Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers

Authors:

Meier B*[1], Volkova N*[2], Hong Y[1], Schofield P[1,3], Campbell PJ [4, 5, 6] Gerstung M[2] and A Gartner[1]
* these authors contributed equally.

Institutions:

[1] Centre for Gene Regulation and Expression, University of Dundee, Dundee, UK.
[2] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, UK.
[3] Division of Computational Biology, University of Dundee, Dundee, UK.
[4] Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.
[5] Department of Haematology, University of Cambridge, Cambridge, UK.
[6] Department of Haematology, Addenbrooke's Hospital, Cambridge, UK.

Correspondence:

Dr Anton Gartner
Centre for Gene Regulation and Expression
University of Dundee
Dow Street
Dundee, DD1 5EH
UK.
Tel: +44 (0) 1382 385809
E-mail: a.gartner@dundee.ac.uk

Dr Moritz Gerstung
European Molecular Biology Laboratory
European Bioinformatics Institute (EMBL-EBI)
Hinxton, CB10 1SA
Cambridgeshire
UK.
Tel: +44 (0) 1223 494636
E-mail: moritz.gerstung@ebi.ac.uk

Keywords: DNA mismatch repair, mutational signatures, hypermutation, replication errors, POLE-4, genome integrity, *C. elegans*

1. ABSTRACT

Throughout their lifetime cells are subject to extrinsic and intrinsic mutational processes leaving behind characteristic signatures in the genome. One of these, DNA mismatch repair (MMR) deficiency leads to hypermutation and is found in different cancer types. While it is possible to associate mutational signatures extracted from human cancers with possible mutational processes the exact causation is often unknown. Here we use *C. elegans* genome sequencing of *pms-2* and *mlh-1* knockouts to reveal the mutational patterns linked to *C. elegans* MMR deficiency and their dependency on endogenous replication errors and errors caused by deletion of the polymerase ε subunit *pole-4*. Signature extraction from 215 human colorectal and 289 gastric adenocarcinomas revealed three MMR-associated signatures one of which closely resembles the *C. elegans* MMR spectrum. A characteristic difference between human and worm MMR deficiency is the lack of elevated levels of NCG>NTG mutations in *C. elegans,* likely caused by the absence of cytosine (CpG) methylation in worms. The other two human MMR signatures may reflect the interaction between MMR deficiency and other mutagenic processes, but their exact cause remains unknown. In summary, combining information from genetically defined models and cancer samples allows for better aligning mutational signatures to causal mutagenic processes.

2. INTRODUCTION

Cancer is a genetic disease associated with the accumulation of mutations. A major challenge is to understand mutagenic processes acting in cancer cells. Accurate DNA replication and the repair of DNA damage are important for genome maintenance. The identification of cancer predisposition syndromes caused by defects in DNA repair

genes was important to link the etiology of cancer to increased mutagenesis. One of the first DNA repair pathways associated with cancer predisposition was the DNA mismatch repair (MMR) pathway. MMR corrects mistakes that arise during DNA replication. Mutations in MMR genes are associated with hereditary non-polyposis colorectal cancer (HNPCC), also referred to as Lynch Syndrome [1-5]. Moreover, MMR deficiency has been observed in a number of sporadic cancers.

DNA mismatch repair is initiated by the recognition of replication errors by MutS proteins, initially defined in bacteria. In *S. cerevisiae* and mammalian cells, two MutS complexes termed MutSα and MutSβ comprised of MSH2/MSH6 and MSH2/MSH3, respectively (Table 1), are required for DNA damage recognition albeit with differing substrate specificity [6-8]. Binding of MutS proteins to the DNA lesion facilitates subsequent recruitment of the MutL complex. MutL enhances mismatch recognition and promotes a conformational change in MutS through ATP hydrolysis to allow for the sliding of the MutL/MutS complex away from mismatched DNA [9, 10]. DNA repair is initiated in most systems by a single-stranded nick generated by MutL (MutH in *E. coli*) on the nascent DNA strand at some distance to the lesion [11, 12]. Exonucleolytic activities in part conferred by Exo1 contribute to the removal of the DNA stretch containing the mismatch followed by gap filling via lagging strand DNA synthesis [13-19]. The most prominent MutL activity in human cells is provided by the MutLα heterodimer, MLH1/PMS2 (Table 1) [20, 21]. However, human MLH1 is also found in heterodimers with PMS1 and MLH3, called MutLβ and MutLγ. Of these two only MutLγ is thought to have a minor role in MMR [21]. In *C. elegans*, MLH-1 and PMS-2 are the only MutL homologs encoded in the genome (Table 1).

A large number of studies have analyzed mutations arising in DNA mismatch repair deficient cells at specific genomic loci or in reporter constructs. Analysis of microsatellite loci in *mlh1* deficient colorectal cancer cell lines suggested rates of repeat expansions or contractions between 8.4x 10^{-3} to 3.8x 10^{-2} per locus and generation [22, 23]. Estimates using *S. cerevisiae* revealed a 100- to 700-fold increase in DNA repeat tract instability in *pms2*, *mlh1* and *msh2* mutants [24] and a ~5-fold increase in base substitution rates [25]. *C. elegans* assays using reporter systems or selected, PCR-amplified regions revealed a more than 30-fold increased frequency of single base substitutions in *msh-6*, a 500-fold increase in mutations in A/T homopolymer runs and a 100-fold increase in mutations in dinucleotide repeats [26-28], akin to the frequencies observed in yeast and mammalian cells [23, 24]. Recently, whole genome sequencing approaches using diploid *S. cerevisiae* started to provide a genome-wide view of MMR deficiency. *S. cerevisiae* lines carrying an *msh2* deletion alone or in conjunction with point mutations in one of the three replicative polymerases, Polα/primase, Polδ, and Polε, were propagated over multiple generations to determine the individual contribution of replicative polymerases and MMR to replication fidelity [29-31]. These analyses observed an average base substitution rate of 1.6 x 10^{-8} per base pair per generation in *msh2* mutants and an increased rate in mutants in which *msh2* and one of the replicative polymerases was mutated [29]. A synergistic increase in mutagenesis was also recently observed in childhood tumors in which MMR deficiency and mutations in replicative polymerase ε and δ needed for leading and lagging strand DNA synthesis occurred [32].

In human cancer samples 30 mutational signatures (referred to as COSMIC signatures from here on) have been uncovered by mathematical modeling across a large number of cancer genomes representing more than 30 tumor types

(http://cancer.sanger.ac.uk/cosmic/signatures) [33, 34]. These signatures are largely define by the relative frequency of the six possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G), occurring in a sequence context defined by their adjacent 5' and 3' base. Of these, COSMIC signatures 6, 15, 20, 21 and 26, have been associated with MMR deficiency with several MMR signatures being present in the same tumor sample. It is not clear if these MMR signatures are conserved across evolution and how they reflect MMR defects. Therefore, MMR 'signatures' deduced from defined monogenic MMR defective backgrounds, we refer to as mutational patterns to distinguish them from computationally derived signatures, could contribute to refine signatures extracted from cancer genomes.

Here we investigate the genome-wide mutational impact of the loss of the MutL mismatch repair genes *mlh-1* and *pms-2* in the nematode *C. elegans*. Furthermore, we address the contribution of a deletion of *pole-4*, a non-essential accessory subunit of the leading-strand DNA polymerase Polε, to mutation profiles and hypermutation. We find that mutation rates for base substitutions and small insertions or deletions (indels) are ~100 fold higher in MMR mutants compared to wild-type and are further increased ~2-3 fold in combination with *pole-4* deficiency. Mutation patterns encompass single base substitutions with a preponderance of transition mutations and 1 base pair indels largely occurring in mono- and polynucleotide repeats. A subset of base substitutions occurs within repetitive DNA sequences, likely a consequence of replication slippage. Given the overall high frequency of base substitutions we were able to extract a robust mutational signature from *C. elegans* MMR mutants. Extracting main signatures from 504 colorectal and stomach cancers defined a *de novo* signature with high similarity to the *C. elegans* MMR mutational pattern. Stratification based on the status of MMR genes confirmed an

enrichment of this signature in cancers with defects in MMR genes. We postulate that additional cancer signatures may reflect a combination of DNA repair or replication deficiencies in conjunction with MMR defects.

3. RESULTS

3.1. Mutation rates and profiles of *mlh-1*, *pms-2* and *pole-4* single mutants grown over 20 generations

We previously established *C. elegans* mutation accumulation assays and demonstrated that defects in major DNA damage response and DNA repair pathways, which included nucleotide excision repair, base excision repair, DNA crosslink repair, DNA end-joining and apoptosis did not lead to overtly increased mutation rates when lines were propagated for 20 generations [35]. We now extend these studies to MMR deficiency conferred by MutL mutations. The experimental setup takes advantage of the 3-4 days life cycle of *C. elegans* and its hermaphroditic reproduction of self-fertilization. This allows the propagation of clonal *C. elegans* lines, which in each generation pass through a single cell bottleneck provided by the zygote. The *C. elegans* genome does not encode obvious MutLβ and γ subunits (PMS1 and MLH3 homologues, respectively), while the MutLα subunits MLH-1 and PMS-2 can be readily identified using homology searches (Table 1). Given that replicative polymerases are essential for viability we focused on one of the non-essential subunits of the Polε leading strand polymerase termed POLE-4.

We detected an average of 5 base substitutions and 4 insertions or deletions in wild-type *C. elegans* lines propagated for 20 generations (Figure 1A, 1B). In contrast, *mlh-1* and

pms-2 mutants carried an average of 844 and 871 unique mutations, respectively, 288 and 309 of which were base substitutions (Figure 1A) and 556 and 563 indels, defined as small insertions and deletions of less than 100 base pairs (Figure 1B). No increased number of structural variants or chromosomal rearrangements was observed in *mlh-1* and *pms-2* mutants (data not shown). The nature of single nucleotide changes and the overall mutation burden were congruent across independent lines of the same genotype and mutation numbers linearly increased from F10 to F20 generation lines (Figure 1, Table 2). Thus, despite the high mutation frequency observed in MMR mutants, we did not observe evidence for secondary mutations that altered mutation profiles or frequencies. In contrast to *mlh-1* and *pms-2* mutants, *pole-4* lines exhibited mutation rates and profiles not significantly different from wild-type (Figure 1, Table 2), a finding we confirmed for *pole-4* lines propagated over 40 generations (data not shown).

3.2. Mutation rates and patterns in *pole-4; pms-2* double mutants

To further investigate the role of *pole-4* and the genetic interaction with MMR deficiency, we generated *pole-4; pms-2* double mutants. Wild-type, *pms-2*, *pole-4* and *pole-4; pms-2* strains obtained as siblings from a cross of highly backcrossed *pms-2* and *pole-4* single mutants were grown for 10 generations and analyzed for mutation frequency and profiles. *pms-2* mutant strains carried an average of 145 base substitution and 340 indels over 10 generations, roughly half the number we observed in the F20 generation (Figures 1C, 1D, Table 2). In comparison, the number of single base substitutions and indels was increased ~4.4 fold and ~1.5 fold in *pole-4; pms-2* double mutants to an average of 637 and 564 (Table 2). No increased frequency of structural variants and chromosomal rearrangements was observed in *pms-2* single or *pole-4; pms-2* double mutants (Table 2). We could not readily propagate *pole-4; pms-2* beyond

the F10 generation, suggesting that a mutation burden higher than ~500-700 single base substitutions (Figure 1C) in conjunction with the 500-600 indels (Figure 1D) might be incompatible with organismal reproduction. These numbers are in line with us not being able to stably propagate *mlh-1* and *pms-2* single mutant lines for 40 generations. The multiplicative effect on mutation burden detected in *pole-4; pms-2* double mutants while no increased mutation rate is observed in *pole-4* alone suggests that replication errors occur at increased frequency in the absence of *C. elegans pole-4* but are effectively repaired by MMR.

Given the *C. elegans* genome size of 100 million base pairs, assuming that it takes 15 cell divisions to go through the *C. elegans* life cycle, and taking into account that heterozygous mutations can be lost during self-fertilization, we estimated the overall mutation rate for wild-type of 8.34×10^{-10} (95% CI: 6.46×10^{-10} to 1.06×10^{-9}) per base pair and germ cell division and of 1.38×10^{-9} (95% CI: 1.14×10^{-9} to 1.67×10^{-9}) for *pole-4* mutants, based on mutations observed in F10 and F20 generations. These wild-type estimates are consistent with our previous findings [35], the variation likely being due to the low number of mutations observed in these genetic backgrounds. In contrast, estimated mutation rates for *mlh-1* and *pms-2* were 5.12×10^{-8} (95% CI: 4.93×10^{-8} to 5.32×10^{-8}) and 5.22×10^{-8} (95% CI: 5.06×10^{-8} to 5.39×10^{-8}) per base pair and cell division, respectively, and *pole-4; pms-2* double mutants exhibited a mutation rate of 1.33×10^{-7} (95% CI: 1.28×10^{-7} to 1.38×10^{-7}).

The genome-wide mutation rates observed in the absence of *C. elegans* MutLα proteins *MLH-1* and *PMS-2* are in line with mutation rates previously determined for *C. elegans* MutS and *S. cerevisiae* MMR mutants [24-28]. However, unlike in mammalian cells

[36, 37], the two *C. elegans* MutLα mutants, *mlh-1* and *pms-2,* exhibited almost identical mutation rates and profiles, suggesting that the inactivation of the MutLα heterodimer is sufficient to induce a fully penetrant MMR phenotype *in C. elegans*. These results are consistent with the absence of readily identifiable PMS1 MutLβ and MLH3 MutLγ homologs in the *C. elegans* genome (Table 1). Our finding that *pole-4* mutants do not show increased mutation rates is surprising given that mutation of the budding yeast homolog of the POLE-4 polymerase ε accessory subunit termed Dpb3 leads to increased mutation rates [38]. Enhanced mutation rates have also been reported for hypomorphic mutants of the Polε catalytic subunit in yeast and mice, and in humans such mutations are associated with an increased predisposition to colorectal cancer [39-41].

3.3 Distribution and sequence context of base substitutions

We next wished to determine the mutational patterns associated with DNA mismatch repair and combined MMR *pole-4* deficiency. Given the complementarity of base pairing, 6 possible base changes, namely C>T, C>A, C>G and T>A, T>C and T>G can be defined. T>C and C>T transitions were present more frequently than T>A, T>G, C>A and C>G transversions in *mlh-1* and *pms-2* single and *pole-4; pms-2* double mutants (Figure 1A, 1C, Table 2). A similar preponderance of T>C and C>T transitions was previously observed in *S. cerevisiae msh2* mutants and in MMR defective human cancer lines [30, 33, 42]. Analyzing these base substitutions within their 5' and 3' sequence context, we found no enrichment of distinct 5' and 3' bases associated with T>C transitions prominent in *mlh-1* and *pms-2* single mutants. In contrast, T>A transversions occurred with increased frequency in an A<u>T</u>T context, C>T transitions in a G<u>C</u>N context and C>A transversions in a N<u>C</u>T context (Figure 1E).

Analysis of the broader sequence context of T>A transversions in an A<u>T</u>T context revealed that > 90% of substitutions occurred in homopolymer sequences; the majority (> 75%) in the context of two adjoining A and T homopolymers (Suppl. Figure 1A). Similarly, an increased frequency of base substitution at the junction of adjacent repeats has recently been reported in *S. cerevisiae* MMR mutants, giving rise to the speculation that such base substitutions may be generated by double slippage events [31]. To further analyze base changes we visually searched for base changes occurring in repeat sequences. We found several examples in which one or several base substitutions had occurred that converted a repeat sequence such that it became identical to flanking repeats consistent with polymerase slippage across an entire repeat (Suppl. Figure 1B-D). Such mechanisms could lead to the equalization of microsatellite repeats a phenomenon referred to as microsatellite purification [43].

While we could not define mutational patterns specifically associated with *pole-4* loss due to the low number of mutations, the profile of *pole-4; pms-2* double mutants differed from mismatch repair single mutants. Most strikingly, in addition to C>T transitions in a G<u>C</u>N context, T>C transitions were generated with higher frequency accounting for >50% of all base changes (Figure 1C and 1E). These T>C substitutions appeared largely independent of their sequence context with an underrepresentation of a flanking 5' cytosine (Figure 1E). Interestingly, T>C changes not embedded in a distinct sequence context have also been reported for MMR-deficient tumor samples containing mutations in the lagging strand polymerase Polδ [32], but not in *S. cerevisiae* and human tumors with a combined MMR and Polε deficiency [30, 32]. No obvious

chromosomal clustering of base substitutions was observed in *pms-2* and *pole-4; pms-2* grown for 10 generations (Suppl. Figure 2A).

3.4 Sequence context of insertions and deletions associated with MMR deficiency

The majority of mutation events observed in *mlh-1* and *pms-2* single and *pole-4; pms-2* double mutants were small insertions/deletions (indels) (Figure 1B, 1D). Indel numbers were increased 1.5 fold in *pole-4; pms-2* double compared to *pms-2* single mutants (Figure 1D, Table 2). Over 90% of all indels in single and double mutant backgrounds constituted 1 bp insertions or deletions, most 1bp indels occurring within homopolymer runs (Figure 1B, 1D, Table 2). 2 bp indels accounted for 3-5% of all indels and affected homopolymer runs as well as dinucleotide repeat sequences at similar frequency (Figure 1B, 1D, Table 2) as recently also reported for MMR defective *S. cerevisiae* strains [29]. Trinucleotide repeat instability is associated with a number of neurodegenerative disorders, such as fragile X syndrome, Huntington's disease and Spinocerebellar Ataxias [44]. Based on our analysis, trinucleotide repeat runs are present in the *C. elegans* genome at > 200 fold lower frequency than homopolymer runs (see Material and Methods). Combining F20 and F10 generation samples of *mlh-1* and *pms-2* mutants, we observed between 0.5 and 1.3 trinucleotide indels per 10 generations (data not shown) precluding an estimation of mutation rates for these lesions.

Mutation clustering along chromosomes was not evident for 1 bp indels in F10 *pms-2* and *pole-4; pms-2* lines beyond a somewhat reduced occurrence in the center of *C. elegans* autosomes (Suppl. Figure 2B, panels bottom left and right), which correlated with a reduced homopolymer frequency in these regions (Suppl. Figure 2B, top panel).

3.5 Dependency of 1 bp indel frequency on homopolymer length

Given the high number of indels arising in homopolymer repeats we aimed to investigate the correlation between the frequency of indels and the length of the homopolymer in which they occurred. We identified 3.433.785 homopolymers, defined as sequences of 4 or more identical bases with the longest homopolymer being comprised of 35 Ts in the *C. elegans* genome (Figure 2A, Suppl. Table 1, Materials and Methods). 47% of genomic homopolymers are comprised of As, 47% of Ts and 3% each account for Gs and Cs (Figure 2A). A and T homopolymer frequencies decrease continuously with increasing homopolymer length; C and G homopolymer frequencies decrease up to homopolymer lengths of 8 bp, followed by roughly consistent numbers for homopolymers of 8-17 bp length and decreasing frequencies with longer homopolymers (Figure 2A). Although we identified slightly higher overall numbers of homopolymer runs in the genome, this base specific size distribution is consistent with previous reports (Suppl. Table 1) [45]. Plotting the frequency of all observed 1 bp indels in our MMR mutant backgrounds in relation to the length of the homopolymer in which they occur, we found that the likelihood of indels increased with homopolymer length of up to 9-10 base pairs, and trails off in longer homopolymers (Figure 2B, top panel). Given that the frequency of homopolymer tracts decreases with length (Figure 2A) we normalized for homopolymer number (Figure 2B, bottom panel). Using this approach we still observed the highest indel frequency in 9-10 nucleotide homopolymers. These results are consistent with observations in budding yeast [31]. To assess the variability of the frequency estimation, we applied an additive model (Materials and Methods) which supported a rapid increase for homopolymers up to length 9 followed by a drop or plateau in indel frequency for longer homopolymer with decreasing confidence (Figure 2C). Firm conclusions about indel frequencies in homopolymers >13 bp are precluded

by the lack of statistical power based on low numbers of long homopolymers in the genome and too few observed indel events (Figure 2B). In summary, our data suggest that replicative polymerase slippage occurs more frequently with increasing homopolymer length, with a peak for homopolymers of 9-10 nucleotides, followed by reduced slippage frequency in slightly longer homopolymers.

3.6 Relation of *C. elegans* MMR patterns to MMR signatures derived from human colorectal and gastric adenocarcinoma samples

To assess how our findings relate to mutation patterns occurring in human cancer we analyzed whole exome sequencing data from the TCGA colorectal adenocarcinoma project (US-COAD) (http://icgc.org/icgc/cgp/73/509/1134) [46] and the TCGA gastric adenocarcinoma project (US-STAD) (https://icgc.org/icgc/cgp/69/509/70268) [47]. These cancer types are commonly associated with MMR deficiency. These datasets contain single nucleotide (SNV) and indel variant calls from 215 and 289 donors, respectively. Having observed high 1bp indel frequencies associated with homopolymer repeats in *C. elegans pms-2* and *mlh-1* mutants (Figure 1B and 1D, Figure 2B), we also considered indels in our analysis of human mutational signatures.

De novo signature extraction across both tumor cohorts combined using a Poisson version of the non-negative matrix factorization (NMF) algorithm (Material and Methods) revealed eight main mutational signatures (Figure 3A). Comparing these signatures to existing COSMIC signatures by calculating the similarity score between their base substitution profiles shows that many have a counterpart in the COSMIC database with high similarity (Table 3, Suppl. Figure 3B), validating our results. We labeled three of these signatures as “MMR-1-3”. MMR-1 shares similarity with MMR-

related COSMIC signature 20, MMR-2 with COSMIC signature 15 and MMR-3 with COSMIC signatures 21 and 26 (Table 3, Suppl. Figure 3B) ([33], http://cancer.sanger.ac.uk/cosmic/signatures). Additional signatures identified in the tumor samples were characteristic of *POLE* mutations ("POLE") ([33], http://cancer.sanger.ac.uk/cosmic/signatures). C>T mutations in a CpG base context result from 5meC deamination and are referred to as "Clock-1 (5meC)". "Clock-2" is present in the majority of samples and likely reflects the background mutation rates. 17-like is of unknown etiology predominantly found in stomach cancers and related to COSMIC signature 17 [33]. Finally, we identified a signature of SNP contamination characterized by high overlap of somatic mutations and SNPs from 1000 Genomes database (http://www.internationalgenome.org/home) and lower nonsynonymous to synonymous ratio, as expected for germline variants.

To confirm the link between signatures MMR-1-3 and mismatch repair deficiency, we correlated their occurrence with the presence of putative high-impact somatic mutations (see Material and Methods) in MMR genes. We considered mutations in *MLH1*, *MLH3, PMS2*, *MSH2*, *MSH3* and *MSH6*, as well as deletions in *EPCAM,* as well as *MLH1* promoter hypermethylation [48] (Material and Methods). *EPCAM* is the 5'gene of *MSH2* and deletions of the last few exons of this gene have been shown to inhibit *MSH2* expression [49]. We found that 43 out of the 215 (20%) samples in the COAD cohort and 59 out of the 289 (20%) samples in the STAD cohort contained high-impact mutations or epigenetic silencing in MMR genes. Comparing MMR deficient and proficient samples, we found that the fraction of MMR-1, MMR-2 and MMR-3 signatures was significantly higher in MMR defective cancer samples (P-values 4.6×10^{-28}, 3.2×0^{-10} and 4.8×10^{-7}, respectively) (Figure 3B). Overall, consistent with the results from *C. elegans* the presence of MMR mutations also correlated with an

increased frequency of 1bp indels in MMR deficient samples (one-tailed *t*-test *p*-value of 2.2×10^{-25}) (Figure 3C), of which 60% to 85% occurred in homopolymer runs (Suppl. Figure 4B). Indels were predominately associated with MMR-1 (Figure 3A). We noted that a small number of cancers without obvious MMR gene mutations displayed a high contribution of MMR-1-3 to their overall mutation spectra (Figure 3B, 3D). Similarly, amongst the samples without obvious mutations in MMR genes there are several samples with high number of indels (Figure 3C). It is reasonable to postulate that these tumors might be associated with epigenetic inactivation of other MMR genes or due to mutation of unknown genes affecting MMR.

Individual signatures often represent the most extreme ends of the mutational spectrum; a typical tumor, however, is usually represented by a linear combination of multiple processes. We thus aimed to cluster tumor samples based on the similarity of their mutational profiles using a *t*-SNE representation (Material and Methods). We found that many tumors cluster into distinct subgroups, usually characterized by specific signature combinations (Figure 3D). A clear cluster of samples with variable combinations of signatures MMR-1-3 emerged that was highly enriched for MMR deficient samples. MMR-1 occurs in the majority of these tumors, while MMR-2 and MMR-3 occur in a small number of tumors that contain high numbers of mutations. Interestingly, the most severely hypermutated samples are largely described by a single signature (MMR-2 in purple, MMR-3 in orange). Minor contributions of MMR-2 and MMR-3 in other tumors may reflect the tendency of the signature calculation method to extract signatures predominantly from the most extreme cases, and impart them on other samples.

A second cluster represented by two tumors (bottom right) is MMR defective, but most mutations observed in these tumors fall within the POLE signature (brown).

Consistently, these samples also carry pathogenic *POLE* mutations (data not shown). In addition, a number of tumor samples outside of these clusters and dispersed over the similarity map carry MMR mutations. These tumors may have acquired MMR deficiency very late in their development or may harbor MMR mutations that do not cause functional defects.

The Clock-1/COSMIC signature 1 is associated with (5meC) deamination and has been described to occur across all cancer types and is correlated with the age at the time of cancer diagnosis [33]. Interestingly, we observed an increased number of mutations assigned to the Clock-1 signature in human MMR deficient samples (P-value = 4.3×10^{-22}), which suggests an increased rate of spontaneous cytosine deamination under MMR deficient conditions. Previous research has suggested an interaction between base excision repair (BER) and MMR pathways in the repair of deaminated bases [50-52]. Notably COSMIC signatures 6 and 20, both associated with defective MMR, show high rates of C>T mutations in NCG contexts possibly reflecting the mixture between a MMR mutational footprint and the acceleration of cytosine deamination.

To compare human and *C. elegans* MMR footprints we first determined mutational patterns from *mlh-1* and *pms-2* single mutants as well as from the *pole-4; pms-2* double mutant (Material and Methods). *mlh-1* and *pms-2* mutational patterns are nearly identical with a cosine similarity of 0.97. In contrast the *pole-4; pms-2* mutational pattern shows a very different relative contributions of C>T and T>C mutations (Figure 4A, top panels). We next adjusted for the differential trinucleotide frequencies in the *C. elegans* genome and the human exome (Figure 4A bottom panels, 4B). Of the three human MMR-associated *de novo* signatures, only MMR-1 displayed similarity to *C. elegans* MMR substitution patterns with a cosine similarity of 0.84 to *pms-2* and of 0.81 to *mlh-1*

mutant profiles (Table 3, Figure 4C). A notable difference in the *C. elegans pms-2* and *mlh-1* patterns in comparison to MMR-1 are a reduced level of C>T mutations in a N<u>C</u>G context (Figure 4C, stars) and a high frequency of T>A mutation in an ATT context. The first is likely due to the lack of spontaneous deamination of 5methyl-C, a base modification that is absent in *C. elegans* [53], the latter likely due to a higher frequency of poly-A and poly-T homopolymers in the *C. elegans* genome versus the human exome (Figure 2A, Suppl. Figure 4A). Thus excluding C>T's in NCG contexts from the analysis, MMR-1 shows a similarity of 0.92 to *pms-2* and 0.90 to *mlh-1* patterns. The similarity is further supported by both *C. elegans* MMR and the human MMR-1 mutational footprints include a large number of one base pair insertions and deletions. None of the human signatures showed notable similarity to the *pole-4; pms-2* mutation pattern.

4. DISCUSSION

Here we characterized the mutational landscape associated with *C. elegans* MutL MMR deficiency. Out of a large number of DNA repair deficiencies we analyzed including nucleotide excision repair, base excision repair, homologous recombination, DNA-end-joining, polymerase θ-dependent microhomology-mediated end-joining, and the Fanconi Anemia pathway, MMR deficiency leads to the by far highest mutation rate ([35], unpublished data). This mutation rate is only surpassed by that of the *pole-4; pms-2* double mutant in which mutation rates are further increased 2-3 fold. Genome maintenance is highly efficient as evidenced by a wild-type *C. elegans* mutation rate in the order of 8×10^{-10} per base and cell division. It thus appears that DNA repair pathways act highly redundantly, and that it may require the combined deficiency of

multiple DNA repair pathways to trigger excessive mutagenesis. Equally a latent defect in DNA replication integrity might only become apparent in conjunction with a DNA repair deficiency. Indeed the increased mutation burden detected in the *pole-4; pms-2* double mutant while no increased mutation rate is observed in *pole-4* alone uncovers a latent role of *pole-4.* It appears that replication errors occur at increased frequency in the absence of *C. elegans pole-4* but are effectively repaired by MMR.

Out of the signatures associated with MMR deficiency in cancer cells, only MMR-1 is related to the mutational pattern found in *C. elegans mlh-1* and *pms-2* mutants. Taking into account the controlled nature of the *C. elegans* experiment we postulate that MMR-1 reflects a 'basal' conserved mutational process of DNA replication errors repaired by MMR. Consistent with this we find that MMR-1 occurs in almost all MMR defective tumors, except for very few tumors that show hypermutation. In these cases we suggest that akin to the *pole-4; pms-2* double mutant, mutational footprints can be attributed to the failed repair of lesions originating from mutations in DNA repair or DNA replication genes. For instance in MMR defective lines also carrying *POLE* catalytic subunit mutations the mutational landscape is overwhelmed by the POLE signature [32]. Likewise we postulate that the MMR-2 and MMR-3 signatures could be attributed to hypermutation, not directly linked to MMR defects. Thus we think that MMR-1 reflects a 'basal' conserved mutational process in humans and *C. elegans*, but that human MMR deficiency also includes an element of failing to repair C>T changes linked to CpG demethylation a process that does not occur in *C. elegans.* This signature, "Clock-1", together with MMR-1 explains the majority of mutations occurring in MMR defective cancers not apparently affected by hypermutation.

Matching mutational signatures to DNA repair deficiency has a tremendous potential to

stratify cancer therapy tailored to DNA repair deficiency. This approach appears advantageous over DNA repair gene genotyping, as mutational signatures provide a read-out for cellular repair deficiency associated with either genetic or epigenetic defects. Following on from our study we expect that analyzing DNA repair defective model organisms and human cell lines, alone or in conjunction with defined genotoxic agents, will contribute to more precisely define mutational signatures occurring in cancer genomes and to relate them to their etiology.

5. MATERIAL AND METHODS

5.1 *C. elegans* strains, crosses, maintenance and propagation.

C. elegans mutants *pole-4(tm4613) II*, *pms-2(ok2529) V* and *mlh-1(ok1917) III* were backcrossed 6 times against the wild-type N2 reference strain TG1813 [35] and frozen for glycerol stocks. Backcrossed strains were then grown for 20 generations as described previously [35]. To obtain the *pole-4 II; pms-2 V* double mutant, freshly backcrossed *pole-4(tm4613)* and *pms-2(ok2529)* lines were crossed to each other and F2 individuals were genotyped to obtain the four genotypes N2 wild-type (TG3551), *pms-2* (TG3552), *pole-4* (TG3553) and *pole-4; pms-2* (TG3554). 5 to 10 lines were propagated per genotype for mutation accumulation over 10 or 20 generations. Genomic DNA was prepared from the initial (P0 or F1) and two to three lines of the final generation per genotype and used for whole-genome sequencing as described [35].

5.2 DNA sequencing, variant calling and post-processing.

Illumina sequencing, variant calling and post-processing filters were performed as described previously [35] with the following adjustments. WBcel235.74.dna.toplevel.fa was used as the reference genome (http://ftp.ensembl.org/pub/release-

74/fasta/caenorhabditis_elegans/dna/). Alignments were performed with BWA-mem and mutations were called using CaVEMan, Pindel and BRASSII [54, 55], each available at (https://github.com/cancerit). Post-processing of mutation calls was performed across a large dataset of 2202 sequenced samples using filter conditions described previously [35] followed by removal of duplicates with identical base changes in the corresponding samples. Finally, an additional filtering step using deepSNV [56, 57] was used to correct for wrongly called base substitutions, events due to algorithm-based sequence misalignment of ends of sequence reads covering 1bp indels (see Figure 2C).

5.3 Estimating mutation rates.

With progeny arising through self-fertilization in *C. elegans*, selection of single animals among the progeny leads to an equal probability of ¼ of newly acquired heterozygous germline mutations either being lost or manifested as homozygous in every generation. Mutation rates were calculated using maximum likelihood methods, assuming 15 cell divisions per generation [35], and considering that mutations have a 25% chance to be lost, a 50% chance to be transmitted as heterozygous, and a 25% chance to become homozygous, thus becoming fixed in the line during each round of *C. elegans* self-fertilization.

5.4 Analysis of homopolymer sequences in *C. elegans.*

Homopolymers, di- and tri-nucleotide runs encoded in the *C. elegans* genome, defined here as repetitive DNA regions with a consecutive number of identical bases or repeated sequence of n≥4, were identified from the reference genome WBcel235.74 using an in house script based on R packages Biostrings and GenomicRanges [58, 59]. Overall, 3433785 homopolymers with a length of 4 bases or more with the longest homopolymer

being comprised of 35 T's were identified with this method. Similarly, we identified 37000 dinucleotide repeats and 16000 trinucleotide repeats. Matching the genomic position of 1 bp indels observed in *pms-2* and *pole-4; pms-2* mutants to the genomic positions of homopolymers we defined 1bp indels that occurred in homopolymer runs and the length of the homopolymer in which they occurred. Generalised additive models with a spline term were used to correlate the frequency of 1 bp indels occurring in homopolymer runs with the frequency of the respective homopolymer in the genome.

5.5 Comparison of *C elegans* mismatch repair mutation patterns to cancer signatures.

Variant calls for whole-exome sequencing data from the colorectal adenocarcinoma (COAD-US) and gastric adenocarcinoma (STAD-US) projects were taken from ICGC database (http://icgc.org). Mutational counts and contexts were inferred from variant tables using [60]. After combining indels and substitutions into vectors of length 104, we extracted the signatures using the Brunet NMF with KL-divergence as implemented in [57], which is equivalent to a Poisson NMF model. The number of signatures was chosen based on the saturation of both the Akaike Information Criterion (AIC) [61] and the residual sum of squares (RSS). The AIC is calculated as $AIC = 2k \cdot (n + N) - 2logL$, where k is the number of parameters (with n being the number of signatures and N being the number of samples), and L is the maximized model likelihood. As the L would naturally increase with the addition of parameters, we performed signature extraction for different ranks and chose the one where AIC and also RSS decrease slows down to avoid oversegmentation (Suppl. Figure 3A).

Similarity between the signatures was calculated via cosine similarity:

$$similarity(S_1, S_2) = \frac{< S_1, S_2 >}{||S_1|| \cdot ||S_2||},$$

where $< S_1, S_2 >$ is a scalar product of signature vectors. When compared to 96-long substitution signatures, indels were omitted from 104-long *de novo* signatures.

Stochastic nearest neighbor representation (*t*-SNE) [62] was obtained using R-package "tSNE" [63] using cosine similarity as distance measure between mutational profiles. In order to confirm the link between signatures MMR-1-3 and MMR deficiency, we determined samples carrying putative high-impact mutations (missense, frameshift or splice region variants, disruptive in-frame deletions or insertions, stop gained, start lost, and stop lost) in at least one of the MMR-related genes. RNA-seq gene expression data and methylation data for the two cancer cohorts were obtained from the ICGC database [46, 47]. Hypermethylation of *MLH1* was defined by correlating the methylation values of 43 CpG sites in the CpG island "chr3:37034228-37035356" overlapping the *MLH1* gene promoter with *MLH1* expression level. 28/43 CpG sites showed absolute correlation values higher than 0.5, correlation between expression counts and average value for these 28 CpG sites is –0.6 (Suppl. Figure 7A) (P-value <0.0001). The average methylation cut-off of 0.3 was chosen as a value well separating methylated and non-methylated cases based on expression value (Suppl. Figure 7B).

To extract the signatures of individual factors from respective *C. elegans* samples, we used additive Poisson model with multiple factors for every trinucleotide context and indel type (104 in total):

$$Y_j \sim \mathrm{Pois}(\lambda_j), \qquad j = 1, \dots, 104$$

$$E[Y_j] = \lambda_j = N\square(\beta_{j,b} + X_{g_1}\beta_{j,g_1} + X_{g_2}\beta_{j,g_2} + X_{g_1:g_2}\beta_{j,g_1:g_2}),$$

$$\beta_{j,\cdot} \geq 0,$$

where Y_j is a number of particular mutations happening in a given context (out of 104 types), N is the number of generations, and g_1 and g_2 – gene knock-outs, b - background contribution, and coefficient $X_{\ldots} \in \{0,1\}$ indicates the presence of a particular factor.

We calculated maximum likelihood estimates of all 104-long vectors $\beta_{j,\cdot}$ for every signature using an algorithm equivalent to non-negative matrix factorization with a fixed multiplier and KL-divergence as distance measure [64, 65].

For comparison of *C. elegans* and human mutational signatures, signatures acquired in *C. elegans* were adjusted by multiplying the probability for 96 base substitutions by the ratio of respective trinucleotide counts observed in the human exome (hg19, the counts pre-calculated in [66]) to those in the *C. elegans* reference genome (Figure 3B). COSMIC signatures were also adjusted to exome nucleotide counts as they were mostly derived from whole exomes [33, 34] and the comparison of *de novo* signatures to COSMIC is more valid in exome space. All signatures were further normalized so that the vector of probabilities sums up to 1.

Relative and absolute contributions of every signature to the samples from the united dataset were tested for association with MMR mutations using one-tailed Student t-test. All p-values were adjusted for multiple testing correction using Bonferroni procedure.

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust COMSIG consortium grant RG70175 and by a Wellcome Trust Senior Research award to AG (090944/Z/09/Z). Pieta Schofield and the Data Analysis Group, Dundee, were funded by the "Wellcome Trust Technology Platform " Strategic Award 097945/Z/11/Z. We would like to thank Steve Hubbard for help with statistics. Sequencing data are available at the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP020555.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

REFERENCES

[1] C.E. Bronner, S.M. Baker, P.T. Morrison, G. Warren, L.G. Smith, M.K. Lescoe, M. Kane, C. Earabino, J. Lipford, A. Lindblom, et al., Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer, Nature, 368 (1994) 258-261.
[2] R. Fishel, M.K. Lescoe, M.R. Rao, N.G. Copeland, N.A. Jenkins, J. Garber, M. Kane, R. Kolodner, The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer, Cell, 75 (1993) 1027-1038.
[3] M. Miyaki, M. Konishi, K. Tanaka, R. Kikuchi-Yanoshita, M. Muraoka, M. Yasuno, T. Igari, M. Koike, M. Chiba, T. Mori, Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer, Nature genetics, 17 (1997) 271-272.
[4] N.C. Nicolaides, N. Papadopoulos, B. Liu, Y.F. Wei, K.C. Carter, S.M. Ruben, C.A. Rosen, W.A. Haseltine, R.D. Fleischmann, C.M. Fraser, et al., Mutations of two PMS homologues in hereditary nonpolyposis colon cancer, Nature, 371 (1994) 75-80.
[5] N. Papadopoulos, N.C. Nicolaides, Y.F. Wei, S.M. Ruben, K.C. Carter, C.A. Rosen, W.A. Haseltine, R.D. Fleischmann, C.M. Fraser, M.D. Adams, et al., Mutation of a mutL homolog in hereditary colon cancer, Science, 263 (1994) 1625-1629.
[6] J. Genschel, S.J. Littman, J.T. Drummond, P. Modrich, Isolation of MutSbeta from human cells and comparison of the mismatch repair specificities of MutSbeta and MutSalpha, The Journal of biological chemistry, 273 (1998) 19895-19901.
[7] J.T. Drummond, G.M. Li, M.J. Longley, P. Modrich, Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells, Science, 268 (1995) 1909-1912.
[8] Y. Habraken, P. Sung, L. Prakash, S. Prakash, Binding of insertion/deletion DNA mismatches by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3, Current biology : CB, 6 (1996) 1185-1187.
[9] D.J. Allen, A. Makhov, M. Grilley, J. Taylor, R. Thresher, P. Modrich, J.D. Griffith, MutS mediates heteroduplex loop formation by a translocation mechanism, The EMBO journal, 16 (1997) 4467-4476.
[10] S. Gradia, D. Subramanian, T. Wilson, S. Acharya, A. Makhov, J. Griffith, R. Fishel, hMSH2-hMSH6 forms a hydrolysis-independent sliding clamp on mismatched DNA, Molecular cell, 3 (1999) 255-261.
[11] F.A. Kadyrov, L. Dzantiev, N. Constantin, P. Modrich, Endonucleolytic function of MutLalpha in human mismatch repair, Cell, 126 (2006) 297-308.
[12] F.A. Kadyrov, S.F. Holmes, M.E. Arana, O.A. Lukianova, M. O'Donnell, T.A. Kunkel, P. Modrich, Saccharomyces cerevisiae MutLalpha is a mismatch repair endonuclease, The Journal of biological chemistry, 282 (2007) 37181-37190.
[13] J. Genschel, P. Modrich, Mechanism of 5'-directed excision in human mismatch repair, Molecular cell, 12 (2003) 1077-1086.
[14] J. Genschel, L.R. Bazemore, P. Modrich, Human exonuclease I is required for 5' and 3' mismatch repair, The Journal of biological chemistry, 277 (2002) 13302-13311.
[15] E.M. Goellner, C.D. Putnam, R.D. Kolodner, Exonuclease 1-dependent and independent mismatch repair, DNA repair, 32 (2015) 24-32.
[16] P. Szankasi, G.R. Smith, A role for exonuclease I from S. pombe in mutation avoidance and mismatch correction, Science, 267 (1995) 1166-1169.
[17] D.X. Tishkoff, A.L. Boerger, P. Bertrand, N. Filosi, G.M. Gaida, M.F. Kane, R.D. Kolodner, Identification and characterization of Saccharomyces cerevisiae EXO1, a

gene encoding an exonuclease that interacts with MSH2, Proceedings of the National Academy of Sciences of the United States of America, 94 (1997) 7487-7492.
[18] M.J. Longley, A.J. Pierce, P. Modrich, DNA polymerase delta is required for human mismatch repair in vitro, The Journal of biological chemistry, 272 (1997) 10917-10921.
[19] N. Constantin, L. Dzantiev, F.A. Kadyrov, P. Modrich, Human mismatch repair: reconstitution of a nick-directed bidirectional reaction, The Journal of biological chemistry, 280 (2005) 39752-39761.
[20] T.A. Prolla, S.M. Baker, A.C. Harris, J.L. Tsao, X. Yao, C.E. Bronner, B. Zheng, M. Gordon, J. Reneker, N. Arnheim, D. Shibata, A. Bradley, R.M. Liskay, Tumour susceptibility and spontaneous mutation in mice deficient in Mlh1, Pms1 and Pms2 DNA mismatch repair, Nature genetics, 18 (1998) 276-279.
[21] E. Cannavo, G. Marra, J. Sabates-Bellver, M. Menigatti, S.M. Lipkin, F. Fischer, P. Cejka, J. Jiricny, Expression of the MutL homologue hMLH3 in human cells and its role in DNA mismatch repair, Cancer research, 65 (2005) 10759-10766.
[22] N.P. Bhattacharyya, A. Skandalis, A. Ganesh, J. Groden, M. Meuth, Mutator phenotypes in human colorectal carcinoma cell lines, Proceedings of the National Academy of Sciences of the United States of America, 91 (1994) 6319-6323.
[23] M.G. Hanford, B.C. Rushton, L.C. Gowen, R.A. Farber, Microsatellite mutation rates in cancer cell lines deficient or proficient in mismatch repair, Oncogene, 16 (1998) 2389-2393.
[24] M. Strand, T.A. Prolla, R.M. Liskay, T.D. Petes, Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair, Nature, 365 (1993) 274-276.
[25] Y. Yang, R. Karthikeyan, S.E. Mack, E.J. Vonarx, B.A. Kunz, Analysis of yeast pms1, msh2, and mlh1 mutators points to differences in mismatch correction efficiencies between prokaryotic and eukaryotic cells, Molecular & general genetics : MGG, 261 (1999) 777-787.
[26] N.P. Degtyareva, P. Greenwell, E.R. Hofmann, M.O. Hengartner, L. Zhang, J.G. Culotti, T.D. Petes, Caenorhabditis elegans DNA mismatch repair gene msh-2 is required for microsatellite stability and maintenance of genome integrity, Proceedings of the National Academy of Sciences of the United States of America, 99 (2002) 2158-2163.
[27] M. Tijsterman, J. Pothof, R.H. Plasterk, Frequent germline mutations and somatic repeat instability in DNA mismatch-repair-deficient Caenorhabditis elegans, Genetics, 161 (2002) 651-660.
[28] D.R. Denver, S. Feinberg, S. Estes, W.K. Thomas, M. Lynch, Mutation rates, spectra and hotspots in mismatch repair-deficient Caenorhabditis elegans, Genetics, 170 (2005) 107-113.
[29] S.A. Lujan, A.B. Clark, T.A. Kunkel, Differences in genome-wide repeat sequence instability conferred by proofreading and mismatch repair defects, Nucleic acids research, 43 (2015) 4067-4074.
[30] S.A. Lujan, A.R. Clausen, A.B. Clark, H.K. MacAlpine, D.M. MacAlpine, E.P. Malc, P.A. Mieczkowski, A.B. Burkholder, D.C. Fargo, D.A. Gordenin, T.A. Kunkel, Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition, Genome research, 24 (2014) 1751-1764.
[31] G.I. Lang, L. Parsons, A.E. Gammie, Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast, G3, 3 (2013) 1453-1465.

[32] A. Shlien, B.B. Campbell, R. de Borja, L.B. Alexandrov, D. Merico, D. Wedge, P. Van Loo, P.S. Tarpey, P. Coupland, S. Behjati, A. Pollett, T. Lipman, A. Heidari, S. Deshmukh, N. Avitzur, B. Meier, M. Gerstung, Y. Hong, D.M. Merino, M. Ramakrishna, M. Remke, R. Arnold, G.B. Panigrahi, N.P. Thakkar, K.P. Hodel, E.E. Henninger, A.Y. Goksenin, D. Bakry, G.S. Charames, H. Druker, J. Lerner-Ellis, M. Mistry, R. Dvir, R. Grant, R. Elhasid, R. Farah, G.P. Taylor, P.C. Nathan, S. Alexander, S. Ben-Shachar, S.C. Ling, S. Gallinger, S. Constantini, P. Dirks, A. Huang, S.W. Scherer, R.G. Grundy, C. Durno, M. Aronson, A. Gartner, M.S. Meyn, M.D. Taylor, Z.F. Pursell, C.E. Pearson, D. Malkin, P.A. Futreal, M.R. Stratton, E. Bouffet, C. Hawkins, P.J. Campbell, U. Tabori, C. Biallelic Mismatch Repair Deficiency, Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers, Nature genetics, 47 (2015) 257-262.
[33] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A. Aparicio, S. Behjati, A.V. Biankin, G.R. Bignell, N. Bolli, A. Borg, A.L. Borresen-Dale, S. Boyault, B. Burkhardt, A.P. Butler, C. Caldas, H.R. Davies, C. Desmedt, R. Eils, J.E. Eyfjord, J.A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinsk, N. Jager, D.T. Jones, D. Jones, S. Knappskog, M. Kool, S.R. Lakhani, C. Lopez-Otin, S. Martin, N.C. Munshi, H. Nakamura, P.A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J.V. Pearson, X.S. Puente, K. Raine, M. Ramakrishna, A.L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T.N. Schumacher, P.N. Span, J.W. Teague, Y. Totoki, A.N. Tutt, R. Valdes-Mas, M.M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L.R. Yates, J. Zucman-Rossi, P. Andrew Futreal, U. McDermott, P. Lichter, M. Meyerson, S.M. Grimmond, R. Siebert, E. Campo, T. Shibata, S.M. Pfister, P.J. Campbell, M.R. Stratton, Signatures of mutational processes in human cancer, Nature, (2013).
[34] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, P.J. Campbell, M.R. Stratton, Deciphering signatures of mutational processes operative in human cancer, Cell reports, 3 (2013) 246-259.
[35] B. Meier, S.L. Cooke, J. Weiss, A.P. Bailly, L.B. Alexandrov, J. Marshall, K. Raine, M. Maddison, E. Anderson, M.R. Stratton, A. Gartner, P.J. Campbell, C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency, Genome research, 24 (2014) 1624-1636.
[36] X. Yao, A.B. Buermeyer, L. Narayanan, D. Tran, S.M. Baker, T.A. Prolla, P.M. Glazer, R.M. Liskay, N. Arnheim, Different mutator phenotypes in Mlh1- versus Pms2-deficient mice, Proceedings of the National Academy of Sciences of the United States of America, 96 (1999) 6850-6855.
[37] A. Baross-Francis, N. Makhani, R.M. Liskay, F.R. Jirik, Elevated mutant frequencies and increased C : G-->T : A transitions in Mlh1-/- versus Pms2-/- murine small intestinal epithelial cells, Oncogene, 20 (2001) 619-625.
[38] A. Aksenova, K. Volkov, J. Maceluch, Z.F. Pursell, I.B. Rogozin, T.A. Kunkel, Y.I. Pavlov, E. Johansson, Mismatch repair-independent increase in spontaneous mutagenesis in yeast lacking non-essential subunits of DNA polymerase epsilon, PLoS genetics, 6 (2010) e1001209.
[39] S.A. Lujan, J.S. Williams, Z.F. Pursell, A.A. Abdulovic-Cui, A.B. Clark, S.A. Nick McElhinny, T.A. Kunkel, Mismatch repair balances leading and lagging strand DNA replication fidelity, PLoS genetics, 8 (2012) e1003016.
[40] C. Palles, J.B. Cazier, K.M. Howarth, E. Domingo, A.M. Jones, P. Broderick, Z. Kemp, S.L. Spain, E. Guarino, I. Salguero, A. Sherborne, D. Chubb, L.G. Carvajal-Carmona, Y. Ma, K. Kaur, S. Dobbins, E. Barclay, M. Gorman, L. Martin, M.B. Kovac, S. Humphray, C. Consortium, W.G.S. Consortium, A. Lucassen, C.C. Holmes, D.

Bentley, P. Donnelly, J. Taylor, C. Petridis, R. Roylance, E.J. Sawyer, D.J. Kerr, S. Clark, J. Grimes, S.E. Kearsey, H.J. Thomas, G. McVean, R.S. Houlston, I. Tomlinson, Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas, Nature genetics, 45 (2013) 136-144.
[41] T.M. Albertson, M. Ogawa, J.M. Bugni, L.E. Hays, Y. Chen, Y. Wang, P.M. Treuting, J.A. Heddle, R.E. Goldsby, B.D. Preston, DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice, Proceedings of the National Academy of Sciences of the United States of America, 106 (2009) 17101-17104.
[42] F. Supek, B. Lehner, Differential DNA mismatch repair underlies mutation rate variation across the human genome, Nature, 521 (2015) 81-84.
[43] B. Harr, B. Zangerl, C. Schlotterer, Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from Drosophila, Molecular biology and evolution, 17 (2000) 1001-1009.
[44] J.R. Brouwer, R. Willemsen, B.A. Oostra, Microsatellite repeat instability and neurological disease, Bioessays, 31 (2009) 71-83.
[45] D.R. Denver, K. Morris, A. Kewalramani, K.E. Harris, A. Chow, S. Estes, M. Lynch, W.K. Thomas, Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of Caenorhabditis elegans, Journal of molecular evolution, 58 (2004) 584-595.
[46] N. Cancer Genome Atlas, Comprehensive molecular characterization of human colon and rectal cancer, Nature, 487 (2012) 330-337.
[47] N. Cancer Genome Atlas Research, Comprehensive molecular characterization of gastric adenocarcinoma, Nature, 513 (2014) 202-209.
[48] P. Hsieh, K. Yamane, DNA mismatch repair: molecular mechanism, cancer, and ageing, Mech Ageing Dev, 129 (2008) 391-407.
[49] M.J. Ligtenberg, R.P. Kuiper, T.L. Chan, M. Goossens, K.M. Hebeda, M. Voorendt, T.Y. Lee, D. Bodmer, E. Hoenselaar, S.J. Hendriks-Cornelissen, W.Y. Tsui, C.K. Kong, H.G. Brunner, A.G. van Kessel, S.T. Yuen, J.H. van Krieken, S.Y. Leung, N. Hoogerbrugge, Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1, Nature genetics, 41 (2009) 112-117.
[50] A. Bellacosa, Functional interactions and signaling properties of mammalian DNA mismatch repair proteins, Cell Death Differ, 8 (2001) 1076-1092.
[51] I. Grin, A.A. Ishchenko, An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation, Nucleic acids research, 44 (2016) 3713-3727.
[52] R. Tricarico, S. Cortellino, A. Riccio, S. Jagmohan-Changur, H. Van der Klift, J. Wijnen, D. Turner, A. Ventura, V. Rovella, A. Percesepe, E. Lucci-Cordisco, P. Radice, L. Bertario, M. Pedroni, M. Ponz de Leon, P. Mancuso, K. Devarajan, K.Q. Cai, A.J. Klein-Szanto, G. Neri, P. Moller, A. Viel, M. Genuardi, R. Fodde, A. Bellacosa, Involvement of MBD4 inactivation in mismatch repair-deficient tumorigenesis, Oncotarget, 6 (2015) 42892-42904.
[53] E.L. Greer, M.A. Blanco, L. Gu, E. Sendinc, J. Liu, D. Aristizabal-Corrales, C.H. Hsu, L. Aravind, C. He, Y. Shi, DNA Methylation on N6-Adenine in C. elegans, Cell, 161 (2015) 868-878.
[54] K. Ye, M.H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads, Bioinformatics, 25 (2009) 2865-2871.

[55] P.J. Stephens, P.S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D.C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G.R. Bignell, L.R. Yates, E. Papaemmanuil, D. Beare, A. Butler, A. Cheverton, J. Gamble, J. Hinton, M. Jia, A. Jayakumar, D. Jones, C. Latimer, K.W. Lau, S. McLaren, D.J. McBride, A. Menzies, L. Mudie, K. Raine, R. Rad, M.S. Chapman, J. Teague, D. Easton, A. Langerod, M.T. Lee, C.Y. Shen, B.T. Tee, B.W. Huimin, A. Broeks, A.C. Vargas, G. Turashvili, J. Martens, A. Fatima, P. Miron, S.F. Chin, G. Thomas, S. Boyault, O. Mariani, S.R. Lakhani, M. van de Vijver, L. van 't Veer, J. Foekens, C. Desmedt, C. Sotiriou, A. Tutt, C. Caldas, J.S. Reis-Filho, S.A. Aparicio, A.V. Salomon, A.L. Borresen-Dale, A.L. Richardson, P.J. Campbell, P.A. Futreal, M.R. Stratton, The landscape of cancer genes and mutational processes in breast cancer, Nature, 486 (2012) 400-404.
[56] M. Gerstung, C. Beisel, M. Rechsteiner, P. Wild, P. Schraml, H. Moch, N. Beerenwinkel, Reliable detection of subclonal single-nucleotide variants in tumour cell populations, Nature communications, 3 (2012) 811.
[57] M. Gerstung, E. Papaemmanuil, P.J. Campbell, Subclonal variant calling with multiple samples and prior knowledge, Bioinformatics, 30 (2014) 1198-1204.
[58] M. Lawrence, W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M.T. Morgan, V.J. Carey, Software for computing and annotating genomic ranges, PLoS Comput Biol, 9 (2013) e1003118.
[59] H. Pagès, P. Aboyoun, R. Gentleman, S. DebRoy, Biostrings: String objects representing biological sequences, and matching algorithms., R package version, 2.42.1 (2016).
[60] F. Blokzijl, R. Janssen, R. Van Boxtel, E. Cuppen, MutationalPatterns: an integrative R package for studying patterns in base substitution catalogues, bioRxiv, (2016).
[61] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, 2nd International Symposium of Information Theory, Akademiai Kiado, Budapest, (1973) 267-281.
[62] L. van der Maaten, G. Hinton, Visualizing high-dimensional data using t-SNE, Journal of Machine Learning Research, 9 (2008) 2579-2605.
[63] J. Donaldson, t-SNE: T-Distributed Stochastic Neighbor Embedding for R, R package (2016) version 0.1-3.
[64] D.D. Lee, H.S. Seung, Algorithms for Non-negative Matrix Factorization, Advances in Neural Information Processing, 13 (2000).
[65] A.T. Cemgil, Bayesian Inference for Nonnegative Matrix Factorisation Models, Computational Intelligence and Neuroscience, 2009 (2009) 17 pages.
[66] R. Rosenthal, N. McGranahan, J. Herrero, B.S. Taylor, C. Swanton, DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution, Genome Biol, 17 (2016) 31.
[67] J.T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nature biotechnology, 29 (2011) 24-26.

Figure 1

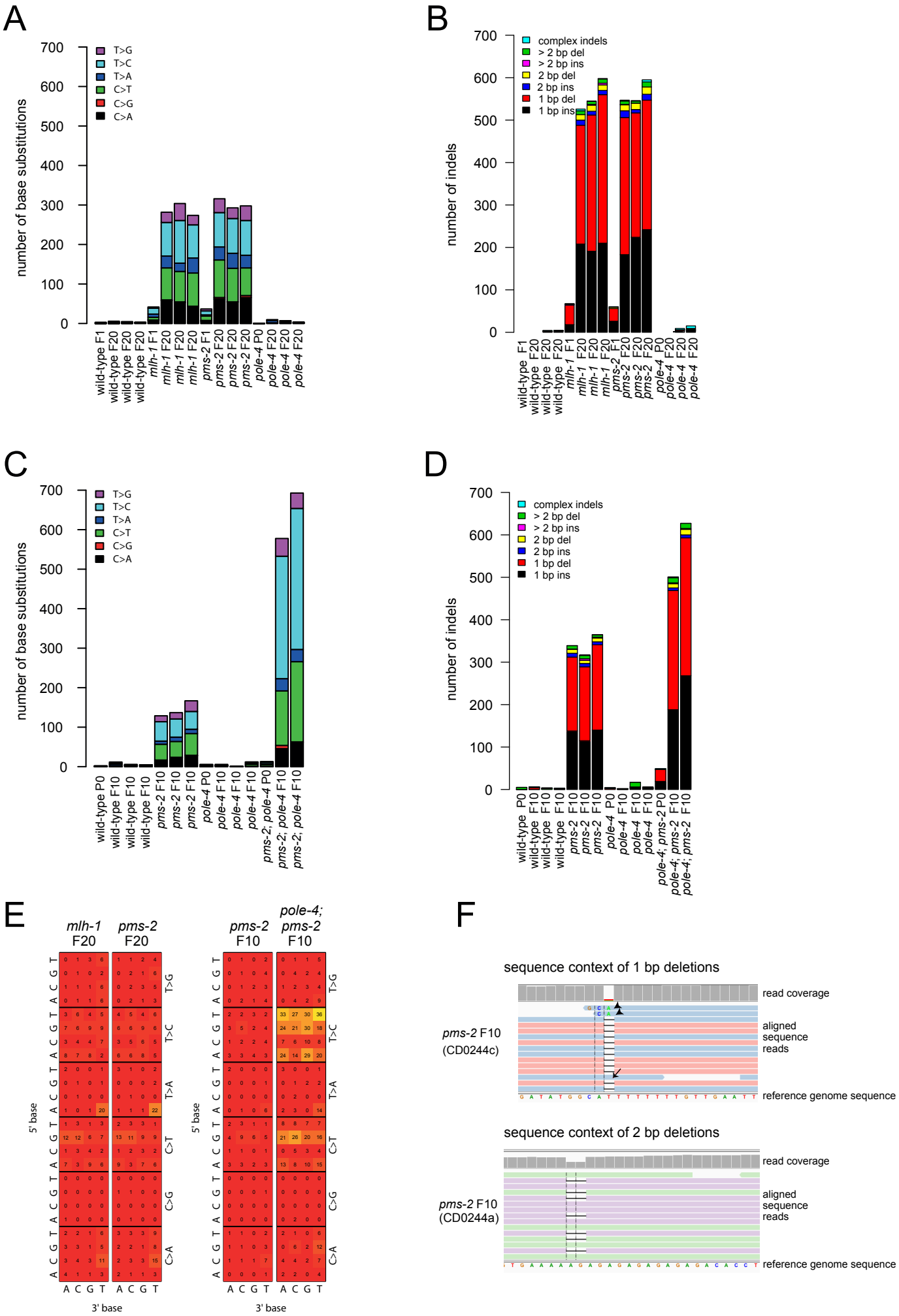

A
number of base substitutions
T>G
T>C
T>A
C>T
C>G
C>A
wild-type F1
wild-type F20
mlh-1 F1
mlh-1 F20
pms-2 F1
pms-2 F20
pole-4 P0
pole-4 F20
B
number of indels
complex indels
> 2 bp del
> 2 bp ins
2 bp del
2 bp ins
1 bp del
1 bp ins
C
wild-type P0
wild-type F10
pms-2 F10
pole-4 P0
pole-4 F10
pms-2; pole-4 P0
pms-2; pole-4 F10
D
pms-2 P0
pole-4 F10
pms-2 F10
pole-4; pms-2 P0
pole-4; pms-2 F10
E
mlh-1 F20
pms-2 F20
pms-2 F10
pole-4; pms-2 F10
5' base
3' base
F
sequence context of 1 bp deletions
pms-2 F10 (CD0244c)
read coverage
aligned sequence reads
reference genome sequence
sequence context of 2 bp deletions
pms-2 F10 (CD0244a)

Figure 1. Mutations in *C. elegans* wild-type and mutant strains grown for 10 or 20 generations. Identical base substitutions as well as indels occurring in the same genomic location among samples of the entire dataset (duplicates) were excluded from the analysis, thus only reporting mutations unique to each individual sample. **(A)** Number and types of base substitutions acquired in wild-type, *mlh-1*, *pms-2,* and *pole-4* single mutants during propagation for 20 generations. Shown are mutations in the initial parental (P0) or one first generation (F1) line and 3 independently propagated F20 lines per genotype. **(B)** Number and types of indels acquired in wild-type, *mlh-1*, *pms-2,* and *pole-4* single mutants during propagation for 20 generations. Shown are indels identified in the initial parental (P0) or F1 generation and in 3 independently propagated F20 lines per genotype. **(C)** Base substitution number and types acquired in wild-type, *pms-2*, *pole-4*, and *pole-4; pms-2* mutants during propagation for 10 generations. Mutations in the initial parental (P0) line and 2-3 independently propagated F10 lines for each genotype are shown. **(D)** Number and type of indels acquired in wild-type, *pms-2 and pole-4* single, and *pole-4; pms-2* double mutants during propagation for 10 generations. **(E)** Heat map of frequency of the six possible base substitutions in their 5' and 3' sequence context observed in *pms-2* and *mlh-1* single mutants after propagation for 20 generations and in *pms-2* single and *pole-4*; *pms-2* double mutants after propagation for 10 generations. Number and types of mutations are shown as mean mutations identified across all individual lines of each genotype. **(F)** Examples of sequence contexts around one and two base indels. Sequence reads aligned to the reference genome WBcel235.74 are visualized in Integrative Genomics Viewer [67]. A one base pair deletion (top panel) and a two base pair deletion (bottom panel) are shown. A subset of sequence reads, which end close to an indel, is erroneously aligned across the indel resulting either in wild-type bases (arrows) or base changes (arrowhead). Such wrongly called base

substitutions were removed during filtering using the deepSNV package [56, 57] (see Material and Methods).

Figure 2

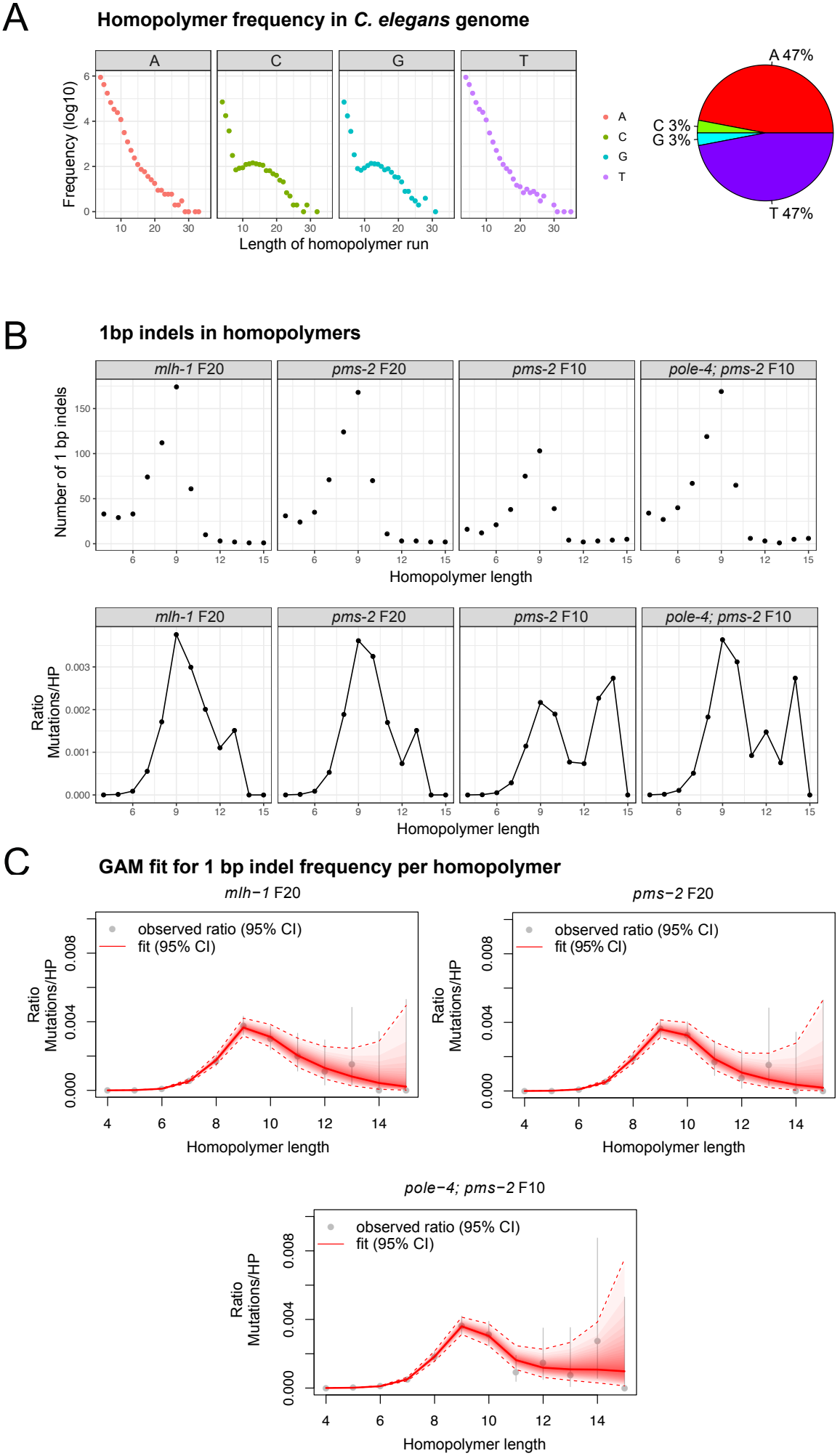

A
Homopolymer frequency in C. elegans genome
A
C
G
T
Frequency (log10)
Length of homopolymer run
A 47%
C 3%
G 3%
T 47%
B
1bp indels in homopolymers
mlh-1 F20
pms-2 F20
pms-2 F10
pole-4; pms-2 F10
Number of 1 bp indels
Homopolymer length
Ratio Mutations/HP
C
GAM fit for 1 bp indel frequency per homopolymer
mlh−1 F20
pms−2 F20
pole−4; pms−2 F10
observed ratio (95% CI)
fit (95% CI)
Ratio Mutations/HP
Homopolymer length

Figure 2. Correlation between homopolymer length and the frequency of +1/-1 bp indels. **(A)** Distribution of homopolymer repeats encoded in the *C. elegans* genome by length and DNA base shown in $\log_{10}$ scale (left panel) and the relative percentage of A, C, G and T homopolymers in the genome (right panel). **(B)** Average number of 1 bp indels in homopolymer runs (top panel) and averaged frequency of 1 bp indels per homopolymer in relation to homopolymer length (bottom panel) in MMR mutant backgrounds indicated. **(C)** Generalized additive spline model (GAM) fit for the ratio of 1bp indels normalized to the frequency of homopolymers in the genome. The average frequency observed across three lines is depicted as a grey dot; grey bars indicate the 95% confidence interval. The red line indicates best-fit. Red dotted lines represent the corresponding 95% confidence interval.

Figure 3

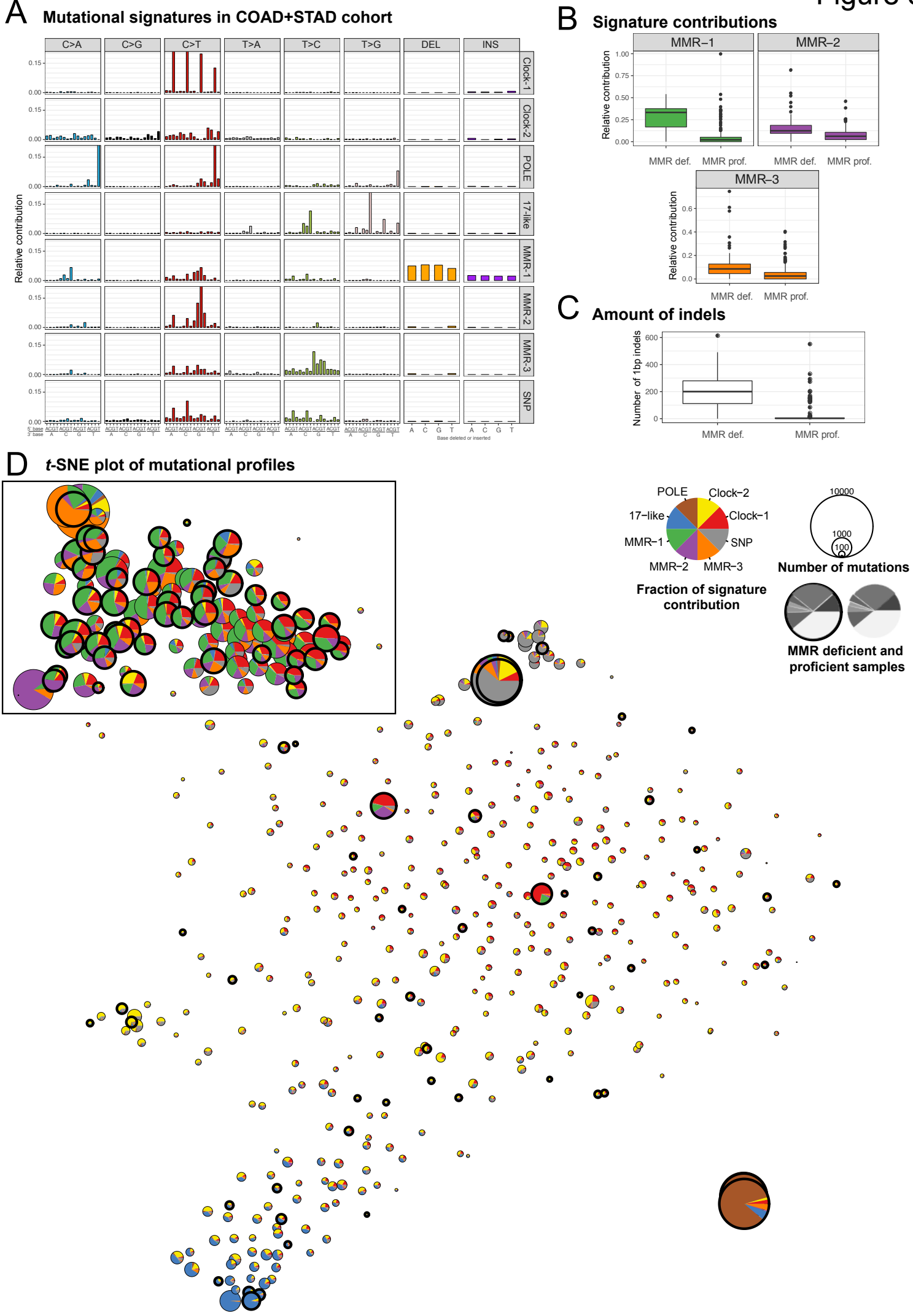
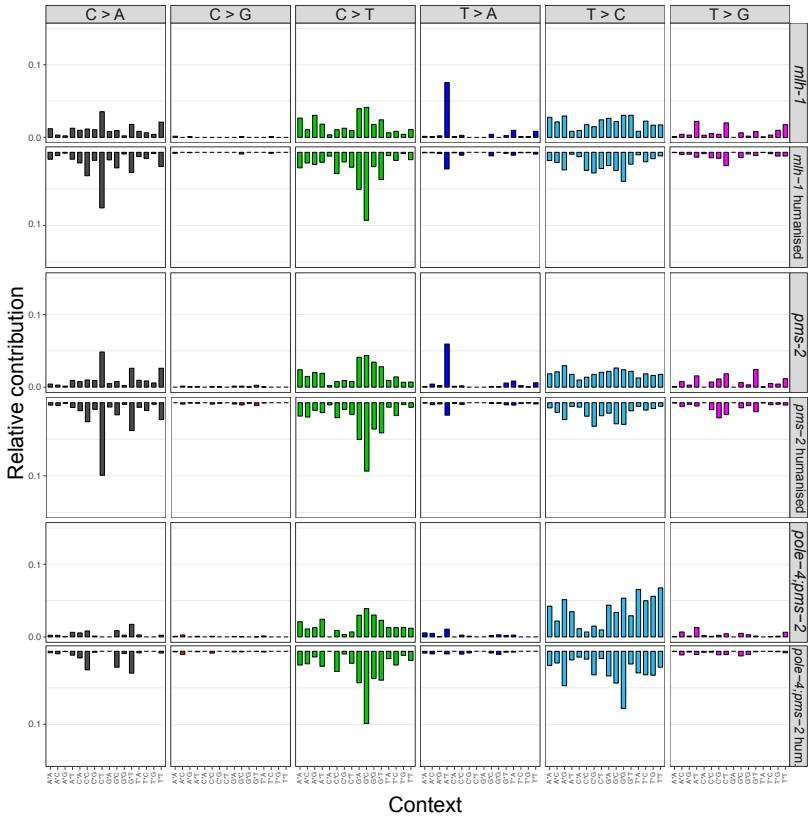

A
Mutational signatures in COAD+STAD cohort
C>A
C>G
C>T
T>A
T>C
T>G
DEL
INS
Clock-1
Clock-2
POLE
17-like
MMR-1
MMR-2
MMR-3
SNP
Relative contribution
Base deleted or inserted
B
Signature contributions
MMR−1
MMR−2
MMR−3
Relative contribution
MMR def.
MMR prof.
C
Amount of indels
Number of 1bp indels
MMR def.
MMR prof.
D
t-SNE plot of mutational profiles
POLE
Clock−2
17−like
Clock−1
MMR−1
SNP
MMR−2
MMR−3
Fraction of signature contribution
10000
1000
100
Number of mutations
MMR deficient and proficient samples

Figure 3. *De novo* identification of signatures from human colorectal and gastric adenocarcinoma projects (COAD-US and STAD-US) samples, contribution of *de novo* signatures to samples with and without MMR gene inactivation. (A) Mutational signatures including base substitutions and 1bp indels derived from the combination of COAD-US and STAD-US datasets. **(B)** Contribution of MMR-1-3 signatures to cancer samples with and without defects (mutations or *MLH1* promoter hypermethylation) in DNA mismatch repair genes. Box plot with outliers shown as individual filled circles. **(C)** Number of 1bp indels in COAD-US and STAD-US samples with and without somatic (epi)mutations in DNA mismatch repair genes. Box plot with outliers shown as individual filled circles. **(D)** Two-dimensional representation of COAD and STAD mutational profiles. The size of each circle reflects the mutation burden. Predicted MMR deficiency is highlighted by a black outline, while the color of segments reflects signature decomposition of each sample. The black frame shows the cluster of MMR deficient samples.

Figure 4

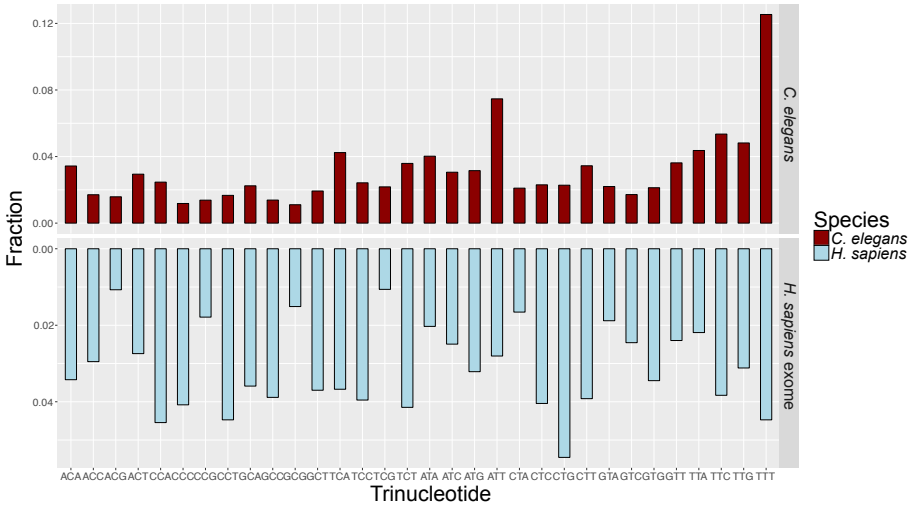
A

Mutational patterns from *C. elegans*



B

Trinucleotide context comparison



C

Comparison of MMR-1 and *C. elegans* MMR patterns

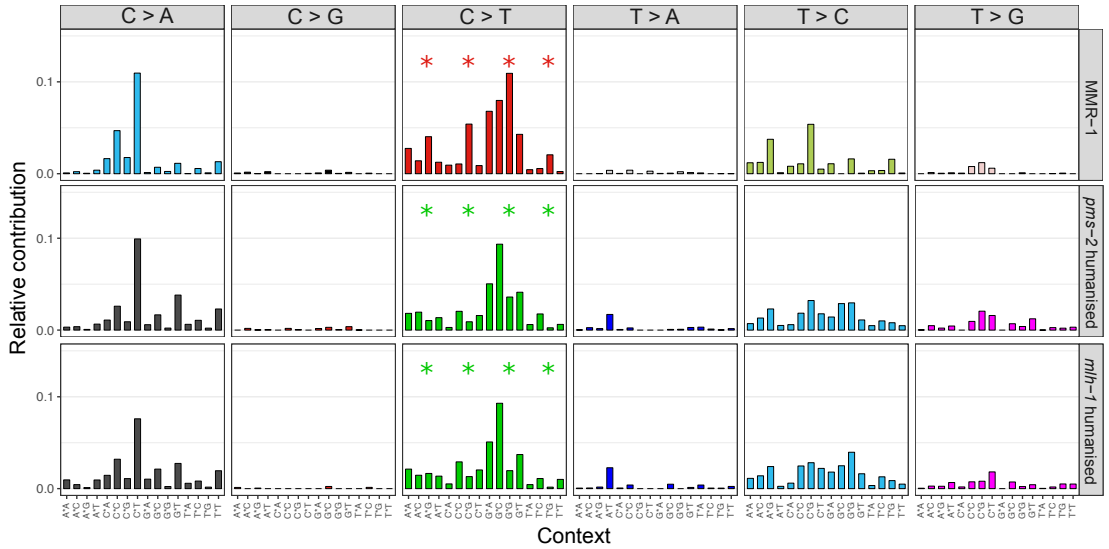


Figure 4. Mutational patterns derived from *C. elegans* MMR mutants and their comparison to *de novo* human signature MMR-1. (A) Base substitution patterns of *C. elegans mlh-1, pms-2* and *pole-4; pms-2* mutants and their corresponding humanized versions (mirrored). **(B)** Relative abundance of trinucleotides in the *C. elegans* genome (red) and the human exome (light blue). **(C)** MMR-1 base substitution signature versus *pms-2* and *mlh-1* mutational patterns, all adjusted to human whole exome trinucleotide frequency. Stars highlight the difference in C>T transitions at CpG sites, which occur at lower frequency in *C. elegans*.

TABLES

Table 1. MutS and MutL complexes in *E. coli*, *C. elegans* and *H. sapiens*.

	complexes	*Escherichia coli*	*Caenorhabditis elegans*	*Homo sapiens*
MutS	MutS*α* MutS*β*	MutS homodimer	MSH-2/ MSH-6 -	MSH2/ MSH6 MSH2/ MSH3
MutL	MutL*α* MutL*β* MutL*γ*	MutL homodimer	MLH-1/PMS-2 - -	MLH1/ PMS2 MLH1/ PMS1 MLH1/ MLH3

Table 2. Average number of unique mutations identified genome-wide in wild-type, *pms-2* and *pole-4* single mutants and *pole-4; pms-2* double mutants grown over 10 generations.

genotype	mutation type		average number observed	average # as % of all mutations
wild-type	base subs	transitions	3.7	
		transversions	6.3	
	indels	1 base insertions	0.3	
		1 base deletions	1.7	
		2 base insertions	0	
		2 base deletions	0.7	
		larger insertions	0	
		larger deletions	1.3	
		complex indels	0.3	
	SVs		0	
pms-2	base subs	transitions	91.3	18.8 %
		transversions	53.7	11.0 %
	indels	1 base insertions	131	26.9 %
		1 base deletions	183	37.6 %
		2 base insertions	8	1.6 %
		2 base deletions	8.3	1.7 %
		larger insertions	3.3	0.7 %
		larger deletions	7	1.4 %
		complex indels	0.7	0.1 %
	SVs		0	0 %
pole-4	base subs	transitions	6	
		transversions	9.3	
	indels	1 base insertions	1	
		1 base deletions	1	
		2 base insertions	0.7	
		2 base deletions	0.7	
		larger insertions	0	
		larger deletions	4.7	
		complex indels	0.3	
	SVs		0.3	
pole-4; pms-2	base subs	transitions	502	41.8 %
		transversions	134.5	11.2 %
	indels	1 base insertions	228	19.0 %
		1 base deletions	303	25.2 %
		2 base insertions	6.5	0.5 %
		2 base deletions	11.5	1.0 %
		larger insertions	2	0.2 %
		larger deletions	12	1.0 %
		complex indels	1	0.1 %
	SVs		0	0 %

base subs = base substitutions
indels = insertions/deletions
SVs = structural variants
Values shown in % are rounded to one digit after the decimal point.

1 **Table 3. Cosine similarity values for the comparison between humanized *C. elegans* derived MMR mutation patterns with *de novo***

2 **signatures and selected COSMIC signatures (both adjusted to human whole-exome trinucleotide frequencies)**.

3

		C. elegans patterns			COSMIC signatures										
					1	2	3	5	6	10	15	17	20	21	26
					mlh-1	0.28	0.09	0.56	0.71	0.44	0.16	0.44	0.27	0.77	0.56	0.63
					pms-2	0.28	0.09	0.53	0.68	0.48	0.18	0.50	0.25	0.77	0.48	0.61
					pole-4; pms-2	0.26	0.12	0.45	0.68	0.40	0.13	0.50	0.17	0.56	0.69	0.74
		mlh-1	*pms-2*	*pole-4; pms-2*											
De novo signatures	Clock 1	0.18	0.17	0.15		0.98	0.11	0.14	0.53	0.86	0.36	0.52	0.20	0.53	0.42	0.46
	Clock 2	0.36	0.32	0.30		0.24	0.62	0.75	0.66	0.18	0.13	0.12	0.19	0.31	0.19	0.27
	POLE	0.19	0.21	0.15		0.32	0.26	0.19	0.31	0.24	0.97	0.18	0.18	0.24	0.17	0.23
	17-like	0.23	0.21	0.13		0.06	0.04	0.22	0.26	0.08	0.06	0.06	0.98	0.12	0.12	0.20
	MMR-1	0.81	0.85	0.63		0.59	0.08	0.43	0.71	0.79	0.24	0.70	0.20	0.85	0.43	0.65
	MMR-2	0.36	0.43	0.42		0.67	0.04	0.14	0.49	0.86	0.21	0.99	0.12	0.30	0.22	0.49
	MMR-3	0.62	0.56	0.75		0.39	0.09	0.34	0.62	0.42	0.16	0.35	0.18	0.53	0.95	0.91
	SNP	0.52	0.47	0.55		0.73	0.14	0.48	0.83	0.69	0.25	0.38	0.29	0.79	0.63	0.71

4