bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MSIsensor-RNA: microsatellite instability detection for bulk and single-cell gene expression data

Peng Jia[1,2,†], Xuanhao Yang[1,2,†], Xiaofei Yang[1,3], Tingjie Wang[1,3], Yu Xu[4], and Kai Ye[1,2,4,5,*]

[1]MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China

[2]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China

[3]School of Computer Science, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

[4]School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China

[5]Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China

*Correspondence: kaiye@xjtu.edu.cn (Ye K)

[†]These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background: Microsatellite instability (MSI) is an indispensable biomarker in cancer immunotherapy. Currently, MSI scoring methods by high-throughput omics methods have gained popularity and demonstrated better performance than the gold standard method for MSI detection. However, MSI detection method on expression data, especially single-cell expression data is still lacking, limiting the scope of clinical application and prohibiting investigation of MSI at single cell level.

Results: Herein, we developed MSIsensor-RNA, an accuracy, robust, adaptable, and standalone software, to detect MSI status by its associated genes' expression values. We demonstrated the favorable performance and promise of MSIsensor-RNA in both bulk and single-cell gene expression data in multiplatform technologies including RNA-seq, Microarray, and single-cell RNA-seq.

Conclusions: MSIsensor-RNA is a versatile, efficient, and robust method for MSI status detection from both bulk and single-cell gene expression data in clinical researches and applications. MSIsensor-RNA is available at https://github.com/xjtu-omics/msisensor-rna.

Keywords: microsatellite instability, cancer, gene expression, multiplatform, single-cell RNA-seq, RNA-seq, Microarray

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

Microsatellite instability (MSI) refers to hypermutations of microsatellite sites due to inactivating alterations of mismatch repair (MMR) genes in malignancies [1, 2]. Currently, MSI is an indispensable pan-cancer biomarker in cancer immunotherapy therapy and prognosis, and it is routinely examined in multiple cancer types, particularly in colorectal cancer (CRC), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC) [2-5]. For example, MSI positive patients are often resistant to 5-fluorouracil treatment but have a better outcome for immune checkpoint blockade treatment [4, 5].

In clinical settings, MSI detection mainly relies on the gold-standard experimental method, MSI-PCR [6], which is laborious and time-consuming. With the advancement of next-generation-sequencing technology, numerous features of genomics, epigenomics, transcriptomics, and histology are investigated, and novel MSI computational algorithms have been developed for a variety of scenarios [7-20]. Genomics-based methods quantify MSI according to genetic mutations at microsatellite sites, which achieve high accuracy and are becoming popular in clinical MSI detection. For example, MSIsensor [9] detects MSI with high concordance as 99.4% on MSK-IMPACT panel [21]. Epigenomics-based method MIRMMR [18] detects MSI using methylation levels in MMR pathway with 0.97

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AUC. In addition, transcription levels of MSI-associated genes exhibit correlation with MSI, hinting possibility of MSI detection using transcriptomics data [15-17]. Besides these high-throughput technologies, deep learning algorithms were also applied to hematoxylin and eosin-stained slides to detect MSI [19, 20]. However, all these MSI methods detected MSI at a sample level, lacking cell-level measuring of MSI. Recently, single-cell RNA-seq (scRNA-seq) technology enables investigation of cell specific transcriptome and sheds light on tumor heterogeneity and tumor stages. In particular, the single-cell and spatial transcriptome enable the dynamic analysis of MSI in the complex tumor microenvironment, such as in metastatic and recurrent cancer [22]. However, current MSI detection methods designed for bulk gene expression data do not perform well on scRNA-seq samples. For example, the only software for gene expression data, PreMSIm [16], only provided fixed signatures and a fixed model for all cancers, which limits the widely application of the methods. Moreover, the normalized method in PreMSIm also leads to poor performance with abnormal samples. Here, we developed MSIsensor-RNA, a robust method for MSI-associated genes detection and MSI evaluation for both bulk gene expression data and single-cell RNA-seq data.

Implementation

Dataset. We downloaded RNA-seq data of 1,428 TCGA samples across CRC, STAD,

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and UCEC from TCGA Research Network (https://portal.gdc.cancer.gov) and obtained their MSI status determined by gold standards (**Table S1**). We obtained 141 RNA-seq samples of ICGC from ICGC data portal (https://dcc.icgc.org), and their MSI status reported by MIMcall [23]. Another 106 RNA-seq samples with the matched MSI status were downloaded from public publication of Clinical Proteomic Tumor Analysis Consortium (CPTAC) [24]. We also downloaded Microarray data and their MSI status of 1,468 samples across CRC and STAD from GEO dataset (https://www.ncbi.nlm.nih.gov/geo). For scRNA-seq data, we got the gene expression data and their MSI status from 133 CRC samples in two recent publications [25, 26].

Overall design. The pipeline of MSIsensor-RNA consists of data preprocessing, informative genes selection, model training, and model testing (**Fig. 1 and Fig. S1**). First, we preprocess the expression values of samples from Microarray, bulk RNA-seq, and scRNA-seq. Next, we select an informative gene set for MSI detection from 1,428 TCGA samples. Then we used these TCGA samples to train a machine learning model for each cancer type for MSI scoring. Finally, we applied the trained model to independent databases to test the performance of the MSIsensor-RNA for each cancer type.

Data preprocessing. In MSIsensor-RNA, we accept Microarray expression value,

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FPKM, TPM, and RESM read count as input. All values of expression matrix were added 1 and followed by log2 transformed. Then, for each sample or cell, expression values were normalized as a Gaussian distribution with 0 mean and 1 standard deviation. For scRNA-seq sample, to obtain accurate MSI status, we only included high-quality cells with at least 20% genes detected for MSI detection. If the number of high-quality cells was less than 20, we sort all cells by the ratio of detected genes in descending order, and the top 20 cells would be utilized for MSI detection. To solve the dropout problem of scRNA-seq, we imputed zero values by the average of the gene expression value in the given sample.

Selection of informative genes. We select informative genes for MSI classification in terms of stability, discrimination, and generalization. Firstly, we remove ribosomal genes, mitochondrial genes, and genes with low FPKM in TCGA dataset. Secondly, we selected genes with discriminative gene expression signatures between MSI samples and MSS samples. We perform rank-sum tests for expression values between MSI samples and MSS samples for each gene, and only genes with *P* value < 0.01 are included for the following analysis. Furthermore, we compute the fold of *i*th gene by:

$$F^i = \left| \log 2 \left(\frac{\frac{1}{n} \sum_{j=1}^{n} G_j^i}{\frac{1}{m-n} \sum_{k=n+1}^{m} G_k^i} \right) \right|$$

where *m* is the sample number for informative genes selection, *n* is the MSI sample

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

number, G_j^i is the gene expression value of *i*th for *j* sample. We only select genes with fold > 0.5 for candidate informative genes. Finally, we keep genes with more generalization ability for MSI detection. We calculate the area under the receiver operating characteristic curve (AUC) of the gene expression value and only genes with AUC > 0.65 are kept for next step. We also calculate the 10-fold cross validation score of SVM and random forest, and only first quartet genes are included the finial informative gene set (**Fig. S2**).

Machine learning model training and testing. We build a support vector machine (SVM) model to classify the MSI status for CRC, STAD, and UCEC in TCGA dataset. Firstly, we utilized SOMTE [27] to correct the imbalance between MSI and MSS in each cancer type by amplifying the MSI samples. Then, we utilized the expression values from correct data as input to train SVM model for MSI classification. To evaluate the performance of MSIsensor-RNA, we tested the trained model with 1,848 independent samples of multiplatform including 247 RNA-seq, 1,468 Microarray, and 133 scRNA-seq samples. For a scRNA-seq sample, we calculated the MSI score with SVM model for each high-quality cell. Then the average cell MSI score is used to evaluate the MSI status of a scRNA-seq sample.

PreMSIm running. To compare performance of MSIsensor-RNA with the only standalone software PreMSIm, we also apply the data of Microarray, RNA-seq, and

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

scRNA-seq from 1,848 independent samples to PreMSIm. For Microarray and RNA-seq samples, we test PreMSIm with two modes: PreMSIm-all and PreMSIm-split. In PreMSIm-all, we integrate all input samples to PreMSIm normalized module and predicted module. PreMSIm-split referred to input samples one database for each run.

Performance comparison of MSIsensor-RNA and PreMSIm. In MSIsensor-RNA, the predicted MSI probability by the SVM model was used to score the MSI status. The probabilities were further transformed to MSI status by the Youden index [28]. We first compared the MSIsensor-RNA score between MSI and MSS samples to test the performance of MSIsensor-RNA in multiplatform by rank sum test. To further evaluate the performance of two MSI detected methods, we calculated AUC, accuracy, F-score, precision, sensitivity, and specificity of MSIsensor-RNA and PreMSIm in different sequencing technologies.

Robustness testing of MSIsensor-RNA and PreMSIm. To test the performance of MSIsensor-RNA and PreMSIm at different normalized methods, we tested these two methods with FPKM, TPM, and read counts format of TCGA samples and calculated the AUC, F1-score, accuracy, precision, sensitivity, and specificity of each normalized method. To overcome the bias of different normalized methods and sequencing technology, we normalized the input data of each sample to a Gaussian

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distribution with 0 mean and 1 standard deviation. However, in PreMSIm, the normalization process was performed by genes, which means the normalized input data of a sample would be influenced by other samples in the bulk. Here, we tested the PreMSIm in two ways. Firstly, we input TCGA samples by three cancer types and calculated the performance of predicted MSI result. Secondly, we input all TCGA samples together to evaluate its performance. We further compared the MSI result and performance of these two ways and found that the performance of PreMSIm was affected by the way input was provided.

Results

The workflow of MSIsensor-RNA includes four modules (**Fig. 1 and Fig. S1**). First, we preprocess the expression value of Microarray, bulk RNA-seq, and scRNA-seq data. Then, we select a set of informative genes for MSI detection. Next, we train a support vector machine (SVM) model to estimate MSI scores using gene expression values of the selected informative genes. Finally, we apply the trained model to predict MSI score for either one clinical sample or a single cell (**Table S1**). For a given scRNA-seq sample, we also developed a model to report MSI status of this sample by integrating MSI scores of cells within.

MSIsensor-RNA accepts a variety of expression data including FPKM, RESM normalized read count, TPM, or microarray expression format as input. Input

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression values were added 1 and then log2 transformed following Z-score normalization per sample or cell. In particular, for single cell module of MSIsensor-RNA, we only included high-quality cell in following steps, and the missing values of each gene in high-quality cells were imputed by the average of the gene expression value in this sample.

The informative gene selection module consists of three key steps (**Fig. S2**): (i) removing mitochondrial genes and ribosomal genes; (ii) filtering of genes, of which expression values do not differ significantly between MSI and MSS samples; (iii) keeping genes, of which expression values have high generalized scores for MSI detection (online methods). We applied the gene selection module to 1,428 samples based on the gene expressions (FPKM values) from three MSI-popular cancer types (CRC, STAD, and UCEC) in TCGA dataset and finally obtained 109 informative genes for MSI classification. We also performed this step for each type of CRC, STAD, and UCEC, yielding 397, 206, and 86 informative genes, respectively (**Fig. S4 and Table S2-S5**). We found that only eight informative genes are detected in all three cancer types. Of which, we found that *MLH1* was the most important informative gene for MSI detection, as confirmed by previous reports [15-17] (**Fig. S5**).

To assess the performance of MSIsensor-RNA in bulk sample data, we first

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trained tumor-specific models for CRC, STAD, and UCEC, as well as a model for all three MSI-popular cancer types in the TCGA dataset. Then we compared the two kinds of models (tumor-specific and MSI-popular) with the standalone software, PreMSIm, in terms of the area under the curve (AUC) of the receiver operating characteristic (ROC), accuracy, sensitivity, and specificity in 1,715 (1468 Microarray and 247 bulk RNA-seq samples) independent samples. Notably, MSIsensor-RNA normalizes the expression value of informative genes for each sample independently, while PreMSIm must normalize each gene for multiple samples at the same time. Thus, we examined PreMSIm with all samples normalized together (PreMSIm-all) or by database (PreMSIm-split).

For Microarray data, we computed MSI status by MSIsensor-RNA and PreMSIm in 1,468 samples from 12 GEO accessions. The result showed that MSIsensor-RNA predicted MSI with 0.952 AUC, while PreMSIm only performed 0.628 AUC in PreMSIm-split and 0.912 AUC in PreMSIm-all mode (**Fig. 2A, S6, S7; Table S6 and S7**). Meanwhile, MSIsensor-RNA achieved much higher sensitivities than PreMSIm-split, and preMSI-all (MSIsensor-RNA: 0.968, PreMSIm-split: 0.912, PreMSIm-all: 0.384) and comparable specificities with PreMSIm-split, and preMSI-all (MSIsensor-RNA: 0.843, PreMSIm-split: 0.912, PreMSIm-all: 0.873).

To evaluate the performance using bulk RNA-seq data, we compared

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MSIsensor-RNA and two modes of PreMSIm on 247 independent samples from ICGC and CPTAC. We noticed that MSIsensor-RNA achieved 0.997 AUC in tumor-specific model and 0.985 AUC in MSI-popular model, which were significantly greater than PreMSIm-all (0.5) and PreMSIm-split (0.870) (**Fig. 2B, S8, S9; Table S8 and S9**). In addition, MSIsensor-RNA performed much better than PreMSIm for both sensitivity (MSIsensor-RNA with tumor-specific model: 0.951, MSIsensor-RNA with MSI-popular model: 0.973, PreMSIm-split: 0.834, PreMSIm-all: 0.25) and specificity (MSIsensor-RNA with tumor-specific mode: 1, MSIsensor-RNA with MSI-popular model: 0.923, PreMSIm-split: 0.906, PreMSIm-all: 0.75). To further investigate the robustness of MSIsensor-RNA for different input data types, we evaluated the performance of MSIsensor-RNA and PreMSIm with FPKM, read count, and TPM normalized samples in TCGA as input. We found that MSIsensor-RNA achieved 0.982 ± 0.040 AUC indicating the robustness of MSIsensor-RNA regardless of the measurements of gene expression (**Table S10**).

To assess the performance of MSIsensor-RNA and PreMSIm in scRNA-seq samples, we applied the trained model of MSIsensor-RNA to 23,902 high-quality cells from 133 samples to obtain sample specific MSI status and compared to the ratio of cells labeled as MSI by PreMSIm. The result showed MSIsensor-RNA detected MSI for scRNA-seq samples with 0.958 AUC, 0.9231 sensitivity, and

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

0.9362 specificity, while PreMSIm with 0.4969 AUC, 1 sensitivity, and 0.0319 specificity (**Fig. 2A**, **S10**; **Table S11 and S12**). The sample level MSI scores based on scRNA-seq was significantly different between MSI and MSS samples by MSIsensor-RNA (rank-sum test, $P = 1.01\times10^{-16}$) while no significant difference was detected for PreMSIm (rank-sum test, $P = 0.9547$) (**Fig. 2B**). Having established the effectiveness of MSIsensor-RNA on scRNA-seq sample, we investigated cell-level MSI. We computed the MSI scores of 21,438 high-quality cells from 100 samples (GSE178341) and found cell-type dependent MSI scores. For example, MSI scores of epithelial and immune cells in MSI samples were greater than that in MSS samples while no significant difference was detected between MSI and MSS for stromal cells (**Fig. 2C, S11 and Table S13**). This indicated the potential of MSIsensor-RNA to assess MSI at the single-cell level, providing a novel measurement for the investigation of tumorigenic process.

Discussion

Microsatellite instability is important for the prognosis assessment of both 5-FU chemotherapy [4] and immunotherapy [5]. In addition to gold-standard experimental methods [6], MSI status is also evaluated according to genomic sequencing data [7-14], gene expression data [15-17], methylation data [18], and H&E-stained slides [19, 20]. Compared to variants in microsatellite regions, gene expression values are

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

more directly reflective of the features of MSI and easier to obtain. In this study, we developed a robust method, MSIsensor-RNA, for MSI detection with gene expression data. MSIsensor-RNA provided informative gene selections, model training, and MSI detection modules. MSIsensor-RNA is able to process data from multiple platforms, including Microarray, RNA-seq, and single cell RNA-seq. Compared to the standalone method PreMSIm, MSIsensor-RNA also provided modules for informative gene selection and model training so that users could apply MSIsensor-RNA for different cancer types. MSIsensor-RNA also improved the normalization method of the data, yielding a more robust result than PreMSIm (Fig. 2). In addition, MSIsensor-RNA facilitates the evaluation of MSI status at the single cell level, which will be critical to better understanding the mechanism of MSI in cancer immunotherapy in the future.

In most MSI detection methods, such as MSIsensor [10] and MSIsensor-pro [11], MSI is quantified according to genetic mutations at microsatellite sites, the consequence of MSI rather than the deficiency of the MMR system, the direct cause of MSI. In this study, a set of MSI-associated genes was identified, and their expression values were used for MSI evaluation. We found that *MLH1* is the most important gene in all tested cancer types. In addition, unexpected expression of *MLH1* is commonly seen in Lynch syndrome [29]. Thus, we test the performance of MSIsensor-RNA for samples with abnormal *MLH1* expression. We train a model

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

based on all informative genes and tested it by samples with simulated abnormal *MLH1* gene expression (**Table S14**). We found that the model achieved 0.974 and 0.972 AUCs when we set the *MLH1* expression value as the maximum and minimum of all gene expression values, respectively. Furthermore, when *MLH1* was excluded from the informative gene set, MSIsensor-RNA also achieved a 0.977 AUC, indicating the robustness of MSIsensor-RNA for MSI detection.

We demonstrate that MSIsensor-RNA achieved higher performance than other methods based on gene expression and comparable performance compared to DNA-based methods (**Table S15**). In our study design, MSIsensor-RNA detects MSI according to the gene expression signature of genes on MSI associated pathways, while MSIsensor evaluates MSI by computing the ratio of somatic microsatellite mutations. Although MSIsensor achieved slightly higher performance than MSIsensor-RNA, it cannot replace the applications of MSIsensor-RNA in gene expression data. Currently, MSIsensor-RNA reports favorable performance in all three MSI-popular cancers, including colorectal cancer, stomach adenocarcinoma, and uterine corpus endometrial carcinoma. The MSI features are different in different cancer types. Thus, the model obtained low performance when the testing samples were inconsistent with training samples in cancer types (**Table S16-S18**). Therefore, the performance of MSIsensor-RNA in other cancer types needs further validation in the future.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusions

MSIsensor-RNA is a cross-platform, efficient, and robust method for MSI status determination from both bulk and single-cell gene expression data. We demonstrated the effectiveness and robustness of MSIsensor-RNA across different platforms, hinting its potential in clinical research. Moreover, MSIsensor-RNA enables single-cell level MSI evaluation, providing a new tool to discover the role of MSI in tumorigenic process and to monitor cell-level dynamic changes during immunotherapy.

Availability and requirements

Project name: msisensor-rna

Project home page: https://github.com/xjtu-omics/msisensor-rna

Operating system(s): Unix System or Docker

Programming language: python

Other requirements: python packages including numpy, pandas and scikit-learn.

License: Custom License (see at homepage)

Any restrictions to use by non-academics: MSIsensor-RNA is free for

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

non-commercial use by academic, government, and non-profit/not-for-profit institutions. A commercial version of the software is available and licensed through Xi'an Jiaotong University. For more information, please contact with pengjia@stu.xjtu.edu.cn or kaiye@xjtu.edu.cn.

Abbreviations

MSI: Microsatellite Instability

MMR: Mismatch Repair

CRC: Colorectal Cancer

STAD: Stomach Adenocarcinoma

UCEC: Uterine Corpus Endometrial Carcinoma

NGS: Next Generation Sequencing

ROC: Receiver Operating Characteristic

AUC: Area Under the Curve

scRNA-seq single-cell RNA sequencing

Declarations

Ethics approval and consent to participate

Not applicable.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Consent for publication

Not applicable.

Availability of data and materials

MSIsensor-RNA is a python program which is available at https://github.com/xjtu-omics/msisensor-rna; Supplementary information: Supplementary data are available at xxxx online.

Competing interests

The authors declare that they have no competing interests.

Funding

K.Y. and X.Y. are supported by the National Natural Science Foundation of China (32125009, 32070663 and 62172325), by the National Key R&D Program of China (2022YFC3400300), the Key Construction Program of the National “985” Project, the Fundamental Research Funds for the Central Universities (xzy012020012), and the Natural Science Basic Research Program of Shaanxi (2021GXLH-Z-098).

Authors' contributions

KY conceived and designed the study. PJ collected the data. PJ, XUY, XIY, TW, and YX discussed and developed the method. PJ and XUY implemented the method. PJ and KY wrote the manuscript. The authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

We thank everyone in the Ye lab at Xi'an Jiaotong University for helpful comments and discussion. We also thank the producers of data we used.

References

1. Yamamoto H, Imai K: **Microsatellite instability: an update**. *Arch Toxicol* 2015, **89**(6):899-921.

2. Hause RJ, Pritchard CC, Shendure J, Salipante SJ: **Classification and characterization of microsatellite instability across 18 cancer types**. *Nat Med* 2016, **22**(11):1342-1350.

3. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS *et al*: **Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer**. *Science* 2015, **348**(6230):124-128.

4. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE *et al*: **Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer**. *N Engl J Med* 2003, **349**(3):247-257.

5. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D *et al*: **PD-1 Blockade in Tumors with Mismatch-Repair Deficiency**. *N Engl J Med* 2015, **372**(26):2509-2520.

6. Baudrin LG, Deleuze JF, How-Kit A: **Molecular and Computational Methods for the Detection of Microsatellite Instability in Cancer**. *Front Oncol* 2018, **8**:621.

7. Janavicius R, Matiukaite D, Jakubauskas A, Griskevicius L: **Microsatellite instability detection by high-resolution melting analysis**. *Clin Chem* 2010, **56**(11):1750-1757.

8. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN *et al*: **A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of**

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* 1998, **58**(22):5248-5257.

9. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, Roychowdhury S: **Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS**. *Oncotarget* 2016, **8**(5).

10. Huang MN, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG: **MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations**. *Sci Rep* 2015, **5**:13321.

11. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L: **MSIsensor: microsatellite instability detection using paired tumor-normal sequence data**. *Bioinformatics* 2014, **30**(7):1015-1016.

12. Jia P, Yang X, Guo L, Liu B, Lin J, Liang H, Sun J, Zhang C, Ye K: **MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability**. *Genomics Proteomics Bioinformatics* 2020, **18**(1):65-71.

13. Hempelmann JA, Scroggins SM, Pritchard CC, Salipante SJ: **MSIplus for Integrated Colorectal Cancer Molecular Testing by Next-Generation Sequencing**. *J Mol Diagn* 2015, **17**(6):705-714.

14. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC: **Microsatellite instability detection by next generation sequencing**. *Clin Chem* 2014, **60**(9):1192-1199.

15. Maruvka YE, Mouw KW, Karlic R, Parasuraman P, Kamburov A, Polak P, Haradhvala NJ, Hess JM, Rheinbay E, Brody Y *et al*: **Analysis of somatic microsatellite indels identifies driver events in human tumors**. *Nat Biotechnol* 2017, **35**(10):951-959.

16. Ratovomanana T, Cohen R, Svrcek M, Renaud F, Cervera P, Siret A, Letourneur Q, Buhard O, Bourgoin P, Guillerm E *et al*: **Performance of Next-Generation Sequencing for the Detection of Microsatellite Instability in Colorectal Cancer With Deficient DNA Mismatch Repair**. *Gastroenterology* 2021, **161**(3):814-826 e817.

17. 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.

18. Li L, Feng Q, Wang X: **PreMSIm: An R package for predicting**

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

microsatellite instability from the expression profiling of a gene panel in cancer. *Comput Struct Biotechnol J* 2020, **18**:668-675.

19. Danaher P, Warren S, Ong S, Elliott N, Cesano A, Ferree S: **A gene expression assay for simultaneous measurement of microsatellite instability and anti-tumor immune activity**. *Journal for immunotherapy of cancer* 2019, **7**(1):15.

20. Foltz SM, Liang WW, Xie M, Ding L: **MIRMMR: binary classification of microsatellite instability using methylation and mutations**. *Bioinformatics* 2017, **33**(23):3799-3801.

21. Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, Heij LR, Tan X, Richman SD, Krause J *et al*: **Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning**. *Gastroenterology* 2020, **159**(4):1406-1416 e1411.

22. Kather JN, Pearson AT, Halama N, Jager D, Krause J, Loosen SH, Marx A, Boor P, Tacke F, Neumann UP *et al*: **Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer**. *Nat Med* 2019, **25**(7):1054-1056.

23. Middha S, Zhang L, Nafa K, Jayakumaran G, Wong D, Kim HR, Sadowska J, Berger MF, Delair DF, Shia J *et al*: **Reliable Pan-Cancer Microsatellite Instability Assessment by Using Targeted Next-Generation Sequencing Data**. *JCO Precis Oncol* 2017, **2017**(1):1-17.

24. Longo SK, Guo MG, Ji AL, Khavari PA: **Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics**. *Nat Rev Genet* 2021, **22**(10):627-644.

25. Fujimoto A, Fujita M, Hasegawa T, Wong JH, Maejima K, Oku-Sasaki A, Nakano K, Shiraishi Y, Miyano S, Yamamoto G *et al*: **Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types**. *Genome Res* 2020.

26. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, Dou Y, Zhang Y, Shi Z, Arshad OA *et al*: **Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities**. *Cell* 2019, **177**(4):1035-1049 e1019.

27. Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, Bejnood A, Dionne D, Ge WH, Xu KH *et al*: **Spatially organized multicellular immune hubs in human colorectal cancer**. *Cell* 2021, **184**(18):4734-4752 e4720.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, Vanhecke J, Verbandt S, Hong H, Min JW *et al*: **Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer**. *Nat Genet* 2020, **52**(6):594-603.

29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: Synthetic Minority Over-sampling Technique**. *Journal of Artificial Intelligence Research* 2002, **16**:321-357.

30. Schisterman EF, Perkins NJ, Liu A, Bondell H: **Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples**. *Epidemiology* 2005, **16**(1):73-81.

31. Chen W, Hampel H, Pearlman R, Jones D, Zhao W, Alsomali M, Knight D, Frankel WL: **Unexpected expression of mismatch repair protein is more commonly seen with pathogenic missense than with other mutations in Lynch syndrome**. *Hum Pathol* 2020, **103**:34-41.

Figure legends

Fig. 1. Workflow of MSIsensor-RNA. MSIsensor-RNA includes four modules: data preprocessing, informative gene selection, SVM model training, and testing. MSIsensor-RNA selects informative genes and trains SVM model by RNA-seq samples from TCGA. MSI scores are predicted by the trained model for Microarray, RNA-seq, and scRNA-seq samples.

Fig. 2. Performance of MSIsensor-pro. A-C. AUC of MSIsensor-RNA and PreMSIm in Microarray (A), RNA-seq (B), and scRNA-seq (C) samples. Tumor-specific: MSI results with tumor specific model; MSI-popular: MSI results with three MSI-popular cancer types. **D.** Boxplot of MSIsensor-RNA score in scRNA-seq samples. **E.** Violin plot of MSIsensor-RNA score of different cell types

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in scRNA-seq samples. Epithelial, stromal, and immune cell types are defined in Pelka et al.[25]

Supplementary tables

Table S1. Overview of samples in this study.

Table S2. Details of informative genes in CRC.

Table S3. Details of informative genes in STAD.

Table S4. Details of informative genes in UCEC.

Table S5. Details of informative genes in three MSI-popular cancers.

Table S6. MSI results of Microarray samples by MSIsensor-RNA and PreMSIm.

Table S7. MSI detection performance of MSIsensor-RNA and PreMSIm in Microarray samples.

Table S8. MSI results of RNA-seq samples by MSIsensor-RNA and PreMSIm.

Table S9. MSI detection performance of MSIsensor-RNA and PreMSIm in RNA-seq samples.

Table S10. MSI detection performance of MSIsensor-RNA and PreMSIm in different normalized samples.

Table S11. MSI results of scRNA-seq samples by MSIsensor-RNA.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S12. MSI detection performance of MSIsensor-RNA and preMSIm in scRNA-seq samples.

Table S13. MSI results of scRNA-seq cells by MSIsensor-RNA.

Table S14. Performance of MSIsensor-RNA with abnormal MLH1 expression values.

Table S15. Performance of MSIsensor-RNA and MSIsensor in TCGA dataset.

Table S16. AUC of MSIsensor-RNA with inconsistent training and testing samples.

Table S17: Performance of train models for cancer with low frequency MSI.

Table S18: Performance of MSIsensor-RNA for cancer with low frequency MSI by 5-fold cross validation.

Supplementary figures:

See in supplementary materials.

Figures:

Fig. 1

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

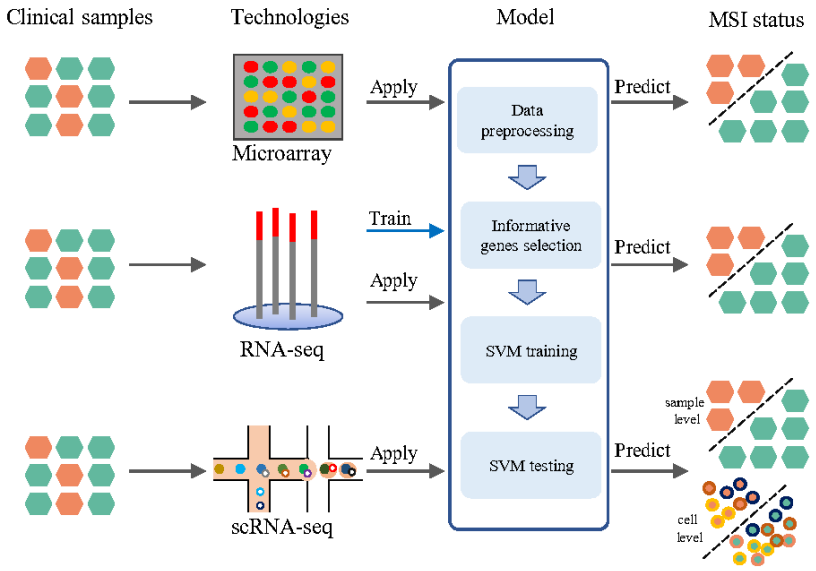
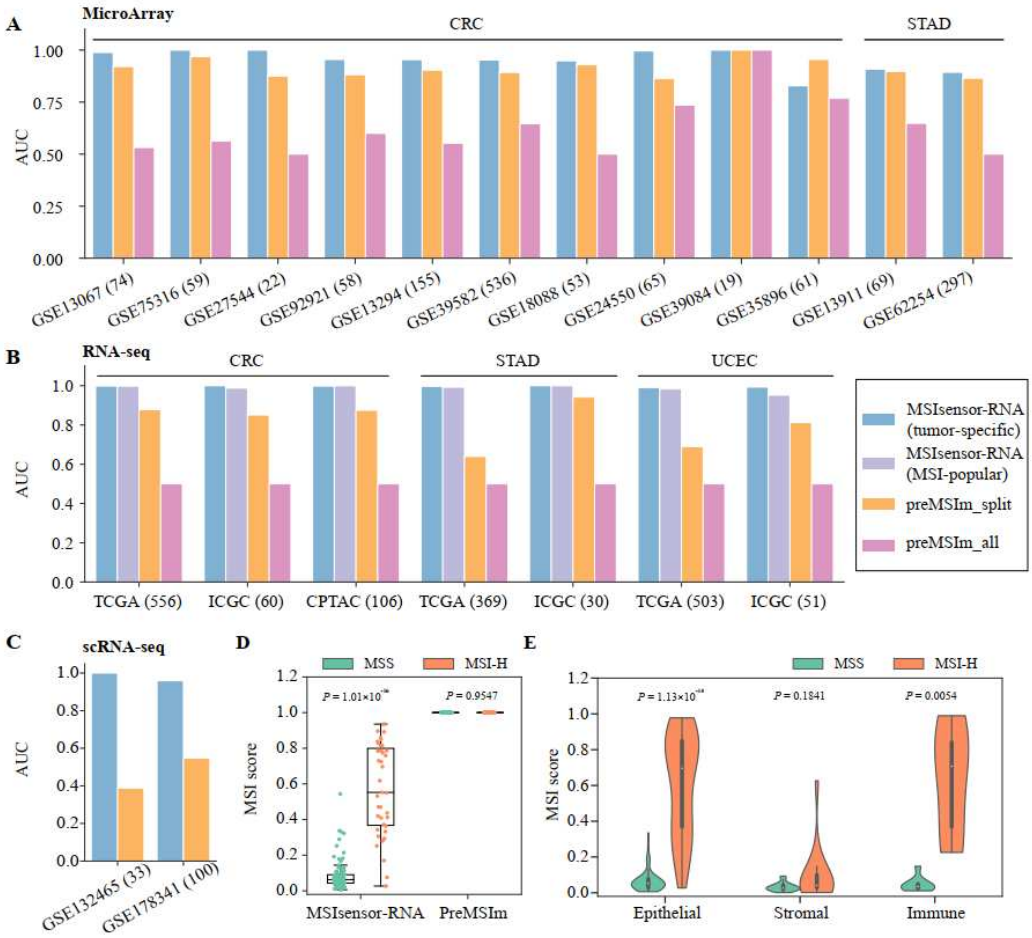
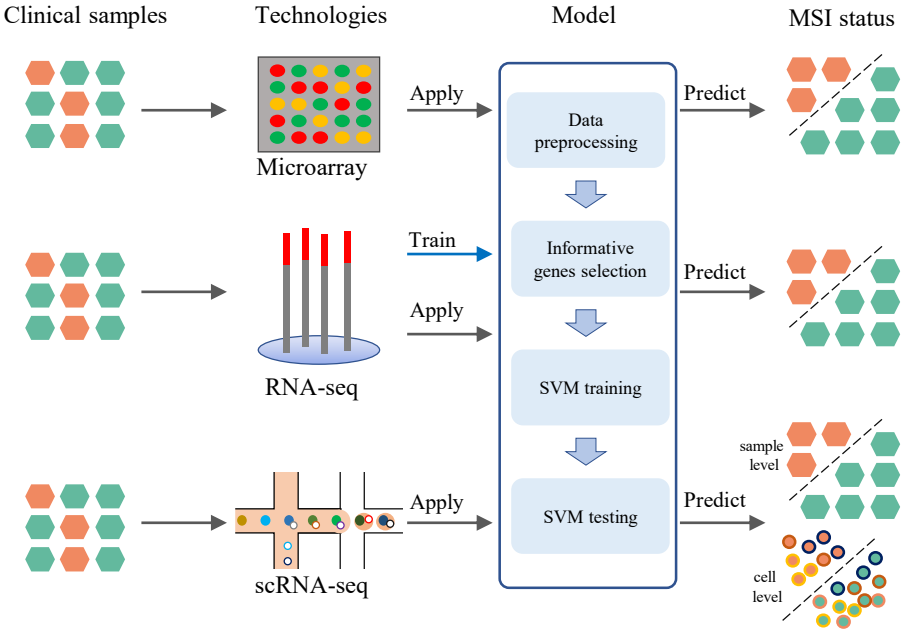


Fig. 2




Clinical samples
Technologies
Model
MSI status
Microarray
RNA-seq
scRNA-seq
Apply
Train
Apply
Apply
Data preprocessing
Informative genes selection
SVM training
SVM testing
Predict
Predict
Predict
sample level
cell level

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.09.561613; this version posted October 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.