Feature selection by replicate reproducibility and non-redundancy

Tümay Capraz [1,2] and Wolfgang Huber [1]

[1]Genome Biology Unit, EMBL, Heidelberg, 69117, Germany.
[2]Faculty of Biosciences, University of Heidelberg, Heidelberg, Germany.

Abstract

Motivation: A fundamental step in many analyses of high-dimensional data is dimension reduction. Two basic approaches are introduction of new, synthetic coordinates, and selection of extant features. Advantages of the latter include interpretability, simplicity, transferability and modularity. A common criterion for unsupervised feature selection is variance or dynamic range. However, in practice it can occur that high-variance features are noisy, that important features have low variance, or that variances are simply not comparable across features because they are measured in unrelated numeric scales or physical units. Moreover, users may want to include measures of signal-to-noise ratio and non-redundancy into feature selection.
Results: Here, we introduce the RNR algorithm, which selects features based on (i) the reproducibility of their signal across replicates and (ii) their non-redundancy, measured by linear independence. It takes as input a typically large set of features measured on a collection of objects with two or more replicates per object. It returns an ordered list of features $i_1, i_2, \ldots, i_k$, where feature i_1 is the one with the highest reproducibility across replicates, i_2 that with the highest reproducibility across replicates after projecting out the dimension spanned by i_1, and so on. Applications to microscopy based imaging of cells and proteomics experiments highlight benefits of the approach.
Availability: The RNR method is implemented in the R package *FeatSeekR* and is available via Bioconductor (Huber *et al.*, 2015) under the GPL-3 open source license.
Contact: tuemay.capraz@embl.de

1 Introduction

Many biological datasets can be represented as a numeric matrix whose rows correspond to measured features and columns to objects of interest (e.g., cells, biological specimen). Here we consider settings where for each object we have two or more replicate measurements. Examples include RNA-Seq transcriptomics, mass spectrometry proteomics and microscopy-based cell morphology, where the features are levels of transcripts or proteins, or morphological descriptors of shape and texture of cells or cell compartments. The number of features can be in the thousands, but typically not all of them are informative (some are dominated by noise), and some are redundant of each other (they measure essentially the same underlying, relevant variable, in different ways). In this case, it can be desirable to reduce the dimensionality of the data. This simplifies subsequent computations, including quality control, and increases interpretability. Additionally, it facilitates extensions of the dataset by additional experiments, integration of multiple datasets and standardization.

Dimensionality reduction can be considered in supervised and unsupervised settings; here we focus on the latter. There are two basic, not necessarily mutually exclusive, approaches: one is to introduce new variables that are linear or non-linear functions of several or all of the original variables, and map into a lower dimensional space; the other is feature selection. Examples for the first approach include singular value decomposition (SVD) and principal component analyis (PCA) (Jolliffe, 1986), factor analysis (Singh *et al.*, 2018; Argelaguet *et al.*, 2018), t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) and uniform manifold approximation and projection (UMAP) (McInnes *et al.*, 2018). Such combination of the original features into new features (by weighted average or some other mathematical function) allows cancelling out of random noise in individual features, and sometimes the new features are meaningful "latent" variables. Here we focus on the second approach, which can facilitate interpretation and across-dataset comparisons, and is attractively simple.

1.1 Related work

Unsupervised feature selection can be broadly categorized into embedded and filter methods. Embedded methods incorporate feature selection into the model fitting process and can be both supervised and unsupervised. An example for an unsupervised embedded method is sparse clustering where a penalization term is added to the clustering objective function (Witten and Tibshirani, 2010). Feature selection based on filtering uses properties of the data to prioritize features. Typically, features are ranked according to a statistic and the user defined top n features are selected. Different statistics of the data were previously used. These include mutual information and variance (Ferreira and Figueiredo, 2012; Guyon and Elisseeff, 2003), entropy (Varshavsky *et al.*, 2006) or methods that minimize the reconstruction error of a projection of the original features induced by a selected subset of features (Wang *et al.*, 2015). Here we introduce FeatSeekR, an unsupervised feature selection filter method, which uses replicate reproducibility as selection criterion. We were motivated for this work by Fischer *et al.* (2015), who devised a special case of our current method to use it on microscopy data, but only cursorily mentioned it in the supplement of their paper, without self-contained description, validation or software.

2 Approach

We posit that features carrying scientifically important information should be more correlated between replicates. The algorithm iteratively selects features with the highest reproducibility across replicates, after projecting out those dimensions from the data that are linearly spanned by the previously selected features. Thus, each newly selected feature has a high degree of uniqueness.

3 Methods

The method pursues two aims. First, it selects features with high correlation between replicates, second, it aims to select features that are non-redundant between each other. We propose the following iterative, greedy procedure.

3.1 The FeatSeekR algorithm

Let $\mathbf{X} \in \mathbb{R}^{pn}$ be a $p \times n$ data matrix with n observations each for p real-valued features. The columns of $\mathbf{X}$ represent repeated measurements on $k < n$ biological conditions. The replication structure is encoded by the n-vector $\mathbf{r} \in \{1, \ldots, k\}^n$, such that $\{j \,|\, r_j = c\}$ are the indices of those columns in $\mathbf{X}$ that contain measurements for the c-th condition. For instance, if each condition was measured twice and replicates are next to each other in $\mathbf{X}$, then $\mathbf{r} = (1, 1, 2, 2, 3, 3, \ldots)$. We assume that most conditions have two or more replicates, but conditions with only one replicate are permitted. $\mathbf{X}$ may contain a small fraction of observations missing at random.

We label the iterations of the algorithm by the index $t = 0, 1, 2, 3, \ldots$ and denote by S_t the set of features selected up until iteration t. Thus, the elements of S_t are integers from 1 to p. Its complement $\bar{S}_t = \{1, \ldots, p\} \setminus S_t$ is the set of features not selected up until iteration t. The algorithm is greedy, so $S_t \subset S_{t+1}$. The initial selection S_0 is either the empty set $\emptyset$, or a set of features already pre-selected by the user based on criteria of their choice.

In iteration step t, we fit a linear model for each not previously selected feature $i \in \bar{S}_t$ as a function of the selected features

$$\mathbf{X}_{i\cdot} = \mathbf{X}_{S_t\cdot}\, \boldsymbol{\beta}_i + \boldsymbol{y}_i, \tag{1}$$

where $\mathbf{X}_{i\cdot}$ is the i-th row of $\mathbf{X}$, containing the observations of feature i, and $\mathbf{X}_{S_t\cdot}$ is the $|S_t| \times n$ matrix obtained by subsetting from $\mathbf{X}$ the rows corresponding to S_t. $\mathbf{X}_{S_t\cdot}$ contains the already selected features. $\boldsymbol{\beta}_i \in \mathbb{R}^{|S_t|}$ is a vector of regression coefficients for these features, and $\boldsymbol{y}_i \in \mathbb{R}^n$ the vector of residuals. We fit all the free parameters on the right hand side of Eqn. (1) by linear regression, i.e., by minimizing the L_2-norm of $\boldsymbol{y}_i$.

We then use the residuals from these fits, $\hat{\boldsymbol{y}}_i$ to represent the current (i.e., at step t) *non-redundant* information contributed by feature i. To measure replicate reproducibility of this non-redundant information, we use the F-statistic

$$F_i = \frac{W_{\text{between},i}}{W_{\text{within},i}}. \tag{2}$$

To compute F_i, first define the overall mean $\bar{y}_i$ and the mean $\bar{y}_{i,c}$ across replicates within condition c of $\hat{\boldsymbol{y}}_i$:

$$\bar{y}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{y}_{ij}, \tag{3}$$

$$\bar{y}_{i,c} = \frac{1}{n_c} \sum_{\{j \,|\, r_j = c\}} \hat{y}_{ij}, \tag{4}$$

where $n_c = |\{j \,|\, r_j = c\}|$ is the number of replicates for condition c, and we have dropped the dependence of these quantities on t to unburden the notation. Numerator and denominator of the F-statistic (2) are then:

$$W_{\text{between},i} = \frac{1}{k-1} \sum_{c=1}^{k} n_c \left(\bar{y}_{i,c} - \bar{y}_i\right)^2, \tag{5}$$

$$W_{\text{within},i} = \frac{1}{n-k} \sum_{c=1}^{k} \sum_{\{j \,|\, r_j = c\}} \left(\hat{y}_{ij} - \bar{y}_{i,c}\right)^2. \tag{6}$$

At the end of iteration step t, we select the feature i^* with highest F_i and proceed to the next iteration with $S_{t+1} =$

$S_t \cup \{i^*\}$ until the user defined maximum number of selected features.

This procedure provides us with a list of features ranked by reproducibility and non-redundancy. A pseudocode representation is given in Algorithm 1.

Algorithm 1 FeatSeekR algorithm

Input: $\mathbf{X}$, max_features, S_0 (default: $S_0 = \emptyset$)
Output: $S_t \subset \{1, \ldots, p\}$
  Initialize:
  $t \leftarrow 0$
  while $t \leq$ max_features **do**
    $\bar{S}_t \leftarrow \bar{S}_t \setminus S_t$
    # Multivariate linear regression of each feature not
    # in the selection S_t on those in S_t
    $\mathbf{P} \leftarrow \mathbf{X}_{S_t}(\mathbf{X}_{S_t}^T \mathbf{X}_{S_t})^{-1} \mathbf{X}_{S_t}^T$
    for $i \in \bar{S}$ **do**
      $\hat{\mathbf{X}}_{i\cdot} \leftarrow \mathbf{P}\mathbf{X}_{i\cdot}$
      $\hat{\boldsymbol{y}}_i \leftarrow \mathbf{X}_{i\cdot} - \hat{\mathbf{X}}_{i\cdot}$
      $\mathbf{X}_{i\cdot} \leftarrow \hat{\boldsymbol{y}}_i$
    end for
    # Add feature with highest F-statistic
    # between residuals to S_t and remove from $\bar{S}_t$.
    $i^* \leftarrow \arg\max_i F_i$ with $i \in \bar{S}_t$
    $S_{t+1} \leftarrow S_t \cup \{i^*\}$
    $t \leftarrow t + 1$
  end while

3.2 Evaluation selected of feature subsets by fraction of explained variance

Optionally, to inform data-adaptive stopping in lieu of a predetermined value for the number of selected features, we can consider the fraction of variance of the dataset that is explained by the currently selected feature subset. We first model each feature $\mathbf{X}_{i\cdot}$ of the original dataset $\mathbf{X}$ as a function of the selected features $\mathbf{X}_{S_t}$ analogous to Equation (1). We then get the fraction of explained variance R_i^2 of each feature $\mathbf{X}_{i\cdot}$ by calculating:

$$R_i^2 = 1 - \frac{\sum_{j=1}^{n} \hat{y}_{ij}}{\sum_{j=1}^{n} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})} \tag{7}$$

where $\bar{\mathbf{X}}_{i\cdot}$ is the mean of feature i. We finally get the fraction of variance explained of the whole dataset, by averaging R^2 over all features.

4 Results

4.1 Simulations

To demonstrate the algorithm we generated two synthetic datasets. The first dataset is characterized by a small number of underlying, "latent" variables that are noisily measured each by several observed features. In the second case we simulated data for a two class clustering problem and compared our method to the widely used variance based feature selection.

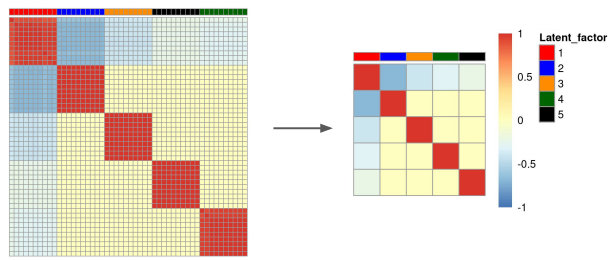

Fig. 1. Left: correlation matrix of simulated data before feature selection. Right: correlation matrix of first 5 selected features.

4.1.1 Selecting a set of non-redundant features

To model redundant groups of features, we followed a step-by-step process. First, we generated an $l \times n$ matrix, denoted as $\mathbf{M}$, where each element M_{ij} was drawn independently from a standard normal distribution, i.e., $M_{ij} \sim \mathcal{N}(0, 1)$, with $l = 5$ representing the number of groups and $n = 500$ representing the number of samples.

Next, we applied the Gram-Schmidt process to orthonormalize the rows of $\mathbf{M}$, resulting in an orthonormal matrix $\mathbf{Q}$. For each group, we generated 10 features denoted as $X_{j\cdot}$, where j ranges from 1 to 10. Each feature was obtained by scaling the corresponding row $\mathbf{Q}_{i\cdot}$, where i ranges from 1 to l, by a random number drawn independently from a standard normal distribution, i.e., $X_{j\cdot} = \alpha_{ij}\mathbf{Q}_{i\cdot}$, with $\alpha_{ij} \sim \mathcal{N}(0, 1)$. This process yielded a 50×500 matrix denoted as $\mathbf{Q}'$, representing the 10 features for each of the 5 groups.

Finally, we created three replicates of $\mathbf{Q}'$ by incorporating random noise. The noise was sampled independently from a standard normal distribution, and added element-wise to $\mathbf{Q}'$. The three replicates were then concatenated, resulting in the final data matrix $\mathbf{X} \in \mathbb{R}^{50 \times (3n)}$. Figure 1 shows the correlation matrices of these simulated data, and of five features selected by FeatSeekR. This result indicates that the algorithm is able to identify non-redundant features in this synthetic setting.

4.1.2 Simulation of two class data

We constructed a $p \times n$ data matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$, where $n = 500$ samples are divided into two classes and $p = 50$ features exhibit distinct signal-to-noise ratios. The mean values were assigned as follows: $\mu_1, \ldots, \mu_{n/2} = 1$ for samples in class 1, and $\mu_{n/2+1}, \ldots, \mu_n = 2$ for samples

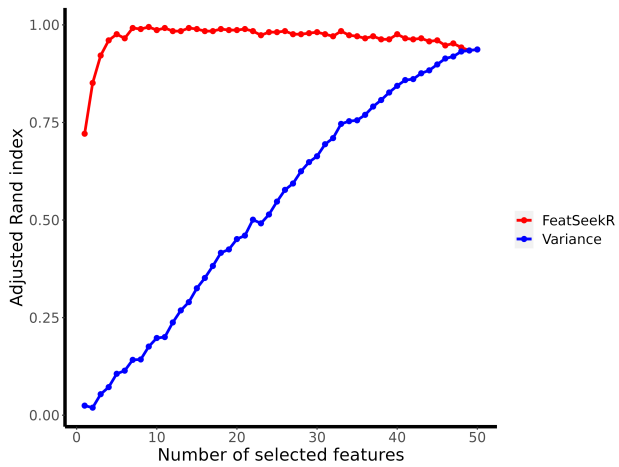

Fig. 2. Performance of feature subsets selected by FeatSeekR and based on variance.

in class 2. For each feature i, we generated a vector $\mathbf{X}_{i\cdot} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_i)$, where σ_i increased linearly from 0.1 to 2. We produced two replicates of $\mathbf{X}_{i\cdot}$ by incorporating independent random noise $\boldsymbol{y}_i \sim \mathcal{N}(0, \sigma_i)$. The final data matrix $\mathbf{X}_{\text{final}} \in \mathbb{R}^{p\times(2n)}$ was obtained by concatenating the two replicates, resulting in $2n = 1000$ samples, with each sample comprising repeated measurements from the corresponding replicates.

We ranked the features using two methods: FeatSeekR and based on their variance. We evaluated the feature selection via the performance of subsequent k-means clustering with $k = 2$ in recovering the two classes. For this, we calculated the adjusted Rand index (Morey and Agresti, 1984) between the clustering result and the known class labels.

Figure 2 shows the adjusted Rand index as a function of number of selected features. As might be expected, in each case the performance improves with increasing number of features, as that increasingly allows the noise to cancel itself out. However, selection by FeatSeekR achieves the same with a smaller number of selected features than the variance based selection. This result shows that feature selection based on total variance is not always an optimal criterion, as it conflates signal and noise, whereas FeatSeekR can disentangle these (see also Figure S1).

4.2 Applications to biological datasets

4.2.1 Microscopy based image data from combinatorial knockout screens

Generic feature sets for microscopy based cytometry sets try to cover a wide range of information, ranging from general features such as intensity quantiles, object shapes (McQuin *et al.*, 2018; Pau *et al.*, 2010) or more abstract textural features introduced by (Haralick *et al.*, 1973). As the produced features are designed to cover as much general information as possible, redundancy can be high. Additionally, not all features capture relevant biological information in every type of experiment. Consequently extracted features are not always reproducible between repeated measurements and sometimes carry mainly noise. Here we used FeatSeekR to rapidly identify unique features with reproducible signal between measurements in two biological image datasets.

We used data from (Laufer *et al.*, 2013), who performed combinatorial gene knock-outs using siRNA, followed by imaging of the cells and image segmentation and feature extraction using the R package EBImage (Pau *et al.*, 2010). A summary of datasets we used is shown in Table 1.

Analogous to our idealized example in Figure 1, the extracted features formed groups of high correlation within and lower correlation between (Figure S2). The grouping was partially interpretable as some groups broadly corresponded to different colour channels (fluorescent labels) or cellular compartments. This supports the idea that the effective dimension of the data matrix is substantially lower than the number of features $p = 202$ and that feature selection is a plausible approach to these data. We used FeatSeekR to select a set of features that explained more than 70% of the variance of the original dataset. The selection comprised of 5 features, a substantial reduction (Figure 3A). The overall low correlation between the selected features confirmed their low redundancy. We note that selected feature sets are dependent on the starting set of features and hypothesized that *Cell actin majoraxis* and *Cell actin eccentricity* contain redundant information. Indeed, when calling FeatSeekR with preselection of the *Cell actin eccentricity* feature, *Cell actin majoraxis* was not selected. This illustrates that multiple selections are equally admissible and can be influenced by a user defined preselected set of features.

4.2.2 Mass spectrometry-based proteomics data

Next we applied FeatSeekR to spectral features of a proteomics dataset from (Collins *et al.*, 2017), where the authors investigated the reproducibility of a mass spectrometry based proteomics measurement across multiple international sites. In this type of experiment, proteins are usually first digested to peptides, separated via LC-MS and their mass spectra are subsequently recorded. To identify individual peptides, mass spectra are either matched to a database or to a library of spectra of known peptides (Aebersold and Mann, 2016). Beforehand, features such as retention time, intensities and mass accuracies are extracted. The matching to the reference is then done based on these extracted features (Röst *et al.*, 2014). We used measurements of 4 sites as replicates, leading to 24835

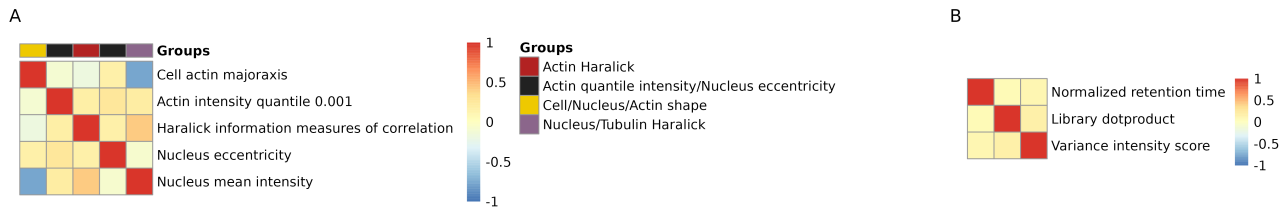

Fig. 3. Correlation matrix of selected features of (Laufer et al., 2013) that explain at least 70% of the variance of the original dataset. Features are colored according to their feature clusters.

peptide assays (samples), 37 features and 4 replicates (see Table 1).

We observed that not all of these automatically extracted spectral features are equally reproducible and informative across sites. Furthermore, the dataset consists of several correlated redundant feature clusters. For example peak features related to retention time, distance to the reference library or p-value related features form very distinct clusters (Figure S3). We used FeatSeekR to select the most reproducible features that explain at least 70% of the total variance. Figure 3B shows that we identified the most reproducible features of the redundant and correlated feature clusters. The selected features cover both peptide retention time, as well as information related to their mass spectra.

Table 1. Summary of the used biological datasets.

Dataset	Samples	Features	Replicates
Laufer et al. 2013	5820	202	2
Collins et al. 2017	24835	37	4

5 Discussion

We present a framework for feature selection which selects features based on their reproducibility between replicates while keeping redundancy low. In contrast to existing filtering based feature selection methods we make use of replicated measurements and are able to effectively separate biological signal from noise. Additionally, FeatSeekR is capable of performing feature selection on ragged data, where not all samples have the same number of replicates. We show on synthetic data that FeatSeekR is able to find exactly one feature per underlying latent factor. We highlighted its utility as a preprocessing step for clustering, by selecting more informative features and removing more noisy ones. Furthermore, we show the application of our method to biological data, derived from microscopy based imaging of cells and proteomics experiments. Our algorithm finds feature sets of biological datasets that achieve a good trade-off between captured information and redundancy.

In practice, feature selection can serve different purposes, such as reduction of storage space, computation time or increase in downstream performance (e.g. classification accuracy). If FeatSeekR is used to improve performance in a machine learning context, feature selection should be incorporated in the cross validation procedure (Ambroise and McLachlan, 2002).

To guide the selection process, we provide diagnostic tools to analyze and visualize information content in biological datasets, within the FeatSeekR package.

The objective that motivates feature selection with FeatSeekR does not lead to a single optimal selection. Instead, it is compatible with multiple selections that are, for all practical purposes, equally admissible. Thus, even if the implementation by FeatSeekR returns a single selection, this should be viewed as a reasonable proposal, not as a unique solution. FeatSeekR uses a greedy algorithm and is not based on a global optimality criterion. Formulating such a global optimality criterion and associated algorithms remains a direction for future research.

Acknowledgements

This work was funded by the European Research Council (ERC Synergy Project DECODE under grant agreement no. 810296).

References

Aebersold, R. and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, **537**(7620), 347–355.

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, **99**(10), 6562–6566.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, **14**(6), e8124.

Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., Chan, D. W., Gibson, B. W., Gingras, A.-C., Held, J. M., *et al.* (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of swath-mass spectrometry. *Nature communications*, **8**(1), 1–12.

Ferreira, A. J. and Figueiredo, M. A. (2012). An unsupervised approach to feature discretization and selection. *Pattern Recognition*, **45**(9), 3048–3060.

Fischer, B., Sandmann, T., Horn, T., Billmann, M., Chaudhary, V., Huber, W., and Boutros, M. (2015). A map of directional genetic interactions in a metazoan cell. *Elife*, **4**, e05464.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, **3**(Mar), 1157–1182.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610–621.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., *et al.* (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, **12**(2), 115–121.

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer.

Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013). Mapping genetic interactions in human cancer cells with rnai and multiparametric phenotyping. *Nature methods*, **10**(5), 427–431.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., Doan, M., Ding, L., Rafelski, S. M., Thirstrup, D., *et al.* (2018). Cellprofiler 3.0: Next-generation image processing for biology. *PLoS biology*, **16**(7), e2005970.

Morey, L. C. and Agresti, A. (1984). The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, **44**(1), 33–37.

Pau, G., Fuchs, F., Sklyar, O., Boutros, M., and Huber, W. (2010). Ebimage—an r package for image processing with applications to cellular phenotypes. *Bioinformatics*, **26**(7), 979–981.

Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L., *et al.* (2014). Openswath enables automated, targeted analysis of data-independent acquisition ms data. *Nature biotechnology*, **32**(3), 219–223.

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., and Lê Cao, K.-A. (2018). Diablo: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv*, page 067611.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(11).

Varshavsky, R., Gottlieb, A., Linial, M., and Horn, D. (2006). Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**(14), e507–e513.

Wang, S., Pedrycz, W., Zhu, Q., and Zhu, W. (2015). Unsupervised feature selection via maximum projection and minimum redundancy. *Knowledge-Based Systems*, **75**, 19–29.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, **105**(490), 713–726.

6 Supplementary Figures

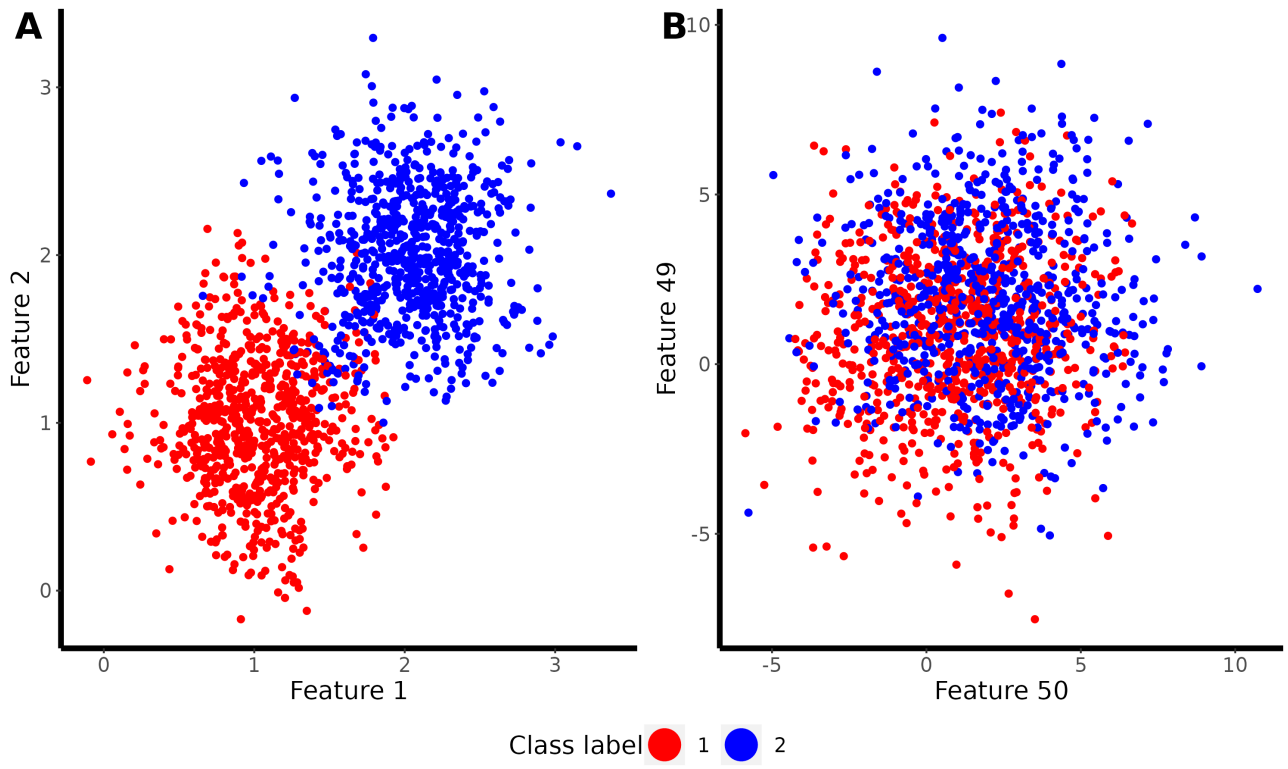

Fig. S1. Scatter plots of simulated 2 class data. Features in (A) have lower variance than in (B) and show better class separation.

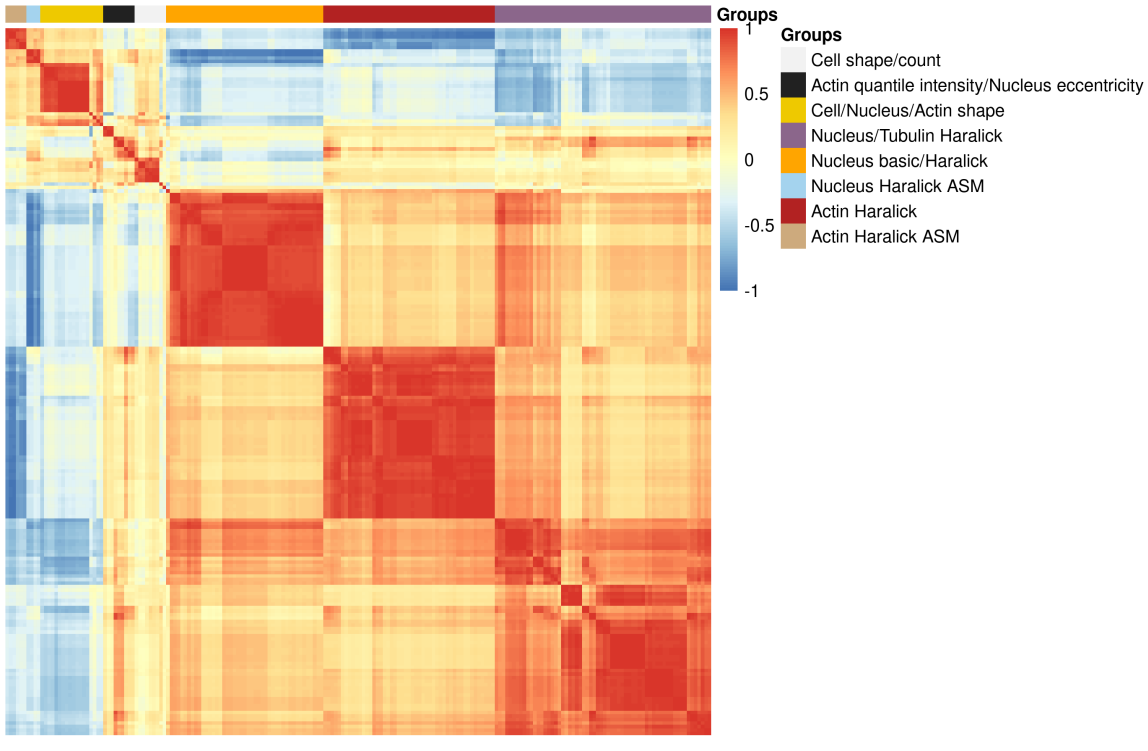

Fig. S2. Correlation matrix of the features from (Laufer et al., 2013)

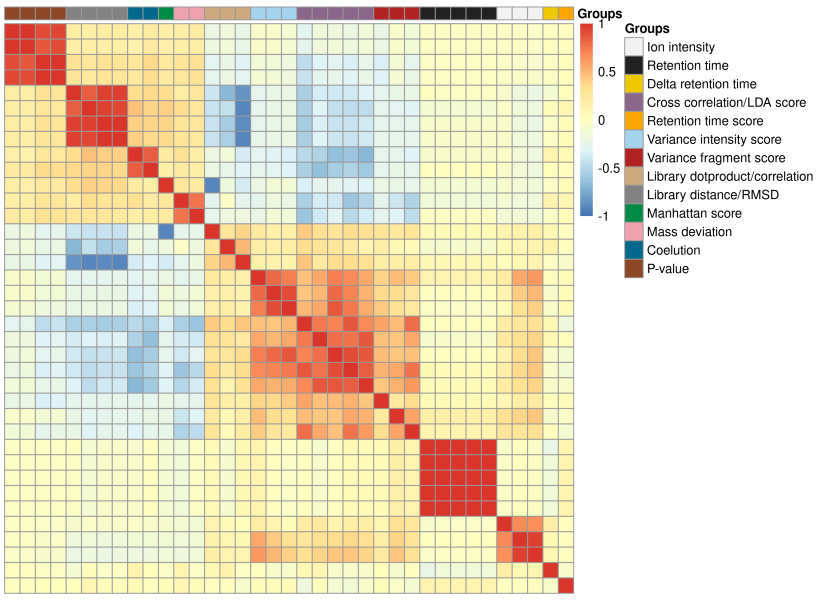

Fig. S3. Correlation matrix of the features from (Collins et al., 2017)