The Journal of Molecular Diagnostics, Vol. 27, No. 1, January 2025

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

A Systematic Method to Detect Next-Generation Sequencing—Based Microsatellite Instability in Plasma Cell-Free DNA

plasmaMSI

Fengchang Huang,*† Lili Zhao,‡ Hongyu Xie,‡ Tiancheng Han,‡ Jian Huang,† Xiaoqing Wang,‡ Jun Yang,† Yuanyuan Hong,‡ Jingchao Shu,‡ Jianing Yu,‡ Qingyun Li,‡ Ji He,‡ Weizhi Chen,‡ Yu S. Huang,‡ and Wenliang Li*§

From the Kunming Medical University, Kunming; the Department of Surgical Oncology,† The First Affiliated Hospital of Kunming Medical University, Kunming; Genecast Biotechnology Co., Ltd.,‡ Wuxi; and the Department of Colorectal Surgery,§ The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming, China*

Accepted for publication October 10, 2024.

Address correspondence to Wenliang Li, M.D., Kunming Medical University, 519 Kunzhou Rd., Kunming, Yunnan 650032, China.
E-mail: liwenliang@kmmu.edu.cn.

Microsatellite instability (MSI) detection using tumor tissue is a well-established prognostic and predictive biomarker for certain types of cancers. However, tumor tissue samples are less convenient to obtain than blood plasma samples. The main challenge facing next-generation sequencing—based MSI detection in blood plasma samples is the ultralow signal/noise ratio in plasma cell-free DNA (cfDNA). To address the challenge, plasmaMSI, a highly accurate cfDNA MSI detection method, is introduced with three novel performance-improving features: i) a set of stringent locus selection criteria to select loci with high robustness and compatibility across sequencing platforms; ii) a new deduplication strategy that greatly improves the signal/noise ratio for MSI detection; and iii) an MSI calling algorithm that customizes the baseline for each test sample based on its duplication rate. Through analytical validation in diluted cell line samples, the limit of detection of plasmaMSI was determined to be 0.15%. Furthermore, in analyzing 95 evaluable cfDNA samples from patients with gastrointestinal cancers, plasmaMSI exhibited a positive percentage agreement of 92.9% (39/42) and a negative percentage agreement of 100% (53/53) with tissue MSI-PCR. plasmaMSI provides novel solutions to key challenges in cfDNA MSI detection that have not been addressed by existing methods. It has also been systematically validated and is already used in clinical testing for patients with cancer. *(J Mol Diagn 2025, 27: 62—73; https://doi.org/10.1016/j.jmoldx.2024.10.002)*

Microsatellite instability (MSI), a hypermutable phenotype caused by DNA mismatch repair deficiency, is an established National Comprehensive Cancer Network—recommended biomarker for the diagnosis and prognosis of patients with colorectal cancer[1,2] and other solid tumors.[3] However, tumor tissue samples are less convenient to obtain than blood plasma samples in a clinical setting. Noninvasive cell-free DNA (cfDNA) assays have been developed to help expand the clinical reach of MSI analysis. Recently, several cfDNA-based MSI detection methods have been applied to predict the patient response to immune checkpoint inhibitors.[4,5] Compared with electrophoresis-based MSI-PCR, which evaluates a few microsatellite loci, next-generation sequencing (NGS)—based MSI assays can assess many more microsatellite loci, boosting the statistical power for accurate MSI detection, especially in samples with low cancer cell fractions. Compared with immunohistochemistry (IHC), which assesses the expression of mismatch repair proteins,[6] NGS-based MSI assay can be part of a single mutation-checking

Supported in part by National Natural Science Foundation of China grant 31660312 (W.L.).

F.H. and L.Z. contributed equally to this work.

https://doi.org/10.1016/j.jmoldx.2024.10.002

NGS test, and it handles both tumor tissue and blood plasma samples.

The MSI calling algorithms used by NGS-based tissue MSI assays, which are broadly based on comparing the repeat lengths of different MSI alleles between a test sample and a normal sample (or a baseline of normal samples),[7−10] have been extended to NGS-based plasma MSI assays. But the low fraction of cell-free tumor-origin DNA fragments in plasma, typically three to four orders of magnitudes lower than that of tumor DNA in tissue,[5] is a significant challenge to overcome. It requires not only improvement in experimental techniques, such as using unique molecular identifiers to suppress errors and increasing the sequencing depth, but also a more sophisticated algorithm to discriminate true MSI signals from noises (aka contracting/shorter and expanding/longer alleles), which are mostly errors introduced during PCR amplification and sequencing.

Many existing NGS-based plasma cfDNA MSI detection methods[5,11−13] use a baseline-dependent statistical framework for assigning a probability to both locus- and sample-level MSI status. bMSISEA[11] used 100 white blood cells (WBCs) to establish a microsatellite stability (MSS) baseline for each locus. Willis et al[5] (2019) used a baseline of healthy donor samples to determine the locus-level threshold. Wang et al[13] (2020) used MSS blood samples as the baseline in the locus selection step. Then, these methods typically combine information of all loci and derive a threshold based on the fraction of unstable loci to call the test sample MSI or MSS. plasmaMSI uses a similar approach.

However, existing methods fail to recognize two issues that affect MSI calling significantly: the large impact of the duplication ratio on the noise level (ie, fraction of contracting and expanding alleles, conceptually similar to stutter ratio in MSI-PCR) and the importance of matching the duplication ratio between the test sample and the baseline; and the adverse impact of canonical greatest sum of sequenced base qualities (GSBQ) deduplication strategy, designed for detecting SNV/insertion/deletion variants, on detecting the true microsatellite (MS) allele signal. To address these issues, plasmaMSI uses a suite of locus selection rules, which can be applied to any panel to yield mononucleotide loci of length 10 to 13 bp and can work with mainstream sequencing platforms (validated on four sequencers from Illumina, San Diego, CA; and BGISEQ, Shenzhen, China). Second, plasmaMSI adopts a new MSI-tailored majority-voting deduplication strategy that results in higher signal/noise ratio in uncovering the true MS allele than the canonical deduplication strategy used by tools such as Picard-MarkDuplicates. Note that duplicate ratio is unique to NGS and has no equivalence in MSI-PCR as PCR cannot distinguish different duplicate families. Third, plasmaMSI uses Kullback-Leibler divergence (KLD) to better measure the divergence from a duplication-level−matching

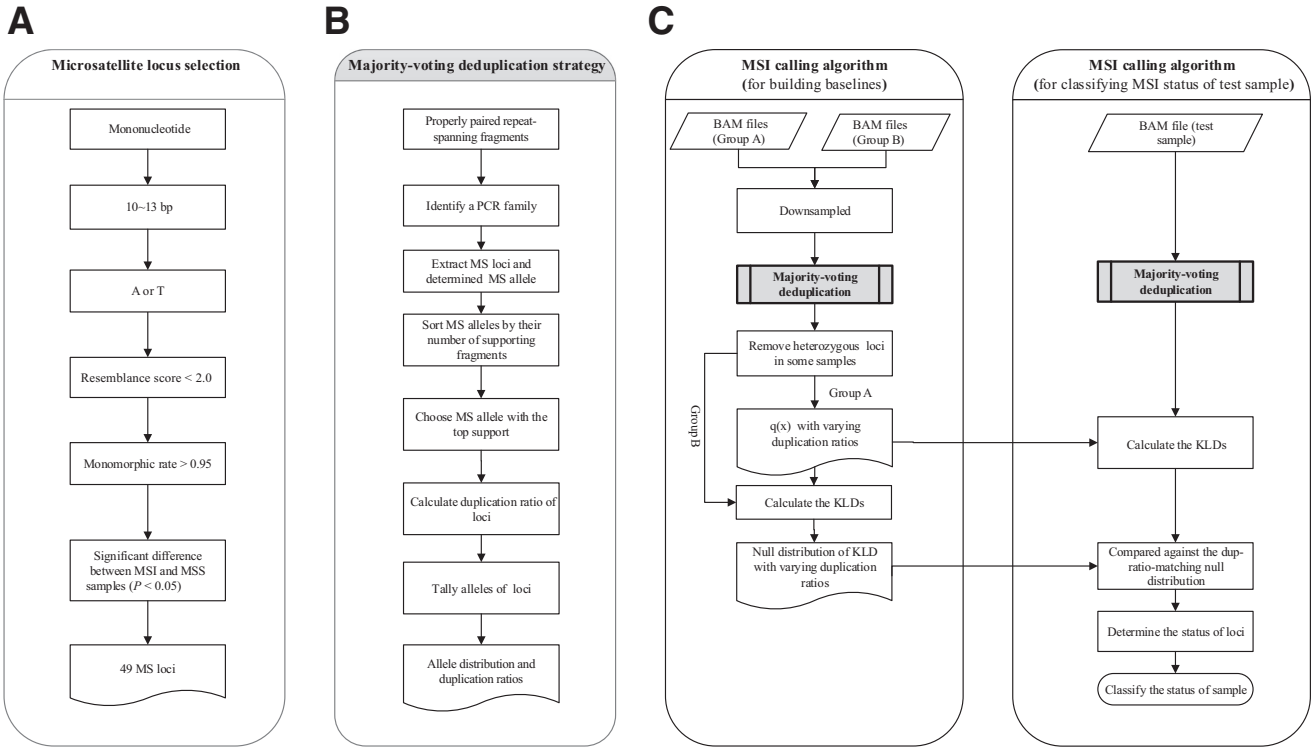

Figure 1 Key components of plasmaMSI: microsatellite (MS) locus selection (**A**), the novel majority-voting deduplication (**B**), and the MS instability (MSI) calling algorithm for baseline construction and MSI calling for a test sample (**C**). The **gray shaded boxed areas** represent the majority-voting deduplication. Dup ratio, duplication ratio; KLD, Kullback-Leibler divergence; MSS, MS stability.

baseline/healthy distribution (Figure 1). Through benchmark analyses and clinical validation, plasmaMSI is shown to be an accurate and robust method for plasma cfDNA MSI detection.

Materials and Methods

Study Design and Sample Information

A total of 259 patients with cancer and 108 healthy donors were recruited from March 2018 to June 2021, and their specimens were shipped to and processed in the Genecast Biotechnology laboratory (Wuxi, China) (Table 1 and Supplemental Table S1). Samples were prepared and processed following an established protocol.[14] Specifically, target enrichment was conducted with a pan-cancer panel (543 genes, 2.09 Mbp). cfDNA and WBCs were obtained from 173 to 198 individuals, respectively. Tumor tissue specimens were available for 149 patients. A training set of 148 peripheral WBC samples and 40 tumor tissue samples (20 positive and 20 negative by MSI-PCR) were used exclusively for microsatellite locus selection. A total of 70 cfDNA samples were used to assess the locus-level baseline distribution. Four previously established MSI cell lines (22Rv1, LNCAP, RL952, and DLD-1; COBIOER Inc., Nanjing, China) were used to assess the analytical accuracy. For clinical validation, plasma cfDNA samples from 103 patients with gastrointestinal cancer were used. Tumor tissue specimens for all 103 patients were tested by either MSI-PCR or IHC. The tissue, blood cell, and cfDNA samples of 50 positive patients (4 with stage I, 11 with stage II, 29 with stage III, 4 with stage IV, and 2 not available) were sequenced and analyzed via an NGS bioinformatics pipeline (see *Bioinformatics Pipeline* for details) to call point mutations and produce the maximum variant allele frequency (max-VAF). To investigate how the sequencing coverage varies between sequencing platforms in response to the varying repeat lengths, an independent set of six tumor tissue samples was used.

Table 1 Characteristics of Clinical Samples

Characteristic	Patients with MSI (n = 50)	Patients with MSS (n = 53)
Age, mean (range), years	50 (23–86)	53 (27–88)
<40	6 (12.00)	3 (5.66)
40–49	10 (20.00)	4 (7.54)
50–59	12 (24.00)	11 (20.75)
≥60	22 (44.00)	32 (60.37)
NA	0 (0.00)	3 (5.66)
Sex		
Female	24 (48.00)	18 (33.96)
Male	26 (52.00)	32 (60.37)
NA	0 (0.00)	3 (5.66)
Cancer type		
Colon cancer	40 (80.00)	25 (47.16)
Rectal cancer	4 (8.00)	16 (30.18)
Gastric cancer	6 (12.00)	9 (16.98)
NA	0 (0.00)	3 (5.66)
Stage of disease		
I	4 (8.00)	4 (7.54)
II	11 (22.00)	17 (32.07)
III	29 (58.00)	17 (32.07)
IV	4 (8.00)	12 (22.64)
NA	2 (4.00)	3 (5.66)

Data are given as number (percentage) of patients unless otherwise indicated.

MSI, microsatellite instability; MSS, microsatellite stability; NA, not available.

Sample Processing

MSI-PCR was performed using the MSI Analysis System (Promega, Madison, WI). For IHC, the antibodies used were postmeiotic segregation increased 2 (PMS2), MutL protein homolog 1 (MLH1), MutS homolog 6 (MSH6), and MutS homolog 2 (MSH2) (ZSGB-BIO, Beijing, China), and Autostainer Link 48 (Agilent, Santa Clara, CA) was used for staining. For the cell line experiments, a fixed amount of genomic DNA (35 ng) was sonicated and then titrated with NA12878 (COBIOER Inc.) to yield five concentrations (0.13%, 0.25%, 0.5%, 1.0%, and 2.0%), each with five technical replicates. cfDNA was extracted from plasma samples with MagMAX (Thermo Fisher, Waltham, MA) per the manufacturer's instructions. Target enrichment was performed with the HyperCap Target Enrichment Kit (Roche, Basel, Switzerland) following the manufacturer's protocol. A fixed mass of products from eight cycles of precapture PCR was fed into 14 cycles of post-capture PCR. The locus selection step was performed on the basis of a 2.09-Mbp, 543-gene pan-cancer panel using tissue specimens sequenced to an average depth of 871. Afterwards, a 100-Kbp gastrointestinal cancer-specific panel was designed, which included the selected MSI loci, and used to enrich all the genomic DNA and cfDNA samples in this study. The resulting libraries were sequenced to an average depth of 25,144 (95% CI, 5890–46,069), corresponding to an average duplication ratio (dup ratio) of 0.84 (95% CI, 0.56–0.93). In addition to locus selection, the 2.09-Mbp panel was used for the targeted enrichment of tissue samples involved in the estimation of tumor content. All libraries were sequenced in 2 × 100-bp paired-end mode on MGI-T7 (BGISEQ). The six tumor tissue samples used for comparison among sequencers were prepared and sequenced by four sequencers: three samples sequenced by NovaSeq6000/NextSeq500 sequencers (Illumina Inc.), and the other three sequenced by MGI-T7 and MGI-2000 sequencers (BGISEQ) with identical experimental conditions.

Bioinformatics Pipeline

For both plasma cfDNA and tissue genomic DNA samples, adaptors were trimmed from raw read pairs using

Trimmomatic version 0.36.[15] Clean reads were mapped against the human reference genome (build hg19; University of California, Santa Cruz), aligned using Burrows-Wheeler Aligner bwa-mem version 0.7.12[16] and sorted using SAMtools version 1.7.[17] A detailed comparison of the majority-voting–based deduplication strategy and the canonical strategy used by Picard MarkDuplicates version 2.1.0 (*https://broadinstitute.github.io/picard*) is described in *The Canonical GSBQ Deduplication Strategy versus MSI-Tailored Majority-Voting Deduplication Strategy*. MarkDuplicates was performed in the default mode, followed by local insertion/deletion realignment using GATK version v3.7 (*https://gatk.broadinstitute.org*). For each microsatellite locus, all fragments spanning the locus, defined as fragments that fully cover the microsatellite plus 2 bp in both 5′ and 3′ directions, were extracted from the realigned Binary Alignment Map file. Following deduplication, the repeat lengths of each locus were assessed and tallied for KLD calculation. Twice the max-VAF of tumor-informed SNVs called in plasma cfDNA was used to approximate the circulating tumor DNA content in the clinical samples. The calling methods and filtering criteria have been described previously.[14] If no variant is called or the max-VAF is below the limit of detection (refer to *Results* for details), the circulating tumor DNA content was considered too low, and the sample was regarded as unevaluable.

plasmaMSI was benchmarked against MSIsensor-ct (latest available version, version 1.0, run according to its instructions).[12] Specifically, models for all 12 loci in the tested panel were used for MSI status prediction on clinical samples.

Microsatellite Locus Selection

The following are the steps to select high-performance microsatellite loci (Figure 1A). i) Only mononucleotide *A* or *T* repeats of length between 10 and 13 bp are selected. Longer repeats are less compatible with different sequencers (Supplemental Figure S1). ii) The resemblance score (RS) of neighboring nucleotides to the selected MS locus must be <2.0. RS measures the similarity of neighboring nucleotides to the MS locus nucleotide:

$$RS = \sum_{d=S}^{E}\left(1 - |d|\frac{1}{10}I(nt_d = nt_{ms})\right) \quad (1)$$

$$RS_{up} = RS(S = -10, E = -1) \quad (2)$$

$$RS_{down} = RS(S = 1, E = 10) \quad (3)$$

where nt_{ms} the repeat unit nucleotide of the mononucleotide MS locus. nt_d is the neighboring nucleotide with distance d from the MS locus under consideration. A positive (or negative) d indicates the nucleotide being downstream (or upstream) of the locus. $I(nt_d = nt_{ms})$, an indicator function, takes value 1 when nt_d is equal to nt_{ms} and value 0 when they are not equal. RS is essentially an MS locus quality measure. A high RS score is a strong indicator of poor alignment in the vicinity of an MSI locus, which renders the derived MS locus length distribution highly inaccurate. plasmaMSI only includes MS loci with both RS_{up} and RS_{down} <2.0. iii) Only monomorphic MS loci in healthy samples are selected. A locus is defined as monomorphic if >95% of 118 WBC samples have a central/signal ratio >0.6. iv) Only loci that can distinguish between MSI and MSS samples are selected. Wilcoxon rank sum test is applied for each MS locus comparing its KLD (details in *MSI Calling Algorithm*) values in 20 MSI versus 20 MSS tissue samples, and loci with $P < 0.05$ are selected.

MSI-Tailored Majority-Voting Deduplication Strategy

The MSI-tailored deduplication strategy includes the following steps (Figure 1B). i) Only properly paired fragments covering one MS locus (so-called locus-spanning fragments/reads) are retained. ii) A PCR family consists of fragments with the identical start and end positions. iii) On the basis of its repeat length, each fragment within the PCR family is classified to be the reference allele, the shorter/contracting allele, or the longer/expanding allele. iv) Rank all MS alleles within the PCR family by the number of supporting fragments. v) The MS allele with the top support is selected to represent the MS allele of the entire PCR family. If more than one MS alleles have the identical top support, the whole PCR family is discarded.

For one MS locus, its dup ratio is calculated as follows:

$$dup\_ratio = \frac{\sum_{i=1}^{N}(S_i - 1)}{\sum_{i=1}^{N} S_i} \quad (4)$$

where N is the number of duplicate families covering the MS locus and S_i is the size of family i.

MSI Calling Algorithm

Following deduplication, the MSI status for each locus is determined via a two-stage statistical test by first deriving a KLD with respect to a dup ratio–matching baseline and then producing a P value by comparing the KLD against its dup ratio–matching null distribution (Figure 1C). KLD measures the loss of information when a proxy distribution $q(x)$ is used to approximate $p(x)$. The higher the KLD, the more divergent the proxy distribution $q(x)$ is from $p(x)$.[18] In the case of plasmaMSI, $p(x)$ is the allele frequency distribution of a test sample and $q(x)$ is that of the baseline. The KLD is calculated as such:

$$D_{KL}(p|q) = \sum_{x} p(x) \cdot log\frac{p(x)}{q(x)} \quad (5)$$

where $p(x)$ and $q(x)$ are the frequencies of a given allele x.

The $q(x)$ for each locus is based on a group of 70 MSS samples. Per observation, as the dup ratio increases, the central/signal ratio and noise ratio of the baseline increases and decreases, respectively (Supplemental Figure S2), which suggests the dup ratio has a strong impact on the allele length distribution of an MS locus. Hence, a series of baselines with varying duplication ratios are built. First, 70 samples are downsampled to a grid of dup ratios from 0.1 to 0.99 with a step size of 0.01, randomly divided into two equivalent groups of 35 samples, group A and B. Let us focus on one MS locus L. At each dup ratio, the allele frequency of an observed allele of locus L is averaged across all samples in group A. Arranging the allele frequencies of all alleles of locus L into a vector ordered by the allele length produces $q(x)$ for locus L. In group B, the allele length distribution for locus L in a sample is regarded as $p(x)$ and its KLD is derived with respect to $q(x)$. The valid KLD values of locus L from 35 samples in group B are fitted by a γ distribution, which is considered as the KLD null distribution for locus L and is the basis for producing the P value for any MS locus in a test sample.

For a locus in a test sample, the dup ratio is first calculated. Second, the corresponding dup ratio−matching $q(x)$ is used to calculate its KLD score, and its P value is calculated against the dup ratio−matching null distribution. If the P value is < 0.005, the locus is considered unstable. Last, if the number of valid loci of the sample is ≥ 15 and the fraction of unstable loci (MSI score) is above a clinically determined threshold (0.13 in this clinical study), the sample is called MSI.

In summary, the locus selection rules, the majority-voting deduplication strategy, and the KLD method with dup ratio−matching baselines are three key components of plasmaMSI.

Ethical Approval and Consent to Participate

This study was approved by the First Affiliated Hospital of Kunming Medical University Ethics Committee (2017-L3). The informed consent was obtained from the study participants or their representatives before enrollment. All experimental methods were performed in accordance with relevant guidelines and regulations, such as the Declaration of Helsinki.

Availability of Software, Data, and Materials

plasmaMSI source code is freely available (*https://github.com/zhaolili-genecast/plasmaMSI*, last accessed July 4, 2024). The raw sequence data reported in this article have been deposited in the Genome Sequence Archive[19] at the National Genomics Data Center,[20] China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number HRA001849 (*https://ngdc.cncb.ac.cn/gsa-human/browse/HRA001849*, last accessed May 23, 2024).

Results

Selection of Microsatellite Loci in a 2.09-Mbp Panel

Locus selection rules aim to improve the overall accuracy in two aspects: the discriminant power of an MS locus to distinguish MSI samples from MSS samples; and how well an MS locus performs across different experimental (PCR) procedures and different sequencers. The locus selection rules are applied to an existing 2.09-Mbp panel.

Regarding the first aspect, mononucleotide repeats are more affected by somatic mutations of mismatch repair deficiency than dinucleotides or longer repeat units. Hence, mononucleotide repeats are a better indicator of mismatch repair deficiency than nonmononucleotide ones. MS loci of shorter repeats are also less likely to manifest instability than loci of longer repeats,[9,21–23] but using more mononucleotide-repeat loci can compensate for the lower sensitivity of individual loci (Supplemental Figure S3). To further improve the MSI detection power, loci that show statistically significant KLD ($P < 0.05$, rank sum test) among the 20 MSI and 20 MSS samples are selected.

Regarding the second aspect, *A*/*T* repeats are known to be less mutable during PCR than *C*/*G* repeats.[24] Furthermore, experiments were conducted to explore the possible drawbacks associated with using longer repeats. The mononucleotide repeat length was found to be positively correlated with the noise ratio, defined as the sum of the ratios of contracting and expanding alleles to all reads (Figure 2). The noise ratio is conceptually similar to stutter ratio in MSI-PCR. Long simple repeats are also known to be followed by bases of lower base quality and thus lower coverage.[25,26] To evaluate how stable sequencing coverage of MS loci is across different sequencers, and how it affects MSI detection, an independent set of six MSS tissue samples were run on four sequencers from two companies (MGI-T7 and MGI-2000 from BGISEQ, and NovaSeq6000 and NextSeq500 from Illumina). In particular, for NextSeq500, once the repeat length was beyond 15 bp, the sequencing coverage of the locus can drop to an unacceptably low level (Supplemental Figure S1). The poor showing of NextSeq500 is largely attributed to its dual-color fluorescence mechanism.

Considering the aforementioned factors, a total of 49 mononucleotide (A or T) loci with length from 10 to 13 bp were chosen and used in subsequent stages (Supplemental Table S2).

The Canonical GSBQ Deduplication Strategy versus MSI-Tailored Majority-Voting Deduplication Strategy

Deduplication is a process that selects one fragment among all duplicates that can best represent the original fragment. The canonical GSBQ deduplication strategy used by virtually all deduplication tools, including Picard MarkDuplicates, ideal for SNV/insertion/deletion calling, is

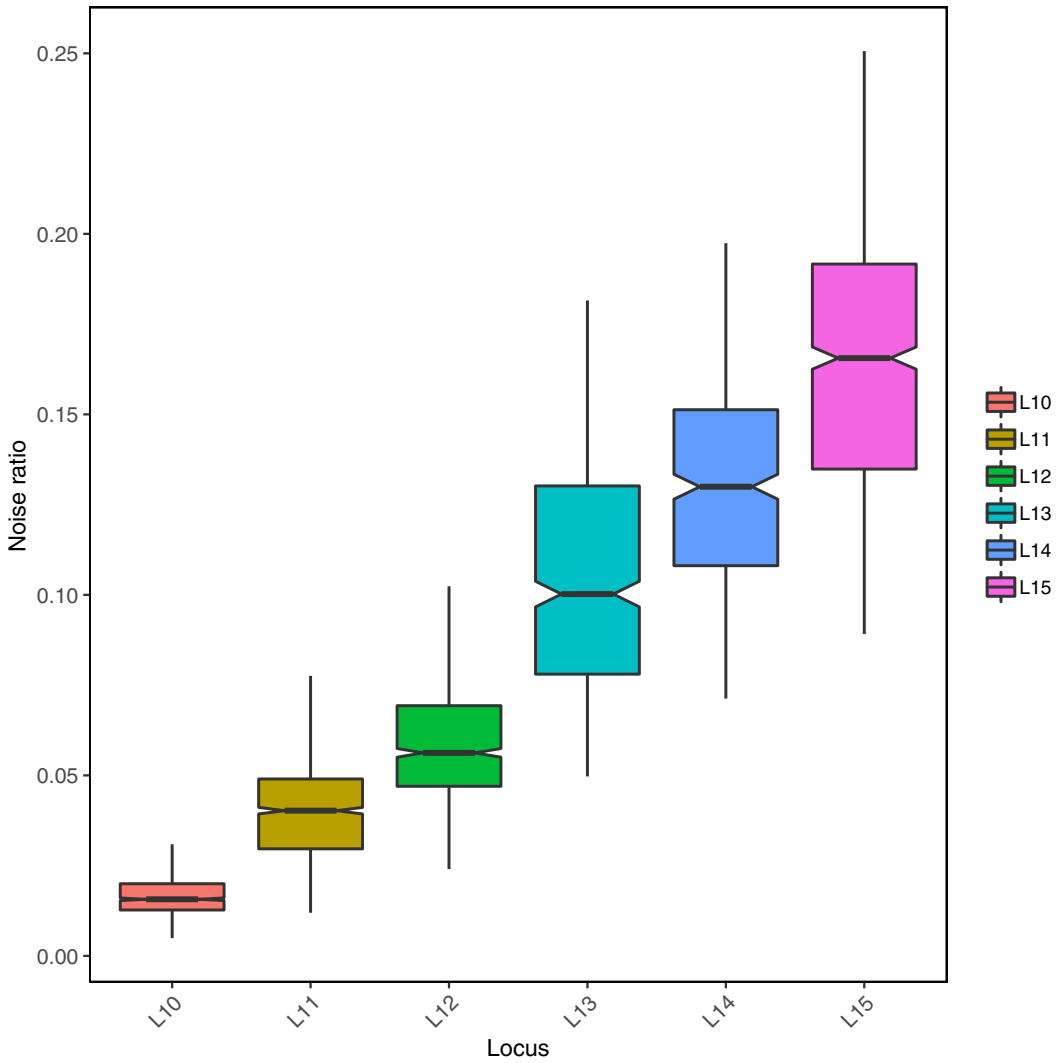

Figure 2 The ratio of the sum of contracting and expanding alleles, a proxy for noise ratio, positively correlates with the mononucleotide repeat length. Each box plot shows the noise ratio (ratio of the sum of contracting and expanding alleles, equivalent to stutter ratio in microsatellite instability PCR) calculated at six mononucleotide loci (10 to 15 bp), each with 118 points representing white blood cell samples from patients with solid tumors. The within-group median is indicated by the middle bar in a box. The whiskers correspond to the 1.5 interquartile range.

found to be unsuitable for MSI calling. Namely, the GSBQ strategy selects the fragment with the greatest sum of sequenced base qualities among all duplicates. Different from SNV/insertion/deletion loci, an MS locus is a long stretch of repeats (homopolymers in this study), which are unfriendly for NGS sequencers.[18] For NGS sequencers, the ability to call bases correctly decreases sharply for bases after an MS locus, which are reflected in the poor quality scores for these bases.[25,26] In addition, the longer the repeat is, the more the quality score decreases. Thus, among all duplicates that span an MS locus, the shorter the repeat within one duplicate fragment is, the higher its sum of base qualities is. This phenomenon results in the GSBQ strategy preferentially selecting the fragment with a shorter repeat (aka the contracting allele) to represent the entire duplicate family. Processing over all duplicate families covering one MS locus, the GSBQ strategy tends to produce an allele spectrum that contains more contracting alleles.

Given that GSBQ would result in a biased allele spectrum, a new majority-voting deduplication strategy is generated to mitigate the bias for better MSI detection. The majority-voting deduplication strategy is based on the observation that the original MS allele (aka signal) remains the most frequent allele among all duplicates and shall be easier to be identified as duplicate ratio (sequencing depth) increases (Figure 3). Figure 3 compares allele spectrum of 7 MS loci (7 columns) of a healthy sample (GCS209) with deduplication-free, GSBQ, and majority-voting deduplication strategies applied. The allele spectrum of each locus is represented by three ratios: the signal/central/reference ratio (=fraction of reads supporting the original MS allele), the contraction ratio (=fraction of reads supporting contracting/shorter alleles), and the expansion ratio (=fraction of reads

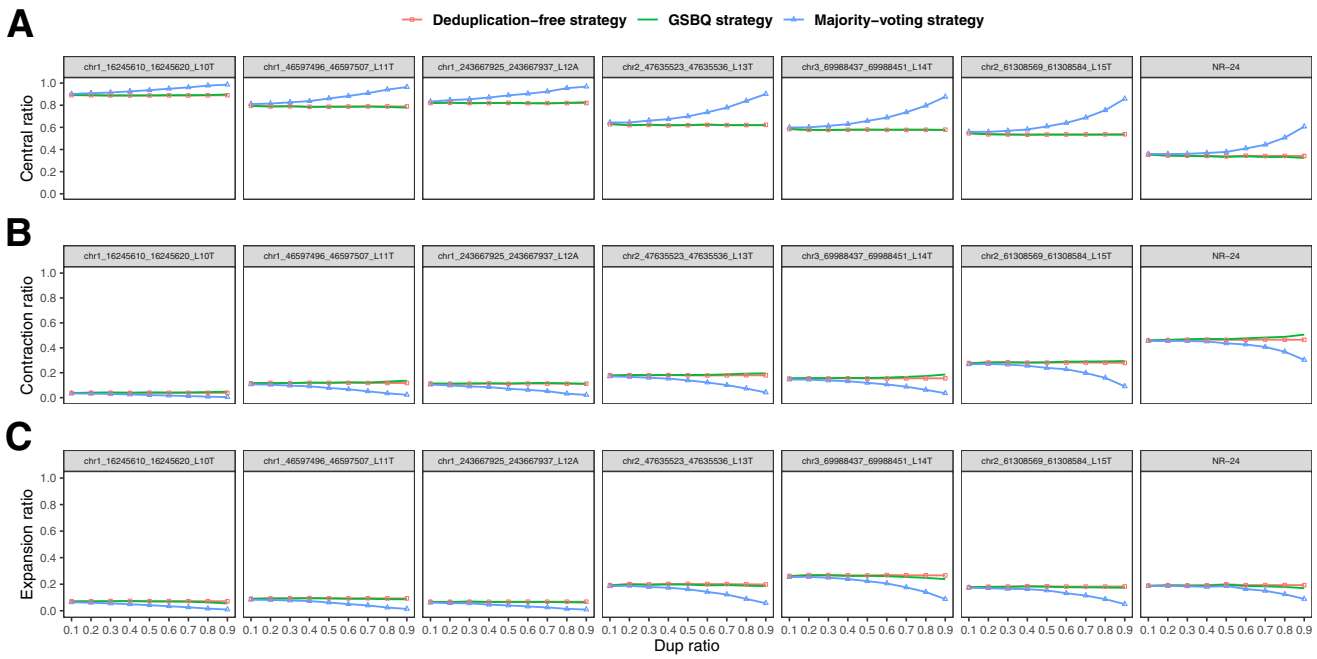

Figure 3 The relationship between the central/signal ratio (**A**), the contraction ratio (**B**), and the expansion ratio (**C**) versus the duplication ratio (Dup ratio) with three deduplication strategies: greater sum of base quality (GSBQ) by Picard MarkDuplicates (green), deduplication free (red), and majority voting (blue). The deduplication-free strategy means no deduplication is performed. A range of lower dup ratio samples were generated by downsampling one high-coverage high dup ratio plasma sample. Chr, chromosome.

supporting expanding/longer alleles). For the deduplication-free strategy, all three ratios remain flat as the dup ratio increases. The canonical GSBQ strategy results in a lower central ratio and a higher contraction ratio as the dup ratio increases (Figure 3), which means it decreases the signal/noise ratio as more data are becoming available. Finally, for the majority-voting strategy, as the dup ratio increases, the noise ratios (the contraction ratio and the expansion ratio) decrease and the central/signal ratio increases, which means this strategy can boost the signal/noise ratio. This result of majority-voting strategy makes sense because a higher dup ratio implies more repeat sequencing data per locus, and logically it should boost the signal for the authentic allele against noises. In conclusion, the majority-voting strategy is better than the canonical GSBQ strategy because the former can increase the signal/noise ratio for MSI detection as more data become available.

Analytical and Clinical Validation of plasmaMSI

To analytically assess the performance of plasmaMSI, 70 cfDNA samples from healthy donors were tested. The maximum MSI score of 0.13 from this healthy cohort is used as the threshold in subsequent analyses (Figure 4). Analytical sensitivity was then evaluated using four established cell lines diluted to five concentrations, ranging from 0.13% to 2%, with 20 data points for each titration level (19 for 0.13% because of technical failure in one experiment). The sensitivity was 79% (15/19) at a minimum of 0.13% and 100% for higher concentrations. The 95% limit of detection was determined to be 0.15% through probit regression (Figure 5).

The clinical accuracy of plasmaMSI was evaluated by comparing its classification results of 103 cfDNA gastrointestinal cancer patient samples with the results of tissue MSI-PCR or IHC (Figure 6A). A mixed set of tissue MSI-PCR or IHC is not a perfect benchmark because there is slight inconsistency among the two. It was used because of facility constraints at different hospitals. For seven of all samples (7/103, 6.8%), no somatic variant with VAF $\geq$0.2% was detected and, thus, their MSI state was considered unevaluable.[5] One sample was excluded because of the lack of circulating tumor DNA content information. Among 42 evaluable MSI samples and 53 MSS samples (95/103), plasmaMSI exhibited a positive percentage agreement (PPA) of 92.9% (39/42; 95% CI, 79.4%−98.1%) and negative percentage agreement of 100% (53/53; 95% CI, 92%−100%) with tissue MSI-PCR, with an overall percentage agreement of 96.8% (92/95; 95% CI, 90.3%−99.2%) and a PPA of 100% (39/39; 95% CI, 88.9%−100%) for patients across all clinical stages. The performance does vary among different cancer types and stages, although PPA estimates for some cancer type or stage have a large 95% CI because of the low number of samples (Supplemental Table S3). Moreover, three of the seven unevaluable samples can be correctly called by plasmaMSI (data not shown). In

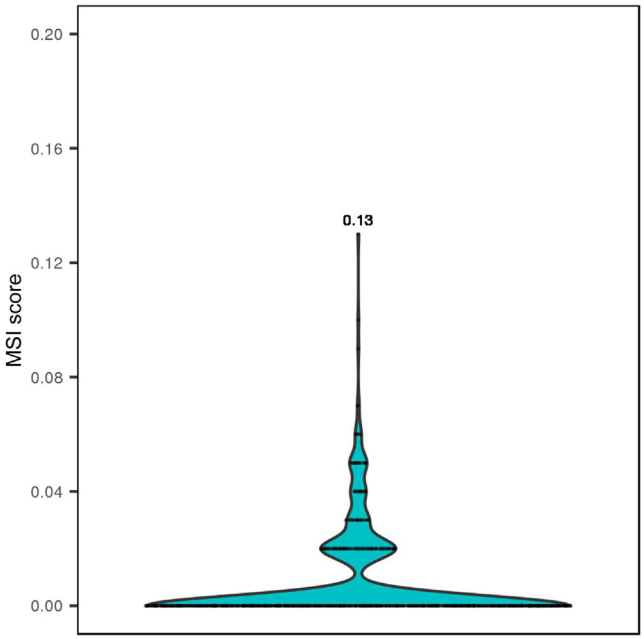

Figure 4 Violin plot showing the microsatellite instability (MSI) score distribution for 70 cell-free DNA samples from healthy donors. The maximum MSI score, 0.13, is used as the threshold to call MSI (>0.13) or microsatellite stability.

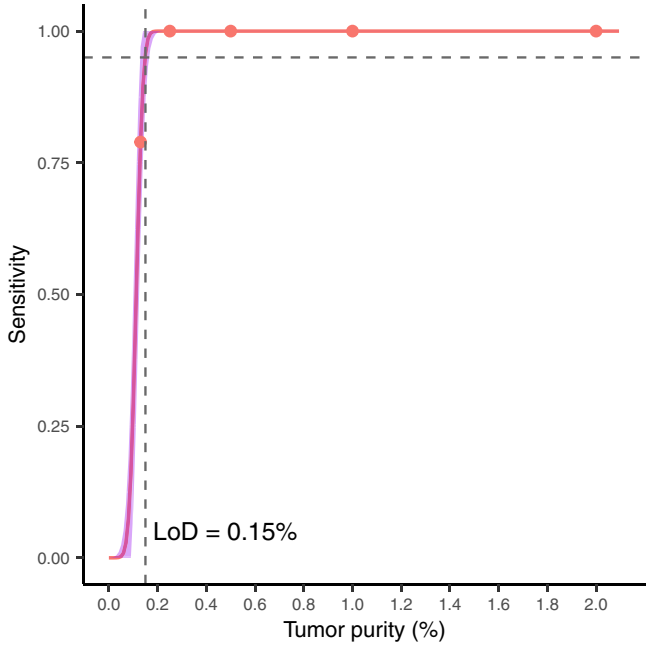

Figure 5 The analytical sensitivity of plasmaMSI versus the tumor purity. Probit regression was performed on sensitivities calculated on four diluted cell line samples (22RV1, DLD, LNCAP, and RL952) at five circulating tumor DNA concentrations (0.13%, 0.25%, 0.5%, 1.0%, and 2.0%, each with four to five replicates); the 95% CI is shown by purple shading. The **dashed line** indicates the 95% limit of detection (LoD) that corresponds to 0.15% tumor purity.

addition, the concordance of plasmaMSI with tissue MSI results for 49 (42 evaluable + 7 unevaluable) clinical cfDNA samples of various max-VAF values was shown in Figure 6B. The vertical dashed line at 0.075% [log (max-VAF) approximately −3.1] indicates the threshold for evaluable samples, and the horizontal dashed line at 0.13 is the MSI-score threshold. Together, these results demonstrate the excellent performance of plasmaMSI and its potential for diagnosing MSI cfDNA samples.

plasmaMSI was benchmarked against MSIsensor-ct, one of the mainstream cfDNA-based MSI detection methods. MSIsensor-ct has a predefined set of loci, in which only 12 of them exist in the tested panel. To have a fair comparison, plasmaMSI was restricted to use only 12 randomly sampled loci (Supplemental Table S4). For 50 MSI tissue samples, plasmaMSI correctly predicted 84.0% (42/50), whereas MSIsensor-ct correctly predicted 16.0% (8/50) (Figure 7A). For 53 MSS tissue samples, both plasmaMSI and MSIsensor-ct correctly predicted all patient-matching cfDNA samples as MSS (Figure 7B). Overall, MSIsensor-ct showed a PPA of 16.0% (8/50; 95% CI, 7.6%−29.7%) and negative percentage agreement of 100% (53/53; 95% CI, 91.6%−100%).

In addition, PPAs for the same cohort using fewer MS loci have been investigated (Supplemental Table S5). The result showed that the minimum number of MS loci to achieve similar performance as 49 locus is approximately 18 (Supplemental Figure S3). However, many samples were excluded because of inadequate number of quality control−passing MS loci, especially when the number of MS loci falls below 30 for this small cohort (approximately 100 samples). In real clinical settings, it is recommended to use the 49-locus panel to cover as many patients as possible.

Discussion

In plasmaMSI, the proposed locus selection criteria were designed to achieve both high accuracy and wide compatibility among sequencers. The majority-voting deduplication strategy greatly improves the signal/noise ratio in detecting MSI than the canonical GSBQ strategy used by Picard MarkDuplicates. The advantage of combining a dup ratio−matching baseline with a full-spectrum divergence metric (KLD) is also demonstrated. All three features contribute to the overall performance of plasmaMSI.

Some of the locus selection rules were not previously considered or explicitly formulated. First, in the absence of a patient-matching normal sample, polymorphic loci should first be screened for specificity,[6,27,28] and ideally, samples from the same population should be used.[29] In this study, a set of WBC samples from the Chinese population was used to select monomorphic loci. Note that if patient-matching normal samples are available, they can also be used. Second, the compatibility of MSI loci across sequencing platforms has not been assessed by existing methods. In general, loci shorter than 15 bp are likely to be compatible with

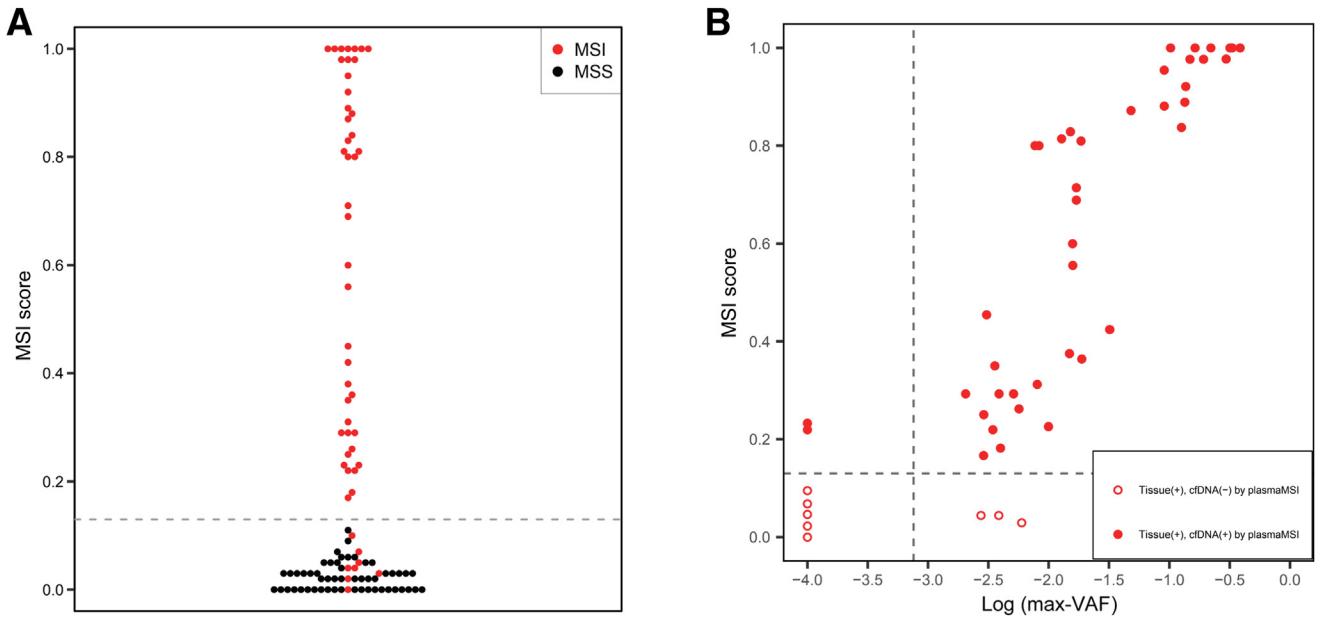

Figure 6 **A:** The distribution of microsatellite instability (MSI) score by plasmaMSI for 103 clinical cell-free DNA (cfDNA) samples in a beeswarm plot. The **dashed line** is the MSI-score threshold of 0.13. **B:** Concordance between cfDNA plasmaMSI results and their corresponding tissue MSI-PCR results for 49 MSI clinical patients at different maximum variant allele frequencies (max-VAFs). Solid and open circles indicate the MSI and microsatellite stability (MSS) status, respectively, called by plasmaMSI. The **horizontal dashed line** is the MSI score threshold of 0.13. The **vertical dashed line** at 0.075% [log (max-VAF) approximately −3.1] indicates the threshold for evaluable samples.

mainstream NGS platforms (Supplemental Figure S1). The thresholds of loci selection rules can be customized for specific applications. For example, choosing an RS (details in *Microsatellite Locus Selection*) threshold >2.0 would result in a greater number of loci that are less specific, and vice versa. Furthermore, a larger panel would facilitate a more stringent threshold. Finally, including more discriminant loci should potentially increase sensitivity, although this should be performed after careful cohort analysis because prior literature has indicated performance can plateau after a certain number of loci or even decrease in certain cancer types.[30–32] Including more diverse samples to select discriminant loci will also likely further improve the performance.

Regarding the deduplication strategy, existing NGS-based MSI detection methods have either explicitly stated that they use the canonical GSBQ strategy by Picard to deduplicate[4,33] or neglect to mention whether or how they deduplicate sequencing data.[8,12,34] plasmaMSI is the first method addressing the signal dampening issue by GSBQ for MSI detection. Deduplication is fundamentally a process to distinguish signals from noise. The signal/noise ratio of MSI detection is a function of the dup ratio, repeat locus characteristics, and deduplication strategy used. By addressing this issue, the majority-voting strategy adopted by plasmaMSI is shown to be superior over the canonical GSBQ strategy.

Clinical cfDNA samples vary in the amount of input DNA, the number of cycles in PCR amplification, the library complexity, and sequencing depth. The dup ratio serves as a proxy that accounts for the effects of all these factors. Matching a baseline to a test sample based on its dup ratio is essentially constructing a background allele length distribution (ie, calibrated for the specific test sample). This approach can potentially be applied to any baseline-dependent MSI calling algorithm. The baseline distribution should be built for a fixed and stable experimental protocol and inspected for outliers and abnormal variances. The dup ratio is only an approximation and is unable to represent all different factors. It has been reported that post-PCR contraction or expansion ratio can be affected by i) the repeat unit length, ii) specific nucleotides in the repeat unit, iii) the number of repeats, and iv) polymerase types.[35] To use plasmaMSI or any other MSI detection tool, the number of PCR cycles should first be assessed. In the current study, the number of PCR cycles was set to be 22 (8 cycles of precapture PCR plus 14 cycles of post-capture PCR), far below the 47-cycle threshold.[26]

plasmaMSI differs from other methods in additional aspects: the use of KLD, a distribution-based metric, that captures the abundance and inherent order of all alleles; and the adoption of a dup ratio−matching baseline to model noises. Other methods typically control the noises by excluding noisy alleles.[11,13] plasmaMSI enables characterization of the stutters/noises in the MS allele spectrum with better clarity, as evidenced by the strong performance of plasmaMSI in clinical and analytic validation experiments. An exact comparison with published methods is difficult because of different MS loci used by different methods. plasmaMSI was benchmarked against MSIsensor-ct version

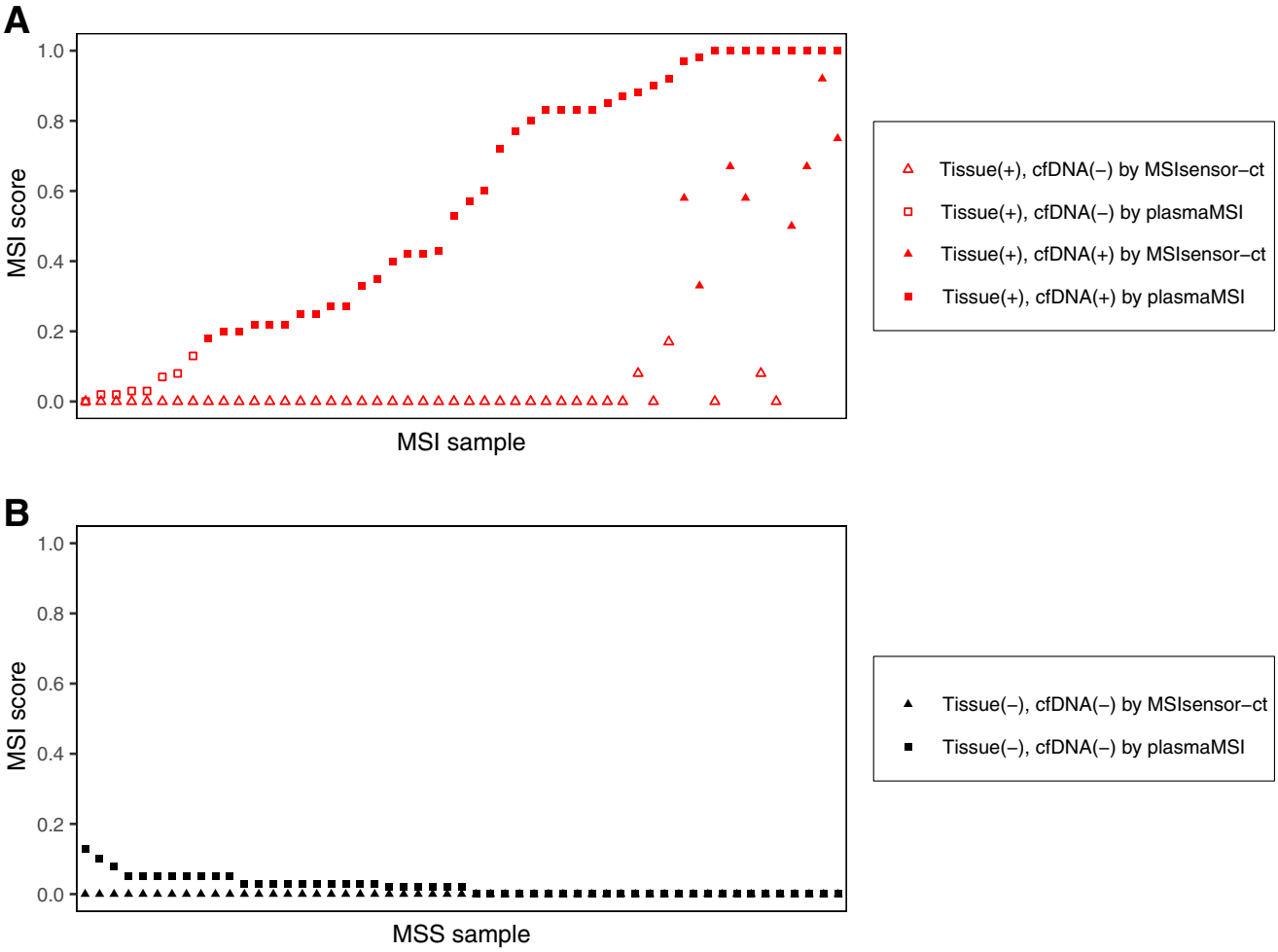

Figure 7 Comparison of microsatellite instability (MSI) scores calculated by plasmaMSI (squares) and MSIsensor-ct (triangles) on cell-free DNA (cfDNA) samples and their concordance with tissue results on 103 clinical samples. **A:** For tissue MSI (+) samples, solid shapes represent samples whose cfDNA were called MSI (+) by the designated method and hollow shapes represent samples whose cfDNA were called microsatellite stability (MSS; −) by the designated method. **B:** A similar plot for tissue MSS (−) samples.

1 by matching the number of MS loci used and showed significantly better performance. The poor performance of the latter can be explained by the adoption of a black-box machine-learning model, which probably incorporated the sequencing-platform−specific or experiment-batch effects into its model during training. Attempts to benchmark against other methods were unsuccessful because of the unavailability of tools or lack of the corresponding baselines. However, based on their reported performances (Supplemental Table S6), plasmaMSI, with a limit of detection of 0.15% and clinical sensitivity and specificity of 92.9% and 100%, respectively, is on par with or outperforms existing methods.

Conclusion

Regarding future improvement, the 49-locus panel has been mainly validated in gastrointestinal cancers and is likely to perform poorly in some other cancer types. Therefore, instead of selecting loci from a gastrointestinal cancer cohort, the method of plasmaMSI should be applied to a cohort of more diverse cancer types to form cancer-specific or pan-cancer MS panel. It is also straightforward to extend plasmaMSI to sequencing data with unique molecular identifiers, which will further increase its overall performance. In conclusion, a novel plasma cfDNA-based MSI detection method for the noninvasive diagnosis of cancer was developed and validated.

Acknowledgments

We thank the patients and staff at the First Affiliated Hospital of Kunming Medical University.

Author Contributions

W.L., Y.S.H., W.C., and J.He designed and supervised the study; F.H., J.Hu., J.Ya., and Q.L. provided and processed

the clinical samples; L.Z. and T.H. wrote the plasmaMSI source code; F.H., L.Z., T.H., J.S., H.X., J.Yu, and Y.S.H. formulated the algorithms and analyzed the data; X.W. and Y.H. performed the wet laboratory experiments; L.Z., Y.S.H., H.X., and W.L. cowrote the manuscript; and all authors approved the final version of the manuscript.

Disclosure Statement

L.Z., H.X., T.H., X.W., Y.H., J.S., J.Yu, Q.L., Y.S.H., W.C., and J.He are employees of Genecast Biotechnology Co., Ltd.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2024.10.002*.

References

1. Hampel H, Frankel W, Panescu J, Lockman J, Sotamaa K, Fix D, Comeras I, La Jeunesse J, Nakagawa H, Westman JA, Prior TW, Clendenning M, Penzone P, Lombardi J, Dunn P, Cohn DE, Copeland L, Eaton L, Fowler J, Lewandowski G, Vaccarello L, Bell J, Reid G, de la Chapelle A: Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. Cancer Res 2006, 66:7810−7817
2. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, Biedrzycki B, Donehower RC, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Duffy SM, Goldberg RM, de la Chapelle A, Koshiji M, Bhaijee F, Huebner T, Hruban RH, Wood LD, Cuka N, Pardoll DM, Papadopoulos N, Kinzler KW, Zhou S, Cornish TC, Taube JM, Anders RA, Eshleman JR, Vogelstein B, Diaz LA Jr: PD-1 blockade in tumors with mismatch-repair deficiency. N Engl J Med 2015, 372: 2509−2520
3. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al: Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science 2017, 357:409−413
4. Georgiadis A, Durham JN, Keefer LA, Bartlett BR, Zielonka M, Murphy D, White JR, Lu S, Verner EL, Ruan F, Riley D, Anders RA, Gedvilaite E, Angiuoli S, Jones S, Velculescu VE, Le DT, Diaz LA Jr, Sausen M: Noninvasive detection of microsatellite instability and high tumor mutation burden in cancer patients treated with PD-1 blockade. Clin Cancer Res 2019, 25:7024−7034
5. Willis J, Lefterova MI, Artyomenko A, Kasi PM, Nakamura Y, Mody K, Catenacci DVT, Fakih M, Barbacioru C, Zhao J, Sikora M, Fairclough SR, Lee H, Kim KM, Kim ST, Kim J, Gavino D, Benavides M, Peled N, Nguyen T, Cusnir M, Eskander RN, Azzi G, Yoshino T, Banks KC, Raymond VM, Lanman RB, Chudova DI, Talasaz A, Kopetz S, Lee J, Odegaard JI: Validation of microsatellite instability detection using a comprehensive plasma-based genotyping panel. Clin Cancer Res 2019, 25:7035−7045
6. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S: A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 1998, 58:5248−5257
7. Baudrin LG, Deleuze JF, How-Kit A: Molecular and computational methods for the detection of microsatellite instability in cancer. Front Oncol 2018, 8:621
8. 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, 18:65−71
9. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ: A molecular portrait of microsatellite instability across multiple cancers. Nat Commun 2017, 8:15180
10. Kim TM, Laird PW, Park PJ: The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 2013, 155: 858−868
11. Cai Z, Wang Z, Liu C, Shi D, Li D, Zheng M, Han-Zhang H, Lizaso A, Xiang J, Lv J, Wu W, Zhang Z, Zhang Z, Yuan F, He S, Sun J: Detection of microsatellite instability from circulating tumor DNA by targeted deep sequencing. J Mol Diagn 2020, 22:860−870
12. Han X, Zhang S, Zhou DC, Wang D, He X, Yuan D, Li R, He J, Duan X, Wendl MC, Ding L, Niu B: MSIsensor-ct: microsatellite instability detection using cfDNA sequencing data. Brief Bioinform 2021, 22:bbaa402
13. Wang Z, Zhao X, Gao C, Gong J, Wang X, Gao J, Li Z, Wang J, Yang B, Wang L, Zhang B, Zhou Y, Wang D, Li X, Bai Y, Li J, Shen L: Plasma-based microsatellite instability detection strategy to guide immune checkpoint blockade treatment. J Immunother Cancer 2020, 8:e001297
14. Xia L, Mei J, Kang R, Deng S, Chen Y, Yang Y, Feng G, Deng Y, Gan F, Lin Y, Pu Q, Ma L, Lin F, Yuan Y, Hu Y, Guo C, Liao H, Liu C, Zhu Y, Wang W, Liu Z, Xu Y, Li K, Li C, Li Q, He J, Chen W, Zhang X, Kou Y, Wang Y, Wu Z, Che G, Chen L, Liu L: Perioperative ctDNA-based molecular residual disease detection for non-small cell lung cancer: a prospective multicenter cohort study (LUNGCA-1). Clin Cancer Res 2022, 28:3308−3317
15. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014, 30:2114−2120
16. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754−1760
17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: 1000 Genome Project Data Processing Subgroup: The sequence alignment/map format and SAMtools. Bioinformatics 2009, 25:2078−2079
18. Kullback SLR: On information and sufficiency. Ann Math Stat 1951, 22:79−86
19. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, Tang B, Dong L, Ding N, Zhang Q, Bai Z, Dong X, Chen H, Sun M, Zhai S, Sun Y, Yu L, Lan L, Xiao J, Fang X, Lei H, Zhang Z, Zhao W: GSA: genome sequence archive. Dev Reprod Biol 2017, 15:14−18
20. National Genomics Data Center Members and Partners: Database resources of the National Genomics Data Center in 2020. Nucleic Acids Res 2020, 48:D24−D33
21. Bacher JW, Flanagan LA, Smalley RL, Nassif NA, Burgart LJ, Halberg RB, Megid WM, Thibodeau SN: Development of a fluorescent multiplex assay for detection of MSI-high tumors. Dis Markers 2004, 20:237−250
22. Bonneville R, Krook MA, Kautto EA, Miya J, Wing MR, Chen HZ, Reeser JW, Yu L, Roychowdhury S: Landscape of microsatellite instability across 39 cancer types. JCO Precis Oncol 2017, 2017:PO.17.00073
23. Hause RJ, Pritchard CC, Shendure J, Salipante SJ: Classification and characterization of microsatellite instability across 18 cancer types. Nat Med 2016, 22:1342−1350
24. Fujimoto A, Fujita M, Hasegawa T, Wong JH, Maejima K, Oku-Sasaki A, Nakano K, Shiraishi Y, Miyano S, Yamamoto G, Akagi K, Imoto S, Nakagawa H: Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. Genome Res 2020, 30:334−346
25. Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, Khayat MM, Mahmoud M, Laaguiby PK, Herbert ZT, Warner D, Grills GS, Jen J, Levy S, Xiang J, Alonso A, Zhao X, Zhang W, Teng F, Zhao Y, Lu H, Schroth GP, Narzisi G, Farmerie W, Sedlazeck FJ, Baldwin DA, Mason CE: Performance

assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. Nat Biotechnol 2021, 39: 1129–1140
26. Stoler N, Nekrutenko A: Sequencing error profiles of Illumina sequencing instruments. NAR Genom Bioinform 2021, 3:lqab019
27. Buhard O, Suraweera N, Lectard A, Duval A, Hamelin R: Quasi-monomorphic mononucleotide repeats for high-level microsatellite instability analysis. Dis Markers 2004, 20:251–257
28. Hatch SB, Lightfoot HM Jr, Garwacki CP, Moore DT, Calvo BF, Woosley JT, Sciarrotta J, Funkhouser WK, Farber RA: Microsatellite instability testing in colorectal carcinoma: choice of markers affects sensitivity of detection of mismatch repair-deficient tumors. Clin Cancer Res 2005, 11:2180–2187
29. Buhard O, Cattaneo F, Wong YF, Yim SF, Friedman E, Flejou JF, Duval A, Hamelin R: Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. J Clin Oncol 2006, 24: 241–251
30. 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 2017, 8: 7452–7463
31. Papke DJ Jr, Nowak JA, Yurgelun MB, Frieden A, Srivastava A, Lindeman NI, Sholl LM, MacConaill LE, Dong F: Validation of a targeted next-generation sequencing approach to detect mismatch repair deficiency in colorectal adenocarcinoma. Mod Pathol 2018, 31: 1882–1890
32. Long DR, Waalkes A, Panicker VP, Hause RJ, Salipante SJ: Identifying optimal loci for the molecular diagnosis of microsatellite instability. Clin Chem 2020, 66:1310–1318
33. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC: Microsatellite instability detection by next generation sequencing. Clin Chem 2014, 60:1192–1199
34. 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, 30:1015–1016
35. Raz O, Biezuner T, Spiro A, Amir S, Milo L, Titelman A, Onn A, Chapal-Ilani N, Tao L, Marx T, Feige U, Shapiro E: Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. Nucleic Acids Res 2019, 47:2436–2445