











# Inherited *MUTYH* mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells

Philip S. Robinson <sup>1,2</sup>, Laura E. Thomas <sup>3</sup>, Federico Abascal <sup>1</sup>, Hyunchul Jung<sup>1</sup>, Luke M. R. Harvey<sup>1</sup>, Hannah D. West<sup>4</sup>, Sigurgeir Olafsson<sup>1</sup>, Bernard C. H. Lee<sup>1,5</sup>, Tim H. H. Coorens <sup>1</sup>, Henry Lee-Six <sup>1</sup>, Laura Butlin <sup>4</sup>, Nicola Lander<sup>4</sup>, Rebekah Truscott<sup>4</sup>, Mathijs A. Sanders<sup>1,6</sup>, Stefanie V. Lensing<sup>1</sup>, Simon J. A. Buczacki<sup>7</sup>, Rogier ten Hoopen <sup>8</sup>, Nicholas Coleman<sup>9,10</sup>, Roxanne Brunton-Sim<sup>11</sup>, Simon Rushbrook<sup>11,12</sup>, Kourosh Saeb-Parsy <sup>13,14</sup>, Fiona Lalloo<sup>15</sup>, Peter J. Campbell<sup>1</sup>, Iñigo Martincorena <sup>1</sup>, Julian R. Sampson<sup>4</sup> & Michael R. Stratton <sup>1</sup>✉

Cellular DNA damage caused by reactive oxygen species is repaired by the base excision repair (BER) pathway which includes the DNA glycosylase *MUTYH*. Inherited biallelic *MUTYH* mutations cause predisposition to colorectal adenomas and carcinoma. However, the mechanistic progression from germline *MUTYH* mutations to *MUTYH*-Associated Polyposis (MAP) is incompletely understood. Here, we sequence normal tissue DNAs from 10 individuals with MAP. Somatic base substitution mutation rates in intestinal epithelial cells were elevated 2 to 4-fold in all individuals, except for one showing a 31-fold increase, and were also increased in other tissues. The increased mutation burdens were of multiple mutational signatures characterised by C > A changes. Different mutation rates and signatures between individuals are likely due to different *MUTYH* mutations or additional inherited mutations in other BER pathway genes. The elevated base substitution rate in normal cells likely accounts for the predisposition to neoplasia in MAP. Despite ubiquitously elevated mutation rates, individuals with MAP do not display overt evidence of premature ageing. Thus, accumulation of somatic mutations may not be sufficient to cause the global organismal functional decline of ageing.

<sup>1</sup>Cancer, Ageing and Somatic Mutation (CASMS), Wellcome Sanger Institute, Hinxton CB10 1SA, UK. <sup>2</sup>Department of Paediatrics, University of Cambridge, Cambridge CB2 0QQ, UK. <sup>3</sup>Institute of Life Science, Swansea University, Swansea SA28PP, UK. <sup>4</sup>Institute of Medical Genetics, Division of Cancer and Genetics, Cardiff University School of Medicine, Cardiff, UK. <sup>5</sup>Hereditary Gastrointestinal Cancer Genetic Diagnosis Laboratory, Department of Pathology, The University of Hong Kong, Queen Mary Hospital, Pokfulam, Hong Kong. <sup>6</sup>Department of Haematology, Erasmus University Medical Centre, 3015 CN Rotterdam, The Netherlands. <sup>7</sup>Nuffield Department of Surgical Sciences, Medical Sciences Division, University of Oxford, Oxford, UK. <sup>8</sup>Department of Oncology, University of Cambridge, Cambridge, UK. <sup>9</sup>Department of Pathology, University of Cambridge, Cambridge, UK. <sup>10</sup>Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. <sup>11</sup>Norfolk and Norwich University Hospital, Norwich, UK. <sup>12</sup>Norwich Medical School, University of East Anglia, Norwich, UK. <sup>13</sup>Department of Surgery, University of Cambridge, Cambridge, UK. <sup>14</sup>Cambridge NIHR Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge, UK. <sup>15</sup>Manchester Centre for Genomic Medicine, Saint Mary's Hospital, Oxford Road, Manchester, UK. ✉email: [mrs@sanger.ac.uk](mailto:mrs@sanger.ac.uk)

The genomes of all normal human cells are thought to acquire mutations during the course of life. However, the mutation rates of normal cells and the processes of DNA damage, repair and replication that underlie them are incompletely understood<sup>1–8</sup>. A ubiquitous source of potential mutations is DNA damage caused by reactive oxygen species (ROS) which are formed as by-products of aerobic metabolism<sup>9</sup>. ROS cause a variety of DNA lesions, the most common being 8-oxoguanine (8-OG)<sup>10</sup>. As a consequence of mispairing with adenine during DNA replication, 8-OG can cause G:C > T:A (referred to as C > A for brevity) transversion mutations<sup>11</sup>. Under normal circumstances, 8-OG and its consequences are efficiently mitigated by the Base Excision Repair (BER) pathway effected by DNA glycosylases; oxoguanine DNA glycosylase (OGG1) removes 8-OG<sup>12</sup> and MutY DNA glycosylase (MUTYH) removes adenines misincorporated opposite 8-OG<sup>13</sup>.

Mutations in *MUTYH* engineered in experimental systems can impair its glycosylase activity, reducing its ability to excise mispaired bases and leading to an increased rate of predominantly C > A mutations<sup>14–18</sup>. *MUTYH* mutations inherited in the germline in humans cause an autosomal recessive syndrome (MUTYH-associated polyposis, MAP) characterised by intestinal adenomatous polyposis and an elevated risk of early onset colorectal and duodenal cancer<sup>19–22</sup>. The age of onset and the burden of intestinal polyps are highly variable between individuals, ranging from 10s to 100s leading to a substantially increased incidence of colorectal cancer<sup>23–27</sup>. Risks of other cancer types are also thought to be increased<sup>28</sup>.

Colorectal adenomas and carcinomas from individuals with MAP show a predominance of C > A mutations consistent with the presence of an elevated mutation rate attributed to defective *MUTYH* function<sup>29–33</sup>. However, whether there is an increased mutation rate in normal cells from individuals with biallelic germline *MUTYH* mutations is unknown. If present in normal cells, understanding the magnitude of the increase in mutation rate, the tissues and cell types in which it occurs, the proportion of cells which show it, the mutational processes responsible and the effects of early neoplastic change would provide insight into the genesis of the elevated cancer risk observed in these individuals.

In this study we perform whole-genome sequencing of normal cells from individuals with MAP. Using whole-genome sequencing we characterise the mutation rates and mutational processes in healthy tissues at a near-single-cell resolution. This study identifies the mutational processes that are associated with neoplastic transformation and are likely to underpin the increased cancer risk observed in this population of high-risk individuals.

## Results

**Clinical information.** Ten individuals aged 16 to 79 years with biallelic germline *MUTYH* mutations were studied. These included five missense mutation homozygotes (four *MUTYH*<sup>Y179C+/+</sup>, one *MUTYH*<sup>G286E+/+</sup>), three compound heterozygotes for the same pair of missense mutations (*MUTYH*<sup>Y179C+/- G396D+/-</sup>), and two siblings homozygous for a nonsense mutation (*MUTYH*<sup>Y104\*+/+</sup>). These *MUTYH* germline mutations have all been previously recognised as predisposing to MAP<sup>22,23</sup>. All 10 individuals had colorectal polyposis, with between 16 and >100 colonic adenomas, six were known to have duodenal polyps, five had colorectal cancer and one developed jejunal and pancreatic neuroendocrine cancer (Supplementary Data 1).

**Mutation rates in normal intestinal stem cells.** An intestinal crypt is constituted predominantly of a population of epithelial cells arising from a single recent common ancestor<sup>34–36</sup>. The

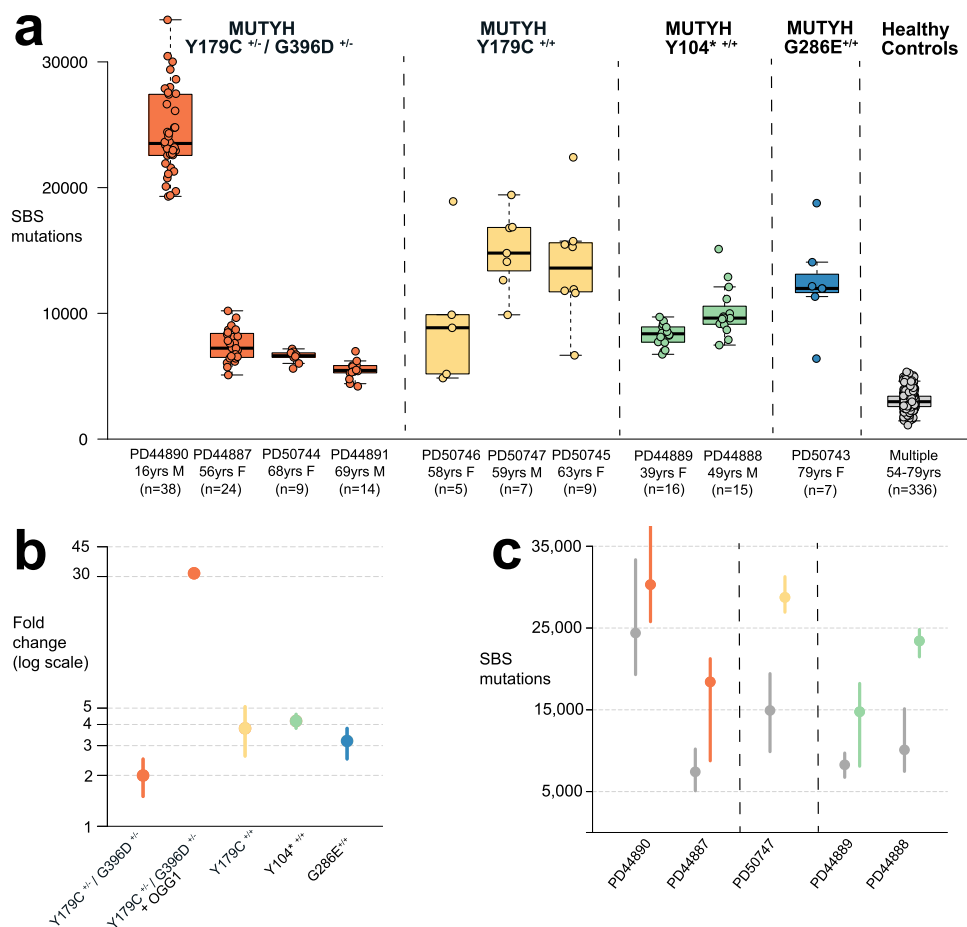
somatic mutations which have accumulated over the course of the individual's lifetime in the ancestral crypt stem cell are present in all its descendant cells<sup>3</sup>. Thus, by sequencing individual crypts, somatic mutations present in the ancestral stem cell can be identified. Using laser-capture microdissection, 144 individual normal intestinal crypts (large intestine  $n = 107$  and small intestine  $n = 37$ ) were isolated from the 10 individuals with germline *MUTYH* mutations (Supplementary Data 2). DNA libraries were prepared from individual crypts using a bespoke low-input DNA library preparation method<sup>37</sup> and were whole-genome sequenced at a mean 28-fold coverage.

The single base substitution (SBS) mutation burdens of individual crypts ranged from a median for each individual of 2294 to 33,350, equating to mutation rates of 92–1446 SBS/year, 2–31-fold higher than normal crypts from wild-type individuals (~46 SBS/year) (Fig. 1b, Methods)(linear mixed-effects model 95% confidence interval (C.I.), 69–1520 SBS/yr). Therefore, all normal crypts from all MAP individuals studied showed elevated somatic mutation rates (Fig. 1a, b).

Differences in mutation rate were observed between individuals with MAP (Fig. 1b). A 31-fold higher rate of SBS accumulation than in wild-type crypts<sup>3</sup> was observed in PD44890, a 16 year old male with *MUTYH*<sup>Y179C+/- G396D+/-</sup> who had an aggressive clinical phenotype with a very large number of adenomas and two different primary cancers at an early age. By contrast, the nine other individuals showed only 2- to 4-fold increases in mutation rate compared to wild type. The reason for this substantial difference is not clear. However, in addition to the *MUTYH* mutations, PD44890 carried two heterozygous germline missense variants in *OGG1* (Supplementary Fig. 1a, b), one inherited from the father and the other from the mother (Supplementary Fig. 2). One of these mutations, R46Q, is reported to impair *OGG1* activity in experimental systems<sup>38,39</sup> and has been observed as somatically mutated in human cancer<sup>40</sup>. Germline *OGG1* mutations are not currently recognised as causing cancer predisposition in humans<sup>41</sup>. However, if either or both of these mutations results in defective 8-OG excision they could account for the substantially elevated mutation rate in PD44890, particularly in the context of defective *MUTYH* activity. The brother of PD44890 shared the same *MUTYH* and *OGG1* germline mutations and demonstrated a similar early onset clinical phenotype, whereas the parents of these siblings were heterozygous for the *OGG1* and *MUTYH* variants and did not show adenomas or cancers.

There was also evidence of differences in mutation rates between the various *MUTYH* germline genotypes studied (Fig. 1b). Excluding the outlier individual PD44890, mutation rates were lower in individuals with the compound heterozygous *MUTYH*<sup>Y179C+/- G396D+/-</sup> (93 SBS/year, 95% C.I. 68–116) than individuals with *MUTYH*<sup>Y179C+/+</sup> (177 SBS/year, 95% C.I. 121–236), *MUTYH*<sup>Y104\*+/+</sup> (193 SBS/year, 95% C.I. 173–212) or *MUTYH*<sup>G286E+/+</sup> (145 SBS/year, 95% C.I. 117–172) ( $P = 10^{-10}$ ,  $P = 10^{-7}$ ,  $P = 10^{-23}$  and  $P = 10^{-13}$  respectively). The results, therefore, indicate that different *MUTYH* genotypes confer differentially elevated mutation rates and that the extent of the mutation rate increase can be modified by other factors.

SBS mutation rates in coding exons in normal intestinal crypts from MAP individuals were also elevated compared to wild-type individuals (Supplementary Fig. 3a, b). These increases were, however, slightly smaller than those observed in the genome-wide mutation rate (Supplementary Fig. 3a, b). Nonsense, missense and synonymous mutation rates were all increased compared with wild-type crypts, with the greatest increase observed in nonsense mutations (~10-fold more nonsense than wild-type vs ~3.5-fold more missense and ~2.6-fold more synonymous) (Supplementary Fig. 3c). This is attributable to the mutational



**Fig. 1 Somatic mutation burdens in cells with *MUTYH* mutations.** Elevated mutation burdens in normal intestinal cells with *MUTYH* mutations. **a** Genome-wide single base substitution (SBS) mutation burden of individual intestinal crypts ( $n = 144$  biologically independent samples) grouped according to patient. Each dot represents an individual intestinal crypt. *MUTYH* genotypes are displayed separately. Boxplots display median, inter-quartile range (IQR) from 1<sup>st</sup> to 3<sup>rd</sup> quartiles and whiskers extend from the last quartile to the last data point that is within 1.5x IQR. **b** Fold-change in SBS rate in intestinal crypts ( $n = 144$ ) with *MUTYH* mutations compared with wild-type controls<sup>3</sup> ( $n = 445$ ). Fold changes are represented by the dot, whiskers represent the 95% confidence interval (Methods). Dots are coloured according to germline genotype: orange; *MUTYH*<sup>Y179C+/- G396D+/-</sup> ( $n = 47$ ), *MUTYH*<sup>Y179C+/- G396D+/-</sup> & OGG1 ( $n = 38$ ), yellow; *MUTYH*<sup>Y179C+/+</sup> ( $n = 21$ ), green; *MUTYH*<sup>Y104+/+</sup> ( $n = 31$ ), blue; *MUTYH*<sup>G286E+/+</sup> ( $n = 7$ ). **c** Genome-wide single base substitution burden in histologically normal crypts (grey) and adenoma crypts (orange, yellow and green) arranged by patient and germline mutation. Data was available for 5 individuals who had adenoma glands sequenced. Dot represents the median and whiskers indicate the range from lowest to highest mutation burden per patient. Source data are provided as a Source Data file.

signatures present (see below) and the tendency of specific mutations at particular trinucleotide contexts to preferentially generate protein-truncating mutations<sup>42,43</sup>.

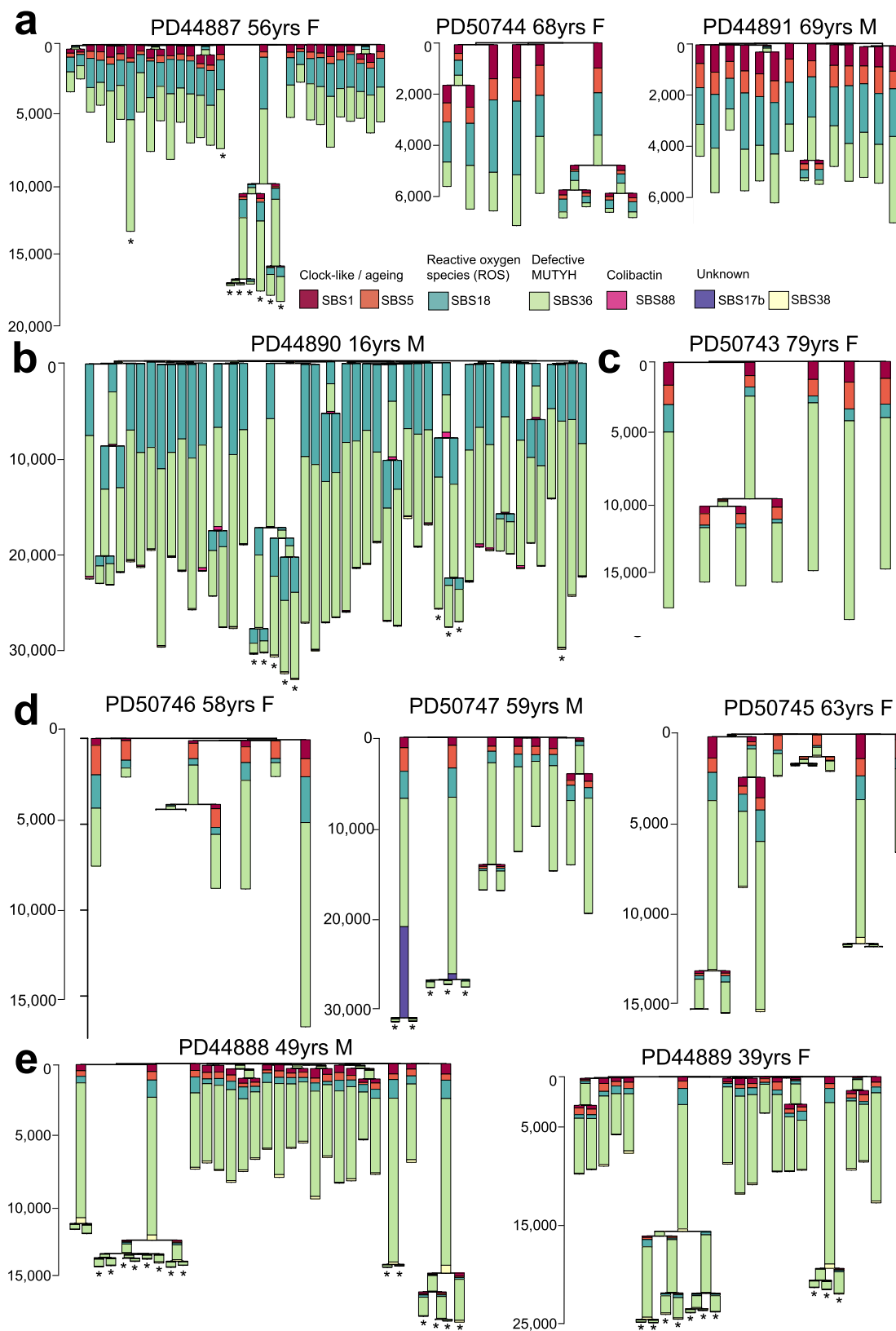
Neoplastic glands from 13 intestinal adenomas from five individuals with MAP showed SBS mutation burdens that were, on average, ~2-fold higher (range 1.2 to 2.5-fold) (Fig. 1c) than normal crypts from the same individuals sampled at the same time. Therefore, the elevated mutation rate observed in histologically normal intestinal crypts in individuals with germline *MUTYH* mutations is further increased during the process of neoplastic transformation, as previously observed in wild type individuals<sup>44,45</sup>.

Small insertion and deletion (ID) mutations accumulated at a rate of 2.1 ID/yr (linear mixed-effects model, 95% confidence interval (C.I.) 1.2–3.0,  $P = < 10^{-4}$ ), which is higher than in wild-type controls (1.3 ID/yr, linear mixed-effects model, C.I. 0.54–2.0,  $P = 0.0011$ )<sup>3,42</sup>. The cause of this modestly elevated ID rate is not clear. In two MAP individuals additional mutational processes could explain the higher burdens observed in these cases. In PD44890 the high ID mutation rate (ID rate 6/yr) was, at least partially, explained by the presence of an additional ID generating

mutational process associated with exposure to the mutagen colibactin produced by a strain of *E.Coli*<sup>3,46,47</sup> present in the colonic microbiome of some people (see below). In PD50747 (ID rate 6/yr), a previously undescribed sporadic ID signature IDA was identified which was not present in other MAP individuals (described below). Structural rearrangements and copy number changes were only observed in a small number of normal intestinal crypts, at similar frequencies to those in wild-type controls (Supplementary Data 2)<sup>3</sup>. Telomere shortening occurred at similar rates in individuals with *MUTYH* mutations compared to wild-type controls (Methods).

**Mutational signatures.** Mutational signatures were extracted from the combined catalogues of SBS mutations from all normal and neoplastic intestinal crypts and glands using two independent methods. We then decomposed each de novo extracted signature into known COSMIC reference mutational signatures. Finally, we used these decompositions to estimate the contribution of each reference signature to each sample (Methods, Supplementary Note). Three de novo extracted signatures, N1-N3, accounted for the majority of mutations, all of which were mainly characterised





**Fig. 3 Phylogenetic trees and mutational signatures in intestinal cells with germline *MUTYH* mutations.** Phylogenetic trees per-individual reconstructed from SBS mutations in individual intestinal crypts showing the number of SBS mutations per branch. Stacked barplots are overlaid onto each branch to represent the proportion of each mutational signature contributing to that branch. Phylogenetic trees are arranged by *MUTYH* germline mutation; **a** *MUTYH*<sup>Y179C+/-</sup> *G396D+/-* **b** *MUTYH*<sup>Y179C+/-</sup> *G396D+/-* with *OGG1* germline mutations, **c** *MUTYH*<sup>G286E+/+</sup>, **d** *MUTYH*<sup>Y179C+/+</sup> and **e** *MUTYH*<sup>Y104+/+</sup>. Adenoma glands bearing cancer driver mutations are indicated with an asterisk '\*'. Source data are provided as a Source Data file.



mutations, Fig. 3b). The SBS88 mutation burdens were consistent with those previously seen in wild type individuals indicating that *MUTYH* is unlikely to be implicated in the genesis of SBS88.

The increased SBS mutation burdens in normal crypts from individuals with *MUTYH* germline mutations appeared to be due to the contributions of SBS18 and SBS36 mutations (Fig. 3a–e). The proportions of SBS18 and SBS36, however, differed between *MUTYH* germline genotypes. SBS18 accounted for a substantially higher proportion of mutations in crypts and glands from individuals with the *MUTYH*<sup>Y179C+/- G396D+/-</sup> genotype ( $n = 85$  crypts) than in individuals with the *MUTYH*<sup>Y179C+/-</sup>, *MUTYH*<sup>Y104\*+/-</sup> and *MUTYH*<sup>G286E+/-</sup> genotypes ( $n = 59$  crypts, Supplementary Fig. 4a). Since *MUTYH*<sup>Y104\*</sup> causes *MUTYH* protein truncation, it is conceivable that SBS36 is the consequence of complete loss of *MUTYH* function and therefore that this is also effected by *MUTYH*<sup>Y179C</sup> and *MUTYH*<sup>G286E</sup>. Conversely, *MUTYH*<sup>G396D</sup> may retain partial activity<sup>14,21</sup> and thus generates a signature more closely resembling SBS18 which is found in normal tissues with fully active *MUTYH*.

The de novo extracted mutational signature N2, which primarily contributes to the mutational spectra of crypts from PD44890 (*MUTYH*<sup>Y179C+/- G396D+/-</sup>), resembled reference signature SBS18 (<https://cancer.sanger.ac.uk/cosmic/signatures>) but showed differences, notably with over representation of C > A mutations at GCA and, to a lesser extent, CCA and ACA trinucleotides (mutated base underlined) (Fig. 2a, b and Supplementary Fig. 1). A signature reported in human cells with in vitro engineered biallelic *OGGI* deletion is also primarily characterised by C > A mutations at GCA and ACA trinucleotides<sup>51</sup>. It is, therefore, possible that mutagenesis due to the germline *OGGI* variant(s) in PD44890 (see above) is superimposed on the mutational signature produced by the *MUTYH* germline mutations to generate N2 (see Supplementary Information for further analysis and discussion).

The mutational signatures in adenoma glands were similar to those seen in normal crypts from the same individuals (Fig. 3a, b, d, e). SBS36 and SBS18 were principally responsible for the increased mutation burdens observed in adenomas compared to normal crypts.

Candidate cancer driver mutations, defined as known or likely oncogenic hotspot mutations and truncating mutations in tumour suppressor genes (Methods, Supplementary Data 3), were observed in 15% of normal crypts (22/144), more than double the rate observed in wild-type crypts from comparable healthy controls; 6% (25/449)<sup>3,42</sup>. A substantial proportion of candidate drivers (16/22) were nonsense mutations, mirroring the broader exome-wide increase in nonsense mutations (Supplementary Fig. 3c), and reflecting the proclivity of certain mutation types to generate truncating mutations<sup>29,42,43</sup>. The mutational spectrum of driver mutations in normal crypts and neoplastic glands resembled the genome-wide spectra with substantial contributions from SBS18 and SBS36 (Supplementary Fig. 5). Hence, the mutational processes resulting from defective *MUTYH* activity appear to promote the accumulation of putative cancer driver mutations in normal and neoplastic tissues<sup>52,53</sup>.

Three known ID signatures were identified. ID1 and ID2 are characterised predominantly by insertions and deletions of single T bases at T mononucleotide repeats which are associated with strand slippage during DNA replication and are seen in most human cancers and normal tissues<sup>1–4,7,33</sup>. ID18 is associated with colibactin exposure, is found in normal intestinal stem cells and certain cancers, usually associated with SBS88<sup>3,47</sup>. ID1 was the dominant signature in normal cells whereas ID2 predominated in neoplastic cells (Supplementary Fig. 6). ID18 was principally observed in samples from PD44890 (16 years old with the high mutation burden and *OGGI* mutations) and is responsible for the

elevated ID rate in this individual (Supplementary Fig. 7). A further ID signature, IDA, identified in PD50747, was characterised by single C insertions at C mononucleotide repeats (Supplementary Figs. 7, 31 and 35). IDA was present in both normal crypts (~5% of total ID burden) and to a greater extent in adenoma glands (~20% of total ID burden). The cause of this previously undescribed signature is unclear but may be associated with previous capecitabine treatment in this individual and seems unlikely to be related to germline *MUTYH* mutations.

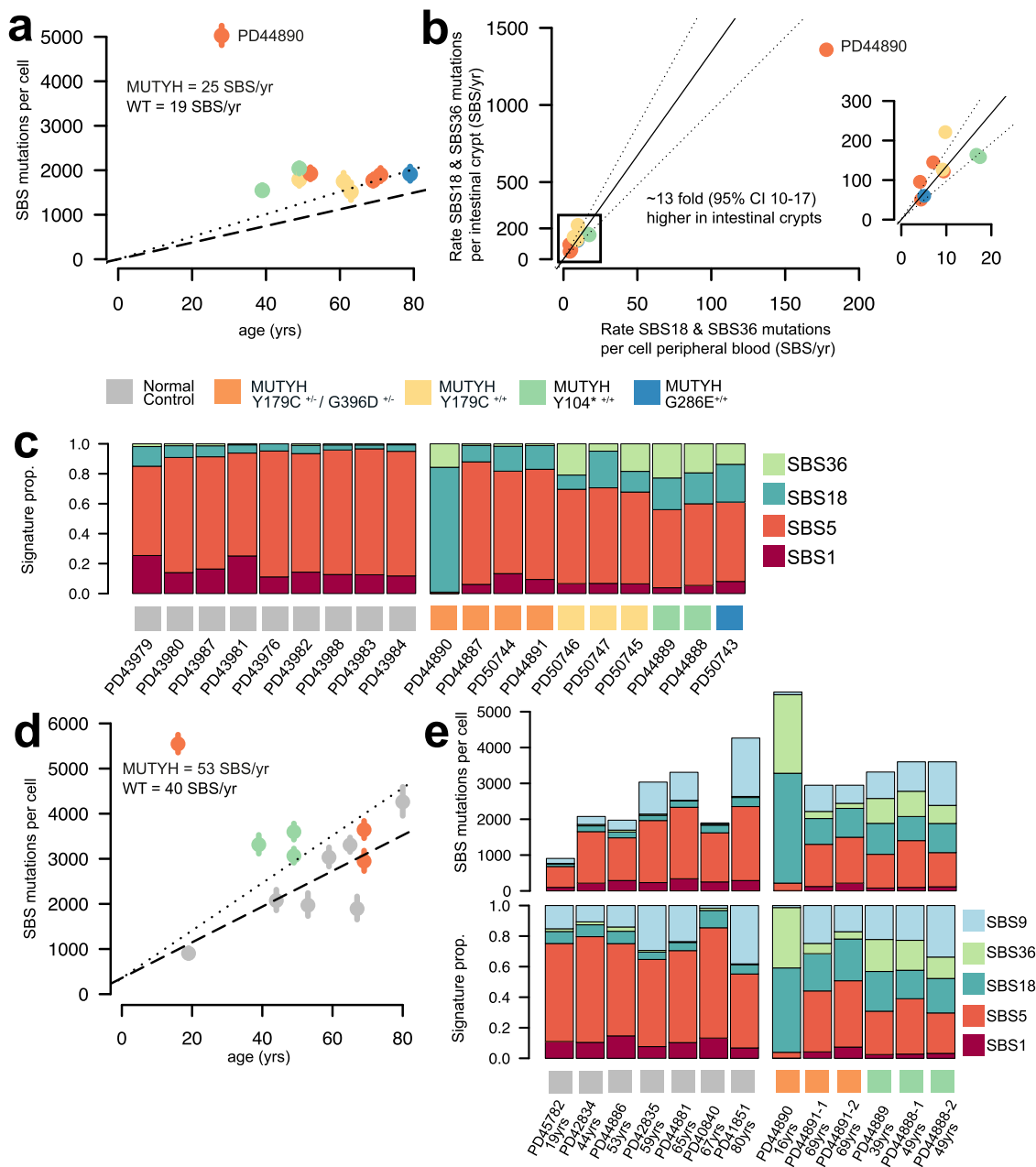
**Mutations in other cell types.** To investigate whether the elevated mutation rates and mutational signatures observed in intestinal epithelium caused by defective *MUTYH* are present in other cell types, peripheral blood and tissue lymphocyte DNAs from individuals with biallelic *MUTYH* mutations were whole genome sequenced using a duplex sequencing method (NanoSeq)<sup>50</sup> that allows mutation calling from single DNA molecules and thus accurately discovers somatic mutations in tissues in which multiple clonal lineages are intimately mixed.

The blood cell SBS mutation rates of all individuals with *MUTYH* mutations were higher than wild-type controls ( $n = 15$  granulocyte samples from 9 healthy individuals aged 20–80 yrs) (Fig. 4)(25 SBS/yr vs 19 SBS/yr, linear mixed-effects model,  $R^2 = 0.89$ , *MUTYH*; 95% C.I., 19–31,  $P = 10^{-7}$  and wild-type; 95% C.I., 14–24,  $P = 10^{-6}$ ). The relative increases in blood mutation rates were lower than in intestinal crypts from each individual (Fig. 4b). Nevertheless, the relative increases paralleled the differential increases observed between individuals in intestinal crypts. SBS mutation rates in tissue lymphocytes were modestly raised compared with wild-type healthy individuals (Fig. 4d) (53 SBS/yr vs 40 SBS/yr, linear mixed-effects model,  $R^2 = 0.68$ , *MUTYH*; 95% C.I., 21–85,  $P = 0.01$  and wild-type; 95% C.I., 13–66,  $P = 0.01$ ). The signatures associated with defective *MUTYH*, SBS18 and SBS36, contributed the excess mutations in all samples (Fig. 4c, e). An additional mutational signature was seen in lymphocytes. SBS9, which is associated with DNA polymerase  $\epsilon$  mediated somatic hypermutation and is a key process in the physiological maturation of B-cells, was observed in most lymphocyte samples indicating that the lymphocyte cell populations contained mature B-cells (Fig. 4e).

## Discussion

This study shows elevated base substitution somatic mutation rates due to SBS18 and/or SBS36 in normal tissues from individuals with *MUTYH* mutations. The results are compatible with all intestinal, and potentially all other cells in the body, showing elevated mutation rates. The relative increases in mutation rate and mutational signature composition differed between individuals, probably due to different *MUTYH* mutations and perhaps to other modifying influences.

We have previously highlighted the capability of normal human cells to tolerate substantially elevated mutation rates<sup>42</sup>. Carriers of *POLE* and *POLD1* exonuclease domain germline mutations exhibited elevated somatic mutation burdens without evident cellular or organismal consequences, other than an increased cancer risk<sup>42,54</sup>. This capability is confirmed in *MUTYH* germline mutation carriers. It is further emphasised by the observation of a 31-fold genome-wide elevated base substitution mutation burden in the 16 year old PD44890, which would confer a “mutational age” of ~500 years, without overt evidence of premature ageing. The increase in mutation burden in coding exons is lower than genome-wide in *POLE/POLD1* mutation carriers. Similarly, in individuals with *MUTYH* mutations there is a smaller increase of coding exon than genome-wide mutation burdens (Supplementary Fig. 3a, b). Nevertheless, in PD44890 the increase is still



**Fig. 4 Mutation burdens and mutational signatures in blood and immune cell populations.** **a** SBS mutation burden in peripheral blood per cell (x-axis) plotted against the age of the individual in years (y-axis). Dots are coloured according to the individual's germline mutation; orange; *MUTYH*<sup>Y179C+/-/G396D+/-</sup>, yellow; *MUTYH*<sup>Y179C+/-</sup>, green; *MUTYH*<sup>Y104+/-</sup> and blue; *MUTYH*<sup>G286E+/-</sup>. Whiskers represent the 95% confidence interval. Dashed line represents the mutation rate in wild-type (WT) normal control samples, dotted line represents the mutation rate in *MUTYH* samples (linear mixed-effects-model)<sup>50</sup>. **b** Mutation rate of *MUTYH* associated mutational signatures; SBS18 and SBS36 per cell for peripheral blood (SBS/yr)(x-axis) against the SBS18 & SBS36 mutation rate of normal intestinal crypts (SBS/yr)(y-axis). Each dot represents one individual and they are coloured according to the individual's germline mutation; orange; *MUTYH*<sup>Y179C+/-/G396D+/-</sup>, yellow; *MUTYH*<sup>Y179C+/-</sup>, green; *MUTYH*<sup>Y104+/-</sup> and blue; *MUTYH*<sup>G286E+/-</sup>. The rate of *MUTYH* associated mutational processes is ~13x fold (linear model, 95% C.I.; 10-17) higher in intestine vs blood. Black line indicates the ratio, and dotted lines the 95% C.I.. Plot inset shows the mutation rate for *n* = 9 patients excluding the outlier, PD44890. **c** Stacked bar plot displaying the mutational signature contribution in each peripheral blood sample organised by patient. Coloured squares indicate the *MUTYH* germline mutation. Normal control data from granulocytes sequenced using the same method (data from Abascal et al 2021)<sup>50</sup>. Significantly higher proportion of SBS18 and SBS36 is observed in individuals with *MUTYH* mutations vs normal healthy controls (two-sided Wilcoxon rank sum, *P* = 0.00004). **d** SBS mutation burden in intestinal lymphocyte cells from wild-type healthy individuals (grey) and individuals with *MUTYH* mutations (coloured according to the germline *MUTYH* genotype) plotted against age (years). Dots represent median values and whiskers represent the 95% confidence interval. Dashed line indicates the rate of increase of SBS burden in wild-type lymphocytes (40 SBS/yr, linear mixed-effects model, *R*<sup>2</sup> = 0.68, 95% C.I., 13-66, *P* = 0.01) and dotted line indicates the rate of increase in SBS burden in lymphocytes from individuals with *MUTYH* mutations (53 SBS/yr, linear mixed-effects model, *R*<sup>2</sup> = 0.68, 95% C.I., 21-85, *P* = 0.01). **e** Stacked bar plots showing the absolute (above) and relative (below) contributions of each mutational signature in tissue lymphocytes from wild-type healthy individuals and individuals with *MUTYH* mutations. Source data are provided as a Source Data file.

~29-fold, and therefore equivalent to a “mutational age” of ~450 years. Whilst lesser increases in mutation rates compared to wild-type individuals were observed in other tissues from PD44890, ~8-fold in white blood cells and ~7 fold in tissue lymphocytes, these still conferred substantially elevated “mutational ages” in the absence of features of premature ageing. Thus, direct deleterious effects of base substitutions accumulated over the course of a lifetime may not be an important cause of ageing.

The elevated mutation rate in normal intestinal epithelium likely contributes to the increased risk of colorectal adenomas and cancers in individuals with *MUTYH* mutations. Indeed, there appears to be a correlation between the extent of elevation of mutation rate and the rate of acquisition of colorectal adenomas. Individuals with the *MUTYH*<sup>Y104<sup>\*</sup>+/+</sup> and *MUTYH*<sup>Y179C+/+</sup> genotypes exhibited greater increases in somatic mutation rates than individuals with the *MUTYH*<sup>Y179C+/- G396D+/-</sup> genotype. Previous detailed clinical phenotyping of large series indicates that individuals with biallelic truncating mutations or *MUTYH*<sup>Y179C+/+</sup> show higher rates of accumulation of adenomas and earlier age of onset of carcinoma<sup>22,23</sup> than *MUTYH*<sup>Y179C+/- G396D+/-</sup>. The correlation between elevation of mutation rate and severity of clinical phenotype is further highlighted by individual PD44890 (16 years of age, *MUTYH*<sup>Y179C+/- G396D+/-</sup>) who exhibited a substantially higher mutation rate than others of this genotype, and showed a much accelerated rate of colorectal adenoma development (Supplementary Data 1, 2). We previously described ~7-fold elevated genome-wide base substitution mutation rates in intestinal cells of *POLE* germline exonuclease domain mutation carriers<sup>42</sup>. *POLE* mutation carriers, however, show lower colorectal adenoma rates than *MUTYH* biallelic mutation carriers who generally only show 2–4 fold increased mutation rates. This apparent discrepancy may, however, be explained by the genomic distribution of mutations. In *POLE* mutation carriers there is relative sparing of coding sequences, with only a three to four-fold increase in exonic mutations in intestinal cells, whereas this sparing is less pronounced in *MUTYH* mutation carriers leading to similar increases in exonic mutation rates (Supplementary Fig. 8a, b). These observations lead to the proposition that measurements of somatic coding mutation rates undertaken early in life could, in future, be used to refine individual cancer risk predictions for *POLE/POLD* and *MUTYH* germline mutation carriers.

As for many other cancer predisposition syndromes, it is unclear why *MUTYH* germline mutations lead particularly to intestinal neoplasia. Elevated somatic mutation rates are also found in white blood cells in MAP individuals (and may therefore be present in other tissues) although the increases appear lesser in extent than in intestinal cells. The propensity to generate SBS18 mutations appears greater in wild type intestinal cells than in other cell types<sup>49</sup> and this may also be contributory. Extending this study to investigate somatic mutagenesis in a greater number of tissues with varying cancer incidence and in larger cohorts of individuals may offer further insight into the role mutation rates and mutational signatures play in tissue specific cancer risk.

In summary, we report elevated somatic base substitution rates characterised by distinctive mutational signatures in normal tissues from individuals with MAP. These findings underscore previous observations that elevated somatic base substitution rates are largely tolerated by cells and do not overtly accelerate the process of ageing. It is likely, however, that increased mutation rates in normal intestinal cells throughout life lead to increased rates of accumulation of driver mutations and, hence, the progression of neoplastic clones culminating in cancer.

## Methods

**Ethical approval and study participants.** This research complies with all relevant ethical regulations. MAP patients were recruited as part of Wales Research Ethics

Committee (REC) 12-WA0071 and 15-WA0075 and samples collected were approved for use in this project by REC 18/ES/0133. Normal healthy controls were recruited as part of the following UK Research Ethics Committee (REC) studies; 15/WA/0131, 15/EE/0152, 18/ES/0133 and 08/h0304/85 + 5.

Informed consent was obtained from all participants and no monetary compensation was offered for their participation. Consent was obtained for publication of demographics including age, sex, phenotypic features and other potentially identifiable data. A complete list of study participants and tissue samples is summarised in Supplementary Data 1, 2.

**DNA extraction from bulk samples.** Frozen whole blood underwent DNA extraction using the Genra Puregene Blood Kit (Qiagen). Briefly, 1–2 ml of frozen blood were thawed, lysed in RBC lysis solution and centrifuged. Cell pellet was resuspended in cell lysis solution and incubated at 37 °C for 2 h. RNA and protein were degraded using RNase A solution and protein precipitation solution. DNA was precipitated with isopropanol.

**Tissue preparation.** Tissues were embedded in Optimal Cutting Temperature (OCT) compound, frozen histological sections were cut at 25–30 µm and mounted on polyethylene naphthalate (PEN) slides and fixed in 70% ethanol for 5 minutes followed by two washes with phosphate buffered saline for 1 min each. Slides were manually stained in haematoxylin and eosin using a conventional staining protocol. A subset of samples were fixed in RNAlater (Sigma Aldrich) according to manufacturer’s instructions. Fixed tissue samples were embedded in paraffin using a Tissue-Tek tissue processing machine (Sakura). No formalin was used in the preparation, storage, fixation or processing of samples. Processed tissue blocks were embedded in paraffin wax, sectioned to 10 µm thickness and mounted onto PEN slides (Leica). Tissue slides were stained using a standard haematoxylin and eosin (H&E) protocol. Slides were temporarily cover-slipped and scanned on a Nano-Zoomer S60 Slide Scanner (Hamamatsu), images were viewed with NDP.View2 software (Hamamatsu).

**Histopathological review of tissues.** All tissues studied were carefully reviewed using the following approach: (1) tissue blocks were reviewed by a pathologist / clinician at the time of sampling and classified based on their macroscopic appearance as being normal or adenoma. (2) Following histological sectioning and high-resolution scanning, tissue sections were categorised as being normal, adenoma or cancer. Slides that were indeterminate were referred to a gastrointestinal pathologist for review. (3) Prior to laser-capture microdissection, each crypt was carefully inspected using the 40x digital scan and classified as being normal or dysplastic / adenoma. Initial review, in stages 2 and 3, was principally undertaken by an experienced clinician with a special interest in gastrointestinal pathology. Glands were classified as adenomatous if they bore histological features indicating them to be dysplastic or likely dysplastic and harboured a cancer “driver” mutation.

**Laser capture microdissection.** Laser capture microdissection was undertaken using a LMD7000 microscope (Leica) into a skirted 96-well PCR plate. Cell lysis was undertaken using 20 µl proteinase-K PicoPure® DNA Extraction kit (Arc-turus®). Samples were incubated at 65 °C for 3 h followed by proteinase denaturation at 75 °C for 30 min. Thereafter samples were stored at –20 °C prior to DNA library preparation.

**Low-input DNA library preparation and sequencing.** DNA library preparation of micro-dissected tissue samples was undertaken using a bespoke low-input enzymatic-fragmentation-based library preparation method<sup>2–4,37</sup>. This method was employed as it allows for high quality DNA library preparation from a very low starting quantity of material (from 100–500 cells). In brief, gDNA was purified from cell lysates using bead purification. Enzymatic fragmentation, end-repair and dA-tailing was performed using NEBNext Ultra II FS DNA Library Prep Kit (New England BioLabs). Indexing and PCR amplification was subsequently performed (12 cycles). DNA library concentration was assessed after library preparation and used to guide choice of samples to take forward to DNA sequencing. The minimum library concentration was 5 ng/µL and libraries with >15 ng/µL were preferentially chosen. 150 bp paired-end Illumina reads were prepared with Unique Dual Index barcodes (Illumina). DNA sequencing was undertaken on a NovaSeq 6000 platform using an XP kit (Illumina). Samples were multiplexed in pools of 6–24 samples. Pools were sequenced to achieve a coverage of ~30x per sample.

**Mutation calling and post-processing filters.** Sequencing reads were aligned to NCBI human genome GRCh37 using the Burrow-Wheeler Alignment (BWA-MEM). Single Base Substitutions (SBS) were called using the ‘Cancer Variants through Expectation Maximization’ algorithm (CaVEMan)<sup>55</sup>. Mutations were called using an unmatched normal synthetic bam file to retain early embryonic and somatic mutations. Post-processing filters were applied to remove low-input library preparation specific artefacts and germline mutations using a previously described method<sup>1,2,37,56</sup>. Filters applied were: (1) common single nucleotide polymorphisms were removed by filtering against a panel of 75 unmatched normal samples<sup>57</sup> (2) to remove mapping artefacts, mutations were required to have a minimum median



read alignment score of mutant reads ( $ASMD \geq 140$ ) and fewer than half of the reads supporting the mutation should be clipped ( $CLPM = 0$ ) (3) a filter to remove overlapping reads that result from the relatively short insert size, which could lead to double counting of variant reads; and (4) a filter to remove cruciform DNA structures that can arise during the low-input library preparation method.

Next, we applied multiple filters to remove germline variants and potential artefacts whilst retaining bona fide embryonic and somatic variants. This approach has been detailed in previous publications and the code for these filters can be found at [https://github.com/TimCoorens/Unmatched\\_NormSeq](https://github.com/TimCoorens/Unmatched_NormSeq). Mutations were aggregated per patient and a read pile-up was performed using an in-house algorithm (cgpVAF) to tabulate the read count of mutant and reference reads per sample for each mutation locus. Germline mutations were filtered out using an exact binomial test which distinguishes germline from somatic variants using the aggregate read counts across all samples from the same patient<sup>1,56</sup>. In brief, the read depth across all samples from each individual was calculated (median in this study 496-fold). This high coverage yields a very precise estimate of the true VAF of each mutation. While the VAF estimates of the earliest embryonic SBS and germline variants from samples sequenced at 30x might overlap, the VAFs from the much higher coverage achieved by aggregating all samples from each individual, are distinguishable using statistical testing. To achieve this, the beta-binomial test was applied. The over dispersion parameter ( $\rho$ ) threshold for genuine variants of  $\rho > 0.1$  was used.

Phylogenetic trees were created using MPBoot (version 1.1.0 bootstrapped – 1000 times) and mutations were mapped to branches using maximum likelihood assignment.

Indels (ID) were called using Pindel<sup>58</sup> using the same synthetic unmatched normal sample employed in SBS mutation calling. ID calls were filtered to remove calls with a quality score of  $< 300$  ('Qual'; sum of mapping qualities of the supporting reads) and a read depth of less than 15. Thereafter, ID filtering was performed in a similar manner as SBS to remove germline variants and library preparation / sequencing artefacts.

**Copy-number alteration calling.** Somatic copy-number variants (CNVs) were called using the Allele-Specific Copy number Analysis of Tumours (ASCAT) algorithm<sup>59</sup> as part of the in the ascatNGS package<sup>60</sup> (<https://github.com/Crick-CancerGenomics/ascat>). Bulk blood samples or phylogenetically unrelated normal samples were used as matched normals. ASCAT was run with default parameters. A bespoke filtering algorithm - ascatPCA - was used to reduce the number of false-positive calls that can arise when analysing genome sequences from normal tissue (<https://github.com/hj6-sanger/ascatPCA>). ascatPCA extracts a noise profile by aggregating the LogR ratio from across a panel of normal unrelated samples and subtracts this signature from that observed in the sample being analysed using principal component analysis.

**Structural variant calling.** Whole-genome sequences were analysed for somatic structural variants (SVs) using the Genomic Rearrangement Identification Software Suite (GRIDSS). In preparation for this analysis, genomes were remapped to Human Genome Version 38 and GRIDSS was run using the same matched normal as used for CNV analysis. Coordinates for SV calls were subsequently converted back to GRCh37. SV calls in L1 transposon donor regions and fragile sites were excluded from the final SV analysis.

**Mutational signature analysis.** The R package, HDP (<https://github.com/nicolaroberts/hdp>), based on the hierarchical Dirichlet process<sup>61</sup>, was used to extract mutational signatures. Analysis of mutational signatures using this package has been applied to normal tissues previously<sup>1–4</sup>. In brief, this nonparametric Bayesian method models categorical count data using the hierarchical Dirichlet process. A hierarchical structure is established using patients as the first tier (parent nodes) and individual samples as the second tier (dependent nodes). Uniform Dirichlet priors were applied across all samples. The algorithm creates a mutation catalogue for each sample and infers the distribution of signatures in any one sample using a Gibbs sampler. We performed mutational signatures analysis per-branch, counting each branch of the phylogenetic tree as a distinct sample to avoid double counting of mutations. Since the MCMC process scales linearly with the number of counts, we randomly subsampled each branch to a maximum of 2500 total substitutions. Branches with fewer than 100 mutations were excluded from the mutational signature extraction. No reference signatures were included as priors.

Next, to estimate the contribution of each mutational process, mutational signatures were refitted to all mutation counts using the R package sigfit (<https://github.com/kgori/sigfit>)<sup>62</sup>. To avoid overfitting, a limited subset of reference mutational signatures were included for each patient corresponding to the HDP signatures that were identified in that individual.

Ageing signatures SBS1 and SBS5 are present in all normal intestinal crypts<sup>3</sup>. Lower than expected burdens of SBS1 and SBS5 were observed in most individuals in this study due to: (1) the inherent challenges of accurately estimating mutation burden in highly mutated samples and (2) the appreciable contamination of reference signatures with SBS1 and SBS5. To partially address this, we used the extracted HDP component corresponding to SBS36 in the refitting stage which has

lower SBS1 and SBS5 contamination than the COSMIC reference SBS36 signature. Nevertheless, in individual PD44890 where SBS18 and SBS36 exposures are many tens of times greater than the normal mutation rate, the estimates of SBS1 and SBS5 are substantially lower than would be expected.

Settings / parameters used for mutational signature extraction:  
 SBS Signature Extraction – Hierarchical Dirichlet Process (HDP)  
 Chains: 20 MCMC chains  
 Iterations: 40,000  
 Burn-in: 20,000  
 Samples: 100 / chain  
 Signature components identified: 9  
 Component Names: HDP N0–HDP N8  
 SBS Signature Extraction – SigProfiler  
 input\_type: vcf  
 startProcess: 1  
 endProcess: 15  
 totalIterations: 1000  
 cpu: –1  
 hierarchy: True  
 refgen: GRCh37  
 genome\_build: GRCh37  
 mtype: ['default']  
 init: random

**Mutational signature extraction.** Signature extraction was performed using two independent methods; the Hierarchical Dirichlet Process (HDP) and SigProfiler. De novo extraction was performed to extract / identify mutational signatures. HDP signature extraction yielded 9 signature components (N0–N8) (Supplementary Figs. 9–17). Components showing close similarity to known reference signatures and were retained as their reference signature (HDP N0 as SBS5, HDP N1 as SBS36, HDP N5 as SBS5, HDP N6 as SBS38, HDP N8 as SBS17b). To deconvolute the other signature components and equate them to known COSMIC reference signatures, an expectation maximisation algorithm was used. Decomposed signature components are shown in Supplementary Figs. 18–21. HDP N2 was broken down into SBS18 and SBS36, HDP N3 into SBS1, SBS18 and SBS36 and HDP N7 into SBS18 and SBS88. A further component, HDP N4 was unable to be fully deconvoluted into known mutational signatures so was retained in its original format for the next stage of analysis.

Signatures were re-fitted to the mutation counts for each branch of the phylogenetic tree to establish the absolute contributions of each mutational signature using the R package SigFit (<https://github.com/kgori/sigfit>). To prevent overfitting, a limited subset of reference signatures was used corresponding to HDP components identified in that patient. Furthermore, any signatures occurring at less than 10% exposure were excluded to prevent over-fitting. Therefore 7 mutational signatures were refitted; SBS1, SBS5, SBS17b, SBS18, SBS36, SBS38 and SBS88.

**Validation of mutational signatures.** To validate the mutational signatures extracted using the HDP method, we used the non-negative matrix factorisation (NNMF) based algorithm SigProfiler. Using the same input data, SigProfiler generated 5 signature components (SigProfiler A–E, Supplementary Fig. 23). SigProfiler generated fewer signature components than HDP (5 vs 9). SigProfiler components SigProfiler.A, SigProfiler.B, SigProfiler.C which accounted for the majority of mutations in the data set, had clear counterparts among the HDP signature components (HDP N1–3) (Supplementary Figs. 22–26). Additional signature components were stably extracted by HDP but not by SigProfiler.

**ID mutational signature analysis.** ID mutational signatures were extracted using the HDP method identifying a total of 5 signatures (HDP N0–N4) (Supplementary Figs. 27–31). Component HDP N1 bore close similarity to COSMIC reference signature ID1, component HDP N2 to ID1 and ID2; and HDP N3 to ID18 (Supplementary Fig. 32–34). Component HDP N4 had no clear comparator among known reference signatures, and deconvolution was unable to adequately recapitulate the original signature component.

ID mutational signature extraction was validated using SigProfiler, which identified 4 signature components that closely correspond to those identified by HDP (Supplementary Fig. 35).

**Cancer driver mutations.** Cancer driver mutations were identified using two methods aiming to identify genes and mutations in this cohort that are subject to positive selection. Firstly, to identify mutations in cancer genes under positive selection in an unbiased manner, we ran a modified dNdS method<sup>63</sup>. To avoid double-counting of mutations, only unique mutations (SBS and ID) mapped to branches of the phylogenetic trees were analysed. dNdScv was run using the following parameters; max\_coding\_muts\_per\_sample = 5000 and max\_muts\_per\_gene\_per\_sample = 20. Genes with a qval of  $< 0.05$  were considered to be under positive selection.

A second phase of cancer driver mutation analysis was undertaken, identifying mutations in this cohort that are codified in cancer mutation databases and exhibit

characteristic traits of driver mutations, an approach that has been previously employed in the study of normal tissues<sup>1,2</sup>. Analysis was restricted to somatic mutations (SBS and ID) in protein coding regions and mutations were filtered using lists of known cancer genes; mutations in samples from intestinal epithelium were filtered using a list of 90 genes associated with colorectal cancer which includes genes that are common in small bowel adenocarcinoma<sup>3</sup>. Samples from all other tissues, including blood, were filtered using a pan-cancer list of 369 driver genes<sup>63</sup>. Genes were then characterised according to their predominant molecular behaviour: dominant, recessive or intermediate (those demonstrating aspects of both types of behaviour), using the COSMIC Cancer Gene Census<sup>64</sup>. Potential hotspot mutations were annotated using the cBioportal MutationMapper database ([https://www.cbioportal.org/mutation\\_mapper](https://www.cbioportal.org/mutation_mapper)). Mutations meeting the following criteria were considered to be driver mutations: truncating mutations (those that cause a shortened RNA transcript i.e. nonsense, essential splice-site, splice region and frameshift ID) in recessively acting genes, known activating hotspot mutations in dominant (and recessive) genes. Lastly, mutations that were in neither of the above categories but were characterised by the MutationMapper database as being 'likely oncogenic' were also included in the final driver mutation catalogue. We then compared the frequency of driver mutations in histologically normal crypts with *MUTYH* mutations to a cohort of  $n = 445$  normal intestinal crypts<sup>3</sup> from wild-type individuals that were analysed using the same method.

**Generation and processing of data from the wild-type control cohort.** Data generated in this study was compared to a cohort of healthy individuals with no germline *MUTYH* mutation. The wild-type cohort was generated as part of a previously published study<sup>3</sup> and comprises  $n = 445$  normal intestinal crypts that were processed using the same laboratory methods employed in this paper. While most of the bioinformatic analysis in the original paper followed the same pipeline employed in this study, there were some small differences in the methods used for filtering. Therefore, we re-filtered the mutations in the wild-type cohort using the same parameters employed in the *MUTYH* cohort. Mutation burden estimates were corrected for sensitivity in the same manner as the *MUTYH* cohort (described below). This was particularly important as the wild-type cohort was sequenced at a lower median coverage than the *MUTYH* cohort (~16-fold vs ~28-fold). In the original study, Lee-Six et al report a mutation burden of 43.6 SBS/yr in normal crypts. Here, using data that was re-filtered, we obtain an estimate of the mutation rate that is highly concordant (46 SBS/yr).

**Mutation calling sensitivity.** The sensitivity of calling mutations in a genome sequence is strongly influenced by the depth of sequencing coverage and clonality of the sample. Natural variation in sequencing coverage and the clonality of samples may, therefore, influence the sensitivity to call mutations and hence the genome-wide mutation burden estimate. To account for these differences, we calculated the sensitivity of mutation calling from its two principal determinants, sequencing coverage and clonality using a previously validated method<sup>3,37,47</sup>. First, we sampled a range of sequence coverage values from a Poisson distribution centred around the mean coverage for the sample being analysed. Next, using these values we simulated the number of sequencing reads at each site using a truncated binomial distribution based on the median VAF of each sample. The sensitivity was then calculated as the fraction of simulated mutation calls with 4 or more reads, which is the minimum number of reads that the SBS mutation calling algorithm, CaVEMan, requires to call a mutation. The genome-wide mutation SBS burden was then corrected by dividing it by the estimated sensitivity to give the corrected SBS mutation count.

**Mutation burden estimates, modelling and fold-changes.** Statistical modelling was performed to assess the mutation rate associated with each germline genotype. A linear mixed-effects model was used to assess the SBS mutation rate for each of the main *MUTYH* germline genotypes and the re-filtered wild-type control data. Fold-changes in the SBS mutation rate were calculated by dividing the modelled mutation rates for the five different *MUTYH* germline mutation groupings by the modelled mutation rate for the wild-type group.

**Telomere content estimation.** Telomere attrition is a hallmark of ageing observed in normal cells<sup>65,66</sup>. To assess whether telomere shortening is altered in normal tissues from individuals with MAP compared with wild-type controls, we used a bioinformatic method to assess the telomere content of DNA sequencing files called TelomereHunter. TelomereHunter has previously been applied to the study of telomere biology in cancer samples and normal tissues<sup>42,67,68</sup>. We applied TelomereHunter to estimate the telomere content in normal intestinal crypts with *MUTYH* mutations ( $n = 144$ ) and in wild-type controls ( $n = 445$ ). Next, we used linear mixed-effects modelling to assess the rate of telomere attrition in the *MUTYH* and wild-type cohorts. No significant difference was observed, thus implying that telomere maintenance is not overtly dysregulated in individuals with germline *MUTYH* mutations.

**Modified duplex sequencing (NanoSeq).** DNA from bulk blood samples from individuals with germline *MUTYH* mutations was extracted as outlined above.

Samples from normal healthy control was obtained and processed using the following method. Whole blood was diluted with PBS and mononuclear cells (MNC) were isolated using lymphoprep™ (STEMCELL Technologies) density gradient centrifugation. The red blood cell and granulocyte fraction of the blood was then removed. The MNC fraction was depleted of red blood cells by lysis steps involving 3 incubations at room temperature for 20 mins/10 mins/10 mins respectively with RBC lysis buffer (BioLegend). Tissue lymphocytes were isolated from Peyer's patches in intestinal mucosa using laser capture microdissection and subjected to protein lysis as outlined above. Cell lysates were processed and whole genome sequenced using the NanoSeq protocol.

Our modified duplex sequencing method, called NanoSeq, relies on blunt-end restriction enzymes to fragment the genome in order to avoid errors associated to the filling of 5' overhangs and the extension of internal nicks during end repair after sonication. Our modified method has error rates  $< 5e-9$ <sup>50</sup>.

Given the uneven frequencies of trinucleotides in the digested genome, the strong filtering of common SNPs sites (typically occurring at CpG), and the strong dependence of mutation rates on trinucleotide contexts, our estimates of mutation burdens are normalised and projected onto genomic trinucleotide frequencies.

Let  $t$  denote the count of a given trinucleotide of type  $i = 1 \dots 32$ . The frequency of each trinucleotide is calculated separately for the genome  $f_i^g$  and for the NanoSeq experiment  $f_i^e$  where (Formula 1):

$$f_i = \frac{t_i}{\sum_{i=1}^{32} t_i} \quad (1)$$

The ratio of genomic to experimental frequencies for a given trinucleotide is (Formula 2):

$$r_i = \frac{f_i^g}{f_i^e} \quad (2)$$

There are  $j = 1 \dots 6$  classes of substitution where the mutated base is a pyrimidine. Let  $s_{ij}$  denote the count of substitution  $j$  in trinucleotide context  $i$ , giving a total of 96 substitution classes. Each substitution count is corrected as follows (Formula 3):

$$s'_{ij} = s_{ij} r_i \quad (3)$$

The corrected substitution counts provide a substitution profile projected onto the human genome, and are also used to calculate the corrected mutation burden (Formula 4):

$$\beta' = \frac{\sum_{i=1}^{32} \sum_{j=1}^6 s'_{ij}}{\sum_{i=1}^{32} t_i} \quad (4)$$

Software used in this study is publicly available at the following locations:

- Mutation calling algorithms are available at <https://github.com/cancerit>.
- Code for filtering mutation calls is available at [https://github.com/TimCoorens/Unmatched\\_NormSeq](https://github.com/TimCoorens/Unmatched_NormSeq).
- Software for mutational signature analysis is available at <https://github.com/nicolaroberts/hdp> and <https://github.com/kgori/sigfit> and <https://github.com/AlexandrovLab>.
- Software for analysis of duplex / NanoSeq data is provided at <https://github.com/cancerit/NanoSeq>.
- Parameters used for these various pieces of software have been included in the manuscript methods section and supplementary information.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Source data are provided with this paper. Raw DNA sequencing data are deposited in the European Genome-Phenome Archive (EGA) with accession codes: [EGAD00001007958](https://ega-archive.org/submission) and [EGAD00001007997](https://ega-archive.org/submission). To ensure the data is used for academic and research purposes, the DNA sequencing data are available via controlled access. Applications to access the data should be directed to the WTSI CGP Data access committee via the contact details listed at the above links. Indefinite access to the data will be made upon request. Further details of the access policy are available at <https://ega-archive.org/submission>. The cBioPortal MutationMapper database was accessed at: [https://www.cbioportal.org/mutation\\_mapper?standaloneMutationMapperGeneTab=ATM](https://www.cbioportal.org/mutation_mapper?standaloneMutationMapperGeneTab=ATM). The COSMIC Cancer Gene Census is available to download at: <https://cancer.sanger.ac.uk/census>. There are no restrictions to accessing the MutationMapper or COSMIC databases.

## Code availability

Code/software required to reproduce the analyses in this paper are available online and are listed in Methods. Code required to reproduce the figures in this manuscript are available online<sup>69</sup> at: <https://github.com/PhilipsRobinson/mutyh>; <https://doi.org/10.5281/zenodo.6504797>.

Received: 12 October 2021; Accepted: 14 June 2022;

Published online: 08 July 2022

## References

- Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
- Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
- Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
- Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
- Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **917**, 911–917 (2018).
- Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
- Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
- Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
- Balaban, R. S., Nemoto, S. & Finkel, T. Mitochondria, oxidants, and aging. *Cell* **120**, 483–495 (2005).
- Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J.* **17**, 1195–1214 (2003).
- Cheng, K. C., Cahill, D. S., Kasai, H., Nishimura, S. & Loeb, L. A. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G → T and A → C substitutions. *J. Biol. Chem.* **267**, 166–172 (1992).
- Rosenquist, T. A., Zharkov, D. O. & Grollman, A. P. Cloning and characterization of a mammalian 8-oxoguanine DNA glycosylase. *Proc. Natl Acad. Sci. USA* **94**, 7429–7434 (1997).
- McGoldrick, J. P., Yeh, Y. C., Solomon, M., Essigmann, J. M. & Lu, A. L. Characterization of a mammalian homolog of the *Escherichia coli* mutY mismatch repair protein. *Mol. Cell. Biol.* **15**, 989–996 (1995).
- Komine, K. et al. Functional complementation assay for 47 MUTYH variants in a MutY-disrupted *Escherichia coli* strain. *Hum. Mutat.* **36**, 704–711 (2015).
- Ruggieri, V. et al. Loss of MUTYH function in human cells leads to accumulation of oxidative damage and genetic instability. *Oncogene* **32**, 4500–4508 (2013).
- Wooden, S. H., Bassett, H. M., Wood, T. G. & McCullough, A. K. Identification of critical residues required for the mutation avoidance function of human MutY (hMYH) and implications in colorectal cancer. *Cancer Lett.* **205**, 89–95 (2004).
- Kundu, S., Brinkmeyer, M. K., Livingston, A. L. & David, S. S. Adenine removal activity and bacterial complementation with the human MutY homologue (MUTYH) and Y165C, G382D, P391L and Q324R variants associated with colorectal cancer. *DNA Repair* **8**, 1400–1410 (2009).
- Parker, A. R. et al. Cells with pathogenic biallelic mutations in the human MUTYH gene are defective in DNA damage binding and repair. *Carcinogenesis* **26**, 2010–2018 (2005).
- Sampson, J. R., Jones, S., Dolwani, S. & Cheadle, J. P. MutYH (MYH) and colorectal cancer. *Biochem Soc. Trans.* **33**, 679–683 (2005).
- Sampson, J. R. et al. Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet* **362**, 39–41 (2003).
- Al-Tassan, N. et al. Inherited variants of MYH associated with somatic G → T: A mutations in colorectal tumors. *Nat. Genet.* **30**, 227–232 (2002).
- Collaborative Group on Duodenal Polyposis in MAP; Thomas, L. E. et al. Duodenal adenomas and cancer in MUTYH-associated polyposis: an international cohort study. *Gastroenterology* **160**, 952–954 e954 (2021).
- Nielsen, M. et al. Analysis of MUTYH genotypes and colorectal phenotypes in patients With MUTYH-associated polyposis. *Gastroenterology* **136**, 471–476 (2009).
- Lubbe, S. J., Di Bernardo, M. C., Chandler, I. P. & Houlston, R. S. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J. Clin. Oncol.* **27**, 3975–3980 (2009).
- Win, A. K. et al. Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. *Gastroenterology* **146**, 1208–1211 e1201-1205 (2014).
- Theodoratou, E. et al. A large-scale meta-analysis to refine colorectal cancer risk estimates associated with MUTYH variants. *Br. J. Cancer* **103**, 1875–1884 (2010).
- Cleary, S. P. et al. Germline MutY human homologue mutations and colorectal cancer: a multisite case-control study. *Gastroenterology* **136**, 1251–1260 (2009).
- Vogt, S. et al. Expanded extracolonic tumor spectrum in MUTYH-associated polyposis. *Gastroenterology* **137**, 1976–1985 e1971-1910 (2009).
- Thomas, L. E. et al. Burden and profile of somatic mutation in duodenal adenomas from patients with familial adenomatous- and MUTYH-associated polyposis. *Clin. Cancer Res.* **23**, 6721–6732 (2017).
- Rashid, M. et al. Adenoma development in familial adenomatous polyposis and MUTYH-associated polyposis: Somatic landscape and driver genes. *J. Pathol.* **238**, 98–108 (2016).
- Viel, A. et al. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine* **20**, 39–49 (2017).
- Pilati, C. et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- Ritsma, L. et al. Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. *Nature* **507**, 362–365 (2014).
- Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
- Barker, N. et al. Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449**, 1003–1007 (2007).
- Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* <https://doi.org/10.1038/s41596-020-00437-6> (2020).
- Audebert, M. et al. Alterations of the DNA repair gene *OGG1* in human clear cell carcinomas of the kidney. *Cancer Res* **60**, 4740–4744 (2000).
- Audebert, M., Radicella, J. P. & Dizdaroglu, M. Effect of single mutations in the *OGG1* gene found in human tumors on the substrate specificity of the OGG1 protein. *Nucleic Acids Res* **28**, 2672–2678 (2000).
- Forbes, S. A. et al. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
- Mur, P. et al. Germline variation in the oxidative DNA repair genes *NUDT1* and *OGG1* is not associated with hereditary colorectal cancer or polyposis. *Hum. Mutat.* **39**, 1214–1225 (2018).
- Robinson, P. S. et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* **53**, 1434–1442 (2021).
- Temko, D. et al. Somatic *POLE* exonuclease domain mutations are early events in sporadic endometrial and colorectal carcinogenesis, determining driver mutational landscape, clonal neoantigen burden and immune response. *J. Pathol.* **245**, 283–296 (2018).
- Lin, S. H. et al. The somatic mutation landscape of premalignant colorectal adenoma. *Gut* **67**, 1299–1305 (2018).
- Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–45 (2018).
- Pleguezuelos-Manzano, C. et al. Mutational signature in colorectal cancer caused by genotoxic pks(+) *E. coli*. *Nature* **580**, 269–26 (2020).
- Olafsson, S. et al. Somatic evolution in non-neoplastic IBD-affected colon. *Cell* **182**, 672–684 e611 (2020).
- Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
- Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381 (2020).
- Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405 (2021).
- Zou, X. Q. et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat. Cancer* <https://doi.org/10.1038/s43018-021-00200-0> (2021).
- Xie, Y. et al. Deficiencies in mouse Myh and Ogg1 result in tumor predisposition and G to T mutations in codon 12 of the K-ras oncogene in lung tumors. *Cancer Res* **64**, 3096–3102 (2004).
- Jones, S. et al. Increased frequency of the k-ras G12C mutation in MYH polyposis colorectal adenomas. *Br. J. Cancer* **90**, 1591–1593 (2004).
- Palles, C. et al. Germline mutations affecting the proofreading domains of *POLE* and *POLD1* predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **45**, 136–143 (2013).
- Jones, D. et al. 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 (2016).
- Coorens, T. H. H. et al. Embryonal precursors of Wilms tumor. *Science* **366**, 1247–124 (2019).
- Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
- Raine, K. M. et al. cgpPindel: Identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.17.11–15.17.12 (2015).



59. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
60. Raine, K. M. et al. ascatNgs: Identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinforma.* **56**, 15.19.11–15.19.17 (2016).
61. Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* **101**, 1566–1581 (2006).
62. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. Preprint at <https://www.biorxiv.org/content/10.1101/372896v2> (2020).
63. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 e1021 (2017).
64. Sondka, Z. et al. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
65. Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
66. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
67. Feuerbach, L. et al. TelomereHunter - in silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics* **20**, 272 (2019).
68. Sieverling, L. et al. Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat. Commun.* **11**, 733 (2020).
69. Robinson, P. S. Inherited *MUTYH* mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells. *Zenodo* <https://doi.org/10.5281/zenodo.6504797> (2022).

## Acknowledgements

We thank the staff of Wellcome Sanger Institute Sample Logistics, Genotyping, Pull-down, Sequencing and Informatics facilities for their contribution including Laura O'Neill, James Hewinson, Yvette Hooks, Stephen Gamble, Calli Latimer and Kirsty Roberts for their support with sample management and laboratory work. In addition; Moritz Gerstung and Harald Vöhringer for help with analysis, advice and discussions. James Chan (Cambridge University Hospitals), Geraint Williams and Emma Short (Cardiff University) for assistance with histopathological review. We thank the clinical and research administration teams at recruitment sites for their assistance with sample collection and patients and their families for their time as study participants. This work was supported by the Wellcome Trust [206194]. P.S.R. is supported by a Wellcome Clinical PhD fellowship. Sample collection and research governance were supported by grants to Wales Gene Park from Health and Care Research Wales (H.C.R.W.). L.E.T. was supported by a postdoctoral Fellowship from HCRW and H.D.W. by a postdoctoral Fellowship from Ser Cymru.

## Author contributions

P.S.R., M.R.S., J.R.S. and L.E.T. conceived the study design. J.R.S., L.E.T., F.L., L.B., N.L., H.D.W., R.B.S., S.R., R.t.H., N.C., S.J.A.B. and K.S.P. recruited individuals, collected samples and curated sample and clinical data. P.S.R., B.C.H.L., R.T. and S.V.L. undertook laboratory work. F.A., I.M., S.V.L., L.M.R.H., T.H.H.C. and M.A.S. developed bespoke DNA library preparation, sequencing and bioinformatic methods. F.A., I.M., L.M.R.H., H.L.S. and S.O. contributed and analysed control data. P.S.R., F.A., I.M., S.O. and H.J. performed data analysis. M.R.S., P.J.C. and I.M. oversaw statistical analysis. M.R.S. and J.R.S. oversaw the study. All authors were involved in the preparation and review of the manuscript.

## Competing interests

P.J.C. is a founder, consultant, and stockholder of Mu Genomics Ltd. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-31341-0>.

**Correspondence** and requests for materials should be addressed to Michael R. Stratton.

**Peer review information** *Nature Communications* thanks Joao Pedro de Magalhaes and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022