Review

Computational advances of tumor marker selection and sample classification in cancer proteomics

Jing Tang a,b, Yunxia Wang b, Yongchao Luo b, Jianbo Fu b, Yang Zhang b,c, Yi Li b, Ziyu Xiaob, Yan Lou d, Yunqing Qiud, Feng Zhu a,b,c

a Department of Bioinformatics, Chongqing Medical University, Chongqing 400016, China
b College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
c School of Pharmaceutical Sciences and Innovative Drug Research Centre, Chongqing University, Chongqing 401331, China
d Zhejiang Provincial Key Laboratory for Drug Clinical Research and Evaluation, The First Affiliated Hospital, Zhejiang University, Hangzhou 310000, China

A R T I C L E   I N F O

Article history:
Received 7 April 2020
Received in revised form 6 July 2020
Accepted 8 July 2020
Available online 17 July 2020

Keywords:
Cancer proteomics
Tumor marker selection
Sample classification
Computational methods

A B S T R A C T

Cancer proteomics has become a powerful technique for characterizing the protein markers driving transformation of malignancy, tracing proteome variation triggered by therapeutics, and discovering the novel targets and drugs for the treatment of oncologic diseases. To facilitate cancer diagnosis/prognosis and accelerate drug target discovery, a variety of methods for tumor marker identification and sample classification have been developed and successfully applied to cancer proteomic studies. This review article describes the most recent advances in those various approaches together with their current applications in cancer-related studies. Firstly, a number of popular feature selection methods are overviewed with objective evaluation on their advantages and disadvantages. Secondly, these methods are grouped into three major classes based on their underlying algorithms. Finally, a variety of sample separation algorithms are discussed. This review provides a comprehensive overview of the advances on tumor marker identification and patients/samples/tissues separations, which could be guidance to the researches in cancer proteomics.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction ........................................................................................................ 2012
2. Considerations concerning study design. .................................................. 2014
3. Feature selection for tumor marker identification ..................................... 2014
   3.1. Filter methods for tumor marker identification from proteomic data. ........ 2014
       3.1.1. Univariate filter methods ................................................................. 2015
       3.1.2. Multivariate filter methods .......................................................... 2017
   3.2. Wrapper methods for tumor marker identification from proteomic data. ... 2018
   3.3. Embedded methods for tumor marker identification from proteomic data. ... 2019
4. Sample clustering and visualization ............................................................... 2020
5. Conclusions. .................................................................................................... 2021

Abbreviations: ANN, Artificial Neural Network; ANOVA, Analysis of Variance; CFS, Correlation-based Feature Selection; DAPC, Discriminant Analysis of Principal Component; DT, Decision Trees; EDA, Estimation of Distribution Algorithm; FC, Fold Change; GA, Genetic Algorithms; GR, Gain Ratio; HC, Hill Climbing; HCA, Hierarchical Cluster Analysis; IG, Information Gain; LDA, Linear Discriminant Analysis; LIMMA, Linear Models for Microarray Data; MBF, Markov Blanket Filter; MWW, Mann–Whitney–Wilcoxon test; OPLS-DA, Orthogonal Partial Least Squares Discriminant Analysis; PCA, Principal Component Analysis; PLS-DA, Partial Least Square Discriminant Analysis; RF-RFE, Random Forest with Recursive Feature Elimination; RF, Random Forest; SA, Simulated Annealing; SAM, Significance Analysis of Microarrays; SBE, Sequential Backward Elimination; SFS, and Sequential Forward Selection; SOM, Self-organizing Map; SU, Symmetrical Uncertainty; SVM-RFE, Support Vector Machine with Recursive Feature Elimination; SVM, Support Vector Machine; sPLS-DA, Sparse Partial Least Squares Discriminant Analysis; tSNE, Student t Distribution; χ², Chi-square.

* Corresponding authors at: College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China (F. Zhu).
E-mail addresses: quyq@zju.edu.cn (Y. Qiu), zhufeng@zju.edu.cn (F. Zhu).

https://doi.org/10.1016/j.csbj.2020.07.009
2001-0370 © 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

In 2015, cancer caused over eight million deaths, making it the second leading cause of death in the world, nearly one-sixth deaths caused by cancer according to the report of the World Health Organization [1,2]. The impact of cancer on economy is significant and growing [3–5], and the cancer is common in late manifestations and inaccessible diagnosis and treatment [6]. If detected early, cancer can respond to effective treatments, leading to lower morbidity, less treatment spending and larger chances of survival [7]. Thus, an effective early diagnosis and treatment can improve the cure rate of many cancers and alleviate the burden of cancer patients [8].

It is important to develop more effective early diagnosis strategies for cancer, and biomarkers are urgently needed to diagnose various types of cancers, assess the severity of diseases, and discover the corresponding therapeutics [9–15]. A biomarker can be a protein, a polypeptide or a metabolite whose levels change with the stage of cancer, as well as the messenger RNA or other kinds of nucleic acids [16–19]. It is well known that there is much less understanding of the pathogenesis of cancer at the proteomic level than gene mutations level [20,21]. Since proteome is a functional translation of genome and a rich source of biomarkers, huge amounts of time and money are needed for proteomics to develop biomarker [22]. Gene is merely a “formulation” of the cell, and the protein encoded by the gene is ultimately a functional participant in normal and cancer physiology [23]. Thus, while the genetic background contributes in part to the susceptibility and development of cancer [24], cancer can now be considered as a proteome disease and has more links to the post-transcriptional steps [25,26]. In terms of patient prognosis, there is an urgent need for protein markers distinguishing cancer patients from normal individuals. In addition, biomarkers used in cancer surveillance may affect the development of new therapy [27]. Protein biomarkers obtained from biological samples can not only provide starting point for finding links between disease and biological pathway, but also play an important role in advancing cancer medical research through the early diagnosis of cancer and prognosis of treatment interventions [28]. An ideal biomarker should be with the following characteristics: (1) good consistency and (2) high reproducibility on same phenotyping biological samples, and (3) good classification performance at distinguishing cases from controls across studies [29–31].

Recently, a well-rounded analysis of human genome, transcriptome, proteome and metabolome has made significant advance, making a significant contribution to the discovery of tumor biomarkers [32]. Proteomics can provide the qualitative and quantitative information on proteins based on a multitude of complex...
biological samples [33,34]. It is also a powerful technique for identifying potential proteomic biomarkers, facilitating the discovery of anticancer drug targets and providing new insight into mechanism underlying complicated cancers [35–37,17]. The workflow of cancer biomarker discovery based on quantitative proteomics is illustrated in Fig. 1. In the proteomics study, it is critical for assessing the performance of feature selection methods using the external validations strategy [38]. The external validation implies the presence of a test dataset of samples that has not been employed for constructing the model, and could be considered the suitable strategy for avoiding overfitting problem [38].

Due to the sparsity of features in a big proteomic data, feature selection methods are applied to identify significant proteins/peptides (or features) between distinct groups, which is a crucial step for cancer classification, diagnosis and prognosis in the comparative proteomics study [38–42]. Currently, diverse feature selections have been developed and applied to the analyzed proteomic data in various cancer-associated studies [41]. However, due to the lack of robustness of feature selection methods, the robustness and consistency of the biomarker sets discovered in most cases is ambiguous [43,44].

This review describes recent computational advances made in the field of cancer proteomics from biomarker candidates’ identification perspectives. A variety of popular feature selection techniques applied to cancer proteomic data are overviewed with critical assessments of both challenges and potentials. The main aim of this review is to make practitioners aware of the benefits and the necessity of applying feature selection techniques.

2. Considerations concerning study design

Tumor heterogeneity, design of study, sample size sand selection of the reference sample should be considered for discovering tumor biomarker in proteomic study [45–47]. For example, the use of normal/healthy controls as a reference group also needs to be made cautiously [47]. The inter-individual heterogeneity among tumors could affect proteins biomarker expression [48] and could result in inappropriate “normal controls” samples employed for tumor biomarker discovery [48]. Moreover, a relatively large numbers of samples employed could provide the biomarker validation in study design [49].

3. Feature selection for tumor marker identification

Feature selection techniques have become quite popular as well as required in tumor marker identification [50–53]. The application of feature selection techniques not only facilitates the identification of optimal differential oncological protein features [54–56], but also improves cancer risk stratification and prediction [57,58]. The available feature selection techniques can be organized into three categories based on their theories of screening variables and classifying distinct groups: filter methods, wrapper methods and embedded approaches. Moreover, one of the major problems in proteomics dataset is the treating of missing values. The imputation methods widely applied in proteomics data included the background imputation, Bayesian principal component imputation, censored imputation, k-nearest neighbor imputation, local least squares imputation and singular value decomposition [59].

As shown in Fig. 2, a common taxonomy of feature selection methods in different categories is provided, and the corresponding advantages and disadvantages of these methods are discussed with examples of the most influential techniques. Due to the various types of feature selection techniques available for tumor marker identification (Table 1), it is a great challenge to identify the optimal approach for analyzing any tumor marker-related study. The extensive application of each feature selection methods in current proteomic study together with popular sample classification methods are shown in Fig. 3. These common feature selection techniques currently available for tumor marker identification are overviewed as follow.

3.1. Filter methods for tumor marker identification from proteomic data

Filter methods involve in choosing the significantly differential features based on discriminating metric (e.g. p-value) that are relatively independent of classification. This metric well reflects the quality of each feature in terms of its discriminative power. The features would be remained when the metric values were within a specific criterion, and features beyond the thresholding condition would be eliminated. However, it should be noted that the feature selection based on p-value could be misleading when the difference is not based on regulation, but on the mere capacity of a cell to produce enough of the protein. Moreover, multicollinearity...
Different feature selection methods for identifying the markers in proteomics research.

| Methods                                      | Extensive application of each feature selection methods in current proteome research                                                                                                                                 |
|----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| **Filter-based feature selection methods**   |                                                                                                                                                                                                                                                                    |
| 1 Fold change                                 | Used to identify the protein markers for early screening detection and monitoring invasive breast cancer progression (PLoS One. 10: e0141876, 2015).                                                                                                                        |
| 2 Entropy-based Filters                       | Applied to discover biomarkers from peptide profiling based on proteomic techniques (Nat Protoc. 2:588–602, 2007).                                                                                                                                             |
| 3 Analysis of variance                        | Used to discover the biomarkers associated with the orthostatic hypotension based on the proteome study (Hypertension. 71: 465–72, 2018).                                                                                                                    |
| 4 Chi-square                                  | Applied to reveal a prognostic signature in oral cancer in combining discovery and targeted proteomics (Nat Commun. 9:3598, 2018).                                                                                                                               |
| 5 Student’s t-distribution                    | Performed to reveal the differential proteins between samples with and without primary dysmenorrhea (F1000Res. 7:59, 2018).                                                                                                                                    |
| 6 Significance Analysis of Microarrays        | Used to analyze significance between two biological samples in quantitative proteomic study (BMC Bioinformatics. 9:187, 2008).                                                                                                                             |
| 7 Linear Models for Microarray Data           | Applied to find significantly affected proteins by cathepsin A via proteomic profiling (J Proteome Res. 15:3188–95, 2016).                                                                                                                                  |
| 8 Linear discriminant analysis                | Used to find circulating lipids and coagulation cascade in septic shock progression based on proteomics data (Sci Rep. 8:5681, 2018).                                                                                                                   |
| 9 Mann–Whitney–Wilcoxon test                  | Used to identify the candidate biomarkers of prostate cancer in proteomic study (BMC Bioinformatics. 16:169, 2015).                                                                                                                                         |
| 10 Correlation-based feature selection        | Used to screen the features for investigating the periplasmic expression of soluble proteins in Escherichia coli (Sci Rep. 6:21844, 2016).                                                                                                                     |
| 11 Markov blanket filter                      | Used to detect proteomic biomarker in the recurrent ovarian cancer study from high-resolution mass spectrometry data (IEEE Trans Inf Technol Biomed. 13:195–206, 2009).                                        |
| 12 Partial Least Square Discriminant Analysis| Applied to select the discriminative proteins between the medullary sponge kidney and idiopathic calcium nephrolithiasis patients (Int J Mol Sci. 20:5517, 2019).                                                     |
| 13 Orthogonal Partial Least Squares Discriminant Analysis | Used to identify candidate biomarkers significantly discriminated cholangiocarcinoma from normal and periductal fibrosis patients (PLoS One. 14:e0221024, 2019).                                      |
| 14 Sparse Partial Least Squares Discriminant Analysis | Applied to discovery the protein biomarkers in hypertrophic cardiomyopathy based on proteomics profiling (J Cardiovasc Transl Res. 12:569–79, 2019).                                      |
| 15 Discriminant Analysis of Principal Component | Used to find conditions for the subsequent identification of biomarkers and stress proteins (PLoS One. 11:e0165504, 2016).                                                                                      |
| **Wrapper-based feature selection methods**   |                                                                                                                                                                                                                                                                    |
| 1 Sequential forward selection               | Used for protein mass spectrometry in the disease diagnosis and biomarker identification (BMC Bioinformatics. 6:68, 2005).                                                                                                                               |
| 2 Sequential backward elimination             | Used to select k-spaced amino acid pairs to identify the pupylated proteins and pupylation sites based on proteomic data (J Theor Biol. 336:11–7, 2013).                                                                                                           |
| 3 Simulated annealing                         | Applied to find a biomarker in prostate proteomic pattern based on prostate protein mass spectrometry data (Carcinobio. 195–200, 2008).                                                                                                                  |
| 4 Hill climber                                | Employed to tumor detection and discriminate tumor samples from non-tumor ones in proteomic study (Artif Intell Med. 32:71–83, 2004).                                                                                                                   |
| 5 Genetic algorithm                           | Used to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease (Lancet. 359:572–7, 2002).                                                                                                                              |
| 6 Estimation of distribution algorithm        | Applied to identify the prognosis biomarker of ovarian cancer in metabolomic study (Metabolomics. 7:614–22 2011)                                                                                                                                                    |
| **Embedded feature selection methods**        |                                                                                                                                                                                                                                                                    |
| 1 Decision Tree                               | Used to automatically determine proteomic biomarkers and predictive models in the diagnosis of rheumatoid arthritis and inflammatory bowel diseases (Bioinformatics, 21:3138–45, 2005).                                           |
| 2 Support Vector Machine                      | Used to find proteins which differ between the breast cancer subtypes based on proteomic profiles (Nat Commun. 7:10259, 2016).                                                                                                                                |
| 3 Weighted naive Bayes                        | Used to discover serum protein biomarkers for the diagnosis of active pulmonary tuberculosis based on proteomic profiling (J Clin Microbiol. 59:3057–71, 2017).                                                                                          |
| 4 SVM Recursive Feature Elimination           | Used to identify the biomarkers for predicting the early recurrence of ovarian cancer based on serum proteomic profiling (Genome Inform. 16:195–204, 2005).                                                                                                  |
| 5 Artificial Neural Networks                  | Applied to identify signature sequences for detecting hepatocellular carcinoma and its subtypes (Clin Chem. 49:752–60, 2003).                                                                                                                             |
| 6 Random Forest                               | Used to identify the protein biomarkers in non-small cell lung cancer for early stage asymptomatic patients (Cancer Genomics Proteomics. 16:229–44, 2019).                                                                                                        |
| 7 RF Recursive Feature Elimination            | Applied to preliminarily screening the differential proteins for discovering biomarkers of hepatocellular carcinoma (J Proteomics. 225:103780, 2020).                                                                                                      |

Implies a strong correlation between features that affect the target vectors simultaneously. The multicollinearity problem could lead to instability of some feature selection method [60]. The type of filter methods is relative of great medical simplicity compared to other wrapper and embedded methods. In particular, these filter methods can be organized into two categories based on the number of variables analyzed: (1) univariate and (2) multivariate methods. The former type refers to methods where only one variable is used for filter analysis, which includes Fold Change (FC), Entropy-based Filters, Analysis of Variance (ANOVA), Chi-square ($\chi^2$), Euclidian Distance, T-test, Information Gain (IG), Significance Analysis of Microarrays (SAM), Linear Models for Microarray Data (LIMMA) and Mann–Whitney–Wilcoxon test (MWW test). The latter type refers to methods where at least two variables are considered for building modes, which include Correlation-based Feature Selection (CFS), Markov Blanket Filter (MBF), Linear Discriminant Analysis (LDA), Partial Least Square Discriminant Analysis (PLS-DA), Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA), Sparse Partial Least Squares Discriminant Analysis (sPLS-DA), Discriminant Analysis of Principal Component (DAPC) and Relief. As reported previously, the univariate methods could be used to evaluate a single biomarker with respect to recurrence and outcome, and multivariate methods were employed to develop a prognostic classification panel for disease recurrence [61]. The application of univariate and multivariate methods usually results in numerous significance features. Advantages of filter techniques are computationally simple.
and fast, and they are independent of the classification algorithm. A common disadvantage of filter methods is that they ignore the interaction with the classifier. Explanation and summary on each filter method are provided in Table 2. And the detailed descriptions and the corresponding applications of cancer-regulated studies are as follow:

3.1.1. Univariate filter methods

Fold Change (FC) is to compare the absolute value change between means of two groups, and it is calculated as the ratio or log of the ratio of the mean metabolite levels between two groups. A threshold needs to be defined, if fold change value exceeds this threshold, the variable will be reported as significant. FC has been widely applied in multiple omics analysis to predict survival of patients with metastatic colorectal cancer [62], and detect changes in the metabolomics profiles of men before and after androgen deprivation therapy for prostate cancer [63].

Entropy-based Filters are filter-based feature ranking techniques which include three classes: Information Gain (IG), Gain Ratio (GR) and Symmetrical Uncertainty (SU). IG selects features based on contribution of information related to the class variable without considering feature interactions. GR is non-symmetrical measure that is introduced to compensate for the bias of the IG, and SU criterion compensates for the inherent bias of IG [64]. Entropy-based Filters have been applied to identify the features related to glioma subtypes [65] and discover biomarkers in Arabidopsis thaliana [66].

A collection of statistical models and their associated estimation procedures was developed by Ronald Fisher called Analysis of Variance (ANOVA) [67]. This method mainly focuses on analyzing the differences among different groups average or variance values in a specific protein [68,69]. When comparing with multiple groups average values for statistical significance, it is more conservative than the traditional multiple two-sample t-tests method, and is therefore applicable for an extensive range of practical problems [70]. The ANOVA has been applied to investigate pro-inflammatory cytokines and liver functional markers [71] and explore the cellular resistance through quantitative proteomic analysis [72].

When a hypothetical feature is actual independent of the class value, Chi-square ($\chi^2$) becomes a popular statistical test for weighting divergence distribution [73]. It is known that the $\chi^2$ test is used to judge the independence of two events and its behavior is erratic for very small counts of cases [74]. Using the $\chi^2$ statistic for feature selection is similar to importing a hypothesis testing about the distribution of classes [75]. Chi-square has been applied to study the incidence of injury to parathyroid glands during surgery for papillary thyroid carcinoma [76]. Main function of t-test is to judge whether the average of the two groups is statistically different [77]. T-test is suitable for samples with a normal distribution [78]. Commonly, the log transformation was needed for obtaining a more symmetric distribution prior to t-test analysis. Moreover, normalization is also a necessary step for proteomics analysis that aims to reduce systematic bias and make samples more compara-

![Feature selections applied in cancer proteomics and popular sample classification methods.](image-url)
ble [79–84]. As reported, the variance stabilization normalization (VSN) method has a built-in transformation and could perform well in the statistical analysis [59,85]. However, t-test with multiple testing correction might not be appropriate when studied sample size is too small (less than 6) [16]. The Multiple testing correction is a necessary step to manage the problem of false positive when performing multiple statistical tests in proteomics study [38]. Popularly applied multiple testing correction methods included the Bonferroni correction, Holm correction and Benjamini-Hochberg correction. The t-test has been applied to distinguish prostate cancer from normal and benign conditions [87]. SAM can assign a grade to every protein according to the distribution of individual proteins [87]. SAM has been applied to identify specific serum proteome features associated with hepatocellular carcinoma [89].

A package called Linear Models for Microarray Data (LIMMA) focuses on differential expression gene analysis of file generated by simple or complex microarray experiments [90]. Matching a linear model to the gene or protein expression data is a central idea of this method. The functions of input and normalization can be performed with the small variables sets. The advantages of sPLS-DA are that variable selection and modeling are allowed in one step and can address multiclass problems.

Table 2
Fifteen filter feature selection methods available and popular in cancer proteome study.

| Methods                                         | Abbr. | Packages in R (Function) | Brief Descriptions                                                                 | Reference          |
|-------------------------------------------------|-------|--------------------------|------------------------------------------------------------------------------------|--------------------|
| Fold change                                     | FC    | metabolomics             | The FC is a popular and simple statistical method for the absolute value change between means of two groups and it is often applied together with other parametric/non-parametric methods | Clin Cancer Res. 18:3677–85, 2012 |
| Entropy-based Filters                           | Entropy | FSelector (entropy, based) | The Entropy filter method is similar to the Chi-square, which screens variables regardless of possible interactions. And its performance depended on the characteristics s of studied data. | Comput Chem Eng. 22:613–626, 1998s |
| Analysis of variance                            | ANOVA | ANOVA,TFNs (janova)      | The ANOVA is a powerful statistical method of parametric/non-parametric and linear. And focus on analyzing the differences among multiple groups average values in a specific protein. The \( \chi^2 \) is a popular statistical test of non-parametric, and difficult interpreting when the present of a large number of categories in the variables. | Br J Math Stat Psychol. 68:23–42, 2015 |
| Chi-square                                      | \( \chi^2 \) | stats (chiq, test)       | The \( \chi^2 \) test is parametric method to judge whether the average of the two groups is statistically different. The SAM is non-parametric and linear statistical technique. And rely on samples when carries out variable specific t-tests and ignore single variable | Am J Orthod Dentofacial Orthop. 150:1066–1067, 2016 |
| Student’s t-distribution                       | T-test | stats (t.test)           | The T-test is parametric method to judge whether the average of the two groups is statistically different. | J Natl Acad Sci USA. 98:5116–21, 2001 |
| Significance Analysis of Microarrays            | SAM   | same (sam)               | The SAM is non-parametric and linear statistical technique. And rely on samples when carries out variable specific t-tests and ignore single variable | Nucl Acids Res. 43:e47, 2015 |
| Linear Models for Microarray Data               | LIMMA | limma (lmFit)            | The LIMMA is an advanced statistical method based on parametric and linear models. And analyze the reverse-phase protein array data | Methods Mol Biol. 1362:175–84, 2016 |
| Linear discriminant analysis                    | LDA   | MASS (lda)               | The LDA is statistical algorithm based on the linear theory existing among different variables. And it is well suitable for study the small number of features. | J Wound Ostomy Continence Nurs. 24:12, 1997 |
| Mann–Whitney–Wilcoxon test                      | MWW   | test (wilcox, test)      | The MWW test is a nonparametric test of the null hypothesis. And it is well suitable when the assumptions of the t test or the data is ordered are not met | Information Sciences. 282:111–135, 2014 |
| Correlation-based feature selection             | CFS   | FSelector (cf)           | The CFS is a parametric and linear algorithm that have a heuristic search strategy. Its goal is to finalize a subset of the characteristics of these groups that are less correlated but highly. | Methods Mol Biol. 1362:175–84, 2016 |
| Markov blanket filter                           | MBF   | MXX (mmmmb)              | The MBF is a based on discretized features selection statistical method, which can be applied for obtaining good classification performance with the small variables sets. | Med Phys. 42:2421–30, 2015 |
| Partial Least Square Discriminant Analysis      | PLS-DA | ropls (opls)             | The PLS-DA is supervised multivariate statistical approach and belongs to type of linear two-class classifier. And can be maximized interval between preformed classes. | Chemometrics&Intelligent Laboratory Systems. 58:2:199–130, 2001. |
| Orthogonal Partial Least Squares Discriminant Analysis | OPLS-DA | ropls (opls)             | The OPLS-DA is an advanced multivariate statistical method for one or more classes problems. And it is well suitable for classification of proteomic dataset when the variables often existed multi-collinear and noisy problems. | J. Chemometrics. 20:341–351, 2006 |
| Sparse Partial Least Squares Discriminant Analysis | sPLS-DA | ropls (opls)             | The sPLS-DA is an advanced multivariate statistical algorithm for screening the variables and may miss out on variables between small sample sizes. The advantages of sPLS-DA are that variable selection and modeling are allowed in one step and can address multiclass problems. | BMC Bioinformatics. 12:253, 2011 |
| Discriminant Analysis of Principal Components   | DAPC  | adegenet (dpc)           | The DAPC is nonparametric and linear method. DAPC retains all property of DA without being dragged down by its restrictions. | BMC Genet. 11:94, 2010 |
As reported, its performance would be dramatically improved with sample size increasing to a certain degree (more than 12) when compared with the other methods (e.g., PCDA) [16,99]. The MWW test has been applied to identify candidate biomarkers of prostate cancer in proteomic study [100]. This method is implemented based on the linear theory existing among features. The LDA has been applied to identify a set of physicochemical properties for facilitating the classification of metal-based colloids [101].

3.1.2. Multivariate filter methods

Feature subsets among the features can be selected by Correlation-based Feature Selection (CFS) relying on the degree of redundancy [102]. The evaluator’s goal is to finalize a subset of the characteristics of these groups that are less correlated but highly related to the class [103]. In particular, the subset evaluators search iteratively and add features using a numeric measure, for instance, conditional entropy [104]. As a multivariate filter, the CFS considers the relationship between features not noticed by non-univariate filters [103]. CFS is applied to select first-rate feature subset and is bonded with search methods, for example, bidirectional search, forward selection, genetic search and best-first search [105]. CFS has been used in the subtype discovery of pediatric acute lymphoblastic leukemia by cancer microarray file [106]. Based on discretized features, a feature selection method named Markov Blanket Filter (MBF) is raised [107]. As a critical evaluation of MBF, contradictory and counterintuitive nature leads to undesirable properties on the small sample size applications which have function of classifying for microarray gene expression data [108]. Linear Discriminant Analysis (LDA) is applied to identify differences or discriminant features among different samples groups and pattern classification [109], which is quite suitable for studying the small number of features [110].

Swedish statistician Herman O. A. Wold and his son Svante Wold proposed and developed Partial Least Square Discriminant Analysis (PLS-DA) [111]. This method belongs to linear two-class classifier and maximize interval between predefined classes [112]. The PLS-DA method will not produce the most accurate decision boundary once the sample sizes are unequal [111]. This method is now usually applied by chemometricians who incorporate it into most pack-ages and is therefore cited in metabolomics papers [113]. The PLS-DA combined other univariate methods has been applied to select the discriminative proteins between the medullary sponge kidney and idiopathic calcium nephrolithiasis patients [114].

A supervised multiple regression analysis called Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) can be applied to identify the difference between different data sets of X and Y [115]. The OPLS methods is an extended Partial Least Squares (PLS) method that combines an orthogonal signal correction filter to discriminate data changes from predictive quantitative response predictions that are useful from orthogonal predictions [116]. This method is a powerful tool to analyze qualitative data structures and predicted results are similar to the results of standard PLS-DA [117]. OPLS-DA has a primary advantage as for interpretation of the models due to its predictability [118]. The OPLS-DA has been applied to identify candidate biomarkers significantly discriminated cholangiocarcinoma from normal and periductal fibrosis patients [119]. Moreover, similar to the partial least squares discriminant analysis, the Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) aims to achieve discrimination analysis based on PLS regression [120]. When adding a Lasso penalization to select variables, sPLS-DA can select variable and reduce dimension simultaneously [120]. sPLS-DA may tend to miss out on variables that only distinguish between small sample sizes, assuming effect size [121]. The advantage of sPLS-DA is that variable selection and modeling are allowed in one step, and interpretability is improved through valuable graphical output [122].

In 2010, Jombart and colleagues proposed Discriminant Analysis of Principal Components (DAPC) for inferring the number of genetically related individuals [123]. DAPC and previously mentioned PLS-DA choose small-size feature sets precisely but miss many true positive features relevant to the spiked proteins [16]. This multivariate statistical approach divides variance of the sample into between-group and within-group to maximize discrimination between groups. DAPC first converts the data using Principal Component Analysis (PCA) and then uses Discriminant Analysis (DA) to identify the cluster [123]. As a new methodological approach, DAPC retains all property of DA without being dragged down by its restrictions [123]. A major advantage of DAPC in the Bayesian Cluster approach is the ability to generate graphs that infer kinship between clusters. DAPC can be used to decompose potential structures in more complex demographic models [124].

3.2. Wrapper methods for tumor marker identification from proteomic data

The wrapper method is one of the most popular feature selection methods proposed by Kohavi and John in 1997 [125]. In the wrapper method, a search for an optimal set of features is made using the induction algorithm as a black box [126]. It provides a simple and powerful technique to solve the challenge of feature selection. The wrapper methods look for the best subset of features based on their predictive power. Wrapper methods are often criticized because they train various new subsets into corresponding models, and need a large amount of computation, but it is not necessarily the case. Researchers can design an effective search strategy. Using this strategy does not necessarily mean sacrificing predictive performance. But it usually provides the best performing feature set for a particular type of model. Generally, different wrapper methods are used to select the best subset. The commonly used methods are Genetic Algorithms (GA), Hill Climbing (HC), Sequential Backward Elimination (SBE), Estimation of Distribution Algorithm (EDA), Simulated Annealing (SA), and Sequential Forward Selection (SFS). Explanation and summary on each wrapper method are provided in Table 3. And the detailed descriptions and the corresponding application of cancer-regulated studies are as follow:

A valid and commonly applicable function minimization method is the Genetic Algorithm (GA) [127]. This method achieves the results of global optimization based on the genetic theory of “survival of the fittest”. Unlike traditional search methods, GA begins with a set of finite-length encoded alphabet strings not a real set of parameters [128]. Selection, crossover, and mutation are operators of GA [129]. Through the selection step, a number of individuals which have strong adaptability are found out from the group, and then the individual copy number is determined based on the selection method [130]. Generally speaking, it is necessary to select the suitable crossover and mutation probabilities based on practical problems [131]. GA had been used to maximize functional value in chronic dialysis patients [132]. Moreover, the most famous local search algorithm may be Hill Climbing (HC) [133]. The main idea of climbing is divided into three steps: first, randomly generate a state, and then find the best assessment value to transfer to the neighbor, and finally if the third step has reached a rigid local minimum, restart other randomly created states. These steps need to be repeated until the solution or a local optimum is found [134,135]. The HC has been applied to predict the protein structure prediction combining with genetic algorithm [136].
tive function $J(Y-x-)$ [137,138]. These features are considered non-monotonic. Because SBS is addressing large subsets in most cases, it performs best while the first-rank feature subset exists many features [139]. The primary disadvantage of SBS is that it is not possible to re-evaluate the reusability of this feature when a feature is removed [140]. Sequential Forward Selection (SFS), a bottom-up search technique, works in opposite directions from SBE. SFS is first based on a null feature set and then gradually adds features picked by some assessment functions to minimize Mean Square Error (MSE) [141]. Sharma et al. proposed SFS [142] and attempted to conquer the shortcoming of traditional feature selection method whereby a weakly ranked protein that could condense to classification accuracy with a suitable subset of proteins has eliminated from the selection [143]. In each iteration, the remaining available features that do not exist in the feature set are selected into the feature set [144]. Therefore, the extended feature set will generate a minimal classification deviation compared to the previously added features [145]. Despite this, SFS is widely applied because of its simplicity and the fact that it only contains F core steps, and many variants and applications are proposed based on the SFS.

Estimation of Distribution Algorithm (EDA), also named Probabilistic Model-building Genetic Algorithm (PMBGA), is a widely used random optimization method [146]. Firstly, EDA learns the explicit probability model which is a promising solution that is currently found, and then generates a new solution by sampling the created model [147]. EDA belongs to a class of evolutionary algorithm. Unlike most traditional evolutionary algorithms which use implicit distributions defined by one or more variant operators to generate new candidate solution, EDA uses Bayesian networks multivariant normal [148]. EDA is an excellent solution to a number of challenging problems [149]. EDA now has been applied to the HP model on a cubic lattice and predict the native structures of relatively small proteins therefore helps in drug design and proteomics [150–152]. Moreover, the Simulated Annealing (SA) is a specific method for a number of type optimization problems [153]. SA has now been applied to help identify HER2 and hormonal receptor expression states of breast cancer patients [154].

### 3.3. Embedded methods for tumor marker identification from proteomic data

Embedded methods are commonly used in feature selection in order to select the optimal subset of features which play a great role in the process of constructing the classifier [155]. A low minimum sample size of around 20 samples was suggested as sufficient to perform robust power analysis [156]. It is clear that sample imbalance has a great effect on identifying differential

### Table 3

| Method | Abbr. | Packages in R (Function) | Brief Description | References |
|--------|-------|---------------------------|------------------|------------|
| Sequential Forward Selection | SFS | Dprep (sff) | SFS is the only method containing F core steps of linear parameter and is a bottom-up search technique. Many variants and applications are proposed based on SFS. | Pattern recognition and signal processing 41–60, 1978 |
| Sequential Backward Elimination | SBE | MAES (stepAIC) | SBE is a linear and parameter algorithm that performs well due to its addressing of large subsets. SBE works in opposite direction. | Pattern recognition and signal processing 41–60, 1978 |
| Simulated Annealing | SA | GenSA (GenSA) | SA is a specific parameters and linear method of approximating the global optimization problems. It is strict with various conditions, but sometimes the cost-increasing can be acceptable. | Soviet Phys Cryst. 5:905–916, 1979 |
| Hill Climbing | HC | FSelector (hill.climbing.search) | HC is a local search linear and parameter algorithm, which does not need to find the global maximum. Combining with genetic algorithm, HC has been used to predict protein structure. | Proteome Sci. 1:519, 2011 |
| Genetic Algorithm | GA | GA (ga) | GA is an effective and universally applicable function of parameters and linear/unlinear, which does not scale well with complexity. | Mitochondrial DNA 25:231–7, 2014 |
| Estimation of Distribution Algorithm | EDA | Copulaedas (Copula) | EDA is a parameter approach and a widely used random optimization. It belongs to the class of evolutionary algorithms. | ISCE: 3:1787–97, 2011 |

### Table 4

| Methods | Abbr. | Packages in R (Function) | Brief Descriptions | Reference |
|---------|-------|--------------------------|-------------------|-----------|
| Decision Tree | DT | dtree (pce) | The DT is a supervised algorithm for classification and prediction. And it is suitable for nonlinear proteomic data. | Machine Learning. 4:161–86, 1989 |
| Support Vector Machine | SVM | e1071 (svm) | The SVM is a supervised learning model for classification and regression. And it is suitable any linear or nonlinear proteomic data, but also complex and time consuming for large scale proteomic data. | Bioinformatics. 21:47–56, 2005 |
| Weighted naive Bayes | WNB | CORElearn (CoreModel) | The WNB is a supervised learning method based on the Bayes theory. And it is only suitable for the proteomic data, where all variables are independent of each other. | Appl Environ Microbiol. 73:5261–7, 2007 |
| SVM Recursive Feature Elimination Artificial Neural Networks | SVM-RFE | MCRestimate (varSel.svm. rfe) | The SVM-RFE is popular supervised embedded method. And it is well suitable for addressing overfitting risk when the number of variables is high. | Anal Chem. 80: 7562–70, 2008 |
| Random Forest | AN | neuroanet (neuralnet) | The ANN is supervised embedded method. And its performance on consistent depended on the characteristic of studied data. | Nat Rev Urol. 10:174–82, 2013 |
| RF Recursive Feature Elimination | RF | randomForest (randomForest) | The RF is popular supervised embedded method. And it is well suitable for any proteomic dataset for identifying the small variables set. | Isprs Journal of Photogrammetry & Remote Sensing. 67:93–104, 2012 |
| | RF | varSelRF (varSelRF) | The RF-RFE is popular supervised embedded method. And it is well suitable for addressing the non-linear or linear variables of proteomic dataset. | Metabolomics. 18: 3677–85, 2012 |
features. For example, the imbalance sample in the dataset could lead to model overfitting or underfitting. To deal with the overfitting problem, the dataset could be randomly divided into training (60%), validation (20%), and test (20%) data sets, or training (70%) and test (30%) data sets in case validation is not required [157].

In this section, six methods are listed as follows: Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), Bayesian Classifier and Artificial Neural Network (ANN), Random Forest with Recursive Feature Elimination (RF-RFE) and Support Vector Machine with Recursive Feature Elimination (SVM-RFE) [84]. Explanation on each method is provided in Table 4, and detailed descriptions and the corresponding application of cancer-regulated studies are as follow:

Decision Tree learner is a common decision support tool that resembles a tree structure in which each non-leaf node represents a test on a specific characteristic, each branch means the result of the test, and each leaf node denotes a class label [158]. Through an iterative process, the Decision Tree learns from a series of training examples, selects an attribute and then splits a given set of examples based on the values of that attribute [159]. Advantages of the method can be summed up in two aspects. First, the construction of the classifier does not require much of professional knowledge and its reliability in diagnosis can be verified through both expert knowledge and testing data [160]. Besides, its inductive classification step is simple and fast [159].

Random Forest developed by Leo Breiman is an algorithm that integrates multiple trees through the idea of integrated learning [161]. Its basic unit is Decision Tree which takes the advantage of a bootstrap sample of the data [162]. The method is excellent in accuracy [163] and generates an internal unbiased estimate of the generalization error as the forest building progresses. Moreover, when comparing with SVM, it performs better in the aspect of classification tasks [164]. Random Forest with Recursive Feature Elimination (RF-RFE) conducts a recursive backward feature elimination procedure. It begins with all the features. In each iteration, a Random Forest is constructed to measure the features’ importance and the feature with least importance is removed. This procedure is repeated until there is no feature left. Finally, the features are ranked according to the deleted sequence, and the top ranked feature is the last deleted one. RF-RFE has been applied to reveal the deregulation of fatty acids metabolism in hepatocellular carcinoma and chronic liver [165,166], and investigate hepatitis C by using metabolomics as a tool [63].

Support Vector Machine (SVM) is a kind of learning machines constructed according to structural risk minimization principle on the basis of Vector Machine theory, which hopes to find a segmentation surface that can separate the data points apart [167]. One of the challenges of this method is how to choose the proper kernel function [168]. When it comes to the advantages of SVM, it performs well in dealing with high-dimensional data sets with very few training examples [169]. Moreover, as family of simple "probabilistic classifiers" based on Bayes theory, the Bayesian Classifier can be applied to predict the probability that a given instance pertains to a class [170]. It assumes that all the attributes are independent of each other [171]. In theory, the naive Bayesian classifier has the minimum error rate when compared to other classification methods. However, this is not always the case in practice due to the above assumption. Even so, the Naive Bayesian classifier has the high precision and speed in the application of large database [172].

Support Vector Machine with Recursive Feature Elimination (SVM-RFE) is an embedded approach that uses the norm of the weights w to rank the variables. Initially, all data is taken, and a classifier is computed. The norm of w is then computed for each of the features and the feature with the smallest norm is eliminated. This process is repeated until all features are ranked. For selecting features relevant to the spiked compounds, linear SVM-RFE performs poorly even if the classification error is relatively low [16]. SVM-RFE has been applied to determine differences between healthy subjects and patients suffering from Streptococcus pneumoniae [173], and interrogate the serum metabolome of early-stage ovarian cancer patients and age-matched control women [174].

Artificial Neural Networks (ANN) is an algorithm mathematical that simulates the behavior characteristics of biological neural networks, including their structure and functionalities [175]. The basic unit of every artificial neural network is artificial neuron which is a simple mathematical function [176]. This kind of network relies on the complexity of the system and can process information by adjusting the weight of the interconnection between a large number of nodes (neurons) within the system [177]. Requiring less formal statistical training, ANN performs well in detecting the complex non-linear relationship between dependent and independent variables [178].

In addition to the above feature selection methods, deep learning is also a type of machine learning approaches which enables to identify the highly complex patterns in omics data [179–181]. Common deep learning methods included the convolutional neural networks, recurrent neural networks and deep neural networks [182]. Until now, deep learning has been widely applied in cancer proteomics [183]. For example, the deep learning approach has been applied for the diagnosis of pancreatic cancer in proteomic data [183].

4. Sample clustering and visualization

Clustering or unsupervised modeling is a powerful technique for discovering the molecular classification of patient’s cancer tissues [184] and identifying subtype specific characterization (e.g. subtype specific cancer therapeutic targets) [185,186]. The underlying theories of clustering technique are considering the similarities among different samples based on high dimension proteins or peptides, which can provide information of the same or differential patterns on protein expression level, aiming at clustering those biologically similar samples together [187]. Therefore, the unsupervised approaches are widely applied to characterize and analyze the complex proteomic dataset. Currently, there are five different approaches for sample separation of cancer proteomic dataset, including three clustering methods and two popular reducing dimension methods [188]. In cluster analysis, Hierarchical Cluster Analysis (HCA), Self-organizing Map (SOM) and K-means Clustering are three prominent representatives in the analysis of proteomic data [189]. Principal Component Analysis (PCA) and the state of art stochastic neighbor embedding analysis based on Student t Distribution (t-SNE) [190,184] are the most commonly used unsupervised approaches of reduction dimensions in cancer proteomic studies [188]. Explanation on these methods is provided in Table 5. Detailed descriptions and the corresponding application of cancer-regulated studies are as follow.

The Principal Component Analysis (PCA) aims at interpreting the first few principal components that often explain thousands of the variables, and is often applied strategy of reducing dimension for high dimensional cancer proteomic data [191]. Particularly, a multivariate proteomic dataset can first be transformed based on the linear relationship between different proteins or peptides [188], and then determine the most significant Principal Components (PCs) according to the obtained variances [191]. In PCA analysis, the scatterplots (e.g. scores plot and loadings plot) are often used for visualizing correlation of the biological samples or the importance of proteins/peptides to discriminate those differential samples, which are often visualized based on two axes (two
Principal Component Analysis

- **Abbr.**: PCA
- **Packages in R (Function)**: ropls (pco)
- **Brief Descriptions**: The PCA is unsupervised, non-parametric and linear method of reduce dimension method. And unsuitable the dataset when the proteins or peptides were nonlinear relationship.
- **Reference**: Clin Biochem. 56:55–61, 2016

**t-Distributed Stochastic Neighbor Embedding**

- **Abbr.**: t-SNE
- **Packages in R (Function)**: Rtsne (Rtsne)
- **Brief Descriptions**: The t-SNE is unsupervised, non-parametric and nonlinear method of reduce dimension method. And unsuitable the dataset when the proteins or peptides were linear relationship.
- **Reference**: J Mach Learn Res. 9:25579–605, 2008

**Self-Organizing Map**

- **Abbr.**: SOM
- **Packages in R (Function)**: som (som)
- **Brief Descriptions**: The SOM is unsupervised, non-parametric and nonlinear method of clustering. And suitable for any proteomic the dataset.
- **Reference**: IEEE Trans Neural Netw. 11(3):574–85, 2000

**Hierarchical Cluster Analysis**

- **Abbr.**: HCA
- **Packages in R (Function)**: Cluster (agnes)
- **Brief Descriptions**: The HCA is unsupervised, non-parametric and nonlinear method of clustering. And suitable for the any datasets.
- **Reference**: Front Bioeng Biotechnol. 3:23, 2015

**K-Means Clustering**

- **Abbr.**: K-Means
- **Packages in R (Function)**: skmeans (skmeans)
- **Brief Descriptions**: The K-Means is unsupervised, non-parametric and nonlinear method of clustering. And unsuitable the dataset when the number of clusters is not known.
- **Reference**: Proteomics. 16:1613–21, 2016

5. Conclusions

Tumor marker selection and sample classification are key issues in current cancer proteomics study. The biomarker sets identified via feature selection methods were inconsistency in several publications. This article reviewed the popular feature selection methods currently applied in tumor maker selection and discussed several sample classification algorithms available for cancer proteomic. These advances made the MS-based proteomics technology more widely applied to identify diagnostic, prognostic, and therapeutic biomarkers for anticancer drug discovery. Moreover, the pathogenetic process of a specific disease could be comprehensively and better understood using a panel of biomarkers, under this circumstance, the wrapper and embedded feature selection methods were suitable for selecting a small the feature panel in the tumor marker identifications.

Consent for publication

Not applicable.
Ethics approval
Not applicable.

Authors’ contributions
F.Z. conceived the idea and supervised the work. J.T., Y.W. and Y. L. performed the research and wrote the manuscript. J.F., Y.Z., Y.L., Z.X., Y.L. and Y.Q. provide the review of cancer proteomics. All authors read and approved the final manuscript.

Declaration of competing interests
The authors declare that they have no competing interests.

Acknowledgements
Funded by the National Key Research and Development Program of China (2018YFC0910500), National Natural Science Foundation of China (81872798 & U1909208), Fundamental Research Funds for Central University (2018QNA7023, 10611CDJXX238826, 2018CDQYSGG007 & CDJZR14468081), Key R&D Program of Zhejiang Province (2020C03010), & Leading Talent of “Ten Thousand Plan” - National High-Level Talents Special Support Plan.

References
[1] Malvezzi M, Caroli G, Bertuccio P, Negri E, La Vecchia C. Relation between mortality trends of cardiovascular diseases and selected cancers in the European Union. 2019. Focus on cohort and period effects. Eur J Cancer 2018;101:341–55.

[2] Arora D, Chaudhary R, Singh A. System biology approach to identify potential receptor for targeting cancer and biomolecular interaction studies of indole [2,1-a]isoquinoline derivative as anticancerous drug candidate against it. Interdiscip Sci Comput Life Sci 2017;11:125–34.

[3] Reddy MM, Kar SS. Unconditional probability of dying and age-specific mortality rate because of major non-communicable diseases in India: time trends from 2001 to 2013. J Postgrad Med 2018;65:12.

[4] Guo Y, Wang H, Li Y, Song Y, Chen C, Luo Y, et al. Genome of Helicobacter pylori strain XX274, an isolate from a tibetan patient with gastric cancer in China. J Bacteriol 2012;194:4146–7.

[5] Fu J, Tang J, Wang Y, Cui X, Yang Q, Hong J, et al. Discovery of the consistently well-performed analysis chain for SWATH-M5 based pharmacometric quantification. Front Pharmacol 2019;8:681.

[6] Collins LG, Haines C, Perkel R, Enck RE. Lung cancer: diagnosis and management. Am Fam Physician 2007;75:56–63.

[7] Yilmaz S, Dursun M, Canoruc F, Yilmaz S, Dursun M, Canoruc F. A patient with gastric cancer tumor: treatment and surveillance options. Turk J Gastroenterol 2005;16:180–1.

[8] Zhang L, Ozzo J, Warner R, Divino C. Review of the pathogenesis, diagnosis, and management of type 1 gastric cancer tumor. World J Surg 2011;35:1879–86.

[9] Zhang A, Sun H, Yang S, Wang P, Han Y, Wang X. Metabolomics in diagnosis and biomarker discovery of colorectal cancer. Cancer Lett 2014;345:17–20.

[10] Singh S, Singh DB, Singh A, Gautam B, Ram G, Dwiwedi S, et al. An approach for identification of novel drug targets in Streptococcus pyogenes SF370 through pathway analysis. Interdiscip Sci 2016;8:388–94.

[11] Ahmad S, Naivid A, Akhtar AS, Azam SS, Wadood A, Perez-Sanchez H. Subtractive genomics, molecular docking and molecular dynamics simulation revealed LpxC as a potential drug target against multi-drug resistant Klebsiella pneumoniae. Interdiscip Sci 2018;1:19.

[12] Li S, Li J, Ning L, Wang S, Niu Y, Jin N, et al. In silico identification of protein S-palmitoylation sites and their involvement in human inherited diseases. J Chem Inf Model 2015;55:2015–23.

[13] Qu KY, Gao F, Guo F, Zou Q. Taxonomy dimension reduction for colorectal cancer prediction. Comput Biol Med 2019;103:180848.

[14] Liao ZJ, Li DP, Wang XR, Li LS, Zou Q. Cancer diagnosis through isomeric expression with machine learning method. Curr Bioinform 2018;13:57–63.

[15] Li YH, Li XX, Hong J, Wang YX, Fu JB, Yang H, et al. Clinical trials, progression-speed differentiating features and significance rule of the innovative targets of first-in-class drugs. Brief Bioinform 2020;21:649–62.

[16] Christin C, Hoefslott HC, Smilde AK, Hoefsloot HC, Smilde AK, Hoekman B. Supe...
Louis E, Adriaensen P, Guedens W, Vanhove K, Vandeuren K, Darqueennes K, et al. Metabolic phenotyping of human blood plasma: a powerful tool to discriminate between cancer types?. Ann Oncol 2016;27:178–84.

Duan L, Yobas L. Label-free multiplexed electrical detection of cancer markers on a microchip featuring an integrated fluidic diode array. ACS Nano 2018;12:7892–900.

Bertini I, Chanfreau H, Martinez D, Garcia D, Lacunza E, Abba MC. BioPat: a software for human cancer biomarker discovery. Bioinformatics 2014;30:1782–4.

Zdunski K, Ziołkowski P, Alhind C, Agrawal A, Agrawal S, Blomqvist C, et al. Nuclear osteopontin-c is a prognostic breast cancer marker. Br J Cancer 2015;112:739–38.

Shen W, Le S, Li Y, Hu F. SeqKtit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE 2016;11:e0163962.

Tang Y, Mou M, Zhang L, Liu Y, Luo Y. MetaFS: performance assessment of biomarker discovery in metabolomics. Brief Bioinform 2020. https://doi.org/10.1093/bib/bbaa105.

Ybarra N, Jeyaseelan K, Faria S, Kopek N, Brisebois P, et al. Bayesian subtyping of breast cancer patients. J Thorac Cardiovasc Surg 2012;144:1344–50.

Koller D, Sahami M, editors. Toward optimal feature selection. Thirteenth International Conference on Machine Learning. AAAI Press 1996:249–56.
