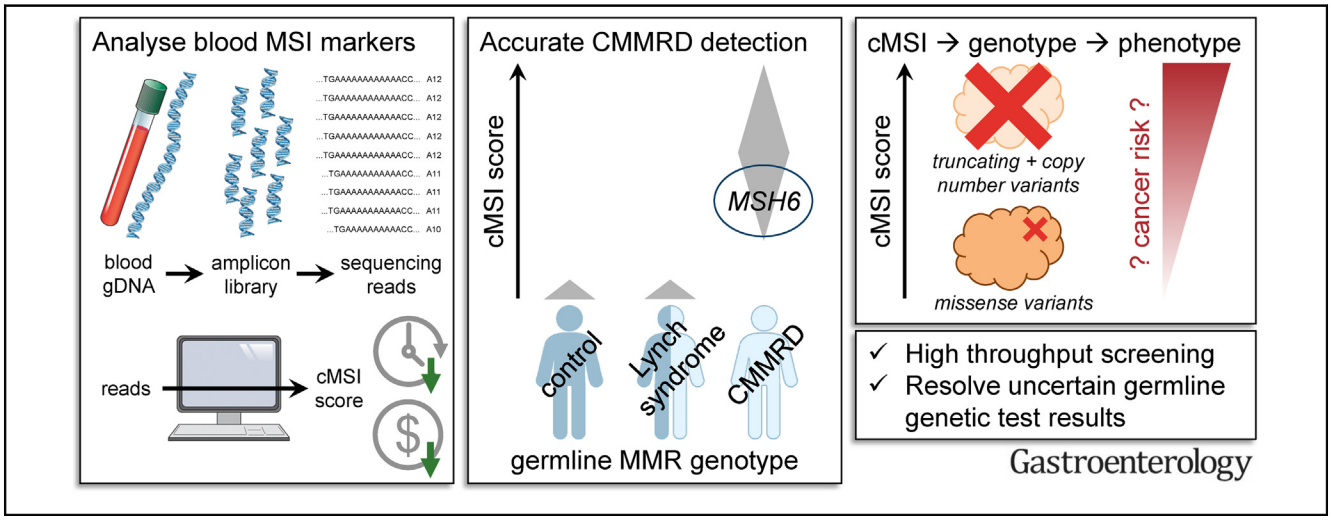
Constitutional Microsatellite Instability, Genotype, and Phenotype Correlations in Constitutional Mismatch Repair Deficiency

Gastroenterology 2023;164:579–592

Constitutional Microsatellite Instability, Genotype, and Phenotype Correlations in Constitutional Mismatch Repair Deficiency

Richard Gallon,[1] Rachel Phelps,[1] Christine Hayes,[1] Laurence Brugieres,[2] Léa Guerrini-Rousseau,[2,3] Chrystelle Colas,[4,5] Martine Muleris,[6] Neil A. J. Ryan,[7,8] D. Gareth Evans,[9] Hannah Grice,[10] Emily Jessop,[10] Annabel Kunzemann-Martinez,[10,11] Lilla Marshall,[10] Esther Schamschula,[12] Klaus Oberhuber,[12] Amedeo A. Azizi,[13] Hagit Baris Feldman,[14] Andreas Beilken,[15] Nina Brauer,[16] Triantafyllia Brozou,[17] Karin Dahan,[18] Ugur Demirsoy,[19] Benoît Florkin,[20] William Foulkes,[21,22,23,24] Danuta Januszkiewicz-Lewandowska,[25] Kristi J. Jones,[26,27] Christian P. Kratz,[15] Stephan Lobitz,[28] Julia Meade,[29] Michaela Nathrath,[30,31] Hans-Jürgen Pander,[32] Claudia Perne,[33] Iman Ragab,[34] Tim Ripperger,[35] Thorsten Rosenbaum,[36] Daniel Rueda,[37] Tomasz Sarosiek,[38] Astrid Sehested,[39] Isabel Spier,[33] Manon Suerink,[40] Stefanie-Yvonne Zimmermann,[41] Johannes Zschocke,[12] Gillian M. Borthwick,[1] Katharina Wimmer,[12] John Burn,[1] Michael S. Jackson,[10] and Mauro Santibanez-Koref[10]

[1]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [2]Department of Children and Adolescents Oncology, Gustave Roussy, Université Paris-Saclay, Villejuif, France; [3]Team "Genomics and Oncogenesis of Pediatric Brain Tumors," INSERM U981, Gustave Roussy, Université Paris-Saclay, Villejuif, France; [4]Département de Génétique, Institut Curie, Paris, France; [5]INSERM U830, Université de Paris, Paris, France; [6]Sorbonne Université, Institut National de la Santé et de la Recherche Médicale, Centre de Recherche Saint-Antoine, Paris, France; [7]The Academic Women's Health Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [8]Department of Gynaecology Oncology, Royal Infirmary of Edinburgh, Edinburgh, United Kingdom; [9]Division of Evolution, Infection and Genomics, University of Manchester, Manchester, United Kingdom; [10]Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [11]Centre for Inflammation and Tissue Repair, University College London, London, United Kingdom; [12]Institute of Human Genetics, Medical University of Innsbruck, Innsbruck, Austria; [13]Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Vienna, Austria; [14]The Genetics Institute and Genomics Center, Tel Aviv Sourasky Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel; [15]Department of Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany; [16]Pediatric Oncology, Helios-Klinikum, Krefeld, Germany; [17]Department of Pediatric Oncology, Hematology and Clinical Immunology, University Children's Hospital, Medical Faculty, Heinrich Heine University, Duesseldorf, Germany; [18]Centre de Génétique Humaine, Institut de Pathologie et Génétique, Gosselies, Belgium; [19]Department of Pediatric Oncology, Kocaeli University, Kocaeli, Turkey; [20]Department of Pediatrics, Citadelle Hospital, University of Liège, Liège, Belgium; [21]Program in Cancer Genetics, Departments of Oncology and Human Genetics, McGill University, Montreal, Quebec, Canada; [22]Department of Human Genetics, McGill University, Montreal, Quebec, Canada; [23]Department of Medical Genetics, McGill University Health Centre, Montreal, Quebec, Canada; [24]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada; [25]Department of Pediatric Oncology, Hematology and Transplantation Medical University, Poznan, Poland; [26]Department of Clinical Genetics, Western Sydney Genetics Program, Children's Hospital at Westmead, Sydney, New South Wales, Australia; [27]University of Sydney School of Medicine, Sydney, New South Wales, Australia; [28]Gemeinschaftsklinikum Mittelrhein, Department of Pediatric Hematology and Oncology, Koblenz, Germany; [29]Division of Pediatric Hematology/Oncology, Department of Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania; [30]Pediatric Hematology and Oncology, Klinikum Kassel, Kassel, Germany; [31]Department of Pediatrics, Pediatric Oncology Center, Technische Universität München, Munich, Germany; [32]Institut für Klinische Genetik, Olgahospital, Stuttgart, Germany; [33]Institute of Human Genetics, Medical Faculty, University of Bonn and National Center for Hereditary Tumor Syndromes, University Hospital Bonn, Bonn, Germany; [34]Pediatrics Department, Hematology-Oncology Unit, Faculty of Medicine, Ain Shams University, Cairo, Egypt; [35]Department of Human Genetics, Hannover Medical School, Hannover, Germany; [36]Department of Pediatrics, Sana Kliniken Duisburg, Duisburg, Germany; [37]Hereditary Cancer Laboratory, University Hospital Doce de Octubre, i+12 Research Institute, Madrid, Spain; [38]Department of Oncology, Luxmed Onkologia, Warsaw, Poland; [39]Department of Pediatrics and Adolescent Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark; [40]Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands; and [41]Department of Pediatric Hematology and Oncology, Children's Hospital, University Hospital, Frankfurt, Germany



BACKGROUND & AIMS: Constitutional mismatch repair deficiency (CMMRD) is a rare recessive childhood cancer predisposition syndrome caused by germline mismatch repair variants. Constitutional microsatellite instability (cMSI) is a CMMRD diagnostic hallmark and may associate with cancer risk. We quantified cMSI in a large CMMRD patient cohort to explore genotype–phenotype correlations using novel MSI markers selected for instability in blood. **METHODS:** Three CMMRD, 1 Lynch syndrome, and 2 control blood samples were genome sequenced to >120× depth. A pilot cohort of 8 CMMRD and 38 control blood samples and a blinded cohort of 56 CMMRD, 8 suspected CMMRD, 40 Lynch syndrome, and 43 control blood samples were amplicon sequenced to 5000× depth. Sample cMSI score was calculated using a published method comparing microsatellite reference allele frequencies with 80 controls. **RESULTS:** Thirty-two mononucleotide repeats were selected from blood genome and pilot amplicon sequencing data. cMSI scoring using these MSI markers achieved 100% sensitivity (95% CI, 93.6%–100.0%) and specificity (95% CI 97.9%–100.0%), was reproducible, and was superior to an established tumor MSI marker panel. Lower cMSI scores were found in patients with CMMRD with MSH6 deficiency and patients with at least 1 mismatch repair missense variant, and patients with biallelic truncating/copy number variants had higher scores. cMSI score did not correlate with age at first tumor. **CONCLUSIONS:** We present an inexpensive and scalable cMSI assay that enhances CMMRD detection relative to existing methods. cMSI score is associated with mismatch repair genotype but not phenotype, suggesting it is not a useful predictor of cancer risk.

Keywords: Pediatric Cancer; Functional Test; Replication Error Repair; Constitutional Mutation Burden.

The DNA mismatch repair (MMR) system is conserved across all 3 domains of life. It mediates the repair of base-to-base mismatches and small insertion-deletion loops generated during DNA replication, while signaling to the wider DNA damage response. The MMR system also detects base mispairings caused by base modifications, such as cytosine deamination and guanine methylation.[1,2] MMR function can be lost in a variety of neoplasias, affecting approximately 1 in 4 endometrial cancers and 1 in 7 colorectal cancers (CRCs).[3,4] MMR-deficient tumors are often hypermutated and display high levels of microsatellite instability (MSI), a molecular phenotype defined as the accumulation of insertion and deletion (indel) variants in short tandem repeat sequences.[5] This elevated mutation rate has been proposed to drive tumorigenesis through secondary mutation of onco- and tumor suppressor genes.[6–12]

Individuals with Lynch syndrome (LS) carry a germline pathogenic variant (PV) affecting 1 of the 4 principal MMR genes (*MLH1*, *MSH2*, *MSH6*, or *PMS2*) and have an increased lifetime risk of adult-onset cancer, in particular CRC, endometrial cancer, and other tumors of the gastrointestinal and genitourinary tracts.[13] LS is one of the most common hereditary cancer predisposition syndromes, affecting approximately 1 in 300 individuals in the general population.[14] Constitutional mismatch repair deficiency (CMMRD) is a far rarer childhood cancer predisposition syndrome caused by germline variants affecting both alleles of *MLH1*, *MSH2*, *MSH6*, or *PMS2*, with an estimated birth incidence of 1 per million.[15] The constitutional loss of MMR function in all tissues is associated with an exceptionally high cancer

Abbreviations used in this paper: **AUC, area under curve; bp, base pair; CMMRD, constitutional mismatch repair deficiency; cMSI, constitutional microsatellite instability; CNV, copy number variant; CRC, colorectal cancer; gDNA, genomic DNA; indel, insertion and deletion; LS, Lynch syndrome; MMR, mismatch repair; MNR, mononucleotide repeat; MSI, microsatellite instability; PBL, peripheral blood leukocyte; PCR, polymerase chain reaction; PV, pathogenic variant; RAF, reference allele frequency; ROC, receiver operator characteristic; smMIP, single-molecule molecular inversion probe; smSequence, single-molecule sequence; WGS, whole genome sequencing.**

Most current article

0016-5085
https://doi.org/10.1053/j.gastro.2022.12.017

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Constitutional mismatch repair deficiency is a childhood cancer predisposition syndrome characterized by microsatellite instability in normal tissues, the level of which may associate with genotype and aid cancer risk stratification.

NEW FINDINGS

We derive a novel microsatellite instability marker panel with increased sensitivity for blood analysis, and show that constitutional microsatellite instability correlates with gene and variant type but not age at first cancer in these patients.

LIMITATIONS

Constitutional mismatch repair deficiency is exceptionally rare and, despite this being one of the largest cohorts analyzed for constitutional microsatellite instability, the power of analysis is limited by sample number.

CLINICAL RESEARCH RELEVANCE

The microsatellite instability assay provides a highly accurate, low-cost, and scalable blood test for constitutional mismatch repair deficiency that achieved 100% sensitivity and 100% specificity and clear separation of patients from controls. Constitutional microsatellite instability is not likely to be a useful predictor for cancer risk stratification in constitutional mismatch repair deficiency syndrome.

BASIC RESEARCH RELEVANCE

In this syndrome, there is a genotype–phenotype correlation involving the gene affected and the type of variant, implying that both impact mismatch repair function. Interestingly, constitutional microsatellite mutation burden does not associate with age at disease onset, suggesting environmental and/or other genetic factors may be more significant contributors to tumorigenesis than mutation rate.

risk; median age at onset is younger than 10 years. This characteristically includes high-grade brain tumors and hematologic malignancies, as well as LS-associated cancers in approximately one-third of cases.[16] CMMRD is also associated with several non-neoplastic features, the most frequent of which are café-au-lait macules reminiscent of neurofibromatosis type 1.[16] Other features can include localized skin hypopigmentation, multiple developmental venous anomalies, pilomatrixoma, and defective immunoglobulin class switch recombination.[17,18] The CMMRD cancer phenotype may depend on which MMR gene is affected in the patient's germline. In a review of 146 published cases, a comparison of *MLH1*- and *MSH2*-associated CMMRD with *PMS2*-associated CMMRD found hematologic malignancies were 1.77-fold more prevalent in the former ($P = .04$), whereas brain tumors were 1.75-fold more frequent in the latter ($P = .01$). Furthermore, *MLH1*- and *MSH2*-associated CMMRD cancers tended to occur earlier than those associated with *MSH6* or PMS2,[17] which reflects the MMR gene–phenotype correlation seen in LS.[13]

For CMMRD diagnosis, assays of MMR function in non-neoplastic tissues provide important ancillary tests to help interpret ambiguous results from genetic testing, in particular variants of uncertain significance[17] and variants in *PMS2*, the MMR gene affected in the majority of patients with CMMRD,[17] for which specialist techniques are required to resolve exon 12–15 variants from those in the closely related *PMS2CL* pseudogene.[19]

Immunohistochemistry of MMR proteins is one such ancillary test, but it cannot detect missense PVs that retain protein expression, and is typically used to assess non-neoplastic tissues in the context of resected tumor material when a lack of staining in all cells may be interpreted as a technical failure.[17] Methylation tolerance and ex vivo MSI are highly sensitive methods to detect CMMRD, but require immortalization and culture of patient primary lymphocytes.[20] CMMRD is also characterized by increased MSI in non-neoplastic tissues, but polymerase chain reaction (PCR) fragment length analysis traditionally used in tumors has too low a sensitivity to detect this constitutional MSI (cMSI).[20,21] Early adaptations to improve the sensitivity of this method either used laborious small pool PCR[17] or analyzed dinucleotide repeats that are insensitive to MSH6 deficiency.[22] More recently, cMSI has been detected by massively parallel sequencing, with several assays separating all CMMRD from control and LS blood samples analyzed.[21–24] Although Chung et al[25] demonstrated that low-pass, whole genome sequencing (WGS) at 1× coverage also accurately detects CMMRD, these assays can require millions of sequence reads per sample,[21,24,25] which may limit scalability for screening when laboratories do not have access to high-capacity sequencing platforms.

We previously published an amplicon sequencing–based assay of 24 mononucleotide repeats that generates a cMSI score for each sample, with higher scores indicating higher cMSI burden. It achieved separation of all CMMRD from control and LS blood samples analyzed,[23] and its method is scalable, low cost, and portable to diagnostic laboratories.[26,27] However, the difference in cMSI score between CMMRD and control samples was minimal, representing a continuum rather than 2 distinct groups. Interestingly, we observed relatively low cMSI scores in CMMRD cases homozygous for a hypomorphic *PMS2* splice-site variant (NM_000535.5(*PMS2*):c.2002A>G) typified by an attenuated phenotype more similar to early-onset LS than classical CMMRD.[23,28] This observation suggested cMSI burden may correlate with CMMRD genotype and/or phenotype, in line with the assumption that the malignant (and nonmalignant) features of CMMRD are, to varying extents, linked to constitutional mutation rate. However, more comprehensive analyses were precluded by the limited cohort size of 32 patients. Exploration of such correlations could broaden our understanding of how MMR deficiency contributes to malignant transformation, aid variant interpretation, and allow risk stratification to guide clinical management of CMMRD.[17,29]

We aimed to first increase the separation of CMMRD patient blood samples from controls by our cMSI assay, and subsequently explore the association of cMSI burden with

CMMRD genotype and phenotype using a larger cohort. The assay originally used markers selected for MSI analysis of tumors,[23,30] which we hypothesized could limit its sensitivity for cMSI analysis. For example, tumors may have different mechanisms and frequencies of microsatellite mutation caused by dysregulated replication,[31] a possible mutator phenotype,[32] and a common lineage whereby cancer subclones are more likely to share variants than the thousands of clones represented in healthy peripheral blood.[33] Therefore, new MSI markers selected for blood analysis were desirable. Here, we identify potentially informative MSI markers from high-depth WGS of CMMRD patient blood, and use amplicon sequencing of a refined marker panel to quantify cMSI burden in more than 50 patients with CMMRD.

Materials and Methods

Patient Samples and Ethical Approval

Anonymized CMMRD peripheral blood leukocyte (PBL) genomic DNAs (gDNAs) were sourced from the Medical University of Innsbruck, Innsbruck, Austria (n = 31), University of Manchester, Manchester, United Kingdom (n = 1), Gustave Roussy Cancer Campus, Villejuif, France (n = 9), Institut Curie, Université de Recherche Paris Sciences et Lettres, Paris, France (n = 4), and Cancer Centre de Recherche Saint-Antoine, Sorbonne University, Paris, France (n = 13). MMR variants were classified according to InSiGHT criteria, version 2.4 (https://www.insight-group.org/criteria/). For patients with 1 or more variants of unknown significance, the diagnosis had been confirmed by assessment of MMR function in non-neoplastic tissues, including assays of germline/constitutional MSI[22,23] and/or ex vivo MSI and methylation tolerance.[20]

Anonymized PBL gDNAs from patients with a CMMRD-like phenotype, according to the C4CMMRD clinical scoring system,[17] who tested negative for germline MMR PVs (CMMRD-negative) were sourced from the Medical University of Innsbruck (n = 8).

Anonymized control PBL gDNAs of patients tested for non-cancer–related conditions were sourced from the Medical University of Innsbruck (n = 73) or as excess diagnostic material from the Northern Genetics Service, Newcastle upon Tyne Hospitals National Health Service Foundation Trust, Newcastle upon Tyne, United Kingdom (n = 50).

Anonymized, genetically diagnosed, LS PBL gDNAs were sourced from the Cancer Prevention Programme Bioresource, Newcastle University, Newcastle upon Tyne, United Kingdom (n = 40).

Anonymized CRC samples were sourced as excess diagnostic material from the Northern Genetics Service as 10-μm formalin-fixed, paraffin-embedded tissue curls of resected tumors (n = 192) or pre-extracted gDNAs from nonfixed endoscopic biopsies (n = 16). Formalin-fixed, paraffin-embedded CRC gDNAs were extracted using the GeneRead DNA FFPE Kit (Qiagen).

Each contributing institution obtained the consent of the individual and/or their legally responsible guardian for use of CMMRD, CMMRD-negative, LS, and control PBL samples in research. MSI analysis of excess diagnostic control PBL and CRC samples was approved by the National Health Service Health Research Authority (REC reference 13/LO/1514).

Samples were divided across several cohorts during selection of novel MSI markers and validation of the new assay, as described in the text and depicted in Supplementary Figure 1. PBL gDNA sample and patient details are provided in Supplementary Table 1.

Genome Sequencing and Variant Analysis

Samples were prepared for WGS by 3-cycle PCR amplification using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs), and were sequenced to >120× coverage on a NovaSeq (Illumina). Reads were aligned to human reference genome build hg19 using BWA mem[34] and BAM files were generated using SAMtools view, sort, and index.[35] Variants were called by a somatic variant calling pipeline and panel of reference control genomes using GATK 4 MuTect2, followed by GetPileupSummaries, CalculateContamination, and FilterMutectCalls, with PCR_indel_model set to NONE.[36] Microsatellites were considered to contain a germline variant if the variant allele with highest frequency had a binomial probability $>10^{-7}$ of equaling 0.5 or 1 (representing heterozygosity or homozygosity, respectively).

For MSI marker selection, microsatellite variants flagged as germline and/or identified in the panel of reference genomes were excluded. Variants annotated as clustered_events, multiallelic, slippage, or PASS, and when the total variant allele frequency was <0.25 (to further exclude potential germline variants), were retained and visually inspected using Integrative Genomics Viewer.[37] Microsatellites with variants captured by high-quality read alignments, not embedded within conserved repetitive elements, and that had higher variant allele frequencies in patients with CMMRD than in controls were selected for further assessment by amplicon sequencing.

Single-Molecule Molecular Inversion Probe Design and Amplicon Sequencing

Single-molecule molecular inversion probes (smMIPs) were designed using MIPgen[38] to amplify MSI markers with capture sizes between 100 and 160 base pair (bp), and an 8N molecular barcode with 4N adjacent to both extension and ligation arms (Supplementary Table 2).

MSI markers were amplified from samples using a published smMIP and high-fidelity polymerase-based protocol.[23] Amplicons were purified using AMPure XP beads (Beckman Coulter), quantified using a QuBit fluorometer 2.0 (Invitrogen), diluted to 4 nM using 10 mM Tris-HCl (pH 8.5), and pooled into 4-nM sequencing libraries. Sequencing libraries were sequenced using custom sequencing primers[23] on a MiSeq (Illumina) to a target depth of 5000×, following manufacturer's protocols.

Microsatellite Amplicon Sequence Analysis and Microsatellite Instability Scoring

Amplicon sequence reads were aligned to human reference genome build hg19 using BWA mem.[34] cMSI analysis of PBL samples followed our previously published analysis pipeline.[23] In brief, reads sharing the same molecular barcode were grouped and the microsatellite length represented in the majority (>50%) of reads was defined as the single-molecule

sequence (smSequence) for each group to reduce PCR and sequencing error for low-frequency variant detection. Groups containing only 1 read or without a majority were discarded. Microsatellite reference allele frequencies (RAFs) in smSequences were used to generate a cMSI score for each sample by comparison with RAFs of 80 known control samples. For any sample, MSI markers with a RAF <0.75 (probable germline variants) or with <100 smSequences were excluded from cMSI scoring. For MSI analysis of CRCs, the samples were divided between 2 cohorts to train and validate a previously published naïve Bayesian MSI classifier, which assesses both the frequency and allelic bias of microsatellite deletions in sequencing reads to generate a tumor MSI score.[30]

Statistical Analyses and Data Availability

All analyses used R, version 4.0.2 (https://www.r-project.org/). Comparisons of 2 sample groups used the Mann-Whitney test. Comparisons of more than 2 sample groups used the Kruskal-Wallis test. Correlation of variables that could be assumed to have a linear relationship used Pearson's r, whereas Spearman's ρ was used for variables when a monotonic, but not necessarily linear, relationship could be assumed. For pairwise analyses of cMSI score in patients sharing the same genotype, the significance of the correlation was assessed using a permutation test that takes into account that individual cMSI scores may be used in multiple pairs. CIs for sensitivity and specificity estimates used a binomial distribution.

Genome sequence BAM and amplicon sequence FASTQ files are available from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) using study IDs PRJEB39601 and PRJEB53321, respectively.

Results

Genome Sequencing of Blood Identifies High-Sensitivity Microsatellite Instability Markers

Three CMMRD (2 *PMS2*- and 1 *MSH6*-associated), 1 LS (*MLH1*-associated), and 2 control blood samples were whole genome sequenced (Supplementary Figure 1). An LS sample was included, as highly sensitive MSI analysis and single-base MMR assays have previously detected reduced MMR function in blood and cell lines with 1 dysfunctional MMR allele.[39–41] The frequency of mononucleotide repeat (MNR) variants was increased in *PMS2*-associated and *MSH6*-associated CMMRD samples relative to control and LS samples, whereas variants in longer motif microsatellites were only increased in the *PMS2*-associated CMMRD samples (Supplementary Figures 2*A* and *B*). To derive a novel marker panel for cMSI analysis, the WGS data were filtered for microsatellites displaying an increase in nongermline variant alleles in the CMMRD samples compared with the controls. This identified more than 2000 loci of interest, the majority of which were 11- to 16-bp A-homopolymers. Manual review of these loci short-listed 121 MNRs as candidate MSI markers (Figure 1). Longer motif microsatellites were excluded, as these did not show increased variants in the *MSH6*-associated CMMRD blood sample compared with controls, consistent with MSI only affecting MNRs in MSH6-deficient tissue.[22,42] smMIPs were designed to capture these 121 MNRs and were assessed by smMIP amplicon sequencing of 3 control samples. Of these, 91 smMIPs (capturing 98 MNRs) generated sufficient reads to be taken forward (Figure 1).

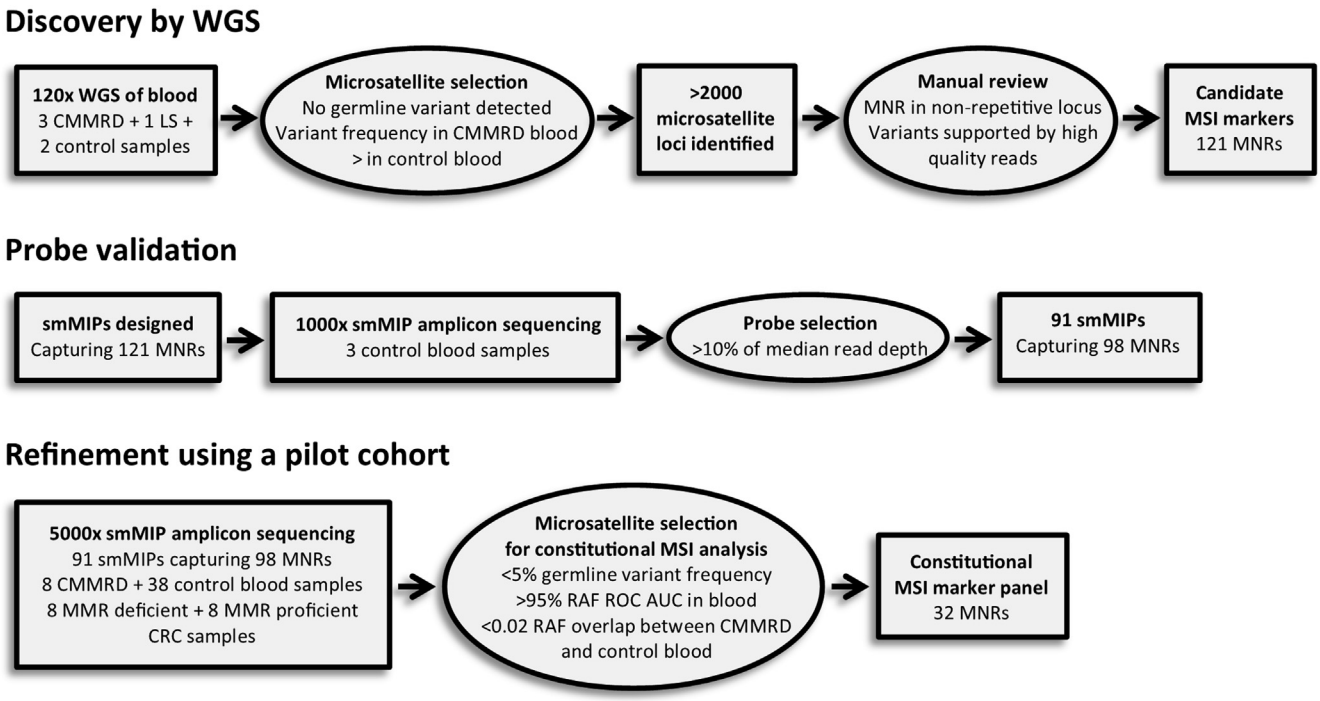


Figure 1. Flow chart of MSI marker selection.

The MSI marker panel was refined based on the ability of candidate MNRs to discriminate between MMR-deficient and MMR-proficient tissues using smMIP-amplicon sequencing of a pilot cohort of 8 CMMRD and 38 control PBL gDNAs, and 8 MMR-deficient and 8 MMR-proficient CRC gDNAs (Supplementary Figure 1). All except 7 control PBL samples had been analyzed previously using the 24 tumor-derived MNRs of the original MSI assay,[23] allowing comparison of marker sets. The new MSI markers had much greater differences in RAF between MMR-deficient and MMR-proficient samples in both CRCs ($P = 1.8 \times 10^{-5}$) and PBLs ($P = 2.2 \times 10^{-8}$; Supplementary Figures 2*C* and *D*), indicating they are more sensitive to MMR deficiency than the original MSI markers. Based on these data, the candidate markers were refined to a panel of the most discriminatory 32 MNRs for cMSI analysis (Figure 1; Supplementary Table 2).

New Microsatellite Instability Markers Enhance Detection of Constitutional Mismatch Repair Deficiency

The 32 new MSI markers were amplified and sequenced from 80 control PBL gDNAs to provide a reference for cMSI scoring, and a blinded cohort consisting of PBL gDNAs from 57 patients with CMMRD, 8 CMMRD-negative patients (CMMRD-like phenotype but no germline MMR PVs), and 43 control individuals. Forty LS PBL gDNAs (10 for each MMR gene) were also analyzed to investigate whether increased cMSI is specific to biallelic loss of MMR function. One sample from the blinded cohort failed to amplify, and was later revealed to be a CMMRD case. All other sample amplicons were sequenced and a cMSI score was generated for each. Markers with low (<100) smSequence counts were observed in only 4 samples from the blinded cohort; 2 had a single low count marker and the other 2 had <100 smSequences in ≥17 MSI markers with equivalent results on repeat amplification and sequencing, suggesting poor sample quality. On unblinding, these 2 poor-quality samples were revealed to be CMMRD cases.

The cMSI score identified CMMRD with 100% sensitivity (56 of 56; 95% CI, 93.6%–100.0%) and 100% specificity (171 of 171; 95% CI, 97.9%–100.0%), including the 2 poor-quality CMMRD samples. There was a clear separation of all CMMRD samples from control, LS, and CMMRD-negative samples (Figure 2*A*, Supplementary Table 1). cMSI score was associated with affected MMR gene ($P = 1.2 \times 10^{-3}$); patients with MSH6 deficiency had significantly lower cMSI scores than patients with MSH2 deficiency ($P = 2.4 \times 10^{-4}$) or PMS2 deficiency ($P = 6.0 \times 10^{-3}$), and a trend for lower scores than patients with MLH1 deficiency ($P = .05$, multiple testing significance at $P < 1.67 \times 10^{-2}$). LS cMSI scores were not significantly different from controls ($P = .17$), but it was notable that 6 scores (3.7–11.3) were greater than the highest control score (3.6). CMMRD-negative samples overall had marginally higher cMSI scores than controls ($P = .02$), with 2 scores (4.1 and 5.3) being greater than the highest control score (3.6). As these high-scoring LS and CMMRD-negative samples had much lower cMSI scores than the CMMRD samples, and due to unavailability of cancer data or MMR variant identity in the patients with LS, these were not analyzed further. To assess cMSI assay reproducibility, residual DNA samples available from 25 patients with CMMRD and 33 controls were re-amplified, sequenced, and scored, and a strong correlation was found between initial and repeat cMSI scores ($r = 0.994$, $P < 10^{-15}$) (Figure 2*B*). There was no significant correlation in cMSI score between control repeats ($r = 0.105$, $P = .56$), suggesting differences in cMSI score between controls is mostly random technical variation. Although unlikely to affect sample classification, small but significant differences were observed between controls of different amplification and sequencing batches (maximum difference in median control cMSI score = 0.94, $P = 1.2 \times 10^{-8}$, Supplementary Figure 3).

Fifty CMMRD and 75 control samples were also analyzed using the original 24 MSI markers.[23] The new MSI markers had greater RAF-based receiver operator characteristic (ROC) area under curve (AUC) values for CMMRD detection than the original set ($P = 9.0 \times 10^{-14}$) (Supplementary Figure 4). The new MSI markers were longer (range, 11–15 bp vs 7–12 bp; $P = 1.9 \times 10^{-7}$) and there was a positive correlation between marker length and ROC AUC ($\rho = 0.730$, $P = 1.8 \times 10^{-10}$). However, comparison of markers of equivalent size (11–12 bp) found higher ROC AUCs for the new markers than the original ($P = 2.5 \times 10^{-4}$; Figure 3*A*). The new MSI markers were ranked by RAF ROC AUC to separate CMMRD from control samples (Supplementary Table 2) and the most discriminatory 24 new MSI markers gave a large cMSI score separation of 15.3 between CMMRD and control samples, compared with the 0.1 cMSI score overlap when using the original 24 MSI markers (Figure 3*B*). Using only 3 new MSI markers gave 100% accurate CMMRD detection (Supplementary Figure 5). The new MSI markers also enhanced tumor MSI classification of CRCs compared with the original set (Supplementary Figures 6*A*–*D*). Despite differences in variant allele frequencies and indel size between CRCs and blood (Supplementary Figure 7), MSI marker RAF ROC AUCs for the detection of MMR deficiency were correlated between the 2 tissue types ($\rho = 0.715$, $P = 9.0 \times 10^{-5}$).

Constitutional Mismatch Repair Deficiency Constitutional Microsatellite Instability Burden Is Associated With Mismatch Repair Variant But Not Age at Tumor Onset

There was a breadth of cMSI scores between patients with CMMRD with deficiency of the same MMR gene, suggesting potential genotype or phenotype correlations with cMSI burden. Variants were labeled as 1 of 3 types according to their effect on protein sequence - truncating and copy number variants (CNVs), splicing variants, and missense variants. Truncating and intragenic CNVs (ie, deletions or duplications of 1 or more exons) were grouped together due to their direct disruption of protein structure and/or expression (Supplementary Table 1). There were 3 exceptions; NM_000179.2(*MSH6*):c.2426_2428del was labeled as a single amino acid deletion (1AAdel),

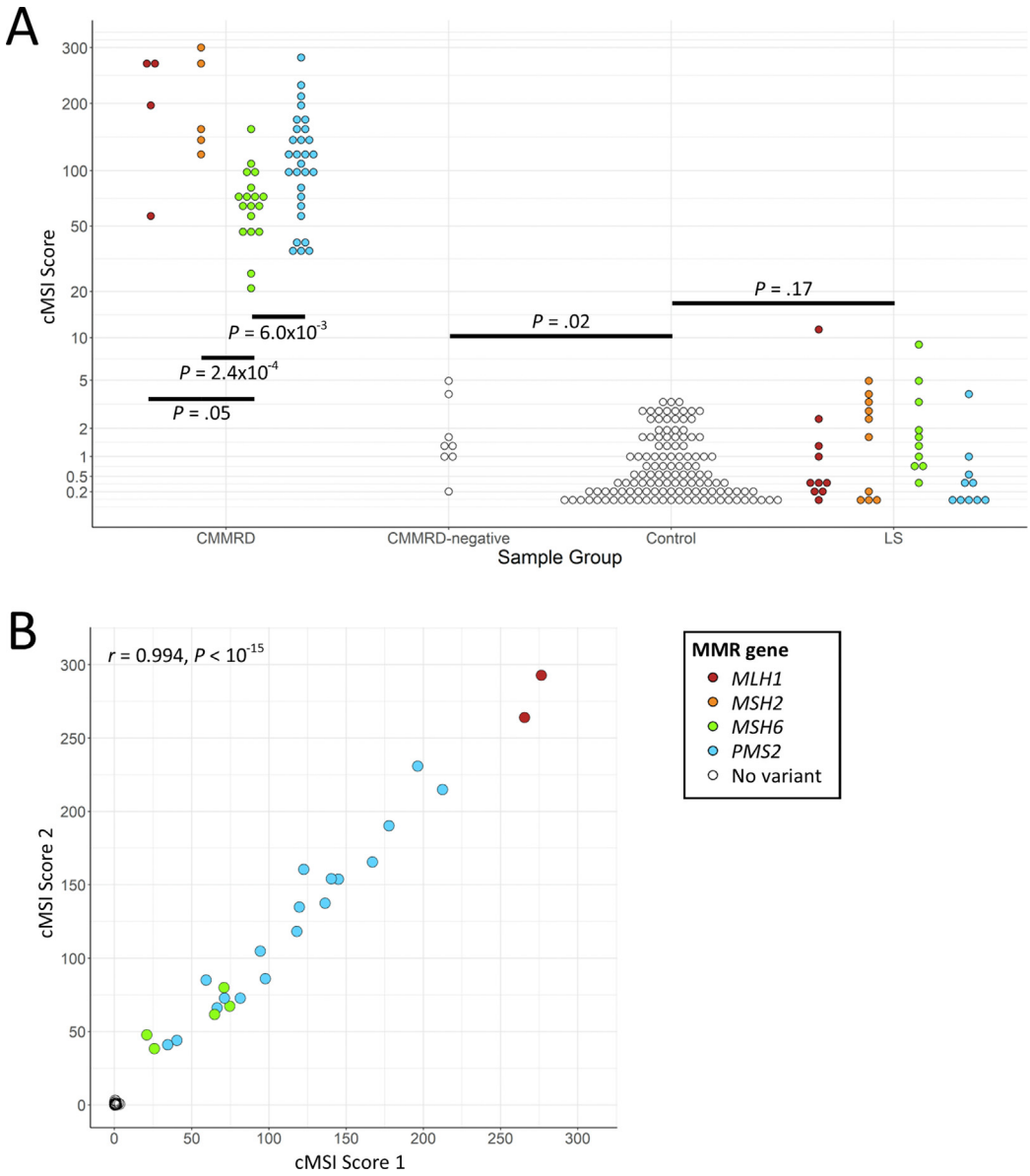


Figure 2. Sample cMSI scores. The cMSI scores of a blinded cohort of 56 CMMRD (*MLH1* n = 4, *MSH2* n = 5, *MSH6* n = 18, *PMS2* n = 29), 8 CMMRD-negative, and 43 control PBL gDNAs, 80 reference control PBL gDNAs, and 40 LS (*MLH1* n = 10, *MSH2* n = 10, *MSH6* n = 10, *PMS2* n = 10) PBL gDNAs, derived from 32 new MSI markers using the amplicon sequencing and MSI scoring method of Gallon et al.[23] CMMRD-negative refers to patients with a CMMRD-like phenotype but no germline MMR variants. The y-axis is scaled based on a logit transformation (*A*). A comparison of initial and repeat cMSI scores of 25 CMMRD (*MLH1* n = 2, *MSH6* n = 5, *PMS2* n = 18) and 33 control samples with residual sample available (*B*).

NM_000179.2(*MSH6*):c.1763_1771dup was labeled as a triple amino acid duplication (3AAdup), and NM_000535.5(*PMS2*):c.2002A>G was labeled as a splicing (missense) variant. The latter creates a novel splice-site causing a p.(Ile668*) truncation, but blood cells from these patients residually express full-length and translatable *PMS2* messenger RNA containing the p.(Ile668Val) missense variant.[28] Patients were grouped by their variant types and cMSI scores were found to be different between the groups ($P = 3.0 \times 10^{-3}$; Figure 4*A*). No increase in cMSI score had been observed in LS blood samples compared with controls, suggesting that cMSI score is determined predominantly by the least disrupted MMR allele. In general, missense variants have more variable effect on protein function than truncating variants or intragenic CNVs. Therefore, patients with CMMRD with at least 1 missense variant (excluding those patients with NM_000535.5(*PMS2*):c.2002A>G due to its splicing effect) were compared with the rest of the cohort and were found to have significantly lower cMSI scores ($P = 7.4 \times 10^{-3}$; Figure 4*A*). Conversely, patients with biallelic truncating variants or intragenic CNVs had significantly higher cMSI scores than those without ($P = .02$; Figure 4*A*). The frequency of mono- or biallelic missense variants and the frequency of biallelic truncating variants/intragenic CNVs were both equivalent between MMR genes ($P = .54$, $P = .61$, respectively), indicating these differences were not

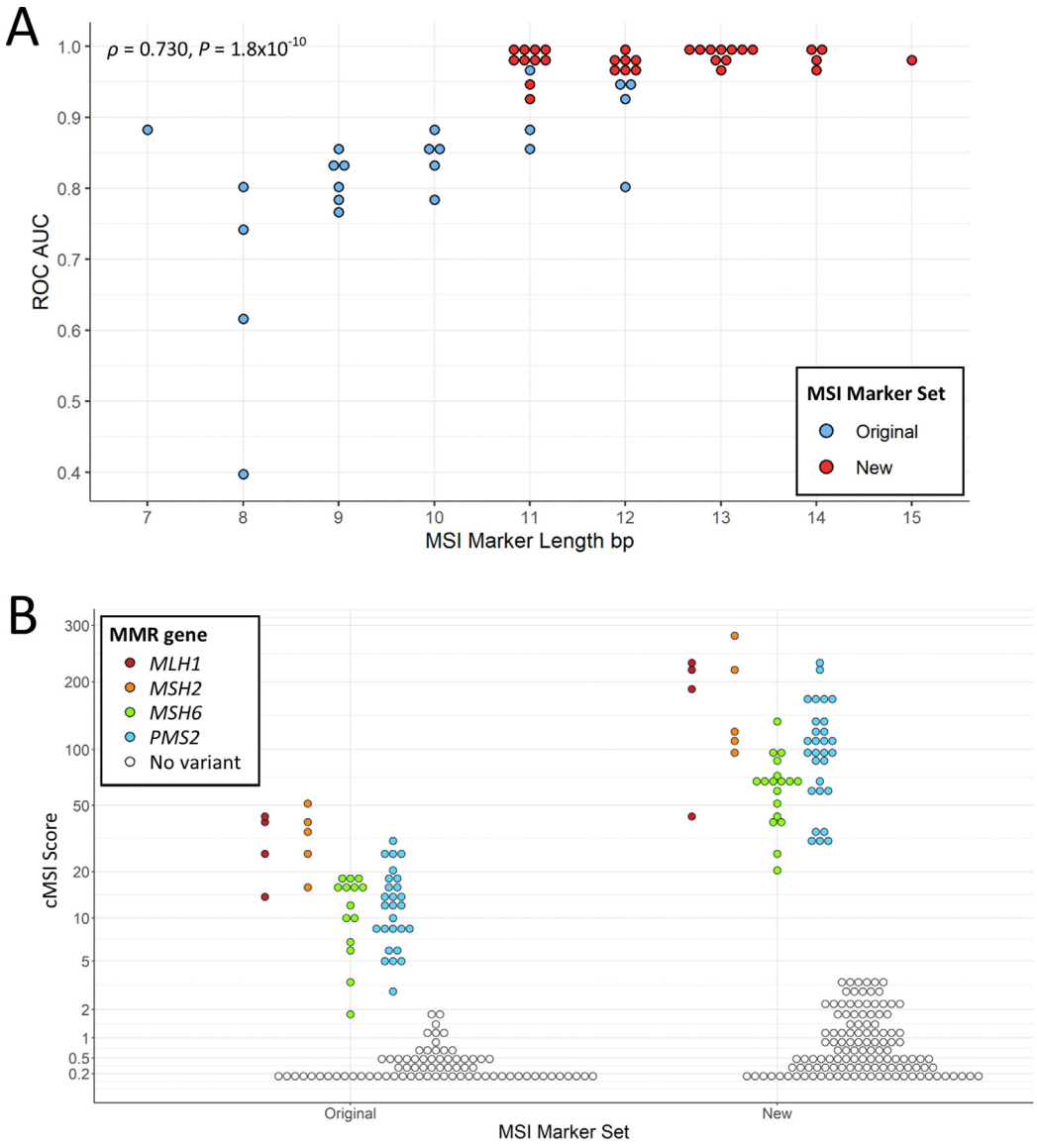


Figure 3. MSI marker characteristics and performance. A comparison of the length of each MSI marker and its ROC AUC to discriminate between CMMRD and control PBL samples (*A*). A comparison of cMSI score of 50 CMMRD (*MLH1* n = 4, *MSH2* n = 5, *MSH6* n = 14, *PMS2* n = 27) and 75 control PBL samples using either the original 24 tumor-derived MSI markers or an equivalent number of the most discriminatory of the new blood-derived MSI markers. The y-axis is scaled based on a logit transformation (*B*).

due to an overrepresentation of variant types in any 1 gene. To further assess whether MMR variants associate with cMSI burden, cMSI score between patients sharing the same genotype were compared. Twelve pairwise comparisons between siblings of 8 CMMRD families were possible, together with 10 pairwise comparisons between 5 unrelated patients homozygous for the recurrent variant NM_000535.5(*PMS2*):c.2007-2A>G. cMSI scores were positively correlated between pairs ($r = 0.744$, permutation test $P = 2.9 \times 10^{-4}$; Figure 4*B*).

A clinical history of tumor diagnoses was available for all patients with CMMRD (n = 56). Five patients had no cancer history and for another the age at tumor diagnosis was unknown, meaning age at first cancer could be compared with cMSI score in 50 patients (Supplementary Table 1). cMSI score was not significantly correlated with age at first tumor overall ($\rho = -0.154$, $P = .29$; Figure 5), or in subgroup analyses of MSH6-deficient patients ($\rho = -0.342$, $P = .20$) and PMS2-deficient patients ($\rho = -0.013$, $P = .95$). It is possible that cMSI burden is associated with the onset of specific tumor types, as there is evidence that both sporadic MMR-deficient and CMMRD-related brain and hematologic malignancies have reduced MSI compared with cancers within the LS spectrum.[3,21] However, no significant correlation was found between cMSI score and age at onset of brain tumors ($\rho = -0.167$, $P = .32$), hematologic malignancies ($\rho = -0.285$, $P = .27$), or LS-associated tumors ($\rho = -0.143$, $P = .58$). There was also no significant association of age at first tumor with affected MMR gene ($P = .48$) or type of variant ($P = .38$).

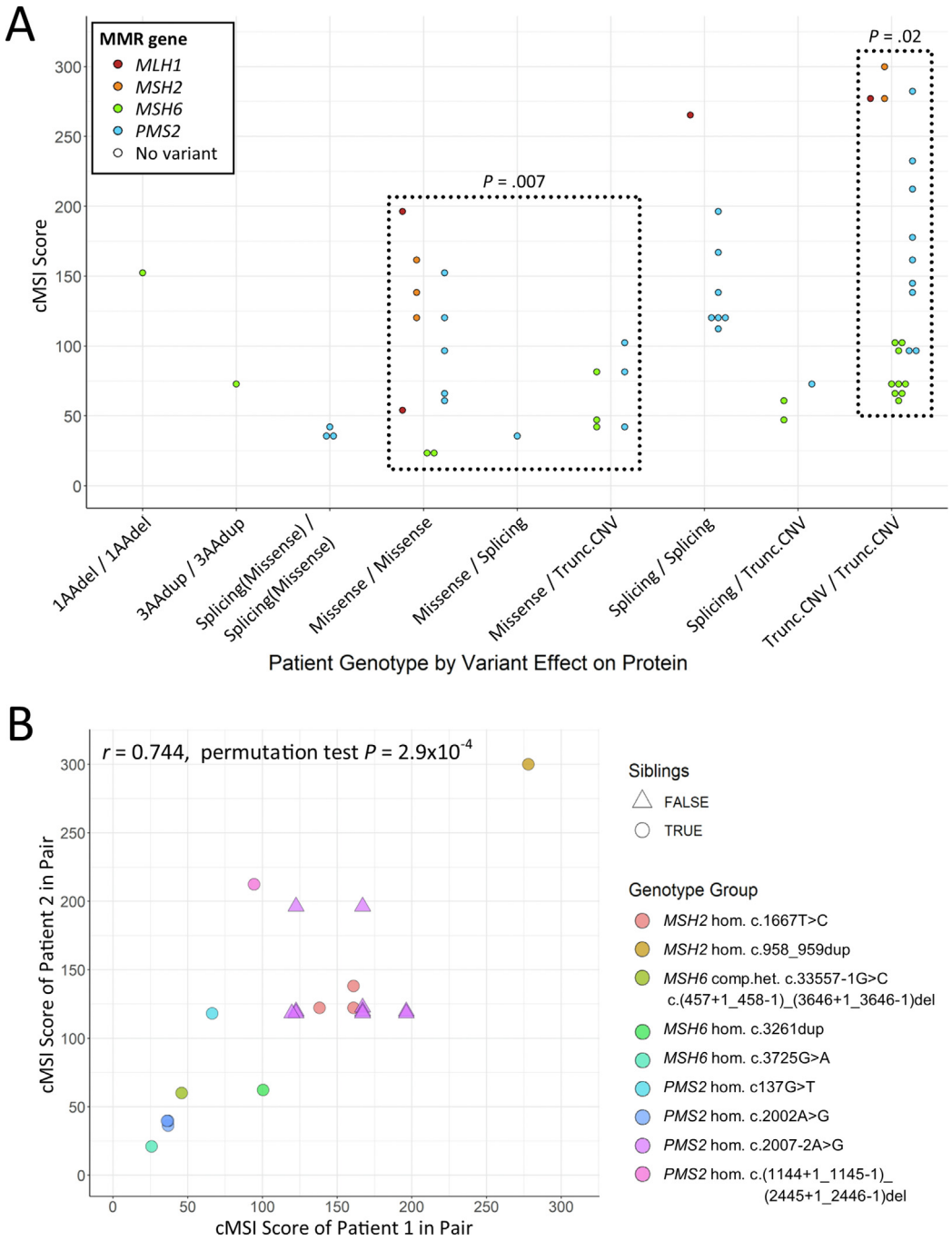


Figure 4. Sample cMSI scores by patient genotype. The cMSI scores of 56 patients with CMMRD (*MLH1* n = 4, *MSH2* n = 5, *MSH6* n = 18, *PMS2* n = 29) grouped by the type of germline MMR variant according to effect on protein sequence. The *dotted boxes* highlight patients with mono- or biallelic missense variants and patients with biallelic truncating or copy number variants. AA, amino acid; Trunc., truncating (*A*). A pairwise comparison of the cMSI scores of patients with CMMRD who share the same MMR genotype. hom., homozygous; comp.het., compound heterozygous (*B*).

Other factors that might affect cMSI burden include age at sample collection[39,43] and contaminating tumor cells or DNA. Age at sample collection was not significantly correlated with cMSI score among 30 patients with CMMRD with data available ($\rho = -0.310$, $P = .10$; Supplementary Figure 8*A*), but was correlated with age at first tumor ($r = 0.727$, $P = 3.9 \times 10^{-5}$) as expected, given CMMRD diagnoses are typically made at or after presentation of malignancy. Similarly, cMSI score was not significantly correlated with age at sample collection in 50 controls with data available ($P = .65$) or in the 40 patients with LS ($P = .28$). For 27 patients with CMMRD, it was also known whether a tumor was present at the time of sample collection; the cMSI scores of the 18 patients with a tumor were not significantly different from those without ($P = .50$, Supplementary Figure 8*B*).

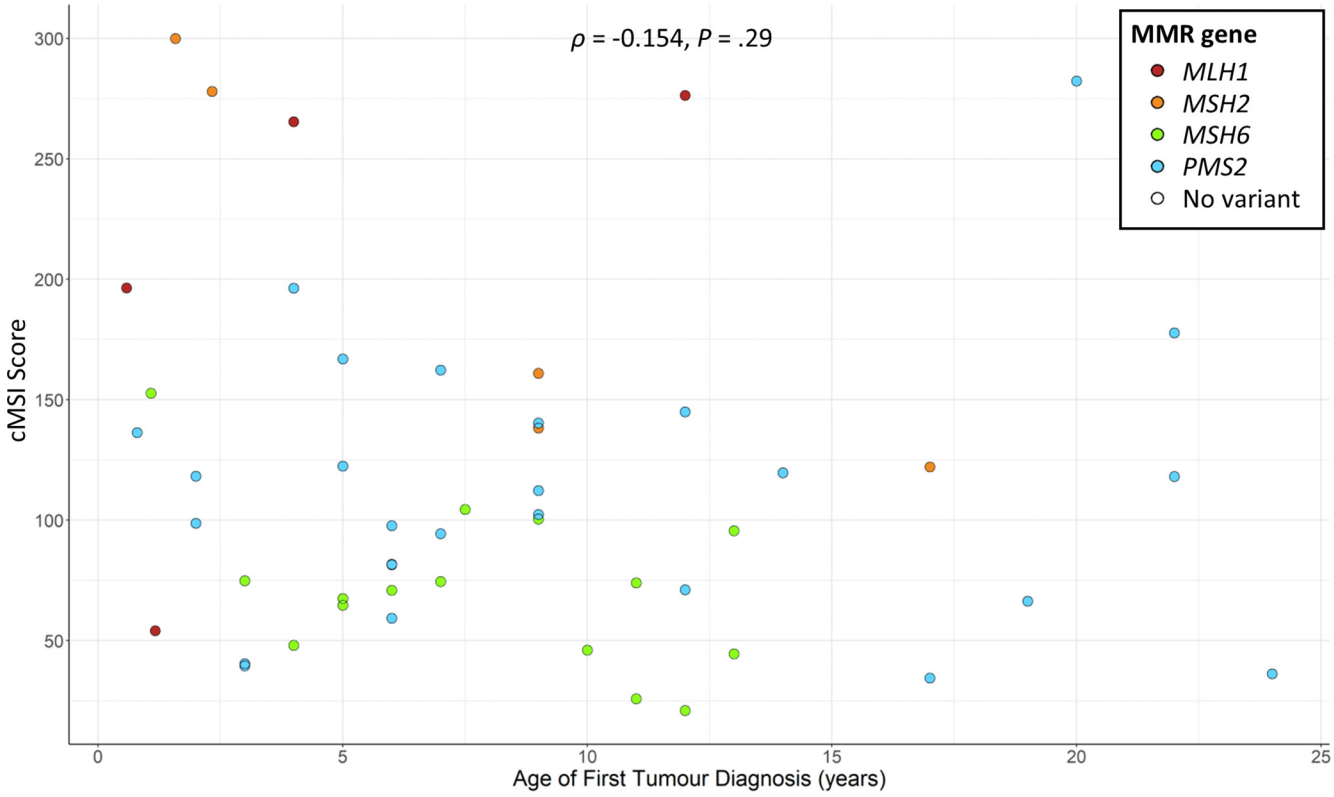


Figure 5. Associations of disease phenotype with cMSI score. The cMSI score and age at first tumor of 50 patients with CMMRD (*MLH1* n = 4, *MSH2* n = 5, *MSH6* n = 16, *PMS2* n = 25).

Discussion

In this study, novel MSI markers were selected from blood WGS to enhance an existing amplicon sequencing–based cMSI assay, achieving excellent separation of CMMRD samples from controls. MNRs were used, as these showed increased instability in WGS data from both *MSH6*- and *PMS2*-associated CMMRD blood samples, whereas increased instability in longer motif microsatellites was found in the *PMS2*-associated CMMRD blood samples only. This is consistent with MSH6 being involved in the repair of single base (but not larger) indel loops, whereas PMS2 is active across MMR as the endonuclease within the MutL complex.[1,2] Hence, increased MSI is typically observed in MNRs only in MSH6-deficient tissues.[22,42] The new MSI markers were longer than the original set, ranging between 11 and 15 bp, which is equivalent to the most sensitive and specific A-homopolymers identified in The Cancer Genome Atlas tumor exome sequencing data.[44] This suggests that a microsatellite's diagnostic utility may simply be a function of its length. However, the new blood-derived MSI markers of 11–12 bp have significantly higher ROC AUCs than the original tumor-derived set of the same length, confirming this new selection has identified more discriminatory markers regardless of their size. The new MSI markers also enhanced detection of MMR deficiency in CRCs, suggesting that they will be sensitive irrespective of tissue type, despite our initial hypothesis that some may be more sensitive in blood than in tumors. However, the original tumor-derived set had also been selected to be $\leq$12 bp to minimize PCR and sequencing error, and to have a *single nucleotide polymorphism* within 30 bp to allow the allelic bias of microsatellite deletions to be used in tumor MSI classification.[30] Therefore, different marker selection criteria preclude clear conclusions regarding whether specific microsatellites are more sensitive to mutation in MMR-deficient tumors compared with MMR-deficient blood.

Sequencing-based MSI analysis of non-neoplastic tissues to detect CMMRD has now been demonstrated with a variety of methods.[21,23–25] The smMIP amplicon-sequencing cMSI assay used here is relatively inexpensive, with total reagent and sequencing costs of $25–$50 per sample, based on analysis of 32 MNRs in 80 or 12 samples on a MiSeq, version 3 or version 2 micro kit, respectively. These costs could be reduced, as only 3 MSI markers were required for separation of CMMRD from control samples (Supplementary Figure 5). The method is scalable from functional testing of a few samples to high-throughput screening for CMMRD, as demonstrated in a study of more than 700 children with neurofibromatosis type 1-like phenotypes but negative for *NF1* or *SPRED1* germline PVs, in whom CMMRD is a differential diagnosis.[26] Assay limitations include its use of custom sequencing primers, which prevents combining the amplicon library with others, and its validation on a MiSeq, which may not be the sequencing platform of choice. Batch effects were also observed, although these are unlikely to affect sample classification, given the very clear separation of CMMRD samples (including individuals homozygous for hypomorphic MMR variants) from control, LS, and CMMRD-

negative samples, as well as the highly reproducible cMSI scores. A possible influence of batch on sample classification was reduced by spreading the reference controls for cMSI score calculation across 5 sequencing runs (Supplementary Table 1). Despite this, for clinical use it would be pertinent to include a set of control samples on all runs to monitor batch effects. Hence, the present cMSI assay could be a valuable asset to CMMRD diagnostics and screening studies. Our results also support reclassification of 8 MMR variants of unknown significance (Supplementary Table 1) as pathogenic, at least in the context of CMMRD.

The cMSI scores of patients with CMMRD were associated with genotype. Previously, González-Acosta et al[24] reported a reduced cMSI burden in *MSH6*- vs *MSH2*-associated patients with CMMRD using an alternative amplicon sequencing assay. We have shown that this is also true for *MSH6*- vs *PMS2*-associated CMMRD and that there is a similar trend comparing *MSH6*- with *MLH1*-associated CMMRD. A reduced cMSI burden of *MSH6*- compared with *PMS2*-associated CMMRD was also observed in our WGS data, and is consistent with WGS of CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9-knockout cell lines, which showed a reduced indel frequency in MSH6- compared with MLH1-, MSH2-, or PMS2-deficient cells.[45] The redundancy for 1-bp indel repair between MSH2-MSH6 (MutSα) and MSH2-MSH3 (MutSβ) heterodimers[1,2] likely explains the reduced frequency of MNR variants in the constitutional tissues of *MSH6*-associated CMMRD. We also observed genotype–phenotype correlations with respect to the type of MMR variant and cMSI score, with missense variants and truncating and/or intragenic CNVs being associated with lower and higher cMSI scores, respectively. To our knowledge, this is a novel observation for MMR genes and could have implications for our understanding of how MMR genotype influences mutation rate. It would be interesting, for example, to explore whether MMR missense variants are associated with reduced MSI in MMR-deficient tumors, and whether this has any association with clinical course.

No significant correlation of MMR genotype or cMSI score with age at first tumor was observed among the 56 patients with CMMRD analyzed. Wimmer et al[17] previously found differences in the incidence of central nervous system tumors and hematologic malignancies and age at first tumor by affected MMR gene in CMMRD, but analyzed a larger cohort of 146 patients. In LS, it is well established that the MMR genes are associated with distinct cancer spectra and risks.[13] With respect to variant type, Suerink et al[46] found both CRC and endometrial cancer occurred earlier in LS carriers of *PMS2* variants that are predicted to cause loss of RNA expression compared with those that retain expression. Ryan et al[47] similarly reported an association between truncating *MLH1* PVs and earlier onset of LS endometrial cancer. Otherwise there are very limited data supporting an effect for type or position of MMR PVs on clinical phenotype in LS.[48] Therefore, although a correlation of MMR genotype with disease penetrance is probable in CMMRD, it is seemingly much weaker than that with cMSI burden, and hence was not observable with our limited cohort size and method. Consequently, both MMR genotype and cMSI score are, at most, weak predictors of age at tumor onset in CMMRD, and may not be clinically useful for risk stratification. However, it remains an intriguing observation that some of the samples with lowest cMSI score include the 3 patients homozygous for the hypomorphic Inuit founder variant NM_000535.5(*PMS2*):c.2002A>G, who have residual expression of functional PMS2 and phenotypes more similar to early-onset LS than classical CMMRD.[28]

A link between mutation rate and cancer risk is based, in part, on the increased mutation burden and rate of tumors compared with healthy tissue,[32] as well as positive correlations of tissue-specific cancer incidence with stem cell division rate[49] and cumulative mutation burden.[50] Further supporting this link, increased mutation burdens in the normal intestinal crypts of cancer predisposition syndromes associated with germline *POLE* and *POLD1* PVs[51] and germline biallelic *MUTYH* PVs[52] have been discovered very recently, as well as increases in mutation rate in primary mammary cells of *BRCA1/2* PV carriers.[53] The question then remains, why are cMSI score and disease phenotype not more strongly correlated in CMMRD? A key limitation of our study is the restricted subgroup or multivariate analyses that might disentangle possible confounding variables. For example, older patients at the time of sampling will likely have higher cMSI burdens, as has been observed in the general population and LS using single-molecule PCR techniques.[39,43] Therefore, when using cMSI score as an estimate of constitutional mutation rate, the positive correlation between age at sampling and age at first tumor within our cohort is likely to confound detection of a negative correlation between constitutional mutation rate and cancer onset. Different patient ages at sampling may also impact other analyses, for example, weakening the correlation in cMSI score between patients sharing the same genotype. Analyzing constitutional mutation rate directly may be superior, but would require alternative methods to quantify, for example, serial sampling of individuals or use of models, which have their own limitations. Furthermore, repair of microsatellite indels is only one of several functions of the MMR system, which includes repair of single-base substitutions and induction of cell cycle arrest and apoptosis.[1,2] Disruption of these pathways may be more significant than repair of microsatellite indels in tumorigenesis,[54] as may environmental and genetic modifiers of cancer risk. Familial modifiers are known to have large effects on cancer risk in LS[55] and genetics may be of particular importance in CMMRD, given parental consanguinity is seen in approximately one-half of CMMRD families.[17] We found no evidence that cMSI score was influenced by presence of a tumor at the time of blood sampling in the patients with CMMRD, but some CMMRD-negative and LS samples had marginally increased cMSI scores. We could not analyze these further due to a lack of cancer data, but future exploration of the effect of contaminating MSI-H circulating tumor cells or DNA on cMSI analysis may be warranted.

In summary, we have analyzed cMSI burden in a relatively large cohort of patients with CMMRD, given the rarity of the syndrome, combining novel MSI markers and a simple amplicon-sequencing method to enhance CMMRD diagnostics. Our data showed an MMR genotype–phenotype correlation with both the gene affected and the type of variant influencing cMSI burden, suggesting MMR genotype could also have implications for tumor mutation burden. However, no association of cMSI score with clinical phenotype was found, implying that environmental and/or other genetic factors could be more significant contributors to tumorigenesis than an increased constitutional mutation rate. Therefore, although cMSI score is a useful diagnostic biomarker, it likely cannot be used to stratify cancer risk in CMMRD, as we initially hypothesized.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2022.12.017.

References

1. Kunkel T, Erie D. DNA mismatch repair. Annu Rev Biochem 2005;74:681–710.
2. Jiricny J. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol 2006;7:335–346.
3. **Gallon R, Gawthorpe P,** Phelps RL, et al. How should we test for Lynch syndrome? A review of current guidelines and future strategies. Cancers (Basel) 2021;13:406.
4. Ryan NAJ, Glaire MA, Blake D, et al. The proportion of endometrial cancers associated with Lynch syndrome: a systematic review of the literature and meta-analysis. Genet Med 2019;21:2167–2180.
5. Campbell B, Light N, Fabrizio D, et al. Comprehensive analysis of hypermutation in human cancer. Cell 2017;171:1042–1056.e10.
6. **Wang J, Sun L**, Myeroff L, et al. Demonstration that mutation of the type II transforming growth factor beta receptor inactivates its tumor suppressor activity in replication error-positive colon carcinoma cells. J Biol Chem 1995;270:22044–22049.
7. Ionov Y, Yamamoto H, Krajewski S, et al. Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution. Proc Natl Acad Sci U S A 2000;97:10872–10877.
8. Duval A, Hamelin R. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res 2002;62:2447–2454.
9. **Deacu E, Mori Y**, Sato F, et al. Activin type II receptor restoration in ACVR2-deficient colon cancer cells induces transforming growth factor-beta response pathway genes. Cancer Res 2004;64:7690–7696.
10. **Lee J, Li L**, Gretz N, et al. Absent in Melanoma 2 (AIM2) is an important mediator of interferon-dependent and -independent HLA-DRA and HLA-DRB gene expression in colorectal cancers. Oncogene 2012;31:1242–1253.
11. Sekine S, Mori T, Ogawa R, et al. Mismatch repair deficiency commonly precedes adenoma formation in Lynch syndrome-associated colorectal tumorigenesis. Mod Pathol 2017;30:1144–1151.
12. **Ahadova A, Gallon R**, Gebert J, et al. Three molecular pathways model colorectal carcinogenesis in Lynch syndrome. Int J Cancer 2018;143:139–150.
13. **Dominguez-Valentin M, Sampson JR, Seppälä TT**, et al. Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. Genet Med 2020;22:15–25.
14. Win A, Jenkins M, Dowty J, et al. Prevalence and penetrance of major genes and polygenes for colorectal cancer. Cancer Epidemiol Biomarkers Prev 2017;26:404–412.
15. Suerink M, Ripperger T, Messiaen L, et al. Constitutional mismatch repair deficiency as a differential diagnosis of neurofibromatosis type 1: consensus guidelines for testing a child without malignancy. J Med Genet 2019;56:53–62.
16. Wimmer K, Rosenbaum T, Messiaen L. Connections between constitutional mismatch repair deficiency syndrome and neurofibromatosis type 1. Clin Genet 2017;91:507–519.
17. Wimmer K, Kratz C, Vasen H, et al. Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium 'care for CMMRD' (C4CMMRD). J Med Genet 2014;51:355–365.
18. Shiran S, Ben-Sira L, Elhasid R, et al. Multiple brain developmental venous anomalies as a marker for constitutional mismatch repair deficiency syndrome. Am J Neuroradiol 2018;39:1943–1946.
19. van der Klift HM, Tops CM, Bik EC, et al. Quantification of sequence exchange events between PMS2 and PMS2CL provides a basis for improved mutation scanning of Lynch syndrome patients. Hum Mutat 2010;31:578–587.
20. **Bodo S, Colas C, Buhard O**, et al. Diagnosis of constitutional mismatch repair-deficiency syndrome based on microsatellite instability and lymphocyte tolerance to methylating agents. Gastroenterology 2015;149:1017–1029.
21. **Chung J, Maruvka YE**, Sudhaman S, et al. DNA polymerase and mismatch repair exert distinct microsatellite instability signatures in normal and malignant human cells. Cancer Discov 2021;11:1176–1191.
22. Ingham D, Diggle C, Berry I, et al. Simple detection of germline microsatellite instability for diagnosis of constitutional mismatch repair cancer syndrome. Hum Mutat 2013;34:847–852.
23. Gallon R, Muhlegger B, Wenzel S, et al. A sensitive and scalable microsatellite instability assay to diagnose constitutional mismatch repair deficiency by sequencing of peripheral blood leukocytes. Hum Mutat 2019;40:649–655.
24. **González-Acosta M, Marín F, Puliafito B**, et al. High-sensitivity microsatellite instability assessment for the

detection of mismatch repair defects in normal tissue of biallelic germline mismatch repair mutation carriers. J Med Genet 2020;57:269–273.
25. **Chung J, Negm L**, Bianchi V, et al. Genomic microsatellite signatures identify germline mismatch repair deficiency and risk of cancer onset. J Clin Oncol 2023; 41:766–777.
26. **Perez-Valencia JA, Gallon R**, Chen Y, et al. Constitutional mismatch repair deficiency is the diagnosis in 0. 41% of pathogenic NF1/SPRED1 variant negative children suspected of sporadic neurofibromatosis type 1. Genet Med 2020;22:2081–2088.
27. Gallon R, Sheth H, Hayes C, et al. Sequencing-based microsatellite instability testing using as few as six markers for high-throughput clinical diagnostics. Hum Mutat 2020;41:332–341.
28. Li L, Hamel N, Baker K, et al. A homozygous PMS2 founder mutation with an attenuated constitutional mismatch repair deficiency phenotype. J Med Genet 2015;52:348–352.
29. Durno C, Boland C, Cohen S, et al. Recommendations on surveillance and management of biallelic mismatch repair deficiency (BMMRD) syndrome: a consensus statement by the US Multi-Society Task Force on Colorectal Cancer. Gastrointest Endosc 2017; 85:873–882.
30. **Redford L, Alhilal G**, Needham S, et al. A novel panel of short mononucleotide repeats linked to informative polymorphisms enabling effective high volume low cost discrimination between mismatch repair deficient and proficient tumours. PLoS One 2018;13:e0203052.
31. Hanahan D, Weinberg R. Hallmarks of cancer: the next generation. Cell 2011;144:646–674.
32. Loeb LA. Human cancers express a mutator phenotype: hypothesis, origin, and consequences. Cancer Res 2016; 76:2057–2059.
33. Biasco L, Pellin D, Scala S, et al. In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. Cell Stem Cell 2016;19:107–119.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25:1754–1760.
35. **Li H, Handsaker B**, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–2079.
36. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–1303.
37. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011; 29:2426.
38. Boyle E, O'Roak B, Martin B, et al. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics 2014; 30:2670–2672.
39. Coolbaugh-Murphy M, Xu J, Ramagli L, et al. Microsatellite instability in the peripheral blood leukocytes of HNPCC patients. Hum Mutat 2010;31:317–324.
40. Kansikas M, Kasela M, Kantelinen J, et al. Assessing how reduced expression levels of the mismatch repair genes MLH1, MSH2, and MSH6 affect repair efficiency. Hum Mutat 2014;35:1123–1127.
41. Kasela M, Nyström M, Kansikas M. PMS2 expression decrease causes severe problems in mismatch repair. Hum Mutat 2019;40:904–907.
42. You J, Buhard O, Ligtenberg M, et al. Tumours with loss of MSH6 expression are MSI-H when screened with a pentaplex of five mononucleotide repeats. Br J Cancer 2010;103:1840–1845.
43. Coolbaugh-Murphy M, Xu J, Ramagli L, et al. Microsatellite instability (MSI) increases with age in normal somatic cells. Mech Ageing Dev 2005;126:1051–1059.
44. Maruvka Y, Mouw K, Karlic R, et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. Nat Biotechnol 2017;35:951–959.
45. Zou X, Koh GCC, Nanda AS, et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. Nat Cancer 2021;2:643–657.
46. Suerink M, van der Klift HM, Ten Broeke SW, et al. The effect of genotypes and parent of origin on cancer risk and age of cancer development in PMS2 mutation carriers. Genet Med 2016;18:405–409.
47. Ryan NAJ, Morris J, Green K, et al. Association of mismatch repair mutation with age at cancer onset in Lynch syndrome: implications for stratified surveillance strategies. JAMA Oncol 2017;3:1702–1706.
48. Peltomäki P. Update on Lynch syndrome genomics. Fam Cancer 2016;15:385–393.
49. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science 2015;347:78–81.
50. Hao D, Wang L, Di LJ. Distinct mutation accumulation rates among tissues determine the variation in cancer risk. Sci Rep 2016;6:19458.
51. **Robinson PS, Coorens THH, Palles C**, et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. Nat Genet 2021;53:1434–1442.
52. Robinson PS, Thomas LE, Abascal F, et al. Inherited MUTYH mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells. Nat Commun 2022;13:3949.
53. **Sun S, Brazhnik K**, Lee M, et al. Single-cell analysis of somatic mutation burden in mammary epithelial cells of pathogenic BRCA1/2 mutation carriers. J Clin Invest 2022:132.
54. Gupta D, Heinen CD. The mismatch repair-dependent DNA damage response: mechanisms and implications. DNA Repair (Amst) 2019;78:60–69.
55. Variation in the risk of colorectal cancer in families with Lynch syndrome: a retrospective cohort study. Lancet Oncol 2021;22:1014–1022.

Author names in bold designate shared co-first authorship.

Received September 8, 2022. Accepted December 12, 2022.

Correspondence
Address correspondence to: Richard Gallon, PhD, MBioch, Cancer Prevention Research Group, Translational and Clinical Research Institute, Faculty of

Medical Sciences, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, NE1 3BZ, United Kingdom. e-mail: richard.gallon@newcastle.ac.uk.

Acknowledgments
The authors thank all of the patients and their families who provided samples for this study. The authors thank the Genomics Core Facility and Bioinformatics Support Unit, Newcastle University, Newcastle upon Tyne, United Kingdom, for their support of genome and amplicon sequencing, and genome sequence analysis, as well as the research group of Prof Joris Veltman, Newcastle University, Newcastle upon Tyne, United Kingdom, for providing a panel of control blood whole genome sequencing data for variant calling. The authors thank Cancer Research UK for their support through the Aspirin for Cancer Prevention (AsCaP) (C569/A24991) and CaPP (C1297/A15394) groups, and The Barbour Foundation (UK charity 328081). Collection of clinical data and biological samples for French patients was supported by la Fondation Gustave Roussy campaign: Guérir Le Cancer de l'Enfant au 21ème siècle. The study received nonfinancial support from the European Reference Network on genetic tumour risk syndromes (ERN GENTURIS) - Project ID No 739547. ERN GENTURIS is partly co-funded by the European Union within the framework of the Third Health Programme "ERN-2016—Framework Partnership Agreement 2017–2021". The authors thank the AsCaP steering committee members for their support: Prof Jack Cuzick, Queen Mary University of London (chair), Prof Frances Balkwill, Queen Mary University of London, Prof Tim Bishop, University of Leeds, Prof Sir John Burn, Newcastle University, Prof Andrew T. Chan, Harvard School of Medicine, Dr Colin Crooks, University of Nottingham, Prof Chris Hawkey, University of Nottingham, Prof Ruth Langley, University College London, Ms Mairead McKenzie, Independent Cancer Patients' Voice, Dr Belinda Nedjai, Queen Mary University of London, Prof Paola Patrignani, Università G. d'Annunzio di Chieti-Pescara, Prof Carlo Patrono, Catholic University of the Sacred Heart, Rome, Dr Bianca Rocca, Catholic University of the Sacred Heart, Rome, and Dr Samuel Smith, University of Leeds. John Burn is a National Institute for Health and Care Research (NIHR) Senior Investigator and thanks the NIHR for their support. D. Gareth Evans is supported by the NIHR Manchester Biomedical Research Centre (IS-BRC-1215-20007).

CRediT Authorship Contributions
Richard Gallon, PhD (Conceptualization: Equal; Data curation: Lead; Formal analysis: Equal; Funding acquisition: Supporting; Investigation: Equal; Methodology: Equal; Software: Equal; Supervision: Equal; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Rachel Phelps, MSc (Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Supporting).

Christine Hayes, BSc (Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Supporting).

Laurence Brugieres, MD (Data curation: Supporting; Resources: Equal; Writing – review & editing: Supporting).

Léa Guerrini-Rousseau, MD (Data curation: Supporting; Resources: Equal; Writing – review & editing: Supporting).

Chrystelle Colas, MD, PhD (Data curation: Supporting; Resources: Equal; Writing – review & editing: Supporting).

Martine Muleris, HDR (Data curation: Supporting; Resources: Equal; Writing – review &editing: Supporting).

Neil A. J. Ryan, MBChB, PhD (Resources: Equal; Writing – review & editing: Supporting).

D. Gareth Evans, MD (Resources: Equal; Writing – review & editing: Supporting).

Hannah Grice, BSc (Investigation: Supporting; Writing – review & editing: Supporting).

Emily Jessop, BSc (Investigation: Supporting; Writing – review & editing: Supporting).

Annabel Kunzemann-Martinez, MSc (Investigation: Supporting; Writing – review & editing: Supporting).

Lilla Marshall, BSc (Investigation: Supporting; Writing – review & editing: Supporting).

Esther Schamschula, PhD (Data curation: Supporting; Formal analysis: Supporting; Writing – review & editing: Supporting).

Klaus Oberhuber, Dipl.MTA (Data curation: Supporting; Resources: Supporting; Writing– review & editing: Supporting).

Amedeo A. Azizi, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Hagit Baris Feldman, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Andreas Beilken, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Nina Brauer, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Triantafyllia Brozou, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Karin Dahan, MD, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Ugur Demirsoy, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Benoît Florkin, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

William Foulkes, MBBS, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Danuta Januszkiewicz-Lewandowska, MD, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Kristi J. Jones, MD, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Christian P. Kratz, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Stephan Lobitz, MD, MSc (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Julia Meade, MD (Data curation: Supporting; Resources: Supporting; Writing – review& editing: Supporting).

Michaela Nathrath, MD (Data curation: Supporting; Resources: Supporting; Writing –review & editing: Supporting).

Hans-Jürgen Pander, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Claudia Perne, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Iman Ragab, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Tim Ripperger, MD, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Thorsten Rosenbaum, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Daniel Rueda, PhD (Data curation: Supporting; Resources: Supporting; Writing –review & editing: Supporting).

Tomasz Sarosiek, MD, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Astrid Sehested, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Isabel Spier, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Manon Suerink, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Stefanie-Yvonne Zimmermann, MD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Johannes Zschocke, MD, PhD (Data curation: Supporting; Resources: Supporting; Writing – review & editing: Supporting).

Gillian M. Borthwick, PhD (Data curation: Supporting; Funding acquisition: Supporting; Project administration: Equal; Resources: Supporting; Writing – review & editing: Supporting).

Katharina Wimmer, PhD (Data curation: Supporting; Formal analysis: Supporting; Resources: Equal; Supervision: Equal; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

John Burn, MD (Conceptualization: Equal; Data curation: Supporting; Formal analysis: Supporting; Funding acquisition: Lead; Project administration: Supporting; Resources: Equal; Supervision: Equal; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

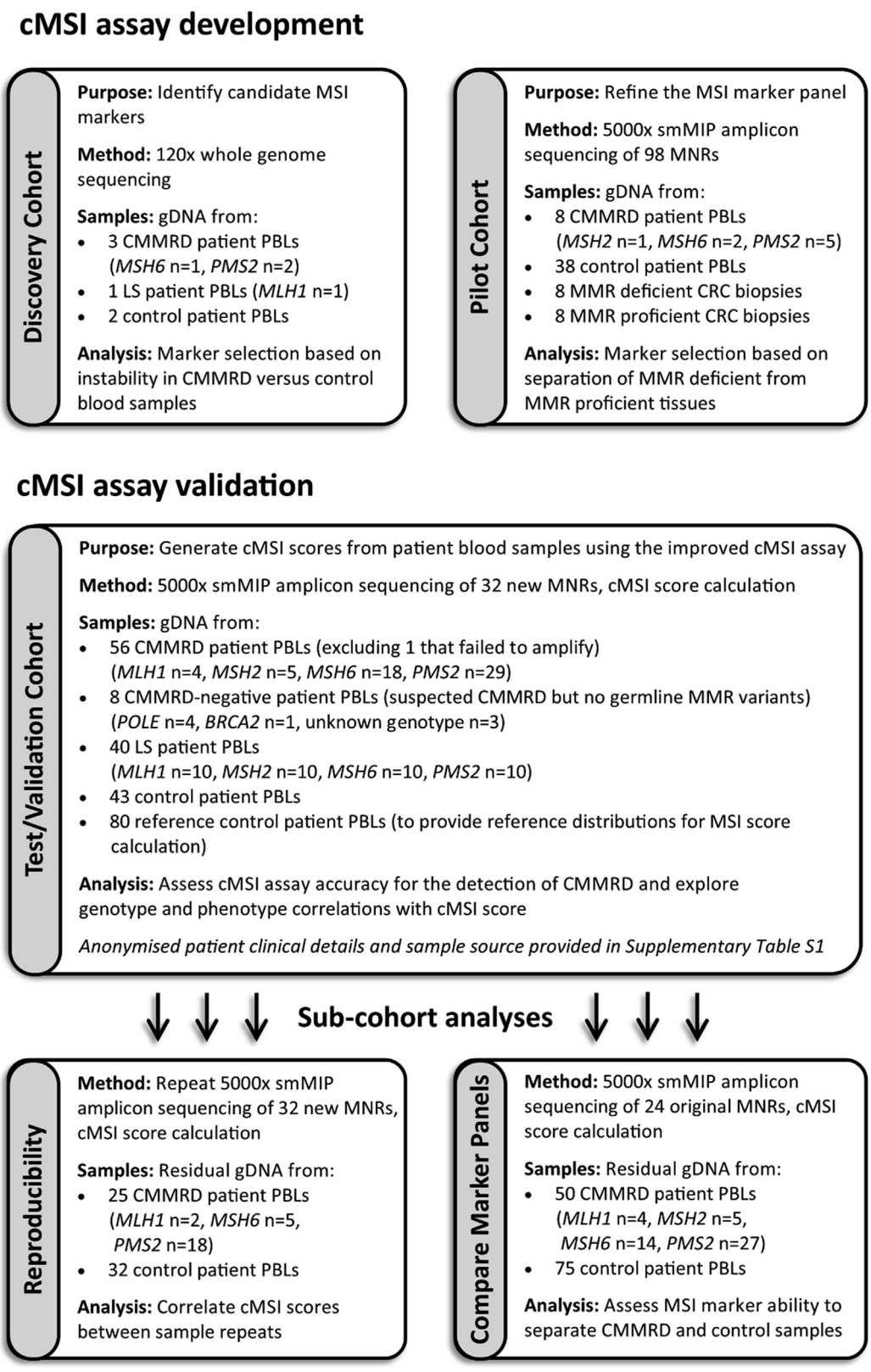
Michael S. Jackson, PhD (Conceptualization: Equal; Data curation: Supporting; Formal analysis: Equal; Funding acquisition: Supporting; Investigation: Equal; Methodology: Equal; Project administration: Equal; Supervision: Equal; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

Mauro Santibanez-Koref, MD (Conceptualization: Equal; Data curation: Supporting; Formal analysis: Equal; Funding acquisition: Supporting; Investigation: Equal; Methodology: Equal; Project administration: Equal; Software: Equal; Supervision: Equal; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

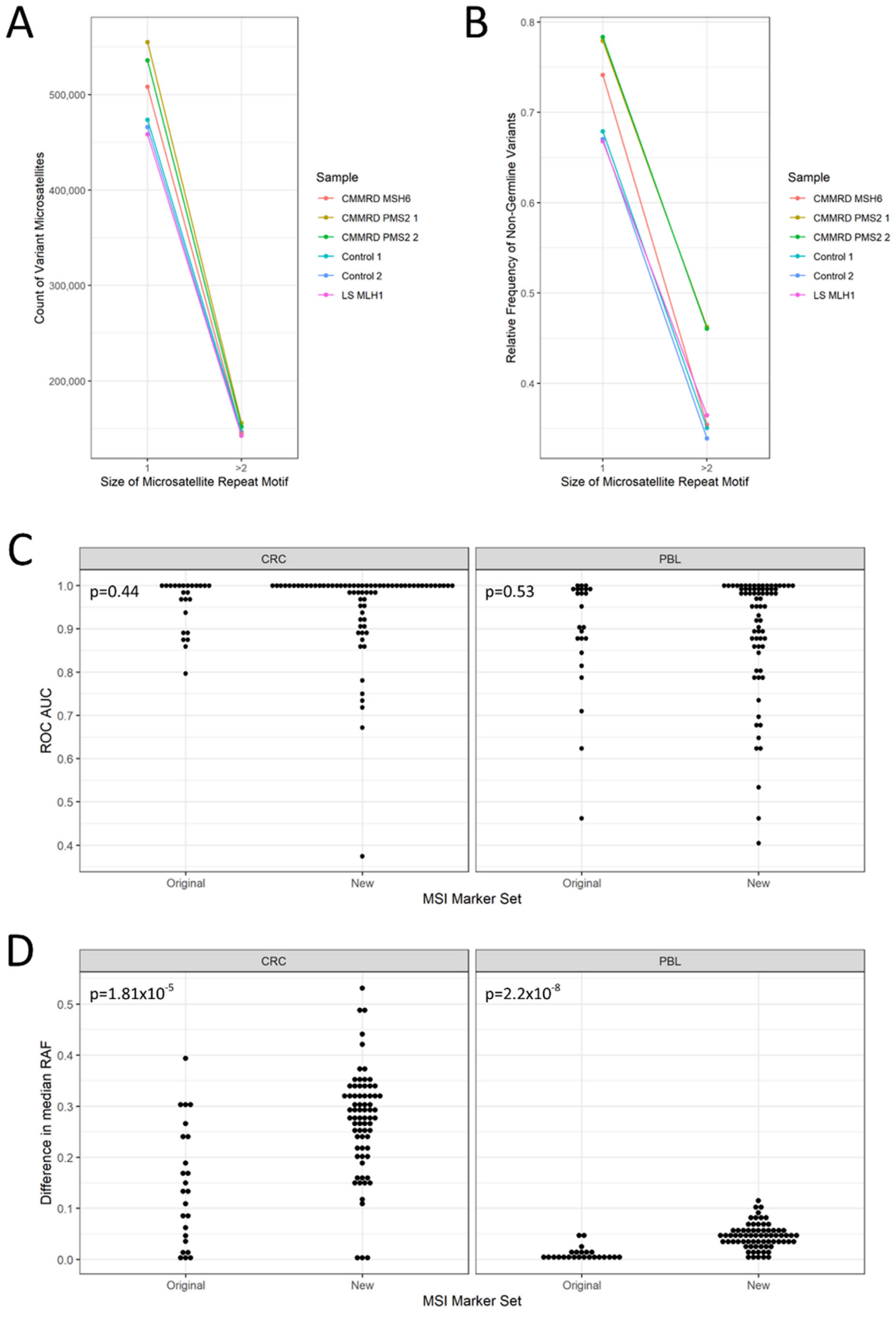
Conflicts of interest
These authors disclose the following: Richard Gallon, John Burn, Michael S. Jackson, and Mauro Santibanez-Koref are named inventors on patents covering the microsatellite instability markers analyzed: WO/2018/037231 (published March 1, 2018), WO/2021/019197 (published February 4, 2021), and GB2114136.1 (filed October 1, 2021). The remaining authors disclose no conflicts.

Funding
This study was funded through the Aspirin for Cancer Prevention (AsCaP) group, Cancer Research UK Grant Code: C569/A24991. Cancer Research UK had no role in study design, analysis, or interpretation. The AsCaP group is led by its senior executive board: Prof J. Burn, Prof A. T. Chan, Prof J. Cuzick, Dr B. Nedjai, and Prof Ruth Langley.

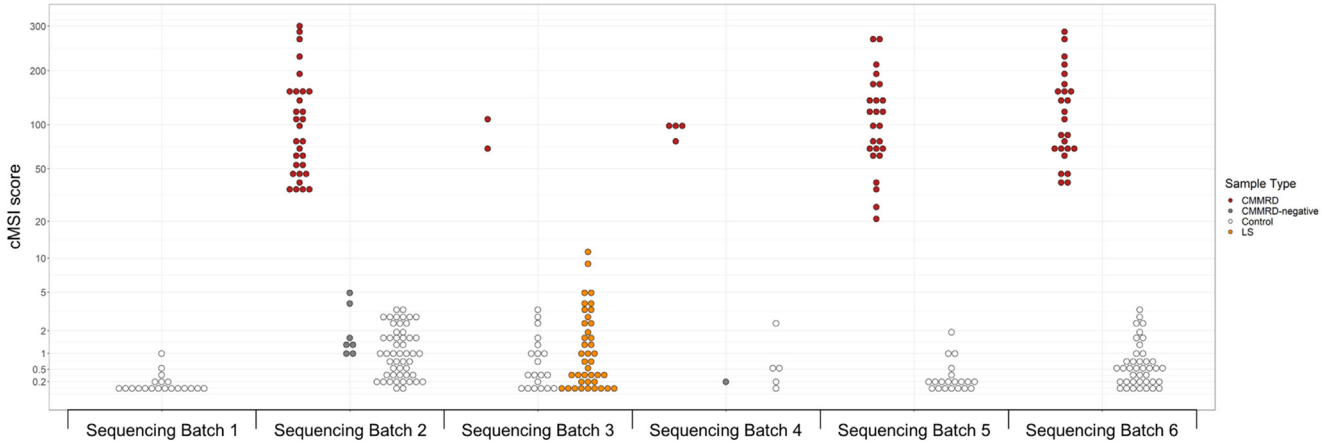
Data Availability
Genome sequence BAM and amplicon sequence FASTQ files are available from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) using Study IDs PRJEB39601 and PRJEB53321, respectively.



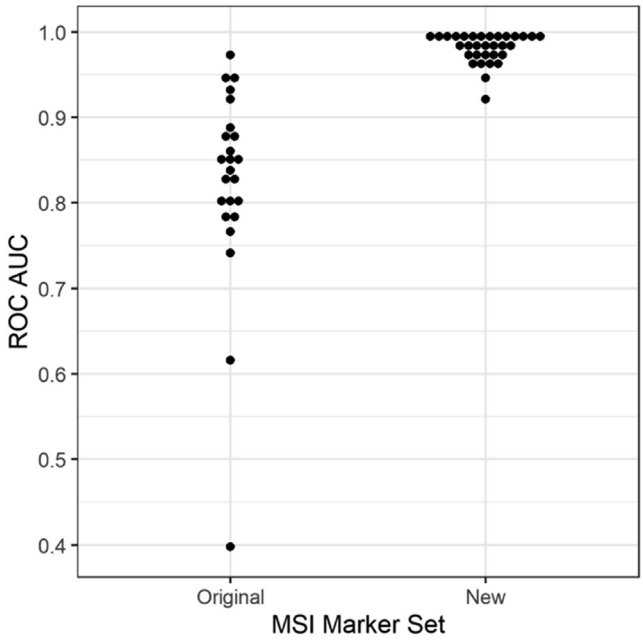
Supplementary Figure 1. Cohort descriptions for cMSI assay development and validation.


A
Count of Variant Microsatellites
500,000
400,000
300,000
200,000
1
>2
Size of Microsatellite Repeat Motif
Sample
CMMRD MSH6
CMMRD PMS2 1
CMMRD PMS2 2
Control 1
Control 2
LS MLH1
B
Relative Frequency of Non-Germline Variants
0.8
0.7
0.6
0.5
0.4
1
>2
Size of Microsatellite Repeat Motif
Sample
CMMRD MSH6
CMMRD PMS2 1
CMMRD PMS2 2
Control 1
Control 2
LS MLH1
C
CRC
PBL
p=0.44
p=0.53
ROC AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Original
New
Original
New
MSI Marker Set
D
CRC
PBL
p=1.81x10-5
p=2.2x10-8
Difference in median RAF
0.5
0.4
0.3
0.2
0.1
0.0
Original
New
Original
New
MSI Marker Set

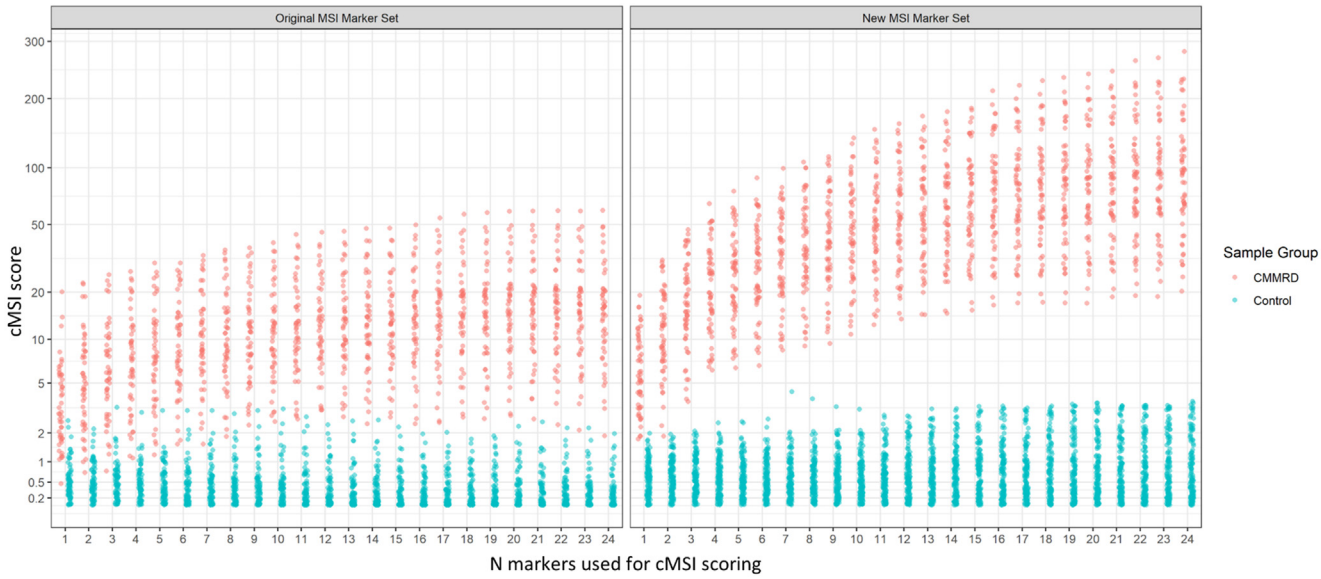
Supplementary Figure 2. Two *PMS2*- and 1 *MSH6*-associated CMMRD, 1 *MLH1*-associated LS, and 2 control blood samples were whole genome sequenced. There was a 1.13- to 1.21-fold and 1.07- to 1.11-fold increase, respectively, in the frequency of MNR variants in *PMS2*-associated and *MSH6*-associated CMMRD samples relative to control and LS samples. However, a consistent increase in variants of longer motif microsatellites was observed in the *PMS2*-associated CMMRD samples only (1.03- to 1.09-fold), and not the MSH6-associated CMMRD samples (0.99- to 1.02-fold) relative to LS and control samples (*A*). These variants include PCR error, sequencing error, germline variants, and somatic variants. To better assess the somatic signal, probable germline variants were identified, and the relative frequency of nongermline variants was assessed. In both *PMS2*-associated and *MSH6*-associated CMMRD samples, there was an increase in the relative frequency of nongermline MNR variants compared with LS and control samples (1.15- to 1.17-fold and 1.09- to 1.11-fold, respectively). An increase in relative frequency of nongermline variants in longer motif microsatellites was observed in the *PMS2*-associated CMMRD samples only (1.26- to 1.36-fold) and not the *MSH6*-associated CMMRD samples (0.97- to 1.04-fold) relative to LS and control samples (*B*). A panel of 98 MSI markers was selected from the WGS data and were further refined by their ability to discriminate between MMR-deficient and MMR-proficient tissues using smMIP amplicon sequencing of a pilot cohort of 8 CMMRD and 38 control blood samples, as well as 8 MMR-deficient and 8 MMR-proficient formalin-fixed, paraffin-embedded CRCs (see the main article and Figure 1). Twenty-seven of the 98 new MSI markers were excluded, as >10% of the pilot PBL samples had an RAF <0.75, indicative of a germline length variant. There was no significant difference in the ROC AUC values based on microsatellite RAF between the remaining 71 new and 24 original MSI markers to detect MMR deficiency in either pilot CRCs ($P = .44$) or pilot PBLs ($P = .53$) (*C*). However, the difference between the median RAFs of MMR-deficient and MMR-proficient samples was significantly greater for the new markers in both CRCs ($P = 1.8 \times 10^{-5}$) and PBLs ($P = 2.2 \times 10^{-8}$) (*D*).



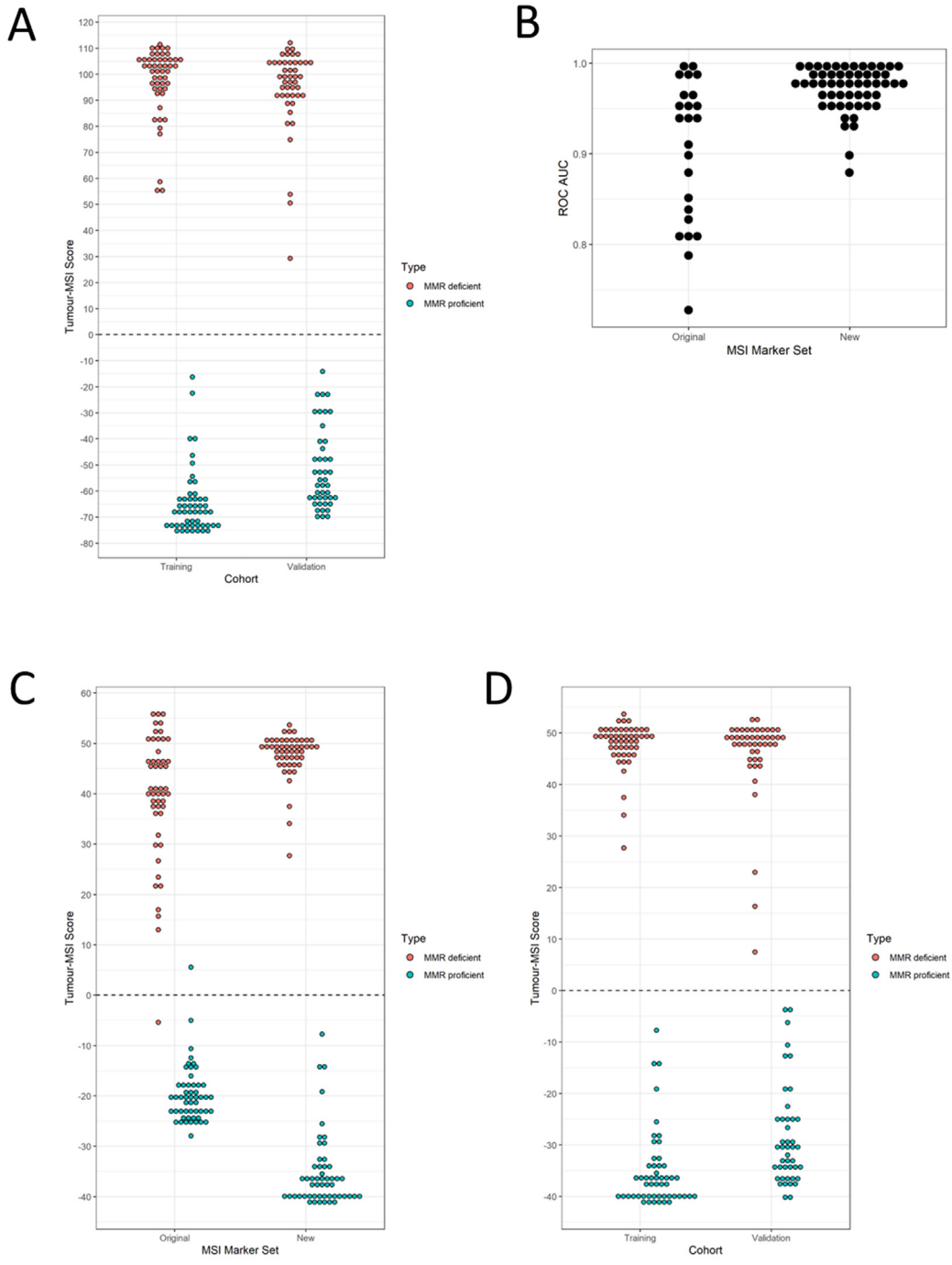
Supplementary Figure 3. The cMSI scores of blood samples by sequencing batch. Note that data for repeat amplification and sequencing of samples are shown.

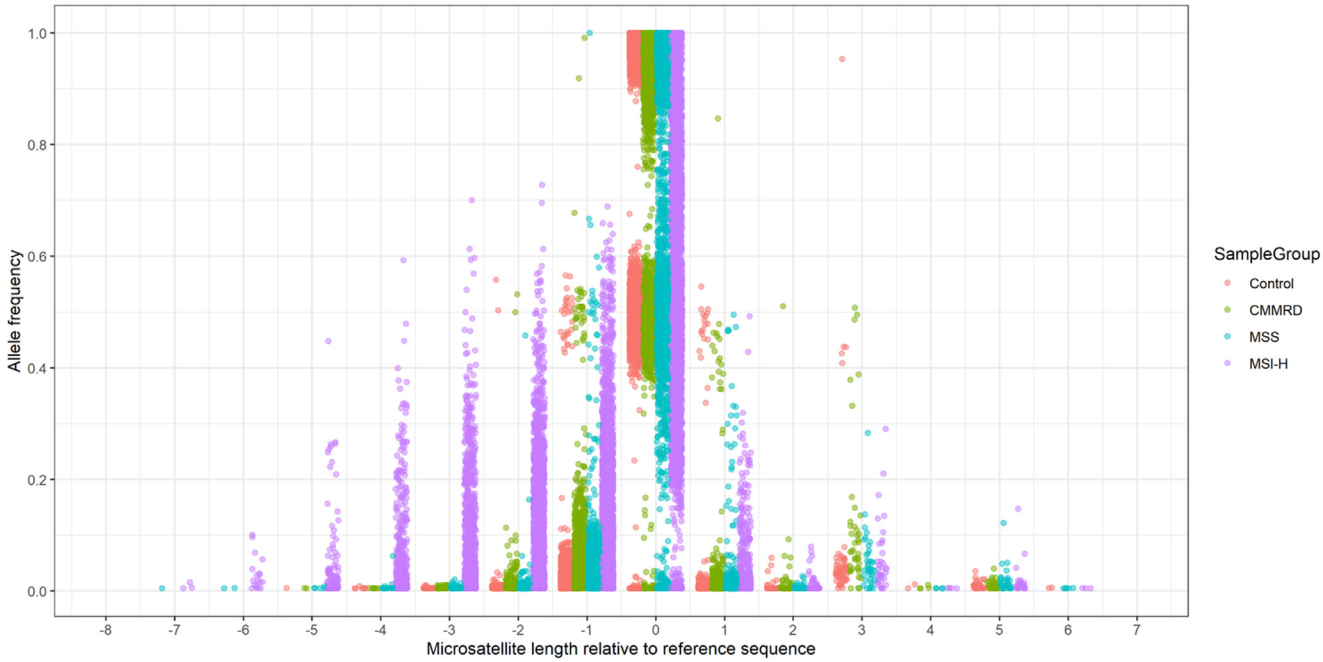


Supplementary Figure 4. The ROC AUC values calculated from the ability of each MSI marker to separate CMMRD from control PBL DNA samples using microsatellite RAF comparing new and original marker sets.



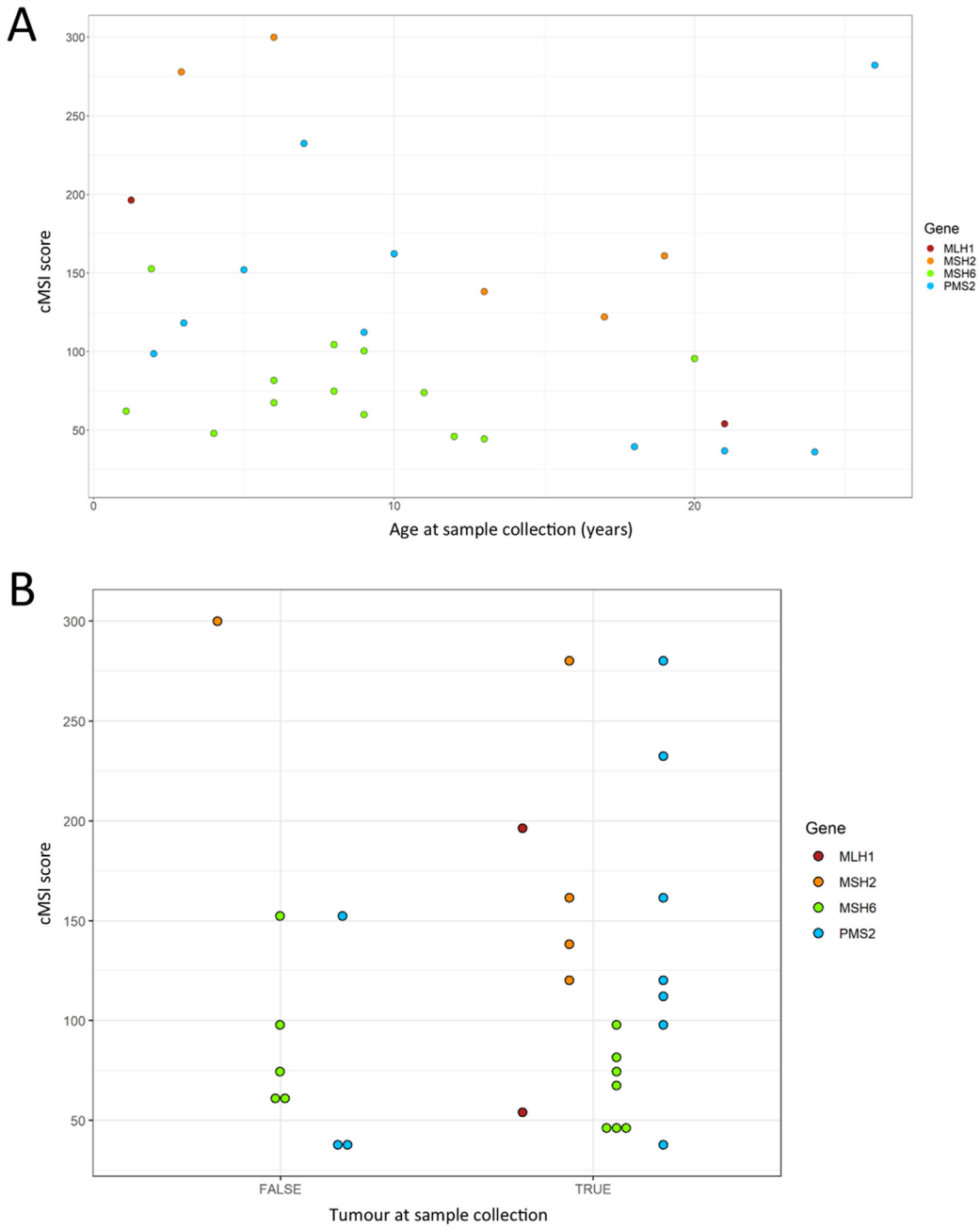
Supplementary Figure 5. The cMSI scores of blood samples using reduced panels of the most discriminatory number of the original MSI markers (*left panel*) and most discriminatory number of the new MSI markers (*right panel*).


A
Tumour-MSI Score
Cohort
Training
Validation
Type
MMR deficient
MMR proficient
B
ROC AUC
MSI Marker Set
Original
New
C
Tumour-MSI Score
MSI Marker Set
Original
New
Type
MMR deficient
MMR proficient
D
Tumour-MSI Score
Cohort
Training
Validation
Type
MMR deficient
MMR proficient



Supplementary Figure 7. The microsatellite allele length and allele frequency distribution of the 24 original and 32 new MSI markers in 75 control blood samples, 50 CMMRD blood samples, 52 microsatellite stable (MSS) CRCs, and 50 MSI-high (MSI-H) CRCs, for which sequence data from both marker sets were available.

Supplementary Figure 6. As a further test of diagnostic utility of the new MSI markers, a larger panel of 54 was selected from the 98 MNRs analyzed in the pilot cohort based on <5% germline variant frequency in the PBL samples and visual inspection of the microsatellite allele distributions in the CRC samples. The 32 MNRs of the cMSI assay described in the main article were also included (Supplementary Table 2). Selection was not as stringent as for the cMSI marker panel to provide a larger, exploratory marker set to facilitate comparisons with the original MSI markers and prime future research. Also, a larger panel of MSI markers could be used, as we have shown previously that smSequences provide no benefit to CRC MSI classification (Gallon et al[27]). Therefore, lower read depths of 3000× can be used, and hence more MSI markers assessed for equivalent cost. The 54 new MSI marker panel was smMIP-amplified and sequenced in 192 CRCs of known MSI status using the MSI Analysis System, version 1.2 (Promega) as a reference test. Custom R scripts were used to extract microsatellite variants from reads.[30] The microsatellite deletion frequencies and allelic bias (if a heterozygous neighboring *single nucleotide polymorphism* was available to discriminate between paternal and maternal alleles) in sequence reads generated from a training cohort of 50 MSI-high (MSI-H) and 52 MSS CRCs were used to train a naïve Bayesian classifier according to Redford et al.[30] The remaining 90 CRCs (46 MSI-H, 44 MSS) formed the validation cohort. A tumor MSI score was generated for each sample using the trained classifier. Tumor MSI scores >0 indicate a higher probability that the sample is MMR-deficient than MMR-proficient, and the inverse for scores <0. Tumor MSI scoring achieved 100% sensitivity (50 of 50; 95% CI, 92.9%–100.0%) and 100% specificity (52 of 52; 95% CI, 93.2%–100.0%) in the training cohort and 100% sensitivity (46 of 46; 95% CI, 92.3%–100.0%) and 100% specificity (44 of 44; 95% CI, 92.0%–100.0%) in the validation cohort (*A*). Training cohort samples were also analyzed by the original MSI markers. Each marker's ability to separate MMR-deficient and MMR-proficient CRCs by microsatellite RAF in the training cohort data was assessed. RAF ROC AUCs of the new MSI markers were greater than the RAF ROC AUCs of the originals ($P = 8.31 \times 10^{-5}$) (*B*). To compare tumor MSI classification by marker set with an equivalent number of MSI markers, the new MSI markers were ranked by ROC AUC and the most discriminatory 24 were used to re-score the training cohort samples, achieving 100% accuracy as for the full 54-marker panel (*C*). Scoring of the training cohort by the original MSI markers misclassified 2 CRCs—1 MMR-deficient and 1 MMR-proficient—achieving 98% sensitivity (49 of 50; 95% CI, 89.4%–99.9%) and 98% specificity (51 of 52; 95% CI, 89.7%–99.9%) (*C*). MMR-deficient CRCs had more positive tumor MSI scores when using new vs original MSI markers ($P = 3.16 \times 10^{-4}$) and MMR-proficient CRCs had more negative scores when using new vs original MSI markers ($P = 2.23 \times 10^{-14}$), demonstrating a greater score separation with the new MSI markers. The most discriminatory 24 new MSI markers also classified the validation cohort with 100% accuracy as for the full 54-marker panel (*D*).



Supplementary Figure 8. The cMSI score and age of sample collection of 30 patients with CMMRD when this was known (*A*). The cMSI score by whether the patient had a tumor at the time of blood sample collection for 27 patients with CMMRD when this was known (*B*).