

## Germline loss-of-function variants in the base-excision repair gene *MBD4* cause a Mendelian recessive syndrome of adenomatous colorectal polyposis and acute myeloid leukaemia

**Claire Palles<sup>1\*</sup>, Edward Chew<sup>2\*</sup>, Judith E. Grolleman<sup>3\*</sup>, Sara Galavotti<sup>1\*</sup>, Christoffer Flensburg<sup>2\*</sup>, Erik A.M. Jansen<sup>3</sup>, Helen Curley<sup>1</sup>, Laura Chegwiddden<sup>1</sup>, Edward Arbe Barnes<sup>4</sup>, Ashish Bajel<sup>5</sup>, Kitty Sherwood<sup>6</sup>, Lynn Martin<sup>1</sup>, Huw Thomas<sup>7</sup>, Demetra Georgiou<sup>8</sup>, Florentia Fostira<sup>9</sup>, Yael Goldberg<sup>10,11</sup>, David J. Adams<sup>12</sup>, Simone A.M. van der Biezen<sup>3</sup>, Michael Christie<sup>2</sup>, Mark Clendenning<sup>13,14</sup>, Constantinos Deltas<sup>15</sup>, Aleksandar J. Dimovski<sup>16</sup>, Dagmara Dymerska<sup>17</sup>, Jan Lubinski<sup>17</sup>, Khalid Mahmood<sup>13,14</sup>, Rachel S. van der Post<sup>18</sup>, Mathijs Sanders<sup>19</sup>, Jürgen Weitz<sup>20</sup>, Jenny C. Taylor<sup>21</sup>, Clare Turnbull<sup>22</sup>, Lilian Vreede<sup>3</sup>, Tom van Wezel<sup>23</sup>, Celina Whalley<sup>1</sup>, Claudia Arnedo Pac<sup>24</sup>, Gulio Caravagna<sup>22</sup>, William Cross<sup>25</sup>, Daniel Chubb<sup>22</sup>, Anna Frangou<sup>4</sup>, Andreas Gruber<sup>26</sup>, Ben Kinnersley<sup>22</sup>, Boris Noyvert<sup>1</sup>, David Church<sup>4</sup>, Trevor Graham<sup>27</sup>, Richard Houlston<sup>22</sup>, Nuria Lopez-Bigas<sup>24</sup>, Andrea Sottoriva<sup>25</sup>, David Wedge<sup>26</sup>, Genomics England Research Consortium<sup>a</sup>, The CORGI Consortium<sup>b</sup>, WGS500 Consortium<sup>c</sup>, Mark A. Jenkins<sup>27,28</sup>, Roland P. Kuiper<sup>3,29</sup>, Andrew W. Roberts<sup>2,5,27</sup>, Marjolijn J.L. Ligtenberg<sup>3,18</sup>, Nicoline Hoogerbrugge<sup>3</sup>, Viktor H. Koelzer<sup>30</sup>, Andres Dacal Rivas<sup>31</sup>, Ingrid M. Winship<sup>28,32</sup>, Clara Ruiz Ponte<sup>33</sup>, Daniel D. Buchanan<sup>13,14,28</sup>, **Ian P.M. Tomlinson<sup>6#</sup>, Ian J. Majewski<sup>2,34#</sup>, Richarda M. de Voer<sup>3#</sup>****

1. Institute of Cancer and Genomic Sciences, College of Medical and Dental Science, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
2. Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052 Australia
3. Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, 6525 GA, Nijmegen, The Netherlands
4. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
5. Peter MacCallum Cancer Center and Royal Melbourne Hospital, Victorian Comprehensive Cancer Centre, Parkville, VIC, Australia
6. Edinburgh Cancer Research Centre, IGMM, University of Edinburgh, Crewe Road, Edinburgh EH4 2XR, UK
7. Family Cancer Clinic, St Mark's Hospital, Watford Road, Harrow HA1 3UJ, UK
8. Genomic Medicine, Imperial College Healthcare Trust and North West Thames Regional Genetics Service, UK
9. Molecular Diagnostics Laboratory, NCSR Demokritos, Athens, Greece
10. Raphael Recanati Genetic Institute, Rabin Medical Center – Beilinson Hospital, Petach Tikva, Israel
11. Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
12. Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
13. Colorectal Oncogenomics Group, Department of Clinical Pathology, Melbourne Medical School, The University of Melbourne, Parkville, Victoria, Australia
14. University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, Victoria, Australia
15. Center of Excellence in Biobanking and Biomedical Research and Molecular Medicine Research Center, University of Cyprus Medical School, Nicosia, Cyprus
16. Center for Biomolecular Pharmaceutical Analyzes, UKIM Faculty of Pharmacy, 1000, Skopje, Republic of Macedonia
17. Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, 70-111, Szczecin, Poland
18. Department of Pathology, Radboud Institute for Molecular Life Sciences, Radboud university medical center, 6525 GA, Nijmegen, The Netherlands
19. Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands
20. Department of Surgical Research, Universitätsklinikum Carl Gustav Carus, Technische Universität Dresden, 01307, Dresden, Germany
21. Oxford NIHR Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
22. Institute of Cancer Research, Cotswold Road, Sutton, Surrey SM2 5NG, UK
23. Department of Pathology, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands
24. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
25. Cancer Institute, 72 Huntley Street, University College London, London, WC1E 6BT, UK
26. Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, M1 7DN, UK
27. Barts Cancer Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK
28. Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Melbourne, Australia
29. Princess Máxima Center for Pediatric Oncology, 3584 CT, Utrecht, The Netherlands
30. Department of Pathology and Molecular Pathology, University Hospital Zurich, University of Zurich, Switzerland

31. Servicio de Digestivo, Hospital Lucus Augusti, Instituto de Investigación Sanitaria de Santiago (IDIS), Lugo, Galicia, Spain
32. Department of Medicine, Melbourne Medical School, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia
33. Fundación Pública Galega de Medicina Xenómica SERGAS, Grupo de Medicina Xenómica-USC, Instituto de Investigación Sanitaria de Santiago (IDIS), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Santiago de Compostela, Galicia, Spain
34. University of Melbourne, Department of Medical Biology, 1G Royal Parade VIC 3052 Australia

<sup>a, b, c</sup> See appendix

\* *These authors contributed equally*; # *These authors jointly supervised this project*

**Short title:** *MBD4*-associated neoplasia syndrome

**Authors for correspondence:**

Ian P. M. Tomlinson - [ian.tomlinson@igmm.ed.ac.uk](mailto:ian.tomlinson@igmm.ed.ac.uk)

Ian J. Majewski - [majewski@wehi.edu.au](mailto:majewski@wehi.edu.au)

Richarda M. de Voer – [richarda.devoer@radboudumc.nl](mailto:richarda.devoer@radboudumc.nl)

**ABSTRACT**

Inherited defects in base-excision repair (BER) predispose to adenomatous polyposis and colorectal cancer (CRC), yet our understanding of this important DNA repair pathway remains incomplete. By combining detailed clinical, histological and molecular profiling, we reveal biallelic germline loss-of-function (LOF) variants in the BER gene *MBD4* to predispose to adenomatous polyposis and –uniquely amongst CRC predisposition syndromes– to myeloid neoplasms. Neoplasms from *MBD4*-deficient patients almost exclusively accumulate somatic CpG>TpG mutations, resembling mutational signature SBS1. *MBD4*-deficient adenomas harbour mutations in known CRC driver genes, although *AMER1* mutations were more common and *KRAS* mutations less frequent. We did not find an increased risk for colorectal tumours in individuals with a monoallelic *MBD4* LOF variant. We suggest that this condition should be termed *MBD4*-associated neoplasia syndrome (MANS) and that *MBD4* is included in testing for the genetic diagnosis of polyposis and/or early-onset AML.

**Keywords (4-5):** polyposis; colorectal cancer; 5<sup>1</sup>-methylcytosine deamination; mutational signature; mutator phenotype

## INTRODUCTION

Several syndromes characterized by adenomatous polyposis and colorectal cancer (CRC) are caused by high-penetrance germline pathogenic variants that directly impair the correction of mispaired bases within DNA. These syndromes include the dominantly inherited condition polymerase proofreading-associated polyposis (PPAP) caused by pathogenic variants in the polymerase proofreading domains of *POLE* and *POLD1*,<sup>1</sup> and the recessively inherited conditions constitutional mismatch-repair deficiency, caused by variants in genes involved in mismatch repair (*MSH2*, *MLH1*, *PMS2*, *MSH6*),<sup>2,3</sup> and *MUTYH*-associated polyposis and *NTHL1*-associated tumour syndrome, caused by genes involved in base excision repair (BER) (*MUTYH* and *NTHL1*).<sup>4,5</sup> In each case, there are also increased risks of specific extra-colonic tumours. Although the mechanistic connection between these germline variants and colorectal tumours is not entirely clear, it is likely that the resulting increased somatic mutation rate in turn causes more somatic mutations in CRC driver genes such as *APC*, *KRAS* and *TP53*. Previous studies have linked the defect in each DNA repair gene to specific mutational signatures in tumours that align with their repair activity.<sup>6-9</sup>

Despite the identification of roughly 15 colorectal polyposis predisposition genes to date, genetic testing currently fails to identify a cause for a significant proportion of patients who develop multiple colorectal adenomas. Whilst some of these cases may have polygenic or non-genetic origins, it is important to identify any remaining high-risk polyposis predisposition genes in order to prevent CRC in probands and their relatives. Here, we used whole-genome sequencing (WGS) to identify the BER gene *MBD4* as a novel colorectal polyposis predisposition gene with recessive inheritance. By molecular profiling we describe how biallelic inactivating germline variants in *MBD4* predispose to both colorectal polyposis and AML. Furthermore, our data suggest that individuals with monoallelic loss-of-function (LOF) variants in *MBD4* are not at increased risk of polyposis and/or CRC.

## METHODS

### *Patient and control cohorts*

#### *i) Discovery cohort:*

Oxford-Illumina WGS500 project: constitutional DNA from 35 probands from the UK CORGI study with at least 10 colonic adenomas before age 60 were whole-genome sequenced as previously described.<sup>1</sup>

#### *ii) Replication cohorts:*

- a. Polyposis and AML: 13 unrelated patients with synchronous or metachronous colorectal polyposis and/or CRC and myelodysplastic syndrome (MDS) and/or AML. Samples were screened for *MBD4* using molecular inversion probes (MIP,  $n=11$ , **Supplemental Methods**) or whole-exome sequencing (WES;  $n=2$ ). Another 13 cases with synchronous or metachronous colorectal polyposis and/or CRC and MDS and/or AML were identified in the UKBiobank (see below).
- b. Polyposis and/or CRC: 1,493 patients with unexplained polyposis, familial and/or young CRC and >10 adenomatous polyps (with or without CRC), or CRC in combination with other tumour(s) were recruited. This study was approved by the local medical ethics committee (CMO; study number 2015/2172 of the Radboudumc Nijmegen). All samples were screened using MIP-sequencing (**Supplemental Methods; Supplementary Table 1**).
- c. Eighty-seven families comprised of 198 CRC and early-onset polyp affected individuals were selected for WGS/WES from participants of the Australasian Colorectal Cancer Family Registry (ACCFR): 1) met the definition of Familial Colorectal Cancer Typex(FCCTX) ( $n=55$ ); 2) Amsterdam II clinical criteria (AMII) with two of the defining triad being CRC-affected ( $n=4$ ); or 3) >2 CRC-affected blood relatives family members within families in the same blood line but not meeting FCCTX or AMII criteria (MCF)( $n=28$ ).

#### *iii) Population cohorts:*

- a. UKBiobank: we searched the WGS/WES data available for 200,000 participants in UKBiobank (October 26th 2020 release) for germline coding or splicing variants in *MBD4*. This research was conducted under UKBiobank application number 8508.
- b. UK 100,000 Genomes Project (100KGP): we searched the WGS data available (ISAAC pipeline) for 17,243 Caucasian participants of the rare diseases programme (v6 release, participants selected for phenotypes with no increased risk of cancer), 2,438 Caucasian CRC patients included in the cancer programme (v8 release) and 283 Caucasian multiple bowel polyps patients (143 v6 rare diseases, 140 pilot project) for germline coding or splicing variants in *MBD4*.
- c. GnomAD: we searched the full and non-Finnish European populations of the gnomAD database (v2.1) for germline coding or splice site variants in *MBD4*.

#### *Identification of germline variants*

WGS of constitutional DNA was performed using the Illumina HiSeq platform as described previously.<sup>1</sup> Targeted *MBD4* mutation screening was performed (i) using a custom MIP capture panel and sequenced on a NextSeq500 (Illumina) system, or (ii) using Sanger sequencing based on custom PCR amplicons. Variants predicted to result in a premature stop codon (nonsense or frameshifts) or alternative splicing/splice site defect were classed as LOF variants (details provided in the **Supplemental Methods**). All *MBD4* mutations were validated using independent Sanger sequencing reactions.

#### *Tumour sequencing and somatic variant calling*

Tumour samples and their origins from fresh-frozen or formalin fixed paraffin embedded (FFPE) tissue are listed in **Supplementary Table 2**. In brief, we used the i) Agilent SureSelectXT Human All Exon V6 (Agilent Technologies), ii) Agilent SureSelect Clinical Research Exome version 2 (Agilent Technologies) or iii) Illumina TruSeq exome (Illumina)

enrichment kits to produce whole-exome libraries that were sequenced using a NextSeq500 (Illumina) or NovaSeq 6000 (Illumina) (**Supplementary Table 2**).

Somatic variants were called using SuperFreq (v1.3.2),<sup>10</sup> Strelka (v2.9.2) and/or Mutect2 (GATK v4.1.0.0) with matched germline samples (**Supplementary Table 2**). Somatic mutations were annotated as described previously.<sup>7</sup> Full details of the somatic variant calling, removal of FFPE artefacts and validations are provided in the **Supplementary Methods**.

BAM files from previously published multi-region WES data (generated using the SeqCapEx Exome Enrichment Kit v3.0 (Nimblegen/Roche)) from nine sporadic fresh-frozen adenomas were available (**Supplementary Table 2**).<sup>11</sup> Somatic variants were re-called using Mutect2 as above.

#### *Somatic mutation analysis*

For each adenoma, high confidence somatic mutations were identified as described previously with minor modifications.<sup>7</sup> In brief, somatic variants covered by  $\geq 15$ x sequencing reads,  $\geq 10\%$  variant allele frequency, and  $\geq 6$  variant reads, and (for variants called by Mutect2) with  $\geq 2$  variant reads per read pair (to exclude FFPE artefacts) were included. WEHI-2 received an allogeneic bone marrow transplant and variants contributed by the haematopoietic stem cell donor were removed based on clonal tracking in SuperFreq.<sup>10</sup> For somatic variants identified by SuperFreq, a mean read depth  $> 30$  across all samples was required and excluded variants if they were detected above 3% VAF (supported by at least 2 reads) in a sample where the clone was deemed absent (clonality  $< 1\%$ ). A representative set of somatic mutations were validated by Sanger sequencing or by processing micro-dissected material from the adenomas with the TruSight Tumor 26 Kit (Illumina).

For each adenoma the number of somatic mutations per Mbp was calculated, with the total capture interval estimated using bases with  $\geq 15$  read depth. The mutation spectrum and the number of CpG>TpG transitions was determined. A linear model describing the number of CpG>TpG transitions as a function of age was analysed in base R. Methylation status of the sites with somatic mutation was assessed in public whole genome bisulfite sequencing

(WGBS) data from normal sigmoid colon from the Roadmap Epigenomics Consortium et al. (as per<sup>12</sup>). The contribution of mutational signatures to the somatic mutation spectrum was inferred using the R package MutationalPatterns<sup>13</sup> in combination with COSMIC-v3 mutational signatures.<sup>14,15</sup>

All non-synonymous somatic variants in cancer driver genes reported by The Cancer Gene Census (CGC v92) in COSMIC were extracted for each of the sequenced adenomas. Next, driver genes were prioritised with their previous associations as a colorectal cancer driver by TCGA and Dietlein et al.<sup>16,17</sup> To compare driver genes and mutation types, genes that were mutated in at least two independent adenomas from either the MBD4-deficient cases or sporadic adenomas were plotted in an oncoprint. In addition, we extracted all somatic mutations present in *APC* from the TCGA Colon Adenocarcinoma (COAD) and Rectum Adenocarcinoma (READ), and from a series of adenomas from patients without germline mutation *APC* mutations.

## RESULTS

### *Identification of a colorectal polyposis patient with a biallelic germline MBD4 loss-of-function variant*

We performed WGS of constitutional DNA from 35 unrelated patients with at least 10 colonic adenomas before the age of 60, but without pathogenic variants in the known Mendelian CRC genes.<sup>1</sup> By prioritizing coding germline variants with a high probability of LOF effects we identified a homozygous 4-bp frameshift deletion in *MBD4* (NM\_003925.2:c.612\_615del; NP\_003916.1:p.(Ser205ThrfsTer9)) in a single patient (D:II-1, **Figure 1A**). D:II-1 had no family history of polyposis or CRC and homozygosity analysis did not suggest consanguinity. We found both parents and a sibling to be heterozygous for the *MBD4* frameshift variant (**Figure 1A**). Re-sequencing of 100 controls from the Mediterranean island from which D:II-1's family originate did not reveal additional carriers. This variant appears once in 250,800 alleles in gnomAD v2.1.1.



*MBD4* encodes a BER glycosylase that repairs G:T mismatches resulting from the deamination of 5'-methylcytosine (5mC). Immunostaining for MBD4 revealed absence of protein expression in adenomas from D:II-1 (**Figure 1B**). Furthermore, *MBD4* mRNA was expressed in a lymphoblastoid cell line from D:II-1 (**Supplemental Figure 1A**), but no protein was detectable (**Figure 1C**).

D:II-1 presented with rectal bleeding and approximately 60 tubular colorectal adenomas ranging in size from <3mm to 1.5cm diameter at age 36. The largest adenomas were removed and showed evidence of high-grade dysplasia. Following colonoscopy, the patient underwent proctocolectomy at age 47, which identified 50-60 adenomas between the caecum and splenic flexure (mostly in the caecum) and 20-30 rectal adenomas, including some of 1cm-4cm diameter. No gastric or duodenal polyps were identified upon upper gastrointestinal endoscopy screening. Seven months after he underwent proctocolectomy, he presented with pancytopenia and was diagnosed with MDS, which progressed to AML three months later.

#### *Colorectal phenotype of previously reported AML patients with inherited MBD4 deficiency*

During the analysis of D:II-1, in an independent, contemporaneous study, we linked inherited *MBD4*-deficiency to early-onset acute myeloid leukaemia (AML).<sup>12</sup> In our prior work, two of the three AML patients with *MBD4*-deficiency were noted to have developed colonic polyps, whereas this was not evaluated in the third. We have subsequently obtained more comprehensive clinicopathological information on the colorectal phenotypes of those two cases. Patient WEHI-2 (previously WEHI-AML-2) developed a cumulative 17 colorectal polyps over a period of 22 years, with the first polyp excised at the age of 18 years. Histological assessment classified all available polyps ( $n=12$ ) as tubular adenomas with mild-to-moderate dysplasia, with the majority ( $n=7$ ) found in the rectum. A moderately differentiated adenocarcinoma was found in the ascending colon at age 40 and she underwent a right hemicolectomy. EMC-1 (previously EMC-AML-1) developed multiple colonic polyps and



underwent a hemicolectomy at age 31 years, although no polyp counts were reported and tissue blocks were unavailable for histological re-assessment.

### *Somatic mutation burden, spectra and signatures in MBD4-deficient colorectal adenomas*

To investigate the somatic mutation burden and spectrum, we performed WES on nine fresh frozen and two FFPE adenomas from D:II-1, and eight FFPE adenomas from WEHI-2. Some adenomas were micro-dissected to obtain multiple regions of the polyps and normal bowel (**Figure 2A-B; Supplemental Figure 1B-C**). Almost all somatic mutations in both MBD4-deficient cases were single nucleotide substitutions (SNVs), with a very small proportion of small indels. The fresh frozen adenomas of D:II-1 exhibited an elevated total somatic mutation burden (median 11.1 mutations/Mb [range 8.5-23.3]) compared to a set of nine sporadic adenomas (median 1.8 mutations/Mb [range 1.0-3.1]) (**Figure 2C, Supplemental Figure 2A and Supplemental Table 2**).<sup>11</sup> The mutation burden per adenoma increased for both cases with age (**Figure 2D**). The vast majority of the mutations were CpG>TpG transitions (>95%), which was significantly higher ( $P=2.9 \times 10^{-7}$ ) than in sporadic colorectal tumours (33% in nine fresh frozen sporadic adenomas, 14% in 468 fresh frozen CRCs from TCGA; **Figure 2E**).

We performed *de novo* signature analysis on the fresh-frozen adenomas from D:II-1 and separately from the FFPE adenomas from WEHI-2 (excluding P9, see below). The extracted signature profiles were nearly identical to each other (cosine similarity 0.998) and to COSMIC SBS1-v3 (cosine similarity 0.984 and 0.975 for D:II-1 and WEHI-2, respectively; **Figure 2F**). SBS1-v3 accounted for nearly all somatic mutations detected in D:II-1's adenomas and most of WEHI-2's adenomas (**Supplemental Figure 2B**). The exception was a WEHI-2 rectal polyp (P9) measuring 2cm diameter. All four physically distinct regions of this adenoma had a lower percentage of CpG>TpG transitions (55-77%) than the other adenomas and a substantial proportion of mutations were CA>AA transversions (**Figure 2C**). Using CpG>TpG transitions as a molecular clock, we found that this shift in mutational profile likely coincided with treatment for AML (**Supplemental Figure 3**).

Signature SBS1 is proposed to result from the failure to detect and/or remove G:T mismatches that result from deamination of 5mC. We assessed the methylation status of CpG sites that were mutated in adenomas from D:II-1 and WEHI-2 in whole genome bisulfite sequencing (WGBS) data from normal sigmoid colon generated by the Epigenome Atlas (<http://www.roadmapepigenomics.org/>). The mutated sites identified in the adenomas from D:II-1 and WEHI-2 were nearly all methylated in normal colon tissue (>96% of sites mutated in both D:II-1 and WEHI-2, compared to 58% across all exonic CpG sites) (**Figure 2G**).

Next, we compared the mutation profiles of all adenomas of D:II-1 and WEHI-2 to (i) the mutation profiles from the AML samples from EMC-1, WEHI-1 and WEHI-2, (ii) to myeloid progenitors from *Mbd4*<sup>-/-</sup> mice,<sup>12</sup> and (iii) to a human HAP1 *MBD4*<sup>ko</sup> cell line. The somatic mutation profiles and mutation rates were relatively consistent across these diverse samples (**Supplemental Figure 4A-B**).

#### *Somatic driver mutations in MBD4-deficient adenomas*

AMLs in *MBD4*-deficient individuals harbour a characteristic set of driver mutations,<sup>12</sup> and we sought to determine whether this was also true in colorectal adenomas from both D:II-1 (*n*=11) and WEHI-2 (*n*=8). In general, the spectra of driver genes were similar to sporadic CRC and adenomas, although indels were under-represented (most clearly seen for *APC*) (**Supplemental Table 3**). CRC driver genes with recurrent somatic mutations in our patients' adenomas included *APC* (*n*=19 adenomas), with a significant enrichment of *APC* p.(Arg1450\*), a CpG>TpG transition, compared to sporadic adenomas (Fisher exact, *P*=.0002). In addition, in adenomas from *MBD4*-deficient cases we observed a significant enrichment of somatic mutations in *TTN* (*n*=13 adenomas; Fisher exact, *P*=.0418) and *AMER1* (*n*=12 adenomas; Fisher exact, *P*=.039) (**Figure 3A**). Compared to sporadic adenomas, less *KRAS* mutations were observed (3/19 in adenomas from *MBD4*-deficient cases vs 7/9 in sporadic adenomas; Fisher exact, *P*=.0028). Strikingly, *APC* p.Arg1450\* mutations were found in ten out of eleven adenomas from D:II-1 and all of the adenomas sequenced by WES from WEHI-2 (**Figure 3A**). The preponderance of the *APC* p.Arg1450\* mutations unlikely reflects the presence of a

shared pre-neoplastic progenitor clone, as the mutation was absent in samples of normal intestine from D:II-1 and WEHI-2 (**Supplemental Table 4**). In sporadic CRCs from TCGA and adenomas from FAP-negative individuals ( $n=31$  individuals), p.(Arg1450\*) is the most frequent somatic *APC* mutation (32/486 and 8/93 tumours, respectively), but is found in a smaller proportion compared to adenomas from the MBD4-deficient cases ( $P<.00001$ ) (**Figure 3B**). Public WGBS data indicates similarly high level methylation at multiple CpG hotspot sites within *APC*, suggesting that methylation differences do not explain the over-representation of p.(Arg1450\*). Overall, 85.4% and 90.7% of CRC driver mutations in adenomas from D:II-1 and WEHI-2, respectively, were CpG>TpG transitions compared to 36.8% of mutations in nine sporadic adenomas. These data further indicate that 5mC deamination accounts for the vast majority of driver mutations contributing to the development of adenomas in MBD4-deficient cases.

*Most patients with both colorectal tumours and AML do not have germline MBD4 variants*

Synchronous or metachronous colorectal polyposis and AML comprise a rare clinical phenotype. In total, we had access to 26 cases with a colorectal and an MDS and/or AML phenotype (see materials and methods) and none had biallelic or monoallelic germline LOF *MBD4* variants.

*Germline LOF MBD4 variants in patients with colorectal polyps and/or CRC*

We sought to determine how common MBD4 deficiency is in patients with polyposis and/or CRC. Using a targeted screening strategy, we sequenced the coding sequence of *MBD4* in constitutional DNA from 1,490 European patients with unexplained colorectal polyps (with or without CRC), familial and/or young CRC, or CRC in combination with other tumour(s) (details in **Supplemental Table 1**). In addition, we assessed 87 families, comprising of 198 CRC- and/or early-onset polyp patients for which the germline was sequenced using WGS/WES for LOF variants in *MBD4*. Finally, we also reviewed 283 constitutional genome sequences from patients with multiple bowel polyps from the pilot programme of the 100KGP. No cases with

biallelic germline LOF variants in *MBD4* were identified, but we identified five with heterozygous LOF variants (**Supplemental Table 1; Supplemental Figure 5**): two nonsense variants (p.(Trp479\*); Family A), p.(Arg546\*); Family C), one frameshift (p.(Ile1111Tyrfs\*16); Family F), one synonymous variant that results in defective splicing (c.1410A>C; Family E, **Supplemental Figure 6A-C**) and one canonical splice site (c.1562-1G>T; Family B, **Supplemental Figure 6D-F**). Compared to the frequency of *MBD4* LOF variant alleles in European non-Finnish gnomAD population this is a significant enrichment (5/3,720 vs 48/129,200; OR =3.62 (95% CI 1.12-9.06), Fisher exact,  $P=0.016$ ).

Where possible, we investigated whether the heterozygous *MBD4* LOF variants segregated with the polyposis and/or CRC phenotypes in the respective families. Only one variant (p.(Trp479\*), Family A) showed complete co-segregation with the polyposis phenotype (**Supplemental Figure 5**). For the other families the segregation was discordant or could not (fully) be assessed (**Supplemental Figure 5**).

To further investigate the impact of heterozygous germline LOF *MBD4* variants on the risk of developing CRC, we studied patients in the CRC Domain of the 100KGP (in which recruitment was not formally enriched for any special features). No carriers of biallelic germline LOF *MBD4* variants were identified. Four Caucasian patients with one LOF allele were identified; this frequency ( $1.64 \times 10^{-3}$ , 4/2,438) was comparable to the frequency observed in a control set of phenotypically disease-free, unrelated participants selected from the 100,000 Genomes Rare Diseases project ( $1.22 \times 10^{-3}$  (21/17,243), indicating no evidence of enrichment in CRC cases (OR =1.34 (95% CI 0.46-3.93), Fisher exact,  $P=0.59$ ). In the UKBiobank, the frequencies of predicted LOF *MBD4* variants for participants with CRC, colorectal polyps, AML, MDS and all other participants were  $8.49 \times 10^{-4}$  (2/2,357),  $1.94 \times 10^{-4}$  (1/5,116)  $4.5 \times 10^{-3}$  (1/221),  $9.6 \times 10^{-3}$  (1/104),  $8.64 \times 10^{-4}$  (167/193,255) respectively (as of the October 2020 release of WES data). This data supports no enrichment in any of these groups (Fisher exact,  $P>0.05$  for all comparisons).

In contrast to our findings for CRC, germline LOF variants in *MBD4* have been linked to a 9-fold (Fisher exact,  $P=2.00 \times 10^{-5}$ ) higher incidence of uveal melanoma. Recent reports have shown that these uveal melanomas exhibit loss of the wild-type *MBD4* allele in the tumours.<sup>18-20</sup> We whole-exome sequenced four colorectal tumors from three individuals with a monoallelic LOF *MBD4* variant and accessed the WGS data of the four CRCs with a monoallelic LOF *MBD4* variant from the 100KGP (**Supplemental Table 1**). No “second hits” were identified in *MBD4* in any of the tumours. Mutational signature analyses were inconsistent across the four tumours from the three monoallelic *MBD4* LOF variant carriers. In two cases (Family A and F) the mutational signatures were comparable to TCGA COAD samples and the sporadic adenomas (**Supplemental Figure 7**). However, the two tumours from B:II-1 exhibited a higher contribution of SBS1, more comparable to the *MBD4*-deficient cases (**Supplemental Figure 7**). However, re-sequencing of *MBD4* in the germline using long-read sequencing did not reveal any additional variants, and Western blot and quantitative real-time PCR suggested the wild-type allele was expressed in blood-derived cell lines from B:II-1 (**Supplemental Figure 6D-F**). We also assessed blood cells from heterozygous *Mbd4* knockout mice and found no evidence of an elevated mutation rate (**Supplementary Figure 4A-B**). Overall, we found no clear evidence to support a large effect of monoallelic *MBD4* LOF variants on the risk of polyposis and/or CRC.

## DISCUSSION

Our data demonstrate that biallelic germline loss-of-function variants in *MBD4* cause a Mendelian recessive neoplasia syndrome comprising the rare phenotype of both AML and adenomatous colorectal polyposis. Previously, we reported multiple polyps in some AML patients with *MBD4* deficiency, but details of the polyps were not available at that time and hence their importance could not be demonstrated. Here, through independent identification of *MBD4*-deficiency in a patient with multiple adenomas and more detailed phenotypic and molecular analysis of tumours, we show that inherited *MBD4* deficiency causes a true adenomatous polyposis, with a colorectal phenotype similar to attenuated FAP and to patients

with germline pathogenic variants in *MUTYH*, *NTHL1*, *POLE* and *POLD1*. We suggest the name *MBD4*-associated neoplasia syndrome (MANS) for this condition. The absence of functional *MBD4* leads to an accumulation of somatic CpG>TpG mutations arising from spontaneous deamination of 5-methylcytosine, resulting in a somatic mutation spectrum very similar to COSMIC signature v3 SBS1. These mutations include characteristic driver mutations that support the development of adenomas and myeloid leukaemia.

Three out of four individuals with MANS thus far identified developed colorectal polyps early in life, and all four developed MDS/AML. We suggest that *MBD4* coding and splice regions are added to the gene panels currently used for the clinical genetic diagnostic testing of either early-onset AML, or CRC/polyp cases or families. In addition, it would be reasonable to test any individual with multiple polyps and MDS/AML. The pathogenic *MBD4* variants reported to date have included truncating and splice site variants, as well as an in-frame deletion affecting the glycosylase domain of *MBD4*. However, it remains possible that missense and other variants impair the function of *MBD4*.<sup>20</sup>

For *MBD4*, we estimate that the frequency of biallelic LOF variants in the general population (based on gnomAD and 100KGP) may be as low as one in 1 in 5-6 million individuals. As is typical of very rare cancer syndromes, it is difficult to estimate age-dependent risks of AML and colorectal tumours in MANS with any precision, although adenomatous colorectal polyps may be present in the second decade of life. While subject to revision as information on the features of MANS accrues, we propose two-yearly colonoscopy from age 18-20 or the date of diagnosis, a regimen often used for other polyposis syndromes caused by deficiency of *MUTYH* or *NTHL1*.<sup>21,22</sup> Regular colonoscopy may avoid the need for prophylactic or therapeutic colectomy, especially in patients with relatively low adenoma burdens. Upper gastrointestinal endoscopy is warranted in almost all other polyposis syndromes and it may be prudent to survey the stomach and duodenum at the current time, despite the lack of small bowel or gastric lesions reported in our patients. At least one of the AMLs in our cases developed from MDS and we have also observed clonal haematopoiesis in some cases.<sup>12</sup> We

suggest regular follow-up full blood counts for MANS cases if their initial presentation is with adenomatous polyposis. If the individual presents with AML, then we suggest genetic testing for any family member being considered as a haematopoietic stem cell donor, in keeping with current expert recommendations for managing inherited predisposition to myeloid malignancy.<sup>23,24</sup>

Although carriers of heterozygous LOF *MBD4* variants are susceptible to uveal melanoma,<sup>18-20</sup> our data shows no convincing evidence that such individuals have a large increased risk of developing polyposis and/or CRC. Complete inactivation of *MBD4* is observed in uveal melanoma, where wildtype *MBD4* is lost due to monosomy of chromosome 3, whereas in CRC this is not a frequent event.<sup>25</sup> Prior analysis of tumours from heterozygous *MBD4* cases in TCGA revealed no significant increase in CpG>TpG mutations,<sup>12</sup> unless the wildtype allele was inactivated, and this was also consistent with results generated in the *Mbd4* knockout mouse. We cannot, however, rule out the possibility that individuals heterozygous for an *MBD4* LOF variant have a small increased risk of CRC and/or polyposis, as has been described for heterozygous LOF *MUTYH* variant carriers.<sup>26</sup> We identified a single heterozygous LOF *MBD4* case with a substantial number of somatic variants that fit with an *MBD4* deficiency in the absence of a second hit in *MBD4* (germline or somatic), but it remains possible that the wildtype allele was silenced.<sup>27</sup> On balance, it seems unlikely that heterozygous *MBD4* LOF carriers would benefit from colorectal surveillance beyond population screening, unless the individual's phenotype or family history suggests otherwise.

In conclusion, constitutional deficiency of *MBD4* causes a rare genetic syndrome that is marked by the development of adenomatous polyposis and AML. Genetic testing for MANS could be implemented readily and rapidly by incorporating *MBD4* into existing gene panels. *MBD4* deficiency results in an elevated mutation burden with a mutation spectrum very similar to COSMIC signature SBS1-v3. A high mutational burden is associated with a good prognosis from CRC, and we speculate that MANS CRCs may also respond well to immune checkpoint



inhibitors, as has been reported for PD-1 inhibitors in MBD4-deficient uveal melanomas.<sup>18,19</sup> It is possible that such a strategy could be used to treat neoplasia in MANS.

## **Acknowledgements**

We thank Dr. Robbert D.A. Weren, Dr. Janet Vos and Eveline Kamping for technical assistance, the Genome Technology Platform for MIP sequencing support, Dr. Christian Gillissen for the use of the annotation pipeline in the Radboudumc, Nijmegen, and Dr. Peggy Manders for the use of samples from the Dutch Parelsnoer Institute Biobank Hereditary Colorectal Cancer. We acknowledge expert technical assistance from the Austin Molecular Laboratory and the Australian Genome Research Facility for providing expert technical assistance with genomic analysis of cancer samples. We thank members of the Colorectal Oncogenomics Group and the participants and staff from the Colon-CFR in particular, Maggie Angelakos, Samantha Fox and Allyson Templeton for their support of this study. Computation and bioinformatics were also provided by Melbourne Bioinformatics on its Peak Computing Facility. Furthermore, we thank the Birmingham Genomics Service at The University of Birmingham, for the generation of the sequencing and methylation array data. Some of the computations described in this paper were performed using the University of Birmingham's BlueBEAR HPC service, which provides a High Performance Computing service to the University's research community. We thank all participants of the CORGI study for their collaboration. The author(s) would further like to acknowledge networking support by the Cooperation in Science and Technology Action CA17118, supported by the European Cooperation in Science and Technology. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. The colon schematics presented in Figure 2 was sourced from Servier SMART Medical Art (<https://smart.servier.com/>) and modified for us under a CC BY 3.0 license.

**Disclosures:** The authors declare they hold no conflict of interest with respect to this work.

**Grant support:** This study was funded by grants from the Dutch Cancer Society (KUN2015-7740; 12174/2019-1), the Sacha Swarttouw-Hijmans foundation, Cancer Research UK C6199, the EU ERC (EVOCAN), the National Health and Medical Research Council of Australia (NHMRC project 1145912, program 1113577, investigator 1174902 grants) and the Cancer Council of Victoria (1181108), with fellowship support from the Victorian Cancer Agency (IJM MCRF15018), the Alfred Felton Bequest (IJM)

and the Leukaemia Foundation of Australia (Bill Long Charitable Trust PhD Clinical Scholarship to EC). CP acknowledges funding from Bowel Cancer UK (grant 18PG0010). Research was also made possible through the Australian Cancer Research Foundation, Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRISS. DDB is supported by a NHMRC R.D. Wright Career Development Fellowship (GNT1125268) and NHMRC Emerging Leadership Fellowship (GNT1194896). MAJ is supported by NHMRC Leadership Fellowship. The ACCFR is supported in part by funding from the National Cancer Institute (NCI), National Institutes of Health (NIH) (award U01 CA167551). The research was also supported in part by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford. JCT discloses that the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

**Contributors:** Study supervision: CP, IPMT, IJM, RMdV. Analysis and drafting: CP, EC, JEG, SG, CF, IPMT, IJM, RMdV. Data support: EAMJ, HC, LC, EAB, AB, KS, LM, HT, DG, FF, YG, DJA, SAMvdB, MC, MC, CD, JAD, DD, JL, KM, RSvdP, MS, JW, JCT, CT, LV, TvW, CW, CA, GC, WC, DC, AF, AG, BK, BN, DC, TG, RH, NLB, AS, DW, MAJ, RPK, AWR, MJLL, NH, VHK, ADR, IMW, CRP, DDD. Critical revision: all authors. Shared last authors: IPMT, IJM, RMdV.

### Appendix:

**<sup>a</sup>Genomics England Research Consortium:** Ambrose, JC<sup>1</sup>, Arumugam, P<sup>1</sup>, Bleda, M<sup>1</sup>, Boardman-Pretty, F<sup>1,2</sup>, Boustred, CR<sup>1</sup>, Brittain, H.<sup>1</sup>, Caulfield, MJ.<sup>1,2</sup>, Chan, GC<sup>1</sup>, Fowler, T<sup>1</sup>, Giess A<sup>1</sup>, Hamblin, A.<sup>1</sup>, Henderson, S.<sup>1,2</sup>; Hubbard, TJP<sup>1</sup>, Jackson, R<sup>1</sup>, Jones, LJ<sup>1,2</sup>; Kasperaviciute, D<sup>1,2</sup>, Kayikci, M<sup>1</sup>, Kousathanas, A<sup>1</sup>, Lahnstein, L<sup>1</sup>, Leigh, SEA<sup>1</sup>, Leong, IUS<sup>1</sup>, Lopez, FJ<sup>1</sup>, Maleady-Crowe, F<sup>1</sup>, Moutsianas, L<sup>1,2</sup>, Mueller, M<sup>1,2</sup>, Murugaesu, N<sup>1</sup>, Need, AC<sup>1,2</sup>, O'Donovan P<sup>1</sup>, Odhams, CA<sup>1</sup>, Patch, C<sup>1,2</sup>, Perez-Gil, D<sup>1</sup>, Pereira, MB.<sup>1</sup>, Pullinger, J<sup>1</sup>, Rahim, T<sup>1</sup>, Rendon, A<sup>1</sup>, Rogers, T<sup>1</sup>, Savage, K<sup>1</sup>, Sawant, K<sup>1</sup>, Scott, RH<sup>1</sup>, Siddiq, A<sup>1</sup>, Sieghart, A<sup>1</sup>, Smith, SC<sup>1</sup>, Sosinsky, A<sup>1,2</sup>, Stuckey, A<sup>1</sup>, Tanguy M<sup>1</sup>, Thomas, ERA<sup>1,2</sup>, Thompson, SR<sup>1</sup>, Tucci, A<sup>1,2</sup>, Walsh, E<sup>1</sup>, Welland, MJ<sup>1</sup>, Williams, E<sup>1</sup>, Witkowska, K<sup>1,2</sup>, Wood, SM<sup>1,2</sup>.

<sup>1</sup> Genomics England, London, UK; <sup>2</sup> William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK.

**<sup>b</sup>The CORGI Consortium:** Dr. Kai Ren Ong (Birmingham Women's Hospital, Birmingham), Prof. Andrew Beggs (Institute of Cancer and Genomic Sciences, University of Birmingham), Dr. Alan Donaldson (St. Michael's Hospital, Dr. Ruth Armstrong, Addenbrooke's NHS Trust, Cambridge), Dr. Carole Brewer (Royal Devon & Exeter Hospital (Heavitree)), Exeter, Prof. Jayantha Arnold (Ealing Hospital, Middlesex), Dr. Munaza Ahmed (Great Ormond Street Hospital, London), Dr. Louise Izatt (Guy's Hospital, London), Dr. Andrew Latchford (St Mark's Hospital, Harrow and Division of Surgery and Cancer, Imperial College London), Dr. Dorothy Halliday (Nuffield Orthopaedic Hospital, Oxford), Peter Risby (The Oxford Genomic Medicine Centre, Oxford), Dr. Paul Brennan (The James Cook University Hospital, Middlesbrough), Dr. Alison Kraus (Chapel Allerton Hospital, Leeds), Dr. Julian Barwell (Leicester Royal Infirmary, Leicester), Dr. Lynn Greenhalgh (Liverpool Women's Hospital), Prof. D. Gareth Evans (University of Manchester, Manchester), Kate Green (University of Manchester, Dr. Timothy Simmons (Institute of Genetic Medicine, International Centre for Life, Newcastle upon Tyne), Dr. Rachel Harrison (Nottingham University Hospitals, Nottingham), Prof. Ragunath (Queen's Medical Centre Campus, Nottingham), Prof. Brian Davidson (Royal Free Hampstead NHS Trust, University Dept. of Liver Medicine & Transplantation, London), Dr. Zoe Kemp (The Royal Marsden, Sutton), Dr. Helen Hanson (St George's University, London), Dr. Katie Snape (St George's University, London), Prof. Anneke Lucassen (Princess Anne Hospital, Southampton), Dr. Kevin J Monahan (West Middlesex University Hospital, Middlesex), Prof. Patrick Morrison (City Hospital Campus, Belfast).

**<sup>c</sup>WGS500 Consortium:** Peter Donnelly<sup>1</sup>, John Bell<sup>2</sup>, David Bentley<sup>3</sup>, Gil McVean<sup>1</sup>, Peter Ratcliffe<sup>1</sup>, Jenny Taylor<sup>1,4</sup>, Andrew Wilkie<sup>4,5</sup>, John Broxholme<sup>1</sup>, David Buck<sup>1</sup>, Jean-Baptiste Cazier<sup>6</sup>, Richard Cornall<sup>1</sup>, Lorna Gregory<sup>1</sup>, Julian Knight<sup>1</sup>, Gerton Lunter<sup>1</sup>, Ian Tomlinson<sup>7</sup>, Andrew Wilkie<sup>4,5</sup>, Christopher Allan<sup>1</sup>, Moustafa Attar<sup>1</sup>, Angie Green<sup>1</sup>, Lorna Gregory<sup>1</sup>, Sean Humphray<sup>3</sup>, Zoya Kingsbury<sup>3</sup>, Sarah Lambie<sup>1</sup>, Lorne Lonie<sup>1</sup>, Alistair Pagnamenta<sup>1</sup>, Paolo Piazza<sup>15</sup>, Amy Trebes<sup>1</sup>, John Broxholme<sup>1</sup>, Richard Copley<sup>8</sup>, Simon Fiddy<sup>9</sup>, Russell Grocock<sup>3</sup>, Edouard Hatton<sup>1</sup>, Chris Holmes<sup>1</sup>, Linda Hughes<sup>1</sup>, Peter Humburg<sup>1</sup>, Alexander Kanapin<sup>10</sup>, Stefano Lise<sup>11</sup>, Hilary Martin<sup>12</sup>, Lisa Murray<sup>3</sup>, Davis McCarthy<sup>13</sup>, Andy Rimmer<sup>14</sup>, Natasha Sahgal<sup>1</sup>, Ben Wright<sup>1</sup>, Chris Yau<sup>16</sup>.

<sup>1</sup>The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK; <sup>2</sup>Office of the Regius Professor of Medicine, Richard Doll Building, Roosevelt Drive, Oxford, OX3 7LF, UK; <sup>3</sup>Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex, CB10

1XL, UK; <sup>4</sup>NIHR Oxford Biomedical Research Centre, Oxford, UK; <sup>5</sup>Weatherall Inst of Molecular Medicine, University of Oxford; John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK; <sup>6</sup>Centre for Computational Biology, Haworth Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK; <sup>7</sup>Edinburgh Cancer Research Centre, IGMM, University of Edinburgh, Crewe Road, Edinburgh EH4 2XR, UK; <sup>8</sup>UPMC Paris 6, CNRS UMR 7009, Villefranche-sur-Mer; <sup>9</sup>Oxford Nanopore Technology, Edmund Cartwright House, 4 Robert Robinson Avenue, Oxford Science Park, Oxford, OX4 4GA, UK; <sup>10</sup>CRUK Oxford Centre, Department of Oncology, Old Road Campus Research Building, Roosevelt Drive Oxford, OX3 7DQ, UK; <sup>11</sup>Centre for Evolution and Cancer, The Institute of Cancer Research, Brookes Lawley Building, 15 Cotswold Road, Sutton, SM2 5NG, UK; <sup>12</sup>Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; <sup>13</sup>EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK; <sup>14</sup>Genomics plc, King Charles House, Park End Street, Oxford, OX1 1JD, UK; <sup>15</sup>Imperial College London, Commonwealth Building, Hammersmith Campus, Du Cane Road, London, W12 0NN, UK; <sup>16</sup>Division of Informatics, Imaging & Data Sciences, University of Manchester, UK

## SUPPLEMENTAL FILES:

### Supplementary Methods

**Supplemental Figure 1:** MBD4 expression in lymphoblastoid cells and histology of polyps from MBD4 deficient cases

**Supplemental Figure 2:** Somatic burden and mutational signature refitting of adenomas of MBD4 deficient cases.

**Supplemental Figure 3:** Clonal evolution of adenomas in WEHI-2.

**Supplemental Figure 4:** Mutation profiles and cosine similarity of observed mutations in various samples.

**Supplemental Figure 5:** Pedigrees of heterozygous *MBD4* loss-of-function variant carriers and co-segregation analyses.

**Supplemental Figure 6:** Analyses of MBD4 heterozygous splice site variants.

**Supplemental Figure 7:** Mutational signature refitting analyses of tumours from heterozygous MBD4 loss-of-function variant carriers.

**Supplemental Table 1:** Patient cohort inclusion and results of the molecular inversion probe screening.

**Supplemental Table 2:** Details on the tumours subjected to sequencing.

**Supplemental Table 3:** Somatic driver mutations in genes associated with colorectal cancer development in MBD4-deficient and sporadic polyps.

**Supplemental Table 4:** Validation of somatic driver mutations in by panel sequencing in polyps and normal tissues from WEHI-2.

## LEGENDS TO FIGURES:

**Figure 1: MBD4 deficiency in individual with colorectal adenomas and acute myeloid leukaemia. A)** Pedigree of Family D with a homozygous *MBD4* loss-of-function variant. For

colorectal adenomas, we show the cumulative tumour numbers from age at first presentation and screening colonoscopy to age at last contact. Abbreviations: Ads: adenomas; AML: acute myeloid leukemia; yoa: years of age; MT: mutation; WT: wild-type. **B)** Representative MBD4 IHC of an adenoma from patient D:II-1 (upper panels) and of a normal colon with wild type MBD4 (lower panels) stained with anti-MBD4 antibody. The D:II-1 (upper panels) show a tubular adenoma with low grade dysplasia showing typical nuclear changes (pencil shaped nuclei, crowding and pseudostratification). **C)** Western blot analysis of MBD4 expression in lymphoblastoid cells from patient D:II-1 (-) and treated with an siRNA targeting MBD4 (+). The colorectal cancer cell line HCT116 was used a positive control. Two central lanes were left empty. Anti-alpha-Tubulin was used as loading control.

**Figure 2: Somatic mutation burden and analysis of polyps of MBD4 deficient cases. A- B)** Polyps were removed from different parts of the large intestine and rectum from the two MBD4-deficient cases, with age in years specified in brackets. Polyps that were formalin-fixed and paraffin-embedded (FFPE) are indicated with an asterisk (\*). Sequencing was performed from multiple regions for some polyps, indicated with a hashtag (#) (H&E shown in Supplemental Figure 1). **C)** Somatic mutation rate for each polyp, FFPE samples indicated with asterisks. The colour of the bars represents mutations in different sequence contexts, with red showing CG>TG mutations, blue showing CA>AA mutations (primarily detected in WEHI-2 P9), and grey representing other base contexts. The median value is presented for samples that had multi-region sequencing. **D)** The number of somatic CG>TG mutations detected in WES data is plotted as function of age. The linear fit is shown, together with an error estimate. **E)** We assessed the contribution of deamination of 5mC to MBD4-deficient samples by comparing the number of CG>TG mutations to all other single nucleotide mutations. The plot compares MBD4 deficient polyps and AMLs<sup>12</sup> to sporadic polyps, and to colon and rectal cancers from TCGA.<sup>16</sup> MSI: microsatellite instability; MSI-H: MSI-high; MSS: microsatellite stable; "MSS" includes both MSS and MSI-low samples; TCGA NA: no MSI data available. **F)** Extracted *de novo* signature SBS1<sup>MBD4</sup> C>T panel from all polyps from D:II-1(left) and polyps

P1-P8 of WEHI-2 (middle) , and the C>T panel from COSMIC SBS1-v3 (right). **G)** Fraction of mutated CpG sites that are methylated in normal sigmoid colon (beta value>0.5 in WGBS data from the Roadmap Epigenomics Consortium). Each point summarises WES results from a sample and includes all sites with sufficient coverage in WGBS ( $n=177-1507$  CG>TG mutations) and the median value is shown with a horizontal line. The grey line shows the fraction of methylated CG sites across all exons.

**Figure 3: Driver gene mutation analysis of polyps from MBD4-deficient cases. A)** Oncoprint of driver gene mutation analysis of genes frequently mutated in colorectal cancer. For each polyp the number and type of somatic mutation is shown. \* Polyp P10 from WEHI-2 was sequenced using a targeted panel. **B)** Map of the somatic mutations observed in *APC* in all polyps from the two MBD-4 deficient cases and the TCGA COAD and READ datasets. *APC* p.(Arg1450\*) is the most frequently observed somatic mutation. Black lollipops indicate protein truncating variants and green lollipops indicate missense variants.

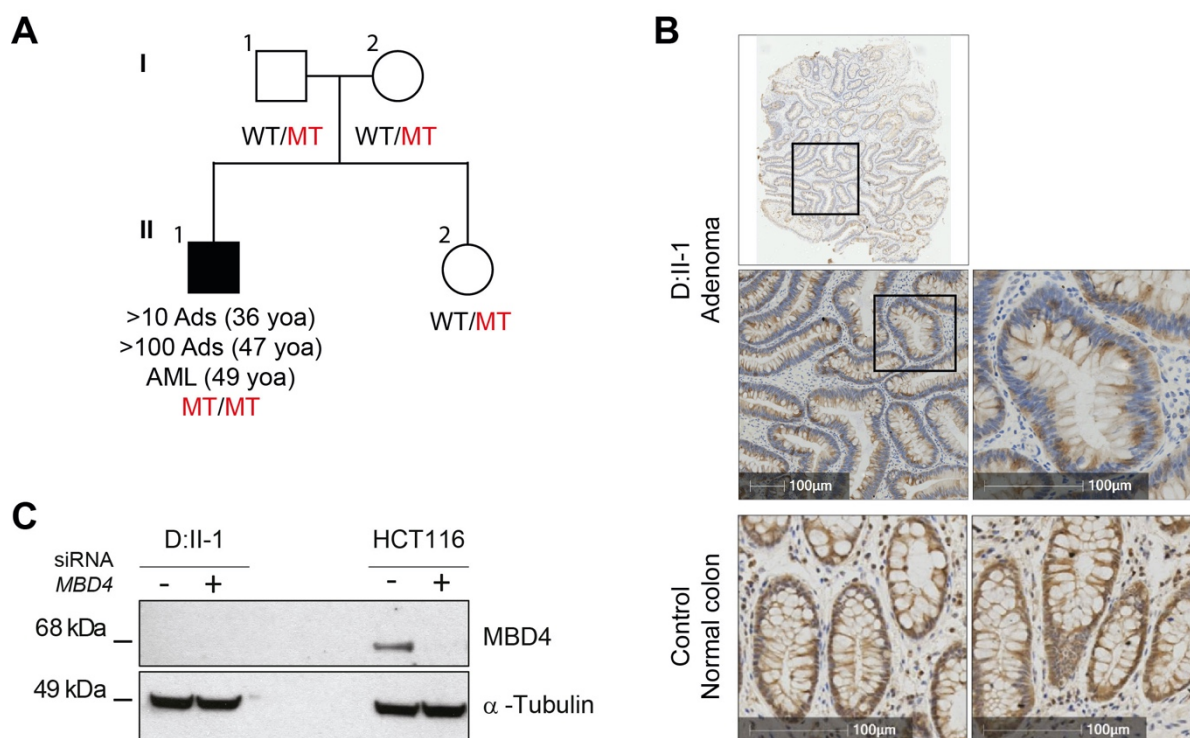
## REFERENCES

1. Palles C, Cazier JB, Howarth KM, et al. Germline mutations affecting the proofreading domains of *POLE* and *POLD1* predispose to colorectal adenomas and carcinomas. *Nat Genet* 2013;45:136-44.
2. Will O, Carvajal-Carmona LG, Gorman P, et al. Homozygous *PMS2* deletion causes a severe colorectal cancer and multiple adenoma phenotype without extraintestinal cancer. *Gastroenterology* 2007;132:527-30.
3. Wimmer K, Kratz CP, Vasen HF, 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.
4. Weren RD, Ligtenberg MJ, Kets CM, et al. A germline homozygous mutation in the base-excision repair gene *NTHL1* causes adenomatous polyposis and colorectal cancer. *Nat Genet* 2015;47:668-71.
5. Al-Tassan N, Chmiel NH, Maynard J, et al. Inherited variants of *MYH* associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* 2002;30:227-32.
6. Drost J, van Boxtel R, Blokzijl F, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* 2017;358:234-238.
7. Grolleman JE, de Voer RM, Elsayed FA, et al. Mutational Signature Analysis Reveals *NTHL1* Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell* 2019;35:256-266 e5.
8. Pilati C, Shinde J, Alexandrov LB, et al. Mutational signature analysis identifies *MUTYH* deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol* 2017;242:10-15.



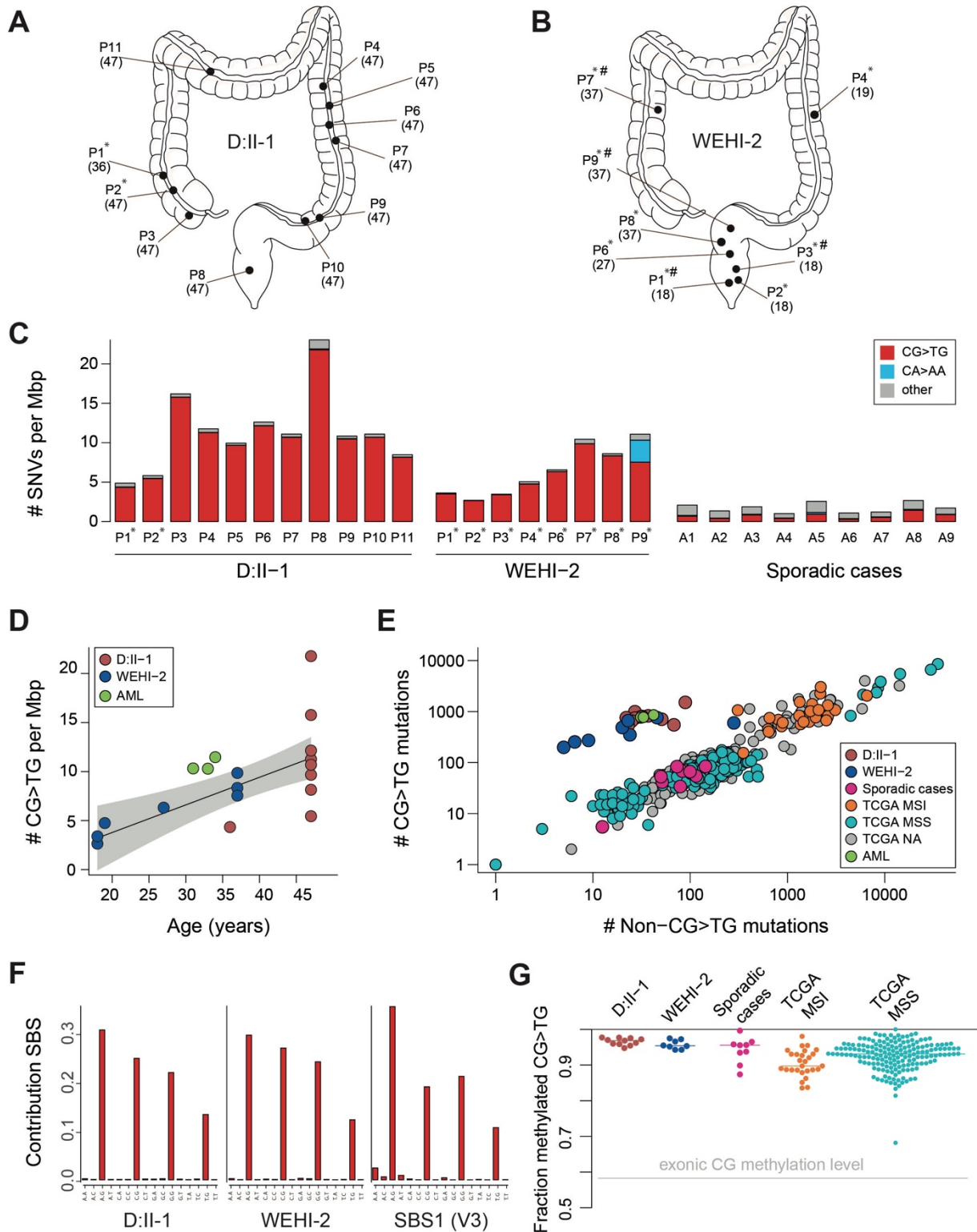
9. Viel A, Bruselles A, Meccia E, et al. A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* 2017;20:39-49.
10. Flensburg C, Sargeant T, Oshlack A, et al. SuperFreq: Integrated mutation detection and clonal tracking in cancer. *PLoS Comput Biol* 2020;16:e1007603.
11. Cross W, Kovac M, Mustonen V, et al. The evolutionary landscape of colorectal tumorigenesis. *Nature ecology & evolution* 2018;2:1661-1672.
12. Sanders MA, Chew E, Flensburg C, et al. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. *Blood* 2018;132:1526-1534.
13. Blokzijl F, Janssen R, van Boxtel R, et al. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* 2018;10:33.
14. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature* 2020;578:94-101.
15. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 2019;47:D941-d947.
16. Muzny DM, Bainbridge MN, Chang K, et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;487:330-337.
17. Dietlein F, Weghorn D, Taylor-Weiner A, et al. Identification of cancer driver genes based on nucleotide context. *Nat Genet* 2020;52:208-218.
18. Johansson PA, Stark A, Palmer JM, et al. Prolonged stable disease in a uveal melanoma patient with germline MBD4 nonsense mutation treated with pembrolizumab and ipilimumab. *Immunogenetics* 2019.
19. Rodrigues M, Mobuchon L, Houy A, et al. Outlier response to anti-PD1 in uveal melanoma reveals germline MBD4 mutations in hypermutated tumors. *Nat Commun* 2018;9:1866.
20. Derrien AC, Rodrigues M, Eeckhoutte A, et al. Germline MBD4 mutations and predisposition to uveal melanoma. *J Natl Cancer Inst* 2020.
21. Nielsen M, Infante E, Brand R. MUTYH polyposis: GeneReviews, 2012.
22. Kuiper RP, Nielsen M, De Voer RM, et al. NTHL1 Tumor Syndrome. Seattle (WA): University of Washington, Seattle: GeneReviews, 2020.
23. Godley LA, Shimamura A. Genetic predisposition to hematologic malignancies: management and surveillance. *Blood* 2017;130:424-432.
24. University of Chicago Hematopoietic Malignancies Cancer Risk T. How I diagnose and manage individuals at risk for inherited myeloid malignancies. *Blood* 2016;128:1800-1813.
25. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;2:401-4.
26. Win AK, Dowty JG, Cleary SP, et al. Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. *Gastroenterology* 2014;146:1208-11.e1-5.
27. Howard JH, Frolov A, Tzeng CW, et al. Epigenetic downregulation of the DNA repair gene MED1/MBD4 in colorectal and ovarian cancer. *Cancer Biol Ther* 2009;8:94-100.

## Figure 1



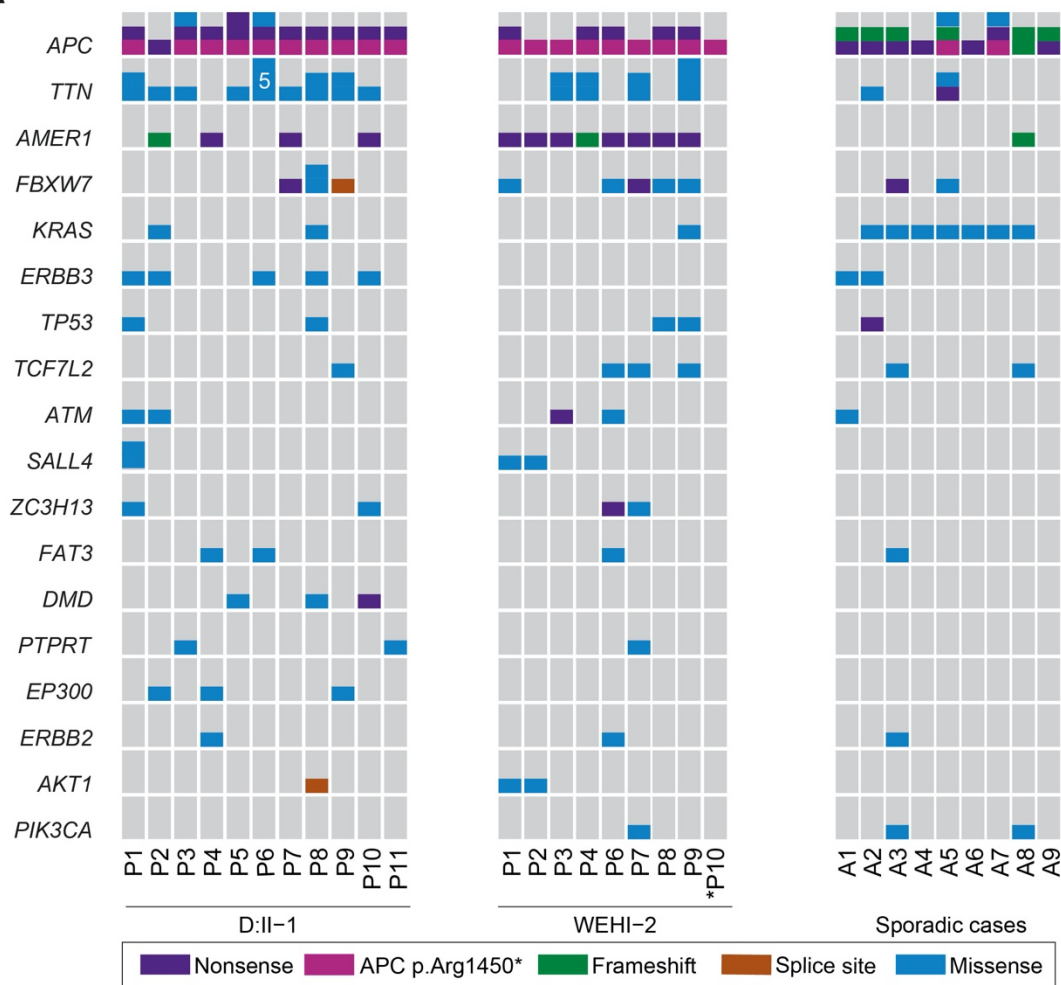


**Figure 2**



**Figure 3**

**A**



**B**

