Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway

Jing Yang[1]◔, Lei Chen[2]◔, Xiangyin Kong[1]*, Tao Huang[3]*, Yu-Dong Cai[4]*

1 The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Jiao Tong University School of Medicine (SJTUSM) and Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai, People's Republic of China, **2** College of Information Engineering, Shanghai Maritime University, Shanghai, People's Republic of China, **3** Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, United States of America, **4** Institute of Systems Biology, Shanghai University, Shanghai, People's Republic of China

Abstract

Cancer is a serious disease that causes many deaths every year. We urgently need to design effective treatments to cure this disease. Tumor suppressor genes (TSGs) are a type of gene that can protect cells from becoming cancerous. In view of this, correct identification of TSGs is an alternative method for identifying effective cancer therapies. In this study, we performed gene ontology (GO) and pathway enrichment analysis of the TSGs and non-TSGs. Some popular feature selection methods, including minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS), were employed to analyze the enrichment features. Accordingly, some GO terms and KEGG pathways, such as biological adhesion, cell cycle control, genomic stability maintenance and cell death regulation, were extracted, which are important factors for identifying TSGs. We hope these findings can help in building effective prediction methods for identifying TSGs and thereby, promoting the discovery of effective cancer treatments.

Citation: Yang J, Chen L, Kong X, Huang T, Cai Y-D (2014) Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway. PLoS ONE 9(9): e107202. doi:10.1371/journal.pone.0107202

Editor: William B. Coleman, University of North Carolina School of Medicine, United States of America

Received June 1, 2014; **Accepted** August 7, 2014; **Published** September 10, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the National Basic Research Program of China (2011CB510102, 2011CB510101), the National Natural Science Foundation of China (31371335, 61202021, 61373028, 11371008, 81030015), Innovation Program of Shanghai Municipal Education Commission (12ZZ087, 12YZ120), the Shanghai Educational Development Foundation (12CG55) and the grant from "The First-class Discipline of Universities in Shanghai." The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: xykong@sibs.ac.cn (XK); tohuangtao@126.com (TH); cai_yud@126.com (YDC)

◔ These authors contributed equally to this work.

Introduction

Currently, cancer is the second most common cause of death, following cardiovascular disease. Cancer that originates from the epithelial cells or mesenchymal cells is characterized by uncontrolled cell proliferation. In malignancy, cancer cells invade adjacent normal tissues and metastasize through blood circulation, lymphokinesis or body cavity transfer. In this process, proteins that are coded by tumor suppressor genes (TSGs) play vital roles in the mechanisms associated with cellular growth, DNA damage, apoptosis and metabolic regulation [1].

It has been reported that tumor suppressor inactivation and haploinsufficiency occur at several different levels in tumor patients. In the past decades, many classic TSGs have been widely identified, which are silenced by recurrent LOH (loss of heterozygosity) and physical deletion in the tumor genome. Increasing evidence has shown the abnormal DNA methylation or histone modifications, and non-coding RNA affect the expression of TSGs at the epigenetic level and post-transcriptional level, respectively [2,3].

The first identified TSG was retinoblastoma protein (Rb), which was identified by studies of familial retinoblastoma in early childhood. Based on this, the "two-hit" hypothesis was introduced by Knudson in 1971 [4,5]. As a guardian to the normal cell cycle, the Rb protein is responsible for the G1/S checkpoint and maintains regular cell growth. In addition to loss of heterozygosity, the high frequent mutations or partial deletions are mainly located in exon13~exon17 of Rb and have been found in various cancer types, especially in lung cancer, breast cancer, osteosarcoma and bladder cancer, with a frequency ranging from 15% to 50% [6–10]. Like Rb, the p53 protein family as a key element of the tumor suppression network, exerts much of its growth arrest in the cell cycle and induces apoptosis. Changes to p53 are involved in various cancers. Genetic variation mainly missense mutations, in p53 are often regarded as the driver mutations that confer apoptosis evasion and abnormal cell growth of tumor cells, especially those that originate from the epithelial tissue. More than 86% of point mutations occur in the evolutionary conservative regions, especially four mutation hotspots [11,12]. In addition, p53 is silenced via LOH in the genome and hypermethylation at the epigenetic level in cancer patients [13,14].

Like Rb and p53, some tumor suppressor proteins control cell behaviors directly by arresting cell proliferation, disturbing the cell cycle and inducing apoptosis, and these are called the gatekeepers. The destiny of a cell is also affected indirectly by some tumor suppressor proteins that are associated with mutation accumulation and genome stability maintenance such as BRCA1 and BRCA2, which are also referred to as caretakers [15,16].

Additionally inherited mutations of BRCA1 and BRCA2 (breast cancer 1/2) are associated with patients who have hereditary breast cancer, accounting for 5–10% of all breast cancer patients [17]. Loss function of their products causes abnormal homologous recombination and genome instability, which increases the susceptibility to breast and ovarian cancer [18].

Unlike the activated oncogene, suppression of TSGs occurs more frequently, providing evidence for understanding the initiation and progress of various cancers. The identification and subsequent activation of TSGs can facilitate controlling cell proliferation, restraining the biological activity of cancer. In this study, we attempted to investigate the characteristics of TSGs. The TSGs retrieved from the web-based database, TSGene (tumor suppressor gene database), facilitated our investigation of TSGs. These genes were called 'positive genes' and all of the remaining genes in the STRING were selected as 'negative genes'. Gene Ontology (GO) is an acknowledged bioinformatics tool for representing gene product properties across all species by defined GO terms, the function of the genes and their products were represented by the GO terms and predicted by the GO annotation effectively [19,20]. In contrast, the Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive database based on known molecular interaction networks and is usually used to understand biological pathways and systems [21]. In view of this, the enrichment scores of the GO terms and KEGG pathways were used to encode all genes investigated in this study. Minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS) [22] combined with a prediction engine were employed to analyze these features. The analysis of the extracted GO terms and KEGG pathways suggests that they are related to TSGs. In addition, the extracted GO terms and KEGG pathways were used to predict the novel TSGs, indicating that they may help build effective computational methods for identifying TSGs.

Materials and Methods

Dataset

We compiled 716 human TSGs in the TSGene database (http://bioinfo.mc.vanderbilt.edu/TSGene/download.cgi), which were collected from two resources: public databases and literature reports. In detail, 187 (human) and 170 (human) known TSGs were retrieved from UniProtKB (28 January, 2012) and the TAG database (http://www.binfo.ncku.edu.tw/TAG/GeneDoc.php) (29 March, 2012), respectively, with only 41 overlapped genes by mapping to the Entrez gene symbols. By combining two exhaustive searches, PubMed and Gene Reference Into Function (GeneRIF) [23,24], and after overlapping and synonymous genes with same the Entrez gene ID were filtered, 637 protein-coding TSGs and 79 non-coding TSGs were identified [25]. Because the encoding method described in Section "Encoding method" employed the neighbors of each investigated TSG in the STRING, we obtained 615 genes with their ensembl protein IDs in the STRING. These genes were termed 'positive genes' and are given in Table S1. The remaining 17,985 ensembl protein IDs in the STRING were considered 'negative genes'.

The number of negative genes was much larger than that of the positive genes. This is an imbalanced dataset. Inspired by some studies dealing with this type of data [26,27], we divided the 17,985 negative genes into six datasets, $A_1,A_2,\ldots,A_6$, where $A_1,A_2,\ldots,A_5$ contained 3,075 negative genes and, A_6 contained 2,610 negative genes. The 615 positive genes were put into each of these datasets, comprising six new datasets, $S_1,S_2,\ldots,S_6$, *i.e.*, S_i ($i=1,2,3,4,5,6$) consisting of genes in A_i ($i=1,2,3,4,5,6$) and 615 positive genes.

Encoding method

To analyze the characteristics of the TSGs, it is very important to encode each gene with its essential properties. GO is an acknowledged bioinformatics tool for representing gene product properties across all species by defined GO terms, while KEGG is a comprehensive database based on known molecular interaction networks and usually includes the biological pathway and system information [21]. Therefore, we selected GO terms and KEGG pathways to code each gene. TSGs have a strong relationship with some GO terms and KEGG pathways. On the other hand, the enrichment method of GO can reflect the relationship between the genes and GO terms [28]. It is reasonable to use this method to encode genes and analyze the relationship of the TSGs and GO terms. Furthermore, this method can also be extended to KEGG pathways [29] to find the relationship between the genes and KEGG pathways.

GO enrichment. For one gene g and one GO term GO_j, the GO enrichment score is defined as the $-\log_{10}$ of the hypergeometric test P value [28−30] of a gene set G containing g's direct neighbors in the protein-protein interaction network of STRING and GO term GO_j, which can be calculated by:

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

where N is the number of overall proteins in human, M is the number of proteins annotated to the GO term GO_j, n is the number of proteins in G, and m is the number of proteins in G, which are annotated to the GO term GO_j. The high score for one gene and one GO term implies that the gene and GO term have a

Table 1. The number of remaining features after using Cramer's coefficient to exclude non-essential features.

Dataset	Number of remaining features
S_1	3,347
S_2	3,837
S_3	4,632
S_4	4,270
S_5	4,956
S_6	6,661

doi:10.1371/journal.pone.0107202.t001

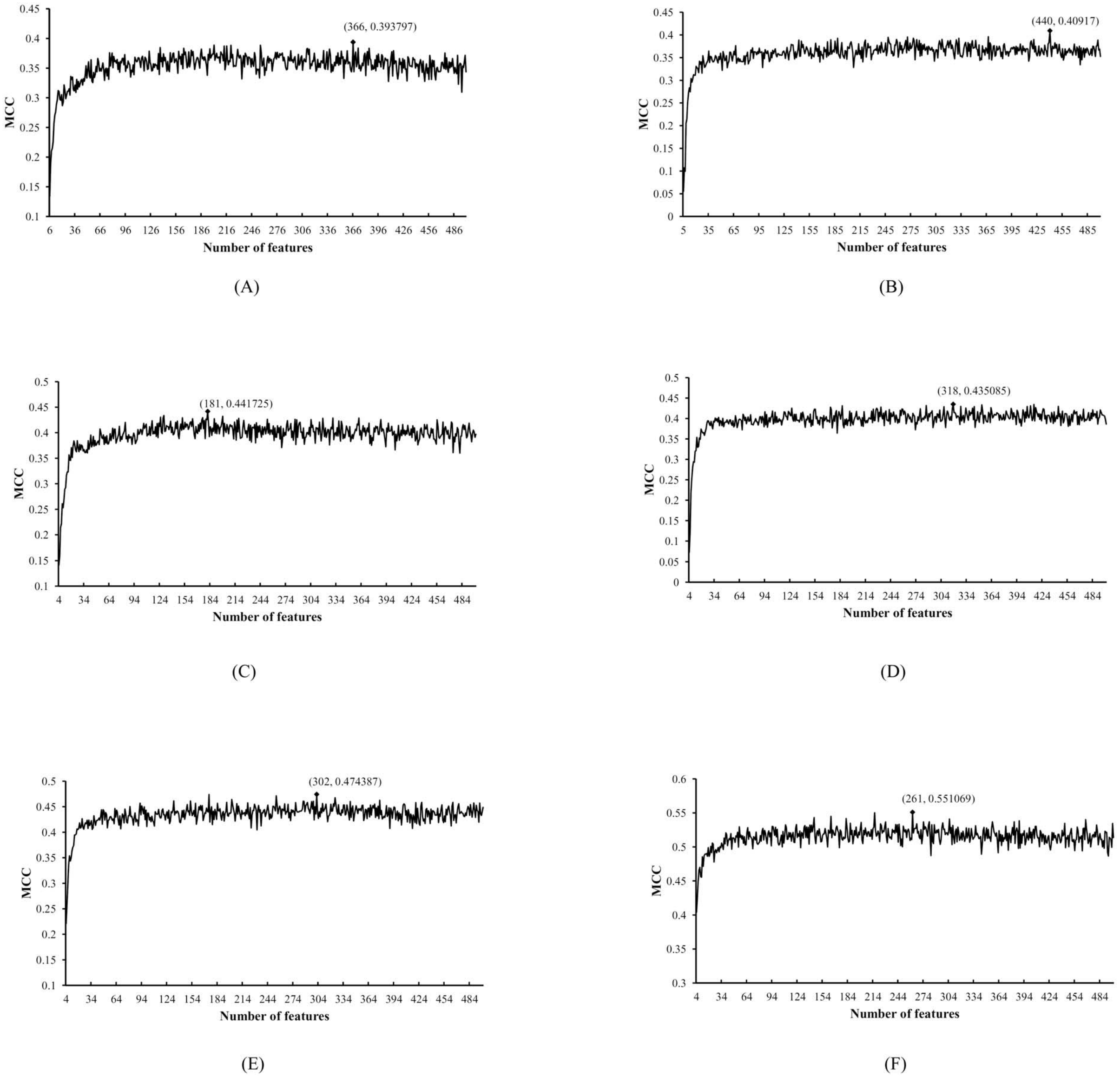

Figure 1. Six IFS-curves for six datasets. In detail, (A) shows the IFS-curve for the dataset S_1; (B) shows the IFS-curve for the dataset S_2; (C) shows the IFS-curve for the dataset S_3; (D) shows the IFS-curve for the dataset S_4; (E) shows the IFS-curve for the dataset S_5; (F) shows the IFS-curve for the dataset S_6. The Y-axis represents the Matthews's correlation coefficient (MCC) and the X-axis represents the number of features participating in the classification model.

doi:10.1371/journal.pone.0107202.g001

Table 2. The number of features in the optimal feature set for each dataset and the MCC values obtained by using these features.

Dataset	Number of features in the optimal feature set	Maximum MCC value
S_1	366	0.3938
S_2	440	0.4092
S_3	181	0.4417
S_4	318	0.4351
S_5	302	0.4744
S_6	261	0.5511
Mean		0.4509

doi:10.1371/journal.pone.0107202.t002

special relationship. The 12,877 GO terms induced 12,877 GO enrichment scores.

KEGG enrichment. For one gene g and one KEGG pathway P_j, the KEGG enrichment score is defined as the $-\log_{10}$ of the hypergeometric test P value [29] of a gene set G containing g's direct neighbors in the protein-protein interaction network of STRING and KEGG pathway P_j, which can be computed as follows:

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

where N is the number of overall proteins in human, M is the number of proteins in the KEGG pathway P_j, n is the number of proteins in G, m is the number of proteins in both G and P_j. Additionally, the higher the KEGG enrichment score for g and P_j, the stronger the relationship between them. The 239 KEGG pathways induced 239 features of KEGG enrichment scores.

Each of the 12,877 GO enrichment scores or each of the 239 KEGG enrichment scores can be considered a dimension. Accordingly, each gene g can be represented by a vector in 12,877+239 = 13,116-D space, which is formulated as:

$$\begin{aligned} v_g = (&S_{\mathrm{GO}}(g,\mathrm{GO}_1),\ldots,S_{\mathrm{GO}}(g,\mathrm{GO}_{12877}),\\ &S_{\mathrm{KEGG}}(g,P_1),\ldots,S_{\mathrm{KEGG}}(g,P_{239}))^{\mathrm{T}} \end{aligned} \qquad (3)$$

Prediction method

Dagging is a well-known meta classifier. The main idea of this classifier is to integrate multiple classifiers derived from a single learning algorithm that is trained by disjoint samples of the original dataset [31]. The brief description of this method is as follows. For a training dataset $\Im$ with samples $s_1,s_2,\ldots,s_n$, construct k disjoint subsets by randomly taking n' samples in $\Im$, without replacement, such that $kn' \leq n$. These subsets were used to train a basic classifier (*e.g.,* support vector machine) and derive k classification models, $M_1,M_2,\ldots,M_k$. For a query sample, each of these models M_i ($1 \leq i \leq k$) provides a predicted result. The predicted result of dagging integrated these results by majority voting.

In Weka 3.6.4 [32], the classifier "Dagging" implements the dagging classifier mentioned above. Here, it was adopted as the prediction engine. For convenience, it was run with its default parameters. In detail, the SMO (Sequential Minimal Optimization), which implements John Platt's sequential minimal optimization algorithm for solving the optimization problem during the training of a support vector classifier using polynomial or Gaussian kernels [33,34], is set as the basic classifier, and k is set to 10.

Evaluation method

Ten-fold cross-validation is a widely used cross-validation method for evaluating the performance of different classification models [35−38]. Compared to the Jackknife test [39,40], the 10-fold cross-validation test requires less computing time and provides similar results for a given dataset. Therefore, the current study adopted this cross-validation method to evaluate the performance of the prediction method.

To represent the predicted results of a two-class classification problem, a confusion matrix was often employed, which contained the following four entries: true positives (TP), true negative (TN), false positives (FP), and false negative (FN) [41,42]. Based on these values, the prediction accuracy (ACC), specificity (SP), sensitivity (SN) [42] and Matthews's correlation coefficient (MCC) [43] were often used to evaluate the predicted results, which can be computed by

$$\begin{cases} \mathrm{ACC} = \dfrac{TP+TN}{TP+TN+FP+FN} \\ \mathrm{SP} = \dfrac{TN}{TN+FP} \\ \mathrm{SN} = \dfrac{TP}{TP+FN} \\ \mathrm{MCC} = \dfrac{TP\cdot TN - FP\cdot FN}{\sqrt{(TN+FN)\cdot(TN+FP)\cdot(TP+FN)\cdot(TP+FP)}} \end{cases} \qquad (4)$$

As mentioned in Section "Dataset", five datasets were constructed in this study to reduce the size difference of the 'positive genes' and 'negative genes'. However, each dataset still had very different class sizes. In detail, the number of 'negative genes' was at least 4 times as many as that of 'positive genes'. Thus, the ACC is not appropriate for evaluating the predicted results on the whole. MCC, as a balanced measure even if the classes are of very different sizes, was employed as the key measurement.

Feature selection method

As mentioned in Section "Encoding method", each gene was represented by 13,116 features of the enrichment scores, which indicated the relationship between the genes and GO terms or KEGG pathways. TSGs are related to some GO terms and KEGG pathways. To identify key GO terms and KEGG pathways, some feature selection methods were employed in this study. The procedure of the feature selection method included two stages: (I) Cramer's coefficient [44,45], which used to discard non-essential features and (II) minimum redundancy maximum relevance (mRMR), incremental feature selection (IFS) [22] and Dagging [31] for further selection.

The Cramer's coefficient [44,45], derived from the Pearson Chi-square test [46], is a statistical measure of two variables. Its value is between 0 and 1. According to the fact that a high Cramer's coefficient of two variables indicates a strong association of two variables, features with low Cramer's coefficients to samples' class labels were deemed non-essential features. Here, we used 0.1 as the threshold and features with Cramer's coefficients lower than 0.1 were excluded.

The second stage of the feature selection involved the mRMR, IFS and Dagging. In detail, the mRMR method sorted the remaining features in two lists, while the IFS and Dagging were used to extract key features based on the feature lists obtained by the mRMR method. The mRMR method, proposed by *Peng et al.* [22], has two criteria: Max-Relevance and Min-Redundancy, producing the following two feature lists: (I) MaxRel feature list and (II) mRMR feature list. The MaxRel feature list sort features only based on the Max-Relevance criterion, while the mRMR feature list sort features based on both the Max-Relevance and Min-Redundancy. In this study, these two lists were formulated as follows:

$$\begin{cases} \text{MaxRel features list} : F_{\mathrm{MaxRel}} = [f_1^M, f_2^M, \cdots, f_N^M] \\ \text{mRMR features list} : F_{\mathrm{mRMR}} = [f_1^m, f_2^m, \cdots, f_N^m] \end{cases} \qquad (5)$$

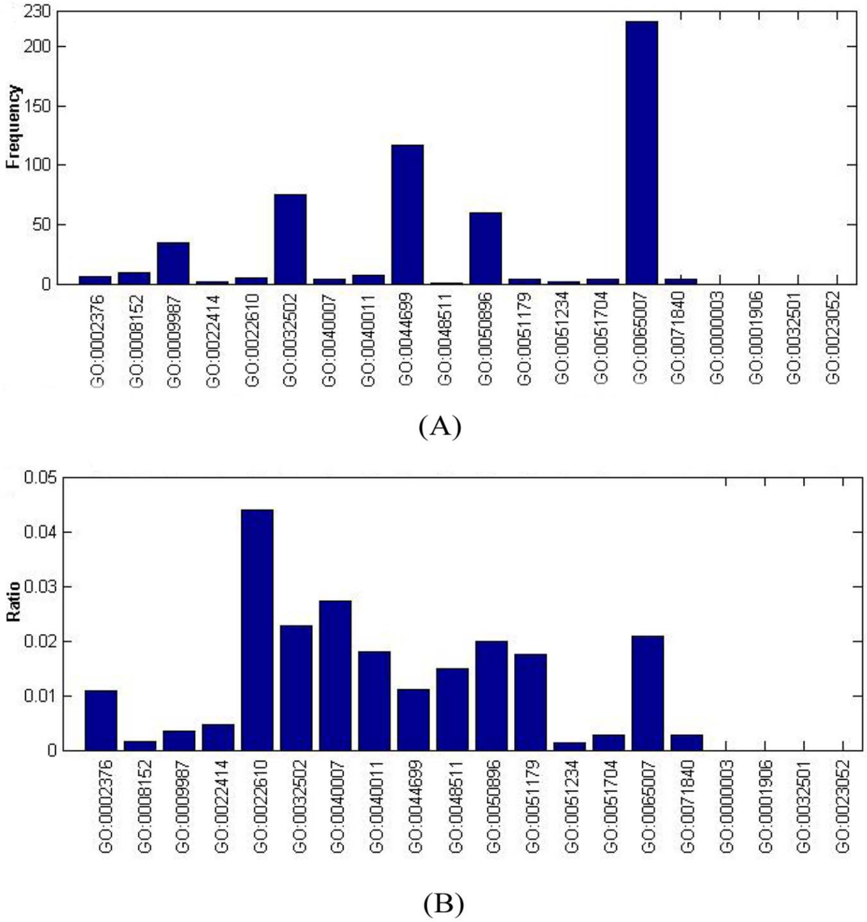

Figure 2. Frequency and ratio of GO terms of biological process in *OS*. (A) Frequency of GO terms of biological process in *OS*. (B) Ratio of GO terms of biological process in *OS*.
doi:10.1371/journal.pone.0107202.g002

where N is the total number of features. The mRMR method has been widely used in recent years to analyze complicated biological problems [36,47−52]. Since the mRMR feature list was built with both the Max-Relevance and Min-Redundancy criteria in mind, it was used to extract important features by combining the IFS and Dagging. This procedure was as follows:

(I) Construct N feature set from the mRMR features list F_{mRMR}, say $F^1_{\text{mRMR}},F^2_{\text{mRMR}},\ldots,F^N_{\text{mRMR}}$, such that $F^i_{\text{mRMR}}=[f_1^m,f_2^m,\cdots,f_i^m](1\le i\le N)$, *i.e.* F^i_{mRMR} consisted of the first i features in F_{mRMR}.

(II) For each F^i_{mRMR}, Dagging was conducted on samples represented by features in F^i_{mRMR}, evaluated by 10-fold cross-validation, thereby obtaining ACC, SP, SN and MCC (cf. **Eq. 4**).

(III) The feature set that can produce the maximum MCC is the optimal feature set. Additionally, an IFS-curve was plotted with the MCC value as its Y-axis and the superscript i of F^i_{mRMR} (the number of features that participate in the classification) as its X-axis.

Results and Discussion

Results of the feature selection

As mentioned in Section "Dataset", 6 datasets, $S_1,S_2,\ldots,S_6$, were constructed. For each, we calculated the Cramer's coefficients of the features and the samples' class labels. Then, the features with Cramer's coefficients lower than 0.1 were excluded. The remaining features were kept for the further selection. The number of remaining features for each dataset is shown in **Table 1**.

The mRMR method, IFS method and Dagging were used to analyze the remaining features for each dataset S_i. The mRMR program, downloaded from http://research.janelia.org/peng/proj/mRMR/, was executed on each dataset S_i, in which each sample was represented by the remaining features. For convenience, the mRMR method was conducted with its default parameters. As mentioned in Section "Feature selection method", the MaxRel features list and mRMR features list were obtained for each dataset S_i. However, to reduce the computation time, we only obtained the first 500 features in each of the two feature lists, which are summarized in Table S2.

The IFS method and classifier Dagging were executed according to the mRMR features list for each dataset S_i, which was evaluated by 10-fold cross-validation. The SNs, SPs, ACCs and MCCs obtained for each dataset S_i are given in Table S3. For easy observation, we plotted an IFS-curve for each dataset S_i. The six IFS-curves are shown in **Figure 1**; the maximum MCCs for datasets $S_1,S_2,\ldots,S_6$ were 0.3938, 0.4092, 0.4417, 0.4351, 0.4744, and 0.5511, respectively. These values are listed in **Table 2**, in which the numbers of the features used to obtain these maximum MCCs are also listed. In detail, by using the first

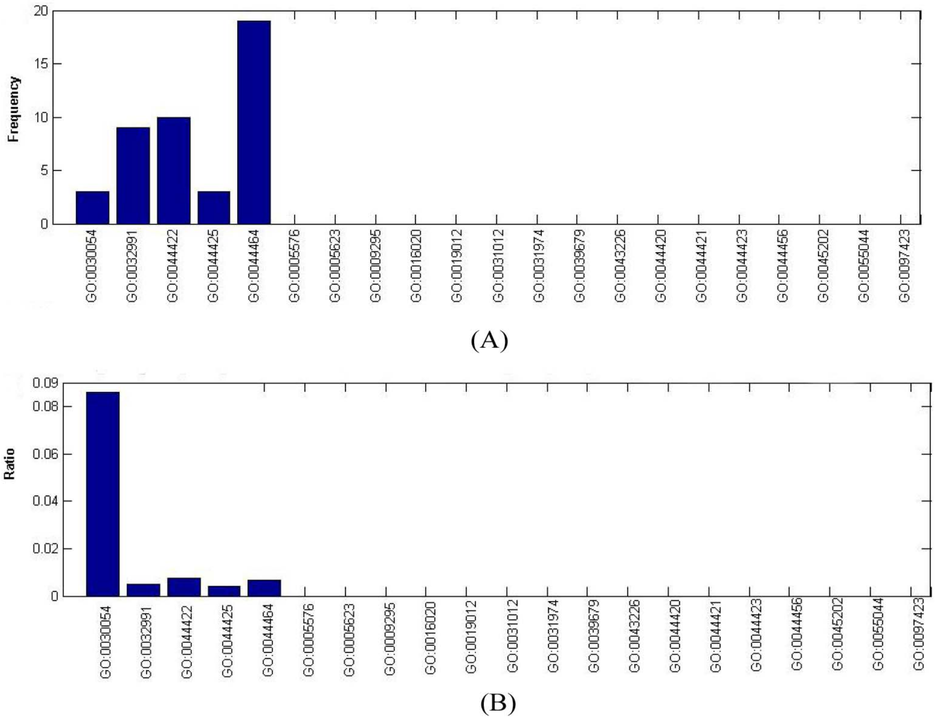

Figure 3. Frequency and ratio of GO terms of cellular component in *OS*. (A) Frequency of GO terms of cellular component in *OS*. (B) Ratio of GO terms of cellular component in *OS*.

doi:10.1371/journal.pone.0107202.g003

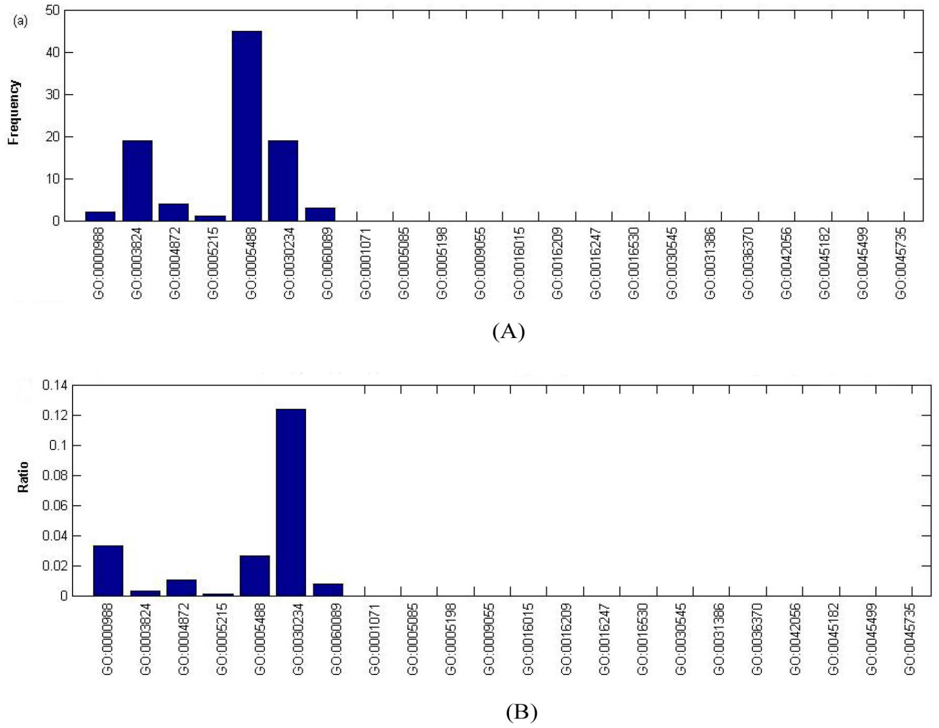

Figure 4. Frequency and ratio of GO terms of molecular function in *OS*. (A) Frequency of GO terms of molecular function in *OS*. (B) Ratio of GO terms of molecular function in *OS*.

doi:10.1371/journal.pone.0107202.g004

Table 3. Top forty putative tumor suppressors based on features in the total optimal feature set.

Ensembl ID	Number of key tumor suppressor functions[a]	Gene symbol
ENSP00000297261	353	SHH
ENSP00000324806	353	GSK3B
ENSP00000389184	345	MARK2
ENSP00000264657	338	STAT3
ENSP00000355069	338	PAX2
ENSP00000293549	337	WNT1
ENSP00000353483	331	MAPK8
ENSP00000263253	331	EP300
ENSP00000218894	327	SUPT20H
ENSP00000328181	327	NOG
ENSP00000228872	327	CDKN1B
ENSP00000338548	325	FGF1
ENSP00000250003	322	MYOD1
ENSP00000206249	322	ESR1
ENSP00000245451	321	BMP4
ENSP00000352514	317	RUNX2
ENSP00000348986	316	INS-IGF2
ENSP00000263025	315	MAPK3
ENSP00000354558	313	MTOR
ENSP00000363822	311	AR
ENSP00000361066	310	NCOA3
ENSP00000339004	309	FOXG1
ENSP00000320604	309	FAXDC2
ENSP00000338018	308	HIF1A
ENSP00000278385	308	CD44
ENSP00000216797	306	NFKBIA
ENSP00000222330	304	GSK3A
ENSP00000255465	304	CCNA1
ENSP00000222726	303	HOXA5
ENSP00000334458	303	GATA4
ENSP00000264498	303	FGF2
ENSP00000323588	302	SOX2
ENSP00000392858	299	TNF
ENSP00000302665	299	IGF1
ENSP00000338297	298	-
ENSP00000362649	297	HDAC1
ENSP00000318977	297	GEN1
ENSP00000343745	296	DICER1
ENSP00000265165	294	LEF1
ENSP00000415481	293	PROM1

[a] The value in this column is the number of features in the total optimal feature set whose values are greater than $-\log_{10}(0.05)$.

doi:10.1371/journal.pone.0107202.t003

366, 440, 181, 318, 302, and 261 features in the mRMR features lists of the six datasets (see Table S3), respectively, the MCCs calculated by **Eq. 4** were 0.3938, 0.4092, 0.4417, 0.4351, 0.4744, and 0.5511, respectively. Accordingly, six optimal feature sets, OS_1, OS_2, …, OS_6 can be obtained by selecting the first 366, 440, 181, 318, 302, and 261 features in six mRMR feature lists of six datasets, respectively.

Analysis of the GO terms in the total optimal feature set

As mentioned in Section "Results of the feature selection", six optimal feature sets were obtained. We took the union operation of these sets and obtained a new dataset denoted by OS ($OS = OS_1 \cup \cdots \cup OS_6$) and termed the total optimal feature set, consisting of 708 enrichment features of the GO terms and 9 enrichment features of the KEGG pathways, which are available in Table S4. The analysis of 708 GO terms is described below.

Seven hundred and eight GO terms can be divided into the following three parts: (1) Biological Process (BF); (2) Cellular Component (CC); and (3) Molecular Function (MF). We mapped the 708 GO terms to the children terms of three GO domains. As we can see in **Figures 2–4**, the GO terms in the *OS* were significantly enriched in some specific children terms with a high frequency and high ratio, which is defined as "the number of each GO term"/"the scale of the number of its children terms".

Biological process GO terms. The top five biological process GO terms of the frequency shown in **Figure 2(A)** are GO: 0065007: biological regulation (221), GO: 0044699: single-organism process (117), GO: 0032502: developmental process (75), GO: 0050896: response to stimulus (60) and GO: 0009987: cellular process (35). The top five biological process terms with large base numbers that perform fundamental functions in organisms and tumor suppressor proteins may be functional disturbance in health maintenance of cancer patients.

For the ratio of the biological process GO terms shown in **Figure 2(B)**, the top five are GO: 0022610: biological adhesion (4.39%, 5/114), GO: 0040007: growth (2.72%, 4/147), GO: 0032502: developmental process (2.28%, 75/3294), GO:0065007: biological regulation (2.09%, 221/10551) and GO:0050896: response to stimulus (2.0%, 60/3001). The GO terms biological adhesion and response to stimulus should be noted and relevant TS proteins act in the alarm reaction and have protective roles in tumorigenesis and the metastasis process. The GO term single-organism process involved in death and cell proliferation is highlighted too, although its percentage is not high. The destiny of an organism is critically regulated by the cell cycle and apoptosis in which TSGs play an important part. TSGs act like brakes on a car and are involved in maintenance of the cell cycle checkpoints and apoptosis induction [53].

Cells are under constant attack by various agents and oncogenic DNA variants form because of endogenous (normal cell metabolite) and exogenous agents (chemical species and physical mutagens). To maintain genome stability, TSGs participated in multiple mechanisms to repair DNA damage and arrest cell proliferation. In DNA double-strand break repair (DSBR), several TS genes, including ATM, NBS1, BRCA1 and BRCA2, are activated by DNA damage to induce cell cycle checkpoint arrest and DSB repair complex formation [54]. The highly conserved DNA mismatch repair (MMR) proteins composed of MSH2, MLH1, PMS1 and PMS2 tumor suppressor proteins in people, are required to correct base mismatches that are formed in response to exogenous or endogenous substances. If the expression of MLH1 or MSH2 is suppressed, cells lose the ability to perform mismatch repair and have resistance to alkylation mutagens that would normally activate G2/M checkpoint or apoptosis [55]. In nucleotide excision repair (NER), the DNA repair genes are regulated by p53 to remove bulky DNA adducts including pyrimidine dimmers induced by UV [56]. Normal, unrepaired DNA variants promote cells apoptosis.

Normally, cell proliferation is tightly regulated in different periods of the cell cycle. The pRb (retinoblastoma protein), known as the first TSG, maintains the G1/S checkpoint through its regulation of the E2F family. Inactivation of pRb, which caused by mutations, promoter methylation or interaction with oncoproteins, results in loss of control of the checkpoint R, allowing for uncontrolled cell proliferation [57,58]. In addition, cancer cells inhibit the expression of many other tumor suppressor proteins to gain malignant proliferation ability. For example, with mutations or the low expression of TGF-βR II (transforming growth factor βreceptor II) and its downstream proteins Smad2/3/4 (SMAD family member 2/3/4), cancer cells will be insensitive to the proliferation inhibition of TGF [59,60]. Similar to pRb, the INK4 (cyclin -dependent kinase inhibitor, *e.g.,* p16INK4A) family, which is regulated by TGF-β, can block CDK, causing cell growth arrest in a different period of the cell cycle. The dysfunction of INK4 or TGF-βR II will allow cells to pass through the checkpoint abnormally and accumulate variations [61,62].

Apoptosis, known as programmed cell death, can be initiated by two distinct signaling pathways, BCL2 induced and death receptor induced, which ultimately converge in the caspase cascade. The most famous TSG, p53, is mutated in ~50% of human cancers and related to some tumor suppression network [14]. p53 is a transcriptional regulator that can be activated by DNA damage, certain oncogenes and other cytotoxic stress signals, triggering cell cycle arrest (G1/S checkpoint), DNA repair and apoptosis. Dysfunction of p53 caused by mutations or methylation prevents the damage-induced cell cycle arrest and apoptosis [63,64]. As a TSG, PTEN (phosphatase with tensin homology) negatively regulates the PI3K (the phosphatidylinositol 3-kinase) pathway, preventing inappropriate metabolism via effects on TOR and promoting cell proliferation via effects on proapoptotic proteins [65]. CYLD(cylindromatosis), first identified as a TSG in the familial cylindromatosis, is a DUB (deubiquitinase) of the USP subfamily. Multiple myeloma patients with dysfunction of CYLD have abnormal activation of NF-kB and cell cycle and apoptosis dysfunction [66,67]. The insufficient activation of caspase 8 (apoptosis-related cysteine peptidase), a key TS gene in the caspase cascade, leads to the interruption of signal transduction from death receptors, inducing normal apoptosis [68,69].

Many tumor cell types acquire the capacity to invade and metastasize though loss of cell-cell adhesion or cell-ECM (extracellular matrix) junctions. The silencing or suppression of E-cadherin, which is regulated by promoter methylation, histone methylation, transcriptional repression or frequent mutations cause EMT (epithelial-mesenchymal transition), disruption of cell contacts, tumor cell detachment and invasion [70,71]. Integrins, a family of heterodimeric transmembrane proteins, mediate cell–ECM (extracellular matrix) interactions. Aberrant integrin can induce the activation of proteolytic enzymes and cause degradation of the extracellular matrix and basement membrane, promoting tumor cells metastasis [72]. MMPs (matrix metalloproteinase) are endopeptidases that are involved in the breakdown of the extracellular matrix; they are regulated by inhibitors, TIMPs (Tissue Inhibitor of Metalloproteinases). Loss of function of TIMPs, which are TSGs, may cause a MMP/TIMP equilibrium shift into a malignant status [73,74].

Except the features above, which help us comprehend the relevance between tumor suppressors and specific GO terms or pathways, some rare investigated terms were highlighted such as metabolic process (GO:0008152), reproductive process (GO:0022414), locomotion (GO:0040011), localization (GO:0051179)/establishment of localization (GO:0051234) and multi-organism process (GO:0051704). These features remind us tumor suppressors participate in protein localization intracellular, various cells migration and locomotion intercellular, complex metabolic process and multi-organism process in the whole organism. Particularly, in some tumor types, tumor suppressors play key roles in reproductive process, usually related to hormone and hormone receptors. These features are not studies deeply as others, but need more attention to mine novel tumor suppressors.

Cellular component GO terms. It can be seen from **Figure 3(A)** that the top five CC GO terms with regard to frequency are GO:0044464: cell part (19), GO:0044422: organelle part (10), GO:0032991: macromolecular complex (9), GO:0030054: cell junction (3), and GO:0044425: membrane part (3), which also have a corresponding high percentage. Their ratios (cf. **Figure 3(B)**) are GO: 0030054: cell junction (8.57%, 3/35), GO: 0044422: organelle part (0.73%, 10/1361), GO: 0044464: cell part (0.67%, 19/2823), GO: 0032991: macromolecular complex (0.49%, 9/1824), and GO:0044425: membrane part (0.41%, 3/724). Cell junction is a cellular component that forms connections between two cells or between a cell and the extracellular matrix. As discussed above, TSGs such as E-cadherin and integrin play critical roles in tumor cell adhesion and metastasis. Additionally, organelles, including the mitochondria, ribosomes and UPS (ubiquitin-proteasome system), participate in the biological process involved in carcinogenesis. Many macromolecular complexes consist of tumor suppressor protein inside cells, such as TSgene SMAD2/3(SMAD family member 2/3) in the SMAD protein complex [75] and SMARCB1(SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1) in the Swi/Snf complex [76].

Molecular function the GO terms. It can be observed from **Figure 4(A)** that the five highest frequency of MF GO terms are GO: 0005488: binding (45), GO: 0003824: catalytic activity (19), GO: 0030234: enzyme regulator activity (19), GO:0004872: receptor activity (4), and GO:0060089: molecular transducer activity (3). On one hand, these high frequency MF GO terms consist of a huge number of proteins that perform basic biological functions; on the other hand, the catalytic activity and enzyme regulator are involved in most vital biological processes, including cell proliferation, DNA damage repair and apoptosis. The cell junction requires protein binding and enzymes catalyze, which can involve biological processes such as phosphorylation, acetylation, the cell-extracellular matrix link and cell cycle control. The transcription factor Dp (DPDP-polypeptide) forms a complex with E2F1 to regulate its binding to DNA and the expression of certain genes (such as myc) catalyzed by enzymes [77]. Genomic instability is essential in almost all tumor factors, and mutations in ATM (ataxia telangiectasia mutated) which belongs to the PI3/PI4-kinase family, leave DSBs (DNA double-strand breaks) unrepaired [78]. The receptor proteins transduce extracellular or intracellular messenger to the biological effectors, triggering a serial biochemical reaction. The typical receptor protein and tumor suppressors in the TGF-β signaling pathway are TGF-βR II and BMPR2 (bone morphogenetic protein receptor, type II (serine/threonine kinase)) [79]. The five most common MF GO terms (cf. **Figure 4(B)**) are GO: 0030234: enzyme regulator (12.33%, 19/154), GO: 0000988: protein binding transcription factor activity (3.28%, 2/61), GO: 0005488: binding (2.64%, 45/1703), GO:0004872: receptor activity (1.02%, 4/391), and GO:0060089: molecular transducer activity (0.74%, 3/405). The corresponding percentages of the top five MF terms are similar to the top MF frequency, which are associated with the BP percentage and CC percentage and participate in tumorigenesis at different level.

Directed acyclic graph (DAG) analysis of the GO children terms. To further understand the function of the selected GO terms, we analyzed the directed acyclic graph of the GO children terms. We found that the GO children terms clustered in several particular modules under the primary GO terms discussed above. In addition to cell adhesion, the cellular response to UV-induced DNA damage and subsequent activated apoptotic signaling pathway and cell cycle regulation, phosphate metabolism, signal transduction and some molecular complex were highlighted in the biological modules.

The phosphorus utilization including phosphorylation and dephosphorylation catalyzed by kinases and phosphatases, respectively, is a key mechanism in a number of vital cellular pathways such as the cell cycle, cell proliferation and apoptosis. Mutations or low expression in certain TSGs, such as PTP (protein tyrosine phosphatase), should bring the phosphorylation/dephosphorylation ratio out of balance [80,81].

Cancer is a disease of aberrant signal transduction. In the functioning biological system, tumor suppressors keep the signaling cascades in balance, such as for the TGF-βR II and Smad2/3/4 in TGFβ signaling pathways [59,60] and ptch1 protein (patched 1) in hedgehog pathway [82].

In addition, some molecular complex and enzyme activity should be noticed. The SWI/SNF complex, which contains a subunit from the BAF family, mediated chromatin remodeling in cell differentiation, proliferation and DNA repair. Several components of the SWI/SNF complex, such as BAF47, function as tumor suppressors, and BRM and BRG1 act as putative tumor suppressors, which is evidenced by frequently loss of heterozygosity [83].

Analysis of the KEGG pathways in the total optimal feature set

Nine KEGG pathway terms in the *OS*, were hsa04115 (p53 signaling pathway), hsa00100 (steroid biosynthesis), hsa05213 (endometrial cancer), hsa05216 (thyroid cancer), hsa05218 (melanoma), hsa05219 (bladder cancer), hsa05220 (chronic myeloid leukemia), hsa05221 (acute myeloid leukemia) and hsa05223 (non-small cell lung cancer). As discussed above, p53 participates in cell death regulation and cell cycle control as a key central element. Aberrant genetic inactivation or diminished expression of p53 was found in the most of KEGG cancers terms. In addition to Rb in bladder cancer and chronic myeloid leukemia [7,84–86], abnormal PTEN was also found in thyroid cancer and endometrial cancer [7,84–86]. In melanoma, chronic myeloid leukemia and non-small cell lung cancer patients, there is reported silence or suppression of ink4a/arf leading to cell cycle disorder and sustained cellular proliferation [7,84–86]. Steroids and steroid metabolism have markedly influenced in some cancer types, such as breast cancer and prostate cancer, which may mediate the apoptosis network [87,88].

Unlike oncogenes, TSGs act as guardians regulating the network of cell cycle and apoptosis factors involved in controlling cell fate. Furthermore, maintaining genomic stability and balanced cell adhesion demand that the TSGs perform normal physiological functions.

Analysis of candidate tumor suppressors predicted based on optimal features

We try to predict the novel TSGs based on features in the total optimal feature set, *i.e.*, the key functions that defines tumor suppressor. For each 'negative gene', we counted the number of key tumor suppressor functions that it was annotated onto. The genes with great number of key tumor suppressor functions were considered as candidate tumor suppressors, since they shared similar functions with the known tumor suppressors. Since oncogene and tumor suppressor are two sides of a coin, their functions are difficult to distinguish. To better prioritize candidate tumor suppressor, we removed the 330 oncogenes from oncogene family of GSEA MSigDB (Molecular Signatures DATAbase, http://www.broadinstitute.org/gsea/msigdb/gene_families.jsp) and 251 oncogenes from HGNC (HUGO Gene Nomenclature Committee, http://www.genenames.org/) with the oncogene as the keyword. MSigDB is an online database, which collected annotated genes sets for GSEA analyze and categorize genes into gene family to provide a functional overview. HGNC is a collection of unique symbols and names for genes, ncRNA genes and pseudogenes. Subsequently, the overlap genes between these genes and the 'negative genes' were filtered out, 17,553 ensembl protein IDs remain in the end, which are available in Table S5.

Our study performs the gene enrichment and pathway enrichment analysis, providing a support to identify novel tumor suppressor in these features and pathways. In **Table 3**, we revealed a list of novel tumor suppressor genes, which shared at least 293 key annotations with known tumor suppressors. It has been demonstrated part of them play suppressive roles in tumorigenesis and more genes need verification by functional evidence and a larger clinical pathological characteristics data set. There are many tumor suppress genes proved partly, such as EP300 [89–91], GATA4 [92], ESR1 [93] and NFKBIA [94,95], which still need a large clinic data validation and functional research.

Glycogen synthase kinase 3 beta (GSK3β) belongs to the glycogen synthase kinase subfamily. GSK2β regulated Wnt signaling and PI3K/Akt pathway negatively, which play key roles in cell cycle, anti-apoptosis and invasion [96,97]. It has been identified suppression of GSK3β in many tumor types including, oral squamous cell carcinoma (OSCC), lung cancer, cutaneous squamous cell carcinoma and esophageal carcinoma [98–101]. Inhibition of constitutively active GSK3β leads to epithelial-mesenchymal transition (EMT) transition during tumorigenesis [102]. In vitro, GSK3β play a negative regulator of myeloid cell leukemia-1(Mcl-1), which has anti-apoptotic function and is correlated to the poor prognosis of breast cancer patients [98,103,104]. Although there are some controversial reports, GSK3β is a putative tumor suppressor and need more studies [105,106].

Homeobox A5 (HOXA5) is belonging to a DNA-binding transcription factor family, homeobox genes cluster A, and regulates organism gene expression, adult differentiation and embryonic development in organism. It has been observed a frequently increased methylation of the HOXA5 promoter region in various tumor tissues [107–109] and is related to decreased expression [107,110]. In addition, HOXA5 up-regulates p53 transcription through binding to a target element in its promoter [111]. These evidences document that HOXA5 is a putative tumor suppressor for tumorigenesis. But it still warrants further functional studies that how HOXA5 suppress tumorigenesis in animal model and in clinic.

Holliday Junction 5′ Flap Endonuclease, previous named gen endonuclease homolog 1 (GEN1) is an enzyme, evolved in Holliday junctions (HJs) formation during homologous recombination and DNA repair. The activity of Yen1, the ortholog of GEN1, is inhibited by phosphorylation events in the G1/S transition, keep inactive through S-phase and G2, and activated by dephosphorylation at the later stages of mitosis [112,113]. Similarly, in the early stages of the cell cycle, GEN1 is excluded from the nucleus, and access chromatin and HJs [113]. GEN1 participates in some specific features: cell cycle, DNA repair and phosphorylation/dephosphorylation, which involved in many tumor suppressors. In Bloom's syndrome cells, depletion of GEN1 results in severe chromosome abnormalities [114]. It has been identified rare recessive at-risk alleles of GEN1 in breast cancer by Ekatherina Sh [115–117], and two somatic frameshift mutations in breast cancer cell lines and primary tumors through exome sequencing [114]. Above all, GEN1 is a novel tumor suppressor akin to some other DNA repair genes, BRCA1 and BRCA2 in breast cancer, although there is rare study prove GEN1 make a high-appreciable contribution to breast cancer. In future study, it would be focus on the methylation or LOH level and anti-tumorigenesis mechanism to explore function of GEN1.

Besides these genes discussed above, our study reveals more novel candidate tumor suppressors including SHH, STAT3, SUPT20H and GSK3A, which are highlighted and need more focus and research in future cancer research.

Conclusions

This study summarizes the enrichment analysis of TSGs. The features of the GO and KEGG pathway enrichment scores were used to encode the investigated genes and some feature selection methods were employed to analyze these features. The analysis of the 708 GO terms and 9 KEGG pathways implies that they are strongly related to the determination of TSGs. We hope that effective methods based on these GO terms and KEGG pathways can be built to identify TSGs.

Supporting Information

Table S1 List of 615 tumor suppressor genes.
(PDF)

Table S2 List of the MaxRel features lists and mRMR features lists obtained by mRMR method for each dataset.
(PDF)

Table S3 List of the SNs, SPs, ACCs and MCCs obtained by IFS and Dagging for each dataset S_i.
(PDF)

Table S4 List of 717 Features in the final optimal feature set.
(PDF)

Table S5 List of the novel tumor suppressors predicted based on features in the total optimal feature set.
(PDF)

Author Contributions

Conceived and designed the experiments: TH YDC. Performed the experiments: LC YDC. Analyzed the data: JY LC XK TH. Contributed reagents/materials/analysis tools: JY TH. Contributed to the writing of the manuscript: JY LC.

References

1. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646–674.
2. Sherr CJ (2004) Principles of tumor suppression. Cell 116: 235–246.
3. Shlien A, Malkin D (2010) Copy number variations and cancer susceptibility. Curr Opin Oncol 22: 55–63.
4. Knudson AG, Jr. (1971) Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68: 820–823.
5. Lee WH, Bookstein R, Hong F, Young LJ, Shew JY, et al. (1987) Human retinoblastoma susceptibility gene: cloning, identification, and sequence. Science 235: 1394–1399.
6. Brandau S, Bohle A (2001) Bladder cancer. I. Molecular and genetic basis of carcinogenesis. Eur Urol 39: 491–497.
7. Giacinti C, Giordano A (2006) RB and cell cycle progression. Oncogene 25: 5220–5227.
8. Pietilainen T, Lipponen P, Aaltomaa S, Eskelinen M, Kosma VM, et al. (1995) Expression of retinoblastoma gene protein (Rb) in breast cancer as related to established prognostic factors and survival. Eur J Cancer 31A: 329–333.
9. Wadayama B, Toguchida J, Shimizu T, Ishizaki K, Sasaki MS, et al. (1994) Mutation spectrum of the retinoblastoma gene in osteosarcomas. Cancer Res 54: 3042–3048.

10. Burkhart DL, Sage J (2008) Cellular mechanisms of tumour suppression by the retinoblastoma gene. Nat Rev Cancer 8: 671−682.
11. Volkenandt M, Schlegel U, Nanus DM, Albino AP (1991) Mutational analysis of the human p53 gene in malignant melanoma. Pigment Cell Res 4: 35−40.
12. Nigro JM, Baker SJ, Preisinger AC, Jessup JM, Hostetter R, et al. (1989) Mutations in the p53 gene occur in diverse human tumour types. Nature 342: 705−708.
13. Muller PA, Vousden KH (2013) p53 mutations in cancer. Nat Cell Biol 15: 2−8.
14. Chuikov S, Kurash JK, Wilson JR, Xiao B, Justin N, et al. (2004) Regulation of p53 activity through lysine methylation. Nature 432: 353−360.
15. Levitt NC, Hickson ID (2002) Caretaker tumour suppressor genes that defend genome integrity. Trends Mol Med 8: 179−186.
16. Levine AJ (1997) p53, the cellular gatekeeper for growth and division. Cell 88: 323−331.
17. Nicoletto MO, Donach M, De Nicolo A, Artioli G, Banna G, et al. (2001) BRCA-1 and BRCA-2 mutations as prognostic factors in clinical practice and genetic counselling. Cancer Treat Rev 27: 295−304.
18. Scully R, Livingston DM (2000) In search of the tumour-suppressor functions of BRCA1 and BRCA2. Nature 408: 429−432.
19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25−29.
20. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322: 881−888.
21. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109−114.
22. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27: 1226−1238.
23. Lu Z, Cohen KB, Hunter L (2007) GeneRIF quality assurance as summary revision. Pac Symp Biocomput: 269−80.
24. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, et al. (2014) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 42(D1): D7–D17.
25. Zhao M, Sun J, Zhao Z (2013) TSGene: a web resource for tumor suppressor genes. Nucleic acids research 41: D970-D976.
26. He Z, Huang T, Shi X, Hu L, Chen L, et al. (2011) Computational Analysis of Protein Tyrosine Nitration; 2010. pp. 35−42.
27. Chen L, Qian ZL, Fen KY, Cai YD (2010) Prediction of Interactiveness Between Small Molecules and Enzymes by Combining Gene Ontology and Compound Similarity. Journal of Computational Chemistry 31: 1766−1776.
28. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 8: R3.
29. Huang T, Zhang J, Xu ZP, Hu LL, Chen L, et al. (2012) Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. Biochimie 94: 1017−1025.
30. Chen L, Li B-Q, Feng K-Y (2013) Predicting Biological Functions of Protein Complexes Using Graphic and Functional Features. Current Bioinformatics 8: 545−551.
31. Ting KM, Witten IH (1997) Stacking bagged and dagged models; San Francisco, CA. pp. 367−375.
32. Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann Pub.
33. Platt J, editor (1998) Fast training of support vector machines using sequential minimal optimization. Cambridge, MA: MIT Press.
34. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation 13: 637−649.
35. Kohavi R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection; San Mateo. pp. 1137−1143.
36. Li B-Q, Feng K-Y, Chen L, Huang T, Cai Y-D (2012) Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. PLoS ONE 7: e43927.
37. Ding C, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17: 349−358.
38. Martin S, Roe D, Faulon J-L (2005) Predicting protein−protein interactions using signature products. Bioinformatics 21: 218−226.
39. Chen L, Zeng WM, Cai YD, Feng KY, Chou KC (2012) Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities. PLoS ONE 7: e35254.
40. Chen L, Lu J, Zhang N, Huang T, Cai Y-D (2014) A hybrid method for prediction and repositioning of drug Anatomical Therapeutic Chemical classes. Molecular BioSystems 10: 868−877.
41. Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16: 412−424.
42. Chen L, Feng KY, Cai YD, Chou KC, Li HP (2010) Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. BMC bioinformatics 11: 293.
43. Matthews B (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405: 442−451.
44. Cramér H (1946) Mathematical Methods of Statistics: Princeton university press.
45. Kendall M, Stuart A (1979) The Advanced Theory of Statistics, vol. 2, Inference and Relationship. New York: Macmillan.
46. Harrison KM, Kajese T, Hall HI, Song R (2008) Risk factor redistribution of the national HIV/AIDS surveillance data: an alternative approach. Public health reports 123: 618−627.
47. Zhang Y, Ding C, Li T (2008) Gene selection algorithm by combining reliefF and mRMR. BMC genomics 9: S27.
48. Chen L, Zeng W-M, Cai Y-D, Huang T (2013) Prediction of Metabolic Pathway Using Graph Property, Chemical Functional Group and Chemical Structural Set. Current Bioinformatics 8: 200−207.
49. Li Z, Zhou X, Dai Z, Zou X (2010) Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm. BMC bioinformatics 11: 325.
50. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 3: 185−205.
51. Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology 281: 18−23.
52. Mohabatkar H, Mohammad Beigi M, Abdolahi K, Mohsenzadeh S (2013) Prediction of Allergenic Proteins by Means of the Concept of Chous Pseudo Amino Acid Composition and a Machine Learning Approach. Medicinal Chemistry 9: 133−137.
53. Delbridge AR, Valente LJ, Strasser (2012) A The role of the apoptotic machinery in tumor suppression. Cold Spring Harb Perspect Biol 4: a008789.
54. Dasika GK, Lin SC, Zhao S, Sung P, Tomkinson A, et al. (1999) DNA damage-induced cell cycle checkpoints and DNA strand break repair in development and tumorigenesis. Oncogene 18: 7883−7899.
55. Young LC, Hays JB, Tron VA, Andrew SE (2003) DNA mismatch repair proteins: potential guardians against genomic instability and tumorigenesis induced by ultraviolet photoproducts. J Invest Dermatol 121: 435−440.
56. Smith ML, Ford JM, Hollander MC, Bortnick RA, Amundson SA, et al. (2000) p53-mediated DNA repair responses to UV radiation: studies of mouse cells lacking p53, p21, and/or gadd45 genes. Mol Cell Biol 20: 3705−3714.
57. Dannenberg JH, van Rossum A, Schuijff L, te Riele H (2000) Ablation of the retinoblastoma gene family deregulates G(1) control causing immortalization and increased cell turnover under growth-restricting conditions. Genes Dev 14: 3051−3064.
58. Sage J, Mulligan GJ, Attardi LD, Miller A, Chen S, et al. (2000) Targeted disruption of the three Rb-related genes leads to loss of G(1) control and immortalization. Genes Dev 14: 3037−3050.
59. Derynck R, Zhang Y, Feng XH (1998) Smads: transcriptional activators of TGF-beta responses. Cell 95: 737−740.
60. Yang J, Song K, Krebs TL, Jackson MW, Danielpour D (2008) Rb/E2F4 and Smad2/3 link survivin to TGF-beta-induced apoptosis and tumor progression. Oncogene 27: 5326−5338.
61. Beausejour CM, Krtolica A, Galimi F, Narita M, Lowe SW, et al. (2003) Reversal of human cellular senescence: roles of the p53 and p16 pathways. EMBO J 22: 4212−4222.
62. Mitrea DM, Yoon MK, Ou L, Kriwacki RW (2012) Disorder-function relationships for the cell cycle regulatory proteins p21 and p27. Biol Chem 393: 259−274.
63. Vousden KH, Lane DP (2007) p53 in health and disease. Nat Rev Mol Cell Biol 8: 275−283.
64. Vazquez A, Bond EE, Levine AJ, Bond GL (2008) The genetics of the p53 pathway, apoptosis and cancer therapy. Nat Rev Drug Discov 7: 979−987.
65. Slomovitz BM, Coleman RL (2012) The PI3K/AKT/mTOR pathway as a therapeutic target in endometrial cancer. Clin Cancer Res 18: 5856−5864.
66. Annunziata CM, Davis RE, Demchenko Y, Bellamy W, Gabrea A, et al. (2007) Frequent engagement of the classical and alternative NF-kappaB pathways by diverse genetic abnormalities in multiple myeloma. Cancer Cell 12: 115−130.
67. Zhang XJ, Liang YH, He PP, Yang S, Wang HY, et al. (2004) Identification of the cylindromatosis tumor-suppressor gene responsible for multiple familial trichoepithelioma. J Invest Dermatol 122: 658−664.
68. Mandruzzato S, Brasseur F, Andry G, Boon T, van der Bruggen P (1997) A CASP-8 mutation recognized by cytolytic T lymphocytes on a human head and neck carcinoma. J Exp Med 186: 785−793.
69. Kim HS, Lee JW, Soung YH, Park WS, Kim SY, et al. (2003) Inactivating mutations of caspase-8 gene in colorectal carcinomas. Gastroenterology 125: 708−715.
70. Espada J, Peinado H, Lopez-Serra L, Setien F, Lopez-Serra P, et al. (2011) Regulation of SNAIL1 and E-cadherin function by DNMT1 in a DNA methylation-independent context. Nucleic Acids Res 39: 9194−9205.
71. van Roy F, Berx G (2008) The cell-cell adhesion molecule E-cadherin. Cell Mol Life Sci 65: 3756−3788.
72. Hood JD, Cheresh DA (2002) Role of integrins in cell invasion and migration. Nat Rev Cancer 2: 91−100.
73. Bourboulia D, Stetler-Stevenson WG (2010) Matrix metalloproteinases (MMPs) and tissue inhibitors of metalloproteinases (TIMPs): Positive and negative regulators in tumor cell adhesion. Semin Cancer Biol 20: 161−168.

74. Roy R, Yang J, Moses MA (2009) Matrix metalloproteinases as novel biomarkers and potential therapeutic targets in human cancer. J Clin Oncol 27: 5287–5297.
75. Fleming NI, Jorissen RN, Mouradov D, Christie M, Sakthianandeswaren A, et al. (2013) SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. Cancer Res 73: 725–735.
76. Lee RS, Roberts CW (2013) Linking the SWI/SNF complex to prostate cancer. Nat Genet 45: 1268–1269.
77. Milton A, Luoto K, Ingram L, Munro S, Logan N, et al. (2006) A functionally distinct member of the DP family of E2F subunits. Oncogene 25: 3212–3218.
78. Shiloh Y, Ziv Y (2013) The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. Nat Rev Mol Cell Biol 14: 197–210.
79. Piccirillo SG, Reynolds BA, Zanetti N, Lamorte G, Binda E, et al. (2006) Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells. Nature 444: 761–765.
80. Julien SG, Dube N, Hardy S, Tremblay ML (2011) Inside the human cancer tyrosine phosphatome. Nat Rev Cancer 11: 35–49.
81. Jacob ST, Motiwala T (2005) Epigenetic regulation of protein tyrosine phosphatases: potential molecular targets for cancer therapy. Cancer Gene Ther 12: 665–672.
82. Merchant AA, Matsui W (2010) Targeting Hedgehog–a cancer stem cell pathway. Clin Cancer Res 16: 3130–3140.
83. Reisman D, Glaros S, Thompson EA (2009) The SWI/SNF complex and cancer. Oncogene 28: 1653–1668.
84. Singhal S, Vachani A, Antin-Ozerkis D, Kaiser LR, Albelda SM (2005) Prognostic implications of cell cycle, apoptosis, and angiogenesis biomarkers in non-small cell lung cancer: a review. Clin Cancer Res 11: 3974–3986.
85. Knowles MA (2006) Molecular subtypes of bladder cancer: Jekyll and Hyde or chalk and cheese? Carcinogenesis 27: 361–373.
86. Renneville A, Roumier C, Biggio V, Nibourel O, Boissel N, et al. (2008) Cooperating gene mutations in acute myeloid leukemia: a review of the literature. Leukemia 22: 915–931.
87. Sharifi N, Auchus RJ (2012) Steroid biosynthesis and prostate cancer. Steroids 77: 719–726.
88. Risbridger GP, Davis ID, Birrell SN, Tilley WD (2010) Breast and prostate cancer: more similar than different. Nat Rev Cancer 10: 205–212.
89. Kim MS, Lee SH, Yoo NJ, Lee SH (2013) Frameshift mutations of tumor suppressor gene EP300 in gastric and colorectal cancers with high microsatellite instability. Hum Pathol 44: 2064–2070.
90. Gayther SA, Batley SJ, Linger L, Bannister A, Thorpe K, et al. (2000) Mutations truncating the EP300 acetylase in human cancers. Nat Genet 24: 300–303.
91. Tamura G (2006) Alterations of tumor suppressor and tumor-related genes in the development and progression of gastric cancer. World J Gastroenterol 12: 192–198.
92. Hellebrekers DM, Lentjes MH, van den Bosch SM, Melotte V, Wouters KA, et al. (2009) GATA4 and GATA5 are potential tumor suppressors and biomarkers in colorectal cancer. Clin Cancer Res 15: 3990–3997.
93. Hishida M, Nomoto S, Inokawa Y, Hayashi M, Kanda M, et al. (2013) Estrogen receptor 1 gene as a tumor suppressor gene in hepatocellular carcinoma detected by triple-combination array analysis. Int J Oncol 43: 88–94.
94. Bredel M, Scholtens DM, Yadav AK, Alvarez AA, Renfrow JJ, et al. (2011) NFKBIA deletion in glioblastomas. N Engl J Med 364: 627–637.
95. Sigglekow ND, Pangon L, Brummer T, Molloy M, Hawkins NJ, et al. (2012) Mutated in colorectal cancer protein modulates the NFkappaB pathway. Anticancer Res 32: 73–79.
96. Katoh M (2007) Networking of WNT, FGF, Notch, BMP, and Hedgehog signaling pathways during carcinogenesis. Stem Cell Rev 3: 30–38.
97. Thornton TM, Pedraza-Alva G, Deng B, Wood CD, Aronshtam A, et al. (2008) Phosphorylation by p38 MAPK as an alternative pathway for GSK3beta inactivation. Science 320: 667–670.
98. Ma C, Wang J, Gao Y, Gao TW, Chen G, et al. (2007) The role of glycogen synthase kinase 3beta in the transformation of epidermal cells. Cancer Res 67: 7756–7764.
99. Suzuki M, Shinohara F, Endo M, Sugazaki M, Echigo S, et al. (2009) Zebularine suppresses the apoptotic potential of 5-fluorouracil via cAMP/PKA/CREB pathway against human oral squamous cell carcinoma cells. Cancer Chemother Pharmacol 64: 223–232.
100. Zheng H, Saito H, Masuda S, Yang X, Takano Y (2007) Phosphorylated GSK3beta-ser9 and EGFR are good prognostic factors for lung carcinomas. Anticancer Res 27: 3561–3569.
101. Lu Z, Liu H, Xue L, Xu P, Gong T, et al. (2008) An activated Notch1 signaling pathway inhibits cell proliferation and induces apoptosis in human esophageal squamous cell carcinoma cell line EC9706. Int J Oncol 32: 643–651.
102. Yan D, Avtanski D, Saxena NK, Sharma D (2012) Leptin-induced epithelial-mesenchymal transition in breast cancer cells requires beta-catenin activation via Akt/GSK3- and MTA1/Wnt1 protein-dependent pathways. J Biol Chem 287: 8598–8612.
103. Ding Q, He X, Xia W, Hsu JM, Chen CT, et al. (2007) Myeloid cell leukemia-1 inversely correlates with glycogen synthase kinase-3beta activity and associates with poor prognosis in human breast cancer. Cancer Res 67: 4564–4571.
104. Farago M, Dominguez I, Landesman-Bollag E, Xu X, Rosner A, et al. (2005) Kinase-inactive glycogen synthase kinase 3beta promotes Wnt signaling and mammary tumorigenesis. Cancer Res 65: 5792–5801.
105. Cao Q, Lu X, Feng YJ (2006) Glycogen synthase kinase-3beta positively regulates the proliferation of human ovarian cancer cells. Cell Res 16: 671–677.
106. Yang J, Takahashi Y, Cheng E, Liu J, Terranova PF, et al. (2010) GSK-3beta promotes cell survival by modulating Bif-1-dependent autophagy and cell death. J Cell Sci 123: 861–870.
107. Strathdee G, Sim A, Soutar R, Holyoake TL, Brown R (2007) HOXA5 is targeted by cell-type-specific CpG island methylation in normal cells and during the development of acute myeloid leukaemia. Carcinogenesis 28: 299–309.
108. Shiraishi M, Sekiguchi A, Oates AJ, Terry MJ, Miyamoto Y (2002) HOX gene clusters are hotspots of de novo methylation in CpG islands of human lung adenocarcinomas. Oncogene 21: 3659–3662.
109. Maroulakou IG, Spyropoulos DD (2003) The study of HOX gene function in hematopoietic, breast and lung carcinogenesis. Anticancer Res 23: 2101–2110.
110. Houghton J, Stoicov C, Nomura S, Rogers AB, Carlson J, et al. (2004) Gastric cancer originating from bone marrow-derived cells. Science 306: 1568–1571.
111. Raman V, Martensen SA, Reisman D, Evron E, Odenwald WF, et al. (2000) Compromised HOXA5 function can limit p53 expression in human breast tumours. Nature 405: 974–978.
112. Matos J, West SC (2014) Holliday junction resolution: Regulation in space and time. DNA Repair (Amst) 19: 176–181.
113. Matos J, Blanco MG, Maslen S, Skehel JM, West SC (2011) Regulatory control of the resolution of DNA recombination intermediates during meiosis and mitosis. Cell 147: 158–172.
114. Wechsler T, Newman S, West SC (2011) Aberrant chromosome morphology in human cells defective for Holliday junction resolution. Nature 471: 642–646.
115. Kuligina E, Sokolenko AP, Mitiushkina NV, Abysheva SN, Preobrazhenskaya EV, et al. (2013) Value of bilateral breast cancer for identification of rare recessive at-risk alleles: evidence for the role of homozygous GEN1 c.2515_2519delAAGTT mutation. Fam Cancer 12: 129–132.
116. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, et al. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet Chapter 10: Unit 10 11.
117. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. Science 318: 1108–1113.