BBA - Molecular Basis of Disease xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

BBA - Molecular Basis of Disease

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

Identification of the functional alteration signatures across different cancer types with support vector machine and feature analysis☆

ShaoPeng Wang, YuDong Cai*

School of Life Sciences, Shanghai University, Shanghai 200444, People's Republic of China

ARTICLE INFO

Keywords:
Cancer prediction
Support vector machine
Gene ontology
KEGG
Monte-Carlo feature selection

ABSTRACT

Cancers are regarded as malignant proliferations of tumor cells present in many tissues and organs, which can severely curtail the quality of human life. The potential of using plasma DNA for cancer detection has been widely recognized, leading to the need of mapping the tissue-of-origin through the identification of somatic mutations. With cutting-edge technologies, such as next-generation sequencing, numerous somatic mutations have been identified, and the mutation signatures have been uncovered across different cancer types. However, somatic mutations are not independent events in carcinogenesis but exert functional effects. In this study, we applied a pan-cancer analysis to five types of cancers: (I) breast cancer (BRCA), (II) colorectal adenocarcinoma (COADREAD), (III) head and neck squamous cell carcinoma (HNSC), (IV) kidney renal clear cell carcinoma (KIRC), and (V) ovarian cancer (OV). Based on the mutated genes of patients suffering from one of the aforementioned cancer types, patients they were encoded into a large number of numerical values based upon the enrichment theory of gene ontology (GO) terms and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. We analyzed these features with the Monte-Carlo Feature Selection (MCFS) method, followed by the incremental feature selection (IFS) method to identify functional alteration features that could be used to build the support vector machine (SVM)-based classifier for distinguishing the five types of cancers. Our results showed that the optimal classifier with the selected 344 features had the highest Matthews correlation coefficient value of 0.523. Sixteen decision rules produced by the MCFS method can yield an overall accuracy of 0.498 for the classification of the five cancer types. Further analysis indicated that some of these features and rules were supported by previous experiments. This study not only presents a new approach to mapping the tissue-of-origin for cancer detection but also unveils the specific functional alterations of each cancer type, providing insight into cancer-specific functional aberrations as potential therapeutic targets. This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

1. Introduction

Cancer is regarded as a malignant proliferative disease that can occur in many tissues and organs in humans [1,2]. As a systemic disease, the symptoms of cancer are not restricted to the sites of tumorigenesis [2]. The proliferative, invasive and metastatic characteristics of cancer have been associated with a high mortality rates [3–5]. In 2012, 14.1 million new cancer cases were diagnosed, and at the same time, approximately 8.2 million people died of such disease. Based on statistical prediction, by 2025, more than 19.3 million people may be diagnosed with cancer, demonstrating that cancer is one of the major threats to human life [6].

It is well known that the early diagnosis of cancers can greatly increase the chances of successful treatment and survival of patients. Cell-free DNA (cfDNA) has been recognized as a potential non-invasive cancer biomarker since the discovery of *TP53* mutations in the urinary sediments of bladder cancer patients and the detection of mutated *RAS* gene in the blood of cancer patients [7–9]. The liquid biopsy of cfDNA in plasma or serum could avoid the need for tumor tissue biopsies and allow the cfDNA to be monitored during the progression and the treatment of cancers. Information about the tissue-of-origin from the liquid biopsies are important for locating and diagnosing the primary cancers early but require knowledge of the cancer-specific or tissue-specific variations. For example, tissue-specific DNA methylation, cell-specific nucleosome occupancy pattern and cancer-specific mutation signatures are now available to characterize these biopsies [10–14].

☆ This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

* Corresponding author.
E-mail address: caiyudong@staff.shu.edu.cn (Y. Cai).

https://doi.org/10.1016/j.bbadis.2017.12.026
Received 17 October 2017; Received in revised form 4 December 2017; Accepted 15 December 2017

Meanwhile, specific mutation patterns have been identified as genetic characteristics to identify tumor types. For example, *ALK* gene rearrangement and its over-expression have been confirmed to be associated with non-small cell lung cancer and anaplastic large cell lymphoma [15]. Therefore, *ALK* gene and its expression products may serve as a core biomarker for the diagnostic and prognostic evaluation of these two cancer types [16]. The identification and clinical application of confirmed tumor genetic markers (mutation patterns) provide a new method to diagnose tumor types and distinguish them from each other. However, these mutated genes or gene products do not function in isolation but interact with each other in cellular networks and processes [17]. Thus, it is a more robust approach to identify the core unique characteristics of various tumor types at the level of biological processes rather than mutation signatures.

Unlike genes that are represented by specific gene names and symbols in computational biology, the biological processes are described by multiple bioinformatics initiatives based upon different point cuts. There are two core bioinformatics initiatives that contribute to the identification and description of functional biological processes and pathways in humans and across different species: gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [18,19]. GO compiles bioinformatics initiative describing genes and gene products by clustering their interactions with each other and annotating their respective contribution to certain biological processes [18,20]. In addition, the KEGG pathways provide a new approach to investigate biological processes. KEGG pathway terms cluster the functional genes into identified functional pathways, reflecting the real contribution of such genes to the living organism [19]. Therefore, during the identification of core unique biological factors in different tumor types, both GO terms and KEGG terms evaluate the differences from the point of view of integrated biological processes in a more comprehensive and convincing manner.

In this study, we applied a pan-cancer analysis to five different types of cancers: (I) breast cancer (BRCA), (II) colorectal adenocarcinoma (COADREAD), (III) head and neck squamous cell carcinoma (HNSC), (IV) kidney renal clear cell carcinoma (KIRC), and (V) ovarian cancer (OV). We obtained the somatic mutations found in these five cancer types from TCGA (The Cancer Genome Atlas) through the cBio cancer genomics portal [21–23]. Based upon the obtained mutated genes, patients with each aforementioned cancer type were encoded into a large number of numerical values using the enrichment theory of GO terms and the KEGG pathway [24–27]. The Monte-Carlo Feature Selection (MCFS) method [28] was adopted to analyze the GO term features and KEGG pathway features, yielding a feature list and sixteen decision rules. This feature list was used for the incremental feature selection (IFS) method to discover the most appropriate features for building the optimal classifier using the classic machine learning algorithm, support vector machine (SVM) [29,30], which could distinguish the five types of cancers with the best performance. This optimal SVM-based classifier provided a Matthews correlation coefficient value of 0.523 and an overall accuracy of 0.619. With regard to the sixteen decision rules, they can provide more clues to understanding the specific functional alterations of each cancer type than the classifier mentioned above, although it yielded a low overall accuracy of 0.498. Finally, important GO terms and KEGG pathways involved in the decision rules and optimal SVM-based classifier were extensively analyzed according to previous experimental results. Our study not only shed light on the mapping of the tissue-of-origin for cancer detection but also classified the functional alteration signatures of the five types of cancers, providing insight into the cancer-specific functional aberrations as potential therapeutic targets.

Table 1
The number of samples in each of the five cancer types.

Cancer type	Full name	Number of samples
BRCA	Breast cancer	513
COADREAD	Colorectal adenocarcinoma	499
HNSC	Head and neck squamous cell carcinoma	306
KIRC	Kidney renal clear cell carcinoma	473
OV	Ovarian cancer	456
Total	–	2247

2. Materials and methods

2.1. Materials

The mutational data in different types of cancers were downloaded from the cBioPortal for Cancer Genomics (http://cbio.mskcc.org/cancergenomics/pancan_tcga/) [23], which contained the mutations in eleven cancer types. Because many cancer types only have very few samples compared with others, cancer types with less than 300 samples were excluded. The remaining five major cancer types included (I) BRCA, (II) COADREAD, (III) HNSC, (IV) KIRC, and (V) OV. The numbers of samples for these five cancer types are listed in Table 1.

2.2. The functional profiles of mutations

There have been many ways to describe a protein, such as the protein sequence based features [31] and secondary structure based features [32]. But the most direct one was the functional annotation of a protein from databases like GO and KEGG. There were limitations of direct binary annotation of whether a protein had a specific function. Such binary functional features will be very sensitive to the mis-annotations in the database. Therefore, the enrichment scores which considered the significance of overlap between a gene set and a GO or KEGG function in the genome background, will be more robust and give a quantitative measurement of function rather than a binary qualitative measurement [33]. In this study, we used the GO and KEGG enrichment scores [24–27] of mutated genes to measure the similarity of the functional effects caused by mutations between cancer patients.

2.2.1. GO enrichment score

For a given cancer patient p and one GO term GO_j, let G_{GO} denote the set of annotated genes of GO_j and $G(p)$ denote the set of mutated genes of cancer patient p. The GO enrichment score of p and GO_j is defined as the hypergeometric test P value [24–27,34–37] on $G(p)$ and G_{GO}, which can be computed with the following equation:

$$S_{\mathrm{GO}}(p, GO_j) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right) \quad (1)$$

where N and M denote the total number of human genes and the number of genes in G_{GO}, respectively; n and m represent the number of mutated genes in $G(p)$ and the number of genes both in $G(p)$ and G_{GO}, respectively. The higher the score, the stronger the functional effects of mutations in patient p on the GO term GO_j are. Overall, 19,997 GO terms were used in this study, inducing 19,997 GO enrichment scores for each cancer patient.

2.2.2. KEGG enrichment score

A similar approach was adopted to define the KEGG enrichment score, which can measure the associations between patients and KEGG pathways. Let G_{KEGG} denote the set of annotated genes of one KEGG pathway K_j. The KEGG enrichment score of p and K_j is defined as the hypergeometric test P value [24–27,34–37] on $G(p)$ and G_{KEGG}. This score can be calculated using the following equation:

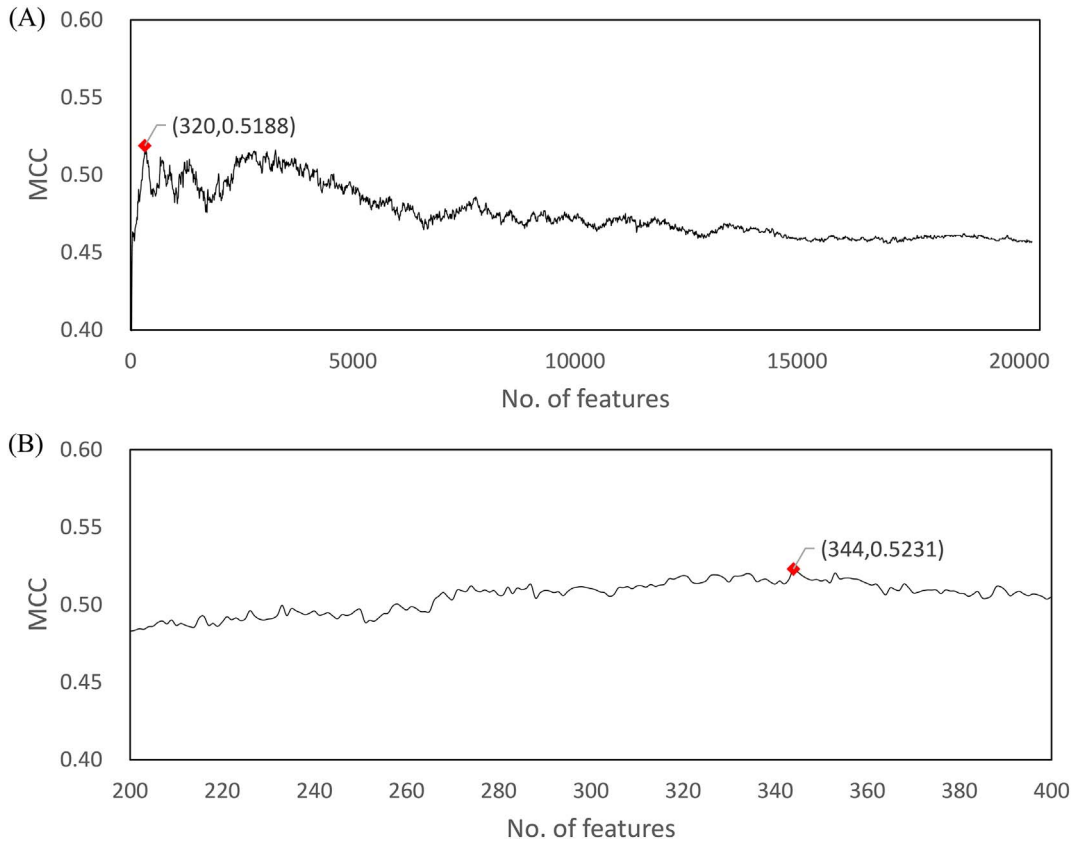

Fig. 1. The IFS curve illustrates the association between the number of features used in the SVM classifiers and *MCC* yielded by the corresponding classifier. (A) The feature number was set to be a multiple of ten. The highest *MCC* value is labeled with a red square on the curve. (B) The feature number was between 200 and 400. The highest *MCC* value is labeled with a red square on the curve.

$$S_{\mathrm{KEGG}}(p, K_j) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right) \tag{2}$$

where the parameters N and n have the same definitions as those in Eq. (1), while M and m denote the number of genes in G_{KEGG} and the number of genes both in $G(p)$ and G_{KEGG}, respectively. Similarly, a large KEGG enrichment score reflects the strong functional effects of mutations of patient p on the KEGG pathway K_j. Overall, 296 KEGG pathways were used in this study, resulting in 296 KEGG enrichment scores for each cancer patient.

Accordingly, based on the 19,997 GO enrichment scores and 296 KEGG enrichment scores, each cancer patient can be represented by 20,293 (19,997 + 296) enrichment scores. All computational analysis would be performed on these scores.

2.3. Feature ranking and decision rule identification

To extract important GO terms and KEGG pathways capable of distinguishing the five cancer types, some advanced computational methods are necessary due to the large-scale data. In this study, we applied the MCFS method [28] to analyze 20,293 features mentioned in Section 2.2, yielding a feature list and some decision rules for further analysis. Compared to some classic feature selection methods, such as minimal redundancy maximal relevance (mRMR) [38–44] and maximum relevance maximum distance (MRMD) [45,46], which always select key features by analyzing the original dataset, the MCFS method determines the importance of a feature by investigating its contribution for building decision trees. A detailed description and analysis of the MCFS method can be found in [28].

To evaluate each feature, the MCFS method first builds a large number of decision trees. Briefly, given an integer m that is much smaller than the total number of features, randomly construct s feature sets such that each set consists of m features. For each feature set, randomly select training samples and testing samples from the original dataset, on which a decision tree can be built. This procedure is executed t times, i.e., t decision trees are built for each feature set. Overall, $s{\cdot}t$ decision trees can be built. The importance of each feature g is evaluated by these decision trees, called relative importance (RI), which can be computed according to the following equation:

$$\mathrm{RI}_g = \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_g(\tau)} \mathrm{IG}(n_g(\tau)) \left(\frac{\text{no. in } n_g(\tau)}{\text{no. in } \tau}\right)^v \tag{3}$$

where $wAcc$ is the weighted accuracy of the decision tree τ, $\mathrm{IG}(n_g(\tau))$ is the information gain of the node $n_g(\tau)$, (no. in τ) is the number of samples in tree τ, u and v are fixed real numbers, and (no. in $n_g(\tau)$) is the number of samples in node $n_g(\tau)$. It is clear that a feature that is assigned a large RI value is more important. Thus, all features can be ranked in a list by the decreasing order of their RI values, i.e., features with high RI values obtain high ranks in the list. In this study, 20,293 features were considered. Using the MCFS method, we can obtain a feature list, denoted as *FL*, containing 20,293 ranked features.

In addition to constructing a feature list, the MCFS method can also yield some decision rules. Each sample can be classified following these rules. To obtain these rules, this method first extracts most informative features that are the top p% features in the feature list. Then, n subsets are generated by randomly selecting samples from the original dataset, where each sample is represented by informative features. Rule-based classifiers (e.g., the rough sets [47]) can be built for each dataset, which consist of a number of IF-THEN rules. As described in [48], the Johnson Reducer algorithm implemented by ROSETTA software is used to generate reducts (the minimal sets of features for a classification task) and

the associated rules for each randomly produced dataset. Here, some decision rules were obtained by executing the MCFS method on samples of the five cancer types. A detailed analysis of these rules can help us understand the specific functional alterations of each cancer type.

2.4. Classification algorithm

Using the MCFS method, a feature list can be obtained. The IFS method can be performed to select important features for building the optimal classifier. This method first constructs a series of feature sets, say F_1, F_2, …, F_N, where F_i contains the first i features in the ranked feature list *FL*. For each feature set, all samples are represented by features in the set, and a classification algorithm is executed on these samples. The quality of the predicted results is used to evaluate the importance of the feature set. Obviously, the feature set yielding the best quality is most important and called optimal feature set. Features in this set are called optimal features. Additionally, the classifier using these optimal features to represent samples is termed as optimal classifier.

As mentioned above, a classification algorithm should be set before executing the IFS method. Here, we selected the SVM algorithm [29,30], one of the most classic machine learning algorithms, as the classification algorithm. Since it was first published, the SVM algorithm has been applied in several fields, including traditional classification and regression problems [49–51] and has always shown good generalizability on problems such as hand-writing recognition [52] and face detection [53]. The SVM model can be built on a small size dataset, whereas it still has good generalization performance. In this algorithm, the non-linear separable samples in the training dataset are always mapped to a higher-dimensional space using the kernel trick. In the higher-dimensional space, the positive and negative samples can be linearly separated by a hyper-plane with a maximum margin. For a new sample to be classified, the sample is also mapped to the same higher dimension, and its class depends on which side of the hyper-plane it falls to.

To quickly implement SVM, a tool, named “SMO”, in Weka [54] was employed in this study. This tool implements a type of SVM that is optimized by sequential minimum optimization (SMO) [55]. For convenience, it was executed with its default parameters.

2.5. Measurements

According to the IFS method, a SVM-based classifier can be built on a feature subset. A series of classifiers was constructed. The 10-fold cross-validation (10-CV) [56] was adopted to evaluate their performance. Although it is less accurate than the Jackknife cross-validation (J-CV) [57], the 10-CV test is a computation-, and time-saving approach that can yield similar results on a large dataset, in which samples are encoded by multiple features.

To rate the predicted results yielded by one SVM-based classifier, we employed some measurements. For each cancer type, we calculated the prediction accuracy (*ACC*), which was defined as

$$ACC_i = \frac{n_i}{N_i} \ (i = 1, 2, 3, 4, 5) \tag{4}$$

where n_i is the number of samples that are correctly predicted in the i-th cancer type, and N_i is the total number of samples for this cancer type. In addition, we further calculated the overall accuracy (*TACC*) to evaluate the prediction abilities of classifiers on the whole, which was defined by

$$TACC = \sum_{i=1}^{5} n_i / \sum_{i=1}^{5} N_i \tag{5}$$

In addition, to reduce the influence of the class sizes on *TACC*, the Matthews correlation coefficient (*MCC*) [58] was also calculated to compare the performance of each SVM-based classifier. It is known that the original *MCC* is a balanced measurement for binary classification even if the dataset is unbalanced. In this study, five types of cancers, i.e., five classes, were considered. Thus, the *MCC* in multiclass [59] was employed, which is more complex than that of a binary classification. Similar to the original *MCC*, it also can give a balanced assessment when the sizes of classes are quite different. To date, it has been applied to evaluate several constructed classification models [41,60–63].

Given a classification problem involving n samples, say s_1, s_2, …, s_n, and N classes, denoted as 1, 2, …, N. According to the true class of each sample, a matrix with n rows and N columns, denoted by M, can be constructed. Its element M_{ij} is set to 1 if s_i belongs to class j or 0 otherwise. The predicted results derived from a SVM-based classifier can be used to construct another matrix T. It also has n rows and N columns, and each element in T can be defined as

$$T_{ij} = \begin{cases} 1 & if \ s_i \ is \ predicted \ to \ be \ in \ class \ j \\ 0 & otherwise \end{cases} \tag{6}$$

Accordingly, compute the covariance function between matrices M and T, which is calculated with the following equation:

$$cov(M, T) = \frac{1}{N} \sum_{k=1}^{N} cov(M_k, T_k) = \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{N} (M_{ik} - \overline{M_k})(T_{ik} - \overline{T_k}) \tag{7}$$

where M_k and T_k are the k-th column of matrices M and T, and $\overline{M_k}$ and $\overline{T_k}$ are the mean value of numbers in M_k and T_k, respectively. The *MCC* for multiclass classification problem can be computed by the equation:

$$MCC = \frac{cov(M, T)}{\sqrt{cov(M, M)\, cov(T, T)}} \tag{8}$$

Consistent with the original *MCC*, the *MCC* in multiclass also ranges between -1 and 1. A larger value of *MCC* obtained from Eq. (8) indicates a better prediction performance for a SVM-based classifier. In this study, *MCC* is used as a major measurement to evaluate the prediction abilities of SVM-based classifiers.

3. Results

In Section 2.2, the GO terms and KEGG pathways were clustered and used to represent the patients with one of five types of cancers based upon the enrichment theory. These features were analyzed by the MCFS method, yielding a feature list *FL* that is provided in **Supplementary Material S1** and sixteen decision rules listed in Table 2.

The feature list *FL* was next analyzed with the IFS method to extract an optimal feature set that can support SVM for yielding the best performance. The original IFS method always tests all possible feature sets. However, this method is time-consuming to perform due to the large number of possible feature sets and our limited computational power. To save on the computational resources, we first tested some special feature sets that contained the top i features in *FL*, where i was a multiple of ten. For each of these feature sets, a SVM-based classifier was built, executed on all samples that were represented by features in this set, and evaluated by the 10-CV test. The *ACC* for each cancer type, *TACC*, and *MCC* values were calculated according to the predicted results. The prediction performances for all SVM-based classifiers are listed in **Supplementary Material S2**. The IFS curve in Fig. 1(A) illustrates the association between the number of features that were used to build SVM-based classifiers and the corresponding *MCC* values yielded by these classifiers, which demonstrates the highest *MCC* value of 0.5188 when the top 320 features were selected. It is easy to conclude that if some more refined tests on the feature sets with sizes approximately 320 are performed, a better feature set could be found. Thus, we further tested the feature sets with sizes between 200 and 400. The measurements mentioned in Section 2.5 were calculated and are listed in **Supplementary Material S3.** Additionally, the association between *MCC* and the number of used features is illustrated in Fig. 1(B),

Table 2
The sixteen decision rules identified by the MCFS method.

Classification	Rules	Features	Criteria
Head and neck squamous cell carcinoma (HNSC)	Rule 1	GO:1901533 (negative regulation of hematopoietic progenitor cell differentiation)	≥ 1.445
		hsa04973 (Carbohydrate digestion and absorption)	≥ 0.075
	Rule 2	GO:0000788 (nuclear nucleosome)	≥ 0.505
		GO:0000788 (nuclear nucleosome)	≤ 3.132
		GO:0019003 (GDP binding)	≤ −0.323
	Rule 3	GO:0061462 (protein localization to lysosome)	≥ 1.169
		GO:0061462 (protein localization to lysosome)	≤ 2.859
		hsa04662 (B cell receptor signaling pathway)	≤ −0.014
		hsa00533 (Glycosaminoglycan biosynthesis - keratan sulfate)	≤ −0.190
	Rule 4	GO:0050821 (protein stabilization)	≥ −0.391
		GO:0050821 (protein stabilization)	≤ 0.797
		GO:0050999 (regulation of nitric-oxide synthase activity)	≤ −0.452
		GO:0031647 (regulation of protein stability)	≤ −0.095
	Rule 5	GO:0072176 (nephric duct development)	≥ 2.722
		GO:0070242 (thymocyte apoptotic process)	≥ 0.292
		GO:0070242 (thymocyte apoptotic process)	≤ 1.017
(Ovarian cancer) OV	Rule 6	GO:0031049 (programmed DNA elimination)	≥ 0.846
		hsa05214 (Glioma)	≤ 1.044
	Rule 7	hsa05211 (Renal cell carcinoma)	≤ −0.699
		GO:0042771 (intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator)	≥ 0.445
		GO:0043015 (gamma-tubulin binding)	≥ 1.807
Kidney renal clear cell carcinoma (KIRC)	Rule 8	hsa05220 (Chronic myeloid leukemia)	≤ −0.876
		GO:0097201 (negative regulation of transcription from RNA polymerase II promoter in response to stress)	≥ 2.395
	Rule 9	hsa05223 (Non-small cell lung cancer)	≤ −0.915
		GO:0030197 (extracellular matrix constituent, lubricant activity)	≥ 2.412
	Rule 10	hsa05223 (Non-small cell lung cancer)	≤ −0.915
		hsa04120 (Ubiquitin mediated proteolysis)	≥ 0.458
		GO:0051234 (establishment of localization)	≤ −0.214
		GO:0031647 (regulation of protein stability)	≤ 0.758
Colorectal adenocarcinoma (COADREAD)	Rule 11	GO:0030275 (LRR domain binding)	≥ −0.270
		GO:0019003 (GDP binding)	≥ −0.357
	Rule 12	hsa05216 (Thyroid cancer)	≥ −0.566
		GO:0010764 (negative regulation of fibroblast migration)	≥ 1.608
		GO:1902804 (negative regulation of synaptic vesicle transport)	≤ 8.769
	Rule 13	hsa05216 (Thyroid cancer)	≥ −0.600
		GO:0070411 (I-SMAD binding)	≥ 0.255
	Rule 14	hsa05216 (Thyroid cancer)	≥ −0.534
		GO:0048642 (negative regulation of skeletal muscle tissue development)	≥ 1.229
		hsa04726 (Serotonergic synapse)	≥ 0.175
	Rule 15	GO:0072075 (metanephric mesenchyme development)	≥ 1.091
		GO:0032525 (somite rostral/caudal axis specification)	≥ 3.734
Breast cancer (BRCA)	Rule 16	other conditions	

from which we can see that the highest *MCC* is 0.5231 that was obtained using the top 344 features in *FL*. Thus, we termed these 344 features the optimal features and the corresponding SVM-based classifier the optimal SVM-based classifier. In addition, the *ACC*s for BRAC, COADREAD, HNSC, KIRC and OV obtained with this optimal classifier were 0.4620, 0.7054, 0.4477, 0.7886 and 0.6404, respectively, and the *TACC* was 0.6190.

As mentioned above, another result of the MCFS method is a group of decision rules that are listed in Table 2. Using these rules for identifying the five cancer types, the *TACC* can be 0.4978. This result demonstrates quite good accuracy because the accuracy yielded at random is only 0.2 (1/5). Additionally, unlike the SVM-based classifier, the process by which the cancer type of a sample is identified can be easily observed. A detailed analysis of these rules can help us understand the specific functional alterations of each cancer type. Our detailed analysis on these rules as well as some optimal features is discussed.

4. Discussion

4.1. Analysis of decision rules yielded by the MCSF method

As previously described, we identified sixteen rules (see Table 2) for distinguishing the five cancer types, among which five are for HNSC, two for OV, three for KIRC, five for COADREAD and the last one for BRCA. These rules involved 22 GO terms and ten KEGG pathways, some of which have been shown to have discriminating power for the five cancers in previous studies.

GO:0000788, the nuclear nucleosome, is a molecular component that has been widely known to play a role in epigenetic regulation [64]. The nucleosome remodeling by histone modification and DNA methylation regulate various biochemical pathways essential for tumorigenesis [64]. A recent study found that the nucleosome positioning varies among different cell types and that the cell-free DNA nucleosome occupancies correlate with the nuclear architecture, gene structure, and expression observed in cells, suggesting that these occupancies could inform the cell type of origin [10].

GO:0050821 is a biological process named protein stabilization that has been assigned as a biomarker for head and neck squamous cell carcinoma in this study. A previous study reported that the anti-apoptotic protein BCL-x(L) was implicated in head and neck cancer and that BCL-x(L) induction appears to be due to protein stabilization rather than transcriptional activation [65], which supports the identification of this GO term as a basis of classification. Similarly, another GO term, **GO:0050999**, nitric-oxide (NO) synthase regulator activity, also contributes to identifying head and neck cancer. An increased level of inducible nitric oxide synthase (iNOS) expression and activity has been found in the tumor cells of head and neck cancer [66,67]. Furthermore,

a previous study of head and neck cancer identified tumor associated immune cells (e.g., macrophages, T/NK-cells) are a source of mediators that may induce the iNOS/NO pathway inside tumor cells. The tumor-associated macrophages can produce high levels of NO that may radiosensitize bystander tumor cells [68]. These results suggest the specific role of this biological process in neck and head cancer.

Regarding ovarian cancer, two GO terms were served as rules in this study and are supported by previous findings. The first one, **GO:0042771**, is a biological process named intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator. In previous studies, the suppression of genes involved in the extrinsic apoptotic pathway was observed in a particular ovarian cancer cell line [69], and the ovarian cancer cell survival, anoikis resistance and peritoneal metastases may be due to the inhibition of the intrinsic apoptotic pathway [70], suggesting the specific role of this biological process in ovarian cancer. The second one, **GO:0043015,** is a molecular function named gamma-tubulin binding. Previous studies support a model by which the BRCA1 ubiquitin ligase, the breast and ovarian cancer specific tumor suppressor, modifies both gamma-tubulin and a second centrosomal protein that controls the localization of gammaTuRC to the centrosome. The loss of BRCA1 would result in centrosome hyperactivity, supernumerary centrosomes and, possibly, aneuploidy [71]. These findings associate the ovarian cancer and this GO term through the cancer suppressor gene *BRCA1*.

GO:0030275 describes the binding of the leucine-rich repeat (LRR) domain. It is well known that the leucine-rich repeat is a protein structural motif that contributes to the formation of α/β horseshoe fold [72]. Such protein structure has been widely identified during the initiation and progression of different tumor types [72]. Concerning the five different tumor types of interest, such biological process has been reported to play variable roles. In breast cancer, head and neck squamous cell carcinoma and clear cell renal cell carcinoma, functional proteins containing leucine-rich repeat contribute to the act as crucial tumor-suppressors [73–75]. However, in colorectal adenocarcinoma, leucine-rich repeat containing gene LGR5 have been reported to be related to a poor prognosis and chemotherapy resistance. Also in ovarian cancer, leucine-rich repeats play a dual role in tumorigenesis, either oncogenic or tumor suppressing, reflecting the complex role of our predicted biological process in tumorigenesis [76]. Considering the different roles of the LRR domain in different cancer types, such biological process may allow us to distinguish the five tumor types.

In addition to the GO terms, we found one KEGG pathway **hsa04120** (Ubiquitin mediated proteolysis) for which previous studies have supported its role in kidney renal clear cell carcinoma. One previous study showed that *SMURF1* (Smad ubiquitin regulatory factor 1) (SMURF1), a E3 ubiquitin ligase for ubiquitination and proteasomal degradation, promoted cell growth and metastasis in clear cell renal cell carcinoma [77]. Another study found that ubiquitin-like with PHD and RING finger domain 1 (UHRF1), a multi-domain ubiquitin E3 ligase, plays critical roles in regulating DNA methylation and histone ubiquitination, is frequently overexpressed in human clear cell Renal Cell Carcinoma (ccRCC) tissues, promotes the non-degradative ubiquitination of p53, and suppresses the p53 pathway activation and p53-dependent apoptosis in ccRCC cells [78]. These results strongly associate this ksa04120 pathway with this cancer type.

4.2. Analysis of optimal features

In addition to the decision rules, we identified the optimal feature set with 344 features, on which an optimal SVM-based classifier was built, yielding the highest *MCC* value (0.5231) for classifying the five cancer types. However, it is quite difficult to analyze all optimal features. By examining the *MCC* values listed in **Supplementary Material S2**, we found that only using the first 40 features, the MCC can reach a value of 0.4630. Therefore, we focused on these 40 features, which are listed in Table 3. Among these 40 features, ten features have clearly been shown to have discriminating power based on previous studies.

GO:0019002, the top GO term in the feature list yielded by the MCFS method, describes selective and non-covalent interactions with guanosine monophosphate (GMP). In certain cells, cyclic GMP (cGMP) can be synthesized from guanosine triphosphate (GTP) by guanylyl cyclase and mediate hormonal signaling [79,80]. As a hormone associated cancer, cGMP signaling pathways have been widely confirmed to contribute to the initiation and progression of breast cancer [81,82]. It has also been confirmed by in vitro experiments that cGMP interactions may also contribute to apoptosis in human colon adenocarcinoma [83]. Regarding ovarian cancer and head and neck squamous cell carcinoma, recent studies also validate the core regulatory role of GMP interactions during the tumorigenesis of such tumor types [84,85]. However, in kidney renal clear cell carcinoma, no reports confirmed the oncogenic role of cGMP interactions during the initiation and progression of kidney renal clear cell carcinoma, although cGMP functions have been confirmed in other renal carcinoma types [86].

GO:0061428 is the biological process that negatively regulates RNA polymerase II-regulated transcription in response to hypoxia. It is well known that hypoxia is a common feature of the tumor microenvironment. This condition alters gene expressions, facilitating the tumor survival and progression [87]. The main molecular drivers of this response are hypoxia inducible factors, known as HIFs [88]. Previous studies found that the roles of HIFs vary in different tumors [89], suggesting that these differences could be used for the classification of different cancers.

GO:2000270 is termed negative regulation of fibroblast apoptotic process. Fibroblasts have been shown to participate in human tumorigenesis by providing a permissive environment for the proliferation and survival of epithelial cells, and by remodeling the ECM to promote tumor growth and invasiveness [90,91]. Cancer associated fibroblasts (CAFs) were found to vary in abundance among different types of cancers. For example, breast, prostate, and pancreatic cancers contain high numbers of CAFs, whereas brain, renal, and ovarian cancers demonstrate fewer CAFs [92,93]. Additionally, the molecular features are diverse in different cancers and cell types [94], suggesting the discriminating power of this GO term.

GO:0044028 and **GO:0044029** refer to DNA hypomethylation and hypomethylation of CpG islands, respectively. Previous studies have demonstrated that different cancers may have specific methylation profiles [95,96]. This knowledge has been utilized to identify the tissues contributing to the circulating DNA pool [11,12]. These results suggest the capacity of DNA hypomethylation in distinguishing different cancer types.

GO:0038028 as a functional candidate GO terms that describes the insulin receptor signaling pathway via phosphatidylinositol 3-kinase. Insulin receptor signaling pathway and the related phosphatidylinositol 3-kinase (PI3K) signaling pathway have been confirmed to contribute to the proliferative regulation and have been reported to contribute to the initiation and progression of various tumor types. Regarding the five cluster of cancer types, it has been confirmed that as the core proliferative regulatory signaling pathways, insulin receptor signaling pathway and related PI3k pathway contribute to the initiation and progression of breast cancer, head and neck squamous cell carcinoma, kidney renal clear cell carcinoma and ovarian cancer [97–100]. However, there is no direct evidence for the contribution of these two signaling pathways during colorectal adenocarcinoma tumorigenesis. Therefore, the biological process described by GO: 0038028 may distinguish colorectal adenocarcinoma from the other four cancer types.

GO:0031052 describes the normal DNA rearrangements induced by regulated cleavage of the genome. Such biological process has also been confirmed to act differently in different tumor types. DNA rearrangements have been widely confirmed in various tumor types. In breast cancer and ovarian cancer, the rearrangement of BRCA1 and BRCA2 is quite significant and has been regarded as one of the driver ariants for the initiation and progression of these two tumor types [101,102].

Table 3
The top 40 ranked features identified for further analysis based upon a literature review.

Features	RI	Category	Definition
GO:0019002	0.1909	BP	the selective and non-covalent interactions with guanosine monophosphate
GO:0061428	0.1485	BP	negative regulation of transcription from RNA polymerase II promoter in response to hypoxia
GO:0044028	0.1484	BP	DNA hypomethylation
GO:0044029	0.1475	BP	hypomethylation of CpG island
GO:0005943	0.1438	CC	phosphatidylinositol 3-kinase complex, class IA
GO:0038028	0.1374	BP	the insulin receptor signaling pathway via phosphatidylinositol 3-kinase
GO:0030275	0.1270	BP	the binding of leucine-rich repeat (LRR) domain
GO:0030891	0.1210	CC	VCB complex
GO:2000270	0.1165	BP	negative regulation of fibroblast apoptotic process
GO:0097201	0.1138	BP	negative regulation of transcription from RNA polymerase II promoter in response to stress
GO:0000153	0.0986	CC	cytoplasmic ubiquitin ligase complex
GO:0019003	0.0946	MF	GDP binding
GO:0051000	0.0931	BP	positive regulation of nitric-oxide synthase activity
GO:0052813	0.0913	BP	phosphatidylinositol bisphosphate kinase activity
GO:0046934	0.0906	BP	phosphatidylinositol-4,5-bisphosphate 3-kinase activity
GO:0097651	0.0872	CC	phosphatidylinositol 3-kinase complex, class I
hsa05216	0.0755	KEGG	Thyroid cancer
GO:0032770	0.0668	BP	positive regulation of monooxygenase activity
GO:0031619	0.0661	BP	homologous chromosome orientation involved in meiotic metaphase I plate congression
GO:2000653	0.0660	BP	regulation of genetic imprinting
GO:0051385	0.0657	BP	response to mineralocorticoid
GO:0051455	0.0656	BP	attachment of spindle microtubules to kinetochore involved in homologous chromosome segregation
GO:0033593	0.0649	CC	BRCA2-MAGE-D1 complex
GO:0051316	0.0637	BP	attachment of spindle microtubules to kinetochore involved in meiotic chromosome segregation
GO:0031049	0.0628	BP	programmed DNA elimination
GO:2000301	0.0619	BP	negative regulation of synaptic vesicle exocytosis
GO:0035005	0.0606	MF	1-phosphatidylinositol-4-phosphate 3-kinase activity
GO:0031052	0.0602	BP	chromosome breakage
GO:0090172	0.0601	BP	microtubule cytoskeleton organization involved in homologous chromosome segregation
GO:0003192	0.0597	BP	mitral valve formation
GO:1902804	0.0581	BP	negative regulation of synaptic vesicle transport
GO:2000793	0.0579	BP	cell proliferation involved in heart valve development
GO:0043560	0.0571	BP	insulin receptor substrate binding
GO:2000811	0.0570	BP	negative regulation of anoikis
GO:0010909	0.0560	BP	positive regulation of heparan sulfate proteoglycan biosynthetic process
GO:0043060	0.0554	BP	meiotic metaphase I plate congression
GO:0010908	0.0553	BP	regulation of heparan sulfate proteoglycan biosynthetic process
GO:0051311	0.0551	BP	meiotic metaphase plate congression
GO:0072076	0.0551	BP	nephrogenic mesenchyme development
GO:0072134	0.0541	BP	nephrogenic mesenchyme morphogenesis

What's more, a specific rearrangement of a functional gene TFE3 in Xp11.2 has been confirmed to contribute to the tumorigenesis of renal clear cell carcinoma, validating that such biological process may also be associated with such type of tumor [103]. For colorectal adenocarcinoma, the fusions of ROS1 and ALK have been widely identified [104]. However, during the tumorigenesis of head and neck squamous cell carcinoma, few DNA rearrangements described by GO: 0031052 have been confirmed, implying that the lack of DNA rearrangement may be a potential characteristic for the head and neck squamous cell carcinoma.

GO:0033593 is a cellular component called BRCA2-MAGE-D1 complex. It is well-known that mutations in *BRCA2* gene are strongly associated with breast and ovarian cancer. The previous study discovered that MAGE-D1 is a downstream target of BRCA2 and that BRCA2 suppresses cell proliferation by stabilizing MAGE-D1. Meanwhile, MAGE-D1 protein expression was downregulated in breast carcinoma cell lines, suggesting its involvement in the tumorigenesis of breast cancer [105].

5. Conclusions

In this study, based on the computational method, we identified the GO terms and KEGG pathways that may distinguish the five cancer types. These functional alteration signatures of five different cancers can not only map the tissue-of-origin in cancer detection but also have the potential to be the targets of specific cancer treatments.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbadis.2017.12.026.

Funding

This work was supported by the National Natural Science Foundation of China [31371335], the fund of the key Laboratory of Stem Cell Biology of Chinese Academy of Sciences [201703].

Transparency document

The Transparency document associated this article can be found, in online version.

References

[1] V.J. Cogliano, R. Baan, K. Straif, Y. Grosse, B. Lauby-Secretan, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, C.P. Wild, Preventable exposures associated with human cancers, J. Natl. Cancer Inst. 103 (2011) 1827–1839.
[2] K. Crane, CANCER IN THE DEVELOPING WORLD, Palliative care gains ground in developing countries, J. Natl. Cancer Inst. 102 (2010) 1613–1615.
[3] W.W. Yang, Y. Xia, H.T. Ji, Y.H. Zheng, J. Liang, W.H. Huang, X. Gao, K. Aldape, Z.M. Lu, Nuclear PKM2 regulates beta-catenin transactivation upon EGFR activation, Nature 480 (2011) 118–122.
[4] M. Li, A. Mukasa, M. del Mar Inda, J.H. Zhang, L. Chin, W. Cavenee, F. Furnari, Guanylate binding protein 1 is a novel effector of EGFR-driven invasion in glioblastoma, J. Exp. Med. 208 (2011) 2657–2673.
[5] A. Kobayashi, H. Okuda, F. Xing, P.R. Pandey, M. Watabe, S. Hirota, S.K. Pai, W. Liu, K. Fukuda, C. Chambers, A. Wilber, K. Watabe, Bone morphogenetic protein 7 in dormancy and metastasis of prostate cancer stem-like cells in bone, J. Exp. Med. 208 (2011) 2641–2655.
[6] T. Rosso, P. Bertuccio, C. La Vecchia, E. Negri, M. Malvezzi, Cancer mortality trend analysis in Italy, 1980–2010, and predictions for 2015, Tumori 101 (2015)

664–675.
[7] D. Sidransky, A. Von Eschenbach, Y.C. Tsai, P. Jones, I. Summerhayes, F. Marshall, M. Paul, P. Green, S.R. Hamilton, P. Frost, et al., Identification of p53 gene mutations in bladder cancers and urine samples, Science 252 (1991) 706–709.
[8] G.D. Sorenson, D.M. Pribish, F.H. Valone, V.A. Memoli, D.J. Bzik, S.L. Yao, Soluble normal and mutated DNA sequences from single-copy genes in human blood, Cancer Epidemiol. Biomark. Prev. 3 (1994) 67–71.
[9] V. Vasioukhin, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, M. Stroun, Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia, Br. J. Haematol. 86 (1994) 774–779.
[10] M.W. Snyder, M. Kircher, A.J. Hill, R.M. Daza, J. Shendure, Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin, Cell 164 (2016) 57–68.
[11] K. Sun, P. Jiang, K.C. Chan, J. Wong, Y.K. Cheng, R.H. Liang, W.K. Chan, E.S. Ma, S.L. Chan, S.H. Cheng, R.W. Chan, Y.K. Tong, S.S. Ng, R.S. Wong, D.S. Hui, T.N. Leung, T.Y. Leung, P.B. Lai, R.W. Chiu, Y.M. Lo, Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) E5503–5512.
[12] S. Guo, D. Diep, N. Plongthongkum, H.L. Fung, K. Zhang, K. Zhang, Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA, Nat. Genet. 49 (2017) 635–642.
[13] J.A. Olsen, L.A. Kenna, R.C. Tipon, M.G. Spelios, M.M. Stecker, E.M. Akirav, A minimally-invasive blood-derived biomarker of oligodendrocyte cell-loss in multiple sclerosis, EBioMedicine 10 (2016) 227–235.
[14] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A. Aparicio, S. Behjati, A.V. Biankin, G.R. Bignell, N. Bolli, A. Borg, A.L. Borresen-Dale, S. Boyault, B. Burkhardt, A.P. Butler, C. Caldas, H.R. Davies, C. Desmedt, R. Eils, J.E. Eyfjord, J.A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D.T. Jones, D. Jones, S. Knappskog, M. Kool, S.R. Lakhani, C. Lopez-Otin, S. Martin, N.C. Munshi, H. Nakamura, P.A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J.V. Pearson, X.S. Puente, K. Raine, M. Ramakrishna, A.L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T.N. Schumacher, P.N. Span, J.W. Teague, Y. Totoki, A.N. Tutt, R. Valdes-Mas, M.M. van Buuren, L. van't Veer, A. Vincent-Salomon, N. Waddell, L.R. Yates, I. Australian Pancreatic Cancer Genome, I.B.C. Consortium, I.M.-S. Consortium, I. PedBrain, J. Zucman-Rossi, P.A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S.M. Grimmond, R. Siebert, E. Campo, T. Shibata, S.M. Pfister, P.J. Campbell, M.R. Stratton, Signatures of mutational processes in human cancer, Nature 500 (2013) 415–421.
[15] S.H. Ou, K. Shirai, Anaplastic lymphoma kinase (ALK) signaling in lung cancer, Adv. Exp. Med. Biol. 893 (2016) 179–187.
[16] V. Atsaves, R. Zhang, D. Ruder, Y. Pan, V. Leventaki, G.Z. Rassidakis, F.X. Claret, Constitutive control of AKT1 gene expression by JUNB/CJUN in ALK + anaplastic large-cell lymphoma: a novel crosstalk mechanism, Leukemia 29 (2015) 2162–2172.
[17] M. Greaves, C.C. Maley, Clonal evolution in cancer, Nature 481 (2012) 306–313.
[18] J.A. Blake, K.R. Christie, M.E. Dolan, H.J. Drabkin, D.P. Hill, L. Ni, D. Sitnikov, S. Burgess, T. Buza, C. Gresham, F. McCarthy, L. Pillai, H. Wang, S. Carbon, H. Dietze, S.E. Lewis, C.J. Mungall, M.C. Munoz-Torres, M. Feuermann, P. Gaudet, S. Basu, R.L. Chisholm, R.J. Dodson, P. Fey, H. Mi, P.D. Thomas, A. Muruganujan, S. Poudel, J.C. Hu, S.A. Aleksander, B.K. McIntosh, D.P. Renfro, D.A. Siegele, H. Attrill, N.H. Brown, S. Tweedie, J. Lomax, D. Osumi-Sutherland, H. Parkinson, P. Roncaglia, R.C. Lovering, P.J. Talmud, S.E. Humphries, P. Denny, N.H. Campbell, R.E. Foulger, M.C. Chibucos, M.G. Giglio, H.Y. Chang, R. Finn, M. Fraser, A. Mitchell, G. Nuka, S. Pesseat, A. Sangrador, M. Scheremetjew, S.Y. Young, R. Stephan, M.A. Harris, S.G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P.J. Kersey, M.D. McDowall, D.M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, G.T. Hayman, S.J. Wang, V. Petri, P. D'Eustachio, L. Matthews, R. Balakrishnan, G. Binkley, J.M. Cherry, M.C. Costanzo, J. Demeter, S.S. Dwight, S.R. Engel, B.C. Hitz, D.O. Inglis, P. Lloyd, S.R. Miyasato, K. Paskov, G. Roe, M. Simison, R.S. Nash, M.S. Skrzypek, S. Weng, E.D. Wong, T.Z. Berardini, D. Li, E. Huala, J. Argasinska, C. Arighi, A. Auchincloss, K. Axelsen, G. Argoud-Puy, A. Bateman, B. Bely, M.C. Blatter, C. Bonilla, L. Bougueleret, E. Boutet, L. Breuza, A. Bridge, R. Britto, C. Casals, E. Cibrian-Uhalte, E. Coudert, I. Cusin, P. Duek-Roggli, A. Estreicher, L. Famiglietti, P. Gane, P. Garmiri, A. Gos, N. Gruaz-Gumowski, E. Hatton-Ellis, U. Hinz, C. Hulo, R. Huntley, F. Jungo, G. Keller, K. Laiho, P. Lemercier, D. Lieberherr, A. MacDougall, M. Magrane, M. Martin, P. Masson, P. Mutowo, C. O'Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, A. Shypitsyna, A. Stutz, S. Sundaram, M. Tognolli, C. Wu, I. Xenarios, J. Chan, R. Kishore, P.W. Sternberg, K. Van Auken, H.M. Muller, J. Done, Y. Li, D. Howe, M. Westerfield, G.O. Consortium, Gene ontology consortium: going forward, Nucleic Acids Res. 43 (2015) D1049–D1056.
[19] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation, Nucleic Acids Res. 44 (2016) D457–462.
[20] B.M. Konopka, T. Golda, M. Kotulska, Evaluating the significance of protein functional similarity based on gene ontology, J. Comput. Biol. 21 (2014) 809–822.
[21] M. Deng, J. Bragelmann, J.L. Schultze, S. Perner, Web-TCGA: an online platform for integrated analysis of molecular cancer data sets, BMC Bioinf. 17 (2016) 72.
[22] K. Tomczak, P. Czerwinska, M. Wiznerowicz, The cancer genome atlas (TCGA): an immeasurable source of knowledge, Contemp. Oncol. 19 (2015) A68–77.
[23] E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobsen, C.J. Byrne, M.L. Heuer, E. Larsson, Y. Antipin, B. Reva, A.P. Goldberg, C. Sander, N. Schultz, The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov. 2 (2012) 401–404.
[24] Y.H. Zhang, C. Chu, S. Wang, L. Chen, J. Lu, X. Kong, T. Huang, H. Li, Y.D. Cai, The use of gene ontology term and KEGG pathway enrichment for analysis of drug half-life, PLoS One 11 (2016) e0165496.
[25] L. Chen, C. Chu, J. Lu, X. Kong, T. Huang, Y.D. Cai, Gene ontology and KEGG pathway enrichment analysis of a drug target-based classification system, PLoS One 10 (2015) e0126492.
[26] J. Yang, L. Chen, X. Kong, T. Huang, Y.D. Cai, Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway, PLoS One 9 (2014) e107202.
[27] Z. Li, B.Q. Li, M. Jiang, L. Chen, J. Zhang, L. Liu, T. Huang, Prediction and analysis of retinoblastoma related genes through gene ontology and KEGG, Biomed. Res. Int. 2013 (2013) 304029.
[28] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, J. Komorowski, Monte Carlo feature selection for supervised classification, Bioinformatics 24 (2008) 110–117.
[29] D. Meyer, F. Leisch, K. Hornik, The support vector machine under test, Neurocomputing 55 (2003) 169–186.
[30] Corinna Cortes, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[31] T. Huang, X. Shi, P. Wang, Z. He, K. Feng, L. Hu, X. Kong, Y. Li, Y. Cai, K. Chou, Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks, PLoS One 5 (2010) e10972.
[32] T. Huang, P. Wang, Z.Q. Ye, H. Xu, Z. He, K.Y. Feng, L. Hu, W. Cui, K. Wang, X. Dong, L. Xie, X. Kong, Y.D. Cai, Y. Li, Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties, PLoS One 5 (2010) e11900.
[33] T. Huang, C. Wang, G. Zhang, L. Xie, Y. Li, SySAP: a system-level predictor of deleterious single amino acid polymorphisms, Protein Cell 3 (2012) 38–43.
[34] T. Huang, Y. Ji, D. Hu, B. Chen, H. Zhang, C. Li, G. Chen, X. Luo, X.W. Zheng, X. Lin, SNHG8 is identified as a key regulator of epstein-barr virus(EBV)-associated gastric cancer by an integrative analysis of lncRNA and mRNA expression, Oncotarget 7 (2016) 80990–81002.
[35] T. Huang, C.-L. Liu, L.-L. Li, M.-H. Cai, W.-Z. Chen, Y.-F. Xu, P.F. O'Reilly, L. Cai, L. He, A new method for identifying causal genes of schizophrenia and anti-tuberculosis drug-induced hepatotoxicity, Sci. Rep. 6 (2016) 32571.
[36] L. Chen, Y.H. Zhang, M.Y. Zheng, T. Huang, Y.D. Cai, Identification of compound-protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds, Mol. Gen. Genomics. 291 (2016) 2065–2079.
[37] L. Chen, Y.-H. Zhang, G. Lu, T. Huang, Y.-D. Cai, Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways, Artif. Intell. Med. 76 (2017) 27–36.
[38] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[39] Q. Ni, L. Chen, A feature and algorithm selection method for improving the prediction of protein structural classes, Comb. Chem. High Throughput Screen. 20 (2017) 612–621.
[40] L. Chen, C. Chu, T. Huang, X. Kong, Y.-D. Cai, Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models, Amino Acids 47 (2015) 1485–1493.
[41] L. Chen, Y.-H. Zhang, T. Huang, Y.-D. Cai, Gene expression profiling gut microbiota in different races of humans, Sci. Rep. 6 (2016) 23075.
[42] L. Liu, L. Chen, Y.H. Zhang, L. Wei, S. Cheng, X. Kong, M. Zheng, T. Huang, Y.D. Cai, Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection, J. Biomol. Struct. Dyn. 35 (2017) 312–329.
[43] L. Chen, C. Chu, K. Feng, Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization, Comb. Chem. High Throughput Screen. 19 (2016) 136–143.
[44] L. Chen, S. Wang, Y.-H. Zhang, J. Li, Z.-H. Xing, J. Yang, T. Huang, Y.-D. Cai, Identify key sequence features to improve CRISPR sgRNA efficacy, IEEE Access (2017), http://dx.doi.org/10.1109/ACCESS.2017.2775703.
[45] Q. Zou, J. Zeng, L. Cao, R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346–354.
[46] Q. Zou, S. Wan, Y. Ju, J. Tang, X. Zeng, Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy, BMC Syst. Biol. 10 (2016) 114.
[47] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publisher, 1992.
[48] M.D.M.J.D.b.K.D.J.K.J. Komorowski, Discovering networks of interdependent features in high-dimensional problems, Big Data Analysis: New Algorithms for a New Society, Springer International Publishing, 2016, pp. 285–304.
[49] K.-B. Duan, S.S. Keerthi, Which Is the best multiclass SVM method? An empirical study, in: N. Oza, R. Polikar, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, Vol. 3541 Springer, Berlin Heidelberg, 2005, pp. 278–285.
[50] Y. Lee, Y. Lin, G. Wahba, Multicategory support vector machines, J. Am. Stat. Assoc. 99 (2004) 67–81.
[51] B.E. Boser, I.M. Guyon, V.N. Vapnik, A Training Algorithm for Optimal Margin Classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM, Pittsburgh, Pennsylvania, USA, 1992, pp. 144–152.
[52] Y. Lecun, L.D. Jackel, L. Bottou, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, U.A. Muller, E. Sackinger, P. Simard, Learning algorithms for classification: a comparison on handwritten digit recognition, Neural Networks: The Statistical

Mechanics Perspective, 1995.
[53] E. Osuna, R. Freund, F. Girosit, Training support vector machines: an application to face detection, Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, 1997, pp. 130–136.
[54] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (2004) 2479–2481.
[55] J. Platt, Sequential Minimal Optimizaton: A Fast Algorithm for Training Support Vector Machines, Technical Report MSR-TR-98-14, (1998).
[56] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence, Vol. 14 Lawrence Erlbaum Associates Ltd., 1995, pp. 1137–1145.
[57] L. Chen, C. Chu, Y.-H. Zhang, M.-Y. Zheng, L. Zhu, X. Kong, T. Huang, Identification of drug–drug interactions using chemical interactions, Curr. Bioinforma. (2017), http://dx.doi.org/10.2174/1574893611666160618094219.
[58] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.
[59] J. Gorodkin, Comparing two K-category assignments by a K-category correlation coefficient, Comput. Biol. Chem. 28 (2004) 367–374.
[60] L. Chen, J. Li, Y.-H. Zhang, K. Feng, S. Wang, Y. Zhang, T. Huang, X. Kong, Y.-D. Cai, Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method, J. Cell. Biochem. (2017), http://dx.doi.org/10.1002/jcb.26507.
[61] P.W. Zhang, L. Chen, T. Huang, N. Zhang, X.Y. Kong, Y.D. Cai, Classifying ten types of major cancers based on reverse phase protein array profiles, PLoS One 10 (2015) e0123147.
[62] D. Albanese, C. De Filippo, D. Cavalieri, C. Donati, Explaining diversity in metagenomic datasets by phylogenetic-based feature weighting, PLoS Comput. Biol. 11 (2015) e1004186.
[63] S.E. Seemann, J. Gorodkin, R. Backofen, Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments, Nucleic Acids Res. 36 (2008) 6355–6362.
[64] S. Biswas, C.M. Rao, Epigenetics in cancer: fundamentals and beyond, Pharmacol. Ther. 173 (2017) 118–134.
[65] J. Ruddel, V.E. Wennekes, W. Meissner, J.A. Werner, R. Mandic, EGF-dependent induction of BCL-xL and p21CIP1/WAF1 is highly variable in HNSCC cells—implications for EGFR-targeted therapies, Anticancer Res. 30 (2010) 4579–4585.
[66] O. Gallo, E. Masini, L. Morbidelli, A. Franchi, I. Fini-Storchi, W.A. Vergari, M. Ziche, Role of nitric oxide in angiogenesis and tumor progression in head and neck cancer, J. Natl. Cancer Inst. 90 (1998) 587–596.
[67] A. Franchi, O. Gallo, M. Paglierani, I. Sardi, L. Magnelli, E. Masini, M. Santucci, Inducible nitric oxide synthase expression in laryngeal neoplasia: correlation with angiogenesis, Head Neck 24 (2002) 16–23.
[68] M. De Ridder, G. Van Esch, B. Engels, V. Verovski, G. Storme, Hypoxic tumor cell radiosensitization: role of the iNOS/NO pathway, Bull. Cancer 95 (2008) 282–291.
[69] A. Ptak, A. Wrobel, E.L. Gregoraszczuk, Effect of bisphenol-A on the expression of selected genes involved in cell cycle and apoptosis in the OVCAR-3 cell line, Toxicol. Lett. 202 (2011) 30–35.
[70] C.S. Mak, M.M. Yung, L.M. Hui, L.L. Leung, R. Liang, K. Chen, S.S. Liu, Y. Qin, T.H. Leung, K.F. Lee, K.K. Chan, H.Y. Ngan, D.W. Chan, MicroRNA-141 enhances anoikis resistance in metastatic progression of ovarian cancer through targeting KLF12/Sp1/survivin axis, Mol. Cancer 16 (2017) 11.
[71] S. Sankaran, D.E. Crone, R.E. Palazzo, J.D. Parvin, BRCA1 regulates gamma-tubulin binding to centrosomes, Cancer Biol. Ther. 6 (2007) 1853–1857.
[72] M.H. Wu, C. Huang, K. Gan, H. Huang, Q. Chen, J. Ouyang, Y.L. Tang, X.L. Li, Y.X. Yang, H.D. Zhou, Y.H. Zhou, Z.Y. Zeng, L. Xiao, D. Li, K. Tang, S.R. Shen, G.Y. Li, LRRC4, a putative tumor suppressor gene, requires a functional leucine-rich repeat cassette domain to inhibit proliferation of glioma cells in vitro by modulating the extracellular signal-regulated kinase/protein kinase B/nuclear factor-kappa B pathway, Mol. Biol. Cell 17 (2006) 3534–3542.
[73] N. Yokdang, J. Hatakeyama, J.H. Wald, C. Simion, J.D. Tellez, D.Z. Chang, M.M. Swamynathan, M. Chen, W.J. Murphy, K.L. Carraway, C. Sweeney, LRIG1 opposes epithelial-to-mesenchymal transition and inhibits invasion of basal-like breast cancer cells, Oncogene 35 (2016) 2932–2947.
[74] W.H. Yang, Y.H. Su, W.H. Hsu, C.C. Wang, J.L. Arbiser, M.H. Yang, Imipramine blue halts head and neck cancer invasion through promoting F-box and leucine-rich repeat protein 14-mediated Twist1 degradation, Oncogene 35 (2016) 2287–2298.
[75] A.G. Kondratov, L.A. Stoliar, S.M. Kvasha, V.V. Gordiyuk, Y.M. Zgonnyk, A.V. Gerashchenko, A.F. Vozianov, A.V. Ryndtch, E.R. Zabarovsky, V.I. Kashuba, Methylation pattern of the putative tumor-suppressor gene LRRC3B promoter in clear cell renal cell carcinomas, Mol. Med. Rep. 5 (2012) 509–512.
[76] B.H. Sorensen, C.S. Dam, S. Sturup, I.H. Lambert, Dual role of LRRC8A-containing transporters on cisplatin resistance in human ovarian cancer cells, J. Inorg. Biochem. 160 (2016) 287–295.
[77] M. Ke, L. Mo, W. Li, X. Zhang, F. Li, H. Yu, Ubiquitin ligase SMURF1 functions as a prognostic marker and promotes growth and metastasis of clear cell renal cell carcinoma, FEBS Open Bio 7 (2017) 577–586.
[78] J. Ma, J. Peng, R. Mo, S. Ma, J. Wang, L. Zang, W. Li, J. Fan, Ubiquitin E3 ligase UHRF1 regulates p53 ubiquitination and p53-dependent cell apoptosis in clear cell Renal Cell Carcinoma, Biochem. Biophys. Res. Commun. 464 (2015) 147–153.
[79] V.S. Sharma, D. Magde, Activation of soluble guanylate cyclase by carbon monoxide and nitric oxide: a mechanistic model, Methods 19 (1999) 494–505.
[80] G. Schultz, W. Rosenthal, Principles of transmembranous signal transduction in the action of hormones and neurotransmitters, Arzneimittelforschung 35 (1985) 1879–1885.
[81] P.F. Windham, H.N. Tinsley, cGMP signaling as a target for the prevention and treatment of breast cancer, Semin. Cancer Biol. 31 (2015) 106–110.
[82] H.C. Wen, C.P. Chuu, C.Y. Chen, S.G. Shiah, H.J. Kung, K.L. King, L.C. Su, S.C. Chang, C.H. Chang, Elevation of soluble guanylate cyclase suppresses proliferation and survival of human breast cancer cells, PLoS One 10 (2015) e0125518.
[83] K. Shailubhai, H.H. Yu, K. Karunanandaa, J.Y. Wang, S.L. Eber, Y. Wang, N.S. Joo, H.D. Kim, B.W. Miedema, S.Z. Abbas, S.S. Boddupalli, M.G. Currie, L.R. Forte, Uroguanylin treatment suppresses polyp formation in the Apc(Min/+) mouse and induces apoptosis in human colon adenocarcinoma cells via cyclic GMP, Cancer Res. 60 (2000) 5151–5157.
[84] T.R. Tuttle, M.L. Mierzwa, S.I. Wells, S.R. Fox, N. Ben-Jonathan, The cyclic GMP/protein kinase G pathway as a therapeutic target in head and neck squamous cell carcinoma, Cancer Lett. 370 (2016) 279–285.
[85] J.C. Wong, M. Bathina, R.R. Fiscus, Cyclic GMP/protein kinase G type-I alpha (PKG-I alpha) signaling pathway promotes CREB phosphorylation and maintains higher c-IAP1, livin, survivin, and Mcl-1 expression and the inhibition of PKG-I alpha kinase activity synergizes with cisplatin in non-small cell lung cancer cells, J. Cell. Biochem. 113 (2012) 3587–3598.
[86] E.S. Robinson, E.V. Khankin, T.K. Choueiri, M.S. Dhawan, M.J. Rogers, S.A. Karumanchi, B.D. Humphreys, Suppression of the nitric oxide pathway in metastatic renal cell carcinoma patients receiving vascular endothelial growth factor-signaling inhibitors, Hypertension 56 (2010) 1131–1136.
[87] B. Muz, P. de la Puente, F. Azab, A.K. Azab, The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy, Hypoxia 3 (2015) 83–92.
[88] S. Rocha, Gene regulation under low oxygen: holding your breath for transcription, Trends Biochem. Sci. 32 (2007) 389–397.
[89] B. Keith, R.S. Johnson, M.C. Simon, HIF1alpha and HIF2alpha: sibling rivalry in hypoxic tumour growth and progression, Nat. Rev. Cancer 12 (2011) 9–22.
[90] B. Elenbaas, R.A. Weinberg, Heterotypic signaling between epithelial tumor cells and fibroblasts in carcinoma formation, Exp. Cell Res. 264 (2001) 169–184.
[91] N.A. Bhowmick, E.G. Neilson, H.L. Moses, Stromal fibroblasts in cancer initiation and progression, Nature 432 (2004) 332–337.
[92] A. Neesse, P. Michl, K.K. Frese, C. Feig, N. Cook, M.A. Jacobetz, M.P. Lolkema, M. Buchholz, K.P. Olive, T.M. Gress, D.A. Tuveson, Stromal biology and therapy in pancreatic cancer, Gut 60 (2011) 861–868.
[93] N.R. Smith, D. Baker, M. Farren, A. Pommier, R. Swann, X. Wang, S. Mistry, K. McDaid, J. Kendrew, C. Womack, S.R. Wedge, S.T. Barry, Tumor stromal architecture can define the intrinsic tumor response to VEGF-targeted therapy, Clin. Cancer Res. 19 (2013) 6943–6956.
[94] A. Orimo, R.A. Weinberg, Heterogeneity of stromal fibroblasts in tumors, Cancer Biol. Ther. 6 (2007) 618–619.
[95] A.F. Fernandez, Y. Assenov, J.I. Martin-Subero, B. Balint, R. Siebert, H. Taniguchi, H. Yamamoto, M. Hidalgo, A.C. Tan, O. Galm, I. Ferrer, M. Sanchez-Cespedes, A. Villanueva, J. Carmona, J.V. Sanchez-Mut, M. Berdasco, V. Moreno, G. Capella, D. Monk, E. Ballestar, S. Ropero, R. Martinez, M. Sanchez-Carbayo, F. Prosper, X. Agirre, M.F. Fraga, O. Grana, L. Perez-Jurado, J. Mora, S. Puig, J. Prat, L. Badimon, A.A. Puca, S.J. Meltzer, T. Lengauer, J. Bridgewater, C. Bock, M. Esteller, A DNA methylation fingerprint of 1628 human samples, Genome Res. 22 (2012) 407–419.
[96] C. Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M.J. Ziller, V. Amin, J.W. Whitaker, M.D. Schultz, L.D. Ward, A. Sarkar, G. Quon, R.S. Sandstrom, M.L. Eaton, Y.C. Wu, A.R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R.A. Harris, N. Shoresh, C.B. Epstein, E. Gjoneska, D. Leung, W. Xie, R.D. Hawkins, R. Lister, C. Hong, P. Gascard, A.J. Mungall, R. Moore, E. Chuah, A. Tam, T.K. Canfield, R.S. Hansen, R. Kaul, P.J. Sabo, M.S. Bansal, A. Carles, J.R. Dixon, K.H. Farh, S. Feizi, R. Karlic, A.R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T.R. Mercer, S.J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R.C. Sallari, K.T. Siebenthall, N.A. Sinnott-Armstrong, M. Stevens, R.E. Thurman, J. Wu, B. Zhang, X. Zhou, A.E. Beaudet, L.A. Boyer, P.L. De Jager, P.J. Farnham, S.J. Fisher, D. Haussler, S.J. Jones, W. Li, M.A. Marra, M.T. McManus, S. Sunyaev, J.A. Thomson, T.D. Tlsty, L.H. Tsai, W. Wang, R.A. Waterland, M.Q. Zhang, L.H. Chadwick, B.E. Bernstein, J.F. Costello, J.R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J.A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes, Nature 518 (2015) 317–330.
[97] L.M. Dillon, J.R. Bean, W. Yang, K. Shee, L.K. Symonds, J.M. Balko, W.H. McDonald, S. Liu, A.M. Gonzalez-Angulo, G.B. Mills, C.L. Arteaga, T.W. Miller, P-REX1 creates a positive feedback loop to activate growth factor receptor, PI3K/AKT and MEK/ERK signaling in breast cancer, Oncogene 34 (2015) 3968–3976.
[98] K.H. Limesand, A.M. Chibly, A. Fribley, Impact of targeting insulin-like growth factor signaling in head and neck cancers, Growth Hormon. IGF Res. 23 (2013) 135–140.
[99] X. He, J. Wang, E.M. Messing, G. Wu, Regulation of receptor for activated C kinase 1 protein by the von Hippel-Lindau tumor suppressor in IGF-I-induced renal carcinoma cell invasiveness, Oncogene 30 (2011) 535–547.
[100] P.J. Beltran, F.J. Calzone, P. Mitchell, Y.A. Chung, E. Cajulis, G. Moody, B. Belmontes, C.M. Li, S. Vonderfecht, V.E. Velculescu, G.R. Yang, J.W. Qi, D.J. Slamon, G.E. Konecny, Ganitumab (AMG 479) inhibits IGF-II-dependent ovarian cancer growth and potentiates platinum-based chemotherapy, Clin. Cancer Res. 20 (2014) 2947–2958.
[101] A. Kwong, J.W. Chen, V.Y. Shin, J.C.W. Ho, F.B.F. Law, C.H. Au, T.L. Chan, E.S.K. Ma, J.M. Ford, The importance of analysis of long-range rearrangement of BRCA1 and BRCA2 in genetic diagnosis of familial breast cancer, Cancer Genet. 208 (2015) 448–454.

[102] P.A. James, S. Sawyer, S. Boyle, M.A. Young, S. Kovalenko, R. Doherty, J. McKinley, K. Alsop, V. Beshay, M. Harris, S. Fox, G.J. Lindeman, G. Mitchell, Large genomic rearrangements in the familial breast and ovarian cancer gene BRCA1 are associated with an increased frequency of high risk features, Familial Cancer 14 (2015) 287–295.

[103] J.C. Hodge, K.E. Pearce, X.K. Wang, A.E. Wiktor, A.M. Oliveira, P.T. Greipp, Molecular cytogenetic analysis for TFE3 rearrangement in Xp11.2 renal cell carcinoma and alveolar soft part sarcoma: validation and clinical experience with 75 cases, Mod. Pathol. 27 (2014) 113–127.

[104] D.L. Aisner, T.T. Nguyen, D.D. Paskulin, A.T. Le, J. Haney, N. Schulte, F. Chionh, J. Hardingham, J. Mariadason, N. Tebbutt, R.C. Doebele, A.J. Weickhardt, M. Varella-Garcia, ROS1 and ALK fusions in colorectal cancer, with evidence of intratumoral heterogeneity for molecular drivers, Mol. Cancer Res. 12 (2014) 111–118.

[105] X.X. Tian, D. Rai, J. Li, C. Zou, Y. Bai, D. Wazer, V. Band, Q. Gao, BRCA2 suppresses cell proliferation via stabilizing MAGE-D1, Cancer Res. 65 (2005) 4747–4753.