“The replication stress response suppresses mutation rates in mismatch repair deficient budding yeast and human cancers”

Amruta Shrikhande[1*], Simon G. Sørensen[2,3*], Judith Offman[4], Barbara Domanska[5], Jakob S. Pedersen[2,3,6**] and Eva R. Hoffmann[1**]

[1]DNRF Center for Chromosome Stability, Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

[2]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Palle Juul-Jensens Boulevard 99, 8200, Aarhus N, Denmark.

[3]Department of Clinical Medicine, Aarhus University, Palle Juul-Jensens Boulevard 82, 8200, Aarhus N, Denmark.

[4]Comprehensive Cancer Centre, School of Cancer and Pharmaceutical Sciences, Faculty of Life Sciences & Medicine, King's College London, London, UK.

[5]Cliniphy, Bellingham House, 2 Huntingdon Street, St. Neots, Cambridgeshire PE19 1BG, United Kingdom.

[6]Bioinformatics Research Center (BiRC), Aarhus University, 8000, Aarhus C, Denmark.

*These authors contributed equally to the work.

**Corresponding authors: eva@sund.ku.dk, jakob.skou@clin.au.dk

Lead Contact: eva@sund.ku.dk

Abstract

Elevated mutation rate is a hallmark of mismatch repair (MMR) deficient cells and tumours. This includes microsatellite instability (MSI), which is caused by insertions and deletions in mono- and dinucleotide repeats. MSI rates, however, are highly variable across MMR-deficient tumours. Here we show that mutation rates are genetically regulated in MMR-deficient cells. A genome-wide deletion screen in budding yeast revealed that 3% gene deletions caused mutation rates to be further elevated, whereas 11% reduced mutation rates. The genes causing an elevation are enriched for DNA repair and replication processes, whereas deletion of genes implicated in transcriptional processes reduce mutation rates. A pan-cancer analysis of MSI revealed that mutations in replication stress response (*ATR*, *TOPBP1*, *CHEK1*) or DNA repair genes (*RAD50*, *TOP3A*) was associated with extreme rates of MSI in MMR-defective tumours, but not when mutated on their own. Since replication stress, DNA damage and repair activities are cell-type specific, this may account for highly variable mutation rates associated with different MMR-deficient tumours.

Keywords: Microsatellite instability, DNA damage response, Mismatch repair, Replication stress, Non-homologous end joining, Genomewide yeast screen, Immunotherapy, Synthetic lethality, Whole genome sequencing, Cancer

40 **INTRODUCTION**

41 Mismatch repair (MMR) is highly conserved across species and removes incorrectly inserted
42 bases and small insertion/deletions from the newly synthesised strand[1-3]. MMR is temporally
43 coupled to the replication fork and loss of MMR results in elevated mutation rates.
44 Microsatellites, especially mono- and dinucleotide repeats, are particularly sensitive to MMR
45 deficiency, resulting in insertions and deletions, referred to as microsatellite instability (MSI)[4].
46 MSI was originally associated with colorectal cancers in families with inherited mutations in
47 MMR genes (Lynch Syndrome)[5-8]. Subsequently, mono- and dinucleotide repeat panels were
48 used to detect MSI for diagnostic purposes of Lynch Syndrome[4]. Furthermore, the types of
49 tumours found in Lynch syndrome families and sporadic cancers defective for MMR also
50 expanded[9-12]. More recently, MSI is used as a biomarker for response to immunotherapy due to
51 high levels of neoantigen formation[13,14].
52
53 MMR-deficient tumours, however, are characterised by highly variable incidence of MSI across
54 tissue type. This ranges from nearly 90% of colorectal cancers to 0% of brain tumours[15,16]. The
55 tissue-specific MSI incidence, and mutation rates more generally, is poorly understood, but
56 suggests that both the formation and repair of mono- and dinucleotide repeats might have cell
57 type-specific origins. Most of our current knowledge of mutation rates comes from budding
58 yeast, where it is possible to conduct genome-wide deletion screens of non-essential genes.
59 Such studies revealed that DNA replication and repair processes regulate mutation rates[17,18].
60 Targeted co-mutations of MMR genes have revealed epistatic, additive, and synergistic effects
61 of a limited number of genes in yeast and human cells, however, the extent of genetic regulation
62 of mutation rates in MMR-deficient cells and human tumours remains unclear[19,20].
63
64 Here we used 4,847 yeast deletion strains and 6,057 human tumours to systematically explore
65 the regulatory mechanisms of mutation rate in MMR-deficient cells. In yeast, co-mutations in

66 genes affecting replication causes elevated mutation rates. Consistent with this, human tumours
67 defective for MMR with co-mutations in the cellular response to replication stress *(ATR*, *CHEK1*,
68 *TOPBP1*) are associated with extremely elevated mutation rates. However, mutations in
69 replication stress response genes alone are not associated with MSI. Our results suggest that a
70 combination of elevated replication stress and loss of MMR combined result in hypermutated
71 tumours in mono- and dinucleotide repeats.
72
73 **RESULTS**
74 **A systematic gene deletion screen for changed mutation rates in MMR-defective yeast**
75 **cells.**
76 Since DNA replication and repair pathways are highly conserved across species, we
77 investigated genome-wide regulation of mutation rates in yeast with MMR deficiency. In budding
78 yeast, cells defective or hypomorphic for core MMR genes (*yMSH2, yMLH1*, *yPMS2*),
79 accumulate errors by the replicative DNA polymerases (yPOL3/hPOLD and yPOL2/hPOLE).
80 This results in the accumulation of base substitutions and indels[21-24] that can be detected using
81 the *CAN1* reporter gene. *CAN1* is an arginine transporter, whose mutation results in resistance
82 to the toxic analogue L-canavanine. MMR-defective strains accumulate 66% single base
83 substitutions and 34% insertions/deletions in *CAN1*[25].Our dilution experiments showed that a 5-
84 fold change in mutation rate can be detected robustly by qualitative assessment in both wild
85 type and MMR-compromised yeast strains (**Supplementary Fig. S1**).
86
87 To explore the genetic landscape that regulates mutation rates in MMR-compromised cells, we
88 transformed 4,847 yeast deletion strains in non-essential genes with a low-copy number
89 plasmid (*CEN ARS*, *LEU2*) that expresses human MLH1 and screened for changed mutation
90 rates[26] (**Fig. 1a**). MLH1 forms a heterodimer with PMS1 in yeast (PMS2 in human; MutLα) that
91 cleaves the newly synthesised strand upon mismatch detection by MSH2-MSH3 (MutSα) or

92 MSH2-MSH6 (MutSβ) heterodimers. This facilitates removal of the incorrectly inserted base and
93 prevents mutation[27-29]. Expression of hMLH1 interferes with yeast MMR, since it forms non-
94 functional complexes with yeast PMS1, resulting in hypomorphic MMR and elevated mutation
95 rates[30-32].
96
97 Of the 4,847 deletion strains, 175 failed to grow on the selective media for plasmid uptake. 161
98 strains (3% of all strains) showed increased mutation rates while 542 (11%) displayed a
99 decrease in mutation rate compared to the wild-type strain expressing hMLH1 (**Fig. 1a**;
100 **Supplementary Tables S1-S4**). We detected important controls, including deletion of *CAN1*,
101 which is expected since all *can1*Δ cells grow on canavanine-containing medium. MMR gene
102 deletions were also detected (**Fig. 1b**). This is expected since hMLH1 renders yeast cells
103 hypomorphic for mismatch repair[32] and deletion *yMSH2*, *yMLH1*, or *yPMS1* further increase
104 mutation rates. Gene ontology analysis using DAVID[33,34] showed strong enrichment for DNA
105 repair and cellular responses to DNA damage in the strains where mutation rates were further
106 elevated (**Fig. 1b**). In contrast, processes involved in a wide range of cellular metabolism,
107 especially transcription, were enriched in strains that displayed decreased mutation rates (**Fig.**
108 **1c**). These observations suggest genetic regulation of mutation rates in MMR-defective cells.
109
110 **Deletion of transcriptional regulators decreases mutation rates in MMR-defective cells.**
111 The decrease in mutation rate in some of the deletion strains is consistent with the finding that
112 some tumour types do not display MSI even when MMR is lost in a sporadic or hereditary
113 manner[15,16]. However, to what extent the 542 genes are caused by plasmid loss and/or an
114 artefact of the assay (canavanine resistance) is unclear, especially since genes involved in
115 arginine biosynthetic processes were highly enriched (**Fig. 1c**). We performed western blots on
116 20 of the strains with decreased mutation rates. Although variable in expression, we did not
117 detect complete loss of hMLH1 protein (**Supplementary Fig. S2; Methods**). To validate directly

118 whether deletions of the genes identified could lower mutation rates when MMR is defective, we
119 randomly selected 348 of the deletion strains and crossed these to complete deletions of three
120 different MMR strains (*msh2*Δ*, mlh1*Δ, *pms1*Δ). 341 of the 348 (98%) of the selected double
121 mutants (*gene*Δ + *mmr*Δ) strains showed low mutation rates in combination with all three MMR
122 deletions (**Supplementary Fig. S3a)**. To further ascertain whether mutation rates can be
123 lowered independently of the *CAN1* reporter gene, we conducted further studies on four genes
124 involved in transcriptional regulation (*TFB5*, *SAC3*, *DEF1*, and *THO2*), because dysregulation of
125 transcription is implicated in replication stress and subsequent mutagenesis[35] (**Fig. 1c;**
126 **Supplementary Tables S1, S2; Methods**). Quantitative mutation rate assays of *CAN1* as well
127 as a *LYS2* template containing a homopolymeric A run (n = 14) revealed that co-mutation of
128 either of these four genes lowered mutation rates in a *mlh1*Δ strain (**Supplementary Fig.**
129 **S3b,c**). We also validated that these genes do not act by interfering with arginine metabolism or
130 by causing slow growth (**Supplementary Fig. S3d**). Dysregulation of transcription can trigger
131 transcription-coupled nucleotide excision repair. However, deletion of either *RAD26* (human
132 *CSB*) or *RAD10* (human *XPG*) did not affect mutation rates in the *mlh1* strain **(Supplementary**
133 **Fig. S3b)**. We infer that regulation of transcription elevates the endogenous mutation rates that
134 MMR has an important role in reducing. This further supports the notion mutation rates can be
135 low, despite MMR deficiency.
136
137 **Deletion of genes implicated in response to replication stress elevate mutation rates.**
138 For the 161 deletion strains with elevated mutation rates, 16 had DNA repair and damage
139 response (3.6-fold enrichment, Fisher's Exact test, $p<1.7x10^{-4}$, one-tailed, **Fig. 1b;**
140 **Supplementary Tables S3, S4**). 15 of the 16 identified genes showed elevated mutation rates
141 when hMLH1 was expressed in quantitative mutation rate assays (**Fig. 1d**). Amongst the 15
142 genes, 11 showed elevated mutation rates above wild-type cells expressing hMLH1. This
143 included *rad27*Δ, which was previously shown to elevate mutation rates in MMR-defective

144 yeast[25,36], as well as novel genes including components of the SHU complex, *ELG1*, *SLX4*,
145 *HUG1*, and *TRM2*, all of which have functions in response to DNA replication stress[37-42] (**Fig.**
146 **1d**; **Supplementary Tables S3, S4**). Our findings suggest that regulation of DNA replication,
147 and potentially replication stress, is a major regulator of mutation rates in yeast when MMR is
148 defective. We also observed an elevation in mutation rates when *TEL1*, the yeast ortholog of
149 *ATM*, was mutated in MMR-defective yeast cells (**Fig. 1c; Supplementary Table S3, S4**). This
150 is in agreement with the function of *ATM* as a central DNA damage response gene in human
151 cancers[43,44] and its association with microsatellite instability[45]. In summary, our genetic screen
152 across 4,847 deletion strains in budding yeast revealed cellular processes that cause up- and
153 down- regulation of mutation rates in MMR-defective cells.
154

155 **Systematic detection of MSI across 6,057 whole cancer genomes**

156 To investigate whether DNA replication and repair processes are associated with elevated
157 mutation rates in hereditary and sporadic human tumours, we performed an unbiased screen of
158 6,057 pan-cancer whole genomes for elevated microsatellite instability (MSI) of mono- and
159 dinucleotide repeats. In brief, we first searched for cancer types with elevated MSI occurrence,
160 and for DDR genes that were frequently mutated in MSI tumours. Secondly, we evaluated the
161 rate of MSI as a continuous measure in association with MMR loss-of-function mutation
162 (monoallelic or biallelic), other DDR gene mutations, and mutations in both MMR and other DDR
163 process genes. The data were obtained from the Pan-Cancer Analysis of Whole Genomes
164 (PCAWG; 2,560 genomes from predominantly primary tumours) and Hartwig Medical
165 Foundation (HMF; 3,497 genomes from metastatic solid tumours)[46,47] (**Fig. 2a; Supplementary**
166 **Table S5)**. The two data sets consist of tumours from 32 major cancer types of both sporadic
167 and hereditary origins. The HMF tumours were metastatic, whereas most PCAWG tumours
168 were confined to their tissue of origin. For each tumour, we determined the genome-wide

169 number of indels in mono- and dinucleotide repeat DNA (**Methods; Fig. 2b; Supplementary**
170 **Table S6**). We observed a high correlation (Spearman correlation of 0.89) between the number
171 of mono- and dinucleotide repeat indels (**Supplementary Fig. S4a**), and our counts correlated
172 (Spearman correlation = 0.81) with the result from the MSIseq algorithm[48]. MSIseq was used for
173 evaluating MSI in the original genomic analyses of HMF[46] (**Supplementary Fig. S4b**), and
174 summarises the number of indels in repeats (1-5 bases) across a pre-determined set of more
175 than 22 million sequences across the genome (**Methods**). We developed our own code for
176 summarising mono- and dinucleotide repeat indels to capture all indels and to focus on MMR.
177 This further allowed us to differentiate between repeat mono- and dinucleotide indels: notably,
178 we generally observed ten times higher rate of indels in mononucleotide than dinucleotide
179 repeats (**Supplementary Fig. S4c**), but did not observe any particular association to either type
180 of indels across our analysis. The repeat indels were enriched in repeats with a median length
181 of nine bases, although we detected repeat indels in sequences down to lengths of three bases,
182 with shorter repeat sequences being more frequent across the genome (**Supplementary Fig.**
183 **S5**). We generally did not observe repeat indels in very long repeat sequences (longer than 20
184 bases), likely due to technical challenges of calling such event. This meant that we did not
185 observe any repeat indels in the Bethesda Panel regions, suggesting that either PCR assays,
186 deeper NGS sequencing, or longer NGS reads would be needed for this type of validation [11].

187

188 In order to classify tumours as MSI or MSS (microsatellite stable), we established MSI
189 thresholds for the number of indels based on pre-annotated MSI tumours from the HMF[46] and
190 PCAWG datasets[49] (**Fig. 2b; Methods**). Using this method, we captured all previously
191 identified MSI tumours[46,49] (**Fig. 2b**). In total, 108 tumours from individual tumours (1.8%)
192 contained indels in their mono- and dinucleotide repeats above the MSI threshold, while the
193 remaining 5,949 were considered MSS (**Fig. 2c**). Importantly, we observed no correlation

194 between the increased indels in mono- and dinucleotide repeat and ploidy changes across the
195 genome (**Supplementary Fig.S6).**
196
197 **Increased MSI in small intestinal, prostate, endometrial, and colon cancers.**
198 We determined the proportion of tumours with MSI across all cancer types using the primary
199 site of disease of each tumour (**Fig. 3a; Supplementary Table S7**). Consistent with previous
200 findings[50,51], tumours showed a variable range of MSI (0% to 25%) with elevated MSI in small
201 intestinal (25%), endometrial (14%), colon (5.2%), and prostate (3.9%) cancers, compared to
202 the pan-cancer average (1.8%; one-tailed binomial test; q ≤0.05 by Benjamin-Hochberg False
203 discovery rate control (FDR); no. of tests = 15). These proportions are lower than in colorectal
204 and endometrial tumours found in Lynch syndrome[50-56], because the tumours in our and others
205 datasets[50,51] are not pre-selected for hereditary mutations in MMR. In contrast, 24 major cancer
206 types show a very low proportion (< 1%) of MSI. This included breast (0.82%), pancreatic
207 (0.52%), and ovarian cancers (0.80%).
208
209 We next determined whether the high MSI in small intestinal, endometrial, prostate, and
210 colorectal cancers was associated with defects in MMR genes. We detected pathogenic
211 variants (somatic and germline) across mismatch repair genes, as well as in *POLD1* and *POLE,*
212 for each tumour type (**Fig. 3b, c**). To this end we used the Combined Annotation Dependent
213 Depletion (CADD) scoring system[57] along with the ClinVar registry. The CADD system is an
214 algorithm that has been trained to predict pathogenicity of any variant across the genome based
215 on more than 60 annotations, including how conserved the site is across species, transcription
216 factor binding, and presence in exonic DNA. We included variants with a CADD phred score
217 above 25 as being pathogenic, which is equivalent to including the 0.3% variants with the
218 highest pathogenicity probability, genome-wide[57]. The ClinVar registry is manually curated for
219 each variant, but only holds information on a small part of the genome and is biased towards

220 known cancer genes. We considered ClinVar as the true positive and found that 87% of the
221 pathogenic and likely pathogenic variants were also identified by CADD (**Supplemental Table**
222 **S8; Methods**).
223
224 We inferred loss-of-function (LOF) whenever one or more pathogenic variants were observed
225 and included monoallelic pathogenic variants as evidence of hypomorphic MMR. Although
226 expected to lead to some false pathogenic event, we chose this strategy since MMR genes are
227 known to be frequently deactivated by a combination of variants and epigenetic events, for
228 which we did not have appropriate data[58,59]. Our approach avoided the reliance on determining
229 gene activity from the presence of mRNA or protein. Given the large number of tumours
230 available, any true statistical association with increased MSI should be detected despite the
231 noise. Consistent with this expectation, 66% of all tumours with MSI contained mutations in
232 MMR compared to 12% for MSS tumours (**Fig. 3b**). Amongst the four tumour types associated
233 with an elevated MSI, we detected considerable variation in the incidence of inactivating MMR
234 mutations. 74% (20 of 27) of colon, 62% (13 of 21) of prostate, and 27% (3 of 11) of endometrial
235 tumours with MSI carried a pathogenic MMR variant (defective MMR, dMMR). For small
236 intestinal cancers, the numbers were too low to estimate the association in a meaningful way (3
237 of 3 contained a mutation in an MMR gene). For the four tumour types, the spectrum of MMR
238 genes that contained a pathogenic variant also differed (**Fig. 3c**). Our findings suggest high
239 variability in the association of pathogenic MMR variants and MSI across tumours.
240
241 **Several DDR genes are enriched for mutations in MSI tumours.**
242 Next, we conducted an unbiased analysis of MSI and pathogenic mutations in 80 core DNA
243 damage response and repair genes amongst a total of 735 (**Supplementary Table S9**)[60-63].
244 Amongst these 80 genes, 95 of the 108 MSI tumours (88%) contained pathogenic variants
245 (**Supplementary Fig. S7**). For each gene, we identified tumours with pathogenic somatic and

246 germline variants and investigated the overrepresentation of MSI (**Fig. 4; Supplementary**
247 **Table 10; Methods**).
248
249 Among all included tumours, 108 of the 6,057 (1.8%) displayed MSI. This was elevated to 11%
250 for tumours with pathogenic germline events in *MSH6* and *PMS2* (**Fig. 4a**). We likewise
251 observed an overrepresentation of MSI among tumours with somatic mutations in the MMR
252 genes *MLH1, MSH3* and *MSH6,* as well as *RAD50* and *TOP3A* (**Fig. 4b**). Tumours with
253 pathogenic variants in *MLH1, MSH3, MSH6, RAD50,* or *TOP3A* showed MSI ranging from 55%
254 (*MLH1*) to 77% (*MSH6*) compared to the median rate of 23% across tumours with mutations in
255 the 735 DDR genes (14% among PCAWG tumours, 25% among HMF tumours). The increased
256 MSI in somatic tumours compared to those with germline-inherited variants is consistent with
257 previous findings from independent cancer genomes[50]. Our observations support a model in
258 which genes, beyond MMR genes, are associated with elevated MSI; in particular, *RAD50* and
259 *TOP3A*.
260
261 **Predicted pathogenic variants in DDR genes are associated with elevated MSI in MMR-**
262 **defective tumours.**
263 Several MSI tumours contained pathogenic variants in multiple MMR genes; this was
264 particularly notable for somatic co-mutations in *MSH3* and *MSH6* (31% overlap). Furthermore**,**
265 we observed frequent co-mutation of MMR genes with *RAD50* and *TOP3A* (**Fig. 5a;**
266 **Supplementary Fig. S8; Supplementary Table 11)**. This included previously observed co-
267 mutations with *RAD50 (*ref 64) as well as novel *TOP3A* co-mutations. Consistent with prior
268 findings, 81% of pathogenic variants in *RAD50* among MSI tumours were mono- and
269 dinucleotide repeat indels[64,65]. In contrast such indels only accounted for 2% of the pathogenic
270 *RAD50* variants in MSS tumours. Among MMR genes, this incidence of mono- and dinucleotide
271 repeat indels was only surpassed by *MSH3* (85%), and most closely followed by *MSH6* (48%)

272 (**Supplementary Fig. S9; Supplementary Table 12**), which form heterodimers with MSH2
273 during mismatch detection[66]. These observations suggest that MSI and loss of MMR and
274 *RAD50* may co-occur.
275
276 The high level of co-mutations led us to investigate whether co-mutated tumours, defective in
277 MMR and another DDR gene, showed a quantitative effect on MSI. To do this, we considered
278 MSI as a continuous measure rather than a binary state[51], making it possible to investigate
279 genetic modifiers of MSI in co-mutated tumours. Across 764 tumours with monoallelic
280 pathogenic variants in mismatch repair/*POLE/POLD1*, we searched for an elevation in the
281 number of indels in mono- and dinucleotide repeats when tumours had additional pathogenic
282 variants in any of the 80 core DNA repair genes **(Fig. 5b**). We observed an elevation in MSI
283 when two MMR components were co-mutated, consistent with our conclusions when
284 considering MSI as a binary trait. Furthermore, although the variants are pathogenic, the
285 mutations predominantly induce a hypomorphic MMR phenotype (**Fig. 5b; Supplementary Fig.**
286 **S10)**. This may be due to the wild-type allele being expressed and/or some overlapping
287 functions of the different MMR heterodimers (e.g. MSH2-MSH3 and MSH2-MSH6).
288
289 As in the yeast screen, we observed that several replication stress response genes showed
290 elevated MSI, exclusively when co-mutated with MMR. This included *ATR*, *BRCA2, CHEK1,*
291 *TOPBP1*, *MRE11*, *RAD50,* and *TOP3A*; **Fig. 5b**; **Supplementary Table 13**). This may
292 potentially be due their role in restarting stalled or collapsed replication forks[67-71]. For several of
293 these genes we saw an elevated rate of mutagenic events that were repeat indels
294 (**Supplementary Fig. S9; Supplementary Table 12**), suggesting that these genes are sensitive
295 to frameshift mutations. *BRCA2* has previously been shown to be co-mutated with MMR in
296 colorectal cancers[70] and our findings extend these observations by showing that their co-
297 mutation is associated with further elevation in the rate of MSI compared to defective MMR

298 alone. Finally, *ATM* showed a significant elevation in MSI, independently of mismatch repair
299 deficiency (**Fig. 5b**; **Supplementary Table 14**). In short, our observations across human
300 cancers are consistent with those in budding yeast. Taken together, the two approaches support
301 a model that tumours with pathogenic events in both MMR and genes important for regulation of
302 replication stress have a higher MSI rate (several orders of magnitudes) than tumours that are
303 only defective in MMR.
304
305 **DISCUSSION**
306 In this study, we set out to explore the genetic regulation of mutation rates in yeast and human
307 tumours with defective MMR. Our yeast deletion screen of 4,847 strains identified that mutation
308 rates can be both up- and downregulated, when MMR is defective, depending on co-mutations
309 in other cellular processes (**Fig. 1A-C**). By treating microsatellite instability (MSI) in human
310 tumours in a quantitative fashion, we also found this to be the case in MMR-defective tumours
311 (**Fig. 5b**). These findings may explain why certain MMR-defective tumours have low MSI[15,16],
312 although we do not have sufficient power in our data set to investigate this phenomenon.
313
314 Including both mono- and biallelic mutations of MMR genes in our analysis and treating MSI as
315 a continuous variable revealed a subgroup of tumours with a two- to three-orders of magnitude
316 elevation in MSI rate, compared to MMR-defective tumours alone (**Fig. 5b**). This included co-
317 mutations in replication response genes *(ATR, TOPBP1*, and *CHEK1*). We also identified genes
318 important for replication stress in our yeast screen. We propose a model in which replication
319 stress and MMR-deficiency collaboratively cause an extreme rate of MSI, possibly mainly in
320 somatic tumours that are subject to a high level of replication stress[72,73]. Since *MSH3*, *MSH6,*
321 and *RAD50* as well as *CHEK1* were highly mutated in mono- and dinucleotide repeats in
322 somatic tumours, these genes may themselves be subject to inactivation replication errors.

323 Thus, replication stress and MMR deficiency may both increase the risk of frameshift errors in
324 the other mechanism, causing a cyclic upregulation of repeat instability (**Fig. 6**).
325
326 Since replication stress is cell- and cancer-type specific our findings explain, at least in part, why
327 MSI is associated differentially with cancer types **(Fig3a)**[51,74]. Furthermore, they may have
328 clinical implications since the replication stress response is a major target for drug development
329 alone or in combination with other DNA damaging agents[60]. In particular, in tumours with MMR
330 defects, inhibition of replication stress genes, including *ATR, RAD50,* or *CHEK1* might further
331 induce MSI[75], thus increasing neoantigens that are important in acquiring response to
332 immunotherapy[14]. In terms of prognosis, our approach using MSI as a continuous score (the
333 number of mono- and dinucleotide repeat indels) may better reflect the nature of the MSI
334 phenotype[20,48,76], compared to the binary classification[11]. We have demonstrated that the
335 continuous score helped identify subgroups of tumours and mechanisms with extremely
336 elevated MSI. It remains to be determined whether a higher rate of MSI is linked to elevated
337 production of neoantigen and improved response to immunotherapy[14].
338
339 **ONLINE METHODS**
340 **QUANTIFICATION AND STATISTICAL ANALYSIS**
341 The data analysis was based on 6,057 whole cancer genomes (**Fig. 1a)**; 2,583 genomes from
342 the Pan-Cancer analysis of Whole Genomes (PCAWG; ICGC study ID. EGAS00001001692),
343 3,515 genomes from the Hartwig Medical Foundation (HMF; Acc. Nr. DR-044). The genomes
344 were obtained in the form of pre-called variants (.vcf), based on matched tumour-normal
345 samples, allowing the detection of both somatic and germline pathogenicity. Both datasets span
346 a multitude of tissues of disease, and we have grouped samples by their primary site of disease.
347 In multiple cases, this meant translating the cancer type (as assigned in each study
348 respectively) into a tissue. The assigned primary tissue of disease may be seen in

349 **Supplementary Table 5** along with sample IDs and donor IDs. The PCAWG dataset has been
350 created in an international collaboration, providing a span in ethnicity. In contrast, the HMF
351 dataset is Dutch and consists of Dutch donors. PCAWG consists of tumours both with and
352 without metastasis whereas HMF exclusively consists of donors with tumours harbouring
353 metastasis[46,47].
354
355 **Sample curation**
356 In the HMF data set, the first available tumour sample was selected whenever multiple tumour
357 samples exist for the same patient. Furthermore, two HMF samples were removed from the
358 analysis, as the donors retracted their consent. In the PCAWG data, the official whitelist
359 (available at https://dcc.icgc.org/releases/PCAWG/donors_and_biospecimens) [47] was used to
360 filter away samples with technical errors. We end up with a single sample per donor within the
361 two studies, in total 6,057 tumours (2,560 PCAWG tumours and 3,497 HMF tumours)
362
363 **Variant curation**
364 We have investigated genetic events, both somatic and germline, across all 6,057 samples. We
365 have filtered away variants with .vcf non-PASS status as well as somatic variants with variant
366 allele frequencies below 0.20, as these variants are likely to be recent and not have had a
367 strong influence on the accumulation of mono- and dinucleotide repeat indels. In the case of
368 PCAWG variants, several mutational callers have been used; we discarded all variants that
369 were only called by a single caller. We discarded variants occurring in more than 50 of either the
370 HMF or the PCWG tumours to avoid possible artefacts or SNPs with commonality in sub-
371 populations. We also discarded germline variants with gnomAD population frequencies (V2.1.1)
372 above 0.5% [47]. We used snpEff [77] to annotate the predicted effect of each variant, which was
373 used for **Supplementary Fig. S7** and **Supplementary Fig. S8.**
374

375 **Annotating microsatellite instability**

376 To get a consistent measure of MSI across the two datasets, we counted the number of
377 insertions and deletions in simple repeat sequences. This included: Single-base indels, in which
378 the indel sequence was similar to the flanking three or more bases on either side of the
379 mutation, and dinucleotide indels in which the dinucleotide indel sequence occurs in two copies
380 on either side of the mutation. We chose this approach to be able to discover possible
381 differences between mono- and dinucleotide counts, but found high correlation between the two
382 (**Supplementary Fig. S4a**), with mononucleotide events happening at a much higher rate than
383 dinucleotide events (**Supplementary Fig. S4a,c**).

384

385 We evaluated our counts by comparing them to microsatellite indel counts acquired using the
386 MSIseq software[48] (**Supplementary Fig. S4b**). In short, MSIseq counted all short insertions and
387 deletions across a pre-defined set of 22,295,304 stretches of simple repetitive DNA. These
388 stretches entail regions with four or more repeats of dinucleotides, trinucleotides, or
389 tetranucleotides, as well as homopolymer runs with lengths of five bases or more. The set of
390 regions was found via the MSIseq github page, via the link
391 http://steverozen.net/data/Hg19repeats.rda on February 12th, 2021[48].

392

393 **Identifying tumours with microsatellite instability**

394 Most studies of MSI separate tumours into groups of either microsatellite unstable or stable. To
395 achieve similar data, we decided to set thresholds on the number of mono- and dinucleotide
396 repeat indels in each sample. To account for possible differences between the PCAWG and
397 HMF datasets, we evaluated a separate threshold for each study:
398 For the PCAWG data, we included the results from a prior study by Fujimoto *et al.*[49] where MSI
399 status of the PCAWG cancer samples was annotated by counting the number of indels within
400 microsatellite regions. Due to the technical challenges of correctly calling indels in repetitive

401 DNA, Fujimoto *et al*. have scrutinised the variant calls. Furthermore, they have defined the
402 microsatellite regions using the combined efforts of multiple microsatellite and tandem repeat
403 identification tools: MsDetector[78], Tandem Repeat Finder[79], and MISA[80]. From this, Fujimoto *et*
404 *al.* have identified several samples with an increased number of mutated microsatellites. We
405 annotated our count of mono- and dinucleotide repeat indels onto all samples and set the
406 threshold at the lowest count among samples called as MSI by Fujimoto *et al.* The threshold for
407 MSI status was set at 9,649 Mono- and dinucleotide repeat indels, and we detected 22 tumours
408 with more mono- and dinucleotide repeat indels than this (**Fig. 1c**).
409
410 For the HMF samples, we included the results from Priestley *et al*.[46] , where the authors used
411 MSIseq to count mono- and dinucleotide repeat indels and then used PCR-analysis of the 5
412 marker Bethesda panel (BAT25, BAT26, NR21, NR24, and MONO27; **Supplementary Table**
413 **15**) on 48 selected samples in order to decide a threshold. Due to the continuous expansion of
414 the HMF database, we had access to samples that were not annotated by Priestley *et al.* To get
415 a coherent annotation of all our samples, we selected the MSI threshold to be the lowest
416 number of mono- and dinucleotide repeat indels among the MSI samples annotated by Priestely
417 *et al.* [46]. The threshold for the HMF tumours is 11,338 mono- and dinucleotide repeat indels. We
418 detected a total of 86 tumours with more mono- and dinucleotide repeat indels than this (**Fig.**
419 **2c**). The per-sample count of mono- and dinucleotide repeat indels and annotation of MSI state
420 may be found in **Supplementary Table 6**. This discrepancy between our set of MSI tumours
421 and the set presented by Priestley *et al.* stems from the increase in available samples in the
422 HMF database.
423
424 Several studies have investigated the use of PCR based and next-generation-sequencing
425 based detection of MSI in tumours, including the well-established Bethesda panel of five repeat
426 DNA regions[4,11,81]. Interestingly, we observed no indels in any of the five Bethesda regions. We

427 suggest that this is due to the length of the Bethesda panel markers and the technology used for
428 testing: The PCR-based technique used for Bethesda panel tests is sensitive to the length-
429 changes of such repeat runs whereas NGS is inappropriate for calling short indels in highly
430 repetitive DNA due to the challenge of assembly. Indeed, we generally saw very few indels in
431 repeats similar to those of the Bethesda panel (**Supplementary Fig. S5b**).
432
433 **Annotating pathogenic events across genes**
434 We combined a set of 735 genes with relation to the DNA damage response from previous work
435 by Knijnenburg *et al.,* Pearl *et al.,* and Olivieri *et al.* We excluded *TP53* from the gene-list, due to
436 the high number of mutations in the gene, which might cause skewed null-distributions for
437 mutation rates. Of these 735 genes, we considered 80 genes to be core DDR genes annotated
438 by Knijnenburg *et al.,* (added *POLD1*) (**Supplementary Table 9**). The 735 genes are generally
439 used for background modelling, and the 80 for hypothesis testing against the background.
440
441 For each of the 735 genes, we annotated all variants, germline and somatic, base substitutions
442 and indels, within the regions of these genes with CADD v1.6 scores[57] and ClinVar
443 pathogenicity annotations when available. All variants with a CADD phred score above 25 were
444 considered pathogenic, unless the ClinVar annotation presented conflicting evidence. Of all
445 possible variants in the genome, 0.3% receive a CADD score of 25 or more. All variants with a
446 ClinVar annotation of pathogenicity were considered pathogenic regardless of CADD-phred
447 score.
448
449 A CADD-phred pathogenicity threshold of 25 is conservative in contrast to suggested thresholds
450 and most gene-specific thresholds, which were suggested in prior studies[82,83]. We selected this
451 threshold to have a high level of confidence in the set of pathogenic variants, and achieve an
452 average rate of agreement with ClinVar in 87% (**Supplementary Table 8**). This means that

453 87% of the variants that were deemed pathogenic by ClinVar were also deemed pathogenic by
454 CADD annotation.
455
456 A single hit by a pathogenic variant within a gene was considered sufficient for the gene to be
457 annotated as LOF. We chose this strategy, as we could rarely detect bi-allelic hits and because
458 we cannot rule out alternative types of secondary hits, e.g. through transcriptional regulation[84].
459 Given the large size of the data, we expected that any true, biological signal would overcome
460 the statistical noise of erroneous LOF annotation. In support of this, we were able to identify the
461 expected association between MMR gene mutations and elevated MSI.
462
463 **Annotating mismatch repair deficiency**
464 Samples with pathogenic variants in one or more of the genes *MSH2, MSH3, MSH6, MLH1,*
465 *MLH3, PMS1, PMS2, EXO1, POLE,* and *POLD1* were denoted as mismatch repair deficient,
466 consistent with the set of genes used by Cortes-Ciriano *et al.*[50].
467
468 **Testing the association with cancer types (Fig. 3a)**
469 To test the prevalence of MSI across cancer types, we counted the number of tumours with MSI
470 for each type of cancer. The null-hypothesis was that there should be no significant deviation
471 from the pan-cancer rate of MSI, being 108/6,057 = 1.8%. We used a one-tailed binomial test,
472 with the alternative hypothesis being greater, to test this null-hypothesis, performing a total of 15
473 tests (one per cancer type with more than one MSI samples). Sample sizes, p-values, and q-
474 values (FDR-adjusted p-values) may be found for each cancer type in **Supplementary Table 7**.
475
476 **Testing for elevated rates of MSI across tumours mutated in DNA damage response**
477 **genes (Fig. 4a)**
478 We counted the number of MSI tumours and MSS tumours with pathogenic hits in each gene of

479 the 735 DDR genes and made a test to search for genes with a surprising proportion of MSI
480 tumours among the mutated tumours. The null-hypothesis was that the proportion of mutated
481 samples harbouring MSI would be the same across all genes. For germline hits, we assumed
482 the MSI proportion to be equal to the overall proportion of MSI tumours, 0.8% in PCAWG, 2.5%
483 in HMF.
484
485 For the somatic rate of MSI, we needed to consider the impact of somatic hypermutation or
486 other processes impacting the somatic mutation rate: For each study, we took the median
487 proportion of MSI tumours across the subset of the 735 genes that had three or more mutated
488 tumours. A threshold was set at three since genes that were mutated in very few cases had
489 extremely varying mutation rates. For PCAWG, the null-proportion of MSI tumours was found to
490 be 14%, in HMF it was 25%. We used the established null-rates to test enrichment of MSI
491 tumours among tumours mutated in the 80 core DDR genes, as we assumed that these null-
492 rates would be the expected rate of mutations in any gene given the hypermutated nature of
493 MSI tumours. For each gene, we did a one-tailed binomial test with the alternative hypothesis
494 being greater, on the number of MSI tumours with pathogenic hits, with germline and somatic
495 tests separately, and PCAWG and HMF tested separately. We combined the p-values for each
496 gene from the two studies using Fisher's method; if a mutated gene is only observed in one of
497 the two studies, the p-value was set to 1 in the opposite study. We ran a total of 63 tests on the
498 somatic data and 34 tests on the germline data (one test per gene with one or more mutated
499 MSI tumours and at least two MSS tumours. All sample sizes and p/q-values may be found in
500 **Supplementary Table 10**.
501
502 **Enrichment of mono- and dinucleotide repeat indels across MMR defective tumours (Fig.**
503 **5b)**
504 We used an unpaired Wilcoxon rank-sum test to association with elevated incidence of mono-

505 and dinucleotide repeat indels in MMR/*POLE*/*POLD1*-mutated tumours across the 80 core DDR
506 genes. For each these genes, we compared the number of mono- and dinucleotide repeat
507 indels between gene mutants and gene wild-types. This was done across MMR deficient
508 tumours (**Supplementary table 13**) and MMR proficient tumours (**Supplementary Table 14)**
509 separately. Through this, we could discover which genes had similar association to enrichment
510 of mono- and dinucleotide repeat indels regardless of the status of the mismatch repair
511 mechanism. P-values were FDR corrected across the 80 tests.
512

513 **EXPERIMENTAL METHODS**

514 **Mutagenesis Screen in yeast**

515 The genomic screen was carried out in 4,847 *S. cerevisiae* strains with reduced MMR activity.
516 To reduce the MMR activity in S. cerevisiae cells, a human mismatch repair gene *MLH1*
517 (*hMLH1*) was expressed in yeast using a low copy plasmid. The low-copy *hMLH1* expression
518 vector, pCI-ML10, is a CENARS LEU2 vector that expresses *hMLH1* derived from wild-type
519 *hMLH1* cDNA. This is expected to lead to a dominant mutator phenotype in yeast [32]. To
520 transform the EUROSCARF deletion library the protocol described by Gietz and Schiesl was
521 followed [85] with the following adjustments: The library was spotted on YPD plates. The spots
522 were then transferred to fresh YAPD plates and incubated for 2 days at 30 degrees. After two
523 hours of heat shocking, the cells were washed twice in 100 µl dH_2O, resuspended in 200 µl
524 permissive media and incubated for 48 hours. Cells were then resuspended in 50 µl dH_2O and
525 transferred onto permissive plates using a 96-prong replicator. After 48 hours at 30°C cells were
526 transferred onto canavanine plates and incubated for 72h, where strains were scored for
527 mutator phenotypes. Deletions were scored as resulting in either slow growth, an increase or a
528 decrease in the mutation rate compared to cell growth expected from wildtype (WT) cells
529 expressing *hMLH1*. Strains in any of these categories were further analysed in the secondary
530 screen. For the secondary screen transformants of interest were patched in quadruplicate on

531 permissive plates, together with the WT BY4741 strain expressing either empty plasmid as a
532 negative and *hMLH1* as a positive control. After 24 hours at 30°C cells were transferred to
533 canavanine plates and incubated for 72 hours. Canavanine resistance was scored in
534 comparison with the WT controls.
535

536 **Fluctuation test**

537 The fluctuation test was performed using the method of median[86]. First, strains were streaked
538 for single colonies, and plates were incubated at 30°C for three days. 5 independent colonies
539 were inoculated in 5ml complete media (or selective media for plasmid in case strains have
540 plasmid) and grown until they reached stationary phase at 30°C shaking at 200 rpm. Cells were
541 spun down at 1,258 × *g* for 5 minutes and re-suspended in 800 µL of sterile water. The volume
542 of each cell pellet was carefully noted after resuspension in water by pipette. The solution was
543 then diluted for different strains to achieve countable numbers of colonies. 100 µl solutions were
544 plated out on complete plates and canavanine plates. After incubation for three days, the
545 number of colonies was noted. The mutation rate for each strain was calculated using the
546 equation $R_o = M \times (1.24 + \ln(M))$, where R_o is the median number of canavanine-resistant
547 colonies, M is the average number of canavanine-resistant colonies per culture. The M value
548 was determined by interpolation and used to estimate the mutation rate, R, as R= M/N, where N
549 is the average number of cells per culture. An average mutation rate was calculated from
550 mutation rates obtained from the three independent experiments.

551 **Enrichment analysis using DAVID 6.8**

552 The DAVID 6.8 database[33,34] was used to identify biological processes that are enriched in the
553 163 non-essential genes that showed an increase in mutation rate in the hypomorphic MMR
554 background. The enrichment analysis was carried out using the 'Functional Annotation Chart'
555 tool for biological processes (GOTERM_BP_DIRECT). The thresholds were set as Count= 2

556 and significance was evaluated using Fisher's exact test (alternative greater), with EASE
557 adjustment of 1 as is default in DAVID 6.8[87].

558 **Sensitivity assay**

559 to test the detection power for *CAN1* forward mutation rate assay qualitatively WT, *mlh1*Δ,
560 *exo1*Δ and MLH1-E31A strains were diluted two-fold, five-fold, and ten-fold and plated on
561 canavanine containing media. Mutation rates were qualitatively assessed qualitatively.
562

563 **Western blot analyses**

564 The primary antibodies used in this study were: anti-hMLH, 1:500 in BSA-PBSTween;PGK1,
565 1:2500 PBS-Tween. The samples were separated on 8% SDS–PAGE gel and transferred to a
566 membrane by semi-dry transfer using Trans-Blot® SD Semi-Dry Transfer Cell. The membranes
567 were blocked with 5% BSA PBS-tween for 30 min, incubated overnight with primary antibodies,
568 washed three times with PBS-tween for 10 min each, incubated with the secondary antibody
569 (anti-mouse IgG, 1:5000 in PBS-tween) for 60 min and , washed three times with PBS-tween for
570 10 min each, and proteins were detected using ECL (Thermo scientific, # 1859698)
571

572 **Data availability**

573 The analysis was based on two pre-existing datasets, that of the Pan-Cancer analysis of Whole
574 Genomes and that of the Hartwig Medical Foundation. PCAWG data (ICGC study ID
575 EGAS00001001692) may be accessed through gbGaP and ICGC DACO, as instructed on this
576 site https://docs.icgc.org/pcawg/data/. The HMF data used in this project may be found by
577 accession code DR-044 and can be obtained upon request at the Hartwig Medical Foundation
578 (https://www.hartwigmedicalfoundation.nl/en).
579

580 **Code availability**

581 All code used for the analysis is available at https://github.com/Meiomap/PanCan_MSI/

582

583 **Acknowledgements**

584 We are grateful to the Pan-Cancer Analysis of Whole Genomes (PCAWG), The International
585 Cancer Genome Consortium (ICGC), and The Cancer Genome Atlas (TCGA) for access to
586 whole cancer genomes. We would also like to thank the Hartwig Medical Foundation (HMF) and
587 the Center for Personalized Cancer Treatment (CPCT) for creating and providing access to
588 metastatic whole cancer genome data. ERH and AS were funded by the Novo Nordisk
589 Foundation (NNF15OC0016662), Cancer Research UK (C23210/A7574), and the Danish
590 National Research Foundation (Center grant, DNRF115). JSP and SGS were funded by the
591 Independent Research Fund Denmark | Medical Sciences (8021-00419B), Aarhus University
592 Research Foundation (AUFF-E-2020-6-14), and a PhD stipend from Aarhus University. We
593 thank Amy V. Kaucher and Jason Alexander Halliwell for valuable comments on the manuscript.

594

595 **Author contributions**

596 The project was designed by ERH and AS. AS, JO, BD performed yeast experiments under
597 supervision of ERH, and SGS performed data analysis under supervision of JSP. SGS, AS, and
598 ERH wrote the manuscript. All authors accepted the manuscript.

599

600 **Competing interests**

601 The authors declare no competing interests.

602

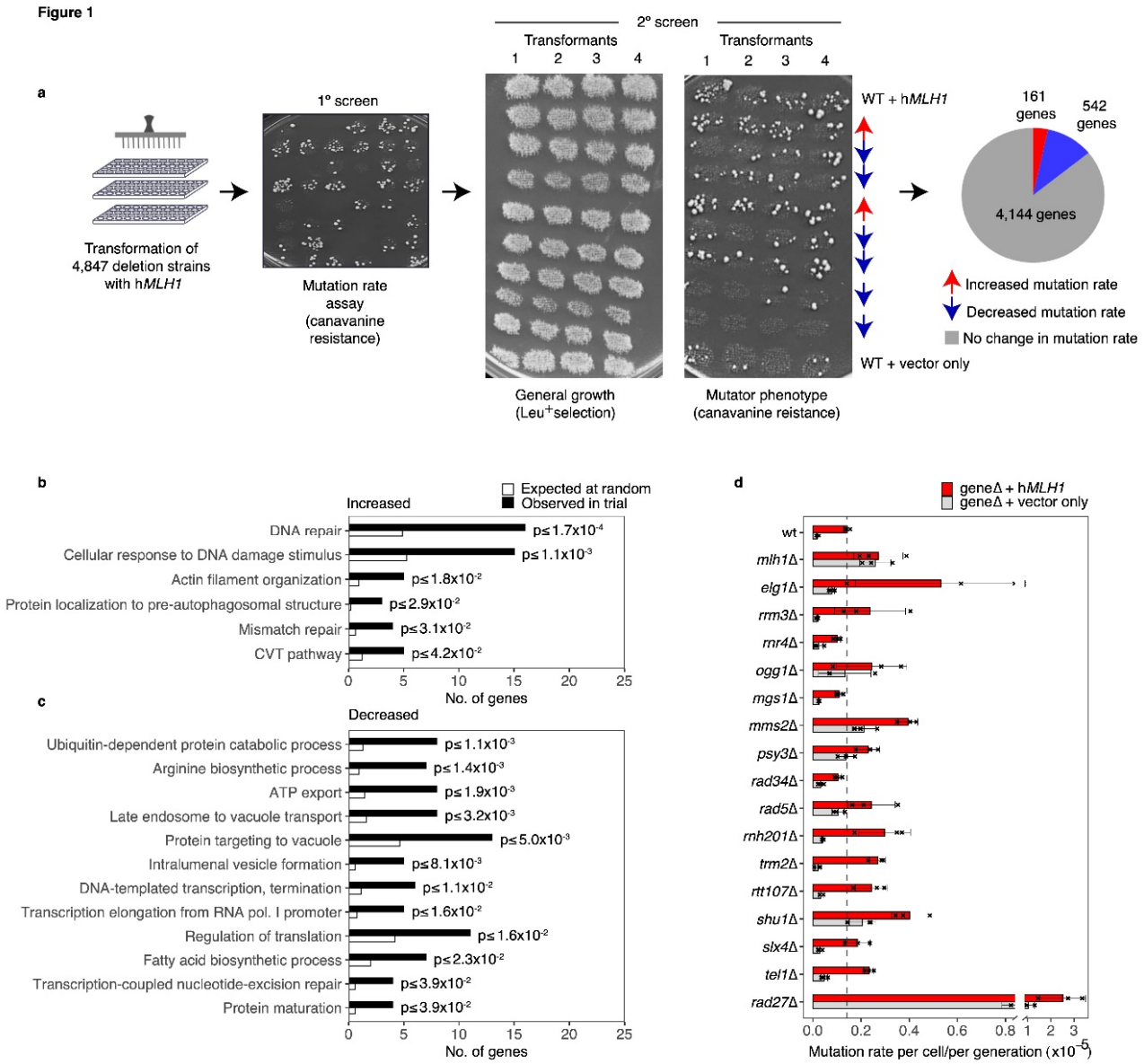

Figure 1: Genome-wide screen for mutants with changed mutation rate in hypomorphic MMR strains in yeast.

a Deletion library (4,847 deletion yeast strains) was transformed with *hMLH1* expression plasmid, and mutation rates were assayed by canavanine resistance (**methods**). In the 1st screen, 1,478 strains showed altered resistance to canavanine. In the 2nd screen, the experiment was repeated for four independent transformants of each of the 1,478 strains. Arrows indicate altered mutation rates, compared to a hypomorphic MMR background (WT

611 +h*MLH1*). The 2nd screen revealed 163 deletion strains with elevated mutation rates, and 523

612 with decreased. Control wild-type with plasmid is also shown (WT +vector). Biological processes

613 (annotated using the DAVID database[33,34] that are

614 **b** enriched amongst deletions resulting in increased mutation rate or

615 **c** decreased mutation rate (Fisher's exact test (alt: greater) with EASE adjustment (**Methods;**

616 **Supplementary Tables 1-4**). Processes with $p \leq 0.0.5$ were included.

617 **d** Average mutation rates (error bars indicate 95% confidence interval) per cell for deletion

618 strains when either vector only or *hMLH1* is expressed. Mutation rates were determined using

619 method of medians [88]. Bars indicate gene of interest deletion with vector (grey) and gene of

620 interest deletion with *hMLH1* (red; **Supplementary Table 1**).

621

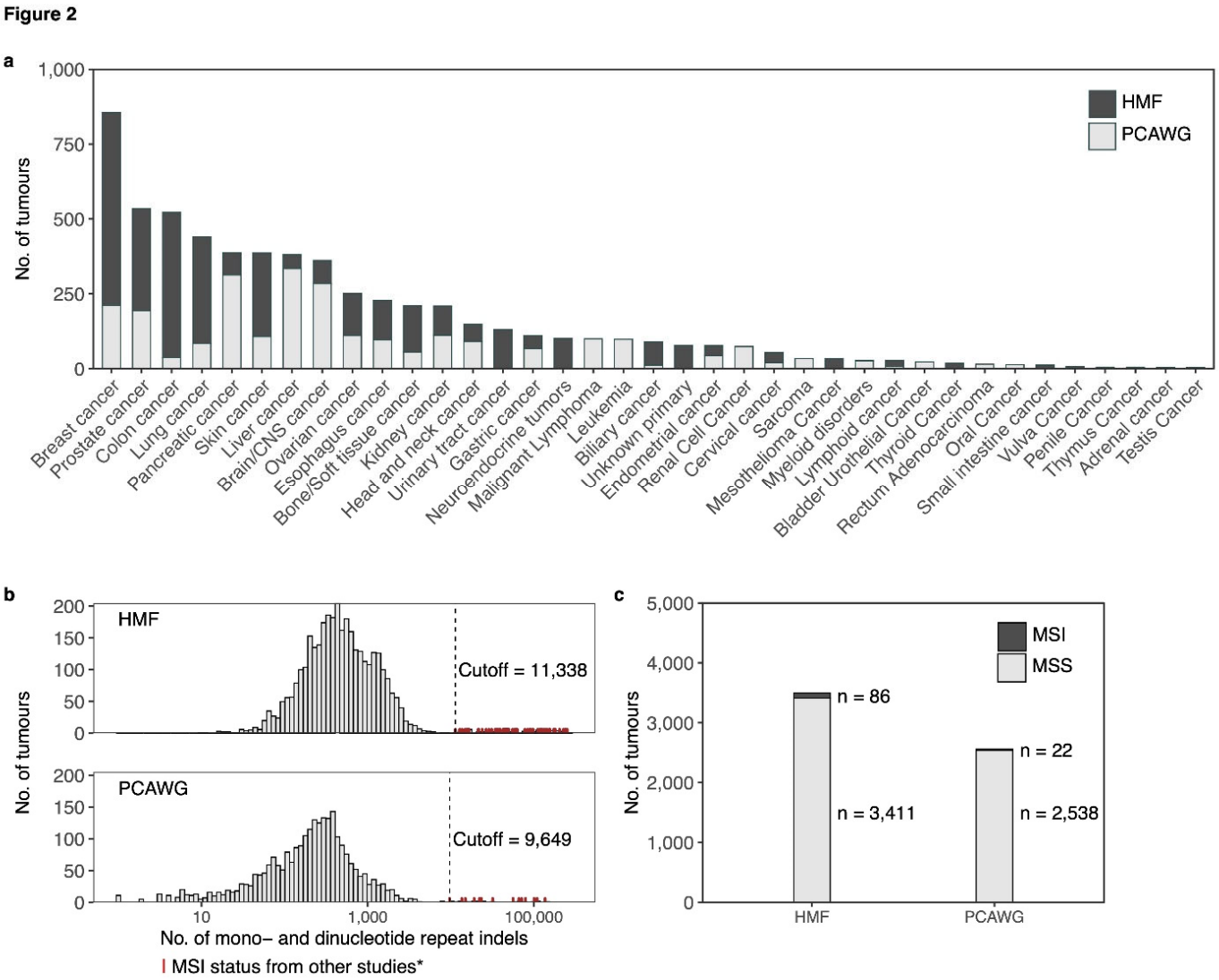

Figure 2. Mono- and dinucleotide insertions and deletions across ~6,000 whole cancer genomes

a Number of tumours by cancer type and study, either Hartwig Medical Foundation (HMF; grey) or Pan Cancer Analysis of Whole Genomes (PCAWG; black).

b Number of mono- and dinucleotide repeat indels per tumour (y-axis), red bars mark tumours that have been annotated as microsatellite unstable in former studies (in PCAWG by Fujimoto *et al*.[49]; in HMF by Priestley *et al*[46]). Dashed lines indicate the selected threshold between microsatellite instability (MSI) and microsatellite stability (MSS).

631 **c** number of inferred MSI (black) and MSS (grey) tumours across the HMF- and PCAWG data

632 set.

633

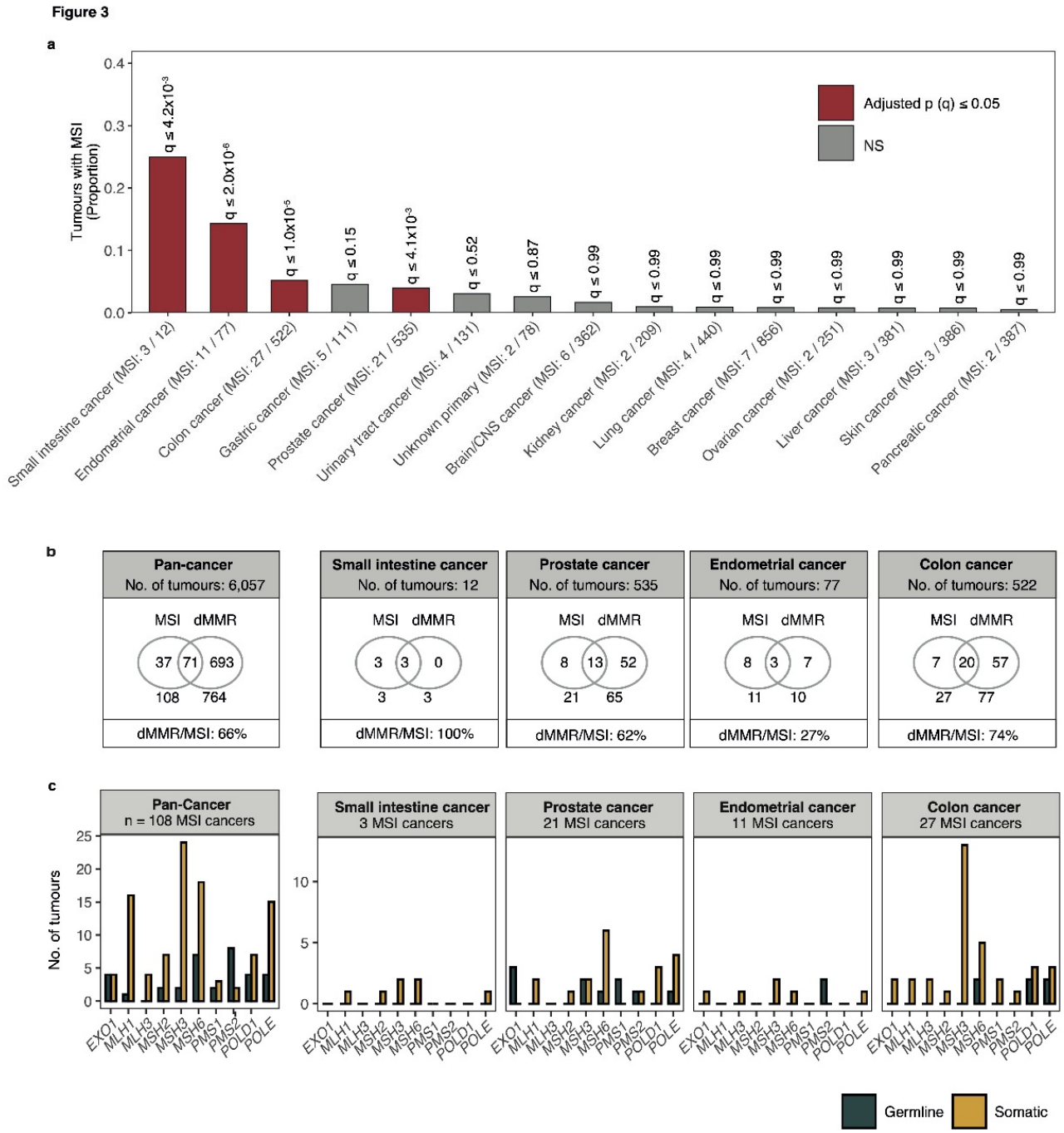

Figure 3. Some cancer types show high MSI incidence

a Incidence of MSI across cancer types (y-axis). Cancer types with elevated proportions of MSI compared to the overall average of 1.8% (108 / 6,057) are shown in red (binomial tests, one-

638 sided (greater); FDR adjusted p-values (q-values); No. of tests = 15, one per cancer type with

639 one or more MSI tumour).

640 **b** Venn diagrams of co-occurrence of MSI and pathogenic mutations in mismatch repair genes,

641 *POLE* and *POLD1*.

642 **c** Variants in mismatch repair genes, *POLE,* or *POLD1* in tumours displaying MSI.

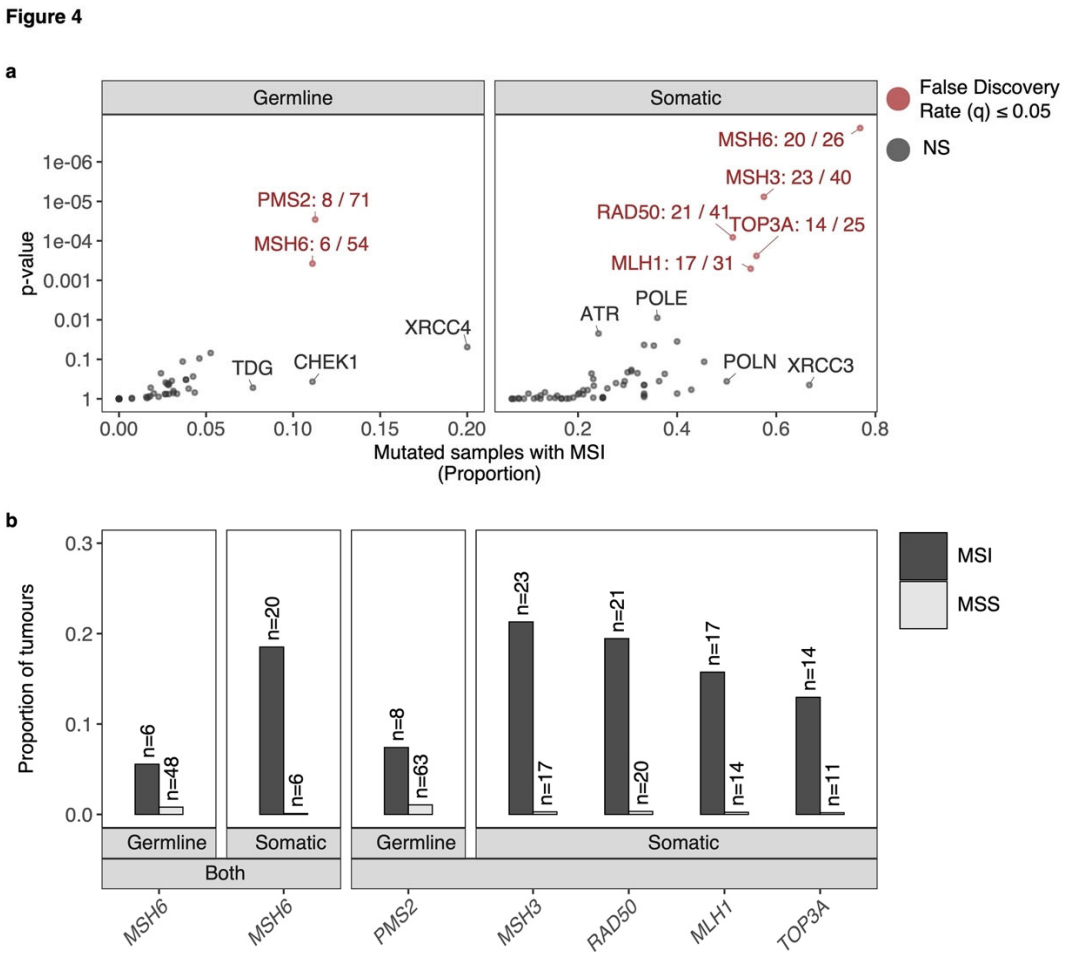

Figure 4. MSI tumours display more frequent pathogenic mutations in some DDR genes

a Pathogenic germline and somatic variants in DDR genes amongst tumours with microsatellite instability statistically explored using one-tailed (greater) binomial testing between the observed rate of MSI (x-axis) among tumours with mutations in each gene and the per-study expected rate (Rate of MSI 735 DDR genes, tests across 80 core DDR genes; **Methods**).Genes with q-values ≤ 0.05 (red; FDR adjusted p-values for multiple testing, 63 tests of somatic association; 34 tests of germline) are considered significant.

b Incidence of pathogenic variants across all MSI or MSS tumours (y-axis) in DDR genes with q<0.05 (Red; Fig. 4a)**.** Numbers indicate the number of tumours with pathogenic events in the gene.

Figure 5

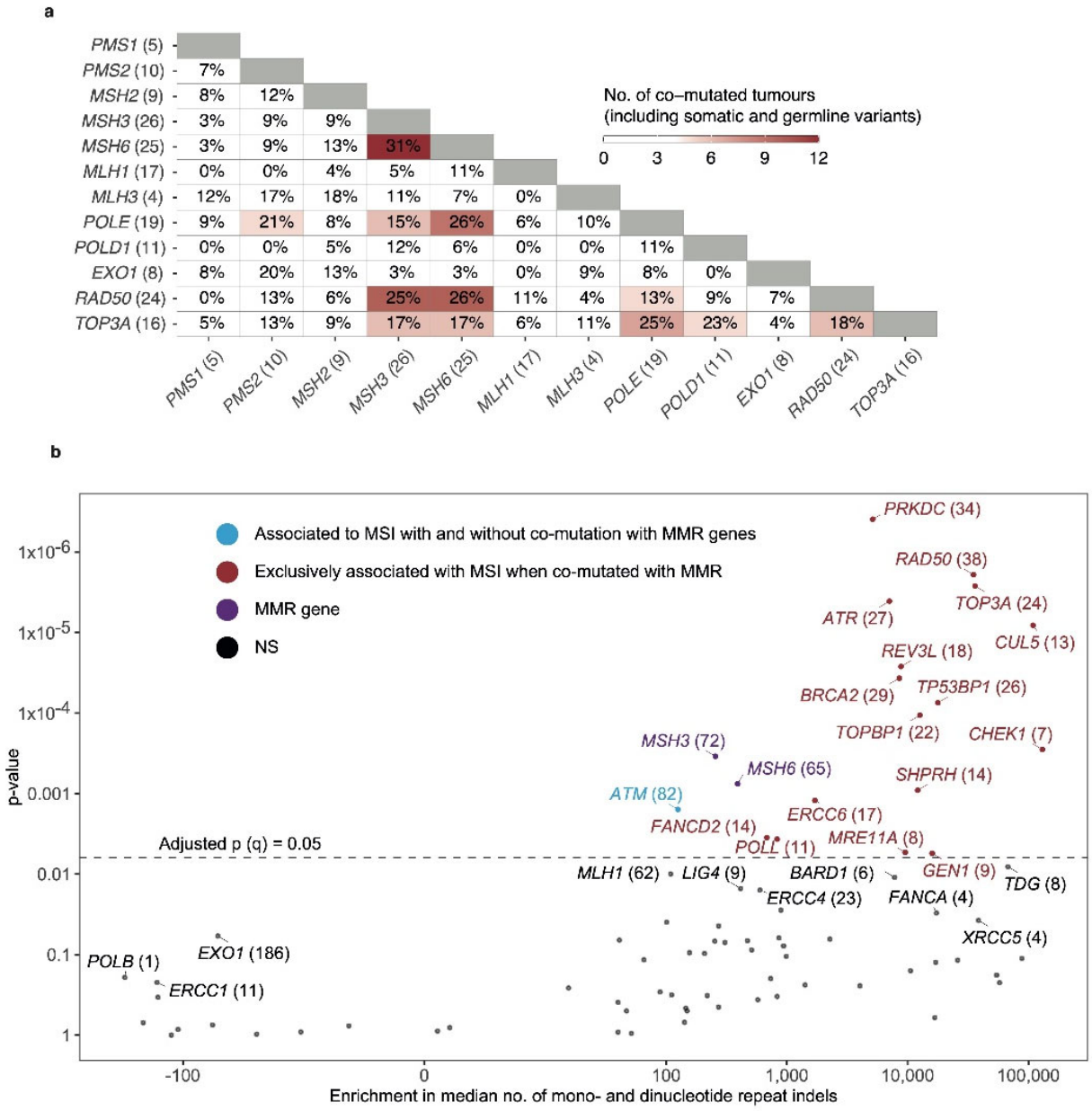

Figure 5. Co-mutation between MMR genes and core DNA damage response genes associates with amplification of mono- and dinucleotide repeat indels

a Co-mutations, including both somatic and germline variants, in MMR genes, *POLE, POLD1, RAD50,* and *TOP3A* across the 108 tumours with MSI. Colour indicates the number of co-mutated tumours, parenthesis indicate the number of mutated tumours of each gene.

660 Percentage in each tile indicates the proportion of tumours with co-mutation out of the set of
661 tumours mutated across the two genes being considered.
662 **b** Enrichment in median number of mono- and dinucleotide repeat indels (x-axis) among tumour
663 with MMR/*POLE/POLD1* deficiency (n = 764 tumours) and pathogenic events in core DDR
664 genes, statistically explored using a one-tailed Wilcoxon rank-sum test (un-paired; greater; y-
665 axis). Cases with FDR-adjust p-values (q; no. of tests = 78) ≤ 0.05 are coloured and the number
666 of tumours mutated in each gene is given in parenthesis. Purple genes are MMR/*POLE/POLD1*;
667 Red genes were exclusively associated with elevated MSI in MMR mutated tumours; Blue
668 genes were found to be both associated with MSI in cancers with and without MMR co-
669 mutations. Black genes had no significant association with altered MSI.
670

Figure 6
a

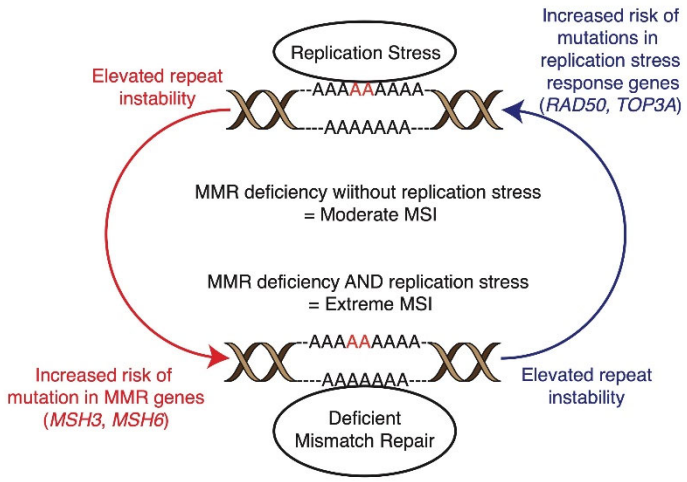

Figure 6. A suggested model of MSI generation

a LOF in either replication stress response or mismatch repair elevated mutation rates in repeat sequences, and LOF in either mechanism may induce LOF in the other. Our data suggest that LOF in both mechanisms may lead to higher MSI than LOF in either alone.

679 **References**

680 1 Smith, B. T., Grossman, A. D. & Walker, G. C. Visualization of mismatch repair in
681 bacterial cells. *Mol. Cell* **8**, 1197-1206, doi:10.1016/s1097-2765(01)00402-6 (2001).
682 2 Kunkel, T. A. & Erie, D. A. DNA mismatch repair. *Annu. Rev. Biochem.* **74**, 681-710,
683 doi:10.1146/annurev.biochem.74.082803.133243 (2005).
684 3 Hombauer, H., Campbell, C. S., Smith, C. E., Desai, A. & Kolodner, R. D. Visualization
685 of eukaryotic DNA mismatch repair reveals distinct recognition and repair intermediates.
686 *Cell* **147**, 1040-1053, doi:10.1016/j.cell.2011.10.025 (2011).
687 4 Murphy, K. M. *et al.* Comparison of the microsatellite instability analysis system and the
688 Bethesda panel for the determination of microsatellite instability in colorectal cancers. *J.*
689 *Mol. Diagn.* **8**, 305-311, doi:10.2353/jmoldx.2006.050092 (2006).
690 5 Kunkel, T. A. Nucleotide repeats. Slippery DNA and diseases. *Nature* **365**, 207-208,
691 doi:10.1038/365207a0 (1993).
692 6 Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous
693 somatic mutations in simple repeated sequences reveal a new mechanism for colonic
694 carcinogenesis. *Nature* **363**, 558-561, doi:10.1038/363558a0 (1993).
695 7 Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO*
696 *Precis Oncol* **2017**, doi:10.1200/PO.17.00073 (2017).
697 8 Jiricny, J. The multifaceted mismatch-repair system. *Nat. Rev. Mol. Cell Biol.* **7**, 335-346,
698 doi:10.1038/nrm1907 (2006).
699 9 Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology*
700 **138**, 2073-2087.e2073, doi:10.1053/j.gastro.2009.12.064 (2010).
701 10 Benatti, P. *et al.* Microsatellite instability and colorectal cancer prognosis. *Clin. Cancer*
702 *Res.* **11**, 8332-8340, doi:10.1158/1078-0432.CCR-05-1030 (2005).
703 11 Umar, A. *et al.* Revised Bethesda Guidelines for hereditary nonpolyposis colorectal
704 cancer (Lynch syndrome) and microsatellite instability. *J. Natl. Cancer Inst.* **96**, 261-268
705 (2004).
706 12 Peltomäki, P. *et al.* Microsatellite instability is associated with tumors that characterize
707 the hereditary non-polyposis colorectal carcinoma syndrome. *Cancer Res.* **53**, 5853-
708 5855 (1993).
709 13 Xiao, Y. & Freeman, G. J. The microsatellite instable subset of colorectal cancer is a
710 particularly good candidate for checkpoint blockade immunotherapy. *Cancer Discov.* **5**,
711 16-18, doi:10.1158/2159-8290.CD-14-1397 (2015).
712 14 Llosa, N. J. *et al.* The vigorous immune microenvironment of microsatellite instable colon
713 cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discov.* **5**, 43-51,
714 doi:10.1158/2159-8290.CD-14-0863 (2015).
715 15 Gylling, A. H. S. *et al.* Differential cancer predisposition in Lynch syndrome: insights from
716 molecular analysis of brain and urinary tract tumors. *Carcinogenesis* **29**, 1351-1359,
717 doi:10.1093/carcin/bgn133 (2008).
718 16 Peltomäki, P. Update on Lynch syndrome genomics. *Fam. Cancer* **15**, 385-393,
719 doi:10.1007/s10689-016-9882-8 (2016).
720 17 Tran, H. T., Gordenin, D. A. & Resnick, M. A. The 3'-->5' exonucleases of DNA
721 polymerases delta and epsilon and the 5'-->3' exonuclease Exo1 have major roles in
722 postreplication mutation avoidance in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **19**,
723 2000-2007, doi:10.1128/MCB.19.3.2000 (1999).
724 18 Morrison, A., Johnson, A. L., Johnston, L. H. & Sugino, A. Pathway correcting DNA
725 replication errors in Saccharomyces cerevisiae. *EMBO J.* **12**, 1467-1473 (1993).
726 19 Huang, M.-E., Rio, A.-G., Nicolas, A. & Kolodner, R. D. A genomewide screen in
727 Saccharomyces cerevisiae for genes that suppress the accumulation of mutations. *Proc.*
728 *Natl. Acad. Sci. U. S. A.* **100**, 11529-11534, doi:10.1073/pnas.2035018100 (2003).

20 Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat Cancer* **2**, 643-657, doi:10.1038/s43018-021-00200-0 (2021).
21 Wierdl, M., Dominska, M. & Petes, T. D. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* **146**, 769-779 (1997).
22 Kunkel, T. A. The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J. Biol. Chem.* **260**, 5787-5796 (1985).
23 Streisinger, G. & Owen, J. Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. *Genetics* **109**, 633-659 (1985).
24 Tran, H. T., Keen, J. D., Kricker, M., Resnick, M. A. & Gordenin, D. A. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. *Mol. Cell. Biol.* **17**, 2859-2865, doi:10.1128/mcb.17.5.2859 (1997).
25 Huang, M. E., Rio, A. G., Nicolas, A. & Kolodner, R. D. A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations. *Proc Natl Acad Sci U S A* **100**, 11529-11534, doi:10.1073/pnas.2035018100 (2003).
26 Gallina, I. *et al.* Cmr1/WDR76 defines a nuclear genotoxic stress body linking genome integrity and protein quality control. *Nat Commun* **6**, 6533, doi:10.1038/ncomms7533 (2015).
27 Kadyrov, F. A., Dzantiev, L., Constantin, N. & Modrich, P. Endonucleolytic function of MutLalpha in human mismatch repair. *Cell* **126**, 297-308, doi:10.1016/j.cell.2006.05.039 (2006).
28 Marsischky, G. T., Filosi, N., Kane, M. F. & Kolodner, R. Redundancy of Saccharomyces cerevisiae MSH3 and MSH6 in MSH2-dependent mismatch repair. *Genes Dev.* **10**, 407-420, doi:10.1101/gad.10.4.407 (1996).
29 Prolla, T. A., Christie, D. M. & Liskay, R. M. Dual requirement in yeast DNA mismatch repair for MLH1 and PMS1, two homologs of the bacterial mutL gene. *Mol. Cell. Biol.* **14**, 407-415, doi:10.1128/mcb.14.1.407-415.1994 (1994).
30 Takahashi, M. *et al.* Functional analysis of human MLH1 variants using yeast and in vitro mismatch repair assays. *Cancer Res.* **67**, 4595-4604, doi:10.1158/0008-5472.CAN-06-3509 (2007).
31 Clyne, M. *et al.* The G67E mutation in hMLH1 is associated with an unusual presentation of Lynch syndrome. *Br. J. Cancer* **100**, 376-380, doi:10.1038/sj.bjc.6604860 (2009).
32 Shimodaira, H. *et al.* Functional analysis of human MLH1 mutations in Saccharomyces cerevisiae. *Nat. Genet.* **19**, 384-389, doi:10.1038/1277 (1998).
33 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).
34 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).
35 Datta, A. & Jinks-Robertson, S. Association of increased spontaneous mutation rates with high levels of transcription in yeast. *Science* **268**, 1616-1619, doi:10.1126/science.7777859 (1995).
36 Tishkoff, D. X., Filosi, N., Gaida, G. M. & Kolodner, R. D. A novel mutation avoidance mechanism dependent on S. cerevisiae RAD27 is distinct from DNA mismatch repair. *Cell* **88**, 253-263, doi:10.1016/s0092-8674(00)81846-2 (1997).
37 Azvolinsky, A., Dunaway, S., Torres, J. Z., Bessler, J. B. & Zakian, V. A. The S. cerevisiae Rrm3p DNA helicase moves with the replication fork and affects replication of all yeast chromosomes. *Genes Dev.* **20**, 3104-3116, doi:10.1101/gad.1478906 (2006).

38 Kanellis, P., Agyei, R. & Durocher, D. Elg1 forms an alternative PCNA-interacting RFC complex required to maintain genome stability. *Curr. Biol.* **13**, 1583-1595, doi:10.1016/s0960-9822(03)00578-5 (2003).
39 Branzei, D., Seki, M. & Enomoto, T. Rad18/Rad5/Mms2-mediated polyubiquitination of PCNA is implicated in replication completion during replication stress. *Genes Cells* **9**, 1031-1042, doi:10.1111/j.1365-2443.2004.00787.x (2004).
40 Meurisse, J. *et al.* Hug1 is an intrinsically disordered protein that inhibits ribonucleotide reductase activity by directly binding Rnr2 subunit. *Nucleic Acids Res.* **42**, 13174-13185, doi:10.1093/nar/gku1095 (2014).
41 Ohouo, P. Y., Bastos de Oliveira, F. M., Almeida, B. S. & Smolka, M. B. DNA damage signaling recruits the Rtt107-Slx4 scaffolds via Dpb11 to mediate replication stress response. *Mol. Cell* **39**, 300-306, doi:10.1016/j.molcel.2010.06.019 (2010).
42 Choudhury, S. A., Asefa, B., Webb, A., Ramotar, D. & Chow, T. Y. K. Functional and genetic analysis of the Saccharomyces cerevisiae RNC1/TRM2: evidences for its involvement in DNA double-strand break repair. *Mol. Cell. Biochem.* **300**, 215-226, doi:10.1007/s11010-006-9386-1 (2007).
43 Cremona, C. A. & Behrens, A. ATM signalling and cancer. *Oncogene* **33**, 3351-3360, doi:10.1038/onc.2013.275 (2014).
44 Jin, M. H. & Oh, D.-Y. ATM in DNA repair in cancer. *Pharmacol. Ther.* **203**, 107391, doi:10.1016/j.pharmthera.2019.07.002 (2019).
45 Kim, H. S., Choi, S. I., Min, H. L., Kim, M. A. & Kim, W. H. Mutation at intronic repeats of the ataxia-telangiectasia mutated (ATM) gene and ATM protein loss in primary gastric cancer with microsatellite instability. *PLoS One* **8**, e82769, doi:10.1371/journal.pone.0082769 (2013).
46 Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210-216, doi:10.1038/s41586-019-1689-y (2019).
47 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).
48 Huang, M. N. *et al.* MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. *Sci. Rep.* **5**, 13321, doi:10.1038/srep13321 (2015).
49 Fujimoto, A. *et al.* Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res.*, doi:10.1101/gr.255026.119 (2020).
50 Cortes-Ciriano, 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, doi:10.1038/ncomms15180 (2017).
51 Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22**, 1342-1350, doi:10.1038/nm.4191 (2016).
52 Eric Bronner, C. *et al.* Mutation in the DNA mismatch repair gene homologue hMLH 1 is associated with hereditary non-polyposis colon cancer. *Nature* **368**, 258-261, doi:10.1038/368258a0 (1994).
53 Papadopoulos, N. *et al.* Mutation of a mutL homolog in hereditary colon cancer. *Science* **263**, 1625-1629, doi:10.1126/science.8128251 (1994).
54 Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-1038, doi:10.1016/0092-8674(93)90546-3 (1993).
55 Plazzer, J. P. *et al.* The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam. Cancer* **12**, 175-180, doi:10.1007/s10689-013-9616-0 (2013).

56 Thompson, B. A. *et al.* Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat. Genet.* **46**, 107-115, doi:10.1038/ng.2854 (2014).
57 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886-D894, doi:10.1093/nar/gky1016 (2019).
58 Wheeler, J. M. *et al.* Mechanisms of inactivation of mismatch repair genes in human colorectal cancer cell lines: the predominant role of hMLH1. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 10296-10301, doi:10.1073/pnas.96.18.10296 (1999).
59 Cunningham, J. M. *et al.* Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res.* **58**, 3455-3460 (1998).
60 Pearl, L. H., Schierz, A. C., Ward, S. E., Al-Lazikani, B. & Pearl, F. M. G. Therapeutic opportunities within the DNA damage response. *Nat. Rev. Cancer* **15**, 166-180, doi:10.1038/nrc3891 (2015).
61 Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, 239-254.e236, doi:10.1016/j.celrep.2018.03.076 (2018).
62 Olivieri, M. *et al.* A Genetic Map of the Response to DNA Damage in Human Cells. *Cell* **182**, 481-496.e421, doi:10.1016/j.cell.2020.05.040 (2020).
63 Ruth, K. S. *et al.* Genetic insights into the biological mechanisms governing human ovarian ageing. *bioRxiv*, doi:10.1101/2021.01.11.20248322 (2021).
64 Duval, A. *et al.* Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers. *Hum. Mol. Genet.* **10**, 513-518, doi:10.1093/hmg/10.5.513 (2001).
65 Ikenoue, T. *et al.* Frameshift mutations at mononucleotide repeats in RAD50 recombinational DNA repair gene in colorectal cancers with microsatellite instability. *Jpn. J. Cancer Res.* **92**, 587-591, doi:10.1111/j.1349-7006.2001.tb01134.x (2001).
66 Genschel, J. & Modrich, P. Mechanism of 5'-directed excision in human mismatch repair. *Mol. Cell* **12**, 1077-1086, doi:10.1016/s1097-2765(03)00428-3 (2003).
67 Lee, J. & Dunphy, W. G. The Mre11-Rad50-Nbs1 (MRN) complex has a specific role in the activation of Chk1 in response to stalled replication forks. *Mol Biol Cell* **24**, 1343-1353, doi:10.1091/mbc.E13-01-0025 (2013).
68 Flynn, R. L. & Zou, L. ATR: a master conductor of cellular responses to DNA replication stress. *Trends Biochem Sci* **36**, 133-140, doi:10.1016/j.tibs.2010.09.005 (2011).
69 Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat Cell Biol* **16**, 2-9, doi:10.1038/ncb2897 (2014).
70 Wang, J. *et al.* Next-generation sequencing reveals heterogeneous genetic alterations in key signaling pathways of mismatch repair deficient colorectal carcinomas. *Mod. Pathol.* **33**, 2591-2601, doi:10.1038/s41379-020-0612-2 (2020).
71 Lomonosov, M., Anand, S., Sangrithi, M., Davies, R. & Venkitaraman, A. R. Stabilization of stalled DNA replication forks by the BRCA2 breast cancer susceptibility protein. *Genes Dev.* **17**, 3017-3022, doi:10.1101/gad.279003 (2003).
72 Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability--an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220-228, doi:10.1038/nrm2858 (2010).
73 Di Micco, R. *et al.* Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature* **444**, 638-642, doi:10.1038/nature05327 (2006).
74 Buckley, A. R., Ideker, T., Carter, H., Harismendy, O. & Schork, N. J. Exome-wide analysis of bi-allelic alterations identifies a Lynch phenotype in The Cancer Genome Atlas. *Genome Med.* **10**, 69, doi:10.1186/s13073-018-0579-5 (2018).

75 Menoyo, A. *et al.* Somatic mutations in the DNA damage-response genes ATR and CHK1 in sporadic stomach tumors with microsatellite instability. *Cancer Res.* **61**, 7727-7730 (2001).
76 Baudrin, L. G., Deleuze, J.-F. & How-Kit, A. Molecular and Computational Methods for the Detection of Microsatellite Instability in Cancer. *Front. Oncol.* **8**, 621, doi:10.3389/fonc.2018.00621 (2018).
77 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695 (2012).
78 Girgis, H. Z. & Sheetlin, S. L. MsDetector: toward a standard computational tool for DNA microsatellites detection. *Nucleic Acids Res.* **41**, e22, doi:10.1093/nar/gks881 (2013).
79 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
80 Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583-2585, doi:10.1093/bioinformatics/btx198 (2017).
81 Hempelmann, J. A. *et al.* Microsatellite instability in prostate cancer by PCR or next-generation sequencing. *J Immunother Cancer* **6**, 29, doi:10.1186/s40425-018-0341-y (2018).
82 van der Velde, K. J. *et al.* GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biol.* **18**, 6, doi:10.1186/s13059-016-1141-7 (2017).
83 Itan, Y. *et al.* The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat. Methods* **13**, 109-110, doi:10.1038/nmeth.3739 (2016).
84 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).
85 Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31-34, doi:10.1038/nprot.2007.13 (2007).
86 Lea, D. E. & Coulson, C. A. The distribution of the numbers of mutants in bacterial populations. *J. Genet.* **49**, 264-285, doi:10.1007/BF02986080 (1949).
87 Hosack, D. A., Dennis, G., Jr., Sherman, B. T., Lane, H. C. & Lempicki, R. A. Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4**, R70, doi:10.1186/gb-2003-4-10-r70 (2003).
88 Lea, D. E. & Coulson, C. A. The distribution of the numbers of mutants in bacterial populations. *J Genet* **49**, 264-285, doi:10.1007/BF02986080 (1949).
89 Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747-1756, doi:10.1101/gr.239244.118 (2018).