Bioinformatics, 2024, **40(9)**, btae548
https://doi.org/10.1093/bioinformatics/btae548
Advance Access Publication Date: 10 September 2024
Original Paper

OXFORD

Data and text mining

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, 69117, Germany
*Corresponding author. Genome Biology Unit, EMBL, Heidelberg, 69117, Germany. E-mail: tuemay.capraz@embl.de (T.C.)

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 unsupervized 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 dependence. 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 highlight benefits of the approach.

Availability and implementation: The RNR method is available via Bioconductor (Huber W, Carey VJ, Gentleman R et al. (Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods 2015;12:115–21.) in the R package *FeatSeekR*. Its source code is also available at https://github.com/tcapraz/FeatSeekR under the GPL-3 open source license.

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 specimens). 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.

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 a smaller number of new variables that are linear or non-linear functions of the original variables; the other is feature selection. Examples for the first approach employ singular value decomposition, principal component analysis (Jolliffe 1986), and numerous versions of (non-linear) multi-dimensional scaling. As the new variables are smooth functions of the original features, random noise can cancel out. Sometimes they are meaningful "latent" variables. Here, however, we focus on feature selection, which can facilitate interpretation and integration of multiple datasets, 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 summary statistic; the user chooses a number n and selects the top n features. Different summary statistics are commonly used. These include mutual information and variance (Guyon and Elisseeff 2003, Ferreira and Figueiredo 2012), entropy (Varshavsky *et al.* 2006), or methods that minimize reconstruction error (Wang *et al.* 2015). Here, we introduce FeatSeekR, an unsupervised filter method that 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.

Received: 8 August 2023; **Revised:** 5 August 2024; **Editorial Decision:** 4 September 2024; **Accepted:** 6 September 2024

2 Approach

We posit that features carrying scientifically important information should be 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, and second, it aims to select features that are non-redundant between each other. We propose the following iterative, greedy forward procedure.

3.1 The FeatSeekR algorithm

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ 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 and/or objects. 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 forward, 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 the vector of coefficients for the regression of feature i on $\mathbf{X}_{S_t \cdot}$, and $\boldsymbol{y}_i \in \mathbb{R}^n$ the vector of residuals. We fit the free parameters on the right hand side of Equation (1) by linear regression, i.e. by minimizing the L_2-norm of $\boldsymbol{y}_i$.

We then use the residuals $\hat{\boldsymbol{y}}_i$ 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 $\hat{\boldsymbol{y}}_i$:

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

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

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

where $n_c = |\{j | r_j = c\}|$ is the number of replicates for condition c, and we have hidden the dependence of these quantities on t in Equations (1–6) to unclutter the notation. Numerator and denominator of the F-statistic (2) are then:

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

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

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.

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:

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

where $\bar{\mathbf{X}}_i$ 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 was characterized by a small number of underlying, "latent" variables that were 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 variance-based feature selection.

4.1.1 Selecting non-redundant features

We generated an $l \times (n/3)$ matrix $\mathbf{M}$ by drawing each element M_{ij} independently from the standard normal distribution; $l = 5$ represents the number of groups and $n/3 = 500$ the number of objects. We applied the Gram–Schmidt process to orthonormalize the rows of $\mathbf{M}$, resulting in an orthonormal matrix $\mathbf{Q}$.

Next, for each group i ($i \in \{1 \dots l\}$), we generated redundant features by scaling $\mathbf{Q}_{i\cdot}$ by each of 10 random numbers $\alpha_{ij} \sim \mathcal{N}(0,1)$ ($j \in \{1 \dots 10\}$) drawn independently from the standard normal distribution, i.e.: $X_{10(i-1)+j,\cdot} = \alpha_{ij}\mathbf{Q}_{i\cdot}$. This process yielded a 50×500 matrix we denote as $\mathbf{Q}'$.

Finally, we created three replicates of $\mathbf{Q}'$ by adding random numbers from the standard normal distribution element-wise to $\mathbf{Q}'$, three times. The three replicates were concatenated, resulting in a final 50×1500 matrix $\mathbf{X}$.

Figure 1 shows the correlation matrices of $\mathbf{X}$ and of the first 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 Finding informative features in two-class data

We generated a $p \times n$ data matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$, where $n/2 = 500$ observations were divided into two classes and $p = 50$ features exhibit distinct signal-to-noise ratios. The mean values were assigned as follows: $\mu_1, \dots, \mu_{n/2} = 1$ for observations in Class 1 and $\mu_{n/2+1}, \dots, \mu_n = 2$ for observations in Class 2. To add correlation between features, we generated a covariance matrix Σ with a Toeplitz structure, where the first row was a sequence from 0 to 0.08 and the remaining rows were generated by shifting the first row by one element to the right, and we set $\Sigma = \Sigma^T\Sigma$. To increase the signal-to-noise ratio of the features, we linearly increased the diagonal of Σ_{ij} with $i = j$ as a function of i from 0.1 to 4. We then simulated the data matrix by sampling $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. In this simulation setting, the two classes serve as 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 Fig. 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 (Pau *et al.* 2010, McQuin *et al.* 2018), 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 information in every type of experiment. Consequently, some features are dominated by fluctuations that are irrelevant for the assay at hand, and not reproducible between repeated measurements. Here, we used

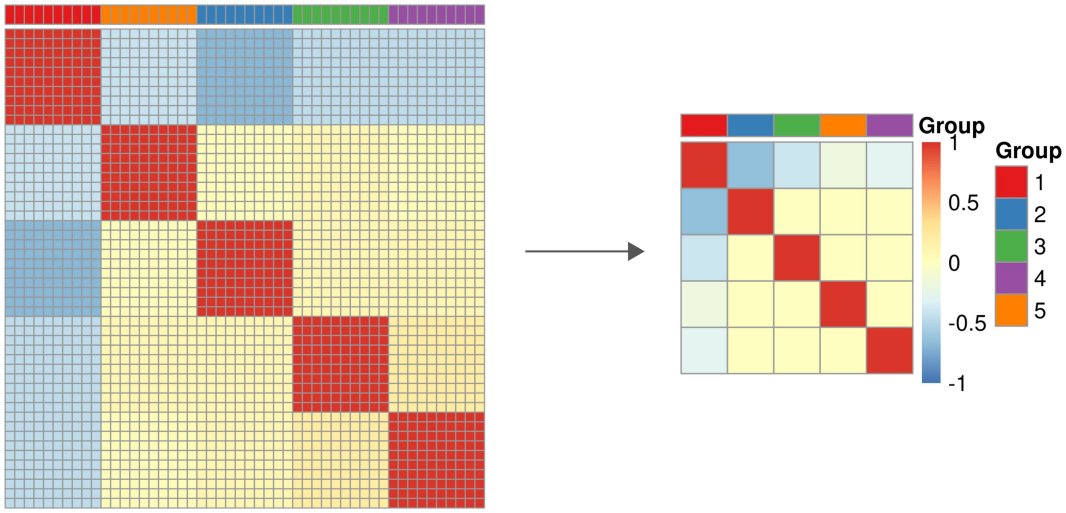

Figure 1. Left: correlation matrix of simulated data. Right: correlation matrix of first five selected features

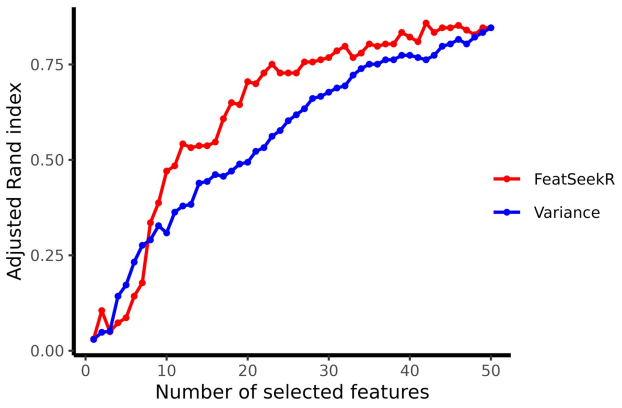

Figure 2. Performance of feature subsets selected by *FeatSeekR* and based on variance

FeatSeekR to 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-downs in human cells using siRNA, followed by imaging, 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 Fig. 1, the extracted features formed groups of high correlation within and lower correlation between (Fig. S2). The grouping was partially interpretable as some groups broadly corresponded to different color 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 five features, a substantial reduction (Fig. 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. For instance, the features 'Cell actin majoraxis' and 'Cell actin eccentricity' are highly correlated, and when calling *FeatSeekR* with preselection of 'Cell actin eccentricity', '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 liquid chromatography-mass spectrometry, 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 four sites as replicates, leading to 99 340 peptide

Table 1. Summary of the used biological datasets.

Dataset	Observations	Features	Replicates
Laufer *et al.* 2013	11 640	202	2
Collins *et al.* 2017	99 340	37	4

assays (observations), 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 (Fig. S3). We used *FeatSeekR* to select the most reproducible features that explained at least 70% of the total variance. Fig. 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.

5 Discussion

We present a framework for feature selection that 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 conditions or observed objects have the same number of replicate observations. 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 and computation time, or better performance of downstream machine learning methods. 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). In such cases, parameters of the feature selection, in particular, the number of selected features, can also be considered (hyper)parameters that can be tuned in the cross-validation.

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 unique optimal selection. Conceptually, it is compatible with multiple selections that are, for practical purposes, equally admissible. Thus, even if the implementation by *FeatSeekR* returns a single selection, this should be viewed as a representative proposal, not as a unique solution. *FeatSeekR* uses a greedy forward algorithm and is not based on a global optimality criterion.

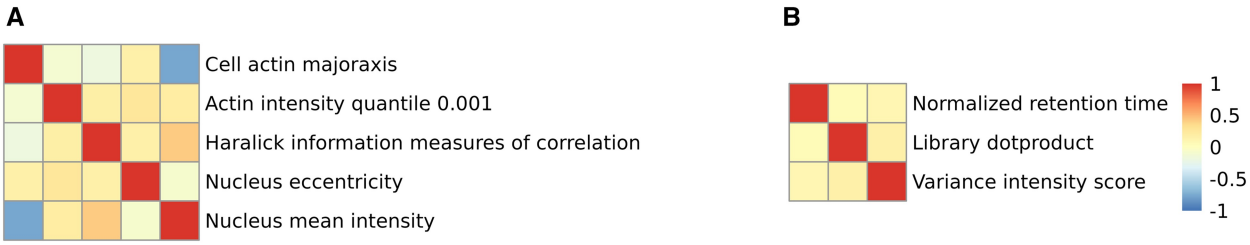

Figure 3. Correlation matrix of selected features of (A) (Laufer *et al.* 2013) and (B) (Collins *et al.* 2017) that explain at least 70% of the variance of the original dataset. Features are colored according to their feature clusters

Formulating such a global optimality criterion and associated algorithms remains a direction for future research.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by funding from the European Research Council (ERC Synergy Project DECODE) [grant agreement no. 810296].

Data availability

The data used in this article can be accessed via the R package *HD2013SGI* on Bioconductor and via the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) with the data set identifier PXD004886.

References

Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature* 2016;**537**:347–55.

Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* 2002;**99**:6562–6.

Collins BC, Hunter CL, Liu Y *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of swath-mass spectrometry. *Nat Commun* 2017;**8**:291–12.

Ferreira AJ, Figueiredo MA. An unsupervised approach to feature discretization and selection. *Pattern Recognition* 2012;**45**:3048–60.

Fischer B, Sandmann T, Horn T *et al.* A map of directional genetic interactions in a metazoan cell. *Elife* 2015;**4**:e05464.

Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;**3**:1157–82.

Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Trans Syst Man Cybern* 1973;SMC-3:610–21.

Huber W, Carey VJ, Gentleman R *et al.* Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods* 2015;**12**:115–21.

Laufer C, Fischer B, Billmann M *et al.* Mapping genetic interactions in human cancer cells with rnai and multiparametric phenotyping. *Nat Methods* 2013;**10**:427–31.

Jolliffe IT. Principal components in regression analysis. In: *Principal Component Analysis*, Springer Series in Statistics. New York, NY: Springer, 1986: 129–155. https://doi.org/10.1007/978-1-4757-1904-8_8

McQuin C, Goodman A, Chernyshev V *et al.* Cellprofiler 3.0: next-generation image processing for biology. *PLoS Biol* 2018;**16**:e2005970.

Morey LC, Agresti A. The measurement of classification agreement: an adjustment to the rand statistic for chance agreement. *Educ Psychol Meas* 1984;**44**:33–7.

Pau G, Fuchs F, Sklyar O *et al.* Ebimage—an r package for image processing with applications to cellular phenotypes. *Bioinformatics* 2010;**26**:979–81.

Röst HL, Rosenberger G, Navarro P *et al.* Openswath enables automated, targeted analysis of data-independent acquisition ms data. *Nat Biotechnol* 2014;**32**:219–23.

Varshavsky R, Gottlieb A, Linial M *et al.* Novel unsupervised feature filtering of biological data. *Bioinformatics* 2006;**22**:e507–13.

Wang S, Pedrycz W, Zhu Q *et al.* Unsupervised feature selection via maximum projection and minimum redundancy. *Knowl Based Syst* 2015;**75**:19–29.

Witten DM, Tibshirani R. A framework for feature selection in clustering. *J Am Stat Assoc* 2010;**105**:713–26.

Bioinformatics, 2024, 40, 1–5
https://doi.org/10.1093/bioinformatics/btae548
Original Paper