Profiling microsatellite instability using RNA sequencing data

Abstract 179

J. Otoničar[1], M. Levstek[1], M. Žganec[2], R. Luštrik[1], J. Kokošar[2], M. Uhlik[2], M. Štajdohar[2], L. Ausec[2], S. Singh[1]

[1]Genialis, d.o.o., Ljubljana, Slovenia, [2]Genialis Inc., Boston MA, United States

Introduction

Microsatellite instability (MSI) as a biomarker for cancer treatment

Microsatellite instability (MSI) is a hyper-mutable phenotype defined by a unique set of alterations in microsatellites (MS) caused by defective DNA mismatch repair (MMR) in the cancer cells. MSI and MSS (microsatellite stable, non-MSI) have emerged as important biomarkers for predicting response to immune checkpoint inhibitor (ICI) therapy in different indications.

The prevalence of MSI varies across tumor types, with the highest rates seen in Endometrial cancer (EC) (25–30 %), Colorectal cancer (CRC) (15–20 %), Gastric cancer (GC) and Ovarian cancer (OC) (5–10 %). MSI tumors are more responsive to ICIs, with overall response rates (ORRs) ranging from 40–70 % across tumor types.

Current MSI detection methods →

- Immunohistochemical (IHC) and PCR-based assays are standard methods for MSI diagnosis.
- The Bethesda guidelines suggest that instability in two out of five poly-A loci (two mononucleotide and three dinucleotide) can indicate an MSI tumor via PCR.
- IHC staining evaluates the expression of four clinically relevant MMR proteins (MLH1, MSH2, MSH6, PMS2) in tumor and non-tumor nuclei. The absence of one or more of these proteins in matched normal tissue suggests an MSI tumor.
- These methods are fast, inexpensive, and readily available by different diagnostic providers and reference laboratories.

Limitations of current approaches →

- The IHC test is not a direct phenotypic evaluation of MSI.
- IHC tests may miss cases where MMR deficiency results from gene inactivation beyond the four proteins assayed.
- Many PCR-based tests require both tumor and matched normal samples for accurate diagnosis.
- Some mutations may not lead to the loss of protein expression.
- Individual tests provide limited information from a comparable amount of biomaterial that could be used in comprehensive sequencing assays.
- Traditional testing is impractical in tumor forms where MSI is rare, despite the presence of some degree of MSI in most solid tumor types, hindering the identification of MSI patients across solid tumor types who could benefit from checkpoint inhibitors.

Solution: Pan-cancer MSI prediction using RNA-sequencing

- RNA sequencing is increasingly used in Precision Oncology for the identification of multiple complex disease markers using a single sequencing assay.
- Genialis developed an RNA-Seq-based pipeline for MSI characterization in various cancer indications. The workflow identifies mutations and gene expression, and classifies samples into MSI or MSS (*Figure 1*).
- This approach utilizes only tumor samples and eliminates the need for matched normal samples.

Methods

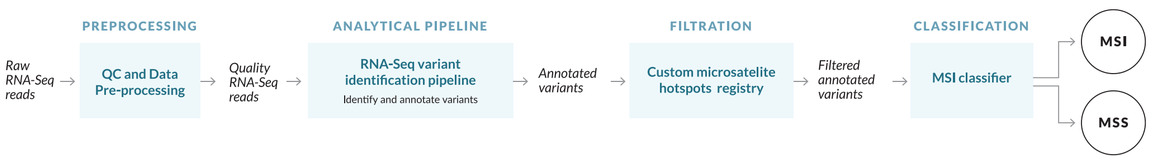
Data

Dataset	Tissue	MSS	MSI	Ethnicity	Accession	Tissue origin
Chatila et al. [1]	Colorectal (109)	104	5	US collection sites	GSE209746	FFPE
DiGuardo et al. [2]	Colorectal (42), Endometrial (35), Ovarian (1), NA (11)	19 (tumor)	70 (tumor)	Non-asian	GSE146889	FFPE
Guo et al. [3]	Colorectal (18)	6 (non-MSI)	12	Chinese	GSE222202	Fresh frozen, FFPE
Mun et al. [4]	Gastric (80)	76	4	Korean	GSE122401	Fresh frozen
Park et al. [5]	Colorectal (145)	124	21	Korean	GSE180440	Fresh frozen
Partner RWD	Gastric (45)	40	5	Korean	Private	FFPE

◀ ***Table 1.***

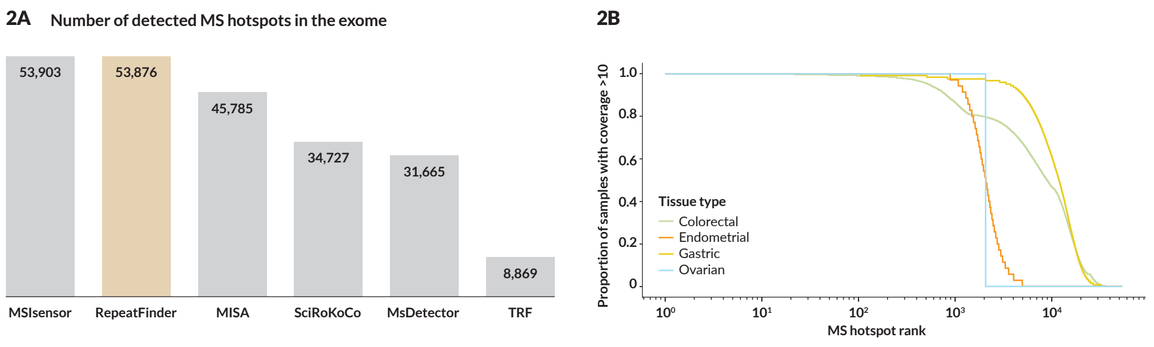
Overview of MSI datasets used in this study.

Algorithm



◀ ***Figure 1.***

An overview of the MSI prediction pipeline. This pipeline includes preprocessing steps to clean and normalize raw RNA-Seq data, followed by analytical steps to identify and annotate variants. The algorithm leverages a proprietary MS hotspot registry to classify samples into MSI or MSS.



◀ ***Figure 2.***

2A (Left): *A comprehensive catalog of MS in the human exome derived from six distinct MS detection tools. Of the six tools, RepeatFinder was tied for first in comprehensiveness, and selected for inclusion in the pipeline thanks to its permissive open source license.*

2B (right): *A rank plot of MS hotspots by mean coverage for each of the four studied tissue types.*

Details of the analysis

Genetic Variant Identification and Filtering: A combination of filters are applied based on the type of mutational event observed, read coverage, quality of detection, and other measurable characteristics of the variant.

Variant Annotation: High-quality variants are annotated using a combination of public and proprietary databases, providing additional context and information about these genetic variations. Some public databases integrated are dbSNP, ClinVar, and SnpEff.

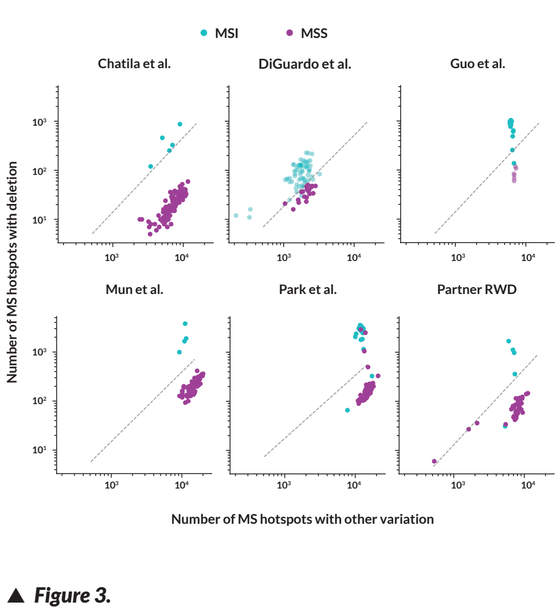
MS hotspot catalog: Six tools were benchmarked to call MS sites in the human genome (*Figure 2A*). After thorough evaluation, only those from RepeatFinder were retained, and filtered to MS sites detectable by RNA-Seq (exome regions). Each of these 53,876 hotspots were tested for statistical significance of the difference in the frequency of observed deletions between all MSI and MSS patients. Thousands of these hotspots have sufficient coverage in all four tissue types (*Figure 2B*).

MSI Classification: A logistic regression model was trained on 486 samples across six datasets (*Table 1*) to classify samples into MSI and MSS classes. The model was trained in leave-one-dataset-out cross-validation on the total number of alterations within the MS hotspots. The model outputs predicted probabilities of the two classes. The predicted class is the one with the higher probability.

Results

Logistic regression model distinguishes MSI and MSS samples based on MS hotspots

The logistic regression model readily separated MSI and MSS samples based on the number of MS hotspots with deletion and those with other variations (*Figure 3*). For each dataset, the decision boundary was trained on the remaining five datasets, as per the leave-one-dataset-out cross-validation approach.



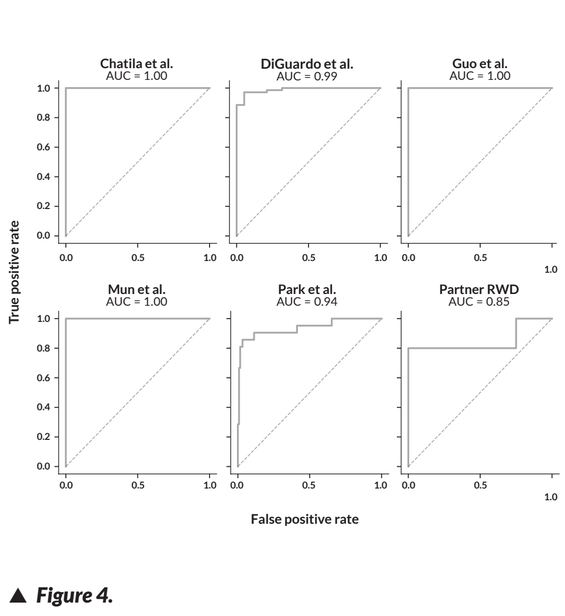
▲ ***Figure 3.***

MSI and MSS samples in the six datasets can be separated based on the number of MS hotspots with deletion and other types of variation. The decision boundary of the logistic regression model trained in leave-one-dataset-out cross-validation is plotted for each dataset as a dashed line.

Note that the number of MS hotspots with deletion and other variations are only loosely correlated. A model that uses only deletions was found to perform worse compared to a model that includes these additional features (data not shown).

Model performance

The performance of the logistic regression model (*Table 2*) as given by the ROC AUC ranged from 0.85 to 1.00, with a mean of 0.96, comparable to current state of the art detection techniques [6]. The confusion matrices that displayed Clinical (reference) versus Predicted MSI/MSS status (not shown) revealed a small bias towards false negative predictions.



▲ ***Figure 4.***

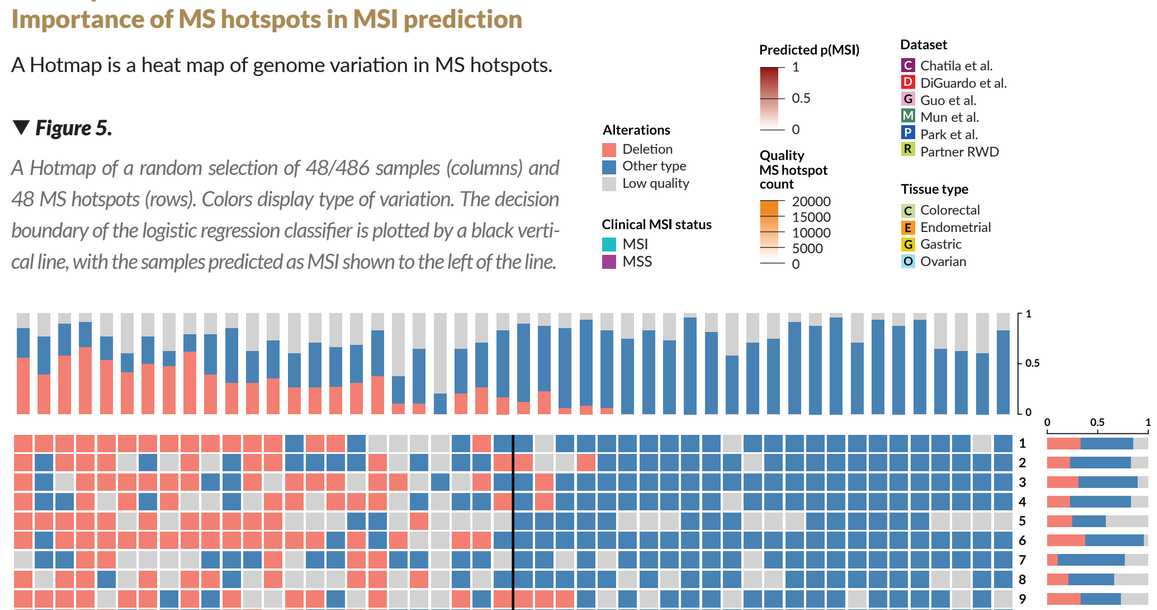
Receiver operating characteristic and the associated ROC AUC value computed in leave-one-dataset-out cross-validation for each of the six datasets.

Hotmap: Importance of MS hotspots in MSI prediction

A Hotmap is a heat map of genome variation in MS hotspots.

▼ ***Figure 5.***

A Hotmap of a random selection of 48/486 samples (columns) and 48 MS hotspots (rows). Colors display type of variation. The decision boundary of the logistic regression classifier is plotted by a black vertical line, with the samples predicted as MSI shown to the left of the line.



▼ ***Table 2.***

Model performance metrics computed in leave-one-dataset-out cross-validation.

Dataset	ROC AUC	Sensitivity (recall)	Specificity	PPV (precision)	NPV	Accuracy
Chatila et al. [1]	1.00	1.00	1.00	1.00	1.00	1.00
DiGuardo et al. [2]	0.99	0.96	0.95	0.99	0.86	0.96
Guo et al. [3]	1.00	0.92	1.00	1.00	0.86	0.94
Mun et al. [4]	1.00	1.00	1.00	1.00	1.00	1.00
Park et al. [5]	0.94	0.81	0.98	0.85	0.97	0.95
Partner RWD	0.85	0.80	0.98	0.80	0.98	0.96

Discussion and Conclusions

A machine learning model that uses mutational profiles determined from RNA-Seq data was trained to predict MSI in cancer patients. The model employs a novel technique for selecting features as a subset of MS hotpots based on the types of alterations, and consists of a simple logistic regression architecture. The model was trained on a dataset of 486 patients compiled from six different studies. The achieved results demonstrate an ROC AUC of > 0.94 for all but one dataset.

The model computes unbiased MSI predictions across various tissue types, making it applicable in different cancer contexts. Notably, it does not require a matched normal sample. It outputs the probability of MSI which indicates an estimate of confidence in the call and allows for post-prediction decision-making based on patient-specific risk factors. It excels at single-sample predictions, eliminating the need for population-based recalibration. The simplicity of the model also facilitates interpretability of predictions.

This MSI classifier is a research tool for the retrospective analysis of RNA-Seq data without the need for matched normal samples. In the future, this classifier will be evaluated in prospective analytical settings and may be incorporated into panels of RNA-Seq based biomarker tests operated by diagnostic and NGS service providers.

References

1. Chatila WK, Kim JK, Walch H, Marco MR, Chen CT, Wu F, et al. Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer. Nat Med. 2022 Aug;28(8):1646–55.
2. DiGuardo MA, Davila JI, Jackson RA, Nair AA, Fadra N, Minn KT, et al. RNA-Seq Reveals Differences in Expressed Tumor Mutation Burden in Colorectal and Endometrial Cancers with and without Defective DNA-Mismatch Repair. J Mol Diagn. 2021 May;23(5):555–64.
3. Guo L, Wang Y, Yang W, Wang C, Guo T, Yang J, et al. Molecular Profiling Provides Clinical Insights Into Targeted and Immunotherapies as Well as Colorectal Cancer Prognosis. Gastroenterology. 2023 Aug;165(2):414-428.e7.
4. Mun DG, Bhin J, Kim S, Kim H, Jung JH, Jung Y, et al. Proteogenomic Characterization of Human Early-Onset Gastric Cancer. Cancer Cell. 2019 Jan 14;35(1):111-124.e10.
5. Park DY, Choi C, Shin E, Lee JH, Kwon CH, Jo HJ, et al. NTRK1 fusions for the therapeutic intervention of Korean patients with colon cancer. Oncotarget. 2016 Feb 16;7(7):8399–412.
6. Chen ML, Chen JY, Hu J, Chen Q, Yu LX, Liu BR, et al. Comparison of microsatellite status detection methods in colorectal carcinoma. Int J Clin Exp Pathol. 2018 Mar 1;11(3):1431–8.