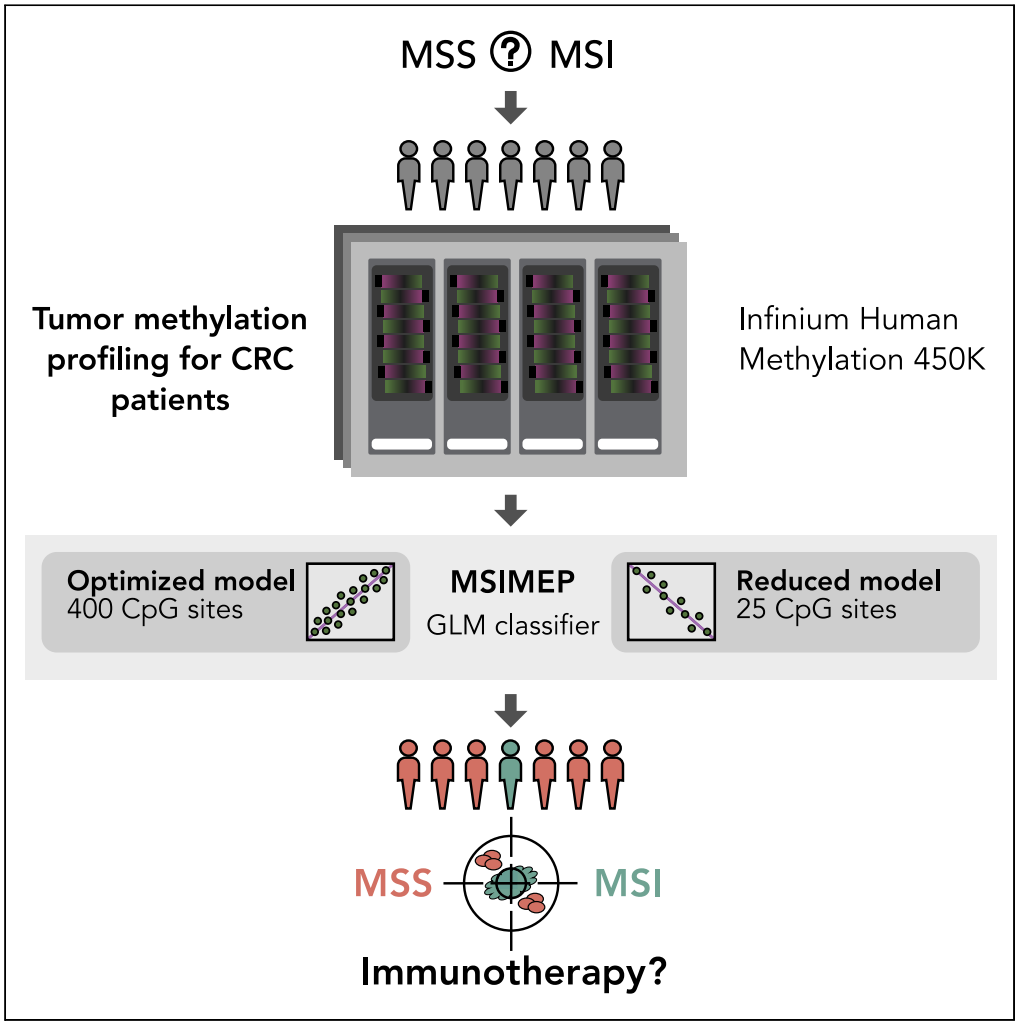
iScience

CellPress
OPEN ACCESS

Article

MSIMEP: Predicting microsatellite instability from microarray DNA methylation tumor profiles



Martín Santamarina-García, Jenifer Brea-Iglesias, Jesper Bertram Bramsen, ..., Jose M.C. Tubio, Rafael López-López, Juan Ruiz-Bañobre

rafael.lopez.lopez@sergas.es (R.L.-L.)
juan.ruiz.banobre@sergas.es (J.R.-B.)

Highlights

MSIMEP, first DNA methylation-only based microsatellite instability predictor

Optimized and reduced versions with 400 and 25 predictors, respectively

Consistent performance across many different colorectal cancer cohorts

Better performance than an MLH1 promoter methylation-based model

Santamarina-García et al., iScience 26, 106127
March 17, 2023
https://doi.org/10.1016/j.isci.2023.106127

iScience

CellPress
OPEN ACCESS

Article

MSIMEP: Predicting microsatellite instability from microarray DNA methylation tumor profiles

Martín Santamarina-García,[1] Jenifer Brea-Iglesias,[1,2] Jesper Bertram Bramsen,[3] Mar Fuentes-Losada,[4,5] Francisco Javier Caneiro-Gómez,[6] José Ángel Vázquez-Bueno,[7] Héctor Lázare-Iglesias,[6] Natalia Fernández-Díaz,[4,5] Laura Sánchez-Rivadulla,[8] Yoel Z. Betancor,[1,5] Miriam Ferreiro-Pantín,[1,5] Pablo Conesa-Zamora,[9] José Ramón Antúnez-López,[6] Masahito Kawazu,[10,11] Manel Esteller,[12,13,14,15] Claus Lindbjerg Andersen,[3] Jose M.C. Tubio,[1] Rafael López-López,[4,5,15,*] and Juan Ruiz-Bañobre[1,4,5,15,16,*]

SUMMARY

Deficiency in DNA MMR activity results in tumors with a hypermutator phenotype, termed microsatellite instability (MSI). Beyond its utility in Lynch syndrome screening algorithms, today MSI has gained importance as predictive biomarker for various anti-PD-1 therapies across many different tumor types. Over the past years, many computational methods have emerged to infer MSI using either DNA- or RNA-based approaches. Considering this together with the fact that MSI-high tumors frequently exhibit a hypermethylated phenotype, herein we developed and validated MSIMEP, a computational tool for predicting MSI status from microarray DNA methylation tumor profiles of colorectal cancer samples. We demonstrated that MSIMEP optimized and reduced models have high performance in predicting MSI in different colorectal cancer cohorts. Moreover, we tested its consistency in other tumor types with high prevalence of MSI such as gastric and endometrial cancers. Finally, we demonstrated better performance of both MSIMEP models vis-à-vis a MLH1 promoter methylation-based one in colorectal cancer.

INTRODUCTION

Microsatellites are short tandem repeat DNA sequences spread throughout the human genome. Because of their highly repetitive nature, these sequences have a higher propensity for acquiring mutations. Deficiency in DNA mismatch repair (MMR) activity results in a hypermutator phenotype, termed microsatellite instability (MSI), characterized by the presence of single nucleotide substitutions or insertion-deletion mutations within these microsatellites.[1] The MMR deficiency resulting from germline mutations or epigenetic alterations in any of the MMR genes (MLH1, MSH2, MSH6, and PMS2), as well as deletions in the EPCAM gene, is the cause of Lynch syndrome (LS) and its variants.[2] One of the most aggressive, highly penetrant childhood cancer predisposition syndromes, the constitutional MMR deficiency syndrome, is caused by homozygous germline mutations in any of the four MMR genes. Furthermore, LS can result from mosaic germline MLH1 epimutations. In contrast, biallelic MLH1 promoter methylation is primarily the key somatic event responsible for the loss of MLH1 expression in ~75% of sporadic cancers with MSI.[1,2] Although MMR deficiency/MSI determination has been classically the first step in LS screening algorithms, today MSI has gained importance as predictive biomarker for various anti-PD-1 therapies across many different tumor types (tumor-agnostic indication) and particularly in colorectal and endometrial cancers.[3] Therefore, evaluation of MSI, through either PCR-based assays or immunohistochemistry, has become a routine clinical practice for various cancer types. With the emergence of next-generation sequencing-based technologies, alternate computational methods to infer MSI using DNA-targeted, whole exome or whole genome sequencing data (e.g., MSIsensor, MSIsensor-pro, mSINGS, MOSAIC, and MANTIS) have been developed.[4–8] Other computational approaches, instead of observing microsatellites directly to evaluate MSI, use orthogonal prediction methods based on gene expression[9–13] or a combination of DNA methylation levels and mutations in MMR pathway genes (MIRMMR).[14] Based on this and taking into consideration the fact that MSI-high tumors frequently exhibit a hypermethylated phenotype,[15] herein we propose an MSI prediction method based on microarray DNA methylation tumor tissue profiling. We explore the

[1]Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain

[2]Translational Oncology Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Álvaro Cunqueiro Hospital, 36213 Vigo, Spain

[3]Department of Molecular Medicine, Aarhus University Hospital, 8200 Aarhus, Denmark

[4]Department of Medical Oncology, University Clinical Hospital of Santiago de Compostela (SERGAS), University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain

[5]Translational Medical Oncology Group (ONCOMET), Health Research Institute of Santiago de Compostela (IDIS), University Clinical Hospital of Santiago de Compostela, University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain

[6]Department of Pathology, University Clinical Hospital of Santiago de Compostela, University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain

[7]Department of Pathology, Complejo Hospitalario Universitario de Ferrol, 15405 Ferrol, Spain

[8]Department of Gynaecology and Obstetrics, Complejo Hospitalario Universitario de Ferrol, 15405 Ferrol, Spain

Continued

underlying features that predict MSI and develop a reduced version with potential applicability to liquid biopsy samples.

RESULTS

Model development and optimization

First, to find differentially methylated CpG probes between MSI and MSS tumors, we interrogated Infinium Human Methylation 450K array data obtained from 388 colorectal primary tumors from the pooled The Cancer Genome Atlas (TCGA) Colon Adenocarcinoma (COAD)/Rectum Adenocarcinoma (READ) cohort. After carrying out methylation probes filtering, data imputation, and quality control (see STAR Methods), we identified 780 probes with an absolute $\Delta\beta>0.3$ and a B-H false discovery rate (FDR)-adjusted p value<0.05. Of these probes 758 (97.2%) were hypermethylated and 22 (2.8%) hypomethylated (Figure S1). Then, these 780 differentially methylated CpG probes were used to train 7 supervised machine learning models on the same dataset. The generalized linear model (GLM) implemented in GLMNET yielded the best performance after five repeats of 10-fold cross-validation (accuracy = 0.98, kappa = 0.92). This performance was reached when setting *alpha* = 0.1 and *lamda* = 0.0182 as tuned parameters. GLM performance was significantly better when compared against the second (Wilcoxon test Holm family-wise error rate (FWER)-adjusted p value = 0.041) and third (Wilcoxon test Holm FWER-adjusted p value = 0.017) best alternative models (Figure 1A).

Once the best model was selected, we carried out recursive feature elimination (RFE) for model optimization (Figure S2). Optimized model (*alpha* = 0.1, *lamda* = 0.0182, accuracy = 0.98, kappa = 0.92, area under the receiver operating characteristic curve (AUROC) = 0.98, sensitivity = 0.88, and specificity = 1.00), hereafter MSIMEP, is composed of 400 CpG probes (Figure 1B and Table 1). Importantly, those cases incorrectly classified by our model presented a MANTIS score statistically significantly closer to 0.4,[8] the average distance threshold established to differentiate MSI from MSS tumors by that computational method, than those properly classified (MSS vs failed, Wilcoxon test p value<0.001; failed vs MSI, Wilcoxon test p value = 0.009) (Figure S3). Importantly, MSIMEP performance was consistent across different clinical and molecular subgroups (Figures S4 and S5).

Additionally, considering a potential applicability to liquid biopsy samples and assuming a limited sacrifice of accuracy (see STAR Methods), we developed a reduced version of MSIMEP composed of 25 CpG probes (Figures S1, S4, and S6).

Even though MSIMEP is intended for colorectal cancer (CRC), we were interested to test its performance in other tumor types with high prevalence of MSI such as gastric and endometrial cancers. For this purpose, we interrogated Infinium Human Methylation 450K array data from the TCGA-Stomach Adenocarcinoma (STAD) and TCGA-Uterine Corpus Endometrial Carcinoma (UCEC) cohorts. The optimized model yielded accuracy = 0.84, kappa = 0.56, AUROC = 0.94, sensitivity = 0.9, specificity = 0.82, and precision = 0.52, and accuracy = 0.88, kappa = 0.70, AUROC = 0.92, sensitivity = 0.72, specificity = 0.95, and precision = 0.86 in STAD and UCEC, respectively (Figure S7). Although far from perfect, reduced MSIMEP yielded accuracy = 0.75, kappa = 0.43, AUROC = 0.93, sensitivity = 0.94, specificity = 0.71, and precision = 0.41, and accuracy = 0.89, kappa = 0.73, AUROC = 0.93, sensitivity = 0.82, specificity = 0.91, and precision = 0.81 in STAD and UCEC, respectively (Figure S7).

Underlying features of the MSIMEP model

The composition of the MSIMEP is based on a 400 CpG probe set spread through the 22 human autosomes (Figure 2A). Chromosomes 3, 1, and 6 have the largest number of MSIMEP probes (41, 34, and 28 CpGs, respectively), and the highest density is present on chromosome 19 (21 CpGs, 0.37 CpGs/Mb). Overall, MSIMEP probes tend to be located on high GC content regions (88%), which reflects not only the composition of the Illumina 450K Infinium Array but also the intrinsic CpG island methylator phenotype of tumors with MSI.

From the genomic perspective, 63 MSIMEP probes (15.8%) are located at intergenic regions, whereas 337 probes (84.3%) are associated with coding genes. Remarkably, 34 MSIMEP probes (8.5%) are associated with human cancer genes (COSMIC), including 19 probes associated with genes involved in cancer syndromes such as MLH1, EXT1, WRN, and AXIN2. Twenty-four MSIMEP probes (6%) are associated with tumor suppressor genes, 13 probes (3.3%) with fusion genes, and six probes (1.5%) with proto-oncogenes (Figure 2A). The top-five relevant genes for the MSIMEP predictive model are MLH1, TRIP10, C1orf95, EXT1, and GPR160, respectively (Figure 2B). In addition, when evaluating the composition of MSIMEP

[9]Department of Clinical Analysis, Santa Lucía University Hospital, 30202 Cartagena, Spain

[10]Chiba Cancer Center, Research Institute, 260-0801 Chiba, Japan

[11]Division of Cellular Signaling, National Cancer Center Research Institute, 104-0045 Tokyo, Japan

[12]Josep Carreras Leukaemia Research Institute (IJC), 08916 Badalona, Barcelona, Spain

[13]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

[14]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), 08907 Barcelona, Spain

[15]Centro de Investigación Biomédica en Red Cáncer (CIBERONC), 28029 Madrid, Spain

[16]Lead contact

*Correspondence: rafael.lopez.lopez@sergas.es (R.L.-L.), juan.ruiz.banobre@sergas.es (J.R.-B.)

https://doi.org/10.1016/j.isci.2023.106127

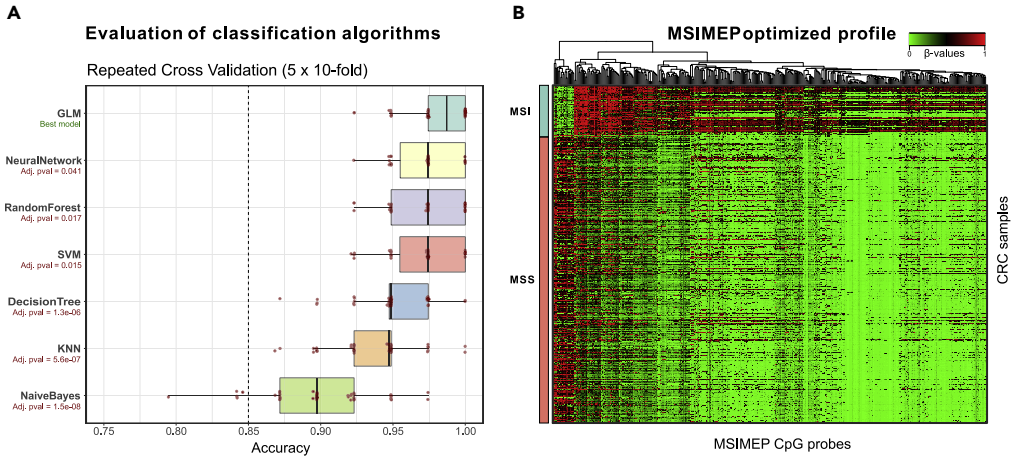


Figure 1. Development of the MSIMEP classifier

(A) Evaluation of alternative algorithms for MSI/MSS phenotype prediction from array methylation tumor profiles. Model performance was assessed through 5 rounds of 10-fold cross-validation on the TCGA COAD/READ dataset, and an optimal algorithm (GLM) was selected based on significant improvement in classification accuracy. Alternative algorithms (Neural Network, Random Forest, SVM, Decision Tree, KNN, and Naive Bayes) are displayed sorted by average accuracy, with Holm FWER-adjusted p values showing statistical differences against GLM.

(B) Heatmap for the methylation tumor profile (β-values) of TCGA COAD/READ patients for the CpG probes incorporated into the MSIMEP optimized model. CpG probes (x axis) are displayed following hierarchical clustering criteria, and TCGA COAD/READ patients (y axis) are grouped by MSI/MSS phenotype (green = MSI, red = MSS).

from a functional perspective, 250 probes (62%) are located at regulatory elements. Compared with the Infinium Human Methylation 450K array, MSIMEP presents a higher proportion of regulatory regions at both genic (χ^2 test p value = 0.003) and intergenic regions (χ^2 test p value = 0.012).

Considering their relevance for gene structure, 219 probes (54.8%) are located within a proximal promoter region (between 1.5 kb upstream of transcription start site and the first exon). One hundred and nine probes (26.3%) are located within the gene body, and 13 probes (3.3%) are located within the 3′UTR. When comparing MSI vs MSS tumors from TCGA COAD/READ cohort, hypermethylated probes are predominantly located on promoter regions (Fisher exact test p value<0.001, hypermethylated vs hypomethylated), whereas hypomethylated probes are mainly located on gene bodies (Fisher exact test p value<0.001, hypermethylated vs hypomethylated) (Figure 2C).

According to the CpG landscape, 231 MSIMEP probes are located at CpG islands (57.8%), and 107 probes are located on their surroundings: 89 on CpG shores (22.3%) and 18 on CpG shelves (4.5%). Sixty-two MSIMEP probes are located on open sea genomic regions (15.5%). Again, when comparing MSI vs MSS tumors from

Table 1. MSIMEP optimized model performance in different cohorts

Cohort	Cancer type	Phase	Sample size (n)	MSI (%)	MSS (%)	Accuracy	Kappa	AUROC	Sensitivity	Specificity	Precision
TCGA COAD/READ	CRC	Training	387	15	85	0.98	0.92	0.98	0.88	1.00	0.99
External I	CRC	Validation	79	14	86	0.97	0.90	0.99	1.00	0.97	0.85
External II	CRC	Validation	262	15	85	0.95	0.83	0.95	0.88	0.97	0.83
External III	CRC	Validation	81	11	89	0.86	0.52	0.86	0.89	0.86	0.44
External IV	CRC	Validation	95	100	NA	NA	NA	NA	0.86	NA	1.00
External Pooled	CRC	Validation	517	30	70	0.93	0.83	0.93	0.88	0.95	0.88
TCGA-STAD	STAD	Exploratory	394	18	82	0.84	0.56	0.94	0.90	0.82	0.52
TCGA-UCEC	UCEC	Exploratory	420	31	69	0.88	0.70	0.92	0.72	0.95	0.86

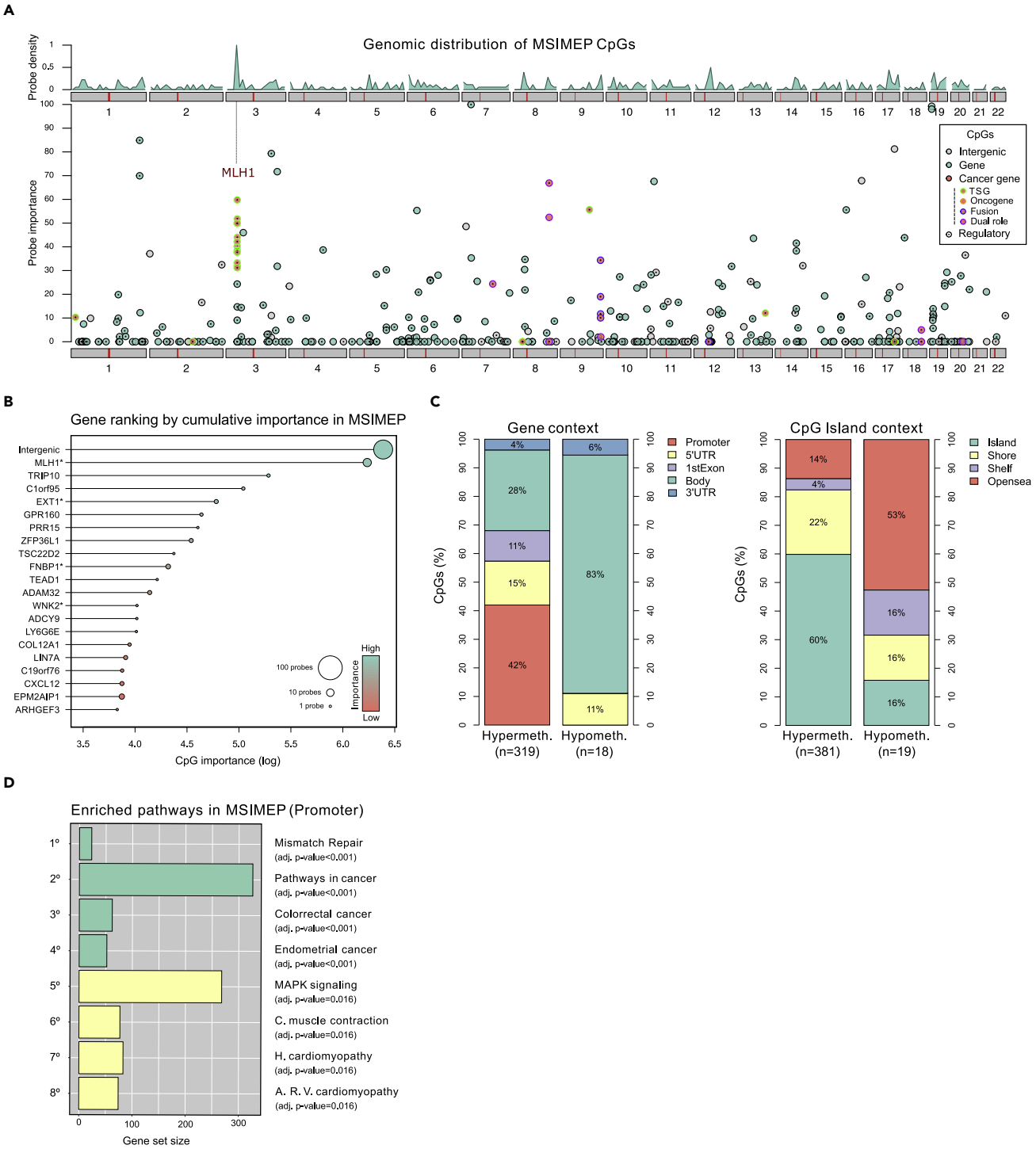


Figure 2. Features of the MSIMEP model

(A) Genomic distribution of MSIMEP CpG probes. The x axis shows the CpG coordinates across the human autosomes (chromosomes 1–22). The y axis shows relative density of MSIMEP CpG probes (top), and the relative importance of each CpG for the GLM classification model (bottom). Probes are classified as genic (green), intergenic (gray), or cancer-related (red). Cancer-related probes are further classified according to the role/s in cancer (tumor suppressor gene, oncogene, fusion gene, or dual role) of the associated gene. Probes associated with regulatory elements are highlighted with a central dot. Location of the MLH1 gene is shown.

(B) Top 20 genes profiled by MSIMEP, ranked by the cumulative importance for the GLM model of the associated CpG probes. Dot size reflects the number of probes associated with each gene. Cancer genes are highlighted with an asterisk.

Figure 2. *Continued*
(C) Comparison of the distribution of hypermethylated and hypomethylated MSIMEP CpG probes (MSI vs MSS tumors from TCGA COAD/READ cohort) according to gene and CpG island context regions.
(D) Pathway enrichment analysis showing overrepresented KEGG terms in MSIMEP at promoter level (n = 134 CpG probes), ranked by B-H FDR-adjusted p values.

TCGA COAD/READ cohort, hypermethylated probes are predominant on CpG islands (Fisher exact test p value<0.001, hypermethylated vs hypomethylated), whereas hypomethylated probes typically are on open sea regions (Fisher exact test p value<0.001, hypermethylated vs hypomethylated) (Figure 2C).

The composition of the MSIMEP reduced model is based on a 25 CpG probe set (Table 2, Figure S1), conformed by 4 probes located at intergenic regions, and 21 CpG probes associated with genes. Among those probes associated with genes, 9 CpGs are attributed to 3 relevant cancer genes (MLH1, EXT1, and WNK2, all well-established [Tier I] tumor suppressor genes).

Finally, to fully characterize MSIMEP from a biological viewpoint, we carry out a gene set enrichment analysis. As expected, when evaluated at promoter level, this analysis revealed an enrichment of important pathways involved in mismatch repair (B-H FDR-adjusted p value<0.001), colorectal and endometrial cancers (B-H FDR-adjusted p value<0.001), and MAPK signaling (B-H FDR-adjusted p value = 0.015) (Figure 2D). When evaluated either all CpG probes or those located at body gene level, mismatch repair (B-H FDR-adjusted p value<0.001), and pathways in cancer (B-H FDR-adjusted p value<0.001) appear among the most significant pathways, confirming again the biological correlation of MSIMEP with the MSI phenotype in cancer. Other significantly enriched pathways are described in Figure S8. Moreover, through computational immune cell deconvolution techniques we confirmed that MSIMEP-predicted MSI CRCs preserve the same immune cell context than those MSI cases originally classified through standard PCR-based assays in the TCGA COAD/READ cohort (Figure S9).

External validation of MSIMEP models

To further confirm the accuracy of MSIMEP models in predicting MSI status in CRC samples, we evaluated its performance in four independent CRC cohorts.

First, we tried MSIMEP in the external cohort I. In this cohort, MSIMEP model yielded promising results when evaluated in either invasive front (accuracy = 0.99, kappa = 0.95, AUROC = 1.00, sensitivity = 1.00, specificity = 0.99, and precision = 0.92), or in luminal (accuracy = 0.99, kappa = 0.95, AUROC = 0.99, sensitivity = 1.00, specificity = 0.99, and precision = 0.92) or center (accuracy = 0.97, kappa = 0.90, AUROC = 0.99, sensitivity = 1.00, specificity = 0.97, and precision = 0.85) regions (Table 1, Figure 3A). Importantly, the concordance of predictions among the three different tumor regions per case was extremely high (Cochran's Q test p value = 0.607, no statistically significant differences found in predictions by tumor region) (Figure 3B).

Next, we evaluated MSIMEP in cohorts II and III, showing again consistent results (cohort II: accuracy = 0.95, kappa = 0.83, AUROC = 0.95, sensitivity = 0.88, specificity = 0.97, and precision = 0.83; cohort III: accuracy = 0.86, kappa = 0.52, AUROC = 0.86, sensitivity = 0.89, specificity = 0.86, and precision = 0.44) (Table 1, Figures 4A and 4B). Moreover, we assessed the predictive potential of MSIMEP in cohort IV, which is only composed of MSI-H cases. Despite that this represents a challenging cohort due to the lack of MSS cases, MSIMEP yielded a sensitivity of 0.86 (Table 1, Figure 4C).

Additionally, we evaluated MSIMEP performance across different important clinical and molecular subgroups in a pooled CRC cohort (external cohorts I to IV). As expected, MSIMEP yielded consistent results, even when stratified by MLH1 promoter methylation status, a challenging predictive scenario considering the high relative importance of MLH1 CpG probes in MSIMEP model (Figure S10).

MSIMEP reduced model performance was also robust in all the external cohorts (I to IV) (Table 3, Figures 3A and 4) and across all subgroups in the pooled cohort (Figure S11).

Altogether, these results suggest that MSIMEP models can be used to predict MSI status from microarray DNA methylation profiles from both fresh-frozen FF and formalin-fixed paraffin-embedded (FFPE) tumor CRC samples.

Table 2. Description of the 25 CpG probes of the MSIMEP reduced model

CpG probe	Chromosome	Position	Strand	Gene	Gene region	Methylation status[a] (MSI vs MSS)
cg15592945	1	226736711	F	C1orf95	1st Exon	Hypermethylated
cg03745431	1	226736713	F	C1orf95	1st Exon	Hypermethylated
cg17621259	3	37035168	F	MLH1	TSS200	Hypermethylated
cg14671526	3	37035200	R	MLH1	TSS200	Hypermethylated
cg27331401	3	37035207	F	MLH1	TSS200	Hypermethylated
cg06590608	3	37035228	F	MLH1	TSS200	Hypermethylated
cg11224603	3	37035282	R	MLH1	1st Exon	Hypermethylated
cg14598950	3	37035355	F	MLH1	1st Exon	Hypermethylated
cg04563996	3	57093519	R	ARHGEF3	5′UTR	Hypomethylated
cg15048832	3	150130775	R	TSC22D2	Body	Hypermethylated
cg12350863	3	169758289	F	GPR160	5′UTR	Hypermethylated
cg20288341	6	31683131	F	LY6G6E	TSS1500	Hypermethylated
cg06226516	7	13072314	R	NA	IGR	Hypermethylated
cg02200207	7	29605237	R	PRR15	5′UTR	Hypermethylated
cg05313153	8	119122430	F	EXT1	1st Exon	Hypermethylated
cg21602557	8	119122878	R	EXT1	1st Exon	Hypermethylated
cg13563298	9	95948059	R	WNK2	Body	Hypermethylated
cg27240158	11	12698019	R	TEAD1	5′UTR	Hypermethylated
cg19524009	13	52735075	R	NEK3	TSS1500	Hypermethylated
cg06223834	16	4103161	R	ADCY9	Body	Hypermethylated
cg09294739	16	55218782	R	NA	IGR	Hypomethylated
cg11582717	17	63452460	R	NA	IGR	Hypermethylated
cg13860006	19	6741178	R	TRIP10	Body	Hypermethylated
cg06410591	19	6741181	R	TRIP10	Body	Hypermethylated
cg05232694	20	48809539	R	NA	IGR	Hypermethylated

F, forward; IGR, intergenic region; NA, not applicable; R, reverse.

[a]Methylation status is described based on the comparison of MSI vs MSS tumors from TCGA COAD/READ cohort.

Comparison of MSIMEP models with a MLH1 promoter methylation-based model

Taking into consideration the role of biallelic MLH1 promoter methylation in most of the sporadic CRCs with MSI, we compared the performance of our MSIMEP models with a MLH1 promoter methylation-based model (Table S1) in a pooled CRC cohort (external cohorts I to IV). As expected, MSIMEP optimized and reduced models yielded a better sensitivity (0.88 for both MSIMEP models vs 0.71 for MLH1 promoter methylation-based model) at the expense of a slightly lower specificity than the MLH1 promoter methylation-based model (0.95 and 0.92 for MSIMEP optimized and reduced models, respectively, vs 0.96 for MLH1 promoter methylation-based model) (Figure 5). Congruent results were obtained when stratified by MLH1 promoter methylation status, confirming the increased sensitivity of MSIMEP models to detect those non-MLH1 dependent cases (Figure S12).

DISCUSSION

We developed MSIMEP, a computational tool for predicting MSI status from microarray DNA methylation profiles of CRC samples. We demonstrated the accuracy and robustness of MSIMEP by testing it in multiple CRC cohorts with varying tumor sample types (either FF or FFPE) from two different Illumina platforms (either Infinium Human Methylation 450K or Infinium MethylationEPIC arrays). Additionally, we tested its performance in other tumor types with high prevalence of MSI such as gastric and endometrial cancers, obtaining consistent results.

As described by others,[16] DNA methylation profiling represents a solid approach for blood-based liquid biopsy due to, among other aspects, the better limit of detection of methylation-based approaches and the advantage of avoiding somatic mutations derived from normal tissues, benign diseases, and clonal

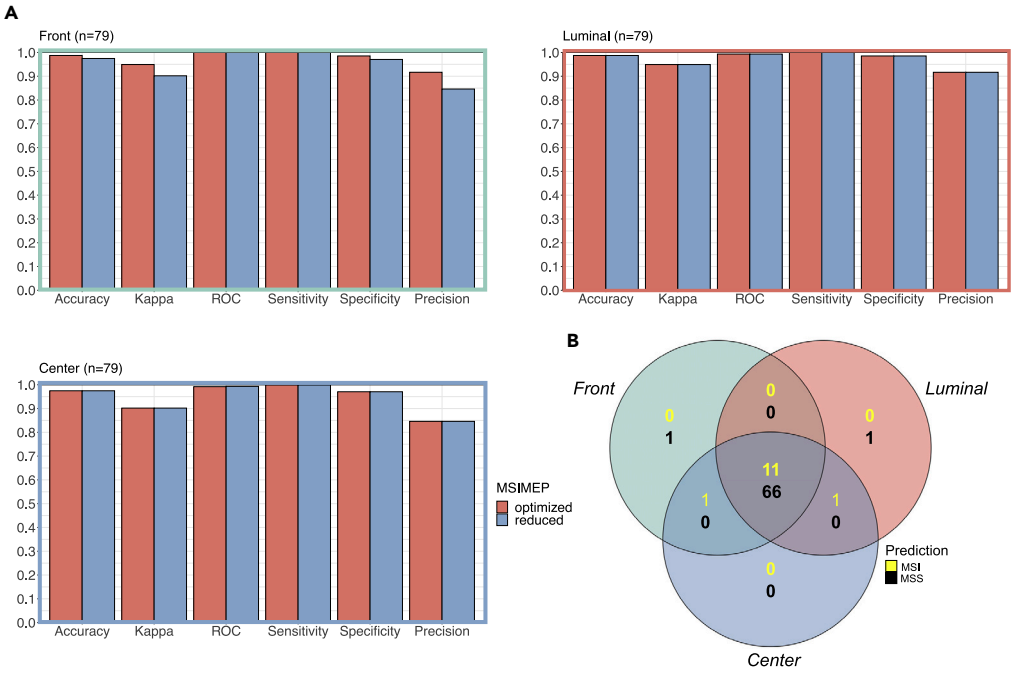


Figure 3. MSIMEP evaluation in CRC external cohort I

(A) Barplots showing complementary metrics (accuracy, kappa, AUROC, sensitivity, specificity, and precision) for the evaluation of MSIMEP classification capacity of MSI/MSS patients across different CRC regions (invasive front, luminal, and center). Models: MSIMEP optimized (red) and MSIMEP reduced (blue).

(B) Venn diagram showing the concordance in MSI/MSS prediction across different CRC regions from external cohort I patients. Cohort: external I (n = 79, 14% MSI).

hematopoiesis of indeterminate potential.[16,17] Acknowledging this fact, we developed and validated the MSIMEP reduced model, which could be adapted, and facilitate its translation to the clinic as a cost-effective liquid biopsy test.

To explore the underlying features of our MSI predictive model, we assessed the relative importance of the different predictors of MSIMEP in their genomic context. Although probes located at intergenic regions numerically represented only 15.8% of the total of predictors included in MSIMEP, taken as a whole, they accumulated a major relative importance in the model. Although the biological interpretation of this finding is complex and out of the scope of the present study, considering that intergenic regions may contain important functional sequences such as promoters and other regulatory elements, their role in the MSI phenotype and its regulation by DNA methylation deserve further investigation. Regarding probes located in genic regions, their highest proportion was in MLH1, reflecting the role of the methylation of its promoter as the primary somatic event responsible for the majority of sporadic MSI-H CRCs.[18]

Additionally, confirming our expectations, (1) GSEA revealed an enrichment of pathways involved not only in mismatch repair but also in colorectal and endometrial cancers and (2) computational immune cell deconvolution techniques confirmed a consistent immune cell context between MSIMEP predicted and PCR-based MSI CRCs.

Last, and after confirming MSIMEP performance across different important clinical and molecular subgroups, we carried out a face-to-face comparison of our MSIMEP models with a MLH1 promoter methylation-based model in a pooled CRC cohort composed of external cohorts I to IV, which confirmed a higher sensitivity of MSIMEP to detect MSI samples, mainly due to an increased capacity to detect those non-MLH1-dependent cases.

In summary, MSIMEP could have useful applications in both basic and translational research by providing a cost-effective technique to characterize MSI status alongside methylation profiles in CRC. Future studies are needed to open MSIMEP to a wider range of tumor types, platforms, and biospecimens.

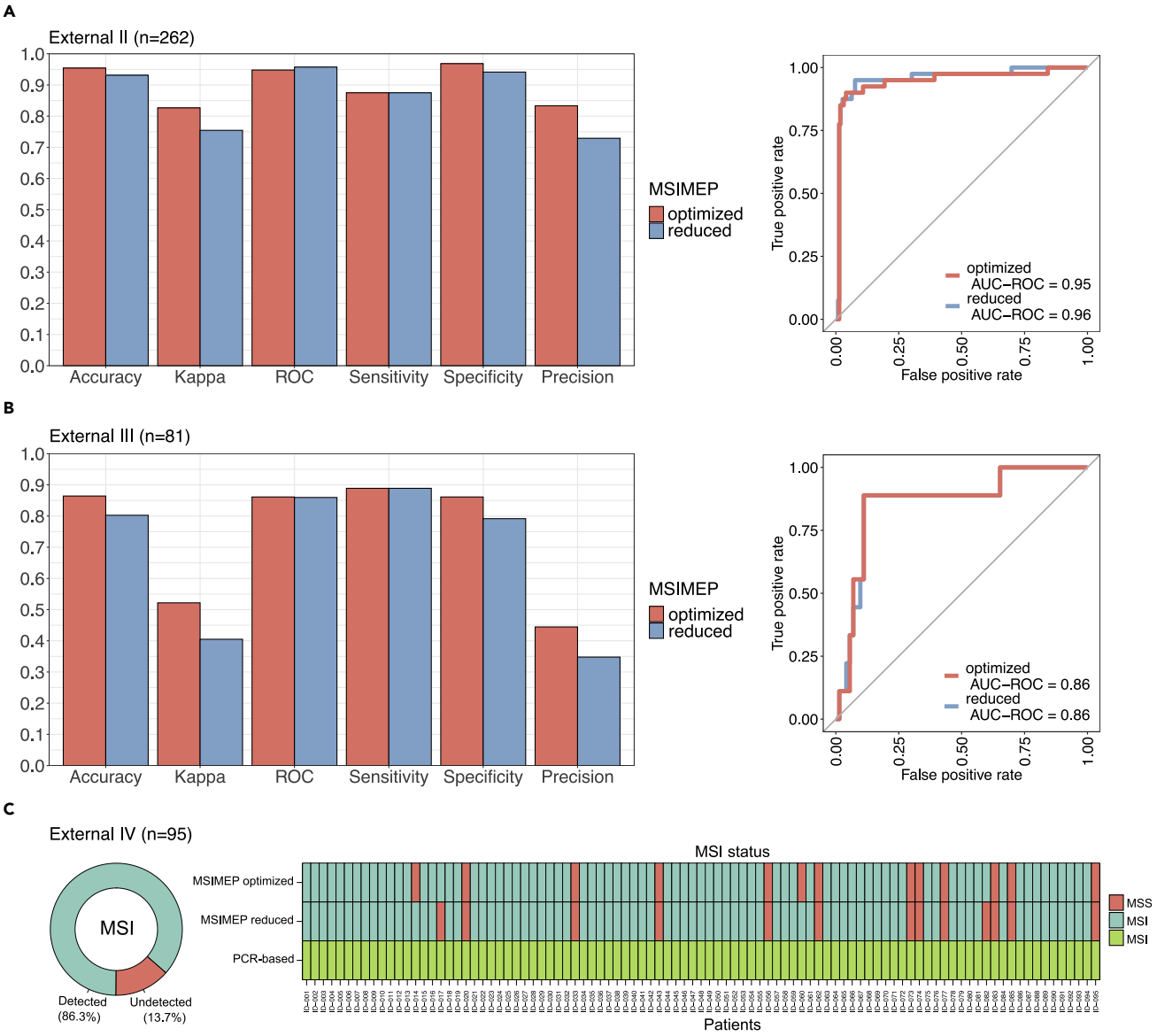


Figure 4. MSIMEP evaluation in CRC external cohorts II–IV

(A–C) Barplots showing complementary metrics (accuracy, kappa, AUROC, sensitivity, specificity, and precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients. ROC curves showing MSIMEP classification performance at alternative predictive thresholds. Models: MSIMEP optimized (red) and MSIMEP reduced (blue). Circle chart showing the fraction of detected cases in a cohort of MSI patients (ID-001 to ID-095). (A) Cohort External II (n = 262, 15% MSI), (B) External cohort III (n = 81, 11% MSI), and (C) external cohort IV (n = 95, 100% MSI).

Limitations of the study

Our tool MSIMEP has several limitations. First, because the training data used for its development comes from Infinium Human Methylation 450K arrays, it should be adapted before moving to other types of platforms. Second, because MSIMEP is focused on the prediction of MSI in CRC, its performance in other tumor types such as gastric cancer or those neoplasms with low prevalence of MSI deserves further investigation. Third, the evaluation of MSI status of tumors by using distinct PCR-based assays introduced bias that only could be solved with a centralized reevaluation. All these aspects, together with the assessment of the MSIMEP reduced model in circulating tumor DNA samples, represent areas of improvement to be prioritized to realize the translation of MSIMEP to the daily clinical practice. Furthermore, it would be of interest to evaluate if MSIMEP offers an increased value in detecting potential candidates for anti-PD-1 therapies beyond PCR-based MSI status.

Table 3. MSIMEP reduced model performance in different cohorts

Cohort	Cancer type	Phase	Sample size (n)	MSI (%)	MSS (%)	Accuracy	Kappa	AUROC	Sensitivity	Specificity	Precision
TCGA COAD/READ	CRC	Training	387	15	85	0.98	0.92	0.99	0.86	1.00	0.95
External I	CRC	Validation	79	14	86	0.97	0.90	0.99	1.00	0.97	0.85
External II	CRC	Validation	262	15	85	0.93	0.75	0.96	0.88	0.94	0.73
External III	CRC	Validation	81	11	89	0.80	0.40	0.86	0.89	0.79	0.35
External IV	CRC	Validation	95	100	NA	0.86	NA	NA	0.86	NA	1.00
External pooled	CRC	Validation	517	30	70	0.91	0.78	0.93	0.88	0.92	0.82
TCGA-STAD	STAD	Exploratory	394	18	82	0.75	0.43	0.93	0.94	0.71	0.41
TCGA-UCEC	UCEC	Exploratory	420	31	69	0.89	0.73	0.93	0.82	0.91	0.81

NA, not applicable.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - TCGA cohorts
  - External cohorts
- METHOD DETAILS
  - TCGA datasets annotation
  - Analysis of DNA methylation arrays
  - Machine-learning predictive models
  - Model description
  - MSIMEP exploratory analysis in alternative TCGA cohorts
  - MSIMEP validation in external colorectal cancer cohorts
  - Deconvolution of immune cell populations
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106127.

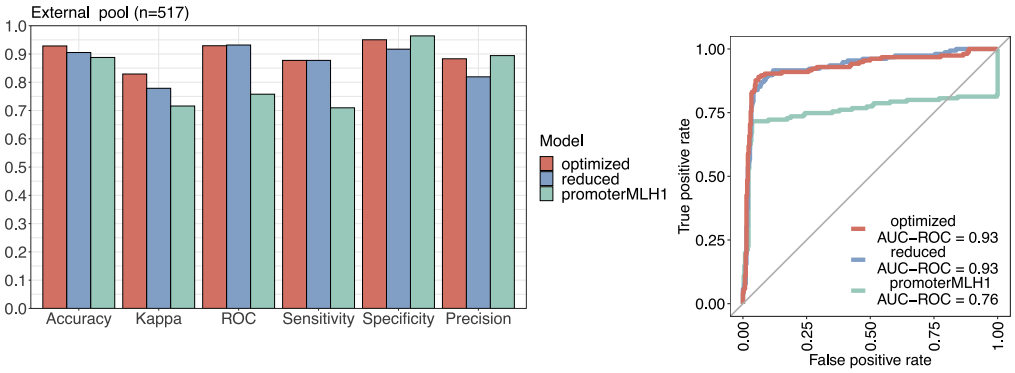


Figure 5. Comparison of the performance of three GLM models in a pooled CRC cohort (external cohorts I to IV)

Barplots showing complementary metrics (accuracy, kappa, AUROC, sensitivity, specificity, and precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients. ROC curves showing MSIMEP classification performance at alternative predictive thresholds. Models: MSIMEP optimized (red), MSIMEP reduced (blue), and promoter MLH1 (green). Cohort: external pool (n = 517, 30% MSI).

ACKNOWLEDGMENTS

This work was supported by a 2020 TTD Research Grant from the Spanish Cooperative Group for the Treatment of Digestive Tumors (TTD) to J.R.-B. M.S.-G. was supported by a predoctoral fellowship from Xunta de Galicia (ED481A-2017/299). J.B.-I. was supported by a predoctoral fellowship from the Spanish Association Against Cancer (AECC). Y.Z.B. was supported by a Programa Investigo 2022 research contract from the Consellería de Emprego e Igualdade and is supported by a predoctoral fellowship from Xunta de Galicia (ED481A 2022/491). M.F.-P. is supported by a Santander Investigación predoctoral research contract from the University of Santiago de Compostela. J.R.-B. was supported by a Río Hortega fellowship (CM19/00087) and is supported by a Juan Rodés contract (JR21/00019), both from the Institute of Health Carlos III. The results shown here are in part based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We thank the Supercomputing Centre of Galicia (CESGA) for providing complementary computational resources.

AUTHOR CONTRIBUTIONS

Conceptualization: M.S.-G. and J.R.-B.; software, M.S.-G., J.B.-I., Y.Z.B., and J.R.-B.; validation, M.S.-G. and J.R.-B.; formal analysis, M.S.-G., J.B.-I., and J.R.-B.; investigation: all authors; resources: all authors; data curation, all authors; writing – original draft, M.S.-G. and J.R.-B.; writing – review & editing, all authors; visualization, M.S.-G., Y.Z.B., and J.R.-B; supervision, J.R.-B.; project administration, J.R.-B.; funding acquisition, J.R.-B.

DECLARATION OF INTERESTS

The authors declare no competing interests related to the work described in this article.

Received: August 20, 2022
Revised: December 15, 2022
Accepted: January 31, 2023
Published: February 3, 2023

REFERENCES

1. Jiricny, J. (2006). The multifaceted mismatch-repair system. Nat. Rev. Mol. Cell Biol. *7*, 335–346. https://doi.org/10.1038/nrm1907.

2. Giardiello, F.M., Allen, J.I., Axilbund, J.E., Boland, C.R., Burke, C.A., Burt, R.W., Church, J.M., Dominitz, J.A., Johnson, D.A., Kaltenbach, T., et al. (2014). Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US multi-society task force on colorectal cancer. Am. J. Gastroenterol. *109*, 1159–1179. https://doi.org/10.1038/ajg.2014.186.

3. Ruiz-Bañobre, J., and Goel, A. (2019). DNA mismatch repair deficiency and immune checkpoint inhibitors in gastrointestinal cancers. Gastroenterology *156*, 890–903. https://doi.org/10.1053/j.gastro.2018.11.071.

4. Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M.D., Wendl, M.C., and Ding, L. (2014). MSIsensor: microsatellite instability using paired tumor-normal sequence data. Bioinformatics *30*, 1015–1016. https://doi.org/10.1093/bioinformatics/btt755.

5. Jia, P., Yang, X., Guo, L., Liu, B., Lin, J., Liang, H., Sun, J., Zhang, C., and Ye, K. (2020). MSIsensor-pro: fast, accurate, and matched-normal-sample-free detection of microsatellite instability. Dev. Reprod. Biol. *18*, 65–71. https://doi.org/10.1016/j.gpb.2020.02.001.

6. Salipante, S.J., Scroggins, S.M., Hampel, H.L., Turner, E.H., and Pritchard, C.C. (2014). Microsatellite instability detection by next generation sequencing. Clin. Chem. *60*, 1192–1199. http://clinchem.aaccjnls.org/content/60/9/1192.abstract.

7. Hause, R.J., Pritchard, C.C., Shendure, J., and Salipante, S.J. (2016). Classification and characterization of microsatellite instability across 18 cancer types. Nat. Med. *22*, 1342–1350. https://doi.org/10.1038/nm.4191.

8. Kautto, E.A., Bonneville, R., Miya, J., Yu, L., Krook, M.A., Reeser, J.W., and Roychowdhury, S. (2017). Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget *8*, 7452–7463. https://doi.org/10.18632/oncotarget.13918.

9. Li, L., Feng, Q., and Wang, X. (2020). PreMSIm: an R package for predicting microsatellite instability from the expression profiling of a gene panel in cancer. Comput. Struct. Biotechnol. J. *18*, 668–675. https://doi.org/10.1016/j.csbj.2020.03.007.

10. Chen, T., Zhang, C., Liu, Y., Zhao, Y., Lin, D., Hu, Y., Yu, J., and Li, G. (2019). A gastric cancer LncRNAs model for MSI and survival prediction based on support vector machine. BMC Genom. *20*, 846. https://doi.org/10.1186/s12864-019-6135-x.

11. Fu, Y., Qi, L., Guo, W., Jin, L., Song, K., You, T., Zhang, S., Gu, Y., Zhao, W., and Guo, Z. (2019). A qualitative transcriptional signature for predicting microsatellite instability status of right-sided Colon Cancer. BMC Genom. *20*, 769. https://doi.org/10.1186/s12864-019-6129-8.

12. Danaher, P., Warren, S., Ong, S., Elliott, N., Cesano, A., Ferree, S., Pačínková, A., and Popovici, V. (2019). A gene expression assay for simultaneous measurement of microsatellite instability and anti-tumor immune activity. J. Immunother. Cancer *7*, 15. https://doi.org/10.1186/s40425-018-0472-1.

13. Pačínková, A., and Popovici, V. (2019). Cross-platform data analysis reveals a generic gene expression signature for microsatellite instability in colorectal cancer. BioMed Res. Int. *2019*, 6763596. https://doi.org/10.1155/2019/6763596.

14. Foltz, S.M., Liang, W.-W., Xie, M., and Ding, L. (2017). MIRMMR: binary classification of microsatellite instability using methylation and mutations. Bioinformatics *33*, 3799–3801. https://doi.org/10.1093/bioinformatics/btx507.

15. Weisenberger, D.J., Siegmund, K.D., Campan, M., Young, J., Long, T.I., Faasse, M.A., Kang, G.H., Widschwendter, M., Weener, D., Buchanan, D., et al. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is

tightly associated with BRAF mutation in colorectal cancer. Nat. Genet. *38*, 787–793. https://doi.org/10.1038/ng1834.

16. Liu, M.C., Oxnard, G.R., Klein, E.A., Swanton, C., Seiden, M.V., Liu, M.C., Oxnard, G.R., Klein, E.A., Smith, D., Richards, D., Yeatman, T.J., et al. (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann. Oncol. *31*, 745–759. https://doi.org/10.1016/j.annonc.2020.02.011.

17. Mustjoki, S., and Young, N.S. (2021). Somatic mutations in "benign" disease. N. Engl. J. Med. *384*, 2039–2052. https://doi.org/10.1056/NEJMra2101920.

18. Ruiz-Bañobre, J., and Goel, A. (2021). Chapter Seven - genomic and epigenomic biomarkers in colorectal cancer: from diagnosis to therapy. In Novel Approaches to Colorectal Cancer, *151*, F.G. Berger and C. R. B. T.-A. in C. R. Boland, eds (Academic Press), pp. 231–304. https://doi.org/10.1016/bs.acr.2021.02.008.

19. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. https://doi.org/10.1101/gr.229102.

20. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer. *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

21. R Core Team (2021). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing). https://www.R-project.org/.

22. Morris, T.J., Butcher, L.M., Feber, A., Teschendorff, A.E., Chakravarthy, A.R., Wojdacz, T.K., and Beck, S. (2014). ChAMP: 450k chip analysis methylation pipeline. Bioinformatics *30*, 428–430. https://doi.org/10.1093/bioinformatics/btt684.

23. Kuhn, M. (2008). Building predictive models in R using the caret package. J. Stat. Softw. *28*, 1–26. https://doi.org/10.18637/jss.v028.i05.

24. Ren, X., and Kuan, P.F. (2019). methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. Bioinformatics *35*, 1958–1959. https://doi.org/10.1093/bioinformatics/bty892.

25. John, C.R. (2020). MLeval: Machine Learning Model Evaluation (Comprehensive R Archive Network). https://cran.r-project.org/web/packages/MLeval/index.html.

26. Chen, H., and Boutros, P.C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinf. *12*, 35. https://doi.org/10.1186/1471-2105-12-35.

27. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782. https://doi.org/10.1038/s41587-019-0114-2.

28. Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. https://doi.org/10.1038/nature11252.

29. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209. https://doi.org/10.1038/nature13480.

30. Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73. https://doi.org/10.1038/nature12113.

31. Martínez-Cardús, A., Moran, S., Musulen, E., Moutinho, C., Manzano, J.L., Martinez-Balibrea, E., Tierno, M., Élez, E., Landolfi, S., Lorden, P., et al. (2016). Epigenetic homogeneity within colorectal tumors predicts shorter relapse-free and overall survival times for patients with locoregional cancer. Gastroenterology *151*, 961–972. https://doi.org/10.1053/j.gastro.2016.08.001.

32. Mattesen, T.B., Rasmussen, M.H., Sandoval, J., Ongen, H., Árnadóttir, S.S., Gladov, J., Martinez-Cardus, A., Castro de Moura, M., Madsen, A.H., Laurberg, S., et al. (2020). MethCORR modelling of methylomes from formalin-fixed paraffin-embedded tissue enables characterization and prognostication of colorectal cancer. Nat. Commun. *11*, 2025. https://doi.org/10.1038/s41467-020-16000-6.

33. Namba, S., Sato, K., Kojima, S., Ueno, T., Yamamoto, Y., Tanaka, Y., Inoue, S., Nagae, G., Iinuma, H., Hazama, S., et al. (2019). Differential regulation of CpG island methylation within divergent and unidirectional promoters in colorectal cancer. Cancer Sci. *110*, 1096–1104. https://doi.org/10.1111/cas.13937.

34. Benhamida, J.K., Hechtman, J.F., Nafa, K., Villafania, L., Sadowska, J., Wang, J., Wong, D., Zehir, A., Zhang, L., Bale, T., et al. (2020). Reliable clinical MLH1 promoter hypermethylation assessment using a high-throughput genome-wide methylation array platform. J. Mol. Diagn. *22*, 368–375. https://doi.org/10.1016/j.jmoldx.2019.11.005.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA Colon Adenocarcinoma	GDC Data Portal	TCGA-COAD
TCGA Rectum Adenocarcinoma	GDC Data Portal	TCGA-READ
TCGA Colon and Rectal Cancer	UCSC Xena Portal	TCGA-COADREAD
TCGA Stomach Carcinoma	GDC Data Portal	TCGA-STAD
TCGA Uterine Corpus Endometrial Carcinoma	GDC Data Portal	TCGA-UCEC
External cohort I	GEO	GSE69550
External cohort II	EGA	EGAS00001004293
External cohort III	ArrayExpress	E-GEOD-68060
External cohort IV	JGA	JGAS00000000113
UCSC Genome Browser GRCh37	Kent et al., 2002[19]	https://genome.ucsc.edu
COSMIC v96 - Cancer Gene Census	Sondka et al., 2018[20]	https://cancer.sanger.ac.uk/cosmic
Software and algorithms		
R	R Core Team, 2021[21]	https://www.R-project.org/
ChAMP v2.24.0	Morris et al., 2014[22]	https://github.com/YuanTian1991/ChAMP
caret v6.0-91	Kuhn, 2008[23]	https://github.com/topepo/caret
methylGSA v1.12.0	Ren and Kuan, 2019[24]	https://github.com/reese3928/methylGSA
MLeval v0.3	John, 2020[25]	https://CRAN.R-project.org/package=MLeval
VennDiagram v1.7.1	Chen and Boutros, 2011[26]	https://github.com/cran/VennDiagram
CIBERSORTx	Newman et al., 2019[27]	https://cibersortx.stanford.edu/
MSIMEP	This paper	https://gitlab.com/mobilegenomesgroup/msimep

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Juan Ruiz-Bañobre (juan.ruiz.banobre@sergas.es).

Materials availability

This study did not generate new unique reagents or material.

Data and code availability

- Methylation data analyzed in this study is publicly available at Genomic Data Commons (GDC) data portal (TCGA-COAD, TCGA-READ, TCGA-STAD, and TCGA-UCEC) (Table S2A), Gene Expression Omnibus (GEO) data repository (GEO accession number: GSE69550) (External cohort I) (Table S2B), European Genome-phenome Archive (EGA accession number: EGAS00001004293) (External cohort II) (Table S2C), ArrayExpress Archive of Functional Genomics Data repository (accession number: E-GEOD-68060) (External cohort III) (Table S2D), and Japanese Geno-type-phenotype Archive (JGA) (JGA accession number: JGAS00000000113) (External cohort IV) (Table S2E). Accession codes for all these required datasets are included in the key resources table.
- All the original code, including MSIMEP package and associated models, has been deposited on our GitLab repository (https://gitlab.com/mobilegenomesgroup/msimep) and is publicly available.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

TCGA cohorts

Methylation datasets were downloaded from the GDC Data Portal (Project IDs TCGA-COAD, TCGA-READ, TCGA-STAD and TCGA-UCEC). Experimental Plat-form was Infinium Human Methylation 450K from Illumina. Downloaded data corresponded to 353 TCGA-COAD samples, 106 TCGA-READ samples, TCGA-STAD 397 samples and TCGA-UCEC 485 samples (Data S1). Only 450K methylation profiles from fresh-frozen (FF) primary solid tumors samples were included.

Clinical information was downloaded from the GDC Data Portal [Project IDs: TCGA-COAD (458 cases), TCGA-READ (172 cases), TCGA-STAD (443 cases), TCGA-UCEC (560 cases)]. Gene expression and single nucleotide variant information was downloaded from GDC Data Portal (TCGA-COAD) (Data S2 and Data S3). Additionally, CIMP status information of samples from TCGA COAD/READ cohort was downloaded from the UCSC Xena Portal (dataset ID: TCGA.COADREAD.sampleMap/COADREAD_clinicalMatrix).

MSI status was evaluated through PCR-based assays as previously described.[28–30] For each patient MSI status was encoded as a binary phenotype after considering MSS and MSI-Low profiles as MSS, and MSI-High profiles as MSI, in order to maximize the model predictive capacity toward relevant cases, as recommended by previous studies.[7]

Among downloaded cases, only those from primary solid tumors with available DNA methylation beta (β)-values and MSI status information were used. From the pooled TCGA-COAD/READ cohort, 387 out of 459 cases fulfilled these criteria (Table S2A). From TCGA-STAD and TCGA-UCEC cohorts, 394 out of 397 cases and 420 out of 485 cases fulfilled these criteria, respectively.

External cohorts

External cohort I

Methylation dataset was downloaded from the Gene Expression Omnibus (GEO) data repository (GEO accession number: GSE69550). Experimental Plat-form was Infinium Human Methylation 450K from Illumina. Downloaded data correspond to 79 formalin-fixed paraffin-embedded (FFPE) colorectal primary tumors. Three regions per primary tumor were selected: the region nearest the digestive tract surface (luminal), the central bulk (center), and the invasive front (front). MSI status was provided by the corresponding author upon request (Table S2B).[31]

External cohort II

Methylation dataset was downloaded from the European Genome-phenome Archive (EGA accession number: EGAS00001004293). Experimental Plat-form was Infinium Human Methylation 450K from Illumina. Downloaded data correspond to 262 FF colorectal primary tumors. MSI status and clinical characteristics were provided by the corresponding author upon request (Table S2C).[32]

External cohort III

Methylation dataset and MSI status information were downloaded from the ArrayExpress Archive of Functional Genomics Data repository (accession number: E-GEOD-68060). Experimental Plat-form was Infinium Human Methylation 450K from Illumina. Downloaded data correspond to 81 FF colorectal primary tumors. Clinical characteristics were provided by the corresponding author upon request (Table S2D).

External cohort IV

Methylation dataset was downloaded from the Japanese Geno-type-phenotype Archive (JGA) (JGA accession number: JGAS00000000113). Experimental Plat-form was Infinium MethylationEPIC array from Illumina. Downloaded data correspond to 95 FF MSI-H colorectal primary tumors. MSI status and clinical characteristics were provided by the corresponding author upon request (Table S2E).[33]

METHOD DETAILS

TCGA datasets annotation

Annotation for methylation datasets was obtained with the ChAMP v2.24.0 R package[31] retrieving relevant information: probe type, strand, genomic coordinates, CpG island context (island ID and associated

features) and gene context (gene symbol and associated features). Gene promoter was defined including probes located up to 1.5 kb upstream from the transcription start site (TSS1500 and TSS200 features). Annotation for cancer genes was obtained from the Cancer Gene Census database (COSMIC v96 release), retrieving relevant information (such as tier, role in cancer, and cancer syndrome) for each cancer gene. Genome-wide annotation for regulatory elements (curated regulatory regions, TFBS and regulatory polymorphisms) was obtained from the Open Regulatory Annotation track available at the UCSC Genome Browser database.

Analysis of DNA methylation arrays

Analysis of DNA methylation arrays was performed with the ChAMP v2.24.0 R package.[22] Methylation probe filtering was performed with *champ.filter*, setting the exclusion of (I) non-CpG probes, (II) Multi-Hit probes, (III) probes matching SNPs, and (IV) probes located in chromosomes X and Y. Methylation data imputation was performed with *champ.impute* under combined method (partial removal followed by KNN imputation) setting ProbeCutoff = 0.2, SampleCutoff = 0.2 and k = 5 parameters. Quality control steps were performed with *champ.QC* generating mdsplots, densityPlots and dendrograms for methylation distribution. Normalization of Type-I and Type-II probes was performed with *champ.norm* under peak-based correction (PBC) method. Identification of differential methylation positions was performed with *champ.DMP* by comparing MSI vs MSS samples, applying Benjamini-Hochberg (B-H) False Discovery Rate (FDR) p value adjustment after limma analysis. MLH1 promoter methylation status was defined as previously described.[34] Briefly, if β-value for all four MLH1 CpG sites cg23658326, cg11600697, cg21490561, and cg00893636 were greater than or equal to 0.18, 0.27, 0.11, and 0.10, respectively, the sample was considered hypermethylated for MLH1 promoter. Otherwise, the sample was considered non-hypermethylated.

Machine-learning predictive models

The selection of predictors was done by retaining those probes from pooled TCGA COAD/READ dataset with an absolute delta β-value ($\Delta\beta$)>0.3 between MSI and MSS, and B-H FDR adjusted p value<0.05. Both hypermethylated and hypomethylated relevant probes were included as predictors, and patient binary MSI/MSS status was considered as response variable.

Model development was performed with caret v6.0-91 R package,[23] training the following machine-learning classifiers: GLM (glmnet), KNN (knn), SVM (svmLinear), NaiveBayes (naive_bayes), NeuralNetwork (nnet), DecissionTree (rpart), and RandomForest (cforest).

Classifier performance was evaluated through 5 repeats of 10-fold cross-validation, and hyperparameter tuning was done through grid search. Average model accuracies were compared through Friedman Rank-Sum Test and Pair-wise Wilcoxon Rank-Sum Test (paired, p.adjust.method = "holm"), and the less complex model among those with significantly better performance was selected.

RFE was used to decrease model complexity, selecting an optimized model and a reduced version with suitable performance (less than 1% accuracy loss). Outer resampling method with 3 repeats of 2-fold cross-validation was applied. Predictors retained during the backward selection steps corresponded to 700, 600, 500, 400, 300, 200, 100, 50, 25, 10, 5 and 2 CpG probes.

Model description

Relative probe importance for the classification model was measured with caret *varImp*, considering the absolute value of the coefficients corresponding to the optimized GLM model. Relative gene importance was assessed by adding the absolute importance of individual CpG probes associated with each gene, normalized by the total number of genes represented in the model.

Gene set enrichment analysis was performed with the methylGSA v1.12.0 R package,[24] which implements logistic regression adjusting for the number of methylation probes. Enrichment in KEGG categories was evaluated with *methylglm* function. CpGs present in alternative gene context regions of the Infinium Human Methylation 450K array were considered (“group = promoter”, “group = body” and “group = all”). Gene sets composed by 20–500 members were displayed and ranked according to their B-H FDR-adjusted p value.

MSIMEP exploratory analysis in alternative TCGA cohorts

Prediction of MSI/MSS phenotype in TCGA-UCEC and TCGA-STAD cohorts was performed using the caret *predict* function, by assessing the methylation profiles for the same 400 predictors (MSIMEP optimized model), or 25 predictors (MSIMEP reduced model), and considering patient binary MSI/MSS status as response variable. Methylation β-values were processed under a common framework (see Analysis of DNA methylation arrays). Accuracy, Kappa, ROC, Sensitivity, Specificity, and Precision metrics were calculated from the predicted class probabilities (*predict* "type = prob") and associated confusion matrix comparing predicted and expected MSI/MSS status. ROC curves were generated using the MLeval v0.3 R package.[25]

MSIMEP validation in external colorectal cancer cohorts

Prediction of MSI/MSS phenotype in External cohort I, External cohort II, External cohort III, and External cohort IV was performed using the caret *predict* function, by assessing the methylation profiles for the same 400 predictors (MSIMEP optimized model), or 25 predictors (MSIMEP reduced model), and considering patient binary MSI/MSS status as response variable. Methylation β-values were processed under a common framework (see Analysis of DNA methylation arrays). Accuracy, Kappa, ROC, Sensitivity, Specificity, and Precision metrics were calculated from the predicted class probabilities (caret:predict "type = prob") and associated confusion matrix comparing predicted and expected MSI/MSS status. ROC curves were generated using the MLeval v0.3 R package.[25] Venn diagram was generated using the VennDiagram v1.7.1 R package.[26]

Deconvolution of immune cell populations

The CIBERSORTx deconvolution algorithm[27] was used to infer immune cell infiltration from TCGA-COAD and TCGA-READ bulk RNA-seq data. CIBERSORTx job type: "Impute cell fractions" was launched in absolute mode on the LM22 signature, with active B mode batch correction and quantile normalization disabled (recommended conditions for bulk RNA-seq mixtures), executed with 500 permutations. Low-quality samples with p value for deconvolution >0.05 were discarded from downstream analyses. Relative cell proportions were obtained by normalizing the CIBERSORTx output to the sample-level sum of cell scores rendering percentages of immune infiltration. Relative cell proportions were compared between predicted MSIMEP MSI CRCs and MSI cases diagnosed through standard PCR-based assays. Profiled immune cell types were B cells (naive and memory), plasmatic cells, T cells (CD8, CD4 naive, CD4 memory resting, CD4 memory activated, follicular helper, regulatory and gamma-delta), NK cells (resting and activated), monocytes, macrophages (M0, M1 and M2), dendritic cells (resting and activated), mast cells (resting and activated), eosinophils, and neutrophils.

QUANTIFICATION AND STATISTICAL ANALYSIS

Comparisons between patient and disease characteristics were carried out using X^2 or Fisher exact test (categorical variables), and T-Student or Wilcoxon tests (continuous variables). All p values were 2-sided, and those less than 0.05 were considered statistically significant. Statistical details can be found in table and figure legends, results and method details. All statistical analyses were performed using R.