OXFORD

Briefings in Bioinformatics, 2022, **23(4)**, 1–15
https://doi.org/10.1093/bib/bbac253
Advance access publication date: 27 June 2022
Problem Solving Protocol

ConSIG: consistent discovery of molecular signature from OMIC data

Fengcheng LI, Jiayi YIN, Mingkun LU, Qingxia YANG, Zhenyu ZENG, Bing ZHANG, Zhaorong LI, Yunqing QIU, Haibin DAI, Yuzong CHEN and Feng ZHU

Corresponding authors: Prof. Feng ZHU, College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China; E-mail: zhufeng@zju.edu.cn; Prof. Yuzong CHEN, Graduate School at Shenzhen, Tsinghua University, Shenzhen; Email: chenyuzong@sz.tsinghua.edu.cn

Abstract

The discovery of proper molecular signature from OMIC data is indispensable for determining biological state, physiological condition, disease etiology, and therapeutic response. However, the identified signature is reported to be highly inconsistent, and there is little overlap among the signatures identified from different biological datasets. Such inconsistency raises doubts about the reliability of reported signatures and significantly hampers its biological and clinical applications. Herein, an online tool, ConSIG, was constructed to realize consistent discovery of gene/protein signature from any uploaded transcriptomic/proteomic data. This tool is unique in ***a***) integrating a novel strategy capable of significantly enhancing the consistency of signature discovery, ***b***) determining the optimal signature by collective assessment, and ***c***) confirming the biological relevance by enriching the disease/gene ontology. With the increasingly accumulated concerns about signature consistency and biological relevance, this online tool is expected to be used as an essential complement to other existing tools for OMIC-based signature discovery. ConSIG is freely accessible to all users without login requirement at https://idrblab.org/consig/

Keywords: Omics, Signature discovery, Feature selection, Signature consistency, Classification accuracy

Introduction

Discovery of proper molecular signatures from OMIC data is indispensable in the determination of the biological state, physiological condition, disease mechanism, and therapeutic response [1–6]. However, the identified signature is found to be highly 'inconsistent' [7–9]. Particularly, there is little overlap among the signatures identified from different datasets of certain biological study [10, 11]. Such 'inconsistency' issue raises doubts about the reliability of the reported signatures, and greatly hampers their applications in biological sciences and clinical investigation [12–15].

To address this issue, a novel feature selection strategy was proposed and successfully validated using the transcriptomic data of *schizophrenia* patients [16]. As shown in Figure 1*a*, this strategy was constructed by: ***α***) integrating the repeated random sampling with consensus scoring, and ***β***) evaluating ranking consistency among multiple datasets [16, 17], which can maximally avoid the erroneous elimination of the molecular features [17]. Because of its ability to effectively enhance signature's consistency, this strategy and its underlying theory have attracted broad interest from and been adopted by diverse communities, including biochemical method [18, 19], plant sciences [20, 21], molecular biology [22, 23], pharmaceutical sciences [24, 25], genetics [26, 27], *etc.*

So far, various tools have been available online to conduct OMIC-based signature discovery [28–34]. Some of them utilize classic univariate strategies for feature selection, such as GEPIA2 [28], ImaGEO [29] and MAINE [30]. Some others integrate the multivariate ones for eliminating non-significant molecular features, including MetaboAnalyst [31], CausalMGM [32], OmicsAnalyst [33], NOREVA [34, 35], *etc.* These tools are found to be popular in their research communities, and the 'clas-

Fengcheng LI, Jiayi YIN, Mingkun LU, Qingxia YANG are PhD/MD candidates of College of Pharmaceutical Sciences in Zhejiang University, China. They are interested in bioinformatics.
Zhenyu ZENG, Zhaorong LI, Bing ZHANG are senior experts on big data/AI product and solution in Alibaba Cloud. Their current focus includes Digital Health, Health Economics and Knowledge Graph.
Yunqing QIU is professor of the First Affiliated Hospital in Zhejiang University, China. He specializes in precision medicine, diagnosis and treatment of liver disease and system biology.
Haibin DAI is professor of the Second Affiliated Hospital in Zhejiang University, China. He specializes in the precision medicine of ischemic brain damage and stroke accompanied by diabetes.
Yuzong CHEN is professor of the State Key Lab of Chemical Oncogenomics in Tsinghua University, China. His research group (http://bidd.group/) has been working in the fields of AI-based drug design.
Feng ZHU is a tenured professor of College of Pharmaceutical Sciences in Zhejiang University, China. His research group (https://idrblab.org/) has been working in the fields of OMIC-based drug discovery.
Received: April 5, 2022. **Revised:** May 9, 2022. **Accepted:** May 31, 2022

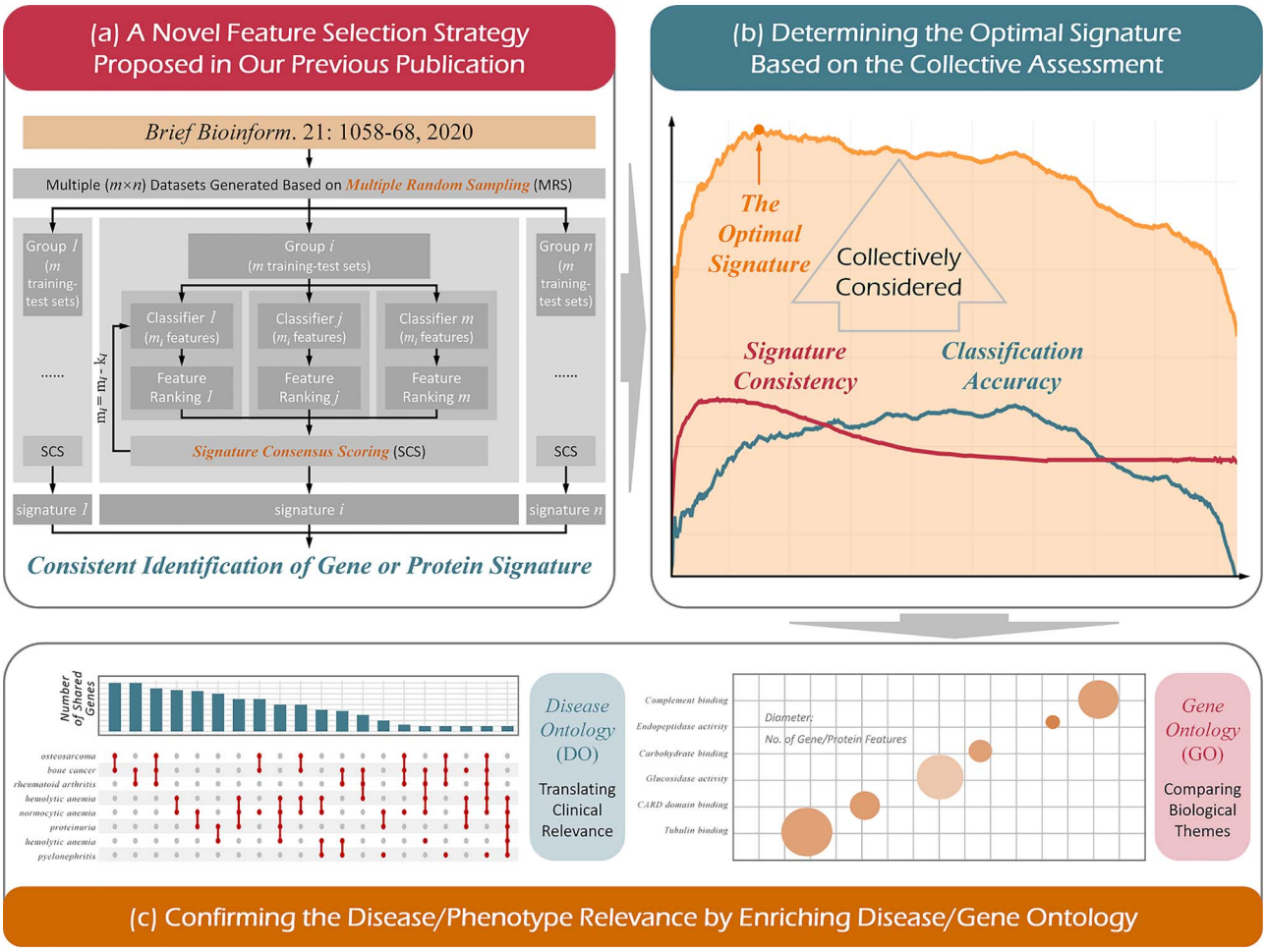

Figure 1. Schematic illustration of the unique functions provided by ConSIG. (***a***) the integration of novel feature selection strategy to substantially enhance the consistency of signature discovery; (***b***) the determination of the optimal signature by collectively assessing *signature consistency* & *classification accuracy*; (***c***) the confirmation of disease or phenotype relevance using enrichment analyses that are based on disease (DO) or gene (GO) ontology.

sification accuracy' is frequently adopted as their primary criterion for model assessment [36–39]. However, the existing tools do not integrate any strategy that can enhance the robustness of signature discovery, and they do not provide any assessment of 'signature consistency', which should be collectively evaluated together with 'classification accuracy' [40, 41]. Moreover, with increasingly accumulated concerns about the biological relevance [42–45] of identified signature, it is important to have a tool that describes the relevance between the identified signature and the studied phenotype [46, 47]. In other words, it is urgently needed to have a tool that provides these key functions to facilitate the OMIC-based signature discovery [40–42].

Herein, an online tool named ConSIG was thus constructed to realize the consistent identification of gene/protein signature from any uploaded transcriptomic/proteomic data. As shown in Figure 1, this online tool works by ***a***) integrating the novel strategy proposed in our previous publication [16] to effectively guarantee the consistent identification of molecular signatures, ***b***) determining the optimal signature by collectively evaluating signature consistency and classification accuracy, and **c**) confirming biological relevance by enriching disease/gene ontologies [42, 46]. To the best of our knowledge, ConSIG is unique in realizing the applications of our novel strategy (proposed in our previous work [16]) to any genomic/transcriptomic/proteomic studies relevant to signature discovery. With the increasingly accumulated concern about both consistency [14, 48] and biological relevance [42, 49] in signature identification, this new online tool is expected to be popular and used as an essential complement to other existing tools for OMIC-based signature discovery. ConSIG is now freely accessible without any login requirement at https://idrblab.org/consig/

Materials and methods

Collection of benchmark datasets for performance assessment

As described in Table 1, four proteomic/transcriptomic benchmarks were collected for analyzing the performances of the studied strategies according to the following criteria [50, 51]: (*a*) datasets should be comparative studies with only one control & case group;

Table 1. Four benchmark datasets collected for analyzing the performances of the studied strategies, which included two proteomic and another two transcriptomic datasets. The information of features, cases and controls were explicitly described, and relatively diverse experiment platforms were applied for data generation, such as triple time-of-flight mass spectrometer (MS), quadrupole orbitrap MS, and Affymetrix gene chip

Dataset Type	ID of Studied Datasets	Experimental Platform of Studied Datasets	No. of Features	Description on the Samples of Studied Datasets
Proteomic Benchmarks	IPX0001256000 [53]	Liquid Chromatography Triple TOF 5600 Mass Spectrometer	1103 Proteins	21 urine samples from female donors 28 urine samples from male donors
	PXD006129 [54]	Q Exactive Hybrid Quadrupole Orbitrap Mass Spectrometer	3243 Proteins	15 samples from western-style diet-fed mice 14 samples from chow diet-fed mice
Transcriptomic Benchmarks	GSE31192 [57]	Affymetrix Gene Chip Human Genome U133 Plus 2.0	12,128 RNAs	20 samples from pregnancy associated breast cancer patients 13 samples from pregnancy non-associated breast cancer patients
	GSE23878 [58]	Affymetrix Gene Chip Human Genome U133 Plus 2.0	20,212 RNAs	35 samples from colorectal cancer patients 24 samples from healthy individuals

(*b*) datasets should be derived from different biological research directions; (*c*) the sample size for each group (control and case) in a study should be six at least and the sample size should be at least 20 for multiple random sampling [52]; (*d*) feature names of the dataset should be genes/proteins with identifiable UniProt ID or ENTREZID for enrichment analysis. Particularly, two proteomic datasets (IPX0001256000 [53] and PXD006129 [54]) were collected from the latest version of iProX [55] and PRIDE [56], respectively. IPX0001256000 described the expression levels of 1103 proteins among 49 urine samples (21 samples from female donors & 28 samples from male donors), and the PXD006129 provided the concentrations of 3243 proteins among 29 samples (15 samples from western-style diet-fed mice & 14 samples from chow diet-fed ones). The remaining two transcriptomic datasets provided in Table 1 (GSE31192 [57] and GSE23878 [58]) were collected from Gene Expression Omnibus [59]. GSE31192 showed the expression profiles of 12,128 RNAs among 33 individuals (20 from the patients with pregnancy-associated breast cancer & 13 from the patients with non-pregnancy-associated breast cancer), and GSE23878 demonstrated the intensity levels of 20,212 RNAs among 59 samples (35 colorectal cancer samples & 24 healthy individual samples). As described in Table 1, diverse experimental platforms were applied for generating these datasets [60, 61].

Classical feature selection strategies discussed in this study

As shown in Table 2, six classical feature selection strategies were included for analyses in this study, which could be divided into two types (univariate and multivariate) [62–64]. Three typical univariate strategies were discussed, such as fold change (FC), univariate *t*-test (*t*-test), and Wilcoxon rank-sum test (Wilcox) [65]. The FC is a widely used strategy reflecting variation between case and control groups through calculating the ratio of mean feature intensity between groups [66]. The *t*-test is a classical method based on the statistical measurement of *p*-value, which considers a feature as 'significant' when its corresponding *p*-value is less than 0.05 [41]. Wilcox is known as a powerful alternative to *t*-test, which utilizes the magnitude-based ranks to assess the significant differences between groups [67]. Moreover, three multivariate strategies analyzed in this study included correlation-based feature selection (CFS), partial least squares-discriminant analysis (PLS-DA), and ReliefF (REF) [68–70]. The CFS assesses the performance of a set of features based on the prediction ability of each feature in the set and the correlation between the feature set and prediction ability [71]. PLS-DA establishes the model to predict sample groups or discriminative variable selection, which predicts the features maximizing the difference between predetermined samples [72]. The REF is an individual evaluation method, which detects feature dependencies by estimating each feature based on the identification of differences between features and neighbors [73].

Metrics for signature consistency and classification accuracy

One of the key features of ConSIG lay in its unique function of identifying the optimal signature. To realize such function, two independent criteria of '*signature consistency*' and '*classification accuracy*' were collectively considered [16, 17] to assess the performance of any feature list, and two types of metrics were employed. Detailed descriptions were provided as follows.

Metrics employed for assessing signature consistency

Herein, the relative-weighted consistency (CW*rel*) was employed to assess *signature consistency* [74], which had been frequently adopted as a well-established metric for evaluating the robustness among various lists of identified features [75–77]. Particularly, CW*rel* is a metric calculated based on multiple feature lists [74]. It is represented by the ratio between the number of occurrences of each feature in each feature list and the total number of occurrences of all features in all lists [78–81].

Table 2. Six classical feature selection strategies included for analyses in this study. These strategies could be divided into two types (univariate and multivariate). Three typical univariate strategies analyzed in this study were: fold change (FC), univariate t-test (t-test), and Wilcoxon rank-sum test (Wilcox). Three typical multivariate strategies were: correlation-based feature selection (CFS), partial least squares-discriminant analysis (PLS-DA), and ReliefF (REF). Detailed descriptions of the mechanism underlying each studied feature selection strategy were also provided.

Strategy Type	Strategy Abbreviation	Full Name of Studied Strategies	The Description of the Mechanism Underlying Studied Strategies
Univariate Strategies	FC	Fold Change [66]	Calculate the FC of each feature by taking the ratio of mean intensities between the cases and controls, and rank the features according to the FC values of all features.
	t-test	Univariate *t*-test [41]	Rank all features based on their statistical differences measured by p-values, and consider a feature as 'significant' when its corresponding *p*-value is less than 0.05.
	Wilcox	Wilcoxon Rank-sum Test [67]	Apply the magnitude-based ranks to establish the significant difference between the case and control groups, which is a non-parametric version of univariate *t*-test.
Multivariate Strategies	CFS	Correlation-based Feature Selection [71]	Sort the feature subsets based on their predictive power and internal correlation, and consider the subset of strong predictive power and low correlation as significant.
	PLS-DA	Partial Least Squares Discriminant Analysis [72]	Establish a model to predict case and control groups, and perform a discriminative feature selection based on the predictive ability of the established model.
	REF	ReliefF [73]	Weight each feature based on how well it can distinguish between case and control groups, and choose the features that can be most distinguished between groups.

As reported, CW*rel* works much better than other metrics for consistency assessment, since it has a superior ability to deal with the problem of 'subset-size bias' [74]. The value of CW*rel* is between 0 and 1, where 1 indicates that all evaluated lists of identified features are identical.

Metrics employed for evaluating classification accuracy

Meanwhile, the area under the receiver operating characteristic (AUC) and *Matthews* correlation coefficient (MCC) was employed for evaluating the *classification accuracy* [82–86], which had been frequently adopted as the primary criterion for model evaluation [36–38]. Particularly, AUC assesses the diagnostic accuracy of a classifier constructed based on the identified signature [11, 83]. Compared with AUC, MCC is known as a more balanced metric, making it useful for unbalanced datasets [87–89]. In other words, AUC together with MCC was simultaneously provided for the assessment of *classification accuracy*. AUC takes values between 0 and 1, while MCC takes values between −1 and 1. For the two metrics, 1 indicates the best authenticity of the test (for AUC) and the perfect classification of all samples (for MCC) [82–85].

All in all, the optimal signature was identified by collectively considering *signature consistency* and *classification accuracy*. A larger CW*rel* value implies a more stable list of identified features and better *signature consistency*. A larger AUC/MCC denotes higher *classification accuracy* [90, 91]. As shown in Figure 1*b*, the feature list giving the largest CW*rel* value is not, for most cases, identical to that offering the highest AUC/MCC. Therefore, both metric types were collectively considered in ConSIG to identify the optimal molecular signature (as illustrated in Figure 1*b*).

Confirmation of biological relevance by enrichment analyses

With increasingly accumulated concerns about the biological relevance [42–44] of the identified signature, two types of enrichment analysis were enabled in ConSIG (as illustrated in Figure 1*c*) to discover the relationship between identified signature and a phenotype of interests [46, 92]. These two types of enrichment analysis were based on disease (DO) and gene (GO) ontologies.

DO annotates genes/proteins in the context of disease indication, which has been frequently used to translate molecular findings from the high-throughput data to clinical relevance [93–95]. ConSIG integrated the *enrichDO* function in an *R* package entitled DOSE [96] to enrich all genes/proteins in the identified signature. The *enrichDO* function is based on '*Disease Ontology*' database [93] and *TTD* database of our research group [62, 97, 98] that offer comprehensive annotation of human and non-human genes/proteins to >10,000 disease ontology terms. By integrating such databases into enrichment analyses [99–101], ConSIG is able to describe the disease relevance of the identified signature to the largest extent. Particularly, DO-based enrichment results in ConSIG were presented using the bubble map [102], the gene-concept network [102], and the upset plot [96].

GO annotates genes/proteins in the context of biological process, molecular function and cellular component by a directed acyclic graph structure, which has been widely adopted as popular way to search for shared functions among genes/proteins by incorporating biological knowledge from gene ontology [103]. ConSIG integrated the *enrichGO* function in the clusterProfiler package [102] of *R* environment to enrich all genes/proteins in the

identified signature. The *enrichGO* function is based on the well-known knowledge base of '*Gene Ontology*' [103] that gives a comprehensive annotation of gene/protein to >45,000 terms. In ConSIG, GO-based enrichment result was shown using the bubble map [102], the gene-concept network [102], and the enrichment map [104].

Implementation details and functional modules of online server

The official website of ConSIG (https://idrblab.org/consig/) is deployed on a server with 128 GB RAM, and CPU E7–4820 × 32 cores. This server runs a CentOS Linux 7.6 operating system, an Apache Tomcat servlet container, and an Apache HTTP web server 2.2.15. The web interface was constructed using *R* v3.6.2 and *R* package Shiny v1.7.1 running on the Shiny-server of v1.5.7.907. A variety of *R* packages were utilized in the background process, including VIM, e1071, ggplot2, DOSE, topGO, clusterProfiler, pathview, enrichplot, mixOmics and pROC [104–106]. ConSIG can be readily accessed by all users without login requirements using various web browsers, including Google Chrome, Mozilla Firefox, Safari, and Internet Explorer (10 or later).

ConSIG is able to handle datasets in common formats including txt, xlsx, tab-delimited and csv. The row names of the input files required by ConSIG should be sample ID and the column names should be gene/protein feature ID. The ID of samples and features should be unique and defined by user's preference, but if the enrichment analysis is required, the feature ID of the input file should be annotated with UniProt ID or ENTREZID. In particular, the second column indicates the category label (case or control) for each sample, with the caveat that the number of samples in each category should be no less than three.

Due to the cost of web connection and the shared nature of the computational resource, the web-based server was expected to be slower compared with the standalone application. To figure out the time cost of ConSIG, a proteomic benchmark PXD005144 [107] having 66 samples of patients with pancreatic cancer and 36 samples of patients with chronic pancreatitis was collected for assessment. According to our evaluation, it took about two minutes for ConSIG to finish the entire identification process (as illustrated in Figure 1, from *signature discovery* to *signature optimization*, then to *enrichment analysis*). Among each step of the entire process, *signature discovery* is the most time-consuming one, and a functional module was therefore deployed to the online tool for enabling the real-time monitoring of the identification progress (Figure 2). In other words, this module gives a dynamic estimation and real-time monitoring of the remaining time costs, which is valuable for users who upload a relatively large dataset. Moreover, it may take a much longer time when handling datasets with a huge number (>10,000) of gene/protein features, and a user-specific hyperlink (Figure 2) is therefore provided to enable later retrieval of the result once the process is finished.

Results and discussion

Consistent discovery of the molecular signature from OMIC data

To illustrate the levels of consistencies of feature selection strategies when discovering molecular signatures, four benchmarks (shown in Table 1) were analyzed. All these benchmarks were case–control studies, and six classical strategies (as described in Table 2) together with ConSIG were systematically assessed by identifying the differential signature using the benchmarks. As shown in the 'Metrics for Signature Consistency and Classification Accuracy' section of **Materials and Methods**, the metric of relative-weighted consistency (CW*rel*) was employed to assess signature consistency [74]. In particular, for each benchmark dataset, 20 sub-datasets were *first* generated via randomly selecting (for 20 times) half of the entire samples, which followed the standardized bootstrapping process enabling a random sampling with replacement [108, 109]. *Then*, all strategies were applied to 20 sub-datasets, and 20 lists of feature ranking were therefore identified for each strategy. *Third*, 20 groups of features ranked in the top 50% were selected from those 20 lists of feature ranking, and CW*rel* was used to assess the consistency among those 20 groups of features. *Finally*, this step above was repeated by another 49 times (from top 49%, to 48%, and finally to top 1%), which resulted in 50 CW*rel* values for each strategy and each benchmark. The parameters involved in the process of accessing consistency were determined as follows: (I) half of the entire samples were selected for generating the widest range of possible sample subsets; (II) randomly selecting half of all samples 20 times is a trade-off between computational efficiency and representativeness of consistency; (III) the top 50% percent from the ranked features were selected to encompass all the commonly used optimal feature subset selection situations. As illustrated in Figure 3, the levels of consistency of all strategies were provided, and the ConSIG (blue color) was found to produce the best consistency in all benchmarks regardless of the signature size (Top n%). For different datasets, the levels of consistency elevation varied greatly. Compared with the best-performing classical strategy, CW*rel* of ConSIG was elevated to 1.56 ~ 2.77 times, 1.10 ~ 1.54 times, 1.26 ~ 1.53 times, and 1.03 ~ 1.14 times for PXD006129, IPX0001256000, GSE31192, and GSE23878, respectively. This result showed the superior performance of ConSIG in consistently discovering the gene/protein signature from OMIC data.

For a more in-depth understanding of the consistency elevation achieved by ConSIG, the stability among the signatures identified from 20 sub-datasets was further evaluated. As demonstrated in Figure 4, the features identified from 20 sub-datasets randomly generated based on PXD006129 benchmark varied significantly. Taking the sub-figure in the upper right corner as an example, *x*-axis illustrated the number of sub-datasets that identified an identical feature (from 1 to 20), and

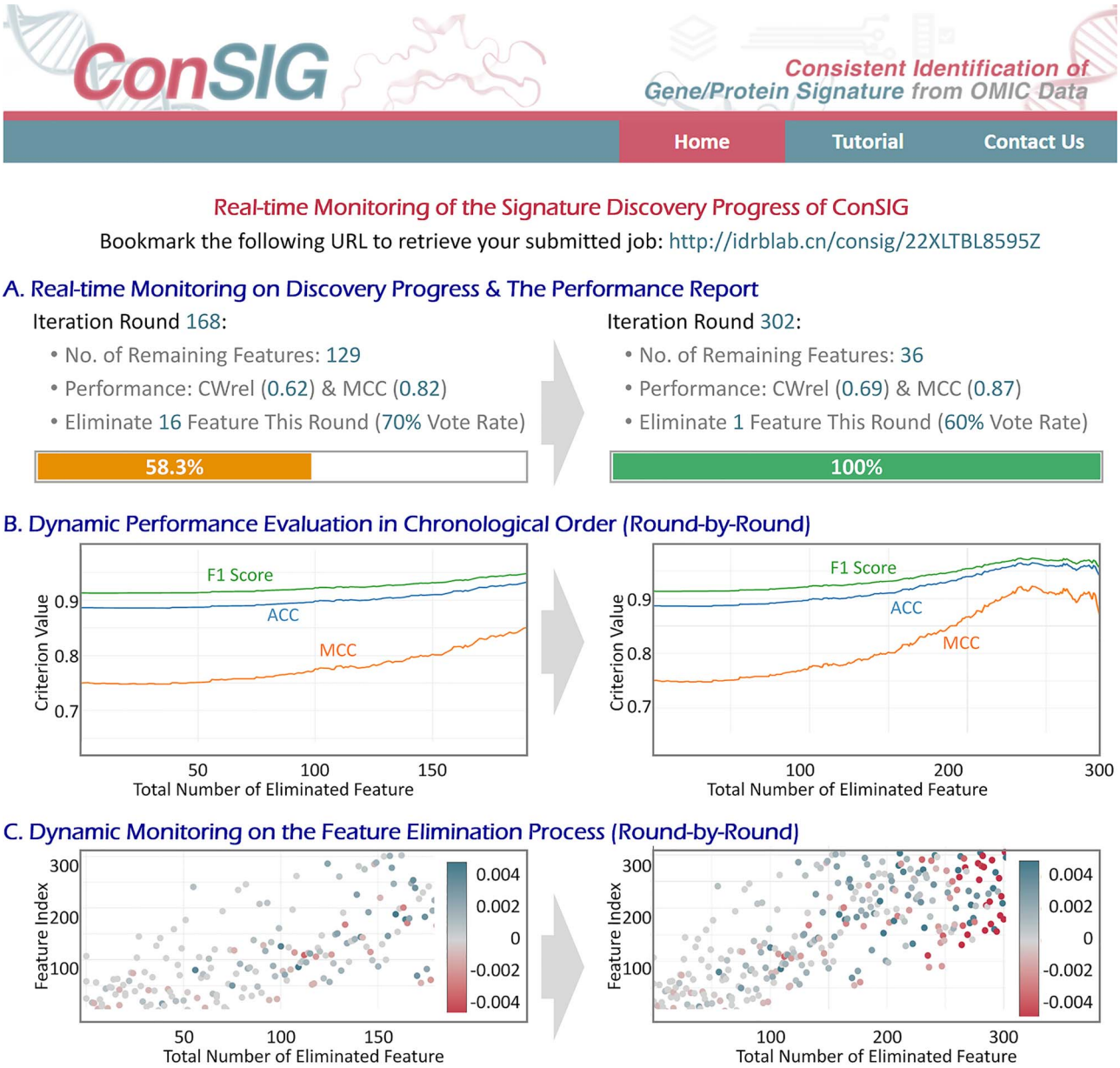

Figure 2. Screenshot of the functional module deployed to the online ConSIG for enabling real-time monitoring of the signature identification progress. This module gives a dynamic estimation and real-time monitoring of the remaining time costs, which is valuable for users who upload a large dataset. Moreover, it may take a much longer time when handling datasets with a huge number (>10,000) of gene/protein features, and a user-specific hyperlink is therefore provided to enable later retrieval of the result once the process is finished. The information displayed online included three sections: real-time monitoring of discovery progress and the performance report, dynamic performance evaluation in chronological order, and monitoring of the feature elimination process.

y-axis demonstrated the percentage of features among all features that were identified by 20 sub-datasets. As provided in this sub-figure, a large number (29.1%, the highest bar on the right-most side) of the identified features were simultaneously found by all 20 sub-datasets, which indicated a significant elevation of consistency compared with classical strategies (shown in the remaining 6 sub-figures on the right column of Figure 4, only a very small fraction (<3.0%) of the identified features could be simultaneously found by all sub-datasets).

Moreover, to have a systematical view of the signature consistency, a new metric named 'percent' was further calculated, which indicated the percentage of features simultaneously found by over half (>10) of all 20 sub-datasets. The larger the 'percent' value is, the more consistent the studied strategy is. As shown in Figure 4 (seven sub-figures on the right side), ConSIG showed superior consistency ('percent' = 60.1%) comparing with the classical ones ('percent' = 23.5 ~ 37.8%). In the meantime, the values of CW*rel* in all sub-figures could reach the same conclusion. That is to say, ConSIG showed superior consistency comparing with the classical ones, since the CW*rel* of ConSIG (0.72) was much higher than that of the classical ones (0.17 ~ 0.38).

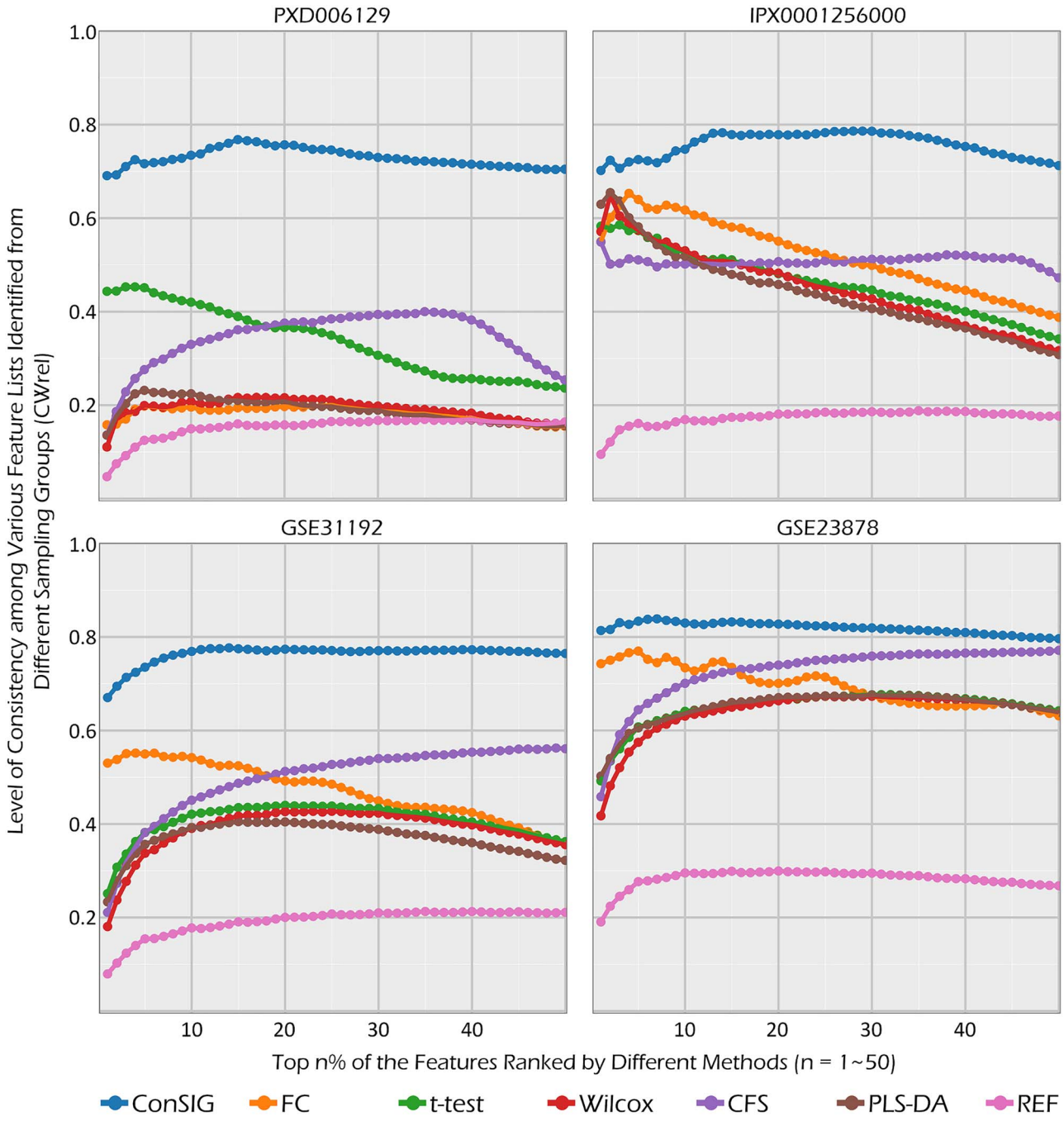

Figure 3. The consistency of seven studied strategies when discovering the molecular signature. For each benchmark, 20 sub-datasets were *first* generated by randomly selecting half of the entire samples. *Then*, all strategies were applied to 20 sub-datasets, and 20 lists of feature ranking were thus identified for each strategy. *Third*, 20 groups of features ranked in the top 50% were selected from those 20 lists of feature ranking, and CW*rel* was used to assess the consistency among those 20 groups of features. *Finally*, the above step was further repeated by another 49 times (from top 49%, to 48%, finally to top 1%), which led to 50 CW*rel* values for each strategy and each dataset. As described, ConSIG was found to produce the best consistency in all datasets regardless of the signature size (Top n%). In different datasets, the levels of consistency elevation varied greatly.

Figure 4 also illustrated the consistency variations induced by different signature sizes (Top 10%, 20%, 30%, and 40%). As well-known, the Top n% of the ranked features were frequently selected as markers in bioinformatics studies [110–112]. As shown in Figure 4, with the decrease of n%, the consistency of each strategy reduced gradually as assessed using both 'percent' and the height of the bar on the right-most side of each sub-figure. Moreover, CW*rel* of ConSIG (0.72 ~ 0.76) was significantly and robustly higher than that of the classical ones (0.15 ~ 0.42). Besides PXD006129, the features identified from 20 sub-datasets randomly generated based on the other three benchmarks (IPX0001256000, GSE31192, and GSE23878) were explicitly shown in **Supplementary Figure S1**, **S2**, and **S3**, respectively. All in all, despite the heterogeneity among 20 different sub-datasets, ConSIG was able to consistently discover the gene/protein signature from OMIC data.

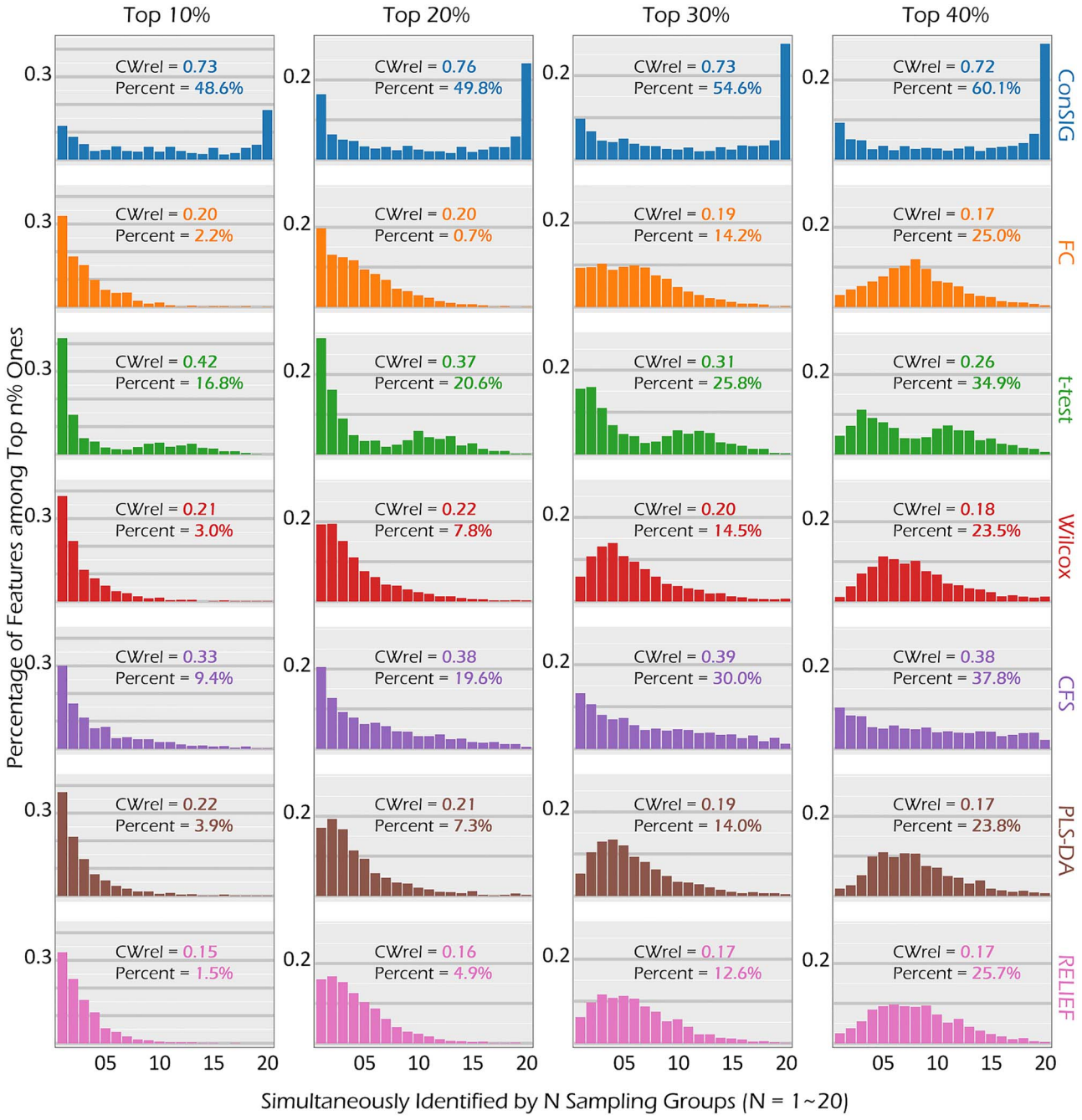

Figure 4. Consistency variation of seven feature selection strategies induced by various signature sizes (Top 10%, 20%, 30%, and 40%). The features identified based on 20 sub-datasets randomly generated based on PXD006129 varied greatly. The x-axis illustrated the number of sub-datasets that identified the identical feature (from 1 to 20), and the y-axis demonstrated the percentage of features among all features that were identified by 20 sub-datasets. To have the systematical view on signature consistency, the metric 'percent' was calculated, which indicated the percentage of features simultaneously found by over half (>10) of all those sub-datasets. The larger the 'percent' value is, the more consistent the studied strategy is.

Identification of the optimal signature using collective assessment

With the elevation of *signature consistency* by ConSIG, it is of great interest to further investigate how *classification accuracy* is affected. In other words, such elevation of consistency should not be accompanied by an obvious sacrifice of the predictive capacity between different phenotypes [82–85]. The optimal signature was therefore identified in ConSIG by collectively assessing both criteria of *signature consistency* and *classification accuracy* (illustrated in Figure 1*b*). As metrics for assessing both criteria of *signature consistency* and *classification accuracy* are positively correlated with the assessed criteria and have the same range of values, ConSIG thus directly summed the values of two types of metrics to fairly and collectively consider both criteria without any bias or preference. As shown in the 'Metrics for Signature Consistency and Classification Accuracy' section of **Materials and Methods**, two measures of the area under the receiver operating characteristic (AUC) and *Matthews*

correlation coefficient (MCC) were employed for evaluating the *classification accuracy* [82–85]. Compared with AUC, MCC is known as a more balanced metric, making it useful for unbalanced datasets [87, 88] that were frequently encountered in OMIC datasets such as IPX0001256000 and GSE31192 in Table 1. Therefore, the CW*rel* together with both AUC and MCC were calculated and collectively considered in this study to discover the optimal signature.

Herein, two benchmarks (GSE31192 & IPX000125600 in Table 1) were collected for calculating the metrics of both criteria (*signature consistency* and *classification accuracy*). Particularly, for each benchmark dataset, 20 sub-datasets were *first* generated through randomly selecting half of the entire samples [108, 109]. *Then*, those six classical strategies (provided in Table 2) and ConSIG were applied to each of the sub-datasets, and 20 lists of feature ranking were therefore identified for each strategy. *Third*, 20 groups of top-ranked features (Top n%, n = 1, 3, 5, 10, and 20; those columns shown in Figure 5) were selected from those 20 lists of feature ranking, and CW*rel* was calculated to measure *signature consistency*. *Fourth*, the same groups of top-ranked features as that selected above were also used to build classifiers using 5-fold cross-validation based on their corresponding sub-dataset, and the AUC and MCC were calculated to measure the *classification accuracy* [113–115]. It is important to emphasize that the principle underlying the calculation of each metric determines that only one CW*rel* is generated for each dataset and a total of 20 AUCs/MCCs were generated for all 20 sub-datasets. *Finally*, the collective assessment of each strategy was shown as the scatter plots with error bars in Figure 5 (the *x*-axis indicated the values of CW*rel*, and the *y*-axis gave the values of either AUC or MCC). As illustrated, under all circumstances, ConSIG showed superior consistencies comparing with the classical ones (the blue points indicating ConSIG were far closer to 1 comparing with all other strategies as assessed by CW*rel*). Moreover, the ConSIG showed comparative (even better, under most circumstances) classification accuracy comparing with the classical ones (the blue error bars, under most circumstances, were closer to 1 comparing with other strategies as assessed by either AUC or MCC). Since IPX0001256000 and GSE31192 in Table 1 were both unbalanced datasets, it is recommended to use MCC as the primary metric [116–118]. In other words, the performance of ConSIG on *classification accuracy* was also obviously better than other strategies (the second and fourth rows in Figure 5).

As reported, some cutoffs of metrics were defined to classify the performance of feature selection strategy. Particularly, the number of 0.5 was proposed to divide CW*rel* into good (≥0.5) and poor (<0.5) consistency, the number of 0.7 was defined to classify MCC into good (≥0.7) and poor (<0.7) predictive ability, and the number of 0.8 was considered to categorize AUC into good (≥0.8) and poor (<0.8) classification accuracy according to experiences from previous publications [76, 119, 120]. Based on these cutoffs, each sub-figure in Figure 5 was further colored to different zones. The strategies within the green zone indicated that they performed 'good' for both *signature consistency* and *classification accuracy*, the strategy within orange zone showed that it performed 'good' under one criterion but 'poor' under the other, and the strategies in red zone denoted that they performed 'poor' under both criteria. As described in Figure 5, ConSIG was always in the green zone regardless of the datasets and the signature sizes, which was very different from other strategies. Moreover, the error bars (measuring AUC/MCC) of ConSIG gave a much smaller variation comparing with other strategies, which further showed its superior stability in *classification accuracy* regardless of analyzed dataset. All in all, ConSIG showed its unique ability to discover the optimal signature based on collective assessment.

Confirmation of biological relevance based on enrichment analysis

For biologists and clinicians who study on OMIC data, the identification of consistent and accurate signature is not the end of the story, and further confirmation of the biological relevance of the identified signature is usually required [121–123]. Thus, the essential function to confirm biological relevance was realized in ConSIG based on an enrichment analysis using disease/gene ontologies [42, 46]. Particularly, two types of enrichment analysis were enabled in ConSIG (as illustrated in Figure 1*c*) to discover the relationships between identified signature and a phenotype of interests [46]. These two types of analysis were based on the databases of the disease (DO) and gene (GO) ontologies [93, 103]. In ConSIG, DO & GO-based enrichment results were visualized using bubble map [102], upset plot [96], enrichment map [104], and gene-concept network [102].

To assess the biological relevance of various signatures identified by ConSIG and other classical strategies, the benchmark dataset GSE23878 was collected, which had been used to discover the differentially expressed genes [58] between the patients with colorectal carcinoma (CRC) and the healthy individuals. *First*, six classical strategies (as described in Table 2) together with ConSIG were applied for selecting the differential genes from GSE23878. *Then*, the top-100 ranked genes identified by each strategy were recorded, which resulted in seven lists of gene biomarkers. *Third*, to illustrate the variations among these seven marker lists, two Venn-plots were drawn to describe the differences between the marker list of ConSIG and that of three univariate strategies (Figure 6*a*) & three multivariate ones (Figure 6*b*). As shown, there were only 12 markers simultaneously identified by ConSIG and three univariate strategies (FC, *t*-test, and Wilcox), and no marker was constantly discovered by ConSIG and three multivariate ones (CFS, PLS-DA, and REF). Among the top-100 gene markers, about 50% of them were solely identified by their own strategy,

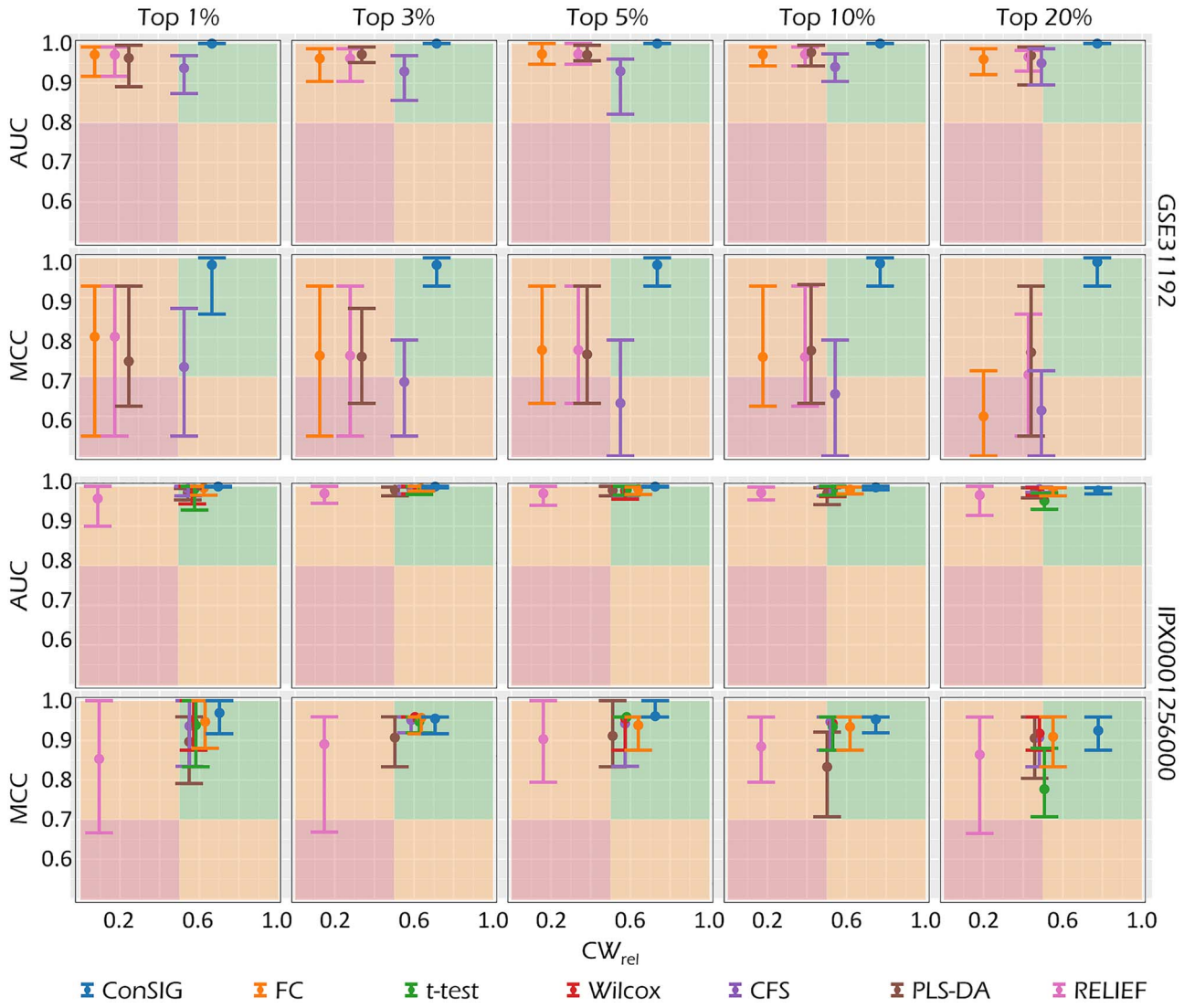

Figure 5. Performance comparison on signature identification using collective evaluation among seven studied strategies. Two benchmarks (GSE31192 and IPX000125600 described in Table 1) were collected to calculate the metrics of both criteria '*signature consistency*' and '*classification accuracy*'. The *x*-axis denoted the value of CWrel (measuring *signature consistency*), and *y*-axis gave the value of either AUC or MCC (measuring *classification accuracy*). Based on the cutoffs previously reported, the number of 0.5 was proposed to divide CW*rel* into good (≥0.5) and poor (<0.5) consistency, the number of 0.7 was defined to classify MCC to good (≥0.7) and poor (<0.7) predictive ability, and the number of 0.8 was considered to categorize AUC into good (≥0.8) and poor (<0.8) classification accuracy (59,87,88). Based on the cutoffs, each sub-figure was further colored to different zones. The strategy within the green zone indicated that its performed 'good' under both *signature consistency* and *classification accuracy*, the strategy within the orange zone denoted that it performed 'good' under one criterion but 'poor' under the other, and the strategies within the red zone denoted that they performed 'poor' under both criteria.

which indicated that there is significant variation among signatures identified by different strategies.

Because of the above variation among different signatures, it was of great interest to assess how different signatures affect their biological relevance [124–126]. Therefore, the top-100 ranked markers were *first* enriched using disease ontology (DO) database [93]. *Then*, the Top-5 DO terms (of the most significant *p*-value) enriched for each strategy were recorded, which led to a total of $5 \times 7 = 35$ DO terms. As shown in Figure 6, a comparison between the enriched DO terms for ConSIG and that for three univariate strategies (Figure 6*c*) & three multivariate ones (Figure 6*d*) was given. For a reasonable signature discovery, the identified DO terms should be closely related to the studied phenotype. In this case, a good strategy should be able to enrich DO terms that are closely related to 'colorectal carcinoma', since it is the key phenotype of the studied benchmark GSE23878 [58]. Therefore, to assess the disease relevance, a reputable database titled MalaCards was introduced [127]. MalaCards is an integrated compendium of annotated diseases, and it enables the intelligent matching between an input DO term and its built-in disease annotations [128—]. In other words, by matching each of the enriched DO terms with 'colorectal carcinoma', MalaCards returned a score indicating the relevance between a DO term and 'colorectal carcinoma'. The larger the score is, the closer the relevance is. As shown in **Figure 6*c*** and **6*d***, MalaCards scores for all enriched DO terms were provided in the right-most column, and all enriched terms

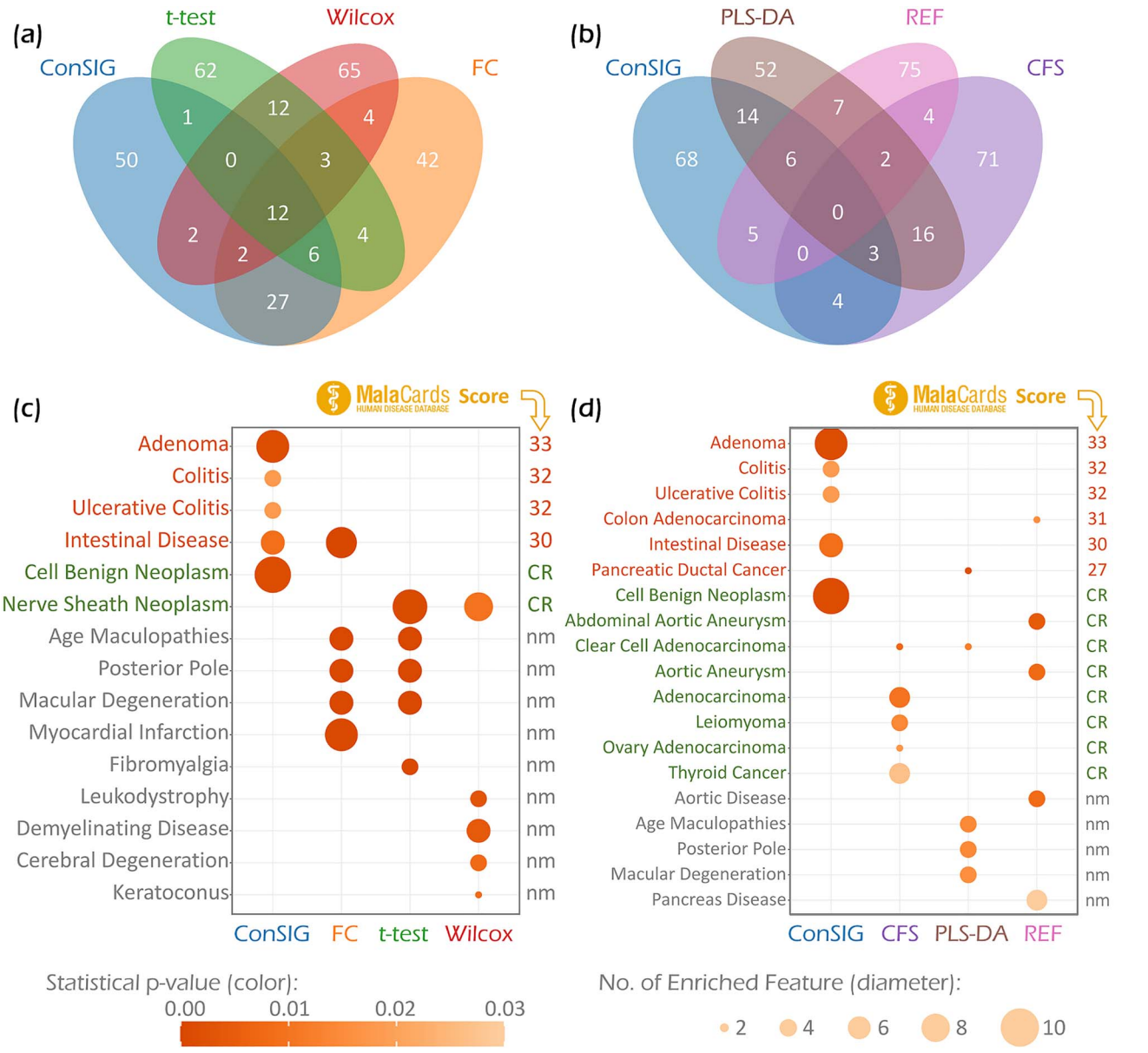

Figure 6. Disease relevance of these signatures identified by different strategies. The Venn plots were used to describe the differences between the markers of ConSIG and that of three univariate strategies (***a***) & three multivariate ones (***b***). The top-5 DO terms (of the most significant *p*-value) enriched for each strategy were recorded, which led to a total of $5 \times 7 = 35$ DO terms. A comparison between those enriched DO terms for ConSIG and that for univariate strategies (***c***) & multivariate ones (***d***) was provided. A good strategy should be able to enrich DO terms that are closely related to 'colorectal carcinoma', since it is the key phenotype of the studied benchmark GSE23878 (50). Therefore, the MalaCards database was introduced to match each of the enriched DO terms with 'colorectal carcinoma'. MalaCards returned a score indicating the relevance between a DO term and 'colorectal carcinoma'. The larger the score is, the closer the relevance is. CR: cancer-related; nm: not matched.

were ranked based on their MalaCards scores. CR indicated 'cancer-related', and nm denoted 'not matched'. In other words, some DO terms were not matched to 'colorectal carcinoma' using MalaCards, which indicated a weak or no relation. As shown, the DO terms identified by ConSIG gave much closer relation to 'colorectal carcinoma', because they were ranked at the top of **Figure 6*c*** and ***6d*** comparing with the terms discovered by classical strategies (FC, *t*-test, Wilcox, CFS, PLS-DA, and REF).

Conclusion and perspectives

A novel feature selection strategy was proposed and successfully validated in our previous study [16]. The underlying theory of this strategy had attracted broad interest from and been used by a wide range of research fields. However, our previous study only focused on the transcriptomic data of *schizophrenia* patients, which asked for the broad and user-friendly applications to other OMIC-related direction. To make this new strategy public to all users (especially those with little computational skills),

the ConSIG was developed to ***a***) integrate the novel strategy proposed in our previous study to effectively guarantee the consistent identification of gene/protein signature, ***b***) discover optimal signature by evaluating both *signature consistency* & *classification accuracy*, and ***c***) confirm biological relevance by enriching disease & gene ontologies. ConSIG has been comprehensively and adequately validated in proteomics and transcriptomics data with different research directions and is expected to be used as an essential complement to other existing tools for OMIC-based signature discovery, which can be accessed by all users without login requirement at https://idrblab.org/consig/

Fundings

Funded by Natural Science Foundation of Zhejiang Province (LR21H300001); National Natural Science Foundation of China (81,872,798 & U1909208); Leading Talent of the 'Ten Thousand Plan' - National High-Level Talents Special Support Plan of China; Fundamental Research Fund for Central Universities (2018QNA7023); 'Double Top-Class' University Project (181,201*194232101); Key R&D Program of Zhejiang Province (2020C03010). This work was supported by Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; Information Technology Center of Zhejiang University.

Conflict of Interest

None declared.

Key Points

- ConSIG introduces a novel strategy proposed in our previous study to research community by effectively guaranteeing the consistent discovery of molecular signature
- ConSIG determines the optimal signature by collectively assessing *signature consistency* and *classification accuracy* based on a variety of evaluating metrics
- ConSIG confirms the biological relevance by enriching both disease and gene ontologies

References

1. Wozniak JM, Mills RH, Olson J, *et al.* Mortality risk profiling of staphylococcus aureus bacteremia by multi-omic serum analysis reveals early predictive and pathogenic signatures. *Cell* 2020;**182**:1311–27.
2. Hou Y, Zhou Y, Hussain M, *et al.* Cardiac risk stratification in cancer patients: a longitudinal patient-patient network analysis. *PLoS Med* 2021;**18**:e1003736.
3. Malfatti MC, Gerratana L, Dalla E, *et al.* APE1 and NPM1 protect cancer cells from platinum compounds cytotoxicity and their expression pattern has a prognostic value in TNBC. *J Exp Clin Cancer Res* 2019;**38**:309.
4. Ghezzi P, Floridi L, Boraschi D, *et al.* Oxidative stress and inflammation induced by environmental and psychological stressors: a biomarker perspective. *Antioxid Redox Signal* 2018;**28**:852–72.
5. Hou Y, Zhou Y, Gack MU, *et al.* Multimodal single-cell omics analysis identifies epithelium-immune cell interactions and immune vulnerability associated with sex differences in COVID-19. *Signal Transduct Target Ther* 2021;**6**:292.
6. Codrich M, Dalla E, Mio C, *et al.* Integrated multi-omics analyses on patient-derived CRC organoids highlight altered molecular pathways in colorectal cancer progression involving PTEN. *J Exp Clin Cancer Res* 2021;**40**:198.
7. Gentles AJ, Newman AM, Liu CL, *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* 2015;**21**:938–45.
8. Bidard FC, Pierga JY, Soria JC, *et al.* Translating metastasis-related biomarkers to the clinic-progress and pitfalls. *Nat Rev Clin Oncol* 2013;**10**:169–79.
9. Wilmanski T, Rappaport N, Earls JC, *et al.* Blood metabolome predicts gut microbiome alpha-diversity in humans. *Nat Biotechnol* 2019;**37**:1217–28.
10. Mistry M, Gillis J, Pavlidis P. Genome-wide expression profiling of schizophrenia using a large combined cohort. *Mol Psychiatry* 2013;**18**:215–25.
11. Li YH, Xu JY, Tao L, *et al.* SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS One* 2016;**11**:e0155290.
12. Lu Y, Brommer B, Tian X, *et al.* Reprogramming to recover youthful epigenetic information and restore vision. *Nature* 2020;**588**:124–9.
13. Li F, Zhou Y, Zhang X, *et al.* SSizer: determining the sample sufficiency for comparative biological study. *J Mol Biol* 2020;**432**:3411–21.
14. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* 2018;**19**:129–47.
15. Yang Q, Wang Y, Zhang Y, *et al.* NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res* 2020;**48**:W436–48.
16. Yang Q, Li B, Tang J, *et al.* Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform* 2020;**21**:1058–68.
17. Yang QX, Wang YX, Li FC, *et al.* Identification of the gene signature reflecting schizophrenia's etiology by constructing artificial intelligence-based method of enhanced reproducibility. *CNS Neurosci Ther* 2019;**25**:1054–63.
18. Li J, Wei L, Guo F, *et al.* EP3: an ensemble predictor that accurately identifies type III secreted effectors. *Brief Bioinform* 2021;**22**:1918–28.
19. Zhang Z, Cui F, Cao C, *et al.* Single-cell RNA analysis reveals the potential risk of organ-specific cell types vulnerable to SARS-CoV-2 infections. *Comput Biol Med* 2021;**140**:105092.
20. Huang Q, Zhang J, Wei L, *et al.* 6mA-RicePred: a method for identifying DNA N (6)-methyladenine sites in the rice genome based on feature fusion. *Front Plant Sci* 2020;**11**:4.
21. Niu M, Lin Y, Zou Q. sgRNACNN: identifying sgRNA on-target activity in four crops using ensembles of convolutional neural networks. *Plant Mol Biol* 2021;**105**:483–95.

22. Ru X, Wang L, Li L, *et al.* Exploration of the correlation between GPCRs and drugs based on a learning to rank algorithm. *Comput Biol Med* 2020;**119**:103660.
23. Li J, He S, Guo F, *et al.* HSM6AP: a high-precision predictor for the homo sapiens N6-methyladenosine (m 6 a) based on multiple weights and feature stitching. *RNA Biol* 2021;**18**:1882–92.
24. Li Q, Zhou W, Wang D, *et al.* Prediction of anticancer peptides using a low-dimensional feature model. *Front Bioeng Biotechnol* 2020;**8**:892.
25. Fu J, Luo Y, Mou M, *et al.* Advances in current diabetes proteomics: from the perspectives of label- free quantification and biomarker selection. *Curr Drug Targets* 2020;**21**:34–54.
26. Chen Z, Shen Z, Zhao D, *et al.* Genome-wide analysis of LysM-containing gene family in wheat: structural and phylogenetic analysis during development and defense. *Genes* 2020;**12**:31.
27. Meng C, Jin S, Wang L, *et al.* AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front Bioeng Biotechnol* 2019;**7**:224.
28. Tang Z, Kang B, Li C, *et al.* GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* 2019;**47**:W556–60.
29. Toro-Dominguez D, Martorell-Marugan J, Lopez-Dominguez R, *et al.* ImaGEO: integrative gene expression meta-analysis from GEO database. *Bioinformatics* 2019;**35**:880–2.
30. Gruca A, Henzel J, Kostorz I, *et al.* MAINE: a web tool for multi-omics feature selection and rule based data exploration. *Bioinformatics* 2021;**38**:1773–5.
31. Pang Z, Chong J, Zhou G, *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res* 2021;**49**:W388–96.
32. Ge X, Raghu VK, Chrysanthis PK, *et al.* CausalMGM: an interactive web-based causal discovery tool. *Nucleic Acids Res* 2020;**48**:W597–602.
33. Zhou G, Ewald J, Xia J. OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. *Nucleic Acids Res* 2021;**49**:W476–82.
34. Fu J, Zhang Y, Wang Y, *et al.* Optimization of metabolomic data processing using NOREVA. *Nat Protoc* 2022;**17**:129–51.
35. Li B, Tang J, Yang Q, *et al.* NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* 2017;**45**:W162–70.
36. Jung S, Lee H, Myung S, *et al.* A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* 2022;**601**:211–6.
37. Tyanova S, Albrechtsen R, Kronqvist P, *et al.* Proteomic maps of breast cancer subtypes. *Nat Commun* 2016;**7**:10259.
38. Zheng A, Lamkin M, Zhao H, *et al.* Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat Mach Intell* 2021;**3**:172–80.
39. Chen Z, Zhao P, Li C, *et al.* iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res* 2021;**49**:e60.
40. Shah MS, DeSantis TZ, Weinmaier T, *et al.* Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* 2018;**67**:882–91.
41. Christin C, Hoefsloot HC, Smilde AK, *et al.* A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol Cell Proteomics* 2013;**12**:263–76.
42. Ghezzi P, Davies K, Delaney A, *et al.* Theory of signs and statistical approach to big data in assessing the relevance of clinical biomarkers of inflammation and oxidative stress. *Proc Natl Acad Sci U S A* 2018;**115**:2473–7.
43. Ghezzi P. Environmental risk factors and their footprints in vivo - a proposal for the classification of oxidative stress biomarkers. *Redox Biol* 2020;**34**:101442.
44. Mangiapane G, Parolini I, Conte K, *et al.* Enzymatically active apurinic/apyrimidinic endodeoxyribonuclease 1 is released by mammalian cells through exosomes. *J Biol Chem* 2021;**296**:100569.
45. Cheng F, Zhao J, Wang Y, *et al.* Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat Genet* 2021;**53**:342–53.
46. Yuan N, Chen Y, Xia Y, *et al.* Inflammation-related biomarkers in major psychiatric disorders: a cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. *Transl Psychiatry* 2019;**9**:233.
47. Xue W, Fu T, Deng S, *et al.* Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram. *ACS Chem Nerosci* 2022;**13**:340–51.
48. Dai Q, Bao CH, Hai YB, *et al.* MTGIpick allows robust identification of genomic islands from a single genome. *Brief Bioinform* 2018;**19**:361–73.
49. Zhang S, Amahong K, Zhang C, *et al.* RNA-RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection. *Brief Bioinform* 2021;**23**:bbab397.
50. Wang R, Zheng X, Wang J, *et al.* Improving bulk RNA-seq classification by transferring gene signature from single cells in acute myeloid leukemia. *Brief Bioinform* 2022;**23**.
51. Yang SQ, Wang YX, Chen Y, *et al.* MASQC: next generation sequencing assists third generation sequencing for quality control in N6-Methyladenine DNA identification. *Front Genet* 2020;**11**.
52. Kong R, Xu X, Liu X, *et al.* 2SigFinder: the combined use of small-scale and large-scale statistical testing for genomic island detection from a single genome. *BMC Bioinformatics* 2020;**21**:159.
53. Shao C, Zhao M, Chen X, *et al.* Comprehensive analysis of individual variation in the urinary proteome revealed significant gender differences. *Mol Cell Proteomics* 2019;**18**:1110–22.
54. Schroeder BO, Birchenough GMH, Stahlman M, *et al.* Bifidobacteria or fiber protects against diet-induced microbiota-mediated colonic mucus deterioration. *Cell Host Microbe* 2018;**23**:27–40.
55. Chen T, Ma J, Liu Y, *et al.* iProX in 2021: connecting proteomics data sharing with big data. *Nucleic Acids Res* 2022;**50**:D1522–7.
56. Perez-Riverol Y, Bai J, Bandla C, *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;**50**:D543–52.
57. Harvell DM, Kim J, O'Brien J, *et al.* Genomic signatures of pregnancy-associated breast cancer epithelia and stroma and their regulation by estrogens and progesterone. *Horm Cancer* 2013;**4**:140–53.
58. Uddin S, Ahmed M, Hussain A, *et al.* Genome-wide expression analysis of middle eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy. *Am J Pathol* 2011;**178**:537–47.
59. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.
60. Fu J, Tang J, Wang Y, *et al.* Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front Pharmacol* 2018;**9**:681.
61. Fu J, Zhang Y, Liu J, *et al.* Pharmacometabonomics: data processing and statistical analysis. *Brief Bioinform* 2021;**22**:bbab138.
62. Yang H, Qin C, Li YH, *et al.* Therapeutic target database update 2016: enriched resource for bench to clinical drug target and

targeted pathway information. *Nucleic Acids Res* 2016;**44**:D1069–74.
63. Liu X, Xu Y, Wang R, *et al.* A network-based algorithm for the identification of moonlighting noncoding RNAs and its application in sepsis. *Brief Bioinform* 2021;**22**:581–8.
64. Onesime M, Yang ZY, Dai Q. Genomic Island prediction via Chi-Square test and random Forest algorithm. *Comput Math Methods Med* 2021;2021.
65. Wang YX, Xu YJ, Yang ZY, *et al.* Using recursive feature selection with random Forest to improve protein structural class prediction for low-similarity sequences. *Comput Math Methods Med* 2021;**2021**.
66. Guo L, Lobenhofer EK, Wang C, *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol* 2006;**24**:1162–9.
67. Wilcoxon F. Individual comparisons of grouped data by ranking methods. *J Econ Entomol* 1946;**39**:269.
68. Zhang YN, Liu RJ, Wang X, *et al.* Boosted binary Harris hawks optimizer and feature selection. *Engineering with Computers* 2021;**37**:3741–70.
69. Hu J, Chen HL, Heidari AA, *et al.* Orthogonal learning covariance matrix for defects of grey wolf optimizer: insights, balance, diversity, and feature selection. *Knowledge-Based Systems* 2021;**213**.
70. Zhang YN, Liu RJ, Heidari AA, *et al.* Towards augmented kernel extreme learning models for bankruptcy prediction: algorithmic behavior and comprehensive analysis. *Neurocomputing* 2021;**430**:185–212.
71. Chuang LY, Yang CH, Wu KC, *et al.* A hybrid feature selection method for DNA microarray data. *Comput Biol Med* 2011;**41**:228–37.
72. Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J* 2013;**4**:e201301009.
73. Urbanowicz RJ, Meeker M, La Cava W, *et al.* Relief-based feature selection: introduction and review. *J Biomed Inform* 2018;**85**:189–203.
74. Somol P, Novovicova J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans Pattern Anal Mach Intell* 2010;**32**:1921–39.
75. Lopez NC, Garcia-Ordas MT, Vitelli-Storelli F, *et al.* Evaluation of feature selection techniques for breast cancer risk prediction. *Int J Environ Res Public Health* 2021;**18**:10670.
76. Song X, Waitman LR, Hu Y, *et al.* Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J Am Med Inform Assoc* 2019;**26**:242–53.
77. Piles M, Bergsma R, Gianola D, *et al.* Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. *Front Genet* 2021;**12**:611506.
78. Tan MS, Cheah PL, Chin AV, *et al.* A review on omics-based biomarkers discovery for Alzheimer's disease from the bioinformatics perspectives: statistical approach vs machine learning approach. *Comput Biol Med* 2021;**139**:104947.
79. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* 2006;**103**:5923–8.
80. Shiri I, Sorouri M, Geramifar P, *et al.* Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients. *Comput Biol Med* 2021;**132**:104304.
81. Petkovic M, Slavkov I, Kocev D, *et al.* Biomarker discovery by feature ranking: evaluation on a case study of embryonal tumors. *Comput Biol Med* 2021;**128**:104143.
82. Eddowes PJ, Sasso M, Allison M, *et al.* Accuracy of FibroScan controlled attenuation parameter and liver stiffness measurement in assessing steatosis and fibrosis in patients with nonalcoholic fatty liver disease. *Gastroenterology* 2019;**156**:1717–30.
83. Goh WWB, Wong L. Advanced bioinformatics methods for practical applications in proteomics. *Brief Bioinform* 2019;**20**:347–55.
84. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, *et al.* Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health* 2020;**2**:179–91.
85. Tang ZQ, Han LY, Lin HH, *et al.* Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation. *Cancer Res* 2007;**67**:9996–10003.
86. Xu J, Li F, Leier A, *et al.* Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Brief Bioinform* 2021;**22**:bbab083.
87. Nielsen AB, Thorsen-Meyer HC, Belling K, *et al.* Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the danish national patient registry and electronic patient records. *Lancet Digit Health* 2019;**1**:78–89.
88. Bedon L, Cecchin E, Fabbiani E, *et al.* Machine learning application in a phase i clinical trial allows for the identification of clinical-biomolecular markers significantly associated with toxicity. *Clin Pharmacol Ther* 2022;**111**:686–96.
89. Chen YZ, Wang ZZ, Wang Y, *et al.* nhKcr: a new bioinformatics tool for predicting crotonylation sites on human nonhistone proteins based on deep learning. *Brief Bioinform* 2021;**22**:bbab146.
90. Chen HL, Wang G, Ma C, *et al.* An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. *Neurocomputing* 2016;**184**:131–44.
91. Xia WQ, Zheng LY, Fang JB, *et al.* PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Comput Biol Med* 2022;**145**:105465.
92. Hu LF, Hong GL, Ma JS, *et al.* An efficient machine learning approach for diagnosis of paraquat-poisoned patients. *Comput Biol Med* 2015;**59**:116–24.
93. Schriml LM, Arze C, Nadendla S, *et al.* Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 2012;**40**:D940–6.
94. Li CY, Hou LX, Sharma BY, *et al.* Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. *Comput Methods Programs Biomed* 2018;**153**:211–25.
95. Liu L, Zhao D, Yu FH, *et al.* Ant colony optimization with Cauchy and greedy levy mutations for multilevel COVID 19 X-ray image segmentation. *Comput Biol Med* 2021;**136**.
96. Yu G, Wang LG, Yan GR, *et al.* DOSE: an R/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2015;**31**:608–9.
97. Wang Y, Zhang S, Li F, *et al.* Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res* 2020;**48**:D1031–41.
98. Zhu F, Shi Z, Qin C, *et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;**40**:D1128–36.

99. Fu T, Li F, Zhang Y, *et al.* VARIDT 2.0: structural variability of drug transporter. *Nucleic Acids Res* 2022;**50**:D1417–31.
100. Wang X, Li F, Qiu W, *et al.* SYNBIP: synthetic binding proteins for research, diagnosis and therapy. *Nucleic Acids Res* 2022;**50**: D560–70.
101. Yin J, Sun W, Li F, *et al.* VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res* 2020;**48**:D1042–50.
102. Yu G, Wang LG, Han Y, *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**:284–7.
103. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.
104. Luo W, Brouwer C. Pathview: an R/bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 2013;**29**:1830–1.
105. Rohart F, Gautier B, Singh A, *et al.* mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;**13**:e1005752.
106. Robin X, Turck N, Hainard A, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:77.
107. Saraswat M, Joenvaara S, Seppanen H, *et al.* Comparative proteomic profiling of the serum differentiates pancreatic cancer from chronic pancreatitis. *Cancer Med* 2017;**6**:1738–51.
108. Garge NR, Bobashev G, Eggleston B. Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC Bioinformatics* 2013;**14**:125.
109. White AM, Philogene GS, Fine L, *et al.* Social support and self-reported health status of older adults in the United States. *Am J Public Health* 2009;**99**:1872–8.
110. Zhang X, Fan M, Wang D, *et al.* Top-k feature selection framework using robust 0-1 integer programming. *IEEE Trans Neural Netw Learn Syst* 2021;**32**:3005–19.
111. Freitas AA. Investigating the role of simpson's paradox in the analysis of top-ranked features in high-dimensional bioinformatics datasets. *Brief Bioinform* 2020;**21**:421–8.
112. Xu X, Zhang HL, Liu QP, *et al.* Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. *J Hepatol* 2019;**70**:1133–44.
113. Xia JF, Chen HL, Li Q, *et al.* Ultrasound-based differentiation of malignant and benign thyroid nodules: an extreme learning machine approach. *Comput Methods Programs Biomed* 2017;**147**: 37–49.
114. Zhang Q, Wang ZY, Heidari AA, *et al.* Gaussian Barebone Salp swarm algorithm with stochastic fractal search for medical image segmentation: a COVID-19 case study. *Comput Biol Med* 2021;**139**.
115. Zhang Y, Zhang HX, Zheng QC. In silico study of membrane lipid composition regulating conformation and hydration of influenza virus B M2 channel. *J Chem Inf Model* 2020;**60**:3603–15.
116. Zhang S, Amahong K, Sun X, *et al.* The miRNA: a small but powerful RNA for COVID-19. *Brief Bioinform* 2021;**22**:1137–49.
117. Zhang Y, Ying JB, Hong JJ, *et al.* How does chirality determine the selective inhibition of histone deacetylase 6? A lesson from Trichostatin a enantiomers based on molecular dynamics. *ACS Chem Nerosci* 2019;**10**:2467–80.
118. Yin J, Li F, Zhou Y, *et al.* INTEDE: interactome of drug-metabolizing enzymes. *Nucleic Acids Res* 2021;**49**:D1233–43.
119. Levitsky J, Asrani SK, Klintmalm G, *et al.* Discovery and validation of a biomarker model (PRESERVE) predictive of renal outcomes after liver transplantation. *Hepatology* 2020;**71**: 1775–86.
120. Tawfik DS, Gould JB, Profit J. Perinatal risk factors and outcome coding in clinical and administrative databases. *Pediatrics* 2019;**143**:e20181487.
121. Shen Z, Kuang S, Zhang Y, *et al.* Chitosan hydrogel incorporated with dental pulp stem cell-derived exosomes alleviates periodontitis in mice via a macrophage-dependent mechanism. *Bioact Mater* 2020;**5**:1113–26.
122. Tang J, Fu J, Wang Y, *et al.* ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform* 2020;**21**:621–36.
123. Tang J, Fu J, Wang Y, *et al.* Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol Cell Proteomics* 2019;**18**:1683–99.
124. Lin B, Zhang H, Zheng Q. How do mutations affect the structural characteristics and substrate binding of CYP21A2?An investigation by molecular dynamics simulations. *Phys Chem Chem Phys* 2020;**22**:8870–7.
125. Liu X, Zheng X, Wang J, *et al.* A long non-coding RNA signature for diagnostic prediction of sepsis upon ICU admission. *Clin Transl Med* 2020;**10**:e123.
126. Shi BB, Ye H, Zheng L, *et al.* Evolutionary warning system for COVID-19 severity: Colony predation algorithm enhanced extreme learning machine. *Comput Biol Med* 2021;**136**.
127. Rappaport N, Twik M, Plaschkes I, *et al.* MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* 2017;**45**:D877–87.