PLOS ONE

RESEARCH ARTICLE

Targeted next-generation sequencing-based detection of microsatellite instability in colorectal carcinomas

Yunbeom Lee[1☯], Ji Ae Lee[2,3☯], Hye Eun Park[2], Hyojun Han[1], Yuhnam Kim[1], Jeong Mo Bae[2,4], Jung Ho Kim[4], Nam-Yun Cho[2], Hwang-Phill Kim[5], Tae-You Kim[5,6], Gyeong Hoon Kang[2,3]*

1 Celemics, Inc. Seoul, Korea, **2** Laboratory of Epigenetics, Cancer Research Institute, Seoul National University College of Medicine, Seoul, Korea, **3** Department of Pathology, Seoul National University College of Medicine, Seoul, Korea, **4** Department of Pathology, Seoul National University Hospital, Seoul, Korea, **5** Laboratory of Cancer Epigenetics, Cancer Research Institute, Seoul National University College of Medicine, Seoul, Korea, **6** Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Korea

☯ These authors contributed equally to this work.
* ghkang@snu.ac.kr

Citation: Lee Y, Lee JA, Park HE, Han H, Kim Y, Bae JM, et al. (2021) Targeted next-generation sequencing-based detection of microsatellite instability in colorectal carcinomas. PLoS ONE 16(2): e0246356. https://doi.org/10.1371/journal.pone.0246356

Editor: Alvaro Galli, CNR, ITALY

Received: October 1, 2020

Accepted: January 18, 2021

Published: February 1, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: This study was supported by Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea, in the form of a grant awarded to GHK (HI14C1277) and Celemics, Inc. in the form of salaries for YL, HH, and YK. The specific roles of these authors are articulated in the 'author contributions' section. The funders had no role in

Abstract

In the present study, we developed a computational method and panel markers to assess microsatellite instability (MSI) using a targeted next-generation sequencing (NGS) platform and compared the performance of our computational method, mSILICO, with that of mSINGS to detect MSI in CRCs. We evaluated 13 CRC cell lines, 84 fresh and 119 formalin-fixed CRC tissues (including 61 MSI-high CRCs and 155 microsatellite-stable CRCs) and tested the classification performance of the two methods on 23, 230, and 3,154 microsatellite markers. For the fresh tissue and cell line samples, mSILICO showed a sensitivity of 100% and a specificity of 100%, regardless of the number of panel markers, whereas for the formalin-fixed tissue samples, mSILICO exhibited a sensitivity of up to 100% and a specificity of up to 100% with three differently sized panels ranging from 23 to 3154. These results were similar to those of mSINGS. With the application of mSILICO, the small panel of 23 markers had a sensitivity of $\geq$95% and a specificity of 100% in cell lines/fresh tissues and formalin-fixed tissues of CRC. In conclusion, we developed a new computational method and microsatellite marker panels for the determination of MSI that does not require paired normal tissues. A small panel could be integrated into the targeted NGS panel for the concurrent analysis of single nucleotide variations and MSI.

Introduction

Microsatellites, also known as short tandem repeats, are tracts of tandemly repeated DNA motifs that range from 1 to 6 bp and are typically repeated 10–60 times. Microsatellites are encountered throughout the human genome at a frequency of one microsatellite locus per 2,000 bp, accounting for 3% of the human genome [1]. Approximately 8% of microsatellites

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have read the journal's policy and have the following competing interests: YL, HH, and YK are employees of Celemics, Inc. This does not alter our adherence to PLOS ONE policies on sharing data and materials. The authors have submitted a patent on analysis method and compositions of genetic markers for detecting microsatellite-instability in cancer via next generation sequencing. There are no other patents, products in development or marketed products associated with this research to declare.

are found in coding regions [2]. and the most common microsatellites of coding regions are A homopolymers [3]. Microsatellites tend to undergo slipped strand mispairing during DNA replication, which is repaired by mismatch repair (MMR) enzymes. Microsatellite instability (MSI) is characterized by genome-wide alterations in the number of repeated DNA bases in microsatellites due to defective DNA MMR. High rates of repeat number alterations and increased rates of single nucleotide variations feature MSI-high tumor cells [4]. The MSI-high molecular phenotype is found in colorectal cancers (CRCs) at a frequency of 8–15% and correlates with morphological phenotype [5].

MSI testing of DNA samples obtained from paired tumor and normal tissues is traditionally performed using the National Cancer Institute (NCI)'s five-marker panel (BAT25, BAT26, D2S123, D5S346, and D17S250). MSI status is classified as MSI-high (MSI-H; instability at 2 or more markers) and microsatellite-stable (MSS; instability at 1 marker or none). Recent data suggest that the use of mononucleotide repeats increases the sensitivity of the detection of MSI compared to the use of dinucleotide repeats [6–8]. The revised Bethesda guideline recommended the use of a new panel, the pentaplex panel of five mononucleotide repeats for the detection of MSI [9]. There has been a growing demand to develop more sensitive solutions for the detection of MSI with a larger number of microsatellite loci. Next-generation sequencing (NGS) allows for the analysis of a greater number of microsatellite markers than PCR-based detection of MSI (PCR-MSI). Furthermore, simultaneous analysis of both MSI and SNVs or indels is possible from a single assay of NGS.

NGS-based assessment of MSI (NGS-MSI) has pursued two approaches based on the type of interrogated microsatellite markers, including one approach using microsatellite loci located within the captured gene sequences [10] and the other approach capturing dedicated specific microsatellite marker sites that are not included in targeted gene capture sequencing data. Several computational methods for NGS-MSI have been developed, including MSISensor, mSINGS, MANTIS, and Cortes-Ciriano, which are based on the comparison of the repeat length distribution of microsatellites [11]. Of these computational methods, mSINGS does not need paired normal tissue samples for the detection of MSI in tumor samples [11]. Instead, mSINGS compares tumor-only samples to a pre-constructed baseline-control. In the present study, we developed a new computational method, mSILICO, which is run on tumor samples without matched normal tissue samples. We designed 3,154 capture probes for dedicated microsatellite sites. mSINGS and mSILICO were compared regarding the sensitivity and specificity for the detection of MSI in CRCs.

Materials and methods

This study was approved by the Institutional Review of Board of Seoul National University Hospital (IRB No. H-1605-080-761). The written informed consent was obtained from CRC patients prior to participating in the present study. This study was conducted in compliance with the principles of the Declaration of Helsinki and its later amendments. Fresh tissue samples of paired CRC tumors and adjacent mucosa were obtained from patients (n = 84) who underwent surgical resection due to CRC in Seoul National University Hospital in 2018. Genomic DNA was extracted from the paired tissue samples using a Qiagen kit (Qiagen, Hilden, Germany). After review of electronic medical records from CRC patients who underwent surgery in 2018, 40 cases of MSI-high CRC and 79 cases of non-MSI-high CRC were selected. After microscopic examination of glass slides, a 1 cm-sized tumor area with the highest tumor purity and a normal tissue area were marked. The corresponding areas were marked on the unstained recut slides and then subjected to deparaffinization. The marked tumor and normal tissue areas were separately scraped into microcentrifuge tubes with a knife blade. Genomic

DNA was extracted using the Qiagen FFPE Kit. Cell lines (n = 13) were purchased from Korean Cell Line Bank (Seoul, Korea), and genomic DNA was extracted from the cell lines after culture. The cells were cultured in a 5% CO_2 humidified atmosphere at 37˚C using DMEM (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 10% fetal bovine serum, 100 μg/mL penicillin and 100 μg/mL streptomycin.

PCR-MSI using Bethesda's five-marker panel

Genomic DNA samples obtained from paired tumor and normal tissue samples were subjected to PCR with fluorescently labeled oligonucleotide primers for Bethesda's five microsatellite loci (BAT25, BAT26, D2S123, D5S346, and D17S250), and then the PCR products were analyzed by capillary electrophoresis on an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA). Instability at the examined locus was defined by altered length of the PCR product in the tumor sample compared with the length of the PCR product in the paired normal sample. MSI status was classified as follows: MSI-H (instability at 2 or more microsatellite loci) and MSS (instability at 1 locus or none).

Immunohistochemistry for MMR proteins

Immunohistochemistry was performed to evaluate the expression of the MLH1, MSH2, MSH6, and PMS2 proteins as described previously [12]. The criteria for the microscopic interpretation of each IHC marker were based on our previous study [12].

Library preparation and next-generation sequencing

The NGS panel was designed for simultaneous detection of MSI status and mutations in 50 CRC-related genes, including 40 genes associated with the WNT, p53, RTK-RAS, TGF-β, and PI3K pathways [13] (S1 Table). Custom RNA probes were designed for target enrichment sequencing (Celemics Inc, Seoul, Korea) and covered all unions of reported exons of 50 genes (total count of regions was 753). The panel covers 777,937 bases of the human genome (hg19). All transcripts of genes reported in UCSC were included as targets to thoroughly detect SNVs, small insertion-deletion mutations, and structural variants. In addition, the panel contains special RNA probes that capture dedicated specific microsatellite marker sites (n = 3,154) for the detection of MSI.

Genomic DNA was sheared and processed for Illumina sequencing. The process included the following steps: end repair, dA tailing, adaptor ligation and pre-PCR for the indexed NGS library. Capture probes were hybridized in buffer to capture target regions of the C5 gene through the use of the Celemics Target Enrichment Kit. After the capture and washing processes were completed, the captured library was amplified by post-PCR. The PCR products were sequenced on the NextSeq 500 platform by Illumina Inc.

'BCL2FASTQ' version 2.19.1.403 (Illumina) was used to demultiplex the base-call image files into individual sequence read files (FASTQ format). All options and parameters followed the default setting. Sequencing adaptors were removed by "AdapterRemoval ver. 2.2.2" [14], after low quality bases (below quality 0 and N sequence) were removed by native python code. All sequencing reads were aligned to the hg19 human genome by BWA-MEM (Burrows-Wheeler Aligner) software. The program used the Burrows-Wheeler Transform algorithm to index the human genome sequence to calculate the constant complexity of each sequencing read. Post-align and recalibration processes were performed by 'Picard' ver. 1.115 (http://broadinstitute.github.io/picard) and 'GenomeAnalysisToolKit (GATK ver. 4.0.4.0)' [15]. We performed variant calling with a GATK haplotype caller. All detailed parameters and options followed GATK best practices.

MSI marker panels

To select microsatellite markers, we went through a series of processes. First, we referred to previously reported studies that conducted NGS-based determination of MSI [16–19] and retrieved 3,154 marker candidates. Next, to construct an intermediate-sized MSI marker panel, we randomly selected 230 markers from 3,154 markers. Finally, we looked for overlap of the marker candidates across the studies [17, 20], which revealed 18 shared markers. In addition to these 18 markers, three Bethesda markers and two markers from Salipante et al.'s study [16] were added, resulting in 23 markers (S2 Table). Because our NGS panel was designed for simultaneous detection of MSI status and mutations in 50 CRC-related genes, two markers of the Salipante's study are already present on the genomic sequence covered by the 50-gene panel. Thus, we included these two markers in the 23-marker panel.

Computational methods for the determination of MSI by NGS

Instability at a specific microsatellite marker was determined using calculation formula which measures the skewness in the distribution of read lengths on a specific microsatellite marker. We collected all the reads mapped on each microsatellite marker and counted the length of each read. The read length was normalized by reference genome (hg19) length, which is set to 0 if the read length is the same as that of the reference. Markers with a sequence coverage ($\leq$20x) were filtered out and excluded from the following analysis. Pearson's skewness coefficient (PSC) was used to quantify the skewness of each length dataset of each marker, "PSC = 3 X ((mean − median)/standard deviation)". PSC compares the distribution of the dataset with a normal distribution. The larger the coefficient value, the larger the distribution of the dataset differs from a normal distribution. When we looked at the distribution of the observed PSC values from all the microsatellite markers and CRC tissues, more than half of the observed PSC values from MSI-H CRCs were more than 1 or less than -1, whereas approximately 10% of the observed PSC values from MSS CRCs were more than 1 or less than -1 (S1 Fig). Based on such a finding, we defined a marker as unstable If the PSC value was more than 1 or less than -1. When over a certain fraction of all valid markers were unstable, the tumor sample was finally called MSI-H.

Statistical analysis

The statistical analysis was performed with SPSS software for Windows, vesion 25.0 (IBM, Chicago, IL, USA). McNemar's test was used to compare the sensitivities and specificities of mSILICO and mSINGS.

Results

Based on our observations that in capillary electrophoresis of fluorochrome-labeled microsatellite sequences, MSI-H tumors display a more skewed distribution of the PCR products than those of MSS tumors, we developed a computational method for NGS-based determination of MSI, called mSILICO, which determines the skewness in the distribution of read lengths on each microsatellite locus and calls instability of each locus based on PSC$>$1 or $<$-1. To evaluate mSILICO's performance, we compared its performance with that of mSINGS in the detection of MSI in CRCs. Immunohistochemistry for MMR proteins, including MLH1, MSH2, MSH6, and PMS2, was performed to confirm MSI statuses which were determined in CRC tissues by NCI's five-marker panel (BAT25, BAT26, D2S123, D5S346, and D17S250). There was no discrepancy between expression status of MMR proteins and MSI status determined by NCI's panel.

Comparison of sensitivity and specificity in the detection of MSI according to the number of MSI markers

Eighty-four paired fresh CRC tissues of which MSI status had been determined by PCR-MSI were recruited. Sixteen cases were positive for MSI, and the others were negative for MSI. Thirteen cell lines, including five MSI-H cell lines, were also included for the analysis of MSI on the NGS platform. First, the MSI detection performances of mSILICO and mSINGS were compared in 13 CRC cell lines (Fig 1A) and 84 fresh CRC tissues (Fig 1B) for 3,154

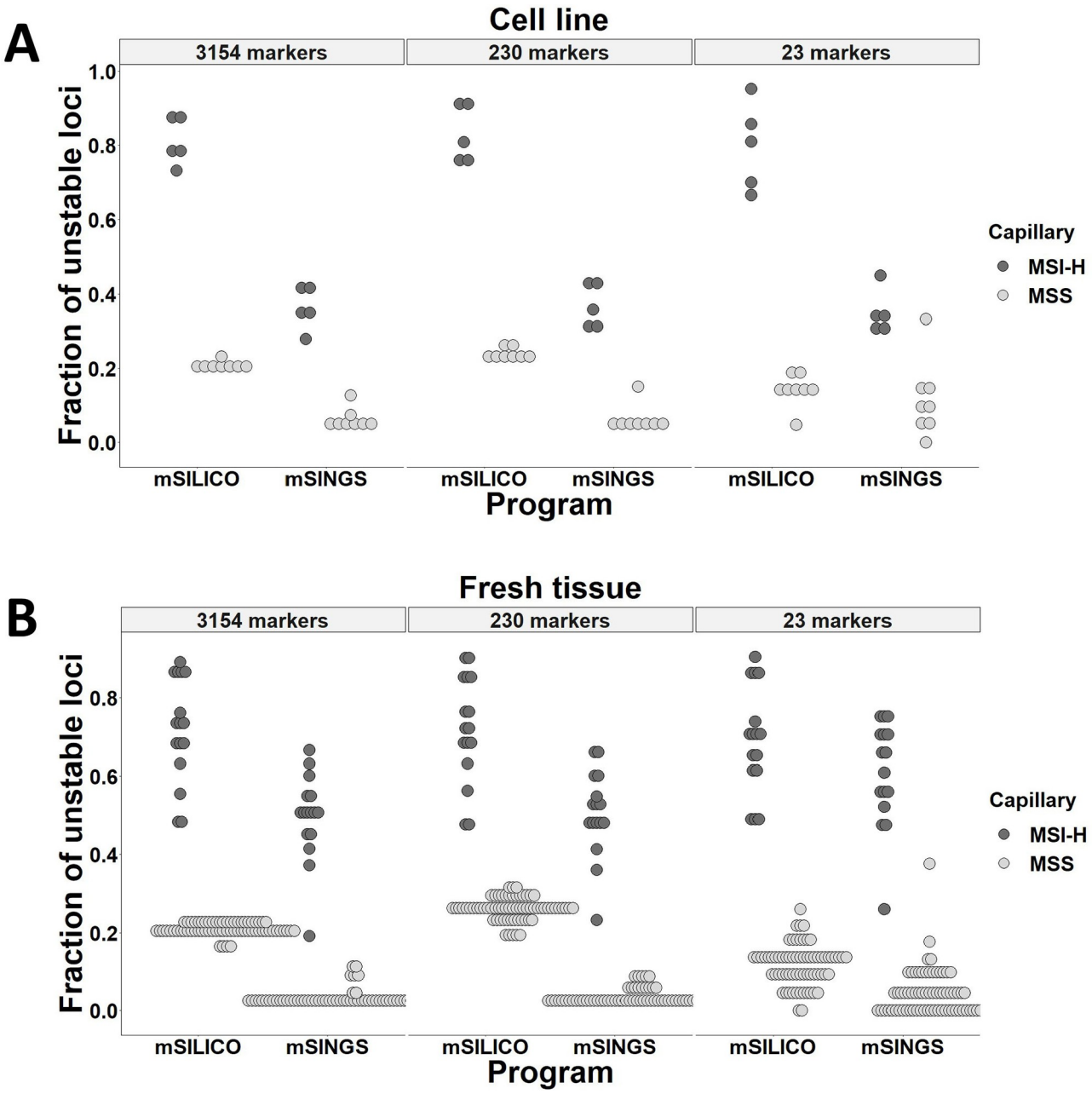

Fig 1. Comparison of the performances of mSILICO and mSINGS in the detection of microsatellite instability. A total of 3,154, 230, and 23 markers were analyzed for their instabilities in colorectal cancer cell lines (A) and fresh colorectal cancer tissues (B).

https://doi.org/10.1371/journal.pone.0246356.g001

microsatellite markers. Regardless of the type of computational method, MSI-H and MSS CRC tumors and cell lines were distinctly distributed along the axis representing the fraction of unstable markers. Cell lines tended to be more widely separated in the fraction of unstable markers with application of mSILICO than with that of mSINGS.

To evaluate whether the number of interrogated microsatellite markers could affect the sensitivity and specificity in the detection of MSI, two additional panels of microsatellite loci that included 230 and 23 loci were assessed for their performance in the detection of MSI by mSINGS and mSILICO (Fig 1A and 1B). For the 230-marker panel, MSI-H and MSS tumors were distinctly distributed regardless of the computational method. However, for the 23-marker panel, partial overlap was seen between MSI-H and MSS fresh CRC tissues and between MSI-H and MSS CRC cell lines with application of mSINGS. One (12.5%) of eight MSS CRC cell lines was misdiagnosed into MSI-H and one (1.5%) of 68 MSS CRC tissues was misplaced into MSI-H. In contrast, no overlap was found between MSI-H and MSS CRC tissues or between MSI-H and MSS CRC cell lines with the application of mSILICO.

For fresh tissue and cell line samples, a fraction of $\geq$0.4 ($\geq$40% unstable loci) and a fraction of $\geq$0.2 ($\geq$20% unstable loci) were used as cut-offs for MSI-H in mSILICO and mSINGS computational methods, respectively. These cut-off values were determined empirically. When the MSI-PCR results were used as the reference, the 3,154-, 230-, and 23-marker panels showed a sensitivity of 100% and a specificity of 100% in both cell lines and fresh tissues with the application of mSILICO. However, with the application of mSINGS, the 3,154-, 230-, and 23-marker panels had sensitivities of 100% and specificities of 100%, 100%, 87.5%, respectively, in cell lines and sensitivities of 93.8%, 100%, and 100%, respectively, and specificities of 100%, 100%, and 98.5%, respectively, in fresh tissues (Table 1).

Table 1. Performance of computational methods (mSINGS and mSILICO) and microsatellite marker panels (3,154, 230 and 23 markers) in CRC cell lines and fresh tissue samples.

			Cell line	Fresh tissue
3,154	mSINGS	Sensitivity	100%	93.8%
		Specificity	100%	100%
		Accuracy	100%	98.8%
	mSILICO	Sensitivity	100%	100%
		Specificity	100%	100%
		Accuracy	100%	100%
230	mSINGS	Sensitivity	100%	100%
		Specificity	100%	100%
		Accuracy	100%	100%
	mSILICO	Sensitivity	100%	100%
		Specificity	100%	100%
		Accuracy	100%	100%
23	mSINGS	Sensitivity	100%	100%
		Specificity	87.5%	98.5%
		Accuracy	92.3%	98.8%
	mSILICO	Sensitivity	100%	100%
		Specificity	100%	100%
		Accuracy	100%	100%

No difference in sensitivities and specificities of 23-, 230-, 3,154-marker panels between mSINGS and mSILICO (by McNemar test)

https://doi.org/10.1371/journal.pone.0246356.t001

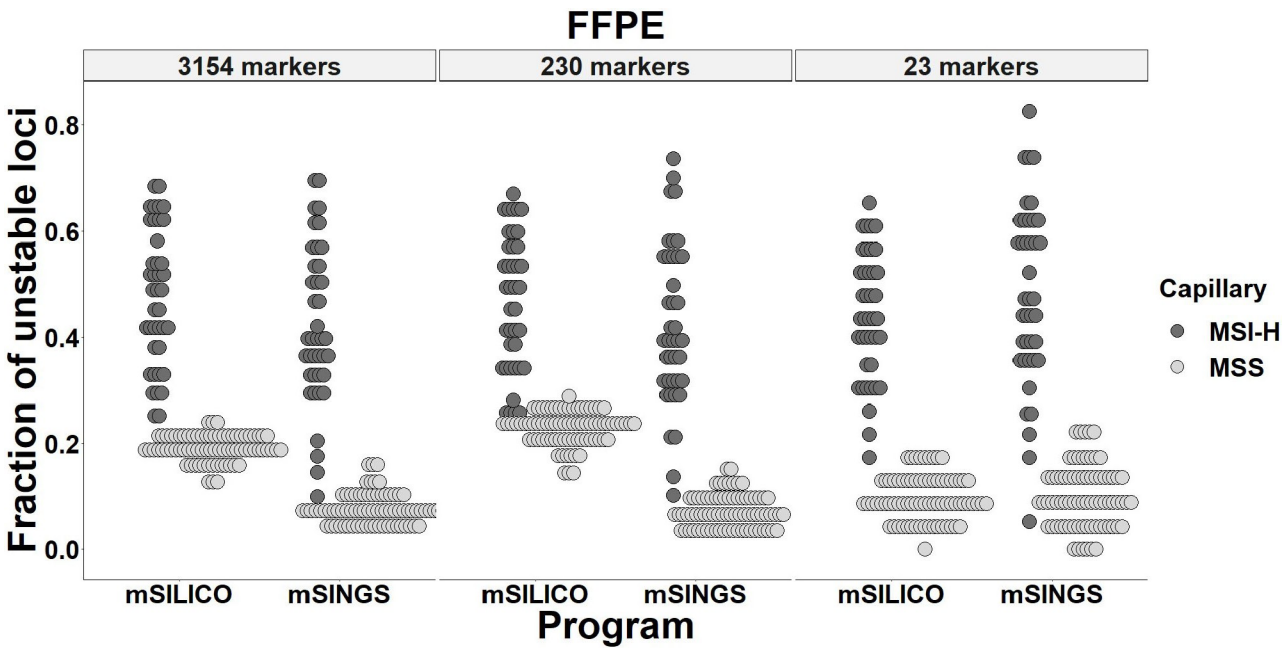

Fig 2. mSILICO and mSINGS were evaluated for their performance in the detection of microsatellite instability in formalin-fixed paraffin-embedded (FFPE) tissues of colorectal cancers using three different marker panels.

https://doi.org/10.1371/journal.pone.0246356.g002

The performance of mSILICO in FFPE tissue samples

Next, to identify whether the mSILICO computational method works in FFPE tissue samples as well as in fresh tissue samples or cell lines, two computational methods coupled with three different marker panels were compared to determine their performance in the detection of MSI with FFPE tissue samples. FFPE tissue samples of 40 MSI-H and 77 MSS CRCs whose MSI status using MSI-PCR had been previously determined were retrieved and subjected to MSI-NGS. Regardless of the number of panel markers, partial overlap between MSI-H and MSS tumors was seen in mSINGS (Fig 2). Although no distinct separation between MSI-H and MSS tumors was noted in the 3,154-, 230-, and 23-marker panels with application of mSILICO, mSILICO showed a lower degree of overlap between MSI-H and MSS tumors in the 3,154- and 23-marker panels compared with mSINGS. With a cut-off value of 0.25, the 3,154-, 230- and 23-marker panels had a sensitivity of 100%, 100%, and 95%, respectively, and a specificity of 100%, 77.2% and 100%, respectively, with application of mSILICO. With the application of mSINGS and a cut-off value of 0.20, the 3,154-, 230- and 23-marker panels had sensitivities of 92.5%, 95%, and 95%, respectively, and specificities of 100%, 100%, and 93.7%, respectively (Table 2).

Discussion

In the present study, we developed a new computational method, mSILICO, which utilizes the skewness in the distribution of the read lengths at each microsatellite locus. When compared with mSINGS, mSILICO showed comparable results in the sensitivity and specificity of MSI detection in both fresh tissue or cell line samples and FFPE tissue samples. In our study, the performance of the 23-marker panel was similar to that of the 3,154-marker panel in the detection of MSI regardless of whether genomic DNA samples were extracted from fresh tissues or

Table 2. Performance of computational methods (mSINGS and mSILICO) and microsatellite marker panels (3,154, 230 and 23 markers) in formalin-fixed paraffin-embedded (FFPE) tissue samples.

			FFPE
3,154	mSINGS	Sensitivity	92.5%
		Specificity	100%
		Accuracy	97.5%
	mSILICO	Sensitivity	100%
		Specificity	100%
		Accuracy	100%
230	mSINGS	Sensitivity	95.0%
		Specificity	100%
		Accuracy	98.3%
	mSILICO	Sensitivity	100%
		Specificity	77.2%
		Accuracy	84.9%
23	mSINGS	Sensitivity	95.0%
		Specificity	93.7%
		Accuracy	94.1%
	mSILICO	Sensitivity	95.0%
		Specificity	100%
		Accuracy	98.3%

A significant difference in sensitivity and specificity of 230-marker panel but no differences in 23-marker panel and 3,154-maker panel between mSILICO and mSINGS (by McNemar test)

https://doi.org/10.1371/journal.pone.0246356.t002

formalin-fixed tissues. Such a small panel can be added to existing targeted exome panels, reinforcing the performance of MSI analysis. Although targeted exome sequencing of ≥275 genes (encompassing 757,787 bp of the genome) was demonstrated to differentiate MSI-H from MSS CRCs at a sensitivity and specificity of ≥99% using tumor mutational load [21, 22], information regarding a mutational load obtained from the smaller genomic space sampled by gene panels, is unlikely to clearly differentiate MSI-H CRCs from MSS CRCs, where a panel of 23 microsatellite markers could help to diagnose MSI accurately.

When the performance of mSILICO and mSINGS for MSI-NGS was compared among cell lines, fresh frozen and FFPE tissue samples, the determination power was the lowest in the FFPE group regardless of the computational method. For the cell lines and fresh tissue samples, the MSI-H and MSS groups were clearly distinguished in mSILICO, regardless of the number of panel markers. Rather, in the control method, mSINGS, false-negative or false-positive results occur in sample types of cell lines and fresh tissue. For FFPE samples with the use of 23 or 3,154 markers, mSILICO tended to show higher accuracy than mSINGS, although mSILICO exhibited lower specificity in the 230-marker panel than mSINGS. Another point that favors mSILICO over mSINGS is that while mSINGS needs MSS samples (n = 10–20) to set the baseline, mSILICO can detect MSI without recruiting MSS sample data. Additionally, the running time of mSINGS for determination of MSI status was longer than that of mSILICO (S2 Fig). However, the cut-off value of the unstable fraction was constant in mSINGS regardless of whether tissue samples were fresh or formalin-fixed, whereas mSILICO showed different cut-off values for determining MSI-H depending on whether the tissue sample was fresh or fixed. In our mSILICO results, the cut-off value for determination of MSI differed between FFPE tissues and cell lines/fresh tissues. For fresh tissues, the cut-off value could be

set at 0.4 regardless of panel size (23, 230, and 3,154 markers), whereas for FFPE tissues, the cut-off value could be set at 0.25. Further optimization of the cut-off value should be performed with a large-scale sample of CRC tissue samples.

In DNA samples obtained from cell lines or fresh frozen tissues, the fraction value of unstable loci was clearly separated between MSI-H tumors and MSS tumors, whereas in DNA samples extracted from FFPE tissues, the fraction of unstable loci was distributed with partial overlap between MSI-H tumors and MSS tumors. The tendency of incomplete separation in relation to the use of genomic DNA from FFPE tissue was also found in the control computational method, mSINGS, and in other MSI-NGS studies using mSINGS [16], which suggests that complete differentiation between MSI-H and MSS tumors might be difficult to accomplish with current NGS and analytical techniques. For such a tumor with borderline fraction value, information on tumor mutational load might help to differentiate MSI-H from MSS tumors.

For MSI-H tumors, the fraction of unstable loci tended to be decreased and widely distributed in FFPE tissues compared with that of fresh tissues regardless of the computational method and the number of panel markers. For MSS tumors, the fraction of unstable loci tended to be more widely distributed in FFPE tissues than in fresh tissues, regardless of the computational method and the number of panel markers. The reason why the fraction of unstable loci was more widely distributed in FFPE tissues than in fresh tissues might be related to sequencing artifacts, which can be caused by damage to DNA due to fixation with formalin, sample storage at room temperature and DNA extraction procedures. Of FFPE-associated sequencing artifacts, including G-C bias, increased base-error rate, decreased proportion of properly paired read alignment, strand-split artifact reads, and aberrant detection of copy number gain or loss [23–25], it is unclear which ones are involved in wide distribution of the fraction of unstable loci in association with formalin fixation. When we analyzed the mean number of reads along the markers in each tumor or cell line and the mean number of reads along the samples in each marker, FFPE tissues showed much wider variation in the mean number of reads than fresh tissues or cell lines (S3 Fig). Each marker also exhibited wider variation in the number of reads in FFPE tissue samples than in fresh tissues or cell lines. A lower number of reads might lead to underrepresentation of the alleles, which might generate a tendency toward a decreasing absolute value of PSC in each marker, finally resulting in a decreased fraction of unstable loci in FFPE tissue samples.

The limitation of our study was that we did not analyze the performance of mSILICO and marker panels in other tissue types of cancers in which MSI occurs as frequently as in CRCs. Thus, we do not know whether mSILICO and marker panels work in gastric cancers or endometrial cancers as well as in CRCs. The baseline microsatellite status for this dataset was determined using MSI-PCR, which was performed with Bethesda's five- marker panel. Bethesda's five-marker panel is known to be inferior to pentaplex PCR. Thus, a few missing MSI tumors might be included in MSS tumors. However, mSILICO and mSINGS did not find any MSI-H tumors in MSS tumors.

In summary, we developed a computational method that can differentiate MSI-H tumors from MSS tumors without paired normal tissue samples and even baseline normal samples. A small panel of 23 markers can be added to existing gene panels for targeted exome sequencing, which are not able to diagnose MSS tumors with mutational load alone.

Supporting information

S1 Table. List of genes.
(DOCX)

S2 Table. Twenty-three microsatellite markers.
(DOCX)

S1 Fig. Distributions of Pearson's skewness coefficient values in microsatellite instability-high and microsatellite-stable colorectal cancers.
(TIF)

S2 Fig. The two programs differ in the time it takes because mSINGS requires several more steps.
(TIF)

S3 Fig. The number of reads was obtained in 23 microsatellite markers, and the mean number of reads (A) and standard deviation (SD) were compared among colorectal cancer cell lines, formalin-fixed, paraffin-embedded (FFPE), and fresh colorectal cancer tissues.
(TIF)

Acknowledgments

GHK is the guarantor of this work and, as such, had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All the authors fulfilled the authorship criteria "1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the manuscript or revising it critically for important intellectual content; 3) final approval of the version to be published; and 4) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Author Contributions

Conceptualization: Hwang-Phill Kim, Tae-You Kim, Gyeong Hoon Kang.

Data curation: Hye Eun Park.

Formal analysis: Ji Ae Lee.

Funding acquisition: Hwang-Phill Kim, Tae-You Kim, Gyeong Hoon Kang.

Investigation: Yunbeom Lee, Ji Ae Lee.

Methodology: Hyojun Han, Yuhnam Kim.

Project administration: Nam-Yun Cho.

Software: Yunbeom Lee.

Supervision: Yuhnam Kim, Gyeong Hoon Kang.

Visualization: Yunbeom Lee.

Writing – original draft: Gyeong Hoon Kang.

Writing – review & editing: Jeong Mo Bae, Jung Ho Kim.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409(6822):860–921. https://doi.org/10.1038/35057062 PMID: 11237011

2. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet. 2000; 24(4):400–2. https://doi.org/10.1038/74249 PMID: 10742106

3. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 2004; 21(6):991–1007. https://doi.org/10.1093/molbev/msh073 PMID: 14963101

4. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell. 2013; 155(4):858–68. https://doi.org/10.1016/j.cell.2013.10.015 PMID: 24209623

5. Kim JH, Kang GH. Molecular and prognostic heterogeneity of microsatellite-unstable colorectal cancer. World journal of gastroenterology: WJG. 2014; 20(15):4230–43. https://doi.org/10.3748/wjg.v20.i15.4230 PMID: 24764661

6. Suraweera N, Duval A, Reperant M, Vaury C, Furlan D, Leroy K, et al. Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. Gastroenterology. 2002; 123(6):1804–11. https://doi.org/10.1053/gast.2002.37070 PMID: 12454837

7. Akiyama Y, Sato H, Yamada T, Nagasaki H, Tsuchiya A, Abe R, et al. Germ-line mutation of the hMSH6/GTBP gene in an atypical hereditary nonpolyposis colorectal cancer kindred. Cancer Res. 1997; 57(18):3920–3. PMID: 9307272

8. Bacher JW, Flanagan LA, Smalley RL, Nassif NA, Burgart LJ, Halberg RB, et al. Development of a fluorescent multiplex assay for detection of MSI-High tumors. Dis Markers. 2004; 20(4–5):237–50. https://doi.org/10.1155/2004/136734 PMID: 15528789

9. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Ruschoff J, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst. 2004; 96(4):261–8. https://doi.org/10.1093/jnci/djh034 PMID: 14970275

10. Middha S, Zhang L, Nafa K, Jayakumaran G, Wong D, Kim HR, et al. Reliable Pan-Cancer Microsatellite Instability Assessment by Using Targeted Next-Generation Sequencing Data. JCO Precis Oncol. 2017; 2017. https://doi.org/10.1200/PO.17.00084 PMID: 30211344

11. Baudrin LG, Deleuze JF, How-Kit A. Molecular and Computational Methods for the Detection of Microsatellite Instability in Cancer. Front Oncol. 2018; 8:621. https://doi.org/10.3389/fonc.2018.00621 PMID: 30631754

12. Kim JH, Park HE, Cho NY, Lee HS, Kang GH. Characterisation of PD-L1-positive subsets of microsatellite-unstable colorectal cancers. British journal of cancer. 2016; 115(4):490–6. https://doi.org/10.1038/bjc.2016.211 PMID: 27404452

13. Lee DW, Han SW, Cha Y, Bae JM, Kim HP, Lyu J, et al. Association between mutations of critical pathway genes and survival outcomes according to the tumor location in colorectal cancer. Cancer. 2017; 123(18):3513–23. https://doi.org/10.1002/cncr.30760 PMID: 28513830

14. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016; 9:88. https://doi.org/10.1186/s13104-016-1900-2 PMID: 26868221

15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

16. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection by next generation sequencing. Clin Chem. 2014; 60(9):1192–9. https://doi.org/10.1373/clinchem.2014.223677 PMID: 24987110

17. Zhao H, Thienpont B, Yesilyurt BT, Moisse M, Reumers J, Coenegrachts L, et al. Mismatch repair deficiency endows tumors with a unique mutation signature and sensitivity to DNA double-strand breaks. Elife. 2014; 3:e02725. https://doi.org/10.7554/eLife.02725 PMID: 25085081

18. Pritchard CC, Smith C, Salipante SJ, Lee MK, Thornton AM, Nord AS, et al. ColoSeq provides comprehensive lynch and polyposis syndrome mutational analysis using massively parallel sequencing. J Mol Diagn. 2012; 14(4):357–66. https://doi.org/10.1016/j.jmoldx.2012.03.002 PMID: 22658618

19. Pritchard CC, Salipante SJ, Koehler K, Smith C, Scroggins S, Wood B, et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J Mol Diagn. 2014; 16(1):56–67. https://doi.org/10.1016/j.jmoldx.2013.08.004 PMID: 24189654

20. 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. https://doi.org/10.1038/ncomms15180 PMID: 28585546

21. Stadler ZK, Battaglin F, Middha S, Hechtman JF, Tran C, Cercek A, et al. Reliable Detection of Mismatch Repair Deficiency in Colorectal Cancers Using Mutational Load in Next-Generation Sequencing

Panels. J Clin Oncol. 2016; 34(18):2141–7. https://doi.org/10.1200/JCO.2015.65.1067 PMID: 27022117

22. Nowak JA, Yurgelun MB, Bruce JL, Rojas-Rudilla V, Hall DL, Shivdasani P, et al. Detection of Mismatch Repair Deficiency and Microsatellite Instability in Colorectal Adenocarcinoma by Targeted Next-Generation Sequencing. J Mol Diagn. 2017; 19(1):84–91. https://doi.org/10.1016/j.jmoldx.2016.07.010 PMID: 27863258

23. Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. Nucleic Acids Res. 2019; 47(2):e12. https://doi.org/10.1093/nar/gky1142 PMID: 30418619

24. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem. 2015; 61(1):64–71. https://doi.org/10.1373/clinchem.2014.223040 PMID: 25421801

25. Hosein AN, Song S, McCart Reed AE, Jayanthan J, Reid LE, Kutasovic JR, et al. Evaluating the repair of DNA derived from formalin-fixed paraffin-embedded tissues prior to genomic profiling by SNP-CGH analysis. Lab Invest. 2013; 93(6):701–10. https://doi.org/10.1038/labinvest.2013.54 PMID: 23568031