A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage

Xueqing Zou1,2,3, Gene Ching Chiek Koh1,2,3, Arjun Scott Nanda1,2, Andrea Degasperi1,2,3, Katie Urgo3, Theodoros I. Roumeliotis4, Chukwuma A. Agu3, Cherif Badja1,2,3, Sophie Momen1,2, Jamie Young1, Tauanne Dias Amarante1,2, Lucy Side5,6, Glen Brice7, Vanesa Perez-Alonso8, Daniel Rueda0,9, Celine Gomez3, Wendy Bushell3, Rebecca Harris1,3, Jyoti S. Choudhary4, Genomics England Research Consortium*, Josef Jiricny10,11, William C. Skarnes3,12 and Serena Nik-Zainal1,2,3 □

Mutational signatures are imprints of pathophysiological processes arising through tumorigenesis. We generated isogenic CRISPR–Cas9 knockouts (Δ) of 43 genes in human induced pluripotent stem cells, cultured them in the absence of added DNA damage and performed whole-genome sequencing of 173 subclones. ΔOGG1, ΔUNG, ΔEXO1, ΔRNF168, ΔMLH1, ΔMSH2, ΔMSH6, ΔPMS1 and ΔPMS2 produced marked mutational signatures indicative of them being critical mitigators of endogenous DNA modifications. Detailed analyses revealed mutational mechanistic insights, including how 8-oxo-2′-deoxyguanosine elimination is sequence context specific while uracil clearance is sequence context independent. Mismatch repair (MMR) deficiency signatures are engendered by oxidative damage (C > A transversions) and differential misincorporation by replicative polymerases (T > C and C > T transitions), and we propose a reverse template slippage model for T > A transversions. ΔMLH1, ΔMSH6 and ΔMSH2 signatures were similar to each other but distinct from ΔPMS2. Finally, we developed a classifier, MMRDetect, where application to 7,695 whole-genome-sequenced cancers showed enhanced detection of MMR-deficient tumors, with implications for responsiveness to immunotherapies.

Omatic mutations arising through endogenous and exogenous processes mark the genome with distinctive patterns termed mutational signatures1–4. While there have been advancements in the analytical aspects of deriving mutational signatures from human cancers5–7, there is an emerging need for experimental substantiation, elucidating etiologies and mechanisms underpinning these patterns8–11. Cellular models have been used to systematically study mutagenesis arising from exogenous sources of DNA damage8–11. Next, it is essential to experimentally explore genome-wide mutagenic consequences of endogenous sources of DNA damage in the absence of external DNA-damaging agents.

Lindahl and Nyberg12 noted that water and oxygen, which are essential for living organisms, are two of the most mutagenic molecules to DNA. Their seminal work demonstrated that spontaneous DNA lesions occur through endogenous biochemical activities such as hydrolysis and oxidation. Errors at replication are also an enormous potential source of DNA changes. Fortuitously, our cells are equipped with DNA repair pathways that constantly mitigate this endogenous damage13,14. In this work, we combine CRISPR–Cas9-based biallelic knockouts of a selection of DNA replicative/repair genes in human induced pluripotent stem cells (hiPSCs), whole-genome sequencing (WGS) and in-depth analysis of experimentally generated data to obtain mechanistic insights into mutation formation. It is beyond the scope of this manuscript to study all DNA repair genes. Thus, we have focused on 42 DNA replicative/repair pathway gene knockouts successfully generated through semi-high-throughput methods. We also compared our experimental data with reported cancer-derived signatures.

While there is substantial literature regarding DNA repair pathways and complex protein interactions involved in maintaining genomic integrity15–20, here we focus on directly mapping whole-genome mutational outcomes associated with DNA repair defects, critically, in the absence of applied, external damage. Therefore, this study allows us to identify replicative/repair genes that are fundamentally important to genome maintenance against endogenous DNA damage.

Results

Biallelic knockouts of DNA replicative/repair genes. We knocked out (Δ) 42 DNA replicative/repair pathway genes and an unrelated control gene, ATP2B4 (Fig. 1a,b and Supplementary Table 1). A pilot experiment was performed to standardize the experimental procedures (Methods and Extended Data Fig. 1). In the full-scale study, two knockout genotypes were generated per gene, except for

1Academic Department of Medical Genetics, School of Clinical Medicine, University of Cambridge, Cambridge, UK. 2MRC Cancer Unit, University of Cambridge, Cambridge, UK. 3Wellcome Sanger Institute, Hinxton, UK. 4The Institute of Cancer Research, Chester Beatty Laboratories, London, UK. 5UCL Institute for Women’s Health, Great Ormond Street Hospital, London, UK. 6Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton, UK. 7Southwest Thames Regional Genetics Service, St George’s University of London, London, UK. 8Pediatrics Department, Doce de Octubre University Hospital, i+12 Research Institute, Madrid, Spain. 9Hereditary Cancer Laboratory, Doce de Octubre University Hospital, i+12 Research Institute, Madrid, Spain. 10Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. 11Institute of Biochemistry, ETH Zurich, Zurich, Switzerland. 12William Harvey Research Institute, Queen Mary University of London, London, UK. 13A list of authors and their affiliations appears at the end of the paper.

E-mail: snz@mrc-cu.cam.ac.uk

NATURE CANCER | VOL 2 | JUNE 2021 | 643–657 | www.nature.com/natcancer
EXO1, MSH2, TDG, MDC1 and REV1, for which only one knockout genotype was obtained. All parental knockout lines were grown for 15 d under normoxic conditions (~20% oxygen). For each genotype, two single-cell subclones were derived for WGS, aiming for four sequenced subclones per edited gene (Fig. 1a). For single-genotype genes, three subclones were derived for ΔEXO1 and ΔMSH2 and four were derived for ΔTDG, ΔMDC1 and ΔREV1.

In total, 173 subclones were obtained from 78 genotyped knockout subclones for WGS, aiming for four subclones per genotype (Methods). For each genotype, subclones were cultured for 15 d. The bars and error bars represent means ± s.d. of subclone observations.

All classes of somatic mutation were called, subtracting variation of the primary hiPSC parental clone. Rearrangements were too infrequent to decipher specific patterns. We confirmed that mutational outcomes were neither due to off-target edits nor to the acquisition of new driver mutations (Methods). We verified that knockouts were biallelic, validated the protein loss via mass spectrometry and ensured that subclones in controlled experimental settings, if simply knocking out a gene (in

**Fig. 1 | Mutational consequences of DNA replicative/repair pathway gene knockouts.** a, Experimental workflow from the isolation of gene knockouts to generating subclones for WGS, KO, knockout. b, Forty-three genes were knocked out, including 42 DNA replicative/repair genes and one control gene (ATP2B4), FA, Fanconi anemia; HR, homologous recombination; ICL, interstrand crosslink; MMEJ, microhomology-mediated end joining; NER, nucleotide excision repair; NHEJ, nonhomologous end joining; TLS, translesion DNA synthesis. c, Distinguishing substitution profiles of control subclones and knockout subclones. The green line shows the cosine similarities between the bootstrapped profiles of controls and the aggregated control substitution profile. The axis shows the aggregated substitution number of each genotype of a knockout. d, Distinguishing indel profiles of control subclones and knockout subclones. The light blue line shows the cosine similarities between the bootstrapped indel profiles of controls and the aggregated control indel profile. The axis shows the aggregated indel number of each genotype of a knockout. In c and d, colored dots represent various gene knockouts that have mutational profiles significantly different from those of controls. e, De novo mutation numbers of knockout subclones (n = -2–4; Supplementary Table 2) cultured for 15 d. The bars and error bars represent means ± s.d. of subclone observations.
the absence of providing additional DNA damage) could produce a signature, the gene is critical to maintaining genome stability from endogenous DNA damage. It would manifest an increased mutation burden above the background and/or altered mutation profile (Extended Data Fig. 2). We found that background substitution and indel mutagenesis, associated with growing cells in culture, occurred at ~150 substitutions and around ten indels per genome and were comparable across all subclones (Supplementary Tables 2 and 3).

To address potential uncertainty associated with the relatively small number of subclones per knockout and variable mutation counts in each gene knockout (Methods), we generated bootstrapped control samples with variable mutation burdens (50–10,000). We calculated the cosine similarity between each bootstrapped sample and the background control (ΔATP2B4) mutational signature (means and standard deviations). A cosine similarity close to 1.0 indicates that the mutation profile of the bootstrapped sample is near-identical to the control signature. Cosine similarities could thus be considered across a range of mutation burdens (green line in Fig. 1c and light blue line in Fig. 1d). Next, we calculated cosine similarities between knockout profiles and controls (colored dots in Fig. 1c,d). A knockout experiment that does not fall within the expected distribution of cosine similarities implies a mutation profile distinct from controls (that is, the gene knockout has a signature). For substitution signatures, two additional dimensionality reduction techniques (namely, contrastive principal component analysis and t-distributed stochastic neighbor embedding (t-SNE)) were also applied to secure high-confidence mutational signatures (Extended Data Fig. 3 and Methods). This stringent series of steps would probably dismiss weaker signals and be highly conservative at calling mutational signatures.

We identified nine single-substitution signatures, two double-substitution signatures and six indel signatures. Two gene knockouts, ΔOGG1 and ΔUNG, produced only substitution signatures. Five gene knockouts, ΔMSH2, ΔMSH6, ΔMLH1, ΔPMS2 and ΔPMS1, presented substitution and indel signatures. Two gene knockouts, ARNF168 and ΔEXO1, had substitution and double-substitution signatures. ΔEXO1 also produced an indel signature. The average de novo mutation burden that accumulated for these nine knockouts ranged between 250 and 2,500 for substitutions and five and 2,100 for indels (Fig. 1e). Based on cell proliferation assays, mutation rates for each knockout were calculated and ranged between six and 129 substitutions and 0.39 and 126 indels per cell division (Supplementary Table 4). In the following sections, we dissect these experimentally generated signatures and compare them with one another, and with cancer-derived mutational signatures, to gain insights into the sources of endogenous DNA damage and mutational mechanisms.

Safeguarding the genome from oxidative DNA damage. Oxygen can generate reactive oxygen species and oxidative DNA lesions. The most common is 8-oxo-2′-deoxyguanosine (8-oxo-dG), although over 25 oxidative DNA lesions are known. 8-oxo-dG is predominantly repaired by base excision repair (BER). It has been speculated that a permissive mutational signature observed in cell-based experiments is due to culture-related oxidative damage[31]. It is similar to a mutational signature identified in adrenocortical cancers and neuroblastomas called RefSig18 (ref. 24) or SBS18 (ref. 6). Biallelic to a mutational signature identified in adrenocortical cancers and reported to generate a hypermutated version of a similar signature25.

adenines inappropriately paired with 8-oxo-dG, has also been loss of the MutY DNA glycosylase gene (MUTYH) damage would also result in these characteristic patterns. It is unclear whether other genes responsible for removing oxidative DNA lesions are due to oxidative damage, specifically implicating 8-oxo-dG. We expanded signature channels by considering ±2 bases flanking the mutated base. Higher-resolution assessment of the most dominant peak at TGC > TTC/GCA > GAA in ΔOGG1 showed an almost identical pattern to control samples carrying culture-related signatures and SBS18 (cosine similarity >0.95; Fig. 2c and Extended Data Fig. 4), strengthening the argument that the G > T/C > A transversions observed in cultured cells and SBS18 are indeed caused by 8-oxo-dG-related damage.

The ΔOGG1 signature is qualitatively analogous to the signature of ΔMUTYH-related adrenocortical cancers[32] (recently renamed SBS36), although the latter demonstrates hypermutator phenotypes and has its tallest peak at TCT (Fig. 2c). These similarities are explained by related but distinct roles played by OGG1 and MUTYH in repairing oxidation-related lesions: 8-oxo-dG can pair with C or A during DNA synthesis. However, 8-oxo-G/C mismatches are not mutagenic and oxidized guanines are simply excised by OGG1 (ref. 39). In contrast, 8-oxo-G/A mismatches are first repaired by MutY glycosylase (which removes the A), and repair synthesis by polymerase β or λ inserts a C opposite the oxidized base. The resulting 8-oxo-G/C pair is then excised by OGG1, as outlined earlier. This mechanistic relatedness probably explains why mutational signatures of ΔOGG1 and ΔMUTYH are qualitatively alike, if quantitatively dissimilar. Notably, that simple knockouts of OGG1 or MUTYH can result in overt mutational phenotypes suggests that these genes are indispensable for maintaining the genome against endogenous oxidative damage.

Lastly, we examined ΔOGG1 G > T/C > A mutations correcting for frequencies of the 16 trinucleotides in the reference genome and found that ΔOGG1 is depleted of mutations at GG/CC dinucleotides (Fig. 2c). Yet, previous literature reports 5′-G in GG, and the first two Gs in GG are more likely to be oxidized through intraduplex electron transfer reactions[29,30]. Therefore, one would expect elevated G > T/C > A mutation burdens in GG-rich regions when OGG1 is defunct. Our results may be explained by previous experiments demonstrating that 8-oxo-dG excision rates by OGG1 are sequence context dependent[31]: 8-oxo-dG excision at consecutive 5′-GGs is reported as inefficient compared with 5′-GC/G′-GG and 5′-AGC/G′-GCT because OGG1 employs a bend-and-flip strategy to recognize 8-oxo-dG[30]. Stacked adjacent 8-oxo-dGs have an increased kinetic barrier, preventing flipping out and removal of 8-oxo-dG[30]. While this may explain why OGG1 cannot repair oxidized guanines at GC/CC motifs, it remains unclear how these motifs are repaired as guanine oxidation does occur at such sites. For some GC/CC motifs, we suggest a possible explanation in the section ‘Endogenous DNA damage managed by MMR’.

Maintaining cytosines from deamination to uracil. Deamination involves hydrolytic loss of an amine group. At CpG dinucleotides, deamination of 5-methylcytosine into thymine is a well-studied, universal process[34–36], with C > T at CpGs (RefSig1/SBS1) found in many tumor types. However, hypermutator phenotypes of C > T at CpGs have been reported in cancers with biallelic loss of the methyl-binding domain 4 gene (MBD4)[37]. This example underscores a mutational process that is customarily under tight MBD4 regulation, wherein its knockout uncovers the potential magnitude of unrepaired endogenous deamination.

Spontaneous, hydrolytic deamination of cytosine to uracil occurs more slowly at ~100–500 cells per day[38]. Deamination of cytosine to uracil is rectified by uracil-N-glycosylase (UNG) via
Fig. 2 | Safeguarding the genome from oxidative damage and cytosine deamination. a. Substitution signatures of background mutagenesis (from the control ΔATP2B4), ΔOGG1, ΔUNG, ΔEXO1 and ΔRNFi68. b. Cosine similarity between mutational signatures of gene knockouts and cancer-derived mutational signatures12. c. Odds ratios of C > A occurring at 16 trinucleotides for ΔOGG1 and ΔMUTYH (SBS536). Calculations were corrected for the distribution of trinucleotides in the reference genome. An odds ratio of <1 with a 95% confidence interval of <1 implies that C > A mutations at that particular trinucleotide are less likely to occur (dashed line shows odds ratio of 1). The mutational profiles of C > A at GCA with ±2 flanking bases are shown for ΔATP2B4, ΔOGG1, SBS18 and SBS536. In a and c, bold blue letters represent base substitutions. The red letters in c represent the nucleotides at +2 or −2 positions from the base substitution. d. Odds ratios of C > T occurring at all 16 trinucleotides for ΔUNG and ΔNTHL1 (SBS530). The bold red letters represent base substitutions. e, f. Transcriptional strand asymmetry of the ΔEXO1 signature (e) and ΔRNFi68 signature (f). The insets show the counts of T > A/A > T or T > C/A > G mutations on transcribed and nontranscribed strands. The dots and error bars in e, f represent calculated odds ratios with 95% confidence intervals.

BER+1. Uracils that are not removed before replication can result in C > T mutations (Fig. 2a). There are signatures associated with enhanced APOBEC-related deamination in many cancers (RefSig/SBS 2 and 13). However, the consequence of UNG dysfunction is less clear. The ΔUNG signature comprised mainly C > T transitions. When corrected for reference genome trinucleotide frequencies, no trinucleotide preferences were observed (Fig. 2d), suggesting a general role for UNG activity on all uracils, regardless of sequence context. The ΔUNG signature is most similar to RefSig30 (cosine similarity = 0.88), previously associated with ΔNTHL1 (ref. 65). Both UNG and NTHL1 are BER glycosylases that process aberrant pyrimidines, which may explain the similarities between these signatures. However, when corrected for trinucleotide frequencies, the ΔNTHL1 signature shows preference for ACC, CCC and TCC trinucleotides in contrast with ΔUNG, supporting that they are signatures of different etiologies.

Preserving thymines and adenines from T > C/A > G transitions. Two genes (ΔEXO1 and ΔRNFi68) with wide-ranging roles in repair/checkpoint pathways66–68 showed mutational signatures. ΔEXO1 encodes a 5’ to 3’ exonuclease with RNase H activity. ΔEXO1 generated substitution, double-substitution and indel signatures in hiPSCs (Fig. 2a and Extended Data Fig. 5), consistent with a previous report of ΔEXO1 in HAP1 lines57. In HAP1 cells, ΔEXO1 had stronger C > A components, probably reflecting differences in model systems. ΔEXO1 also produced a double-substitution pattern defined by TCA > AT, T > AA and GG > AA mutations, and an indel signature characterized by 1-base pair (bp) A/T insertions at long (>25-bp) poly[d(A-T)] sequences and 1-bp deletions at short (<5-bp) poly[d(A-T)] or poly[d(C-G)] sequences (Extended Data Fig. 5).

ΔRNFi68 encodes an E3 ubiquitin ligase involved in DNA double-strand break (DSB) repair69 that regulates 53BP1, BRCA1 and RAD18 recruitment to DSBs through ubiquitin-dependent signaling44–46. The substitution signature of ΔRNFi68 has two T > C peaks at ATAT > ACA and 1TA > TCA (Fig. 2a) and shares similarity with ΔEXO1 (cosine similarity = 0.94). Double-substitution patterns were defined by TCA > AA and GC > AA mutations. An indel signature was not observed for ΔRNFi68.

Substitution signatures of ΔEXO1 and ΔRNFi68 are most similar to RefSig5 of cancer-derived signatures (Fig. 2b (cosine similarity = 0.89–0.90) and Extended Data Fig. 6a, b), defined mainly by T > C/A > G substitutions. Additionally, ΔEXO1 and ΔRNFi68 signatures show transcriptional strand bias for T > C/A > G mutations.
Endogenous DNA damage managed by MMR. Knockouts of five genes involved in the MMR pathway—MSH2, MSH6, MLH1, PMS2 and PMS1, produced substitution and indel signatures (Fig. 3a,b). ∆MLH1, ∆MSH2 and ∆MSH6 produced qualitatively identical substitution signatures (cosine similarity = 0.99) characterized by a single strong peak at C>T/CAT/AGG> ATG and multiple peaks of C>T and T>C (Fig. 3a). In contrast, ∆PMS2 generated a signature of predominantly T>C transitions with a predominance at ATA, ATG and CTG (Fig. 3a). The single peak at C>T/CAT/AGG> ATG remained visible in the ∆PMS2 signature, albeit markedly reduced (10–3%). In addition, ∆MSH2, ∆MSH6 and ∆MLH1 generated indel signatures dominated by A/T deletions at long repetitive sequences. In contrast, ∆PMS2 produced similar proportions of A/T insertions and A/T deletions at long repetitive sequences (Fig. 3b and Extended Data Fig. 5a,b). ∆PMS1 generated A/T deletions only at long (≥5 bp) poly(dA–T) sequences and long deletions (>1 bp) at repetitive sequences (Extended Data Fig. 5a).

In-depth analysis of these mutational signatures allowed us to determine putative sources of endogenous DNA damage (Fig. 3c) acted on by MMR.

First, we consistently observed replication strand bias across ∆MLH1, ∆MSH2, ∆MSH6 and ∆PMS2. C>A on the lagging strand (equivalent to G>T leading strand bias); C>T on the leading strand (or G>A lagging) and T>C on the lagging strand (or A>G leading) (Fig. 3d). Similar results were previously reported in yeast and human cancers50–52. Under our experimental settings, where exogenous DNA damage was not administered, mismatches may be generated by DNA polymerases α, β or ε during replication. In the absence of MMR, these lesions become permanently etched as mutations. To understand which replicative polymerases could be causing these mutations, we analyzed putative progeny of all 12 possible base/base mismatches (Extended Data Fig. 7). T/G mismatches are the most thermodynamically stable and represent the most frequent polymerase error53. Our assessment suggests that the predominance of T>C transitions on the lagging strand can only be explained by misincorporation of G by lagging strand polymerases, polymerase α and/or polymerase δ, leading to G/T mismatches (Fig. 3c). Similarly, the observed bias for C>T transitions on the leading strand is likely to be predominantly caused by misincorporation of T on the lagging strand by polymerase α and/or polymerase δ, resulting in T/G mismatches (Fig. 3c).

Second, the C>A predominance could be explained by differential processing of 8-oxo-dGs (Fig. 3c). The principal C>A/G>T peak in MMR-deficient cells occurs at C>T/CAT/AGG>ATG followed by CCC>CAC/GGG>GTG and is distinct from the C>A/G>T peaks observed in ∆OGGI. However, we previously showed a depletion of mutations at CC/GG sequence motifs for ∆OGGI. Intriguingly, the experimental data suggest that the 8-oxo-G/A mismatches can be repaired by MMR, preventing C>A/G>T mutations45. Furthermore, G>T/C>A mutations of MMR-deficient cells occurred most frequently at the second G in 5’T−Gn (n ≥ 3) in ∆MLH1, ∆MSH2 and ∆MSH6 (Fig. 3e and Extended Data Fig. 8). This is consistent with previous reports65 of the classical imprint of guanine oxidation at polyG tracts where the site reactivity in the double-stranded 5’T−GnG’G’G’T sequence is reported as G’2>G’1>G1>G0. These results implicate MMR involvement in repairing 8-oxo-G:A mismatches at G G motifs that perhaps cannot be cleared by OGG1 in BER. As for G>T leading strand bias, studies in yeast have demonstrated that an excess of 8-oxo-dG-associated mutations occurs during leading strand synthesis66. Furthermore, translesion synthesis polymerase η is also more error prone when bypassing 8-oxo-dG on the leading strand67, which would result in increased 8-oxo-G:A mispairs on the leading strand.

Third, T>A transversions at ATT were strikingly persistent in MMR knockout signatures, although with modest peak sizes (<3% normalized signature; Fig. 3a). Additional sequence context information revealed that T>A occurred most frequently at AAATT or TTAAA—junctions of poly(A) and poly(T) tracts (Fig. 3f)68,69. Moreover, the length of 5’- and 3’-flanking homopolymers influenced the likelihood of mutation occurrence: T>A transversions were one to two orders of magnitude more likely to occur when flanked by homopolymers of 5’poly(A)/3’poly(T) (A1,T1) or 5’poly(T)/3’poly(A) (T1,A1) than when there were no flanking homopolymeric tracts (Fig. 3g).

Since polynucleotide repeat tracts predispose to indels due to replication slippage (a known source of mutagenesis in MMR-deficient cells), we hypothesize that T>A transversions observed at abutting poly(A)/poly(T) tracts are the result of reverse template slippage. In this scenario, the polymerase replicating across a mixed repeat sequence such as AAAAAATTTTT, in which the template slipped at one of the As, would incorporate five instead of six Ts opposite the A repeat (red arrow pathway in Fig. 3h). If at this point the template were to revert to its original correct alignment, A/A mismatch would occur, resulting in a T>A transversion. If the slippage remained, this would give rise to a single-nucleotide deletion, a characteristic feature of MMR-deficient cells known as microsatellite instability (MSI) (indel signatures in Fig. 3b).

Gene-specific mutational signatures in MMR deficiency. There are uncertainties regarding which cancer-derived signatures are truly MMR-deficiency (MMRd) signatures. It was suggested that SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and SBS44 are MMRd related49. In an independent analytical exercise, only two MMR-associated signatures were identified41, with variations seen in different tissue types41. An experimental process would help to obtain clarity in this regard41–44.

As described earlier, the substitution and indel patterns of ∆MSH2, ∆MSH6 and ∆MLH1 showed qualitative similarities and were distinct from those of ∆PMS2 (Fig. 3a,b and Extended Data Fig. 9a,b). While the qualitative indel profiles of ∆MSH2, ∆MSH6 and ∆MLH1 were very similar (Fig. 3b), their quantitative burdens were different (Fig. 1e). ∆MLH1 and ∆MSH2 had high indel burdens, whereas ∆MSH6 had half the indel burden (Fig. 1e). Substitution-to-indel ratios showed that ∆MSH2, ∆PMS2 and ∆MLH1 produced similar numbers of substitutions and indels, while ∆MSH6 generated nearly 2.5 times more substitutions than indels (Extended Data Fig. 9c,d). This is in keeping with known protein interactions: MSH2 and MSH6 form the heterodimer MutSα that addresses base–base mismatches and small indels (one or two nucleotides)9,49,70. MSH2 can also heterodimerize with MSH3 to form heterodimer MutSβ, which does not recognize base–base mismatches, but can address indels of 1–15 nucleotides71. This functional redundancy in small indel repair between MSH6 and MSH3 explains the smaller number of indels observed in ∆MSH6 (Fig. 1e and Extended Data Fig. 9d) compared with ∆MSH2 cells. This is consistent with the near-identical MSI phenotypes of Msh2−/− and Msh3−/−Msh6−/− mice42.

Thus, there are clear qualitative differences between the substitution and indel profiles of ∆MSH2, ∆MSH6 and ∆MLH1 versus
ΔPMS2. To validate these MMR gene-specific knockout signatures, we interrogated genomic profiles of normal cells derived from patients with inherited autosomal recessive defects in MMR genes resulting in constitutional MMRd (CMMRD)—a severe, hereditary cancer predisposition syndrome characterized by an increased risk of early-onset (often pediatric) malignancies and cutaneous café au lait macules. hiPSCs were generated from erythroblasts derived from four patients with CMMRD (one PMS2 homozygote, one PMS2

**Fig. 3 | Multiple endogenous sources of DNA damage managed by MMR.** a, b, Substitution (a) and indel signatures (b) for five MMR gene knockouts. The indel signature of ΔPMS1 is shown in Extended Data Fig. 5a. In a, bold letters represent mutated bases. Rep, repetitive tracts; Mh, microhomology-mediated; other, indel without homology or repetitive tracts in flanking sequence. c, Dissection of DNA MMRd mutational signatures: C > A mutations believed to be due to oxidative damage of guanine and the proposed mechanism of how DNA polymerase errors contribute to misincorporated bases that result in C > T and T > C substitutions. All other mismatch possibilities and their outcomes are shown in Extended Data Fig. 7. The red and black strands represent lagging and leading strands, respectively. The strands with arrowheads are nascent strands. d, Replicative strand asymmetry (lagging:leading) observed for mutational signatures generated by four MMR gene knockouts. The dots and error bars represent odds ratios with 95% confidence intervals. e, Relative frequency of occurrence of G > T/C > A in polyG tracts for ΔMSH2. The counts and relative frequencies of occurrence of G > T/C > A in polyG tracts for ΔMSH2 and ΔMLH1 are shown in Extended Data Fig. 8. f, The T > A mutation frequency is highest at junctions of poly(A)poly(T) or poly(T)poly(A). The inset shows that T > A mutations have a striking peak at AT. g, The odds of T > A mutations occurring at poly(A)poly(T) or poly(T)poly(A) are higher than AT sequences flanked by other nucleotides, corrected for sequence context through the whole genome. The data are presented as means ± s.e.m. (n = 2–4; see Supplementary Table 2). h, Putative reverse template slippage model: T > A substitutions at poly(A)poly(T) or poly(T)poly(A) junctions arise due to template strand slippage and subsequent reversal of the slipped template strand. IDL, insertion-deletion loop. The red arrow shows DNA strand slippage synthesis. The red letter A represents the base involved in IDL.
compound heterozygote and two MSH6 homoygotes) and one healthy control. The obtained hiPSC clones were genotyped and expression arrays and cellomics-based immunohistochemistry (IHC) were performed to ensure PSCs were generated (Methods). Parental clones were grown for mutation accumulation, and single-cell subclones were derived and sequenced by WGS (Fig. 4a).

The mutational signatures seen in CMMRD hiPSCs were virtually identical to those of the CRISPR–Cas9 knockouts (Fig. 4b–d). The ΔMSH2 CMMRD patterns carried the same propensity for T > C mutations as ΔAPMS2 and a small contribution of C > A/ΔT > T mutations at the CCT/AGG peak. The ΔMSH6 CMMRD patterns carried an excess of C > T mutations, with pronounced C > A/ΔG > T at CCT/AGG, similar to ΔMLH1, ΔMSH2 and ΔMSH6 clones (Fig. 4c). Indel propensities were also reflected in patient-derived cells (Fig. 4d). Accordingly, the gene specificity of signatures generated in the experimental knockout system was well recapitulated in independent patient-derived stem cell models.

Furthermore, gene-specific MMRd signatures were seen in the International Cancer Genome Consortium cohort of >2,500 primary WGS-sequenced cancers. Biallelic MSH2/MSH6/MLH1 mutant tumors carried the same signature (RefSig MMR1) as ΔMSH2/ΔMSH6/ΔMLH1 clones (Fig. 4e). We also identified biallelic PMS2 mutants in several cancers, including breast and ovarian cancers with mutation patterns (RefSig MMR2) that were indistinguishable from experimentally generated ΔAPMS2 signatures (Fig. 4e).

A mutational signature-based classifier of MMRd. Algorithms to classify MMR-deficient tumors, developed using massively parallel sequencing data, depend on detecting elevated tumor mutational burden or MSI. New knowledge from our experimental data and awareness of tissue-specific signature variation (Fig. 4e) led us to derive an MMRd classifier.

We obtained WGS data on 336 colorectal cancers from patients recruited via the National Health Service-based UK 100,000 Genomes Project (UK100kGP) run by Genomics England. Critically, these samples had accompanying IHC testing of the MMRd status based on protein staining of MSH2, MSH6, MLH1 and PMS2, and 79 out of 336 cases (~24%) were identified as MMR deficient. This cohort of 336 samples was randomly assigned to a training set (180 MMR-proficient samples and 23 MMR-deficient samples) or a test set (77 MMR-proficient samples and 23 MMR-deficient samples). We developed a logistic regression classifier called MMRDetect using new mutational signature-based parameters derived from the experimental insights gained from this study: (1) the exposure of MMR-deficient substitution signatures ($E_{\text{subsig}}$); (2) the cosine similarity between the substitution profile of the tumor and that of MMR knockouts; (3) the mutation burden of indels in repetitive regions ($N_{\text{rep.indel}}$); and (4) the cosine similarity between the repeat-mediated deletion profile of the tumor and that of MMR knockouts (further details are provided in the Methods, Extended Data Fig. 10 and Supplementary Tables 5 and 6). Tenfold cross-validation was conducted in the training set (Extended Data Fig. 10f). As a comparator, we applied another widely used MSI classifier, MSIsq, to the same cohort of 336 colorectal cancers.

Samples with an MMRDetect-calculated probability of <0.7 are defined as MMR deficient by MMRDetect (Extended Data Fig. 10g). In total, 75 of 336 samples were concordantly determined as MMR deficient by MMRDetect, MSIsq and IHC (Fig. 5a and Supplementary Table 5). Eight samples had discordant statuses, including four samples with MMRd deficiency only by IHC, two samples with MMRd by MSIsq and MMRDetect and two samples uniquely called by MSIsq. To understand these discordances, we sought driver mutations. Among these eight samples, two samples missed by IHC had confirmed loss-of-function mutations in MMR genes. Additionally, two cases uniquely called by MSIsq were misclassified and were, in fact, POLE mutated and not MMR deficient (Fig. 5a and Supplementary Table 5). While receiver operating characteristic curves generated by these three methods showed excellent performance across the board, MMRDetect had the highest area under the curve of 1 in this dataset (Fig. 5b).

Next, we compared MMRDetect and MSIsq on another 2,012 colorectal and 713 uterine samples from UK100kGP, 2,610 published primary cancers sequenced by WGS and 2,024 metastatic cancers sequenced by WGS (Supplementary Tables 7–10 and Methods). There was a very high concordance between MMRDetect and MSIsq for classifying tumors (0.970–0.997; Fig. 5c). To understand the discrepancies between the two algorithms, we compared variables that were used by the two classifiers (Fig. 5d) and found that samples uniquely identified as MMR deficient by MSIsq had significantly higher $N_{\text{rep.indel}}$ values and non-MMRd signatures than those identified as MMRd by MMRDetect only ($P < 0.001$; two-sided Mann–Whitney $U$-test; Extended Data Fig. 10h). This indicated a higher likelihood of misclassifying samples with high indel loads caused by non-MMRd mutational processes (that is, false positives) for MSIsq, a known generic problem reported for NGS indel-based classifiers. Many of these samples showed signatures associated with proofreading POLE mutations. This demonstrates that MMRDetect has improved specificity over MSIsq.

Of note, samples identified as MMR deficient only by MMRDetect had significantly lower $N_{\text{rep.indel}}$, $E_{\text{subsig}}$, $E_{\text{deletion}}$ values than samples concordantly identified as MMR deficient by both MSIsq and MMRDetect ($P < 0.001$; two-sided Mann–Whitney $U$-test; Extended Data Fig. 10h), suggesting that MMRDetect has improved sensitivity for MMR-deficient cancers with lower overall MMR-related mutation counts ($E_{\text{deletion}}$). Indeed, of 15 bona fide MMR-deficient breast cancers (a tumor type that is not as proliferative as colon/uterine cancer and has lower mutation numbers in general), MMRDetect identified 13 cases (87%), while MSIsq identified five cases (~33%; the remaining ten samples had lower repeat-mediated indel loads (2.885–18.863) (Supplementary Table 10). The two cases missed by MMRDetect had very low levels of MMRd signatures and were complicated by high levels of APOBEC-related mutagenesis. Thus, MMRDetect has enhanced sensitivity, particularly at detecting MMR-deficient samples with lower mutation burdens (Fig. 5d), although it could miss cases
Articles Nature Cancer

Blood sample from patient

Extraction of erythroblast cells

iPSC reprogramming

iPSCs

Derivation of subclones

Subclone expansion

WGS

PMS2 mutant

MSH6 mutant

C > A C > G C > T T > A T > C T > G

CMMRD3 (PMS2 mutant)

CMMRD77 (PMS2 mutant)

CMMRD89 (MSH6 mutant)

CMMRD94 (MSH6 mutant)

Indel type

\[\text{[−]}Mh \hspace{1cm} \text{[−>1]}\text{Other} \hspace{1cm} \text{[−>1]}\text{Rep} \hspace{1cm} \text{[−C]}\text{Rep} \hspace{1cm} \text{[−T]}\text{Rep} \hspace{1cm} \text{[+]}Mh \hspace{1cm} \text{[+>1]}\text{Other} \hspace{1cm} \text{[+>1]}\text{Rep} \hspace{1cm} \text{[+C]}\text{Rep} \hspace{1cm} \text{[+T]}\text{Rep} \hspace{1cm} \text{Complex} \]

Substitution type

\[321 \hspace{1cm} 9876540321 \hspace{1cm} 987654321 \hspace{1cm} 987654\]

Cancer: PD11356a

Cancer: PD23564a

Subclone expansion

WGS

Derivation of subclones

Blood sample from patient

Extraction of erythroblast cells iPSC reprogramming iPSCs

Subclone expansion

WGS

Derivation of subclones
where MMRd is present at a very low level. We note that the current version of MMRDetect classifier has been trained on highly proliferative colorectal cancers. More sequencing data are required to improve MMRDetect’s detection sensitivity in other tumor types.

**Discussion**

In standardized experiments performed in a diploid, nontransformed human stem cell model, biallelic gene knockouts that produce mutational signatures in the presence of administered DNA damage are indicative of genes that are important for maintaining the genome and protecting it from intrinsic DNA damage sources (Fig. 6). We found substitution, double-substitution and/or indel signatures of nine genes (ΔOGG1, ΔUNG, ΔEXO1, ΔRNFL68, ΔMLH1, ΔMSH2, ΔMSH6, ΔPMS2 and ΔPMS1), suggesting that these proteins are critical guardians of the genome in nontransformed cells. Many gene knockouts did not show mutational signatures under these conditions. This does not mean that they do not transcribe important DNA repair proteins. There may be redundancy, or the gene may be crucial to the orchestration of DNA repair, even if it is not imperative in directly preventing mutagenesis. It is also possible that some knockouts have very low rates of mutagenesis such that statistically distinct signatures cannot be distinguished from background mutagenesis within our experimental time frame. For genes involved in DSB repair, hiPSCs may be a prerequisite before these compromised genes reveal associating with MMRd signatures; and the number of non-MMRd mutations. The numbers of MMR-deficient samples, as determined by MMRDetect only (blue), MSIseq only (pink), both (yellow) and neither (purple) were 34, 20, 587 and 6,718, respectively.

**Fig. 5** | Mutational signature-based MMRd classifier, MMRDetect. **a**. Venn diagram illustrating the concordance of three MMRd detection methods—IHC staining, MSIseq and MMRDetect—on 336 colorectal cancers. IHC staining, MSIseq and MMRDetect identified 79, 79 and 77 MMR-deficient samples, respectively. Details of the eight samples with discordant outcomes from the three methods are provided in Supplementary Table 5. Four samples classified as MMR proficient by MMRDetect and MSIseq had abnormal IHC staining (shown in dark yellow). However, no functional mutations in MMR genes were found. Two samples classified as MMR proficient by MMRDetect and IHC staining were identified as MMR deficient by MSIseq (shown in pink) and did not have MMR gene mutations but had POLE mutations and signatures instead. Two samples classified as MMR deficient by MMRDetect and MSIseq had normal IHC staining (shown in orange). Both had mutations in MMR genes. **b**. Receiver operating characteristic curves of IHC staining, MMRDetect and MSIseq classification. AUC, area under the curve. **c**. Concordance between MSIseq and MMRDetect on 2,012 Genomics England (GEL) colorectal cancers, 713 GEL uterine cancers, 2,024 Hartwig metastatic cancers and 2,610 cancers from the Pan-Cancer Analysis of Whole Genomes (PCAWG) and Sweden Cancerome Analysis Network—Breast (SCANB) projects. The bars show the numbers of samples that were identified as MMR deficient by MSIseq only (pink), MMRDetect only (blue), both (yellow) and neither (purple). **d**. Distribution of three variables among samples that were discordantly (blue and pink) and concordantly (yellow and purple) detected by MSIseq and MMRDetect: the number of repeat-mediated indels; the number of mutations associated with MMRd signatures; and the number of non-MMRd mutations. The numbers of MMR-deficient samples, as determined by MMRDetect only (blue), MSIseq only (pink), both (yellow) and neither (purple) were 34, 20, 587 and 6,718, respectively.
of DNA-damaging agents or using alternative cellular models (for example, cancer lines or permissive cellular models of specific tissue types), could amplify signals, but they could also modify mutational outcomes. This must be taken into consideration when interpreting data. Also, not all genes have been successfully knocked out in this endeavor and could have similarly important roles in directly preventing mutagenesis.

Detailed dissection of experimental signatures revealed interesting mutational insights, including how OGG1 and MMR sanitize oxidized guanines at specific sequence motifs. In contrast, UNG maintains all cytosines from hydrolytic deamination, irrespective of sequence context. Exhaustive assessment of DNA mismatches and their putative outcomes also uncovered precise polymerase errors that are repaired by MMR, including misincorporation of G (resulting in T>C transitions) and misincorporation of T (resulting in G>A/C>T transitions by lagging strand polymerases). We also observed a T>A substitution pattern at abutting poly(A) and poly(T) tracts and postulate a mechanism called reverse template slippage.

While it is known that 8-oxo-dGs can result in G>T mutations, our work demonstrates that the etiology of the culture-related and cancer-derived signature 18 is mainly 8-oxo-dG. We highlight the importance of functional EXO1 and RNFL18 in preventing signature 5, a relatively ubiquitous signature characterized by T>C/A>G transitions. We define gene specificities of MMRd signatures, prove that these are robust in normal stem cells derived from patients with CMMRD and identify gene-specific signatures in human cancers.

Finally, unlike signatures of environmental mutagens that are historical, signatures of repair pathway defects are likely to be ongoing. They could serve as biomarkers in precision medicine13,14,18 (Fig. 6) to identify pathway defects where selective therapeutic strategies are available. Our experiments led to the development of a more sensitive and specific assay to detect MMRd, MMRDetect. Current tumor mutational burden-based assays may have reduced sensitivity to detect MMRd in tissues that do not have high proliferative rates. They may also falsely call MMR-proficient cases as MMR deficient because single components were used for measurement (for example, indel or substitution burdens only). High mutational burdens can be due to different biological processes77. Consequently, assays based on burden alone are unlikely to be adequately specific. As a community, we are at the early stages of seeking experimental validation of mutational signatures. However, we hope that our approach, which leans on experimental data, provides a template for improving biological understanding of how mutational patterns arise, and this, in turn, could help propose improved tools for tumor characterization going forward.

Methods

Cell lines and culture. The hiPSC line used in this study was described previously14. The line was derived at the Wellcome Sanger Institute (Hinxton, United Kingdom). The use of this cell line model was approved by the Proportionate Review subcommittee of the National Research Ethics Service (North West–Liverpool Central) under the project 'Exploring the biological processes underlying mutational signatures identified in induced pluripotent stem cell-lines (iPScs) that have been genetically modified or exposed to mutagens' (reference 14.NW.0129). This is a long-standing iPSc line that is diploid and does not have any known driver mutations. It carries a balanced translocation between chromosomes 6 and 8, grows stably in culture and does not acquire a vast number of karyotypic abnormalities. This was confirmed through mutational and copy number assessment of the WGS data reviewed for all subclones.

Cell culture reagents were obtained from STEMCELL Technologies unless otherwise indicated. Cells were routinely maintained on Vitronectin XF-coated
articles

Generation of donor plasmids for precise gene targeting via homology-directed repair.

Sample collection from patients with CMMRD.

Generation of DNA repair gene knockouts in human iPSCs.

Proteomics analysis.

Delivery of CRISPR-editing plasmids, knockout clone selection and genotyping.

Proliferation assay.

Generation of iPSCs from patients with CMMRD.

Guide RNA (gRNA) design and cloning.

Delivery of CRISPR editing plasmids.
FS microplate reader. Luminescence readings were normalized and presented as luminescence units relative to \( \epsilon \). Supplementary Table 14 shows the statistics of six replicates for each time point per indicated knockout line. The doubling time was calculated based on replicate-averaged readings on the linear portion of the proliferation curve (exponential phase) using the following formula:

\[
24h \times \log(2) \quad \log(\text{final measurement}) - \log(\text{initial measurement})
\]

**Genomic DNA extraction and WGS.** Samples were quantified with a Biotium AccuClear Ultra High Sensitivity dsDNA Quantitative kit (using a Mosquito LV liquid platform, the Bravo WS automation system and a BMG FLUOstar Omega plate reader) and sheared to 500 bp using a Covaris liquid handling platform. Cherrypicked plates were sheared to 450 bp using a Covaris LE220 instrument. Post-chromatography were purified using Agencourt AMPure XP SPRI beads on the Agilent Bravo WS. Libraries were constructed (end repair, A-tailing and ligation) using the Agilent SureSelect kit on the Agilent Bravo WS automation. Kapata FF HotStarMix and Integrated DNA Technologies 96 iPCR tag barcodes were used for PCR set-up on the Agilent Bravo WS automation system. PCR cycles included six standard cycles: (1) incubate at 95 °C for 5 min; (2) incubate at 98 °C for 30 s; (3) incubate at 65 °C for 30 s (4) incubate at 72 °C for 1 min; (5) cycle from (2) five more times; and (6) incubate at 72 °C for 10 min. The post-PCR plate was purified using Agencourt AMPure XP SPRI beads on a Beckman Biomek NX-96 liquid handling platform. Libraries were quantified with a Biotium AccuClear Ultra High Sensitivity dsDNA Quantitation kit (using the Mosquito LV liquid-handling platform, Bravo WS automation system and BMG FLUOstar Omega plate reader), then pooled in equimolar amounts on a Beckman Biomek NX-8 liquid-handling platform and finally normalized to 2.8 nM ready for cluster generation on a cBot and loading on a requested Illumina sequencing Biomek NX-8 liquid-handling platform. Libraries were quantified with a Biotium AccuClear Ultra High Sensitivity dsDNA Quantitation kit (using the Mosquito LV liquid-handling platform, Bravo WS automation system and BMG FLUOstar Omega plate reader), then pooled in equimolar amounts on a Beckman Biomek NX-8 liquid-handling platform and finally normalized to 2.8 nM ready for cluster generation on a cBot and loading on a requested Illumina sequencing platform. Pooled samples were loaded on the Illumina HiSeq X-Ten platform using 150 bp paired-end sequencing reads, sequenced to ~25X coverage. Details of the sequence coverage for all of the clones and subclones are provided in Supplementary Table 2.

**Alignment and somatic variant calling.** Short reads were aligned to the human reference genome GRCh37/final assembly using the BWA-MEM algorithm\(^\text{11}\). Three algorithms, CAVEMan (http://cancerit.github.io/CAVEMan/)\(^\text{13}\), Pindel (http://cancerit.github.io/eggpindel/)\(^\text{14}\) and BRASS (https://github.com/cancerit/BRASS), were used to call somatic substitutions, indels and rearrangements in all subclones, respectively.

**Assurance of the knockout state using WGS data.** First, we examined whether there were CRISPR-Cas9 off-target effects by seeking relevant mutations in other DNA repair genes besides the genes of interest. We also searched for potential off-target sites based on gRNA target sequences using COSMID\(^\text{15}\) and confirmed that there were no off-target hits in knockouts that generated mutation signatures. According to the mutational profile of control subclones, the only confident off-target site was the target sequence for gRNA #1 in the control subclones.

**Determination of gene knockout-associated mutational signatures.** An intrinsic background mutagenesis exists in normal cells grown in culture. Knocking out a DNA repair gene that is involved in repairing endogenous DNA damage may result in increased unrepaired DNA damage and thereby result in mutation accumulation with subsequent rounds of replication. WGS of these knockouts can detect the mutations that occur as a result of being a specified knockout. If the mutation burden and mutational profile of a knockout are significantly different from the control subclones that harbor only the background mutagenesis, it is likely possible that there is gene knockout-associated mutagenesis. Based on this principle, our approach to identify gene knockout-associated mutational signatures involved three steps: (1) determine the background mutational signature; (2) determine the difference between the mutational profile of the knockout and background mutation profiles; and (3) remove the background mutation profile from the mutation profile of the knockout subclone.

Substitution profiles were described according to the classical convention of 96 channels: the product of six types of substitution multiplied by four types of 5’ base (A,C,G,T) and four types of 3’ base (A,C,G,T). Indel profiles were described by type (insertion, deletion or complex), size (1 bp or longer) and flanking sequence (repeat mediated, microhomology mediated or other) of the indel. Here, we used two sets of indel channels. Set one contained 15 channels: a 1-bp C/T insertion at a short repetitive sequence (<5 bp); a 1-bp C/T insertion at a long repetitive sequence (<2 bp); long insertions (>1 bp) at repetitive sequences; microhomology-mediated insertions; a 1-bp C/T deletion at a short repetitive sequence (<5 bp); a 1-bp C/T deletion at a long repetitive sequence (<2 bp); long deletions (>1 bp) at repetitive sequences; microhomology-mediated deletions; and other deletion and complex indels (Extended Data Fig. 5a). Set two contained 45 channels in which the 1-bp C/T indels at repetitive sequences were further expanded according to the exact length of the repetitive sequences (Fig. 3b). Indel channel set one was applied to all knockout subclones, while channel set two was only applied to four MMR gene knockouts (ΔMLH1, ΔAPMS2, ΔMSH2 and ΔMSH6) to obtain a higher resolution of mutational signatures of MMR gene knockouts.

**Identifying background signatures.** The mutational profiles of control subclones were used to determine background mutagenesis. Aggregated substitution profiles of all control subclones (ΔATP2B4) were used as the background substitution mutational signature. Aggregated indel profiles of all subclones containing ≤8 indels were used as the background indel mutational signature.

**Distinguishing mutational profiles of control and gene-edited subclone profiles.** The signal-to-noise ratio affects mutational signature detection. In this study, noise was largely background mutagenesis. The average mutation burden caused by the background mutagenesis in control cells for substitutions and indels were around 150 and 10, respectively, with standard deviations of 10 and 1.4. Signal represents the elevated mutation burden caused by gene knockouts. The average mutation burden in knockouts ranged from 63–2,360 for substitutions and 0–1,212 for indels after 15 d in culture, as shown in Supplementary Table 2.

The costs associated with WGS are prohibitive; thus, we used two to four subclones per knockout. The intrinsic fluctuation of the detected mutation burden in each sample and the limited subclone numbers imposed a greater uncertainty in mutational signature detection. Thus, to distinguish high-confidence mutational signatures from noise, we employed three different methods.

First, we evaluated the similarity of mutational profiles between the control and each gene knockout. According to the mutational profile of control subclones, \( p_{\text{control}} = \left[ p_{\text{control}}^1, p_{\text{control}}^2, \ldots, p_{\text{control}}^l \right] \) for a given number of mutations \( N(0 < N < 10,000) \), one could generate \( L \) bootstrapped samples:

\[
M_{\text{boot}} = \left[ m_{\text{boot},1}, m_{\text{boot},2}, \ldots, m_{\text{boot},L} \right] = \left[ \left[ m_{1,1}, m_{1,2}, \ldots, m_{1,L} \right], \ldots, \left[ m_{L,1}, m_{L,2}, \ldots, m_{L,L} \right] \right]
\]

where \( \sum_{k=1}^{K} m_{k} = N \). One can calculate the cosine similarities \( (s) \) between bootstrapped control samples \( (m) \) and the experimentally obtained control profile \( (p_{\text{control}}) \) to obtain a distribution of cosine similarities, \( P(S) \):

\[
P(S) = \frac{m_{\text{control}} \cdot p_{\text{control}}}{\| m_{\text{control}} \| \| p_{\text{control}} \|}
\]

We can then calculate the cosine similarity \( S_{\text{knockout}} \) between the control profile \( (p_{\text{control}}) \) and knockout profile \( (p_{\text{knockout}}) \). As shown in Fig. 1c,d, when the mutation count is low, the bootstrapped samples are less similar to the actual control profile than the bootstrapped samples with a higher mutation count. Comparing \( S_{\text{knockout}} \) and \( P(S) \) at a given mutation number, \( N_{\text{threshold}} \), one can identify which gene knockout has distinct mutational profiles compared with the control \( (P(1) = 0.01) \) but different from the background.

Second, we used contrastive principal component analysis\(^\text{16}\), which efficiently identified directions that were enriched in the knockouts relative to the background through eliminating confounding variations present in both (Extended Data Fig. 3a), to recognize gene knockout-specific patterns from the background signature.

Third, we used t-SNE\(^\text{17}\), which is a visualization technique for viewing pairwise similarity data resulting from nonlinear dimensionality reduction based on probability distributions. In t-SNE implementation, mutational profiles that are similar to each other are plotted near to each other, whereas profiles that are dissimilar are plotted distant in two-dimensional space (Extended Data Fig. 3b). Subtraction of the background mutational signature from the knockout mutational profile. The experiment-associated mutational signature can then be obtained by subtracting the background mutational signature from the mutational profile of treated subclones through quantile analysis. First, one can generate a set of
bootstrap samples of each treated subclone, to determine the distribution of mutation numbers for each channel. According to the distribution, the upper and lower boundaries (for example, the 99% confidence interval) for each channel can be identified. Then, based on the background mutational signature and averaged mutation burden (as an initial value), one can construct a bootstrapped background profile and subtract it from the centroid of bootstrap subclone samples. Due to data noise, some channels may have negative values, in which case, the negative values are set to zero. Occasionally, the number of mutations in a few channels will fall outside the lower boundary after removing the background profile. To avoid negative values, the background mutation pattern is maintained but the burden is scaled down through an automated iterative process.

**Topography analysis of signatures. Strand bias.** Reference information of replicative strands and replication timing regions was obtained from Repli-seq data from the ENCODE project (https://www.encodeproject.org/). The transcriptional strand coordinates were inferred from the known footprints and the transcriptional direction of protein-coding genes. First, for a given mutational signature, we calculated the expected ratio of mutations between transcribed and nontranscribed strands, or between lagging and leading strands, according to the distribution of the trinucleotide sequences. Second, the observed ratio of mutations between different strands was identified through mapping mutations to the genomic coordinates of all gene footprints (for transcription) or leading/lagging regions (for replication). Third, all mutations were orientated towards pyrimidines as the mutated base (as this has become the convention in the field), for the direction in which strand the mutation was on. Fourth, the level of asymmetry between different strands was measured by calculating the odds ratio of mutations occurring on one strand (for example, the transcribed or leading strand) versus the other strand (for example, the nontranscribed or lagging strand). IntersectBed^24 was used to identify mutations overlapping certain genomic features.

**MMRDetect algorithm.** We trained an MMR deficiency logistic regression-based classifier, called MMRDetect, based on mutational signatures obtained from the experimental work. We obtained mutation data from 336 WGS-sequenced colorectal cancers with accompanying IHC staining of the four MMR proteins from the experimental work. We obtained mutation data from 336 WGS-sequenced colorectal cancers with accompanying IHC staining of the four MMR proteins (MSH2, MSH6, MLH1 and PMS2) from UK100kGP. Within this cohort of 336 colorectal cancers, there were 79 (24%) cancers with abnormal IHC staining indicative of MMR deficiency. The 336 cancers were randomly divided into a training set and a test set using the R function sample(). The training set had 180 colorectal cancers with accompanying IHC staining of the four MMR proteins and 56 MMR-deficient samples. The test dataset had 77 training set and a test set. A tenfold cross-validation was performed on the training data weights (coefficients) of the four variables obtained from training the model using the regression algorithm (glm() function) provided in the R package glmnet was calculated using the two-sided Mann–Whitney test (wilcox.test() in R). The values of different variables were transformed to between 0 and 1 using the formula \( \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)} \).

**Data analysis.** All statistical analyses were performed in R^24. P-values were calculated using the two-sided Mann–Whitney test (wilcox.test() in R). The enrichment of mutational-specific trinucleotide sequences was assessed by calculating the odds ratio between the observed and expected ratios. The 95% confidence interval was calculated to estimate the precision of the odds ratio. All plots were generated using ggplot2 (ref. 44).

**Statistics and reproducibility.** No statistical method was used to predetermine the sample size. We produced two genotypes for each gene and two subclones for each genotype. Four subclones were obtained for most of the genes, except for EXO1 and LMSH2. Off-target gene knockouts and/or mislabeled samples could cause erroneous results by reporting a signature incorrectly; thus, to reduce the likelihood of errors, we excluded these samples. The experiments were not randomized and the investigators were not blinded to allocation during the experiments and outcome assessment. However, this was a systematic experimental study performed with identical conditions across all knockouts; thus, all sequencing data generated afterwards were agnostic and fully comparable with one another.

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

**Data availability**

Raw sequence files are deposited at the European Genome-Phenome Archive with accession numbers EGAS00001000800 and EGAS00001000874. Mutation calls have been deposited at Mendeleev: https://doi.org/10.17632/ynn3ykKNx5. hiPSCs can be obtained directly from the authors. The curated data are available for general browsing from our reference mutational signatures website, Signal (https://signal.mutationsignatures.org). Age information relating to human patient samples is not publicly available as this could compromise privacy and lead to identification of the individuals. Publicly available genomic datasets realanalyzed here to compare the performance of MMRDetect and MSIsoq are available from the European Genome-Phenome Archive (EGAS00001001178), http://dcc.icgc.org/pcawg/ (ref. 21), https://data.mendeley.com/datasets/zmvd4dpyp1/1 (ref. 24), https://resources.hartwigmedicalfoundation.nl/ (ref. 61) and the Genomics England Research Environment (main program version 8) via https://re.extent.co.uk/ovd/. Source data are provided with this paper. All other data supporting the findings of this study are available from the corresponding author upon reasonable request.

**Code availability**

The R code used to generate results presented in Figs. 1–5 and the R source code of MMRDetect can be obtained from https://github.com/Nik-Zainal-Group/COMSIG_KO.git.

Received: 4 August 2020; Accepted: 17 March 2021; Published online: 26 April 2021

**References**

1. Helliday, T., Eshdat, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* 15, 585–598 (2014).
2. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* 500, 415–421 (2013).
3. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* 149, 994–1007 (2012).
4. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* 149, 979–993 (2012).
5. Haradhvila, N. J. et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* 9, 1746 (2018).
6. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature 578*, 94–101 (2020).
7. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in uterine cancers. *Nat. Genet.* 48, 600–606 (2016).
8. Nik-Zainal, S. et al. The genome as a record of environmental exposure. *Mutagenesis* 30, 763–770 (2015).
9. Zou, X. et al. Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* 9, 1744 (2018).
10. Christensen, S. et al. 5-Fluorouracil treatment induces characteristic T·G mutations in human cancer. *Nat. Commun.* 10, 4571 (2019).
11. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* 177, 821–836.e16 (2019).
12. Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* 11, 3610–3618 (1972).
13. Mardis, E. R. The impact of next-generation sequencing on cancer genomics: from discovery to clinic. *Cold Spring Harb. Perspect. Med.* 9, a036269 (2019).
14. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* 15, 353–365 (2018).
15. David, S. S., O’Shea, V. L. & Kundu, S. Base-excision repair of oxidative DNA damage. *Nature* 447, 941–950 (2007).
16. Kunkel, T. A. & Erie, D. A. DNA mismatch repair. *Annu. Rev. Biochem.* 74, 681–710 (2005).
17. Kottemann, K. C. & Smogorzewska, A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. *Nature 493*, 356–363 (2013).
18. Wood, R. D., Mitchell, M., Sguerri, J. & Lindahl, T. Human DNA repair genes. *Science* 291, 1284–1289 (2001).
41. Banerjee, A. & Verdine, G. L. A nucleobase lesion remodels the interaction of DNA via electron transfer: significant lowering of ionization potential and 5′-localization of HOMO of stacked GG bases in B-form DNA. J. Am. Chem. Soc. 118, 7063–7068 (1996).
42. Allgayer, J., Kitsera, N., von der Lippen, C., Epe, B. & Khobta, A. Modulation of base excision repair of 8-oxoguanine by the nucleotide sequence. Nucleic Acids Res. 41, 8559–8571 (2013).
43. Friedman, J. I. & Stivers, J. T. Detection of damaged DNA bases by DNA glycosylase enzymes. Biochemistry 49, 4957–4967 (2010).
44. Lutsenko, E. & Bhagwat, A. S. Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. A model, its experimental support and implications. Mutat. Res. 437, 11–20 (1999).
45. Shen, J. C., Rideout, W. M. 3rd & Jones, P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. 22, 972–976 (1994).
46. Waters, T. R. & San, P. F. Thymine-DNA glycosylase and G to A transition mutations at CpG sites. Mutat. Res. 462, 137–147 (2000).
47. Santos, M. A. et al. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. Blood 132, 1526–1534 (2018).
48. Barnes, D. E. & Lindahl, T. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. Annu. Rev. Genet. 38, 445–476 (2004).
49. Mol, C. D. et al. Crystal structure and mutational analysis of human uracil-DNA glycosylase: structural basis for specificity and catalysis. Cell 80, 869–878 (1995).
50. Groffen, J. E. et al. mutT homologue deficiency causes a multi-tumor phenotype. Cancer Cell 35, 256–266.e5 (2019).
51. Genschel, J. & Modrich, P. Mechanism of 5′-directed excision in human mismatch repair. Mol. Cell. 12, 1077–1086 (2003).
52. Bolderson, E. et al. Phosphorylation of Exo1 modulates homologous recombination repair of DNA double-strand breaks. Nucleic Acids Res. 38, 1821–1831 (2010).
53. Mattirolli, F. et al. RNF168 ubiquitinates K13–15 on H2A/H2AX to drive DNA damage signaling. Cell 150, 1188–1205 (2012).
54. Bohgaki, M. et al. RNF168 ubiquitinates 53BP1 and controls its response to DNA double-strand breaks. Proc. Natl Acad. Sci. USA 110, 20982 (2013).
55. Doi, C. et al. RNF168 binds and amplifies ubiquitin conjugates on damaged chromatin to allow accumulation of repair proteins. Cell 136, 435–446 (2009).
56. Stewart, G. S. et al. The RIDDLE syndrome protein mediates a ubiquitin-dependent signaling cascade at sites of DNA damage. Cell 136, 420–434 (2009).
57. Gupta, S., Gellett, M. & Yang, W. Mechanism of mismatch recognition revealed by human MutSβ bound to unpaired DNA loops. Nat. Struct. Mol. Biol. 19, 72–78 (2012).
49. Palombo, F. et al. GTPase, a 160-kilodalton protein essential for mismatch-binding activity in human cells. Science 268, 1912–1914 (1995).
50. Warren, J. J. et al. Structure of the human MutSα DNA lesion recognition complex. Mol. Cell. 26, 579–592 (2007).
51. Andreanova, M. A., Bazykin, G. A., Nikolaev, S. I. & Selyapryasky, V. B. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. Genome Res. 27, 1336–1343 (2017).
52. Lawara, S. A. et al. Mismatch repair balances leading and lagging strand DNA replication fidelity. PLoS Genet. 8, e1000316 (2012).
53. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. Nat. Commun. 7, 11383 (2016).
54. Aboul-ela, F., Koh, D., Tinoco, I., Jr. & Martin, F. H. Base-base mismatches. Thermodynamics of double helix formation for dCASXAG + dCT3TYG (N = A, C, G, T). Nucleic Acids Res. 13, 4811–4824 (1985).
55. Mazurek, A., Berardini, M. & Fishel, R. Activation of human MutS homologs by 8-oxoguanine DNA damage. J. Biol. Chem. 277, 8260–8266 (2002).
56. Morikawa, M. et al. Analysis of guanine oxidation products in double-stranded DNA and proposed guanine oxidation pathways in single-stranded, double-stranded or quadruplex DNA. Biomolecules 4, 140–159 (2014).
57. Pavlov, Y. I., Newlon, C. S. & Kunkel, T. A. Yeast origins establish a strand bias for replication mutagenesis. Mol. Cell. 10, 207–213 (2002).
58. Mudrak, S. V., Welz-Voegelle, C. & Jinks-Robertson, S. The polymerase η transcription synthesis DNA polymerase acts independently of the mismatch repair system to limit mutagenesis caused by 7,8-dihydro-8-oxoguanine in yeast. Mol. Cell. Biol. 29, 5316–5326 (2009).
59. Meier, B. et al. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. Genome Res. 28, 666–675 (2018).
60. Jiang, G. L., Parsons, D. N., Demir, A. E., Mutagenesis records, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. G3 (Bethesda) 3, 1453–1465 (2013).
61. Drummond, J. T., Li, G. M., Longley, M. J. & Modrich, P. Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. Science 268, 1909–1912 (1995).
62. Palombo, F. et al. MutSβ, a heterodimer of hMSH2 and hMSH3, binds to insertion/deletion loops in DNA. Curr. Biol. 6, 1181–1184 (1996).
63. Wind, N. et al. HNPCC-like cancer predisposition in mice through simultaneous loss of Msh3 and Msh6 mismatch-repair protein functions. Nat. Genet. 23, 359–362 (1999).
64. Poulogiannis, G., Frayling, I. M. & Arends, M. J. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. Histopathology 56, 167–179 (2010).
65. Heinen, C. D. Mismatch repair defects and Lynch syndrome: the role of the basic scientist in the battle against cancer. DNA Repair 38, 127–134 (2016).
66. Agu, C. A. et al. Successful generation of human induced pluripotent stem cell lines from blood samples held at room temperature for up to 48hr. Stem Cell Rep. 5, 660–671 (2015).
67. Ni Huang, M. et al. MSIsense: software for assessing microsatellite instability from catalogs of somatic mutations. Sci. Rep. 5, 13321 (2015).
68. Niu, B. et al. MSIsensor: microsatellite instability detection using paired end short-read normal-tumor sequence data. Nucleic Acids Res. 41, 1015–1016 (2013).
69. Wang, C. & Liang, C. MSIpred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. Sci. Rep. 8, 17546 (2018).
70. Cortes-Ciriazano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. Nat. Commun. 8, 15180 (2017).
71. Salpante, S. J., Sroggins, S. M., Hample, H. L., Turner, E. H. & Pritchard, C. C. Microsatellite instability detection by next generation sequencing. Clin. Chem. 60, 1119–1199 (2014).
72. Hause, R. J., Pritchard, C. C., Shendure, J. & Salpante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. Nat. Med. 22, 1342–1350 (2016).
73. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).
74. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020).
75. Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat. Med. 25, 1526–1533 (2019).
76. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 575, 210–216 (2019).
77. Fujimoto, A. et al. Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. Genome Res. 30, 334–346 (2020).
78. Campbell, B. B. et al. Comprehensive analysis of hypermutation in human cancer. Cell 171, 1042–1056.e6 (2017).
79. Bresan, R. B. et al. Efficient CRISPR/Cas9-assisted gene targeting enables rapid and precise genetic manipulation of mammalian neural stem cells. Development 144, 635–648 (2017).
97. Tate, P. H. & Skarnes, W. C. Bi-allelic gene targeting in mouse embryonic stem cells. Methods 33, 331–338 (2011).
98. Hodgkins, A. et al. WGE: a CRISPR database for genome engineering. Bioinformatics 31, 3078–3080 (2015).
99. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
100. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. Curr. Protoc. Bioinform. 56, 15.10.1–15.10.18 (2016).
101. Raine, K. M. et al. ecppPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. Curr. Protoc. Bioinformatics 52, 15.7.1–15.7.12 (2015).
102. Craddock, T. J., Qiu, P., Lee, C. M., Fine, E. J. & Bao, G. COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. Mol. Ther. Nucleic Acids 3, e214 (2014).
103. The Encode Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
104. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
105. Wickham, H. ggplot2: Elegant Graphics for Data Analysis (Springer, 2009).

Acknowledgements

We thank the Wellcome Sanger Institute Cellular Genetics and Phenotyping Facility for assistance and the CASM IT team, J. Foreman and G. Ping for assistance in carrying out and completing this project. We thank the COMSIG Consortium spearheaded by S. Jackson. In Cambridge, this work was funded by the Cancer Research UK (CRUK) Advanced Clinician Scientist Award (C60100/A23916), the dr. Josef Steiner Cancer Research Award 2019, a Medical Research Council (MRC) Grant-in-Aid to the MRC Cancer Unit, the CRUK Pioneer Award, a Wellcome Strategic Award (WT101126) and Wellcome Sanger Institute faculty funding and supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014) and UK Regenerative Medicine Platform (MR/R015724/1). The work of T.I.R. and J.S.C. was funded by a CRUK Centre grant (reference number C309/A25144). Support for the MMRDetect classifier was enabled by access to data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England (a wholly owned company of the Department of Health and Social Care), funded by the NIHR and NHS England. The Wellcome Trust, CRUK and the MRC have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. The views expressed are those of the author(s) and not necessarily those of the NIHR or Department of Health and Social Care. This publication and the underlying research were facilitated by data that were generated by the Hartwig Medical Foundation (HMF) and Center for Personalized Cancer Treatment (CPCT) in the Netherlands and the International Cancer Genome Consortium.

Author contributions

S.N.-Z. and W.C.S. conceived of the study idea. S.N.-Z., X.Z., C.G.C.K., J.L. and W.C.S. wrote the paper. L.S., G.B., V.P.-A., D.R. and S.N.-Z. collected the clinical samples. G.C.C.K., K.U., T.I.R., C.A.A., W.B. and C.G. performed the laboratory work. X.Z., G.C.C.K., A.S.N., A.D., C.R., S.M., T.D.A., T.I.R., J.S.C. and S.N.-Z. performed data curation and formal analysis. R.H., W.B. and J.Y. performed administrative tasks.

Competing interests

S.N.-Z. holds patents on clinical algorithms of mutational signatures and, during completion of this project, served advisory roles for AstraZeneca, Artios Pharma and the Scottish Genomes Project.

Genomics England Research Consortium

John C. Ambrose13, Prabhu Arumugam13, Emma L. Baple13, Marta Bleda13, Freya Boardman-Pretty13,14, Jeanne M. Boissiere13, Christopher R. Bousted13, Helen Brittain13, Mark J. Caulfield13,14, Georgia C. Chan13, Clare E. H. Craig13, Louise C. Daugherty13, Anna de Burca13, James E. Holman13, Tim J. P. Hubbard13, Kristina ibáñez13,14, Rob Jackson13, Louise J. Jones13,14, Adam Giess13, Joanne M. Hackett13, Dina Halai13, Angela Hamblin13, Shirley Henderson13,14, James E. Holman13, Tim J. P. Hubbard13, Kristina ibáñez13,14, Rob Jackson13, Louise J. Jones13,14, Dalia Kasperavičiute13,14, Melis Kayikci13, Athanasios Kousathanas13, Lea Lahnstein13, Kay Lawson13, Sarah E. A. Leigh13, Ivonne U. S. Leong13, Javier F. Lopez13, Fiona Maleady-Crowe13, Joanne Mason13, Ellen M. McDonagh13,14, Loukas Moutsianas13,14, Michael Mueller13,14, Nirupa Murugaesu13, Anna C. Need13,14, Pter O’Donovan13, Chris A. Odhams13, Andreas Orioli13, Christine Patch13,14, Mariana Buongermino Pereira13, Daniel Perez-Gil13, Dimitris Polychronopoulos13, John Pullinger13, Tahrima Rahim13, Augusto Rendon13, Pablo Riesgo-Ferreiro13, Tim Rogers13, Mina Ryten13, Kevin Savage13, Kushmita Sawant13, Richard H. Scott13, Afshan Siddiq13, Alexander Sieghart13, Damian Smedley13,14, Katherine R. Smith13,14, Samuel C. Smith13,14, Alona Sosinsky13,14, William Spooner13, Helen E. Stevens13, Alexander Stuckey13, Razvan Sultana13, Mélanie Tanguy13, Ellen R. A. Thomas13,14, Simon R. Thompson13, Carolyn Tregidgo13, Arianna Tucci13,14, Emma Walsh13, Sarah A. Watters13, Matthew J. Welland13, Eleanor Williams13, Katarzyna Witkowska13,14, Suzanne M. Wood13,14, Magdalena Zarowiecki13

13The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. 14Genomics England, London, UK.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-021-00200-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-021-00200-0.

Correspondence and requests for materials should be addressed to S.N.-Z.

Peer review information Nature Cancer thanks Daniel Durocher and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Results of pilot study. Three genes were selected for knockout (\(\Delta\)): MSH6, UNG and ATP2B4 (negative control). Two genotypes per gene were obtained and grown in culture to gauge reproducibility of signatures between different genotypes of a gene-knockout. These lines were cultured under normoxic (20%) and hypoxic (3%) states, for defined culture times of ~15, 30 or 45 days. Two single-cell subclones were derived for whole genome sequencing for each parental line (equivalent to four subclones per gene edit). One of the UNG genotypes appeared to be heterozygous, which was excluded in downstream analysis. a, Substitution burden for knockouts of ATP2B4, UNG and MSH6 under hypoxic and normoxic conditions as well as different culturing time. b, The cosine similarities between the mutational profile of each subclone and background signature of culture. c, Indel burden for knockouts of ATP2B4, UNG and MSH6 under hypoxic and normoxic conditions as well as different culturing time. d, The cosine similarities between the mutational profile of each subclone with background signature of culture. Overall, the differences between normoxic and hypoxic conditions were not marked, although normoxic conditions produced slightly more mutations. Time in culture made only a marginal, non-linear difference to burden of mutagenesis. Given the results of the pilot, weighing up the costs and risks associated with prolonged culture time (risk of infection, risk of selection, marked increase in cost of experimental reagents) with the minimal return in terms of mutation number, and also intending to minimize transitions between hypoxic to normoxic conditions while handling cell cultures, we opted to proceed with the full-scale study under normoxic conditions and for 15 days for the rest of study.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Detecting mutational consequences of knockouts in the absence of added external DNA damage. a,b, Schematic illustration of potential components of background signature (a) and possible mutational consequences of the DNA repair gene knockouts for proteins that are critical mitigators of mutagenesis (b). c–e, Mutation burden of whole-genome-sequenced subclones of gene knockouts. c, Substitution, (d) indel and (e) double substitution. Bars represent the mean. Individual data points are shown in orange dots. In all comparative analyses, all gene knockouts were cultured for 15 days and only daughter subclones that were fully clonal (that is, clearly derived from a single cell) were included. \( N = 2-4 \), which is the number of clonal knockout subclones cultured under normoxic condition for 15 days (see Supplementary Table 2). f, 96-channel substitution mutation profiles of 173 gene knockout subclones.
Extended Data Fig. 3 | Results of contrastive principal component analysis and t-SNE. a, Contrastive principal component analysis (cPCA) was employed to discriminate knockout profiles from control profiles (ΔATP2B4). Each figure contains six different genes. Nine gene knockouts separate from the controls. Using this method, ΔADH5 did not separate clearly from ΔATP2B4, indicative of either having no signature or a weak signature. Dot colours indicate the repair/relicative pathway that each gene is involved: in black - control; green - MMR; orange – BER; dark purple – HR and HR regulation; light purple - checkpoint. Each dot represents a subclone. The number of subclones for each gene knockout (N = 2–4) can be found in Supplementary Table 2. b, The t-SNE algorithm was applied to discriminate the mutational profiles of gene knockouts from those of control knockouts. Gene knockouts that produce mutational signatures separate clearly from control subclones and other knockouts which do not have signatures. Subclones of the gene knockouts which produce signatures are clustered together, indicating consistency between subclones.
Extended Data Fig. 4 | Oxidative damage-associated mutational signatures. 

**a**. Relative mutation frequency of G>T/C>A in 256 possible channels which take two adjacent bases 5' and 3' of each mutated base (4×4×4×4=256) for ΔATP2B4, ΔOGG1, a head and neck cancer with strong SBS18 and SBS18.

**b**. Left: tSNE plot of tissue-specific mutational signature 18. Two groups are featured with predominant peaks at TGC>TTC/GCA>GAA (highlighted in green) and AGA>ATA/TCT>TAT (highlighted in purple), respectively. Right: heatmap of 21 tissue-specific mutational signatures at C>A. We compared experimental signatures to previously published cancer-derived signatures, focusing on 21 tissue-specific variations of Signature 18. Interestingly, we found two distinct groups of Signature 18. Signatures of ΔOGG1, cellular models and signatures derived from head and neck tumors, pancreas, myeloid, bladder, uterus, cervix, lymphoid tumors were most similar to each other, with the predominant G>T/C>A peak at TGC>TTC/GCA>GAA. By contrast, an alternative version of this signature with a predominant G>T/C>A peak at AGA>ATA/TCT>TAT was noted in colorectal, esophagus, stomach, bone, lung, CNS, breast, skin, prostate, liver, head and neck tumors (Signature Head_neck_G). ovary, biliary and kidney cancers. Indeed, there are many types of oxidative species which could fluctuate between tissues, variably affecting trinucleotides resulting in the variation observed in Signature 18.
Extended Data Fig. 5 | Indel signatures and double substitution signatures. a, 15-channel Indel signatures. b, 186-channel Indel signatures. c, Aggregated double substitution profile of ΔRNF168 and ΔEXO1.
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Similarities between ΔEXO1, ΔRNF168 signatures and RefSig5 and results of analysis on transcriptional strand bias and distribution of mutations on replication timing domains. **a**, Hierarchical clustering of cancer-derived reference signatures (RefSig) with ΔEXO1 and ΔRNF168 signatures. **b**, Hierarchical clustering of tissue-specific signature 5 with ΔEXO1 and ΔRNF168 signatures. **c**, Transcriptional strand bias in 9 gene knockouts. Pearson’s Chi-Squared test (chisq.test()) was used to calculate the p-value. P-value was corrected using p.adjust(). Unlike mutational signatures of environmental mutagens, we do not observe striking transcriptional strand bias in signatures generated by DNA repair gene knockouts, except for T>C generated by ΔEXO1 and ΔRNF168. Since transcriptional strand bias is largely induced by NER repairing DNA bulky adducts, lack of it indicates that most of the endogenous DNA damage is not particularly bulky or DNA-deforming. **d**, Distribution of mutation density across replication timing domains (separated into deciles) for signatures associated with different gene knockouts. Green bars indicate observed distribution. Blue lines indicate expected distribution with correction of trinucleotide density of each domain. Bars and error bars represent mean ± SD of bootstrapping replicates (n=100).
Extended Data Fig. 7 | Putative outcomes of all possible base-base mismatches. Outcomes from 12 possible base-base mismatches. The red and black strands represent lagging and leading strands, respectively. The arrowed strand is the nascent strand. The highlighted pathways are the ones that generate C>A (blue), C>T (red) and T>C mutations (green) in the ΔMSH2 mutational signature.
Extended Data Fig. 8 | Distribution of G>T/C>A mutations in polyG tracts of ∆MSH2, ∆MSH6 and ∆MLH1. a, Relative frequency of occurrence of G>T/C>A in polyG tracts. b, Occurrence of G>T/C>A in polyG tracts.
Extended Data Fig. 9 | Gene-specific mutational signatures in MMR-deficiency. Proportion of different mutation types of substitution (a) and indel (b) signatures for four MMR gene knockouts. c, The ratio of substitution and indel burden. d, Schematic interpretation of the relative mutation burdens of ∆MSH2 and ∆MSH6.
Extended Data Fig. 10 | Development of MMRDetect. (a)–(e) Distribution of the five parameters across IHC-determined MMR gene abnormal (orange) and MMR gene normal (green) samples. Black dots and error bars represent mean ± SD of the parameters. N_{abnormal} = 79 samples (yellow); N_{normal} = 257 samples (green). a, Exposure of MMRd signatures. b, Cosine similarity between the substitution profile of cancer samples and that of MMR gene knockouts. c, Number of indels in repetitive regions. d, Cosine similarity between the profile of repeat-mediated deletions of cancer sample and that of knockout generated indel signatures. e, The cosine similarity between the profile of repeat-mediated insertion of cancer sample and that of knockout generated indel signatures. P-values were calculated through two-sided Mann-Whitney test. f, Distribution of coefficients from 10-fold cross validation using training data set. Box plots denote median (horizontal line) and 25th to 75th percentiles (boxes). The lower and upper whiskers extend to 1.5× the inter-quartile range. N = 10 iterations. g, MMRDetect-calculated probabilities for 336 colorectal cancers. With cut-off of 0.7, 77 out of 336 were predicted to be MMR-deficient samples (probability < 0.7). Colour bars represent the MSI status determined by IHC staining: red – abnormal; blue – normal. Four samples with abnormal IHC staining have probabilities > 0.7, whilst 2 samples with normal IHC staining have probabilities < 0.7. The four samples were revealed to be false positive cases and the two samples were false negative ones for IHC staining through validation using MSIseq and seeking coding mutations in MMR genes. h, Distribution of the mutation number of repeat-mediated indels, MMRd signatures and non-MMRd signatures across four groups of samples: MMR-deficient samples determined by only MMRDetect (yellow), MMR-deficient samples determined by only MSIseq (purple), MMR-deficient samples determined by both MMRDetect and MSIseq (blue) and non-MMR-deficient samples determined by both MMRDetect and MSIseq (pink). P-values were calculated through two-sided Mann-Whitney test. Numbers of MMR-deficient samples determined by MMRDetect only (blue), MSIseq only (pink), both (yellow) and none (purple) are 34, 20, 587 and 6,718, respectively.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- n/a Confirmed
- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

CaVEMan 1.11 [http://cancerit.github.io/CaVEMan/] (Jones et al., 2016), Pindel 2.10 [http://cancerit.github.io/cggPindel] (Raine et al., 2015) and BRASS 5.3.2 [https://github.com/cancerit/BRASS]

Data analysis

Code can be obtained here: https://github.com/xqzhou/COMSIG_KO.git

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequence files are to be deposited at the European Genome/Phenome Archive with accession numbers EGAS00001000800 and EGAS00001000874. Mutation calls have been deposited at Mendeley: http://dx.doi.org/10.17632/ynn3kkmyx.3

Publicly available genomic datasets reanalyzed here to compare the performance of MMRDetect and MSisoeq are available at European Genome-phenome Archive [EGAS00001001178], http://dx.doi.org/10.10017/gx76f6q, https://data.mendeley.com/datasets/2mn4ctdpwp/1, https://resources.kartwig.medicalfoundation.nl and the Genomics England Research Environments (main programme v8) via https://re.egg.co.uk/ovd/.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

- Sample size
  - This was a systematic CRISPR screen. We aimed to knockout as many genes as possible using high throughput methods.

- Data exclusions
  - Off-target gene knockouts and/or mislabelled samples could cause erroneous results by reporting a signature incorrectly, thus, to reduce the likelihood of errors, we excluded these samples.

- Replication
  - In the manuscript we stated in the main text (first paragraph of results section) that we aimed for four subclones (two subclones for two genotypes of each gene). We stated that the use of successful (in the case of EXO1, MSH2, TDH, MDC1 and REV1) where we only managed one genotype, so for these, we also stated that we obtained 3 subclones for EXO1 and MSH2 and managed four for TDH, MDC and REV1.

- Randomization
  - This was not a clinical trial. We did not have a testing hypothesis to consider for the overall experiment.

- Blinding
  - This was not a clinical trial. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a | n/a |
| ☑ | ☑ |
| Antibodies | Involved in the study |
| ☑ | | |
| Eukaryotic cell lines | ChIP-seq |
| ☑ | | |
| Palaeontology and archaeology | Flow cytometry |
| ☑ | | |
| Animals and other organisms | MRI-based neuroimaging |
| ☑ | | |
| Human research participants | | |
| ☑ | | |
| Clinical data | | |
| ☑ | | |
| Dual use research of concern | | |

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

The human iPSC line used for this study was derived at the Wellcome Sanger Institute (Hinxton, UK). It is a long-standing iPSC line originally derived from a patient with alpha-1-antitrypsin deficiency, for which one of the alleles was corrected. The cell line is diploid and does not have any known driver mutations. It does carry a balanced translocation between chromosomes 6 and 8. It is stably growing in culture and does not acquire a vast number of karyotypic abnormalities.

Authentication

All donor-derived human iPSC lines were assessed by array-comparative genomic hybridization, expression arrays and cellomics immunohistochemistry markers to ensure that they had the necessary pluripotency and novelty scores expected of hiPSCs.

Mycoplasma contamination

This cell line was tested negative for mycoplasma contamination.

Commonly misidentified lines (See ICTAC register)

No commonly misidentified lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics

- MSH3, female, PMS2 mutant
| Population characteristics | MSH77, male, PMS2 mutant  
|                           | MSH89, female, MSH6 mutant  
|                           | MSH94, female, MSH6 mutant  
|                           | MSH159, male, normal control  
| The age information of the human patient samples is not publicly available as this information could compromise privacy and lead to identification of the individuals. |
| Recruitment               | CMMRD patients were recruited at Doce de Octubre University Hospital, Spain, St George’s Hospital in London and Great Ormond Street Hospital under the auspices of the Insignia project. This included two PMS2-mutant patients and two MSH6-mutant patients. A healthy donor was recruited as control. |
| Ethics oversight          | Ethical approval for the generation of hiPSCs from patients and healthy controls was received for the Insignia project under the title “Exploring the biological processes underlying mutational signatures identified in patients with inherited disorders and in patients exposed to mutagens”, with reference number 13/EE/0302, from the East of England Cambridgeshire and Hertfordshire Research Ethics Committee. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.