Genome-wide mapping of spontaneous DNA replication error-hotspots using mismatch repair proteins in rapidly proliferating Escherichia coli

Flavia C Hasenauer, Hugo C Barreto, Chantal Lotton, Ivan Matic

▶ **To cite this version:**

Flavia C Hasenauer, Hugo C Barreto, Chantal Lotton, Ivan Matic. Genome-wide mapping of spontaneous DNA replication error-hotspots using mismatch repair proteins in rapidly proliferating Escherichia coli. Nucleic Acids Research, 2024, 10.1093/nar/gkae1196 . hal-04866534

HAL Id: hal-04866534
https://hal.science/hal-04866534v1

Submitted on 6 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Genome-wide mapping of spontaneous DNA replication error-hotspots using mismatch repair proteins in rapidly proliferating *Escherichia coli*

Flavia C. Hasenauer[1#], Hugo C. Barreto[1#], Chantal Lotton[1#], Ivan Matic[1*]

[1]Université Paris Cité, CNRS, Inserm, Institut Cochin, F-75014 Paris, France

[#]These authors contributed equally to this work

*To whom correspondence should be addressed

Email address of corresponding author : ivan.matic@inserm.fr

ABSTRACT

Fidelity of DNA replication is crucial for the accurate transmission of genetic information across generations, yet errors still occur despite multiple control mechanisms. This study investigated the factors influencing spontaneous replication errors across the *Escherichia coli* genome. We detected errors using the MutS and MutL mismatch repair proteins in rapidly proliferating *mutH*-deficient cells, where errors can be detected but not corrected. Our findings reveal that replication error hotspots are non-randomly distributed along the chromosome and are enriched in sequences with distinct features: lower thermal stability facilitating DNA strand separation, mononucleotide repeats prone to DNA polymerase slippage, and sequences prone to forming secondary structures like cruciforms and G4 structures, which increase likelihood of DNA polymerase stalling. These hotspots showed enrichment for binding sites of nucleoid-associated proteins, RpoB and GyrA, as well as highly expressed genes, and depletion of GATC sequence. Finally, the enrichment of single-stranded DNA stretches in the hotspot regions establishes a nexus between the formation of secondary structures, transcriptional activity, and replication stress. In conclusion, this study provides a comprehensive genome-wide map of replication error hotspots, offering a holistic perspective on the intricate interplay between various mechanisms that can compromise the faithful transmission of genetic information.

INTRODUCTION

Genomes constitute intricate ensembles of complex, interdependent, and highly coordinated genetic information. Consequently, many of the newly generated mutations are deleterious, disrupting the harmonious functioning of cells and thereby affecting growth, survival, and reproduction (1). This underscores the critical importance of accurate genome replication for the faithful transmission of genetic information across generations. For this reason, replicative DNA polymerases have evolved with high insertion accuracy and exonucleolytic proofreading to execute high-fidelity DNA replication (2). Nonetheless, despite efficient replication fidelity control mechanisms, replication errors can still occur. At a low frequency, DNA polymerases incorporate noncomplementary nucleotides, leading to the formation of mismatched base pairs. DNA polymerases can also generate errors by slipping on repetitive nucleotide sequences, resulting in the formation of single-strand loops on either the newly synthesized or template strand. Both types of errors, i.e., mispaired and unpaired bases, may respectively result in substitution and insertion/deletion mutations in the subsequent round of DNA replication if not repaired beforehand. However, effective mismatch repair systems, conserved in all domains of life, are capable of correcting nearly all of these errors. For example, *Escherichia coli* mismatch repair system corrects 99 % of DNA replication errors, thereby reducing replication error rates to approximately 2 per 10^{10} base incorporation events (3).

E. coli mismatch repair primarily involves three dedicated proteins: MutS, MutL, and MutH (4, 5). When replication errors occur, the MutS protein homodimer binds to mismatches, except for C-C mismatches and insertion/deletion loops larger than 4 bases. This process is facilitated by MutS's direct associations with the replisome, specifically through interaction with the β-sliding clamp (DnaN protein dimer) (6). This

interaction is essential for *in vivo* mismatch repair activity (7). Upon binding to a mismatch, the MutS protein homodimer binds ATP and undergoes a conformational change, generating the MutS sliding clamp. Next, the MutS sliding clamp dissociates from the mismatch and diffuses bidirectionally along the adjacent DNA, recruiting a MutL protein homodimer onto the DNA. ATP binding-dependent dimerization of MutL results in the formation of MutL sliding clamps, which may dissociate from MutS sliding freely along the DNA or remain in a MutS-MutL sliding clamp complex. Subsequently, the MutH endonuclease binds to MutL clamps. MutH ensures the fidelity of the repair process by recognizing and incising the first encountered unmethylated GATC site on the newly synthesized strand. The incision site can be located on the 3′ or 5′ side of the mismatch and may be up to several thousand base pairs away from the mismatch (5). Helicase II (UvrD) then unwinds the DNA duplex starting from the incision site, allowing the degradation of the displaced single-strand DNA by different exonucleases and ensuring the irreversibility of the repair process. The repair process is finalized by the replicative DNA polymerase III and DNA ligase, which fills the gap and seals the nick in newly synthesized DNA, respectively. Inactivation of any of the three genes coding for the dedicated mismatch repair in *E. coli*, *mutS*, *mutL*, and *mutH*, results in up to 100-fold increase in spontaneous mutation rates (8).

DNA replication errors constitute a significant source of spontaneous mutations (9, 10). Beyond the inherent limitations of its fidelity control mechanisms, DNA replication fidelity is influenced by various factors, including local sequence context, DNA topology, balance of dNTP pools, conflicts between replication and transcription, and other DNA transactions such as recombination, repair, and translesion synthesis processes (11). Additionally, endogenous mutagens like reactive oxidative species, along with spontaneous chemical modifications of bases such as depurination, can

result in alterations or destruction of their coding information, thereby inducing DNA replication errors. If not repaired, these errors end up causing mutations.

Mechanisms by which spontaneous mutations arise have been studied using different mutation-detecting assays that employ two main methodologies: phenotypic assays and DNA sequencing (12–14). Phenotypic assays identify mutations by observing altered phenotypes resulting from mutations in target genes, while DNA sequencing detects genome-wide mutations. However, the reliability of phenotypic assays is weakened by several factors, including the small size of mutation targets, bias towards specific mutations in a target gene, distortions in mutation frequency caused by fitness effects, and the prevalence of neutral mutations lacking observable phenotypes (15). Whole genome sequencing allows for the avoidance of small target gene biases, but it is generally still affected by fitness-related distortions in mutation frequencies. The most reliable approach for estimating mutation rates is the combination of Mutation Accumulation (MA) with whole genome sequencing, as this method minimizes the influence of fitness effects (3). However, all these assays have limitations. They do not provide information on when a mutation occurs or the state of the cell at the time of mutation emergence. Furthermore, these methods cannot directly quantify replication errors on a *per* cell division or *per* chromosome replication basis. Consequently, such values can only be inferred based on assumptions, which are inherently subject to approximations.

We have addressed these limitations and enhanced the accuracy and reliability of replication error detection in *E. coli* by harnessing the robust natural replication error detection system: the mismatch repair system (16, 17). By tagging MutL protein with a fluorescent protein, we are able to monitor DNA replication errors, which appear as fluorescent foci, in individual living cells. Previously, we have demonstrated that

inactivating the *mutS* gene results in the complete disappearance of MutL fluorescent foci, confirming that these foci are not aggregates of the MutL protein fusions independent of mismatches (10, 16). This assay allows for the determination of the exact moment when a replication error occurs within a single cell, as the MutL fluorescent focus remains visible for a limited time before disappearing as the subsequent DNA replication cycle separates the two DNA strands, effectively fixing the mutations. Importantly, MutL fluorescent foci allow the detection of DNA replication errors irrespective of their future potential impact on phenotype—be it beneficial, neutral, deleterious, or even lethal. Finally, inactivation of the *mutH* gene enables the detection and binding of the MutS and MutL proteins to mismatches but prevents repair from progressing to completion, enabling us to detect nearly all DNA replication errors using MutL-based mutation assay (10, 16). Importantly, while methylation status of GATC sequences is important for strand discrimination, it is irrelevant for the mismatch detection and binding by the MutS and MutL proteins.

Therefore, unlike the above-mentioned commonly used mutation detection assays, which typically have limited time resolution, MutL-binding enables direct detection of DNA replication errors during each individual genome replication. Furthermore, it allows for the detection of emerging mutations, *i.e.,* DNA replication errors before fixation, in contrast to other assays, which only detect fixed mutations that have escaped multiple error-correcting mechanisms. We took advantage of this assay to perform genome-wide mapping of sites of spontaneous DNA replication errors in *E. coli*. We performed a CFP-MutL-Chromatin Immunoprecipitation followed by high-throughput sequencing (ChIP-seq) in populations of *mutH*-deficient cells growing exponentially under optimal conditions without exogenous stress. In these cells, mismatches are detected and tagged, but

cannot be repaired. Genomic regions preferentially enriched by the CFP-MutL ChIP-seq [further in the text: MutL-associated regions (MutL-ARs)] were analyzed for their localization within genes/intergenic regions, replichores, and genome macrodomain regions (18). Additionally, we determined for these regions the GC content, thermal stability, presence of the single-nucleotide and larger repetitive sequences, density of the GATC sites, and the presence of G-quadruplex (G4)-prone sequences (19). Data from published studies were also incorporated in our analysis to investigate whether MutL-ARs were enriched for HupA, HupB, Fis, and the histone-like nucleoid structuring (HN-S) proteins binding sites (20), GapR binding sites (21), ssDNA gaps (22), GATC methylation (23), data from RNA-Seq (24), RNA polymerase subunit (RNAP) RpoB and GyrA binding sites (20), and G4-prone sequences (25). The goal of this study was to identify genomic regions with particular properties and *cis*-acting mechanisms that render some genome regions DNA replication error hot spots.

MATERIALS AND METHODS

Strain construction

Strains used in this study were derived from the *E. coli* K-12 MG1655 strain. For the microscopy analysis, we used the *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-CFP-*mutL*::FRT Δ*mutH*::FRT *dnaN*-mCherry::FRT and the Δ*mutS*::kanamycin-resistance cassette *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-CFP-*mutL*::FRT Δ*mutH*::FRT strains. For the ChIP experiment, we used the *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-CFP-*mutL*::FRT Δ*mutH*::FRT strain. As a control for antibody specificity, we used the MG1655 strain and its Δ*mutS*::streptomycin/spectinomycin-resistance cassette and *mutL218*::Tn*10* Δ*mutH*::kanamycin-resistance cassette derivatives. The construction of the CFP-*mutL*::chloramphenicol-resistance cassette-reporter fusion and the procedure to integrate the MutL reporter fusions into the chromosome under the control of the lactose promoter were previously described in (17). The *mutL*218::Tn*10* (Tn*10* provides tetracycline resistance), Δ*mutH*::kanamycin-resistance cassette, Δ*mutS*::kanamycin-resistance and Δ*mutS*::streptomycin/spectinomycin-resistance cassette alleles are from our laboratory collection. The *dnaN*-mCherry::kanamycin-resistance cassette-reporter fusion was kindly provided by C. Lesterlin (26). These alleles and reporter fusions were introduced into the final strains using P1 transduction (27). Transductants were selected based on appropriate antibiotic resistance. When antibiotic resistance cassettes were flanked by FRT (Flippase Recognition Target) sites, we used Flp recombinase enzyme to remove the resistance cassettes, leaving behind a single FRT "scar" (28).

Growth Conditions

For microscopy and ChIP-Seq assays, bacterial cultures were initiated from glycerol stocks and grown at 37°C in LB medium supplemented with 0.1mM isopropyl-β-D-thiogalactopyranoside (IPTG), with shaking at 150 rpm. The following day, overnight cultures were diluted 1/400 (v/v) in LB supplemented with 0.1 mM IPTG, and incubated at 37°C with shaking until reaching the exponential growth phase (OD_{600nm} = 0.200 to 0.215).

Microscopy and image analysis

Microscopy experiments were carried out as described previously (10). Briefly, exponential phase cultures were washed with a 10^{-2} M $MgSO_4$ solution and inoculated on microscope chamber slides composed of a 10^{-2} M $MgSO_4$ 1.5% agarose matrix. Cells were observed using a 100× objective lens on an Axiovert 200 inverted microscope (Carl Zeiss) equipped with a Photometrics CoolSNAP camera (Princeton Instruments). Images were captured and analyzed using following softwares: MetaMorph version 7.10.2.240 and ImageJ version 1.52C with MicrobeJ version 5.12d plugin. We quantified number of replication forks per cell using *dnaN-mCherry* reporter. Due to a large number of replication forks in rapidly growing cells, it is likely that several forks can be present within one fluorescent spot. However, because chromosomal replication in *E. coli* is highly coordinated in both time and space, the number of replication forks per cell can easily be calculated from the number of DnaN-tagged forks, even if all of them cannot be visualized individually. Therefore, we used the following criteria to determine subpopulations of cells having different growth rates: (i) Cells without DnaN foci are nongrowing cells. (ii) The presence of one DnaN focus means that there are two forks in cell, which occurs in slow-growing cells with an origin/terminus ratio of 2. (iii) If 3 to 6 DnaN foci are visible, there are 6 forks per cell,

which is typical of fast-growing cells with an origin/terminus ratio of 4. (iv) If 7 or more DnaN foci are visible, there are at least 14 forks per cell, which happens in growing cells where origin/terminus ratio is 8, but also in cells suffering from endogenous stressors that perturb DNA replication and cell division, such as those that induce SOS response (29, 30). Cells were considered as filamenting if their length was higher than 2*Median of all observed cells.

ChIP assay

Samples of bacterial cells were processed following the protocol published by Diaz *et al.* (31). Briefly, cells from exponential phase cultures of the *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-CFP-*mutL*::FRT Δ*mutH*::FRT strain were crosslinked with 1% formaldehyde for 5 min (22.5°C, 90 rpm) and subsequently quenched with 0.5 M glycine for 10 min (22.5°C, 90 rpm). Cell pellets were collected by centrifugation (15 min, 4°C, 4000 rpm), washed three times with cold phosphate buffered saline (PBS), and then resuspended in 250 µL of CHIP BUFFER (0.2 M Tris, 1 mM EDTA, protease inhibitor cocktail, 0.5 % SDS). To shear the chromosomes, cells were sonicated using a Diagenode Bioruptor at 30-second intervals for 7 cycles, keeping samples stables at 4°C using a temperature-controlled water bath. After sonication, cell debris was discarded by centrifugation (10 min, 4°C, 14000 g). The supernatant was diluted in dilution buffer (0.1 mM EDTA, 250 mM NaCl and protease inhibitor cocktail) to obtain a SDS concentration lower than 0.2 %, according to the antibody manufacture recommendations. A total of 50 µL was taken out prior to the addition of the antibody as an Input control. Immunoprecipitation (IP) was performed using alpaca anti-CFP nano-antibody fusion to magnetics beads, a ChIP-validated antibody (32) (GFP-Trap® Magnetic Particles M-270 CHROMOTEK, ProteinTech), at a concentration of 12.5 µL of beads per 5 mL of culture. IPs were

rotated overnight at 4 °C. After incubation with the antibody, the beads were magnetically collected and several washes were performed. First, the beads were washed with 300 µL of 2× IP BUFFER (50 mM Hepes-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, 1 mM PMSF) by incubating at 4°C for 10 min with end-over-end rotation. Two additional washes were performed using WASH BUFFER 2 (50 mM Hepes-KOH pH 7.5, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, 1 mM PMSF). Afterwards, an additional washing was performed using the WASH BUFFER 3 (10mM Tris–HCl pH 7.8, 250 mM LiCl, 1Mm EDTA, 0.5% IGEPAL CA-630 (Sigma-Aldrich), 0.1% sodium deoxycholate, 1 mM PMSF), and the final wash was with WASH BUFFER TE (10 mM Tris pH 7.5, 1 mM EDTA). After the washes, 300 µL of elution buffer (50 mM Tris–HCL pH 7.5, 10 mM EDTA, 1% SDS) was added, and the beads were incubated at 65°C for 20 minutes with shaker. The eluates of the IP and the Input were then treated with RNase A (10 mg/mL) and proteinase K (10mg/mL) at 37°C for 1 hour with shaking, followed by overnight incubation at 65°C with gentle agitation. This step at 65 °C is optimal for both the function of proteinase K and reversal of cross-linking. The following day, the supernatant was magnetically collected and DNA was purified using the Quick Kit (Qiagen) according to the manufacturer's instructions. The quality of the extracted DNA was assessed using a NanoDrop spectrophotometer and an Agilent 2100 Bioanalyzer. The quantity of DNA was measured using a Qubit 3 Fluorometer (Life Technologies). As a control for antibody specificity, cultures of the MG1655 strain and its Δ*mutS* and Δ*mutL* Δ*mutH* derivatives lacking the fluorescent epitope were also submitted to ChIP, and the DNA obtained from IP and Input samples was quantified following the same protocol used for the *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-CFP-*mutL*::FRT Δ*mutH*::FRT strain. As expected, no DNA was detected in the IP of

241 the control samples. Only Input and IP samples from the *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-
242 CFP-*mutL*::FRT Δ*mutH*::FRT strain were used for library preparation and sequencing.
243

244 **Library preparation and sequencing for ChIP-seq**

245 Library preparation for ChIP-seq was performed using the True Seq Nano DNA library
246 prep kit with IDT-ILMN True Seq DNA UD indexes (96ind) (Illumina) according to the
247 manufacturer's instruction. ChIP-seq samples were sequenced using the Illumina
248 NextSeq 500/550 Sequencing system at the Biomics Platform, C2RT, Institute
249 Pasteur, Paris, France. Each sample was pair-end sequenced, producing datasets of
250 paired-end 75 bp read pairs.
251

252 **ChIP-seq bioinformatic analysis**

253 Sequencing adapters were removed using version 0.23.4 of *fastp* (33) and raw reads
254 were trimmed bidirectionally by 4 bp window sizes across which an average base
255 quality of 20 was required to be retained. Further retention of reads required at least
256 50% base pairs with phred scores at or above 20. The reads were then mapped to
257 the *E. coli* K-12 MG1655 genome (accession number: NC_000913.3) using version
258 0.7.17-r1188 of *bwa* (34) with default parameters, followed by sorting using version
259 1.19 of *samtools* (35). Identification of MutL-ARs was performed using three different
260 software tools: version 3.0.0 of *MACS3* (36) with default parameters except a cutoff of
261 0.05 and a genome size of 4641652; version 2.40.0 of *mosaics* (37) with default
262 parameters except for a False Discovery Rate threshold of 0.05; and version 0.0.52 of
263 *epic2* (38) with default parameters except for a genome size of 4641652. For all three
264 software tools, each IP sample was normalized to Input samples. The final list of MutL-
265 ARs **(Sup. Table 1)** included only the overlapping regions detected by at least two of

the software tools and with a size lower than 15000 bp. Log2 ratio (IP/Input) after Signal Extraction Scaling (SES) was obtained using *deepTools* tool *bamCompare* (version 3.5.4) (39) with a bin size of 200 bp.

Provenance of data tested for association with MutL-ARs

HupA, HupB, Fis, H-NS, RpoB, and GyrA binding sites were obtained from proChIPdb v1.0.0 (20). GapR binding sites were obtained from Guo *et al.* (21). ssDNA regions were obtained from Pham *et al.* (40). RNA-seq data was obtained from Niccum *et al.* (24). G-quadruplex sequences were obtained from Kaplan *et al.* (25) and identified using G4Hunter (19). Macrodomain regions were obtained from Espeli *et al.* (18). GC content and melting temperature were calculated using 200 bp bins. Melting temperature was estimated with nearest neighbor thermodynamics described by Breslauer *et al.* (41) using the R-package *TmCalculator*. Microsatellites and GATC regions in the *E. coli* genome (accession number: NC_000913.3) were identified using the R-package *Biostrings* version 2.70.1. GATC methylation was obtained from Cohen *et al.* (23). Cruciform prone sequences were identified using CIRI (42).

Gene enrichment analysis

Gene enrichment analysis was performed using the EcoCyc database version 27.5 (43). EcoCyc, which is regularly updated by manual curation, provides enrichment analysis for Gene ontologies, pathways, and transcriptional regulators. As an input for this analysis, we considered the genes within the peak regions detected by the ChIP-seq analysis **(Supp. Table 2)**. For pathways and transcriptional regulators enrichment, the reference gene set used was all genes assigned to any metabolic pathway or directly regulated by a transcriptional factor in *E. coli*. For Gene ontologies, the

reference gene set used was genes with assigned Gene ontology terms. Statistical significance was assessed using a Fisher's exact test, followed by Benjamin-Hochberg correction for multiple comparisons. Only adjusted p-values below 0.05 were considered as statistically significant.

Statistical analysis

Statistical analysis were performed in the statistical software R version 4.3.0 using the R-packages *regioneR* (44) and *rstatix*. Permutation test using the R-package *regioneR* was performed with 1000 permutations, the randomization function *randomizeRegions*, the evaluation function *numOverlaps*, and *E. coli* genome NC_000913.3 as input genome. The parameter count.once was defined as TRUE for testing the against regions with a region size larger than the smallest MutL-AR. For each permutation a new set of MutL-ARs is created, with the exact same size as the original MutL-ARs but randomly distributed in the *E. coli* genome, followed by calculation of the number of overlaps between the criteria tested and the new set of MutL-ARs. Finally, the distribution for the number of overlaps between the MutL-ARs and the regions of interest obtained in the permutations is compared with the observed for the original MutL-ARs. The strength of the association (Z-score) is calculated as the number of standard deviations between the expected mean overlap between the MutL-ARs and the regions of interest obtained from the permutations and the observed overlap between the MutL-ARs and the regions of interest. Principal Component Analysis (PCA) was performed with scaling, using as input data the relative coverage of proteins binding sites in each MutL-AR and the GC content and melting temperature calculated for each MutL-AR. The statistical tests used are indicated in the main text

315 and/or in the figure legends. Statistical significance was defined for $p < 0.05$ in all
316 analysis and calculated as described in the main text and/or figure legends.

RESULTS AND DISCUSSION

Visualizing and quantifying replication errors in individual cells

We first quantified cell-to-cell heterogeneity in DNA replication errors using a fully functional CFP-MutL translation fusion reporter and fluorescent microscopy. We analyzed asynchronous populations of the *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-CFP-*mutL*::FRT Δ*mutH*::FRT *dnaN*-mCherry::FRT that were growing exponentially under optimal conditions in rich LB medium without exogenous stress. Asynchronous bacterial populations may contain cells with different growth rates, resulting in different numbers of replication forks per cell. Although the number of DNA replication errors per cell may vary depending on the number of replication forks, we decided not to synchronize DNA replication using standard methods, such as using replication initiation proteins temperature-sensitive mutants or treatment with hydroxyurea, because they are highly invasive and can disrupt normal cellular functioning. Instead, we opted to visualize DNA replication forks in individual cells using a fluorescently tagged mCherry-DnaN protein. DnaN, the β subunit of DNA polymerase III holoenzyme, forms dimers to create the sliding clamp, facilitating the linkage of the core polymerase to DNA and enabling rapid, processive DNA replication (45). Utilizing this replication fork reporter enabled us to identify subpopulations of cells based on the number of replication forks, and therefore to determine the number of replication errors per cell within each subpopulation.

E. coli initiates bidirectional chromosome replication from a single replication origin (*oriC*) and completes chromosome replication within a termination zone (*ter*). In slowly growing cells, where the *oriC*/*ter* ratio is 2, two replication forks per cell are typically observed. However, in rapidly growing cells, new rounds of replication start before the initial two replication forks reach the *ter* region. Consequently, fast-growing

cells, which divide every 20 minutes, exhibit an *oriC*/*ter* ratio of 4 and possess six replication forks per cell (29). Finally, some cells may harbor more than six replication forks per cell, such as those that are exposed to endogenous stressors. We analyzed 93,861 cells from 53 independent experiments and quantified the number of mCherry-DnaN and CFP-MutL fluorescent foci per cell **(Fig. 1A, Supp. Table 3)**. We found that 0.6%, 6.3%, 78.9%, and 14.2% of cells harbored 0, 2, 6, and greater than or equal to 7 replication forks per cell, respectively **(Fig. 1B)**. Within these four categories, 17.6%, 38.8%, 44.8%, and 55.4% of cells exhibited at least one CFP-MutL focus, respectively **(Fig. 1C)**. We were able to detect up to 10 CFP-MutL foci per cell. Taken together, our data showed that the majority of the cells were rapidly growing, with an *oriC*/*ter* ratio of 4, and that about 46% of all cells had at least one replication error under our experimental condition **(Fig. 1D)**. The distribution of the observed number of MutL foci per cell in the total population reasonably fits a Poisson distribution calculated from the measured number of MutL foci per cell **(Fig. 1D)**. The mean number of MutL foci per cell was 0.6. However, there was a small excess of cells (469 out of 93,861 cells) observed with four or more MutL foci compared to what would be predicted by the Poisson distribution **(Fig. 1E)**. This suggests that there may be an increased likelihood of multiple errors occurring in a small subpopulation of individual cells beyond what would be expected purely from a uniform Poisson distribution of events. We cannot exclude the possibility that, with light microscopy, we may not be able to distinguish between two or more closely positioned foci. Consequently, cells with multiple MutL foci may occur more frequently than observed. Moreover, while this study examined cells growing under optimal conditions, it does not mean that detected spontaneous DNA replication errors were not caused by endogenous stresses that can both damage DNA and impede DNA replication machinery (10). This hypothesis is supported by our

finding of a significantly higher frequency of filamentation in cells with four or more MutL foci (11.5%) compared to cells with fewer than four MutL foci (1.5%) (Proportion test, $p = 9.2 \times 10^{-66}$). Filamentation is known to be induced by various stress conditions (46) that also trigger the SOS response (47). The induction of the SOS response could potentially lead to an increased number of MutL foci due to elevated mutation rates resulting from the activation of error-prone DNA polymerases.

We have previously tested the possibility that MutL fused to different fluorescent proteins might interact with DNA structures other than mismatches by investigating whether MutL fluorescent foci are still present in Δ*mutS* cells. For this we used MutL-eGFP and MutL-mCherry reporters and we found no MutL fluorescent foci in Δ*mutS* cells (10, 16). In the present study, we tested if this is also the case with MutL-CFP and found no foci in 1113 Δ*mutS* Δ*mutH* cells. Therefore, it is unlikely that our results are affected by non-specific binding of MutL-CFP to non-mismatch DNA structures.

Identification of the MutL-ARs

To pinpoint the genome-wide locations of the MutL-ARs, we conducted CFP-MutL-ChIP-seq experiments in the *mutL218*::Tn*10* Δ*lacZ*::P_{lac}-CFP-*mutL*::FRT Δ*mutH*::FRT cells that were growing exponentially under optimal conditions in rich LB medium without exogenous stress. Immunoprecipitation was performed using alpaca anti-CFP nano-antibody, and the enriched DNA samples were sequenced using the Illumina NextSeq technology. We performed two independent experiments. The identification of the MutL-ARs was conducted using three different software tools: MACS3 (36), Mosaics (37), and Epic2 (38). Analysis using only the MACS software tool, which is the most frequently used to detect ChIP-Seq peaks, provided by far the highest number of identified peaks **(Fig. 2A-B)**. However, the use of a single software to detect ChIP-

Seq peaks may lead to the detection of false positives, which could affect the downstream analysis. To mitigate potential false positives inherent in relying solely on one software, and to identify the strongest replication error-prone regions in the genome, we retained for further analysis only those MutL-ARs that were identified by at least two software tools. Although this conservative approach results in a lower number of detected ChIP-Seq peaks, it provides higher confidence that the MutL-ARs detected are genomic regions strongly bound to MutL. Using this stringent criterion, a total of 38 MutL-ARs **(Fig. 2A-B, Supp. Table 1)** were identified, with a median size of 2463 bp **(Supp. Fig. 1)**. From the 38 MutL-ARs, 21 overlapped between the two independent experiments **(Supp. Table 1)**.

The size and shape of the MutL-ARs peaks detected by CFP-MutL-ChIP-seq can be influenced by several factors. *In vitro* studies have shown that MutS and MutL proteins scan DNA bidirectionaly, both from the 5' to 3' and from the 3' to 5', from a mismatch. When these proteins encounter a hemimethylated GATC sequence, the MutH protein cuts the unmethylated strand. The distance between mismatches and MutH incision sites has been reported to extend up to 2 kb *in vitro* (48, 49) and up to approximately 6 kb *in vivo* in wild-type cells (50). *In vivo*, this distance is also determined by an outcome of the competition between Dam methylase and MutH protein. However, in the absence of MutH protein, which is the case in our study, the distance between a mismatch and hemimethylated GATC sites should not affect the size of the MutL-ARs. In addition, when MutS and MutL scanning is not stopped at hemimethylated GATC sites, it can even continue beyond. This is exemplified by our previous observation that the introduction of the MutHE56A mutant protein, which acts as a steric block on the DNA by binding to hemimethylated GATC sites without cleaving them in *mutH*-deficient cells, causes a decrease in the amount of MutL in fluorescent

417 foci (17). Finally, closely located multiple mismatches can generate overlapping repair
418 events, thus also generating long MutL-ARs.

419 The Log2 ratio (IP/Input) distribution of MutL-ARs exhibited diverse patterns,
420 including narrow, broad, and multiple interconnected peaks **(Fig 2C)**. Single peaks
421 likely result from the migration of successive MutS sliding clamps followed by MutL
422 sliding clamps from a single mismatch bidirectionally towards hemimethylated GATC
423 sequences **(Fig. 2C)**. Overlapping peaks likely emerge in replication error-prone
424 regions, where numerous errors at different sites are anticipated to occur. In a
425 population of asynchronously replicating cells, the genomic regions with the highest
426 Log2 ratio (IP/Input) are expected to occur around individual replication error hotspots.
427 This allowed us to narrow down the locations of replication errors within the larger
428 regions of the MutL-ARs **(Fig. 2C)**.

429 Finally, we observed that nine MutL-ARs peaks exhibited asymmetry, with
430 higher levels observed towards *oriC* compared to *ter* **(Fig. 2C)**. While *in vitro* studies
431 suggest that MutL-ARs peaks should be symmetrical on both sides of mismatches
432 (51), *in vivo* results show that the mismatch repair machinery preferentially searches
433 for the hemimethylated GATC site located closer to the replication fork, which
434 advances toward the *ter* region, and that DNA degradation proceeds back toward the
435 mismatch and *oriC* (50). Therefore, the asymmetry in some of the MutL-ARs observed
436 in our study could be explained by this directional search for the hemimethylated GATC
437 sites. This directionality may stem from the interaction of the MutS and MutL proteins
438 with the DNA polymerase β-sliding clamp, which is essential the for *in vivo* mismatch
439 repair activity (7).
440

Global chromosome structure and nucleoid-associated proteins in the MutL-ARs

The *E. coli* chromosome is organized into four distinct insulated macrodomains (MD): Ori, Right, Left, and Ter, and two less constrained regions: nonstructured right (NSR) and left (NSL) (18, 52). We examined the distribution of MutL-ARs within chromosomal MD **(Fig. 3A-B)**. We found that the highest number of MutL-ARs was detected in the Left MD (11 regions), followed in descending order by NSL (7 regions), Ori (7 regions), Right (6 regions), NSR (4 regions), and Ter (3 regions) MDs **(Supp. Fig. 2A)**. We also calculated what fraction of each MD is covered by the total cumulative length of MutL-ARs and found the highest fraction in the NSR MD, followed in descending order by Ori, Left, NSL, Right, and Ter MDs **(Supp. Fig. 2B)**.

The MutL-ARs peaks were identified based on the Log2 ratio (IP/Input), which normalizes the number of ChIP-seq reads in IP *versus* Input samples across all chromosomal positions. Therefore, the observed difference in the number of MutL-ARs between different MDs should not be a direct result of varying DNA amounts in different MDs, such as the descending amounts of DNA from *oriC* towards the *ter* region in replicating genomes. This normalization approach suggests that other factors or mechanisms may be influencing the enrichment of MutL-ARs in different chromosomal regions (53, 54).

Chromosome folding and compaction result from a combination of processes, including the binding of nucleoid-associated proteins (NAPs). NAPs constitute a diverse class of proteins capable of wrapping, bridging, or bending DNA. They play roles in both nucleoid structuring and transcription regulation (55). The binding of NAPs can protect DNA from mutagenic processes, but it can also remodel DNA topology, thereby rendering DNA more prone to mutations (56). Their impact on gene regulation

can also indirectly affect mutation generation, as transcription has been found to be associated with mutagenesis (57). We focused on key *E. coli* nucleoid-associated proteins (NAPs): HupA and HupB, the subunits of the HU protein complex that plays a crucial role in chromosome organization and compaction; Fis protein, which is involved in shaping the nucleoid structure, forming topological domain barriers, and regulating gene expression; and H-NS protein that contributes to chromosomal structuring and acts as a repressor of gene expression. Highlighting that the binding of these proteins may impact mutagenesis, we found that HupA (Permutation test, $p = 0.001$, Z-score = 6.1) and HupB (Permutation test, $p = 0.001$, Z-score = 7.9), Fis (Permutation test, $p = 0.001$, Z-score = 5.3), and HN-S (Permutation test, $p = 0.004$, Z-score = 4.2), binding sites are enriched in the MutL-ARs **(Fig. 3A-B, Supp. Fig. 2C-F).**

It was previously observed that the mutational density in *E. coli* mismatch repair deficient strain was high in regions of the chromosome where gene expression is responsive to NAPs (58). The authors of that study interpreted this enrichment as a consequence of modified chromosome structure rather than being directly related to gene expression itself. Similarly, the enrichment of NAPs binding sites in the MutL-ARs may also result from the role of NAPs' in shaping nucleoid structure, contributing to the increased replication errors. However, we cannot fully exclude that NAPs impact on gene expression may also impact replication fidelity. Direct evidence of the exact mechanisms by which NAP binding may lead to mutagenic consequences requires further research.

Finally, to enhance our understanding of MutL-ARs localization, we also used data on the distribution of DNA positive supercoiling, which accumulates ahead of the replication and transcription complexes (21, 59). In these studies, the *E. coli* genome was mapped using the *Caulobacter crescentus* protein GapR as a probe, which binds

to positive supercoiled regions. We found that GapR binding sites are enriched in the MutL-ARs (Permutation test, p = 0.005, Z-score = 2.7) **(Fig. 3A-B, Supp. Fig. 2G).** While the direct relationship between positive DNA supercoiling and replication fidelity is not fully elucidated, it's clear that the topological state of DNA may impact the replication process by causing local DNA structural changes that affect fork progression (60).

Local DNA sequence properties of the MutL-ARs

Macromolecular machines such as DNA and RNAP must separate two DNA strands to synthesize new strands of DNA or RNA during replication or transcription. This process is facilitated by dedicated proteins like helicases, as well as by the local sequence composition, *i.e.*, AT-rich segments are less stable than GC-rich segments, which determines the stability of double-stranded DNA (dsDNA). For example, *E. coli* genome replication initiation occurs in the AT-rich *oriC* sequence (61). It was also reported that the local DNA thermodynamic stability varies significantly along the *E. coli* chromosome and that the gradient of DNA thermodynamic stability correlates with the polarity of chromosome replication (62). So, we investigated whether the local base pair composition influences the localization of DNA replication errors detected by our MutL-ChIP assay **(Fig. 4A-B)**. We found that the GC content and the melting temperature in the MutL-ARs are significantly lower when compared to the whole genome (Wilcoxon test, $p = 2.6 \times 10^{-19}$ for GC content, $p = 1.3 \times 10^{-40}$ for melting temperature) **(Fig. 4A-B, Supp. Fig. 3A-B)**, suggesting that DNA replication errors occur more frequently in regions with relatively lower thermodynamic stability. We found that HN-S binding sites, which are AT-rich (63), are highly enriched in the MutL-ARs **(Fig. 3A-B, Supp. Fig. 2F)**.

Direct repeats of short DNA sequences, known as microsatellites, are highly prone to mutations due to DNA polymerase slippage during replication. The slippage process happens when the nascent DNA strand dissociates from the template strand and realigns incorrectly during DNA synthesis. This misalignment can result in the insertion or deletion (indel) of repeat units, leading to changes in the number of repeats. It was previously reported that the rate of indel formation at direct mononucleotide repeats in *E. coli* genome increases exponentially with the length of the repeat, with any repeat of 4 nucleotides or more being a potential indel hotspot (3). We observed a significant enrichment of mononucleotide repeats of 4 nucleotides or more in the MutL-ARs (Permutation test, p = 0.001, Z-score = 4.5) **(Fig. 4A-B, Supp. Fig. 3C)**. Given the efficiency of mismatch repair in detecting small loops resulting from misalignments in mononucleotide repeats, our result strongly validates the effectiveness of MutL ChIP-seq in detecting hotspots of DNA replication errors.

Inverted sequence repeats have the potential to generate secondary structures such as hairpins or cruciforms that cause the DNA polymerase to stall. Mismatch repair mechanisms can bind to these structures, particularly when small loops or base pair mismatches occur within the stem portion of the secondary structure. Our analysis showed an enrichment of the cruciform formation-prone sites in the MutL-ARs compared to the whole genome (Permutation test, p = 0.001, Z-score = 6.0) **(Fig. 4A-B, Supp. Fig. 3D)**. Importantly, as stated above, we also found that DNA replication errors occur more frequently in regions with relatively lower GC content and lower melting temperature **(Fig. 4A-B, Supp. Fig. 3A-B)**. As lower melting temperatures facilitate the formation of alternate secondary structures, this may help explain why regions containing direct and inverse DNA repeats are enriched in replication error-hotspots.

Besides lower thermodynamic stability, the formation of secondary structures such as hairpins or cruciforms requires the initial creation of the single-stranded DNA (ssDNA) stretches. For this reason, we investigated whether ssDNA gaps sites were enriched within MutL-ARs. For this analysis, we utilized data from a published study in which regions with high ssDNA coverage were identified using non-denaturing bisulfite treatment combined with deep sequencing of samples treated with RNAse (22). We found a significant enrichment of regions with high ssDNA coverage within MutL-ARs (Permutations test, p = 0.002, Z-score = 3.2) **(Fig. 4A-B, Supp. Fig. 3E)**. How are these MutL-ARs-associated ssDNA regions generated? ssDNA can arise as intermediate structures during DNA replication, recombination, repair, and transcription. Two key findings from the above-cited study support that normal replication activity does not generate detected ssDNA genomic hotspots (22). (i) No significant difference was observed in the amount of ssDNA between leading and lagging replication strands. (ii) The average length of ssDNA stretches was approximately 108 nucleotides, which is considerably shorter than the average length of Okazaki fragments, typically around 1,000-2,000 nucleotides long in *E. coli* (64). Furthermore, the lack of enrichment for RecA-binding sites in these ssDNA regions (40) suggests that RecA-mediated homologous recombination is also unlikely to be responsible for generation of the ssDNA. These findings point to other potential sources for the observed ssDNA regions, such as DNA repair processes or transcription-related events. However, regardless of how ssDNA stretches are generated, themselves or/and proteins involved in their processing may cause DNA replication errors detected by MutL-binding. Further research is needed to elucidate molecular pathways linking ssDNA formation, replication errors and mismatch repair activity *in vivo*.

Adenine methylation, or absence of methylation, within GATC sequences is used by *E. coli* mismatch repair system to discriminate between template and newly synthesized strands (5). For this reason, *E. coli* mismatch repair system is also called methyl-directed mismatch repair system. Adenines in GATC sites in newly synthesized DNA strands are transiently unmethylated because adenine methylation by Dam methylase lags several minutes behind replication, while those in template strands remain methylated. Given the importance of GATC sequences for the mismatch repair system, we compared the number of GATC sites present in the MutL-ARs and the whole genome. Interestingly, we found that GATC sequences are depleted in the MutL-ARs (Permutation test, $p = 0.001$, Z-score = -4.4) **(Fig. 4A-B, Supp. Fig. 3F)**. However, the observed depletion of GATC sequences in these regions is not expected to impact the efficacy of mismatch repair as the distances between different GATC sites vary from 4 bp to 4 kbp (65) but are present at an average every 2.5 kbp (66). Therefore, the mismatch repair machinery can efficiently span the mismatch site and the strand discrimination site, even across the widest spacing of GATC sequences. Importantly, it has also been observed that the majority of GATC sequences in *E. coli* genome are highly methylated during growth of *E. coli* (23), and we found no differences in the GATC methylation when comparing the MutL-ARs with the whole genome (Wilcoxon test, $p = 0.57$) **(Fig. 4A-B, Supp. Fig. 3G)**. In addition to being bound and cut by the MutH protein, hemimethylated GATC sequences in newly replicated DNA are also bound by the SeqA protein (67). On average, several hundred SeqA proteins bind to 200-400 GATC sites behind replication forks. This binding of SeqA to long stretches of DNA has been proposed to prevent the premature separation of newly replicated chromosomes, which is important for protecting the integrity of replication forks, particularly when DNA polymerase is stalled by a roadblock (68). Importantly, it has

been reported that there exists a negative correlation between SeqA binding and RNAP binding (69). This raises the possibility that SeqA binding prevents the rapid re-establishment of transcription complexes after replication disrupts ongoing transcription. This problem could be attenuated by depletion of the GATC sites in the highly expressed regions. This hypothesis is supported by the observation that highly expressed ribosomal RNA-coding operon *rrnC* exhibits a lower GATC content compared to its surrounding DNA (69). Therefore, the observed depletion of the GATC sequences in the MutL-ARs may be a consequence of high transcriptional activity in these regions.

Transcription activity within MutL-ARs

Highly transcribed genomic regions present a challenging environment for the replication machinery due to increased DNA unwinding, formation of secondary structures and altered nucleoid states. All these factors can reduce the fidelity of DNA replication in the highly transcriptionally active regions compared to less transcriptionally active genomic regions (57, 70). For example, high mutation rates in highly expressed genes have been observed in genome-wide studies of *E. coli* (71), *Salmonella typhimurium* (72), *Saccharomyces cerevisiae,* and the human germline (73). Therefore, we investigated whether there is a difference in transcriptional activity between MutL-ARs and the rest of the genome. We first used published RNA-Seq data of *E. coli* growing in rich LB medium (24) and found that highly expressed genes are enriched in MutL-ARs (Wilcoxon test, $p = 2.9 \times 10^{-9}$) **(Fig. 5A-B, Supp. Fig. 4A)**. Because this RNA-Seq data did not include transcription levels of ribosomal RNA-encoding genes and two other genes (24), we conducted additional analysis using a genome-wide map of the RNAP β subunit, RpoB, binding sites (20) and found their

enrichment in MutL-ARs (Permutation test, p = 0.001, Z-score = 4.8) **(Fig. 5A-B, Supp. Fig. 4B)**. Furthermore, it was observed that DNA gyrase cleavage sites are enriched downstream of the highly transcribed genes (74). DNA gyrase relaxes positive supercoils that accumulate ahead of moving RNA and DNA polymerases. We used a genome-wide map of the GyrA binding sites (20), which similarly displayed enrichment in MutL-ARs (Permutation test, p = 0.001, Z-score = 4.4) **(Fig. 5A-B, Supp. Fig. 4C)**. This result corroborates our previous results for GapR, which also interacts with positively supercoiled DNA, in which we observed an enrichment for GapR binding sites in MutL-ARs **(Fig. 3A-B, Supp. Fig. 2G)**.

Transcriptional activity can impact replication fidelity because it requires separation of dsDNA into ssDNA. The non-transcribed strand, which is not protected by the proteins associated with transcription, is expected to be more vulnerable to premutagenic chemical modifications (75, 76), such as deamination (77, 78), oxidation (79) and alkylation (80). These damages in ssDNA can create mismatches in the dsDNA during realignment or replication, which are detected by mismatch repair. Additional mechanisms by which ssDNA may be exposed in cells include DNA repair processes that remove these DNA damages from dsDNA, leading to the creation of ssDNA gaps, which we observed to be enriched within the MutL-ARs **(Fig. 4A-B, Supp. Fig. 3E)**. Besides being chemically unstable, ssDNA facilitates DNA polymerase slippage on mononucleotide repeats and formation of secondary structures, which increase likelihood of DNA polymerase stalling.

Another potential source of replication errors due to transcriptional activity is the collision between DNA replication and transcription machineries. In bacteria, DNA replication and transcription occur simultaneously on a common DNA template, and as the DNA replication machinery moves 10 to 20 times faster than elongating RNAP,

frequent collisions are unavoidable (81). Highly deleterious head-on collisions are greatly avoided because *E. coli* highly expressed genes are oriented to be transcribed in the same direction as chromosome replication (81). Co-directional collisions that cause replication stalling are less deleterious but they do occur at highly expressed loci such as ribosomal RNA operons (82, 83). Both types of collisions are aggravated by the R-loop formation, which is stabilized by the formation of G-quadruplex (G4) sequences within ssDNA stretches (84). We found an enrichment of G4-prone sequences, previously identified by Kaplan *et al* or predicted using the software G4Hunter (19, 25), in the MutL-ARs (Permutation test, p = 0.001, Z-score = 10.6, data from Kaplan *et al*; p = 0.002, Z-score = 3.4, G4Hunter) **(Fig. 5A-B, Supp. Fig. 4D-E)**. However, we observed no difference in the gene transcription orientation relative to replication direction in the MutL-ARs compared to the rest of the genome (Binomial test, p = 0.24) **(Supp. Fig. 4F)**.

Co-directional collisions that cause replication stalling can also be caused by backtracked RNAPs and by transcription terminators. Backtracking is a fundamental property of RNAP implicated in the control of transcription elongation, pausing, termination and proofreading, and fidelity (85). However, RNAPs can also backtrack when encountering roadblocks such as DNA-bound proteins and DNA damage. It was shown that the co-directional collisions between the replication machinery and backtracked transcription elongation complexes can result in double-strand breaks on both plasmid and chromosome in *E. coli* (86). Repairing double-strand breaks is normally a high-fidelity process, but it can switch to a mutagenic mode in stressed cells when error-prone DNA polymerases are involved in the repair process (87). Finally, it was observed that transcription terminators disrupt replication of *E. coli* plasmids when they were co-oriented with replication (88). Therefore, although bacteria possess

multiple mechanisms that have evolved to deal with transcription-replication conflicts (81), they are clearly not failproof.

Despite extensive knowledge about molecular mechanisms involved in transcription-associated mutagenesis cited above, several mutation accumulation (MA) studies have shown weak or no correlation between high transcription and mutation rates (3, 70, 71). Several important differences between our experimental approach and MA experiments may explain these discrepancies: First, we detect emerging mutations, i.e., DNA replication errors before fixation, while MA studies identify only fixed mutations. Second, in our study, replication errors are detected by MutS and MutL proteins that, in the absence of the MutH protein, remain bound to DNA. This likely prevents other DNA repair mechanisms from removing replication errors prior to mutation fixation, which allows us to detect a vast majority of replication errors. In contrast, MA misses errors that are corrected by DNA repair mechanisms other than mismatch repair. Third, our assay can detect even errors that eventually give rise to lethal mutations, which is impossible using the MA approach that requires cell growth and division. Fourth, to study the impact of transcription on mutation rates, we corroborated gene expression levels with RpoB binding and GATC depletion, providing a more comprehensive view of transcriptional activity. MA studies typically relied solely on gene expression data.

Genes and gene functions localized within the MutL-ARs

We have shown that high transcription activity is enriched in MutL-ARs compared to the rest of the genome, and it is known that transcriptional activity can differ depending on the gene function. The highly expressed genes in *E. coli* are primarily involved in the core cellular processes of transcription, translation, and metabolism, all of which

691 are essential for rapid growth and cellular proliferation. Pathway and gene ontology
692 (GO) enrichment analysis of the genes located within the MutL-ARs revealed
693 enrichment of these functional categories, especially those related to rRNA and tRNA-
694 coding genes **(Fig. 6A).** Furthermore, transcription factor enrichment analysis
695 identified an enrichment for several transcription factors that directly regulate the genes
696 located in the MutL-ARs **(Fig. 6A)**, including Fis and HN-S NAPs, further validating the
697 association we observed for their binding sites and the MutL-ARs **(Fig. 3A-B, Supp.**
698 **Fig. 2E-F)**.

699 We also found that there is an enrichment of the intergenic regions within MutL-
700 ARs (Paired Proportion test with Holm's correction, $p < 2 \times 10^{-16}$) **(Fig. 6B)**.
701 Furthermore, the enrichment of intergenic regions is also observed if we consider only
702 the MutL-ARs' summits, *i.e.* the location in the MutL-ARs with the highest Log2 ratio
703 (IP/Input) (Paired Proportion test with Holm's correction, $p = 0.0015$) **(Fig. 6B)**.
704 Because intergenic regions play crucial roles in the regulation of gene expression by
705 providing biding sites for the regulatory proteins, including NAPs, as well as for RNAP,
706 it is plausible that the enrichment of the NAPs' biding sites in the MutL-associated
707 replication error hotspots sites **(Fig. 3A-B, Supp. Fig. 2C-G)** results from their impact
708 on transcription.

709 Highly expressed genes are crucial for bacterial fitness and generally tend to
710 evolve more slowly compared to genes with lower expression levels (89, 90). So, it is
711 intriguing that we found genes coding for tRNA and rRNA operons to be replication
712 error hotspots. However, it was already observed that genes encoding rRNA and
713 ribosomal proteins exhibit a positive correlation between substitution mutations and
714 gene expression level in mismatch repair deficient *E. coli* (91). Additionally, these
715 authors also found that several genes coding for tRNA have high mutation rates

716 primarily due to mononucleotide runs. Could this be explained by the fact that highly
717 expressed genes coding for rRNA and tRNA are replication error-prone, but these
718 errors are particularly efficiently repaired by the mismatch repair system? While there
719 is no direct experimental evidence for this, the observation that the loss of the
720 mismatch repair system leads to a more significant increase in mutation rates in coding
721 DNA compared to noncoding DNA in *E. coli* (3, 92) suggests that this may be the case.
722 Highly expressed genes are also under strong purifying selection because any
723 deleterious mutations in these genes can have significant fitness consequences for the
724 organism, which explains their lower phylogenetic divergence. However, by using our
725 MutL-based assay, we can detect replication errors that are highly deleterious and
726 even lethal, before they have any impact on cell functioning and therefore are invisible
727 to purifying selection.

728 We identified 223 genes within MutL-ARs **(Supp. Table 2**). Importantly, 167 of
729 these genes (71.6%) were previously found mutated in MA study using mismatch
730 repair deficient *E. coli* strain (91). These 223 genes were similarly distributed between
731 two replisomes, each representing one of the two replication arms stretching from *oriC*
732 to the *ter* region, with a small enrichment in genes of the right replication arm (Binomial
733 test, $p = 0.049$) **(Fig. 6C)**. This enrichment most likely results from the localization of 5
734 out of 7 ribosomal operons in the right replication arm. In *E. coli*, highly expressed
735 genes, such as rRNA and tRNA-coding genes, are typically transcribed in the same
736 direction as replication, which helps prevent harmful head-on collisions. However, we
737 found that certain genes within MutL-ARs are transcribed opposite to the direction of
738 replication, which has the potential to stall and disrupt replication. This is not surprising
739 as, unlike in *B. subtilis*, where transcription is co-oriented with replication for 75% of all
740 genes, in *E. coli*, this co-orientation is only 55% (93). We verified and found that there

741 is a lower frequency of genes transcribed opposite to the direction of replication within
742 the MutL-ARs (84 out of 233, 36%) when compared to the rest of the genome (2013
743 out of 4494, 45%) (Binomial test, p = 0.008).

744 Transcription-replication conflicts occur frequently during co-directional
745 encounters and they are not always benign. It was observed that co-directional
746 conflicts at highly transcribed rRNA operons in *B. subtilis* can disrupt replication (82).
747 Importantly, these conflicts were detected under fast growth conditions but not under
748 slow growth conditions, which reduce the transcription of rRNA genes, highlighting the
749 role of high transcriptional activity. We have also previously shown that increasing
750 transcription rates of rRNA operons cause DNA replication blockage, massive DNA
751 breakage at the rRNA operon sites, and increased mutagenesis due to the involvement
752 of low-fidelity DNA polymerases in *E. coli* (83). Besides being responsible for RNAP
753 traffic jams, the genes encoding rRNA and tRNA may also act as replication blockage
754 hotspots due to the presence of inverted DNA repeats within their sequences. These
755 inverted repeats are crucial for the proper folding and three-dimensional structure
756 formation of the RNA molecules, which is essential for their functional roles in protein
757 synthesis. The stem-loop DNA structures formed by the inverted repeats can
758 potentially cause stalling or blockage of the replication machinery.

759 Our findings highlight a delicate trade-off resulting from the intricate balance
760 between necessity to preserve functional integrity over evolutionary timescales and the
761 elevated mutation rates associated with high expression levels that are essential for
762 rapid growth, allowing responses to ecological challenges. This balance is
763 orchestrated by the interplay between evolutionary constraints, such as purifying
764 selection, and the complex molecular processes governing gene expression, DNA
765 replication, and repair mechanisms. High mutation rates associated with high

766 expression levels may, when needed, facilitate rapid evolution of new functions as it
767 was shown for rapid evolution of tRNA coding genes allowing to meet novel translation
768 demands in *S. cerevisiae* (94).

CONCLUDING REMARKS

Our study provides a comprehensive map of replication error hotspots across the *E. coli* genome **(Fig. 7),** along with associated genomic features and proteins. We recognize that these features and proteins are often interconnected and may not act independently **(Fig. 7A)**. For example, Fis binds to DNA and shapes nucleoid structure, which in turn affects RNAP's access to different genomic regions, potentially explaining its proximity to the RNAP β subunit, RpoB, in the PCA Biplot **(Fig. 7A)**. Rather than considering each feature contribution to replication fidelity in isolation, they should be considered within the broader context of dynamic processes, such as global and local chromosomal architecture and transcriptional activity, which collectively shape replication fidelity. Some of them may even not be associated directedly with replication fidelity but with the processes that impact replication fidelity. For instance, transcription activity generates negative supercoiling behind the transcribing enzyme, promoting the formation of DNA structures that are known to promote replication errors, such as cruciforms and G-quadruplexes **(Fig. 7A)**. Although all these features clearly influence the emergence of replication errors, the genomic regions preferentially bound by MutL exhibit idiosyncratic patterns, as demonstrated by the differing presence or absence of specific features in the MutL-ARs **(Fig. 7B-C)**. Acknowledging these higher-level interconnections helps avoid oversimplifying or misinterpreting the roles of individual factors and promotes a more integrated understanding of the mechanisms driving replication errors.

The relevance of our findings is underscored by the observation that the accumulation of genetic diversity in the genomes of natural *E. coli* isolates is predominantly associated with errors in DNA replication (9). Although we used a mismatch repair-deficient *E. coli* strain, the implications remain broadly applicable

794 because mismatch repair-deficient strains are common in natural populations of
795 various bacterial species, including *E. coli* (1, 95). Additionally, the mismatch repair
796 system can become inoperative due to the suppression of MutS protein synthesis by
797 small regulatory RNAs under stress conditions (96, 97), or because the amount of MutL
798 protein becomes insufficient to support efficient mismatch repair due to its titration by
799 an excess of replication errors (98, 99). The replication error hotspots we identified in
800 the *E. coli* genome have the potential to cause mutations, which may impact its ability
801 to colonize, persist, resist to antibiotics or cause disease in the host. However, this
802 potential is expected to fluctuate, contingent upon factors such as environmental
803 conditions affecting the metabolic state of bacterial cells, the presence of stressors,
804 and the specific DNA polymerase involved (54, 100). Therefore, it is imperative to
805 investigate in future studies how the ever-changing environmental factors within *E. coli*
806 natural habitat, the mammalian gut, influence the distribution of replication error
807 hotspots and their mutational consequences.

DATA AVAILABILITY

ChIP-seq sequencing data has been deposited with links to BioProject accession number PRJNA1121661 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/). The code for reproducing the statistical analysis and the figures are available in https://github.com/hugocbarreto/MutL-ChIP-seq and in Zenodo, at https://dx.doi.org/10.5281/zenodo.12625199. The data underlying this article are available in the article, in its online supplementary material, and in Zenodo, at https://dx.doi.org/10.5281/zenodo.12625199.

SUPPLEMENTARY DATA STATEMENT

Supplementary Data are available at NAR Online.

CONFLICT OF INTEREST DISCLOSURE

Authors declare that they have no competing interests.

FUNDING

This work was supported by French “Agence Nationale de la Recherche” grants (ANR-21-CE12-006-01 and ANR-20-AMR-0002). F.C.H. was supported by Labex “Who am I?” Idex ANR-11-IDEX-0005-02 / ANR-11-LABX-0071, post-doctoral fellowship. H.C.B. was supported by DREAM ANR-20-AMR-0002 grant and by the HORIZON-MSCA-2023-PF-01 project number 101148351 - MICROINVADER, funded by the European Union. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible

for them. Biomics Platform, C2RT, Institut Pasteur, Paris, France, was supported by France Génomique (ANR-10-INBS-09) and IBISA grants.

ACKNOWLEDGMENTS

We thank M. Haustant, L. Lemée, and T. Cokelaer from Biomics Platform, C2RT, Institut Pasteur, Paris, for performing library preparation and sequencing for ChIP-seq analysis (Project #16000). We thank M. Guo for kindly providing the GapR binding sites and P. Pham for kindly providing the ssDNA regions. We are grateful to C. Lesterlin (MMSB, Lyon, France) for generous gift of a strain. We are thankful to O. Tenaillon for their valuable advices in the bioinformatics analysis. We thank J. Horton and J. Ibarra for critical reading of the manuscript.

REFERENCES

1. Denamur,E. and Matic,I. (2006) Evolution of mutation rates in bacteria. *Mol Microbiol*, **60**, 820–827.

2. Kunkel,T.A. (2009) Evolving views of DNA replication (in)fidelity. *Cold Spring Harb Symp Quant Biol*, **74**, 91–101.

3. Lee,H., Popodi,E., Tang,H. and Foster,P.L. (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci U S A*, **109**, E2774-2783.

4. Fishel,R. (2015) Mismatch repair. *J Biol Chem*, **290**, 26395–26403.

5. Iyer,R.R., Pluciennik,A., Burdett,V. and Modrich,P.L. (2006) DNA mismatch repair: functions and mechanisms. *Chem Rev*, **106**, 302–323.

6. López de Saro,F.J. and O'Donnell,M. (2001) Interaction of the beta sliding clamp with MutS, ligase, and DNA polymerase I. *Proc Natl Acad Sci U S A*, **98**, 8376–8380.

7. López de Saro,F.J., Marinus,M.G., Modrich,P. and O'Donnell,M. (2006) The beta sliding clamp binds to multiple sites within MutL and MutS. *J Biol Chem*, **281**, 14340–14349.

8. Glickman,B.W. and Radman,M. (1980) Escherichia coli mutator mutants deficient in methylation-instructed DNA mismatch correction. *Proc Natl Acad Sci U S A*, **77**, 1063–1067.

9. Garushyants,S.K., Sane,M., Selifanova,M.V., Agashe,D., Bazykin,G.A. and Gelfand,M.S. (2024) Mutational Signatures in Wild Type Escherichia coli Strains Reveal Predominance of DNA Polymerase Errors. *Genome Biol Evol*, **16**, evae035.

10. Woo,A.C., Faure,L., Dapa,T. and Matic,I. (2018) Heterogeneity of spontaneous DNA replication errors in single isogenic Escherichia coli cells. *Sci Adv*, **4**, eaat1608.

11. Kunkel,T.A. and Bebenek,K. (2000) DNA replication fidelity. *Annu Rev Biochem*, **69**, 497–529.

12. Kondrashov,F.A. and Kondrashov,A.S. (2010) Measurements of spontaneous rates of mutations in the recent past and the near future. *Philos Trans R Soc Lond B Biol Sci*, **365**, 1169–1176.

13. Nishant,K.T., Singh,N.D. and Alani,E. (2009) Genomic mutation rates: what high-throughput methods can tell us. *Bioessays*, **31**, 912–920.

14. Schroeder,J.W., Yeesin,P., Simmons,L.A. and Wang,J.D. (2018) Sources of spontaneous mutagenesis in bacteria. *Crit Rev Biochem Mol Biol*, **53**, 29–48.

879 15. Tenaillon,O. and Matic,I. (2020) The Impact of Neutral Mutations on Genome
880 Evolvability. *Curr Biol*, **30**, R527–R534.

881 16. Elez,M., Murray,A.W., Bi,L.-J., Zhang,X.-E., Matic,I. and Radman,M. (2010)
882 Seeing mutations in living cells. *Curr Biol*, **20**, 1432–1437.

883 17. Elez,M., Radman,M. and Matic,I. (2012) Stoichiometry of MutS and MutL at
884 unrepaired mismatches in vivo suggests a mechanism of repair. *Nucleic Acids*
885 *Res*, **40**, 3929–3938.

886 18. Espeli,O., Mercier,R. and Boccard,F. (2008) DNA dynamics vary according to
887 macrodomain topography in the E. coli chromosome. *Mol Microbiol*, **68**, 1418–
888 1427.

889 19. Brázda,V., Kolomazník,J., Lýsek,J., Bartas,M., Fojta,M., Šťastný,J. and Mergny,J.-
890 L. (2019) G4Hunter web application: a web server for G-quadruplex prediction.
891 *Bioinformatics*, **35**, 3493–3495.

892 20. Decker,K.T., Gao,Y., Rychel,K., Al Bulushi,T., Chauhan,S.M., Kim,D., Cho,B.-K.
893 and Palsson,B.O. (2022) proChIPdb: a chromatin immunoprecipitation
894 database for prokaryotic organisms. *Nucleic Acids Res*, **50**, D1077–D1084.

895 21. Guo,M.S., Kawamura,R., Littlehale,M.L., Marko,J.F. and Laub,M.T. (2021) High-
896 resolution, genome-wide mapping of positive supercoiling in chromosomes.
897 *Elife*, **10**, e67236.

898 22. Pham,P., Shao,Y., Cox,M.M. and Goodman,M.F. (2022) Genomic landscape of
899 single-stranded DNA gapped intermediates in Escherichia coli. *Nucleic Acids*
900 *Res*, **50**, 937–951.

901 23. Cohen,N.R., Ross,C.A., Jain,S., Shapiro,R.S., Gutierrez,A., Belenky,P., Li,H. and
902 Collins,J.J. (2016) A role for the bacterial GATC methylome in antibiotic stress
903 survival. *Nat Genet*, **48**, 581–586.

904 24. Niccum,B.A., Lee,H., MohammedIsmail,W., Tang,H. and Foster,P.L. (2019) The
905 Symmetrical Wave Pattern of Base-Pair Substitution Rates across the
906 Escherichia coli Chromosome Has Multiple Causes. *mBio*, **10**, e01226-19.

907 25. Kaplan,O.I., Berber,B., Hekim,N. and Doluca,O. (2016) G-quadruplex prediction in
908 E. coli genome reveals a conserved putative G-quadruplex-Hairpin-Duplex
909 switch. *Nucleic Acids Res*, **44**, 9083–9095.

910 26. Reuter,A., Hilpert,C., Dedieu-Berne,A., Lematre,S., Gueguen,E., Launay,G.,
911 Bigot,S. and Lesterlin,C. (2021) Targeted-antibacterial-plasmids (TAPs)
912 combining conjugation and CRISPR/Cas systems achieve strain-specific
913 antibacterial activity. *Nucleic Acids Res*, **49**, 3584–3598.

914 27. Miller,J.H. (1992) A short course in bacterial genetics: a laboratory manual and
915 handbook for Escherichia coli and related bacteria Cold Spring Harbor
916 Laboratory Pr, Cold Spring Harbor, NY.

28. Datsenko,K.A. and Wanner,B.L. (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A*, **97**, 6640–6645.

29. Khan,S.R., Mahaseth,T., Kouzminova,E.A., Cronan,G.E. and Kuzminov,A. (2016) Static and Dynamic Factors Limit Chromosomal Replication Complexity in Escherichia coli, Avoiding Dangers of Runaway Overreplication. *Genetics*, **202**, 945–960.

30. Kuzminov,A. (2016) Chromosomal Replication Complexity: A Novel DNA Metrics and Genome Instability Factor. *PLoS Genet*, **12**, e1006229.

31. Diaz,R.E., Sanchez,A., Anton Le Berre,V. and Bouet,J.-Y. (2017) High-Resolution Chromatin Immunoprecipitation: ChIP-Sequencing. *Methods Mol Biol*, **1624**, 61–73.

32. Wang,Z., Zou,L., Zhang,Y., Zhu,M., Zhang,S., Wu,D., Lan,J., Zang,X., Wang,Q., Zhang,H., *et al.* (2023) ACS-20/FATP4 mediates the anti-ageing effect of dietary restriction in C. elegans. *Nat Commun*, **14**, 7683.

33. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

34. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

35. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M., *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.

36. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.

37. Sun,G., Chung,D., Liang,K. and Keleş,S. (2013) Statistical analysis of ChIP-seq data with MOSAiCS. *Methods Mol Biol*, **1038**, 193–212.

38. Stovner,E.B. and Sætrom,P. (2019) epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics*, **35**, 4392–4393.

39. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, **44**, W160-165.

40. Pham,P., Wood,E.A., Cox,M.M. and Goodman,M.F. (2023) RecA and SSB genome-wide distribution in ssDNA gaps and ends in Escherichia coli. *Nucleic Acids Res*, **51**, 5527–5546.

41. Breslauer,K.J., Frank,R., Blöcker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A*, **83**, 3746–3750.

954 42. Miura,O., Ogake,T. and Ohyama,T. (2018) Requirement or exclusion of inverted
955 repeat sequences with cruciform-forming potential in Escherichia coli revealed
956 by genome-wide analyses. *Curr Genet*, **64**, 945–958.

957 43. Keseler,I.M., Gama-Castro,S., Mackie,A., Billington,R., Bonavides-Martínez,C.,
958 Caspi,R., Kothari,A., Krummenacker,M., Midford,P.E., Muñiz-Rascado,L., *et al.*
959 (2021) The EcoCyc Database in 2021. *Front Microbiol*, **12**, 711077.

960 44. Gel,B., Díez-Villanueva,A., Serra,E., Buschbeck,M., Peinado,M.A. and
961 Malinverni,R. (2016) regioneR: an R/Bioconductor package for the association
962 analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**,
963 289–291.

964 45. Moolman,M.C., Krishnan,S.T., Kerssemakers,J.W.J., van den Berg,A., Tulinski,P.,
965 Depken,M., Reyes-Lamothe,R., Sherratt,D.J. and Dekker,N.H. (2014) Slow
966 unloading leads to DNA-bound β2-sliding clamp accumulation in live
967 Escherichia coli cells. *Nat Commun*, **5**, 5820.

968 46. Justice,S.S., Hunstad,D.A., Cegelski,L. and Hultgren,S.J. (2008) Morphological
969 plasticity as a bacterial survival strategy. *Nat Rev Microbiol*, **6**, 162–168.

970 47. Simmons,L.A., Foti,J.J., Cohen,S.E. and Walker,G.C. (2008) The SOS Regulatory
971 Network. *EcoSal Plus*, **3**.

972 48. Grilley,M., Griffith,J. and Modrich,P. (1993) Bidirectional excision in methyl-
973 directed mismatch repair. *J Biol Chem*, **268**, 11830–11837.

974 49. Modrich,P. (1991) Mechanisms and biological effects of mismatch repair. *Annu*
975 *Rev Genet*, **25**, 229–253.

976 50. Hasan,A.M.M. and Leach,D.R.F. (2015) Chromosomal directionality of DNA
977 mismatch repair in Escherichia coli. *Proc Natl Acad Sci U S A*, **112**, 9388–9393.

978 51. Liu,J., Lee,R., Britton,B.M., London,J.A., Yang,K., Hanne,J., Lee,J.-B. and
979 Fishel,R. (2019) MutL sliding clamps coordinate exonuclease-independent
980 Escherichia coli mismatch repair. *Nat Commun*, **10**, 5294.

981 52. Duigou,S. and Boccard,F. (2017) Long range chromosome organization in
982 Escherichia coli: The position of the replication origin defines the non-structured
983 regions and the Right and Left macrodomains. *PLoS Genet*, **13**, e1006758.

984 53. Kivisaar,M. (2019) Mutation and Recombination Rates Vary Across Bacterial
985 Chromosome. *Microorganisms*, **8**, 25.

986 54. Horton,J.S. and Taylor,T.B. (2023) Mutation bias and adaptation in bacteria.
987 *Microbiology (Reading)*, **169**, 001404.

988 55. Lioy,V.S., Junier,I. and Boccard,F. (2021) Multiscale Dynamic Structuring of
989 Bacterial Chromosomes. *Annu Rev Microbiol*, **75**, 541–561.

990 56. Warnecke,T., Supek,F. and Lehner,B. (2012) Nucleoid-associated proteins affect
991 mutation dynamics in E. coli in a growth phase-specific manner. *PLoS Comput*
992 *Biol*, **8**, e1002846.

993 57. Jinks-Robertson,S. and Bhagwat,A.S. (2014) Transcription-associated
994 mutagenesis. *Annu Rev Genet*, **48**, 341–359.

995 58. Foster,P.L., Hanson,A.J., Lee,H., Popodi,E.M. and Tang,H. (2013) On the
996 mutational topology of the bacterial genome. *G3 (Bethesda)*, **3**, 399–407.

997 59. Fu,Z., Guo,M.S., Zhou,W. and Xiao,J. (2024) Differential roles of positive and
998 negative supercoiling in organizing the E. coli genome. *Nucleic Acids Res*, **52**,
999 724–737.

1000 60. Postow,L., Crisona,N.J., Peter,B.J., Hardy,C.D. and Cozzarelli,N.R. (2001)
1001 Topological challenges to DNA replication: conformations at the fork. *Proc Natl*
1002 *Acad Sci U S A*, **98**, 8219–8226.

1003 61. Magnan,D. and Bates,D. (2015) Regulation of DNA Replication Initiation by
1004 Chromosome Structure. *J Bacteriol*, **197**, 3370–3377.

1005 62. Sobetzko,P., Glinkowska,M., Travers,A. and Muskhelishvili,G. (2013) DNA
1006 thermodynamic stability and supercoil dynamics determine the gene expression
1007 program during the bacterial growth cycle. *Mol Biosyst*, **9**, 1643–1651.

1008 63. Fang,F.C. and Rimsky,S. (2008) New insights into transcriptional regulation by H-
1009 NS. *Curr Opin Microbiol*, **11**, 113–120.

1010 64. Ogawa,T. and Okazaki,T. (1980) Discontinuous DNA replication. *Annu Rev*
1011 *Biochem*, **49**, 421–457.

1012 65. Waldminghaus,T. and Skarstad,K. (2009) The Escherichia coli SeqA protein.
1013 *Plasmid*, **61**, 141–150.

1014 66. Brendler,T., Sawitzke,J., Sergueev,K. and Austin,S. (2000) A case for sliding SeqA
1015 tracts at anchored replication forks during Escherichia coli chromosome
1016 replication and segregation. *EMBO J*, **19**, 6249–6258.

1017 67. Helgesen,E., Fossum-Raunehaug,S., Sætre,F., Schink,K.O. and Skarstad,K.
1018 (2015) Dynamic Escherichia coli SeqA complexes organize the newly replicated
1019 DNA at a considerable distance from the replisome. *Nucleic Acids Res*, **43**,
1020 2730–2743.

1021 68. Rotman,E., Khan,S.R., Kouzminova,E. and Kuzminov,A. (2014) Replication fork
1022 inhibition in seqA mutants of Escherichia coli triggers replication fork breakage.
1023 *Mol Microbiol*, **93**, 50–64.

1024 69. Sánchez-Romero,M.A., Busby,S.J.W., Dyer,N.P., Ott,S., Millard,A.D. and
1025 Grainger,D.C. (2010) Dynamic distribution of seqA protein across the
1026 chromosome of escherichia coli K-12. *mBio*, **1**, e00012-10.

70. Lynch,M., Ackerman,M.S., Gout,J.-F., Long,H., Sung,W., Thomas,W.K. and Foster,P.L. (2016) Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet*, **17**, 704–714.

71. Chen,X. and Zhang,J. (2013) No gene-specific optimization of mutation rate in Escherichia coli. *Mol Biol Evol*, **30**, 1559–1562.

72. Lind,P.A. and Andersson,D.I. (2008) Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci U S A*, **105**, 17878–17883.

73. Park,C., Qian,W. and Zhang,J. (2012) Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep*, **13**, 1123–1129.

74. Sutormin,D., Rubanova,N., Logacheva,M., Ghilarov,D. and Severinov,K. (2019) Single-nucleotide-resolution mapping of DNA gyrase cleavage sites across the Escherichia coli genome. *Nucleic Acids Res*, **47**, 1373–1388.

75. Saini,N. and Gordenin,D.A. (2020) Hypermutation in single-stranded DNA. *DNA Repair (Amst)*, **91–92**, 102868.

76. Chan,K., Sterling,J.F., Roberts,S.A., Bhagwat,A.S., Resnick,M.A. and Gordenin,D.A. (2012) Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS Genet*, **8**, e1003149.

77. Polosina,Y.Y. and Cupples,C.G. (2010) Wot the 'L-Does MutL do? *Mutat Res*, **705**, 228–238.

78. Marinus,M.G. (2012) DNA Mismatch Repair. *EcoSal Plus*, **5**.

79. Wyrzykowski,J. and Volkert,M.R. (2003) The Escherichia coli methyl-directed mismatch repair system repairs base pairs containing oxidative lesions. *J Bacteriol*, **185**, 1701–1704.

80. Rasmussen,L.J. and Samson,L. (1996) The Escherichia coli MutS DNA mismatch binding protein specifically binds O(6)-methylguanine DNA lesions. *Carcinogenesis*, **17**, 2085–2088.

81. Merrikh,H., Zhang,Y., Grossman,A.D. and Wang,J.D. (2012) Replication-transcription conflicts in bacteria. *Nat Rev Microbiol*, **10**, 449–458.

82. Merrikh,H., Machón,C., Grainger,W.H., Grossman,A.D. and Soultanas,P. (2011) Co-directional replication-transcription conflicts lead to replication restart. *Nature*, **470**, 554–557.

83. Fleurier,S., Dapa,T., Tenaillon,O., Condon,C. and Matic,I. (2022) rRNA operon multiplicity as a bacterial genome stability insurance policy. *Nucleic Acids Res*, **50**, 12601–12620.

84. Goehring,L., Huang,T.T. and Smith,D.J. (2023) Transcription-Replication Conflicts as a Source of Genome Instability. *Annu Rev Genet*, **57**, 157–179.

85. Nudler,E. (2012) RNA polymerase backtracking in gene regulation and genome instability. *Cell*, **149**, 1438–1445.

86. Dutta,D., Shatalin,K., Epshtein,V., Gottesman,M.E. and Nudler,E. (2011) Linking RNA polymerase backtracking to genome instability in E. coli. *Cell*, **146**, 533–543.

87. Rosenberg,S.M., Shee,C., Frisch,R.L. and Hastings,P.J. (2012) Stress-induced mutation via DNA breaks in Escherichia coli: a molecular mechanism with implications for evolution and medicine. *Bioessays*, **34**, 885–892.

88. Mirkin,E.V., Castro Roa,D., Nudler,E. and Mirkin,S.M. (2006) Transcription regulatory elements are punctuation marks for DNA replication. *Proc Natl Acad Sci U S A*, **103**, 7276–7281.

89. Sharp,P.M., Shields,D.C., Wolfe,K.H. and Li,W.H. (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science*, **246**, 808–810.

90. Sharp,P.M. and Li,W.H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol*, **4**, 222–230.

91. Foster,P.L., Niccum,B.A. and Lee,H. (2021) DNA Replication-Transcription Conflicts Do Not Significantly Contribute to Spontaneous Mutations Due to Replication Errors in Escherichia coli. *mBio*, **12**, e0250321.

92. Foster,P.L., Lee,H., Popodi,E., Townes,J.P. and Tang,H. (2015) Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. *Proc Natl Acad Sci U S A*, **112**, E5990-5999.

93. Wang,J.D., Berkmen,M.B. and Grossman,A.D. (2007) Genome-wide coorientation of replication and transcription reduces adverse effects on replication in Bacillus subtilis. *Proc Natl Acad Sci U S A*, **104**, 5608–5613.

94. Yona,A.H., Bloom-Ackermann,Z., Frumkin,I., Hanson-Smith,V., Charpak-Amikam,Y., Feng,Q., Boeke,J.D., Dahan,O. and Pilpel,Y. (2013) tRNA genes rapidly change in evolution to meet novel translational demands. *Elife*, **2**, e01339.

95. Bjedov,I., Tenaillon,O., Gérard,B., Souza,V., Denamur,E., Radman,M., Taddei,F. and Matic,I. (2003) Stress-induced mutagenesis in bacteria. *Science*, **300**, 1404–1409.

96. Gutierrez,A., Laureti,L., Crussard,S., Abida,H., Rodríguez-Rojas,A., Blázquez,J., Baharoglu,Z., Mazel,D., Darfeuille,F., Vogel,J., *et al.* (2013) β-Lactam antibiotics promote bacterial mutagenesis via an RpoS-mediated reduction in replication fidelity. *Nat Commun*, **4**, 1610.

97. Chen,J. and Gottesman,S. (2017) Hfq links translation repression to stress-induced mutagenesis in E. coli. *Genes Dev*, **31**, 1382–1395.

98. Schaaper,R.M. and Radman,M. (1989) The extreme mutator effect of Escherichia coli mutD5 results from saturation of mismatch repair by excessive DNA replication errors. *EMBO J*, **8**, 3511–3516.

99. Matic,I., Babic,A. and Radman,M. (2003) 2-aminopurine allows interspecies recombination by a reversible inactivation of the Escherichia coli mismatch repair system. *J Bacteriol*, **185**, 1459–1461.

100. Foster,P.L., Niccum,B.A., Popodi,E., Townes,J.P., Lee,H., MohammedIsmail,W. and Tang,H. (2018) Determinants of Base-Pair Substitution Patterns Revealed by Whole-Genome Sequencing of DNA Mismatch Repair Defective Escherichia coli. *Genetics*, **209**, 1029–1042.

FIGURE LEGENDS

Figure 1. Visualization and quantification of replication forks and replication errors in *mutH*-deficient *E. coli*. A) Representative microscopy images showing *E. coli* cells expressing mCherry-DnaN and CFP-MutL fluorescent foci tagging replication forks and replication errors, respectively. Scale bar is 1 μm. **B)** Frequency of *E. coli* cells with 0, 2, 6, or ≥ 7 replication forks per cell, obtained from 93,861 cells. The numbers above the bars indicate the number of cells in each category. **C)** Frequency of MutL foci per cell, in cells with 0, 2, 6, or ≥ 7 replication forks obtained from 93,861 cells. **D)** Frequency of MutL foci per cell, in 93,861 cells. The red line and dots indicate the predicted Poisson distribution. **E)** Q-Q plot of the observed MutL foci per cell obtained from 93,861 cells and the theoretical quartiles when following a Poisson distribution. The red line indicates the expected distribution if the data have a Poisson distribution. For panels B, C, and D, a square root transformation was performed in the y-axis for better visualization of the lower frequencies.

Figure 2. Localization and patterns of the MutL-ARs. A) Circular map of E. coli K-12 MG1655 and **B)** zoomed section showing the localization of MutL-ARs and the Log2 ratio (IP/Input) after scaling with Signal Extraction Scaling (SES). The order of the rings (from outside to inside) is: Combined (MutL-ARs predicted by at least two software), MutL-ARs predicted by MACS3, MutL-ARs predicted by MOSAICS, MutL-ARs predicted by epic2, and the Log2 ratio (IP/Input) after scaling with SES for two independent experiments. The numbers located on the outermost ring indicate genome coordinates in megabase pairs (Mbs). **C, D, and E)** Different shapes of the MutL-ARs peaks. The upper row of panels **C**, **D**, and **E** contains schematic drawings demonstrating how the number of the replication error sites (mismatches), as well as the position of a mismatch relative to the replication fork, may be responsible for the

appearance of the different shapes of MutL-ARs peaks we observed in this study (bottom row). Bifurcating arrows indicate the loading site and the sliding of MutS and MutL protein clamps away from the mismatch site. Schematic drawings in panels **C** and **D** show bidirectional symmetrical sliding of MutS and MutL protein clamps from a single mismatch, as supported by *in vitro* data (51): **C** same localization of a single mismatch in different genomes, resulting in a single-summit MutL-AR peak (bottom), and **D** different but close localization of single mismatches in different genomes resulting, in overlapping MutL-ARs peaks (bottom). Schematic drawing in panel **E** shows asymmetric sliding of MutS and MutL protein clamps away from the mismatch site towards replication fork, as supported by the *in vivo* data (50), resulting in an asymmetric MutL-AR peak with a summit shifted towards replication origin (bottom). Examples of observed MutL-ARs' peaks shown in the bottom row of panels the **C**, **D**, and **E** correspond to peaks number 33, 28 and 32, respectively. Two different colored lines show that the shapes of the MutL-ARs' peaks were practically identical in two independent experiments.

Figure 3. Global chromosome structure and nucleoid-associated proteins binding sites in the MutL-ARs. A) Circular map of *E. coli* K-12 MG1655 and **B)** zoomed section showing the localization of MutL-ARs, the Macrodomains and nucleoid-associated proteins binding sites. The order of the rings (from outside to inside) is: macrodomains, MutL-ARs, HupA, HupB, Fis, H-NS, and GapR binding sites. The numbers located on the outermost ring indicate genome coordinates in megabase pairs (Mbs). NSL, non-structured left. NSR, non-structured right.

Figure 4. Mapping the local DNA sequence properties of the MutL-ARs. A) Circular map of *E. coli* K-12 MG1655 and **B)** zoomed section showing the localization of (from outside to inside) MutL-ARs, GC content per 200 bp (ranging from 0 to 1),

melting temperature per 200 bp (ranging from 70 to 120), number of microsatellites (mononucleotide repeats of 4 nucleotides or more) per 1000 bp (ranging from 0 to 35), number of cruciform-prone sequences per 1000 bp (ranging from 0 to 20), ssDNA location, and number of GATC sites per 1000 bp (ranging from 0 to 23). The numbers located on the outermost indicate genome coordinates in megabase pairs (Mbs).

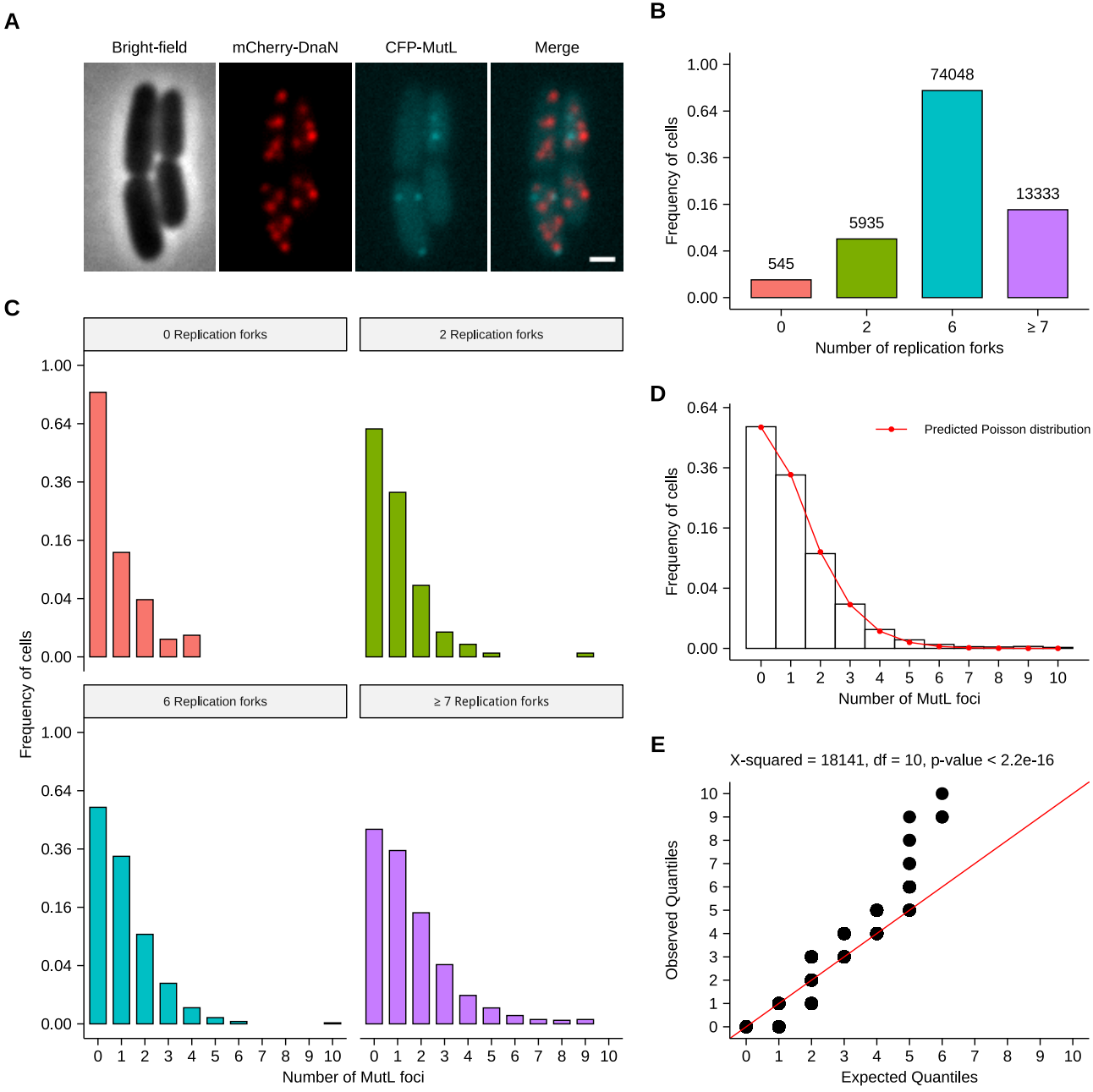
Figure 5. Mapping the transcription activity within MutL-ARs. A) Circular map of *E. coli* K-12 MG1655 and **B)** zoomed section showing the localization of (from outside to inside) MutL-ARs, RNA-Seq (ranging from 0 to 6), Log10 of transcription level), RpoB binding sites, GyrA binding sites, G4-prone sequences from Kaplan *et al.* (23), and G4-prone sequences identified by G4Hunter. The numbers located on the outermost ring indicate genome coordinates in megabase pairs (Mbs).

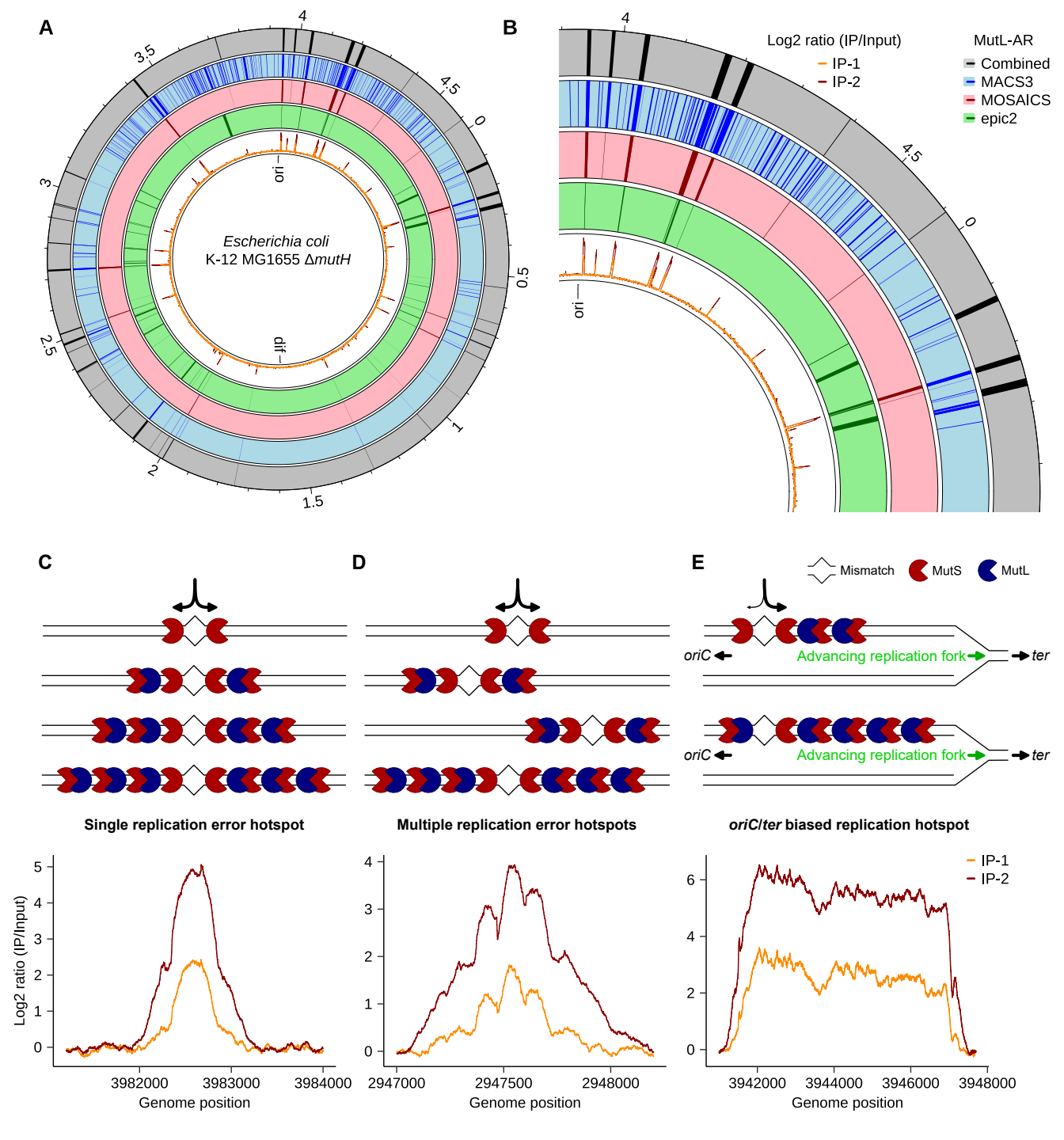
Figure 6. Genes and gene functions localized within the MutL-ARs. A) Enriched Gene Ontology Biological Process (GO - BP), Gene Ontology Cellular Component (GO - CC), Gene Ontology Molecular Function (GP - MF), Pathways, and Transcriptional Regulators (TR (direct)) obtained after gene enrichment analysis of the MutL-ARs using the EcoCyc database. The x-axis indicates the number of genes in the MutL-ARs that belong to the different categories indicated in the y-axis. **B)** Relative frequency of coding region in the genome, in MutL-ARs, and in the MutL-ARs summits. **C)** Relative frequency of genes in the right and left replication arm. For panel B a pairwise Proportion test with Benjamin-Hochberg's correction was used. For panel C a Binomial test was used.

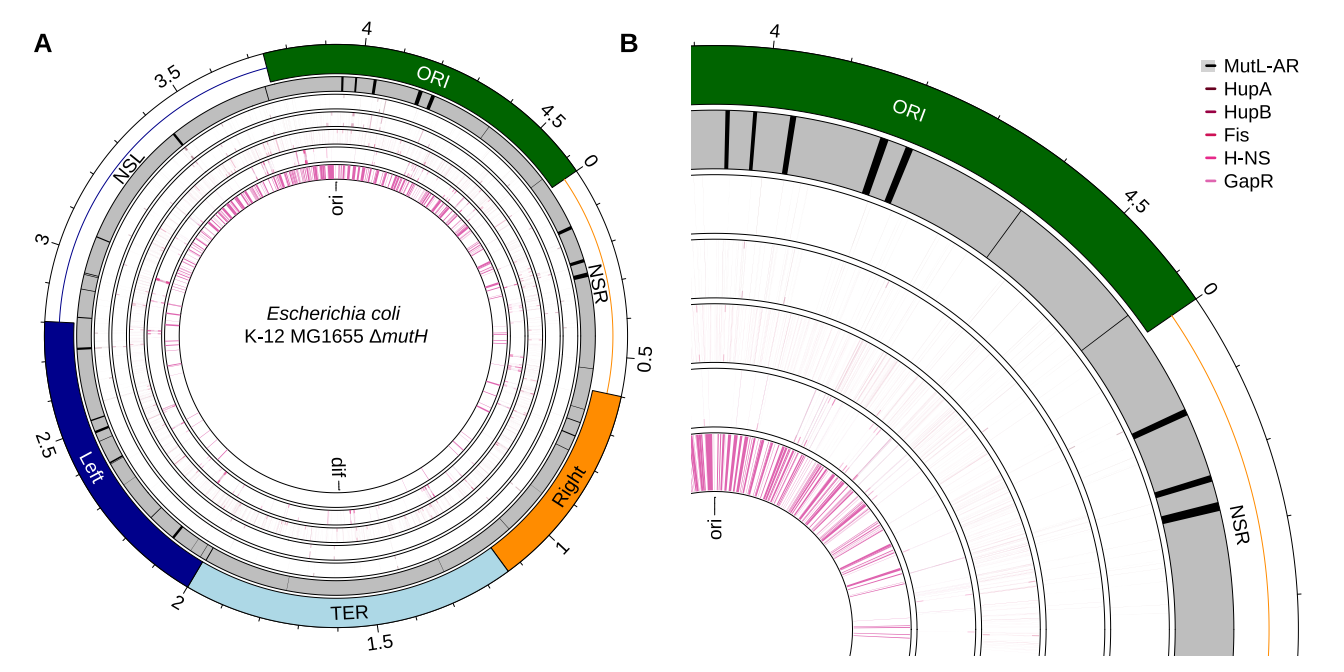
Figure 7. Sequence properties and protein interactions at MutL-associated replication error hotspots. A) Principal Component Analysis (PCA) Biplot of MutL-ARs and associated features. The first two principal components (Dim1 and Dim2) are

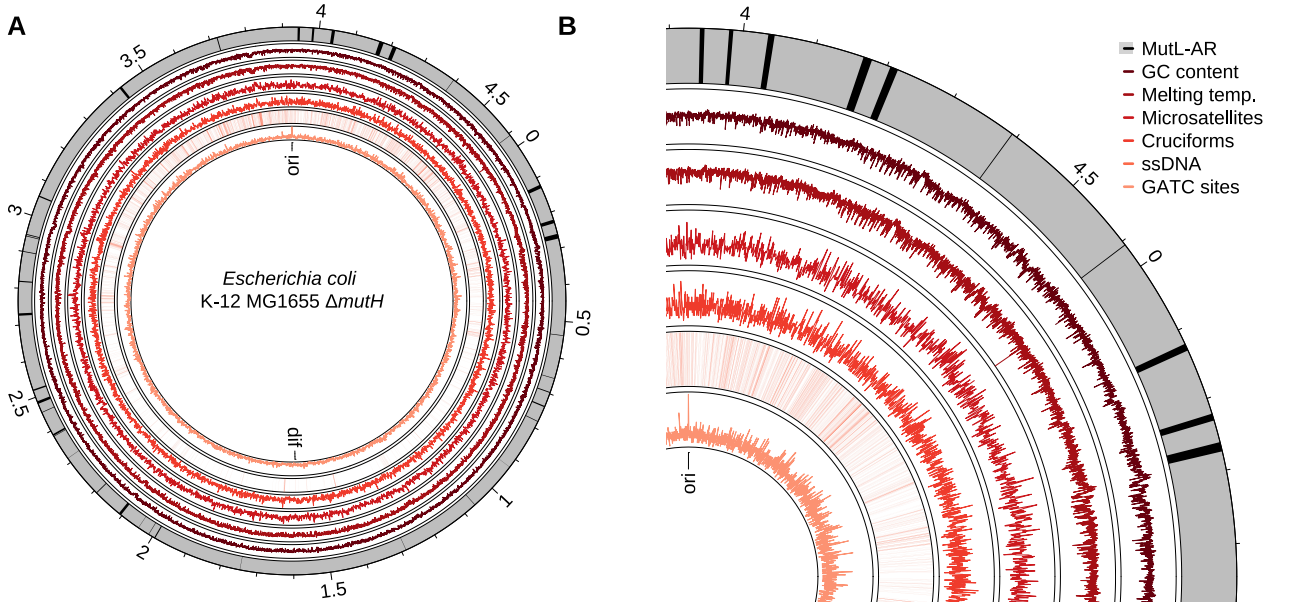
displayed on the axes, explaining the largest amount of variance (52.7%). Dim1 explains 32.5% of the variance and is primarily driven by factors related to DNA structure and stability (GC content, Melting temperature, GyrA, HupA, HupB, and GapR). Dim2 explains 20.2% of the variance and is primarily driven by regulatory elements (H-NS, Fis, Microsatellites and GATC sites). The lower left quadrant features elements that influence or respond to transcription activity, while the lower right quadrant features elements that impact or respond to supercoiling. Each numbered point corresponds to a specific MutL-AR (see Supp. Table 1). Black points indicate MutL-AR that include genes coding for ribosomal and/or tRNAs, while grey points indicate MutL-ARs containing other genes. The blue arrows represent the features associated with MutL-ARs. The direction of each arrow indicates the direction in which each feature increases, while the length of the arrow represents the strength of each feature's influence on the principal components. Arrows that are closely located or point in the same direction suggest a positive correlation between those factors. Arrows pointing in opposite directions indicate factors that are negatively correlated. **B-C)** Two genomic regions showing representative MutL-ARs (number 1 and 6), their DNA sequence properties, distribution of protein binding sites and gene expression levels, which are enriched or depleted in these regions. These examples emphasize the involvement of the high transcription activity and presence of sequences prone to form secondary structures in the localization of replication error hots spots. **B)** The MutL-ARs summit positions colocalize with highly expressed genes and an enrichment of inverted repeats prone to cruciform formation. The intergenic region immediately upstream from the MutL-ARs summit position colocalize with sequences of low thermal stability, ssDNA gap region, sequences enriched for microsatellites, and the localization of Fis, GapR and RpoB binding sites. **C)** The colocalization of the MutL-

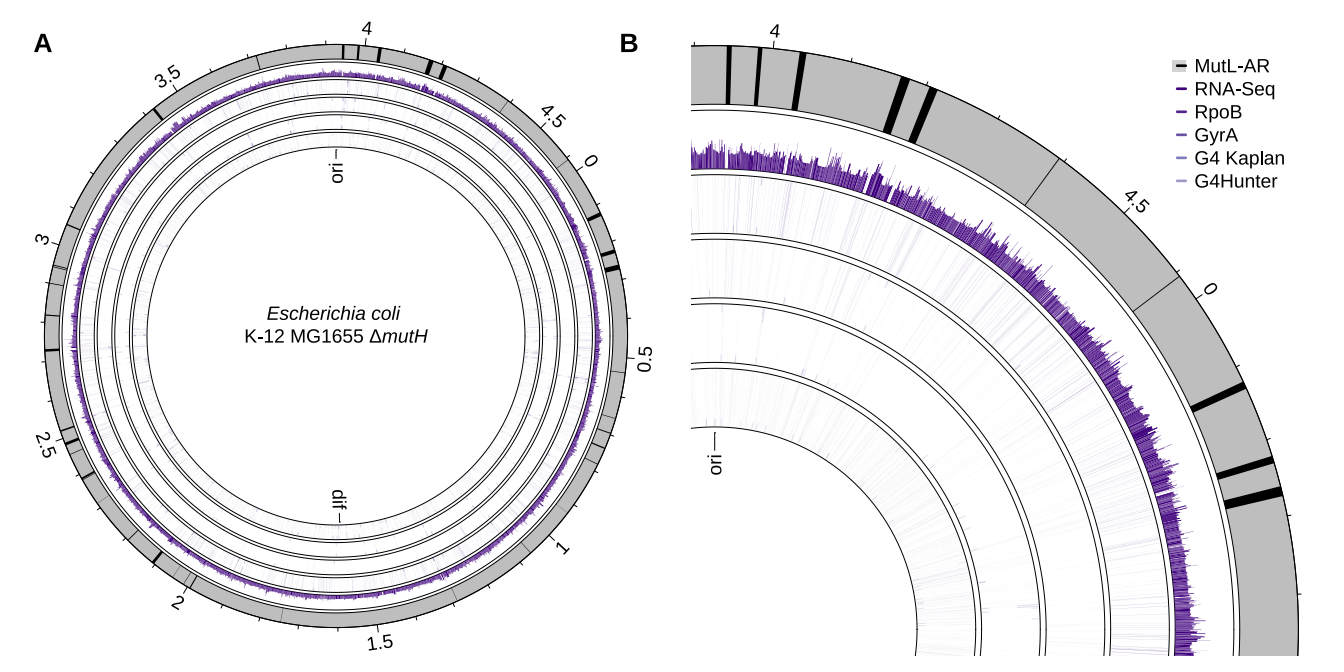
ARs summit position within the tRNA coding gene array with high gene expression, sequences enriched for microsatellites, and inverted repeats prone to cruciform formation, alongside the depletion of the GATC sequences. The intergenic region immediately upstream from the tRNA coding gene array colocalizes with sequences of low thermal stability, and the localization of Fis, HupB and RpoB binding sites. The grey rectangles indicate the MutL-ARs. The grey dashed rectangular lines highlight a portion of the MutL-ARs around the summit of each ChiP-Seq replicate, for which an overlap of several factors shown to be enriched or depleted in the MutL-ARs is observed. Values for RNA-seq represent the Log10 of transcription. GC content and melting temperature are represented is 200 bp bins. Gene arrows are represented to scale.

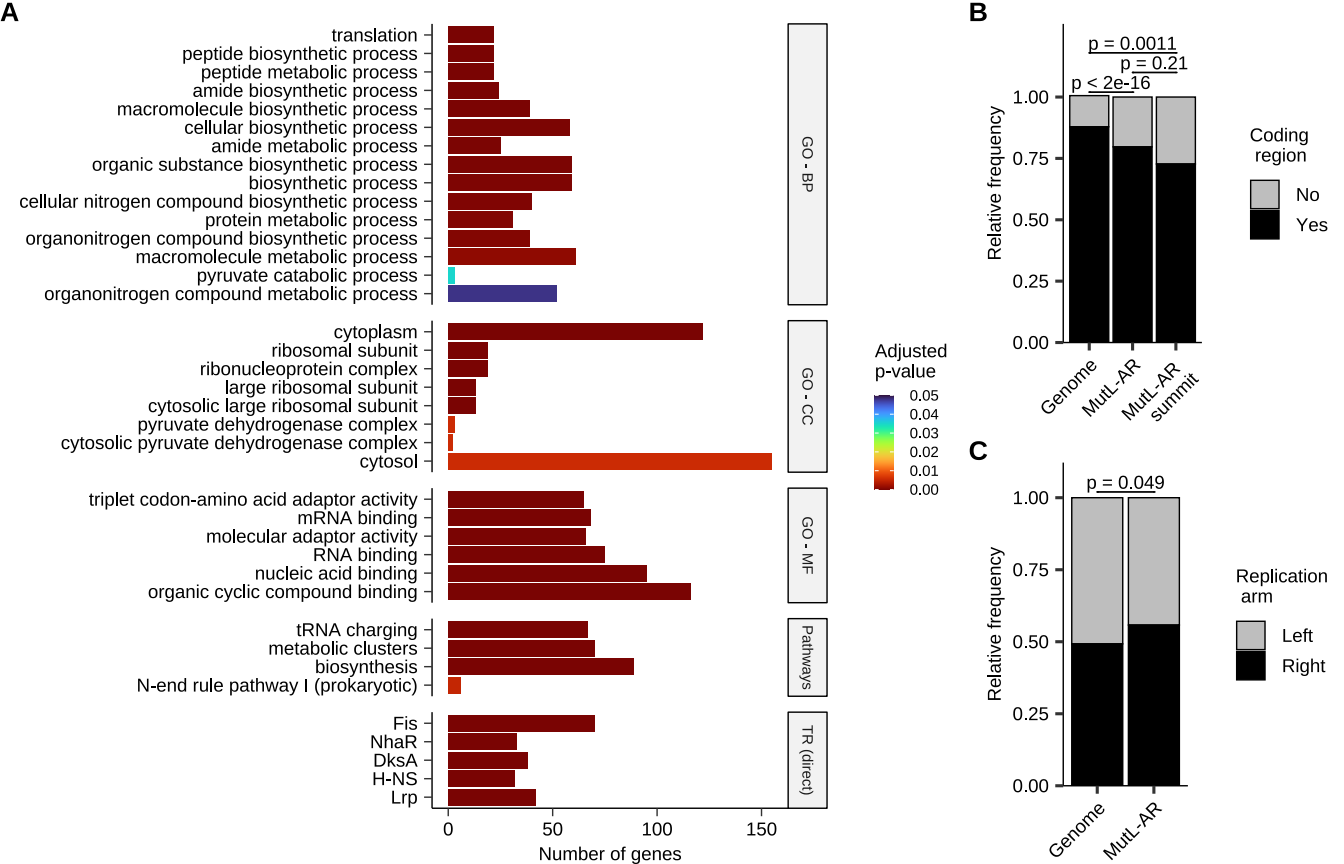

A
Bright-field
mCherry-DnaN
CFP-MutL
Merge
B
Frequency of cells
1.00
0.64
0.36
0.16
0.04
0.00
545
5935
74048
13333
0
2
6
≥ 7
Number of replication forks
C
0 Replication forks
2 Replication forks
6 Replication forks
≥ 7 Replication forks
Frequency of cells
0 1 2 3 4 5 6 7 8 9 10
Number of MutL foci
D
Predicted Poisson distribution
Frequency of cells
Number of MutL foci
E
X-squared = 18141, df = 10, p-value < 2.2e-16
Observed Quantiles
Expected Quantiles

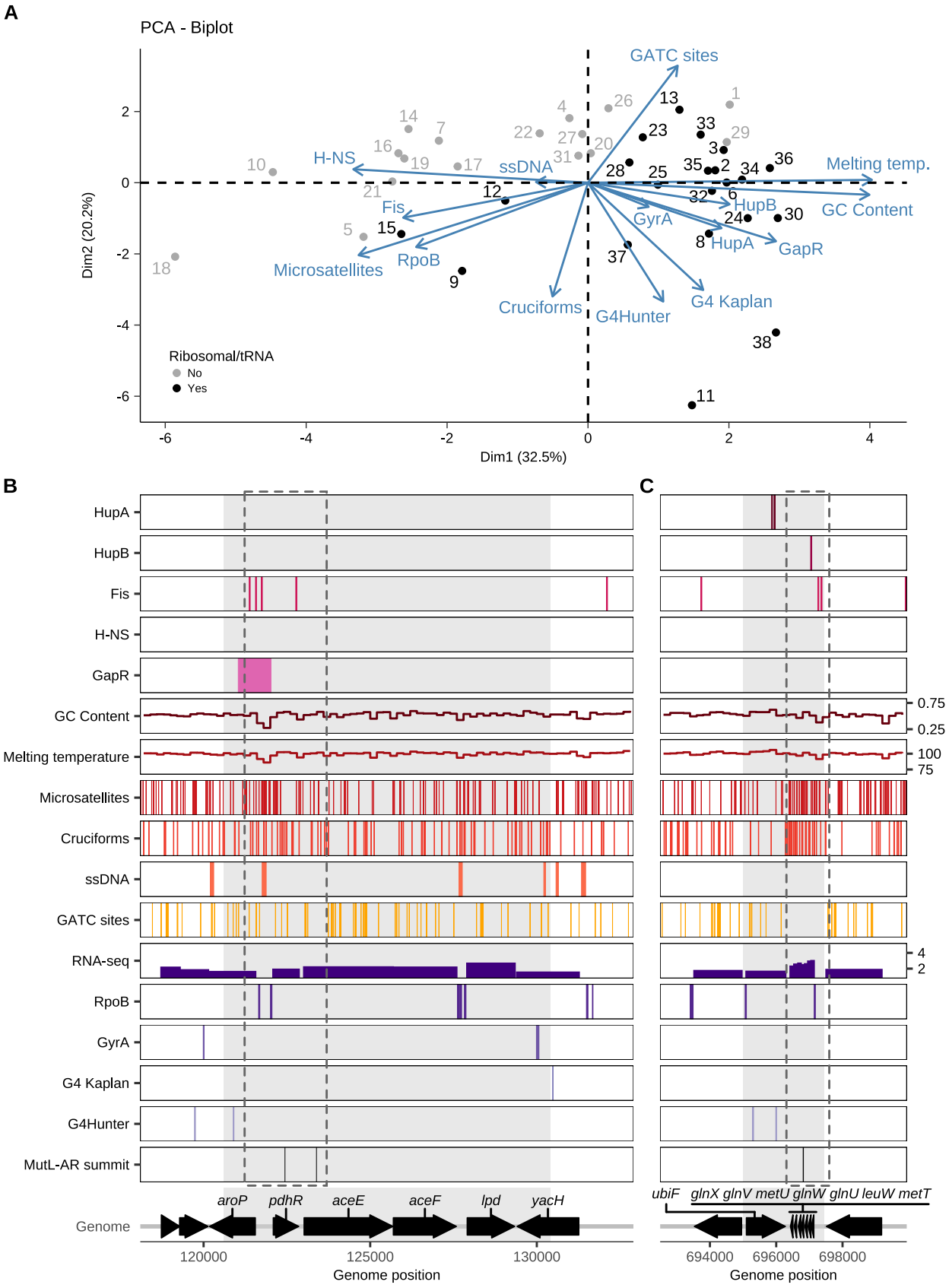

A
B
Escherichia coli
K-12 MG1655 ΔmutH
ori
dif
Log2 ratio (IP/Input)
IP-1
IP-2
MutL-AR
Combined
MACS3
MOSAICS
epic2
C
D
E
Mismatch
MutS
MutL
oriC
Advancing replication fork
ter
Single replication error hotspot
Multiple replication error hotspots
oriC/ter biased replication hotspot
Log2 ratio (IP/Input)
Genome position
3982000
3983000
3984000
2947000
2947500
2948000
3942000
3944000
3946000
3948000


A
B
ORI
NSR
NSL
TER
Left
Right
ori
dif
Escherichia coli
K-12 MG1655 ΔmutH
MutL-AR
HupA
HupB
Fis
H-NS
GapR
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5


A
B
4
4.5
0
0.5
1
1.5
2
2.5
3
3.5
ori
dif
Escherichia coli
K-12 MG1655 ΔmutH
MutL-AR
GC content
Melting temp.
Microsatellites
Cruciforms
ssDNA
GATC sites


A
B
4
3.5
4.5
0
0.5
1
1.5
2
2.5
3
ori
dif
Escherichia coli
K-12 MG1655 ΔmutH
MutL-AR
RNA-Seq
RpoB
GyrA
G4 Kaplan
G4Hunter


A
translation
peptide biosynthetic process
peptide metabolic process
amide biosynthetic process
macromolecule biosynthetic process
cellular biosynthetic process
amide metabolic process
organic substance biosynthetic process
biosynthetic process
cellular nitrogen compound biosynthetic process
protein metabolic process
organonitrogen compound biosynthetic process
macromolecule metabolic process
pyruvate catabolic process
organonitrogen compound metabolic process
GO - BP
cytoplasm
ribosomal subunit
ribonucleoprotein complex
large ribosomal subunit
cytosolic large ribosomal subunit
pyruvate dehydrogenase complex
cytosolic pyruvate dehydrogenase complex
cytosol
GO - CC
triplet codon-amino acid adaptor activity
mRNA binding
molecular adaptor activity
RNA binding
nucleic acid binding
organic cyclic compound binding
GO - MF
tRNA charging
metabolic clusters
biosynthesis
N-end rule pathway I (prokaryotic)
Pathways
Fis
NhaR
DksA
H-NS
Lrp
TR (direct)
0
50
100
150
Number of genes
Adjusted p-value
0.05
0.04
0.03
0.02
0.01
0.00
B
p = 0.0011
p = 0.21
p < 2e-16
Relative frequency
1.00
0.75
0.50
0.25
0.00
Genome
MutL-AR
MutL-AR summit
Coding region
No
Yes
C
p = 0.049
Relative frequency
1.00
0.75
0.50
0.25
0.00
Genome
MutL-AR
Replication arm
Left
Right


A
PCA - Biplot
GATC sites
H-NS
ssDNA
Melting temp.
Fis
HupB
GC Content
GyrA
HupA
GapR
RpoB
Microsatellites
Cruciforms
G4Hunter
G4 Kaplan
Ribosomal/tRNA
No
Yes
Dim2 (20.2%)
Dim1 (32.5%)
B
C
HupA
HupB
Fis
H-NS
GapR
GC Content
Melting temperature
Microsatellites
Cruciforms
ssDNA
GATC sites
RNA-seq
RpoB
GyrA
G4 Kaplan
G4Hunter
MutL-AR summit
Genome
aroP
pdhR
aceE
aceF
lpd
yacH
ubiF
glnX glnV metU glnW glnU leuW metT
Genome position
Genome position

SUPPLEMENTARY MATERIAL

Genome-wide mapping of spontaneous DNA replication error-hotspots using mismatch repair proteins in rapidly proliferating *Escherichia coli*

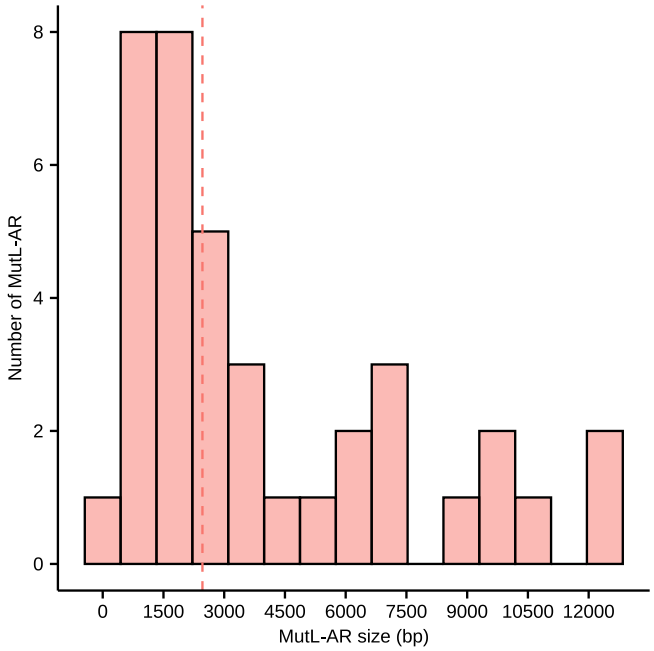
Flavia C. Hasenauer[1#], Hugo C. Barreto[1#], Chantal Lotton[1#], Ivan Matic[1*]

[1]Université Paris Cité, CNRS, Inserm, Institut Cochin, F-75014 Paris, France

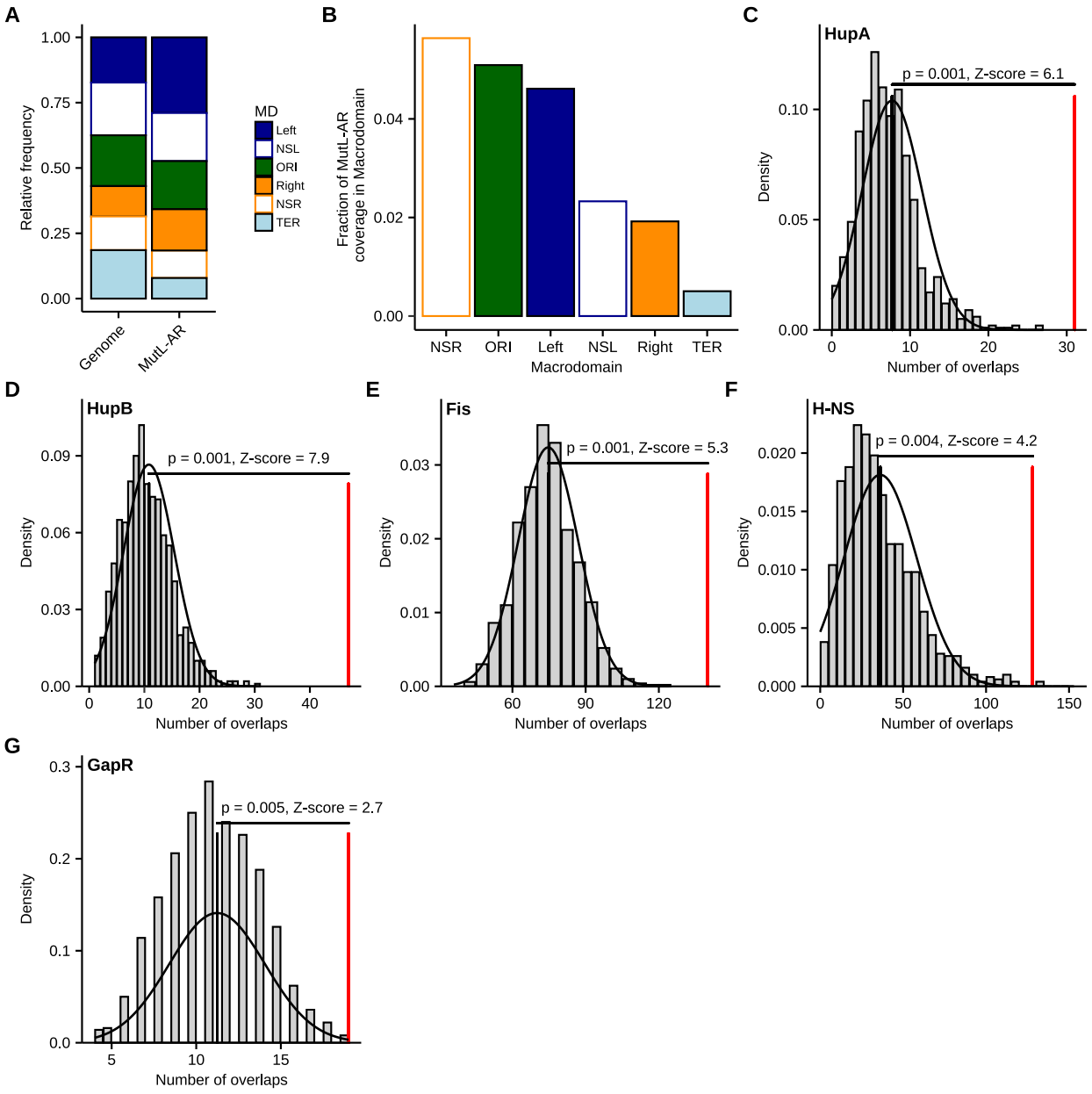
[#]These authors contributed equally to this work

*To whom correspondence should be addressed

Email address of corresponding author : ivan.matic@inserm.fr

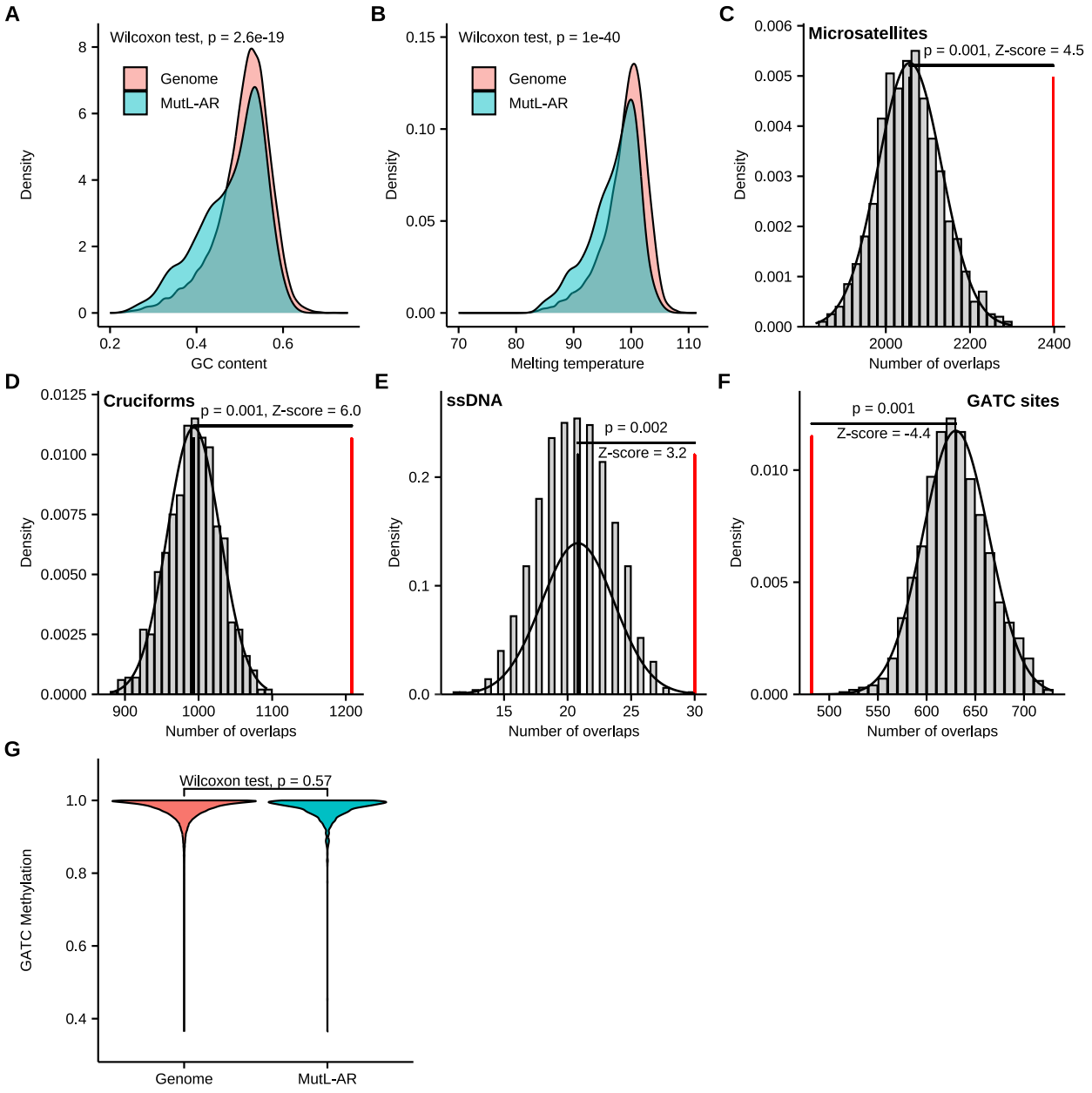


Supplementary Figure 1. Distribution of MutL-AR sizes. Histogram with 15 bins for the MutL-AR size. The dashed red line indicates the median MutL-AR size (2463 bp).



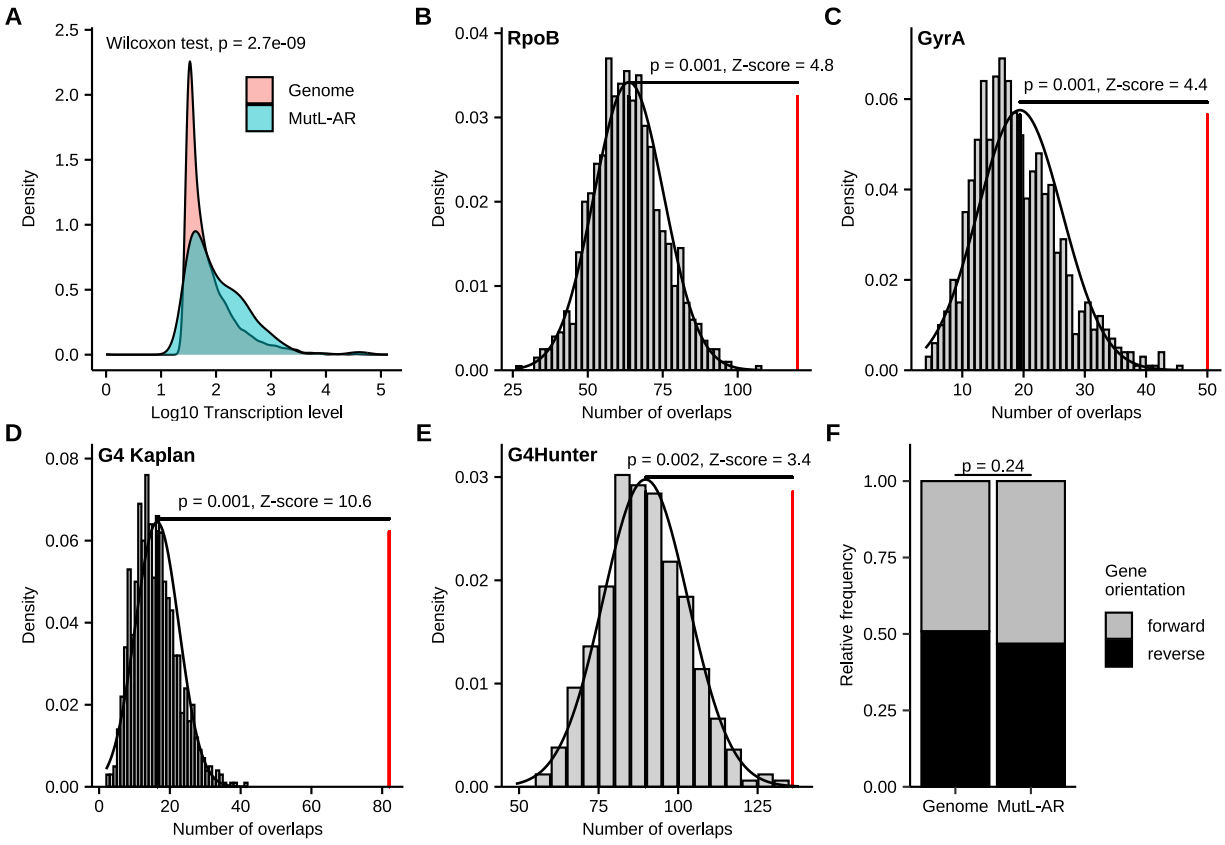
Supplementary Figure 2. MutL-AR distribution in macrodomains and nucleoid-associated proteins binding sites enrichment in MutL-AR. A) Relative frequency of each macrodomain (MD) in the *E. coli* genome and in the MutL-AR. **B)** Fraction of MutL-AR coverage in each macrodomain. **C-G)** Histograms representing the Permutation test for nucleoid-associated proteins binding sites **C)** HupA, **D)** HupB, **E)** Fis, **F)** H-NS, and **G)** GapR. The grey bars indicate the number of overlaps observed after 1000 permutations. The black curve indicates the normal distribution of the permutations. The black vertical line indicates the mean number of overlaps after 1000 permutations and the red line

indicates the observed number of overlaps in the MutL-AR. The y-axis represents the relative likelihood of observing values within different regions of the data range (Density), and the x-axis represents the number of overlaps between the randomized/original MutL-AR and the criterion tested for association (Number of overlaps).



Supplementary Figure 3. Local DNA sequence properties of the MutL-AR. A) Density plot showing the GC content distribution in *E. coli* genome and in MutL-AR. GC content was calculated using 200 bp bins. B**)** Density plot showing the melting temperature distribution in the *E. coli* genome and in MutL-AR. Melting temperature was calculated using 200 bp bins. **C-F)** Histogram representing the Permutation test for **C)** the location of microsatellites (mononucleotide repeats of 4 nucleotides or more), **D)** the location of cruciform regions (obtained using CIRI), **E)** the location of ssDNA, and **F)** the location of GATC sites. **G)** Violin plots showing the frequency of GATC methylation (1) for the

genome and MutL-AR. For panels C, D, E, and F, the grey bars indicate the number of overlaps observed after 1000 permutations. The black curve indicates the normal distribution of the permutations. The black vertical line indicates the mean number of overlaps after 1000 permutations and the red line indicates the observed number of overlaps in the MutL-AR. The y-axis represents the relative likelihood of observing values within different regions of the data range (Density), and the x-axis represents the number of overlaps between the randomized/original MutL-AR and the criterion tested for association (Number of overlaps).



Supplementary Figure 4. Transcription activity within MutL-AR. A) Density plot showing the Log10 transcription levels per gene in *E. coli* genome and in MutL-AR. Transcription level per gene was obtained from Niccum *et al.* (2). **B-E)** Histogram representing the Permutation test for **B)** the location of RpoB binding sites, **C)** the location of GyrA binding sites, **D)** the location of G4-prone sequences from *Kaplan et al*. (3), and **E)** the location of G4-prone sequences identified by G4Hunter. **(F)** Relative frequency of gene orientation in the genome and in MutL-AR. For panels **B-E**, the grey bars indicate the number of overlaps observed after 1000 permutations. The black curve indicates the normal distribution of the permutations. The black vertical line indicates the mean number of overlaps after 1000 permutations and the red line indicates the observed number of overlaps in the MutL-AR. The y-axis represents the relative likelihood of observing values within different regions of the data range (Density), and the x-axis represents the number of overlaps between the randomized/original MutL-AR and the criterion tested for association (Number of overlaps). For panel **F** a Binomial test was used.

References.

1. Cohen,N.R., Ross,C.A., Jain,S., Shapiro,R.S., Gutierrez,A., Belenky,P., Li,H. and Collins,J.J. (2016) A role for the bacterial GATC methylome in antibiotic stress survival. *Nat Genet*, **48**, 581–586.

2. Niccum,B.A., Lee,H., MohammedIsmail,W., Tang,H. and Foster,P.L. (2019) The Symmetrical Wave Pattern of Base-Pair Substitution Rates across the Escherichia coli Chromosome Has Multiple Causes. *mBio*, **10**, e01226-19.

3. Kaplan,O.I., Berber,B., Hekim,N. and Doluca,O. (2016) G-quadruplex prediction in E. coli genome reveals a conserved putative G-quadruplex-Hairpin-Duplex switch. *Nucleic Acids Res*, **44**, 9083–9095.