OXFORD

Briefings in Bioinformatics, 2024, **25(5)**, bbae390
https://doi.org/10.1093/bib/bbae390
Problem Solving Protocol

Performance assessment of computational tools to detect microsatellite instability

Harrison Anthony [1,2] and Cathal Seoighe [1,2,*]

[1]School of Mathematical and Statistical Sciences, University of Galway, Galway H91 TK33, Ireland
[2]The SFI Centre for Research Training in Genomics Data Science, Galway D02 FX65, Ireland
*Corresponding author: Cathal Seoighe, School of Mathematical and Statistical Science, University of Galway, University Road, Galway, Ireland.
E-mail: cathal.seoighe@universityofgalway.ie

Abstract

Microsatellite instability (MSI) is a phenomenon seen in several cancer types, which can be used as a biomarker to help guide immune checkpoint inhibitor treatment. To facilitate this, researchers have developed computational tools to categorize samples as having high microsatellite instability, or as being microsatellite stable using next-generation sequencing data. Most of these tools were published with unclear scope and usage, and they have yet to be independently benchmarked. To address these issues, we assessed the performance of eight leading MSI tools across several unique datasets that encompass a wide variety of sequencing methods. While we were able to replicate the original findings of each tool on whole exome sequencing data, most tools had worse receiver operating characteristic and precision-recall area under the curve values on whole genome sequencing data. We also found that they lacked agreement with one another and with commercial MSI software on gene panel data, and that optimal threshold cut-offs vary by sequencing type. Lastly, we tested tools made specifically for RNA sequencing data and found they were outperformed by tools designed for use with DNA sequencing data. Out of all, two tools (MSIsensor2, MANTIS) performed well across nearly all datasets, but when all datasets were combined, their precision decreased. Our results caution that MSI tools can have much lower performance on datasets other than those on which they were originally evaluated, and in the case of RNA sequencing tools, can even perform poorly on the type of data for which they were created.

Keywords: cancer biomarker; benchmarking; microsatellite instability

Introduction

Microsatellite instability (MSI) is a phenomenon characterized by the accumulation of insertions and deletions (indels) in microsatellite regions found throughout the genome [1]. First described in hereditary non-polyposis colorectal cancer [2, 3], it has now been observed across multiple cancer types [4] and in both sporadic as well as familial cancers [5]. While it has yet to be demonstrated in a laboratory setting exactly how MSI arises, the predominant hypothesis is that defects in the DNA mismatch repair pathway can cause an increase in the number of indels at microsatellite sites [6]. This marked increase in the rate of indels is the primary characteristic of MSI and how it is primarily identified. The identification of MSI is important because cancers with high microsatellite instability (MSI-H) can be good candidates for immune checkpoint inhibitor treatment [7]. Historically, MSI-H has been inferred from the PCR amplification of five microsatellite markers [8, 9]. However, next-generation sequencing (NGS) allows the MSI status to be inferred using a larger number of loci and enables the determination of the MSI status to be incorporated into a comprehensive genome profiling pipeline [10].

A variety of computational tools and algorithms have been developed to determine the MSI status from NGS data. MSI tools here refers to programs that are available for use with NGS data and that can be quickly incorporated into a research or clinical bioinformatics pipeline whereas algorithms would require additional work before they could be used with NGS data. Tools differ primarily in the type of sequencing data used to determine the MSI status and use either DNA, RNA, or multi-omic sequencing data. The majority of tools use DNA sequencing data to compare the indels in microsatellites between a tumor and paired-normal sample. One example, and one of the first MSI tools, MSIsensor [11], uses a relatively simple algorithm that compares tumor and normal k-mer read length counts at each microsatellite and identifies microsatellite sites as unstable if they are significantly different in a χ^2 test. Other tools compare the gene expression values from RNA sequencing data to pre-trained baseline expression values. PreMSIm is an example of a tool that predicts the MSI status by using a k-nearest neighbors classification algorithm based on the expression of 15 genes [12]. Multi-omic tools use more complex algorithms. One example is DeltaMSI [13], which distinguishes unstable loci using a machine learning model built with both immunohistochemistry (IHC) data and NGS data. Other researchers have already created exhaustive lists of these tools and described their methods in detail [14–16], but, in general, all MSI tools classify samples as having MSI-H or as being microsatellite stable (MSS). The output of all tools is a value representing the level of MSI present in a sample – typically reported as the proportion of microsatellite sites that are unstable. Lastly, a threshold is picked to distinguish MSI-H from MSS.

Received: March 4, 2024. **Revised:** June 26, 2024. **Accepted:** July 25, 2024

Table 1. TCGA datasets

Project ID	Cancer	Sequencing	Number of Samples	Number of MSS	Number of MSI-H
COAD	Colon	WGS	56	46	10
ESCA	Esophageal	WGS	2	0	2
STAD	Stomach	WGS	136	107	29
UCEC	Uterine/endometrial	WGS	145	102	43
COAD	Colon	WXS	284	232	52
ESCA	Esophageal	WXS	3	0	3
READ	Rectum	WXS	3	0	3
STAD	Stomach	WXS	292	228	64
UCEC	Uterine/endometrial	WXS	268	196	72
COAD	Colon	RNA	280	230	50
ESCA	Esophageal	RNA	3	0	3
READ	Rectum	RNA	3	0	3
STAD	Stomach	RNA	272	213	59
UCEC	Uterine/Endometrial	RNA	268	196	72

The metadata for all TCGA datasets with paired MSI status (MSS is microsatellite stable, and MSI-H is high microsatellite instability). Whole genome sequencing and whole exome sequencing are abbreviated to WGS and WXS, respectively.

The authors of most tools provide a recommended threshold to distinguish MSI-H from MSS samples; however, the recommended settings and scope of MSI tools are sometimes unclear. For example, MSIsensor [11], was originally tested on whole exome sequencing (WXS) of 242 endometrial tissue samples from The Cancer Genome Atlas (TCGA). The latest version of MSIsensor comes with both WXS and whole genome sequencing (WGS) recommended settings despite having never been tested on WGS samples. This is also the case for the two successor tools MSIsensor-pro [17] and MSIsensor2 [18]. Two other widely used MSI tools, mSINGS [19] and MANTIS [20], both provide recommended thresholds to define MSI-H using WXS data but, like the other tools, it is not clear whether they are applicable to other sequencing types. This is also the case for a more recent tool, MSINGB [21], which does not require the user to set a threshold and simply outputs the status of the tumor sample. The only MSI tools that have been trained and tested on RNA sequencing data, PreMSIm [12] and MSIsensor-RNA [22], do not require the user to pick a threshold and, like MSINGB, will output the MSI status of a sample. They also indicate that they can be used with bulk and single-cell RNA sequencing, as well as microarray panels. However, there has not yet been an independent benchmarking of MSI NGS tools to verify the reported performance metrics of each tool and to determine the factors that affect those performance metrics.

Here we set out to address these issues by benchmarking the leading MSI tools on several datasets derived from different sequencing strategies. These included MSIsensor, which was among the first MSI tools to be created and which has also received FDA approval [11]; MSIsensor2 [18], a tool used to calculate an MSI score from the National Cancer Institute's Genomic Data Commons (GDC); MSIsensor-pro [17], which claims to improve upon the original MSIsensor by including a unique tumor-only algorithm with higher accuracy; mSINGS [19], the first tumor-only MSI tool and a tool that is commonly used with gene panel data; another heavily cited, paired-normal tool, MANTIS [20]; a very recent tool that uses somatic variant information to classify samples, MSINGB [21]; and two tools made solely for use with RNA sequencing data, PreMSIm [12] and MSIsensor-RNA [22]. By measuring tool performance on data derived from a broad range of sequencing methods, we aimed to replicate the high published performance metrics of each tool and determine which tools work best across sequencing types. Our results shed light on tool performance under optimal and non-optimal conditions and potential shortcomings in the underlying algorithms used to classify samples as MSI-H versus MSS.

Methods

Datasets

All TCGA WXS, WGS, and RNA sequencing data was downloaded in BAM format from the GDC Data Portal (https://gdc-portal.nci.nih.gov/) (Table 1). All WXS and WGS samples were subset down to only microsatellite regions to help with storage and processing time. We were able to obtain PCR MSI status for a total of 852 WXS and 321 WGS paired tumor-normal samples as well as 825 tumor-only RNA sequencing samples from the Broad Institute Genome Data Analysis Center (https://gdac.broadinstitute.org). Our TCGA sample list is based on those used in another publication from Cortes-Ciriano et al. [23], which has already compiled a list of all TCGA cohorts that have matched PCR MSI status. For the purposes of binary classification (used by all MSI tools), we treated MSI-L samples as MSS.

All gene panel samples were downloaded in FASTQ format from the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). We aligned the raw SRA FASTQ files to the same human reference genome used by TCGA (GRCh38.p14). The alignment for all samples was done using BWA [24]. In total, we collected 142 TSO-500 gene panel samples uploaded as part of three separate studies [25–27]. We also used 191 samples from the Oncomine 161-marker gene panel [28] and 178 samples from a 6-mononucleotide repeat paired-normal panel [29]. The TSO-500 and Oncomine samples also have an MSI score given by their commercial software, the Illumina TSO-500 pipeline and MSIcall, respectively. We used a threshold cut-off of 15 for the TSO-500 pipeline based on the current pre-print [30]. For the Oncomine samples, only MSI status as determined by MSIcall was given by the authors of the publication [28].

The other additional sequencing datasets we used for testing included DNA and RNA sequencing datasets. They were collected and managed the same way as the gene panel samples and were also from the SRA. However, for the RNA sequencing data, we used a different alignment software, STAR [31], with the same

Table 2. Additional non-TCGA datasets

Project ID	Cancer	Sequencing	Number of Samples	Number of MSS	Number of MSI-H
PRJNA629785	Colorectal	End-seq	34	7	27
PRJNA810563	Pan	6 Marker Panel	178	166	12
SRP008162	Prostate	T/O WXS	21	16	5
PRJNA727917	Colorectal	P/N WXS	21	0	21
PRJNA256024	Prostate	53 Marker Panel	43	30	13
PRJNA701182	Pan	161 Marker Panel	191	185	6
PRJNA841034	Gastric	TSO500	36	34	2
PRJEB57620	Male Breast	TSO500	14	14	0
PRJNA843231	Pan	TSO500	14	11	3
PRJNA748264	Colon	RNA	143	122	21

The metadata for all additional non-TCGA datasets with paired MSI status (MSS is microsatellite stable, and MSI-H is high microsatellite instability). The project ID is the searchable accession number for the Sequence Read Archive. MSI status establisher was the method used by the original publications to determine MSI status for each sample.

GRCh38 reference genome, and then created gene count matrices with FeatureCounts [32]. Gene counts then underwent $\log_2(n+1)$ normalization for use with MSIsensor-RNA and then also scaled to be between 0 and 1 for PreMSIm. The MSI status for each of these datasets was determined with either PCR or IHC, or based on the fact that the samples were from MSI-H tumor cell lines (Table 2). These datasets comprised 29 tumor-only WXS samples [33], 21 paired-normal WXS samples (Bioproject: PRJNA727917), 34 End-seq samples [34], and 143 tumor-only bulk RNA sequencing samples [35] (Table 2).

One tool, MSINGB, required extra data handling before it could be used. For TCGA WXS samples, Mutect2 VCF files were readily available and were downloaded from the GDC. For all other DNA sequencing datasets, we created VCF files following Mutect2 best practice settings [36]. It was run in paired-normal mode unless the sample lacked a paired-normal, then it was run with tumor-only mode. Lastly, we converted the VCF files to MAF format using vcf2maf [37].

MSI tools and software settings

Eight MSI tools were evaluated in total. These were: MSIsensor [11], MSIsensor2 [18], MSIsensor-pro [17], mSINGS [19], MANTIS [20], MSINGB [21], PreMSIm [12], and MSIsensor-RNA [22] (Table 3). Each tool was run with author recommended settings, though MANTIS and mINSGS did not provide WGS recommendations. For MANTIS to work with WGS we had to lower the minimum locus quality to 15, the minimum average per-base read quality to 10, and the minimum coverage threshold to 10. Without these adjustments, MANTIS could not find usable microsatellite sites. We did not adjust any settings for mSINGS or MSINGB and applied the same settings used with WXS. This was done because there were no parameters to help correct for the difference in sequencing depth. As we expected the coverage of microsatellites to be low in RNA sequencing data, we ran MSIsensor2, MSIsensor-pro, mSINGS, with the same settings used with WGS data.

Baselines and microsatellite target sites were created for each tool based on their included reference genome scanner. For all these functions we supplied each tool with the same reference genome. Microsatellite bed files were generated using the default setting for each tool. Baseline files were created by giving each tool the same 20 randomly sampled WXS, WGS, or RNA normal BAM files. We chose 20 because this number has been suggested as the minimum number of required normal files by the authors of MSI-sensor pro (https://github.com/xjtu-omics/msisensor-pro/wiki/Frequently-Asked-Questions), and other researchers have used a similar number of normal samples to create an MSIngs baseline [38–40]. MSIsensor-RNA required an additional seven MSI-H samples to be included in the baseline. For MSIsensor2, MSINGB, and PreMSIm, we used the hg38 baseline files included with these tools as they either lacked a baseline generator [18], or their baseline generator could not be used successfully [21]. We also removed samples that had fewer than five microsatellite sites that passed the thresholds of each tool.

We first determined a threshold cut-off for each tool to classify a sample as MSI-H or MSS then measured performance for each tool. The recommended threshold cut-offs were 3.5 for MSIsensor, 20 for MSIsensor2, 0.2 for mSINGS, and 0.4 for MANTIS. Unfortunately, the authors of MSI-sensor pro do not provide a recommended threshold. Lastly, we used a threshold score of 20 for MSIsensor2 based on the author's recommendations. Although the original MSIsensor publication recommended a threshold of 3.5, more recent publications have used a threshold of 10 [41,42]. We adjusted the threshold to 10 based on this. We also decided to use a threshold cut-off of 10 for MSIsensor-pro as it follows the same scoring system as its predecessor, MSIsensor.

We assessed how well each tool performed on each dataset by creating confusion matrices where the positive values were the MSI-H samples. Then we calculated precision, recall, specificity, and F1 score using equations outlined by Fawcett [43]. To view how these values change across different thresholds we created receiver operating characteristic (ROC) and precision-recall (PR) curves using the R packages MLeval [44], pROC [45], and caret [46]. We cross validated the ROC and PR curves using leave-one-out cross validation. All graphs were generated using R [47] and plotted with ggplot2 [48].

Lastly, we benchmarked the runtime and total memory used by each program on one random WXS, WGS, and RNA sample, where applicable, from TCGA. Runtime was measured using hyperfine [49] using the default settings except for the warmup parameter, which we set to 3, and the runtime of each tool was measured across 10 runs. We then measured total memory with the tool Massif available through Valgrind [50]. We set the pages-as-heap parameter to 'yes' which measures total heap memory used by a program, and the trace-children parameter to 'yes' which also measures the memory used by child processes. All benchmarks were done on a virtual machine running Ubuntu 20.04 LTS with 15 virtual CPUs and 65 gigabytes of RAM. While the virtual machine we created had multiple cores available, we chose to benchmark all tools in single-threaded mode. This is because only MANTIS, MSIsensor-pro, and MSIsensor-RNA were able to be run

Table 3. MSI tool summaries

Tool	Original evaluation data	Algorithm used for MSI detection	Output (MSI score)	Recommended threshold	Requires paired normal
MSIsensor	242 endometrial TCGA WXS samples	χ^2 test between tumor and normal read counts	Percent of unstable microsatellites	3.5	Yes
MSIsensor-pro	1532 pan-cancer TCGA WXS samples	Multinomial distribution model distinguishes MSI sites by comparing probability of polymerase slippage	Percent of unstable microsatellites	None	No
MSIsensor2	117 EGA samples and 10 TSO500 samples (TCGA also used but not numerically described)	Machine learning based (specifics not given)	Percent of unstable microsatellites	20	No
mSINGS	26 TCGA pan-cancer WXS and 298 pan-cancer gene panel samples	Read count differences between tumor sample and baseline normal	Fraction of unstable microsatellites	0.2	No
MANTIS	387 pan-cancer TCGA WXS samples	Absolute stepwise difference between tumor and normal read counts	Average aggregate instability	0.4	Yes
MSINGB	1432 pan-cancer TCGA WXS samples and 1055 pan-cancer non-TCGA WXS samples	NGBoost machine learning model based on somatic mutations	MSI status and probability of the classification	N/A (No score output)	No
PreMSIm	1383 pan-cancer TCGA RNA samples and 2006 gastric/colorectal microarray samples	K-nearest neighbors machine learning model based on gene expression	MSI status and probability of the classification	N/A (No score output)	No
MSIsensor-RNA	1428 pan-cancer TCGA RNA samples, 247 non-TCGA RNA samples, 1468 gastric/colorectal microarray samples, and 133 SC-RNA colorectal samples	Support vector machine learning classifier based on gene expression	MSI status and probability of the classification	N/A (No score output, but there are recommendations for feature selection thresholds)	No

A summary table of all the MSI tools used in this study. Whole genome sequencing and whole exome sequencing are abbreviated to WGS and WXS, respectively. Single-cell RNA sequencing was abbreviated to SC-RNA.

in multi-threaded mode, and we found that changing the multiple threads option either did not greatly improve the runtime of the program or caused it to crash before it could complete all 10 benchmarking runs. We chose not to include runtime or memory usage benchmarks for MSIsensor-RNA, PreMSIm, and MSINGB because they operate on prebuilt models making the time and memory used to run one sample negligible.

Results

MSI tools perform better on WXS than WGS samples

Using the published recommended settings most MSI tools performed better on WXS data than on WGS data (Tables 4 and 5, Fig. 1). The two exceptions were mSINGS and MSINGB, which had low performance metrics on the additional paired-normal and tumor-only WXS datasets (Table 5, Fig. 1). All the MSI tools showed good performance for TCGA WXS data, with again the exception of mSINGS, which had low recall and F1 scores (Table 4, Fig. 1). MSIsensor, MSIsensor-pro, MSIsensor2, and MANTIS all had high area under the curve (AUC) for both the ROC and PR curves (Fig. 2A, B). However, this was not always the case for the TCGA WGS data. Out of all MSI tools, only MSIsensor2 had high values for all performance metrics on WGS data (Table 4, Fig. 1). All other tools had one or more performance metrics substantially lower for WGS than for WXS data. The ROC and PR AUC values were also substantially lower for the WGS data than for the WXS data for all tools, except for MSIsensor2 and MANTIS (Fig. 2A, B, C, D). There were also large drop-offs in AUC when measured in ROC space versus PR space, implying tools might be missing more true positives (Fig. 2C, D). The most notable differences in ROC and PR AUC were seen with mSINGS on WXS data and with MSIsensor and MSIsensor-pro on WGS data (Fig. 2C, D).

MSI tools have widely different performance metrics depending on sequencing type and lack agreement on multiple sequencing types

To further evaluate tool performance, we merged the results of each tool across all datasets they could be tested on and calculated a confusion matrix (Supplementary Table 1, Fig. 1). Out of the paired-normal tools, MSIsensor had the highest performance metrics (Supplementary Table 1). MSIsensor2 and MSIsensor-pro both had the highest performance metrics when testing across all datasets., However, on the combined dataset, MSIsensor-pro had low recall (0.68) and MSIsensor2 had low precision (0.62). Both RNA sequencing tools had very high recall across all RNA

Table 4. A confusion matrix for each MSI tool on all TCGA samples

Dataset	Tool	Number of samples	Number of filtered samples	TP	FP	TN	FN	Precision	Recall	F1 score	Accuracy	Specificity
TCGA WXS	MSIsensor-Pro	852	0	130	0	658	64	1.000	0.670	0.802	0.925	1.000
TCGA WXS	MSIsensor	852	0	172	9	649	22	0.950	0.887	0.917	0.964	0.986
TCGA WXS	MSIsensor2	852	0	188	6	652	6	0.969	0.969	0.969	0.986	0.991
TCGA WXS	mSINGS	849	3	54	1	654	140	0.982	0.278	0.434	0.834	0.998
TCGA WXS	MANTIS	852	0	150	5	653	44	0.968	0.773	0.860	0.942	0.992
TCGA WXS	MSINGB	852	0	185	221	437	9	0.456	0.954	0.617	0.730	0.664
TCGA WGS	MSIsensor-Pro	342	0	1	1	257	83	0.500	0.012	0.023	0.754	0.996
TCGA WGS	MSIsensor	342	0	27	38	220	57	0.415	0.321	0.362	0.722	0.853
TCGA WGS	MSIsensor2	339	3	78	5	253	3	0.940	0.963	0.951	0.976	0.981
TCGA WGS	mSINGS	324	18	21	73	172	58	0.223	0.266	0.243	0.596	0.702
TCGA WGS	MANTIS	341	1	81	151	106	3	0.349	0.964	0.513	0.548	0.412
TCGA WGS	MSINGB	226	0	7	26	184	9	0.212	0.438	0.286	0.845	0.876
TCGA RNA	MSIsensor-pro	825	0	181	39	599	6	0.823	0.968	0.889	0.945	0.939
TCGA RNA	MSIsensor2	825	0	187	306	332	0	0.379	1.000	0.550	0.629	0.520
TCGA RNA	mSINGS	750	75	121	9	572	48	0.931	0.716	0.809	0.924	0.985
TCGA RNA	MSIsensor-RNA	825	0	169	484	154	18	0.259	0.904	0.402	0.392	0.241
TCGA RNA	PreMSIm	825	0	183	338	300	4	0.351	0.979	0.517	0.585	0.470

We abbreviated whole genome sequencing to WGS, whole exome sequencing to WXS, true positive to TP, false positive to FP, true negative to TN, and false negative to FN. Although no MSINGB samples were filtered for TCGA WGS data, 116 samples took too long to process and were left out of the final analysis.

sequencing datasets (0.91, 0.71), but all the other performance metrics were very low (Supplementary Table 1, Fig. 1).

To investigate the reduction in tool performance in WGS data relative to WXS data we compared MSI scores on 321 TCGA samples for which both types of data were available. In general, MSIsensor, MSIensor2, and MSIsensor-pro reported a higher proportion of unstable microsatellites for the WXS compared to the WGS data. In contrast, MANTIS and mSINGS scored samples higher for WGS versus WXS (Fig. 3). Interestingly, the tools that had high ROC and PR AUC scores on WGS samples, MANTIS and MSIsensor2, had significantly different MSI scores between WXS and WGS samples (Paired Wilcoxon rank sum test $P = 2.2 \times 10^{-16}$; $P = 1.3 \times 10^{-7}$; Fig. 3). This was also true to a lesser extent for MSIsensor-pro ($P = 0.006$), but it performed poorly on WGS data. MSIsensor and mSINGS did not have significantly different MSI scores between WGS and WXS samples ($P > 0.05$).

When we applied the tools on gene panel data, we obtained strikingly inconsistent results (Fig. 4A, B). For both the TSO-500 data and 161-marker panel there was only one MSI-H sample in common between all methods tested (Fig. 4A, B). The largest overlaps were observed between MSIsensor2 and mSINGS for the 161-marker panel (25 cases), and between MSIsensor2, MSIsensor-pro, and mSINGS in the case of the TSO-500 data (10 cases). Surprisingly, MSIsensor2 identified 126 unique MSI-H cases in the Oncomine data (Fig. 4A). Overall, there was very little agreement between the evaluated tools and the commercial software included with the gene panels.

Variation in performance by tool, threshold, and across data sets

To determine how well the author recommended WXS settings carry over to other sequencing types, we calculated F1 scores over varying thresholds for each tool across datasets that these tools would have no training exposure to (Fig. 5). All MSI tools were able to achieve a nearly perfect F1 score for all datasets if the correct optimal threshold for that dataset was used (aside from mSINGS on the 6-mononucleotide panel dataset). The only tool that demonstrated high F1 scores across a wide range of thresholds on all datasets was MSIsensor2. By contrast, MSIsensor-pro required very low and dataset-specific threshold values to achieve a high F1 score (Fig. 5). This result was also the case for accuracy, specificity, recall, and precision (Supplementary Figs 2–5).

The published thresholds for MSIsensor, mSINGS, and MANTIS on WXS provided good performance on the additional paired-normal WXS dataset (Fig. 5). However, achieving optimal performance, as measured by the F1 score, on the tumor only WXS dataset with mSINGS would require the use of a very different threshold (close to 0.5, compared to the recommended threshold of 0.2). Our results suggest that different thresholds may be required for use with different sequencing data types.

MSI tools designed for DNA sequencing perform better than MSI tools designed for RNA sequencing on RNA sequencing datasets

We also evaluated the performance of tools designed for DNA when applied to RNA sequencing data. The two tools designed for use with RNA sequencing data, PreMSIm and MSIsensor-RNA, had low-performance metrics on the two RNA sequencing datasets used in our study (Fig. 1). While they could identify the MSI-H cases in the dataset well (recall ≥0.9), all other performance metrics were low (Tables 4 and 5). Interestingly, two of the tools designed for DNA sequencing, MSIsensor-pro and mSINGS had

Table 5. Confusion matrix for all additional datasets

Dataset	Tool	Number of samples	Number of filtered samples	TP	FP	TN	FN	Precision	Recall	F1 score	Accuracy	Specificity
T/O WXS	MSIsensor-Pro	32	0	5	0	22	5	1.000	0.500	0.667	0.844	1.000
T/O WXS	MSIsensor2	32	0	10	0	22	0	1.000	1.000	1.000	1.000	1.000
T/O WXS	mSINGS	32	0	10	22	0	0	0.312	1.000	0.476	0.312	0.000
T/O WXS	MSINGB	30	2	3	4	16	7	0.429	0.300	0.353	0.633	0.800
P/N WXS	MSIsensor-Pro	21	0	12	0	0	9	1.000	0.571	0.727	0.571	NA
P/N WXS	MSIsensor	21	0	20	0	0	1	1.000	0.952	0.976	0.952	NA
P/N WXS	MSIsensor2	21	0	20	0	0	1	1.000	0.952	0.976	0.952	NA
P/N WXS	mSINGS	19	2	2	0	0	17	1.000	0.105	0.190	0.105	NA
P/N WXS	MANTIS	21	0	20	0	0	1	1.000	0.952	0.976	0.952	NA
P/N WXS	MSINGB	21	0	21	0	0	0	1.000	1.000	1.000	1.000	NA
End-seq	MSIsensor-Pro	36	0	29	1	6	0	0.967	1.000	0.983	0.972	0.857
End-seq	MSIsensor2	19	17	15	1	3	0	0.938	1.000	0.968	0.947	0.750
End-seq	mSINGS	36	0	25	6	1	4	0.806	0.862	0.833	0.722	0.143
End-seq	MSINGB	36	0	29	6	1	0	0.829	1.000	0.906	0.833	0.143
6 marker	MSIsensor-Pro	178	0	6	0	166	6	1.000	0.500	0.667	0.966	1.000
6 marker	MSIsensor	178	0	12	6	160	0	0.667	1.000	0.800	0.966	0.964
6 marker	MSIsensor2	174	4	12	0	162	0	1.000	1.000	1.000	1.000	1.000
6 marker	mSINGS	178	0	12	2	164	0	0.857	1.000	0.923	0.989	0.988
6 marker	MANTIS	178	0	3	4	162	9	0.429	0.250	0.316	0.927	0.976
6 marker	MSINGB	174	4	6	87	75	6	0.065	0.500	0.114	0.466	0.463
RNA	MSIsensor-Pro	142	3	17	3	118	4	0.850	0.810	0.829	0.951	0.975
RNA	MSIsensor2	142	3	20	89	32	1	0.183	0.952	0.308	0.366	0.264
RNA	mSINGS	141	4	16	3	118	4	0.842	0.800	0.821	0.950	0.975
RNA	PreMSIm	140	5	19	67	52	2	0.221	0.905	0.355	0.507	0.437
RNA	MSIsensor-RNA	143	2	21	118	4	0	0.151	1.000	0.262	0.175	0.033

A confusion matrix for each MSI tool on all additional samples. We abbreviated whole genome sequencing to WGS, whole exome sequencing to WXS, true positive to TP, false positive to FP, true negative to TN, and false negative to FN.

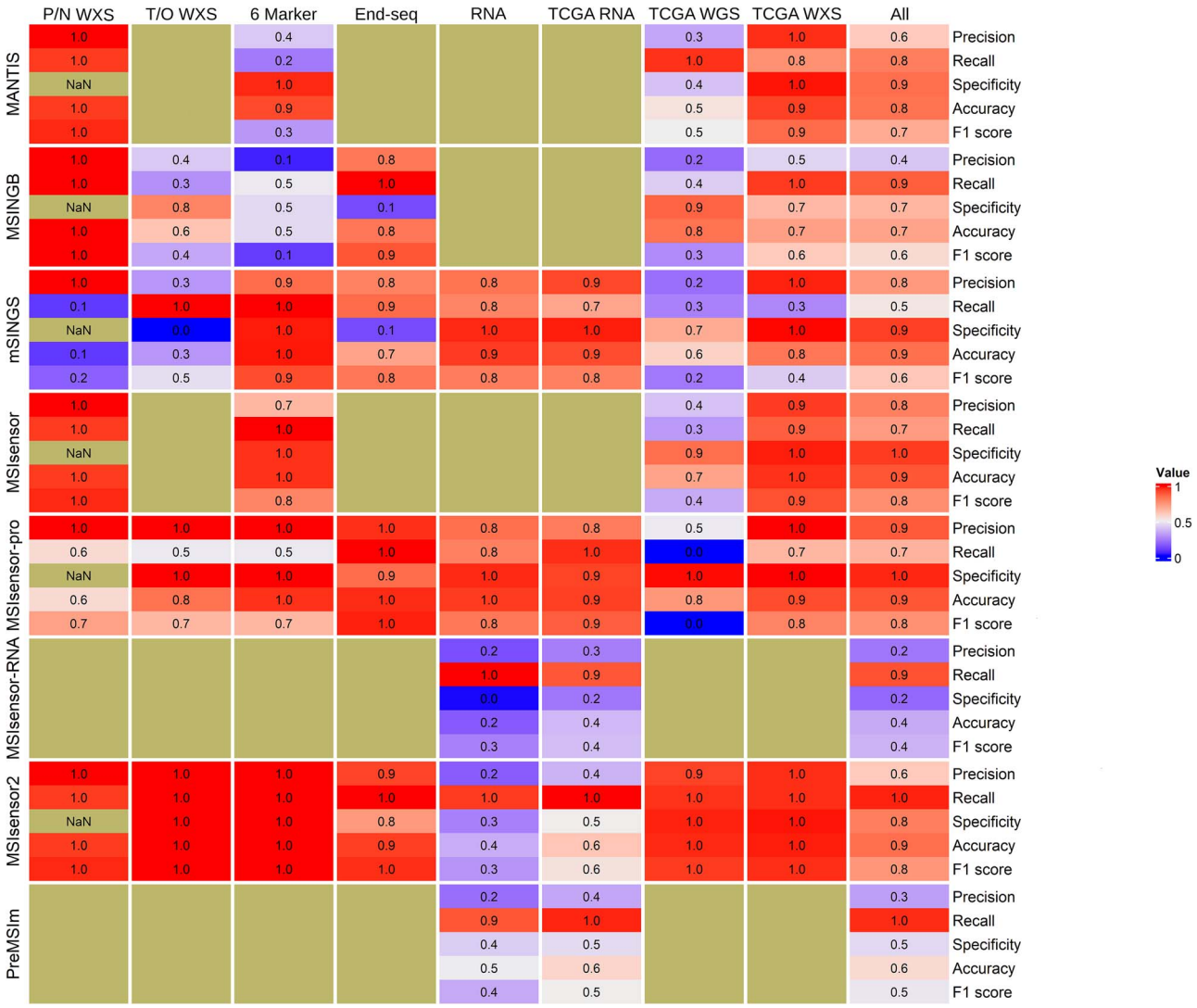

Figure 1. Performance metric heatmap. A heatmap showing all MSI tools and their performance across all datasets where a confusion matrix was created. The author recommended thresholds were used for each tool that provided one. Black tiles are NA values, and black and white striped tiles are instances where the metric could not be calculated. P/N WXS is the additional paired-normal whole exome sequencing dataset, T/O WXS is the additional tumor-only whole exome sequencing dataset. 6 marker stands for the 6-mononucleotide panel. TCGA WGS and WXS are the datasets comprised of whole exome and WGS data from The Cancer Genome Atlas, respectively. The 'All' column is the merged results for each tool.

high F1 scores on both the TCGA RNA sequencing data (0.81–0.90) (Table 4) and on the additional bulk RNA sequencing dataset (0.83–0.84) (Table 5). Similar to the RNA sequencing tools, MSIsensor2 had high recall scores (0.9–1.0) but very low precision scores (0.18–0.37) (Tables 4 and 5). When tested across many thresholds with ROC and PR curves, all three DNA sequencing tools had high values for both ROC-AUC (0.97–.99) and PR-AUC (0.92–0.99) (Supplementary Fig. 1A, B).

Tools vary greatly in runtime and memory usage

Lastly, we measured average runtime across 10 runs and total memory usage for all tools. We found that there were substantial differences in runtime and memory usage between tools (Supplementary Fig. 6, Supplementary Tables 2 and 3). Most notably, MSIsensor2 and MSIsensor-pro were much faster than the other tools, taking an average of 115 and 90.86 seconds, respectively, to process a WGS sample. While the other tools took upwards of 3800 seconds to process both WXS and WGS samples, MANTIS had the worst runtime, taking an average of 14 111 seconds to process a WGS sample and 6410 seconds to process a WXS sample. Although it had the worst average runtime, MANTIS was by far the most efficient in terms of memory usage, using only 23.34 megabytes to process a WGS sample (Supplementary Table 3). All other tools had similar memory usage to one another, except for mSINGS, which required over 30 gigabytes of memory to process either a WGS or WXS sample.

Discussion

We chose eight MSI tools to benchmark. Out of the many tools that are available, these eight have common pipelines making them easy to compare and run in a parallel workflow. Many other MSI NGS methods exist, but we found they were either lacking a downloadable tool [23,51,52], they were deprecated [53], or they did not output an MSI score or status [54]. While we chose to add RNA sequencing and two tools that predict MSI status based on expression, we did so because the other tools used in our study were compatible with these data types. We chose not to

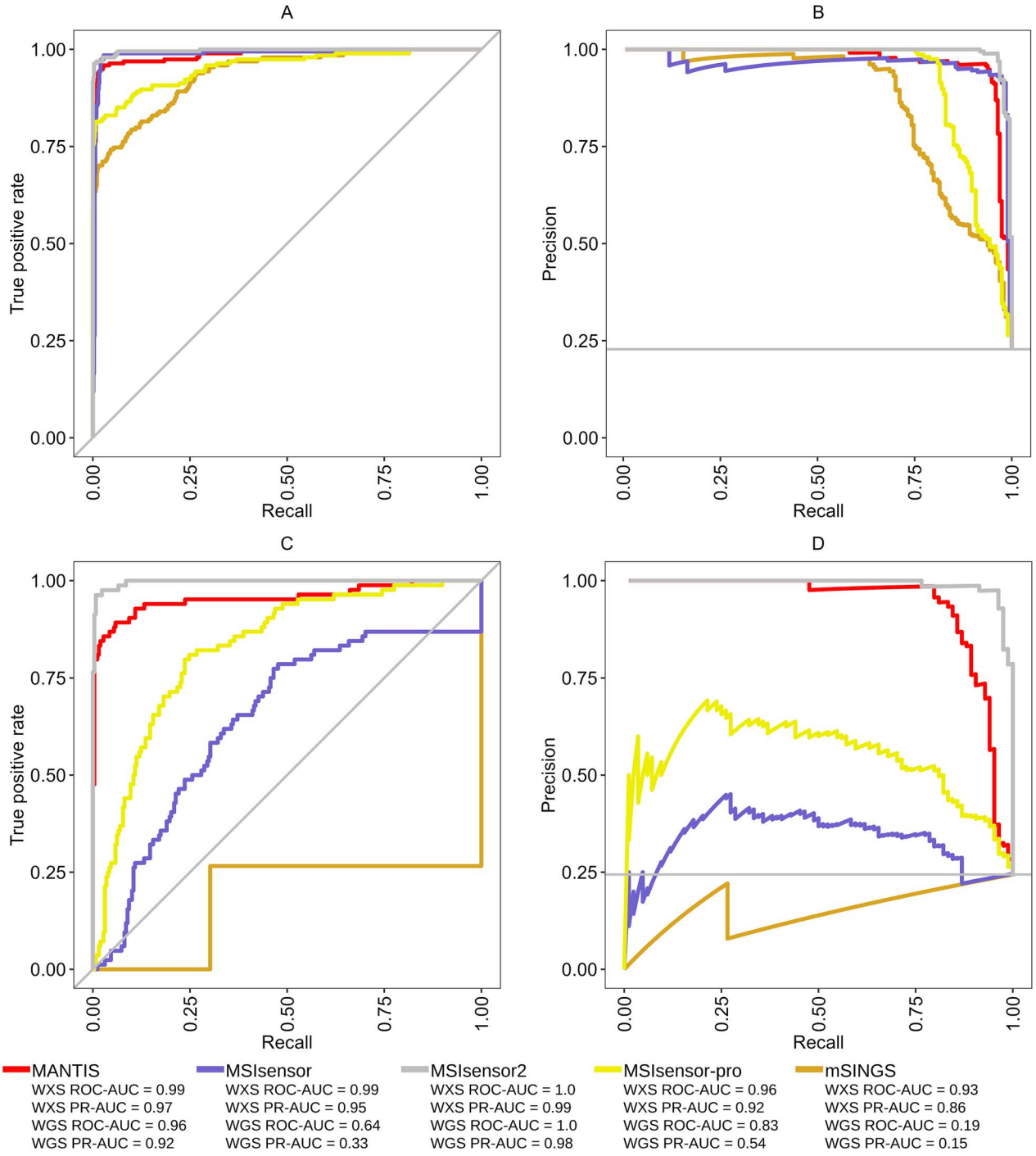

Figure 2. ROC and PR curves for all TCGA samples. All ROC curves and PR curves for TCGA WXS (A, B) and WGS (C, D) samples.

include additional, multi-omic approaches [13, 55, 56] as the data they require is difficult to obtain in large numbers from publicly available repositories.

Although the developers of several MSI tools have reported very high performance metrics, we found that performance can be severely affected by sequencing type and the choice of threshold used to classify samples as MSI-H. Many MSI tools were originally tested using WXS data from TCGA and, indeed, on these data tools often achieved high performance; however, they had worse performance metrics on WGS data (Figs 1, 2A, B, C, D). Further, MSI tools scored the same TCGA individuals differently with WXS compared to WGS data, with some having significantly different MSI scores (Fig. 3). There was also an overall lack of tool agreement on MSI-H status with gene panel data, and some tools required specific thresholds for different sequencing types to achieve a high F1 score. Lastly, we found tools designed for RNA sequencing data not only had low performance metrics on RNA sequencing datasets, but that some tools developed for DNA sequencing data had higher performance metrics on these data.

The authors of multiple tools have suggested they can be used with a variety of sequencing types [17–20], but only one includes more than one sequencing type in its original publication [18]. The

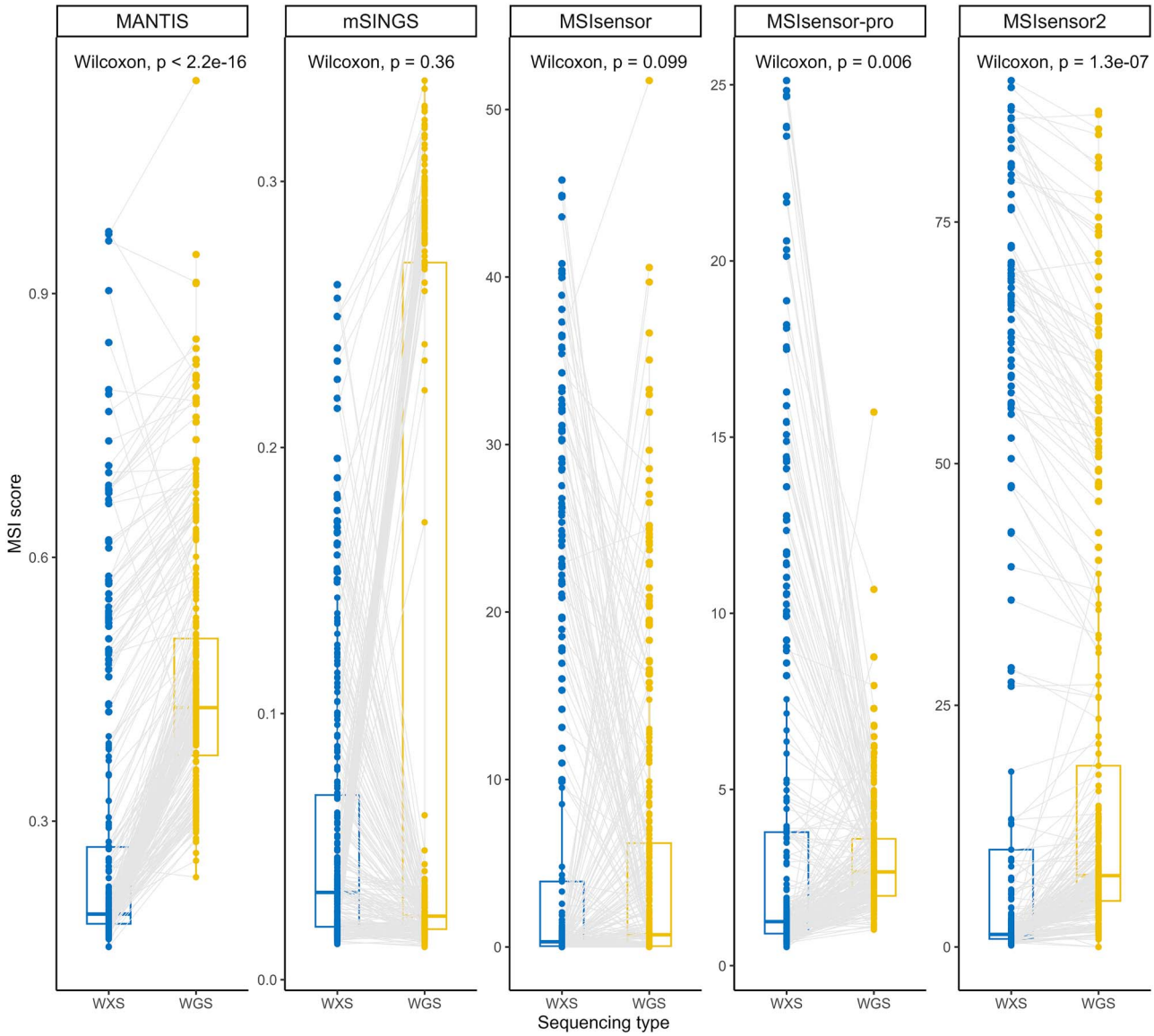

Figure 3. Boxplots of MSI scores for TCGA samples. Box plots of MSI scores for all TCGA samples that had both WGS and WXS samples. Gray lines were drawn between samples that have the same TCGA case ID. Paired Wilcox rank sum test values were calculated between the distribution of WXS and WGS scores for each tool.

lack of validation on different sequencing methods can be problematic as we have shown that some of the tools are not suitable for use with WGS. This is highlighted primarily by the reduction in ROC-AUC and PR-AUC when compared to the TCGA WXS results. The worst performer on WGS data, mSINGS, had an abnormal looking ROC curve due to its MSI scores being noninformative for the data. The horizontal and vertical bars seen, as opposed to usual jagged cut-points, are the result of no increase in true positive rate despite change in threshold and sudden increase in true positive rate, respectively.

The differences in tool performance on WGS and WXS datasets are likely attributable to the number of microsatellites included in the analysis and the overall coverage of microsatellites. The trade-off between these two sequencing strategies is that many more microsatellites are sequenced with WGS, but the sites sequenced with WXS are at a much higher coverage. It could be the case that simply including fewer microsatellites in the analysis while sequencing them at a much higher coverage is a better strategy for detecting MSI. Another factor to consider is that exonic regions contain a higher proportion of trinucleotide and hexanucleotide repeat types, whereas intronic regions contain more of the other repeat types [57]. There is the possibility that microsatellites in exonic regions are more indicative of MSI than intronic regions, but the current paradigm suggests mononucleotide and dinucleotide microsatellites are the most sensitive for MSI detection [58]. From our results, two tools, MANTIS and MSIsensor2 had high ROC and PR AUC scores for both WGS and WXS datasets. MANTIS uses a more complex aggregate scoring approach that calculates normalized distance values as opposed to the locus-by-locus strategy of the other MSI tools. Alongside its aggregate approach to determining MSI status, it includes additional filtering steps, which could potentially help explain its success with WGS data.

While MSIsensor2 performed well on TCGA WGS and WXS data, it failed to process most End-seq samples because no microsatellite sites passed the filter thresholds. In principle, End-seq should be a compatible sequencing method with this tool because it is a genome-wide sampling scheme like WGS that works well with repair-deficient genotypes [59] and it has been

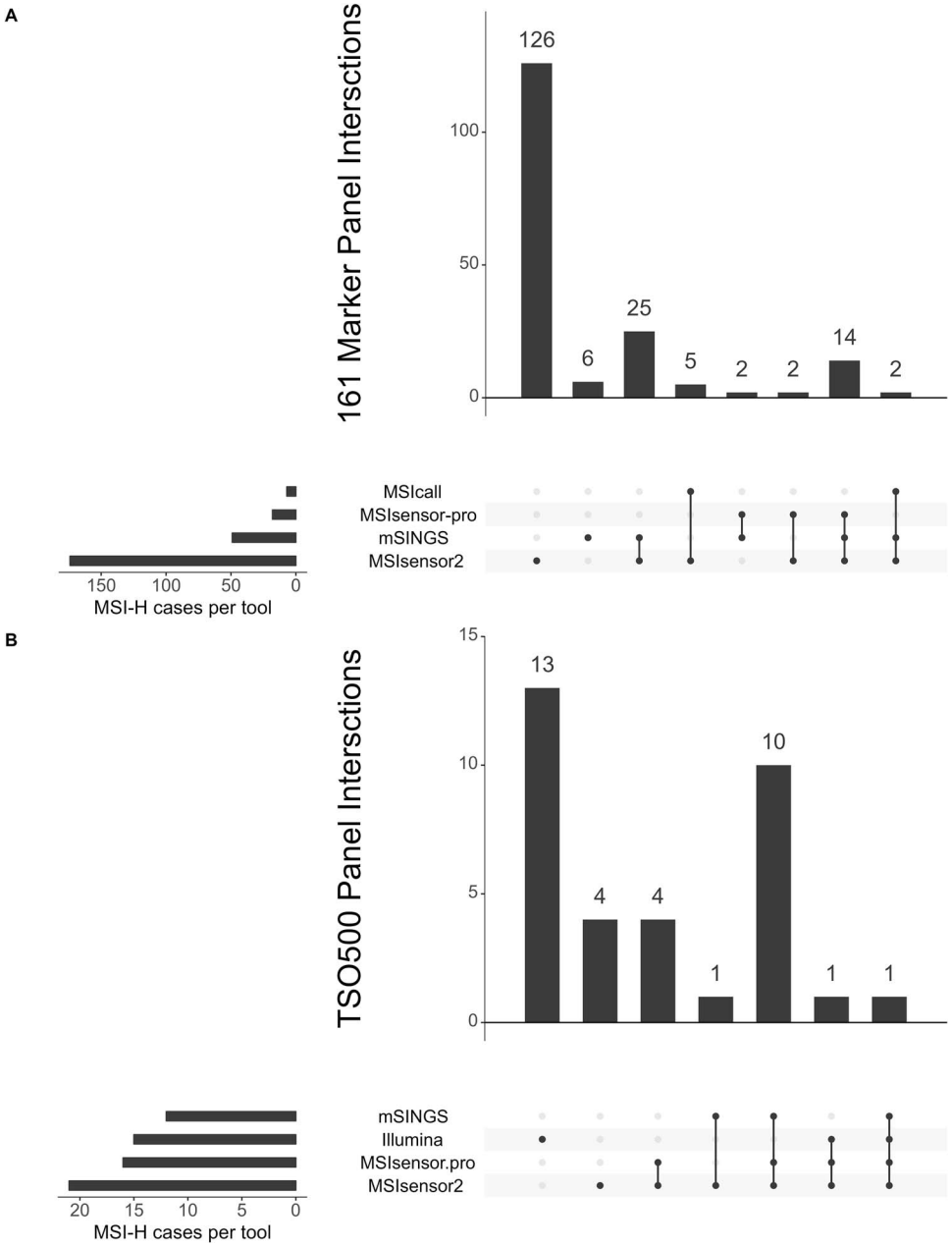

Figure 4. Upset plots for MSI tools on gene panel data. Upset plots showing intersections of MSI-H cases for each tool on 161-marker, Oncomine gene panel data (A) and TSO-500 panel data (B). Illumina and MSIcall are the MSI commercial for software results for the TSO-500 gene panel and 161-marker panel datasets, respectively. Connected dots show tools that agree on an MSI-H case whereas single dots are unique MSI-h cases for that tool. Vertical bars represent the total number of MSI-H cases that are agreed upon by one or more tools, and horizontal bars show the total number of MSI-H cases called by a tool.

shown to have similar read depth to WXS with upwards of 100x coverage [34]. However, the poor performance of MSIsensor2 on this sequencing type suggests caution is needed when using MSI tools on novel datasets.

Aside from the End-seq data, MSIsensor2 showed the best performance across all other datasets. This is surprising, as MSIsensor2 is unpublished and seems to only have been used by the GDC for 'bioinformatics-derived' MSI status (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/). Unfortunately, there is very limited information on the machine learning algorithm used by MSIsensor2 and no references are provided to the samples that have been used to train the hg38 models they provide [18]. Because of this, the high performance metrics seen in our results should be treated with caution. Another striking result we found with MSIsensor2 was the number of unique MSI-H cases identified in the 161-marker gene panel data. The highest estimates of MSI status come from colorectal cancers where occurrence is seen in ~15% of cancers [60]. Therefore, the number of MSI-H cases reported by MSIsensor2 is unlikely to be close to the correct value, as it classified ~74% of these 191 samples as MSI-H. Although the results obtained by MSIsensor2 on all other datasets were promising, this aberrant result is concerning and suggests that further testing on additional data types may be required.

Importantly, this benchmarking analysis highlights the risks of establishing recommended settings for MSI tools based on

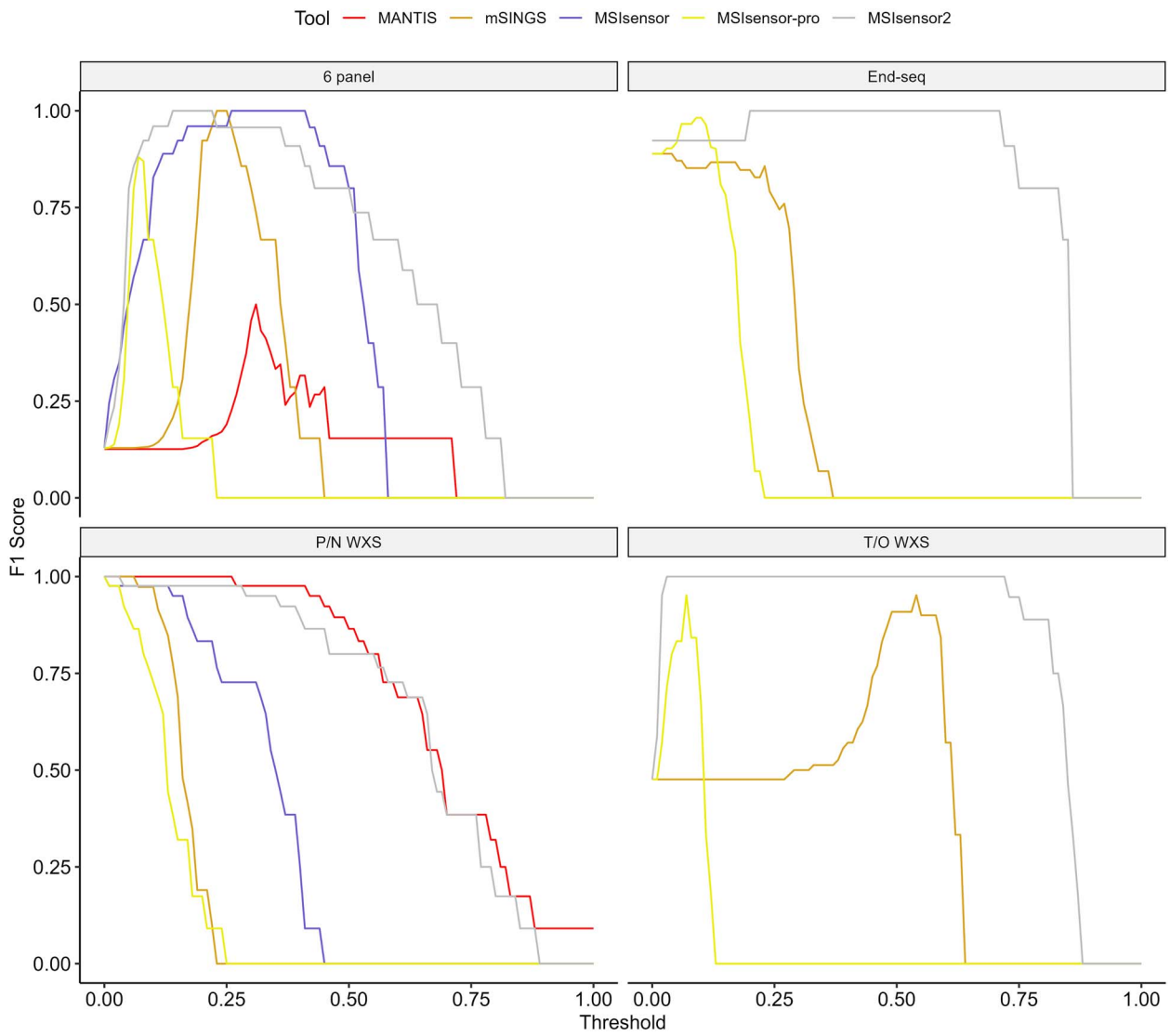

Figure 5. MSI tool performance on additional datasets by threshold. Graphs showing measures of performance (F1 scores) plotted across scaled MSI tool thresholds for each tool on additional datasets. These datasets included paired-normal whole exome sequencing data (P/N WXS), tumor-only whole exome sequencing data (T/O WXS), data from a 6-microsatellite marker panel (6 panel), and end-seq data. Thresholds were put on the same scale by converting MSI-sensor, MSI-sensor pro, and MSI-sensor 2 from percent to decimal.

relatively narrow training cohorts. For example, MSIsensor established its recommended MSI-H threshold based on only 242 endometrial samples [11] (Table 3). While we were able to replicate the ROC-AUC values of these tools on TCGA WXS data, we found the optimal F1 score impossible or difficult to achieve with recommended threshold settings. This has led to situations where researchers try to account for this by simply picking the optimal threshold for their specific dataset [61], or where researchers report difficulty in determining a clear threshold for MSI scores at all [62]. Ideally, thresholds to determine MSI-H status for a given sequencing type should be validated with a large and diverse cohort as was done for MSI-sensor on MSK-impact gene panels [63], which serves as a good example of how to integrate these tools into a precision medicine setting.

Conclusion

Several computational tools have been developed to classify NGS samples as MSI-H or MSS and they have generally been reported to achieve high performance. However, we have found that there are potentially serious issues affecting the reliability of these tools, particularly when applied across diverse sequencing data types. Notably, there was a large drop in performance on WGS relative to WXS samples. Tools were also shown to lack agreement on gene panel samples both with one another and with commercially available software, and that optimal thresholds can change with sequencing type. Lastly, we showed that tools designed for use with RNA sequencing data had poor performance metrics on bulk RNA test datasets and were even outperformed by some MSI tools meant to be used with DNA sequencing data.

Two MSI tools stood out from the others, MSIsensor2 and MANTIS. Both performed very well across most datasets, but the methodology of MSIsensor2 has yet to be published and there is currently no adequate information available on the algorithm it uses. When looking across all datasets the performance metrics of both tools were much lower. It also could not process most of the End-seq samples and classified a very large and implausible number of samples as MSI-H in a 161-marker panel dataset. MANTIS on the other hand did not achieve high performance metrics on a 6-marker panel dataset, still requires a

paired-normal sample, and has a slow runtime in comparison to other MSI tools, limiting its applicability and rendering it hard to integrate into some genome profiling pipelines.

We have shown evidence that when tested outside their optimal settings, MSI NGS tools can have difficulty replicating the performance originally seen on TCGA WXS data. This problem is further compounded by a lack of clear best practice settings, and until more MSI NGS datasets are publicly available for additional testing, it is difficult to validate the applicability of MSI tools on additional sequencing types. Altogether, this research highlights several concerns relating to MSI tools used with NGS data and suggests that, despite the high-performance metrics reported for several tools, there remains a need for a rigorously evaluated tool that can be reliably applied to a wide range of sequencing data types.

Key Points

- Tools designed to determine MSI status performed worse on whole genome sequencing data than on whole exome sequencing data.
- Two tools, MSI-sensor 2 and MANTIS, had excellent performance across all datasets, but each comes with its own drawbacks.
- Optimal choice of MSI-H threshold depends strongly on the data type.
- Tools that infer the MSI status based on read counts performed better on RNA sequencing data than tools that use gene count matrices to infer the MSI status.

Supplementary data

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

Acknowledgements

We would like to thank all other members of the Seoighe lab, Tyler Medina, Mehak Chopra, Sophie Matthews, Dónal O'Shea, Brian O'Sullivan, Noor Khehrah, Siobhan Cleary, and Declan Bennett for their useful input on our research ideas. Additionally, this work was presented to and influenced by members of the Science Foundation Ireland Centre for Research Training in Genomic Data Science program, attendees of the Virtual Institute of Bioinformatics and Evolution conference, and members of the European Association for Cancer Research. All of these peers as well as three anonymous peer reviewers provided valuable feedback that significantly improved our work.

Conflict of interest: None declared.

Funding

This research was made possible with the financial support of Science Foundation Ireland under Grant number 18/CRT/6214.

Data availability

All data used in this research are available from the GDC data portal (https://portal.gdc.cancer.gov/) or from the SRA archive (https://www.ncbi.nlm.nih.gov/sra). All related project ID's or accession numbers are included in Tables 1 and 2 of this research article. The code used to generate results, figures, and tables in this research article are available from our GitHub (https://github.com/harrison-anth/benchmark_msi).

References

1. Ionov Y, Peinado MA, Malkhosyan S. *et al.* Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nat Cell Biol* 1993;**363**(June):558–61. https://doi.org/10.1038/363558a0.
2. Thibodeau SN, Bren G, Schaid D. Microsatellite instability in cancer of the proximal colon. *Science*. 1993;**260**(5109):816–9. https://doi.org/10.1126/science.8484122.
3. Aaltonen LA, Peltomäki P, Leach FS. *et al.* Clues to the pathogenesis of familial colorectal cancer. *Science*. 1993;**260**(5109):812–6 https://doi.org/10.1126/science.8484121.
4. Arzimanoglou II, Gilbert F, Barber HRK. Microsatellite instability in human solid tumors. *Cancer* 1998;**82**(10):1808–20. https://doi.org/10.1002/(SICI)1097-0142(19980515)82:10<1808::AID-CNCR2>3.0.CO;2-J.
5. Lawes DA, SenGupta S, Boulos PB. The clinical importance and prognostic implications of microsatellite instability in sporadic cancer. *Eur J Surg Oncol* 2003;**29**(3):201–12. https://doi.org/10.1053/ejso.2002.1399.
6. Pećina-Šlaus N, Kafka A, Salamon I. *et al.* Mismatch repair pathway, genome stability and cancer. *Front Mol Biosci* 2020;**7**(June):1–12. https://doi.org/10.3389/fmolb.2020.00122.
7. Lee V, Murphy A, Le DT. *et al.* Mismatch repair deficiency and response to immune checkpoint blockade. *Oncologist* 2016;**21**(10):1200–11. https://doi.org/10.1634/theoncologist.2016-0046.
8. Murphy KM, Zhang S, Geiger T. *et al.* Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers. *J Mol Diagnostics* 2006;**8**(3):305–11. https://doi.org/10.2353/jmoldx.2006.050092.
9. Berg KD, Glaser CL, Thompson RE. *et al.* Detection of microsatellite instability by fluorescence multiplex polymerase chain reaction. *J Mol Diagnostics* 2000;**2**(1):20–8. https://doi.org/10.1016/S1525-1578(10)60611-3.
10. Bonneville R, Krook MA, Chen HZ. *et al.* Detection of microsatellite instability biomarkers via next-generation sequencing. *Methods Mol Biol* **2020**;2055(2017):119–32 https://doi.org/10.1007/978-1-4939-9773-2_5.
11. Niu B, Ye K, Zhang Q. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* 2014;**30**(7):1015–6. https://doi.org/10.1093/bioinformatics/btt755.
12. Li L, Feng Q, Wang X. PreMSIm: an R package for predicting microsatellite instability from the expression profiling of a gene panel in cancer. *Comput Struct Biotechnol J* 2020;**18**:668–75. https://doi.org/10.1016/j.csbj.2020.03.007
13. Swaerts K, Dedeurwaerdere F, De Smet D, De Jaeger P, Martens GA. DeltaMSI: artificial intelligence-based modeling of microsatellite instability scoring on next-generation sequencing data. *BMC Bioinformatics* 2023;**24**(1):1–14. https://doi.org/10.1186/s12859-023-05186-3
14. 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.
15. Yamamoto H, Imai K. An updated review of microsatellite instability in the era of next-generation sequencing and precision medicine. *Semin Oncol* 2019. **46**(3):261–70 https://doi.org/10.1053/j.seminoncol.2019.08.003.

16. Yu F, Makrigiorgos A, Leong KW. *et al.* Sensitive detection of microsatellite instability in tissues and liquid biopsies: recent developments and updates. *Comput Struct Biotechnol J* 2021;**19**: 4931–40. https://doi.org/10.1016/j.csbj.2021.08.037.
17. Jia P, Yang X, Guo L. *et al.* MSIsensor-pro: fast, accurate, and matched-normal-sample-free detection of microsatellite instability. *Genom Proteom Bioinform* 2020;**18**(1):65–71. https://doi.org/10.1016/j.gpb.2020.02.001
18. Niu BO, MSI-sensor 2 [Internet]. *GitHub*; 2024. Available from: https://github.com/niu-lab/msisensor2
19. Salipante SJ, Scroggins SM, Hampel HL. *et al.* Microsatellite instability detection by next generation sequencing. *Clin Chem* 2014; **60**(9):1192–9. http://www.clinchem.org/content/vol60/issue9.
20. Kautto EA, Bonneville R, Miya J. *et al.* Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* 2017;**8**(5):7452–63. https://doi.org/10.18632/oncotarget.13918.
21. Chen J, Wang M, Zhao D. *et al.* MSINGB: a novel computational method based on NGBoost for identifying microsatellite instability status from tumor mutation annotation data. *Interdiscip sci – Comput life sci* 2023;**15**(1):100–10. https://doi.org/10.1007/s12539-022-00544-w.
22. Jia P, Yang X, Yang X. *et al.* MSIsensor-RNA: microsatellite instability detection for bulk and single-cell gene expression data. *Genom Proteom Bioinform* 2024. https://doi.org/10.1093/gpbjnl/qzae004.
23. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* 2017;**8**:1–12. https://doi.org/10.1038/ncomms15180
24. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
25. Hsieh Y, Kuo W, Hsu W. *et al. Cancers (Basel)* 2022;**15**(1):269. https://doi.org/10.3390/cancers15010269.
26. Pestinger V, Smith M, Sillo T. *et al.* Use of an integrated pan-cancer oncology enrichment next-generation sequencing assay to measure tumour mutational burden and detect clinically actionable variants. *Mol Diagnosis Ther* 2020;**24**(3):339–49. https://doi.org/10.1007/s40291-020-00462-x
27. Valentini V, Silvestri V, Bucalo A. *et al.* Molecular profiling of male breast cancer by multigene panel testing: implications for precision oncology. *Front Oncol* 2023;**12**(January):1–12. https://doi.org/10.3389/fonc.2022.1092201.
28. Li S, Wang B, Chang M. *et al.* A novel algorithm for detecting microsatellite instability based on next-generation sequencing data. *Front Oncol* 2022;**12**(June):1–8 https://doi.org/10.3389/fonc.2022.916379.
29. Zhao C, Jiang T, Ju JH. *et al.*. TruSight Oncology 500: enabling comprehensive genomic profiling and biomarker reporting with targeted sequencing. *bioRxiv* [Internet]. 2020. Available from: https://www.biorxiv.org/content/early/2020/10/22/2020.10.21.349100.
30. Özdoğan M, Papadopoulou E, Tsoulos N. *et al.* Comprehensive tumor molecular profile analysis in clinical practice. *BMC Med Genomics* 2021;**14**(1):1–21. https://doi.org/10.1186/s12920-021-00952-9
31. Dobin A, Davis CA, Schlesinger F. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
32. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.
33. Lazzari L, Corti G, Picco G. *et al.* Patient-derived xenografts and matched cell lines identify pharmacogenomic vulnerabilities in colorectal cancer. *Clin Cancer Res* 2019;**25**(20):6243–59. https://doi.org/10.1158/1078-0432.CCR-18-3440.
34. van Wietmarschen N, Sridharan S, Nathan WJ. *et al.* Repeat expansions confer WRN dependence in microsatellite-unstable cancers. *Nature* 2020;**586**(7828):292–8. https://doi.org/10.1038/s41586-020-2769-8.
35. Park DY, Choi C, Shin E. *et al.* NTRK1 fusions for the therapeutic intervention of Korean patients with colon cancer. *Oncotarget* 2016;**7**(7):8399–412. https://doi.org/10.18632/oncotarget.6724.
36. Van der Auwera GA, Carneiro MO, Hartl C. *et al.* From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;**43**(SUPL.43):11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43.
37. Kandoth C, Gao J, Qwangmsk Mattioni M. *et al.* mskcc/vcf2maf: vcf2maf v1.6.16 [Internet]. Zenodo; 2018. Available from: https://doi.org/10.5281/zenodo.1185418
38. Zhao L, Shan G, Li L. *et al.* A robust method for the rapid detection of microsatellite instability in colorectal cancer. *Oncol Lett* 2020;**20**(2):1982–8 https://doi.org/10.3892/ol.2020.11702.
39. Kuo AJ, Paulson VA, Hempelmann JA. *et al.* Validation and implementation of a modular targeted capture assay for the detection of clinically significant molecular oncology alterations. *Pract Lab Med*. 2020;**19**(December 2019):e00153. https://doi.org/10.1016/j.plabm.2020.e00153
40. Lee Y, Lee JA, Park HE. *et al.* Targeted next-generation sequencing-based detection of microsatellite instability in colorectal carcinomas. *PloS One* 2021;**16**(2):1–12. https://doi.org/10.1371/journal.pone.0246356.
41. Stadler ZK, Battaglin F, Middha S. *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.
42. Nacev BA, Sanchez-Vega F, Smith SA. *et al.* Clinical sequencing of soft tissue and bone sarcomas delineates diverse genomic landscapes and potential therapeutic targets. *Nat Commun* 2022 Jun;**13**(1):3405.
43. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;**27**(8):861–74 https://doi.org/10.1016/j.patrec.2005.10.010.
44. Christopher M, John R. *Package "MLeval" Machine Learning Model Evaluation*. 2022. Available from: https://cran.r-project.org/package=MLeval
45. Robin X, Turck N, Hainard A. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**(1):77. https://doi.org/10.1186/1471-2105-12-77.
46. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008;**28**(5):1–26 https://www.jstatsoft.org/index.php/jss/article/view/v028i05.
47. R Core Team (2020)Development Core Team. *A Language and Environment for Statistical Computing*, Vol. **3**. Vienna, Austria: R Foundation for Statistical Computing, 2020. Available from: https://www.R-project.org
48. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016. Available from: https://ggplot2.tidyverse.org.
49. Peter D. Hyperfine [Internet]. 2023. Available from: https://github.com/sharkdp/hyperfine

50. Nethercote N, Seward J. Valgrind: a framework for heavyweight dynamic binary instrumentation. *ACM SIGPLAN Not* 2007;**42**(6): 89–100 https://doi.org/10.1145/1273442.1250746.
51. Nowak JA, Yurgelun MB, Bruce JL. *et al.* Detection of mismatch repair deficiency and microsatellite instability in colorectal adenocarcinoma by targeted next-generation sequencing. *J Mol Diagnostics* 2017;**19**(1):84–91. https://doi.org/10.1016/j.jmoldx.2016.07.010.
52. Trabucco SE, Gowen K, Maund SL. *et al.* A novel next-generation sequencing approach to detecting microsatellite instability and Pan-tumor characterization of 1000 microsatellite instability–high cases in 67,000 patient samples. *J Mol Diagnostics* 2019;**21**(6):1053–66. https://doi.org/10.1016/j.jmoldx.2019.06.011
53. Ni Huang M, McPherson JR, Cutcutache I. *et al.* MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. *Sci Rep* 2015;**5**:1–10 1 https://doi.org/10.1038/srep13321.
54. Fujimoto A, Fujita M, Hasegawa T. *et al.* Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res* 2020;**30**(3):334–46. https://doi.org/10.1101/gr.255026.119.
55. Foltz SM, Liang WW, Xie M, Ding L. MIRMMR: binary classification of microsatellite instability using methylation and mutations. *Bioinformatics* 2017;**33**(23):3799–801. https://doi.org/10.1093/bioinformatics/btx507.
56. Cao Y, Wang D, Wu J. *et al.* MSI-XGNN: an explainable GNN computational framework integrating transcription- and methylation-level biomarkers for microsatellite instability detection. *Brief Bioinform* 2023;**24**(6). https://doi.org/10.1093/bib/bbad362.
57. Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* 2003;**4**(2) https://doi.org/10.1186/gb-2003-4-2-r13.
58. Vilar E, Gruber SB. Microsatellite instability in colorectal cancerthe stable evidence. *Nat Rev Clin Oncol* 2010;**7**(3):153–62. https://doi.org/10.1038/nrclinonc.2009.237.
59. Wong N, John S, Nussenzweig A, Canela A. END-seq: an unbiased, high-resolution, and genome-wide approach to map DNA double-strand breaks and resection in human cells. In: Aguilera A, Carreira A, editors. *Methods in Molecular Biology*. New York, NY: Springer US, 2021. p. 9–31. https://doi.org/10.1007/978-1-0716-0644-5_2
60. De' Angelis GL, Bottarelli L, Azzoni C. *et al.* Microsatellite instability in colorectal cancer. *Acta Biomed* 2018;**89**(6):97–101. https://doi.org/10.1038/nrclinonc.2009.237.
61. Zhu L, Huang Y, Fang X. *et al.* A novel and reliable method to detect microsatellite instability in colorectal cancer by next-generation sequencing. *J Mol Diagnostics* 2018;**20**(2):225–31. Available from: https://doi.org/10.1016/j.jmoldx.2017.11.007
62. Rodrigues DN, Rescigno P, Liu D. *et al.* Immunogenomic analyses associate immunological alterations with mismatch repair defects in prostate cancer. *J Clin Invest* 2018;**128**(10):4441–53. https://doi.org/10.1172/JCI121924.
63. Middha S, Zhang L, Nafa K. *et al.* Reliable pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data. *JCO Precis Oncol* 2017;**1**:1–17. https://doi.org/10.1200/PO.17.00084.