FSOL - a workflow for the detection of patient subgroups and affected molecular features in high-throughput omics data

Maike Ahrens[1,2], **Michael Turewicz**[2], **Katrin Marcus**[2], **Helmut E. Meyer**[3], **Caroline May**[2], **Martin Eisenacher***[2], **and Jörg Rahnenführer***[1]

[1]**Department of Statistics, TU Dortmund University**
[2]**Medizinisches Proteom-Center, Ruhr-University Bochum**
[3]**Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Dortmund**
*These authors contributed equally to this work.

ABSTRACT

In personalized medicine, one major goal is the identification of yet unknown patient subgroups with specific gene or protein expression. Different subgroups can indicate different molecular subtypes of a disease. These subtypes might correlate with disease progression, prognosis or therapy response, and the subgroup-specific genes or proteins are potential drug targets. Using high-throughput molecular data, the aim is to characterize the patient subgroup by identifying both the set of samples that shows a distinct expression pattern as well as the set of features that are affected.
We present the new workflow FSOL for the identification of patient subgroups from two sample comparisons (e.g. healthy vs. diseased). First, a pre-filtering based on the univariate score FisherSum (FS) is applied to assess subgroup-specific expression of the features. FS outperforms competing methods in several settings. Second, the selected features are compared regarding the samples that form the affected subgroup. This step uses the OrderedList (OL) method that was originally developed for the comparison of result lists from gene expression studies. We compare our workflow FSOL to a reference workflow based on biclustering using real world and simulated data. On a leukemia data set, a true biological subgroup can be detected with higher stability by FSOL. On simulated data, FSOL shows higher sensitivity and accuracy compared to biclustering especially for small to moderate differences. The exploratory approach FSOL may help in identifying yet unknown mechanisms in pathologic processes and may assist in the generation of new research hypotheses.

Keywords: personalized medicine, subgroup detection, FSOL, biclustering

INTRODUCTION

A common research aim is the identification of novel diagnostic biomarker candidates from a hypothesis-generating high-throughput omics study, where up to ten thousands of features are measured simultaneously. Depending on the technique, features in the data set might represent proteins, DNA copy numbers or mRNA transcripts. Most often, a differential analysis is conducted on a two-group comparison, say healthy versus diseased, to deduce the most promising candidates. For example, Student's t-test or some moderated version of it might be used to compare the gene expression in healthy and diseased subjects. A feature is considered a promising biomarker candidate, if a significant shift between the distributions of both groups is detected, preferable with a large shift and a small overlap of distributions. However, in a large number of applications the prerequisite of homogeneity in both groups is actually not met. For a number of cancer types, e.g. breast cancer (Slamon et al., 1987), lung cancer (Tockman et al., 1997), and prostate cancer (Shah et al., 2004), the observed heterogeneity has been investigated and explained in more detail. It was found that an important factor is the set of the involved oncogenes, i.e. the genes that promote tumour development and growth. The knowledge about those tumour subtypes can directly affect diagnosis and prognosis as well as assist in the development of new personalized therapies.

Besides cancer, there are other diseases that are known to be heterogeneous, e.g. the neurodegenerative disorders Parkinson's disease or Alzheimer's disease. Despite decades of research, previous efforts and experiments did not succeed in the detection of a biomarker that is able to distinguish the group of patients in early stages from healthy controls. New approaches are required to gain insights into yet unknown subgroups in those diseases to hopefully overcome the period of stagnation.

Especially in the context of gene expression microarray studies, several methods were proposed that aim either at the detection of single features with a subgroup indicating expression pattern or at the direct identification of subgroups of samples in a high-throughput data set along with the affected features. Recently, we conducted a comprehensive comparison of different univariate measures for the detection of patient subgroups (Ahrens et al., 2013) comprising real world data and extensive simulation studies. Among others, *outlier sum* (Tibshirani and Hastie, 2007), *profile analysis using kurtosis* (Teschendorff et al., 2006), and *outlier robust t-statistics* (Wu, 2007) were included in the study. The FS method suggested by Ahrens et al. (2013) outperformed existing methods in many settings.

The manuscript of Ahrens et al. (2013) closes with the recommendation to establish a workflow that builds upon the univariate feature selection and additionally includes an appropriate method to combine the selected features and depict the subgroup structure in the data. In the following, we present such an extension that uses an algorithm for the comparison of ordered lists. The application of this specific similarity measure applied to a set of features pre-selected based on FS ranking is the key for the identification of small patient subgroups, even if the difference in expression levels is small. We compare this new approach FSOL to a popular multivariate approach for this purpose, called biclustering. Briefly, it is specifically designed to identify subsets of samples that show a similar expression in a subset of features. Plenty of different biclustering algorithms have been proposed, but here, we focus on a common choice for gene expression analysis (or the like), which is based on the Plaid model by Lazzeroni and Owen (2002). For further reading on biclustering, we recommend the work of Madeira and Oliveira (2004) where different approaches are described and compared. The authors define different types of biclusters and show which methods are suitable for their detection.

The remainder of this work is organized as follows. First, we explain our new workflow FSOL that aims at the detection of patient subgroups and affected molecular features in high-throughput omics data. Second, we outline a reference workflow based on biclustering, that will be compared to FSOL using real and simulated data. After introducing the real data example, we describe the design of the simulation study for subgroup detection, list the different parameter settings and explain the quality criterion we chose to compare both methods. Afterwards, we present and discuss the results of the real data analysis and of the simulation study before we close with a short outlook.

METHODS

FSOL: Novel workflow for subgroup detection

Briefly, the new workflow FSOL for the detection of patient subgroups (SGs) in high-throughput data comprises three steps, namely:

1. pre-selection of features according to univariate FS ranking,
2. grouping of features with respect to the indicated subgroup,
3. nomination of samples for potential subgroup.

Each step will be elucidated individually in the following paragraphs and Figure 1 lists the most important parameters along with the default values that are used in the simulation study.

Step 1. Pre-selection of features according to univariate FS ranking

The idea behind FS is described by means of a comparison of diseased subjects D and control subjects C. For each feature in the data set, all observations are first centered around the respective median of C. Then the cutoff value q, which defaults to the 90 percent quantile of the centered values of D (denoted D'), is determined. FS equals the difference of the summed values in D' and C', respectively, above the cutoff q:

$$FS = w \sum_{d' \in D', d' > q} d' - \sum_{c' \in C', c' > q} c',$$

where d' and c' are the centered expression values in C and D. Owing to the centering, the score captures absolute differences between the highest values and the remaining data irrespective of the location (expression level). The weight w is used for adjustment in case of unbalanced designs. In this work, we only present comparisons with equal sample sizes, and w is set to 1. w also works as a penalty parameter on so-called non-disease-specific (nds) subgroups. Those nds subgroups show a distinct up-regulation not only in the diseased (or experimental) group, but in the control group. As they do not assist in the specific characterization of the sample subgroups in one of the groups (usually the diseased group), the features that indicate nds subgroups are usually not of interest in the subgroup detection setting. In this work, we use a one-sided version of FS searching for up-regulated subgroups. For down-regulated subgroups, one may proceed analogously by selecting features with

Important default parameters for FSOL workflow	
1. Pre-selection of features according to univariate FS ranking	
univariate ranking	FS
number of features for Step 2	$T = 50$
2. Assessment of similarity structure of features w.r.t. the indicated subgroup	
OL weights	$\alpha = 1.1$
number of permutations to estimate p_{OL}	$n_{perm} = 1000$
3. Nomination of samples for potential subgroup	
Scoring of feature groups	
threshold for p_{OL}	$t_{OL} = 0.01$
splitting in graph elements	components
minimum feature group size	$\min_G = 1$
feature group ranking by	median FS
Nomination of samples	
mode	average
minimum proportion	$p > 0.5$

Figure 1. Details on FSOL including the default values of FSOL used for the presented simulation studies. See text for description of parameters and possible adaptations.

a subset of expression values that are lower than in the remaining samples. This is easily done by switching signs of centered data before computing the FS score. A more complex way is to consider both directions in the same analysis: For each feature, both FS versions (up- and down-regulation) are computed, and the higher absolute score is assigned to the feature and used for the pre-selection ranking. In the assessment of similarity of features, one would then use the two-sided version for OrderedList (see Yang et al. (2008) for details). In that case, two features are considered similar, if the same set of samples is found in either tail of the two compared lists, e.g. five samples show highest expression values in feature A and lowest expression values in feature B.

In theory, p-values can be obtained by simulating the distribution of FS under the assumption of the null hypothesis. However, for FSOL the features with largest FS values are selected for the second step of the procedure irrespective of their significance. This is motivated by the idea, that the evidence for a true SG is increased if different features indicate this subgroup, even if the single features are non-significant. The details on this combination of features are described in the next paragraph.

Step 2. Assessment of similarity structure of features w.r.t. the indicated subgroup

A number of already proposed approaches that are based on univariate feature ranking only presented the independent interpretation of highly ranked features by means of biological knowledge. Independent of the chosen method for the univariate assessment of features, we propose an additional step where the pre-selected features are grouped according to the sample subgroups that they indicate. This aims at an **increase of evidence for the indicated sample subgroup and its biological relevance**: Assume that a single feature shows a distinct up-regulation of only three samples. The small number of potential subgroup samples might raise doubt about the biological meaning and one may be more prone to suspect outlier values due to technical issues. However, if the same subgroup shows an up-regulation in additional features the risk of a false positive nomination of the subgroup is reduced. For example, in the analysis of gene expression microarrays, this increase of evidence is especially true if the features are annotated with different genes, that might be known to be involved in the same pathway. In case the 'supporting' feature is a probe set, that is annotated with the same gene, this might not add to the interpretability of the subgroup by pointing out possibly involved pathways, but it does help diminish the concerns of solely technical issues in the measurement of the feature. Furthermore, the grouping of features offers an additional benefit in the context of subgroup detection. It helps to condense the sample set that is taken into account for the subgroup nomination (see Step 3 below for details) which further increases evidence for the nominated sample SG.

However, we want to point out an important issue with regard to the interpretation of evidence: While the indication of a sample subgroup by multiple features does increase the evidence for the subgroup, one should not disregard single features with a distinct subgroup pattern out of hand. A simple explanation for missing confirmation by additional features might be that these were not among the top T features, but were assigned to slightly lower ranks. This is relevant if either multiple SGs are present in the data set or if few SGs affect a large number of features.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1305v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

Original application:
gene lists: rank features by p-value

r	gene list 1	gene list 2	O_r
1	feature A	feature Z	0
2	feature E	feature A	1
3	feature C	feature C	2
4	feature F	feature H	2
...	...	...	...

$\rightarrow$

FSOL: Feature similarity w.r.t. subgroups:
features: rank samples by expression values

r	feature 1	feature 2	O_r
1	sample a	sample z	0
2	sample e	sample a	1
3	sample c	sample c	2
4	sample f	sample h	2
...	...	...	...

Table 1. Illustration of the switch in interpretation to utilize OrderedList algorithm for the assessment of feature similarity with respect to the indicated sample subgroup.

Application of OrderedList to assess similarity of features

For the grouping of features according to the sample subgroup they indicate we make use of an already existing method that was originally developed in order to compare ordered result lists from gene expression studies. The algorithm is called *OrderedList* and is implemented in the eponymous R package Yang et al. (2008). Assume two lists that result from different experiments. Let both lists contain the identical set of features but in possibly different orders. This order might be determined by the rank of a p-value from a differential analysis, for example. Two lists are considered similar if a similar set of features is among the top ranks of both lists. The first step in the computation of OrderedList's similarity score is to determine the number of shared entries among the top r ranks of both lists, i.e. the size O_r of the overlap of top ranks (see Table 1 for short illustration). This information is summarized in a *weighted overlap score*

$$wos = w_\alpha \sum_r O_r, \qquad w_\alpha = \exp(-\alpha r)$$

assigning larger weights to the tails of the list. The choice of the parameter α directly influences the number of top ranks considered for the similarity assessment as well as the individual weights of those ranks. Thus, α should be chosen according to the expected size of the unknown subgroup(s) in the data. For example, our default value of $\alpha = 1.1$ corresponds to ranks 1 to 10 being taken into account. The significance of the observed similarity can be assessed by empirical p-values (p_{OL}) that are obtained via permutation. To this end, one of the lists is shuffled n_{perm} times and the empirical distribution of the observed weighted overlap scores is determined. Then, p_{OL} is the proportion of iterations with a higher observed similarity.

The OL algorithm is directly applicable for the assessment of feature similarity assessment in FSOL: Two features are regarded similar in our context if a similar set of samples has the highest expression values in both features. The direct comparison of the interpretation in the original context and in FSOL is shown in Table 1. The results of Step 2 are gathered in a matrix M_{OL}, where entry (i, j) is the OrderedList p-value p_{OL} for the comparison of those features with ranks i and j in the FS ranking. Thus, if $T = 50$ features have been pre-selected, the similarity structure is summarized in the matrix M_{OL} of size 50×50.

Step 3. Nomination of samples for potential subgroup

The final step in the FSOL workflow is the actual nomination of samples that belong to the indicated sample subgroup(s) based on the feature grouping results in M_{OL}. We here present the workflow as it is used in the simulation study presented later. Options for a more flexible analysis can be found in the result section on the ALL data.

In order to define feature groups that are 'sufficiently' similar in terms of the subgroup they indicate, a threshold $t_{OL} = 0.01$ is defined. Similarity of features is then binarized in the following way: Two features are considered similar (with respect to the indicated subgroup) if their pairwise comparison yields a p-value p_{OL} below t_{OL}. This grouping result is summarized in an adjacency matrix A, with $A_{i,j} = 1$ if the features with FS ranks i and j are regarded similar, and 0 otherwise. The matrix A enables a graphical representation of the splitting of features in distinct groups (see result section on ALL for an example, Figure 3, panel B). The connected groups of features are so-called components, which can be subdivided into maximal cliques, i.e. feature groups where each pair is considered similar according to t_{OL}. The simulations are based on splitting in components, but for single data sets, we suggest to compare both splitting approaches. An advantage of maximal cliques over components is the higher coherence of groups since the similarity measure is not transitive.

Suppose the splitting results in g feature groups $G_1, \ldots, G_g$. The detected feature groups are ranked by means of the median FS score of the individual features in the respective group to determine

the one with the highest subgroup evidence. Additional information may be gained if the indicated subgroups are determined not only for a single feature group but for all g groups, as different groups might indicate the same sample set.

Let G_0 denote the selected feature group of interest. In order to be consistent with the similarity measure, we take into account the number of top ranks r_{max} that are considered in the computation of p_{OL} when nominating the sample subgroup. This parameter should be adjusted with regard to the overall sample size and the size of the sample subgroup that is considered interesting for further evaluation. A sample is nominated by a feature group G_0 if it is among the top r_{max} ranks for a pre-defined number n_{nom} of features in G_0 (thus $n_{nom} \in \{1,\ldots,|G_0|\}$). We suggest a moderate nomination criterion with a minimal proportion of $p = 0.5$, such that the nominated sample is among the top r_{max} ranks in more than half of the features in G_0 (i.e., $n_{nom} = \lfloor p \cdot |G_0| \rfloor + 1$). Alternatively, one could choose a liberal criterion using $n_{nom} = 1$ or the conservative option $n_{nom} = |G_0|$.

Set-up of the comparison of FSOL to a reference workflow

We now provide all means necessary for a comparison of FSOL to the traditional biclustering approach. First, we give some details on the chosen biclustering algorithm. Then, we introduce a real data set that is used to demonstrate the usefulness of FSOL. It also helps to illustrate the high variability in biclustering results, which we consider a major drawback for practical application. As the comparison based on a single real data set is not very conclusive, we also conducted a simulation study. We explain its design and list the different parameter settings before we define an appropriate criterion to assess each methods accuracy in the detection of sample subgroups.

Biclustering using the Plaid model

We give a brief idea of the biclustering algorithm used in this work. We make use of the R package `biclust` (Kaiser et al., 2015) that provides an implementation for the Plaid model biclustering as described in Turner et al. (2005). In the Plaid model, the matrix Y of expression values is considered a sum of so-called *layers*, that are basically linear models in the form of

$$\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$$

for layer k, where i indicates the feature and j the sample (i.e. row and column of Y, resp.) A greedy algorithm adds individual layers to the background layer, if a sum of squared residuals are reduced 'sufficiently'. For K layers, the expression value y_{ij} of sample j in feature i is modeled as

$$y_{ij} = \theta_{ij0} + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk} + \varepsilon_{ij},$$

where ρ_{ik} and κ_{jk} indicate whether feature i and sample j, respectively, are included in layer k. Each sample and each feature can be included in different layers, but can also be included in no layer at all. Each fitted layer corresponds to a so-called bicluster.

Real data example: ALL

To illustrate the application and the benefit of FSOL on a real world data set, it is advisable to choose a supervised setting, where a truly existing sample subgroup is known. This allows the valid assessment of the detection accuracy of the applied method. Therefore, we use the ALL data set, which is provided in the identically-named R package by Li (2009). The data set contains Affymetrix chip data and corresponding clinical data of 128 patients diagnosed with acute lymphoblastic leukemia (ALL for short). The given expression matrix contains normalized expression values of 12,625 features. More details can be found in the ALL manual.

To construct an appropriate data set, the covariates `BT` and `mol.biol` were used. The first one indicates if the patient suffers from a B or T cell leukemia and the latter one specifies the molecular biology of the tumour. The homogeneous group ('controls C') comprises only B-cell leukemia with the same assigned molecular biology (labeled `NEG`), while the heterogeneous group ('diseased D') consists of B-cell patients that mainly express the BCR/ABL fusion gene but also a subgroup of five patients with a different fusion pattern (E2A/PBX1). This results in the comparison

42 NEG vs. (37 BCR/ABL + 5 E2A/PBX1)

For simplicity, this subset of the ALL data set will be refered to as *ALL data* throughout this work. While the true SG is known by 'spiking in' the third molecular subtype, we do not know the exact set of features that is affected by this subgroup. Aside from that, we cannot rule out the possibility of additional patient subgroups, e.g. within the BCR/ABL group. Notably, the plot of the first principal components does not indicate any sample subgroups (see Figure 2, Panel A).

parameter	description	values
n_{sg}	sample size of subgroup	$5, 10$
δ	shift in mean for SG	$1.5, 2, 3, 4, 6$

Table 2. Overview of simulation parameters for the comparison of FSOL and biclustering. In all cases, $p_{sg} = 5$ out of $p = 1000$ features in total are affected by the SG in a group of $n = 40$.

Design of the simulation study

In addition to the ALL data example that demonstrates how FSOL enables the detection of patient subgroups, we present a systematic comparison of FSOL with the biclustering approach by means of a simulation study. In contrast to the real data example, not only all true sample subgroups are known, but also the exact feature set that they might affect. For each setting, we compare the methods with respect to their sensitivity to detect the existing subgroup as well as their accuracy in the actual nomination of the samples for the subgroup.

In each simulation run l, $l = 1, \ldots, L = 500$, we first draw a data set of size $p \times 2n$ from $N(0,1)$, i.i.d. (independent, identically distributed). We choose $p = 1000$ as the number of features in the data set and n denotes the sample size per group. The first n columns represent the samples in the diseased (heterogeneous) group, the remaining n columns the control samples (homogeneous group). Then, a subgroup of n_{SG} samples is induced, which is reflected in the replacement of a submatrix of size $p_{SG} \times n_{SG}$ that contains random numbers from a shifted distribution $N(\delta, 1)$. Thus, the shift parameter δ controls the extent of up-regulation in the subgroup. Table 2 lists the values of simulation parameters in the study.

Assessment of detection accuracy using the Jaccard index

To comprehensively assess the performance of a subgroup detection workflow both the true sample subgroup and the true set of indicating features should be known. This is the case for the simulation studies, but only to a limited extent in the real data example. An appropriate measure to assess and compare the performance of different SG detection workflows is given by the Jaccard index J:

$$J(A,B) = |A \cap B| \; / \; |A \cup B|,$$

where A and B represent the sets of nominated and true SG samples, respectively. For further analyses, also the accuracy of both methods with respect to (w.r.t.) the accuracy in detected feature groups could be compared. However, here we focus on the identification of sample subgroups.

RESULTS AND DISCUSSION

First, we compare the suggested FSOL workflow and the biclustering approach by the example of the ALL data set described earlier. Starting with FSOL, we explain the default workflow listed in Figure 1, but also suggest some alternative options that may be helpful in the daily routine. Although there is a stochastic component involved in FSOL in the simulation of the similarity p-values p_{OL}, the results of different repetitions are very similar (data not shown here). On the contrary, repeated application of the biclustering algorithm generates significantly different results. In fact, the variation between runs is so substantial, that a proper assessment of the biclustering performance is not feasible for the real data example without additional summary methods. Second, we present the results of the simulation study where we evaluate the performance of both methods and compare them directly.

Application of FSOL to ALL data

We here present some options for the exploration of individual data sets that are not applicable for an automated analysis workflow that is required e.g. for the subsequently presented simulations. However, in practice one might prefer this manner to get a more in depth view on the data.

Figure 2 shows a scatter plot of the first two principle components of the ALL data (panel A). This illustration is commonly used as a first check for sample subgroups in a data set. Here, neither the two large groups (tumour types NEG and BCR/ABL) do separate, nor is there any hint for the smaller spiked-in E2A/PBX1 subgroup, which we aim to identify. In contrast, the visual inspection of the heatmap of the **pre-selected top 50 FS features** clearly points out a sample subgroup in the data (Figure 2, panel B): It contains the 5 true subgroup samples and one additional sample. 6 features seem to be affected, among them PBX1, which is involved in the defining gene mutation (E2A/PBX1) of the spiked-in subgroup samples. This shows that already Step 1 of FSOL may help in the detection of sample subgroups and affected feature sets.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1305v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

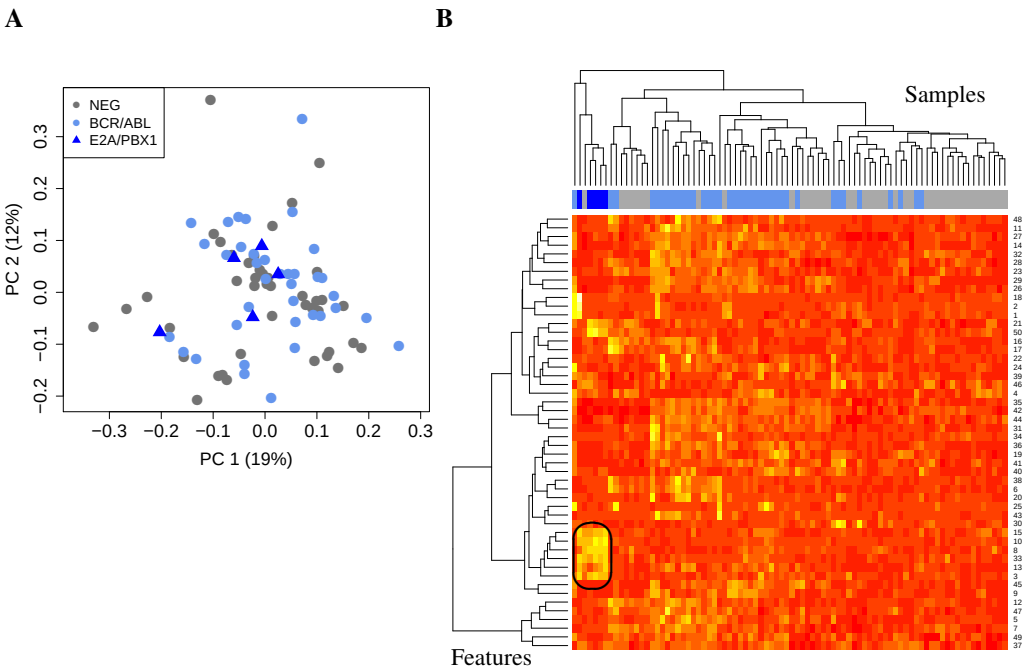

Figure 2. Illustrations of ALL data set. A: Plot of first 2 principal components (proportion of variance in parentheses). The true molecular subgroup (E2A/PBX1) is not separated from the other groups. B: Heatmap based on expression values of top 50 FS features of ALL data, i.e. after Step 1 of FSOL (hierarchical clustering, Euclidean distance, complete linkage). Columns correspond to samples and are color-coded according to tumour type (see Panel A for legend). Rows represent features and are labeled on the right with their FS rank. The most prominent (bright yellow) block is marked by an oval and contains all samples of the SG of interest as well as the mutation defining PBX1 gene.

Figure 3 (panel A) demonstrates the additional benefit of the usage of OrderedList to **measure similarity of pre-selected features w.r.t. the sample subgroups** they indicate. For the graphical representation of Step 2 of FSOL we provide two options (see Figure 3): The more flexible way is to use the heatmap of the $(-\log_{10})$ p-values p_{OL} to identify feature groups. Besides the group that was already identifiable in Figure 2, there is now a larger number of feature groups recognizable, that may point to different sample subgroups. For the analysis of a single data set, one would check for known biological associations between the apparent features in each group of interest at this point. Note that this procedure based on visual examination does not require a threshold for p_{OL} in order to characterize two features as similar or not. It is not feasible for an automated analysis like in our simulation studies, where we set t_{OL} to 0.01. The results of this approach are shown in panel B of Figure 3.

The last step of the workflow is the **identification of the sample set that is nominated as a subgroup by a given feature group**. Here, we have a closer look at one of the maximal cliques we identified in this graph (again Figure 3, panel B, darker grey shading). As this exemplary group of four features is a maximal clique, each pair of features has a p_{OL}-value below 0.01. We consider a sample being nominated for a potential subgroup by this feature set, if it is among the top 10 ranks of at least 3 of the 4 the features. This is translated into $r_{max} = 10$ and $p = 0.5$ using the notation introduced above. Application of these criteria yields a nominated subgroup of seven samples which includes the five samples of the true E2A/PBX1 subgroup and two additional samples. This results in a Jaccard index of $5/7 = 0.71$.

Application of biclustering to ALL data

The inherent optimization of the biclustering algorithm is the reason for a substantial potential of variability. To quantify this variability and assess the reproducibility of biclustering results, we apply the algorithm 1000 times to the ALL data and count the number of biclusters found in each run, see Table 3. Although this number ranges from 0 to 10 (part A of Table 3), results vary less if one focusses on the first bicluster only (if any are detected). This is shown in parts B and C. In B, numbers of features and samples from the first bicluster are listed. Part C tabulates in how many of the biclusters

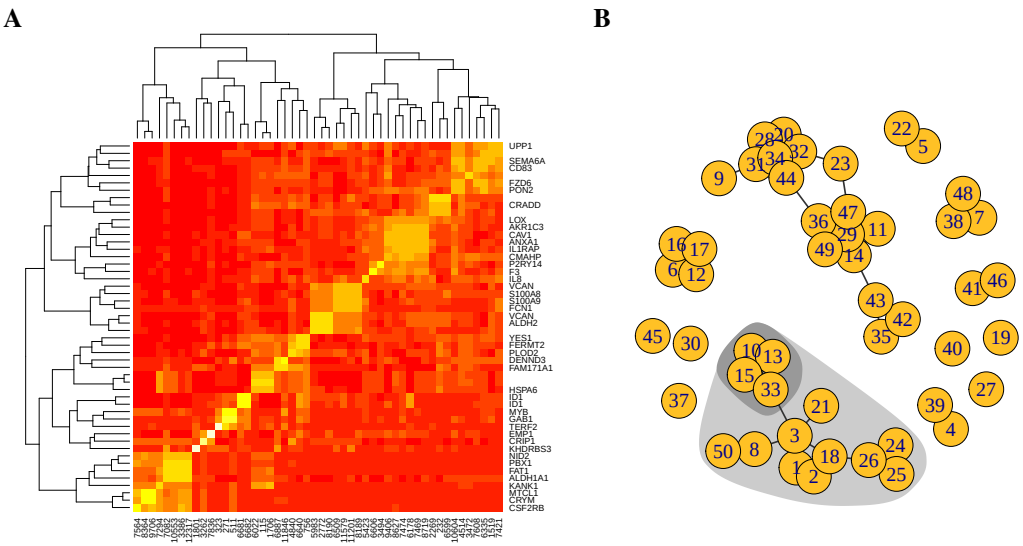

Figure 3. Options for the visualization of FSOL results. A: Application of traditional clustering (again, Euclidean distance, complete linkage) to the matrix of $-\log_{10} p_{OL}$ that contains (transformed) p-values obtained from pairwise feature comparisons using OrderedList. (For computational reasons, p-values equal to zero are set to the minimum positive p-value, here 0.001.) B: Representation of FSOL grouping results suitable for automated analysis. Nodes represent features, where node numbers reflect the feature ranking according to FS. Two nodes are linked by an edge if the observed similarity w.r.t. the indicated subgroup is high. More precisely, an edge is drawn, if $p_{OL} < 0.01$. The light grey shading indicates an exemplary maximal connected component (path exists between each pair of features), the darker shading highlights a maximal clique (edge between each pair of features).

the individual samples are included. Among the 14 samples involved, there are only two of the spiked-in subgroup (with sample IDs 2 and 4), but as stated above, we cannot rule out the possibility of additional true sample subgroups in the data that are detected by the biclustering algorithm.

Simulation study to compare FSOL and biclustering approach

Here, we present the comparison of FSOL and biclustering in detail for one of the parameter settings defined above, while we briefly summarize the results for the other settings. For similarity to the ALL example, we focus on the comparison with group size $n = 40$ and a true sample subgroup of size $n_{SG} = 5$, affecting $p_{SG} = 5$ features.

First, we describe the results obtained by the biclustering approach. Table 4 A gives the numbers of biclusters identified for different shifts δ. For small to moderate shifts, a strikingly large number of runs results in no indication for the existing subgroup. This number of negative results decreases only slowly with larger shifts. For an exploratory approach this lack of sensitivity might not be favorable. The comparison of the samples involved in the best bicluster with the true sample subgroup shows that only for large shifts a reasonable number of true subgroup samples is included in this bicluster. However, for a large shift of $\delta = 6$ the biclustering algorithm returns at least one bicluster in most runs and the determined sample subgroup includes all five spiked-in samples in 390 of 500 runs. Considering the corresponding Jaccard indices that are shown in Figure 4, we can state that only for large shifts, biclustering yields perfect accuracy in a high proportion of runs. For comparison, Figure 4 also contains the presentation of Jaccard indices for a subgroup of $n_{SG} = 10$ within a group of again $n = 40$. Here, we see that biclustering yields sufficient accuracy for a smaller shift of $\delta = 4$ (instead of 6).

For the description of FSOL results, we first list the numbers of truly detected subgroup samples before we directly compare the detection accuracy of FSOL to the accuracy of biclustering with respect to the obtained Jaccard indices. The actual number of feature groups that results from the grouping in Step 2 is much less informative (for $\min_G = 1$) than the number of biclusters, thus the respective numbers are not tabulated here.

Note, that in contrast to biclustering, FSOL always reports a 'best' feature group. This fact is only advantageous if the best feature group indicates the true subgroup in a sufficiently high proportion of cases. Table 5 shows that FSOL does indeed tend to nominate true subgroup samples in the best feature group and the number of truly detected samples increases with the shift δ. To further

A *Number of biclusters (BC) found in 1000 repetitions*

no. of BC found	0	1	2	3	4	5	6	7	8	9	10
frequency	6	138	161	158	153	127	86	65	46	31	29

B *Number of features in best BC*

no. of features	85	290	none*
frequency	77	917	6

Number of samples in best BC

no. of samples	7	10	none*
frequency	917	77	6

C

sample ID	2	4	8	10	24	25	26	28	29	30	31	32	33	42	none*
frequency	917	917	77	917	917	77	77	77	994	994	77	994	77	77	6

Table 3. Assessment of bicluster (BC) stability observed in 1000 repetitions on ALL data (complete feature set). The algorithm was applied to the heterogeneous group where we aim to find a subgroup. A gives the frequencies of numbers of biclusters found in the different runs. B summarizes the size of the best bicluster in terms of samples and features, and C tabulates the frequencies of samples to be found in the best BC. B and C both refer to the respective best bicluster in each run, i.e. the first one selected by the algorithm. Counts and IDs labeled 'none*' indicate the runs where no BC was found.

A

	# biclusters found						
δ	0	1	2	3	4	5	6
1.5	**489**	11	0	0	0	0	0
2	**492**	8	0	0	0	0	0
3	**479**	21	0	0	0	0	0
4	**386**	102	10	2	0	0	0
6	22	**373**	75	24	3	2	1

B

	# true samples					
δ	0	1	2	3	4	5
1.5	**5**	3	2	1	0	0
2	**4**	2	1	0	1	0
3	2	2	4	**7**	3	3
4	2	2	8	19	22	**61**
6	1	0	11	23	53	**390**

Table 4. Simulation results for biclustering approach. Table A gives the numbers of identified biclusters in 500 runs for each of the different shifts δ. For those runs where at least one bicluster was found, the right table B gives the number of true subgroup samples in the best bicluster.

assess the accuracy in the nominated subgroups, we again refer to Figure 4 where the Jaccard indices of both methods are compared directly for increasing shifts. According to the simulation results, FSOL outperforms the biclustering approach considerably in terms of sensitivity and accuracy of the subgroup detection for small to moderate shifts. Only for very large shifts and for a subgroup of 5 samples in a group of size 40, biclustering shows better accuracy.

	# true samples					
δ	0	1	2	3	4	5
1.5	155	**175**	99	35	27	9
2	115	**137**	81	42	62	63
3	37	47	19	8	65	**324**
4	13	12	3	2	13	**457**
6	12	11	4	1	2	**470**

Table 5. Number of true SG samples among the sample set nominated by the best feature group in FSOL.

To sum up, the suggested workflow **FSOL is more sensitive in the detection of small subgroups than the biclustering approach**. This is especially true if the shift amount is small as well. Only for certain combinations of large shifts and/or large subgroups the performance of both approaches can be comparable. An important factor in the explanation of this observation is given in the following: the biclustering algorithm selects a combined set of features and samples as a bicluster only if the impact on a global score that is minimized during the computations is 'sufficiently' large. It is thus influenced by the size of the submatrix involved (number of features times number of samples in the bicluster), more precisely by the proportion of the submatrix to the complete data matrix analysed, as well as by the shift amount. On the other hand, the ability of FSOL to pick up small shifts in smaller subgroups is easily explained by the combination of the univariate pre-selection based on FS, where features with the largest subgroup potential (i.e. highest FS score) are chosen for Step 2, even if the differences are small. The subsequent grouping of those subgroup indicating features according to

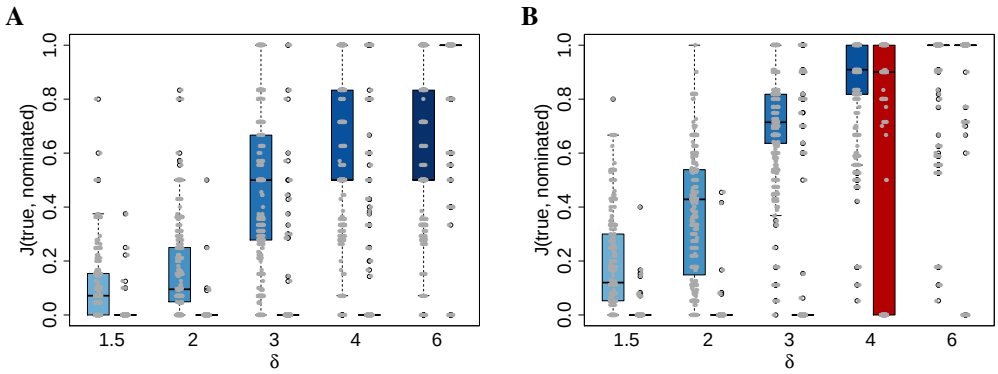

Figure 4. Comparison of FSOL (left boxplots per shift, blue) and biclustering (right boxplots, red) w.r.t. obtained Jaccard indices in different simulation settings. For 500 runs per setting and shift δ, the Jaccard indices of the best feature group of FSOL and of the best bicluster, respectively, are plotted. In both panels group size n equals 40 and subgroup size n_{SG} is 5 in panel A and 10 in panel B, respectively. Note, that in many runs there were no biclusters found at all (see also Table 4), which per definition yields a Jaccard index equal to zero.

the indicated subgroups with a specific similarity measure involves only small variations, yielding reproducible results. The idea behind FSOL is easily comprehensible and provides an understandable graphical representation.

OUTLOOK

An important goal for the future is to apply FSOL to additional real data sets e.g. generated from different omics technologies such as NGS-based DNA-Seq or RNA-Seq data to underpin its benefit for the detection of unknown patient subgroups in different fields. This may also help to assess the need for adjustments in the default parameters of FSOL in future applications, e.g. for different sample sizes. Also the influence of additional filters such as minimal feature group size could be investigated.

Regarding biclustering, the results obtained with the basic Plaid model are too unstable to be useful in practice. To overcome this high variability, ensemble clustering methods have been proposed and implemented (e.g. in the R package `superbiclust` (Khamiakova, 2014)). Those methods aim at the condensation of different biclusters that are resulted in repeated runs, but also within an individual run, as in the Plaid model used here, samples and features can be included in different biclustern in the same run. In future studies, this extended biclustering approach should be compared to FSOL to quantify the improvement over biclustering results.

REFERENCES

Ahrens, M., Turewicz, M., Casjens, S., May, C., Pesch, B., Stephan, C., Woitalla, D., Gold, R., Brüning, T., Meyer, H. E., Rahnenführer, J., and Eisenacher, M. (2013). Detection of patient subgroups with differential expression in omics data: A comprehensive comparison of univariate measures. *PLoS ONE*, 8(11):e79380.

Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., Theron, R., Quintales, L., Leisch, F., and Troyer., E. D. (2015). *biclust: BiCluster Algorithms*. R package version 1.2.0.

Khamiakova, T. (2014). *superbiclust: Generating Robust Biclusters from a Bicluster Set (Ensemble Biclustering)*. R package version 1.1.

Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86.

Li, X. (2009). *ALL: A data package*. R package version 1.4.16.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.

Shah, R. B., Mehra, R., Chinnaiyan, A. M., Shen, R., Ghosh, D., Zhou, M., MacVicar, G. R., Varambally, S., Harwood, J., Bismar, T. A., Kim, R., Rubin, M. A., and Pienta, K. J. (2004). Androgen-independent prostate cancer is a heterogeneous group of diseases: Lessons from a rapid autopsy program. *Cancer Research*, 64(24):9209–9216.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1305v1 | CC-BY 4.0 Open Access | rec: 13 Aug 2015, publ: 13 Aug 2015

Slamon, D., Clark, G., Wong, S., Levin, W., Ullrich, A., and McGuire, W. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177–182.

Teschendorff, A. E., Naderi, A., Barbosa-Morais, N. L., and Caldas, C. (2006). PACK: Profile analysis using clustering and kurtosis to find molecular classifiers in cancer. *Bioinformatics*, 22(18):2269–2275.

Tibshirani, R. and Hastie, T. (2007). Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1):2–8.

Tockman, M., Mulshine, J., Piantadosi, S., Erozan, Y., Gupta, P., Ruckdeschel, J., Taylor, P., Zhukov, T., Zhou, W., Qiao, Y., and Yao, S. (1997). Prospective detection of preclinical lung cancer: results from two studies of heterogeneous nuclear ribonucleoprotein A2/B1 overexpression. *Clinical Cancer Research*, 3(12):2237–2246.

Turner, H., Bailey, T., and Krzanowski, W. (2005). Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data Analysis*, 48(2):235 – 254.

Wu, B. (2007). Cancer outlier differential gene expression detection. *Biostatistics*, 8(3):566–575.

Yang, X., Scheid, S., and Lottaz, C. (2008). *OrderedList: Similarities of Ordered Gene Lists*. R package version 1.38.0.