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

Richard Gallon, Rachel Phelps, Christine Hayes, Laurence Brugieres, Léa Guerrini-Rousseau, Chrystelle Colas, Martine Muleris, Neil A.J. Ryan, D. Gareth Evans, Hannah Grice, et al.

Title

Constitutional microsatellite instability, genotype, and phenotype correlations in Constitutional Mismatch Repair Deficiency

Short title

A constitutional microsatellite instability test

Authors

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], Mauro Santibanez-Koref[10]

1 Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK
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, France / Université de Paris, Paris, France
6 Sorbonne Université, Inserm, Centre de Recherche Saint-Antoine, CRSA, Paris, France
7 The Academic Women’s Health Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
8 Department of Gynaecology Oncology, Royal Infirmary of Edinburgh, Edinburgh, UK
9 Division of Evolution, Infection and Genomics, University of Manchester, Manchester, UK
10 Biosciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK
11 Centre for Inflammation and Tissue Repair, University College London, London, UK
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 (IPG), 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, Australia
27 University of Sydney School of Medicine, Sydney, Australia
28 GK 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, USA
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
41 Department of Pediatric Hematology and Oncology, Children's Hospital, University Hospital, Frankfurt, Germany

Grant support

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, Prof. Ruth Langley.

Abbreviations

AA: amino acid

APC: adenomatous polyposis coli
AUC: area under curve
bp: base pair
CMMRD: constitutional mismatch repair deficiency
cMSI: constitutional microsatellite instability
CNV: copy number variant
CRC: colorectal cancer
CRISPR: clustered regularly interspaced short palindromic repeats
DNA: deoxyribose nucleic acid
gDNA: genomic deoxyribose nucleic acid
indel: insertion-deletion mutation
LS: Lynch syndrome
MLH1: MutL homolog 1
MMR: mismatch repair
MNR: mononucleotide repeat
MSH2: MutS homolog 2
MSH3: MutS homolog 3
MSH6: MutS homolog 6
MSI: microsatellite instability
NF1: neurofibromatosis type 1
PBL: peripheral blood leukocyte
PCR: polymerase chain reaction
PMS2: post-meiotic segregation-1 homolog 2
PMS2CL: PMS2 C-terminal like pseudogene
PV: pathogenic variant
RAF: reference allele frequency
ROC: receiver operator characteristic
smMIP: single molecule molecular inversion probe
smSequence: single molecule sequence
SPRED1: sprouty related EVH1 domain containing 1
VUS: variant of unknown significance
WGS: whole genome sequencing

Correspondence

richard.gallon@newcastle.ac.uk
Dr. 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, UK
+44 (0) 7530 949 298

Disclosures

R. Gallon, J. Burn, M. S. Jackson, and M. Santibanez-Koref are named inventors on patents covering the microsatellite instability markers analysed: WO/2018/037231 (published March 1, 2018), WO/2021/019197 (published February 4, 2021), and GB2114136.1 (filed October 1, 2021). T. Rosenbaum received consulting fees from AstraZeneca and Alexion. The other authors declare no conflicts of interest.

Author Contributions

Study concept and design: R. Gallon, J. Burn, M. S. Jackson, M. Santibanez-Koref.
Provision of clinical samples and data: L. Brugieres, L. Guerrini-Rousseau, C. Colas, M. Muleris, N. A. J. Ryan, D. G. Evans, K. Oberhuber, A. A. Azizi, H. Baris Feldman, A. Beilken, T. Brozou, K. Dahan, U. Demirsoy, B. Florkin, W. Foulkes, T. Imschweiler, D. Januszkiewicz-Lewandowska, K. J. Jones, C. Kratz, S. Lobitz, J. Meade, M. Nathrath, H-J Pander, C. Perne, I. Ragab, T. Ripperger, T. Rosenbaum, D. Rueda, A. Sehested, I. Spier, M. Suerink, S. Tomasz, S-Y. Zimmermann, J. Zschocke, G. Borthwick, K. Wimmer.
Data curation: R. Gallon, R. Phelps, C. Hayes, E. Schamschula, K. Wimmer.
Data generation: R. Gallon, R. Phelps, C. Hayes, H. Grice, E. Jessop, A. Kunzemann-Martinez, L. Marshall.
Data analysis: R.Gallon. R. Phelps, M. S. Jackson, M. Santibanez-Koref.
Data interpretation: R.Gallon, K. Wimmer, J. Burn, M. S. Jackson, M. Santibanez-Koref.
Manuscript writing: R.Gallon. K. Wimmer, M. S. Jackson, M. Santibanez-Koref.
Manuscript review and approval: all authors.

Data Transparency Statement

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.

Abstract

Background & Aims: Constitutional mismatch repair deficiency (CMMRD) is a rare recessive childhood cancer predisposition syndrome caused by germline mismatch repair (MMR) 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, one Lynch syndrome (LS), and two control blood samples were genome sequenced to >120x depth. A pilot cohort of eight CMMRD and 38 control blood samples, and a blinded cohort of 56 CMMRD, eight suspected CMMRD, 40 LS, and 43 control blood samples were amplicon sequenced to 5000x depth. Sample cMSI score was calculated using a published method comparing microsatellite reference allele frequencies to 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 tumour MSI marker panel. Lower cMSI scores were found in CMMRD patients with MSH6 deficiency and patients with at least one MMR missense variant, whilst patients with biallelic truncating/copy number variants had higher scores. cMSI score did not correlate with age at first tumour.

Conclusions: We present a cheap and scalable cMSI assay that enhances CMMRD detection relative to existing methods. cMSI score is associated with MMR 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

Word Count

6998

Manuscript

Introduction

The DNA mismatch repair (MMR) system is conserved across all three domains of life. It mediates the repair of base-to-base mismatches and small insertion-deletion loops generated during DNA replication whilst signalling 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 one in four endometrial cancers and one in seven colorectal cancers (CRCs)[3,4]. MMR-deficient tumours are often hypermutated and display high levels of microsatellite instability (MSI), a molecular phenotype defined as the accumulation of insertion and deletion mutations (indels) in short tandem repeat sequences[5]. This elevated mutation rate has been proposed to drive tumorigenesis through secondary mutation of onco- and tumour suppressor genes[6-12].

Individuals with Lynch syndrome (LS) carry a germline pathogenic variant (PV) affecting one of the four principal MMR genes (*MLH1*, *MSH2*, *MSH6*, or *PMS2*) and have an increased life-time risk of adult-onset cancer, in particular CRC, endometrial cancer, and other tumours of the gastrointestinal and genitourinary tracts[13]. LS is one of the most common hereditary cancer predisposition syndromes, affecting approximately one 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 one per million[15]. The constitutional loss of MMR function in all tissues is associated with an exceptionally high cancer risk, with a median age of onset <10 years. This characteristically includes high grade brain tumours and haematological 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 localised 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 haematological malignancies were 1.77-fold more prevalent in the former (p=0.04) whereas brain tumours were 1.75-fold more frequent in the latter (p=0.01). 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 (VUS)[17] and variants in *PMS2*, the MMR gene affected in the majority of CMMRD patients[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 tumour material where 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 immortalisation and culture of patient primary lymphocytes[20]. CMMRD is also characterised by increased MSI in non-neoplastic tissues but PCR fragment length analysis traditionally used in tumours 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 analysed 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 analysed[21,23,24]. Whilst Chung et al[25] demonstrated that low pass whole genome sequencing (WGS) at 1x coverage also accurately detects CMMRD, these assays can require millions of sequence reads per sample[21,24,25], which may limit scalability for screening where 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 analysed[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 two 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 non-malignant) 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 tumours[23,30], which we hypothesised could limit its sensitivity for cMSI analysis. For example, tumours 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 mutations 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 over 50 CMMRD patients.

Materials and methods

Patient samples and ethical approval

Anonymised CMMRD peripheral blood leukocyte (PBL) genomic DNAs (gDNAs) were sourced from the Medical University of Innsbruck, Innsbruck (n=31), Austria, the University of Manchester, Manchester, UK (n=1), the Gustave Roussy Cancer Campus, Villejuif, France (n=9), the Institut Curie, Université de Recherche Paris Sciences et Lettres, Paris, France (n=4), and the Cancer Centre de Recherche Saint-Antoine, Sorbonne University, Paris, France (n=13). MMR variants were classified according to InSiGHT criteria v2.4 (https://www.insight-group.org/criteria/). For patients with one or more VUS, 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].

Anonymised 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).

Anonymised 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 NHS Foundation Trust, Newcastle-upon-Tyne, UK (n=50).

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

Anonymised CRC samples were sourced as excess diagnostic material from the Northern Genetics Service as 10µm FFPE tissue curls of resected tumours (n=192) or pre-extracted gDNAs from non-fixed endoscopic biopsies (n=16). FFPE CRC gDNAs were extracted using the GeneRead DNA FFPE Kit (QIAGEN).

Consent of the individual and/or their legally-responsible guardian for use of CMMRD, CMMRD-negative, LS, and control PBL samples in research was received by each contributing institution. MSI analysis of excess diagnostic control PBL and CRC samples was approved by the NHS 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 S1. PBL gDNA sample and patient details are given in Supplementary Table S1.

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 >120x 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 equalling 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 where 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 CMMRD patients 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 100bp and 160bp, and an 8N molecular barcode with 4N adjacent to both extension and ligation arms (Supplementary Table S2).

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 4nM using 10mM Tris-HCl, pH 8.5, and pooled into 4nM sequencing libraries. Sequencing libraries were sequenced using custom sequencing primers[23] on a MiSeq (Illumina) to a target depth of 5000x, 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 one 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 to 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 two 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 tumour-MSI score[30].

Statistical analyses and data availability

All analyses used R version 4.0.2 (https://www.r-project.org/). Comparisons of two sample groups used the Mann-Whitney test. Comparisons of more than two 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 Rho was used for variables where 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. Confidence intervals 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 MSI markers

Three CMMRD (two *PMS2*- and one *MSH6*-associated), one LS (*MLH1*-associated), and two control blood samples were whole genome sequenced (Supplementary Figure S1). An LS sample was included as highly sensitive MSI analysis and single-base-mismatch repair assays have previously detected reduced MMR function in blood and cell lines with one 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 S2A-B). To derive a novel marker panel for cMSI analysis, the WGS data were filtered for microsatellites displaying an increase in non-germline variant alleles in the CMMRD samples compared to the controls. This identified over 2000 loci of interest, the majority of which were 11-

16bp 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 to 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 three control samples. Of these, 91 smMIPs (capturing 98 MNRs) generated sufficient reads to be taken forward (Figure 1).

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 S1). All except seven control PBL samples had been previously analysed using the 24 tumour-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 proficient samples in both CRCs ($p=1.8x10^{-5}$) and PBLs ($p=2.2x10^{-8}$; Supplementary Figures S2C-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 S2).

New MSI markers enhance the detection of CMMRD

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 CMMRD patients, eight 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 analysed to investigate if 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 generated for each. Markers with low (<100) smSequence counts were observed in only four samples from the blinded cohort: Two had a single low count-marker, whilst the other two had <100 smSequences in ≥17 MSI markers with equivalent results upon repeat amplification and sequencing, suggesting poor sample quality. On un-blinding, these two poor quality samples were revealed to be CMMRD cases.

The cMSI score identified CMMRD with 100% sensitivity (56/56; 95% CI: 93.6-100.0%) and 100% specificity (171/171; 95% CI: 97.9-100.0%), including the two poor quality CMMRD samples. There was a clear separation of all CMMRD samples from control, LS, and CMMRD-negative samples (Figure 2A, Supplementary Table S1). cMSI score was associated with affected MMR gene ($p=1.2x10^{-3}$); patients with MSH6 deficiency had significantly lower cMSI scores than patients with MSH2 deficiency ($p=2.4x10^{-4}$) or PMS2 deficiency ($p=6.0x10^{-3}$), and a trend for lower scores than patients with MLH1 deficiency (p=0.05, multiple testing significance at $p<1.67x10^{-2}$). LS cMSI scores were not significantly different from controls (p=0.17), but it was notable that six 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=0.02), with two 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 LS patients, these were not analysed further. To assess cMSI assay reproducibility, residual DNA samples available from 25 CMMRD patients 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 2B). There was no significant correlation in cMSI score between control repeats (R=0.105, p=0.56), 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, $p=1.2x10^{-8}$, Supplementary Figure S3).

Fifty CMMRD and 75 control samples were also analysed using the original 24 MSI markers[23]. The new MSI markers had greater RAF-based ROC AUCs for CMMRD detection than the original set ($p=9.0x10^{-14}$, Supplementary Figure S4). The new MSI markers were longer (range 11-15bp versus 7-12bp, $p=1.9x10^{-7}$) and there was a positive correlation between marker length and ROC AUC (Rho=0.730, $p=1.8x10^{-10}$). However, comparison of markers of equivalent size (11-12bp) found higher ROC AUCs for the new markers than the original ($p=2.5x10^{-4}$, Figure 3A). The new MSI markers were ranked by RAF ROC AUC to separate CMMRD from control samples (Supplementary Table S2) and the most discriminatory 24 new MSI markers gave a large cMSI score separation of 15.3 between CMMRD and control samples, compared to the 0.1 cMSI score overlap when using the original 24 MSI markers (Figure 3B). Using only three new MSI markers gave 100% accurate CMMRD detection (Supplementary Figure S5). The new MSI markers also enhanced tumour-MSI classification of CRCs compared to the original set (Supplementary Figures S6A-D). Despite differences in variant allele frequencies and indel size between CRCs and blood (Supplementary Figure S7), MSI marker RAF ROC AUCs for the detection of MMR deficiency were correlated between the two tissue types (Rho=0.715, $p=9.0x10^{-5}$).

CMMRD cMSI-burden is associated with MMR variant but not age of tumour onset

There was a breadth of cMSI scores between CMMRD patients with deficiency of the same MMR gene suggesting potential genotype or phenotype correlations with cMSI-burden. Variants were labelled as one of three types, Truncating/CNV, Splicing, or Missense, according to their effect on protein sequence - truncating and intragenic copy number variants (CNVs; i.e. deletions or duplications of one or more exons) were grouped together due to their direct disruption of protein structure and/or expression (Supplementary Table S1). There were three exceptions: NM_000179.2(*MSH6*):c.2426_2428del was labelled as a single amino acid deletion (1AAdel), NM_000179.2(*MSH6*):c.1763_1771dup was labelled as a triple amino acid duplication (3AAdup), and NM_000535.5(*PMS2*):c.2002A>G was labelled 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* mRNA 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 ($p=3.0x10^{-3}$, Figure 4A). No increase in cMSI score had been observed in LS blood samples compared to 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, CMMRD patients with at least one missense variant (excluding those patients with NM_000535.5(*PMS2*):c.2002A>G due to its splicing effect) were compared to the rest of the cohort and were found to have significantly lower cMSI scores ($p=7.4x10^{-3}$, Figure 4A). Conversely, patients with bi-allelic truncating variants or intragenic CNVs had significantly higher cMSI scores than those without (p=0.02, Figure 4A). The frequency of mono- or bi-allelic missense variants and the frequency of bi-allelic truncating variants/intragenic CNVs were both equivalent between MMR genes (p=0.54, p=0.61, respectively) indicating these differences were not due to an over-representation of variant types in any one 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 eight CMMRD families were possible, together with ten pairwise comparisons between five 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 $p=2.9x10^{-4}$, Figure 4B).

A clinical history of tumour diagnoses was available for all CMMRD patients (n=56). Five patients had no cancer history, and for another the age of tumour diagnosis was unknown, meaning age of first cancer could be compared to cMSI score in 50 patients (Supplementary Table S1). cMSI score was not significantly correlated with age of first tumour overall (Rho=-0.154, p=0.29, Figure 5), or in subgroup analyses of MSH6-deficient patients (Rho=-0.342, p=0.20) and PMS2-deficient patients (Rho=-0.013, p=0.95). It is possible that cMSI-burden is associated with the onset of specific tumour types as there is evidence that both sporadic MMR-deficient and CMMRD-related brain and haematological malignancies have reduced MSI compared to cancers within the LS spectrum[3,21]. However, no significant correlation was found between cMSI score and the age of onset of brain tumours (Rho=-0.167, p=0.32), haematological malignancies (Rho=-0.285, p=0.27), or LS-associated tumours (Rho=-0.143, p=0.58). There was also no significant association of age of first tumour with affected MMR gene (p=0.48) or type of variant (p=0.38).

Other factors that might affect cMSI-burden include age at sample collection[39,43] and contaminating tumour cells or DNA. Age at sample collection was not significantly correlated with cMSI score among 30 CMMRD patients with data available (Rho=-0.310, p=0.10, Supplementary Figure 8A), but was correlated with age of first tumour (R=0.727, $p=3.9x10^{-5}$) 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=0.65) or in the 40 LS patients (p=0.28). For 27 CMMRD patients it was also known if a tumour was present at the time of sample collection; the cMSI scores of the 18 patients with a tumour were not significantly different to those without (p=0.50, Supplementary Figure 8B).

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 11bp and 15bp, which is equivalent to the most sensitive and specific A-homopolymers identified in The Cancer Genome Atlas tumour 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 11-12bp have significantly higher ROC AUCs than the original tumour-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 tumours. However, the original tumour-derived set had also been selected to be ≤12bp to minimise PCR and sequencing error, and to have a SNP within 30bp to allow the allelic-bias of microsatellite deletions to be used in tumour-MSI classification[30]. Therefore, different marker selection criteria preclude clear conclusions regarding whether specific microsatellites are more sensitive to mutation in MMR-deficient tumours compared to MMR-deficient blood.

Sequencing-based MSI analysis of non-neoplastic tissues to detect CMMRD has now been demonstrated with a variety of methods[21,23,24,25]. The smMIP amplicon-sequencing cMSI assay used here is relatively cheap, 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 v3 or v2 Micro kit, respectively. These costs could be reduced as only three MSI markers were required for separation of CMMRD from control samples (Supplementary Figure S5). The method is scalable from functional testing of a few samples to high throughput screening for CMMRD, as demonstrated in a study of >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 five sequencing runs (Supplementary Table S1). 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 eight MMR VUS (Supplementary Table S1) as pathogenic, at least in the context of CMMRD.

CMMRD patient cMSI scores were associated with genotype. Previously, Gonzalez-Acosta et al[24] reported a reduced cMSI-burden in *MSH6*- versus *MSH2*-associated CMMRD patients using an alternative amplicon sequencing assay. We have shown that this is also true for *MSH6*- versus *PMS2*-associated CMMRD and that there is a similar trend comparing *MSH6*- to *MLH1*-associated CMMRD. A reduced cMSI-burden of *MSH6*- compared to *PMS2*-associated CMMRD was also observed in our WGS data, and is consistent with WGS of CRISPR-Cas9-knockout cell lines, which showed a reduced indel frequency in MSH6- compared to MLH1-, MSH2-, or PMS2-deficient cells[45]. The redundancy for 1bp 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 copy number variants 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 if MMR missense variants are associated with reduced MSI in MMR-deficient tumours, and whether this has any association with clinical course.

No significant correlation of MMR genotype or cMSI score with age of first tumour was observed among the 56 CMMRD patients analysed. Wimmer et al[17] previously found differences in the incidence of CNS tumours and haematological malignancies, and the age of first tumour by affected MMR gene in CMMRD, but analysed 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 to those that retain expression. Ryan et al[47] similarly showed an association between truncating *MLH1* PVs and earlier onset of LS endometrial cancer. Otherwise, there is very limited data supporting an effect for type or position of MMR PVs on clinical phenotype in LS[48]. Therefore, whilst 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 of tumour 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 three 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 tumours compared to 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, very recently, 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, 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 of first tumour 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. Analysing 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 half of CMMRD families[17]. We found no evidence that cMSI score was influenced by presence of a tumour at the time of blood sampling in the CMMRD patients, but some CMMRD-negative and LS samples showed marginally increased cMSI scores. We could not analyse these further due to a lack of cancer data, but future exploration of the effect of contaminating MSI-H circulating tumour cells or DNA on cMSI analysis may be warranted.

In summary, we have analysed cMSI-burden in a relatively large cohort of CMMRD patients given the rarity of the syndrome, combining novel MSI markers and a simple amplicon-sequencing method to enhance CMMRD diagnostics. Our data show a MMR genotype-phenotype correlation with both gene affected and the type of variant influencing cMSI-burden, suggesting MMR genotype could also have implications for tumour 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, whilst cMSI score is a useful diagnostic biomarker, it likely cannot be used to stratify cancer risk in CMMRD as we initially hypothesised.

References

1. Kunkel T, Erie D. DNA mismatch repair. Annual Review of Biochemistry 2005;74:681-710.

2. Jiricny J. The multifaceted mismatch-repair system. Nature Reviews. Molecular Cell Biology 2006;7:335-46.

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.

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. Genetics in Medicine 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. Journal of Biological Chemistry 1995;270:22044-9.

7. Ionov Y, Yamamoto H, Krajewski S, et al. Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution. Proceedings of the National Academy of Sciences of the United States of America 2000;97:10872-7.

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 research 2002;62:2447-54.

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 Research 2004;64:7690-6.

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;31:1242-53.

11. Sekine S, Mori T, Ogawa R, et al. Mismatch repair deficiency commonly precedes adenoma formation in Lynch Syndrome-Associated colorectal tumorigenesis. Modern Pathology 2017;30:1144-51.

12. **Ahadova A**, **Gallon R**, Gebert J, et al. Three molecular pathways model colorectal carcinogenesis in Lynch syndrome. International Journal of 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. Genetics in medicine : official journal of the American College of Medical Genetics 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 Epidemiology, Biomarkers and Prevention 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. Clinical Genetics 2017;91:507-19.

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;51:355-65.

18. Shiran S, Ben-Sira L, Elhasid R, et al. Multiple Brain Developmental Venous Anomalies as a Marker for Constitutional Mismatch Repair Deficiency Syndrome. American Journal of Neuroradiology 2018.

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;31:578-87.

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;149:1017-29.

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. Human Mutation 2013;34:847-52.

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. Human Mutation 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 2022:Jco2102873.

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. Journal of Medical Genetics 2015;52:348-52.

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. Gastrointestinal Endoscopy 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;144:646-74.

32. Loeb LA. Human Cancers Express a Mutator Phenotype: Hypothesis, Origin, and Consequences. Cancer Res 2016;76:2057-9.

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;25:1754-60.

35. **Li H**, **Handsaker B**, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.

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;20:1297-303.

37. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24-6.

38. Boyle E, O'Roak B, Martin B, et al. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics 2014;30:2670-2.

39. Coolbaugh-Murphy M, Xu J, Ramagli L, et al. Microsatellite Instability in the Peripheral Blood Leukocytes of HNPCC Patients. Human mutation 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;35:1123-7.

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. British Journal of Cancer 2010;103:1840-5.

43. Coolbaugh-Murphy M, Xu J, Ramagli L, et al. Microsatellite instability (MSI) increases with age in normal somatic cells. Mechanisms of Ageing and Development 2005;126:1051-9.

44. Maruvka Y, Mouw K, Karlic R, et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. Nature Biotechnology 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. Nature 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;18:405-9.

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;15:385-93.

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.

Figures

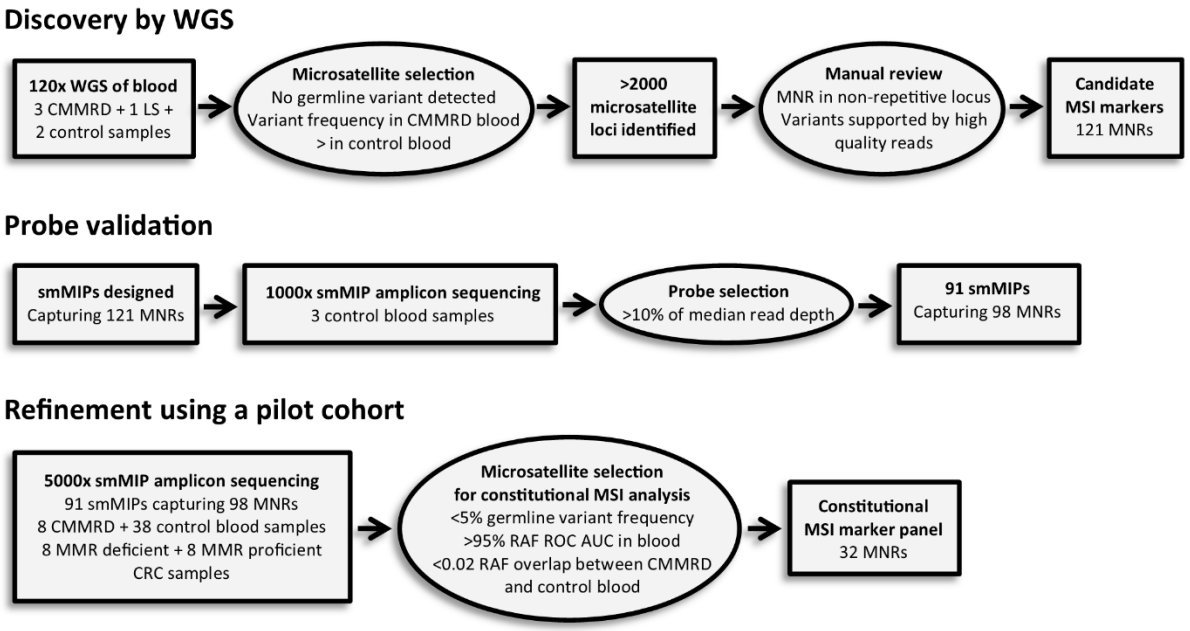

Figure 1: Flow chart of MSI marker selection. CMMRD: constitutional mismatch repair deficiency; CRC: colorectal cancer; LS: Lynch syndrome; MMR: mismatch repair; MNR: mononucleotide repeat; MSI: microsatellite instability; RAF: reference allele frequency; ROC AUC: receiver operator characteristic area under curve; smMIP: single molecule molecular inversion probe; WGS: whole genome sequencing.

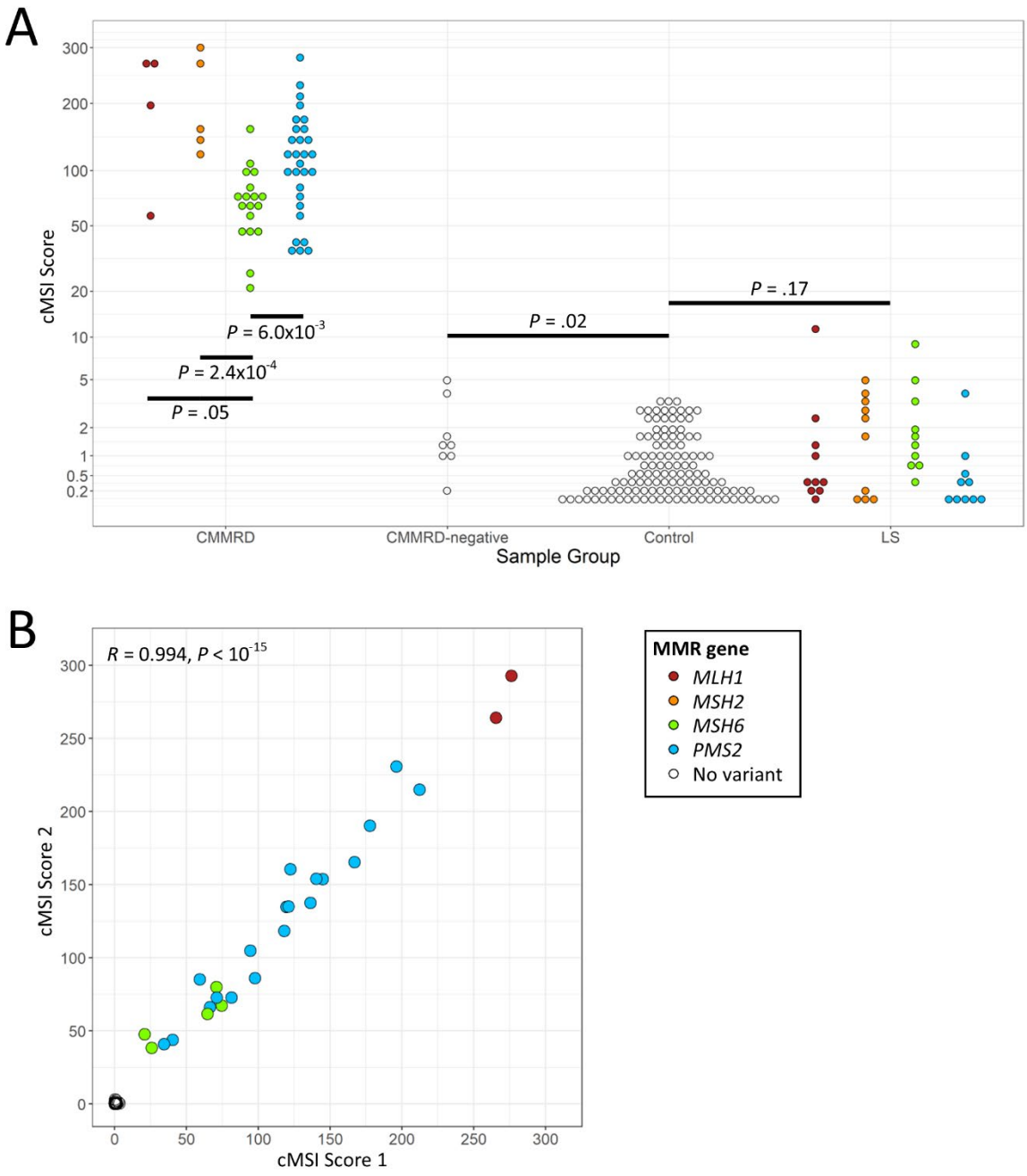

Figure 2: Sample cMSI scores. The cMSI scores of a blinded cohort of 56 CMMRD (*MLH1* n=4, *MSH2* n=5, *MSH6* n=18, *PMS2* n=29), 8 CMMRD-negative, and 43 control peripheral blood leukocyte (PBL) gDNAs, 80 reference control PBL gDNAs, and 40 Lynch syndrome (LS; *MLH1* n=10, *MSH2* n=10, *MSH6* n=10, *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).

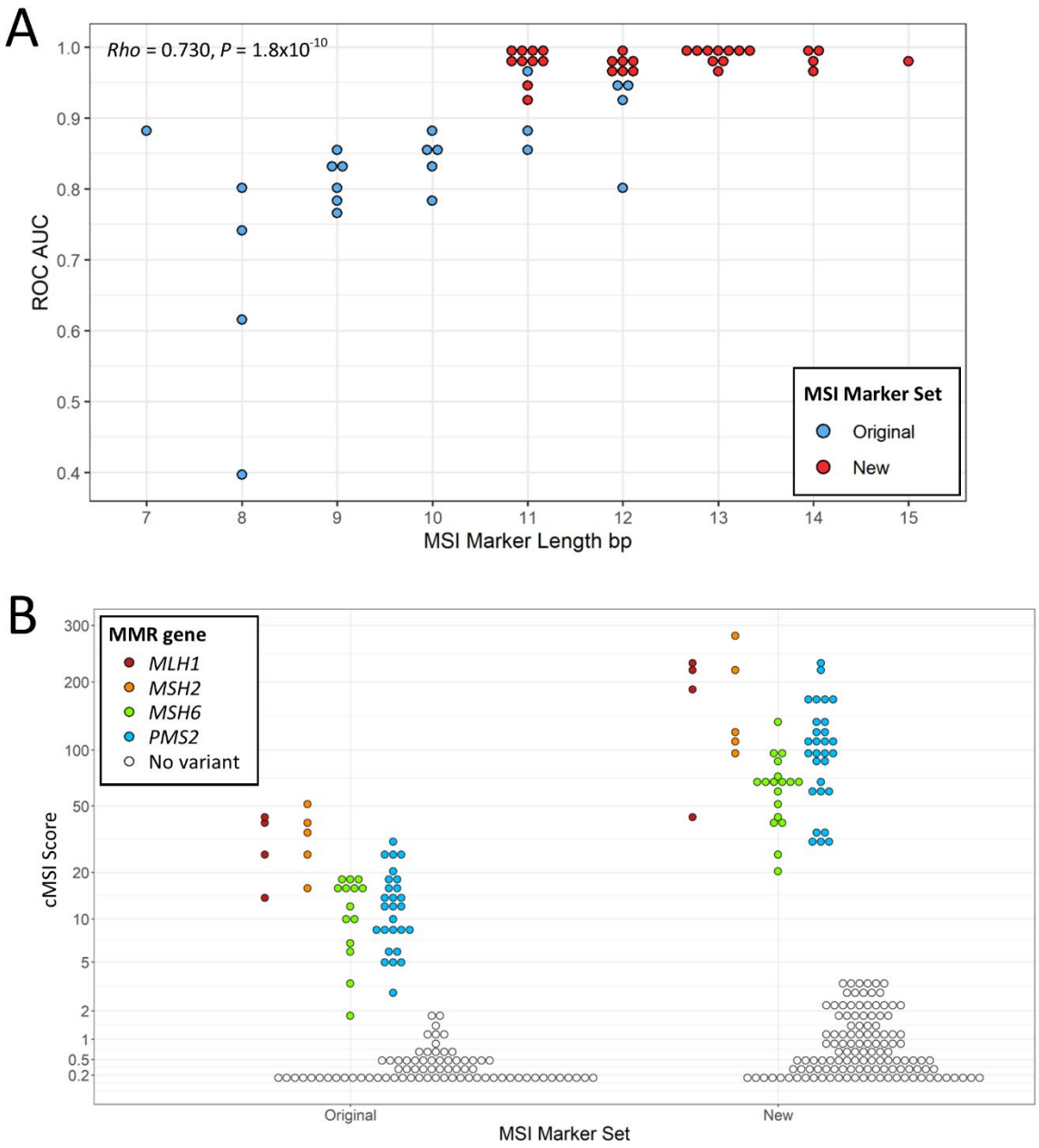

Figure 3: MSI marker characteristics and performance. A comparison of the length of each MSI marker and its receiver operator characteristic area under curve (ROC AUC) to discriminate between CMMRD and control PBL samples (A). A comparison of cMSI score of 50 CMMRD (*MLH1* n=4, *MSH2* n=5, *MSH6* n=14, *PMS2* n=27) and 75 control PBL samples using either the original 24 tumour-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).

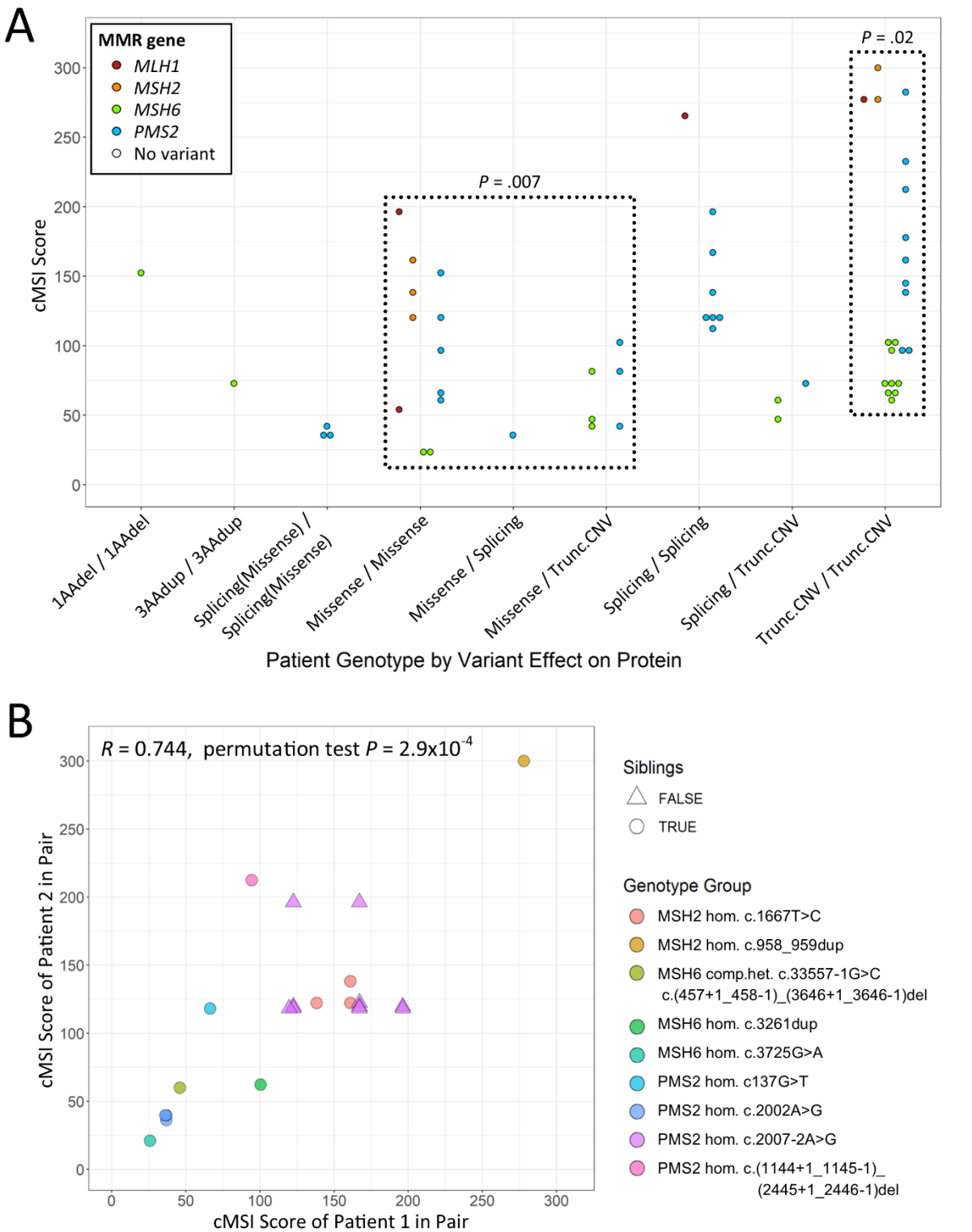

Figure 4: Sample cMSI scores by patient genotype. The cMSI scores of 56 CMMRD patients (*MLH1* n=4, *MSH2* n=5, *MSH6* n=18, *PMS2* n=29) grouped by the type of germline MMR variant according to effect on protein sequence. The dotted boxes highlight patients with mono- or biallelic missense variants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

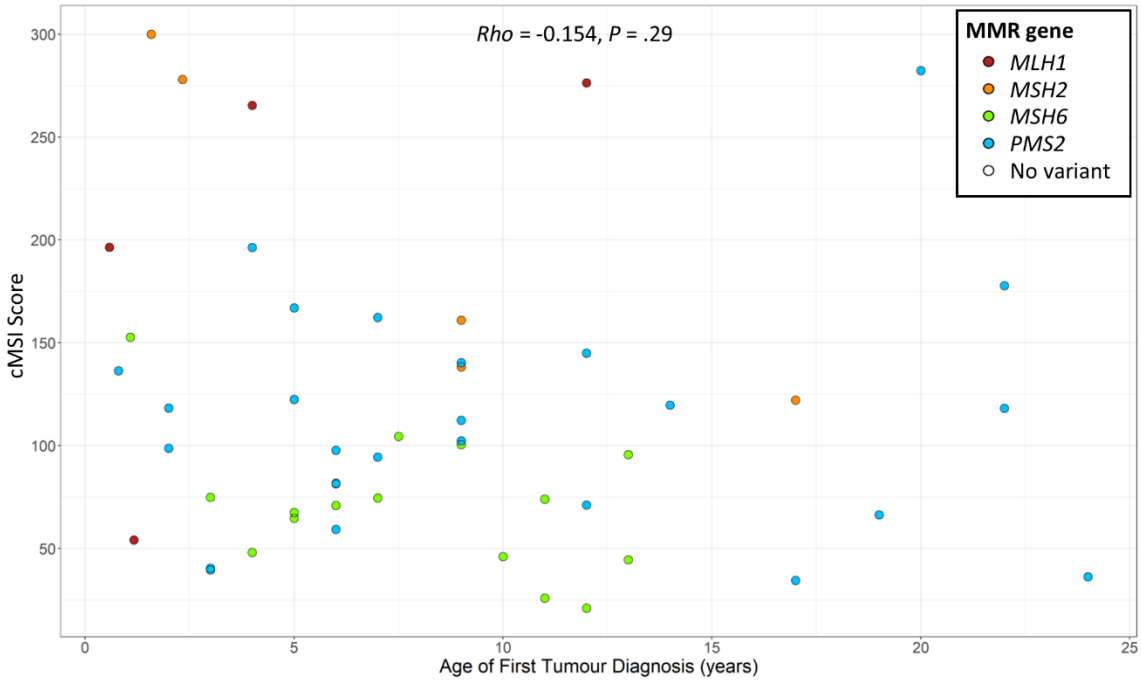

Figure 5: Associations of disease phenotype with cMSI score. The cMSI score and age of first tumour of 50 CMMRD patients (*MLH1* n=4, *MSH2* n=5, *MSH6* n=16, *PMS2* n=25).

Acknowledgements

The authors thank all 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, UK, 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, UK, 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 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 non-financial 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: Professor Jack Cuzick, Queen Mary University of London (Chair), Professor Frances Balkwill, Queen Mary University of London, Professor Tim Bishop, University of Leeds, Professor Sir John Burn, Newcastle University, Professor Andrew T. Chan, Harvard School of Medicine, Dr Colin Crooks, University of Nottingham, Professor Chris Hawkey, University of Nottingham, Professor Ruth Langley, University College London, Ms Mairead McKenzie, Independent Cancer Patients' Voice, Dr Belinda Nedjai, Queen Mary University of London, Professor Paola Patrignani, Università "G. d'Annunzio" di Chieti-Pescara, Professor 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.

J. Burn is a National Institute for Health and Care Research (NIHR) Senior Investigator and thanks the NIHR for their support. D. G. Evans is supported by the National Institute for Health Research (NIHR) Manchester Biomedical Research Centre (IS-BRC-1215-20007).