Computational and Structural Biotechnology Journal 20 (2022) 1567–1579

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

SRIQ clustering: A fusion of Random Forest, QT clustering, and KNN concepts

Jacob Karlström, Mattias Aine, Johan Staaf *, Srinivas Veerla *

Division of Oncology, Department of Clinical Sciences Lund, Lund University, Medicon Village, SE-22381 Lund, Sweden

ARTICLE INFO

Article history:
Received 14 December 2021
Received in revised form 21 March 2022
Accepted 31 March 2022
Available online 04 April 2022

Keywords:
Lung adenocarcinoma
Clustering
Molecular subtypes
Gene expression
Random Forest
QT clustering
KNN

ABSTRACT

Gene expression profiling together with unsupervised analysis methods, typically clustering methods, has been used extensively in cancer research to unravel, e.g., new molecular subtypes that hold promise of disease refinement that may ultimately benefit patients. However, many of the commonly used methods require a prespecified number of clusters to extract and frequently require some type of feature preselection, e.g. variance filtering. This introduces subjectivity to the process of cluster discovery and the definition of putative novel tumor subtypes. Here, we introduce SRIQ, a novel unsupervised clustering method that could circumvent some of the issues in commonly used unsupervised analysis methods. SRIQ incorporates concepts from random forest machine learning as well as quality threshold- and k-nearest neighbor clustering. It is implemented as a Java and Python pipeline including data preprocessing, differential expression analysis, and pathway analysis. Using 434 lung adenocarcinomas profiled by RNA sequencing, we demonstrate the technical reproducibility of SRIQ and benchmark its performance compared to the commonly used consensus clustering method. Based on differential gene expression analysis and auxiliary molecular data we show that SRIQ can define new tumor subsets that appear biologically relevant and consistent compared and that these new subgroups seem to refine existing transcriptional subtypes that were defined using consensus clustering. Together, this provides support that SRIQ may be a useful new tool for unsupervised analysis of gene expression data from human malignancies.

1. Introduction

Cancer is the second leading cause of death globally and estimated to account for 9.6 million deaths in 2018 [1]. A hallmark of cancer is the vast heterogeneity of the disease, between different anatomical sites, between different patients, and within individual tumors. Cancer outcome is gradually improving through better diagnostics, clinical management, and novel therapeutics. To further refine patient prognostication and treatment prediction new tools/methods are needed beyond existing clinical markers, which to a large extent are still based on morphological or single marker analyses of tumor tissue. One molecular method that has shown promise in, e.g., breast cancer is gene expression profiling, where some of the first reported gene signatures now exist as commercial products with approval for use in the clinical setting. In cancer research in general, gene expression profiling has been used extensively to unravel new molecular subtypes as well as to define prognostic and/or treatment predictive gene signatures. Commonly, different unsupervised analysis methods for clustering samples have often been used for subtype discovery, ranging from the simplest form of hierarchical clustering (e.g., defining the original breast cancer subtypes [2]) to more refined methods including e.g. bootstrapping [3]. One of the most commonly used approaches represent variations of consensus/bootstrap clustering in which cluster solutions are arrived at after a large number of repeated clustering loops that reduce the impact of outliers thereby yielding more robust results. A limitation of consensus/bootstrap clustering is the requirement of a predefined parameter, K, for the number of clusters to be defined. Even though there are means to gauge the optimal K [4,5], there is no absolute way of knowing the true number of biological entities and evaluation of clustering results therefore requires in-depth knowledge of the cohort and tumor type in question. In contrast, density based clustering algorithms, e.g., Quality Threshold (QT) clustering [6] and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [7] find clusters without a prior K value, but are limited in the case of high

* Corresponding authors.
E-mail addresses: johan.staaf@med.lu.se (J. Staaf), srinivas.veerla@med.lu.se (S. Veerla).

https://doi.org/10.1016/j.csbj.2022.03.036

dimensional data [8]. Moreover, clustering methods for gene expression data may often restrict the number of input features (genes) for different reasons, which may constrain or make the results dependent on the selected feature set. It is also customary to use all available cases (tumors) for clustering. Even though including a large number of cases increases the clustering efficiency and allows for the discovery of less common tumor phenotypes, it frequently also introduces "fuzzy" borders between the major subgroups as all samples are given equal weighting in the clustering. Traditional clustering methods also do not provide a means of defining core/prototypical cases and due to the issues discussed above, are sensitive to the relative sample composition of the analyzed tumor cohort.

It is well known that distance values of high dimensional data become meaningless because every observation in a dataset is equidistant from all the others, which is famously called the curse of dimensionality [9]. As a result, the clustering process may lead to overfitting and result in unstable clusters. Furthermore, for a high dimensional dataset, the single squared distance matrix values may be highly influenced by the outlier features [10–12]. To reduce the bias resulting from outliers one can use the Jackknife method which calculates the distance value, e.g., a correlation coefficient value (r) between two samples by leaving one feature out at a time for which the maximum of r values is then taken as a final distance value [13]. This procedure is often computationally and memory intensive for large datasets. To overcome the computational issue, the number of features may be reduced by using different filtering methods, e.g. a variance filter, but this may lead to reduced information content. In the present investigation we aimed to develop, benchmark, and evaluate a novel clustering approach termed systematic random forest integration to quality threshold (QT) clustering, SRIQ, which could circumvent some of the issues in unsupervised clustering analysis of tumor gene expression data (Fig. 1). SRIQ incorporates concepts from Random Forest machine learning (bagging and aggregation) [14] to avoid the restriction on the number of input features, as this concept is known for handling large datasets efficiently [14–16] and can reduce the feature dimensionality in a meaningful way. This concept produces squared distance matrices from randomly extracted features subsets and aggregating the results produces robust estimates of distance values, which are not highly affected by the outlier features [15,16]. In addition to this, bagging can also reduce the high variance caused due to missing values in the high dimensional data. Next, QT clustering is used to form "core clusters" (Fig. 1A) while avoiding the requirement for a prespecified number of clusters (K) to be discovered, and finally k-nearest neighbor (KNN) is used to expand core clusters in order to assign subgroup calls to lower confidence or more heterogenous samples (Fig. 1B).

This fusion of concepts makes it possible to select the number of tumor classes based on stability and cluster tightness, and the approach identifies "core clusters" composed of typical samples which are representative of the major biological entities while being less influenced by the entirety of statistical signals present in the data set. For samples falling outside core cluster entities, KNN is used to assign class labels to the remaining samples/cases using core cluster centroids. KNN-expanded core clusters are referred to as "spiral clusters" and provide a means to assign cluster identities to the all cohort samples (Fig. 1A). To assess cluster quality and stability the entire clustering process is iterated, and a pairwise sample co-occurrence matrix is used to evaluate the stability of core and spiral clusters.

To provide a context for evaluating the usefulness of SRIQ in unsupervised analysis of high-dimensional tumor data (whole transcriptome RNA sequencing data) we chose lung adenocarcinoma (LUAD) as a model. LUAD is the most frequent histological type of non-small cell lung cancer (NSCLC) [17], and a highly lethal disease mainly due to late diagnosis. While treatment options for advanced stage LUAD have greatly improved during the last decades, treatments are still often palliative due to nearly inevitable treatment resistance over time. For surgically treated patients, i.e., patients with lower stage tumors treated with a curative intent, there is still a high-risk of metastatic relapse even for tumors of the lowest stage [18]. Consequently, additional prognostic and predictive tools are needed to improve patient outcome. Surgically resected LUAD have been intensively studied using different high-throughput molecular profiling techniques, including gene expression analysis. The latter technique has been used extensively to derive, e.g., prognostic gene signatures as well as transcriptional LUAD subtypes [19–25]. Concerning the latter, three transcriptional subtypes termed the terminal respiratory unit (TRU), the proximal-inflammatory (PI), and the proximal-proliferative (PP) subtypes have been proposed in LUAD based on consensus cluster analysis by The Cancer Genome Atlas (TCGA) consortium [19,22,26]. While these three transcriptional subtypes have been associated with different clinicopathological and molecular features as well as patient outcome, there is still significant heterogeneity within the subtypes and recent studies have proposed other subtyping schemes that intersect the TRU/PI/PP classification [27]. The lack of transcriptional subtype consensus in LUAD indicates that current subtypes may need to be refined through, e.g., analyses of larger more representative cohorts, and/or new analysis methods. As such, LUAD represents a suitable context for assessing how a novel unsupervised analysis method like SRIQ performs in relation to preexisting molecular classifications (TRU/PI/PP) and existing methods (e.g., consensus clustering), and whether it can extract novel biology not currently captured in the existing subtyping schemes.

Based on a well characterized cohort of 434 surgically resected lung adenocarcinoma specimens from the TCGA consortium profiled by RNA sequencing we demonstrate the technical reproducibility of SRIQ and its performance compared to consensus clustering. More interestingly, based on differential gene expression analysis and additional molecular data we demonstrate that SRIQ can define new tumor subsets that appear biologically relevant and consistent in relation to consensus clustering, and that these new subgroups also refine existing transcriptional subtype classifications. Together, this provides support that SRIQ may be a useful new tool for unsupervised analysis of gene expression data from human malignancies.

2. Materials & methods

2.1. SRIQ

The principle of the SRIQ framework is illustrated in Fig. 1B. In brief, we repeatedly (default, t = 10000 number of times) randomly select a given number of genes, termed bag size (default, BagSize = $\sqrt{n}$, n = total genes), from a gene expression matrix and produce between-sample distance matrices (Euclidean or Pearson distance). Next, the average of all random distance matrices is calculated. The QT clustering method [6] is applied on the aggregate matrix using a sliding window of cluster diameter (d), i.e., cluster tightness, ranging from 0 to 1 with an increment interval of 0.01 and a minimum size of the cluster (default, cs=$\sqrt{s}$, s = total samples). At each d cut-off, a distinct number of core cluster solutions are produced without a requirement of a specified K. Subsequently, the KNN method is used to identify nearest neighbors to these core clusters from the remaining samples in the cohort using the core cluster centroids (K = 1 as default). As a result, each core cluster is expanded into what we term a "spiral cluster" (Fig. 1A). To determine the quality and stability of a cluster solution the above pro-

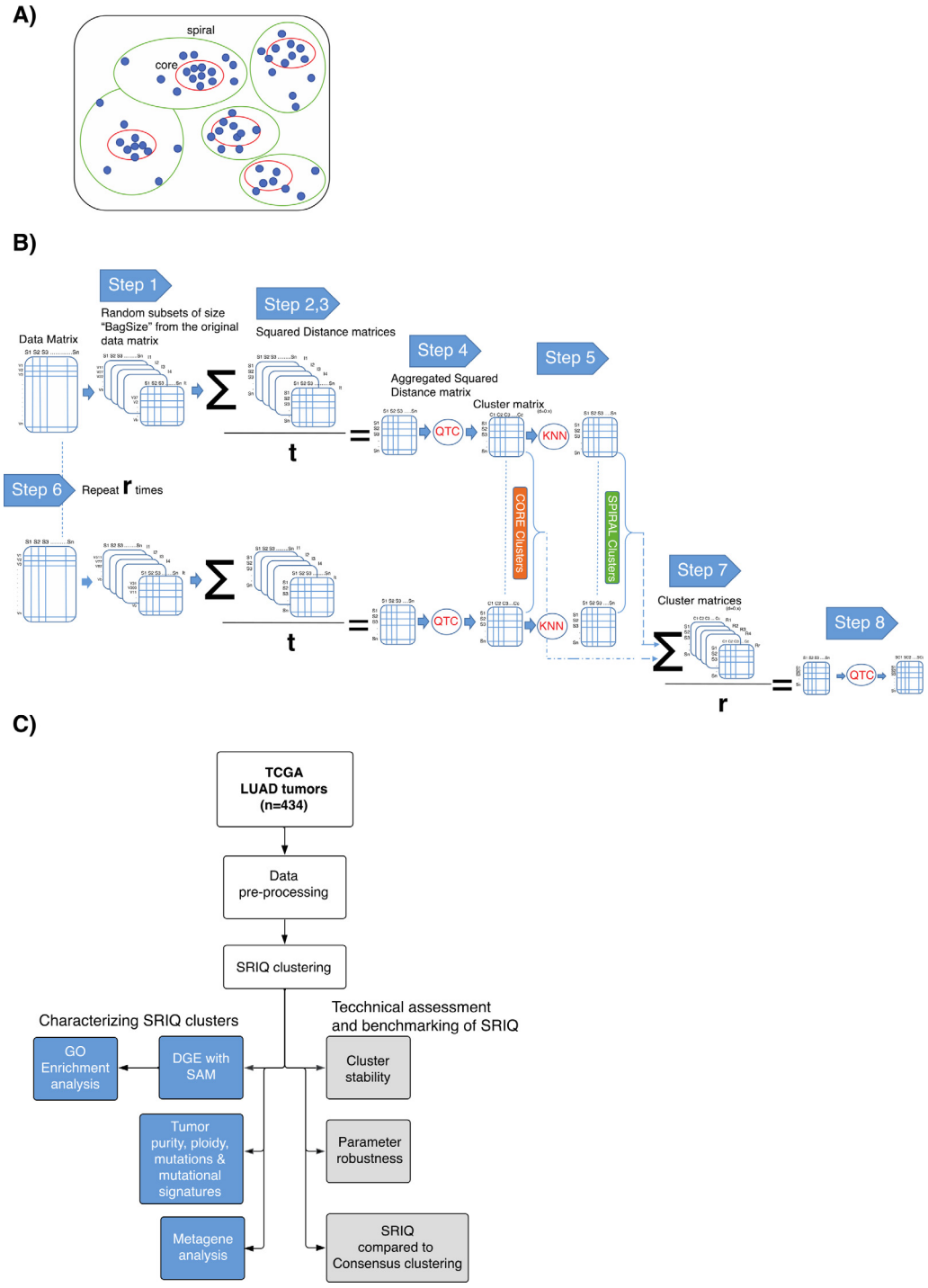


Fig. 1. SRIQ framework and study outline. (**A**) Concept of core and spiral clusters. A blue dot represents a sample, red circles define different core clusters of samples and green circles define different spiral clusters. (**B**) SRIQ framework. (**C**) Study outline. LUAD: lung adenocarcinoma. DGE: differential gene expression analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cedure is repeated (default, r = 10). For each run and diameter, a co-occurrence matrix for each cluster solution (K_x) is generated. Next, the average of the co-occurrence matrix (S) is calculated and converted to a distance matrix (1-S). This final distance matrix for each diameter is subjected to the QT clustering process for a series of quality or stability diameter (QSD), ranging from 0 to 1 with an increment interval of 0.05. This process is stopped at a diameter where it identifies cluster solution (K_x). The resulting clusters are considered final. Subsequently, to calculate quality or stability score (QSS) for cluster solution (K_x) the following formulae is used.

$$QSS = r \times (1 - QSD) \quad (1)$$

The QSS represents the stability/consensus score of sample pairwise cooccurrence among the r repetitive clustering analyses. In this process, some samples may be excluded from being assigned to a cluster, based on alternating cluster associations (instability). We recommend selecting a cluster solution (K_x) with a low cluster

diameter, a high (re-)occurrence across repetitions, and a high QSS. In short, the SRIQ framework includes eight steps as outlined in Fig. 1B: 1) Retrieval of random subsets of the expression matrix using the defined or default BagSize parameter value, 2) calculation of distance matrices for each random subset, 3) calculation of the average across all distance matrices, 4) using step 3 results as the input to the QT clustering method to form core clusters with defined or default parameter values, 5) application of KNN to the core clusters to form spiral clusters, 6) repeating steps 1–5, where for each repetition a co-occurrence matrix is generated for each QSD, 7) calculation of the average co-occurrence matrices from all step 4 (core clusters) results or step 5 (spiral clusters) results for a specific density d, and 8) and applying QT clustering to step 7 results to obtain the final stable core and spiral clusters. SRIQ steps 6–8 are further graphically visualized in the Supplementary Methods File.

The step-by-step pseudocode for SRIQ is provided in the Supplementary Methods File. The SRIQ framework is implemented using the JAVA programming language and utilizes parallel processing wherever repeated calculations are performed during the clustering process. A complete analysis pipeline, including data preprocessing, SRIQ, clustering validation (Silhouette) scores, differential gene expression analysis, and gene ontology (GO) enrichment analysis, written in Jupyter notebook is available at https://github.com/StaafLab/SRIQ. In addition to QSS validity/quality scores, the SRIQ pipeline also outputs Silhouette scores that can be used to decide on cluster solutions.

2.2. Patient cohorts

LUAD gene expression data (n = 60438 transcripts) for 434 cases used in this study was generated through RNA sequencing (RNA-seq) by the TCGA consortium. Expression data was obtained from the GDC portal (https://portal.gdc.cancer.gov/) in FPKM format. Additional clinicopathological data, molecular data, and molecular classifications were obtained from https://gdc.cancer.gov/about-data/publications/panimmune [28], including immune-related classifications from Cibersort [29]. Clinicopathological characteristics of the 434-patient cohort is detailed in Supplementary Table 1.

2.3. Gene expression data preprocessing

Two gene expression datasets were generated from the original RNA-seq data: i) a FPKM expression matrix, from which the median, variance, and log fold change were calculated, and ii) a median centered log2 transformed FPKM (all values below 1 set to 1) matrix, of the top 25,785 most varying transcripts (obtained by removing all transcripts consisting solely of 0's and then removing 65% least varying genes across all samples), used for SRIQ clustering, differential gene expression analysis, and biological metagene analysis. The removal of low varying genes was performed to reduce the sparsity of the gene expression matrix.

2.4. Consensus clustering

As consensus clustering is frequently used to derive molecular subtypes, including LUAD, it was chosen as a benchmark method for SRIQ. Consensus clustering was performed using hierarchical clustering with Pearson correlation as distance metric and ward. D linkage. Number of repetitions were set to 2000 and a resampling schedule of 0.7 was used for both genes and samples. Cluster solutions between 3 and 7 were collected.

2.5. Clustering comparison metrics

We created a metric called Pair Similarity score (PS, not part of SRIQ) to compare similarity and diversity between the SRIQ and consensus clustering solutions. This was done by first creating a binary pairwise similarity matrix and then applying the Jaccard index method to it.

2.6. Lung adenocarcinoma TCGA molecular subtype classification

Classification of the LUAD data according to the proposed gene expression subtypes by TCGA [22]: terminal respiratory unit (TRU), proximal inflammatory (PI), and proximal proliferative (PP) was performed using centroid classification as originally outlined by Wilkerson et al. [19] for the subtypes, using the Wilkerson et al. reported centroid genes matching the TCGA RNA-seq gene set. For each sample the Pearson correlation to each centroid (subtype) was calculated. The subtype with the highest Pearson correlation was assigned to the sample.

2.7. Differentially expressed genes, biological metagenes, and GO enrichment analysis

Differentially expressed genes between SRIQ clusters (one vs rest approach) were obtained by implementation of Significance Analysis of Microarray (SAM) [30] yielding up- and down-regulated genes. For the framework analysis we used SAM to obtain differentially expressed genes with significance parameters, FDR q-value<=5 and a fold change fc>=2. Metagene scores for six metagenes proposed to represent biological processes in lung cancer based on gene network analysis [31] were calculated from expression data for each sample. The metagene score of each sample was calculated as the mean expression of all genes associated with a specific metagene. GO enrichment analysis of differentially expressed genes was performed using the enrichR API endpoint [32–34].

2.8. SRIQ computation time

Computation times for SRIQ in datasets of different sizes and technological platforms (n = 434–3520), including TCGA-LUAD, are provided in Supplementary Table 2.

3. Results

The outline of the study is shown in Fig. 1C incorporating two main branches of analyses: i) technical reproducibility and benchmarking, and ii) assessment of molecular features of derived SRIQ clusters.

3.1. Technical reproducibility and benchmarking of SRIQ

3.1.1. SRIQ clustering of 434 LUAD cases

Application of SRIQ to 434 LUAD cases with expression data for 25,785 transcripts resulted in multiple potential clustering solutions using a BagSize parameter value of 1200 and 10,000 permutations (Fig. 2A, Supplementary Table 1). Of the 434 samples, 382 were classified as core or spiral samples and 52 samples were considered unstable, i.e., alternating between clusters during repetitions. Of the observed cluster solutions, three were of particular interest based on recurrence of cluster solution, occurrence across different diameter solutions, and stability throughout the iterations: i) a three cluster (K3), ii) a five cluster (K5) and iii) a six cluster solution (K6) at cluster diameter cut off ranges of 0.58, 0.60, and 0.63, respectively. Notably, the K6 solution occurred for several

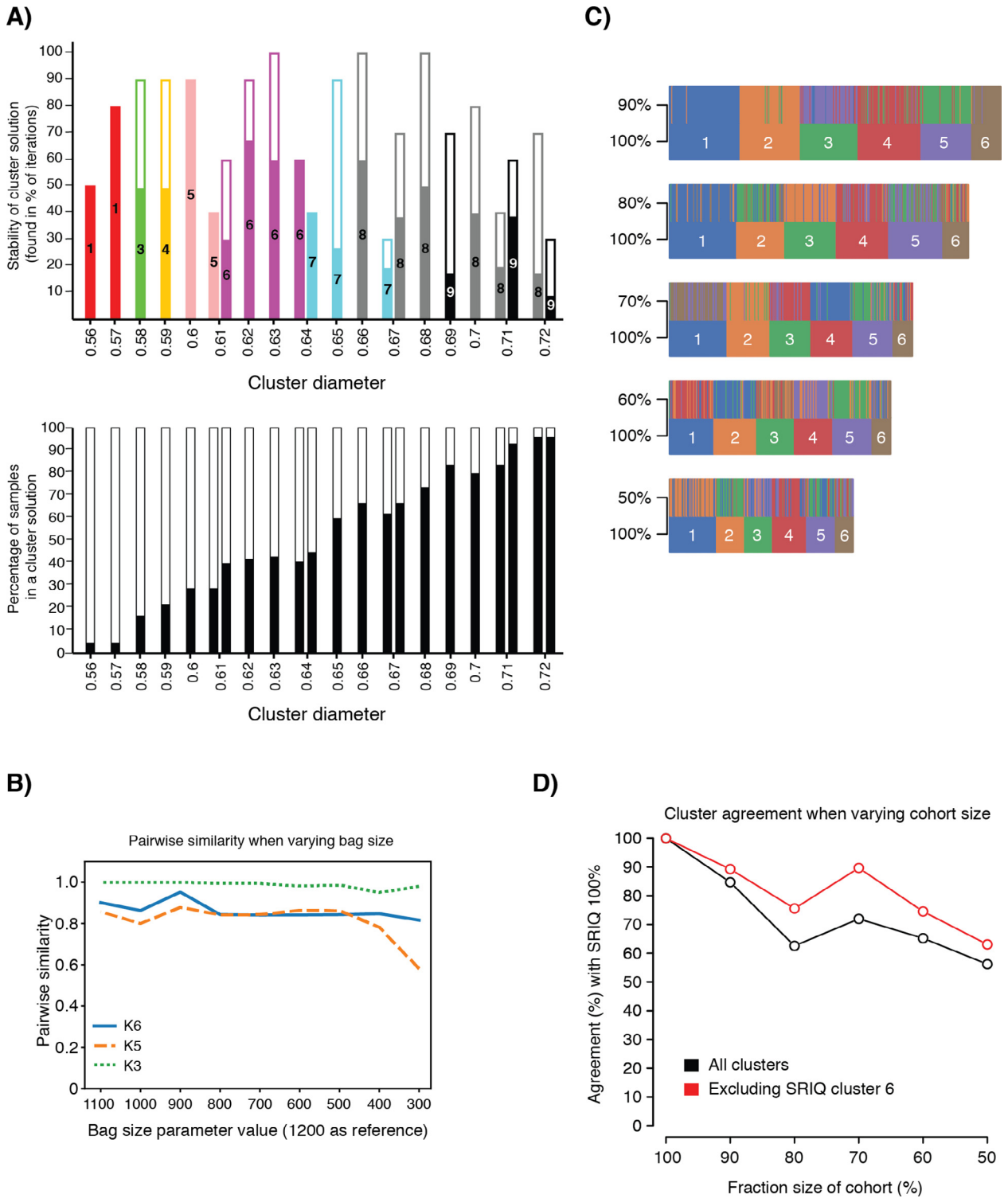


Fig. 2. SRIQ analyses of 434 TCGA LUAD cases and stability analyses. (**A**) SRIQ summary plot using 10,000 permutations and a bag size parameter value of 1200. The height of the bar indicates the % of times the cluster solution has been detected across 10 iterations. The fraction of coloring in the upper bar chart represents the stability score (QSS) of the respective cluster solution as described in the Material and Methods section. The fraction of coloring in the lower bar chart represents the percentage of samples included for corresponding cluster solution above. Only solutions found >2 times (20%) are shown in the panels. (**B**) Pair similarity when varying bag size using the 1200 bag size run as comparison for different K solutions using all samples. (**C**) Sample overlap between SRIQ runs when removing random samples from the original cohort, using 100% of the tumors and a bag size of 1200 as reference solution. E.g., in the panel, 90% corresponds to SRIQ analysis using 90% of all LUAD samples. Samples are ordered according to the 100% solution and colored according to the cluster number set by the 100% solution (irrespective of core/spiral assignment). SRIQ cluster identifiers (colors) can change between runs, thus samples in a 100% cluster may have different cluster assignments in a lower fraction analysis. (**D**) Sample agreement across clusters for the different fraction sizes in C versus the 100% cohort cluster solution as reference, irrespective of core/spiral assignment. Black line corresponds to the agreement (%) from a sample confusion matrix of all six SRIQ clusters between the 100% cohort versus each tested fraction size. To calculate an agreement, we selected for each 100% cluster the corresponding largest cluster in the lower size cohort to be representative. The agreement was then calculated by dividing the number of samples in agreement with the total number of cases from the confusion matrix. Red line corresponds to the same calculation as for the black line, however omitting samples from the smallest SRIQ cluster 100% solution, cluster 6, entirely from all calculations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

diameter cut offs (0.60–0.64) compared to the K3 (0.58) and K5 (0.6, 0.61) solutions (Fig. 2A). At diameter cut-off 0.63, the K6 solution had 100% recurrence across all 10 iterations with a stability score (QSS) of 60% (QSS represents the percentage of seeing sample-sample pairs in the same cluster when repeating the entire clustering analysis). Core cluster samples for this solution were also clearly separated from each other based on UMAP analysis (Fig. 3A) and spiral clusters also showed high recurrence and stability across iterations (Supplementary Fig. 1A).

As seen in Fig. 2A, SRIQ may generate different cluster solutions (e.g., K2, K3, etc.) at the same diameter cut-off, but also similar solutions (e.g. K6) at different diameter cut-offs. For the former, different cluster solutions at the same diameter cut-off often differ by one cluster (e.g. K2 and K3) (Fig. 2A). The reason for this is that one cluster from a lower solution may be divided into two new clusters that do not need to consist of all of the samples of the original cluster, although they usually display some similarity as exemplified in Supplementary Fig. 1B. Furthermore, the number of samples in clusters, as well as number of clusters, typically increases with increased diameter cut-off (Fig. 2A). When both cut-off diameter and the number of samples increases, but K remains the same, three scenarios could happen: i) the existing clusters are expanded upon, ii) one cluster might be split into two new valid clusters and another cluster omitted, and iii) a cluster may be omitted in support of another cluster (Supplementary Fig. 1C). For the purpose of technical reproducibility analysis and algorithm benchmarking we chose to focus onwards on the K6 cluster solution with a cut-off of 0.63, as it showed high stability in both core and spiral cluster solutions (Fig. 2A and Supplementary Fig. 1A), meanwhile appearing at a larger span of cut off diameters than, e.g., a K5 solution. Full data for the selected K6 solution is provided in Supplementary Table 1. Silhouette analysis performed revealed that the selected K6 solution median silhouette score was 0.15, with only a few samples being confounded in other cluster centers (Supplementary Fig. 1D).

3.1.2. Assessment of SRIQ stability for key parameters

Key parameter alterations in clustering algorithms, as well as modifications in cohort composition, can infer dramatic differences in cluster number and composition [35]. To assess the impact of such alterations for SRIQ we focused on cluster alterations inferred by modifications to the BagSize parameter (a key SRIQ parameter representing the number of randomly samples features, i.e., genes, used in each permutation), and modifications inferred by changes in sample numbers (i.e., cohort size/composition).

For the BagSize parameter evaluation we used the results from Fig. 2A as reference (BagSize = 1200) and then performed new runs in which the BagSize value was stepwise reduced from 1100 down to 300 (Fig. 2B). Focusing on three different cluster solutions (K3, K5, and K6) we observed that all solutions showed >80% pairwise similarity to the original run, with the K3 solution showing nearly 100% similarity for each run. Both K5 and K6 showed around 10% lower pairwise similarity, with the K6 solution being more stable than the K5 solution. Furthermore, for the K5 solution the pairwise similarity started to drop at a BagSize of 500 (Fig. 2B).

Next, we changed the sample size for SRIQ by randomly removing an increasing fraction of the original samples, while keeping other parameters intact. As shown in Fig. 2C, apparent overlap between cluster assignments are seen for several clusters even when removing a substantial portion of the total sample number. This is further illustrated in Fig. 2D by calculation of classification agreement. As seen in Fig. 2C-D, small clusters (e.g. SRIQ cluster 6) have a more rapid deterioration in cluster overlaps with a negative effect on overall agreement (Fig. 2D) in smaller cohorts of randomly selected cases compared to the original solution.

3.1.3. Core and spiral cluster samples show similar transcriptional characteristics

A feature of SRIQ is the calling of core and spiral samples, forming core and spiral clusters, respectively (Fig. 1A). Conceptually, it would be expected that spiral samples associated with a specific cluster should have similar, albeit necessarily not as distinct, characteristics as the core samples defining the base cluster. To test this hypothesis, we split each cluster from the K6 solution with cut-off 0.63 (shown in Fig. 2A) into core samples and spiral samples and compared these groups with respect to TCGA LUAD gene expression subtype (TRU/PI/PP), actual TCGA subtype centroid correlation values, as well as expression of six different metagenes representing biological processes previously reported in lung cancer [31].

As shown in Fig. 3B and C, analysis of subtype classification both with respect to discrete subtype status (TRU/PI/PP) and actual subtype Pearson centroid correlation values showed that spiral samples have similar patterns as the core cluster samples. Notably, with the exception of clusters 1 and 2 in the K6 solution, both core and spiral samples within specific clusters are predominantly comprised of a single molecular subtype. Cluster 1 in the K6 solution comprises a mix of TRU and PI cases, whereas cluster 2 comprises a mix of TRU and PP cases. The resemblance of spiral samples with respective core samples within a cluster was further supported when analyzing expression of the six biological metagenes (Fig. 3D). Kruskal-Wallis test revealed that all of the six metagene profiles were statistically significant for core cluster samples ($p < 0.001$) with each cluster showing a distinct metagene expression difference across the six genes. Adding spiral observations still maintains significant expression profiles ($p < 0.05$), however they become less distinct.

3.1.4. Comparison of SRIQ to consensus clustering

To benchmark SRIQ versus another unsupervised algorithm used to derive gene expression subtypes we compared SRIQ clustering to consensus clustering of the same data, focusing on two stable SRIQ solutions, K3 and K6 shown originally in Fig. 2A. A general difference between SRIQ and consensus clustering is that SRIQ can use potentially all expression data (through bagging), whereas prefiltering, e.g. based on expression variance, is commonly used in consensus clustering to reduce the influence of for example non-informative genes and lower the computational requirements. In the SRIQ analysis for this study we did remove non-expressed transcripts across samples, as well as a set of transcripts with very low variation. Still, >25000 transcripts were retained and used for the SRIQ analysis.

In a first comparison we noted that the lower number of genes used for consensus clustering, the more dissimilar the cluster results were to SRIQ. The K3 solutions appear relative similar between the two methods when choosing a feature set > 6000 genes for consensus clustering (see, e.g., Fig. 4A and Supplementary Fig. 2A). For the K6 cluster solution, the two methods differed from each other irrespective of the number of features used in consensus clustering, with <60% similarity for any cluster (Supplementary Fig. 2B and Fig. 4B). As exemplified in Fig. 4B, cluster 1 in the K6 consensus cluster analysis was differentiated by SRIQ into cluster 1 and 3 mainly, whereas consensus cluster 6 was not matching any SRIQ cluster specifically. In summary, as expected the two methods show agreement, but also differences for larger K solutions. To evaluate the biological relevance of the respective solutions more in depth, we proceeded to examine the transcriptional and molecular properties of the solutions.

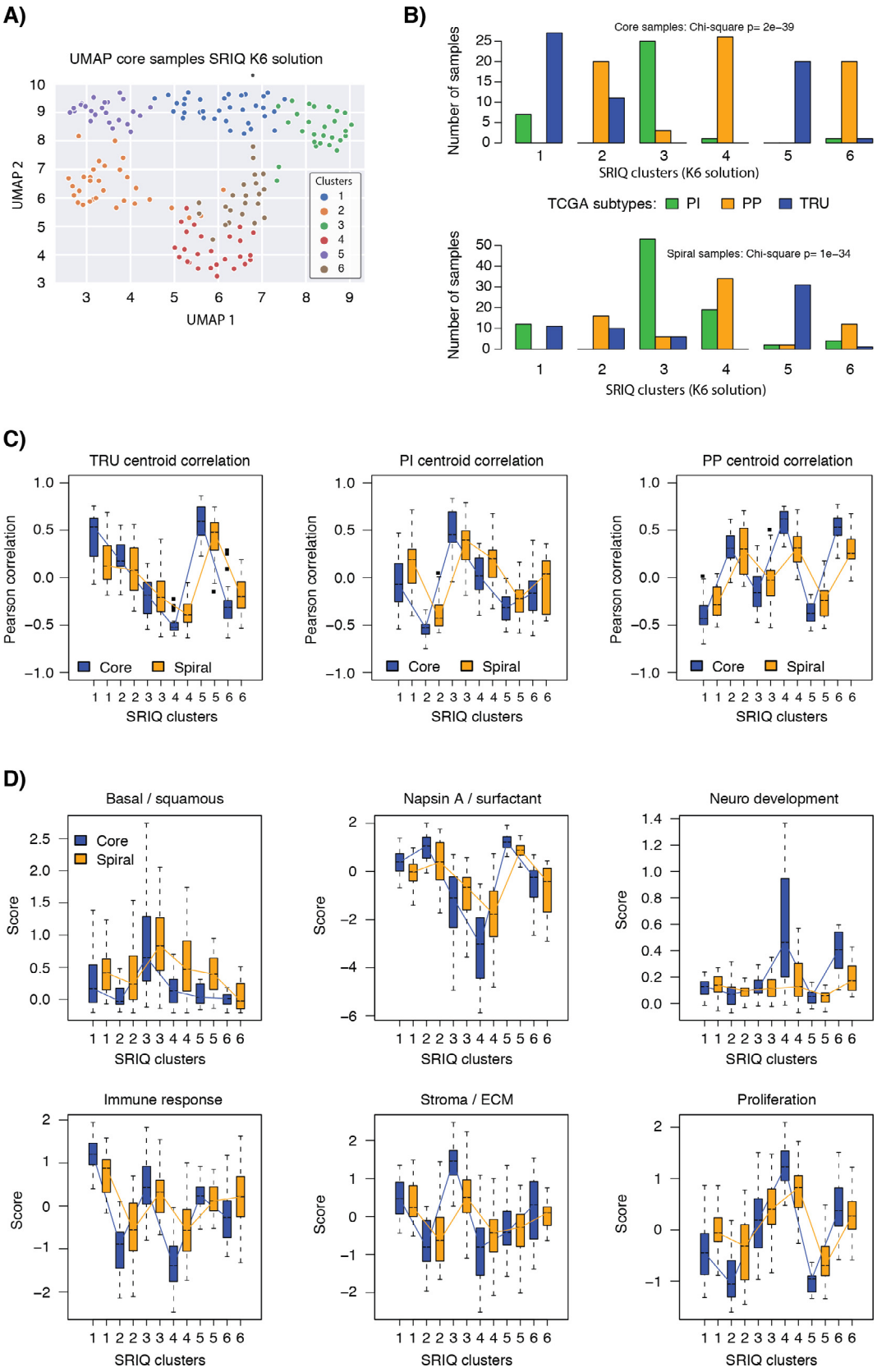
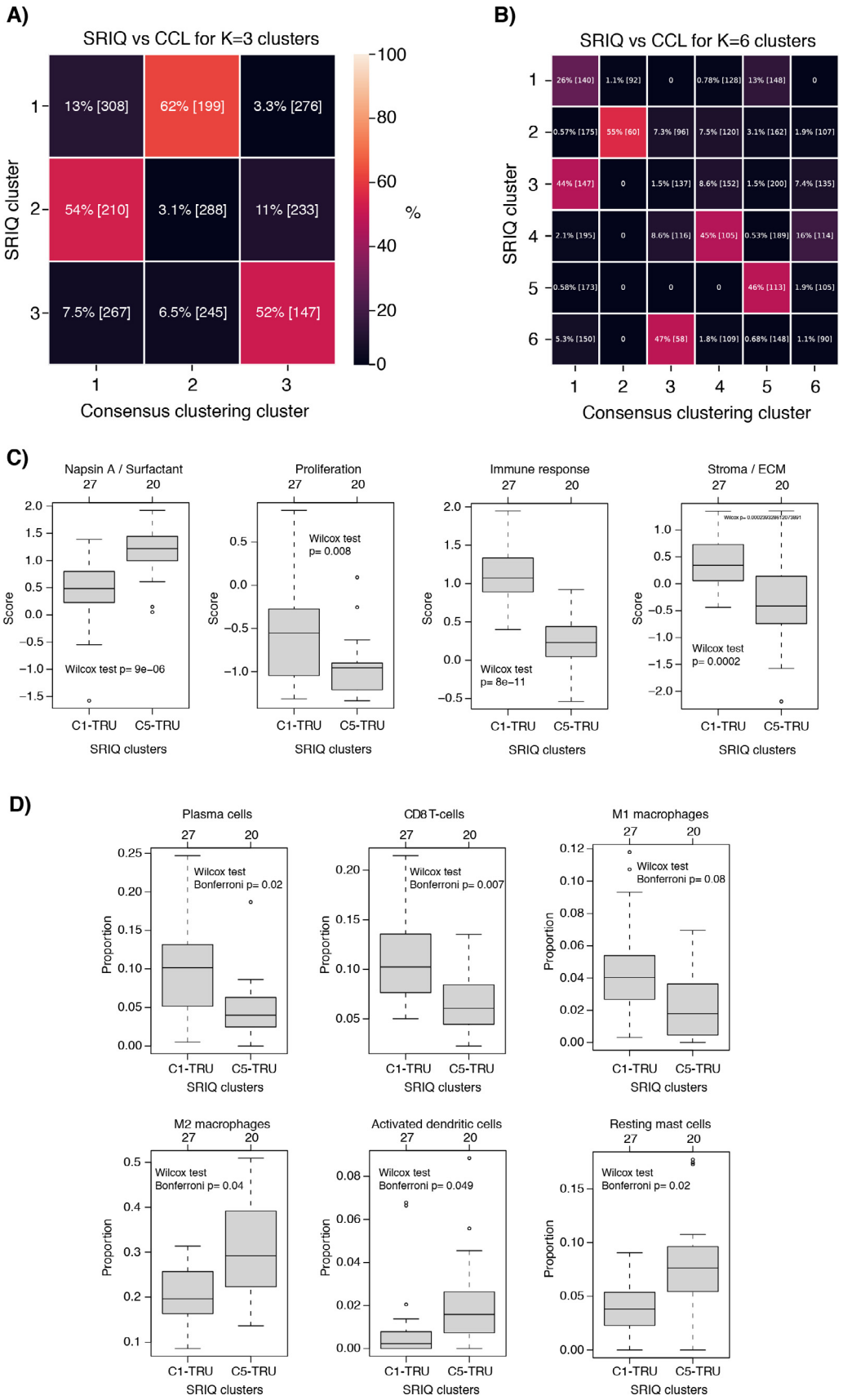


Fig. 3. Characterization of SRIQ K6 solution with respect to TCGA gene expression subtypes and expression of biological metagenes. (**A**) UMAP expression analysis of SRIQ K6 core samples using a bag size of 1200 and a cluster diameter of 0.63 as cut-off in SRIQ for cluster definition. Only the two first UMAP components are shown. (**B**) TCGA gene expression subtype (TRU/PI/PP) distribution in SRIQ K6 core cluster samples (top panel) and spiral samples (lower panel). (**C**) TCGA gene expression subtype (TRU:left, PI:center, PP:right) Pearson centroid correlation values for SRIQ K6 core cluster samples (blue) and spiral samples (orange). (**D**) Expression scores of six biological metagenes versus SRIQ K6 core (blue) and spiral samples (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)


A)
SRIQ vs CCL for K=3 clusters
13% [308]
62% [199]
3.3% [276]
54% [210]
3.1% [288]
11% [233]
7.5% [267]
6.5% [245]
52% [147]
SRIQ cluster
Consensus clustering cluster
%
B)
SRIQ vs CCL for K=6 clusters
SRIQ cluster
Consensus clustering cluster
C)
Napsin A / Surfactant
Wilcox test p= 9e−06
Proliferation
Wilcox test p= 0.008
Immune response
Wilcox test p= 8e−11
Stroma / ECM
Wilcox test p= 0.0002
Score
C1-TRU
C5-TRU
SRIQ clusters
D)
Plasma cells
Wilcox test Bonferroni p= 0.02
CD8 T-cells
Wilcox test Bonferroni p= 0.007
M1 macrophages
Wilcox test Bonferroni p= 0.08
M2 macrophages
Wilcox test Bonferroni p= 0.04
Activated dendritic cells
Wilcox test Bonferroni p= 0.049
Resting mast cells
Wilcox test Bonferroni p= 0.02
Proportion

3.2. Molecular investigation of SRIQ K6 cluster solution

3.2.1. Characterization of SRIQ K6 solution versus molecular subtypes and biological metagenes

To analyze the biological relevance of SRIQ results we focused on LUAD cases classified as core samples by the K6 solution with a cut-off of 0.63, originally shown in Fig. 2A and further explored in Fig. 3A-D.

Briefly, for the SRIQ K6 solution, cluster 1 was defined by TRU and PI classes, high immune response expression, medium number of *TP53* mutations and >80% females. Cluster 2 was characterized by PP and TRU, high expression of the Napsin A surfactant metagene, low expression of the proliferation, immune response, and stroma/ECM metagenes, and a high frequency of *KRAS* and *STK11* mutations (52% and 55%, respectively). Characteristics of cluster 3 core samples were high expression of the stroma/ECM, basal/squamous, and immune response metagenes, 90% PI classified samples, and a high number of *TP53* mutations (64%). Cluster 4 consisted of 96% PP samples with high expression of the proliferation metagene, low expression of the Napsin A/surfactant metagene, and a high *TP53* mutation rate (85%). Cluster 5 consisted solely of TRU classified samples with high expression of the Napsin A/surfactant metagene, low expression of the proliferation metagene, and a proportionally high fraction of all *EGFR* mutated cases (41% of all *EGFR* mutations in core samples). Cluster 6 consisted mainly of PP classified samples with high mutational burden of *KRAS* and *STK11* (59% and 45%, respectively) and mixed expression of several of the biological metagenes.

3.3. SRIQ clustering refines the TRU molecular phenotype

As shown in Fig. 3B, SRIQ differentiates the TCGA molecular subtypes into different clusters. E.g., SRIQ cluster 5 was primarily composed of TRU classified cases, irrespective of whether core (100% TRU, representing 34% of all core samples classified as TRU) or spiral samples were considered (89% TRU). In addition to cluster 5, SRIQ cluster 1 also harbored a high proportion of TRU cases (79% of core samples representing 46% of all core samples classified as TRU and 49% of spiral samples). To investigate the relevance of this stratification by SRIQ, we compared different molecular variables for SRIQ cluster 5 TRU core cases versus cluster 1 TRU core cases. Based on proposed biological metagenes, capturing broad transcriptional tumor intrinsic and microenvironment patterns, significant differences in expression involved both likely tumor related features like expression of Napsin A/surfactant (higher in cluster 5 vs cluster 1) and proliferation (lower in cluster 5 vs cluster 1), but also microenvironment features like immune response and stroma/ECM (lower in cluster 5 vs cluster 1) (Fig. 4C). Substantiating a difference in tumor microenvironment, cluster 5 TRU cases had significantly higher tumor purity on average than cluster 1 TRU cases (Wilcoxon's test p = 0.01). Other interesting features showing trend-like differences, albeit non-significant, was a trend of higher tumor ploidy in cluster 5 cases vs cluster 1 (Wilcoxon's test p = 0.09), while a lower exposure of the mutational signature related to smoking (signature 4 [36]) was observed in cluster 5 (median exposure = 26%) compared to cluster 1 (median exposure = 32%) cases (Wilcoxon's test p = 0.35). Together, this suggests possible refinement of the TRU phenotype based on both tumor intrinsic and tumor microenvironment features.

3.3.1. SRIQ differentiates immune type infiltration within the TRU molecular subtype

The importance of the host immune response to malignant growth has become increasingly important in LUAD based on the clinical introduction of immune checkpoint inhibitors. While the original TCGA molecular subtypes have been proposed to align to some extent with differences in immune response, illustrated by the renaming of the original squamoid subtype [19] to proximal inflammatory (PI) [22], large within subtype heterogeneity still exists. As illustrated in Fig. 4C, SRIQ clustering separates TRU classified core cases into one cluster (cluster 1) with high expression of immune response associated genes and one with lower expression (cluster 5). To explore whether the difference in immune response could be due to different infiltrating immune cell types, we compared 22 different immune cell type fractions obtained by Cibersort analysis of RNA-seq data (obtained from [28]) between the core sample groups. As shown in Fig. 4D, statistical differences, after multiple testing adjustment, between the groups were observed for several cell types, including plasma cells, CD8 T-cells, M1 macrophages (borderline non-significant), M2 macrophages, activated dendritic cells, and resting mast cells, further supporting the SRIQ stratification of TRU classified cases.

3.3.2. Differential gene expression and GO enrichment between clusters

To more in detail identify transcriptional programs characterizing the SRIQ clusters we performed differential gene expression of the core clusters through a "one versus rest" approach using SAM analysis. Differentially expressed genes are illustrated in Fig. 5A, listed in Supplementary Table 3, and enriched GO terms based on gene ontology analysis are listed in Supplementary Table 4 and illustrated in Fig. 5B-C for up- and down-regulated genes, respectively. Briefly, similar to the biological metagene expression patterns these results illustrate the presence of biological processes likely related to both tumor intrinsic factors and differences in the composition of the tumor microenvironment captured by SRIQ. For instance, in the high proliferative core cluster 4 enriched GO terms for up-regulated genes included DNA replication, mitotic spindle and cell cycle terms likely associated with higher intrinsic tumor proliferation, while immune response associated GO terms were enriched in down-regulated genes for this cluster, consistent with the immune cold phenotype suggested by the immune response metagene expression (likely representative of a low lymphocyte infiltrative tumor microenvironment). In opposite, in the immune response high cluster 1 immune response associated GO terms were enriched among up-regulated genes while DNA replication and spindle associated GO terms were enriched in down-regulated genes.

Fig. 4. Comparison of SRIQ to consensus clustering, and refinement of TCGA LUAD molecular subtypes. (**A**) Comparison of the SRIQ K3 (irrespective of core/spiral assignment) and consensus cluster 3 group solutions. First number in each tile is the percentage of common samples and the number within brackets the sum of samples present in both clusters. Consensus clustering was performed using hierarchical clustering, ward's linkage, Pearson correlation, 2000 iterations and 70% resampling for both items (samples) and features (genes) as key parameters, whereas SRIQ was run as described in Fig. 2A. (**B**) Comparison of the SRIQ K6 (irrespective of core/spiral assignment) and consensus cluster 6 group solutions. Consensus clustering was performed as above, whereas SRIQ analysis was performed as described in Fig. 2A. **(C)** Biological metagene expression is significantly different between core TRU-classified samples stratified by SRIQ cluster 1 (C1) and 5 (C5). (**D**) Immune cell types from Cibersort analysis showing statistical difference between core TRU-classified samples stratified by SRIQ cluster 1 (C1) and 5 (C5). P-values calculated using Wilcoxon's test with Bonferroni adjustment (for testing of 22 Cibersort cell types in total).

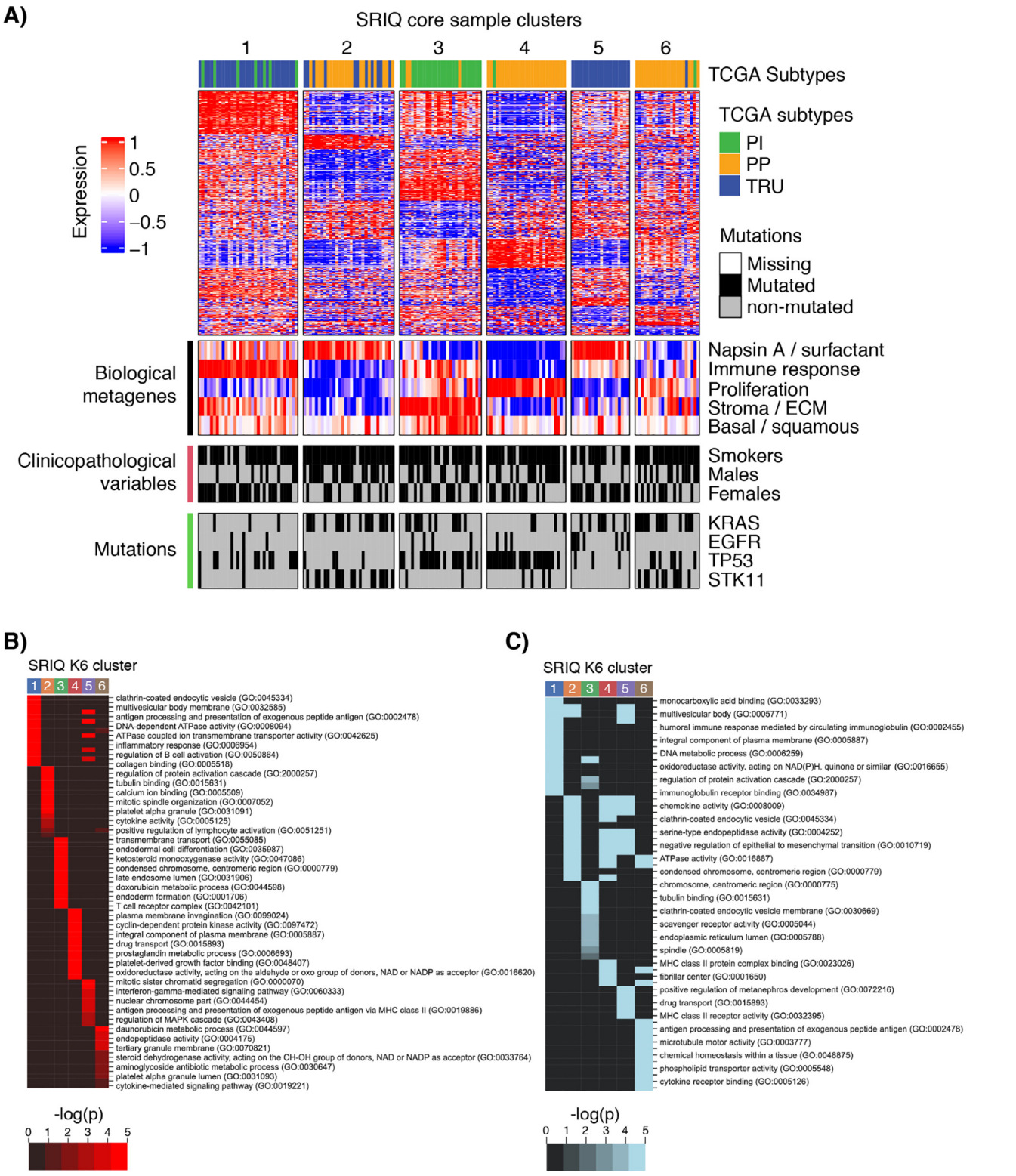


Fig. 5. Differential gene expression patterns and functional analysis for SRIQ K6 core samples. (**A**) Gene expression heatmap of the significantly expressed genes (n = 2417) based on SAM analysis with q-value <= 5 and a fold-change >= 2 for 162 core samples in total for each SRIQ cluster in a one versus rest approach. Annotation bars present biological metagene expression patterns, selected clinical information, and mutations (all variant types) for *KRAS*, *EGFR*, *TP53*, and *STK11*. SRIQ cluster solution was based on bag size 1200 and diameter cut-off 0.63 as shown in Fig. 2A. (**B**) Heatmap of up-regulated GO-terms for significant genes ordered by cluster, with the color representing statistical significance (−log(p)). Only a subset of GO-terms is listed in text, full details are shown in the corresponding Supplementary Table. (**C**) Heatmap of down-regulated GO-terms for significant genes ordered by cluster with the color representing statistical significance (−log(p)). Only a subset of GO-terms is listed in text, full details are shown in the corresponding Supplementary Table.

3.3.3. SRIQ refinement of the proximal-proliferative (PP) TCGA subtype

As seen in the heatmap in Fig. 5A, SRIQ clustering besides stratification of the TRU TCGA subtype also separates the PP TCGA subtype into three main nearly equally large subsets (core clusters 2, 4, and 6). Similar to the TRU case, this division appears associated with both tumor intrinsic properties and tumor microenvironment features likely representing readouts of the former (Supplementary Fig. 3). Briefly, PP cases in cluster 2 appear low proliferative but Napsin A/surfactant positive, driven by often concomitant *KRAS* and *STK11* alterations (e.g. 70% of *KRAS* mutated cases also had *STK11* mutations) but not *TP53* mutations. In contrast, PP cases in cluster 4 appear as the opposite to cluster 2, with infrequent *KRAS* and *STK11* alterations, 84% *TP53* alterations, very low Napsin A/surfactant expression, high tumor proliferation, and an immune cold phenotype. Finally, cluster 6 PP cases appear as the intermediate group between cluster 2 and 4, with frequent *KRAS* and *STK11* alterations, intermediate *TP53* mutations, and a more heterogeneous expression of the biological metagenes.

3.3.4. Transcriptional features of clusters different between SRIQ and consensus clustering

To deepen our comparison between SRIQ and consensus clustering we analyzed the observed differences from Fig. 4B regarding the division of consensus cluster 1 samples into SRIQ cluster 1 and 3. Firstly, when considering the global gene expression pattern UMAP analysis clearly demonstrated the general separation of SRIQ cluster 1 and 3 core samples (Fig. 3A, irrespective of consensus cluster status). Moreover, concerning expression of the biological metagenes, SRIQ cluster 1 and 3 samples (core and spiral) have clearly different patterns for several biological metagenes, including Napsin A/surfactant, immune response, stroma/ECM and proliferation, as well as clearly differentially expressed genes and different proportions of TRU/PP/PI gene expression subtypes as previously shown in Fig. 3B-D and 5. Finally, unsupervised hierarchical clustering of consensus cluster 1 samples specifically showed the heterogeneity of this cluster (Supplementary Fig. 4A), further illustrated by UMAP analysis of the same samples (Supplementary Fig. 4B). Together, these results substantiate that SRIQ in this case appears to stratify samples more appropriately from a molecular standpoint.

4. Discussion

In the current study we report an algorithm, SRIQ, that is a fusion of concepts from machine learning methods (Random Forest and QT- as well as KNN clustering) for unsupervised analysis of gene expression data. Based on technical reproducibility and benchmark analyses combined with deeper molecular correlations of derived clusters in bulk LUAD RNA-seq data, we demonstrate how the method can be applied and show that SRIQ-clustering captures biologically coherent variability in previously reported transcriptional phenotypes.

The reproducibility analyses presented in Fig. 2 demonstrate the importance of selecting a high enough bag size value for SRIQ (Fig. 2B). This is not surprising considering that using smaller feature sets (bag sizes) cause cluster instability in clustering methods due to selection bias, an issue not limited to SRIQ. Moreover, we show that SRIQ cluster identification, up to a point, is robust to reduced dataset sizes (Fig. 2C-D). As a rule in cluster analyses smaller subgroups, can only be detected by a method if their "core samples" are present in sufficient numbers. This represents a universal issue and is therefore not a specific limitation of SRIQ. Importantly, the analyses of SRIQ cluster stability and the benchmarking versus consensus clustering presented in Figs. 2 and 4 may serve as useful examples of how SRIQ users can explore their own data to assure optimal/robust results.

In contrast to conventional clustering algorithms, SRIQ users have an option to investigate both tight clusters, "core clusters", which may be the core subtype representatives of the cohort, and the expanded, "spiral clusters", which constitute samples that are neighbors to core clusters or samples representing admixtures of different cell types or subgroups. Reassuringly, we show that core and spiral cluster samples in the TCGA LUAD dataset had similar broad transcriptional profiles representing mixtures of larger transcriptional programs related to both tumor intrinsic properties (e.g., expression of Napsin A/surfactant genes associated with certain subtypes of LUAD as well as proliferation-related genes) and tumor microenvironment features such as the level of immune and stromal cell infiltration (e.g. expression of immune response and stroma/ECM metagenes) (Fig. 3C-D). What we believe is a second important feature of SRIQ is the possibility of producing cluster solutions at different diameter values in one run, where each solution is analyzed also for performance, accuracy and stability. Thus, a user can in a more straightforward way decide which cluster solution is appropriate to go ahead with in downstream analyses and provide a clear motivation for this choice. Finally, an important conclusion from both the technical benchmarking and the expanded molecular investigation of SRIQ results versus the established consensus clustering algorithm is that while both methods define common patient subsets, SRIQ clustering appears capable of capturing additional variability, allowing for more biologically refined clustering.

An important aim of this study was to go beyond technical validation and algorithmic benchmarking by analyzing derived cluster solutions for biological relevance and comparison with reported gene expression phenotypes in order to provide support for the usefulness of SRIQ as a method. Here, we believe the presented SRIQ results strongly support that further refinement of proposed TCGA subtypes in LUAD is feasible. Importantly, current LUAD subtypes are based on unsupervised analysis (using consensus clustering in the original study [19]) of bulk tumor gene expression data [19,22]. This implies that subtypes reflect tumor specific transcriptional patterns as well as the composition of the non-malignant microenvironment. In this context, SRIQ was able to delineate both the TRU and PP TCGA subtypes into subgroups reflecting different tumor microenvironments, as evidenced by, e.g., expression patterns of the biological metagenes used in this study (Figs. 3–5). With the introduction of immune checkpoint inhibitors in lung cancer the role of the microenvironment and immune response is now acknowledged as an important factor to consider for both treatment prediction and prognostication [37]. Thus, further refinement of LUAD gene expression subtypes may help in determining their future clinical significance/usefulness with respect to current therapy options and patient outcome.

Similar to all algorithms in the field, certain limitations of SRIQ are apparent. Here, we believe the presented SRIQ reproducibility analyses may serve as an example for independent users on how to approach the matter in their own data. Cluster solutions primarily depend on two hyperparameters: i) the minimum number of samples in clusters, and ii) the cluster diameter. As discussed above, SRIQ cannot circumvent the general problem of smaller datasets limiting the possibility to detect small or underrepresented biological subgroups (clusters). Regarding cluster diameter, cluster solutions with lower diameter values generate tight/stable clusters, whereas those with greater diameter values produce more relaxed clusters. Importantly, the diameter needs to be considered in the context of the stability and occurrence of a solution across

iterations. Another important SRIQ parameter is the bag size. Based on our analyses it is evident that not choosing a very small value should in general be sufficient. The combination of a large bag size, high number of permutations, and large datasets (samples and features) can make running SRIQ computation and memory intensive. Despite taking advantage of the symmetrical nature of the pairwise similarity matrix and the parallelism by Java programming, more optimized computing strategies could be employed to facilitate the application of SRIQ to, e.g., single cell sequencing and epigenomic data. Future improvements to the provided SRIQ pipeline could include additional data pre-processing procedures for different data types, additional distance metric methods beside Pearson and Euclidean distance, additional metrics for cluster stability to facilitate cluster solution selection, and flexibility to add metadata to the clusters to, e.g., facilitate the evaluation of cluster solutions.

In summary, we have demonstrated that SRIQ is an unsupervised analysis method for gene expression data that can circumvent problematic issues in unsupervised clustering. In the case of LUAD, it can be used to provide strong support for refinement of previously proposed transcriptional subtypes based on improved separation of subtypes by better capturing tumor intrinsic and microenvironmental transcriptional patterns. As such, we believe that SRIQ can become a valuable new tool for translational cancer research and may in the future also apply to other types of high-dimensional omics data beyond gene expression.

CRediT authorship contribution statement

Jacob Karlström: Visualization, Methodology, Formal analysis, Software, Data curation, Investigation, Writing – original draft. **Mattias Aine:** Data curation, Writing – original draft. **Johan Staaf:** Formal analysis, Investigation, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition, Project administration, Conceptualization. **Srinivas Veerla:** Conceptualization, Formal analysis, Visualization, Data curation, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing.

Acknowledgements

The authors would like to acknowledge Dr Pontus Eriksson at the Division of Oncology, Lund University, Sweden for constructive methodological input.

Funding

This work was supported by the Swedish Cancer Society, the Mrs Berta Kamprad Foundation, Sweden, The Swedish Research Council, and The National Health Services (Region Skåne/ALF).

Conflict of interests statement

Authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.03.036.

References

[1] World Health Organization (WHO) http://www.who.int. Accessed Nov 26 2021.
[2] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.
[3] Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, et al. A review of clustering techniques and developments. Neurocomputing 2017;267:664–81.
[4] Kaufman L, Rousseeuw P. Partitioning around medoids (Program PAM). Wiley Series in Probability and Statistics. Wiley; 1990.
[5] Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Softw 2014;61:1–36.
[6] Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. Genome Res 1999;9:1106–15.
[7] Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks. Science 2014;344:1492–6.
[8] A'lZoubi WA. A survey of clustering algorithms in association rules mining. Int J Comput Sci Inf Technol 2019;11:17–25.
[9] Kriegel H-P, Kröger P, Zimek A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans Knowl Discov Data 2009;3. Article 1.
[10] Zimek A, Schubert E, Kriegel H-P. A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Anal Data Mining 2012;5:363–87.
[11] Reunanen N, Räty T, Lintonen T. Automatic optimization of outlier detection ensembles using a limited number of outlier examples. Int J Data Sci Anal 2020;10:377–94.
[12] Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 2011;27:1986–94.
[13] Kim Y, Kim T-H, Ergün T. The instability of the Pearson correlation coefficient in the presence of coincidental outliers. Finance Res Lett 2015;13:243–57.
[14] Breiman L. Random forests. Machine Learn 2001;45:5–32.
[15] Breiman L. Bagging predictors. Machine Learn 1996;24:123–40.
[16] Petersen ML, Molinaro AM, Sinisi SE, van der Laan MJ. Cross-validated bagged learning. J Multivariate Anal 2007;98:1693–704.
[17] Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. J Thorac Oncol 2015;10:1243–60.
[18] Crino L, Weder W, van Meerbeeck J, Felip E. Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2010;21(Suppl 5):v103–115.
[19] Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS ONE 2012;7:e36530.
[20] Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
[21] Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. J Clin Oncol 2006;24:5079–90.
[22] Cancer Genome Atlas Research N: Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014, 511:543-550.
[23] Planck M, Edlund K, Botling J, Micke P, Isaksson S, Staaf J. Genomic and transcriptional alterations in lung adenocarcinoma in relation to EGFR and KRAS mutation status. PLoS ONE 2013;8:e78614.
[24] Ringner M, Jonsson G, Staaf J. Prognostic and Chemotherapy Predictive Value of Gene-Expression Phenotypes in Primary Lung Adenocarcinoma. Clin Cancer Res 2016;22:218–29.
[25] Ringner M, Staaf J. Consensus of gene expression phenotypes and prognostic risk predictors in primary lung adenocarcinoma. Oncotarget 2016;7:52957–73.
[26] The Cancer Genome Atlas http://cancergenome.nih.gov/. Accessed Nov 26 2021.
[27] Dama E, Melocchi V, Dezi F, Pirroni S, Carletti RM, Brambilla D, et al. An aggressive subtype of stage I lung adenocarcinoma with molecular and prognostic characteristics typical of advanced lung cancers. Clin Cancer Res 2017;23:62–72.
[28] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The immune landscape of cancer. Immunity 2019;51:411–2.
[29] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7.
[30] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–21.
[31] Karlsson A, Jonsson M, Lauss M, Brunnstrom H, Jonsson P, Borg A, et al. Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. Clin Cancer Res 2014;20:6127–40.
[32] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf 2013;14:128.
[33] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90–97.
[34] Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene set knowledge discovery with enrichr. Curr Protoc 2021;1:e90.

[35] Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LDF, et al. Clustering algorithms: a comparative approach. PLoS ONE 2019;14:e0210236.
[36] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415–21.
[37] Altorki NK, Markowitz GJ, Gao D, Port JL, Saxena A, Stiles B, et al. The lung microenvironment: an important regulator of tumour growth and metastasis. Nat Rev Cancer 2019;19:9–31.