Evaluating multiple next-generation sequencing derived tumor features to accurately predict DNA mismatch repair status

Romy Walker[1,2], Peter Georgeson[1,2], Khalid Mahmood[1,2,3], Jihoon E. Joo[1,2], Enes Makalic[4], Mark Clendenning[1,2], Julia Como[1,2], Susan Preston[1,2], Sharelle Joseland[1,2], Bernard J. Pope[1,3], Ryan Hutchinson[1,2], Kais Kasem[5], Michael D. Walsh[6], Finlay A. Macrae[7,8], Aung K. Win[2,4], John L. Hopper[4], Dmitri Mouradov[9,10], Peter Gibbs[9,10,11], Oliver M. Sieber[9,10,12,13], Dylan E. O'Sullivan[14,15], Darren R. Brenner[14,15,16], Steven Gallinger[17,18,19], Mark A. Jenkins[2,4], Christophe Rosty[1,2,20,21], Ingrid M. Winship[7,22], Daniel D. Buchanan [1,2,7#]

[1] Colorectal Oncogenomics Group, Department of Clinical Pathology, Victorian Comprehensive Cancer Centre, The University of Melbourne, Parkville, Victoria, Australia

[2] University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, Victoria, Australia

[3] Melbourne Bioinformatics, The University of Melbourne, Melbourne, Victoria, Australia

[4] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Carlton, Victoria, Australia

[5] Department of Clinical Pathology, Medicine Dentistry and Health Sciences, The University of Melbourne, Parkville, Victoria, Australia

[6] Sullivan Nicolaides Pathology, Bowen Hills, Queensland, Australia

[7] Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, Melbourne, Victoria, Australia

[8] Colorectal Medicine and Genetics, The Royal Melbourne Hospital, Parkville, Victoria, Australia

[9] Personalized Oncology Division, The Walter and Eliza Hall Institute of Medial Research, Parkville, Victoria, Australia

[10] Department of Medical Biology, The University of Melbourne, Parkville, Victoria, Australia

[11] Department of Medical Oncology, Western Health, Victoria, Australia

[12] Department of Surgery, The University of Melbourne, Parkville, Victoria, Australia

[13] Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria, Australia

[14] Department of Oncology, University of Calgary, Calgary, Canada

[15] Department of Community Health Sciences, University of Calgary, Calgary, Canada

[16] Department of Cancer Epidemiology and Prevention Research, Alberta Health Services, Calgary, Canada

[17] Ontario Institute for Cancer Research, Toronto, Ontario, Canada

[18] Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada

[19] Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada

[20] Envoi Specialist Pathologists, Brisbane, Australia

[21] University of Queensland, Brisbane, Australia

[22] Department of Medicine, The University of Melbourne, Parkville, Australia

Running Title: NGS tumor mismatch-repair deficiency

[#]To whom correspondence should be addressed:

Associate Professor Daniel D. Buchanan

Head, Colorectal Oncogenomics Group

Department of Clinical Pathology

The University of Melbourne

Victorian Comprehensive Cancer Centre

305 Grattan Street

Parkville, Victoria, 3010 Australia

Ph: +61 385597004

Email: daniel.buchanan@unimelb.edu.au

Number of text pages: 53 pages

Figures & Tables: 4 figures and 4 tables

Running head: 41 (with characters)

Abstract: 220 words

References: 76 references

Abstract

Identifying tumor DNA mismatch repair deficiency (dMMR) is important for precision medicine. We assessed tumor features, individually and in combination, in whole-exome sequenced (WES) colorectal cancers (CRCs) and in panel sequenced CRCs, endometrial cancers (ECs) and sebaceous skin tumors (SSTs) for their accuracy in detecting dMMR. CRCs (n=300) with WES, where MMR status was determined by immunohistochemistry, were assessed for microsatellite instability (MSMuTect, MANTIS, MSIseq, MSISensor), COSMIC tumor mutational signatures (TMS) and somatic mutation counts. A 10-fold cross-validation approach (100 repeats) evaluated the dMMR prediction accuracy for 1) individual features, 2) Lasso statistical model and 3) an additive feature combination approach. Panel sequenced tumors (29 CRCs, 22 ECs, 20 SSTs) were assessed for the top performing dMMR predicting features/models using these three approaches. For WES CRCs, 10 features provided >80% dMMR prediction accuracy, with MSMuTect, MSIseq, and MANTIS achieving ≥99% accuracy. The Lasso model achieved 98.3%. The additive feature approach with ≥3/6 of MSMuTect, MANTIS, MSIseq, MSISensor, INDEL count or TMS ID2+ID7 achieved 99.7% accuracy. For the panel sequenced tumors, the additive feature combination approach of ≥3/6 achieved accuracies of 100%, 95.5% and 100%, for CRCs, ECs, and SSTs, respectively. The microsatellite instability calling tools performed well in WES CRCs, however, an approach combining tumor features may improve dMMR prediction in both WES and panel sequenced data across tissue types.

Keywords: Colorectal cancer, DNA mismatch repair deficiency, endometrial cancer, Lynch syndrome, microsatellite instability, *MLH1* promoter methylation, sebaceous skin tumor, tumor mutation burden, tumor mutational signatures

Declared conflicts of interest

The authors have no conflicts of interest to declare.

Data availability statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

Funding by a National Health and Medical Research Council of Australia (NHMRC) project grant GNT1125269 (PI- Daniel Buchanan), supported the design, analysis, and interpretation of data. RW is supported by the Margaret and Irene Stewardson Fund Scholarship and by the Melbourne Research Scholarship. DDB is supported by an NHMRC Investigator grant (GNT1194896) and University of Melbourne Dame Kate Campbell Fellowship. PG is supported by the University of Melbourne Research Scholarship. MAJ is supported by an NHMRC Investigator grant (GNT1195099). AKW is supported by an NHMRC Investigator grant (GNT1194392). JLH is supported by the University of Melbourne Dame Kate Campbell Fellowship. OMS is supported by an NHMRC Senior Research Fellowship (GNT1136119). DEO is supported by a Canadian Institutes of Health Research (CIHR) Post-doctoral Fellowship. BP is supported by a Victorian Health and Medical Research Fellowship from the Victorian Government.

Introduction

DNA mismatch-repair (MMR) deficiency (dMMR) is an important molecular phenotype of solid tumors characterized by the presence of microsatellite instability (MSI) and/or loss of expression of one or more of the DNA MMR proteins, MLH1, MSH2, MSH6 and PMS2. Identifying dMMR tumors is important for understanding disease prognosis[1], response to immune checkpoint inhibition therapy[2] and to identify people with Lynch syndrome. Lynch syndrome is the most common inherited cancer predisposition disorder and, therefore, the Evaluation of Genomic Applications in Practice and Prevention Working Group recommends that all newly diagnosed colorectal (CRC) and endometrial cancers (EC) are screened for dMMR to improve the identification of carriers[3,4].

The dMMR mutator phenotype arises in tumors where errors occur during the DNA replication process[5]. Specifically, defects in the components of the MMR system responsible for the recognition of mismatches such as single nucleotide variants (SNVs) and insertion-deletions (INDELs), can lead to the development of numerous frameshift mutations in coding and non-coding microsatellite regions[6]. dMMR is related to biallelic inactivation of one of the MMR genes, resulting from either somatic methylation of the *MLH1* gene promoter region[7] or double somatic MMR gene mutations[8] (sporadic dMMR), or germline pathogenic variants in the MMR genes[9] or deletions in the 3′ end of the *EPCAM* gene[10] (inherited dMMR). CRC, EC and sebaceous skin tumors (SSTs), including sebaceous adenomas, carcinomas and sebaceomas, are tissue types that demonstrate the highest frequencies of dMMR where up to 26%[11], 31%[11] and 31%[12] of these tissue types, respectively, present with the dMMR phenotype, followed by stomach cancer at 19%[11].

The most common approach for identifying dMMR tumors is by assessing MMR protein expression through immunohistochemistry (MMR IHC)[13,14] and/or by testing for high levels of

133 microsatellite instability using polymerase chain reactions (MSI-PCR)[15]. While both screening
134 methodologies are commonly used, each present advantages and limitations. The advantages of
135 performing MMR IHC include simple experimental execution, short turnaround time, low
136 associated costs as well as giving an indication of the defective gene[16]. However, false positive or
137 false negative MMR IHC results can occur due to technical artefacts, variable performance of
138 different MMR antibodies and inherent variability in the interpretation of the staining by different
139 pathologists[16]. Further challenges include the interpretation of weaker staining in less proliferative
140 tissue and heterogenous patterns of MMR protein loss[17–24].

141 While MMR IHC is more widely adopted in the clinical setting, MSI-PCR remains the gold
142 standard for detecting dMMR[16]; to date multiple markers have been identified to call MSI in tumor
143 samples[25]. The limitations for MSI-PCRs include additional laboratory implementation
144 requirements related to tissue DNA extraction and increased labor costs; both can lead to a delay
145 in receiving test results[16]. Nonetheless, MMR IHC and MSI-PCR methodologies have proven to
146 be effective for identifying dMMR in CRC samples[26] with a reported concordance of 91.9%[16], but
147 the accuracy for either of these tools can decrease when applied to different tissue types[27]. As next-
148 generation sequencing (NGS) becomes more widely adopted for precision oncology, there is an
149 increasing need to accurately determine tumor MMR status using NGS data.

150 To date, several tools have been developed to assess MSI from NGS data, including
151 MSISensor[28], MSIseq[29], MANTIS[30] and more recently MSMuTect[31]. To the best of our
152 knowledge, the comparison of these four MSI tools on the same tumors has not yet been performed.
153 In addition to MSI, other tumor features derived from NGS have been shown to be associated with
154 dMMR, such as tumor mutational burden (TMB)[32] and tumor mutational signatures (TMS)[33].

TMB, characterized by high SNV and INDEL counts, is a biomarker for response to immune checkpoint inhibition therapy[34,35] and is increased in dMMR tumors[36].

TMS aggregate tens to thousands of the observed somatic mutations within a tumor into patterns related to the underlying mutational processes[37,38]. The predominant TMS framework, published on the COSMIC website, defines 107 different signature definitions categorized into three distinct subgroups: 1) 78 single base substitutions (SBS) where seven of the SBS signatures (SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and SBS44) are associated with dMMR; 2) 18 small (1 to 50 base pair) insertions and deletions or ID signatures where ID1, ID2, and ID7 are associated with dMMR, and 3) 11 doublet base substitutions or DBS signatures where DBS7 and DBS10 have both been previously associated with dMMR[33]. However, DBS signatures have a reported low prevalence in CRC compared with other tissue types so were excluded from our study[38]. Previously, we have shown that the combination of individual TMS can improve the ability of TMS to discriminate important molecular and genetic subtypes of CRC, including identifying germline biallelic carriers of pathogenic variants in the *MUTYH* gene by combining SBS18 and SBS36[39,40]. We further observed that the combination of ID2 with ID7 (TMS ID2+ID7) was the most informative for differentiating dMMR from pMMR CRCs amongst all possible TMS combinations[39]. To date, the comparison of MSI calling tools, somatic mutation counts, TMB and TMS tumor features for determining the dMMR status in CRC tumors has not yet been undertaken.

In this study, we assessed 104 tumor features derived from whole-exome sequencing (WES) (**Table 1**), consisting of the MSI prediction tools (MSMuTect, MANTIS, MSIseq and MSISensor), TMS (78 SBS and 18 ID signatures), TMS ID2+ID7, TMB and individual SNV and INDEL somatic mutation counts for their accuracy in predicting dMMR status in 300 well-characterized CRCs. Secondly, we investigated whether a combination of these tumor features,

using either a statistical model or a simple approach that added individual features together (additive feature combination), could improve the dMMR prediction accuracy in WES CRC tumors. Finally, we evaluated the effectiveness of the top performing tumor features from the WES analysis, individually and in combination, in an independent set of CRC, EC and SST tumors that had undergone targeted multigene panel sequencing for their dMMR prediction accuracy.

Materials and Methods

Study Cohort

The study population included men and women retrospectively identified from five studies where pMMR or dMMR status was determined by MMR IHC and where an etiology for dMMR status could be defined, namely a sporadic etiology caused by tumor *MLH1* methylation or double somatic MMR mutations, or an inherited etiology caused by a germline MMR gene pathogenic variant (Lynch syndrome). The breakdown of participants included in this study by their dMMR and pMMR status, tissue type and by WES or panel sequencing is shown in **Figure 1**:

1) the ANGELS study (*Applying Novel Genomic approaches to Early-onset and suspected Lynch Syndrome colorectal and endometrial cancers*)[39] recruited participants that were diagnosed with CRC or EC between 2014 – 2021 who were referred from family cancer clinics across Australia (n=79). All ANGELS study participants provided informed consent and the study was approved by the University of Melbourne human research ethics committee (HREC#1750748) and institutional review boards at each family cancer clinic;

2) CRC- or EC-affected participants from the ACCFR (*Australasian Colorectal Cancer Family Registry*) were selected from both population-based and clinic-based recruitment (n=139);

3) CRC-affected participants from the OFCCR (*Ontario Familial Colorectal Cancer Registry*) were population-based patients (<50 years old) recruited from the Cancer Care Ontario, Toronto, Canada (n=53). Study participants from both the ACCFR and OFCCR were recruited between 1998 and 2008, and were included according to the recruitment policy and eligibility criteria previously described[41,42]. Informed consent was obtained from all study participants and the study protocol was approved by the institutional human ethics committee at both study sites;

4) CRC-affected participants from the WEHI study (*Walter and Eliza Hall Institute of Medical Research*) were recruited from the Royal Melbourne Hospital (Parkville, VIC, Australia) and the Western Hospital Footscray (Footscray, VIC, Australia), between Jan 1, 1993, and Dec 31, 2009[39]. All patients provided written informed consent. The study was approved by human research ethics committees at both sites (HREC 12/19) (n = 80);

5) SST-affected participants from the MTS study (*Muir-Torre Syndrome Study*) were referred between July 2016 and September 2021 following clinical diagnostic MMR IHC testing by Sullivan Nicolaides Pathology service in Brisbane[12] or by family cancer clinics in Australia. Informed consent was obtained from the study participants and the study protocol was approved by the human research ethics committee from the University of Melbourne (HREC#1648355) and by the relevant institutional human ethics committees (n = 20).

Tumor Categorization

MMR IHC testing was performed on formalin-fixed paraffin embedded (FFPE) tissues for all four MMR proteins for the ACCFR and OFCCR as previously described[42–44], and a subset of these tumors also underwent MSI-PCR testing as previously described[45]. MMR IHC testing for the ANGELS and MTS studies was part of routine clinical assessment in pathology laboratories

224 across Australia, reported by the duty pathologist. Fresh-frozen tissue specimens from the WEHI
225 study were assessed for MLH1, MSH2 and MSH6 MMR IHC and MSI-PCR tested using BAT25,
226 BAT26, D5S346, D2S123 and D17S250 MSI markers. Germline MMR gene testing (as described
227 in Buchanan *et al.*[43]) and tumor *MLH1* promoter methylation testing by MethyLight (as described
228 in Buchanan *et al.*[46]) were performed on all dMMR tumors showing loss of MLH1/PMS2 protein
229 expression or sole PMS2 loss by IHC. Tumors were considered to have double somatic MMR
230 mutations when they were found to have two pathogenic/likely pathogenic somatic mutations or a
231 single somatic pathogenic/likely pathogenic mutation in combination with presence of loss of
232 heterozygosity. Germline pathogenic variants and somatic MMR gene mutations were confirmed
233 in WES and targeted panel sequencing data prior to analysis. Therefore, for each of the dMMR
234 tumors included in this study we could confirm an inherited or acquired cause for their respective
235 pattern of MMR IHC protein loss. Concurrently, for the pMMR tumors, we did not find evidence
236 of a germline MMR pathogenic variant or double MMR somatic mutation in these tumor samples.
237 All tumors in the study were assigned to one of four categories based on dMMR or pMMR
238 status determined from MMR IHC and/or MSI-PCR and based on the cause for dMMR:
239 **1) dMMR-Lynch syndrome (dMMR-LS)** – identified carrier of a germline pathogenic variant
240 in one of the DNA MMR genes where the corresponding tumor showed commensurate loss of
241 MMR protein expression by IHC;
242 **2) dMMR-*MLH1* methylation (dMMR-MLH1me)** – tumors were positive for methylation of
243 the *MLH1* gene promoter "C region"[47] and showed loss of MLH1 and PMS2 protein expression
244 by IHC without a germline MMR gene pathogenic variant;

3) dMMR-double somatic (dMMR-DS) – tumors harbored two somatic mutations (SNVs and/or loss of heterozygosity) in the same MMR gene that showed loss of protein expression by IHC with no identified pathogenic germline MMR gene variant; and

4) MMR-proficient (**pMMR**) – tumors showed normal expression of all four MMR proteins and did not show presence of double somatic MMR gene mutations or a germline MMR gene pathogenic variant.

The three dMMR subtypes dMMR-LS, dMMR-DS and dMMR-MLH1me were combined as a single dMMR tumor group in downstream analysis.

Whole-Exome and Targeted Panel Sequencing Capture Regions

The targeted panel was based on the design described in Zaidi *et al*.[48] consisting of probes targeting the following regions: 1) 298 genes incorporating key hereditary CRC[49–51] and EC[52] risk genes and genes that are frequently mutated as identified by The Cancer Genome Atlas (TCGA) data[32,53,54], 2) 28 microsatellite loci including the five 'gold standard' MSI markers (BAT25, BAT26, NR-21, NR-24, and MONO-27) currently implemented in routine MSI-PCR diagnostics, 3) 212 homopolymer regions distributed genome-wide to assess for MSI in tumor samples and 4) 56 copy number variants known to be susceptible to copy number changes in CRCs. The panel capture was 2.005 megabases (Mb) in size. The WES capture incorporates all exonic regions within the genome and is 67.296 Mb in size. The panel additionally included capture of intronic regions within the MMR genes, which the WES capture did not cover.

265 ***Next-Generation Sequencing***

266 In total, 300 CRC tumors were sequenced by WES and 71 tumors (29 CRCs, 22 ECs and
267 20 SSTs) were sequenced by the targeted multigene panel (**Figure 1**). FFPE CRC, EC or SST
268 tissues were macrodissected and DNA extracted using the QIAmp DNA FFPE Tissue Kit (Qiagen,
269 Hilden, Germany) according to the manufacturer's instructions. Peripheral blood-derived DNA
270 was extracted using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany) and sequenced as
271 germline references.

272 The WES capture was the Agilent Clinical Research Exome V2 kit (Agilent Technologies
273 Santa Clara, United States) with sequencing performed on an Illumina NovaSeq 6000 comprising
274 150 base pair (bp) paired-end reads performed at the Australian Genome Research Facility[39]. For
275 the WEHI CRCs, exome-enrichment was performed using the TruSeq Exome Enrichment Kit
276 (Illumina, San Diego, United States) and 100 bp paired-end read sequencing performed on an
277 Illumina HiSeq 2000 at the Australian Genome Research Facility[39]. The on-target coverage for the
278 300 WES samples had a median of 323.7 for the FFPE tumor DNA samples and 137.4 for blood-
279 derived DNA samples, with an interquartile range of 111.8 – 426.4 and 100.6 – 204.9, respectively.

280 Library preparation for targeted panel sequencing was performed using the SureSelect^{TM}
281 Low Input Target Enrichment System (Agilent Technologies, Santa Clara, United States) using
282 standard protocol and sequenced on an Illumina NovaSeq 6000 comprising 150 bp paired end
283 reads performed at the Australian Genome Research Facility. The on-target coverage for the 71
284 panel sequenced samples was (median and interquartile range) 919.3 and 694.6 – 1164.9 for FFPE
285 tumor DNA samples and 160.6 and 135.8 – 178.0 for blood-derived DNA samples.

286 ***Bioinformatics Pipeline***

287 For both WES and targeted panel sequenced samples, adapter sequences were trimmed
288 from raw FASTQ files using trimmomatic 0.38[55] and aligned to the GRCh37 human reference
289 genome using Burrows-Wheeler Aligner v. 0.7.12. Germline variants, somatic variants (SNVs)
290 and somatic INDELs were called using Strelka (v. 2.9.2., Illumina) using the recommended
291 workflow[56]. TMS were calculated using the pre-defined set of 78 SBS and 18 ID signatures
292 published on COSMIC as version 3.2 (COSMIC, https://cancer.sanger.ac.uk/signatures/, last
293 accessed date: June 15, 2022)[33]. Variants outside the WES and panel capture regions were
294 excluded and variants with the PASS filter called from Strelka were retained. Additional variant
295 filters included were restrictions to a minimum depth of 50x for germline and tumor samples with
296 a minimum variant allele frequency of 10% as detailed previously[39].

297

298 ***Selection of Features of Interest***

299 The 104 tumor features selected for analysis in this study are shown in **Table 1**. Several
300 tools have been developed to assess MSI from NGS data. Our analysis focused on MSMuTect[31],
301 MANTIS[30], MSIseq[29] and MSISensor[28]. Tumors were classified as having high levels of MSI
302 (MSI-H) or as microsatellite stable (MSS). We assessed all SBS (n=78) and ID (n=18) TMS as
303 described by COSMIC[33], but the DBS TMS were excluded due to their reported low prevalence
304 in CRCs[38]. Previously, we have shown that combining ID2 and ID7 TMS enabled detection of
305 dMMR CRCs[39] and, therefore, was included as a tumor feature in this study. Somatic mutation
306 counts, namely SNVs or INDELs, as well as TMB (SNV and INDEL mutation count combined /
307 Mb) were each included, given previous associations with tumor dMMR status[57].

308

309 ***Feature Performance Evaluation in WES data from CRCs***

310 We assessed the 104 tumor features calculated from WES from 209 pMMR CRCs and 91
311 dMMR CRCs (pMMR:dMMR ratio = 2.3:1) (**Figure 1**). The dMMR CRCs comprised dMMR-LS
312 tumors (n=49), dMMR-*MLH1*me tumors (n=26) and dMMR-DS tumors (n=16). All 300 CRCs
313 were randomly partitioned into a training set (80% of CRCs) and a test set (20% of CRCs), while
314 maintaining the same pMMR:dMMR ratio, using *caret* R package[58]. We performed a 10-fold cross
315 validation approach on the training set (repeated 100x) to calculate the average classification
316 accuracy by fitting a generalized linear model and determining the error rate, specificity,
317 sensitivity, and the area under the curve (AUC) with corresponding 95% confidence intervals
318 (CIs). Based on the unequal distribution of dMMR and pMMR tumors in the WES dataset, the *no*
319 *information rate* was 69.5%, indicating that any feature with this prediction accuracy was
320 equivalent to selecting a dMMR sample by chance.
321
322 Tumor feature analysis of the WES CRC dataset comprised of three different approaches:
323 *A) Individual tumor feature assessment*
324 Each of the 104 tumor features were assessed individually and then ranked by their accuracy in
325 identifying dMMR tumors. Individual CRC tumor features with a prediction accuracy >80% from
326 the WES data were considered good predictors for differentiating dMMR from pMMR tumors and
327 were included in downstream analyses.
328
329 *B) Generation of a statistical model by combining tumor features*
330 We investigated whether combining tumor features using a Lasso penalized regression model[59]
331 could improve the overall dMMR prediction accuracy in CRC. Lasso enables the simultaneous

parameter estimation and variable selection as well as having been shown to reduce overfitting when compared to conventional maximum likelihood regression models. Lasso regression has a tuning parameter called lambda that controls which features are included in the regression model by shrinking the coefficient or "weighting" of individual features within the model towards zero, helping with the exclusion of some of the features from integration into the final model via a penalization process using cross-validation.

C) Applying an additive feature combination count

Our third approach investigated combining the top ranked individual tumor features in an additive approach (additive feature combination). Specifically, the tumor features that achieved a mean prediction accuracy >95% from the WES CRC analysis (from part A), were included in this approach and added together to give an overall count. The bimodal distribution supported a majority vote decision on dMMR status.

Assessment of individual tumor features, the statistical model and additive feature combination approaches derived from the WES analysis on panel sequenced CRCs, ECs, and SSTs

The top individual tumor features determined from (A), best performing Lasso model (B) and the additive feature combination approach (C) were then assessed for their dMMR prediction accuracy in three independent tumor sets comprised of n=29 CRCs, n=22 ECs and n=20 SSTs tested by targeted multigene panel sequencing. The *no information rate* for features analyzed from the panel dataset was at 71.8%, indicating a prediction accuracy of this value was similar to selecting a dMMR sample by chance.

Statistical Analysis

All statistical analyses were done using the R programming language (v.4.1.0). The *tidyverse* package (v.1.3.1.)[60] was used for data import, tidying and visualization purposes and the *caret* (v.6.0-9.0) package[58] was used for cross-validation. Receiving operator curves (ROC) were generated using the *pROC* package (v.1.18.0)[61], with the AUC being determined using the *cvAUC* package (v.1.1.4)[62]. Statistical models were fitted using the Lasso (*glmnet*, v.4.1-3)[63] package. We used the *cutpointr* (v.1.1.1) package[64] for estimation of the best "cut points" or "thresholds" which maximize the Youden-index (true positive rate minus false positive rate over all possible cut points), defined as the most optimal threshold in binary disease classification tasks. Here, the *cutpointr* package determines a recommended threshold that best differentiates dMMR from pMMR cases for each feature and validates its performance using bootstrapping. The average weight for each group was calculated using the *plyr* (v.1.0.7) package[65]. The *ggplot2* (v.3.3.5) package[66] was used for data visualization in combination with *hrbrthemes* (v.0.8.0)[67] for histogram generation and *ggrepel* (v.0.9.1)[68] for histogram annotations. Correlation scores between the dMMR and pMMR groups were estimated by a *heteroscedastic two-tailed t-test*. P-values <0.05 were considered statistically significant. The 95% CIs for the WES data were calculated using the binomial (Clopper-Pearson) "exact" method[69] and for the targeted panel data using the *binom* (v.1.1-1) package[70] in R.

Results

For the initial performance evaluation of 104 tumor features we assessed 209 (69.7%) pMMR CRCs and 91 (30.3%) dMMR CRCs sequenced by WES. The clinicopathological characteristics, pattern of MMR IHC loss and dMMR etiology are summarized in **Supplementary Table 1.** The

mean age at CRC diagnosis (± standard deviation, SD) for the dMMR group was 51 ± 15.0 with 62.6% being female and 49 ± 16.3 with 55.5% being female for the pMMR group. The clinicopathological characteristics, pattern of MMR IHC loss and dMMR etiology for panel sequenced CRC (n=29), EC (n=22) and SST (n=20) tumors are summarized in **Supplementary Table 2**. Within the panel sequenced tumors, the proportion of dMMR for the CRC, EC and SST subsets was 72.4% (21/29), 81.8% (18/22) and 65.0% (13/20), respectively. The predominant dMMR subtype across the CRC WES and targeted panel sequenced tumors was dMMR-LS (53.8% and 66.7%, respectively). Within the dMMR subgroup, the most predominant pattern of loss observed in CRCs and ECs was MLH1/PMS2 (WES CRCs: 65.9%, panel CRCs: 47.6% and ECs: 50.0%), whereas for the SSTs tumors, this was MSH2/MSH6 loss (76.9%). Tumors showing less common patterns of MMR loss including solitary loss of MSH6 or PMS2 by IHC were present in both the WES CRCs (16.5%) and panel sequenced tumors (19.2%), however, sole PMS2 loss cases were absent from the EC and SST cohorts.

Assessment of Tumor Features for dMMR Prediction Accuracy in WES CRCs

A) Individual tumor feature assessment

Twelve of the 104 tumor features derived from WES had a mean dMMR prediction accuracy >80% on the test dataset (**Table 2**). The mean accuracy for the remaining 92 features is shown in **Supplementary Table 3.** The four MSI tools were among the best predictors, with MSMuTect, MSIseq and MANTIS each achieving a mean prediction accuracy of ≥99.0% with MSMuTect achieving the highest accuracy (99.3%, 95% CI: 99.1%-99.5%) (**Table 2**). The combination of TMS ID2+ID7 achieved an accuracy of 96.8% (95% CI: 96.4%-97.2%), and outperformed these signatures individually (**Table 2**). To avoid collinearity issues between the

combined TMS ID2+ID7 variable with the individual TMS ID2 and TMS ID7 features, the latter were excluded from downstream analysis as they provided a lower prediction score. Therefore, the remaining 10 features were considered as the top 10 dMMR predictors and included in subsequent analyses (**Figure 1**).

The mean, SD, and range of values for each of these top 10 dMMR predictive features by MMR status and by dMMR subtype for the 300 WES CRCs are shown in **Supplementary Table 4**. For each of these features, the mean values were significantly different between the dMMR and pMMR CRCs (all $p<1x10^{-12}$ from a *two-tailed t-test*), with TMS ID2+ID7 showing the most significant difference (p-value = $7.775x10^{-98}$), although MSISensor presented with the highest Cohen's *d* effect size of 4.5, indicating that the means of the pMMR and dMMR groups differed by more than four times the SD (**Supplementary Table 4**). The variation in proportion or counts was larger in the dMMR tumors than in the pMMR tumors for all but one of these top 10 features where TMS ID2+ID7 demonstrated a broad range of values in the pMMR CRCs compared with the dMMR CRCs (**Figure 2, Supplementary Table 4**).

The AUCs for the top 10 features when taking all possible thresholds into account are shown in **Supplementary Figure 1.** The MSI prediction tools MSMuTect, MSIseq, and MANTIS as well as INDEL count demonstrated the best AUCs. In addition, we calculated recommended thresholds for each feature for differentiating dMMR from pMMR CRCs using the methodology described in the methods (**Supplementary Table 5**). When applying these thresholds, it was not possible to achieve a complete separation between the dMMR and pMMR tumors for each of the tumor features (**Figure 3**).

Investigation of the CRCs misclassified based on the individual tumor feature analysis demonstrated that the misclassification rate (error rate) for the MSI tools was low with MSMuTect

(2/300), MANTIS (1/300), MSIseq (1/300) and MSISensor (5/300) calling ≤5 incorrectly out of 300 tumors (≤1.7% error rate). Of the CRCs misclassified by the MSI tools, only two tumors were misclassified by more than one MSI tool, both were dMMR-MLH1me CRCs classified as pMMR. Of note, one of these dMMR-MLH1me CRCs was misclassified as a pMMR tumor by 9 out of the top 10 tumor features. The second misclassified dMMR-MLH1me CRC was classified as pMMR by MSMuTect and MSISensor but classified as dMMR by MSIseq and MANTIS (overall 6/10 features classified this CRC as dMMR). For INDEL count, 3/300 were incorrectly classified, where two pMMR CRCs were classified as dMMR. TMS ID2+ID7 had 10/300 incorrect classifications with seven pMMR tumors incorrectly called as dMMR. The remaining features from the top 10 prediction accuracy list demonstrated the following incorrect classifications: SBS20 (34/300), SBS54 (55/300), SBS15 (44/300) and TMB (19/300) encompassing incorrect calls in both directions (dMMR to pMMR and vice versa).

B) Generation of a statistical model by combining tumor features

We assessed whether a combination of features within a statistical model could improve dMMR prediction accuracy. For this, we performed a Lasso penalized logistic regression. Here, after calculating the best lambda value, we found that the combination of TMS ID2+ID7 (coefficient = 5.29), MANTIS (coefficient = 1.70), MSISensor (coefficient = 0.09) with SBS15 (coefficient = 2.25) provided the best prediction accuracy from all possible feature combinations, demonstrating a mean accuracy of 98.3% (95% CI: 0.981-0.986), sensitivity of 0.973 (95% CI: 0.966-0.980) and specificity of 1.000 (95% CI: 1.000-1.000) on the test set.

445 *C) Assessing an additive feature combination count for dMMR prediction*

446 Based on the observation that the top performing tumor features from the individual feature
447 analysis did not all misclassify the same CRCs lead us to explore a novel approach of combining
448 tumor features together to increase the overall accuracy i.e., an additive tumor feature combination
449 approach. This approach used a majority count of individual tumor features to overcome the small
450 inaccuracies that each of the top tumor features displayed individually i.e., if one of these top
451 dMMR predictive tumor features misclassified a CRC then the other top dMMR predictive tumor
452 features would correctly classify the same CRC and, thereby, achieve the correct classification
453 overall. Six of the top 10 features from the 10-fold cross-validation analysis demonstrated a mean
454 prediction accuracy of >95% and thus had the least number of incorrect CRC tumor classifications,
455 consisting of MSMuTect, MANTIS, MSIseq, MSISensor, INDEL count, and TMS ID2+ID7. We
456 applied the recommended threshold for determining dMMR status determined previously for each
457 tumor feature (**Figure 3**, **Supplementary Table 5**) to derive a count out of these six selected
458 features, in which each feature is weighted equally. The results show a bimodal distribution across
459 the 300 CRCs (**Figure 4)** where 0/6 to 2/6 features correctly classified all the pMMR CRCs and
460 4/6 to 6/6 correctly classified all but one of the dMMR tumors with an accuracy of 99.7%. The
461 only exception was the previously mentioned dMMR-MLH1me tumor, which did not meet the
462 recommended thresholds for all six features and thus received a count of 0/6 features suggestive
463 the CRC is pMMR rather than its initial dMMR status.
464
465 A summary of the results from the WES CRC analysis for the three approaches is shown
466 in **Table 3** and **Figure 1**.
467

Assessment of individual tumor features, Lasso statistical model and additive feature combination approaches derived from the WES analysis on panel sequenced CRCs, ECs, and SSTs

To determine the generalizability of the findings from the three approaches performed on the WES CRCs, we tested 71 tumors with targeted panel sequencing data to evaluate performance on both a smaller capture and across different tissue types known to have a high prevalence of dMMR.

A) Evaluation of the top performing individual features from WES analysis on the panel sequenced CRC, EC, and SST tumors

Out of the top 10 dMMR tumor features from the WES CRC analysis, only four achieved a mean dMMR prediction accuracy of >80% in the panel sequenced CRC tumors (**Table 4**). For EC and SST tumors only one feature (MANTIS) and two features (MANTIS and TMS ID2+ID7), respectively, of the top 10 tumor features achieved a mean dMMR prediction accuracy of >80% (**Table 4**). Across the three tissue types, MANTIS demonstrated the highest mean accuracy, achieving 100% (95% CI: 88.1%-100.0%) accuracy in the panel sequenced CRCs, 86.4% accuracy in ECs (95% CI: 65.1%-97.1%) and 85% accuracy in SSTs (95% CI: 62.1%-96.8%) (**Table 4**). MSMuTect and INDEL count performed poorly in all three panel sequenced tissue types compared with their accuracy in the WES CRCs. MSMuTect and INDEL count are features that provide absolute counts that in our data were two orders of magnitude smaller in the panel sequenced tumors compared with the WES CRCs. The reduction in discriminatory ability is likely related to differences in the size (WES: 67.7 Mb and panel: 2.0 Mb) and location (additional coverage of intronic regions of the MMR genes in the panel capture) of the regions covered by the WES and panel captures resulting in a lower somatic mutation count.

The mean, SD, and range of values for each of these top 10 dMMR predicting features by MMR status and by dMMR subtype for each of CRC, EC and SST tissue types are shown in **Supplementary Tables 6A, 6B, 6C** and in **Supplementary Figure 2, Supplementary Figure 3,** and **Supplementary Figure 4,** respectively. The mean values of each of the top 10 predictors were significantly different between the dMMR and pMMR tumors in all three tissue types except for TMS SBS15 in CRCs, MSISensor in ECs, TMB in ECs and SSTs and, TMS SBS20 and TMS SBS54 in SSTs. MSMuTect consistently had the highest Cohen's *d* effect size of all top 10 tumor features for each tissue type with the highest effect size observed in CRCs (3.2), indicating the mean of the dMMR and pMMR subgroups for this feature differ by approximately three SDs.

B) Evaluation of the Lasso statistical model on the panel sequenced CRC, EC, and SST tumors

From WES analysis, the Lasso statistical model comprised of TMS ID2+ID7, MANTIS, MSISensor and SBS15 achieved a mean prediction accuracy of 98.3%. When this model was applied, with the coefficients determined from the WES analysis, on these three independent panel sequenced tissue types, the prediction accuracies were lower (CRC: 89.7%, EC: 68.2% and SST: 85.0%) (**Table 3**).

C) Evaluation of the additive tumor feature combination approach on the panel sequenced CRC, EC, and SST tumors

For each of the top 10 dMMR predictive tumor features we determined the optimal thresholds for the panel sequenced CRCs, ECs, and SSTs (**Supplementary Table 5**) and plotted them by tissue type (CRC - **Supplementary Figure 5**), (EC - **Supplementary Figure 6**), (SST - **Supplementary Figure 7**). The determined thresholds for MANTIS were consistent across both

WES and panel captures as well as across tissue types while the calculated thresholds for MSIseq were consistent for CRC across WES and panel captures but different to the thresholds determined for EC and SST. The remaining eight tumor features showed variability in their determined thresholds across both capture type and tissue type (**Supplementary Table 5**). As such, we applied the thresholds determined for each tissue type for the panel sequenced data in the additive feature combination approach below.

The additive feature combination approach incorporates a count of MSMuTect, MANTIS, MSIseq, MSISensor, INDEL count and TMS ID2+ID7 tumor features to classify a tumor as dMMR. The distribution of the counts of these six tumor features determined for each tumor are shown for CRC (**Supplementary Figure 8**), EC (**Supplementary Figure 9**) and SSTs (**Supplementary Figure 10**). For each tissue type, all the dMMR tumors had ≥3/6 tumor features classify them as dMMR, except for a single dMMR-MLH1me EC (1/71, 1.4%) which scored 0/6 and, therefore, was suggestive of pMMR status. This approach achieved accuracy scores of 100%, 95.5% and 100%, for CRC, EC and SST, respectively (**Table 3**).

A summary of the WES CRC and CRC, EC, and SST panel sequencing results for all three approaches is provided in **Table 3**.

Discussion

In this study, we compared tumor features calculated from next generation sequencing data for their accuracy in predicting dMMR status in 300 CRCs, 91 of which were dMMR determined by immunohistochemistry or MSI-PCR and with an established sporadic or inherited etiology for

their dMMR status. Ten features achieved >80% dMMR prediction accuracy from the WES CRC tumors, with the highest accuracy predictors being the MSI tools MSMuTect, MSIseq and MANTIS, all of which achieved ≥99% accuracy. The combination of TMS ID2+ID7 achieved the highest mean accuracy for dMMR prediction out of the 97 TMS features assessed. When applied to the targeted multi-gene panel setting, the performance of these 10 features was reduced not only in CRC but also for the EC and SST tumors. In addition, we investigated two approaches that combined these top 10 performing tumor features to improve the overall prediction accuracy. The Lasso generated model achieved 98.3% accuracy in WES CRCs although the performance of the model was reduced in the panel sequenced CRC, EC, and SST tumors. For both the WES CRCs and panel sequencing across tissue types, the additive tumor feature combination approach, where having ≥3 of the top 6 tumor features classify a tumor as dMMR, achieved the highest prediction accuracies of the three approaches tested.

To date, multiple tools to detect MSI from NGS data have been developed[71]. NGS based MSI tool development has been constantly evolving since the introduction of MSISensor[28] and mSINGS[72], which were followed by MSIseq[29], MANTIS[30] and MSMuTect[31]. However, to the best of our knowledge, neither a comparison of more than three MSI detection tools on the same tumor sample nor the effectiveness of these MSI tools specifically on SST tumors has been performed to date. Previously, MANTIS has been compared to MSISensor with the former showing superior sensitivity (97.18% vs. 96.48%) and specificity (99.68% vs. 98.73%)[30]. This was supported by our findings, and we additionally showed that across the WES and panel tested CRCs, MANTIS provided the highest dMMR prediction accuracy and was shown to be the top performing feature in the EC and SST tumors as well. Recently, the United States Food and Drug Administration

559 (FDA) approved MSISensor for detecting MSI in metastatic CRCs for selecting patients for
560 immune checkpoint inhibition therapy[73]. In our study, MSISensor had the lowest accuracy (97.7%)
561 in WES CRCs of the four MSI tools tested, incorrectly classifying 5/300 CRCs. Seeking FDA
562 approval for other MSI tools in addition to MSISensor is warranted based on our findings.
563
564 MSMuTect has been trained on 20 different tissue types using WES data and, therefore, it
565 was not surprising it had the highest mean accuracy of the top performing tumor features in our
566 WES CRC analysis. MSMuTect has been designed to accurately detect somatic MSI indels using
567 a count of indels from the captured sequencing region[31]. Thus, the MSI indel count from WES data
568 (67.7 Mb) could be up to ~34x larger than that from panel data (2.0 Mb), which likely explains the
569 poor performance of this tool observed in our panel sequencing data test sets. When we adjusted
570 the MSMuTect threshold for calling dMMR for panel data, MSMuTect showed improved
571 discrimination of dMMR from pMMR tumors. This increase in prediction accuracy was also
572 observed for the INDEL count where adjusting the threshold for panel data improved the overall
573 performance. Adjusting the threshold for panel sequencing data enabled the inclusion of
574 MSMuTect and INDEL count as two of the six tumor features in our additive feature combination
575 approach that ultimately performed well on panel sequenced tumors. Tumor features that calculate
576 a percentage rather than raw counts such as MANTIS, MSISensor, SBS TMS and ID TMS are
577 more adaptable to changes in capture size. For example, our results showed that the calculated
578 thresholds for differentiating dMMR from pMMR for MANTIS were consistent across both WES
579 and panel captures as well as across tissue types. Therefore, we recommend training features that
580 incorporate a count of genomic variants, such as INDELs, SNVs and MSMuTect on the capture
581 size to improve dMMR prediction accuracy.

582

583 While three ID TMS (ID1, ID2 and ID7) are reported to be associated with dMMR[33], our
584 results showed that the combination of ID2 and ID7 TMS achieved the highest dMMR prediction
585 accuracy of any of the TMS features in WES CRC tumors, outperforming ID2 or ID7 alone. Of
586 the seven SBS TMS that are associated with dMMR (SBS6, SBS14, SBS16, SBS20, SBS21,
587 SBS26 and SBS44)[33], only two, TMS SBS15 and TMS SBS20, showed >80% dMMR prediction
588 accuracy in WES CRC tumors, but were shown to be poor predictors in the panel sequenced
589 tumors. Interestingly, TMS SBS54 was one of the top 10 dMMR predictors from the WES CRC
590 analysis, although currently its proposed etiology in COSMIC is related to a "possible sequencing
591 artefact and/or a possible contamination with germline variants"[33]. Another study has shown that
592 SBS15, SBS20 and SBS54 are observed in CRCs with a high immune cytolytic activity (CYT)
593 compared with CYT-low CRCs[74]. CYT-high CRCs have been shown to correlate with an increased
594 somatic mutation load and high levels of MSI[75], this may explain the observation of TMS SBS15,
595 TMS SBS20 and TMS SBS54 demonstrating >80% dMMR prediction accuracy in our WES CRC
596 analysis.

597

598 The combination of tumor features via the Lasso regression model achieved similar mean
599 accuracy as the four MSI tools individually in the WES CRC analysis. The Lasso calculated final
600 model that best distinguished dMMR from pMMR tumors in the WES CRC cohort consisted of
601 TMS ID2+ID7, MANTIS, MSISensor and TMS SBS15. The statistical approach used to determine
602 the final model assigns a 'weight' (coefficient value) or confidence of how well each feature
603 detects dMMR. As per generalized linear modelling methodology, the weight of any given feature
604 is reduced as the model incorporates additional features. Hence, with MANTIS being one of the

best predictors, its weighting was reduced when other features were added to the final model. This resulted in the Lasso model prediction accuracy being lower than MANTIS alone. Of note, since most of the approaches taken (i.e., assessing features individually or in combination) already achieved a very high prediction accuracy of ~99%, alternate modelling approaches such as Random Forest would not result in a significant improvement in dMMR prediction accuracy.

Strengths of our study were a large sample of tumors including dMMR tumors with confirmed sporadic or inherited etiology concordant with MMR IHC and MSI-PCR results for both the WES and panel sequenced datasets. Tumor MMR status combined with identified etiology provided a more reliable reference group of CRCs than would a group based on MMR IHC test results without etiological confirmation given the known challenges that can lead to false positive and negative MMR IHC results[16]. We assessed many tumor features that can be readily derived from NGS data ensuring that our findings have potential to be easily implemented in clinical diagnostics. We applied our findings from WES to panel data to determine the generalizability of our findings to smaller panel captures such as those that are currently used in clinical diagnostics. We showed the applicability of our findings on tissue types that display a high proportion of dMMR phenotype. Our dMMR tumor samples included those with the frequent pattern of MMR IHC namely MLH1/PMS2 loss and MSH2/MSH6 loss but also tumors with solitary MSH6 loss or solitary PMS2 loss, ensuring we covered the spectrum of dMMR tissue types which is particularly relevant given the identified challenges associated with interpretation of solitary MSH6 loss[76].

There were several limitations of our study including testing of only three tissue types. Testing of these tumor features and approaches in other tissue types such as stomach cancer, which also has a high prevalence of dMMR overall and dMMR related to Lynch syndrome, would determine the suitability of these tumor features for inclusion in an additive feature combination approach in a pan-cancer setting. In addition, the sample size for the panel sequenced tumors was limited for all three tissue types, however, there was a high proportion of dMMR in the tumors tested (72.4% for CRC, 81.8% for EC and 65.0% for SST). No tumor feature or approach achieved 100% accuracy in the CRC WES analysis. This was largely related to a single tumor (dMMR-MLH1me) from the WES CRC analysis that was called incorrectly by 9/10 top individual tumor features suggesting the CRC was pMMR. Therefore, we repeated the *MLH1* methylation testing for this tumor using both MethyLight and MS-HRM assays. Both assays found no evidence of *MLH1* methylation in the tumor. These new *MLH1* methylation results and the pMMR classification from our analysis suggest the initial dMMR classification was a false positive. If this CRC would initially have been categorized as a pMMR tumor, then MANTIS and MSIseq would have achieved 100% accuracy in the WES CRC analysis. Furthermore, the identification of an initial tumor misclassification provides strong support for evaluating multiple dMMR prediction tumor features and highlights the advantage of combining these features through an additive feature combination approach.

Conclusion

Our findings provide an important comparison of tumor features for dMMR prediction, highlighting performance differences between capture size and tissue types. Our results demonstrate the high accuracy of multiple individual tumor features including the MSI calling

650 tools MSMuTect, MSIseq, MANTIS and MSISensor, as well as INDEL count and the combination
651 of TMS ID2+ID7 for predicting dMMR status using WES CRCs. Moreover, our findings highlight
652 the benefit of combining these six tumor features in a simple additive feature combination
653 approach to improve dMMR prediction accuracy, particularly in targeted panel sequencing data
654 from CRC, EC, or SST tumors. With the reported inaccuracies of MMR IHC and the increasing
655 application of clinical NGS testing of tumor tissue, accurately deriving dMMR status from this
656 NGS data will have important implications for diagnostics and targeted therapy and likely improve
657 patient outcomes and cancer prevention.
658
659 **Acknowledgements:** We thank members of the Colorectal Oncogenomics Group and members
660 from the Genomic Medicine and Family Cancer Clinic for their support of this manuscript. We
661 thank the participants and staff from the Australasian and Ontario Colorectal Cancer Family
662 Registries (ACCFR/OFCCR) and the ANGELS, Muir-Torre and WEHI studies. We especially
663 thank Maggie Angelakos, Samantha Fox, Allyson Templeton for supporting this study. We thank
664 the Australian Genome Research Facility for their collaboration on this project. We thank A/Prof
665 Sue Finch of the Melbourne Statistical Consulting Platform and Statistical Consulting Centre at
666 the University of Melbourne for guidance with the statistical aspects of this study.

References

1. Gryfe R, Kim H, Hsieh ET, Aronson MD, Holowaty EJ, Bull SB, Redston M, Gallinger S. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. N Engl J Med, 2000, 342:69–77
2. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, Biedrzycki B, Donehower RC, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Duffy SM, Goldberg RM, de la Chapelle A, Koshiji M, Bhaijee F, Huebner T, Hruban RH, Wood LD, Cuka N, Pardoll DM, Papadopoulos N, Kinzler KW, Zhou S, Cornish TC, Taube JM, Anders RA, Eshleman JR, Vogelstein B, Diaz LA. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med, 2015, 372:2509–20
3. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group. Recommendations from the EGAPP Working Group: genetic testing strategies in newly diagnosed individuals with colorectal cancer aimed at reducing morbidity and mortality from Lynch syndrome in relatives. Genet Med, 2009, 11:35–41
4. Green RF, Ari M, Kolor K, Dotson WD, Bowen S, Habarta N, Rodriguez JL, Richardson LC, Khoury MJ. Evaluating the role of public health in implementation of genomics-related recommendations: a case study of hereditary cancers using the CDC Science Impact Framework. Genet Med, 2019, 21:28–37
5. Baretti M, Le DT. DNA mismatch repair in cancer. Pharmacol Ther, 2018, 189:45–62
6. Eshleman JR, Markowitz SD. Mismatch repair defects in human carcinogenesis. Hum Mol Genet, 1996, 5 Spec No:1489–94
7. Young J, Simms LA, Biden KG, Wynter C, Whitehall V, Karamatic R, George J, Goldblatt J, Walpole I, Robin S-A, Borten MM, Stitz R, Searle J, McKeone D, Fraser L, Purdie DR,

Podger K, Price R, Buttenshaw R, Walsh MD, Barker M, Leggett BA, Jass JR. Features of Colorectal Cancers with High-Level Microsatellite Instability Occurring in Familial and Sporadic Settings. Am J Pathol, 2001, 159:2107–16

8. Haraldsdottir S, Hampel H, Tomsic J, Frankel WL, Pearlman R, de la Chapelle A, Pritchard CC. Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. Gastroenterology, 2014, 147:1308-1316.e1
9. Lynch HT, Lynch PM, Lanspa SJ, Snyder CL, Lynch JF, Boland CR. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin Genet, 2009, 76:1–18
10. Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TY, Bodmer D, Hoenselaar E, Hendriks-Cornelissen SJ, Tsui WY, Kong CK, Brunner HG, van Kessel AG, Yuen ST, van Krieken JH, Leung SY, Hoogerbrugge N. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nat Genet, 2009, 41:112–7
11. Bonneville R, Krook MA, Kautto EA, Miya J, Wing MR, Chen H-Z, Reeser JW, Yu L, Roychowdhury S. Landscape of Microsatellite Instability Across 39 Cancer Types. JCO Precis Oncol, 2017, 2017
12. Walsh MD, Jayasekara H, Huang A, Winship IM, Buchanan DD. Clinico-pathological predictors of mismatch repair deficiency in sebaceous neoplasia: A large case series from a single Australian private pathology service. Australas J Dermatol, 2019, 60:126–33
13. Mascarenhas L, Shanley S, Mitchell G, Spurdle AB, Macrae F, Pachter N, Buchanan DD, Ward RL, Fox S, Duxbury E, Driessen R, Boussioutas A. Current mismatch repair

712 deficiency tumor testing practices and capabilities: A survey of Australian pathology
713 providers. Asia Pac J Clin Oncol, 2018, 14:417–25
714 14. Shia J. Immunohistochemistry versus microsatellite instability testing for screening
715 colorectal cancer patients at risk for hereditary nonpolyposis colorectal cancer syndrome.
716 Part I. The utility of immunohistochemistry. J Mol Diagn, 2008, 10:293–300
717 15. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ,
718 Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute
719 Workshop on Microsatellite Instability for cancer detection and familial predisposition:
720 development of international criteria for the determination of microsatellite instability in
721 colorectal cancer. Cancer Res, 1998, 58:5248–57
722 16. Chen M-L, Chen J-Y, Hu J, Chen Q, Yu L-X, Liu B-R, Qian X-P, Yang M. Comparison of
723 microsatellite status detection methods in colorectal carcinoma. Int J Clin Exp Pathol, 2018,
724 11:1431–8
725 17. Chapusot C, Martin L, Bouvier AM, Bonithon-Kopp C, Ecarnot-Laubriet A, Rageot D,
726 Ponnelle T, Laurent Puig P, Faivre J, Piard F. Microsatellite instability and intratumoural
727 heterogeneity in 100 right-sided sporadic colon carcinomas. Br J Cancer, 2002, 87:400–4
728 18. Graham RP, Kerr SE, Butz ML, Thibodeau SN, Halling KC, Smyrk TC, Dina MA, Waugh
729 VM, Rumilla KM. Heterogenous MSH6 loss is a result of microsatellite instability within
730 MSH6 and occurs in sporadic and hereditary colorectal and endometrial carcinomas. Am J
731 Surg Pathol, 2015, 39:1370–6
732 19. Joost P, Veurink N, Holck S, Klarskov L, Bojesen A, Harbo M, Baldetorp B, Rambech E,
733 Nilbert M. Heterogenous mismatch-repair status in colorectal cancer. Diagn Pathol, 2014,
734 9:126

20. McCarthy AJ, Capo-Chichi J-M, Spence T, Grenier S, Stockley T, Kamel-Reid S, Serra S, Sabatini P, Chetty R. Heterogenous loss of mismatch repair (MMR) protein expression: a challenge for immunohistochemical interpretation and microsatellite instability (MSI) evaluation. J Pathol Clin Res, 2019, 5:115–29
21. Pai RK, Plesec TP, Abdul-Karim FW, Yang B, Marquard J, Shadrach B, Roma AR. Abrupt loss of MLH1 and PMS2 expression in endometrial carcinoma: molecular and morphologic analysis of 6 cases. Am J Surg Pathol, 2015, 39:993–9
22. Shia J, Zhang L, Shike M, Guo M, Stadler Z, Xiong X, Tang LH, Vakiani E, Katabi N, Wang H, Bacares R, Ruggeri J, Boland CR, Ladanyi M, Klimstra DS. Secondary mutation in a coding mononucleotide tract in MSH6 causes loss of immunoexpression of MSH6 in colorectal carcinomas with MLH1/PMS2 deficiency. Mod Pathol, 2013, 26:131–8
23. Watkins JC, Nucci MR, Ritterhouse LL, Howitt BE, Sholl LM. Unusual Mismatch Repair Immunohistochemical Patterns in Endometrial Carcinoma. Am J Surg Pathol, 2016, 40:909–16
24. Watson N, Grieu F, Morris M, Harvey J, Stewart C, Schofield L, Goldblatt J, Iacopetta B. Heterogeneous staining for mismatch repair proteins during population-based prescreening for hereditary nonpolyposis colorectal cancer. J Mol Diagn, 2007, 9:472–8
25. Baudrin LG, Deleuze J-F, How-Kit A. Molecular and computational methods for the detection of microsatellite instability in cancer. Frontiers in Oncology, 2018, 8
26. Vasen HFA, Hendriks Y, de Jong AE, van Puijenbroek M, Tops C, Bröcker-Vriends AHJT, Wijnen JTh, Morreau H. Identification of HNPCC by Molecular Analysis of Colorectal and Endometrial Tumors. Dis Markers, 2004, 20:207–13

27. Siemanowski J, Schömig-Markiefka B, Buhl T, Haak A, Siebolts U, Dietmaier W, Arens N, Pauly N, Ataseven B, Büttner R, Merkelbach-Bruse S. Managing Difficulties of Microsatellite Instability Testing in Endometrial Cancer-Limitations and Advantages of Four Different PCR-Based Approaches. Cancers (Basel), 2021, 13:1268
28. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics, 2014, 30:1015–6
29. Ni Huang M, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG. MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. Sci Rep, 2015, 5:13321
30. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, Roychowdhury S. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget, 2017, 8:7452–63
31. Maruvka YE, Mouw KW, Karlic R, Parasuraman P, Kamburov A, Polak P, Haradhvala NJ, Hess JM, Rheinbay E, Brody Y, Koren A, Braunstein LZ, D'Andrea A, Lawrence MS, Bass A, Bernards A, Michor F, Getz G. Analysis of somatic microsatellite indels identifies driver events in human tumors. Nat Biotechnol, 2017, 35:951–9
32. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, Kovar CL, Lewis LR, Morgan MB, Newsham IF, Reid JG, Santibanez J, Shinbrot E, Trevino LR, Wu Y-Q, Wang M, Gunaratne P, Donehower LA, Creighton CJ, Wheeler DA, Gibbs RA, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, Stojanov P, McKenna A, Lander ES, Gabriel S, Getz G, Ding L, Fulton RS, Koboldt DC, Wylie T, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature, 2012, 487:330–7

33. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res, 2019, 47:D941–7

34. Panda A, Betigeri A, Subramanian K, Ross JS, Pavlick DC, Ali S, Markowski P, Silk A, Kaufman HL, Lattime E, Mehnert JM, Sullivan R, Lovly CM, Sosman J, Johnson DB, Bhanot G, Ganesan S. Identifying a Clinically Applicable Mutational Burden Threshold as a Potential Biomarker of Response to Immune Checkpoint Therapy in Solid Tumors. JCO Precis Oncol, 2017, 2017

35. Zheng M. Tumor mutation burden for predicting immune checkpoint blockade response: the more, the better. J Immunother Cancer, 2022, 10:e003087

36. Chang H, Sasson A, Srinivasan S, Golhar R, Greenawalt DM, Geese WJ, Green G, Zerba K, Kirov S, Szustakowski J. Bioinformatic Methods and Bridging of Assay Results for Reliable Tumor Mutational Burden Assessment in Non-Small-Cell Lung Cancer. Mol Diagn Ther, 2019, 23:507–20

37. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DTW, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, et al. Signatures of mutational processes in human cancer. Nature, 2013, 500:415–21

802 38. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A,
803 Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ,
804 McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Getz G, Rozen
805 SG, Stratton MR. The repertoire of mutational signatures in human cancer. Nature, 2020,
806 578:94–101
807 39. Georgeson P, Pope BJ, Rosty C, Clendenning M, Mahmood K, Joo JE, Walker R,
808 Hutchinson RA, Preston S, Como J, Joseland S, Win AK, Macrae FA, Hopper JL, Mouradov
809 D, Gibbs P, Sieber OM, O'Sullivan DE, Brenner DR, Gallinger S, Jenkins MA, Winship IM,
810 Buchanan DD. Evaluating the utility of tumour mutational signatures for identifying
811 hereditary colorectal cancer and polyposis syndrome carriers. Gut, 2021, 70:2138–49
812 40. Georgeson P, Harrison TA, Pope BJ, Zaidi SH, Qu C, Steinfelder RS, Lin Y, Joo JE,
813 Mahmood K, Clendenning M, Walker R, Amitay EL, Berndt SI, Brenner H, Campbell PT,
814 Cao Y, Chan AT, Chang-Claude J, Doheny KF, Drew DA, Figueiredo JC, French AJ,
815 Gallinger S, Giannakis M, Giles GG, Gsur A, Gunter MJ, Hoffmeister M, Hsu L, Huang W-
816 Y, Limburg P, Manson JE, Moreno V, Nassir R, Nowak JA, et al. Identifying colorectal
817 cancer caused by biallelic MUTYH pathogenic variants using tumor mutational signatures.
818 Nat Commun, 2022, 13:3254
819 41. Jenkins MA, Win AK, Templeton AS, Angelakos MS, Buchanan DD, Cotterchio M,
820 Figueiredo JC, Thibodeau SN, Baron JA, Potter JD, Hopper JL, Casey G, Gallinger S, Le
821 Marchand L, Lindor NM, Newcomb PA, Haile RW, Colon Cancer Family Registry Cohort I.
822 Cohort Profile: The Colon Cancer Family Registry Cohort (CCFRC). Int J Epidemiol, 2018,
823 47:387–388i

824 42. Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R, Hall D, Hopper JL,
825 Jass J, Le Marchand L, Limburg P, Lindor N, Potter JD, Templeton AS, Thibodeau S,
826 Seminara D, Colon Cancer Family R. Colon Cancer Family Registry: an international
827 resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiol
828 Biomarkers Prev, 2007, 16:2331–43
829 43. Buchanan DD, Clendenning M, Rosty C, Eriksen SV, Walsh MD, Walters RJ, Thibodeau
830 SN, Stewart J, Preston S, Win AK, Flander L, Ouakrim DA, Macrae FA, Boussioutas A,
831 Winship IM, Giles GG, Hopper JL, Southey MC, English D, Jenkins MA. Tumour testing to
832 identify Lynch syndrome in two Australian colorectal cancer cohorts. J Gastroenterol
833 Hepatol, 2017, 32:427–38
834 44. Walsh MD, Buchanan DD, Pearson S-A, Clendenning M, Jenkins MA, Win AK, Walters RJ,
835 Spring KJ, Nagler B, Pavluk E, Arnold ST, Goldblatt J, George J, Suthers GK, Phillips K,
836 Hopper JL, Jass JR, Baron JA, Ahnen DJ, Thibodeau SN, Lindor N, Parry S, Walker NI,
837 Rosty C, Young JP. Immunohistochemical testing of conventional adenomas for loss of
838 expression of mismatch repair proteins in Lynch syndrome mutation carriers: a case series
839 from the Australasian site of the colon cancer family registry. Mod Pathol, 2012, 25:722–30
840 45. Cicek MS, Lindor NM, Gallinger S, Bapat B, Hopper JL, Jenkins MA, Young J, Buchanan
841 D, Walsh MD, Le Marchand L, Burnett T, Newcomb PA, Grady WM, Haile RW, Casey G,
842 Plummer SJ, Krumroy LA, Baron JA, Thibodeau SN. Quality assessment and correlation of
843 microsatellite instability and immunohistochemical markers among population- and clinic-
844 based colorectal tumors results from the Colon Cancer Family Registry. J Mol Diagn, 2011,
845 13:271–81

46. Buchanan DD, Tan YY, Walsh MD, Clendenning M, Metcalf AM, Ferguson K, Arnold ST, Thompson BA, Lose FA, Parsons MT, Walters RJ, Pearson SA, Cummings M, Oehler MK, Blomfield PB, Quinn MA, Kirk JA, Stewart CJ, Obermair A, Young JP, Webb PM, Spurdle AB. Tumor mismatch repair immunohistochemistry and DNA MLH1 methylation testing of patients with endometrial cancer diagnosed at age younger than 60 years optimizes triage for population-level germline mismatch repair gene mutation testing. J Clin Oncol, 2014, 32:90–100

47. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, Kang GH, Widschwendter M, Weener D, Buchanan D, Koh H, Simms L, Barker M, Leggett B, Levine J, Kim M, French AJ, Thibodeau SN, Jass J, Haile R, Laird PW. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet, 2006, 38:787–93

48. Zaidi SH, Harrison TA, Phipps AI, Steinfelder R, Trinh QM, Qu C, Banbury BL, Georgeson P, Grasso CS, Giannakis M, Adams JB, Alwers E, Amitay EL, Barfield RT, Berndt SI, Borozan I, Brenner H, Brezina S, Buchanan DD, Cao Y, Chan AT, Chang-Claude J, Connolly CM, Drew DA, Farris AB, Figueiredo JC, French AJ, Fuchs CS, Garraway LA, Gruber S, Guinter MA, Hamilton SR, Harlid S, Heisler LE, Hidaka A, et al. Landscape of somatic single nucleotide variants and indels in colorectal cancer and impact on survival. Nat Commun, 2020, 11:3644

49. Belhadj S, Terradas M, Munoz-Torres PM, Aiza G, Navarro M, Capellá G, Valle L. Candidate genes for hereditary colorectal cancer: Mutational screening and systematic review. Hum Mutat, 2020, 41:1563–76

50. Seifert BA, McGlaughon JL, Jackson SA, Ritter DI, Roberts ME, Schmidt RJ, Thompson BA, Jimenez S, Trapp M, Lee K, Plon SE, Offit K, Stadler ZK, Zhang L, Greenblatt MS, Ferber MJ. Determining the clinical validity of hereditary colorectal cancer and polyposis susceptibility genes using the Clinical Genome Resource Clinical Validity Framework. Genet Med, 2019, 21:1507–16

51. Weren RDA, Ligtenberg MJL, Kets CM, de Voer RM, Verwiel ETP, Spruijt L, van Zelst-Stams WAG, Jongmans MC, Gilissen C, Hehir-Kwa JY, Hoischen A, Shendure J, Boyle EA, Kamping EJ, Nagtegaal ID, Tops BBJ, Nagengast FM, Geurts van Kessel A, van Krieken JHJM, Kuiper RP, Hoogerbrugge N. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. Nat Genet, 2015, 47:668–71

52. Spurdle AB, Bowman MA, Shamsani J, Kirk J. Endometrial cancer gene panels: clinical diagnostic vs research germline DNA testing. Mod Pathol, 2017, 30:1048–68

53. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, Benz CC, Yau C, Laird PW, Ding L, Zhang W, Mills GB, Kucherlapati R, Mardis ER, Levine DA. Integrated genomic characterization of endometrial carcinoma. Nature, 2013, 497:67–73

54. Cherniack AD, Shen H, Walter V, Stewart C, Murray BA, Bowlby R, Hu X, Ling S, Soslow RA, Broaddus RR, Zuna RE, Robertson G, Laird PW, Kucherlapati R, Mills GB, Akbani R, Ally A, Auman JT, Balasundaram M, Balu S, Baylin SB, Beroukhim R, Bodenheimer T, Bogomolniy F, Boice L, Bootwalla MS, Bowen J, Bowlby R, Broaddus R, Brooks D, Carlsen R, Cherniack AD, Cho J, Chuah E, Chudamani S, et al. Integrated Molecular Characterization of Uterine Carcinosarcoma. Cancer Cell, 2017, 31:411–23

55. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, 30:2114–20

56. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics, 2012, 28:1811–7

57. Sha D, Jin Z, Budzcies J, Kluck K, Stenzinger A, Sinicrope FA. Tumor Mutational Burden (TMB) as a Predictive Biomarker in Solid Tumors. Cancer Discov, 2020, 10:1808–25

58. Kuhn M, cre, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C, Hunt T. caret: Classification and Regression Training. 2022

59. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological), 1996, 58:267–88

60. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the Tidyverse. Journal of Open Source Software, 2019, 4:1686

61. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 2011, 12:77

62. LeDell E, Petersen M, Laan M van der. cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals. 2022

63. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 2010, 33:1–22

914 64. Thiele C, Hirschfeld G. cutpointr: Improved Estimation and Validation of Optimal Cutpoints
915 in R. Journal of Statistical Software, 2021, 98:1–27
916 65. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical
917 Software, 2011, 40:1–29
918 66. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016
919 67. Rudis B, cre, Kennedy P, Reiner P, support) DW (Secondary axis, Adam X, Fonts) G
920 (Roboto C& TW, Font) I (Plex S, Font) IT (Public S, Barnett J, Leeper TJ, Meys J.
921 hrbrthemes: Additional Themes, Theme Components and Utilities for "ggplot2." 2020
922 68. Slowikowski K, Schep A, Hughes S, Dang TK, Lukauskas S, Irisson J-O, Kamvar ZN, Ryan
923 T, Christophe D, Hiroaki Y, Gramme P, Abdol AM, Barrett M, Cannoodt R, Krassowski M,
924 Chirico M, Aphalo P. ggrepel: Automatically Position Non-Overlapping Text Labels with
925 "ggplot2." 2021
926 69. Clopper CJ, Pearson ES. THE USE OF CONFIDENCE OR FIDUCIAL LIMITS
927 ILLUSTRATED IN THE CASE OF THE BINOMIAL. Biometrika, 1934, 26:404–13
928 70. Dorai-Raj S. binom: Binomial Confidence Intervals for Several Parameterizations. 2022
929 71. Renault V, Tubacher E, How-Kit A. Assessment of Microsatellite Instability from Next-
930 Generation Sequencing Data. Edited by Laganà A. Computational Methods for Precision
931 Oncology, Cham, Springer International Publishing, 2022, pp. 75–100
932 72. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite Instability
933 Detection by Next Generation Sequencing. Clinical Chemistry, 2014, 60:1192–9
934 73. Ratovomanana T, Cohen R, Svrcek M, Renaud F, Cervera P, Siret A, Letourneur Q, Buhard
935 O, Bourgoin P, Guillerm E, Dorard C, Nicolle R, Ayadi M, Touat M, Bielle F, Sanson M,
936 Rouzic PL, Buisine M-P, Piessen G, Collura A, Fléjou J-F, Reyniès A de, Coulet F,

Ghiringhelli F, André T, Jonchère V, Duval A. Performance of Next-Generation Sequencing for the Detection of Microsatellite Instability in Colorectal Cancer With Deficient DNA Mismatch Repair. Gastroenterology, 2021, 161:814-826.e7

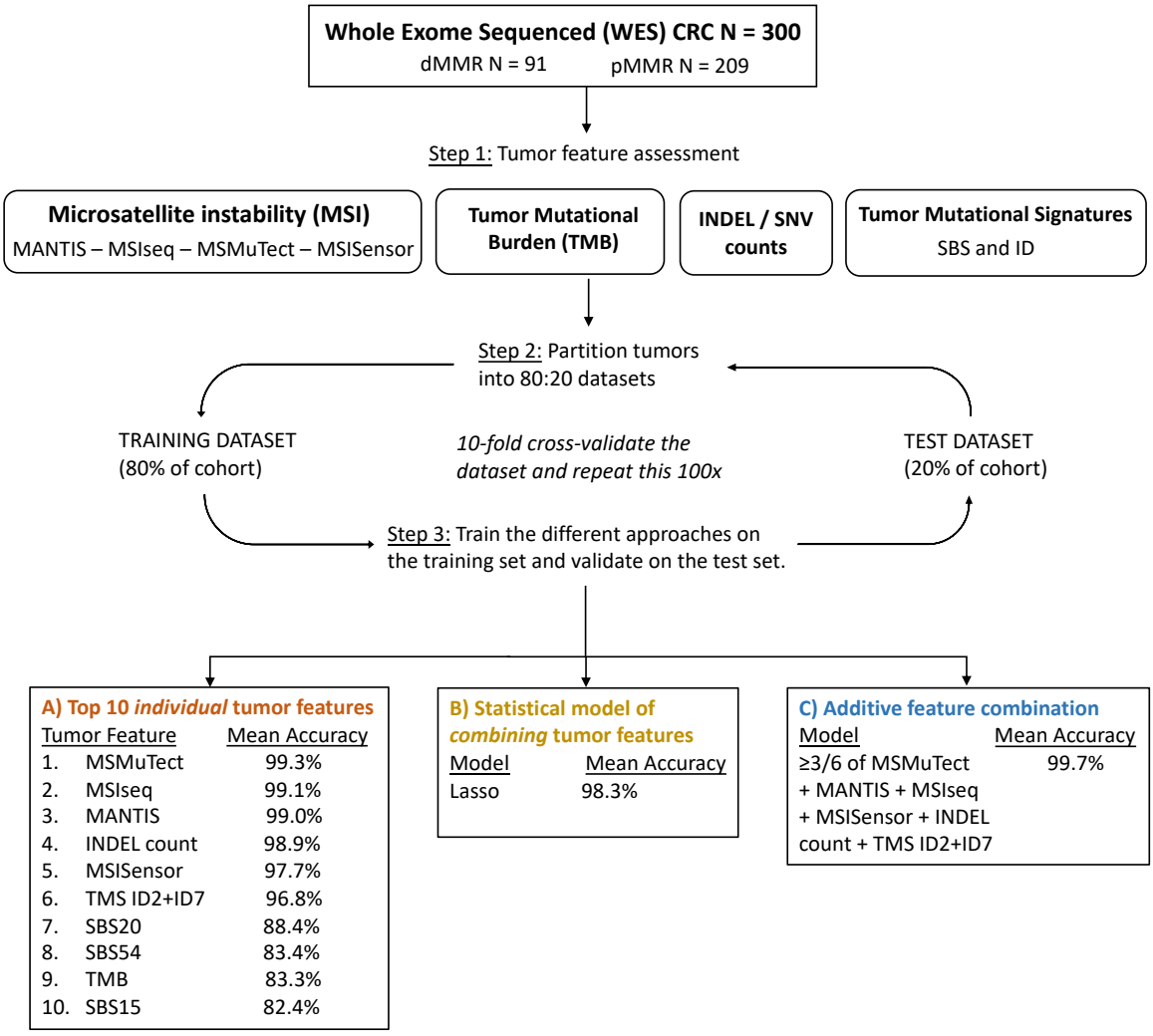
74. Roufas C, Georgakopoulos-Soares I, Zaravinos A. Molecular correlates of immune cytolytic subgroups in colorectal cancer by integrated genomics analysis. NAR Cancer, 2021, 3:zcab005

75. Zaravinos A, Roufas C, Nagara M, de Lucas Moreno B, Oblovatskaya M, Efstathiades C, Dimopoulos C, Ayiomamitis GD. Cytolytic activity correlates with the mutational burden and deregulated expression of immune checkpoints in colorectal cancer. J Exp Clin Cancer Res, 2019, 38:364

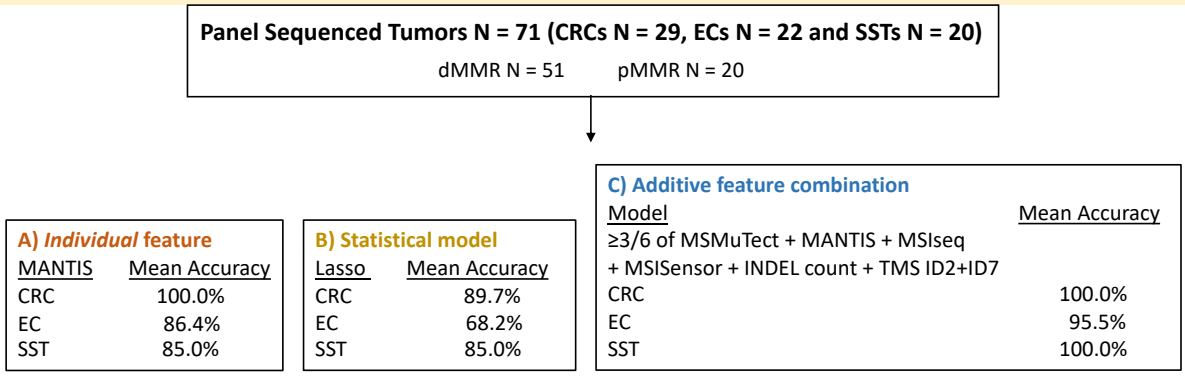
76. Chen W, Pearlman R, Hampel H, Pritchard CC, Markow M, Arnold C, Knight D, Frankel WL. MSH6 immunohistochemical heterogeneity in colorectal cancer: comparative sequencing from different tumor areas. Hum Pathol, 2020, 96:104–11

951 **Figures**

Analysis 1. Assessment of tumor features for dMMR prediction accuracy in WES CRCs



Analysis 2. Assessment of A) individual tumor features, B) statistical model and C) additive feature combination approaches derived from the WES analysis on panel sequenced CRCs, ECs and SSTs



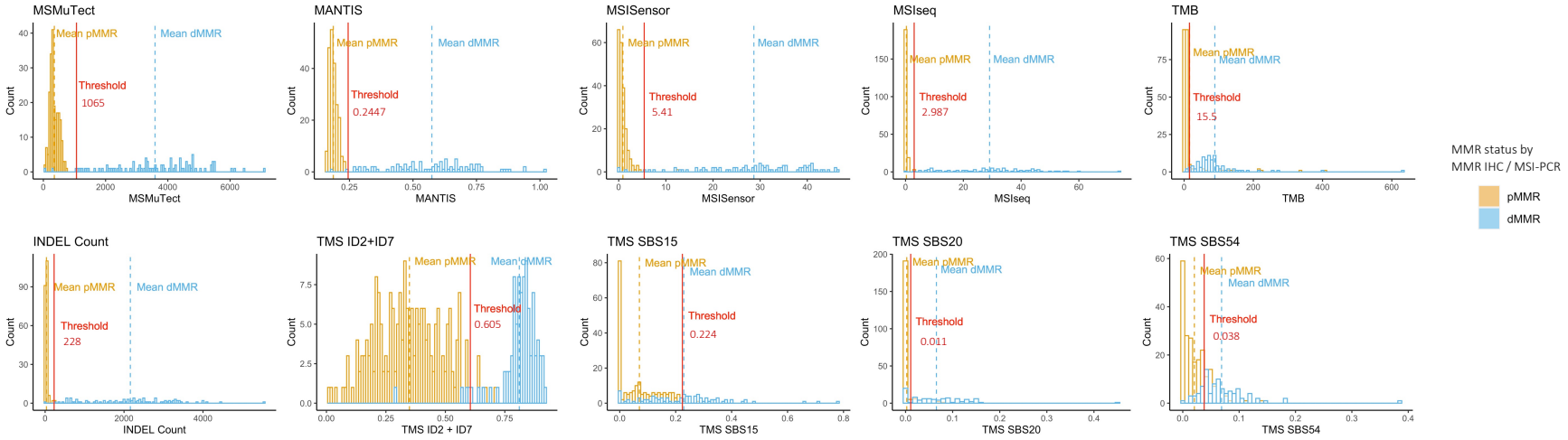
952

Figure 1. Overview of the study design. In total, 300 whole-exome sequenced (WES) colorectal cancers (CRCs) consisting of 91 DNA mismatch repair deficient (dMMR) and 209 DNA mismatch repair proficient (pMMR) tumors were analyzed. We investigated 104 tumor features for their ability to distinguish dMMR from pMMR tumors consisting of four MSI tools, 97 tumor mutational signature definitions (TMS), tumor mutation burden (TMB) calculated as mutations per mega base, somatic insertion / deletion (INDEL) and somatic single nucleotide variant (SNV) counts. We performed a 10-fold cross-validation approach with 100 repeats to calculate the mean accuracy on the test dataset. (A) The top 10 ranked individual tumor features, (B) a Lasso regression model and (C) an additive feature combination approach was tested to determine the benefit of combining tumor features to improve dMMR prediction. The findings from these three approaches were tested on an independent set of targeted panel sequenced tumors of CRC, endometrial cancer (EC) and sebaceous skin tumor (SST) tissue types with reported mean accuracies.


MSMuTect
MANTIS
MSISensor
MSIseq
INDEL Count
TMB
TMS ID2+ID7
TMS SBS15
TMS SBS20
TMS SBS54
pMMR
dMMR-combined
dMMR-LS
dMMR-MLH1me
dMMR-DS
MMR Subtype
Subtype

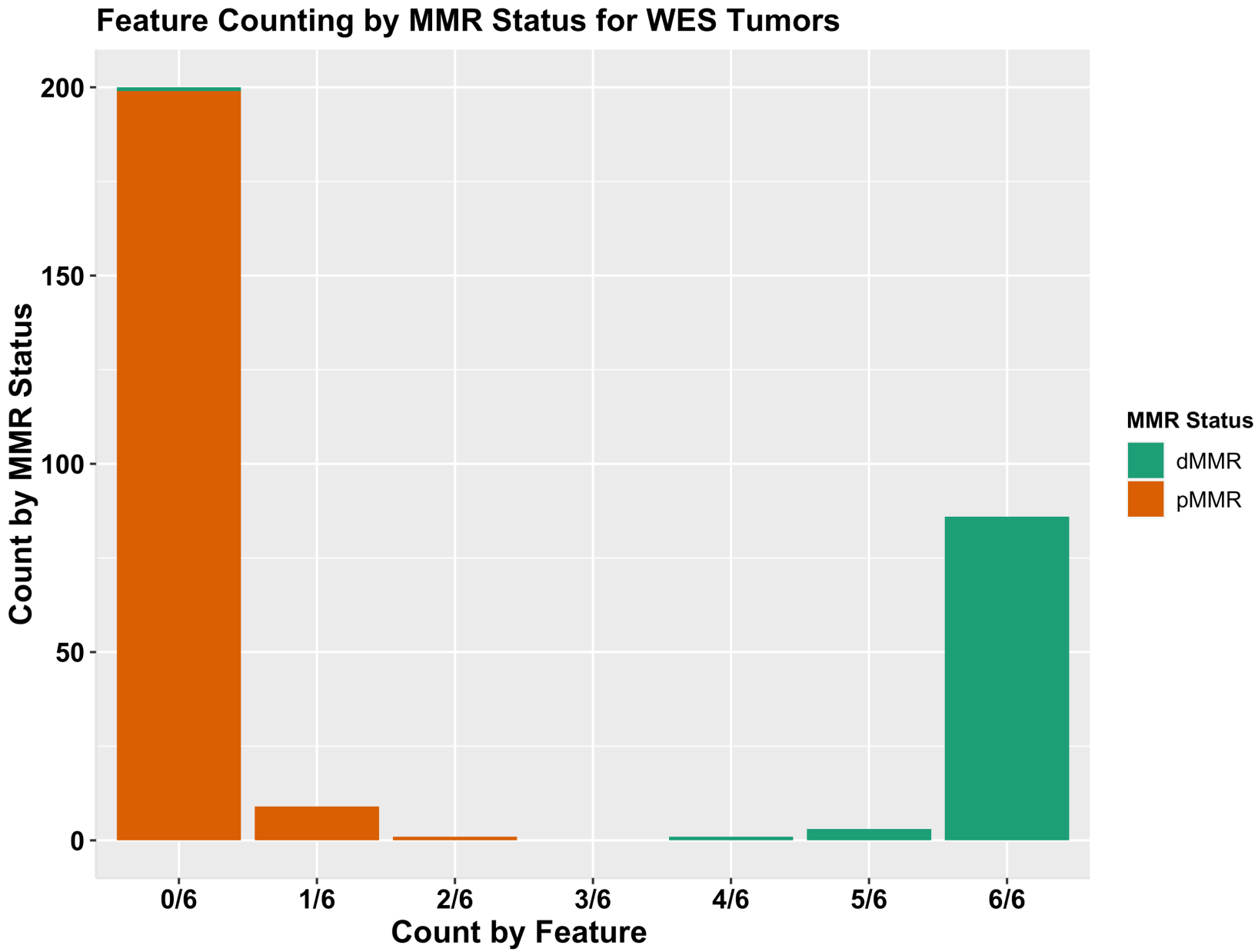
966

Figure 2. Tumor distribution of the top 10 DNA mismatch repair (MMR) deficient (dMMR) predicting features in the whole exome sequenced (WES) colorectal cancers (CRCs) by MMR subtype. Boxplots showing the distribution of tumors by MMR status (MMR-proficient (pMMR) versus dMMR) as well as stratified by dMMR subtype - dMMR-LS (Lynch syndrome), dMMR-DS (double somatic MMR gene mutations) and dMMR-MLH1me (*MLH1* promoter methylation) for each of the top 10 predicting features MSMuTect, MANTIS, MSISensor, MSIseq, INDEL (insertion / deletion) count, TMB (tumor mutation burden calculated as mutations / mega base), TMS (tumor mutational signature) ID2+ID7, TMS SBS15, TMS SBS20 and TMS SBS54 as determined from the WES CRC analysis. ID, small insertions / deletions; SBS, single base substitution.



974

975 **Figure 3**. Determination of thresholds for differentiating DNA mismatch repair (MMR) deficient (dMMR) from MMR-proficient
976 (pMMR) colorectal cancers (CRCs) using whole exome sequencing (WES) data for each of the top 10 performing tumor features. Bar
977 graphs presenting the distribution of tumors after applying the recommended thresholds (red line) for each of the top 10 predicting tumor
978 features MSMuTect, MANTIS, MSISensor, MSIseq, INDEL count, TMB, TMS ID2+ID7, TMS SBS15, TMS SBS20 and TMS SBS54
979 as determined from the WES CRC analysis. Orange coloring indicates pMMR and blue coloring represents dMMR status. ID, small
980 insertions / deletions; SBS, single base substitution.


Feature Counting by MMR Status for WES Tumors
Count by MMR Status
200
150
100
50
0
0/6
1/6
2/6
3/6
4/6
5/6
6/6
Count by Feature
MMR Status
dMMR
pMMR

981

Figure 4. The additive tumor feature combination approach demonstrating the distribution of counts of the top six tumor features by the DNA mismatch repair (MMR) status of the 300 colorectal cancers (CRCs) with whole exome sequencing (WES). Bar graphs presenting the distribution of tumors after applying the additive tumor feature combination approach with the recommended thresholds from the WES CRC analysis using a count of ≥3 out of the top six predictors from the WES CRC analysis, consisting of MSMuTect, MANTIS, MSIseq, MSISensor, INDEL (insertion / deletion) count and TMS (tumor mutational signature) ID2+ID7 (small insertions / deletions) for MMR status calling: MMR-deficient (dMMR) versus MMR-proficient (pMMR).

Tables

Table 1. The breakdown of the 104 tumor features calculated from next generation sequencing analysis included in this study.

Feature Type	*Count*	*Name*	*Reference*
Total	***N = 104***		
Microsatellite instability (MSI) Tools	**N = 4**	MSISensor	Niu *et al.*, 2014
		MSIseq	Huang *et al.*, 2015
		MANTIS	Kautto *et al.*, 2017
		MSMuTect	Maruvka *et al.*, 2017
Tumor mutational signatures (TMS)	**N = 97**	SBS (N = 78)	Tate *et al.*, 2018
		ID (N = 18)	Tate *et al.*, 2018
		ID2+ID7	Georgeson *et al.*, 2021
Somatic mutation counts	**N = 3**	INDELs	
		SNVs	
		TMB (SNVs + INDELs/ MB)	Muzny *et al.*, 2012

The 104 tumor features can be categorized into three distinct groups: microsatellite instability (MSI) tools, tumor mutational signatures (TMS) and somatic mutation counts. These features have previously been shown to be associated with MSI / DNA mismatch repair status as indicated by the provided references. The MSI group consists of four MSI tools namely MSISensor, MSIseq, MANTIS and MSMuTect. TMS consisted of 78 single base substitutions (SBS), 18 small insertions / deletions (IDs) and TMS ID2+ID7. The somatic

mutation count consisted of the single nucleotide variant count, larger insertions / deletions count and the tumor mutation burden (TMB), which was calculated as the combination of SNVs and INDELs counts per megabase.

997 **Table 2.** Performance of the top tumor features demonstrating a prediction accuracy >80% ranked by highest mean accuracy from
998 whole-exome sequenced (WES) colorectal cancers (CRCs).

Tumor Feature	**Mean Accuracy**	**Error Rate**	**95% CI: (Accuracy)**	**Mean Sensitivity**	**95% CI: (Sensitivity)**	**Mean Specificity**	**95% CI: (Specifity)**	**Mean AUC**	**95% CI: (AUC)**
MSMuTect	99.3%	0.7%	99.1% - 99.5%	97.6%	96.9% - 98.3%	100.0%	-	98.8%	98.5% - 99.1%
MSIseq	99.1%	0.9%	98.9% - 99.4%	97.7%	97.0% - 98.3%	99.8%	99.6% - 100.0%	98.7%	98.4% - 99.1%
MANTIS	99.0%	1.0%	98.8% - 99.2%	97.1%	96.4% - 97.7%	99.9%	99.8% - 100.0%	98.5%	98.1% - 98.8%
INDEL count	98.9%	1.1%	98.7% - 99.2%	97.7%	97.0% - 98.3%	99.5%	99.2% - 99.8%	98.6%	98.2% - 98.9%
MSISensor	97.7%	2.3%	97.3% - 98.0%	93.4%	92.4% - 94.5%	99.5%	99.3% - 99.7%	96.5%	96.0% - 97.0%
TMS ID2+ID7	96.8%	3.2%	96.4% - 97.2%	94.2%	93.2% - 95.2%	97.9%	97.5% - 98.4%	96.0%	95.5% - 96.6%

TMS ID2	93.3%	6.7%	92.8% - 93.8%	90.7%	89.5% - 91.9%	94.4%	93.7% - 95.1%	92.6%	92.0% - 93.1%
TMS SBS20	88.4%	11.6 %	87.6% - 89.2%	68.9%	66.6% - 71.2%	97.0%	96.4% - 97.6%	82.9%	81.8% - 84.1%
TMS ID7	87.6%	12.4 %	87.0% - 88.3%	74.2%	72.6% - 75.9%	93.5%	92.8% - 94.2%	83.9%	83.0% - 84.7%
TMS SBS54	83.4%	16.6 %	82.6% - 84.2%	59.4%	57.5% - 61.4%	93.9%	93.1% - 94.7%	76.7%	75.6% - 77.7%
TMB	83.3%	16.7 %	82.6% - 83.9%	57.8%	55.2% - 60.4%	94.5%	93.7% - 95.2%	76.1%	75.0% - 77.3%
TMS SBS15	82.4%	17.6 %	81.5% - 83.3%	58.8%	56.5% - 61.1%	92.8%	91.9% - 93.7%	75.8%	74.6% - 77.0%

The mean accuracy values after 10-fold cross-validation with 100 repeats, error rate, mean sensitivity, mean specificity, and mean area under the curves (AUCs) with corresponding 95% confidence intervals (CIs) are shown for each of the top 10 predicting tumor features MSMuTect, MSIseq, MANTIS, INDEL (insertion / deletion) count, MSISensor, TMS (tumor mutational signature) ID2+ID7, TMS ID2, TMS SBS20, TMS ID7, TMS SBS54, TMB (tumor mutation burden) and TMS SBS15 from the WES CRC analysis. ID, small insertions, and deletions; SBS, single base substitutions.

Table 3. Summary of the best dMMR prediction results by individual tumor feature, Lasso regression model and the additive feature combination approach for the whole-exome sequencing (WES) colorectal cancers (CRCs) and the panel sequenced CRCs, endometrial cancers (ECs) and sebaceous skin tumors (SST).

	Performance of best Individual Feature		Performance of Statistical Model		Performance of Additive Feature Combination Approach	
WES	**Feature**	**Mean Accuracy**	**Lasso**	**Mean Accuracy**	**Feature Combination**	**Mean Accuracy**
CRC	MSMuTect	99.3%	MANTIS + TMS ID2+ID7 + MSISensor + TMS SBS15	98.3%	MSMuTect + MANTIS + MSIseq + MSISensor + INDEL count + TMS ID2+ID7	99.7%
PANEL	**Feature**	**Accuracy**	**Lasso**	**Accuracy**	**Feature Combination**	**Accuracy**
CRC	MANTIS	100.0%	MANTIS + TMS ID2+ID7 + MSISensor + TMS SBS15	89.7%	MSMuTect + MANTIS + MSIseq + MSISensor + INDEL count + TMS ID2+ID7	100.0%
EC	MANTIS	86.4%	MANTIS + TMS ID2+ID7 + MSISensor + TMS SBS15	68.2%	MSMuTect + MANTIS + MSIseq + MSISensor + INDEL count + TMS ID2+ID7	95.5%
SST	MANTIS	85.0%	MANTIS + TMS ID2+ID7 + MSISensor + TMS SBS15	85.0%	MSMuTect + MANTIS + MSIseq + MSISensor + INDEL count + TMS ID2+ID7	100.0%

This table provides the top performing results from A) individual tumor feature, B) statistical model application (Lasso) and C) additive feature combination approach assessments for WES CRCs as well as targeted panel sequenced CRCs, ECs and SSTs.

TMS, tumor mutational signature; ID, small insertions, and deletions; SBS, single base substitution; INDEL count, insertions / deletions.

Table 4. Assessment of top performing tumor features from whole-exome sequenced (WES) colorectal cancers (CRCs) in panel sequenced CRC, endometrial cancer (EC) and sebaceous skin tumor (SST) test sets.

	CRC			EC			SST		
Tumor Feature	Mean Accuracy	95% CI	Error Rate	Mean Accuracy	95% CI	Error Rate	Mean Accuracy	95% CI	Error Rate
MSMuTect	27.6%	12.7% - 47.2%	72.4%	18.2%	5.2% - 40.3%	81.8%	35.0%	15.4% - 59.2%	65.0%
MSIseq	82.8%	64.2% - 94.2%	17.2%	68.2%	45.1%. - 86.1%	31.8%	65.0%	40.8% - 84.6%	35.0%
MANTIS	100.0%	88.1% - 100.0%	0.0%	86.4%	65.1% - 97.1%	13.6%	85.0%	62.1% - 96.8%	15.0%
INDEL count	27.6%	12.7% - 47.2%	72.4%	18.2%	5.2% - 40.3%	81.8%	35.0%	15.4% - 59.2%	65.0%
MSISensor	96.6%	82.2% - 99.9%	3.4%	77.3%	54.6% - 92.2%	22.7%	75.0%	50.9% - 91.3%	25.0%
TMS ID2+ID7	82.8%	64.2% - 94.2%	17.2%	63.6%	40.7% - 82.8%	36.4%	85.0%	62.1% - 96.8%	15.0%

TMS SBS20	69.0%	49.2% - 84.7%	31.0%	50.0%	28.2% - 71.8%	50.0%	40.0%	19.1% - 63.9%	60.0%
TMS SBS54	51.7%	32.5% - 70.6%	48.3%	36.4%	17.2% - 59.3%	63.6%	40.0%	19.1% - 63.9%	60.0%
TMB	44.8%	26.4% - 64.3%	55.2%	31.8%	13.9% - 54.9%	68.2%	35.0%	15.4% - 59.2%	65.0%
TMS SBS15	44.8%	26.4% - 64.3%	55.2%	27.3%	10.7% - 50.2%	72.7%	60.0%	36.1% - 80.9%	40.0%

1014 Table presents the prediction accuracies, error rates and corresponding 95% confidence intervals (CIs) for panel sequenced CRCs, ECs
1015 and SSTs for the top 10 predicting tumor features MSMuTect, MSIseq, MANTIS, INDEL (insertions / deletions count), MSISensor,
1016 TMS (tumor mutational signature) ID2+ID7, TMS SBS20, TMS SBS54, TMB (tumor mutation burden, mutations / mega base) and
1017 TMS SBS15 from WES CRC analysis applied on panel sequenced CRCs, ECs and SSTs. ID, small insertions, and deletions; SBS, single
1018 base substitution.