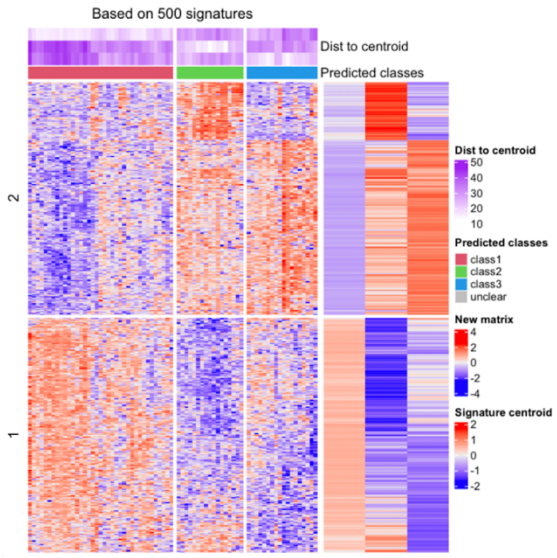
Supplementary file 9. Compare to machine learning methods for predicting class labels of deselected samples

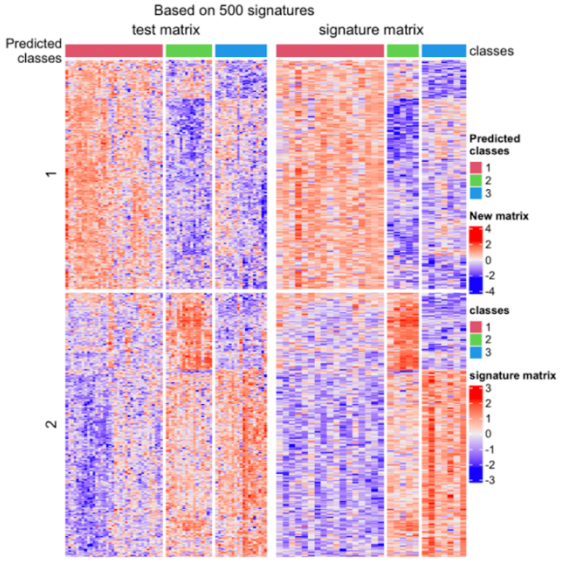
Zuguang Gu <z.gu@dkfz.de>

We proposed a centroid-based method for predicting class labels of deselected samples. *cola* also supports machine learning methods for the prediction. In this document, we compare centroid-based method and SVM/random forest.

When the sample size is huge, we randomly picked a subset of samples as the training set for *cola* classification. The class label prediction for the deselected samples is based on the signature matrix where features in the signature matrix show significant differences between subgroups in the training set. Thus for the machine learning methods, as for the training set, it is very easy to find hyperplanes that fully separate the classes. On the other hand, the classification is based on a randomly sampled subset of the original data, if the features can well separate subgroups in the training set, it is very likely that the features would have very similar patterns in the deselected samples and they can separate subgroups in the deselected samples as well. Therefore, using the centroid-based method or SVM/random forest would give very similar classifications. A similar centroid-based method is used in the *SingleR* package to assign cell types to cells based on the cell marker expression (which can be thought of as signature genes specifically highly expressed in certain cell types).

In the next few heatmaps, we compared performance of the centroid-based method and SVM on the Golub gene expression dataset (https://bioconductor.org/packages/golubEsets/). The dataset contains 72 samples, and we randomly picked 36 samples as a training set for *cola* analysis under three-group classification. Based on this, we predicted the class labels for all 72 samples with top 500 signature genes that are most differentially expressed between the three *cola* groups. The left figure is based on the centroid-based method where the left heatmap shows the expression of signature genes in all 72 samples and the right three-column heatmap shows the centroid of signature gene expression in the three groups in the training set. The right figure is based on SVM where the right heatmap is the training matrix. Basically we can see the two methods give highly similar class predictions.





In the next figures, we counted the agreement of the classifications from centroid-based method, SVM and random forest. They have very high agreement for class prediction.

	svm		
centroid	1	2	3
1	36	0	1
2	0	17	0
3	0	0	18

	randomForest		
centroid	1	2	3
1	37	0	0
2	2	13	2
3	0	1	17