REPORT

www.the-innovation.org

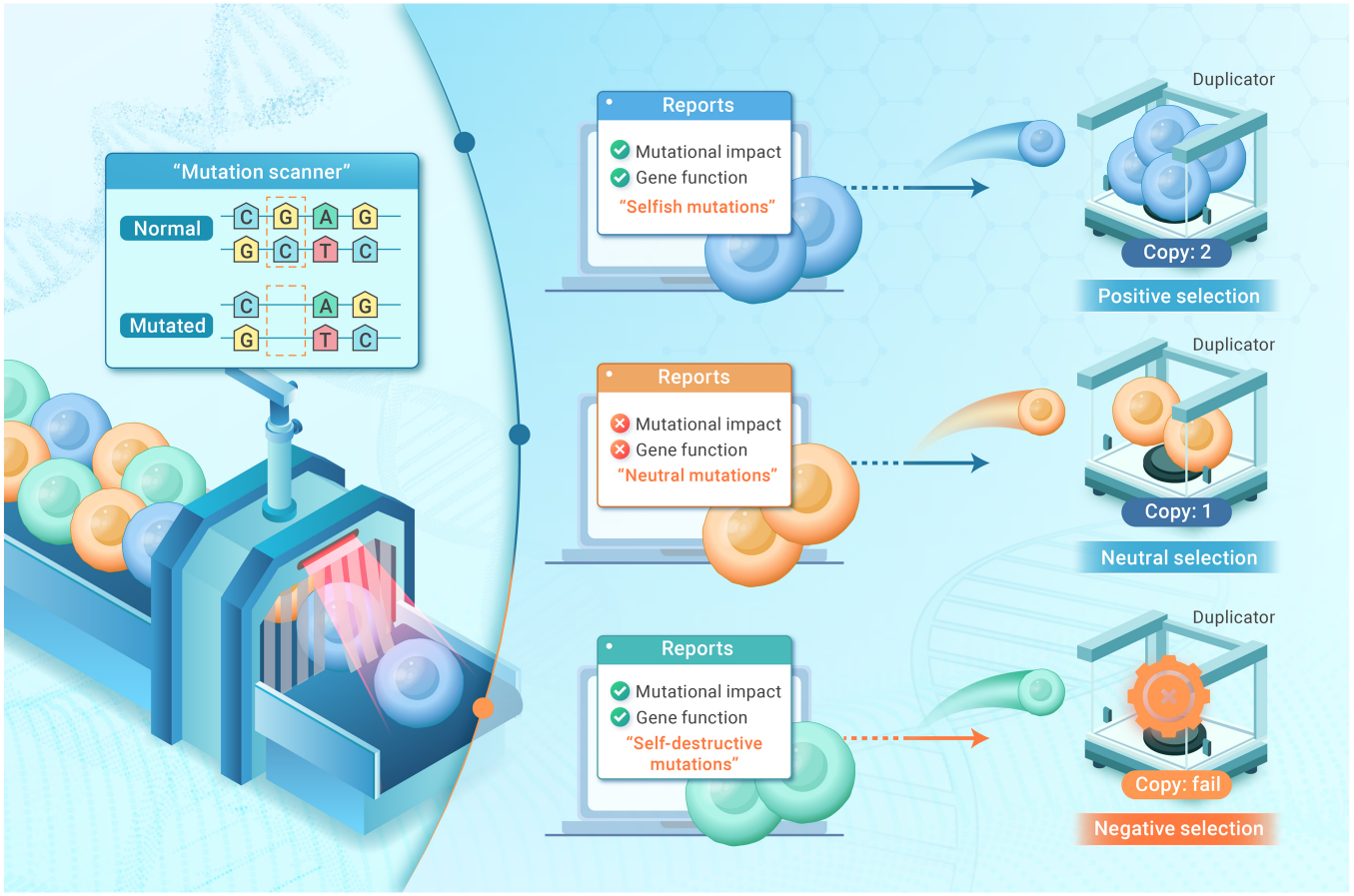
Divergent accumulation patterns of SNVs and INDELs reveal negative selection in noncancerous cells

Lei Zhang,[1,2,*] Moonsook Lee,[3] Xiaoxiao Hao,[3,12] Xiao Ma,[1,4] Chuwei Xia,[1,4] Yiwei Zhao,[1,4] Joseph Ehlert,[5] Zhongxuan Chi,[3] Bo Jin,[6] Ronald Cutler,[3] Alexander Y. Maslov,[3] Albert-László Barabási,[5,7,8] Jan H.J. Hoeijmakers,[9,10,11] Winfried Edelmann,[6] Jan Vijg,[3,*] and Xiao Dong[1,4,*]

*Correspondence: zhan8273@umn.edu (L.Z.); jan.vijg@einsteinmed.edu (J.V.); dong0265@umn.edu (X.D.)

Received: August 1, 2024; Accepted: June 23, 2025; Published Online: June 25, 2025; https://doi.org/10.1016/j.xinn.2025.101008

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- SNVs accumulate linearly during passaging in mismatch-repair-deficient and control cells.
- In contrast, small insertions and deletions (INDELs) reach a plateau after passaging in both genotypes.
- Negative selection acts *in vitro* to prevent mutations with deleterious effects from accumulating.

OPEN ACCESS

REPORT

The Innovation

Divergent accumulation patterns of SNVs and INDELs reveal negative selection in noncancerous cells

Lei Zhang,[1,2,*] Moonsook Lee,[3] Xiaoxiao Hao,[3,12] Xiao Ma,[1,4] Chuwei Xia,[1,4] Yiwei Zhao,[1,4] Joseph Ehlert,[5] Zhongxuan Chi,[3] Bo Jin,[6] Ronald Cutler,[3] Alexander Y. Maslov,[3] Albert-László Barabási,[5,7,8] Jan H.J. Hoeijmakers,[9,10,11] Winfried Edelmann,[6] Jan Vijg,[3,*] and Xiao Dong[1,4,*]

[1]Masonic Institute on the Biology of Aging and Metabolism, University of Minnesota, Minneapolis, MN 55455, USA

[2]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN 55455, USA

[3]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[4]Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455, USA

[5]Department of Physics, Network Science Institute, Northeastern University, Boston, MA 02115, USA

[6]Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[7]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

[8]Department of Network and Data Science, Central European University, 1051 Budapest, Hungary

[9]Department of Molecular Genetics, Erasmus University Medical Center, 3015 GD Rotterdam, the Netherlands

[10]University of Cologne, Faculty of Medicine, Cluster of Excellence for Aging Research, Institute for Genome Stability in Ageing and Disease, 50931 Cologne, Germany

[11]Princess Maxima Center for Pediatric Oncology, Oncode Institute, 3584 CS Utrecht, the Netherlands

[12]Present address: The Big Data Center of Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, Guangdong 510123, China

*Correspondence: zhan8273@umn.edu (L.Z.); jan.vijg@einsteinmed.edu (J.V.); dong0265@umn.edu (X.D.)

Received: August 1, 2024; Accepted: June 23, 2025; Published Online: June 25, 2025; https://doi.org/10.1016/j.xinn.2025.101008

Citation: Zhang L., Lee M., Hao X., et al., (2025). Divergent accumulation patterns of SNVs and INDELs reveal negative selection in noncancerous cells. The Innovation **6(10)**, 101008.

Somatic mutations accumulate with age in human tissues. Clonal amplification of some mutations causes cancers and other diseases. However, it is unclear if random mutation accumulation affects cellular function without clonal amplification. We tested this in cell culture, avoiding the limitation that mutation accumulation *in vivo* leads to cancer. We performed single-cell whole-genome sequencing of fibroblasts from DNA-mismatch-repair-deficient *Msh2*$^{-/-}$ mice and controls after long-term passaging. While maintaining the same growth rates, in the *Msh2*$^{-/-}$ fibroblasts, single-nucleotide variants increased up until >50,000 per cell, with small insertions and deletions plateauing at ~16,000 per cell. We provide evidence for genome-wide negative selection and large-scale mutation-driven population changes, including significant clonal expansion of preexisting mutations and widespread cell-strain-specific hotspots, likely caused by positive selection of mutations in specific genes. Since negative selection to prevent mutations with adverse effects *in vivo* during aging is difficult to envision, these results suggest a causal role of somatic mutations in age-related cell functional decline.

INTRODUCTION

Accumulation of somatic mutations has been proposed as a cause of aging and cancer since the 1950s[1,2] DNA mutations occur spontaneously in every cell of an organism due to errors during repair or replication of a damaged DNA template.[3] However, apart from the very small fraction of mutations that are clonally amplified, typically the cause of cancer, most mutations cannot be detected by bulk sequencing and require single-cell or single-molecule approaches. Using accurate single-cell whole-genome sequencing (scWGS),[4,5] somatic single-nucleotide variants (SNVs) have recently been found to accumulate with age in every human tissue or cell type analyzed, including lymphocytes,[6] hepatocytes,[7] epithelial cells,[8] neurons,[9,10] and cardiomyocytes.[11] Somatic SNV burden ranges from a few hundred to a few thousand mutations depending on cell type and age. While confirming the original hypotheses of somatic mutation accumulation with age, it remains unclear if an increased burden of somatic mutations, in the absence of clonal amplification, has functional consequences for cells and tissues at old age.

If mutation accumulation is indeed a cause of aging, one would expect an upper limit of mutations that cells can tolerate. Here, we tested this using primary fibroblasts from a DNA-mismatch-repair (MMR)-deficient mouse model, i.e., *Msh2*$^{-/-}$ mice. The Msh2 (MutS homolog 2) gene encodes a protein that dimerizes with Msh6 and Msh3 proteins to make MutSα and MutSβ MMR complexes, respectively, and is critical for correcting base mismatches and insertion or deletion mispairs during DNA replication.[12] Such mice are known to have highly increased somatic mutation frequencies and a greatly increased risk of cancer.[13,14] The lifespan of an *Msh2*$^{-/-}$ mouse, 50% of which die within 6 months,[15] is significantly less than that of a wild-type mouse in captivity, which typically lives to about 2–2.5 years, and the expression of *Msh2* is positively correlated with the maximum lifespan across different rodent species.[16] The MMR deficiency would continually drive the generation of SNVs and small insertions and deletions (INDELs) during passaging of these cells, allowing us to test a possible limit of tolerance *in vitro* (schematically depicted in Figure 1). The results show no such limit for SNVs up until at least ~50,000 SNVs per cell, i.e., far exceeding the number of SNVs observed in most tissues upon normal aging. INDEL accumulation, however, reached a limit at ~600 and ~16,000 INDELs per cell in control and *Msh2*$^{-/-}$ cells, respectively. Our results also indicate a strong negative selection against deleterious SNVs and INDELs, suggesting that somatic mutations can adversely affect cell function *in vivo* where selection for a fitness advantage is rarely possible.

MATERIALS AND METHODS

Transgenic mice

Mice nullizygous for the *Msh2* gene were generated and backcrossed into C57BL/6 as described previously.[17] In this study, three *Msh2*$^{-/-}$ mice (4–5 months of age) and two of their wild-type littermates (4–5 months of age) were used. Two additional wild-type nonlittermates were included for cell passaging and apoptosis assay. All procedures involving animals were approved by the Institutional Animal Care and Use Committee (IACUC) of Albert Einstein College of Medicine and performed in accordance with relevant guidelines and regulations.

Bulk DNA extraction and genotyping

Genomic DNA was extracted from the tail of each mouse using the DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's specifications. The concentrations of DNA were quantified using the Qubit High Sensitivity dsDNA Kit (Invitrogen Life Science), and the quality of the DNA was evaluated with 1% agarose gel electrophoresis.

We validated the genotypes of the mouse strains by polymerase chain reaction (PCR) genotyping using the genomic DNA as template. Each reaction contained 1 μL of genomic DNA (10 ng/μL), 1.5 μL of 10× PCR buffer II (Roche), 1.5 μL of $MgCl_2$ (25 mM, Roche), 0.1 μL of Taq Gold (5 U/μL), and Primers A, B, and C (the sequences of the primers are listed in Figure S1). The total reaction volume for PCR was 12.5 μL. PCR conditions were 94°C for 5 min; 40 cycles of 94°C for 45 s, 55°C for 1 min, and 72°C for 1 min; and 72°C for 5 min. The PCR results are shown in the picture of 1% agarose gel electrophoresis (Figure S1).

Lung fibroblast isolation and passaging

Primary lung fibroblasts were isolated following a cell isolation protocol adapted from Seluanov et al.[18] In brief, mouse lung was minced and incubated in DMEM F-12 medium with 0.13 unit/mL Liberase Blendzyme 3 and 1× penicillin/streptomycin at 37°C for 40 min. Dissociated cells were washed, plated in cell culture dishes with complete DMEM F-12 medium and 15% fetal bovine serum (FBS), and cultured at 37°C, 5% CO_2, and 3% O_2. Upon reaching confluence, the cells were split and replated in Eagle's minimal

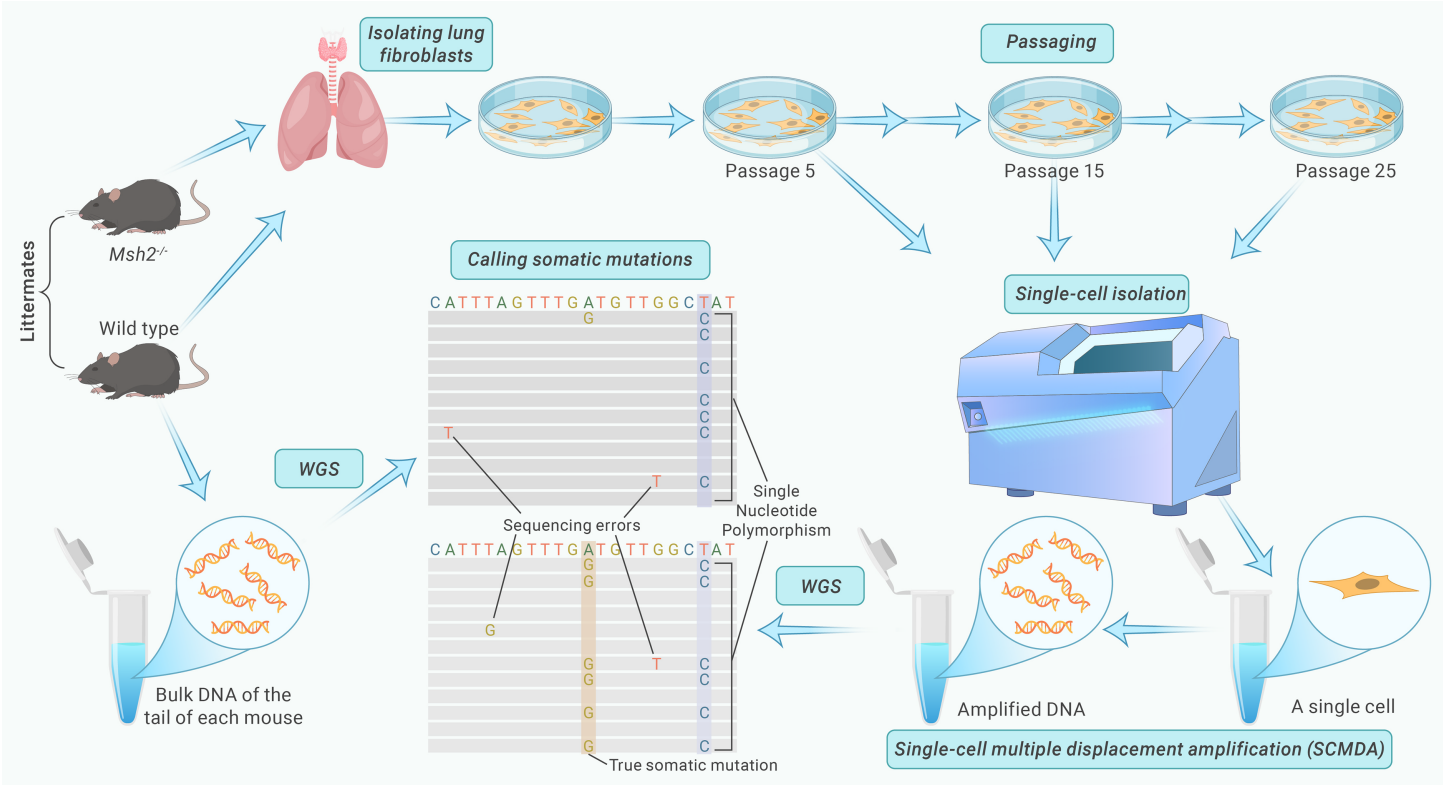


Figure 1. Study design Schematic illustration of the study design. We isolated lung fibroblasts of *Msh2*$^{-/-}$ and wild-type mice and cultured them for 25 passages. *De novo* mutations in fibroblasts in passages 5, 15, and 25 of the cell strains obtained from different animal subjects were analyzed using single-cell whole-genome sequencing and compared to bulk whole-genome sequencing of the tails of the corresponding animals.

essential medium (EMEM) supplemented with 15% FBS and 100 units/mL penicillin and streptomycin. Lung fibroblasts were purified by further passaging in the same medium.

From each animal, we passaged one cell strain. Cells from each cell strain were cultured and passaged in two 10-cm plates with EMEM supplemented with 15% FBS and 100 units/mL penicillin and streptomycin. The initial cell number was 0.5 million or 1 million for each plate each passage. We counted cell numbers during passaging by applying the Cellometer Auto T4 cell counter (Nexcelom), calculated cell population doublings based on the cell number of each cell strain, and plotted the cell proliferation curve.

Apoptosis assay

Apoptosis was assessed using the Guava Annexin Red Kit (FCCH100108, Luminex) and Guava easyCyte flow cytometer (Millipore) following the manufacturer's instructions. Data analysis was performed using GuavaSoft software. Briefly, wild-type and *Msh2*$^{-/-}$ cells were harvested at passages 15 and 25 during cell passaging. Cells were counted and re-suspended at a concentration of approximately 2×10^5 to 5×10^5 cells/mL. For each sample, 100 μL of the cell suspension was transferred to a well of a 96-well plate and mixed with 100 μL of Annexin reagent. The mixture was incubated at room temperature for 20 min in the dark. After incubation, the plate was loaded into the flow cytometer, and the "Nexin Assay Plus" program was run. All Annexin V-positive cells were considered apoptotic. The percentage of apoptotic cells in each strain at both passage 15 and passage 25 is presented in Figure S2. A comparison between wild-type and *Msh2*$^{-/-}$ cells is shown in Figure 2. The Nexin program settings and analysis parameters were kept consistent across all species, experimental conditions, and time points.

Single-cell isolation, whole-genome amplification, library preparation, and sequencing

Single lung fibroblasts were isolated using the CellRaft AIR system (Cell Microsystems) according to the manufacturer's instructions. Isolated single fibroblasts in 2.5 μL PBS were frozen immediately on dry ice and kept at −80°C until amplification.

The isolated single fibroblasts were amplified using single-cell multiple displacement amplification (SCMDA) as described.[5] The amplicons were subjected to quality control using a locus dropout test.[19] Of those passing the quality control, three amplicons per mouse were subjected to library preparation and sequencing with 150-bp paired-end reads on an Illumina HiSeq X Ten sequencer (Novogene, Inc.). Bulk DNA extracted from tails of the same mice was sequenced without amplification and used for filtering out germline polymorphisms during variant calling as described.[5]

Sequence alignment and mutation calling

Raw sequence reads were subjected to quality control using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), adaptor and quality trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), and aligned to mouse reference genome mm10 using bwa mem.[20] PCR duplicates were removed using samtools.[21] The aligned reads were then INDEL realigned and base-pair score quality was recalibrated using GATK.[22] SNVs and INDELs observed in a cell but not present in the corresponding bulk DNA of the tail were called by comparing the aligned sequences of the cell to the bulk using SCcaller (version 2.0)[23]: (1) from genomic regions covered with a minimum depth of 20× in both the cell and the bulk, (2) with default parameters for SNVs, and (3) requiring a variant calling quality ≥30 for INDELs. Mutations from the autosomes were included for analysis. Mutation burden per cell was estimated based on the number of observed mutations adjusting coverage of the genome and variant calling sensitivity. The variant calling sensitivity was estimated using the fraction of germline heterozygous mutations observed in the same single cells.[24]

Bulk RNA sequencing and data analysis

For each cell strain of different passages, total RNA was extracted using the RNeasy Micro Kit (Qiagen) according to the manufacturer's specifications. The concentrations of RNA were quantified with the Qubit RNA HS Assay Kit (Invitrogen Life Science), and the quality of the RNA was evaluated using a bioanalyzer with the Agilent RNA 6000 Pico Kit (Agilent Technologies). The qualified RNA samples (RIN [RNA Integrity Number] ≥7.0, OD260/280 >2.0, concentration ≥20 ng/μL, and volume ≥20 μL) were submitted to Novogene for library preparation and sequencing. The insert size of the double-strand cDNA library is 250–300 bp. The libraries of the RNA samples were sequenced on the Illumina NovaSeq 6000, with 2× 150-bp paired-end reads. The average sequencing amount of raw data of each library was 9.24 Gbp.

Raw sequence reads were subjected to quality control using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), adaptor and quality trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), and aligned to the mouse reference transcriptome mm10 using STAR.[25] Gene expression levels

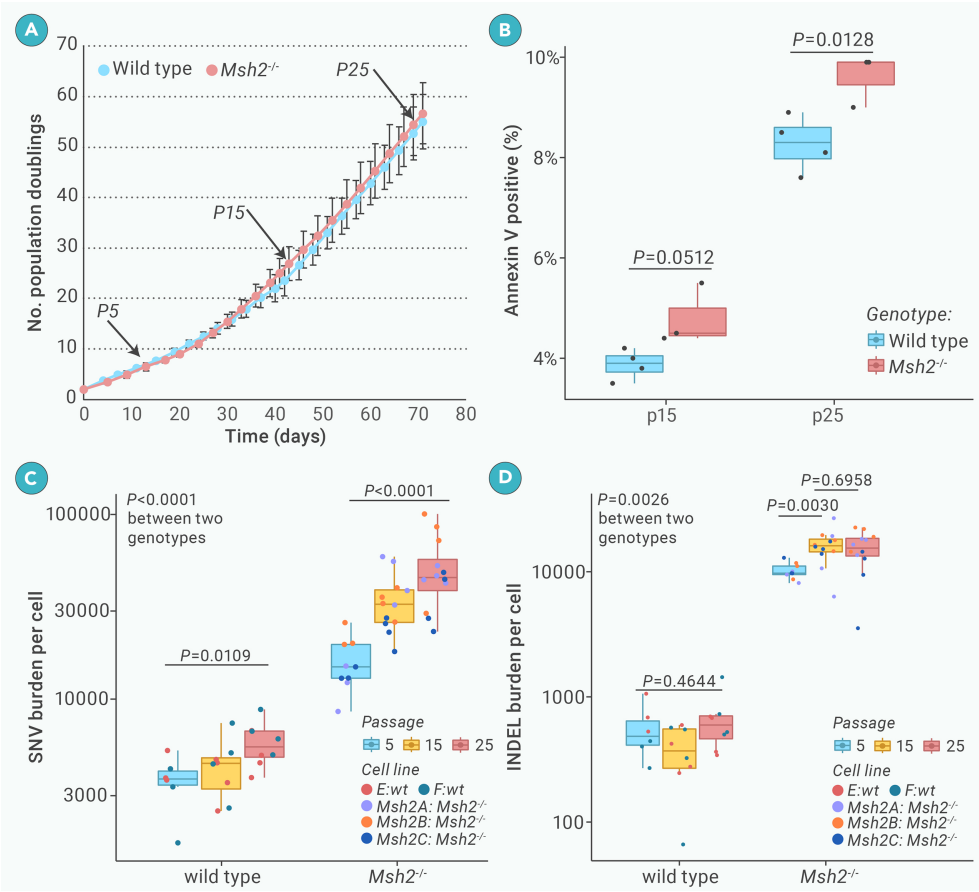


Figure 2. Cell growth and mutation burden (A) Cell growth during passaging. Error bars present SD. (B) Fractions of apoptotic cells. The *p* values were estimated using Student's t test, one-tailed. Individual data points of the two passages are presented in Figure S2. (C and D) (C) SNV burden and (D) INDEL burden per cell on log scales. Each data point represents a cell. The *p* values were estimated using linear mixed-effects models, two-tailed, using the "nlme" package of R. Boxplot elements are defined as follows: center line indicates median, box limits indicate upper and lower quartiles, and whiskers indicate 1.5× interquartile range.

were quantified using RSEM.[26] Expressed protein-coding genes were determined as those with an average transcripts per million (TPM) value ≥1 across all samples.

Single-cell RNA sequencing and data analysis

We performed single-cell RNA sequencing (scRNA-seq) targeting 3,000 cells per sample of passages 5 and 25 of both cell strains E and Msh2A using the 10× Chromium system at the Genomics Core at the Albert Einstein College of Medicine, and sequencing was performed using the Illumina NovaSeq 6000, with 2× 150-bp paired-end reads, by Novogene.

Raw sequencing data were aligned to the mouse reference genome (mm10) using CellRanger. DropletQC (v.0.9) was used to remove empty droplets containing ambient RNA from the gene expression matrices.[27] Scrublet (v.0.2.3) was applied to identify and remove doublets with default settings.[28] The expression matrices were merged and processed in Seurat (v.5.0.3).[29] Cells with fewer than 20,000 or more than 250,000 nCount_RNA, fewer than 4,500 nFeature_RNA, or more than 5% of reads mapping to the mitochondrial genome were further excluded.

Somatic mutations found in scWGS data were mapped to scRNA-seq data by examining sequencing reads that cover the corresponding bases in the scRNA-seq data. Mutated sites detected in fewer than 500 cells were discarded. By integrating cluster and sample labels, we defined five groups: Cluster1_Msh2A_p25, Cluster2_E_p25, Cluster3_Msh2A_p25, Cluster4_E_p5, and C4_Msh2A_p5. The mutant ratio was calculated as the number of cells with a detected mutation divided by the total number of cells exhibiting any signal (mutant or nonmutant). We retained only those positions where a single group had a mutant ratio >0.1, while all others remained <0.05. Thirty-three potential mutations remained. Among them, nine were found to locate in genes that are widely expressed across most cells. Here are the genes affected: in Cluster1_Msh2A_p25, *Pigu* and *Med27*, and in Cluster3_Msh2A_p25, *Med29*, *Ppp2r1a*, *Exosc8*, *Gnb1*, *Ncapd2*, *Sec24c*, and *Luzp1*.

RESULTS

Somatic mutation burden in *Msh2*$^{-/-}$ mouse fibroblasts

Mice nullizygous for the *Msh2* gene were generated and backcrossed into C57BL/6 as described previously.[17] Their genotypes were validated using PCR of the DNA extracted from their tails (Figure S1). Lung fibroblasts isolated from three *Msh2*$^{-/-}$ mice (4–5 months of age) and four wild-type mice, i.e., two wild-type littermates (4–5 months of age) and two additional, nonlittermate wild-type mice (C57BL/6, 6 months of age), were cultured for 25 passages up to a total of 62 population doublings. As shown in Figure 2A, growth rates of the three *Msh2*$^{-/-}$ and four wild-type fibroblast strains were almost identical, with no morphologic evidence for neoplastic transformation. We next assessed the level of apoptosis by Annexin V and observed a significant increase in the percentage of apoptotic cells from passage 15 to passage 25 in both genotypes. At passage 25, the proportion of apoptotic cells was marginally but significantly higher in *Msh2*$^{-/-}$ cells compared to wild-type cells (Figures 2B and S2). These results indicate that *Msh2*$^{-/-}$ cells are more prone to growth defects than wild-type cells; however, this is compensated for by enhanced growth of the surviving cell population.

To quantitively analyze somatic mutation burden, we performed scWGS on 55 single cells at passages 5, 15, and 25 (denoted as P5, P15, and P25, respectively) of the three *Msh2*$^{-/-}$ cell strains and the two wild-type littermate cell strains (Figure 1). Of note, the SCMDA and variant calling procedure (SCcaller) have been designed to avoid artificial mutations, previously the main problem in somatic mutation analysis.[5,23] For each cell strain, we also performed whole-genome sequencing of tail DNA from the same mice to identify germline polymorphisms, which were filtered out in calling *de novo* somatic mutations from the single cells. Depth of sequencing reached on average 27.5× and 21.4× per sample for single cells and bulk DNAs, respectively (Table S1), to ensure that mutations could be identified accurately. One potential challenge is the coverage uniformity of scWGS across the genome. The scWGS protocol that we employed provides whole-genome coverage for all cells (Table S1). While the sequencing depth distribution is uneven across the genome and varies among individual cells, the average coverage among single cells provides coverage comparable to that of bulk WGS (Figure S3A). While this still limits the ability to discover all mutations from each cell, sequencing multiple cells from the same cell population compensates for that.

From the scWGS data on the five cell strains, we identified a total of 192,933 *de novo* mutations, including 147,955 SNVs and 44,978 INDELs, which was sufficient for analyzing mutation burden, spectrum, and distribution across the genome, especially for the *Msh2*$^{-/-}$ strains because of their high mutation frequencies (below). We also plotted the variant allele fraction (VAF; i.e., the number of reads reporting the mutation compared to the total read number at the same loci in the same cell) distribution for all somatic mutations (Figure S3B). The average VAF centers around 50%, indicating overall high accuracy in variant calling. However, we observed a small subset of mutations with lower VAFs (~30%), which could result from either (1) false-positive calls, which cannot be entirely excluded without significantly sacrificing sensitivity, or (2) some true variants, particularly INDELs, that differ from the reference genome and thus align less efficiently.

After correcting for the sensitivity of variant calling and genome coverage (Table S2), we found that, as expected, *Msh2*$^{-/-}$ cells had a significantly higher SNV burden than wild-type cells across all passages ($p < 0.0001$, linear

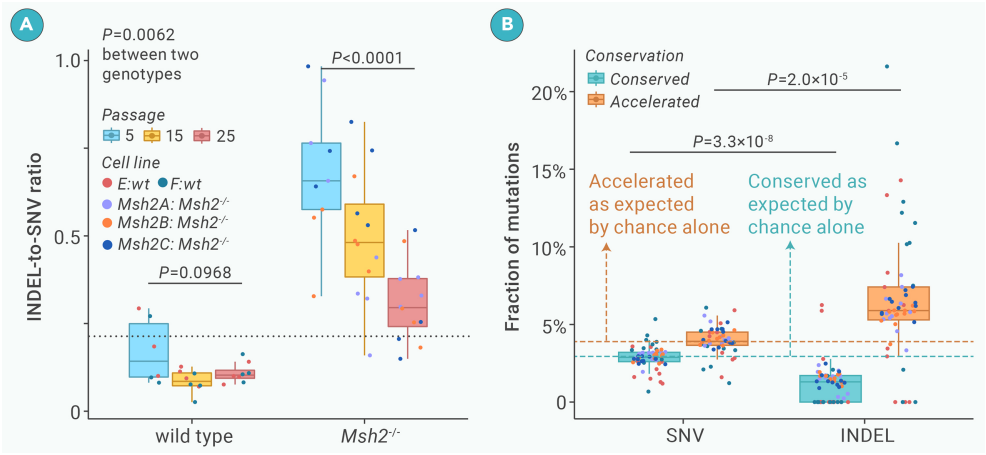


Figure 3. Selection pressure against INDELs (A) The ratio of the number of INDELs to the number of SNVs per cell. (B) The fraction of mutations (SNVs and INDELs combined) at evolutionarily conserved and accelerated sites out of total mutations (wild type and *Msh2*$^{-/-}$ combined). The fractions of SNVs and INDELs at conserved and accelerated sites by chance alone were estimated based on randomly generated mutations using SigProfilerSimulator.[31] We randomly generated the same number of SNVs and INDELs as the observed numbers with also the same mutation signature, performed the same analysis of their conservation scores, and repeated the above two steps 2,000 times to reach stable estimations. Because there is no difference between the values of SNVs and INDELs expected by chance alone, we merged them into two single values as indicated by the two dashed lines (for conserved and accelerated sites separately). Boxplot elements are defined as follows: center line indicates median, box limits indicate upper and lower quartiles, and whiskers indicate 1.5× interquartile range.

mixed-effects model, two-tailed). In wild-type cells, SNV burden increased with passage number in fibroblasts from 3,618 ± 1,185 per cell (average ± SD; P5) to 5,817 ± 1,623 per cell (P25) in the wild-type cells (p = 0.0109, linear mixed-effects model, two-tailed) (Figure 2C), which corresponds to a mutation rate of ~8.2×10^{-9} per base pair per mitosis, almost the same as we estimated earlier for mouse primary fibroblasts (8.1×10^{-9} per base pair per mitosis).[19] In the *Msh2*$^{-/-}$ cells, SNV burden increased from 16,080 ± 5,381 per cell (P5) to 53,146 ± 24,701 per cell (P25) (p < 0.0001, linear mixed-effects model, two-tailed). There was no sign of a plateau between P5 and P25, not even in the *Msh2*$^{-/-}$ cells after acquiring tens of thousands of SNVs per cell. At P5, SNV burden in *Msh2*$^{-/-}$ cells was more than 4-fold higher than in the cells from their littermate controls. Since we did not compare cells at different stages of embryonic development, we do not know how many more somatic mutations were present in the *Msh2*$^{-/-}$ mice from embryogenesis to early adulthood as compared to control mice, but it is safe to say that the original estimates based on reporter genes have been seriously overstated, i.e., 35–550 mutations per 10^5 bp, corresponding to $1–15 \times 10^6$ mutations per cell.[14]

INDELs showed a different pattern of accumulation during passaging compared with SNVs (Figure 2D). As expected, *Msh2*$^{-/-}$ cells had a significantly higher INDEL burden than the wild-type cells across all passages (p = 0.0026, linear mixed-effects model, two-tailed). INDEL burden during passaging only increased by 1.6-fold in the *Msh2*$^{-/-}$ cells between P5 and P15 (10,172 ± 1,506 and 16,150 ± 4,995 INDELs per *Msh2*$^{-/-}$ cell for P5 and P15, respectively; p = 0.0030, linear mixed-effects model, two-tailed), but not between P15 and P25 (16,150 ± 4,995 and 15,370 ± 5,323 INDELs per *Msh2*$^{-/-}$ cell for P15 and P25 separately; p = 0.6958, linear mixed-effects model, two-tailed). In cells from the littermate controls, no significant increase was observed during passaging (565 ± 280 and 660 ± 346 INDELs per cell for P5 and P25 separately; p = 0.4644, linear mixed-effects model, two-tailed). These results suggest that INDEL tolerance reaches an upper limit in both wild-type and *Msh2*$^{-/-}$ cells, but earlier in the control cells. However, INDELs are likely to occur at high frequency in MMR-deficient cells and mostly in repetitive regions, most notably microsatellites, where they are likely to be less toxic (see the following section for analyses). Nevertheless, the plateau of INDEL induction in both WT and *Msh2*$^{-/-}$ cells, but not SNVs, indicates toxicity of the former, without apparently adversely affecting growth rate of primary fibroblasts.

Genome-wide selection against damaging mutations

The results thus far appear to suggest that increased burden of somatic mutations per se, i.e., without clonal amplification, cannot cause cellular degeneration and death. Indeed, somatic mutation burden in tissues of aged humans or mice never reaches levels as observed in the MMR-deficient cells.[30] However, while during *in vivo* aging, when most tissues are not mitotically active, selection against mutations that adversely affect cellular function is difficult to envision, primary fibroblasts expanded *in vitro* offer an immediate mechanism of avoiding adverse somatic mutations by selection against mutations causing growth inhibition. In this respect, INDELs are generally more damaging than SNVs, which are often synonymous with no impact at all. To address the different impacts of INDELs and SNVs in *Msh2*$^{-/-}$ and control cells during passaging we performed three comparisons as follows.

First, to test if the selection against INDELs is significantly stronger than the selection against SNVs, we calculated the ratio of INDEL burden to SNV burden for each single cell. As shown in Figure 3A, there is a trend of decrease in INDEL-to-SNV ratio in fibroblasts of both genotypes: from 0.17 ± 0.09 (P5) to 0.11 ± 0.03 (P25) in the wild-type cells (p = 0.0968, linear mixed-effects model, two-tailed) and from 0.69 ± 0.20 (P5) to 0.31 ± 0.11 per cell (P25) in the *Msh2*$^{-/-}$ cells (p < 0.0001, linear mixed-effects model, two-tailed), i.e., a 2.2-fold decrease. These results indicate negative selection against INDELs during passaging in cells of both genotypes.

Second, to evaluate possible negative selection for both INDELS and SNVs, separately, we utilized phyloP scores,[32,33] with a positive score indicating conservation, i.e., slower evolution than expected, and a negative score indicating acceleration, i.e., faster evolution than expected. We obtained phyloP scores for all bases of the mouse reference genome from the UCSC genome browser.[34] We then defined mutations at evolutionarily conserved sites as those with a phyloP score >0, its original p < 0.05, and a percentile of the phyloP score of the mutated site, as compared to the phyloP scores of its ±500 flanking bases, of >95% (to avoid a potential difference in genome coverage). Mutations at evolutionarily accelerated sites were defined by a phyloP score <0, its original p < 0.05, and a percentile of the phyloP score of the mutated site, as compared to the phyloP scores of its ±500 flanking bases, of <5%.

For both SNVs and INDELs in both wild-type and *Msh2*$^{-/-}$ cells, the fraction of mutations at an evolutionarily conserved site was substantially lower than that at an accelerated site (Figure 3B). However, compared to mutations randomly sampled from the genome, we found that the fractions of SNVs at both conserved and accelerated sites were as expected by chance alone, while the fractions of INDELs were substantially different from the random sampling. A significantly smaller fraction of INDELs (1.2% ± 1.3%) was observed at a conserved site compared with SNVs (2.9% ± 0.8%; $p = 3.3 \times 10^{-8}$, paired Wilcoxon signed-rank tests, two-tailed) or expected based on chance alone. By contrast, a greater fraction of INDELs was found at an accelerated site compared with SNVs (6.6% ± 4.0% and 4.0% ± 0.9% for INDELs and SNVs respectively; $p = 2.0 \times 10^{-5}$, paired Wilcoxon signed-rank tests, two-tailed) or as expected by chance alone. Of note, in 77% of wild-type cells, we did not observe any INDELs at a conserved site. During passaging, no significant change was observed between SNVs and INDELs at accelerated and conserved sites in cells of the two genotypes (linear mixed-effects models, two-tailed; Figures S4A–S4D), with two exceptions. First, we found a marginal increase in INDELs at conserved sites in *Msh2*$^{-/-}$ cells during passaging (p = 0.0455, i.e., no longer significant if adjusting for multiple testing; Figure S4C), while the fraction of INDELs at conserved sites remained much lower than expected by chance alone. Second, there was a significant decrease in SNVs at accelerated sites in *Msh2*$^{-/-}$ cells (p = 0.0011; Figure S4B). Overall, these results indicate negative selection at evolutionarily conserved sites for INDELs during passaging but not for SNVs.

Finally, we performed bulk RNA sequencing of each fibroblast cell strain to determine genes that are transcriptionally active. Using mutation

annotation by ANNOVAR,[35,36] we analyzed mutations that alter protein coding sequences of transcriptionally active genes (Table S3). We calculated the ratio of nonsynonymous to synonymous SNVs in the two genotypes during passaging and found that this ratio remains approximately the same and shows no significant difference from the ratios expected by chance alone (Figures 4A and 4B), suggesting a lack of negative selection. This was confirmed by utilizing the SIFT_4G annotation, which assesses if nonsynonymous SNVs are damaging (Figures S5A–S5C).[37] However, the trend becomes very different considering the most severe types of mutations, i.e., loss-of-function mutations, including frameshifting INDELs, and stop-gain and stop-loss SNVs. Significantly fewer frameshifting INDELs than expected by chance alone were found in these cells during passaging (0.05 ± 0.21 per cell and 3.7 ± 2.6 per cell for wild-type and *Msh2*$^{-/-}$ cells separately), as well as significantly fewer stop-gain SNVs (0.14 ± 0.47 per cell and 1.0 ± 1.5 per cell, separately) or stop-loss SNVs (0 ± 0 per cell and 0.03 ± 0.17 per cell, separately) (Figures 4C, 4D, 4F, 4G, 4I, and 4J). Of note, the ratio of observed frameshifting INDELs to that expected by chance was substantially smaller in *Msh2*$^{-/-}$ cells than in WT cells at passage 5 (Figure 4D), with INDELs no longer accumulating after passage 15 in *Msh2*$^{-/-}$ cells (Figure 2D). In addition, although the sensitivity of INDEL calling is slightly lower than that of SNV calling (Table S2), the observed number of frameshifting INDELs was higher than the total number of stop-gain and stop-loss SNVs, despite a substantial difference in their accumulation during passaging. We also estimated the ratio of each type of loss-of-function mutation to synonymous mutations and compared the ratios to those expected by chance alone. As shown in Figures 4E, 4H, and 4K, most of the ratios were significantly smaller than expected by chance alone, indicating that the limited numbers of loss-of-function mutations are a result of negative selection. This is in keeping with our previous observations that, in human B cells from aged human subjects, on average less than one loss-of-function mutation (including stop-gain, stop-loss, and splicing alteration) per cell was observed.[6] Hence, these results do indicate that negative selection occurs in SNVs also, although this is limited to those SNVs expected to be most severe. This was confirmed by the significantly lower ratio of observed vs. expected frameshifting INDELs.

Each *Msh2*$^{-/-}$ cell strain acquires common and unique mutational signatures during passaging

As shown in studies of human cancers, mutational spectra and signatures suggest specific factors that drive mutagenesis, e.g., oxidative damage or radiation.[38,39] However, connection between mutation signatures and causal factors are often derived computationally. In this study, we had an opportunity to test if passaging and DNA MMR deficiency indeed cause the mutational signatures inferred from human cancers.

First, we compared SNV spectra between the cell strains. As expected, *Msh2*$^{-/-}$ cells are substantially different from wild-type cells, with more C>T and T>C mutations (Figure S6A). However, we noticed substantial variation between the three *Msh2*$^{-/-}$ cell strains: the Msh2A cell strain acquired more T>C mutations, the Msh2C cell strain acquired more C>T mutations, and the Msh2B cell strain was in between (Figure 5A). Of note, their unique mutational spectra became more obvious during passaging (Figure S6B).

Then, we performed SNV signature analyses in two ways, both using the "MutationalPatterns" package of R.[40] First, we performed *de novo* signature extraction and identified three signatures (Figure 5B). Using a cosine correlation cutoff at 0.85 with known mutational signatures of human cancers reported in the COSMIC database,[39] we labeled the three signatures as SBS-A (no similar cancer signature was found), SBS26-like (positively correlated with the COSMIC Single Base Substitution signature #26), and SBS44-like signatures. The SBS26-like signature dominates mutations in the Msh2A cell strain, and its fraction out of all mutations increases with passaging, while the SBS44-like signature is more dominant in the Msh2C cell strain (Figure 5C). Of note, both SBS26 and SBS44 signatures in tumors have been suggested to be the result of DNA MMR deficiency.[39] The SBS-A signature, which was not reported in the COSMIC database, contributes to most mutations in the wild-type cells (Figure 5C) and is likely a result of replication errors. However, SBS-A (characterized by NTT>NGT or NCT mutations; Figure 5B) is very different from the SBS1 signature (characterized by NCG>NTG mutations)[39] in human tumors, which has been associated with cell division.

Second, we refitted COSMIC signatures to the mutations that we observed. When doing that, we found another DNA MMR signature, i.e., SBS21, in the *Msh2*$^{-/-}$ cell strains, but the differences between the *Msh2*$^{-/-}$ cell strains remained (Figure S7). Together, despite confirming that MMR deficiency can indeed cause the corresponding signatures found in human cancers, these results indicate that a single factor, i.e., *Msh2* deficiency, can result in different mutational signatures.

For INDELs, we also performed signature extraction and identified two signatures: an ID2-like signature (positively correlated with the COSMIC small Insertion and Deletion signature #2), which is characterized as a single-base T deletion in repetitive T sequences, and another new signature, termed IDA, which does not correlate with a COSMIC signature (Figure S8A). IDA was mostly found in our wild-type control cells (Figure S8B) and is characterized by either insertion or deletion at repeat regions of multiple homopolymers or repeat units. The ID2-like signature, mostly single base deletions in a long homopolymer of thymines, was predominantly found in our *Msh2*$^{-/-}$ cell strains (Figure S8B). The ID2 signature in human cancers is suggested to be caused by slippage during DNA replication of the template DNA strand and is often found in DNA MMR-deficient tumors.[39] Of note, in the COSMIC database, another INDEL signature, ID7, characterized by 1-bp deletions at homopolymers of both cytosines and thymine and suggested to be a result of MMR deficiency in humans, was not observed here.

Hotspots, mutational overlap, and positive selection

We then tested for mutational hotspots (for SNVs and INDELs together) in the mouse genome by using the "ClusteredMutations" package in R. A substantial number of mutational hotspots were observed in both WT and *Msh2*$^{-/-}$ fibroblasts, but significantly more in the *Msh2*$^{-/-}$ cells (Figure 6A). Surprisingly, mutational hotspots were so obvious, even in wild-type cells, that we could identify them for each individual cell, while in our previous study of human lymphocytes we had to pool mutations observed in tens of cells to discover significant mutational hotspots.[6] We then used a rainfall plot to visualize the distribution of the mutational hotspots across the genome. Again, different cell strains showed substantially different patterns (Figure 6B). The Msh2A strain continuously gained additional mutational hotspots at the end of chromosome 17, while in the Msh2B cell strain, which showed the highest number of mutational hotspots, these spread across the entire reference genome during passaging. Two "super-hotspots" are worth noticing. One is at chr17:86,631,535–90,041,858 bp, found exclusively in the Msh2A cell strain. Interestingly, *Msh2* and *Msh6* genes are located in this region along with over 20 other genes, but all mutations in the hotspots at this region are located at intergenic sequences. The other super-hotspot was found at chr1:170,941,871–170,943,280 bp and was observed in four of the five cell strains (two WT and two *Msh2*$^{-/-}$) but not in the Msh2A strain. This region is entirely intergenic and is part of a long terminal repeat (LTR) element.

Why would each *Msh2*$^{-/-}$ cell strain develop its own unique pattern of mutational hotspots? It is possible that substantial clonal expansion occurred during passaging, and each cell strain was eventually dominated by different clones. To test this, we calculated for each cell in each cell strain (of both WT and *Msh2*$^{-/-}$) the ratio of (1) the mutations overlapping with mutations in other cells of the same passage and cell strain to (2) the mutations found to overlap in all cells of all cell strains. A higher ratio indicates more clonal expansion. As shown in Figures 6C and S9, ratios increase dramatically during passaging in cell strains of both genotypes: from 6.0 ± 6.6 (P5) to 27.3 ± 22.1 (P25) in wild-type cells (p = 0.0192, linear mixed-effects model, two-tailed) and from 1.7 ± 1.9 (P5) to 71.9 ± 58.7 (P25) in *Msh2*$^{-/-}$ cells ($p < 2.2 \times 10^{-16}$, linear mixed-effects model, two-tailed). Although the difference between cells of the two genotypes was not statistically significant (p = 0.3967, linear mixed-effects model, two-tailed), likely due to large cell-to-cell variations, the increase in *Msh2*$^{-/-}$ cells was substantially higher (a 42-fold increase from P5 to P25) than in the wild-type cells (a 4.6-fold increase).

We next investigated whether overlapping mutations identified within the same cell strains occurred in genes known to be cancer drivers in humans, as derived from the COSMIC Cancer Gene Census.[42] Indeed, we found multiple overlapping mutations affecting known "cancer driver" genes across all three

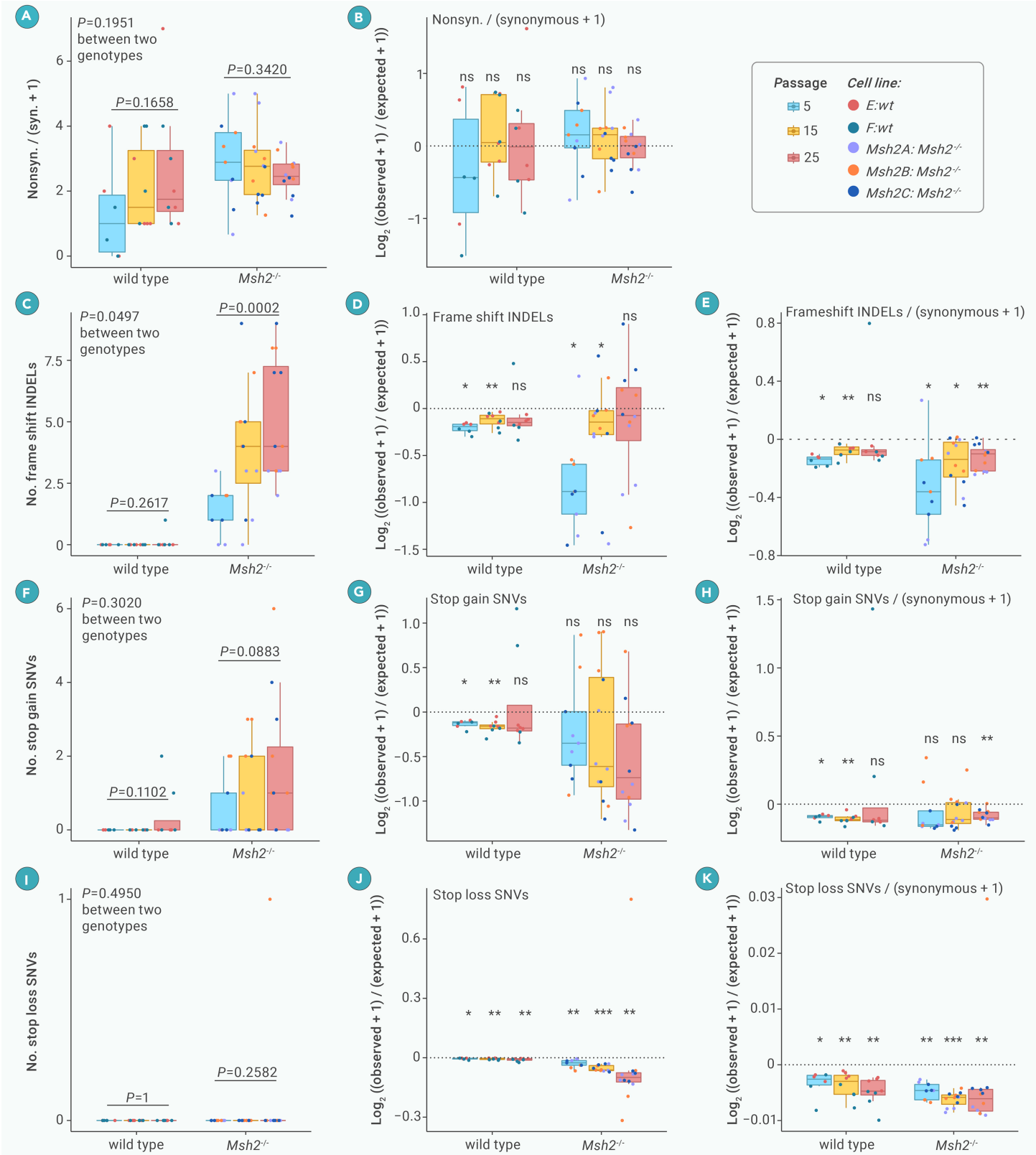


Figure 4. Selection pressure against damaging mutations (A) The ratio of the number of nonsynonymous mutations to synonymous mutations. We added 1 to the denominator values to avoid potential 0. (B, E, H, and K) The observed ratios of the numbers of nonsynonymous, frameshifting, stop-gain, and stop-loss mutations to the numbers of synonymous mutations vs. their corresponding ratios expected by chance alone. (C, F, and I) The numbers of frameshifting, stop-gain, and stop-loss mutations per cell. (D, G, and J) The numbers of observed frameshifting, stop-gain, and stop-loss mutations vs. their corresponding numbers expected by chance alone (in log2-transformed ratios). To estimate the number of mutations expected by chance alone, we first used SigProfilerSimulator[31] to randomly generate the same number of SNVs and INDELs as the observed numbers with also the same mutation signature, then annotated the artificial mutations with ANNOVAR[36] to determine the number of mutations in each functional category, and finally repeated the above two steps 2,000 times to reach stable estimations. Each dot represents a cell. The *p* values in (A), (B), (C), (D), (F), and (H) were estimated using linear mixed-effects models, two-tailed. In (E), (G), and (I), ns represents $p > 0.05$ and *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$, separately, which were estimated using binomial tests, two-tailed. Boxplot elements are defined as follows: center line indicates median, box limits indicate upper and lower quartiles, and whiskers indicate 1.5× interquartile range.

Figure 5. SNV spectra and signatures (A) SNV spectra of each cell strain. Error bars present SD. (B) Three SNV signatures of the fibroblasts identified by *de novo* signature extraction using the "MutationalPatterns" package of R.[40] (C) Contribution of each SNV signature to the total SNVs per cell.

Msh2$^{-/-}$ cell strains, whereas no such mutations were detected in wild-type cells (Table S4). Two observations are noteworthy. First, most mutated genes, including *Plcg1*, *Mtor*, and *Ccnd1*, are involved in cell growth regulation. These mutations may potentially compensate for growth deficiencies caused by other deleterious genomic mutations in the same and/or different cells of the same cell population, thereby resulting in comparable overall growth rates between wild-type and *Msh2*$^{-/-}$ cells. Second, we identified an overlapping mutation in the *Ercc2* gene exclusively in the Msh2A strain. *Ercc2* plays a crucial role in nucleotide excision repair (NER), a critical pathway for repairing DNA damage. In contrast, an *Ercc2* mutation was found in only one cell of the Msh2B strain, and no *Ercc2* mutations were detected in the Msh2C strain. Interestingly, Msh2A and Msh2C strains displayed the greatest differences in mutational signature, while Msh2B exhibited intermediate characteristics (Figure 5C). While the mutational signatures observed here differed from previously reported *Ercc2*-associated signatures found in human cancers,[43] the Ercc2 deficiency in our study arose specifically within an *Msh2*$^{-/-}$ genetic background, potentially differing from genetic contexts typically found in human cancers. These results collectively suggest that the distinct mutation signatures and hotspot variations we observed may reflect consequences of secondary mutations.

To further validate the clonal expansion, we performed scRNA-seq on wild-type (E strain) and *Msh2*$^{-/-}$ (Msh2A strain) cells at passages 5 and 25 using the 10× Genomics platform, analyzing a total of 10,277 single cells that passed quality control. Based on UMAP (Uniform Manifold Approximation and Projection) clustering of the transcriptomic profiles (Figure 6D), we observed that at passage 5, both wild-type and *Msh2*$^{-/-}$ cells grouped together within a single cluster. However, by passage 25, while wild-type cells continued to form a single cohesive cluster, *Msh2*$^{-/-}$ cells separated into two distinct clusters. Several mutations identified by scWGS were found almost exclusively in one of the *Msh2*$^{-/-}$ cell clusters at passage 25. These mutations occurred in genes such as *Med29*, *Ppp2r1a*, *Gnb1*, and *Med27*, which are associated with transcription regulation and cell growth (Figure 6E).

These results confirm the occurrence of substantial clonal expansion during passaging of cells of both genotypes, especially in the *Msh2*$^{-/-}$, with different cell strains taken over by different clones. This process is a likely cause of the different mutational signatures and hotspots observed in different cell strains and very likely has already started during development of these mice before cell isolation. These results also suggest strong positive selection of

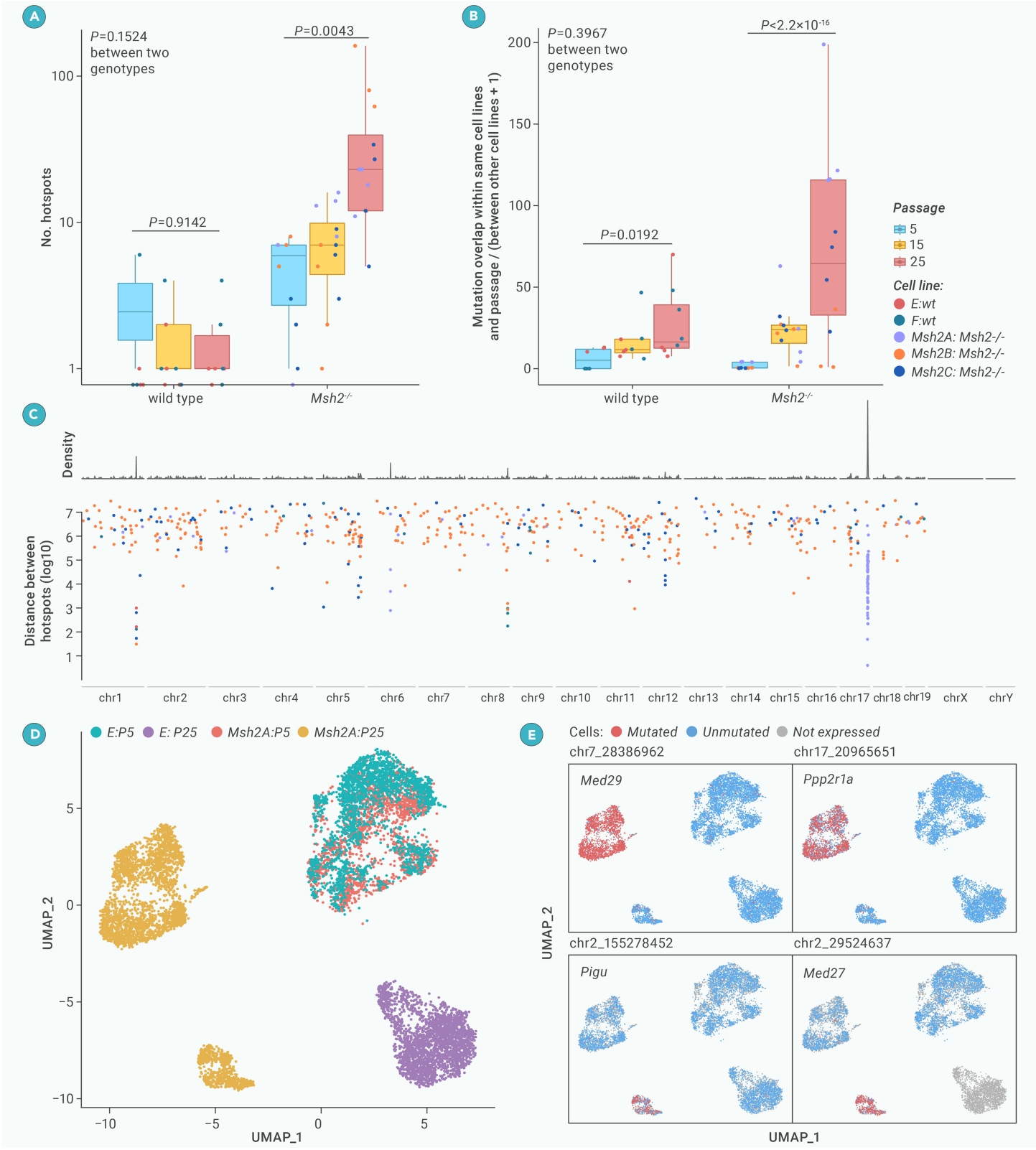


Figure 6. Mutational hotspots and overlap (A) The number of mutational hotspots (SNV and INDELs combined) per cell. (B) A rainfall plot of the distribution of mutational hotspots across the genome. The plot was generated using the "karyoploteR" package of R.[41] Each data point represents a mutational hotspot observed within a single cell. (C) The ratio of the number of overlapping mutations among cells of the same passage and same cell strain (i.e., animal) to the number of overlapping mutations among all cells of all strains. We added 1 to the denominator values to avoid potential 0. Each data point represents a cell. The *p* values were estimated using linear mixed-effects models, two-tailed. Boxplot elements are defined as follows: center line indicates median, box limits indicate upper and lower quartiles, and whiskers indicate 1.5× interquartile range. (D) UMAP plot of the scRNA-seq data. (E) Somatic mutations found in both scWGS and scRNA-seq data of the same cell strains. The top four genes that are found almost exclusively associated with one-cell clusters are shown.

specific cell lineages in the different cell strains, which is frequently observed in tumor cells.[44]

DISCUSSION

With the emergence of advanced high-throughput sequencing methods, including high-accuracy single-cell and single-molecule methods, increased insights are now being obtained in somatic rather than germline mutations as a possible cause of human genetic disease and aging.[3,45] Mutation frequency in somatic cells and tissues appeared to be 1–2 orders of magnitude higher than germline mutation frequency.[19] This is in keeping with the disposable soma theory of aging, which states that reproduction is prioritized over somatic maintenance.[46] This idea is in line with the observed correlation of somatic maintenance and species-specific lifespan.[47] Indeed, we and others recently showed that somatic mutation rate is inversely correlated with species-specific lifespan.[24,48]

Recent findings that somatic mutation burden increases with age in different human tissues[30] support a possible causal role of somatic mutations in the aging process. Indeed, clonally amplified somatic mutations, which are relatively easy to detect by high-depth sequencing, have now been shown to be a cause of a large number of human diseases other than cancer.[45,49] However, what remains unclear is if increased somatic mutation burden per se can cause cellular degeneration and death. In this respect, a key question is if random somatic mutations can rise to a level high enough to infringe on the integrity of the gene regulatory pathways that provide function to the specialized somatic cells in the human body. Here, we present mutation accumulation data for a simplified cell culture model in the form of mouse primary fibroblasts with mutations continuously generated through a defect in DNA MMR.

The first conclusion that can be drawn from our data is that somatic SNVs can accumulate to levels at least six times as high as observed in human postmitotic tissues from aged subjects.[7,9] Our finding that these high numbers of random mutations have no significant effect on growth rate seems to rule out a causal role of somatic mutations in aging. However, in contrast to the situation during normal aging, cell culture systems are subject to selection against deleterious mutations affecting growth. We found ample evidence for such selection in all fibroblast strains studied, including the control, wild-type strains. First, among SNVs we found significant negative selection against stop-loss and stop-gain mutations. Second, while SNV burden never reached plateau levels up until a population doubling level (PDL) of 50–60 (i.e., P25, Figure 2C), INDEL burden did not increase in controls and no longer increased after 20–30 PDL (i.e., P15) in the Msh2-deficient cells. These observations are different from mutations in human tumors, in which positive selection has been shown to outweigh negative selection.[44] However, others have reported evidence for negative selection also during cancer evolution.[50,51]

Of note, in mitotically active human B lymphocytes, we previously found the rate of age-related SNV accumulation in the ~10% functionally active part of the genome to be only half of the genome-wide average.[6] Yet, except for loss-of-function SNVs, which do not increase with age in human lymphocytes, the number of potentially functional SNVs still accumulated with age, even in subjects in their 80s or 90s.[6]

In addition to the evidence for direct selection against deleterious mutations, most notably INDELS, we also found evidence for widespread mutational hotspots and significant clonal expansion. Both differed between the cell strains studied, gradually leading to unique populations in each strain. Together with direct selection against deleterious mutations, such mutational evolution could be responsible for maintaining normal growth rate even after acquiring tens of thousands of SNVs and over 10,000 INDELs in the Msh2-deficient cells.

The fact that somatic mutations, either spontaneous or driven by the MMR defect, show such dramatic evolutionary dynamism in culture strongly suggests they have functional consequences. If they were completely neutral, none of these effects would be expected to occur. However, with some possible exceptions (e.g., the lymphoid and intestinal systems), adult tissues have limited options for negative selection based on growth since most are not mitotically active. Moreover, negative selection against mutations not related to growth rate but adversely affecting critical functions of the host is difficult to imagine. By contrast, positive selection, as seen in clonal amplification of somatic mutations due to a growth or survival advantage, occurs in most if not all tissues during human aging. The best example is clonal hematopoiesis,[52] but such positive selection has been seen in many other tissues and was, as expected, associated with both age and tissue-specific cell proliferation rate.[53]

Regarding mutational signatures, in human cancer studies, the discovery of mutational signatures has typically involved decomposing mutation patterns observed across different samples.[54] The relationship between each signature and its causal factor(s) was predominantly established through correlation analyses. Multiple signatures found in human tumors, including SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44, have been associated with MMR deficiency, and it has remained unclear whether these signatures result from mutations in different genes or even from the same mutation. Our results clarify this causal relationship and demonstrate that distinct mutational signatures can indeed arise from the same initial mutation, i.e., inactivation of *Msh2*, while the differences observed between signatures may be attributable to variations in secondary mutations acting as modifiers.

Some limitations of our current study should be mentioned. One limitation is the driver of the high level of somatic mutagenesis itself. MMR deficiency does not elevate all categories of mutations equally, and it can be argued that the most impactful mutations, including genome structural variation, are not significantly elevated at all, while less damaging mutations, such as regulatory and nonsynonymous SNVs, may be compensated for by concurrent fitness-enhancing mutations. Indeed, this could be one of the reasons for a lack of premature aging in MMR-deficient mice or humans.[55] Another reason could simply be the lack of detailed analysis of premature aging in MMR-deficient mice or humans (who usually die from cancer well before old age), which is not trivial.[56] Another limitation involves our use of mouse fibroblasts. Due to the absence of recent tools capable of fully annotating mutation impact in the mouse genome, our analysis was constrained primarily to clearly deleterious mutations, such as frameshift, stop-gain, and stop-loss mutations, or depended on conservation scores. To address these limitations, our collaborators have treated human fibroblasts with multiple low doses of N-ethyl-N-nitrosourea (ENU), a chemical compound known primarily for inducing point mutations.[57] Their findings demonstrated that individual cells could accumulate approximately 60,000 SNVs with only minor negative effects on cell growth rates. Their data also indicate that cells manage these high mutation loads through selective elimination of variants within gene-coding regions and critical biological pathways involved in cell growth and survival. Finally, our study utilized an *in vitro* model, which does not allow our conclusions to extend to the *in vivo* situation. Indeed, our approach was primarily geared toward exploring the possibility of an upper limit to the number of *de novo* mutations a typical mammalian cell can tolerate.

In summary, our present data uncover the comprehensive landscape of somatic mutations in MMR-deficient mouse primary fibroblasts as compared to wild-type control cells passaged *in vitro*. The results show that the MMR-deficient cell populations maintain high growth rates despite an SNV burden of at least 50,000 mutations per cell, while INDEL burden reaches a plateau of about 16,000 per cell. Further analysis showed extensive somatic evolution, including negative selection to maintain growth rate, possibly by eliminating deleterious mutations. We conclude that in the absence of such selection options, deleterious effects of accumulating somatic mutations to the levels that have been observed *in vivo* is inevitable. Further research on cell populations that can be directly interrogated for a functional relationship between somatic mutation burden and specific cellular functions known to decline with age will provide a more definitive test of a causal relationship between somatic mutations and aging.

RESOURCE AVAILABILITY

Materials availability

This study did not generate new materials.

Data and code availability

Raw sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) under accession number SRA: PRJNA1267087.

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the American Federation for Aging Research (the Sagol Network GerOmic Award for Junior Faculty), the US National Institutes of Health (P01

AG017242, P01 AG047200, P01 AI172501, P01HL132825, P01 HL160476, P30 AG038072, R00 AG056656, U01 ES029519, U01 HL145560, and U19 AG056278), the University of Minnesota (Fesler-Lampert Chair for Aging Studies), the Veteran's Affairs Medical Center of Boston (36C24122N0769), and the European Union's Horizon 2020 Research and Innovation Programme (810115 – DYNASNET). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: L.Z., J.V., and X.D. Experiment: L.Z., M.L., Z.C., and B.J. Data analysis: X.H., X.M., C.X., Y.Z., J.E., R.C., and X.D. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

L.Z., M.L., A.Y.M., J.V., and X.D. are co-founders and shareholders of SingulOmics Corp. J. V. and A.Y.M. are co-founders and shareholders of MutaGenTech, Inc. A.-L.B is co-scientific founder of and is supported by Scipher Medicine, Inc.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.1016/j.xinn.2025.101008.

REFERENCES

1. Failla, G. (1958). The aging process and cancerogenesis. *Ann. N. Y. Acad. Sci.* **71**:1124–1140. DOI:https://doi.org/10.1111/j.1749-6632.1958.tb46828.x.
2. Szilard, L. (1959). ON THE NATURE OF THE AGING PROCESS. *Proc. Natl. Acad. Sci. USA* **45**:30–45. DOI:https://doi.org/10.1073/pnas.45.1.30.
3. Vijg, J. and Dong, X. (2020). Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* **182**:12–23. DOI:https://doi.org/10.1016/j.cell.2020.06.024.
4. Bohrson, C.L., Barton, A.R., Lodato, M.A. et al. (2019). Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat. Genet.* **51**:749–754. DOI:https://doi.org/10.1038/s41588-019-0366-2.
5. Dong, X., Zhang, L., Milholland, B. et al. (2017). Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**:491–493. DOI:https://doi.org/10.1038/nmeth.4227.
6. Zhang, L., Dong, X., Lee, M. et al. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci. USA* **116**:9014–9019. DOI:https://doi.org/10.1073/pnas.1902510116.
7. Brazhnik, K., Sun, S., Alani, O. et al. (2020). Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. *Sci. Adv.* **6**:eaax2659. DOI:https://doi.org/10.1126/sciadv.aax2659.
8. Huang, Z., Sun, S., Lee, M. et al. (2022). Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nat. Genet.* **54**:492–498. DOI:https://doi.org/10.1038/s41588-022-01035-w.
9. Lodato, M.A., Rodin, R.E., Bohrson, C.L. et al. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**:555–559. DOI:https://doi.org/10.1126/science.aao4426.
10. Miller, M.B., Huang, A.Y., Kim, J. et al. (2022). Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**:714–722. DOI:https://doi.org/10.1038/s41586-022-04640-1.
11. Choudhury, S., Huang, A.Y., Kim, J. et al. (2022). Somatic mutations in single human cardiomyocytes reveal age-associated DNA damage and widespread oxidative genotoxicity. *Nat. Aging* **2**:714–725. DOI:https://doi.org/10.1038/s43587-022-00261-5.
12. Li, G.M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**:85–98. DOI:https://doi.org/10.1038/cr.2007.115.
13. de Wind, N., Dekker, M., Berns, A. et al. (1995). Inactivation of the mouse Msh2 gene results in mismatch repair deficiency, methylation tolerance, hyperrecombination, and predisposition to cancer. *Cell* **82**:321–330. DOI:https://doi.org/10.1016/0092-8674(95)90319-4.
14. Hegan, D.C., Narayanan, L., Jirik, F.R. et al. (2006). Differing patterns of genetic instability in mice deficient in the mismatch repair genes Pms2, Mlh1, Msh2, Msh3 and Msh6. *Carcinogenesis* **27**:2402–2408. DOI:https://doi.org/10.1093/carcin/bgl079.
15. Lin, D.P., Wang, Y., Scherer, S.J. et al. (2004). An Msh2 point mutation uncouples DNA mismatch repair and apoptosis. *Cancer Res.* **64**:517–522. DOI:https://doi.org/10.1158/0008-5472.can-03-2957.
16. Lu, J.Y., Simon, M., Zhao, Y. et al. (2022). Comparative transcriptomics reveals circadian and pluripotency networks as two pillars of longevity regulation. *Cell Metab.* **34**:836–856.e5. DOI:https://doi.org/10.1016/j.cmet.2022.04.011.
17. Smits, R., Hofland, N., Edelmann, W. et al. (2000). Somatic Apc mutations are selected upon their capacity to inactivate the beta-catenin downregulating activity. *Genes Chromosomes Cancer* **29**:229–239.
18. Seluanov, A., Vaidya, A. and Gorbunova, V. (2010). Establishing primary adult fibroblast cultures from rodents. *J. Vis. Exp.* **44**:2033. DOI:https://doi.org/10.3791/2033.
19. Milholland, B., Dong, X., Zhang, L. et al. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**:15183. DOI:https://doi.org/10.1038/ncomms15183.
20. Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. DOI:https://doi.org/10.1093/bioinformatics/btp324.
21. Li, H., Handsaker, B., Wysoker, A. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:2078–2079. DOI:https://doi.org/10.1093/bioinformatics/btp352.
22. McKenna, A., Hanna, M., Banks, E. et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**:1297–1303. DOI:https://doi.org/10.1101/gr.107524.110.
23. Zhang, L., Lee, M., Maslov, A.Y. et al. (2024). Analyzing somatic mutations by single-cell whole-genome sequencing. *Nat. Protoc.* **19**:487–516. DOI:https://doi.org/10.1038/s41596-023-00914-8.
24. Zhang, L., Dong, X., Tian, X. et al. (2021). Maintenance of genome sequence integrity in long- and short-lived rodent species. *Sci. Adv.* **7**:eabj3284. DOI:https://doi.org/10.1126/sciadv.abj3284.
25. Dobin, A., Davis, C.A., Schlesinger, F. et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. DOI:https://doi.org/10.1093/bioinformatics/bts635.
26. Li, B. and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.* **12**:323. DOI:https://doi.org/10.1186/1471-2105-12-323.
27. Muskovic, W. and Powell, J.E. (2021). DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biol.* **22**:329. DOI:https://doi.org/10.1186/s13059-021-02547-0.
28. Wolock, S.L., Lopez, R. and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**:281–291.e9. DOI:https://doi.org/10.1016/j.cels.2018.11.005.
29. Hao, Y., Stuart, T., Kowalski, M.H. et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**:293–304. DOI:https://doi.org/10.1038/s41587-023-01767-y.
30. Ren, P., Dong, X. and Vijg, J. (2022). Age-related somatic mutation burden in human tissues. *Front. Aging* **3**:1018119. DOI:https://doi.org/10.3389/fragi.2022.1018119.
31. Bergstrom, E.N., Barnes, M., Martincorena, I. et al. (2020). Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinf.* **21**:438. DOI:https://doi.org/10.1186/s12859-020-03772-3.
32. Siepel, A., Bejerano, G., Pedersen, J.S. et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**:1034–1050. DOI:https://doi.org/10.1101/gr.3715005.
33. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. et al. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**:110–121. DOI:https://doi.org/10.1101/gr.097857.109.
34. Lee, B.T., Barber, G.P., Benet-Pagès, A. et al. (2022). The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res.* **50**:D1115–D1122. DOI:https://doi.org/10.1093/nar/gkab959.
35. Wang, K., Li, M. and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**:e164. DOI:https://doi.org/10.1093/nar/gkq603.
36. Yang, H. and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* **10**:1556–1566. DOI:https://doi.org/10.1038/nprot.2015.105.
37. Vaser, R., Adusumalli, S., Leng, S.N. et al. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* **11**:1–9. DOI:https://doi.org/10.1038/nprot.2015.123.
38. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C. et al. (2013). Signatures of mutational processes in human cancer. *Nature* **500**:415–421. DOI:https://doi.org/10.1038/nature12477.
39. Alexandrov, L.B., Kim, J., Haradhvala, N.J. et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* **578**:94–101. DOI:https://doi.org/10.1038/s41586-020-1943-3.
40. Blokzijl, F., Janssen, R., van Boxtel, R. et al. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**:33. DOI:https://doi.org/10.1186/s13073-018-0539-0.
41. Gel, B. and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**:3088–3090. DOI:https://doi.org/10.1093/bioinformatics/btx346.
42. Sondka, Z., Bamford, S., Cole, C.G. et al. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**:696–705. DOI:https://doi.org/10.1038/s41568-018-0060-1.
43. Kim, J., Mouw, K.W., Polak, P. et al. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**:600–606. DOI:https://doi.org/10.1038/ng.3557.
44. Martincorena, I., Raine, K.M., Gerstung, M. et al. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**:1029–1041.e21. DOI:https://doi.org/10.1016/j.cell.2017.09.042.
45. Mustjoki, S. and Young, N.S. (2021). Somatic Mutations in "Benign" Disease. *N. Engl. J. Med.* **384**:2039–2052. DOI:https://doi.org/10.1056/NEJMra2101920.
46. Kirkwood, T.B. (1977). Evolution of ageing. *Nature* **270**:301–304. DOI:https://doi.org/10.1038/270301a0.

47. Hart, R.W. and Setlow, R.B. (1974). Correlation between deoxyribonucleic acid excision-repair and life-span in a number of mammalian species. *Proc. Natl. Acad. Sci. USA* **71**:2169–2173.
48. Cagan, A., Baez-Ortega, A., Brzozowska, N. et al. (2022). Somatic mutation rates scale with lifespan across mammals. *Nature* **604**:517–524. DOI:https://doi.org/10.1038/s41586-022-04618-z.
49. Erickson, R.P. (2010). Somatic gene mutation and human disease other than cancer: an update. *Mutat. Res.* **705**:96–106. DOI:https://doi.org/10.1016/j.mrrev.2010.04.002.
50. Zapata, L., Pich, O., Serrano, L. et al. (2018). Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.* **19**:67. DOI:https://doi.org/10.1186/s13059-018-1434-0.
51. Bányai, L., Trexler, M., Kerekes, K. et al. (2021). Use of signals of positive and negative selection to distinguish cancer genes and passenger genes. *eLife* **10**:e59629. DOI:https://doi.org/10.7554/eLife.59629.
52. Jaiswal, S. and Ebert, B.L. (2019). Clonal hematopoiesis in human aging and disease. *Science* **366**:eaan4673. DOI:https://doi.org/10.1126/science.aan4673.
53. Yizhak, K., Aguet, F., Kim, J. et al. (2019). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**:eaaw0726. DOI:https://doi.org/10.1126/science.aaw0726.
54. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C. et al. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**:246–259. DOI:https://doi.org/10.1016/j.celrep.2012.12.008.
55. Robinson, P.S., Coorens, T.H.H., Palles, C. et al. (2021). Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* **53**:1434–1442. DOI:https://doi.org/10.1038/s41588-021-00930-y.
56. Franco, I., Revêchon, G. and Eriksson, M. (2022). Challenges of proving a causal role of somatic mutations in the aging process. *Aging Cell* **21**:e13613. DOI:https://doi.org/10.1111/acel.13613.
57. Heid, J., Cutler, R., Sun, S. et al. (2024). Negative selection allows human primary fibroblasts to tolerate high somatic mutation loads induced by N-ethyl-N-nitrosourea. Preprint at *bioRxiv*. DOI:https://doi.org/10.1101/2024.04.07.588286.