9-2-2021

MSIsensor-ct: Microsatellite instability detection using cfDNA sequencing data

Xinyin Han

Shuying Zhang

Daniel Cui Zhou

Dongliang Wang

Xiaoyu He

See next page for additional authors

Authors

Xinyin Han, Shuying Zhang, Daniel Cui Zhou, Dongliang Wang, Xiaoyu He, Danyang Yuan, Ruilin Li, Jiayin He, Xiaohong Duan, Michael C Wendl, Li Ding, and Beifang Niu

Briefings in Bioinformatics, 22(5), 2021, 1–10

https://doi.org/10.1093/bib/bbaa402
Problem Solving Protocol

MSIsensor-ct: microsatellite instability detection using cfDNA sequencing data

Xinyin Han †, Shuying Zhang†, Daniel Cui Zhou, Dongliang Wang, Xiaoyu He, Danyang Yuan, Ruilin Li, Jiayin He, Xiaohong Duan, Michael C. Wendl, Li Ding and Beifang Niu

Corresponding authors: Beifang Niu, Computer Network Information Center, Chinese Academy of Sciences No. 4 South Street, Zhongguancun, Haidian District, Beijing, 100190, P. R. China. Fax: 86-010-58812114; E-mail: niubf@cnic.cn. Li Ding, McDonnell Genome Institute, Washington University in St. Louis, St. Louis,MO 63108, U.S.A. Fax: 314-286-1810; E-mail: lding@wustl.edu.
†These authors contributed equally to this work.

Xinyin Han is a Ph.D. candidate at the Computer Network Information Center, Chinese Academy of Sciences. He is mainly engaged in cancer genomics research focusing on the precise detection of tumor immunotherapy biomarkers. His affiliation is with Computer Network Information Center, Chinese Academy of Sciences, University of the Chinese Academy of Sciences, Beijing 100190, P. R. China.
Shuying Zhang is currently a master student at Computer Network Information Center, Chinese Academy of Sciences. Her research mainly focuses on the cancer genome and bioinformatics. Her affiliation is with Computer Network Information Center, Chinese Academy of Sciences, University of the Chinese Academy of Sciences, Beijing 100190, P. R. China.
Daniel Cui Zhou is a Ph.D. candidate in the Division of Biology and Biomedical Sciences at Washington University in St. Louis. His research focuses on the integration of bulk and single cell omics technologies in cancer. His affiliation is with McDonnell Genome Institute, Washington University in St. Louis, Department of Medicine, Washington University in St. Louis, St. Louis, MO 63108, U.S.A
Dongliang Wang earned his Ph.D. degree at Harbin Medical University. He is now the Chief Medical Officer of ChosenMed Technology (Beijing). His research mainly focuses on the mining and verification of molecular markers for tumor therapy. His affiliation is with ChosenMed Technology (Beijing) Co., Ltd., Beijing 100176, P. R. China.
Xiaoyu He is a Ph.D. candidate at the Computer Network Information Center, Chinese Academy of Sciences. She is mainly engaged in the research of the cancer genome and construction of the Chinese cancer genome database. Her affiliation is with Computer Network Information Center, Chinese Academy of Sciences, University of the Chinese Academy of Sciences, Beijing 100190, P. R. China.
Danyang Yuan is a master student at the Computer Network Information Center, Chinese Academy of Sciences. She is mainly engaged in the research of cancer genomics and leukemia-related bioinformatics. Her affiliation is with Computer Network Information Center, Chinese Academy of Sciences, University of the Chinese Academy of Sciences, Beijing 100190, P. R. China.
Ruilin Li is a Ph. D. at the Computer Network Information Center, Chinese Academy of Sciences. Her research interests include high-performance computing and bioinformatics. Her affiliation is with Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, P. R. China.
Jiayin He received her master's degree in Statistics from the George Washington University. She is currently conducting research work at the Computer Network Information Center, Chinese Academy of Sciences. Her research interests include biostatistics and computational statistics. Her affiliation is with Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, P. R. China.
Xiaohong Duan received her master's degree from Northeast Agricultural University. She is currently in charge of the laboratory at ChosenMed Technology (Beijing). She is mainly engaged in tumor genetic testing. Her affiliation is with ChosenMed Technology (Beijing) Co., Ltd., Beijing 100176, P. R. China.
Michael C. Wendl is an assistant professor at the Department of Genetics, Washington University School of Medicine in St.Louis. His research interests include mathematical and statistical aspects of cancer biology and biophysics. His affiliation is with McDonnell Genome Institute, Department of Genetics, Department of Mathematics, Washington University in St. Louis, St. Louis, MO 63108, U.S.A.
Li Ding is a professor at McDonnell Genome Institute Washington University School of Medicine in St.Louis. Her activities mainly focus focuses on integrating cancer proteogenomics, patient-derived cancer models, functional genomics, and drug treatment studies to advance cancer biology and precision medicine. Her affiliation is with McDonnell Genome Institute, Department of Genetics, Department of Medicine, Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO 63108, U.S.A.
Beifang Niu is a research professor at the Computer Network Information Center, Chinese Academy of Sciences. His research interests include cancer genomic, metagenomics and the development of computational tools for working with data from next generation sequencing technologies. His affiliation is with Computer Network Information Center, Chinese Academy of Sciences, University of the Chinese Academy of Sciences, ChosenMed Technology (Beijing) Co., Ltd., Beijing 100176, P. R. China.

Submitted: 12 October 2020; **Received (in revised form):** 25 November 2020

Abstract

Motivation: Microsatellite instability (MSI) is a promising biomarker for cancer prognosis and chemosensitivity. Techniques are rapidly evolving for the detection of MSI from tumor-normal paired or tumor-only sequencing data. However, tumor tissues are often insufficient, unavailable, or otherwise difficult to procure. Increasing clinical evidence indicates the enormous potential of plasma circulating cell-free DNA (cfNDA) technology as a noninvasive MSI detection approach. Results: We developed MSIsensor-ct, a bioinformatics tool based on a machine learning protocol, dedicated to detecting MSI status using cfDNA sequencing data with a potential stable MSIscore threshold of 20%. Evaluation of MSIsensor-ct on independent testing datasets with various levels of circulating tumor DNA (ctDNA) and sequencing depth showed 100% accuracy within the limit of detection (LOD) of 0.05% ctDNA content. MSIsensor-ct requires only BAM files as input, rendering it user-friendly and readily integrated into next generation sequencing (NGS) analysis pipelines. Availability: MSIsensor-ct is freely available at https://github.com/niu-lab/MSIsensor-ct. Supplementary information: Supplementary data are available at *Briefings in Bioinformatics* online.

Key words: MSI; cfDNA; ctDNA; machine learning

Introduction

Microsatellites (MS) refer to short repetitive DNA fragments in units of 1–6 nucleotide combinations [1]. The deficiency of mismatch repair (MMR) genes during DNA replication results in the unrecoverable chain slip phenomenon of the microsatellite segment. As a consequence, the number of microsatellite repetitions alters, giving rise to microsatellite instability (MSI) [2, 3]. MSI was first discovered in colorectal cancer by Altonen et al. in 1993 and is now confirmed as a critical carcinogenic molecular mechanism [4].

MSI is a sensitive indicator of genetic instability in diverse cancer types, including approximately 28% endometrial cancers, 15% colorectal cancers, and 22% gastric carcinomas [5–8]. It is well documented that MSI can be used to screen for Lynch syndrome, guide the neoadjuvant chemotherapy, and evaluate the prognosis of patients [9, 10]. Moreover, immune checkpoint inhibitors are more beneficial for MSI patients than those who are microsatellite stable (MSS) [11]. In 2017 and 2018, the U.S. Food and Drug Administration (FDA) respectively approved pembrolizumab [12] and nivolumab [13] for the treatment of patients with MSI solid colorectal tumors who had not progressed in previous treatments. This increasing clinical relevance underscores the urgency of accurate assessment of MSI status.

The widely-accepted assays for detecting MSI in the clinic are polymerase chain reaction (PCR) and immunohistochemistry (IHC) of the impaired DNA MMR proteins [14, 15]. Due to excessive reliance on manual operations, both methods are often criticized for insufficient precision [16].

The growth of NGS technologies applied to biological and clinical aspects [17] has prompted development of a number of NGS-based MSI detection methods, some of which are designed for paired normal and tumor samples such as MSIsensor [18] and MANTIS [19], and others like MIAmS [20], MSIsensor-pro [21], mSINGS [22], and MSIpred [23] need only tumor samples. As a result of insufficiency in tumor tissue quantity [24] or the invasive nature of the tissue biopsy [25], a number of patients do not meet the basic requirements for solid tumor sequencing [26]. It is well established that ctDNA allowing high-resolution tracking of cancer progression overcomes many of these limitations [27].

ctDNA carries genomic and epigenome mutation information matching the tumor mutation spectrum, such as copy number variation and DNA methylation [28, 29]. Importantly, the MSI phenotype can be directly assessed from ctDNA, which is highly consistent with that from tissue specimens [26, 30–32]. Therefore, how to accurately detect MSI in ctDNA through NGS to achieve noninvasive diagnoses and early tumor screening has become an increasingly urgent issue [33].

The mutations introduced by clonal hematopoiesis make it challenging to explore MSI signals from plasma cfDNA sequencing data [34, 35], though this problem is effectively solved by the intervention of UMI sequences or paired white blood cell (WBC) sequencing data [36]. To our best knowledge, only bMSISEA is now published for assessing MSI status from cfDNA sequencing data [37].

Although some MSI callers mentioned above can detect MSI status from cfDNA or tumor-only sequencing samples, there are three important limitations. Firstly, the construction of a panel-specific baseline is complicated, which requires paired cfDNA and WBC sequencing samples with known MSI statuses. Secondly, the preliminary analysis of large cohorts is inevitable due to the lack of a stable threshold for determining the MSI status. Finally, existing cfDNA MSI caller performs poorly in samples with ctDNA content $<$0.4% [37].

Here, we present MSIsensor-ct, a novel method for detecting MSI status using plasma cfDNA sequencing data, with 100% sensitivity and specificity on our 39 real samples. The limitation test demonstrates that it can reliably assess MSI status on samples with ctDNA content over 0.05% and sequencing depth over 3000×. Furthermore, its robustness renders MSIsensor-ct compatible with any pan-cancer sequencing panels covering more than 30 out of 1476 classifiers.

Materials and Methods

Catalogue of data and materials

Thirty-nine patients were recruited, concurrently with their paired cfDNA and WBC sequencing specimens obtained from ChosenMed Technology (Supplementary Table 1). Each patient had given written, informed consent, and five of them had verified MSI status by IHC, while the others had MSI status inferred by MSIsensor with filtered methods (Supplementary Material). 1724 paired whole-exome sequencing (WES) data were obtained from the Cancer Genome Atlas (TCGA), European Genome-phenome Archive (EGA, EGAS00001001056 and

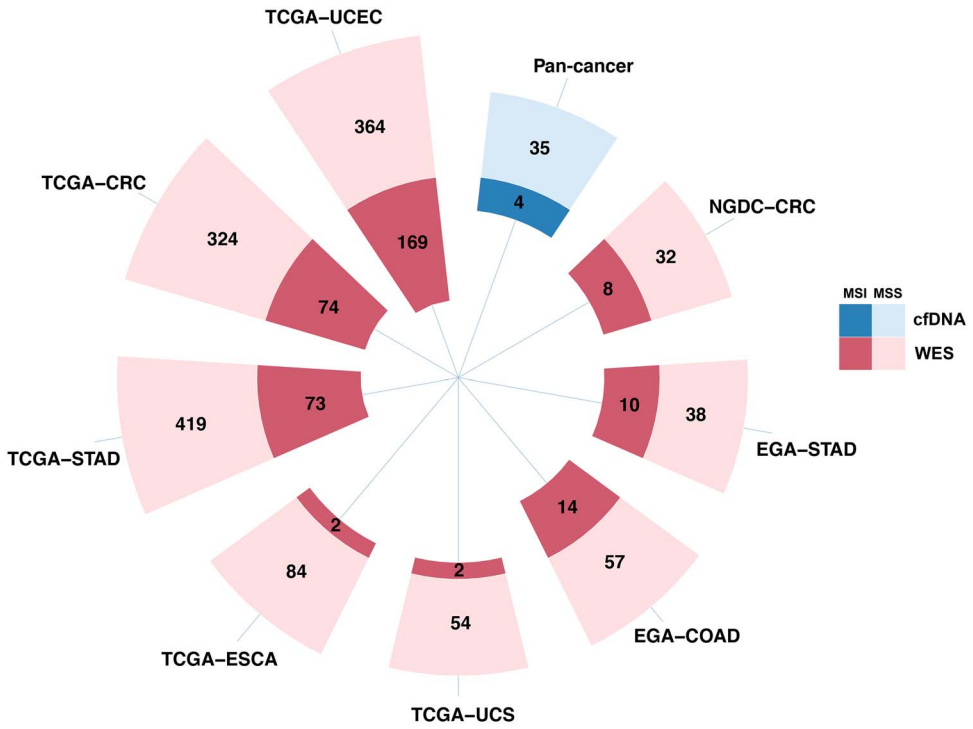

Figure 1. Distribution of MSI/MSS samples across 1763 data. There were 1565 tumor-normal paired WES data from TCGA, 119 from EGA and 40 from NGDC. In addition, 39 patients' targeted plasma cfDNA and paired WBC sequencing data were collected from ChosenMed Technology and the panel used for sequencing contains 599 genes. UCEC refers to uterine corpus endometrial carcinoma, CRC refers to colon adenocarcinoma/rectum adenocarcinoma, STAD refers to stomach adenocarcinoma, ESCA refers to esophageal carcinoma, and UCS refers to uterine carcinosarcoma. Details about 39 cfDNA samples are in Supplementary Table 1.

EGAS00001000288), and National Genomics Data Center (NGDC, PRJCA000610). All samples from public databases were MSI-labeled, and more details about cancer types with the number of MSI/MSS cases are shown in Figure 1.

In consideration of the consistency of MSI in tumor tissues and ctDNA, 1565 WES samples from TCGA were treated as the training set. Furthermore, 1565 simulated cfDNA sequencing data based on TCGA were regarded as an independent testing set, with an average sequencing depth of 10,000× and ctDNA content of 0.1%. Further tests were conducted on simulated datasets based on the 159 EGA and NGDC samples. The details of each dataset are provided in Table 1, and the simulation process is described in the supplementary material. Simultaneously, the 39 genuine cfDNA samples from ChosenMed Technology also played a crucial role in testing the performance of MSIsensor-ct.

Models for MSI calling

MSIsensor performed standard chi-square testing for each site in paired tumor and normal specimens to measure the goodness of fit between the respective allele frequency distributions [38], with the chi-square test p-value subsequently used to classify the stability of distribution in tumor [18, 39]. In this study, allele frequency distributions and p-values of microsatellites in each original TCGA sample were obtained using MSIsensor.

There were over 2.7 billion allele frequency distributions of 1763,163 microsatellites obtained from 1565 tumor samples. Here, the p-value was used to identify the stability of each distribution for the subsequent machine learning step. The stability of a distribution was labeled as 'unstable' for p-value <0.05, otherwise 'stable'.

We preserved sites that meet the following conditions: 1) sequencing depth $\geq 20\times$, 2) presenting in $\geq 10\%$ of samples, 3) the proportion of unstable distributions of a site $\geq 20\%$. Afterwards, there were over 16 million distributions of 25,861 sites remained. Allele frequency distributions and stabilities of a site in different tumor samples were aggregated into a set, which we named a site-set. Hence, 25,861 site-sets were generated (Supplementary Figure 1).

We treated the distributions in each site-set as features and the stabilities as labels. Each site-set was randomly divided into the customary training (80%) and validation (20%). Considering that ctDNA makes up only a small fraction in cfDNA, we removed the allele frequency consistent with the human reference genome in distributions to increase the influence of low-frequency variants.

Subsequently, a machine learning protocol was applied to explore the distribution pattern of each site-set, after which 25,861 site-classifiers were generated. An AUC restriction was set to filter inefficient classifiers, and only those with an AUC exceeding the restriction on validation datasets were retained. In this action, five machine learning protocols were implemented for comparison: Support Vector Machines [40], eXtreme Gradient Boosting [41], Gradient Boosting Decision Tree [42], Logistic Regression [43], and Adaptive Boosting [44]. These were tested in conjunction with different AUC restrictions. Then, the comparative evaluation was performed on the simulated dataset 1 (Table 1 and 2), where XGBoost with 0.85 AUC restriction performed overall best (Table 2).

This machine learning algorithm was incorporated into MSIsensor-ct, a C++ compiled MSI detection tool, with 1476 binary classification models (Supplementary Figure 1).

Table 1. Simulated datasets and their usage

Simulated dataset	Number of specimens	Data source	Simulated sequencing depth (×)	Simulated ctDNA content (%)	Usage	Accuracy	Sensitivity	Specificity	AUC
1	1565	TCGA samples	10,000	0.10	Test sets for 5 machine learning models		Details are in Table 2		
2	159	EGA and NGDC samples		0.10		1	1	1	1
3	159	EGA and NGDC samples	10,000	0.20	Independent test sets, robustness tests and limitation tests	1	1	1	1
4	159	EGA and NGDC samples		0.30		1	1	1	1
5	159	EGA and NGDC samples		0.40		1	1	1	1
6	159	EGA and NGDC samples	1000	0.10		0.7421	1	0.6772	1
7	159	EGA and NGDC samples	1000	0.20		1	1	1	1
8	159	EGA and NGDC samples	1000	0.30		1	1	1	1
9	159	EGA and NGDC samples	1000	0.40		1	1	1	1
10	159	EGA and NGDC samples	2000	0.05		0.7673	1	0.7087	1
11	159	EGA and NGDC samples	2000	0.10		1	1	1	1
12	159	EGA and NGDC samples	3000	0.05		1	1	1	1
13	159	EGA and NGDC samples	3000	0.10	Limitation tests	1	1	1	1
14	159	EGA and NGDC samples	5000	0.05		1	1	1	1
15	159	EGA and NGDC samples	5000	0.10		1	1	1	1
16	159	EGA and NGDC samples	10,000	0.05		1	1	1	1
17	159	EGA and NGDC samples	20,000	0.05		1	1	1	1
18	159	EGA and NGDC samples	20,000	0.10		1	1	1	1
19	159	EGA and NGDC samples	30,000	0.05		1	1	1	1
20	159	EGA and NGDC samples	30,000	0.10		1	1	1	1

This table contains 20 simulated datasets with their usage and test result. Accuracy, sensitivity and specificity are based on the threshold MSIscore = 20%. AUC: Area Under Curve

Quantification of MSI

For each sample, MSIsensor-ct reported the distributions of 1476 classifiable sites and binary classifiers were then applied to classify their stabilities. The percentage of unstable sites is the MSIscore.

Application and comparison

Simulated datasets 2–5 (Table 1) and the 39 cfDNA sequencing data from ChosenMed Technology were used as independent testing sets to verify the MSI calling ability of MSIsensor-ct. In addition, we also tested other MSI callers on the same datasets.

Baselines for MSIsensor-pro (version v1.0.a) and mSINGS (version v3.6) were established according to their detailed operating procedures [21, 22]. However, we failed to establish the baseline for bMSISEA, so it was not evaluated.

Robustness and limitation test

In order to verify the compatibility for panels with various microsatellites, the robustness of MSIsensor-ct was evaluated by randomly extracting site-classifiers from the set of 1476 and testing their MSI calling performance on simulated datasets 2–5 (Table 1). In addition, we explored the LOD of MSIsensor-ct in terms of sequencing depth and ctDNA content on simulated datasets 2, 6–20 (Table 1).

MMR proteins identification using immunohistochemistry

Immunohistochemistry was performed on formalin-fixed paraffin-embedded tissues with monoclonal antibodies against MLH1 (ES05 clone; ZM-0154; ZSGB-BIO, Beijing, China; dilution 1:50), MSH2 (RED2 clone; ZA-0622; ZSGB-BIO, Beijing, China; dilution 1:100), MSH6 (EP49 clone; ZA-0541; ZSGB-BIO, Beijing, China; dilution 1:200), and PMS2 (EP51 clone; ZA-0542; ZSGB-BIO, Beijing, China; dilution 1:50). Optical microscope was used for further interpretation.

All screened sections were photographed with a 40× field of vision, and five fields were then randomly selected under a 400× magnification. Positive cells were identified by the presence of brownish yellow granules in the nucleus, and the absence of staining was the criteria for being negative. Moreover, the expression of a specific protein was described according to the percentage of positive cells. If the percentage of positive cells was less than 25%, the protein was considered to be negative; otherwise, it was considered positive. Immunohistochemistry results were independently scored by two pathologists. If the results of the two pathologists were inconsistent, a senior pathologist reassessed the slides.

Mismatch repair–deficiency (dMMR) is defined as the loss of expression of any MMR protein in tumor tissues, and tumor cells with all four MMR proteins evaluated as positive are considered to be mismatch repair–proficient (pMMR).

Preparation of plasma cfDNA

10 mL of peripheral blood was collected in a cell-free DNA BCT tube, stored at −80 °C, and centrifuged at 1600 g for 15 minutes at 4 °C within 72 hours. Then, 2 ml of the uppermost plasma was transferred to another centrifuge tube. Additional centrifugation at 16,000 g at 4 °C was performed for 10–15 minutes to remove cell debris. Subsequently, the supernatant was transferred to a new RNAase-free tube and store at −80 °C until further DNA extraction. According to the manufacturer's instructions, cfDNA

Table 2. Comparison of the five machine learning protocols with different AUC restrictions

AUC restriction	Simulated data	XGBoost	AdaBoost	GBDT	SVM	LR
0.95	TCGA_CRC_AUC	0.9819	0.9807	0.9659	0.9844	0.9901
	TCGA_ESCA_AUC	0.9940	1.0000	1.0000	1.0000	1.0000
	TCGA_STAD_AUC	0.9896	0.9920	0.9957	0.9835	0.9828
	TCGA_UCEC_AUC	0.9589	0.9467	0.9330	0.9602	0.9544
	TCGA_UCS_AUC	0.9167	0.7092	0.7150	1.0000	0.7450
	All_samples_AUC	0.9656	0.9573	0.9490	0.9690	0.9658
	All_samples_Accuracy	0.8708	0.8867	0.8213	0.9036	0.9545
	All_samples_Sensitivity	0.9547	0.9381	0.9450	0.9450	0.8673
	All_samples_Specificity	0.8480	0.8723	0.7877	0.8923	0.9786
	Number of classifiers	28	30	18	28	27
0.90	TCGA_CRC_AUC	0.9683	0.9587	0.9738	0.9694	0.9470
	TCGA_ESCA_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	TCGA_STAD_AUC	0.9996	0.9997	0.9997	0.9997	0.9954
	TCGA_UCEC_AUC	0.9892	0.9876	0.9886	0.9853	0.9914
	TCGA_UCS_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	All_samples_AUC	0.9820	0.9792	0.9824	0.9797	0.9781
	All_samples_Accuracy	0.9669	0.9674	0.9384	0.9681	0.9708
	All_samples_Sensitivity	0.9591	0.9497	0.9686	0.9589	0.9308
	All_samples_Specificity	0.9689	0.9721	0.9306	0.9704	0.9812
	Number of classifiers	342	313	191	277	298
0.85	TCGA_CRC_AUC	0.9955	0.989	0.9845	0.9444	0.9776
	TCGA_ESCA_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	TCGA_STAD_AUC	1.0000	0.9999	1.0000	1.0000	0.9998
	TCGA_UCEC_AUC	0.9911	0.9877	0.9937	0.9891	0.9891
	TCGA_UCS_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	All_samples_AUC	0.9867	0.9811	0.9843	0.9703	0.9801
	All_samples_Accuracy	0.9738	0.9584	0.9564	0.9589	0.9565
	All_samples_Sensitivity	0.9344	0.9344	0.9563	0.95	0.8969
	All_samples_Specificity	0.9839	0.9646	0.9565	0.9612	0.9719
	Number of classifiers	1476	1431	906	882	1093
0.80	TCGA_CRC_AUC	0.9978	0.9938	0.9959	0.9948	0.9779
	TCGA_ESCA_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	TCGA_STAD_AUC	1.0000	0.9999	1.0000	1.0000	0.9998
	TCGA_UCEC_AUC	0.9925	0.9883	0.9954	0.9889	0.9888
	TCGA_UCS_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	All_samples_AUC	0.9861	0.9818	0.9863	0.9789	0.9802
	All_samples_Accuracy	0.9738	0.9661	0.9693	0.9495	0.9572
	All_samples_Sensitivity	0.8875	0.8656	0.9281	0.9250	0.9031
	All_samples_Specificity	0.9960	0.9920	0.9799	0.9558	0.9711
	Number of classifiers	3558	3548	2433	2026	2560
0.75	TCGA_CRC_AUC	0.9952	0.9947	0.9960	0.9950	0.9916
	TCGA_ESCA_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	TCGA_STAD_AUC	1.0000	0.9999	0.9999	0.9999	1.0000
	TCGA_UCEC_AUC	0.9919	0.9880	0.9959	0.9869	0.9851
	TCGA_UCS_AUC	1.0000	1.0000	1.0000	1.0000	1.0000
	All_samples_AUC	0.9843	0.9805	0.9857	0.9754	0.9795
	All_samples_Accuracy	0.9610	0.9589	0.9732	0.9463	0.9527
	All_samples_Sensitivity	0.8188	0.7986	0.8844	0.8750	0.7813
	All_samples_Specificity	0.9976	0.9983	0.9960	0.9647	0.9968
	Number of classifiers	6142	6025	4464	3598	4487

All samples referred to 1565 simulated sequencing data from TCGA samples, whose sequencing depth was 10,000X and ctDNA content was 0.1%. AUC of all samples was proposed to be the ultimate criterion.
Accuracy, sensitivity and specificity were calculated based on the threshold: MSIscore = 20%.
XGBoost: eXtreme Gradient Boosting
AdaBoost: Adaptive Boosting
GBDT: Gradient Boosting Decision Tree
SVM: Support Vector Machine
LR: Logistic Regression
AUC: Area Under Curve

was extracted from plasma samples with the QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany), and then quantified using a Qubit 4.0 fluorometer with dsDNA HS Assay Kits (Life Technologies).

Next-generation sequencing (NGS) library preparation

The purified cfDNA was adenylate 3′ ends after end repair, and adapters with dual unique molecule identifiers were ligated to both ends of DNA fragments. Target size DNA fragments were

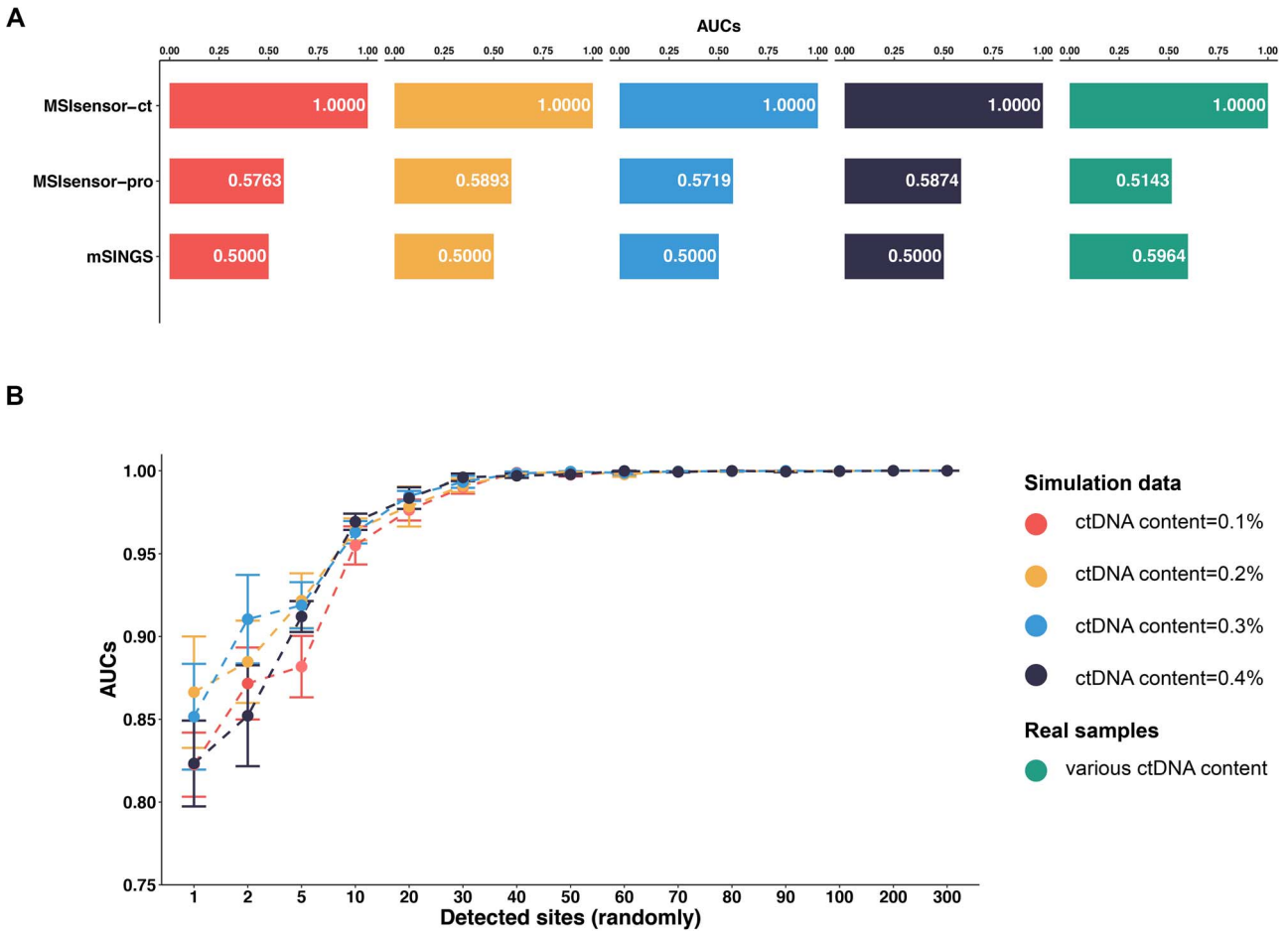

Figure 2. Performance of MSIsensor-ct. A) The comparison involved 636 simulated sequencing data with ctDNA content ranging from 0.1% to 0.4% and 39 real plasma cfDNA sequencing samples with various ctDNA contents. B) Robustness test focused on simulation data sets 2–5. Different numbers of site-classifiers (x-axis) were randomly chosen from 1476 models, and each random draw was conducted five times.

selected by magnetic beads (Beckman) and then amplified by 10-cycle PCR using index labeled primers. After beads-based purification, quantification and validation, equimolar concentrations of each library with different indices were pooled into a set and hybridized at 65 °C; enrichment of coding exons and flanking intronic regions were performed using a custom-designed 599-Gene Panel (ChosenMed Technology); pools were then washed, PCR enriched and purified. The validated DNA libraries were sequenced on NovaSeq 6000 (Illumina) according to the manufacturer's paired-end (2 × 150 bp) instructions. Read pairs were aligned to the human reference genome (hg19, downloaded from the UCSC Genome Browser) by BWA-MEM (version 0.7.11) [40]. SAMtools (version 1.3) [41] was used to generate chromosomal coordinate-sorted BAM files. The reads were realigned and quality recalibrated by Genome Analysis Toolkit (GATK, version 3.6) [42]. The maximum allelic fraction (maxAF) obtained from VarDict (version 11.5.1) [43] was adopted to estimate the ctDNA content.

Results and discussion

On the basis of MSIsensor, a well-established gradient-boosting algorithm, XGBoost, was implemented to explore the distribution patterns of 25,861 MSI site-sets, from which 1476 binary classifiers were obtained to assess the stability of sites according to their allele frequency distributions in cfDNA sequencing data. In addition, the MSI status of each sample was quantified by MSIscore reported by MSIsensor-ct and the threshold determined manually.

MSIsensor-ct accurately detected MSI status using cfDNA sequencing data with extraordinary low ctDNA concentrations in the range of 0.1%–0.4%. Comparison with other MSI callers on the also showed the outperformance of MSIsensor-ct. Further experiments were also conducted to illustrate the robustness and LOD of MSIsensor-ct at a more comprehensive level.

Performance on extra low ctDNA concentration data

We simulated 636 cfDNA sequencing data (Simulated datasets 2–5, Table 1) with extra low ctDNA content (0.1%–0.4%) to examine the performance of MSIsensor-ct, simultaneously, compared with other tumor-only or cfDNA-only MSI callers.

The result showed that MSIsensor-ct correctly distinguished 128 MSI and 508 MSS instances at an average sequencing depth of 10,000× with ctDNA content varying from 0.1% to 0.4%, which outperformed other algorithms. Moreover, for the 39 cfDNA samples across different ctDNA content, the AUC of MSIsensor-ct was uniformly 1.0, whereas none of the other tools surpassed 0.6 (Figure 2.A). Our results showed that while both mSINGS and MSIsensor-pro are validated MSI calling algorithms for tumor-only samples, there is still room for improvement in cfDNA-only samples.

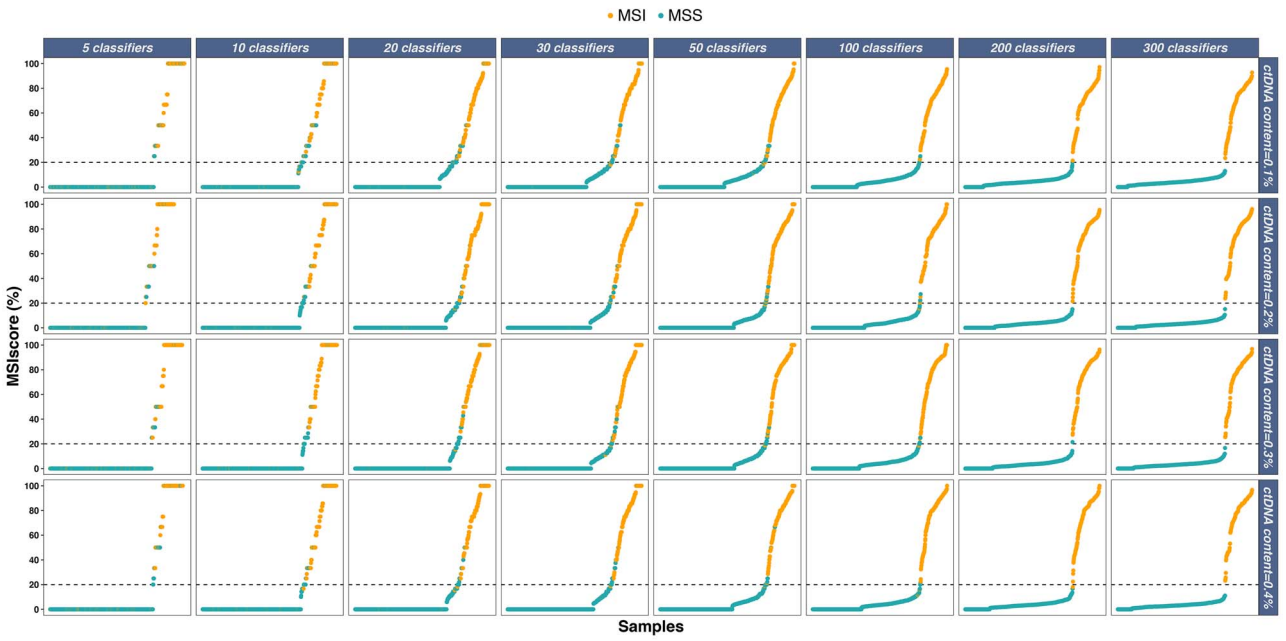

Figure 3. Threshold for MSI in MSIsensor-ct. In the robustness test, we randomly extracted 5, 10, 20, 30, 50, 100, 200, and 300 site-classifiers to explore the compatibility of MSIsensor-ct further. Five experiments were performed for each test. MSI calling accuracy with the threshold of 20% ranged between 88.95–100%, 90.69–100%, 93.21–100% and 92.55–100%, when ctDNA concentrations were 0.1%, 0.2%, 0.3 and 0.4% respectively.

Compatibility and a potential stable threshold for panels

The difference in microsatellites covered by diverse panels makes the compatibility of MSI callers a critical issue. In this regard, several tumor-only or cfDNA-only algorithms require cohort-specific retraining, which involves massive preliminary work: 1) the panel has already served sufficient patients with known MSI/MSS statuses; 2) tumor and normal (cfDNA and WBC) paired sequencing data are always required, while normal samples are sometimes difficult to obtain; 3) large crowd cohort analysis is necessary for determining the panel-specific threshold.

We verified the compatibility of MSIsensor-ct by robustness test on simulated datasets 2–5 (Table 1). Our results revealed that the AUCs of any random 10 site-classifiers attained up to 0.96 and reached 0.99 for 30 site-classifiers (Figure 2.B). In addition, the selected 1476 sites cover 84 MSI sites in the ChosenMed targeted 599-gene NGS panel.

Furthermore, by comparing accuracies obtained at different MSI thresholds (Supplementary Table 3), we identified a stable threshold, MSIscore = 20%, to distinguish MSI versus MSS samples (Figure 3).

The LOD of MSIsensor-ct

Sequencing depth and ctDNA concentration simultaneously affect the accuracy of MSI callers. Therefore, limitation test was conducted to clarify the LOD of MSIsensor-ct.

The results demonstrated that, under the stable threshold of MSIscore = 20%, MSIsensor-ct accurately detected MSI status in cfDNA samples of 0.05% ctDNA content with sequencing depth over 3000× or ctDNA content over 0.2% with at least 1000× sequencing depth (Figure 4).

Execution time and memory usage

Practical considerations of performance, including execution time and maximum memory usage, are always relevant to software [44, 45]. Here, 39 samples were employed to evaluate the computational performance of MSIsensor-ct. On average, each sample required 16 seconds of computing time, with a maximum memory footprint of 2 Mb on a Linux machine running Centos 6.4 with Intel(R) Xeon(R) CPU E5–2680 v2 @ 2.80GHz.

Conclusion

Currently, increasing evidence supports that the MSI phenotype in ctDNA is highly consistent with that in tumor tissues. Here, we present MSIsensor-ct with 1476 site-classifiers obtained by a gradient boost machine learning protocol based on the microsatellites' allele frequency distributions in solid tumor sequencing data.

For MSI detection in cfDNA samples, MSIsensor-ct attained 100% sensitivity and specificity in 39 samples and 17 simulation datasets (#2–4, #7–9, and #11–20). Furthermore, MSIsensor-ct accurately discriminated MSI status on data with ctDNA content at the level of 0.05% and sequencing depth over 3000×, which still poses a challenge for MSI callers, in general.

Different from other MSI callers that require a baseline construction process, MSIsensor-ct only requires 10 sites in a panel that covered with the 1476 site classifiers for MSI detection with AUC attained up to 0.96. The 1476 classifiable sites provided powerful guidance for the customization of sequencing panels on MSI detection. Moreover, there is a stable MSI/MSS division threshold for MSIsensor-ct, which frees the user from massive preliminary work.

In summary, MSIsensor-ct efficiently detects MSI status in cfDNA sequencing samples, and its robustness allows its high compatibility with different sequencing panels. We believe that

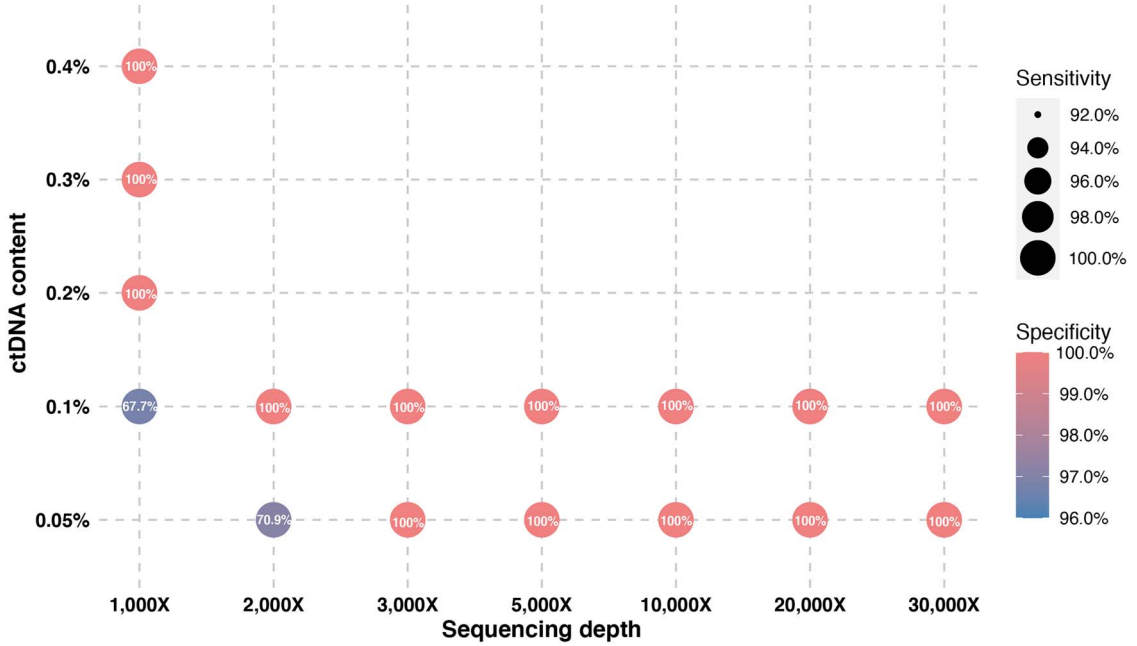

Figure 4. The LOD of MSIsensor-ct limitation. The evaluation involves 16 simulated datasets (#2, #6–20) with different sequencing depths and ctDNA content. The sensitivity and specificity were calculated at the MSIscore threshold of 20%. The sensitivity of all experiments was 100%. The white text shown inside each point refers to the exact value of specificity.

the application of MSIsensor-ct in liquid biopsies or tumor early screening associated with the MSI phenotype will yield substantial clinical benefits.

Supplementary Data

Supplementary materials are available online at *Briefings in Bioinformatics*.

Funding

This work is supported by the National Natural Science Foundation of China [grant number 31771466], the Strategic Priority Research Program of the Chinese Academy of Sciences, China [grant number XDB38040100], Special Information Program of the Chinese Academy of Sciences, China [grant number XXH13506–408].

Author contributions

B.N. and L.D. designed and supervised research. X.H. and S.Z. performed all analysis and wrote the manuscript. D.C.Z. analyzed TCGA samples. X.H. prepared figures and tables. S.Z., D.Y. and X.H. simulated cfDNA sequencing data. B.N. and X.H. contributed to MSIsensor-ct code. X.D. and D.W. conducted the biological experiment. B.N., L.D., M.C.W., R.L. and J.H. revised the manuscript.

Conflict of interest

The authors declare no competing interests.

Key Points

- MSIsensor-ct exhibits high sensitivity and specificity in MSI calling. Our results have demonstrated MSIsensor-ct's MSI calling ability with 100% sensitivity and specificity on 39 plasma cfDNA samples and 17 simulation datasets (#2–4, #7–9, and #11–20).
- MSIsensor-ct possesses extra low detection limits. According to our results, MSIsensor-ct accurately detected the MSI status for a cfDNA sequencing sample with sequencing depth over 3000× and ctDNA content at the level of 0.05%.
- MSIsensor-ct is highly compatible with panels with different microsatellites. The robustness test revealed that any random 10 microsatellites overlapped with 1476 site-classifiers, the AUC for MSI calling could be up to 0.96. In addition, no matter how many site-classifiers were covered, MSIscore = 20% is a stable threshold to distinguish MSI versus MSS samples.
- MSIsensor-ct is a user-friendly tool for installing and operating. MSIsensor-ct requires only BAM files as input and is free from additional baseline establishment, and can be flexibly integrated into the routine next generation sequencing analysis.

References

1. Kelkar YD, Strubczewski N, Hile SE, *et al.* What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at a/T and GT/AC repeats. *Genome Biol Evol* 2010;**2**:620–35.

2. Karran P. Microsatellite instability and DNA mismatch repair in human cancer. In: *Seminars in cancer biology*. Elsevier, 1996, 15–24.
3. Geiersbach KB, Samowitz WS. Microsatellite instability and colorectal cancer. *Arch Pathol Lab Med* 2011;**135**:1269–77.
4. Aaltonen LA, Peltomaki P, Leach FS, *et al*. Clues to the pathogenesis of familial colorectal cancer. *Science* 1993;**260**:812–6.
5. Whelan AJ, Babb S, Mutch DG, *et al*. MSI in endometrial carcinoma: absence of MLH1 promoter methylation is associated with increased familial risk for cancers. *Int J Cancer* 2002;**99**:697–704.
6. Bonneville R, Krook MA, Kautto EA, *et al*. Landscape of microsatellite instability across 39 cancer types. *JCO Precis Oncol* 2017;**1**:1–15.
7. Kuismanen SA, Holmberg MT, Salovaara R, *et al*. Genetic and epigenetic modification of MLH1 accounts for a major share of microsatellite-unstable colorectal cancers. *Am J Pathol* 2000;**156**:1773–9.
8. Vilar E, Gruber SB. Microsatellite instability in colorectal cancer—the stable evidence. *Nat Rev Clin Oncol* 2010;**7**:153.
9. Pino MS, Chung DC. Microsatellite instability in the management of colorectal cancer. *Expert Rev Gastroenterol Hepatol* 2011;**5**:385–99.
10. Shia J. Immunohistochemistry versus microsatellite instability testing for screening colorectal cancer patients at risk for hereditary nonpolyposis colorectal cancer syndrome: part I. *The utility of immunohistochemistry, The Journal of molecular diagnostics* 2008;**10**:293–300.
11. Dudley JC, Lin M-T, Le DT, *et al*. Microsatellite instability as a biomarker for PD-1 blockade. *Clin Cancer Res* 2016;**22**:813–20.
12. Marcus L, Lemery SJ, Keegan P, *et al*. FDA approval summary: pembrolizumab for the treatment of microsatellite instability-high solid tumors. *Clin Cancer Res* 2019;**25**:3753–8.
13. Ganesh K, Stadler ZK, Cercek A, *et al*. Immunotherapy in colorectal cancer: rationale, challenges and potential. *Nat Rev Gastroenterol Hepatol* 2019;**16**:361–75.
14. Kawai A, Healey JH, Boland PJ, *et al*. Prognostic factors for patients with sarcomas of the pelvic bones, cancer: interdisciplinary international journal of the. *American Cancer Society* 1998;**82**:851–9.
15. Umar A, Boland CR, Terdiman JP, *et al*. Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (lynch syndrome) and microsatellite instability. *J Natl Cancer Inst* 2004;**96**:261–8.
16. Hirotsu Y, Nagakubo Y, Amemiya K, *et al*. Microsatellite instability status is determined by targeted sequencing with MSIcall in 25 cancer types. *Clin Chim Acta* 2020;**502**:207–13.
17. Gullapalli RR, Desai KV, Santana-Santos L, *et al*. Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. *J Pathol Informatics* 2012;**3**:40.
18. Niu B, Ye K, Zhang Q, *et al*. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* 2013;**30**:1015–6.
19. Kautto EA, Bonneville R, Miya J, *et al*. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* 2017;**8**:7452.
20. Escudié F, Van Goethem C, Grand D, *et al*. *MIAmS: microsatellite instability detection on NGS amplicons data*. Oxford University Press, 2020;**36**:1915–6.
21. Jia P, Yang X, Guo L, *et al*. MSIsensor-pro: fast, accurate, and matched-normal-sample-free detection of microsatellite instability. *Genomics Proteomics Bioinformatics* 2020;**18**:65–71.
22. Salipante SJ, Scroggins SM, Hampel HL, *et al*. Microsatellite instability detection by next generation sequencing. *Clin Chem* 2014;**60**:1192–9.
23. Wang C, Liang C. MSIpred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. *Sci Rep* 2018;**8**:17546.
24. Schwartzberg LS, Horinouchi H, Chan D, *et al*. Liquid biopsy mutation panel for non-small cell lung cancer: analytical validation and clinical concordance. *NPJ precision oncology* 2020;**4**:1–7.
25. Luddi A, Zarovni N, Maltinti E, *et al*. Clues to non-invasive implantation window monitoring: isolation and characterisation of endometrial exosomes. *Cell* 2019;**8**:811.
26. Willis J, Lefterova MI, Artyomenko A, *et al*. Validation of microsatellite instability detection using a comprehensive plasma-based genotyping panel. *Clin Cancer Res* 2019;**25**:7035–45.
27. Cresswell GD, Nichol D, Spiteri I, *et al*. Mapping the breast cancer metastatic cascade onto ctDNA using genetic and epigenetic clonal tracking. *Nat Commun* 2020;**11**:1–12.
28. Qin Z, Ljubimov VA, Zhou C, *et al*. Cell-free circulating tumor DNA in cancer. *Chin J Cancer* 2016;**35**:1–9.
29. Oikkonen J, Hautaniemi S. Circulating tumor DNA (ctDNA) in precision oncology of ovarian cancer. *Future Medicine* 2019;**20**:1251–3.
30. Mao X, Zhang Z, Zheng X, *et al*. Capture-based targeted ultradeep sequencing in paired tissue and plasma samples demonstrates differential subclonal ctDNA-releasing capability in advanced lung cancer. *J Thorac Oncol* 2017;**12**:663–72.
31. Thierry A, El Messaoudi S, Mollevi C, *et al*. Clinical utility of circulating DNA analysis for rapid detection of actionable mutations to select metastatic colorectal patients for anti-EGFR treatment. *Ann Oncol* 2017;**28**:2149–59.
32. Mayrhofer M, De Laere B, Whitington T, *et al*. Cell-free DNA profiling of metastatic prostate cancer reveals microsatellite instability, structural rearrangements and clonal hematopoiesis. *Genome Med* 2018;**10**:1–13.
33. Deng A, Yang J, Lang J, *et al*. Monitoring microsatellite instability (MSI) in circulating tumor DNA by next-generation DNA-seq. *American Society of Clinical Oncology* 2018;**36**:12025.
34. Razavi P, Li BT, Hou C, *et al*. Cell-free DNA (cfDNA) mutations from clonal hematopoiesis: implications for interpretation of liquid biopsy tests. *American Society of Clinical Oncology* 2017;**35**:11526.
35. Hu Y, Ulrich BC, Supplee J, *et al*. False-positive plasma genotyping due to clonal hematopoiesis. *Clin Cancer Res* 2018;**24**:4437–43.
36. Razavi P, Li BT, Brown DN, *et al*. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat Med* 2019;**25**:1928–37.
37. Cai Z, Wang Z, Liu C, *et al*. Detection of microsatellite instability from circulating tumor DNA by targeted deep sequencing. *J Mol Diagn* 2020;**22**:860–70.
38. Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research (2nd ed.). *J Am Stat Assoc* 1982;**77**:946–7.
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995;**57**:289–300.
40. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
41. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

42. McKenna A, Hanna M, Banks E, *et al*. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**:1297–303.
43. Lai Z, Markovets A, Ahdesmaki M, *et al*. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 2016;**44**:e108–8.
44. Musa JD, Okumoto K. A logarithmic Poisson execution time model for software reliability measurement. In: *Proceedings of the 7th international conference on Software engineering*. 1984, p. 230–8. Citeseer.
45. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. *Bioinformatics* 2013;**29**: 652–3.