PLOS ONE

RESEARCH ARTICLE

Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways

Lei Chen[1,2☯], Yu-Hang Zhang[3☯], ShaoPeng Wang[1], YunHua Zhang[4], Tao Huang[3]*, Yu-Dong Cai[1]*

1 School of Life Sciences, Shanghai University, Shanghai, People's Republic of China, **2** College of Information Engineering, Shanghai Maritime University, Shanghai, People's Republic of China, **3** Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China, **4** Anhui province key lab of farmland ecological conversation and pollution prevention, School of Resources and Environment, Anhui Agricultural University, Hefei, People's Republic of China

☯ These authors contributed equally to this work.
* tohuangtao@126.com (TH); cai_yud@126.com (YDC)

OPEN ACCESS

Citation: Chen L, Zhang Y-H, Wang S, Zhang Y, Huang T, Cai Y-D (2017) Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways. PLoS ONE 12(9): e0184129. https://doi.org/10.1371/journal.pone.0184129

Editor: Bin Liu, Harbin Institute of Technology Shenzhen Graduate School, CHINA

Received: June 16, 2017

Accepted: August 18, 2017

Published: September 5, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by the National Natural Science Foundation of China (31371335), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program, The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the Science Foundation of Anhui (1608085MC58) and the Science and Technology Research Projects

Abstract

Identifying essential genes in a given organism is important for research on their fundamental roles in organism survival. Furthermore, if possible, uncovering the links between core functions or pathways with these essential genes will further help us obtain deep insight into the key roles of these genes. In this study, we investigated the essential and non-essential genes reported in a previous study and extracted gene ontology (GO) terms and biological pathways that are important for the determination of essential genes. Through the enrichment theory of GO and KEGG pathways, we encoded each essential/non-essential gene into a vector in which each component represented the relationship between the gene and one GO term or KEGG pathway. To analyze these relationships, the maximum relevance minimum redundancy (mRMR) was adopted. Then, the incremental feature selection (IFS) and support vector machine (SVM) were employed to extract important GO terms and KEGG pathways. A prediction model was built simultaneously using the extracted GO terms and KEGG pathways, which yielded nearly perfect performance, with a Matthews correlation coefficient of 0.951, for distinguishing essential and non-essential genes. To fully investigate the key factors influencing the fundamental roles of essential genes, the 21 most important GO terms and three KEGG pathways were analyzed in detail. In addition, several genes was provided in this study, which were predicted to be essential genes by our prediction model. We suggest that this study provides more functional and pathway information on the essential genes and provides a new way to investigate related problems.

of Anhui (1604e0302006). 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.

1. Introduction

As is known, genes are the basic molecular unit of heredity. However, the functions of genes have been widely reported to be redundant and reduplicative [1, 2]. Some genes have turned out to be significant for survival, while others seem to be not necessary. To distinguish these two groups of genes and identify the core heretical regulatory factors, a new concept, named essential genes, has been presented. Essential genes refer to a group of fundamental genes necessary for a specific organism to survive in a specific environment [3]. Based on two reliable and widely quoted literatures, essential genes refer to sets of genes that are absolutely required for indispensable for the viability of individual human cell types [4, 5]. Generally, the essential genes encode conservative functional elements, which mainly contribute to DNA replication, gene translation, gene transcription and substance transportation [6]. Compared to other genes, essential genes convey fewer selective advantages and may have decreased fitness, escaping from the natural selection.

Considering the fundamental role of essential genes that participate in organism survival, for centuries, people have concentrated on the identification of essential genes in various organisms. Bacteria are a perfect experimental model for the identification of essential genes [6, 7]. For bacteria, there are mainly two methods that may contribute to the identification of essential genes in the genome-scale: gene direct deletion and transposon-based randomized mutagenesis [8, 9]. Relying on two such functional experimental methods, the essential and non-essential genes in various microorganism species have been identified and distinguished from each other. The well-known pathogen *Helicobacter pylori* is an example. *H. pylori* is a gram-negative pathogen that is usually found in the stomach [10]. Based on whole population mutagenesis and microarray analysis, there have been 1,576 candidate ORFs identified in the genome of *H. pylori*. After screening, there were 1,178 ORFs that were non-essential genes, which left approximately 25% essential genes. However, such screening processes have mostly been applied in prokaryotic organisms and yeast. It is difficult to do similar screenings in multicellular organisms, especially in mammals.

With the development of screening technologies, the essential genes in mice have been generally revealed [11]. Based on microinject KO and nuclear transfer techniques, such research has identified that at least 2,114 genes are essential in mice [12]. However, due to ethical arguments, such an experimental method cannot be repeated in humans. Traditionally, the identification of human essential genes mainly depends on literature searching, which is not comprehensive and is generally biased [13]. In 2016, following the rapid development of genome-editing techniques, especially the CRISPR/Cas9 genome editing system, a further trial on essential gene screening in human cells was reported [14]. In the study, the research group compared the editing efficacy of traditional RNAi techniques and the newly applied CRIPSR/Cas9 system, and further identified a group of experimental validated essential genes and non-essential genes in the human chronic myelogenous leukemia cell line K562, providing a new experimental tool for essential gene identification [14]. Although, in such a study, large series of essential genes have been revealed, it is not the ultimate goal for us to identify essential genes. Actually, the fundamental biological processes the directly affect the physical and pathological processes in human bodies was our focus. However, until now, how essential genes systematically contribute to fundamental biological processes has not been fully revealed and analyzed in detail. Among the computational methods for predicting essential genes, the first one was provided in 1996 by Mushegian and Koonin to identify a minimal set of essential genes from 468 protein-encoding genes in *Mycoplasma genitalium*. Finally, 256 genes were selected and regarded as the possible minimal gene set to maintain the existence of modern-type cells [15]. In prediction methods using machine learning algorithms, some that are

commonly used, including support vector machine (SVM), neural networks, Naïve Bayes, and decision trees, were used to predict essential genes based on biological and topological features. Detailed information on this can be found in the excellent review [16]. Specifically, Zhong *et al*. provided a gene expression programming-based (GEP) method to predict essential genes in *S. cerevisiae* using a combination of features. The constructed classifier was comprehensively evaluated by eight measurements. The results showed that the GEP classifier outperformed other methods using individual features and received a better AUC score than most of the classifiers trained by various machine learning algorithms [17].

In the field of bioinformatics, several computational methods have been proposed to tackle various problems [18–20]. Among them, the biological sequence analysis methods were always used to investigate the DNA-related and protein-related problems [21–24]. On the other hand, the effective clustering of functional genes, mainly based on gene ontology (GO) and KEGG pathways [25, 26], which cluster functional genes into different biological processes, is also useful to tackle these problems. In this study, we investigated the essential and non-essential genes using GO terms and KEGG pathways. Unlike the computational methods mentioned above, we aimed to identify important GO terms and KEGG pathways that can be important indicators for distinguishing essential and non-essential genes, and at the same time a classification model was built using these GO terms and KEGG pathways. The essential/non-essential genes of the human chronic myelogenous leukemia cell line K562 were retrieved from Morgens *et al*.'s study [14]. The relationship between one gene and one GO term or KEGG pathway was encoded into a numeric value using the enrichment theory of GO and KEGG. Then, some popular computational methods, maximum relevance minimum redundancy (mRMR) [27], incremental feature selection (IFS), and a support vector machine (SVM) [28, 29], were employed to analyze involved GO terms and KEGG pathways. As a result, some key GO terms and KEGG pathways were extracted. Among them, the most important ones were extensively analyzed. In addition, a prediction model was proposed to distinguish essential and non-essential genes, which provided nearly perfect performance, with a Matthews correlation coefficient of 0.951. In consideration of the perfect performance of the prediction model, it was applied to identify essential genes from unlabeled ones, yielding several possible essential genes. Considering the crucial role of essential genes investigated in this study, the biological processes described by our identified GO terms and KEGG pathways may be fundamental for cell survival. Therefore, for the first time, we screened out and analyzed the essential genes at functional levels.

2. Materials and methods

2.1 Dataset

We used the gold standard 217 essential genes and 927 non-essential genes that were compiled by Morgens *et al*. [14]. These essential genes were critical for cell growth [30] and are provided in S1 Table. For convenience, essential genes were deemed positive samples, while non-essential genes were regarded as negative samples. The purpose of this study was to explore the functional difference between essential genes and non-essential genes. In detail, we aimed to extract important GO terms and biological pathways that can perfectly discriminate essential and non-essential genes. In view of this, twelve essential genes and 97 non-essential genes were discarded because their enrichment scores on GO terms and biological pathways were not available; i.e., 205 essential genes and 830 non-essential genes were investigated in this study.

2.2 Representation of essential and non-essential genes

To extract important GO terms and pathways that are tightly related to essential genes, all genes, including essential and non-essential genes, should be encoded by GO terms and

pathways. Then they can be analyzed by various computational methods. In this study, we used the enrichment theory [31] of GO terms and KEGG pathways, which can encode the relationship between one gene and one GO term (KEGG pathway) into a numeric value. Its brief description is as follows.

2.2.1 GO enrichment score. For a given gene g and one GO term GO_j, let G_{GO} denote a gene set consisting of annotated genes of GO_j and $G(g)$ denotes another gene set consisting of g and its direct neighboring genes in the protein-protein interaction network reported in STRING. The GO enrichment score of g and GO_j is defined as the hypergeometric test P value [32–37] on $G(g)$ and G_{GO}, which can be computed by:

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

where N and M denote the total number of human genes and the number of genes in G_{GO}; n and m represent the number of genes in $G(g)$ and the number of genes in both $G(g)$ and G_{GO}. The higher the score is, the stronger the functional association of the gene g and GO term GO_j. A total of 17,916 GO terms were used in this study, inducing 17,916 GO enrichment scores.

2.2.2 KEGG enrichment score. A similar method can be applied to obtain the KEGG enrichment score of one gene and one pathway, which can measure the relationship between them. Let G_{KEGG} denote a set consisting of annotating genes of one KEGG pathway K_j. The KEGG enrichment score of g and K_j is defined to be the hypergeometric test P value [32–37] on $G(g)$ and G_{KEGG}, which can be calculated by:

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

where N and n are same as those in Eq 1, M denotes the number of genes in G_{KEGG}, and m denotes the number of genes in both $G(g)$ and G_{KEGG}. Similarly, a large KEGG enrichment score means the strong functional association between the gene g and the KEGG pathway K_j. A total of 279 KEGG pathways were used in this study, resulting in 279 KEGG enrichment scores.

Each essential or non-essential gene can be represented by 17,916 GO enrichment scores and 279 KEGG enrichment scores. In other words, each gene g can be encoded into an 18,195-D vector, formulated as:

$$v(g) = [S_{GO}(g, GO_1), \cdots, S_{GO}(g, GO_{17916}), S_{KEGG}(g, K_1), \cdots, S_{KEGG}(g, K_{279})]^{\mathrm{T}} \quad (3)$$

2.3 Feature evaluation using the mRMR method

As mentioned in Section 2.2, several GO terms and KEGG pathways were employed to represent essential genes and non-essential genes. However, not all of them can provide positive contributions for discriminating them. To analyze them, a reliable and widely used feature selection method; i.e., the mRMR method [27], was employed in this study. This method has been widely applied to analyze several complicated biological problems [35, 37–44]. Two

excellent criteria were used in this method: maximum relevance and minimum redundancy. Each feature was evaluated from two aspects: (1) the relevance to the target and (2) the redundancy to other features. Based on them, a feature list, namely, the mRMR feature list, can be obtained in which features are ranked in a rigorous way. The description of how to obtain this list is described below.

The mRMR method is a mutual information (MI)-based feature selection method, and all evaluations are based on the MI of two variables x and y, which can be computed according to the following equation:

$$I(x,y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{4}$$

where $p(x, y)$ is joint probabilistic density of x and y, $p(x)$ and $p(y)$ are their marginal probabilistic density. A large MI value between x and y indicates the strong association between them. For a dataset in which each sample is represented by N features, let Ω be a set consisting of N features, Ω_s be a set consisting of the already selected features (initially, it is set to an empty set), Ω_t be a set containing the rest features. In the next round, a feature in Ω_t is selected and moved to Ω_s. For each feature f in Ω_t, we computed its relevance to target c by

$$D = I(f,c) \tag{5}$$

In addition, its redundancy to features in Ω_s is also considered, which can be evaluated by:

$$R = \frac{1}{|\Omega_s|}\sum_{f_i\in\Omega_s} I(f,f_i) \tag{6}$$

In particular, if Ω_s is an empty set, R will be set to zero. To integrate the relevance to the target and the redundancy to already selected features, we further compute D-R for each feature in Ω_t. The feature with the maximum D-R value can be found and it is removed from Ω_t to Ω_s. When all features are in Ω_s, the whole procedures stop. All features are ranked according to their selection orders, producing the mRMR feature list, formulated as

$$F = [f_1, f_2, \cdots, f_N] \tag{7}$$

In addition to the mRMR feature list, the mRMR method also yields another feature list, MaxRel feature list, in which features are ranked by the decreasing order of their MI values to targets. It is clear that features receiving high ranks in this list can give key contributions for classification.

2.4 IFS method

The mRMR method ranks all features in the mRMR feature list only. However, which features should be selected to participate in classification is still a problem. Here, the IFS method was employed. Based on the mRMR feature list, F, we can construct a series of feature sets. Each set contained some top features in the list F. For each constructed set, say F', all essential and non-essential genes were represented by features in F'. Then, a classification algorithm was executed on these genes, with its performance evaluated by ten-fold cross-validation [45]. By testing all possible feature sets or some of them, the feature set yielding the best performance can be obtained. This feature set was called the optimal feature set and features in this set were named optimal features. In addition, a prediction model can be constructed using the optimal features to represent genes and the classification algorithm as the prediction engine.

2.5 SVM algorithm

According to the description of the IFS method in Section 2.4, one classification algorithm is necessary. Here, we selected one of the classic machine learning algorithms, the SVM algorithm [28, 29]. The model based on this algorithm can be constructed on a dataset with small size, whereas it can provide good generalization performances. Thus, this algorithm has been widely used in bioinformatics [46–49]. In the algorithm, samples in the dataset are mapped into a higher-dimensional space using the kernel trick, in which the set of positive and negative samples can easily be separated by a hyper-plane with maximum margin. For a query sample, it is also mapped to the same higher dimension and its predicted class depends on which side of the hyper-plane it falls on.

In this study, a type of SVM algorithm was adopted, which is optimized by the sequential minimum optimization (SMO) algorithm [50] proposed by Platt. Different from the traditional optimization methods, the SMO method partitions the original quadratic programming (QP) problem into several smallest sub-QP problems and solves them in an analytical way. To quickly implement this type of SVM, we employed the tool, namely, SMO, in Weka [51], a suite of software collecting several popular machine learning algorithms. For convenience, it was executed using its default parameters.

2.6 Accuracy measurements

Because one gene is either an essential or a non-essential gene, it is a binary classification problem. For the predicted results of a binary classification problem, they can be counted as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In detail, TP/TN represents the number of positive/negative samples that are predicted correctly, and FP/FN represents the number of negative/positive samples that are predicted incorrectly. In addition, these values can induce four measurements: sensitivity (SN), specificity (SP), accuracy (ACC) [52], and Matthews correlation coefficient (MCC) [53], which can be calculated by:

$$SN = \frac{TP}{TP + FN} \tag{8}$$

$$SP = \frac{TN}{TN + FP} \tag{9}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

It is easy to see that SN and SP represent the prediction accuracy of positive and negative samples, respectively. Thus, they are not proper to evaluate the predicted results on the whole. Because the non-essential genes are more than four times as many as essential genes, ACC is also inappropriate. MCC is more proper because it is a balanced measurement even if the sizes of positive and negative samples have a great difference. Furthermore, it has been used as the major measurement in several studies [35, 37, 54–57]. Thus, it was selected as the major method to evaluate the predicted results yielded by different prediction models.

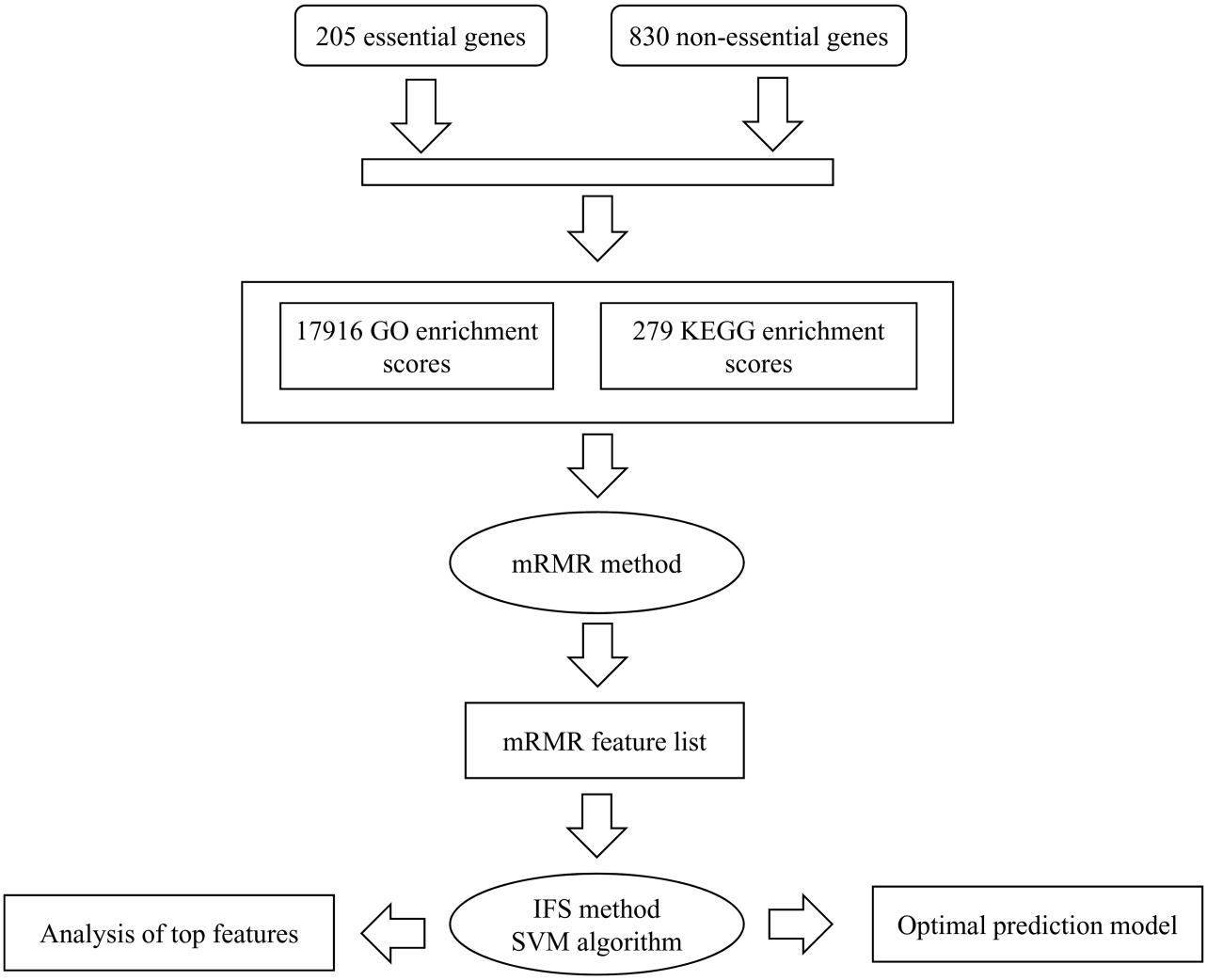

Fig 1. Flow chart of the whole procedure for investigating essential and non-essential genes.

https://doi.org/10.1371/journal.pone.0184129.g001

3. Results and discussion

In this study, we used some computer techniques, such as the mRMR method, IFS method and SVM algorithm, to extract important GO terms or KEGG pathways that can discriminate essential genes and non-essential gene to the extent possible. The whole procedures are illustrated in Fig 1. This section gives detailed results of each procedure and the analysis of the identified GO terms or KEGG pathways.

3.1 Results of the mRMR method

As mentioned in Section 2.2, each essential gene and non-essential gene was represented by several features that were extracted from the GO terms and KEGG pathways. To analyze them, the mRMR method was employed, resulting in an mRMR feature list *F* and a MaxRel feature list, in which all features were ranked in a rigorous way. The obtained mRMR feature list and MaxRel feature list are provided in S2 and S3 Tables, respectively.

3.2 Results of the IFS method

To extract the optimal combination of features that can provide key contributions for discriminating essential and non-essential genes, the IFS method was adopted as mentioned in Section 2.4. However, if all possible feature sets were tested, it would take a lot of time because there were 18,195 features totally. On the other hand, it is impossible that several GO terms and KEGG pathways can indicate the differences between essential and non-essential genes. Thus, we tried feature sets containing features from 5 to 500; i.e., feature sets $F_i = \{f_1, f_2, \cdots, f_i\}$ ($5 \leq i \leq 500$) were tested. For each feature set, we adopted the SVM algorithm, evaluated by ten-fold cross-validation, to examine it, inducing the measurements SN, SP, ACC, and MCC. After testing these 496 feature sets, a series of SNs, SPs, ACCs and MCCs were obtained, which are available in S4 Table. Because MCC was selected as the key measurement, we found a feature set yielding the maximum MCC. To give a clear observation, we plotted a curve, namely, an IFS-curve, which used the MCC as its Y-axis and the number of features participating in classification as its X-axis, as shown in Fig 2. It can be observed that this curve generally follows an increasing trend in the beginning. It is reasonable because increasingly important GO terms and KEGG pathways participated in the classification procedure. The maximum MCC was 0.951 when the first 345 features in the mRMR feature list were used. Thus, the feature set containing these features was called the optimal feature set and these 345 features were termed as optimal features. Furthermore, an optimal prediction model can be constructed using the optimal features to represent genes and SVM as the prediction engine. The detailed performance of this model is shown in Table 1, from which we can see that the SN, SP and ACC were 0.927, 0.999 and 0.985, respectively, indicating that the model is almost a perfect prediction model.

As mentioned above, 345 top features in the mRMR feature list were used to construct the optimal prediction model. Among these 345 features, 342 were derived from 342 GO terms and three were from three KEGG pathways. Three KEGG pathways were hsa03015 (mRNA

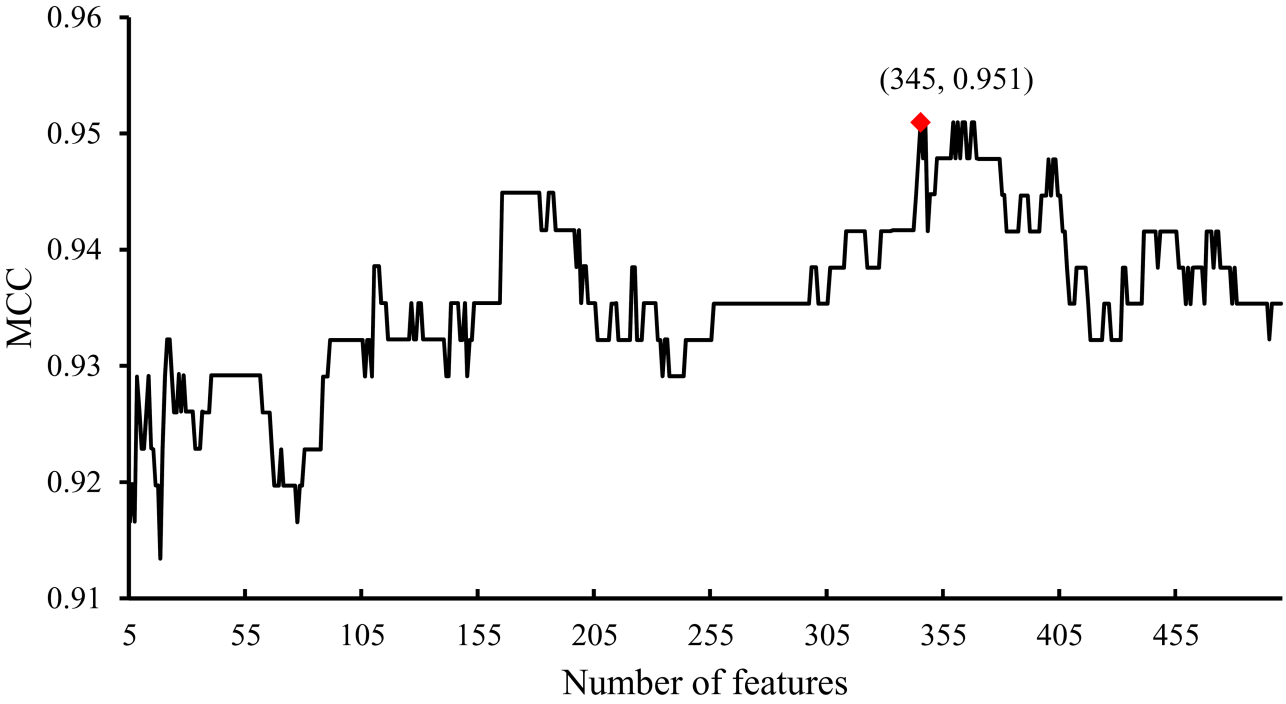

Fig 2. The IFS curve using the MCC as its Y-axis and the number of features participating in classification as its X-axis.

https://doi.org/10.1371/journal.pone.0184129.g002

Table 1. The *SN*, *SP*, *ACC*, and *MCC* yielded by the optimal SVM prediction model and the model using features of KEGG enrichment scores.

Model	Number of features	SN	SP	ACC	MCC
Optimal SVM prediction model	345	0.927	0.999	0.985	0.951
Model using features of KEGG enrichment scores	279	0.873	0.989	0.966	0.891

https://doi.org/10.1371/journal.pone.0184129.t001

surveillance pathway), hsa03013 (RNA transport) and hsa03020 (RNA polymerase). For GO terms, it is known that all GO terms can be clustered into three groups: (1) Biological process (BP); (2) Cellular component (CC) and (3) Molecular function (MF). The distribution of 342 GO terms on these three groups is illustrated in Fig 3, from which we can see that BP GO terms were most, followed by CC GO terms and MF GO terms.

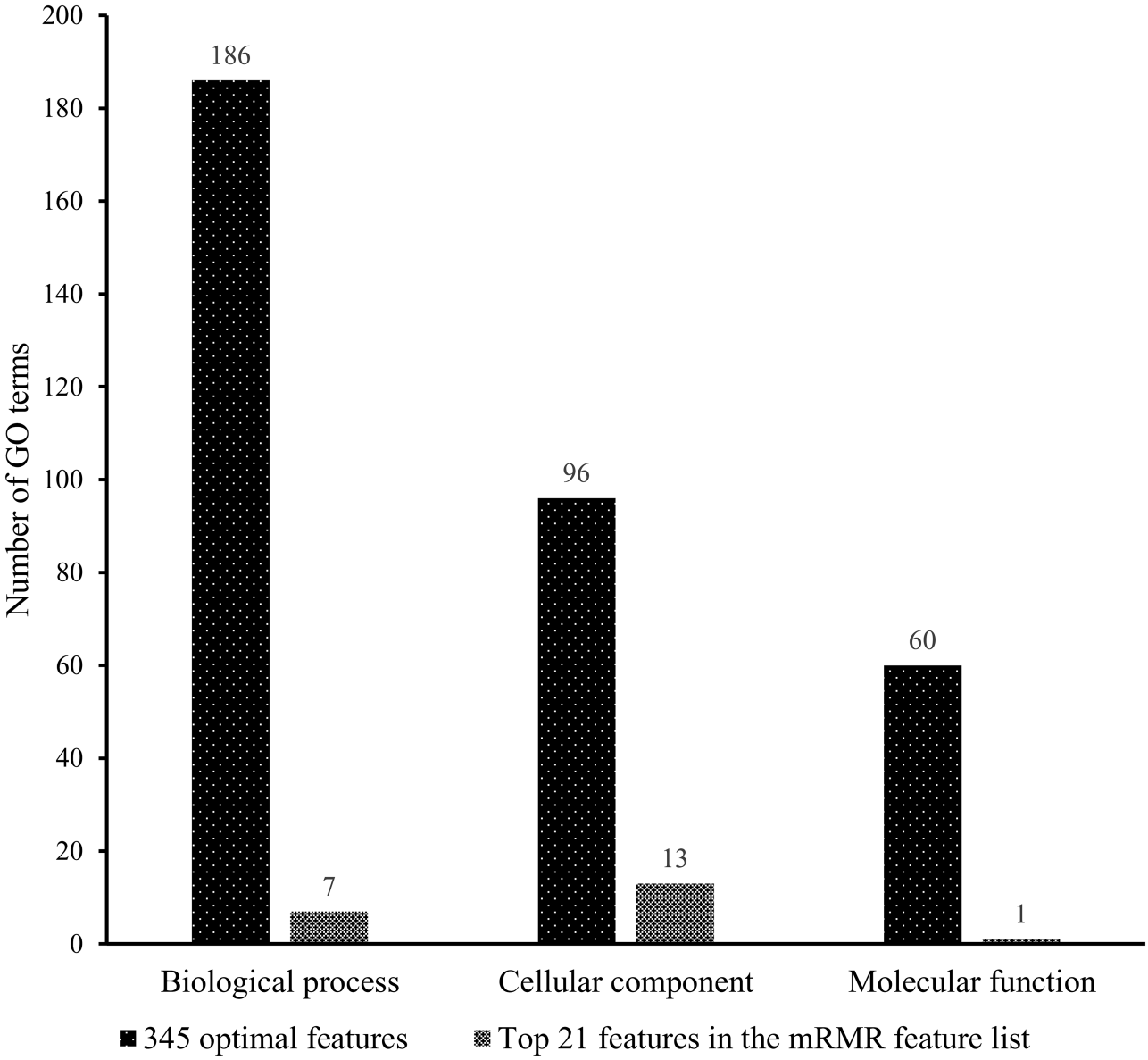

Fig 3. Distribution of the corresponding GO terms of the optimal features and top 21 features in the mRMR features in the three groups.

https://doi.org/10.1371/journal.pone.0184129.g003

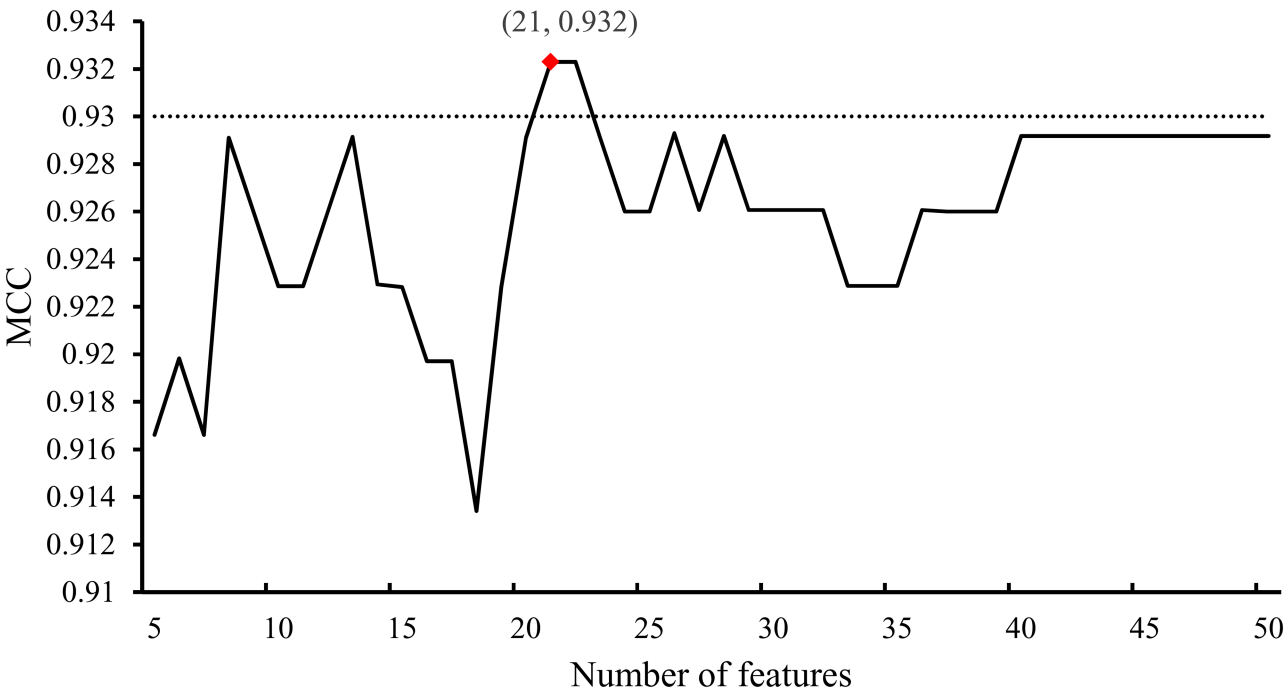

Fig 4. A part of the IFS curve shown in Fig 2.

https://doi.org/10.1371/journal.pone.0184129.g004

3.3 Analysis of important GO terms

A total of 345 optimal features were extracted using the IFS method and SVM algorithm. However, it is difficult to analyze all of these 345 features one by one. It is easy to know that features with higher ranks in the mRMR features are more important. Thus, we tried to analyze only some top features in the mRMR features. To determine these features, we amplified the IFS curve between X-axis 5 and 50, which is shown in Fig 4. It can be seen that the MCC value first exceeds 0.930 when the first 21 features in the mRMR feature list were used. Here, we analyzed the corresponding GO terms or KEGG pathways of these features, which are listed in S5 Table. All 21 features were derived from GO terms, which are listed in Table 2. Like the 342 GO terms mentioned in Section 3.2, the distribution of 21 GO terms on three groups is also illustrated in Fig 3. In detail, seven were BP GO terms, thirteen were CC GO terms and one was an MF GO term.

Among the 342 GO terms which derived the features in the optimal feature set, 186 GO terms turn out to describe biological processes, 60 turn out to describe molecular functions and 96 turn out to describe cellular component (see Fig 3). Therefore, from the overall distribution regardless of the contribution value and degree of correlation, the BP indeed are more informative. However, from the distribution of 21 GO terms shown in Fig 3, CC associated terms are more significant. In fact, CC associated terms are indeed more significant than BP and MF associated terms (with higher ranks in the mRMR feature list). Some literature can support that such fact is true. First, a literature presented by Li *et al*. confirmed that the subcellular location of essential genes is quite significant [58]. Second, considering that most of the essential genes contribute to the maintenance of cell proliferation and survival, it is quite reasonable to predict various subcellular locations as significant factors because most of such proliferation and survival associated biological processes happen in the same place (subcellular

Table 2. Twenty-one GO terms corresponding to the top 21 features in the mRMR feature list.

Rank in mRMR feature list	GO term ID	GO term	Cluster
1	GO:0032991	macromolecular complex	Cellular component
2	GO:0021888	hypothalamus gonadotrophin-releasing hormone neuron development	Biological process
3	GO:0071008	U2-type post-mRNA release spliceosomal complex	Cellular component
4	GO:0044424	intracellular part	Cellular component
5	GO:0000154	rRNA modification	Biological process
6	GO:0043226	organelle	Cellular component
7	GO:0016071	mRNA metabolic process	Biological process
8	GO:0071146	SMAD3-SMAD4 protein complex	Cellular component
9	GO:0072669	tRNA-splicing ligase complex	Cellular component
10	GO:0044422	organelle part	Cellular component
11	GO:0021886	hypothalamus gonadotrophin-releasing hormone neuron differentiation	Biological process
12	GO:0002183	cytoplasmic translational initiation	Biological process
13	GO:0005622	intracellular	Cellular component
14	GO:0015030	Cajal body	Cellular component
15	GO:0030874	nucleolar chromatin	Cellular component
16	GO:0044446	intracellular organelle part	Cellular component
17	GO:0010467	gene expression	Biological process
18	GO:0043227	membrane-bounded organelle	Cellular component
19	GO:1902369	negative regulation of RNA catabolic process	Biological process
20	GO:0044822	poly(A) RNA binding	Molecular function
21	GO:0005737	cytoplasm	Cellular component

https://doi.org/10.1371/journal.pone.0184129.t002

locations) and involve specific functional proteins, which are all described by CC, like post-mRNA release spliceosomal complex and SMAD3-SMAD4 protein complex. Therefore, it is reasonable that more CC associated terms were listed in the top of mRMR feature list because one cellular component may involve in multiple biological processes and molecular function that connected to essentialness but in turn are not. In the following text, all 21 GO terms were analyzed.

3.3.1 Analysis of important CC GO terms. Among the 21 important GO terms, there were thirteen CC GO terms. Such CC GO terms have all been confirmed to be related to the survival and proliferation of leukemia cell line K562. **GO: 0044422** (organelle part) describes any constituent part of an organelle including the nucleus, mitochondria, plastids, vacuoles, vesicles, ribosomes and the cytoskeleton, except for the plasma membrane. Considering the constituent part of such cell organelles may definitely involve the survival of the whole cell, genes that can be enriched in this GO term may definitely be fundamental genes, which can be clustered as essential genes [59]. Recent studies on yeast, cancer cells and Neurospora crassa also confirmed the conclusion, circumstantially proving the accuracy and efficacy of GO: 0044422 being a functional benchmark to distinguish essential and non-essential genes [59–61]. Similar with GO: 0044422 describing the organelle part, **GO: 0043226** (organelle) describing the organized structure of organelles may also distinguish essential from non-essential genes in the same way. As the related GO terms of such two CC GO terms we have analyzed above, three specific GO terms, **GO: 0044446** (intracellular organelle part), **GO: 0005622** (intracellular) **and GO: 0005737** (cytoplasm), may also be optimal classifying standards for the identification of essential and non-essential genes. GO: 0044446, which describes the intracellular organelle part of cells, similar to GO: 0044422, has also been confirmed to contribute to the cell survival processes [62]. GO: 0043229 describes all the intracellular

organelles, which is also the parental GO term for GO: 0044422 we have analyzed in detail above. For GO: 0005622 and GO: 0005737, these two GO terms describe the general intracellular and cytoplasm cell components, respectively. As we have analyzed above, crucial genes that play irreplaceable roles during the formation and normal metabolism processes of cell organelles may definitely be essential genes [63, 64]. In addition, a specific GO term named **GO: 044424** (intracellular part) has also been identified as a specific GO term that involves essential regulatory processes in our model cell line. Describing any constituent part of the living contents of a cell, including GO: 0044422 and GO: 0044446, as like what we have analyzed above. Considering that these two GO terms have both been confirmed to be essential gene associated GO terms, it is reasonable to regard GO: 044424 as another specific essential gene associated GO term.

Apart from such general CC GO terms, we also obtained some detailed CC GO terms that may be essential for the survival of our target cells, the K562 cells. The crucial GO term **GO: 0043227** (membrane-bounded organelle) may tell the differences between essential genes and non-essential genes. The nucleus, mitochondria, plastids, vacuoles, and vesicles are all fundamental cellular organelles for cell survival, as we have analyzed above [59]. Taking a specific membrane-bounded organelle; i.e., mitochondria, as an example, mitochondria produce the energy currency of the cell and have been confirmed to be essential for the survival of both normal cells and pathological cells, including tumor cells [65, 66]. Therefore, GO: 0043227 may definitely distinguish non-essential from the essential genes. **GO: 0071018** (U2-type post-mRNA release spliceosomal complex) has been screened to be a biological process that could be used to distinguish essential from non-essential genes. As is known, in leukemia cells, the specific mutations of the spliceosome have been confirmed to contribute to the initiation and progression of such malignant diseases [67, 68]. Considering the regulatory role of the spliceosome complex in leukemia cells, the spliceosome has been widely reported to be a candidate therapeutic target of such diseases, implying the essential role of this GO term for leukemia cell survival [67, 69]. Apart from GO: 0071008, another crucial functional complex, **GO: 0032991** (macromolecular complex), was also identified as a crucial judgment factor. The macromolecular complex has been widely reported to contribute to regulation of cell migration and drug resistance in leukemia cells, including K562 cell lines [70, 71]. Therefore, GO term GO:0032991 may definitely contribute to the identification of essential genes. As we have analyzed above, the transcription and translation regulatory role may be essential for the target K562 cells [67, 68]. A specific GO term **GO: 0072669** (tRNA-splicing ligase complex) may also be a crucial reference standard for essential or non-essential genes. As is known, the tRNA-splicing ligase complex contributes to the splicing and formation of tRNA molecules [72]. In leukemia cells, the regulation of functional tRNA molecules may involve various fundamental biological processes, including cell death and morphologic changes, implying that such GO terms may also distinguish essential from non-essential genes [73, 74].

In addition, **GO: 0015030** (Cajal body) was also identified. The Cajal body (CB), which can also be named as coiled bodies, are specific spherical sub-organelles that mainly contribute to the regulation of the telomerase assembly and cell cycle [75, 76]. As we have analyzed above, the cell cycle regulators may definitely be essential for the leukemia cell line K562. Also, a specific GO term **GO: 0030874** (nucleolar chromatin) may also contribute to the distinction of essential and non-essential genes. Nucleolar chromatin turns out to reflect the portion of nuclear chromatin associated with the nucleolus [77, 78]. For leukemia cells, such CC GO terms have been confirmed to regulate the morphological alteration and apoptosis of leukemia cells, validating the essential role of genes contributing to such biological process for leukemia cells [79]. **GO: 0071146** (SMAD3-SMAD4 protein complex) turns out to be a significant indicator for essential genes. According to recent publications, this CC GO term has been validated

to contribute to the transforming growth factor-beta signaling pathway, which further participates in the proliferation, survival and transformation of normal and malignant somatic cells, including leukemia cells [80, 81].

3.3.2 Analysis of important BP GO terms. Apart from the CC GO terms, which we analyzed above, we also obtained a group of BP GO terms. Four of the GO terms that can be clustered into RNA associated regulated processes have also been identified as an inspection standard to distinguish essential and non-essential genes. **GO: 0016071** (mRNA metabolic process) is involved in all the chemical reactions and pathways that may involve RNAs. We identified this GO term to distinguish essential genes from non-essential genes. As is known, the expression of certain crucial genes is essential for the initiation and progression of various tumor subtypes including leukemia [82–84]. Since the abnormal expression of certain genes, as we have mentioned above, may definitely initiate the malignant process in leukemia and all the expression biological processes have to rely on the metabolism of RNAs, especially mRNAs, it is reasonable to regard genes contributing to mRNA metabolism as essential genes for the K562 cell line [85–87]. As the reverse processes of such biological processes, **GO: 1902369** (negative regulation of RNA catabolic process) has also been screened out in this study, implying its separating capacity to distinguish essential from non-essential genes. In leukemia cells, recent publications also confirmed that aberrant RNA splicing and editing processes involving these GO terms affect the drug sensitivity of leukemia cells, implying its specific role for cell survival [88]. Apart from that, the catabolic process of RNA has been further validated to regulate starvation- and rapamycin-induced autophagy in K562 cell lines [89]. Similarly, another GO term, **GO: 0000154** (rRNA modification) may also tell the differences between essential and non-essential genes. As is known, the rRNAs contribute to the formation and normal function of ribosomes [90–92]. As the downstream of the transcription processes analyzed above, rRNA associated translation processes are also essential and fundamental for the normal expression of every single protein, not to speak of further biological functions such as cell survival [92, 93]. Further, in leukemia cells, rRNA associated modification has been confirmed to contribute to the elicitation of growth arrest, validating the essential role of this GO term for cell survival [94, 95]. Further, we identified a more detailed biological process that may distinguish essential from non-essential genes, **GO: 0002183** (cytoplasmic translational initiation). This GO term includes the first two amino acids of a protein and the formation of a complex of the ribosome, mRNA, and an initiation complex [96–98]. It has been confirmed that the first two amino acids of a protein are significant for the normal translation of such a protein [99]. Consider a specific gene, SNAT2, as an example. As the neutral amino acid transporter, the depletion of such a gene may be fatal for cells, including leukemia cells, validating the essential role of genes clustered in such GO terms [99, 100]. Therefore, our screened GO term may definitely be essential for the translation of all the candidate mRNAs, implying the essential role of genes that can be clustered in such GO terms for further cell survival.

Further, we also obtained a parental GO term of the GO terms that we analyzed above; i.e., **GO: 0010467** (gene expression). According to the analysis above, the expression processes of certain genes may definitely be essential for the survival of human cells, including the K562 leukemia cell line [87, 94, 101]. Therefore, all BP GO terms have been confirmed to distinguish essential genes from non-essential genes based on recent publications. We also obtained two specific hypothalamus gonadotrophin-releasing hormone neuron associated biological processes; i.e., **GO: 0021886** (hypothalamus gonadotrophin-releasing hormone neuron differentiation) and **GO: 0021888** (hypothalamus gonadotrophin-releasing hormone neuron development). As is known, as a functional component of the neuroendocrine regulation of reproductive function, the hypothalamus gonadotrophin-releasing hormone has been

confirmed to contribute to the survival and proliferation of leukemia cells by interacting with leukemia inhibitory factor and interleukin-6 [102, 103].

3.3.3 Analysis of important MF GO terms. As for molecular functions, only one GO term, named **GO: 0044822** (poly(A) RNA binding), contributes to the poly (A) RNA binding process. Generally, such poly (A) binding processes refer to the regulation of transcription and translation of certain genes. In leukemia, it is significant for the regulation of certain genes at the mRNA level [104–106]. Taking a specific RNA binding gene as an example, the TET RNA-binding proteins, EWSR1 and TAF15, have both been confirmed to be essential for the initiation and progression of the leukemia cell line [104, 107, 108].

3.4 Analysis of important KEGG pathways

Among the 345 features, three of them were about KEGG pathways. They were hsa03020 (RNA polymerase), hsa03015 (mRNA surveillance pathway) and hsa03013 (RNA transport). All three KEGG pathways turned out to involve in RNA metabolism. Among them, **hsa03020**, which describes the RNA polymerase associated pathway, has been screened out to be one of the essential pathways. As is known, RNA polymerase turns out to be the key enzyme that produces primary transcript RNAs in all organisms and many viruses [109]. Considering that gene transcription and expression turn out to be the essential procedures for the production of proteins and are major participants of most biological processes, it is reasonable to regard such KEGG pathways as a specific essential biological process for leukemia cells. The other two KEGG pathways, **hsa03015** and **hsa03013**, have both been related to RNA metabolism, validating the specific role of RNA metabolism for cell survival, especially the survival of leukemia cells, which has been validated by recent publications [110–112]. Describing the mRNA surveillance pathway, hsa03015 turns out to be the quality control mechanism that detects and degrades abnormal mRNAs [113]. For the specific biological functions of such KEGG pathways in leukemia cells, it has been confirmed that Human-T-cell Leukemia Virus type-I (HTLV-1) interferes with the normal mRNA decay processes and initiates the malignant transformation of leukemia cells [114]. Considering that the malignant transformation is significant for the survival of leukemia cells, it is reasonable to regard hsa03015 as a specific essential pathway. For hsa03013, it describes another specific process for the transcription of mRNAs, the RNA transport. Referring to the RNA transport from the nucleus to the cytoplasm, such biological processes describe the previous process of hsa03020 in the complete transcription process in cells, validating such biological processes as essential as hsa03020, as which we have analyzed above [109]. In addition, hsa03013 received the highest ranks in the MaxRel feature list (see S3 Table) among all KEGG pathways, suggesting it may be the most important pathway for classification of essential and non-essential genes.

In summary, these three KEGG pathways all describe RNA metabolism associated biological processes that have been confirmed to be significant for the malignant transformation and cell survival of leukemia cells, according to recent publications.

3.5 Further validation of the proposed prediction model

In this study, we used the enrichment scores of GO terms and KEGG pathways to represent each essential and non-essential gene. However, it can be seen that features of KEGG pathways were much less than those of GO terms. And in the optimal prediction model, only three features of KEGG pathways were involved. To test the contribution of KEGG pathway features, each essential and non-essential gene was represented by 279 features of KEGG pathways calculated by Eq 2. Then, the SVM algorithm was executed on these genes with its performance evaluated by ten-fold cross-validation. The predicted results were counted as SN, SP, ACC and

MCC, listed in Table 1. It can be seen that these values are 0.873, 0.989, 0.966 and 0.891, respectively. They are not better than those yielded by the optimal prediction model. However, it is still quite good.

One of the contributions of this study was to propose a prediction model for classification of essential and non-essential genes. To give more clues for biologists, this model was further adopted to test the unlabeled genes that had not been validated to be essential genes or non-essential genes. As a result, 2,576 genes were predicted to be essential genes, which are provided in S6 Table. It is hopeful that many of them can be validated by solid evidences.

4. Conclusions

Based on the recent screening of essential/non-essential genes in the leukemia cell line K562, we continued the investigation using GO terms and KEGG pathways. Some important GO terms and KEGG pathways were extracted through a number of computational methods, such as the mRMR method, IFS method and SVM algorithm. In addition, a prediction model with nearly perfect performance was built. Some of the most important GO terms were extensively analyzed, and the results partly explain why they are essential for distinguishing essential and non-essential genes. It is hopeful that the identified GO terms and KEGG pathways may give new insights for further study of essential genes. Finally, we hope that our machine learning method can be applied to address other related problems, such as DNA-binding protein prediction [115], detection of tubule boundary [116], methylation site prediction [117], phosphorylation site prediction [118], and protein-protein interaction prediction [119].

Supporting information

S1 Table. Essential and non-essential genes.
(DOCX)

S2 Table. The mRMR feature list yielded by the mRMR method.
(XLSX)

S3 Table. The MaxRel feature list yielded by the mRMR method.
(XLSX)

S4 Table. The measurements yielded by the IFS method on different feature sets.
(DOCX)

S5 Table. Detailed information of the corresponding GO terms of the top 21 features in the mRMR feature list.
(DOCX)

S6 Table. The unlabeled genes that were predicted to be essential genes by the optimal prediction model.
(DOCX)

Acknowledgments

This study was supported by the National Natural Science Foundation of China (31371335), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program, The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the Science Foundation of Anhui (1608085MC58) and the Science and Technology Research Projects of Anhui (1604e0302006).

Author Contributions

Conceptualization: Lei Chen, Tao Huang, Yu-Dong Cai.

Data curation: Lei Chen.

Formal analysis: Yu-Hang Zhang.

Funding acquisition: YunHua Zhang, Yu-Dong Cai.

Methodology: Lei Chen, ShaoPeng Wang, Yu-Dong Cai.

Validation: YunHua Zhang, Tao Huang.

Writing – original draft: Lei Chen, Yu-Hang Zhang, ShaoPeng Wang.

Writing – review & editing: Tao Huang, Yu-Dong Cai.

References

1. O'Neill RS, Clark DV. The Drosophila melanogaster septin gene Sep2 has a redundant function with the retrogene Sep5 in imaginal cell proliferation but is essential for oogenesis. Genome. 2013; 56(12): 753–8. https://doi.org/10.1139/gen-2013-0210 PMID: 24433211
2. Wu YP, Baum M, Huang CL, Rodan AR. Two inwardly rectifying potassium channels, Irk1 and Irk2, play redundant roles in Drosophila renal tubule function. Am J Physiol-Reg I. 2015; 309(7):R747–R56. https://doi.org/10.1152/ajpregu.00148.2015 PMID: 26224687
3. Ning K, Ng HK, Srihari S, Leong HW, Nesvizhskii AI. Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. Bmc Bioinformatics. 2010; 11:505. https://doi.org/10.1186/1471-2105-11-505 PMID: 20939873
4. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015; 163(6): 1515–26. https://doi.org/10.1016/j.cell.2015.11.015 PMID: 26627737
5. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015; 350(6264):1096–101. https://doi.org/10.1126/science.aac7041 PMID: 26472758
6. Commichau FM, Pietack N, Stulke J. Essential genes in Bacillus subtilis: a re-evaluation after ten years. Molecular Biosystems. 2013; 9(6):1068–75. https://doi.org/10.1039/c3mb25595f PMID: 23420519
7. Juhas M, Reuss DR, Zhu B, Commichau FM. Bacillus subtilis and Escherichia coli essential genes and minimal cell factories after one decade of genome engineering. Microbiology+. 2014; 160(Pt 11): 2341–51. https://doi.org/10.1099/mic.0.079376-0 PMID: 25092907
8. Robinson KE, Meers J, Gravel JL, McCarthy FM, Mahony TJ. The essential and non-essential genes of Bovine herpesvirus 1. J Gen Virol. 2008; 89(Pt 11):2851–63. https://doi.org/10.1099/vir.0.2008/002501-0 PMID: 18931083
9. Grazziotin AL, Vidal NM, Venancio TM. Uncovering major genomic features of essential genes in Bacteria and a methanogenic Archaea. FEBS J. 2015; 282(17):3395–411. https://doi.org/10.1111/febs.13350 PMID: 26084810
10. Kibria KM, Hossain ME, Sultana J, Sarker SA, Bardhan PK, Rahman M, et al. The Prevalence of Mixed Helicobacter pylori Infections in Symptomatic and Asymptomatic Subjects in Dhaka, Bangladesh. Helicobacter. 2015; 20(5):397–404. https://doi.org/10.1111/hel.12213 PMID: 25827337
11. White JK, Gerdin AK, Karp NA, Ryder E, Buljan M, Bussell JN, et al. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell. 2013; 154(2): 452–64. https://doi.org/10.1016/j.cell.2013.06.022 PMID: 23870131
12. Liao BY, Zhang JZ. Mouse duplicate genes are as essential as singletons. Trends in Genetics. 2007; 23(8):378–81. https://doi.org/10.1016/j.tig.2007.05.006 PMID: 17559966
13. Georgi B, Voight BF, Bucan M. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. Plos Genetics. 2013; 9(5):e1003484. https://doi.org/10.1371/journal.pgen.1003484 PMID: 23675308
14. Morgens DW, Deans RM, Li A, Bassik MC. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. Nature Biotechnology. 2016; 34(6):634–6. https://doi.org/10.1038/nbt.3567 PMID: 27159373

15. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A. 1996; 93(19):10268–73. Epub 1996/09/17. PMID: 8816789

16. Zhang X, Acencio ML, Lemke N. Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review. Front Physiol. 2016; 7:75. Epub 2016/03/26. https://doi.org/10.3389/fphys.2016.00075 PMID: 27014079

17. Zhong J, Wang J, Peng W, Zhang Z, Pan Y. Prediction of essential proteins based on gene expression programming. BMC Genomics. 2013; 14 Suppl 4:S7. Epub 2013/12/07. PMID: 24267033

18. Wei L, Wan S, Guo J, Wong KK. A novel hierarchical selective ensemble classifier with bioinformatics application. Artif Intell Med. 2017. https://doi.org/10.1016/j.artmed.2017.02.005 PMID: 28245947

19. Chen L, Zeng W-M, Cai Y-D, Feng K-Y, Chou K-C. Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities. PLoS ONE. 2012; 7(4):e35254. https://doi.org/10.1371/journal.pone.0035254 PMID: 22514724

20. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011; 12(1):56–68. https://doi.org/10.1038/nrg2918 PMID: 21164525

21. Liu B, Long R, Chou KC. iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics. 2016; 32(16):2411–8. https://doi.org/10.1093/bioinformatics/btw186 PMID: 27153623

22. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015; 43(W1):W65–71. Epub 2015/05/11. https://doi.org/10.1093/nar/gkv458 PMID: 25958395

23. Liu B, Liu F, Fang L, Wang X, Chou KC. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics. 2015; 31(8):1307–9. Epub 2014/12/17. https://doi.org/10.1093/bioinformatics/btu820 PMID: 25504848

24. Liu B, Liu F, Fang L, Wang X, Chou KC. repRNA: a web server for generating various feature vectors of RNA sequences. Molecular genetics and genomics: MGG. 2016; 291(1):473–81. https://doi.org/10.1007/s00438-015-1078-7 PMID: 26085220

25. Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, et al. Gene Ontology Consortium: going forward. Nucleic acids research. 2015; 43(D1):D1049–D56. https://doi.org/10.1093/nar/gku1179 PMID: 25428369

26. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016; 44(D1):D457–62. Epub 2015/10/18. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

27. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(8):1226–38. https://doi.org/10.1109/TPAMI.2005.159 PMID: 16119262

28. Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing. 2003; 55(1–2):169–86. http://dx.doi.org/10.1016/S0925-2312(03)00431-4

29. Corinna Cortes VV. Support-vector networks. Machine Learning. 1995; 20(3):273–97.

30. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014; 10:733. https://doi.org/10.15252/msb.20145216 PMID: 24987113

31. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol. 2007; 8(1):R3. Epub 2007/01/06. https://doi.org/10.1186/gb-2007-8-1-r3 PMID: 17204154

32. Yang J, Chen L, Kong X, Huang T, Cai YD. Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway. PLoS One. 2014; 9(9):e107202. Epub 2014/09/11. https://doi.org/10.1371/journal.pone.0107202 PMID: 25207935

33. Li Z, Li BQ, Jiang M, Chen L, Zhang J, Liu L, et al. Prediction and analysis of retinoblastoma related genes through gene ontology and KEGG. Biomed Res Int. 2013; 2013:304029. Epub 2013/09/03. https://doi.org/10.1155/2013/304029 PMID: 23998122

34. Huang T, Ji Y, Hu D, Chen B, Zhang H, Li C, et al. 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. 2016; 7(49):80990–1002. Epub 2016/11/12. https://doi.org/10.18632/oncotarget.13167 PMID: 27835598

35. Chen L, Zhang Y-H, Zheng M, Huang T, Cai Y-D. Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds. Molecular Genetics and Genomics. 2016; 291(6):2065–79. https://doi.org/10.1007/s00438-016-1240-x PMID: 27530612

36. Huang T, Liu C-L, Li L-L, Cai M-H, Chen W-Z, Xu Y-F, et al. A new method for identifying causal genes of schizophrenia and anti-tuberculosis drug-induced hepatotoxicity. Scientific Reports. 2016; 6:32571. https://doi.org/10.1038/srep32571 PMID: 27580934

37. Chen L, Zhang Y-H, Lu G, Huang T, Cai Y-D. Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. Artificial Intelligence in Medicine. 2017; 76:27–36. https://doi.org/10.1016/j.artmed.2017.02.001 PMID: 28363286

38. Wang S, Zhang YH, Lu J, Cui W, Hu J, Cai YD. Analysis and Identification of Aptamer-Compound Interactions with a Maximum Relevance Minimum Redundancy and Nearest Neighbor Algorithm. Biomed Res Int. 2016; 2016:8351204. Epub 2016/03/10. https://doi.org/10.1155/2016/8351204 PMID: 26955638

39. Chen L, Chu C, Huang T, Kong X, Cai YD. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. Amino Acids. 2015; 47(7):1485–93. Epub 2015/04/22. https://doi.org/10.1007/s00726-015-1974-5 PMID: 25894890

40. Huang T, Wang M, Cai YD. Analysis of the preferences for splice codes across tissues. Protein & cell. 2015; 6(12):904–7. Epub 2015/10/29. https://doi.org/10.1007/s13238-015-0226-5 PMID: 26507841

41. Li Z, Zhou X, Dai Z, Zou X. Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm. BMC bioinformatics. 2010; 11(1):325.

42. Ni Q, Chen L. A feature and algorithm selection method for improving the prediction of protein structural classes. Combinatorial Chemistry & High Throughput Screening. 2017.

43. Li B-Q, Zheng L-L, Hu L-L, Feng K-Y, Huang G, Chen L. Prediction of linear B-cell epitopes with mRMR feature selection and analysis. Current Bioinformatics. 2016; 11(1):22–31. https://doi.org/10.2174/1574893611666151119215131

44. Mohabatkar H, Mohammad Beigi M, Abdolahi K, Mohsenzadeh S. Prediction of Allergenic Proteins by Means of the Concept of Chous Pseudo Amino Acid Composition and a Machine Learning Approach. Medicinal Chemistry. 2013; 9(1):133–7. PMID: 22931491

45. Kohavi R, editor A study of cross-validation and bootstrap for accuracy estimation and model selection. International joint Conference on artificial intelligence; 1995: Lawrence Erlbaum Associates Ltd.

46. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014; 30(4):472–9. https://doi.org/10.1093/bioinformatics/btt709 PMID: 24318998

47. Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, et al. iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. PLOS ONE. 2014; 9(9):e106691. https://doi.org/10.1371/journal.pone.0106691 PMID: 25184541

48. Chen L, Chu C, Zhang YH, Zhu C, Kong X, Huang T, et al. Analysis of Gene Expression Profiles in the Human Brain Stem, Cerebellum and Cerebral Cortex. PLoS One. 2016; 11(7):e0159395. https://doi.org/10.1371/journal.pone.0159395 PMID: 27434030

49. Zhang PW, Chen L, Huang T, Zhang N, Kong XY, Cai YD. Classifying ten types of major cancers based on reverse phase protein array profiles. PLoS ONE. 2015; 10(3):e0123147. Epub 2015/03/31. https://doi.org/10.1371/journal.pone.0123147 PMID: 25822500

50. Platt J. Sequential Minimal Optimizaton: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14. 1998.

51. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004; 20(15):2479–81. Epub 2004/04/10. https://doi.org/10.1093/bioinformatics/bth261 PMID: 15073010

52. Chen L, Feng KY, Cai YD, Chou KC, Li HP. Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. BMC Bioinformatics. 2010; 11:293. https://doi.org/10.1186/1471-2105-11-293 PMID: 20513238

53. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405(2):442–51. Epub 1975/10/20. PMID: 1180967

54. Liu L, Chen L, Zhang Y-H, Wei L, Cheng S, Kong X-Y, et al. Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection. Journal of Biomolecular Structure and Dynamics. 2017; 35(2):312–29. https://doi.org/10.1080/07391102.2016.1138142 PMID: 26750516

55. Chen L, Chu C, Zhang Y-H, Zheng M-Y, Zhu L, Kong X, et al. Identification of Drug-Drug Interactions Using Chemical Interactions. Current Bioinformatics. 2017.

56. Fang Y, Chen L. A binary classifier for prediction of the types of metabolic pathway of chemicals. Combinatorial Chemistry & High Throughput Screening. 2017; 20(2):140–6.

57. Fan X, Kurgan L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. Journal of Biomolecular Structure and Dynamics. 2014; 32(3):448–64. https://doi.org/10.1080/07391102.2013.775969 PMID: 23534882

58. Li G, Li M, Wang J, Wu J, Wu FX, Pan Y. Predicting essential proteins based on subcellular localization, orthology and PPI networks. BMC Bioinformatics. 2016; 17 Suppl 8:279. https://doi.org/10.1186/s12859-016-1115-5 PMID: 27586883

59. Jedd G. Fungal evo-devo: organelles and multicellular complexity. Trends in cell biology. 2011; 21(1): 12–9. https://doi.org/10.1016/j.tcb.2010.09.001 PMID: 20888233

60. Gupta S, Hastak K, Afaq F, Ahmad N, Mukhtar H. Essential role of caspases in epigallocatechin-3-gallate-mediated inhibition of nuclear factor kappa B and induction of apoptosis. Oncogene. 2004; 23(14):2507–22. https://doi.org/10.1038/sj.onc.1207353 PMID: 14676829

61. Nargang FE, Preuss M, Neupert W, Herrmann JM. The Oxa1 protein forms a homooligomeric complex and is an essential part of the mitochondrial export translocase in Neurospora crassa. J Biol Chem. 2002; 277(15):12846–53. https://doi.org/10.1074/jbc.M112099200 PMID: 11823466

62. Szikra T, Krizaj D. Intracellular organelles and calcium homeostasis in rods and cones. Visual Neurosci. 2007; 24(5):733–43.

63. Papetti M, Herman IM. Mechanisms of normal and tumor-derived angiogenesis. Am J Physiol Cell Physiol. 2002; 282(5):C947–70. https://doi.org/10.1152/ajpcell.00389.2001 PMID: 11940508

64. Linnane AW, Kios M, Vitetta L. The essential requirement for superoxide radical and nitric oxide formation for normal physiological function and healthy aging. Mitochondrion. 2007; 7(1–2):1–5. https://doi.org/10.1016/j.mito.2006.11.009 PMID: 17317335

65. Leblondel G, Allain P. Ca2+ uptake and energy supply of sheep heart mitochondria in presence of some calcium antagonists. Research communications in chemical pathology and pharmacology. 1984; 44(3):499–502. PMID: 6611569

66. Rolfe DF, Brown GC. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. Physiol Rev. 1997; 77(3):731–58. PMID: 9234964

67. Rozovski U, Keating M, Estrov Z. The significance of spliceosome mutations in chronic lymphocytic leukemia. Leukemia Lymphoma. 2013; 54(7):1364–6. https://doi.org/10.3109/10428194.2012.742528 PMID: 23270583

68. MacRae T, Sargeant T, Lemieux S, Hebert J, Deneault E, Sauvageau G. RNA-Seq Reveals Spliceosome and Proteasome Genes as Most Consistent Transcripts in Human Cancer Cells. Plos One. 2013; 8(9):e72884. https://doi.org/10.1371/journal.pone.0072884 PMID: 24069164

69. Kashyap MK, Kumar D, Villa R, La Clair JJ, Benner C, Sasik R, et al. Targeting the spliceosome in chronic lymphocytic leukemia with the macrolides FD-895 and pladienolide-B. Haematologica. 2015; 100(7):945–54. https://doi.org/10.3324/haematol.2014.122069 PMID: 25862704

70. Pillozzi S, Brizzi MF, Bernabei PA, Bartolozzi B, Caporale R, Basile V, et al. VEGFR-1 (FLT-1), beta1 integrin, and hERG K+ channel for a macromolecular signaling complex in acute myeloid leukemia: role in cell migration and clinical outcome. Blood. 2007; 110(4):1238–50. https://doi.org/10.1182/blood-2006-02-003772 PMID: 17420287

71. Kushev D, Gorneva G, Taxirov S, Spassovska N, Grancharov K. Synthesis, cytotoxicity and antitumor activity of platinum(II) complexes of cyclopentanecarboxylic acid hydrazide. Biol Chem. 1999; 380(11):1287–94. https://doi.org/10.1515/BC.1999.164 PMID: 10614821

72. Popow J, Englert M, Weitzer S, Schleiffer A, Mierzwa B, Mechtler K, et al. HSPC117 Is the Essential Subunit of a Human tRNA Splicing Ligase Complex. Science. 2011; 331(6018):760–4. https://doi.org/10.1126/science.1197847 PMID: 21311021

73. Shirafuji N, Takahashi S, Matsuda S, Asano S. Mitochondrial antisense RNA for cytochrome C oxidase (MARCO) can induce morphologic changes and cell death in human hematopoietic cell lines. Blood. 1997; 90(11):4567–77. PMID: 9373268

74. Katoh I, Yasunaga T, Yoshinaka Y. Bovine Leukemia-Virus Rna Sequences Involved in Dimerization and Specific Gag Protein-Binding—Close Relation to the Packaging Sites of Avian, Murine, and Human Retroviruses. J Virol. 1993; 67(4):1830–9. PMID: 8383213

75. Cantarero L, Sanz-Garcia M, Vinograd-Byk H, Renbaum P, Levy-Lahad E, Lazo PA. VRK1 regulates Cajal body dynamics and protects coilin from proteasomal degradation in cell cycle. Scientific Reports. 2015; 5:10543. https://doi.org/10.1038/srep10543 PMID: 26068304

76. Jady BE, Richard P, Bertrand E, Kiss T. Cell cycle-dependent recruitment of telomerase RNA and Cajal bodies to human telomeres. Molecular biology of the cell. 2006; 17(2):944–54. https://doi.org/10.1091/mbc.E05-09-0904 PMID: 16319170

77. Tsang CK, Bertram PG, Ai W, Drenan R, Zheng XF. Chromatin-mediated regulation of nucleolar structure and RNA Pol I localization by TOR. The EMBO journal. 2003; 22(22):6045–56. https://doi.org/10.1093/emboj/cdg578 PMID: 14609951

78. Busch RK, Busch H. Antigenic proteins of nucleolar chromatin of Novikoff hepatoma ascites cells. Tumori. 1977; 63(4):347–57. PMID: 201061

79. Smetana K, Jiraskova I, Mikulenkova D, Klamova H. Nucleolar and cytoplasmic RNA density-concentration in leukemia granulocytic progenitors in human bone marrow biopsies: A short cytochemical note. Acta Histochem. 2011; 113(1):58–61. https://doi.org/10.1016/j.acthis.2009.07.008 PMID: 19698977

80. Park Y, Kim W, Lee JM, Park J, Cho JK, Pang K, et al. Cytoplasmic DRAK1 overexpressed in head and neck cancers inhibits TGF-beta1 tumor suppressor activity by binding to Smad3 to interrupt its complex formation with Smad4. Oncogene. 2015; 34(39):5037–45. https://doi.org/10.1038/onc.2014.423 PMID: 25531329

81. Rachakonda G, Sekhar KR, Jowhar D, Samson PC, Wikswo JP, Beauchamp RD, et al. Increased cell migration and plasticity in Nrf2-deficient cancer cell lines. Oncogene. 2010; 29(25):3703–14. https://doi.org/10.1038/onc.2010.118 PMID: 20440267

82. Doxani C, Voulgarelis M, Zintzaras E. MDR1 mRNA expression and MDR1 gene variants as predictors of response to chemotherapy in patients with acute myeloid leukaemia: a meta-analysis. Biomarkers. 2013; 18(5):425–35. https://doi.org/10.3109/1354750X.2013.808263 PMID: 23805980

83. Mata JF, Silveira VS, Mateo EC, Cortez MAA, Queiroz RGP, Yunes JA, et al. Low mRNA Expression of the Apoptosis-Related Genes CASP3, CASP8, and FAS Is Associated With Low Induction Treatment Response in Childhood Acute Lymphoblastic Leukemia (ALL). Pediatric Blood & Cancer. 2010; 55(1):100–7.

84. Ji YQ, Zhang WG, Wang J, Gu LF. mRNA expression of the XAGE-1 gene in human acute leukemia. International journal of hematology. 2010; 91(2):209–12. https://doi.org/10.1007/s12185-010-0527-7 PMID: 20178013

85. Ravi S, Schilder RJ, Kimball SR. Role of Precursor mRNA Splicing in Nutrient-Induced Alterations in Gene Expression and Metabolism. Journal Of Nutrition. 2015; 145(5):841–6. https://doi.org/10.3945/jn.114.203216 PMID: 25761502

86. Li XR, Schulte P, Godin DV, Cheng KM. Differential mRNA expression of seven genes involved in cholesterol metabolism and transport in the liver of atherosclerosis-susceptible and -resistant Japanese quail strains. Genet Sel Evol. 2012; 44:20. https://doi.org/10.1186/1297-9686-44-20 PMID: 22682430

87. Xin P. mRNA expression of iron metabolism relation genes in macrophages by infection with Salmonella typhimurium. Afr J Microbiol Res. 2011; 5(16):2245–8.

88. Gao PK, Jin Z, Cheng YY, Cao XS. RNA-Seq analysis identifies aberrant RNA splicing of TRIP12 in acute myeloid leukemia patients at remission. Tumor Biology. 2014; 35(10):9585–90. https://doi.org/10.1007/s13277-014-2228-y PMID: 24961348

89. Tekirdag KA, Korkmaz G, Ozturk DG, Agami R, Gozuacik D. MIR181A regulates starvation- and rapamycin-induced autophagy through targeting of ATG5. Autophagy. 2013; 9(3):374–85. https://doi.org/10.4161/auto.23117 PMID: 23322078

90. Root-Bernstein R, Root-Bernstein M. The ribosome as a missing link in prebiotic evolution II: Ribosomes encode ribosomal proteins that bind to common regions of their own mRNAs and rRNAs. Journal Of Theoretical Biology. 2016; 397:115–27. https://doi.org/10.1016/j.jtbi.2016.02.030 PMID: 26953650

91. Wicker-Planquart C, Ceres N, Jault JM. The C-terminal alpha-helix of YsxC is essential for its binding to 50S ribosome and rRNAs. Febs Letters. 2015; 589(16):2080–6. https://doi.org/10.1016/j.febslet.2015.06.006 PMID: 26103561

92. Bachellerie JP, Nicoloso M, Qu LH, Michot B, CaizerguesFerrer M, Cavaille J, et al. Novel intron-encoded small nucleolar RNAs with long sequence complementarities to mature rRNAs involved in ribosome biogenesis. Biochem Cell Biol. 1995; 73(11–12):835–43. PMID: 8721999

93. Noah JW, Shapkina T, Wollenzien P. UV-induced crosslinks in the 16S rRNAs of Escherichia coli, Bacillus subtilis and Thermus aquaticus and their implications for ribosome structure and photochemistry. Nucleic Acids Research. 2000; 28(19):3785–92. PMID: 11000271

94. Negi SS, Brown P. Transient rRNA synthesis inhibition with CX-5461 is sufficient to elicit growth arrest and cell death in acute lymphoblastic leukemia cells. Oncotarget. 2015; 6(33):34846–58. https://doi.org/10.18632/oncotarget.5413 PMID: 26472108

95. Negi SS, Brown P. rRNA synthesis inhibitor, CX-5461, activates ATM/ATR pathway in acute lymphoblastic leukemia, arrests cells in G2 phase and induces apoptosis. Oncotarget. 2015; 6(20):18094–104. https://doi.org/10.18632/oncotarget.4093 PMID: 26061708

96. Appuhamy JADRN, Bell AL, Nayananjalie WAD, Escobar J, Hanigan MD. Essential Amino Acids Regulate Both Initiation and Elongation of mRNA Translation Independent of Insulin in MAC-T Cells and Bovine Mammary Tissue Slices. Journal Of Nutrition. 2011; 141(6):1209–15. https://doi.org/10.3945/jn.110.136143 PMID: 21525255

97. Volkova OA, Kochetov AV. Interrelations between the Nucleotide Context of Human Start AUG Codon, N-end Amino Acids of the Encoded Protein and Initiation of Translation. Journal Of Biomolecular Structure & Dynamics. 2010; 27(5):611–8.

98. Vary TC, Jefferson LS, Kimball SR. Amino acid-induced stimulation of translation initiation in rat skeletal muscle. Am J Physiol-Endoc M. 1999; 277(6):E1077–E86.

99. Gaccioli F, Huang CC, Wang C, Bevilacqua E, Franchi-Gazzola R, Gazzola GC, et al. Amino acid starvation induces the SNAT2 neutral amino acid transporter by a mechanism that involves eukaryotic initiation factor 2 alpha phosphorylation and cap-independent translation. Journal Of Biological Chemistry. 2006; 281(26):17929–40. https://doi.org/10.1074/jbc.M600341200 PMID: 16621798

100. Broer A, Rahimi F, Broer S. Deletion of Amino Acid Transporter ASCT2 (SLC1A5) Reveals an Essential Role for Transporters SNAT1 (SLC38A1) and SNAT2 (SLC38A2) to Sustain Glutaminolysis in Cancer Cells. Journal Of Biological Chemistry. 2016; 291(25):13194–205. https://doi.org/10.1074/jbc.M115.700534 PMID: 27129276

101. Zheng Z, Venkatapathy S, Rao G, Harrington CA. Expression profiling of B cell chronic lymphocytic leukemia suggests deficient CD1-mediated immunity, polarized cytokine response, altered adhesion and increased intracellular protein transport and processing of leukemic cells. Leukemia. 2002; 16(12):2429–37. https://doi.org/10.1038/sj.leu.2402711 PMID: 12454749

102. Dozio E, Ruscica M, Galliera E, Corsi MM, Magni P. Leptin, ciliary neurotrophic factor, leukemia inhibitory factor and interleukin-6: class-I cytokines involved in the neuroendocrine regulation of the reproductive function. Curr Protein Pept Sci. 2009; 10(6):577–84. PMID: 19751193

103. Chatterjee R, Kottaridis PD, McGarrigle H, Goldstone AH. Reversal of fludarabine induced testicular damage in a patient with chronic lymphocytic leukaemia (CLL), by suppression of pituitary-testicular axis using Gonadotrophin releasing hormone (GnRH). Leukemia Lymphoma. 2001; 41(1–2):213–5. https://doi.org/10.3109/10428190109057974 PMID: 11342377

104. Martini A, Janssen H, Bilhou-Nabera C, La Starza R, Corveleyn A, Mecucci C, et al. The TET RNA-binding proteins, EWSR1 and TAF15, are involved in acute lymphoblastic leukemia, through fusion with a new transcription factor, CIZ/NMP4. Blood. 2002; 100(11):528a–a.

105. Comai L, Song YH, Tan CY, Bui T. Inhibition of RNA polymerase I transcription in differentiated myeloid leukemia cells by inactivation of selectivity factor 1. Cell Growth & Differentiation. 2000; 11(1):63–70.

106. Lenzmeier BA, Nyborg JK. In vitro transcription of human T-cell leukemia virus type 1 is RNA polymerase II dependent. J Virol. 1997; 71(3):2577–80. PMID: 9032404

107. Martinelli G, Remiddi C, Visani G, Farabegoli P, Testoni N, Zaccaria A, et al. Molecular Analysis Of Pml-Rar-Alpha Fusion Messenger-Rna Detected by Reverse Transcription-Polymerase Chain-Reaction Assay In Long-Term Disease-Free Acute Promyelocytic Leukemia Patients. British Journal Of Haematology. 1995; 90(4):966–8. PMID: 7669683

108. Maurel S, Mougel M. Murine leukemia virus RNA dimerization is coupled to transcription and splicing processes. Retrovirology. 2010; 7:64. https://doi.org/10.1186/1742-4690-7-64 PMID: 20687923

109. Cramer P, Bushnell DA, Fu JH, Gnatt AL, Maier-Davis B, Thompson NE, et al. Architecture of RNA polymerase II and implications for the transcription mechanism. Science. 2000; 288(5466):640–9. https://doi.org/10.1126/science.288.5466.640 PMID: 10784442

110. Masiero M, Minuzzo S, Pusceddu I, Moserle L, Persano L, Agnusdei V, et al. Notch3-mediated regulation of MKP-1 levels promotes survival of T acute lymphoblastic leukemia cells. Leukemia. 2011; 25(4):588–98. https://doi.org/10.1038/leu.2010.323 PMID: 21263446

111. Fujiwara T, Zhou J, Ye S, Zhao H. RNA-binding protein Musashi2 induced by RANKL is critical for osteoclast survival. Cell Death & Disease. 2016; 7. https://doi.org/10.1038/Cddis.2016.213 PMID: 27441652

112. Palacios F, Yan XJ, Barrientos J, Kolitz JE, Allen SL, Rai KR, et al. Musashi2 RNA binding protein is up-regulated in the proliferative B-cell fraction of chronic lymphocytic leukemia clones. Leukemia Lymphoma. 2015; 56:10–1.

113. Maquat LE. Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. Nature Reviews Molecular Cell Biology. 2004; 5(2):89–99. https://doi.org/10.1038/nrm1310 PMID: 15040442

114. Nakano K, Ando T, Yamagishi M, Yokoyama K, Ishida T, Ohsugi T, et al. Viral interference with host mRNA surveillance, the nonsense-mediated mRNA decay (NMD) pathway, through a new function of HTLV-1 Rex: implications for retroviral replication. Microbes And Infection. 2013; 15(6–7):491–505. https://doi.org/10.1016/j.micinf.2013.03.006 PMID: 23541980

115. Wei L, Tang J, Zou Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. Inform Sciences. 2017; 384:135–44.

116. Su R, Zhang C, Pham TD, Davey R, Bischof L, Vallotton P, et al. Detection of tubule boundaries based on circular shortest path and polar-transformation of arbitrary shapes. Journal of microscopy. 2016; 264(2):127–42. https://doi.org/10.1111/jmi.12421 PMID: 27172164

117. Wei L, Xing P, Shi G, Ji ZL, Zou Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Trans Comput Biol Bioinform. 2017. https://doi.org/10.1109/TCBB.2017.2670558 PMID: 28222000

118. Wei L, Xing P, Tang J, Zou Q. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. IEEE Trans Nanobioscience. 2017. https://doi.org/10.1109/TNB.2017.2661756 PMID: 28166503

119. Wei L, Xing P, Zeng J, Chen J, Su R, Guo F. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif Intell Med. 2017. PMID: 28320624