Computational and Structural Biotechnology Journal 18 (2020) 668–675

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

PreMSIm: An R package for predicting microsatellite instability from the expression profiling of a gene panel in cancer

Lin Li [a,b,c], Qiushi Feng [a,b,c], Xiaosheng Wang [a,b,c,*]

[a] Biomedical Informatics Research Lab, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing 211198, China
[b] Cancer Genomics Research Center, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing 211198, China
[c] Big Data Research Institute, China Pharmaceutical University, Nanjing 211198, China

ARTICLE INFO

Article history:
Received 4 December 2019
Received in revised form 6 March 2020
Accepted 8 March 2020
Available online 19 March 2020

Keywords:
Cancer
Microsatellite instability
Gene expression profiling
Machine learning
Algorithm
Classification

ABSTRACT

Microsatellite instability (MSI) is a genomic property of the cancers with defective DNA mismatch repair and is a useful marker for cancer diagnosis and treatment in diverse cancer types. In particular, MSI has been associated with the active immune checkpoint blockade therapy response in cancer. Most of computational methods for predicting MSI are based on DNA sequencing data and a few are based on mRNA expression data. Using the RNA-Seq pan-cancer datasets for three cancer cohorts (colon, gastric, and endometrial cancers) from The Cancer Genome Atlas (TCGA) program, we developed an algorithm (PreMSIm) for predicting MSI from the expression profiling of a 15-gene panel in cancer. We demonstrated that PreMSIm had high prediction performance in predicting MSI in most cases using both RNA-Seq and microarray gene expression datasets. Moreover, PreMSIm displayed superior or comparable performance versus other DNA or mRNA-based methods. We conclude that PreMSIm has the potential to provide an alternative approach for identifying MSI in cancer.

1. Introduction

Microsatellite instability (MSI) is the molecular feature of the cancers with deficient DNA mismatch repair [1]. MSI is prevalent in several cancer types, including esophageal, gastric, colorectal, and endometrial cancers, and is a useful marker for cancer diagnosis and treatment [2]. Notably, MSI has been recognized as a biomarker for the favorable immune checkpoint blockade therapy response in cancer [3]. Thus, the detection of MSI is significant in clinical practice. The genetic or immunohistochemical tests for MSI are commonly used in clinics [4,5]. In addition, several computational methods have been proposed for the detection of MSI [6–10]. Typically, most of these computational methods for predicting MSI are based on DNA sequencing data. A few methods have been developed for predicting MSI on the basis of mRNA expression data [11–13]. However, few of these methods have been developed into easy-to-use tools. In this study, we developed an algorithm PreMSIm (Predicting MSI from mRNA) for predicting MSI from the expression profiling of a gene panel in cancer. We tested the prediction performance of PreMSIm using a number of RNA-Seq and microarray gene expression profiling datasets. Moreover, we compared the prediction performance of PreMSIm with that of other computational methods.

Abbreviations: AUC, area under the curve; CV, cross validation; GEO, Gene Expression Omnibus; GO, gene ontology; *k*-NN, *k*-nearest neighbor; MSI, microsatellite instability; MSS, microsatellite stability; PPI, protein-protein interaction; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine; TCGA, The Cancer Genome Atlas; XGBoost, extreme gradient boosting; ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THYM, thymoma; THCA, thyroid carcinoma; UCS, uterine carcinosarcoma; UCEC, uterine corpus endometrial carcinoma; UVM, uveal melanoma.

* Corresponding author at: Biomedical Informatics Research Lab, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing 211198, China.

E-mail address: xiaosheng.wang@cpu.edu.cn (X. Wang).

https://doi.org/10.1016/j.csbj.2020.03.007

2. Methods

2.1. Datasets

We downloaded the TCGA (The Cancer Genome Atlas) RNA-Seq datasets (level 3 and RSEM normalized) for six cancer cohorts (esophageal, colon, rectum, gastric, uterine, and endometrial cancers) from the genomic data commons data portal (https://portal.gdc.cancer.gov/), and the pan-cancer from the UCSC Xena project (https://xenabrowser.net/datapages/) (RSEM normalized). In addition, we downloaded 16 microarray gene expression profiling datasets (normalized) for gastric and colorectal cancers from the NCBI gene expression omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). A summary of these datasets is shown in Table 1.

2.2. Classifier development and evaluation

Before the development and evaluation of classifier, all RSEM-normalized RNA-Seq gene expression values were added 1 and then log2-transformed, and all RNA-Seq and microarray gene expression values were scaled to the range [0,1] in both training and test datasets. Within each training set, we utilized the *t*-test to identify the most significant genes in distinguishing MSI-high (MSI-H) cancers from MSI-low/microsatellite stability (MSS) cancers. The top 30 genes with the largest absolute t-scores were selected as predictors in the classification model. After feature selection, we used the *k*-Nearest Neighbors (*k*-NN, $k = 5$) classifier for class prediction. We first used 10-fold cross validation (CV) to evaluate classifier performance. Next, we built the MSI prediction model PreMSIm, which included the TCGA pan-cancer (involving colon, gastric, and endometrial cancers) dataset as the training set, the *k*-NN ($k = 5$) classification algorithm, and 15 gene features. The flowchart for the algorithm is illustrated in Fig. 1A. The 15 gene features were the genes which were commonly selected across all loops of the pan-cancer 10-fold CV model (the top 30 genes selected by t-scores based on the training set in each loop of the 10-CV are presented in Supplementary Table S1). The 15 gene features included *DDX27*, *EPM2AIP1*, *HENMT1*, *LYG1*, *MLH1*, *MSH4*, *NHLRC1*, *NOL4L*, *RNLS*, *RPL22L1*, *RTF2*, *SHROOM4*, *SMAP1*, *TTC30A*, and *ZSWIM3*. Among the 15 genes, three genes *LYG1*, *MSH4*, and *RPL22L1* were more highly expressed in the MSI-H subtype than in the MSI-L/MSS subtype of the TCGA pan-cancer and the other 12 were more lowly expressed in the MSI-H subtype (Fig. 1B). We tested the prediction performance of PreMSIm in numerous RNA-Seq and microarray gene expression profiling datasets. The classification accuracy, sensitivity, specificity, and area under the ROC curve (AUC) were reported.

2.3. Comparison of classification performance

We compared the classification performance between PreMSIm and other DNA- and mRNA-based MSI prediction methods using the TCGA datasets. The grid search using R package "caret" was applied to estimate the parameter *k* for *k*-NN. In addition, we compared the classification performance between *k*-NN and other classification algorithms, including Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost). We also compared the classification performance between PreMSIm with SMOTE [14] for correcting imbalanced classes and PreMSIm without such a correction.

2.4. Pathway, gene ontology (GO), and protein–protein interaction network analysis

We identified the pathways and GO (biological process) terms associated with the 15 gene features using the GeneCards database (https://www.genecards.org/), and investigated their protein–protein interaction (PPI) network using STRING [15].

3. Results

3.1. Classification performance of PreMSIm

Within each of the three individual cancer types (gastric, colon, and endometrial cancers), we obtained considerably high 10-fold CV accuracy, sensitivity, and specificity (Table 2). All sensitivities

Table 1
A summary of datasets.

Platform	Cancer type	Source	Number of samples	Number of MSI-H samples	Number of MSI-L/MSS samples
RNA-seq[a]	Colon cancer	TCGA	281	52	229
	Endometrial cancer		367	123	244
	Esophageal cancer		89	2	87
	Gastric cancer		415	80	335
	Rectum cancer		94	3	91
	Uterine cancer		56	2	54
	Pan-cancer		1383	328	1055
Microarray (GPL570)[b]	Gastric cancer	GSE13911	39	19	20
		GSE62254	300	68	232
	Colorectal cancer	GSE13067	74	11	63
		GSE13294	155	78	77
		GSE18088	53	19	34
		GSE26682	160	18	142
		GSE35896	61	5	56
		GSE39084	70	16	54
		GSE39582	536	77	459
		GSE75316	59	11	48
		GSE92921	58	5	53
Microarray (GPL5175)[c]		GSE24550	65	14	51
Microarray (GPL2986)[d]		GSE25071	46	5	41
Microarray (GPL13158)[b]		GSE27544	22	8	14
Microarray (GPL96)[b]		GSE26682	140	17	123
		GSE41258	168	35	133

Note:
[a] Poly-A.
[b] Affymetrix Oligonucleotide Array.
[c] Agilent Oligonucleotide Array.
[d] Affymetrix Exon Array.

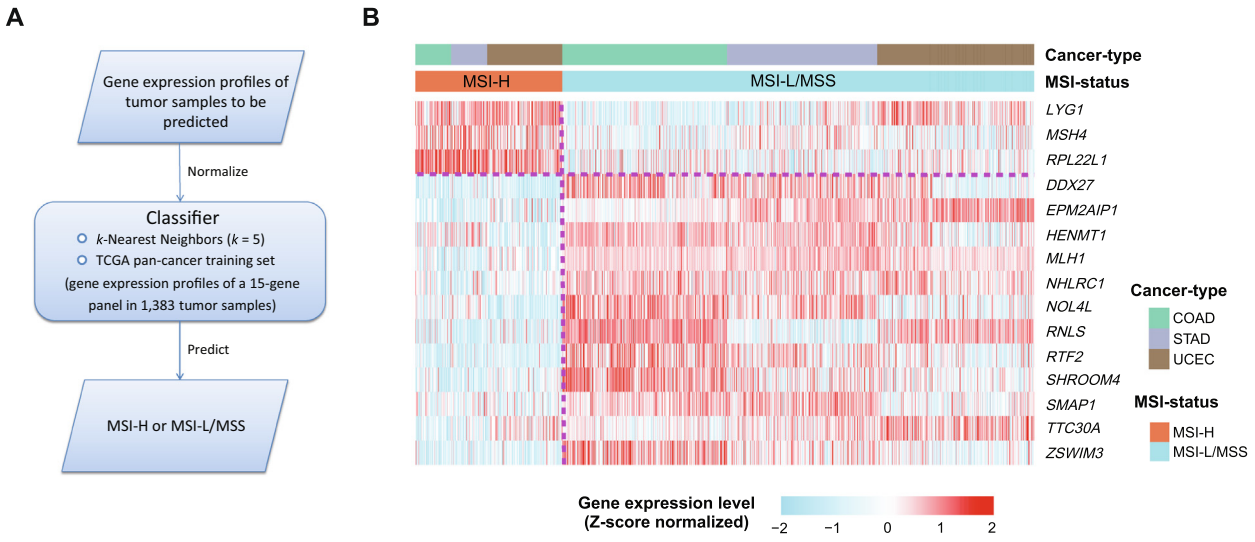

Fig 1. A summary of the PreMSIm algorithm and 15 gene signatures selected. A, Flowchart for the algorithm. B, Heatmap for the expression levels of 15 gene signatures in PreMSIm in the MSI-H and MSI-L/MSS subtypes of the TCGA pan-cancer. MSI-H: MSI-high. MSI-L/MSS: MSI-low/microsatellite stability.

Table 2
The classification performance within TCGA datasets (%).

Cancer type	Accuracy	Sensitivity	Specificity	AUC
Gastric cancer[e]	97	86	99	97
Colon cancer[e]	96	96	97	97
Endometrial cancer[e]	90	86	92	93
Pan-cancer (all samples)[e]	95	85	97	95
Pan-cancer (80% of samples)[e]	94	88	97	97
Pan-cancer (20% of samples)[f]	94	86	96	95

Accuracy

Training	Test		
	Gastric cancer	Colon cancer	Endometrial cancer
Gastric cancer		97	87
Colon cancer	92		80
Endometrial cancer	82	85	

Sensitivity

	Test		
Training	Gastric cancer	Colon cancer	Endometrial cancer
Gastric cancer		83	63
Colon cancer	80		62
Endometrial cancer	94	83	

Specificity

	Test		
Training	Gastric cancer	Colon cancer	Endometrial cancer
Gastric cancer		100	100
Colon cancer	95		90
Endometrial cancer	79	86	

AUC

	Test		
Training	Gastric cancer	Colon cancer	Endometrial cancer
Gastric cancer		94	92
Colon cancer	95		83
Endometrial cancer	95	87	

Note:
[e] 10-fold cross validation.
[f] Validation in the independent test set.

Table 3
The classification performance of PreMSIm (%).

Cancer type [g]	Accuracy	Sensitivity	Specificity	AUC
Esophageal cancer	96	100	95	99
Rectum cancer	91	67	92	79
Uterine cancer	95	100	94	99
Gastric cancer (GSE13911)	90	89	90	89
Gastric cancer (GSE62254)	88	78	91	87
Colorectal cancer (GSE13067)	98	100	95	99
Colorectal cancer (GSE13294)	92	86	99	96
Colorectal cancer (GSE18088)	96	95	97	97
Colorectal cancer (GSE26682-GPL570)	98	83	99	93
Colorectal cancer (GSE26682-GPL96)	70	82	68	81
Colorectal cancer (GSE35896)	92	100	91	98
Colorectal cancer (GSE39084)	93	100	91	99
Colorectal cancer (GSE39582)	90	90	90	94
Colorectal cancer (GSE41258)	77	60	82	82
Colorectal cancer (GSE75316)	95	100	94	99
Colorectal cancer (GSE92921)	95	80	96	89
Colorectal cancer (GSE27544)	91	75	100	94
Colorectal cancer (GSE24550)	88	100	84	94
Colorectal cancer (GSE25071)	87	100	85	98

Note:
[*] These prediction results were obtained by the PreMSIm R package.
[g] The Esophageal, Rectum, and Uterine cancer datasets were from TCGA and the others were from GEO.

were higher than 85% and specificities were higher than 90% in the three cancer types. Notably, in colon cancer, we attained 96% accuracy, 96% sensitivity, and 97% specificity. In the pan-cancer analysis, 95% accuracy, 85% sensitivity, and 97% specificity (10-fold CV) were achieved (Table 2). Moreover, we randomly separated all pan-cancer samples into training (80% of samples) and test sets (20% of samples). In the training set, 94% accuracy, 88% sensitivity, and 97% specificity (10-fold CV) were achieved, and in the test set, the accuracy, sensitivity, and specificity were 94%, 86%, and 96%, respectively (Table 2). Furthermore, we took all samples in an individual cancer type as the training set and all samples in another individual cancer type as the test set. In general, we achieved good classification performance in these experiments (Table 2). For example, when training using the endometrial cancers, the classification accuracy, sensitivity, and specificity were 82% (or 85%), 94% (or 83%), and 79% (or 86%) in testing the gastric (or colon) cancers, respectively. All together, these results demonstrate that the gene expression profiling-based prediction of MSI is fairly accurate.

Furthermore, we used PreMSIm to predict MSI in several other TCGA datasets, including esophageal, rectum, and uterine cancers. In these cohorts, we achieved considerably high accuracy, sensitivity, and specificity in general (Table 3). In addition, we tested PreMSIm in 16 external microarray gene expression profiling datasets. As shown in Table 3, the classification accuracy, sensitivity, and specificity were high or acceptable in most of these datasets. It suggests that PreMSIm is robust in predicting MSI.

3.2. Comparison of PreMSIm with other methods

We compared PreMSIm with three DNA-based MSI predictors, including MOSAIC [16], MANTIS [17], and MSIsensor [8]. Hause et al. used MOSAIC to detect MSI across 18 TCGA cancer types [16]. We predicted MSI from the same cancer types using PreMSIm and found that our prediction results highly overlapped with theirs in pan-cancer (Fisher's exact test, $P = 2.57 \times 10^{-241}$) and multiple individual cancer types, including gastric, endometrial, colon, rectum, and lung (glandular) cancers (Fisher's exact test, $P < 0.05$) (Fig. 2A). In another recent study [17], Bonneville et al. predicted MSI across 33 TCGA cancer types using MANTIS. Their prediction results also highly overlapped with the results yielded by PreMSIm in pan-cancer (Fisher's exact test, $P < 2 \times 10^{-16}$) and diverse individual cancer types, including gastric, endometrial, colon, cervical, esophageal, breast, uterine, rectum, head and neck, lung (glandular), and bladder cancers (Fisher's exact test, $P < 0.05$) (Fig. 2B). Niu et al. used MSIsensor for predicting MSI across 33 TCGA cancer types [8]. We found that the results predicted by PreMSIm and MSIsensor were highly overlapped in pan-cancer (Fisher's exact test, $P < 2 \times 10^{-16}$), as well as in multiple individual cancer types, including gastric, endometrial, colon, cervical, esophageal, lung (glandular), breast, rectum, head and neck cancers (Fisher's exact test, $P < 0.05$) (Fig. 2C). Furthermore, we compared the classification performance between PreMSIm and the three DNA-based methods in six TCGA cancer cohorts with a relatively prevalent MSI subtype, including COAD, STAD, UCEC, READ, ESCA, and UCS. In general, the accuracy and specificity were close between PreMSIm and these DNA-based methods in the six cancer cohorts (Supplementary Fig. S1). The sensitivity of PreMSIm was slightly lower than that of the DNA-based methods in COAD and UCEC while it was close in the other cancer cohorts.

In addition, we compared the prediction performance of PreMSIm with that of two other mRNA-based methods [11,12]. In [11], Danaher et al. used TCGA RNA-seq datasets for colon, stomach, and endometrial cancers as the training set and selected as features the mismatch repair genes *MLH1*, *PMS2*, *MSH2*, and *MSH6* and 10 other genes strongly associated with tumor hypermutation in pan-cancer to predict MSI. The authors reported 90% (or 79%), 90% (or 75%), and 93% (or 70%) sensitivity (or specificity) in COAD, STAD, and UCEC, respectively, compared to 85% (or 99%), 95% (or 99%), and 87% (or 96%) sensitivity (or specificity) achieved by PreMSIm (Fig. 2D). In [12], Pacinkova et al. developed an algorithm for predicting MSI on the basis of a 25-gene expression signature. In all the six datasets tested in that study, PreMSIm outperformed the 25-gene expression signature algorithm with higher AUC (Fig. 2E).

Collectively, these results demonstrate that the classification performance of PreMSIm is superior to or comparable with that of the established algorithms.

3.3. Comparison of k-NN with other classifiers

k-NN is a lazy machine learning algorithm in that it invests more in the prediction than in the learning procedure. The *k* is often an odd number in the binary classification. To investigate whether the *k*-NN ($k = 5$) is an optimal choice for PreMSIm, we

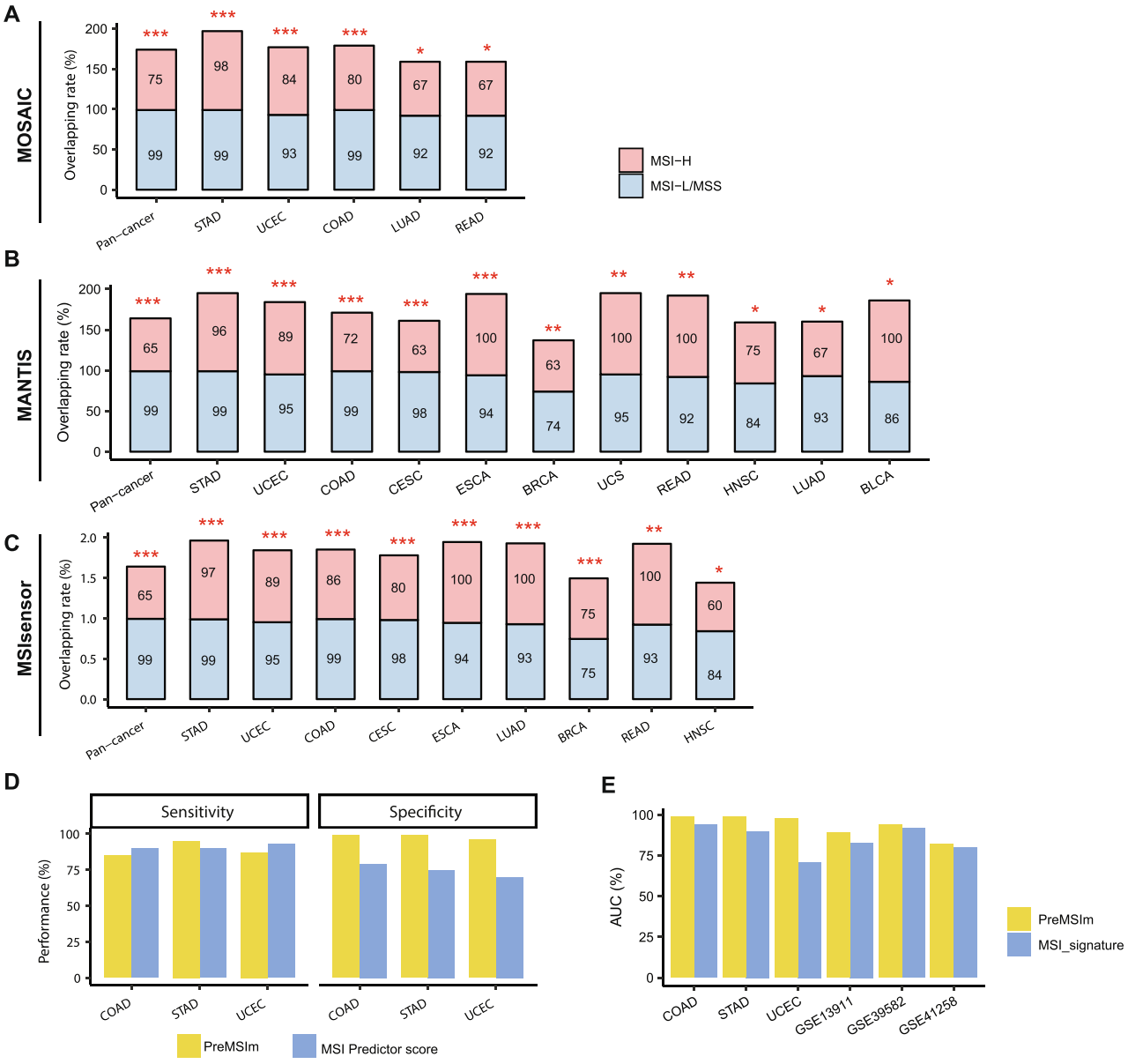

Fig 2. Comparisons of the MSI prediction results by PreMSIm with those by other algorithms. A, B, and C, The overlapping rates of the MSI prediction results between PreMSIm and MOSAIC [16] (A), MANTIS [17] (B), and MSIsensor [8] (C) in the TCGA pan-cancer and multiple individual cancer types. The Fisher's exact test *P*-values are shown. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. D and E, Comparisons of the prediction performance of PreMSIm with that of two other mRNA-based methods by Danaher et al. [11] (D) and by Pacinkova et al. [12] (E), respectively. BLCA: bladder urothelial carcinoma. BRCA: breast invasive carcinoma. CESC: cervical squamous cell carcinoma and endocervical adenocarcinoma. COAD: colon adenocarcinoma. ESCA: esophageal carcinoma. HNSC: head and neck squamous cell carcinoma. LUAD: lung adenocarcinoma. READ: rectum adenocarcinoma. STAD: stomach adenocarcinoma. UCEC: uterine corpus endometrial carcinoma. UCS: uterine carcinosarcoma.

compared the classification performance between different *k*-NNs and between *k*-NN and other commonly used classification algorithms, including RF, SVM, and XGBoost. We first used grid search with 10-fold CV in the TCGA pan-cancer to search for the optimal *k* (s) for *k*-NN, and found that the classification performance was the best when *k* = 5, 7, 9, and 11 (Fig. 3A). Furthermore, we compared the performance between the four different *k*-NNs (*k* = 5, 7, 9, and 11) in predicting MSI using two of the TCGA COAD, STAD, and UCEC datasets as the training set and the other one as the test set. In general, the prediction performance was close between the different *k*-NNs (Fig. 3B). These results indicate that *k*-NN (*k* = 5) is a reasonable choice for PreMSIm. In comparison of *k*-NN (*k* = 5) with RF, SVM, and XGBoost, *k*-NN showed comparable prediction performance with these classifiers (Fig. 3C). Because the number of MSI samples is far less than that of non-MSI samples, we used the SMOTE method [14] for correcting imbalanced classes by amplifying the number of MSI samples by 2-fold. We observed a slightly elevated sensitivity while decreased accuracy and specificity after using SMOTE (Supplementary Fig. S2). These results indicate that the class correction methods may not necessarily improve the performance of PreMSIm.

3.4. Biological characteristics of the 15 gene features

In our prediction model, a total of 15 gene features were used. Pathway analysis showed that these genes were mainly involved in DNA damage repair (*MLH1* and *MSH4*), cell cycle regulation (*MLH1*, *MSH4*, and *HENMT1*), pathways in cancer (*MLH1*), metabolism (*NHLRC1* and *RPL22L1*), and gene expression (*MLH1*, *HENMT1*, and *RPL22L1*) (Table 4). Gene ontology (GO) analysis showed that

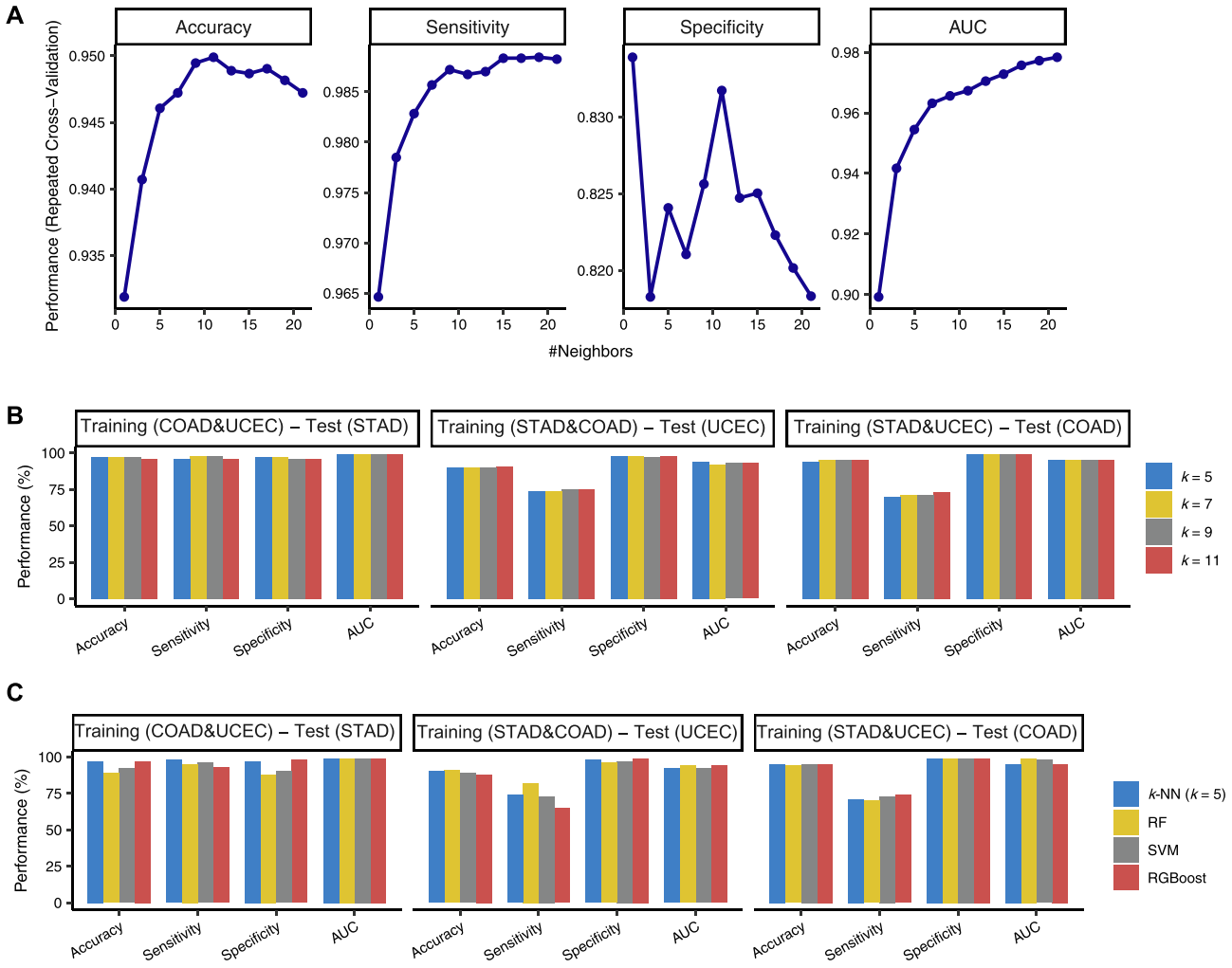

Fig 3. Comparison of *k*-NN with other classifiers. A, The grid search with 10-fold CV in the TCGA pan-cancer to search for the optimal *k*(s) for *k*-NN. B, Comparison of the performance between four different *k*-NNs (*k* = 5, 7, 9, and 11) in predicting MSI. C, Comparison of the performance between *k*-NN (*k* = 5) and the RF, SVM, and XGBoost classifiers. RF: random forest. SVM: support vector machine. XGBoost: extreme gradient boosting.

these genes were involved in the biological processes of DNA repair (*MLH1*, *MSH4*, and *RTF2*), cell cycle (*MLH1*, *MSH4*, *RPL22L1*, and *RTF2*), metabolic process (*NHLRC1*, *HENMT1*, *LYG1*, and *SMAP1*), gene expression regulation (*NHLRC1* and *HENMT1*), biogenesis (*NHLRC1*, *RNLS*, *DDX27*, and *EPM2AIP1*), and cell and organism development (*SHROOM4*, *SMAP1*, and *TTC30A*) (Table 4). Furthermore, network analysis showed that few of these genes interacted with each other, except between *MLH1* and *MSH4* (strong evidence), between *MLH1* and *EPM2AIP1* (weak evidence), and between *NHLRC1* and *EPM2AIP1* (weak evidence) (Supplementary Fig. S3).

4. Discussion

We developed an algorithm (PreMSIm) for predicting MSI from gene expression profiles in cancer. We demonstrated the accuracy and robustness of PreMSIm by testing it in various datasets with varying cancer types and platforms. In the 31 classification results, 23 (74%) had AUC above 0.9 and 30 (97%) had AUC above 0.8 (Tables 2 & 3, Fig. 4). In PreMSIm, we used the *k*-NN (*k* = 5) algorithm which first calculated the Euclidean distance of the expression values of 15 genes between the predicted sample and each sample in the training set and then selected the five samples in the training set with the nearest distance from the predicted sample. The class of the predicted sample was assigned with the class in the majority of the five samples. The 15 genes were mainly involved in DNA damage, cell cycle, and metabolic process pathways or biological processes. Of them, *MLH1* encodes a protein which is a member of seven DNA mismatch repair proteins (MLH1, MLH3, MSH2, MSH3, MSH6, PMS1, and PMS2) [18]. *MLH1* promoter hypermethylation has been associated with MSI in multiple cancer types, such as endometrial [19], colorectal [20], and gastric cancers [21]. Our model shows that this gene is downregulated in MSI cancers, consistent with these previous studies. Another gene *MSH4* interacts with *MLH1* during mammalian meiosis [22]. Network analysis shows that few of these genes are intercorrelated, indicating that the 15 gene features are not likely to be redundant in the classifier. Notably, unlike the other two mRNA-based MSI prediction algorithms in which multiple mismatch repair genes were selected as features [11,12], PreMSIm has only one feature (*MLH1*) being the mismatch repair gene. However, PreMSIm displays superior or comparable performance versus both mRNA-based methods (Fig. 2D&E). This indicates that not all the DNA mismatch repair genes are excellent features for detecting MSI based on their transcriptional expression levels.

A major advantage of mRNA-based over DNA-based MSI prediction algorithms is that the mRNA data is closer to protein and phenotype data than the DNA data. As a result, the mRNA data may indicate the phenotypic change of MSI which would be otherwise

Table 4
Pathways and GO associated with the 15 gene features in PreMSIm.

Gene Symbol	Pathway	GO (BP)
DDX27	NA	ribosome biogenesis; rRNA processing
EPM2AIP1	NA	positive regulation of glycogen biosynthetic process
HENMT1	Gene Expression; Mitotic Prophase; PIWI-interacting RNA (piRNA) biogenesis	RNA methylation; methylation; gene silencing by RNA; piRNA metabolic process; production of siRNA involved in RNA interference
LYG1	NA	metabolic process
MLH1	Mismatch repair; Gene Expression; Meiosis; DNA Damage; Fanconi anemia pathway; Pathways in cancer; Cell Cycle, Mitotic; DNA Double-Strand Break Repair; Regulation of TP53 Activity; DNA damage_Role of Brca1 and Brca2 in DNA repair; Direct p53 effectors	mismatch repair; DNA repair; cellular response to DNA damage stimulus; cell cycle; double-strand break repair via nonhomologous end joining; reciprocal meiotic recombination; somatic hypermutation of immunoglobulin genes; somatic recombination of immunoglobulin gene segments; meiotic chromosome segregation; homologous chromosome segregation; negative regulation of mitotic recombination; meiotic cell cycle
MSH4	Meiosis; Cell Cycle, Mitotic	meiotic cell cycle; reciprocal meiotic recombination
NHLRC1	Glucose metabolism; Ubiquitin mediated proteolysis; Metabolism	protein ubiquitination; autophagy; glycogen biosynthetic process; regulation of protein phosphorylation; glycogen metabolic process; regulation of gene expression; regulation of protein ubiquitination; response to endoplasmic reticulum stress; cellular macromolecule metabolic process; regulation of protein kinase activity; regulation of protein localization to plasma membrane
NOL4L	NA	NA
RNLS	NA	oxidation–reduction process
RPL22L1	Gene Expression; Metabolism; Metabolism of proteins	cytoplasmic translation
RTF2	NA	mitotic DNA replication termination; regulation of DNA stability; site-specific DNA replication termination at RTS1 barrier
SHROOM4	NA	multicellular organism development; actin filament organization; actin cytoskeleton organization
SMAP1	Endocytosis	positive regulation of GTPase activity; regulation of clathrin-dependent endocytosis
TTC30A	Organelle biogenesis and maintenance	cell projection organization
ZSWIM3	NA	NA

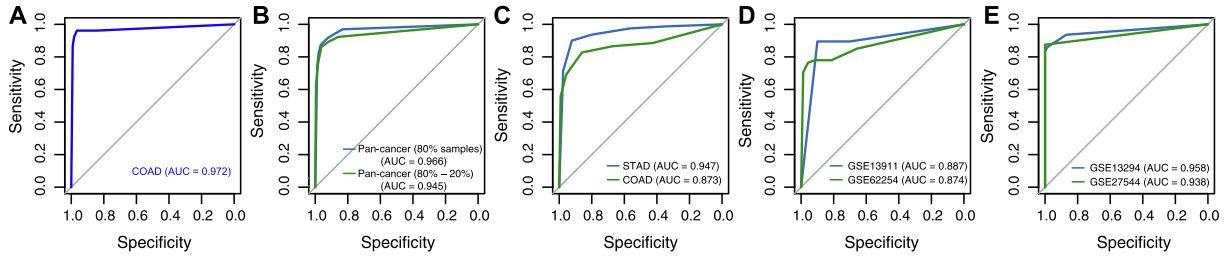

Fig 4. Prediction performance of PreMSIm in predicting MSI. A, ROC curve analysis of TCGA colon cancer. B, ROC curve analysis of pan-cancer. All pan-cancer samples were separated into training (80% of samples) and test sets (20% of samples). In the training set, the 10-fold CV AUC was shown. C, ROC curve analysis of TCGA gastric and colon cancers using the TCGA endometrial cancers as the training set. D and E, ROC curve analysis of two gastric (D) and two colorectal (E) cancer cohorts in which the PreMSIm R package was used to predict MSI. MSI: microsatellite instability. CV: cross validation. AUC: area under the ROC curve. COAD: colon adenocarcinoma. STAD: stomach adenocarcinoma.

difficult to be evaluated based on the DNA data. Thus, the mRNA-based methods are supplementary to the DNA-based methods for detecting MSI in cancer.

Our method has several limitations in detecting MSI. First, because the training data in PreMSIm are RNA-Seq data, the prediction ability of PreMSIm may compromise for microarray data. Although we normalize and scale all gene expression levels into the range [0,1], microarray data have certain properties distinct from RNA-Seq data. Indeed, PreMSIm did not achieve satisfactory results in predicting MSI for some microarray datasets (Table 3). We have tried to use microarray data to build another training set for specially predicting microarray datasets. However, because the performance did not outperform that using the RNA-seq data as the training set, we did not build such an additional training set. Second, because the training data in PreMSIm involve only three cancer cohorts (COAD, STAD, and UCEC) which have a relatively prevalent MSI subtype, the prediction power of PreMSIm for other cancer types could be weaker than that for the three cancer cohorts. Indeed, when we used PreMSIm to predict MSI in each of the 33 TCGA cancer cohorts, we observed unconfident results in some cancer cohorts (Supplementary Table S2). Hence, the improvement of the MSI prediction ability in microarray datasets and the cancer cohorts without prevalent MSI would enhance the utility of PreMSIm. This is the priority for our future study.

In conclusion, PreMSIm is superior to or comparable with the established algorithms, and is a supplementary or alternative tool for predicting MSI in cancer.

5. Availability and implementation

PreMSIm R package is publicly available in the GitHub repository (https://github.com/WangX-Lab/PreMSIm).

Funding

This work was supported by the China Pharmaceutical University (grant numbers 3150120001 to XW).

CRediT authorship contribution statement

Lin Li: Software, Validation, Formal analysis, Investigation, Resources, Data curation, Visualization, Writing - review & editing. **Qiushi Feng:** Software, Validation, Formal analysis, Investigation, Resources, Data curation, Visualization. **Xiaosheng Wang:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.03.007.

References

[1] Vilar E, Gruber SB. Microsatellite instability in colorectal cancer-the stable evidence. Nat Rev Clin Oncol 2010;7(3):153–62.
[2] de la Chapelle A, Hampel H. Clinical relevance of microsatellite instability in colorectal cancer. J Clin Oncol 2010;28(20):3380–7.
[3] Le DT et al. PD-1 blockade in tumors with mismatch-repair deficiency. N Engl J Med 2015;372(26):2509–20.
[4] Umar A et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst 2004;96(4):261–8.
[5] Hegde M et al. ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). Genet Med 2014;16(1):101–16.
[6] Salipante SJ et al. Microsatellite instability detection by next generation sequencing. Clin Chem 2014;60(9):1192–9.
[7] Pang, J., et al., Microsatellite instability detection using a large next-generation sequencing cancer panel across diverse tumour types. 2019.
[8] Niu B et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 2014;30(7):1015–6.
[9] Kautto EA et al. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget 2017;8(5):7452–63.
[10] Kather JN, Pearson AT, Halama N. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25(7):1054–6.
[11] Danaher P, Warren S. A gene expression assay for simultaneous measurement of microsatellite instability and anti-tumor immune activity. J Immunother Cancer 2019;7(1):15.
[12] Pacinkova A, Popovici V. Cross-platform data analysis reveals a generic gene expression signature for microsatellite instability in colorectal cancer. Biomed Res Int 2019;2019:6763596.
[13] Fei F et al. Efficacy and safety of docetaxel combined with oxaliplatin as a neoadjuvant chemotherapy regimen for Chinese triple-negative local advanced breast cancer patients. A prospective, open, and unicentric Phase II clinical trial. Am J Clin Oncol 2013;36(6):545–51.
[14] Chawla NV et al. SMOTE: synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16:321–57.
[15] Szklarczyk D et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47(D1):D607–13.
[16] Hause RJ et al. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med 2016;22(11):1342–50.
[17] Bonneville R et al. Landscape of microsatellite instability across 39 cancer types. JCO Precis Oncol 2017;2017.
[18] Pal T, Permuth-Wey J, Sellers TA. A review of the clinical relevance of mismatch-repair deficiency in ovarian cancer. Cancer 2008;113(4):733–42.
[19] Esteller M et al. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. Oncogene 1998;17(18):2413–7.
[20] Gazzoli I et al. A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the MLH1 gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor. Cancer Res 2002;62(14):3925–8.
[21] Haron NH et al. Microsatellite instability and altered expressions of MLH1 and MSH2 in gastric cancer. Asian Pac J Cancer Prev 2019;20(2):509–17.
[22] Santucci-Darmanin S et al. MSH4 acts in conjunction with MLH1 during mammalian meiosis. FASEB J 2000;14(11):1539–47.