bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Life without mismatch repair

Authors: Mathijs A. Sanders[1,2], Harald Vöhringer[3], Victoria J. Forster[4,5], Luiza Moore[1], Brittany B. Campbell[4,5], Yvette Hooks[1], Melissa Edwards[4,5], Vanessa Bianchi[4,5], Tim H. H. Coorens[1], Timothy M. Butler[1], Henry Lee-Six[1], Philip S. Robinson[1,6], Christoffer Flensburg[7,8], Rebecca A. Bilardi[7,8], Ian J. Majewski[7,8], Agnes Reschke[9], Elizabeth Cairney[10], Bruce Crooks[11], Scott Lindhorst[12], Duncan Stearns[13], Patrick Tomboc[14], Ultan McDermott[1,15], Michael R. Stratton[1], Adam Shlien[4,16,17], Moritz Gerstung[3,18,21], Uri Tabori[4,5,19,21,*], Peter J. Campbell[1,19,21,*]

Affiliations

1. Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK
2. Department of Hematology, Erasmus MC Cancer Institute, 3015 CN Rotterdam, the Netherlands
3. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, UK
4. Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
5. The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
6. Department of Paediatrics, University of Cambridge, Cambridge, Cambridgeshire CB2 0QQ, UK
7. Blood Cells and Blood Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia
8. Department of Medical Biology, University of Melbourne, Parkville, VIC 3052, Australia
9. Division of Pediatric Hematology/Oncology, Stanford Medicine, Stanford, CA 94304, USA
10. Department of Pediatrics, London Health Sciences Centre, London, ON N6A 5W9, Canada
11. Division of Hematology-Oncology, IWK Health, Halifax, NS B3K 6R8, Canada
12. Department of Neurosurgery, Medical University of South Carolina, Charleston, SC 29425, USA
13. Department of Pediatrics, University Hospitals Cleveland, Cleveland, OH 44106, USA
14. Department of Paediatrics, WVU Medicine Children's Hospital, Morgantown, WV 26506, USA
15. Oncology R&D, AstraZeneca, Cambridge CB2 0AW, UK
16. Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
17. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5S 1A8, Canada
18. European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany
19. Division of Hematology/Oncology, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
20. Department of Haematology, University of Cambridge, Cambridge, Cambridgeshire CB2 2XY, UK
21. These authors contributed equally

***Addresses for correspondence and material requests:**

Dr Peter J. Campbell, Cancer Genome Project, Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom. Telephone: +44 (0) 1223 834244. e-mail: pc8@sanger.ac.uk

Dr Uri Tabori, The Hospital for Sick Children, Toronto, ON, Canada. e-mail: uri.tabori@sickkids.ca

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Mismatch repair (MMR) is a critical defence against mutation, but we lack quantification of its activity on different DNA lesions during human life. We performed whole-genome sequencing of normal and neoplastic tissues from individuals with constitutional MMR deficiency to establish the roles of MMR components, tissue type and disease state in somatic mutation rates. Mutational signatures varied extensively across genotypes, some coupled to leading-strand replication, some to lagging-strand replication and some independent of replication, implying that the various MMR components engage different forms of DNA damage. Loss of *MSH2* or *MSH6* (MutSα), but not *MLH1* or *PMS2* (MutLα), caused 5-methylcytosine-dependent hypermutation, indicating that MutSα is the pivotal complex for repairing spontaneous deamination of methylated cytosines in humans. Neoplastic change altered the distribution of mutational signatures, particularly accelerating replication-coupled indel signatures. Each component of MMR repairs 1-10 lesions/day per normal human cell, and many thousands of additional events during neoplastic transformation.

Highlights

- MMR repairs 1-10 lesions/day in every normal cell and thousands more in tumor cells
- MMR patterns and rates are shaped by genotype, tissue type and malignant transformation
- MSH2 and MSH6 are pivotal for repairing spontaneous deamination of methylated cytosine
- Replication indels and substitutions vary by leading versus lagging strand and genotype

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Protecting the integrity of our genome is vital to the normal functioning of somatic cells and reducing their risk of transforming to cancer. The genome is under constant assault from endogenous mutagens, such as oxygen free radicals and aldehydes; from errors of DNA replication, such as those acquired in cell division or DNA repair; and from exogenous mutagens, such as ultraviolet light and chemicals in tobacco smoke. With such diverse threats to our genome, a similarly diverse set of DNA repair mechanisms have evolved, each targeting particular types of DNA damage.

The mismatch repair (MMR) pathway is a major defence against errors introduced during DNA replication by leading- and lagging-strand polymerases (Pol-ε and Pol-δ respectively). The MMR complex MutSα, comprising a heterodimer of MSH2 and MSH6, performs surveillance by sliding along histone-free DNA, engaging mismatched bases, damaged DNA or insertion/deletion loops (Kunkel and Erie, 2015). It then pairs with the MutLα complex, a heterodimer of MLH1 and PMS2, which recruits a cascade of DNA repair factors that ultimately excise a section of the lesion-containing nascent strand (Hombauer et al., 2011; Kadyrov et al., 2006). MutSα and MutLα are the major heterodimers participating in MMR, but others also contribute in eukaryotes. Perhaps most well studied is the MutSβ heterodimer of MSH2 and MSH3, which particularly targets large insertion/deletion loops and single-base-pair deletions (Johnson et al., 1996; Romanova and Crouse, 2013). It remains unclear what the relative activities of these different components are in normal human cells and during transformation.

There are also potential roles of mismatch repair in protecting genomes outside of DNA replication (Crouse, 2016; Meier et al., 2018). While there is a clear differential between template and nascent strands that guides MMR during replication, this differential is not so obvious in interphase. For a given mismatch, if MMR cannot distinguish which strand has the correct sequence, there is a risk that the repair will use the incorrect sequence as the template, fixing the mismatch as a mutation on both strands. This possibility is actually harnessed physiologically during somatic hypermutation of the immunoglobulin loci in B cells (Zubani et al., 2017), but may also introduce somatic mutations predisposing to cancer or trinucleotide repeat expansions (Peña-Diaz et al., 2012).

The detailed biochemistry of MMR pathways has been pieced together from *in vitro* experimental studies and animal models (Johnson et al., 1996; Lujan et al., 2014; Meier et al., 2018; Romanova and Crouse, 2013; Serero et al., 2014; St Charles et al., 2015). What role

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MMR plays in humans *in vivo* has largely been inferred from extracting mutational signatures from cancers with deficiency of one component (Alexandrov et al., 2020; Campbell et al., 2017; Chung et al., 2020; Meier et al., 2018). This has identified several signatures associated with MMR deficiency, and has also demonstrated the role of MMR in protecting strategically important regions of the genome (Frigola et al., 2017; Supek and Lehner, 2015). However, these studies have focused on somatic mutations in established cancers, meaning it can be difficult to disentangle consequences of increased cell division rates, the emergence of additional mutational signatures during transformation and other endogenous mutational processes active before MMR deficiency was acquired (Alexandrov et al., 2020; Haradhvala et al., 2018; Meier et al., 2018).

Common to both experimental studies of MMR and analyses of human cancer is the utility of studying mutational signatures in knock-outs of individual pathway components. MMR deficiency in human cancers is mostly either sporadic (both alleles somatically mutated or silenced) or inherited as an autosomal dominant condition (Lynch syndrome), where one allele is inactivated in the germline and the other allele lost somatically in the tumor. Rarely, however, individuals are born with both alleles of an MMR gene inactivated in the germline, known as constitutional mismatch repair deficiency (CMMRD). These individuals have a substantially increased risk of early-onset cancers, especially cancers of the brain, gastrointestinal system and blood (Shlien et al., 2015; Tabori et al., 2017). Here, we harnessed recently developed protocols to identify mutations in normal somatic cells (Brunner et al., 2019; Ellis et al., 2020; Lee-Six et al., 2019) to evaluate individuals with CMMRD. Studying lifelong mutation accumulation in normal cells, in neoplastic lesions, in different organ systems and in different genotypes provides direct quantification of the importance of MMR in day-to-day life and the types of DNA lesion repaired by individual pathway components in humans.

Results

MMR Pathways Repair 1-10 Lesions/Day in Every Intestinal Epithelial Cell

Ten CMMRD donors, aged between 3 and 32 years, were studied, comprising 2 with *MSH2*, 4 with *MSH6*, 2 with *MLH1* and 2 with *PMS2* biallelic germline deficiency (**Table S1**). Colonic (n=7), small bowel (n=6) and brain (n=3) samples were collected during endoscopy, surgical resection or autopsy. Matched normal and neoplastic intestinal epithelium samples (adenomas or adenocarcinomas) were available for 4 individuals.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Intestinal crypts contain 5-15 stem cells at their base, with single stem cells frequently sweeping to fixation through genetic drift (Lopez-Garcia et al., 2010). As a result, all epithelial cells in individual crypts derive from a single, recent common ancestor, meaning that somatic mutation burdens, rates and signature composition can be accurately estimated by sequencing single crypts (Lee-Six et al., 2019). We therefore isolated 80 normal intestinal crypts, 30 neoplastic intestinal crypts and 25 samples from other tissues by laser capture microdissection, followed by whole genome sequencing (WGS) at a median of 30-fold coverage to investigate somatically acquired mutations. We called single base substitutions (SBS), small insertions and deletions (indels), structural variants (SV), telomere length and copy number variations (CNV) using well-established bioinformatic algorithms (Campbell et al., 2008; Jones et al., 2016; Raine et al., 2015) (**Table S2**). These data were compared to previously published analyses of somatic mutations in normal intestinal epithelium from mismatch repair-proficient individuals (Lee-Six et al., 2019) (**Figure S1**).

The burdens of single base substitutions were considerably increased in the normal intestinal epithelium of CMMRD donors, across all genotypes, compared to healthy individuals (**Figure 1a**). Even with the young age of our cohort, individual normal intestinal epithelial cells frequently had tens of thousands of mutations. The SBS mutation rate varied extensively between individuals, ranging between 250 and 1450 SBS/year (**Figure 1b**), which is 5 to 30-fold higher than the 49 SBS/year rate observed for normal colonic epithelium. However, within an individual, SBS mutation burdens exhibited limited variation, suggesting that genotype and environmental modifiers are stronger influences on mutation rates than cell-to-cell variation.

Small indel mutation burdens were similarly increased in the normal intestinal epithelium of all CMMRD donors, with an even greater magnitude of increase than for base substitutions. Rates ranged between 60 and 1200 indels/cell/year (**Figure 1b**), compared to 1 indel/cell/year for normal intestinal epithelium without MMR deficiency. As seen for base substitutions, between-individual variation was more pronounced in non-transformed epithelial cells than within-individual variation, again attributable to genotype. Comparing the number of base substitutions to indels, we found that *MSH2*, *MLH1* and *PMS2*-deficient cases all exhibited a ratio of approximately 1:1 – namely, one indel is acquired for every base substitution acquired in these normal cells (compared to 1 indel for every 49 substitutions in MMR-proficient normal epithelium; **Figure 1c**). For *MSH6* deficiency, the relative mutation rate was similarly fixed between the two types of lesions, but with a ratio of 1 indel for every 4 substitutions. The lower rate of indels relative to substitutions for *MSH6* deficiency indicates that the MutSβ

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

heterodimer (MSH2 and MSH3) repairs ~3x as many indels as MutSα (MSH2 and MSH6) in normal human intestinal cells.

Comparing the ratio of deletions to insertions, we found that *MSH2* and *MSH6*-deficient intestinal epithelium has a predilection for deletions, especially in polyA/T tracts, while in *MLH1*-deficient cases this ratio is at near-equilibrium and for *PMS2*-deficient cases insertions at polyA/T tracts are the predominant form of indel (**Figure S2a-b**). Upon neoplastic change the burden of deletions considerably outnumbers the insertion burden irrespective of MMRD genotype indicating that increased microsatellite instability (MSI) is predominantly defined by deletions, often at poly-A/T tracts (**Figure S2c**).

Comparing synchronous normal and neoplastic intestinal epithelium from the same individual showed that SBS burdens were 2 to 4-fold higher in neoplastic than normal intestinal epithelium, while indel burdens increased 4 to 10-fold (**Figure 1a**). As expected, telomere lengths of the intestinal crypts of neoplastic biopsies were significantly reduced compared to their normal counterpart (median normal and neoplastic telomere content, Mann-Whitney U test; PD40101, 478 vs. 287, P=0.0025; PD40103, 288 vs 181, P=0.0008; PD47719, 727 vs 514, P=0.008; PD47721, 432 vs 153, P=0.016) – this suggests that the increased cell division associated with malignant change drives both telomere shortening and increases in replication-associated somatic mutation.

Overall, then, in normal intestinal epithelial cells from patients lacking MMR, the combined rates of base substitutions and indels ranged from 500 to 3000 mutations per year per clone – this suggests that the mismatch repair pathway typically repairs an average of 1-10 events per day in every normal epithelial cell throughout the intestine.

MMR Repairs Replication-dependent and Independent DNA Damage

The mismatch repair pathway has variable efficiency for correcting different types of mismatch, being particularly tuned to events not already targeted by the proof-reading capacity of Pol-δ and Pol-ε polymerases (Lujan et al., 2014). Thus, we might expect MMR deficiency to be associated with distinct mutational signatures, with mutation rates varying by specific base changes, indels, local sequence context and replication strand. To assess this, we extracted mutational signatures jointly for all microdissected samples, including both healthy and neoplastic, using the TensorSignatures algorithm (Vöhringer et al., 2020). TensorSignatures extracts mutational signatures from a broad set of local and global

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequence information enabling the coupling of signatures of substitutions with indels as well as replication and transcriptional strand biases.

In total, 6 mutational signatures were extracted from the sequencing data of all microdissected samples, which we designate as signatures MMR-1 to MMR-6 (**Figure 2**, **Table S3**). Although these signatures partially match signatures previously extracted from cancer genomes and experimental models with MMR deficiency (Alexandrov et al., 2020; Meier et al., 2018; Serero et al., 2014; St Charles et al., 2015), the use of normal and transformed clones from cases with congenital deficiency enables clearer delineation of patterns, and isolation of individual components of previously inseparable signatures.

Signature MMR-1 is characterised by C>T substitutions at CG dinucleotides, likely due to deamination of 5-methylcytosine (5mC). It closely matches signature SBS1, extracted from cancer genomes (Alexandrov et al., 2020), which has previously been linked with MMR deficiency (Meier et al., 2018; Supek and Lehner, 2015). MMR-1 has minimal contribution from indels and lacks transcription and replication strand biases. It was broadly present across all samples but was especially prominent in patients with defects in MutSα, namely *MSH2* or *MSH6* deficiency (**Figure 2d**, **Table S4**). Deamination of 5mC to thymine occurs as a spontaneous chemical reaction in DNA (Shen et al., 1994) throughout the cell cycle. Such events occurring in interphase require repair before DNA replication, otherwise the new thymine becomes fixed as a C>T mutation in the genome. The predominance of this signature in patients with MutSα deficiency suggests that a major role of the MSH2/MSH6 complex is to identify G:dT mismatches for repair in interphase, independent of its activities during replication. The paucity of this signature in the normal and neoplastic tissues from *MLH1* and *PMS2*-deficient patients implies that the repair occurs through pathways independent of MutLα, likely through nucleotide or base excision repair.

Signature MMR-2 is defined by both base substitutions and indels, predominantly A/T insertions at polyA/T tracts, with the base substitutions being specifically biased towards leading-strand replication (using the pyrimidine of each base-pair as the reference base). Signature MMR-3 is its counterpart, with substitutions being more frequent on the lagging strand of replication, comprising largely C>T and T>C transitions. In contrast to MMR-1, signatures MMR-2 and MMR-3 were most enriched in those patients with defects in MutLα (*MLH1* or *PMS2*). The distribution of base substitutions in MMR-3 closely matches signature

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

246 SBS26 extracted from cancer genomes, whereas MMR-2 has only partial similarity to SBS15
247 (Alexandrov et al., 2020).

248

249 Signature MMR-4 strongly resembles a combination of SBS17a and SBS17b as seen in
250 cancer genomes of the GI tract, with high rates of T>C in a CTT context. This signature has
251 been associated with 5-fluorouracil treatment and misincorporation of 8-hydroxy-dGTP
252 opposite adenines (Christensen et al., 2019; Pich et al., 2019; Supek and Lehner, 2017).
253 Signature MMR-5 has partial similarity to SBS20 distinguished by C>A and C>T base
254 substitutions and deletions. Interestingly in this signature, C>T events are biased towards
255 leading-strand replication and C>A events towards lagging-strand replication. Finally,
256 signature MMR-6 is nearly completely defined by A/T-base deletions at polyA/T tracts. It is
257 more pronounced in *MSH2* deficiency than *MSH6* deficiency, suggesting it may capture the
258 repair activities of the MutSβ heterodimer (MSH2 and MSH3).

259

260 The observed replication strand biases of MMR-2, MMR-3 and MMR-5 could in principle be
261 of opposite strandedness when the complementary purine bases are considered as the
262 reference, such as G>A on the leading strand instead of C>T on the lagging strand. Further
263 insight can be derived from cases with combined MMRD and proofreading-defective
264 replication polymerases from The Cancer Genome Atlas (TCGA). Given that Pol-δ and Pol-ε
265 are the primary polymerases for lagging and leading-strand replication respectively, any
266 overlap in mutational signatures observed in these cancers and those extracted from the
267 CMMRD microdissections would provide evidence of their replicative origin. Cancers with
268 proofreading-defective POLD1 combined with MMRD have distinctive mutation profiles with a
269 predilection for C>A mutations at CCN as well as C>T and T>C transition mutations (**Figure**
270 **2b**). These mutational patterns resemble a combination of signatures MMR-3 and the C>A
271 component of MMR-5 with lagging-strand bias. Cancers with proofreading-defective POLE
272 combined with MMRD have mutation profiles with strong predilection for C>T mutations
273 predominantly at GCG trinucleotides, but also GCN (**Figure 2c**). MMR-2 and the C>T
274 component of MMR-5 with leading-strand bias recapitulate these mutational patterns. Taken
275 together, then, these data suggest that signature MMR-2 is indeed tied to leading-strand
276 replication; MMR-3 to lagging-strand replication; and MMR-5 has mixed contributions from
277 leading and lagging-strand replication.

278

279 **Mutational Processes Vary with Genotype and Cellular Transformation**

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We reconstructed phylogenetic trees showing clonal relationships among normal and neoplastic intestinal epithelium crypts (n=110). In the normal epithelium of MutSα-deficient CMMRD cases (3 *MSH6*; 1 *MSH2*), the predominant mutational process was that of methylated cytosine deamination, causing C>T mutations at CG dinucleotides (signature MMR-1; **Figure 2a** and **3**). For example, for case PD40101, the median contribution of this signature exceeded 80%, resulting in an average 16,000 methylation-deamination mutations per cell over 32 years of life. A corollary to the predominance of the deamination signature, MMR-1, in these normal epithelial cells is that the replication-dependent MMR signatures contributed considerably fewer mutations. In contrast, in synchronously collected intestinal polyps and colorectal cancers from these same patients, the replication-based signatures, MMR-2, MMR-3 and MMR-5, became much more evident, accounting for the overwhelming majority of excess mutations seen with cellular transformation. Thus, repair of methylated cytosine deamination is numerically the most important activity of MutSα in normal intestinal epithelium, whereas repair of replication-mediated errors becomes a much more significant activity with neoplastic transformation.

Loss of MutLα showed a different distribution of mutational signatures in normal intestinal crypts compared to MutSα-deficiency (**Figure 4**). MMR-2, marked by A/T-base insertions and leading-strand C>T mutations, was common to all MutLα-deficient intestinal crypts sequenced. MMR-3, with C>T and T>C transitions biased towards the lagging-strand, was similarly present in all intestinal crypts, but was more pronounced in *PMS2*-deficient than *MLH1*-deficient cases. This suggests that even among cases with deficiencies of components from the same MMR complex there are differences in the relative activity of mutational processes. Neoplastic change also impacted the composition of operative mutational processes, as seen in PD47719 (**Figure 4a**). Here, MSI was considerably increased in the transformed clone, as evidenced by the massive, sustained contribution of MMR-6 to the excess mutations seen on this branch of the phylogeny, including the most recent branches. The consequence of this was that, while insertions predominated over deletions in the normal epithelium of MutLα-deficient patients including PD47719, deletions considerably outnumbered insertions in this transformed clone.

Investigation of potential driver mutations revealed known oncogenic mutations in *FBXW7*, *ERBB2*, *ERBB3*, *PIK3CA*, *TP53*, *PPP2R1A, ACVR1B, ACVR2A, ATR, AXIN2* and *BAP1* in normal intestinal epithelium (**Figure 4**, **Figure S3a**, **Table S5**). Crypts from polyps and tumors showed the expected complement of *APC* mutations (**Figure S3a**). For all MutSα-deficient

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

neoplastic tissues, *APC* mutations were C>T mutations at CG dinucleotides, indicating that MMR-1 can induce driver mutations. Analysis of the driver mutation landscape across all intestinal crypts indicates that MMR-1 is the primary force of driver mutation acquisition, further complemented by microsatellite instability (**Figure S3b**). While MMR-proficient intestinal epithelial cells rarely acquire driver mutations, constitutionally MMR-deficient cells acquire more mutations, at younger ages, especially in cases of MutSα-deficiency (**Figure S3c**). This confirms that MMR-1 represents a major force of early driver acquisition in MutSα-deficient cells.

Activity of Mutational Processes in Other Normal Tissues

In addition to intestinal epithelium, we microdissected and sequenced 25 samples from 5 CMMRD donors across a wide array of tissues types, including brain cortex, lymphocyte aggregates, intraepithelial nerves and arterial walls. All MMRD genotypes, with the exception of *MLH1*, were represented. Sufficient numbers of SBS and indels were detected for mutational signature extraction (**Figure 5a**, **Table S6**). While mutation burdens varied considerably across the samples based on tissue type and level of clonality, we found that MutSα-deficient donors were characterised by high levels of MMR-1, the 5mC deamination signature, across different tissues; furthermore, this signature was again largely absent from MutLα-deficient samples (**Figure 5b**). This was especially evident in the brain cortex and striatopallidal fibers of the *MSH6*-deficient case PD40100 and to a lesser degree in the brain blood vessel of the *MSH2*-deficient case PD40098 (**Figure 5c**). The only exceptions were large intra-epithelial lymphoid populations in PD40103 (**Figure 5c**), which were enriched for the replication-associated mutational signature MMR-3. The clonality of these samples suggests that single B cells seeded the colonic intra-epithelial and expanded *in situ* to form the observed secondary lymphoid structures, which may explain why replication-mediated signatures predominated. Similar to the colonic epithelium, the tissue samples from MutLα-deficient cases revealed strong contributions of signatures MMR-2 and MMR-3.

Taken together, these data indicate that while MMRD genotype imposes a strong influence on the composition of mutational signatures across tissues, there are also detectable differences based on tissue type that could be the consequence of stem cell dynamics or environmental factors. This suggests that tissue-specific DNA damage influences the effects of MMRD on the observed mutational landscape.

Distribution of Mutations Across the Genome

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The MMR pathway is known to preferentially correct mutations in early replicating regions and even more efficiently in exons, meaning that its deficiency equalizes mutation rates across the genome (Frigola et al., 2017; Supek and Lehner, 2015). In our data, we found that this equalization of mutation rates by replication timing was more pronounced in cases of MutSα-deficiency (loss of either *MSH2* or *MSH6*) than MutLα-deficiency (loss of either *MLH1* or *PMS2*; **Figure S4a-b**). This difference was especially apparent in exonic regions, where MutSα-deficient cases acquired a median 50 exonic substitutions per year compared to a median 19 exonic substitutions per year for cases of MutLα deficiency. Thus, our data demonstrate that the targeting of mismatch repair to early replicating and exonic regions is largely dependent on MutSα rather than MutLα.

Across the genome, we observed that broad domains of late-replicating DNA had high relative mutation rates in cases PD40102, PD47719 and PD47722, all MutLα-deficient patients; a more pronounced variation than seen for MutSα-deficient patients (**Figure 6a-b**). In contrast, cases of MutSα-deficiency exhibit increased relative mutation rates at regions of high CpG density, presumably reflecting the greater opportunity for 5mC deamination in such regions (**Figure 6b**, **Figure S4c**). Indeed, using colonic whole genome methylation data, we found a strong linear relationship between the density of methylated CpG dinucleotides and the numbers of C>T mutations in a CG context (**Figure 6c**). Correcting the relative mutation rates of CG>TG for overall CpG density only partially removed differences across genomic elements; however, correcting for methylated CpG density completely eliminated these differences (**Figure 6d**).

Regions with high CpG density tend to be early replicating and more gene-dense, suggesting that some of the apparent equalization of mutation density in MutSα-deficient cases could be caused by balanced effects of replication-dependent signatures MMR-2 and MMR-3 in late-replicating regions versus MMR-1 in early-replicating regions. We therefore studied the distribution of individual MMR signatures across the genome. In MutSα-deficient intestinal epithelium, MMR-1 contributes to >80% of somatic mutations in early-replicating regions (**Figure S5**, **Table S7**) and >90% in coding sequence (CDS) (**Figure S6**, **Table S8**), whereas replication-dependent mutational processes are more evident in later-replicating regions or outside the coding genome. Mutations arising from MMR-1 in the most functional genomic regions is much more a feature of MutSα-deficiency than MutLα-deficiency (for earliest-replicating genome: effect size 4.25; $CI_{95\%}$=1.94-6.51; *P*=0.01; for CDS: effect size 2.98;

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

384 $CI_{95\%}$=1.52-4.41; *P*=0.01; generalized linear mixed models comparing MutSα-deficient and
385 MutLα-deficient samples).

386

387 In summary, then, mutations arising from deamination of methylated cytosine (MMR-1)
388 concentrate in the most functional coding and early replicating regions of the genome,
389 whereas mutations arising from replication errors (MMR-2, MMR-3, MMR-5) predominantly
390 distribute to late-replicating regions of the genome. In the absence of MutSα, these
391 distributions balance almost exactly, leading to broadly equal mutation rates across the
392 genome, whereas in MutLα-deficiency, the replication-associated signatures are predominant
393 such that mutation rates are higher in late-replicating regions.

394

395 **Verification of Genotype-specific Mutational Signatures in Sporadic Cancers**

396 Our investigation into subjects born without mismatch repair reveals that the mutational
397 consequences are shaped by MMRD genotype, tissue type and neoplastic transformation. To
398 assess whether these observations generalize, we analysed sequencing data from sporadic
399 MMRD tumors, collated from TCGA, CMMRD cancers and cancers sequenced in-house. The
400 germline and somatic landscape were screened for substitutions, indels, SVs and CNVs that
401 perturb any of the 4 genes encoding the core MMR pathway. Cases with disruptive MMR
402 mutations (nonsense, frameshift, splice site, copy number loss or structural variation) present
403 as homozygous, compound heterozygous variants or affecting two MMR genes were
404 identified. For missense substitutions, only clinically relevant variants recorded in ClinVar,
405 InSiGHT or the literature were classified as pathogenic. Cases with documented MMRD or
406 those with high mutation burdens and at least one MMRD-associated mutational signature
407 were selected for further characterization.

408

409 With this approach, we identified 117 cases with an established MMRD genotype, spanning
410 endometrial cancer, colorectal adenocarcinoma and gastric cancer most frequently, but also
411 occasional cases of prostrate adenocarcinoma, breast cancer, brain cancers, sarcomas and
412 other cancers (**Figure 7a**, **Table S9**). Sporadic MMRD cancers were most often defined by
413 the loss of a single MMR component, although disruptive mutations affecting two MMR genes
414 were also seen, with *MSH2* and *MSH6* being the most common co-mutated pair (**Figure 7b**).
415 We used Dirichlet regression on the mutational signature composition to infer associations
416 between mutational signature exposure and MMRD genotype. *PMS2*-deficient cases were
417 used as reference. Replicating the observation made in congenital cases, activity of signature
418 MMR-1, reflecting methylated cytosine deamination, was significantly greater in tumors with

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

loss of *MSH2* (P=1.66x10^{-5}) or *MSH6* (P=1.64x10^{-5}; **Figure 7c**) (**Table S10**). Exposure to MMR-5, found in MutSα and *MLH1*-deficient CMMRD tissues especially upon neoplastic change, was similarly detected in sporadic MMRD cancers with these genotypes (*MLH1*: P=1.3x10^{-3}; *MSH2*: P=1.34x10^{-4}; *MSH6*: P=0.02; **Figure 7c, Table S10**). Comparing MutSα-deficient to MutLα-deficient cancers revealed that the association between MutSα loss and increased MMR-1 activity could be replicated across different cancer subtypes (all, endometrial, colorectal and rare MMRD cancers; **Figure S7a**, **Table S10**). The association between MutSα loss and MMR-5 activity was confirmed in endometrial and colorectal MMRD cancers (**Figure S7b**, **Table S10**).

As previously reported (Touat et al., 2020), defects in *MSH2* and *MSH6* were strongly associated with glioblastoma multiforme in our dataset. As a result, the mutational spectrum of these tumors was dominated by C>T mutations at CG dinucleotides (**Figure 7d**). For colorectal and endometrial cancers, cases included both MutSα-deficient and MutLα-deficient genotypes, and the mutation spectrum differed accordingly. Overall, then, the differences across genotype and tissue type in CMMRD individuals generalize to sporadic MMR deficiency.

Discussion

The mismatch repair pathway has long been recognised as a major bulwark in the defence of our genome – our data now quantify the magnitude of its activity in the day-to-day life of humans. We see rates of 500-3000 base substitutions and indels per year per cell in patients with deficiency of a single component of the pathway, compared to ~50/year in normal individuals (Lee-Six et al., 2019). Furthermore, we observe considerable differences in mutational signatures between individuals inheriting deficiency of MutSα and those without MutLα. Together, these observations suggest that each component of the MMR pathway normally repairs 1-10 potentially mutagenic events every day in every intestinal cell, on average, and the two major components target different forms of DNA damage or replication error. This is in normal cells – once a clone begins to exhibit neoplastic growth, MMR becomes an even more critical defence, likely repairing tens of thousands of additional events over short periods of time.

Mismatch repair is classically described as a replication-linked process – errors introduced by replication polymerases, both substitutions and indels, are recognised by MutSα, triggering recruitment of MutLα, excision of the mismatch from the nascent strand and repair by

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

454 replication from the template strand (Kunkel and Erie, 2015). In our data, the MMR-2 and
455 MMR-3 signatures likely capture this canonical replication-linked repair – the two signatures
456 are found in both MutSα and MutLα deficiency, show clear biases by leading or lagging strand
457 of replication respectively, and vary in intensity across the genome by replication timing. The
458 two signatures reflect the different patterns of replication errors made by leading- and lagging-
459 strand polymerases, with small insertions dominating the leading-strand signature (MMR-2)
460 and T>C transitions dominating the lagging-strand signature (MMR-3). Beyond MutSα, it is
461 known that MutSβ, comprising MSH2 and MSH3, also signals to MutLα to repair errors
462 introduced during replication, especially deletions (Romanova and Crouse, 2013). We find that
463 the MutSβ pathway is numerically more important as the MutSα pathway for preventing indels
464 in normal human cells, evidenced by the observation that the rate of indels in deficiency of
465 *MSH6*, not a component of MutSβ, is a quarter of the rate seen with deficiency of *MSH2*,
466 present in both MutSα and MutSβ, or with deficiencies of *PMS2* and *MLH1*, components of
467 MutLα (**Figure 1c**). The MMR-6 signature in part captures this MutSβ repair process, as it has
468 considerably lower contribution to *MSH6*-deficient cases than *MSH2* deficiency, and is
469 dominated by deletions.

470

471 An unexpected finding is that MutSα is the major repair pathway in human cells against
472 spontaneous deamination of methylated cytosine to thymine, a repair activity that cannot be
473 linked to replication. Our evidence for this is that signature MMR-1 is numerically the dominant
474 process in MutSα deficiency, a signature comprising almost purely C>T substitutions at CpG
475 dinucleotides, with no replication strand or transcription bias, correlating linearly with the
476 fraction of methylated cytosines across the genome. Deamination of methylated cytosine is a
477 spontaneous hydrolytic reaction in double-stranded DNA (Shen et al., 1994). As a
478 spontaneous reaction, it will occur in all phases of the cell cycle, including interphase. Because
479 the deamination generates a thymine, if the G:dT mismatch is not repaired before replication,
480 the event will become fixed as a somatic mutation – hence, this repair activity must occur
481 before replication. Interestingly, we find no increase in the rates of this mutational signature in
482 MutSα-deficient cells after neoplastic transformation, consistent with it being independent of
483 replication. Furthermore, the tumor types with sporadic MMR deficiency showing the highest
484 rates of MMR-1 are those with low intrinsic cell division rates, tissues such as meninges,
485 pleura and glial cells (**Figure 7**).

486

487 In MutSα-deficient intestinal crypts, 350-1150 C>T substitutions at CpG dinucleotides accrued
488 every year. Spontaneous deamination of methylated cytosine is estimated to occur at rates of

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

489 $5.8x10^{-13}$/5mC/second at 37°C (Shen et al., 1994) – with methylated cytosines comprising
490 ~1% of the human genome (Ehrlich et al., 1982), this rate equates to ~1000 deamination
491 events per diploid genome per year. The striking correspondence of the expected number of
492 spontaneous deamination events with our observed mutation rates argues that MutSα is the
493 critical component of our defence against this mutational process. The paucity of MMR-1 in
494 patients with *MLH1* or *PMS2* deficiency implies that MutLα is not the partner for repair of these
495 deamination events. Instead, a strong candidate is *MBD4*, a thymine DNA glycosylase that
496 binds G:T mismatches in a methylated cytosine context (Hendrich et al., 1999). Germline
497 deficiency of *MBD4* causes tumors with remarkably elevated C>T mutation rates at CpG
498 dinucleotides, with mutation rates and mutation spectrum very similar to that of MMR-1 here
499 (Derrien et al., 2020; Pan-cancer Analysis of Whole Genomes Consortium, 2020; Sanders et
500 al., 2018).

501

502 These data demonstrate that studying somatic mutations in normal tissues from natural
503 human knock-outs can enrich observations from experimental model systems and cancers.
504 By comparing mutation rates to normal tissues in healthy controls, we have quantified the day-
505 to-day activity of MMR in normal human cells across tissue types; by comparing mutational
506 signatures and rates across genotypes, we have separated the relative contributions of
507 individual components of the MMR pathway to different types of DNA lesions; by comparing
508 normal and transformed cells from the same patients, we have shown the increased
509 dependence on MMR's protective activity caused by rapid cellular proliferation.

510

511 **ACKNOWLEDGMENTS**

512 We thank the staff of the Wellcome Sanger Institute Sample Logistics, Sequencing and
513 Informatics facilities for their contribution, including Laura O'Neill, Calli Latimer and Kirsty
514 Roberts for their support with sample management; Remco Hoogenboezem and Peter Valk
515 for help with analyses, advice and discussions; Stefan Dentro for help with analyses; and Iñigo
516 Martincorena for advice and discussions. Results presented here are in part based upon data
517 generated by the TCGA Research Network: https://www.cancer.gov/tcga.

518

519 **FUNDING**

520 This work was supported by a Cancer Research UK Grand Challenge Award (C98/A24032)
521 and the Wellcome Trust (206194). MAS is supported by a Rubicon fellowship from NWO
522 (019.153LW.038) and a KWF Kankerbestrijding young investigator grant (12797/2019-2, Bas
523 Mulder Award; Dutch Cancer Foundation). This research has been supported by the National

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Health and Medical Research Council of Australia (Project 1145912, IJM), the Cancer Council of Victoria (IJM), Victorian State Government Operational Infrastructure Support and the Australian Government NHMRC IRIISS. The Tabori lab research is supported by a Stand Up To Cancer - Bristol-Myers Squibb Catalyst Research Grant (SU2C-AACR-CT-07-17). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research Grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. This Tabori lab research is also supported by Meagan's Walk (MW-2014-10), b.r.a.i.n.child Canada, LivWise, a Canadian Institutes for Health Research (CIHR) Grant (PJT-156006), and the CIHR Joint Canada-Israel Health Research Program. PJC was supported by a Wellcome Senior Clinical Fellowship until 2020 (WT088340MA).

AUTHOR CONTRIBUTIONS

M.A.S., U.T. and P.J.C. conceived the study design. V.F., B.B.C., M.E., V.B., A.S. and U.T. recruited individuals, collected samples, and curated clinical data. L.M. and M.A.S. developed bespoke DNA library preparation methods. L.M., Y.H. and T.B. helped with sample preparation, tissue fixation, tissue sectioning and sample submission. H.L. and M.R.S. contributed and analysed control data. C.F., R.B. and I.J.M. provided help and advice on analysing TCGA data. U.M. provided sequencing data. P.S.R. and T.H.H.C. provided a bespoke procedure for unmatched variants filtering. H.S.V. and M.G. performed signature extraction and fitting. M.A.S., H.S.V., M.G. and P.J.C. performed data analysis. P.J.C. and M.G. oversaw statistical analyses. P.J.C. and M.A.S. drafted the manuscript, with input from V.F. and U.T.

CODE AVAILABILITY

Code required to reproduce the analyses in this paper are available online (https://github.com/MathijsSanders/MMR). Mutation calling algorithms are available through GitHub (https://github.com/cancerit). All other code is available from the authors upon request.

DATA AVAILABILITY

Whole genome and exome sequencing BAM files for cases reported in this study have been deposited in the European Genome-Phenome Archive (EGA) under accession code EGAS00001002881 (https://www.ebi.ac.uk/ega/home).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101.

Brunner, S.F., Roberts, N.D., Wylie, L.A., Moore, L., Aitken, S.J., Davies, S.E., Sanders, M.A., Ellis, P., Alder, C., Hooks, Y., et al. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature *574*, 538–542.

Campbell, B.B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., Davidson, S., Edwards, M., Elvin, J.A., Hodel, K.P., et al. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. Cell *171*, 1042-1056.e10.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. *40*, 722–729.

Christensen, S., Van der Roest, B., Besselink, N., Janssen, R., Boymans, S., Martens, J.W.M., Yaspo, M.-L., Priestley, P., Kuijk, E., Cuppen, E., et al. (2019). 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat. Commun. *10*, 4571.

Chung, J., Maruvka, Y.E., Sudhaman, S., Kelly, J., Haradhvala, N.J., Bianchi, V., Edwards, M., Forster, V.J., Nunes, N.M., Galati, M.A., et al. (2020). DNA polymerase and mismatch repair exert distinct microsatellite instability signatures in normal and malignant human cells. Cancer Discov.

Crouse, G.F. (2016). Non-canonical actions of mismatch repair. DNA Repair (Amst). *38*, 102–109.

Derrien, A.-C., Rodrigues, M., Eeckhoutte, A., Dayot, S., Houy, A., Mobuchon, L., Gardrat, S., Lequin, D., Ballet, S., Pierron, G., et al. (2020). Germline MBD4 Mutations and Predisposition to Uveal Melanoma. JNCI J. Natl. Cancer Inst. *113*, 1–8.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., Mccune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. *10*, 2709–2721.

Ellis, P., Moore, L., Sanders, M.A., Butler, T.M., Brunner, S.F., Lee-Six, H., Osborne, R., Farr, B., Coorens, T.H.H., Lawson, A.R.J., et al. (2020). Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. Nat. Protoc. 1–31.

Frigola, J., Sabarinathan, R., Mularoni, L., Muinõs, F., Gonzalez-Perez, A., and López-

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

592 Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. Nat.
593 Genet. *49*, 1684–1692.

594 Haradhvala, N.J., Kim, J., Maruvka, Y.E., Polak, P., Rosebrock, D., Livitz, D., Hess, J.M.,
595 Leshchiner, I., Kamburov, A., Mouw, K.W., et al. (2018). Distinct mutational signatures
596 characterize concurrent loss of polymerase proofreading and mismatch repair. Nat.
597 Commun. *9*.

598 Hendrich, B., Hardeland, U., Ng, H.-H., Jiricny, J., and Bird, A. (1999). The thymine
599 glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature
600 *401*, 301–304.

601 Hombauer, H., Campbell, C.S., Smith, C.E., Desai, A., and Kolodner, R.D. (2011).
602 Visualization of eukaryotic DNA mismatch repair reveals distinct recognition and repair
603 intermediates. Cell *147*, 1040–1053.

604 Johnson, R.E., Kovvali, G.K., Prakash, L., and Prakash, S. (1996). Requirement of the yeast
605 MSH3 and MSH6 genes for MSH2-dependent genomic stability. J. Biol. Chem. *271*, 7285–
606 7288.

607 Jones, D., Raine, K.M., Davies, H., Tarpey, P.S., Butler, A.P., Teague, J.W., Nik-Zainal, S.,
608 and Campbell, P.J. (2016). cgpCaVEManWrapper: Simple Execution of CaVEMan in Order
609 to Detect Somatic Single Nucleotide Variants in NGS Data. In Current Protocols in
610 Bioinformatics, (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp. 15.10.1-15.10.18.

611 Kadyrov, F.A., Dzantiev, L., Constantin, N., and Modrich, P. (2006). Endonucleolytic
612 Function of MutLα in Human Mismatch Repair. Cell *126*, 297–308.

613 Kunkel, T.A., and Erie, D.A. (2015). Eukaryotic Mismatch Repair in Relation to DNA
614 Replication. Annu. Rev. Genet. *49*, 291–313.

615 Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L.,
616 Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., et al. (2019). The landscape of
617 somatic mutation in normal colorectal epithelial cells. Nature *574*, 532–537.

618 Lopez-Garcia, C., Klein, A.M., Simons, B.D., and Winton, D.J. (2010). Intestinal stem cell
619 replacement follows a pattern of neutral drift. Science (80-.). *330*, 822–825.

620 Lujan, S.A., Clausen, A.R., Clark, A.B., MacAlpine, H.K., MacAlpine, D.M., Malc, E.P.,
621 Mieczkowski, P.A., Burkholder, A.B., Fargo, D.C., Gordenin, D.A., et al. (2014).
622 Heterogeneous polymerase fidelity and mismatch repair bias genome variation and
623 composition. Genome Res. *24*, 1751–1764.

624 Meier, B., Volkova, N. V., Hong, Y., Schofield, P., Campbell, P.J., Gerstung, M., and
625 Gartner, A. (2018). Mutational signatures of DNA mismatch repair deficiency in C. elegans

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

626 and human cancers. Genome Res. *28*, 666–675.

627 Pan-cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole
628 genomes. Nature *578*, 82–93.

629 Peña-Diaz, J., Bregenhorn, S., Ghodgaonkar, M., Follonier, C., Artola-Borán, M., Castor, D.,
630 Lopes, M., Sartori, A.A., and Jiricny, J. (2012). Noncanonical Mismatch Repair as a Source
631 of Genomic Instability in Human Cells. Mol. Cell *47*, 669–680.

632 Pich, O., Muiños, F., Lolkema, M.P., Steeghs, N., Gonzalez-Perez, A., and Lopez-Bigas, N.
633 (2019). The mutational footprints of cancer therapies. Nat. Genet. *51*, 1732–1740.

634 Raine, K.M., Hinton, J., Butler, A.P., Teague, J.W., Davies, H., Tarpey, P., Nik-Zainal, S.,
635 and Campbell, P.J. (2015). cgpPindel: Identifying Somatically Acquired Insertion and
636 Deletion Events from Paired End Sequencing. Curr. Protoc. Bioinformatics *52*, 15.7.1-
637 15.7.12.

638 Romanova, N. V., and Crouse, G.F. (2013). Different Roles of Eukaryotic MutS and MutL
639 Complexes in Repair of Small Insertion and Deletion Loops in Yeast. PLoS Genet. *9*,
640 e1003920.

641 Sanders, M.A., Chew, E., Flensburg, C., Zeilemaker, A., Miller, S.E., Al Hinai, A.S., Bajel, A.,
642 Luiken, B., Rijken, M., Mclennan, T., et al. (2018). MBD4 guards against methylation
643 damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML.
644 Blood *132*, 1526–1534.

645 Serero, A., Jubin, C., Loeillet, S., Legoix-Né, P., and Nicolas, A.G. (2014). Mutational
646 landscape of yeast mutator strains. Proc. Natl. Acad. Sci. U. S. A. *111*, 1897–1902.

647 Shen, J. cheng, Rideout, W.M., and Jones, P.A. (1994). The rate of hydrolytic deamination
648 of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. *22*, 972–976.

649 Shlien, A., Campbell, B.B., de Borja, R., Alexandrov, L.B., Merico, D., Wedge, D., Van Loo,
650 P., Tarpey, P.S., Coupland, P., Behjati, S., et al. (2015). Combined hereditary and somatic
651 mutations of replication error repair genes result in rapid onset of ultra-hypermutated
652 cancers. Nat. Genet. *47*, 257–262.

653 St Charles, J.A., Liberti, S.E., Williams, J.S., Lujan, S.A., and Kunkel, T.A. (2015).
654 Quantifying the contributions of base selectivity, proofreading and mismatch repair to
655 nuclear DNA replication in Saccharomyces cerevisiae. DNA Repair (Amst). *31*, 41–51.

656 Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate
657 variation across the human genome. Nature *521*, 81–84.

658 Supek, F., and Lehner, B. (2017). Clustered Mutation Signatures Reveal that Error-Prone
659 DNA Repair Targets Mutations to Active Genes. Cell *170*, 534–547.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

660 Tabori, U., Hansford, J.R., Achatz, M.I., Kratz, C.P., Plon, S.E., Frebourg, T., and Brugieres,
661 L. (2017). Clinical management and tumor surveillance recommendations of inherited
662 mismatch repair deficiency in childhood. Clin. Cancer Res. *23*, e32–e37.

663 Touat, M., Li, Y.Y., Boynton, A.N., Spurr, L.F., Iorgulescu, J.B., Bohrson, C.L., Cortes-
664 Ciriano, I., Birzu, C., Geduldig, J.E., Pelton, K., et al. (2020). Mechanisms and therapeutic
665 implications of hypermutation in gliomas. Nature *580*, 517–523.

666 Vöhringer, H., Hoeck, A. van, Cuppen, E., and Gerstung, M. (2020). Learning mutational
667 signatures and their multidimensional genomic properties with TensorSignatures. BioRxiv
668 850453.

669 Zubani, G.G., Zivojnovic, M., Smet, A. De, Albagli-Curiel, O., Huetz, F., Weill, J.C., Reynaud,
670 C.A., and Storck, S. (2017). Pms2 and uracil-DNA glycosylases act jointly in the mismatch
671 repair pathway to generate Ig gene mutations at A-T base pairs. J. Exp. Med. *214*, 1169–
672 1180.

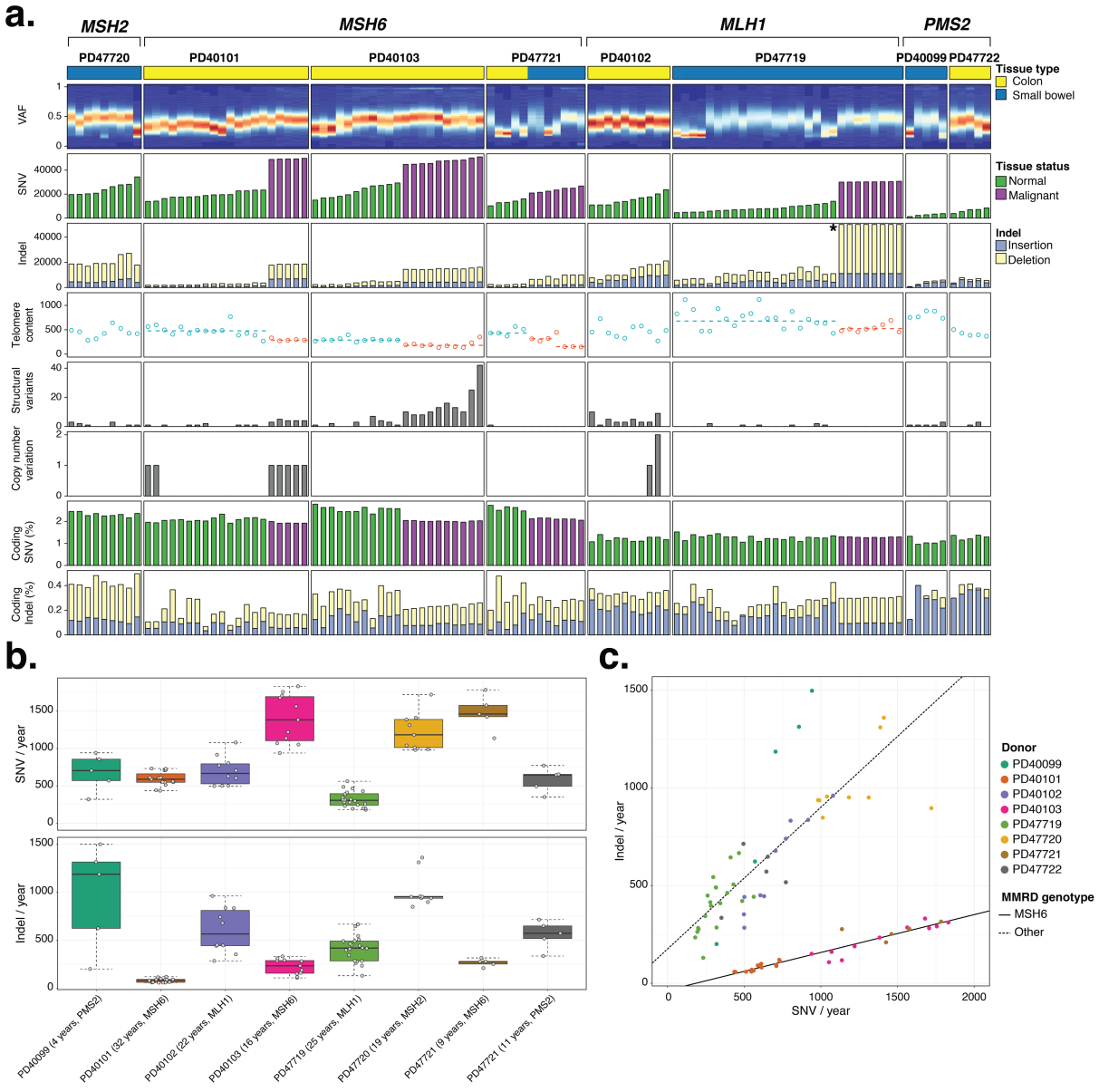
673

674

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

675 **FIGURES**

676



677

678 **Figure 1. Mutation Burden and Acquisition rates in Normal and Neoplastic CMMRD**
679 **Intestinal Crypts.**

680 **(a)** Each column represents an individual sample. From top to bottom: tissue type indicator,
681 VAF distribution heatmap, SNV burden coloured by disease state, indel burden separated on
682 insertions and deletions, telomere content, SV burden, CNV burden, percentage of coding
683 SNVs coloured by disease state, percentage of coding indels separated on insertions and
684 deletions. Dashed horizontal lines indicate median telomere content for matched normal and
685 neoplastic intestinal tissue samples.

686 **(b)** Annual SNV and indel rate for normal intestinal epithelium samples of each CMMRD donor
687 (samples from neoplastic tissues not shown). Boxplots mark median, inter-quartile range

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(IQR) and whiskers extend to 1.5 IQR from last quartile. Donor ID, age and MMRD genotype are displayed below.

(c) Relationship between annual SNV and indel rates. Individual samples are colour-coded based on donor origin. Solid line regresses the relationship for MSH6-deficient samples only. Dashed line regresses the relationship for all other MMRD genotypes.

*, y-axis is limited to 50,000 indels to accommodate visualisation of the indel burden for all samples.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

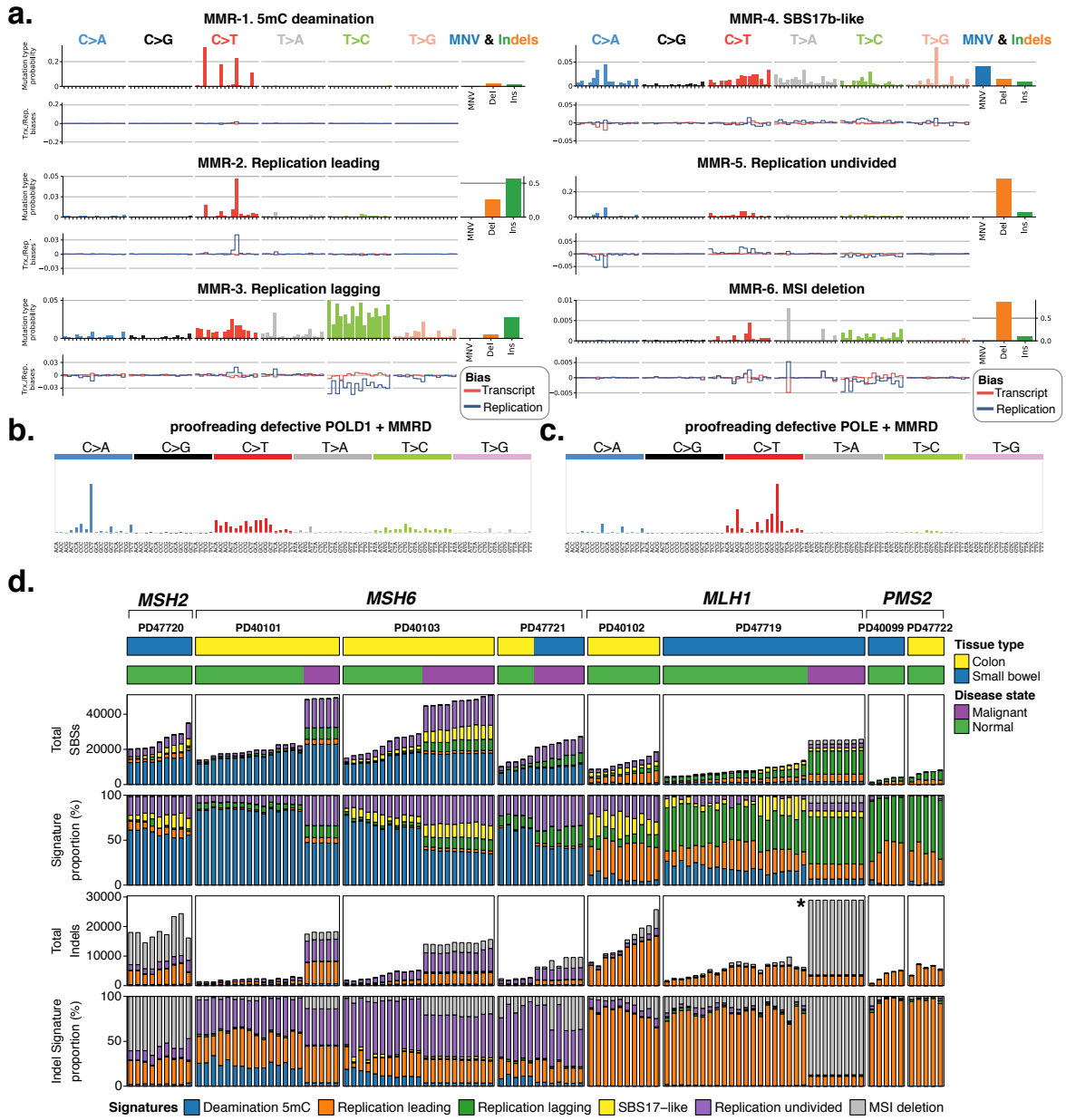


Figure 2. Mutational Signature Analysis Reveals DNA Damage Repair Specificity with Respect to MMRD Genotype and Disease State.

(a) The 6 mutational signatures extracted by TensorSignatures from all CMMRD intestinal crypts and other tissues samples. Mutational catalogues for intestinal crypts were extracted from variant phylogenies (**Figures 2-3**) to prevent duplication of variants presents in multiple related samples. For each signature, the left top panel shows probability for the 96 possible single base substitutions classified by trinucleotide contexts grouped into the 6 possible pyrimidine-based mutations. The right top panel shows mutation type probability for MNVs, insertions and deletions. Total contribution of all possible SNVs and other mutations sums to 1. The bottom panel shows transcriptional (red line) and replication (blue line) strand biases for each possible SBS type and context. Positive values indicate coding and leading strand

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

biases, respectively. Negative values indicate template and lagging strand biases, respectively.

(b) Typical SBS mutational profile for cancers from TCGA marked by both proofreading defective POLD1 and concomitant loss of MMR.

(c) Typical SBS mutational profile for cancers from TCGA marked by both proofreading defective POLE and concomitant loss of MMR.

(d) Each column indicates an individual sample. From top to bottom: tissue type indicator, mutational signature attribution for SBSs, proportional mutational signature attribution for SBSs, mutational signature attribution for indels, proportional mutational signature attribution for indels.

***:** y-axis is limited to 30,000 to accommodate the visualisation of mutational signature attribution for all samples.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

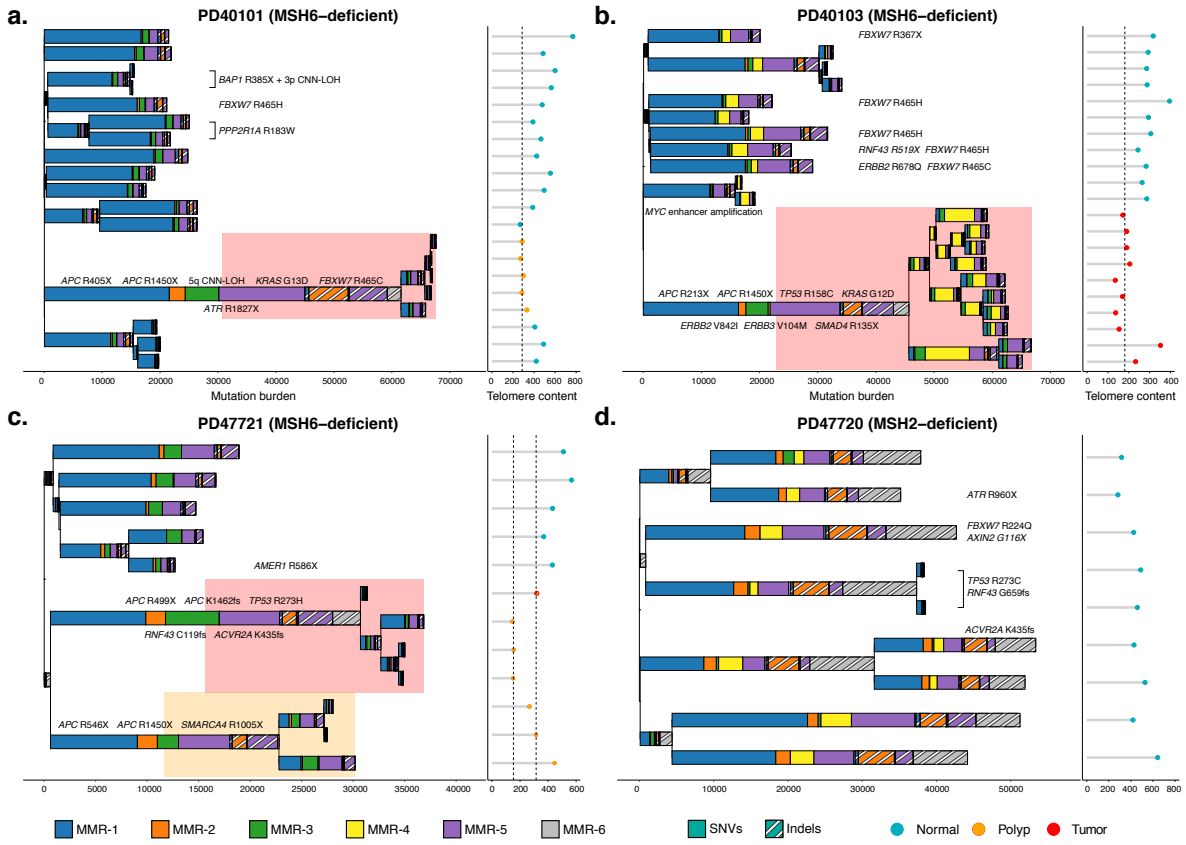


Figure 3. Phylogenetic Mutational Signature Attribution reveals Differential DNA Damage Repair Deficiency in Healthy and Neoplastic MutSα-deficient Intestinal Crypts. (**a-d**) Intestinal crypt phylogenies were inferred from all mutations (SNVs and indels). Branch length is proportional to the number of mutations detected. Branching indicates a historical crypt fission event. Stacked bar plots represent the mutational signature composition contributing to individual branches and colour-coded according to activity. Branches defined by <100 mutations were omitted from signature extraction. Stacked bar plot constituents: solid pattern – SNVs, stripped pattern – indels. Right panel: telomere content for individual crypts coloured by disease state. Dashed line indicates the median telomere content for neoplastic intestinal crypts. Coloured rectangular areas highlight branches formed by intestinal crypts microdissected from neoplastic tissues. Detected driver mutations are listed next to or along branches to indicate their presence among related intestinal crypts. **(a)** PD40101 (MSH6-deficient), phylogeny comprising normal (n = 12) and matched neoplastic intestinal crypts (n = 5). **(b)** PD40103 (MSH6-deficient), phylogeny comprising normal (n = 11) and matched neoplastic intestinal crypts (n = 10). **(c)** PD47721 (MSH6-deficient), phylogeny comprising normal (n = 5) and matched neoplastic intestinal crypts (n = 4, n = 3). Coloured area indicates the branches common to the 2 individual microdissected matched polyps. **(d)** PD47720 (MSH2-deficient), phylogeny comprising only normal intestinal crypts (n = 9).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

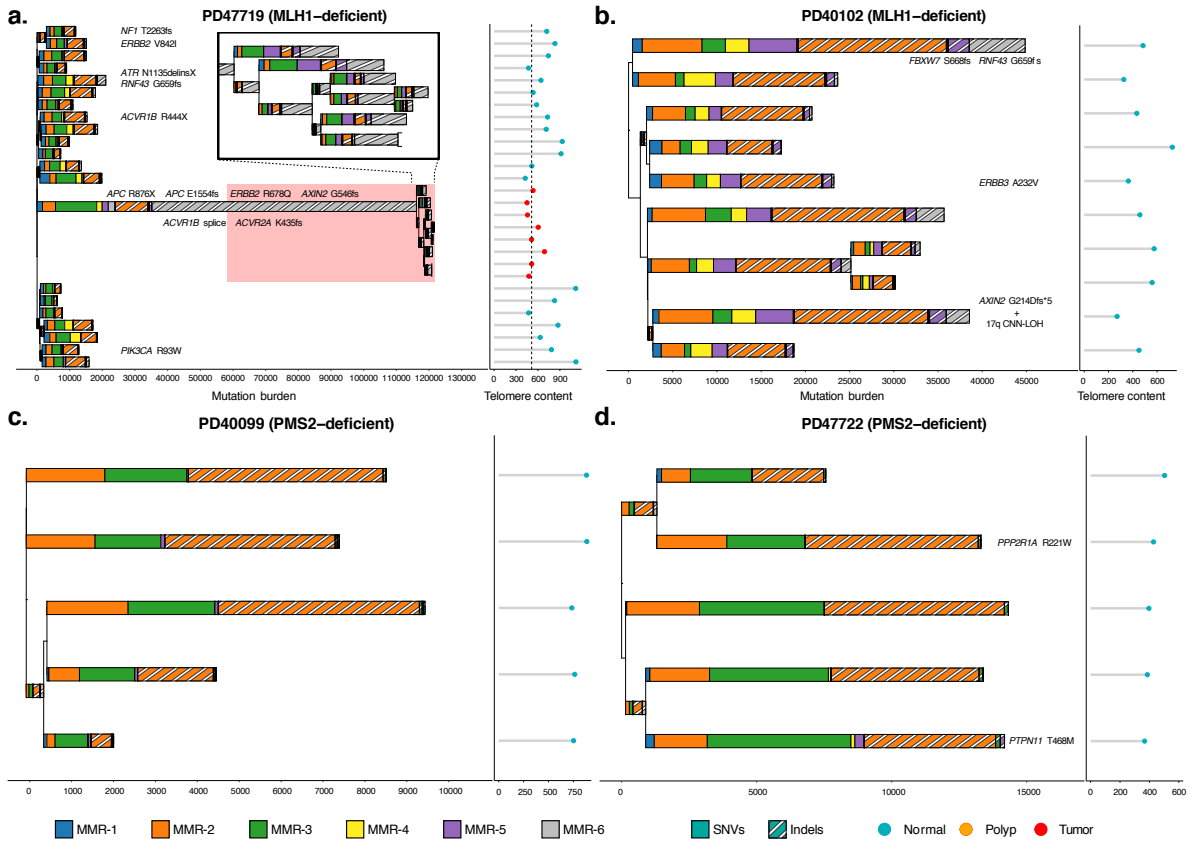


Figure 4. Phylogenetic Mutational Signature Attribution reveals Differential DNA Damage Repair Deficiency in Healthy and Neoplastic MutLα-deficient Intestinal Crypts. (**a-d**) Intestinal crypt phylogenies were inferred from all mutations. Branch length is proportional to the number of mutations detected. Branching indicates a historical crypt fission event. Stacked bar plots represent the mutational signature composition contributing to individual branches and colour-coded according to exposure. Branches defined by <100 mutations were omitted from signature extraction. Stacked bar plot constituents: solid pattern – SNVs, stripped pattern – indels. Right panel: telomere content for individual crypts coloured by disease state. Dashed line indicates the median telomere content for neoplastic intestinal crypts. Coloured rectangular areas highlight branches formed by intestinal crypts microdissected from neoplastic tissues. Detected driver mutations are listed next to or along branches to indicate their presence among related intestinal crypts. **(a)** PD47719 (MLH1-deficient), phylogeny of normal (n = 20) and neoplastic (n = 7) intestinal crypts. Coloured area indicates branches common matched microdissected colon cancer. Inset illustrates branches established in the final stages of colon cancer development. (**b**) PD40102 (MLH1-deficient), phylogeny of only normal (n = 10) intestinal crypts. (**c**) PD40099 (PMS2-deficient), phylogeny of only normal (n = 5) intestinal crypts. (**d**) PD47722 (PMS-deficient), phylogeny of only normal (n = 5) intestinal crypts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5. Mutational Signature Analysis of Other CMMRD Tissues Confirms DNA Damage Repair Specificity Among MMR Components.

(a) Each column indicates an individual sample. From top to bottom: tissue type indicator, VAF distribution heatmap, SNV burden, indel burden distinguished by insertions and deletions, telomere content, SV burden, CNV burden, percentage of coding SNVs, percentage of coding indels distinguished by insertions and deletions.

(b) Total and relative mutational signature contribution for microdissected non-intestinal samples as identified by the indicated colour-coding. Legend is shared with panel (**a**).

(c) Left: typical mutational spectra for different MutLα-deficient tissues. Right: typical mutational spectra for different MutSα-deficient tissues.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

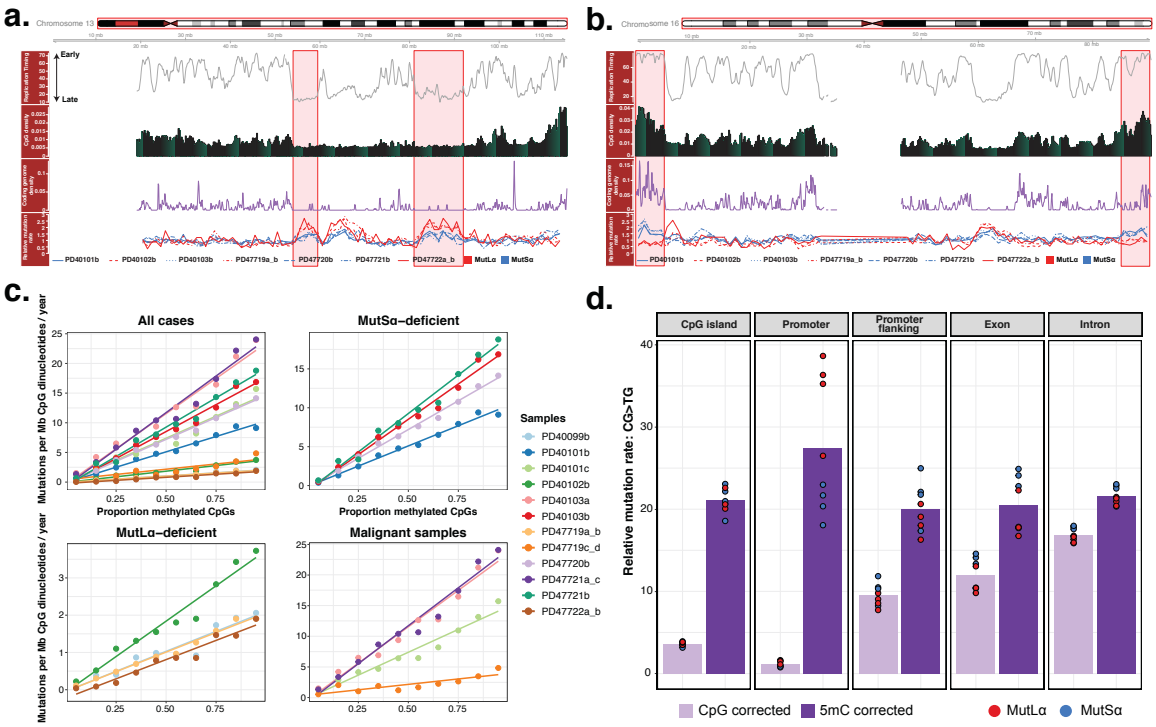


Figure 6. Determinants and Genomic Distribution of DNA Damage Typically Repaired by MMR.

(a) Distribution of determinants and mutations on chromosome 13 across non-overlapping 1 Mb bins. From top to bottom. (I) Average replication timing profile. High values indicates early replication and vice versa. (II) CpG-dinucleotide frequency. (III) Coding genome frequency. (IV) Average relative SNV rate of healthy intestinal crypts across the non-overlapping bins. Line type indicates individual donors whereas line colour indicates which MMR complex is deficient. Highlighted areas indicate broad domains of late genome replication. Case PD40099 was excluded on the basis of low genome-wide mutation counts.

(b) Distribution of determinants and mutations on chromosome 16 across non-overlapping 1 Mb bins. Information included and colour coding is the same as panel (**a**). Highlighted areas indicate areas of high CpG and coding genome density.

(c) Linear relationship between methylation levels and methylation damage. CpG-dinucleotides are grouped on observed methylation levels (methylation level bins) detected in a normal sigmoid colon. Mutation burden was determined across all bins and corrected for total CpG counts in millions and age. Solid line is the results of ordinary linear regression. Data for all microdissected healthy intestinal crypts and deficiency of individual MMR complexes is displayed in the top row and lower-left panel. Data for neoplastic samples are displayed in the lower-right panel.

(d) Relative mutation rates (RMR) of methylation damage (CG>TG) across different genomic elements further distinguished by MMR complex deficiency. RMR values are obtained per

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

795 genomic element by correcting the total CG>TG burden per sample by either CpG density or
796 methylation levels.
797

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

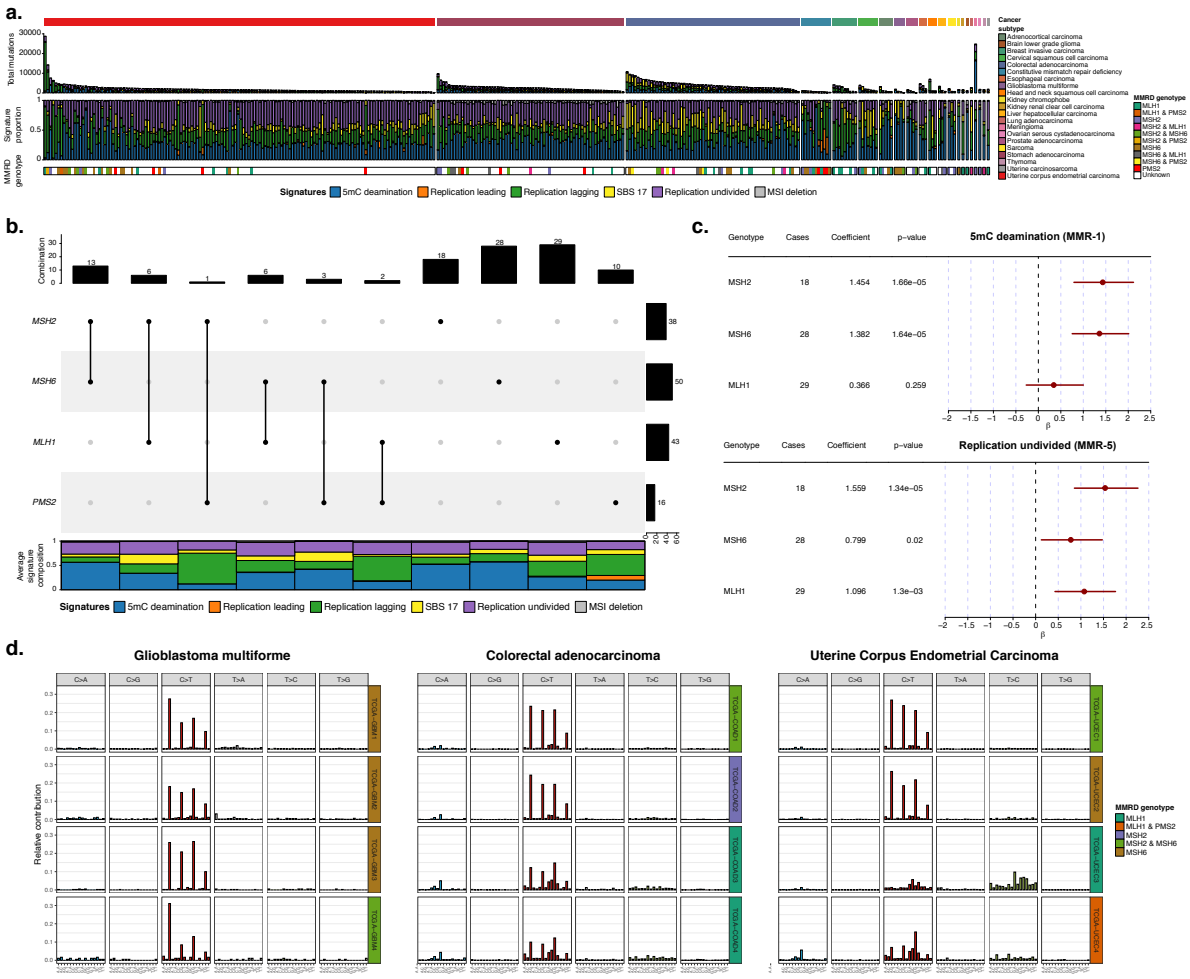


Figure 7. MutSα is a Primary Repair Force Engaging Methylation Damage as Confirmed in MutSα-deficient Cancers.

(a-e) MMRD status was determined across all TCGA cancer subtypes and supplemented by a CMMRD WES cancer cohort. MMRD genotype was determined by screening for loss-of-function or known damaging variants, CNVs and SVs involving MMR components. Cases with MMRD as a consequence of MMR component promoter silencing were listed as unknown.

(a) Mutation burden, mutational signature contribution and MMRD genotype for MMRD cancers present in the combined cohort. Each column represents an individual MMRD cancer. From top to bottom: cancer subtype, mutational signature attribution based on all mutations (SNVs and indels), relative mutational signature composition and MMRD genotype indicator. A white MMRD genotype indicator indicates either an unknown genotype or that MMRD is established through MMR gene promoter silencing.

(b) UpSet plot summarizing MMRD genotype counts across all genotyped MMRD cancers. Top: number of cases with the indicated MMRD genotype combination. Right: number of cases per MMR component lost. Bottom: average mutational signature contribution per indicated MMRD genotype.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(c) Forest plots summarizing the Dirichlet regression results. PMS2-deficient cancers were used as reference. Cancers with components lost from both MMR complexes (MutSα and MutLα) were excluded. Significant associations between MMR component loss and mutational signature activity was seen for the MMR-1 and MMR-5 signatures.
(d) SBS mutational profiles per indicated cancer subtypes across different MMRD genotypes. Coloured label on the right of each mutational profile indicates the established MMRD genotype. Left: mutational profiles of MMRD glioblastoma multiforme enriched for loss of MutSα. Middle: colorectal adenocarcinoma (COAD), top 2 rows are typical mutational profiles for MutSα-deficient cases while the bottom 2 rows are indicative for MutLα-deficient cases. Right: endometrial cancer, top 2 rows are typical mutational profiles for MutSα-deficient cases while the bottom 2 rows are indicative for MutLα-deficient cases.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

Human Tissue Samples

All human biological samples were collected with informed consent by the Biallelic Mismatch Repair Deficiency Syndrome (bMMRD) biobank located at The Hospital of Sick Kids, Toronto, Canada, in accordance with procedures approved by the Local Research Ethics Committee (REC reference 17/NW/0713). Samples were snap-frozen in liquid nitrogen or stored at -80 °C upon collection and stored at -80 °C after transfer to the bMMRD biobank at The Hospital of Sick Kids.

Tissue Preparation

Frozen tissues were first thawed at 4 °C for 10-15 min. Larger biopsies were dissected into smaller sections which were stored at -80 °C to safeguard materials for future experiments. Thawed tissues were fixed in 70% ethanol and embedded in paraffin by the Tissue-Tek VIP 6 AI tissue processing instrument (Sakure) based on an ethanol-only embedding protocol. Sections of 10µm thickness were cut from the paraffine-embedded tissue blocks and mounted onto poly-ethylene naphtholate (PEN)-membrane slides (Leica). Tissue slides were stained with the following haematoxylin and eosin (H&E) protocol in sequential steps: xylene (2 min, twice), ethanol (100%, 1 min, twice), deionized water (1 min, once), Gill's haematoxylin (10-20 s, once), tap water (20 s, twice), eosin (10 s, once), tap water (10-20 s, once), ethanol (70%, 20 s, twice), and xylene (10-20 s, twice). Tissue slides were cover-slipped after H&E staining for scanning with the NanoZoomer S60 Slide Scanner (Hamamatsu), generated images were viewed with NDP.View2 (Hamamatsu).

Laser Capture Microdissection

Areas of interest were identified from the scanned slides. Laser capture microdissection of these areas was performed with the LMD7000 microscope (Leica) and collected into a skirted 96-well PCR plate. Overview pre- and post-dissection images were stored for quality control. For some donors bulk samples from connective or muscle tissues were taken to be used as matched control. Cell lysis was achieved using 20µl of proteinase-K Picopure DNA Extraction kit (Arcturus). Thermocycler protocol: sample incubation at 65 °C for 3 hours followed by proteinase denaturation at 75 °C for 30 minutes. Processed samples were stored at -20 °C prior to submission for DNA library preparation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Low-input DNA Library Preparation and Sequencing

DNA library preparation was undertaken as previously described using a low-input DNA library preparation method (Ellis et al., 2021). This protocol allows for high quality DNA library preparation starting from low material quantities without whole genome amplification. DNA library concentration was assessed after library preparation, a minimum library concentration of 5 ng/µl was used to determine whether to take DNA sequencing forward. All LCM samples were subjected to 2x150bp paired-end whole genome sequencing. Samples from donors PD40098, PD40099, PD40100, PD40101, PD40102 and PD40103 were sequenced on the HiSeq X platform (Illumina), while samples from donors PD47719, PD47720, PD47721 and PD47722 were sequenced on the NovaSeq 6000 platform (Illumina) with the XP kit and using unique dual indexes. Pools of samples were sequenced to achieve a coverage of ≥ 30.

SBS Calling

Sequencing data were aligned to the NCBI human reference genome (GRCh37) using the Burrow-Wheeler Aligner (BWA-MEM) (Li and Durbin, 2009). Duplicated mate-pairs were marked for removal during downstream analysis. Single Base Substitutions (SBS) were called using 'Cancer Variants through Expectation Maximization' (CaVEMan) (Jones et al., 2016). Mutations were called against a synthetic unmatched normal control. Post-calling filters were utilized to remove low-quality variants, alignment errors, low-input DNA library preparation specific artefacts and germline variants as previously described (Lee-Six et al., 2019; Moore et al., 2020). The following filters were applied: (1) common single nucleotide polymorphisms and recurrent artefacts were removed by filtering against a panel of 75 unmatched normal samples (Nik-Zainal et al., 2012), (2) common artefacts and erroneous variants as a consequence of poor alignment were removed by applying a threshold for the median alignment score (ASMD ≥ 140) of reads supporting variants and less than half of the variant-supporting reads could be clipped (CLPM = 0), (3) by counting variant supporting fragments instead of reads to prevent counting variant-supporting mate-pairs double and (4) a filter that detects variants introduced due to incorrect processing of cruciform DNA during DNA library preparation. Software implementing these filtering steps can be found at https://github.com/MathijsSanders/SangerLCMFiltering.

A subsequent set of filters is applied to remove retained germline variants and potential artefacts. Code for performing these filtering steps can be found at https://github.com/TimCoorens/Unmatched_NormSeq. Mutations across all intestinal

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

899 samples were aggregated and read pile-up was performed with AnnotateBAMStatistics
900 (https://github.com/MathijsSanders/AnnotateBAMStatistics) to determine the fragment counts
901 of the mutant and wildtype alleles per mutation. Germline variants were filtered using an exact
902 binomial test. Per donor mutant counts were aggregated across all intestinal samples in
903 conjunction with the total fragment coverage. Taking the fragment variant and coverage
904 counts a variant is retained when it deviates significantly from the expected values from a
905 binomial distribution with p = 0.5. A second filter is used to filter potential remnant artefacts
906 present at lower variant allele frequencies (VAF). For somatic variants a significant
907 overdispersion is present when fitting a beta-binomial distribution. Variants consistently
908 present with similar VAFs across all included samples are likely erroneous and yield little
909 overdispersion. Therefore, a beta-binomial test was applied. Genuine variants were
910 determined by using a overdispersion threshold (rho) of $\geq$ 0.1. This process yielded the final
911 list of somatic variants used for downstream analysis.
912
913 **Indel Calling**
914
915 Indels were called by cgpPindel (Raine et al., 2015) using a synthetic normal, described under
916 'SBS calling', as unmatched control. Indels passing the standard filtering were subjected to a
917 quality score $\geq$ 300 and should be at least covered by 15 reads. Detected indels positioned in
918 large repeats (> 9bp) are flagged by cgpPindel and removed from the final list of somatic
919 variants. The number of mutant reads, wildtype reads and VAF was determined by cgpVAF.
920 After collating these statistics, a similar protocol for determining germline variants and possible
921 artefacts was used as described for SBS.
922
923 **Deletion-to-indel Ratio**
924
925 The deletion:indel ratio was determined for each normal intestinal crypt sequenced to assess
926 the linear relationship between these lesions dependent on the mismatch repair deficiency
927 (MMRD) genotype. A linear mixed model was constructed with lme4 (Bates et al., 2014) to
928 model the number of deletions as a dependent variable on the interaction effect
929 insertions:genotype (the number of insertions and MMRD genotype, fixed effect), donor label
930 (random intercept to account for multiple samples per donor) and genotype-dependent
931 insertions (random slope for insertions based on MMRD genotype). Crypts from neoplastic
932 intestinal tissues have a strong phylogenetic relationship resulting in similar numbers of indels.
933 The average deletion:indel ratio was calculated for these samples as an indication of which
934 indel lesion type are often acquired.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

935

936 **Structural Variant Calling**

937

938 Structural variants (SV) were called using a matched, but phylogenetically unrelated, sample
939 from the same donor with the 'Breakpoints Via Assembly' (BRASS) algorithm (Campbell et
940 al., 2008), with further annotation from GRASS (https://github.com/cancerit/BRASS). SVs
941 were annotated further by AnnotateBRASS
942 (https://github.com/MathijsSanders/AnnotateBRASS) as previously described in detail (Moore
943 et al., 2020). In brief, low-input DNA preparation results in the erroneous introduction of SV at
944 low read counts. AnnotateBRASS determines the following per SV: the number of supporting
945 read pairs, the variance in the alignment position of read pairs, whether read-pairs are clipped
946 or carry an excess of variants not reported in single nucleotide polymorphism databases, are
947 in the correct orientation or whether SV-supporting read-pairs are in regions marked by high
948 proportions of other read-pairs aligning to different parts of the genome (high homology). After
949 annotation SVs are filtered as previously described in detail
950 https://github.com/MathijsSanders/AnnotateBRASS (Moore et al., 2020).

951

952 **Copy Number Variation Calling**

953

954 Coverage and B-allele frequency (BAF) information was extracted for samples included in the
955 study by ConstructASCATFiles (https://github.com/MathijsSanders/ConstructASCATFiles).
956 Single nucleotide polypmorphisms (SNP) with a population frequency greater than 0.01 were
957 used for extracting coverage and BAF statistics across all samples. Coverage and BAF
958 statistics were grouped by donor and assessed for quality via 'QualityControl_and_PCA.R'
959 (https://github.com/MathijsSanders/PREASCAT). For each donor one or more control
960 samples are defined which are assumed to comprise fully diploid cells. This could either be
961 bulk polyclonal samples (connective tissue) or other tissues for which it is certain that the
962 genome is diploid. SNPs with limited coverage across all control samples are excluded from
963 the analysis. Samples are corrected for library size by assuming that heterozygous SNPs with
964 a balanced VAF (~ 0.5) represent diploid genomic areas. The LogR ratio is determined by
965 dividing the library-size corrected coverage for each samples with the median library-size
966 corrected coverage of the controls. The X chromosome is multiplied by 2 for male individuals
967 to correct for sex differences. The low-input DNA library preparation protocol on occasion
968 results in substantial variance in the BAF of heterozygous SNPs in the control samples. The
969 median BAF across the control samples is used to determine the set of heterozygous SNPs
970 for each donor. Principal component analysis (PCA) was applied to normalized LogR values
971 to detect systemic biases present in all samples or the samples of the donor of interest. Among

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such areas are the HLA locus and peri-centromeric regions and regions of high-vs-low CpG density. Final LogR and BAF statistics are calculated with 'construct_ASCAT_file.R' (https://github.com/MathijsSanders/ConstructASCATFiles). This algorithm follows the same approach but includes prinicipal component (PC) regression to assuage any systemic noise present. LogR profiles are centered around 0 by subtracting the mean value followed by PC regression to determine the absolute presence of noise. This fit is extracted from the 0-centered LogR profile followed by restoration to the previous average value by adding the mean value earlier used for subtraction. Finally, the resulting files are used in ASCAT for allele-specific CNV detection (Van Loo et al., 2010).

Telomere Content Estimation

Attrition of the telomeres is a hallmark of cellular aging. Assessment of telomere lengths for all microdissected samples, including neoplastic intestinal epithelium, was achieved by two distinct approaches.

Telomerecat v3.4.0 (Farmery et al., 2018) was used for estimating telomere lengths based on WGS data with length correction enabled and the number of simulations set to 100. In rare instances telomere lengths could not be estimated. One possible explanation is the use of the HiSeq X and NovaSeq 6000 instruments for distinct sets of microdissected samples. A second telomere length algorithm, TelomereHunter (Feuerbach et al., 2019), was used as an orthogonal approach which yielded results for all samples. Strong correlation was observed upon comparing the different telomere length estimation algorithms. Telomeres from neoplastic intestinal epithelium were significant shorter compared to their synchronously collected matched normal tissues indicative that neoplastic change results in increased proliferation rates and shorter telomeres. Telomere content estimates from TelomereHunter were reported because information was available for all samples.

Inference of Phylogenetic Trees

Phylogenies from all mutations, SBS and indels, of normal and neoplastic colonic crypts (n=110) was reconstructed for 8 CMMRD donors across all MMRD genotypes. SBS were called by CaVEMan as described under 'SBS calling', while indels were called by cgpPindel as described under 'Indel calling'. Finally mutations were recalled across all intestinal crypts samples for each donor using AnnotateBAMStatistics (SBS) or cgpVAF (indel). Variants with a VAF $\geq$ 0.3 were labelled as present (1), VAF $\leq$ 0.1 as absent (0) and a VAF between 0.1 and 0.3 (?) as ambiguous. This string of labels was used as input for phylogenetic tree

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reconstruction using a parsimonious approach in MPBoot (Hoang et al., 2018), bootstrap parameter set 1.000, and mutations were mapped to branches of the phylogenic tree using maximum likelihood estimation.

Mutational Signature Extraction

All classes of SBS, multi-nucleotide variants (MNV) and indels were used for mutational signature extraction. For normal and neoplastic intestinal crypts branches from the reconstructed phylogenetic trees were used as input rather than the complete per-sample mutational catalogue. Some intestinal crypts have strong phylogenetic relationship due to crypt fission at a later point in life. Using branches from the phylogenetic trees prevents mutations being duplicated in the same mutational signature extraction procedure. Additionally, using the branches as input gives insight into the changes of the mutational signature composition upon neoplastic change. The full mutational catalogue for non-intestinal epithelium microdissected samples is used due to little genetic kinship to other non-intestinal samples from the same tissue. Mutational signature analysis was performed using TensorSignatures (Vöhringer et al., 2020), which defines mutational signatures in terms of SBS, MNVs as well as indels. SBS are counted strand symmetrically (C>A, C>G, C>T, T>A, T>C and T>G) and were classified based on the 5' and 3' flanking nucleotide context. MVS and indels were classified by variant type and increasing length. TensorSignatures deconvolves mutational signatures further on the basis of genomic properties by grouping SBS by genomic contexts including transcription and replication orientation, epigenetic and nucleosomal context, and clustering status. Transcription states include the coding, template and unknown strand and were derived from GENCODE v19 annotations. Assignment to replication orientation (leading, lagging or unknown strand) were based on consensus annotations from Repli-seq profiles of GM12818, K564, HeLa, HUVEC and HepG2 cell lines (Hansen et al., 2010). Epigenetic states (TssA, TssAFlnk, TxFlnk, Tx, TxWk, EnhG, Enh, Znf/Rpts, Het, TSSBiv, BivFlnk, EnhBiv, ReprPC, ReprPCWk and Quies) stem from consensus ChromHMM annotations derived from the 127 ENCODE cell lines (Roadmap Epigenetic Consortium). SBS assignment to nucleosomal status (minor grooves facing away from and towards histones, and linker DNA between nucleosomes) were based on MNase cut efficiency data, locating conserved nucleosome dyad (midpoint) positions in human lymphoblastoid cell lines (Pich et al., 2018). Clustered SBS were identified using a 2-state Hidden-Markov model (Vöhringer et al., 2020). The analysis was performed for decomposition ranks ranging between 2-20 and the overdispersion parameters: 10, 20, 30, 50, 100, 1,000 and 10,000. The model with 6 (overdispersion = 10.000) mutational signatures was selected as suggested by the Bayesian information criterion estimator. Optimization of each model was

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

performed for 50.000 training epochs using an ADAM optimizer and an exponential decaying learning rate (starting learning rate 0.1).

Phylogenetic Tree Visualisation

The mutational composition per phylogenetic tree branches was estimated by TensorSignatures (Vöhringer et al., 2020). The structure of the reconstructed tree was determined as described under 'Inference of phylogenetic trees'. The fortified tree was plotted by ggtree (Yu, 2020), while the stacked bar plots were drawn upon the tree with ggplot2 and ggpattern (https://github.com/coolbutuseless/ggpattern). Length of branch indicates the number of SBS and indels detected, while the coloured segments of the stacked barplot indicates the contribution of each mutational signature. The SBS and indels were separated for each branch. Coloured solid segments indicate the number of SBS contributed per mutational signature while the coloured striped segments indicate the number of indels contributed per mutational signature. Telomere lengths were added with aplot and ggpubr.

Driver Mutation Calling

Driver mutation analysis was performed by taking two separate approaches. First, dNdScv (Martincorena et al., 2017) was used to identify mutations under positive selection by comparing the observed nonsynonymous:synonymous mutation ratio to the expected value after correcting for mutational context biases and local genomic information, such as replication timing or local epigenetic state. Mutations were extracted from the phylogenetic trees to prevent variant duplication. Settings were kept at default because none of the samples violated the predefined condition of having a moderate amount of coding mutations or an excess of mutations for a single gene. When including all dissected intestinal crypts, irrespective whether normal or neoplastic, yielded only the gene *APC* as significantly enriched (Benjamini-Hochberg corrected q-value $\leq$ 0.05) for disruptive mutations which is not unexpected given the inclusion of neoplastic intestinal tissues. For normal tissues none of the recurrently mutated genes were considered significantly enriched for non-silent mutations, even when performing restricted hypothesis testing.

Yet, there were multiple mutations detected in the normal intestinal crypts with amino acid changes often seen in intestinal cancers as reported by COSMIC (Forbes et al., 2008). In the second phase SBS and indels were collated across the normal and neoplastic intestinal crypts separately. Analysis was restricted to protein-coding changes and essential splice acceptor or

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

donor sites. Mutations were annotated with ANNOVAR (Wang et al., 2010), including COSMIC db v91, and compared to a list of 369 driver genes previously published (Martincorena et al., 2017). All types of premature truncating variants (PTV; frameshifts, stopgains or essential splice-site mutations) were considered sufficient for genes known to act as tumour suppressors or have been associated with frequent PTVs. Missense mutations should be located in hotspot locations previously reported in COSMIC in at least a few cancers. For instance, *FBXW7* is found recurrently mutated in this study, most often at known hotspot positions. However, a few missense mutations detected in this gene were reported in a low number of intestinal cancers. Even less missense mutations were not reported and could therefore not be considered genuine driver mutations. The frequency of driver mutations was compared between samples taken from of MutSα and MutLα deficient tissues. These frequencies were similarly compared to the driver mutation landscape of 445 MMR-proficient healthy intestinal crypts (Lee-Six et al., 2019), which were annotated and filtered using the same approaches.

Genomic Distribution Mutations

Genomic regions were masked based on the ENCODE blacklist to rule out errors due to read misalignment, regions of poor alignment or regions exhibiting anomalous coverage intensity signals in an approach similar as previously published (Supek and Lehner, 2015). The remainder of the unmasked genome was segregated into non-overlapping 100 kb (Figure S4B) or 1 Mb (Figures 6A-B and Figure S4C) windows and those for which a significant proportion (≥ 50%) fails the above criteria are removed from further analysis. The Y chromosome was similarly excluded from analysis. This masking process yielded 26396 100 kb or 2588 1 Mb windows passing the above criteria. Different from previous approaches (Supek and Lehner, 2015), exonic areas are included. SBS densities were calculated per donor and tissue combination. For instance the normal colonic epithelium of case PD40102. SBS can be aggregated across intestinal crypts under the assumption that mutation rates are approximately similar across the different intestinal crypts from the same tissue. SBS are extracted from the phylogenetic trees to prevent mutations being duplicated. For each donor and tissue combination, the SBS density of each window is determined by aggregating the SBS across all microdissected intestinal crypts and dividing the total count by the effective window length after masking. In the end there are 26396 or 2588 SBS density values for each donor and intestinal tissue combination. The relative mutation frequency (RMF) is determined by dividing the SBS density value by the mean SBS density value across all windows per donor and tissue combination.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A DNA replication timing profile was generated from 11 Repli-seq datasets from ENCODE (Consortium, 2012) downloaded through the UCSC genome browser (hg19). The data comprises genome-wide wavelet-smoothed Repli-seq values per 1kb bins for 11 cell lines (BG02ES, BJ, GM12878, Hela-S3, HepG2, HUVEC, IMR90, K562, MCF-7 and NHEK). The median value across all cell lines was calculated per 1kb window and from these median values the median value was calculated for 100 kb or 1 Mb non-overlapping windows. These non-overlapping windows were sorted based on the summarized Repli-seq value and distributed among 5 equally-sized domains, named the Earliest, Early, Median, Late and Latest replication timing domains. The higher the summarized Repli-seq value the earlier it is expected that this region of the genome undergoes replication. Thereafter mutation windows are assigned to the replication timing domains to assess the effect of replication timing on the relative SBS mutation rate.

For each non-overlapping RMF window the density of coding sequence is determined with the EnsDb.Hsapiens.v75 package in R, while the CpG density is determined using the BSgenome.Hsapiens.UCSC.hg19 package in R. RMF, replication timing, coding genome density and CpG density were plotted with Gviz (Hahne and Ivanek, 2016).

Relative Mutation Rate

Whole genome bisulfite sequencing (WGBS) data from 2 normal sigmoid cell colon samples were downloaded from GEO Omnibus (GSM983645 and GSM1010989) and was used for assessing the methylation state of CpGs. The analysis was restricted to CpG sites covered by at least 1 WGBS sample, the average is calculated in case sufficient coverage was present in both samples. The relative mutation rate for CpG or 5mC can be calculated by determining the CpG abundance (RMR-CG) or 5mC (RMR-5mC) abundance as previously described (Sanders et al., 2018). RMR-CG reflects the number of mutations expected in 1Mb of CpGs in a particular genomic feature if 1000 somatic mutations are randomly selected from the mutation catalogue:

$$\mathrm{RMR}_{ij}^{CpG} = \frac{\mu_{ij}}{\left(\frac{\alpha_i}{10^6}\right) \cdot \left(\frac{\tau_j}{10^3}\right)}$$

Where μ_{ij} represents the CG>TG mutation count in feature i for sample j, α_i represents the total number of CpGs in feature i, and τ_j is the total CG>TG mutation burden for sample j.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For RMR-5mC the α_i is replaced with the methylation level sum across feature i:

$$\mathrm{RMR}_{ij}^{5mC} = \frac{\mu_{ij}}{\left(\frac{\omega_i}{10^6}\right) \cdot \left(\frac{\tau_j}{10^3}\right)}$$

Where:

$$\omega_i = \sum_{CpG \in i} \beta_i$$

Where β_i is the average methylation level (ranging between 0 – no methylation – and 1 – full methylation). RMR-5mC corrects for the methylation state of features of interest. For instance, promoters often have little methylation.

To assess the linear relationship between methylation level and mutation probability, CG>TG mutations are binned per sample based on the methylation level detected in the normal sigmoid colon into consecutive bins of 0.1 interval, ranging from 0 (no methylation) to 1 (full methylation) with a 0.1 interval. The total count per bin is divided by 1.000.000 and the age of the donor to assess the annual mutation rate per million methylated CpGs.

Fitting Mutational Signatures for Regions of Interest

For normal intestinal epithelium crypts mutations per region of interest are determined. First, replication timing domains (Early, Earliest, Median, Late, Latest) as described before. Second, across the coding sequence (CDS), all human exons, complete genes and the intergenic space. Mutational signatures extracted from the microdissected samples, described under 'Mutational signature extraction', are fitted to the mutational catalogues of the replication domains or regions of interest with TensorSignatures (Vöhringer et al., 2020). For MutSα-deficient intestinal crypts the contribution of MMR-1 is high in early replication regions and the CDS due to the increase CpG density in these regions. A beta generalized linear mixed model was fitted to test whether MutSα-deficient intestinal crypts have an enrichment of MMR-1 in these regions compared to the MutLα-deficient intestinal crypts. Fitting was done with glmmTMB (Brooks et al., 2017) using the exposure of MMR-1 (proportion) as the dependent variable, MutSα-deficiency as fixed effect (dichotomous variable) and donor as random intercept (multiple samples per tissue). The 95%-CI was determined by 10.000 bootstraps via bootMer from lme4, while the p-value was determined by parametric bootstrapping for 10.000 bootstraps with a full model, containing MutSα-deficiency as fixed effect and a reduced model, without

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MutSα-deficiency as variable and calculating the loglikelihood deviance as comparator.

Assessing MMRD Genotype

MSI status for endometrial, colorectal and stomach cancer from The Cancer Genome Atlas (TCGA) was downloaded from the GDC data portal. For the remainder of the TCGA cohort, including endometrial, colorectal and stomach cancers with a negative MSI status, samples were selected with a high whole exome sequencing (WES) mutation burden (≥ 150 somatic mutations) for further characterization. The mutational catalogues of selected WES were fitted with the standard COSMIC v3 MMR signatures SBS-6, SBS-15, SBS-20, SBS-21 and SBS-26 and SBS-1. Cases with an exposure of ≥ 0.25 for one MMR signature or SBS-1 were selected for genotyping. Mutational profiles were on occasion manually reviewed to ensure correct cases were selected for further analysis. Including the constitutive mismatch repair deficiency (CMMRD) this resulted in 337 possible MMRD cancers.

Tumor-only variant calling was done with Mutect2 (Benjamin et al., 2019). Protein truncating variants (PTV) in *MLH1*, *MSH2*, *MSH6* or *PMS2* were considered disruptive. Missense mutations in these genes in a hypermutation setting can result in incorrect genotyping. Only missense mutations listed as 'likely pathogenic' or 'pathogenic' in the InSiGHT (Thompson et al., 2014) or ClinVar (Landrum et al., 2018) databases were considered clinically relevant. The remainder of missense mutations in the MMR genes were only considered relevant if there was strong evidence in the literature from *in vivo*/*in vitro* assays that a hypermutation phenotype is observed. Transcript fusions or structural variants effecting a MMR gene was considered relevant. Finally, CNV loss as determined by ASCAT (Van Loo et al., 2010) or by local loss of coverage was considered relevant.

Based on sample purity as estimated by ASCAT, cases were only considered genotyped, if the following is found: a clinically relevant homozygous mutation after taking purity into account, compound heterozygous relevant mutations in the same gene, digenic relevant mutations or when a clinically relevant mutation is combined with a CNV or SV. Promoter methylation followed by gene silencing, commonly observed for *MLH1* in CpG island hypermethylation (CIMP) colorectal cancers, was not taken along because the methylator phenotype might be an event secondary to neoplastic change or is transient in nature. This resulted in 117 MMRD cancers being genotyped. For a considerable number of cases the clinically relevant variant is germline, presumably Lynch syndrome. In accordance with TCGA

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

procedures the cases have been anonymized in all figures and tables. The names reflect the TCGA cancer subtype only and additional information on the MMR genes with disruptive variants are included in the supplementary tables.

Dirichlet Regression

The extracted mutational signatures are fitted to the TCGA and CMMRD WES mutational catalogues to determine the mutational signature composition per case. Dirichlet regression (Maier, 2014) was applied to relate changes in the mutational signature composition to the MMRD genotype or the deficiency of a specific MMR complex. Cases with clinically relevant variants in both the MutSα and the MutLα (digenic) complexes were excluded from the analysis. It remains difficult to assess whether the loss of both complexes in cancer has a compounding effect on a mutational level or whether there was an chronological order in the loss of both complexes. Cases with digenic clinically relevant mutations in the same complex, e.g., *MSH2* and *MSH6*, were retained. First, PMS2-deficient cancers were used as the reference when linking changes in the mutational signature composition to the MMRD genotypes. This analysis was performed on all the MMRD cancers fulfilling the listed criteria. Cancers with MutLα-deficiency were used as reference when comparing MutSα-deficiency to MutLα-deficiency for changes in the mutational signature composition. This analyses was performed across all MMRD cancers, all MMRD cancers commonly associated with MMRD (endometrial, colorectal, stomach and prostate cancer), endometrial cancer alone, colorectal cancer alone, stomach cancer alone and all other remaining MMRD cancers.

REFERENCES

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:14065823.

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. bioRxiv, 861054.

Brooks, M.E., Kristensen, K., Van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Machler, M., and Bolker, B.M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R journal *9*, 378-400.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C.*, et al.* (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet *40*, 722-729.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Ellis, P., Moore, L., Sanders, M.A., Butler, T.M., Brunner, S.F., Lee-Six, H., Osborne, R., Farr, B., Coorens, T.H.H., Lawson, A.R.J.*, et al.* (2021). Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. Nat Protoc *16*, 841-871.

Farmery, J.H.R., Smith, M.L., Diseases, N.B.-R., and Lynch, A.G. (2018). Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. Sci Rep *8*, 1300.

Feuerbach, L., Sieverling, L., Deeg, K.I., Ginsbach, P., Hutter, B., Buchhalter, I., Northcott, P.A., Mughal, S.S., Chudasama, P., Glimm, H.*, et al.* (2019). TelomereHunter - in silico estimation of telomere content and composition from cancer genomes. BMC Bioinformatics *20*, 272.

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., and Stratton, M.R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet *Chapter 10*, Unit 10 11.

Hahne, F., and Ivanek, R. (2016). Visualizing genomic data using Gviz and bioconductor. In Statistical genomics (Springer), pp. 335-351.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A *107*, 139-144.

Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A., von Haeseler, A., and Minh, B.Q. (2018). MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol Biol *18*, 11.

Jones, D., Raine, K.M., Davies, H., Tarpey, P.S., Butler, A.P., Teague, J.W., Nik-Zainal, S., and Campbell, P.J. (2016). cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. Curr Protoc Bioinformatics *56*, 15 10 11-15 10 18.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W.*, et al.* (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*, D1062-D1067.

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M.*, et al.* (2019). The landscape of somatic mutation in normal colorectal epithelial cells. Nature *574*, 532-537.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1295 Maier, M.J. (2014). DirichletReg: Dirichlet regression for compositional data in R.
1296 Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies,
1297 H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and
1298 Somatic Tissues. Cell *171*, 1029-1041 e1021.
1299 Moore, L., Leongamornlert, D., Coorens, T.H.H., Sanders, M.A., Ellis, P., Dentro, S.C.,
1300 Dawson, K.J., Butler, T., Rahbari, R., Mitchell, T.J.*, et al.* (2020). The mutational landscape
1301 of normal human endometrial epithelium. Nature *580*, 640-646.
1302 Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K.,
1303 Jones, D., Hinton, J., Marshall, J., Stebbings, L.A.*, et al.* (2012). Mutational processes
1304 molding the genomes of 21 breast cancers. Cell *149*, 979-993.
1305 Pich, O., Muinos, F., Sabarinathan, R., Reyes-Salazar, I., Gonzalez-Perez, A., and Lopez-
1306 Bigas, N. (2018). Somatic and Germline Mutation Periodicity Follow the Orientation of the
1307 DNA Minor Groove around Nucleosomes. Cell *175*, 1074-1087 e1018.
1308 Raine, K.M., Hinton, J., Butler, A.P., Teague, J.W., Davies, H., Tarpey, P., Nik-Zainal, S.,
1309 and Campbell, P.J. (2015). cgpPindel: Identifying Somatically Acquired Insertion and
1310 Deletion Events from Paired End Sequencing. Curr Protoc Bioinformatics *52*, 15 17 11-15 17
1311 12.
1312 Sanders, M.A., Chew, E., Flensburg, C., Zeilemaker, A., Miller, S.E., Al Hinai, A.S., Bajel, A.,
1313 Luiken, B., Rijken, M., McLennan, T.*, et al.* (2018). MBD4 guards against methylation
1314 damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML.
1315 Blood *132*, 1526-1534.
1316 Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate
1317 variation across the human genome. Nature *521*, 81-84.
1318 Thompson, B.A., Spurdle, A.B., Plazzer, J.P., Greenblatt, M.S., Akagi, K., Al-Mulla, F.,
1319 Bapat, B., Bernstein, I., Capella, G., den Dunnen, J.T.*, et al.* (2014). Application of a 5-tiered
1320 scheme for standardized classification of 2,360 unique mismatch repair gene variants in the
1321 InSiGHT locus-specific database. Nat Genet *46*, 107-115.
1322 Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman,
1323 V.J., Marynen, P., Zetterberg, A., Naume, B.*, et al.* (2010). Allele-specific copy number
1324 analysis of tumors. Proc Natl Acad Sci U S A *107*, 16910-16915.
1325 Vöhringer, H., van Hoeck, A., Cuppen, E., and Gerstung, M. (2020). Learning mutational
1326 signatures and their multidimensional genomic properties with TensorSignatures. bioRxiv,
1327 850453.
1328 Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic
1329 variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164.
1330 Yu, G. (2020). Using ggtree to visualize data on tree-like structures. Current protocols in
1331 bioinformatics *69*, e96.
1332

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SUPPLEMENTARY FIGURES

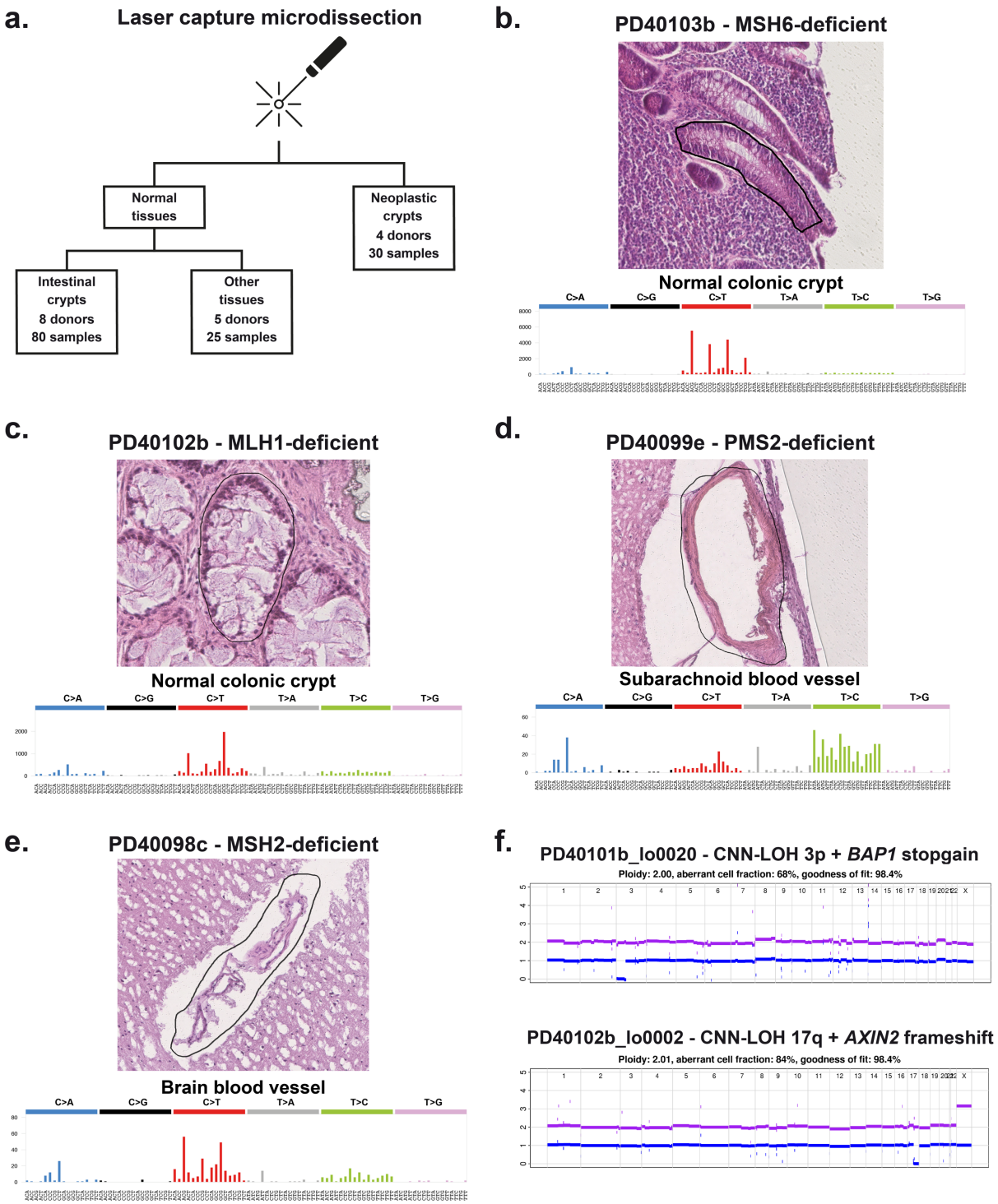


Figure S1. Mutational Profiles and Genetic Lesions Determined from CMMRD Tissues by Laser Capture Microdissection, Related to Figures 1-2 and 5.

(**a**) Sample collection outline for CMMRD tissues by laser capture microdissection.

(**b-e**) Microdissected areas and matched mutational profile for (**b**) a normal colonic crypt from MSH6-deficient donor PD40103, (**c**) a normal colonic crypt from MLH1-deficient donor PD40102, (**d**) a subarachnoid blood vessel collected from a brain sample of PMS2-deficient

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1342 donor PD40099 and (**e**) a blood vessel collected from a brain sample from MSH2-deficient
1343 donor PD40098.
1344 (**f**) Examples of homozygous loss of driver genes through truncating mutations coinciding with
1345 copy number neutral loss of heterozygosity in individual crypts of donors PD40101 and
1346 PD40102.
1347

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

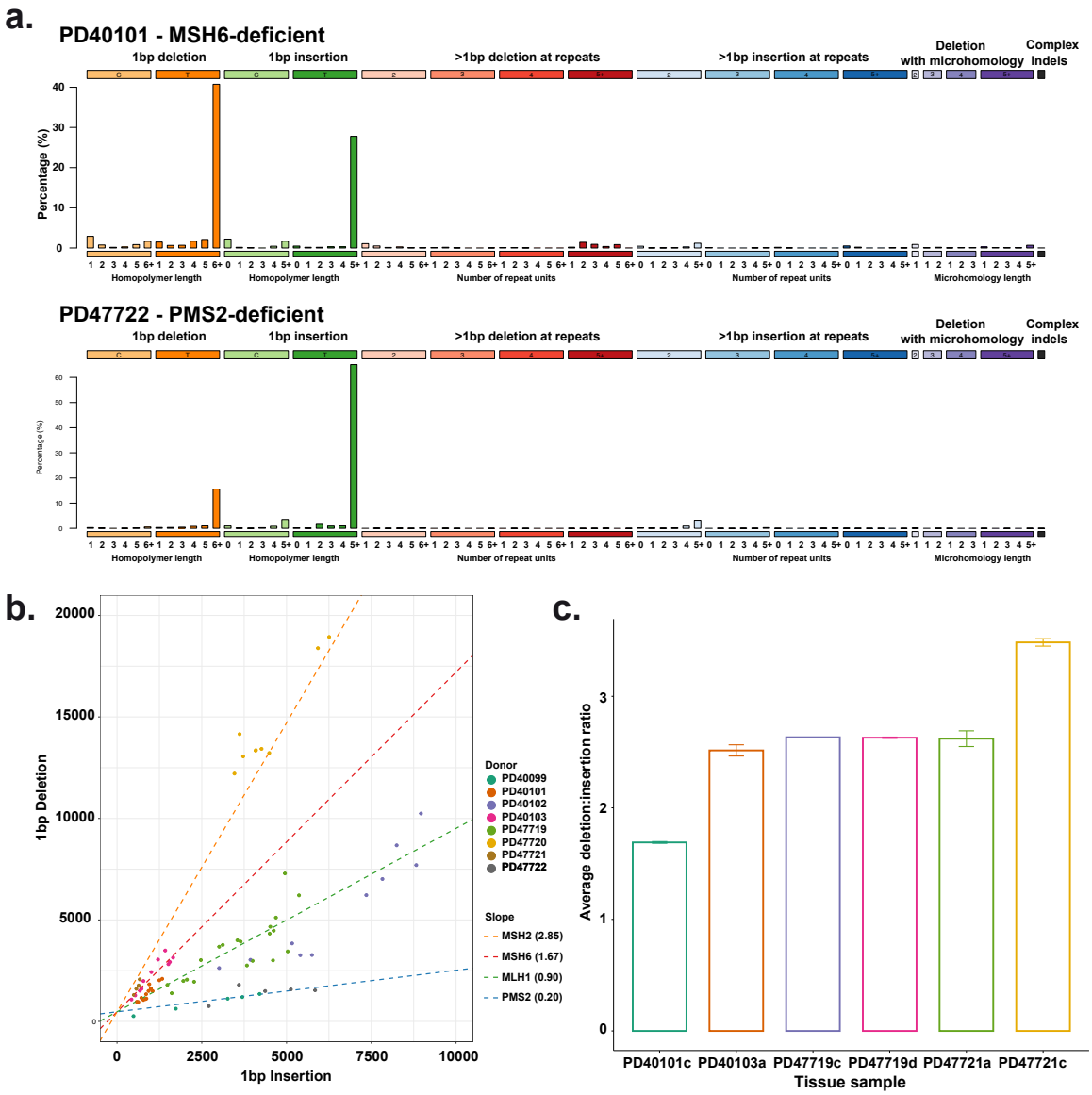


Figure S2. Insertion and Deletion Types and Acquisition Rates Differs Across MMRD Genotype, Related to Figures 1 and 3-4.

(**a**) Typical insertion and deletion (indel) classification profiles observed from normal intestinal crypts from a MSH6-deficient and PMS2-deficient donor. Similar indel classification profiles were observed for other donors with the same MMRD genotype.

(**b**) Linear mixed model fit for 1bp insertions against 1bp deletions with the MMRD genotype (MSH2, MSH6, MLH1 and PMS) included as fixed effect, donor included as random intercept and MMRD genotype included as random slope.

(**c**) Average 1bp deletion-insertion ratios calculated across intestinal crypts from collected neoplastic tissues.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

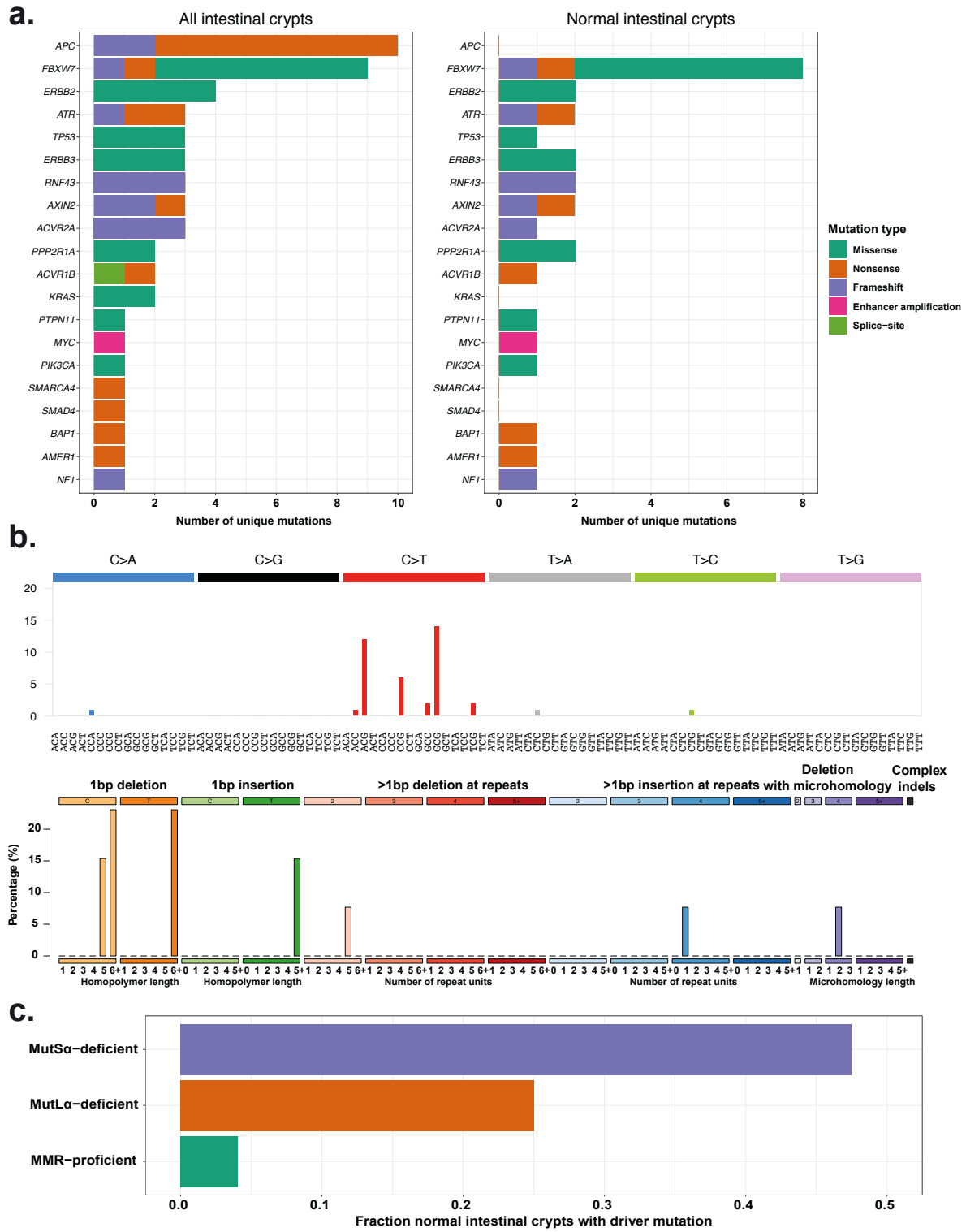


Figure S3. MMRD Driver Mutation Landscape Recapitulates Active Mutational Processes and Is Enriched Compared to the Healthy Population, Related to Figures 3-4.

(**a**) Driver mutation count coloured by mutation type for (Left) all intestinal crypts isolated and (Right) all normal intestinal crypts. Mutation types include: missense, nonsense, essential splice site, frameshifts mutations and copy number changes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**b**) SBS and indel mutational profiles of driver mutations detected in healthy and neoplastic intestinal crypts.

(**c**) Fraction of normal intestinal crypts with driver mutations as detected in MutSα-deficient donors (MSH2 or MSH6, n = 40 crypts), MutLα-deficient donors (MLH1 or PMS2, n = 40 crypts) and MMR-proficient donors (n = 445 crypts).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

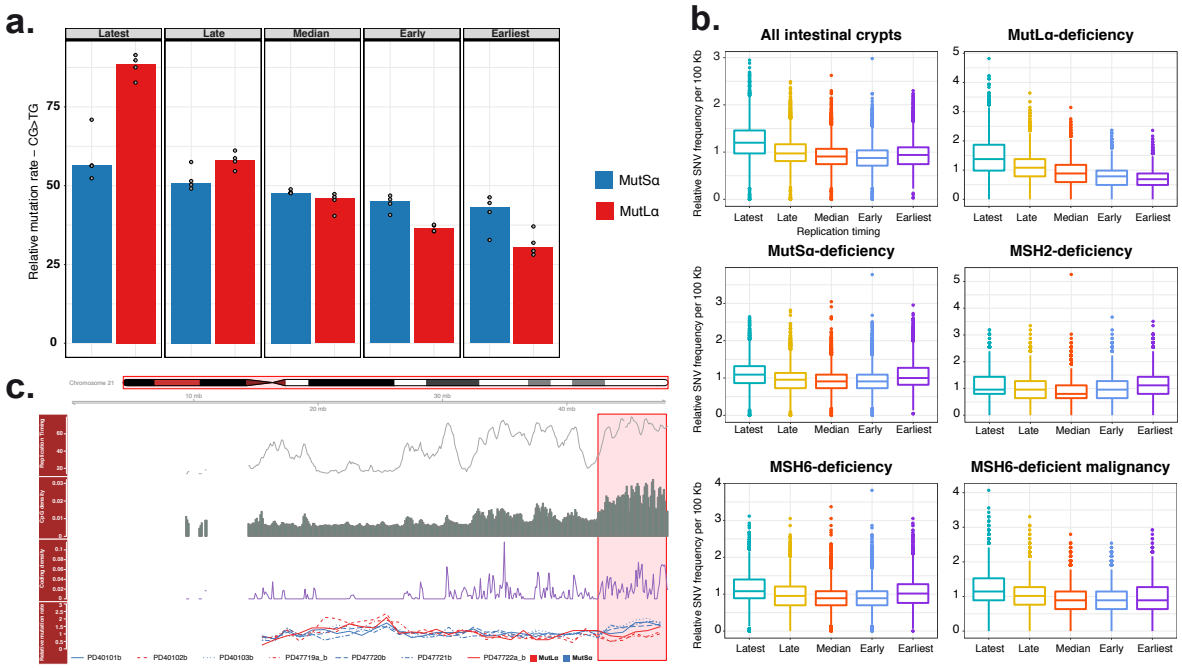


Figure S4. Genomic Distribution of Mutations in Intestinal Crypts is Associated with MMRD Genotype and Disease State, Related to Figures 1, 3-4 and 6.

(**a**) Relative mutations rates (RMR) of CG>TG mutations was calculated for the replication timing domains (latest, late, median, early and earliest) by correcting for CpG-density and sample-specific mutation burden. Calculated RMR values were grouped based on MutSα or MutLα-deficiency to illustrate the changes in CG>TG mutation rates across replication timing domains.

(**b**) Relative SBS frequency boxplots for 100 kb non-overlapping bins across the genome grouped on replication timing domains. Data for healthy intestinal crypts is illustrated for all donors, donors grouped on MMR complex deficiency and loss of individual MMR components. Data for MSH6-deficient neoplastic intestine is illustrated in the bottom-right panel.

(**c**) Distribution of mutations and established determinants of mutation rate for chromosome 21 across non-overlapping 1 Mb bins. From top to bottom. (I) Average replication timing profile. High values indicate early replication and vice versa. (II) CpG-dinucleotide density. (III) Coding genome density. (IV) Average relative SBS frequency of healthy intestinal crypts for individual donors as indicated by line type. Line color indicates the MMR complex lost for each donor. Highlighted areas indicate broad regions of high CpG and coding genome density.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

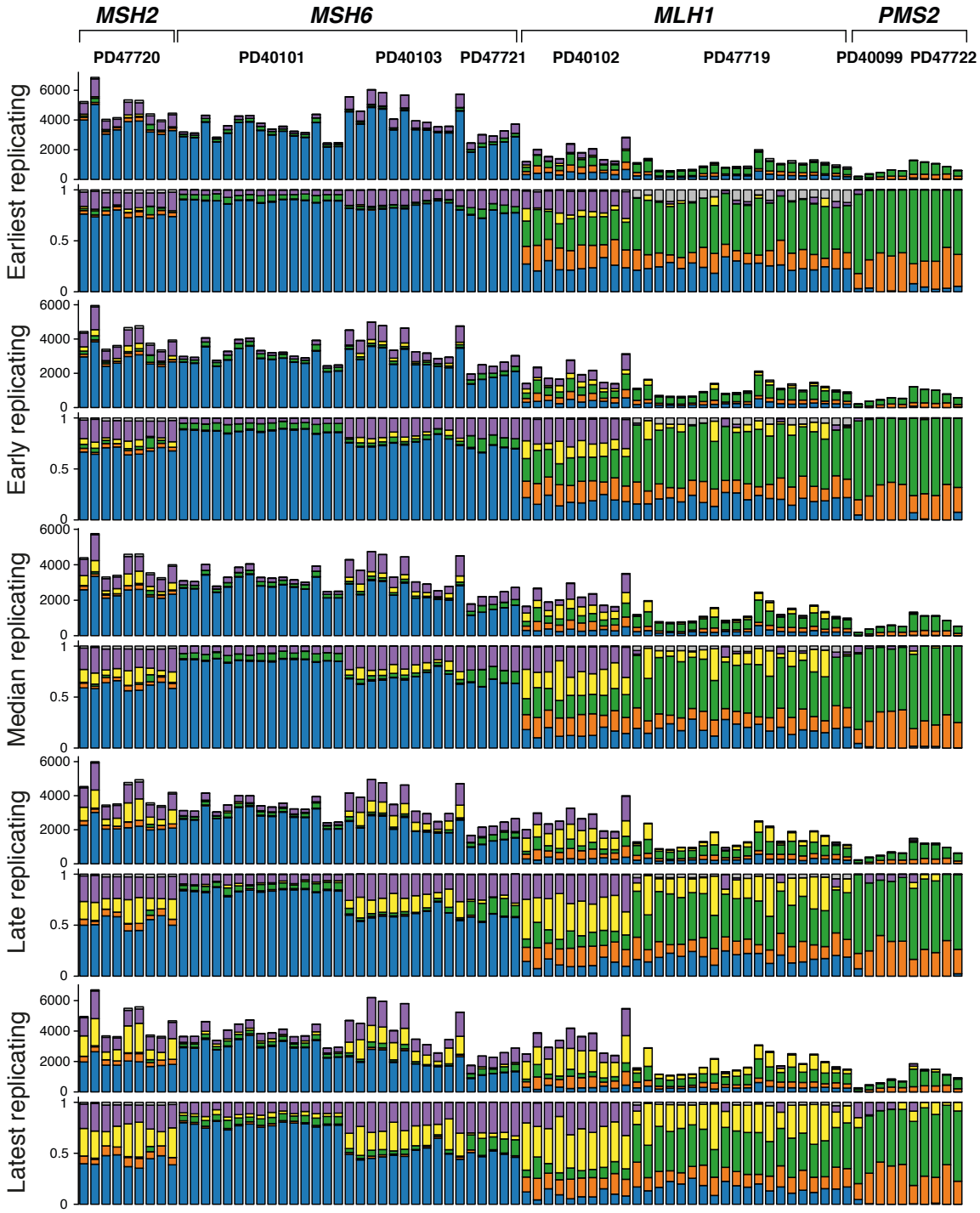


Figure S5. Differential DNA Damage Acquisition Across Replication Timing Domains Reveals the DNA Repair Patterns of Individual MMR components, Related to Figures 2-4 and 6.

Mutations where counted across the replication timing domains. Mutational signatures extracted from the genome-wide normal CMMRD intestinal crypt mutation catalogues (Figure 2) were fitted to each replication timing domain to establish mutational signature composition per sample and domain. Samples are grouped by donor and MMR complex deficiency. From

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1400 top to bottom: Mutational signature attribution based on all mutations (SBSs and indels) and
1401 the relative mutational signature attribution for each replication timing domain, ranging from
1402 earliest, early, median, late, latest.
1403
1404

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

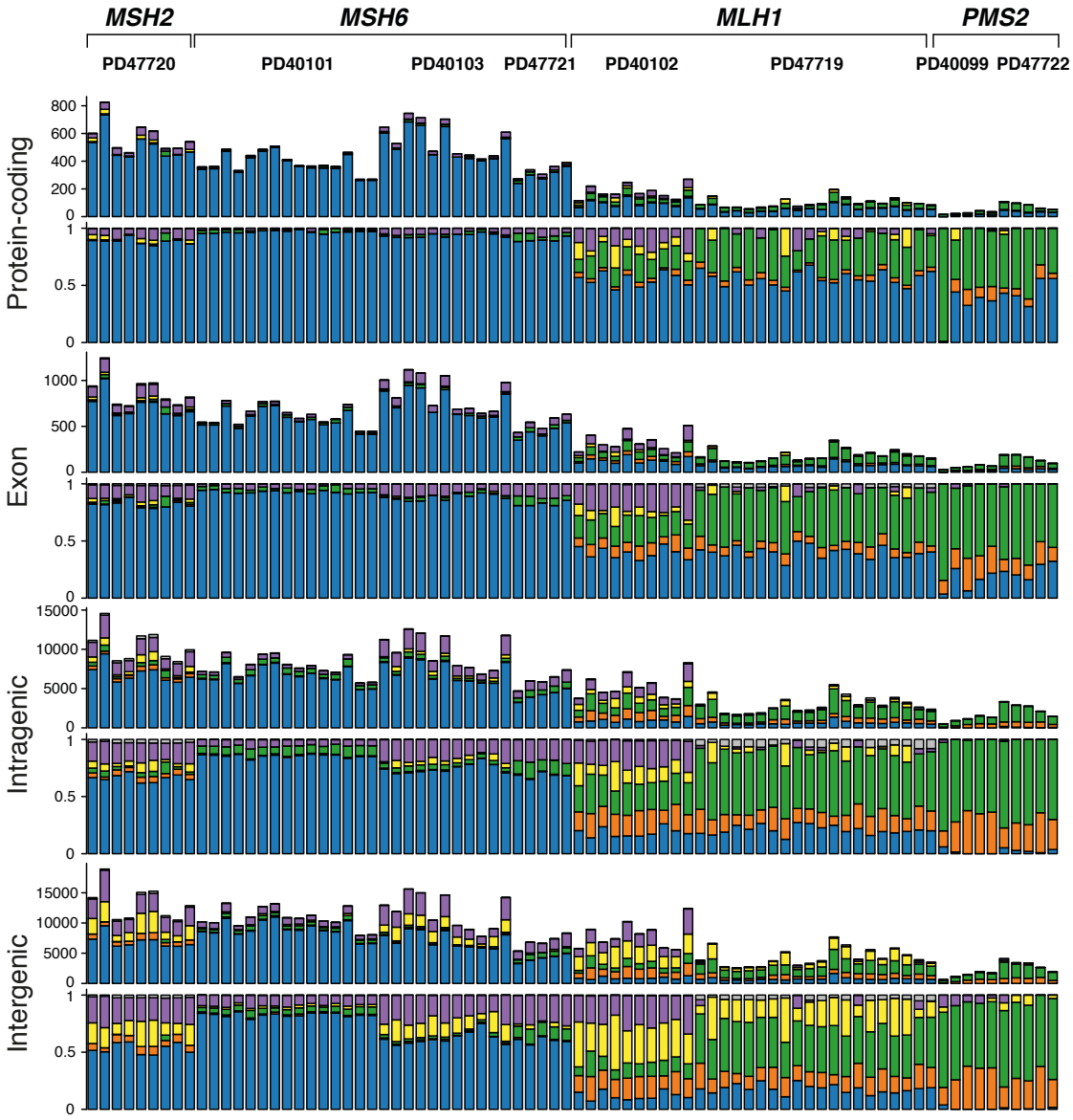


Figure S6. Differential DNA Damage Acquisition Across Genomic Elements Reveals the DNA Repair Patterns of Individual MMR components, Related to Figures 2-4 and 6.

Mutations were counted across different genomic elements, ranging from the coding sequence, all exons (including untranslated regions), all genic constituents (exons and introns) and the intergenic space. This order of genomic elements follows the trend of methylated CpG density. Mutational signatures extracted from genome-wide mutation catalogues (Figure 2) were fitted to each genomic element to establish the mutational signature composition per sample and genomic element. Samples are grouped by donor and MMR complex deficiency. From top to bottom: Mutational signature attribution based on all mutations (SBSs and indels) and the relative mutational signature attribution for each genomic segment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

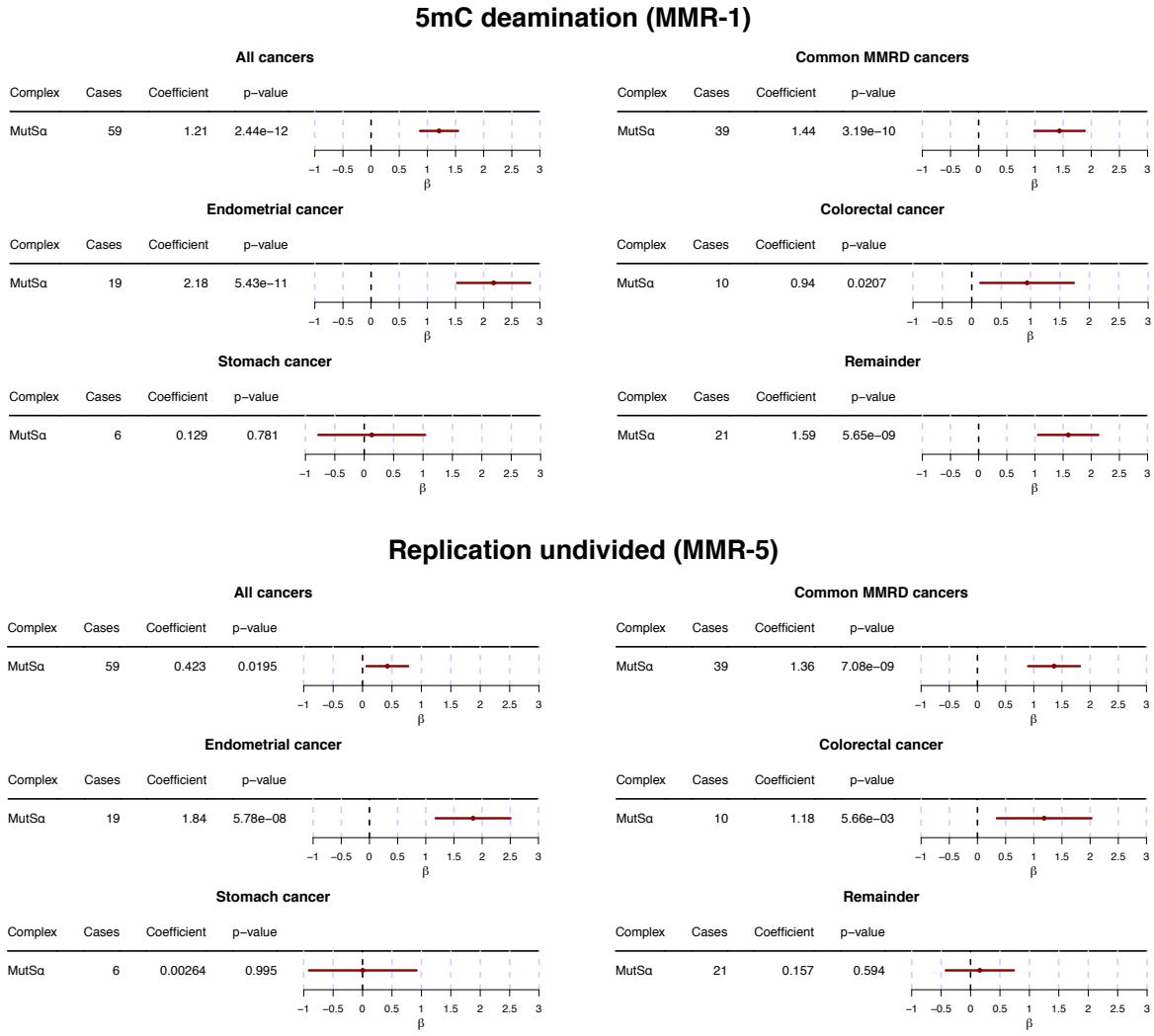


Figure S7. Validation of Mutational Signature Activity in Public Cancer Datasets, Related to Figures 2 and 7.

Dirichlet regression performed on the mutational signature composition established by fitting the extracted mutational signatures to the WES mutational catalogues of MMRD cancers. Cases with loss of both MMR complexes (MutSα and MutLα) were excluded from the analysis. Cancers with MutLα-deficiency were used as reference. Enrichment or significance should be considered relative. Significance of mutational signature enrichment was tested in all MMRD cancers, the cancer subtypes enriched for MMRD (endometrial, colorectal, stomach and prostate cancer), endometrial cancer, colorectal cancer, stomach cancer and the remainder of MMRD cancers.

(**a**) Validation of the association between 5mC deamination signature (MMR-1) activity and MutSα-deficiency in all predefined cancer categories with the exception of MMRD stomach cancer.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.14.437578; this version posted April 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**b**) Validation of the association between 'Replication Undivided' signature (MMR-5) activity and MutSα-deficiency in all predefined cancer categories with the exception of MMRD stomach cancer and the remainder of MMRD cancers.