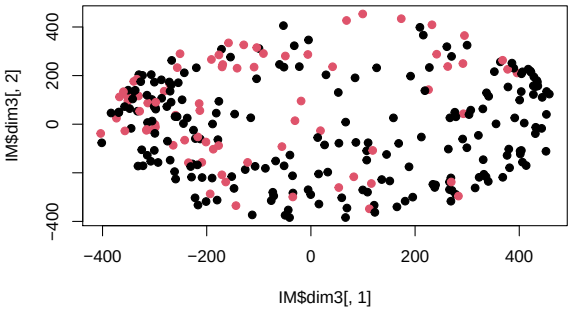


Figure 18: ISOMAP, dimensions 1 and 2.

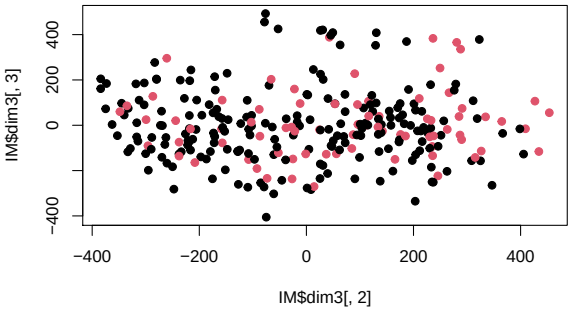


Figure 19: ISOMAP, dimensions 2 and 3.

6.1.5 Sammon mapping

Sammon mapping presented by Sammon (1969). The results are in figures 20 and @ref(fig:sammon2: no distinction can be made between the two groups.

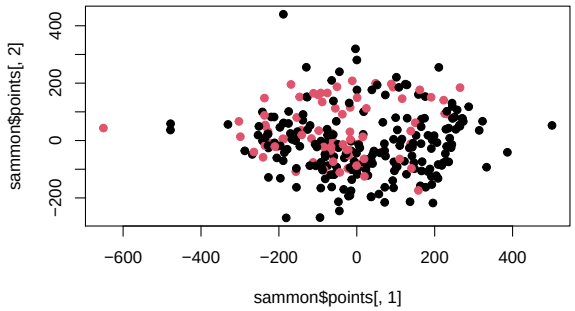


Figure 20: Sammon mapping dimensions 1 and 2.

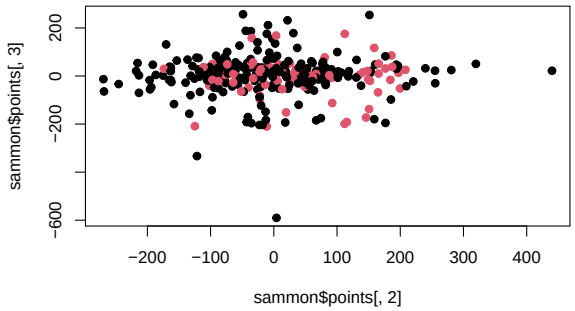


Figure 21: Sammon mapping dimensions 2 and 3.

6.1.6 Diffusion maps

Diffusion mapping was presented by Nadler et al. (2006) and LAfon and Lee (2006). From figure 22, 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 23 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

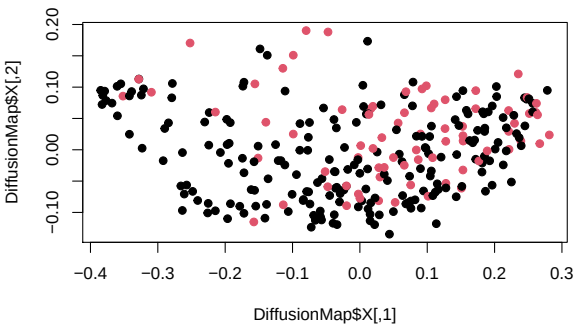


Figure 22: Diffusion map

one group that is less heterogeneous, though far from homogeneous.

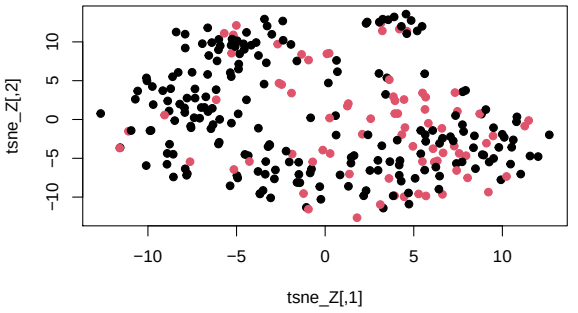


Figure 23: Two dimensional representation of the data through application of t-SNE

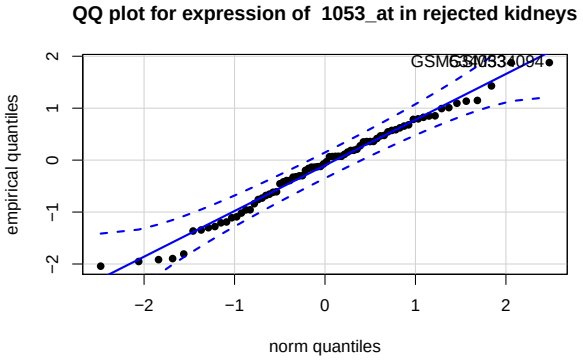
6.1.8 Simulation sparse LDA

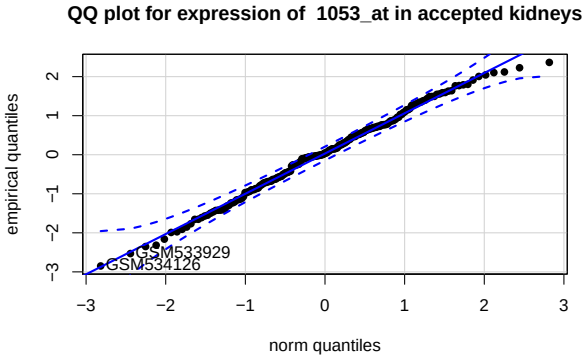
In order to have an idea whether sparse LDA split in 3 parts can be used as an exploratory tool we ran a small simulation. We simulated high dimensional data in such a way that also the split parts were high dimensional (n = 33, p = 102). We kept the number of observations and variables as low as possible 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 as is the case for the real data (allowing for variation over simulation repe-

titions).

The simulation was repeated 50 times and we looked at following measures: correlation between coefficients (mean correlation 0.83, .25$^{\text{th}}$ and .75$^{\text{th}}$ quantiles (0.74, 0.97, higher is better), 6 correlations were not computable because one or both method(s) gave no coefficients), proportion of simulations in which at least 1 variable detected by the full method was also detected by the split method (0.88, higher is better), proportion of variables detected by the full method that also were detected by the split method (mean proportion 0.94, .25$^{\text{th}}$ and .75$^{\text{th}}$ quantiles (1, 1, higher is better), and proportion of variables detected by the split method that weren't detected by the full method (mean proportion 0.5, .25$^{\text{th}}$ and .75$^{\text{th}}$ quantiles (0, 0.75, lower is better).

Although this not a formal way of making a comparison, with these results we feel confident enough to use this adapted approach, albeit only as an exploratory tool. Please note that we don't claim that the adapted method *is* valid.





6.2 QQ plots