Evidence for negative selection against somatic mutations induced in normal fibroblasts by *N*-ethyl-*N*-nitrosourea

Ronald Cutler, Johanna Heid, Shixiang Sun, et al.

Genome Res. published online May 15, 2026
Access the most recent version at doi:10.1101/gr.281376.125

Title

Evidence for negative selection against somatic mutations induced in normal fibroblasts by *N*-ethyl-*N*-nitrosourea

Short Title

Negative selection of mutations in normal cells

Authors

Ronald Cutler[1,2,=,*], Johanna Heid[1,=], Shixiang Sun[1], Moonsook Lee[1], Alexander Y. Maslov[1,3], Lei Zhang[4], Simone Sidoli[2], Xiao Dong[4, 5], Jan Vijg[1,*]

Affiliations

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

[2] Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[3] Laboratory of Applied Genomic Technologies, Voronezh State University of Engineering Technology, Voronezh 394000, Russia

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

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

[=] Authors contributed equally to this work

[*] Correspondence to J.V. (jan.vijg@einsteinmed.edu) & R.C. (ronald.cutler@einsteinmed.edu)

Abstract

Mutations accumulate with age in most human tissues. While some undergo clonal expansion and contribute to disease, the mutational burden tolerated by a normal cell without functional decline remains unknown. Here, we repeatedly treat proliferating human primary fibroblasts with the point mutagen *N*-ethyl-*N*-nitrosourea, and analyze mutation burden by single-cell whole-genome sequencing. Mutation burden increases linearly to ~56,000 single-nucleotide variants per cell, with only a modest reduction in growth rate. We detect negative selection against potentially deleterious coding and non-coding variants, including mutations affecting pathways important for cell growth and maintenance. These findings suggest that selective depletion of harmful variants helps proliferating cells maintain function despite an extreme mutation burden. Because most adult tissues are largely non-dividing and cannot remove damaging mutations through a growth disadvantage, somatic mutations that accumulate during aging may have pronounced functional consequences *in vivo*.

Introduction

Aging is a complex process of continuous decline in cellular and tissue function that is accompanied with increased disease risk (López-Otín et al., 2023). DNA damage is thought to play a central role in this process and may be a universal cause of aging (Schumacher et al., 2021). Among the major molecular consequences of DNA damage during aging are somatic DNA mutations. Somatic mutations are the result of errors during repair or replication of a damaged DNA template and can only be purged via the death of the cell or organism. The accumulation of somatic mutations has since long been implicated as a major cause of aging (Failla, 1958; Szilard, 1959), which is strongly supported by the age-related exponential increase of cancer, a disease known to be caused by mutations (Laconi et al., 2020). Recent single-cell and single-molecule methods now allow accurate quantification of small somatic mutations in normal (non-cancerous) tissues. Application of such methods has provided ample evidence that somatic mutation rate is far higher than the germline mutation rate and that thousands of somatic mutations accumulate with age in most, if not all, human tissues (Ren et al., 2022), including postmitotic tissues such as the brain and heart (Choudhury et al., 2022; Lodato et al., 2018). In addition, evidence has emerged that somatic mutations are a cause of a wide variety of human diseases other than cancer (Erickson, 2010; Li & Williams, 2013).

What thus far remains unknown is the possible functional impact, collectively, of thousands of random mutations accumulating in a non-cancerous cell, for example, by adversely affecting gene-coding or gene-regulatory sequences relevant for the given cell type (Vijg & Dong, 2020). If so, one would expect an upper limit to the burden of random mutations a cell can tolerate or selection against mutations in crucial functional regions. *N*-ethyl-*N*-nitrosourea (ENU) is well suited to test this hypothesis because it is a highly efficient mutagen that induces mutations in a dose-dependent

manner while generating predominantly single-nucleotide variants (SNVs) rather than large-scale alterations. Its mutational spectrum reflects the formation of ethyl adducts such as O6-ethylguanine and O2/O4-ethylthymine and is typically enriched for C:G>T:A, T:A>C:G, and T:A>A:T substitutions (Bronstein et al., 1992; Chen et al., 2000; Yang et al., 1994). Here, we treated actively proliferating human primary fibroblasts with low, repeated doses of ENU to test if and how normal cells can tolerate high levels of somatic mutations.

Results

Repeated Mutagen Treatment of Normal Cells Only Slightly Affects Cell Growth and Death

To determine the impact of increased somatic mutation burden on primary human cells *in vitro*, we treated low-passage fetal lung fibroblasts (IMR-90) multiple times with a sublethal dose of 50 µg/mL of ENU. Each treatment was followed by a recovery period of 7 days, after which 1 million cells (determined by cell counting) were replated and treated again. This process was repeated for 9 cycles (Figure 1A; Methods). We measured cell growth by calculating population doublings (PDs), and found that repeated ENU treatment gradually reduced cell growth over the course of the experiment, such that by cycle 9 control cells had undergone a total of 28.25 ± 1.87 (mean ± standard deviation) PDs compared to 24.1 ± 1.34 PDs for ENU-treated cells (Figure 1B). This effect was more pronounced at the 3-day time point (Figure 1C) than at the 7-day time point (Figure 1D), indicating that ENU causes an early slowing of proliferation that partially recovers over time, consistent with a transient DNA damage response (Dogliotti et al., 1987). To assess whether increased cell death also contributed to the overall reduction in growth, we measured

markers of early and late apoptosis at both time points. ENU-treated cells showed a slight but consistent increase in both early and late apoptosis relative to controls at 3 and 7 days after treatment (Figures 1E-F and 1G-H). Together, these results indicate that the modest decline in cell growth after repeated ENU exposure reflects both an immediate slowing of proliferation and a concurrent increase in apoptosis. To analyze possible effects of a DNA damage response induced by ENU treatment in more detail, we performed bulk mRNA-sequencing analysis on samples collected at cycles 1 and 9 for the control group and cycle 9 for the ENU-treated group (Figures S1A-K; Table S1). As a positive control for the effect of cell passaging, we first compared the control cycle 9 and the control cycle 1 groups, which resulted in 979 significantly differentially expressed genes (DEGs) (Figures 2A; Table S2). We then performed targeted gene set enrichment analysis (GSEA) which showed that cell cycle-related genes were significantly decreased, while apoptosis-related genes were not significantly changed (Figures 2B-C). As cells are known to become senescent during passaging, we also tested for an enrichment of senescence-related genes using the SenMayo gene set (Saul et al., 2022), which were found to be significantly increased (Figure 2D). Furthermore, we found the DEGs to be positively correlated with DEGs found in a previous study on replicative senescence (Figure S1N) (Lackner et al., 2014).

To assess the effects of the repeated ENU treatment, we compared the ENU-treated cycle 9 and the control cycle 9 groups. Unexpectedly, this resulted in only 2 significant DEGs (Figure 2E; Table S2; Supplemental Note 1). However, after performing GSEA (which does not rely on *p*-value thresholds), we found that repeated ENU treatment had a similar, albeit weaker, effect as cell passaging – a significant decrease of cell cycle-related genes and a significant increase in senescence-related genes (Figure 2F, 2H). In contrast to the effect of cell passaging, we found a significant increase in the abundance of apoptosis-related genes in the ENU-treated cells relative

to the controls (Figure 2G), confirming the results of the direct apoptosis measurements (Figures 1G-H).

Taken together, these phenotypic and molecular data suggest that the observed slight decrease in cell growth rate after repeated ENU treatment is, at least in part, due to DNA damage-induced cell cycle inhibition and increased apoptosis. As the resulting effect sizes due to repeated ENU treatment were relatively small, these results demonstrate the tolerance of normal primary cells proliferating in culture to this mutagen, which is in agreement with previous literature (Barbaric et al., 2007).

Repeated Mutagen Treatment of Normal Cells Results in Extremely High Mutation Burden

Somatic DNA mutations in normal cells, which are largely unique to each cell, cannot be distinguished from sequencing artifacts after bulk whole genome sequencing. Hence, we used our established and highly accurate single-cell whole genome sequencing (SCWGS) assay, which has been extensively validated by parallel analysis of kindred single-cell clones (Dong et al., 2017; L. Zhang et al., 2024) (Supplemental Note 2), to analyze mutation burden in 21 single-cells from control and ENU-treated groups collected in 3 independent batches at cycles 3, 6, and 9 from the repeated ENU treatment experiment (Figure 1A, Table S3, Supplemental Methods). Cells for mutation analysis were collected at the 7-day time point after each treatment cycle, as DNA damage responses are abated by that time and all mutations fixed (Dogliotti et al., 1987). We corrected for the less than 100% genome coverage and sensitivity, which provided estimated mutation burdens higher than the observed burdens (Figures S2A-G; Supplemental Methods). We

also corrected for technical artifacts in the control group for cycle 1, which reduced mutation burden somewhat but had no effect at all on the ENU-induced mutation burden (Figures S3A-D; Supplemental Methods). We found that the number of estimated SNVs per cell in the control group was initially 1,996 ± 332 at cycle 1, which then increased to 19,064 ± 823 following 3 ENU treatment cycles, (~7.8-fold increase compared to cycle 1, p-value = 1.54 x 10^{-6}), and then 33,098 ± 1,785 after 6 ENU treatment cycles (~1.7-fold increase compared to cycle 3, p-value = 1.90 x 10^{-5}), and finally reached 55,954 ± 4,066 after the 9th ENU treatment cycle (~1.7-fold increase compared to the cycle 6, p-value = 3.33 x 10^{-9}) (Figures 3A and S2H). The most frequent types of mutations were thymine-to-adenine (T>A) transversions, followed by thymine-to-cytosine (T>C) transitions, both known to be induced by ENU (Barbaric et al., 2007) (Figures S2K-L). Importantly, there was no significant difference between the mutation burden of control cells at cycle 1 and at cycle 9, ruling out that the increase in SNVs in ENU-treated cells was due to passaging (Supplemental Note 3). We also checked for the possibility of clonal expansion by looking at shared mutations between cells within each group, but only detected 8 SNVs that were shared between 2 cells in the ENU-treated group at cycle 6 (Figure S2F), indicating a shared common ancestor, but not clonal expansion. This was expected as extensive clonal expansion cannot occur over the relatively short number of cell divisions. We conclude that our treatment paradigm increased mutational load to extremely high levels, with the increase in SNVs occurring linearly with the treatment cycles (Figure 3B).

While ENU is known to mainly induce base substitution mutations through alkylation damage, a slight increase in small insertions and deletions (INDELs) was also observed after ENU treatment, starting from 123 ± 49 in the control group at cycle 1, to 246 ± 49 in the ENU-treated group at cycle 3 (~1.8-fold increase, p-value = 4.10 x 10^{-3}); however, there was no further increase at cycles

6 and 9 (Figures S2I-J). INDELs have been previously observed to be induced by ENU, albeit at low frequencies (Watson et al., 1998), and could be caused by DNA polymerase slippage during DNA replication at sites of alkylation damage. As with SNVs, there was no significant difference in the amount of estimated INDELs between control cells from cycle 1 and cycle 9. While the lack of an increase in INDELs after cycle 3 in the ENU-treated group may be suggestive of negative selection acting to prevent a further increase, as was seen in a recent study (Zhang et al., 2025), this finding remains inconclusive due to the low amount of observed INDELs per cell (~20 INDELs per cell).

We then characterized the underlying SNV mutational processes in the aggregated mutations from the cells at all cycles within the control and ENU-treated groups. The ENU-treated group showed an expected increase in the relative contribution of T>A transversions, followed by T>C transitions (Figures 4A, S2K, S3A). This is in keeping with previously published ENU-induced mutational spectra by us and others (Barbaric et al., 2007; Maslov et al., 2022). We then extracted SNV mutational signatures *de novo* using non-negative matrix factorization with a reference set of mutations from previous SCWGS studies (Alexandrov et al., 2020; Dong et al., 2017; Huang et al., 2022; Miller et al., 2022) (Figures S3E-F). Two signatures were identified, labeled as SBSA and SBSB, which were dominant in the cells of the ENU-treated and control groups, respectively (Figures 4B and S3G). Importantly, we found that SBSA resembled the SIGNAL human signature associated with ENU-treatment, validating the source of the induced mutations (Figure 4C) (Kucab et al., 2019). We then compared our *de novo* signatures with those in the COSMIC database and found that SBSB was most similar to SBS5, a ‘clock-like’ mutational signature that has been previously found to correlate with cellular replication during aging (Alexandrov et al., 2020;

Hwang et al., 2025) (Figure 4D). Though relatively few INDELs were observed (369 in total), we also provide a characterization of their mutational spectra (Figure S3H; Supplemental Note 4).

Negative Selection Against ENU-Induced Accumulation of Damaging Coding and Non-Coding Variants

Based on the results above, we hypothesized that the sustained cell survival and growth in the face of extremely high mutation burden could be explained by negative selection against damaging mutations. Our dataset presented a unique opportunity to test this hypothesis as there were a total of 108,600 SNVs observed collectively in all cells. Note that this was smaller than the *estimated* number after corrections for sensitivity and genome coverage as mentioned above, but still expected to provide sufficient power for statistical testing of selection pressure (Figure S2H).

To obtain a random background distribution of expected mutations against which to compare the observed mutations in testing for selection pressure, we used the SigProfilerSimulator tool (Bergstrom et al., 2020). This method controls for variation in genomic coverage, mutational signatures, and sequence context of the observed mutations. With this, we simulated 10,000 instances of random mutations for each treatment cycle in both the control and ENU-treated cells (Methods). Following this, both observed and expected SNVs in coding regions were annotated as synonymous (non-protein-altering) or nonsynonymous (protein-altering) variants, using the Ensembl Variant Effect Predictor (VEP) program (Table S4; Supplemental Methods) (McLaren et al., 2016). The simulations of expected mutations performed well in predicting the number of synonymous variants ($R = 0.94$, p-value $= 3.6 \times 10^{-10}$; Figure S4A), which should be under minimal selection. We then measured selection pressure within each group by calculating the

Log_2(observed/expected) (written as Log_2(O/E) hereafter) ratios for each variant type using the frequency of observed and expected variants (Lek et al., 2016). The direction of selection pressure (i.e. neutral, positive, or negative) was ascertained by comparing the resulting Log_2(O/E) ratio to 0, where no difference from 0 is what would be expected under neutral selection (Greenman et al., 2006) (Methods). We then used bootstrapping to determine the threshold for the frequency of observed variants needed to obtain sufficient power for statistically significant results (p-value < 0.05) at various Log_2(O/E) ratios (Figures S4B; Table S5; Methods). Finally, to account for the known effect of DNA repair in depleting the frequency of mutations within transcribed regions through transcription-coupled repair (TCR) and mismatch repair (MMR) (Frigola et al., 2017; Heilbrun et al., 2021), we normalized the Log_2(O/E) ratios of variants within transcribed regions by the Log_2(O/E) ratio of mutations within intronic regions (assumed to be equally transcribed, but under neutral selection).

With this approach, we found that while there were significant cycle-dependent increases in both synonymous and nonsynonymous variant frequencies in the ENU-treated group relative to the controls (Figure 5A), the Log_2(O/E) ratios of nonsynonymous variants at cycles 6 and 9 in the ENU-treated group indicated that the frequency of these damaging variants was significantly lower than what would be expected, consistent with negative selection (Figures 5B and S4C-E, right panels). The negative Log_2(O/E) ratios of nonsynonymous variants at cycles 6 and 9 corresponded to a ~20% reduction in mutation burden, equivalent to a decrease of 14.5 ± 9.1 and 19.9 ± 12.5 variants per cell at cycles 6 and 9, respectively (Figure 5C). In regard to synonymous variants, the Log_2(O/E) ratios of these variants trended towards 0 across all cycles in the ENU-treated group, which would suggest neutral selection (as would be expected for these benign variants). Of note, statistically meaningful results could only be obtained with a sufficiently high number of

mutations, which were only present at the two highest treatment cycles. Indeed, for synonymous mutations, which are generally lower in number than the nonsynonymous events, the frequency even in the highest treatment cycle was slightly below the threshold set by the previously described power analysis, which prevented us from making any definite conclusions (Figures 5B and S4C-E, left panels; Table S5). See Supplemental Note 5 for discussion on the relationship of negative selection strength and mutation burden.

We then sought to verify our finding of negative selection against nonsynonymous variants by using a method which calculates *dN/dS* ratios (i.e. the *dNdScv* program), an alternative measure of selection commonly used in cancer or evolutionary studies (Martincorena et al., 2017). This metric compares the observed number of nonsynonymous mutations (i.e. missense, nonsense, and splice-site) to the number expected under neutral evolution, using synonymous mutations as a neutral reference. As this is effectively the same as calculating the O/E of nonsynonymous mutations over the O/E of synonymous mutations, we calculated the Log_2(O/E) nonsynonymous/synonymous (N/S) to facilitate the comparison. Synonymous variants, assumed to be functionally neutral, occur in the same sequence contexts and at the same depth as nonsynonymous variants, so they serve as an internal baseline for the true mutation rate introduced by ENU. A deficit of nonsynonymous variants relative to this neutral baseline is therefore attributable to negative selection rather than to sequencing bias or differences in mutagenesis. Log_2(O/E) (N/S) ratios were significantly below 0 at cycles 6 and 9 in the ENU-treated group (Figures 5D, S4F-G), indicating negative selection and corroborating our previous result (Figure 5B). By contrast, using the *dNdScv* method, no evidence for negative selection could be found (Figure S5A). The discrepancy is most likely due to important differences between this method and our O/E method. Unlike our O/E method, which empirically estimates expected mutations via genome-wide simulations accounting for coverage

and mutational signatures (Lek et al., 2016), the *dNdScv* method relies on probabilistic modeling to estimate the expected mutations, assumes uniform genomic coverage, and expects abundant mutations in coding regions. Therefore, the assumptions of this method make it less suitable for our SCWGS data type as it is limited by variable genome coverage and a limited number of coding mutations (Figures S2A and 5A). See Supplemental Note 6 for a detailed comparison between the methods.

To identify the contribution of the various types of nonsynonymous variants to the negative selection signal observed in the ENU-treated group, we stratified the nonsynonymous variants into stop retained, missense, start lost, stop gained, and stop lost variants and calculated their respective Log_2(O/E) ratios. This revealed that missense variants made up the majority of nonsynonymous variants in the ENU-treated group (Figure 5E) and also that the negative selection signal was driven almost exclusively by these variants (Figures 5F and S4H-J). See Supplemental Note 7 and Table S6 for details on stop gained variants.

Next, we tested if variants in non-coding regions were under selection pressure. For this, we again used the VEP program to annotate SNVs in non-coding regions, i.e., introns, intergenic, 5' untranslated regions (UTR), 3' UTR, and splice sites, and then quantified their frequencies (Table S4, Supplemental Methods) (McLaren et al., 2016). As with the coding variants, non-coding variants showed cycle-dependent increases in the ENU-treated group relative to the controls, with variants in intronic and intergenic regions accounting for 49.9% and 35.1% of all non-coding variants, respectively (Figure 5G). For variants in intronic regions in the ENU-treated groups at all cycles, Log_2(O/E) ratios were approximately -0.08, significantly less than what would be expected by chance (Figure 5H, Table S5). The vast majority of introns (with the exception of splice sites and other non-coding regulatory elements) do not contain functional sequences and are evolving

at neutrality in the germline. However, they are subject to TCR and MMR (Frigola et al., 2017; Heilbrun et al., 2021). Hence, we interpret the negative Log_2(O/E) ratios in these regions as a signal of TCR rather than negative selection. In intergenic regions for ENU-treated groups at all cycles, we found that the Log_2(O/E) was approximately 0.13, significantly greater than what would be expected by chance and indicative of mutation enrichment (Figure 5H, Table S5). Similar to introns, the vast majority of intergenic regions lack functional sequences and are evolving at neutrality in the germline. However, DNA repair activity in these regions is known to be substantially reduced, which likely explains the positive Log_2(O/E) ratios (Dukler et al., 2022; Yurchenko et al., 2023) (Supplemental Note 8).

Of all the non-coding variants, only Log_2(O/E) ratios of variants occurring in 3' UTR and splice regions in the ENU-treated group at cycle 9 were significantly less than 0 (Figures 5H, S4K-M, Table S5), which is consistent with negative selection and supported by previous studies on germline selection pressures (Denisov et al., 2014; Dukler et al., 2022; Findlay et al., 2024). The Log_2(O/E) ratios of 3' UTRs and splice sites in the ENU-treated group at cycle 9 corresponded to ~20-25% reduction in mutation burden, a decrease of 19.3 ± 12.5 and 6.2 ± 2.7 variants per cell, respectively (Figure 5I).

Taken together, these results show that potentially damaging variants in both coding and a subset of non-coding regions did not accumulate to levels that would be expected, indicating negative selection. This explains, in part, how a population of cells exposed to repeated mutagen treatments was able to prevent a massive decline in growth rate, even as overall SNV burden continued to increase linearly throughout the experiment. Of note, the selection pressure analysis was limited to mutations that were *observed* (i.e. actually physically detected), which are ~20% of the ~56,000 estimated mutations at cycle 9 in the ENU-treated group (Figures 3A and S2H). Hence, these

results have relevance for the numbers of SNVs that are estimated to accumulate during normal aging, e.g., from ~3,000 in human B lymphocytes to ~ 9,000 in human hepatocytes (Ren et al., 2022).

Pathways Supporting Fibroblast Growth and Survival are Protected from Mutation Accumulation

To gain insight into selective pressures at broader levels associated with cellular function, we investigated if gene sets active in proliferating human fibroblasts were under negative selection. To do this, we analyzed the SNVs across annotated functional genomic regions relevant to human lung fibroblasts (i.e. exons, upstream promoters, and enhancers) and stratified these into various gene sets (Boix, 2023; Boix et al., 2021; Carlson M, 2015). We then calculated the frequency of SNVs within these regions as well as the Log_2(O/E) ratios in a similar manner to the previous analyses.

First, we analyzed nonexpressed and expressed genes as determined by our bulk mRNA-sequencing data (Figures S1A-B; Table S7). In keeping with the previous analyses of other genomic regions, we found cycle-dependent increases of SNV frequency in nonexpressed and expressed regions for the ENU-treated group relative to the controls (Figure 6A). Log_2(O/E) ratios of SNVs in expressed genes were significantly below 0 at all cycles in the ENU-treated group even after normalization with intronic regions, consistent with negative selection (Figures 6B and S6A-C, right panels). In contrast, the Log_2(O/E) ratios of SNVs in nonexpressed genes did not significantly deviate from the expectation in the ENU-treated group at cycle 9, consistent with neutral selection (Figures 6B and S6A-C, left panels). Next, we subdivided the expressed genes

into nonessential and essential gene sets based on previous *in vitro* screens and repeated the same analysis as above (Table S7) (Hart et al., 2015; Wang et al., 2015). This showed increased SNV frequency in the ENU-treated group relative to the controls (Figure 6C), along with a Log_2(O/E) ratio significantly below 0 at cycle 9 in the ENU-treated group, consistent with negative selection (Figures 6D and S6D-F, right panels). Of note, this is in agreement with global *dN/dS* ratios on truncating mutations across 17 genes essential from a recent large-scale study sampling the normal oral epithelium (Lawson et al., 2025).

As negative selection is expected to be generally weak due to the diploid nature of the human genome (Martincorena et al., 2017; I. E. Vorontsov et al., 2016), but stronger in haploinsufficient regions (i.e. regions where both alleles are required for function), we subdivided expressed genes into haplosufficient (pLI < 0.1) and haploinsufficient (pLI > 0.9) gene sets based on a large scale study of protein-coding genetic variation (Table S7) (Lek et al., 2016). As expected, analysis of these gene sets showed increased SNV frequency in the ENU-treated group relative to the controls (Figure 6E). Log_2(O/E) ratios were significantly below 0 in both haplosufficient and haploinsufficient gene sets for the ENU-treated cells at cycle 9. However, only the ENU-treated group at cycle 6 showed Log_2(O/E) ratios significantly below 0 for the haploinsufficient gene set and not the haplosufficient gene set, suggesting overall stronger negative selection of haploinsufficient genes, as would be expected (Figures 6F and S6G-I). In further support of this, when the Log_2(O/E) was compared between haplosufficient and haploinsufficient gene sets for ENU-treated cells at cycle 9, we found that the Log_2(O/E) ratio was considerably lower for the haploinsufficient genes (-0.37 ± 0.16) as compared to the haplosufficient genes (-0.28 ± 0.04). In summary, these results suggest that damaging SNVs are selectively purged from functional

genomic regions based on expression status, essentiality, and dosage sensitivity, extending our previous results on negative selection of damaging SNVs in coding and non-coding regions.

Moving to the pathway level, we identified 399 Gene Ontology (GO) biological processes as significantly enriched in the set of expressed genes from our bulk RNA-seq data (Figure S6J, Table S8, Methods). As expected, the mean frequency of SNVs within each pathway showed significant cycle-dependent increases of SNV frequency in the ENU-treated group relative to the controls (Figure 6G). At the global level, Log_2(O/E) ratios of the identified pathways in the ENU-treated group showed a cycle-dependent decrease, suggesting widespread negative selection at the pathway level (Figures 6H, S6K-L, Table S9). At cycle 9 in the ENU-treated group, 6 pathways showed Log_2(O/E) ratios significantly below 0 after multiple hypothesis correction, consistent with negative selection. The pathways under negative selection were characterized as those related to basic cellular processes supporting lung fibroblast cell growth *in vitro* and survival, i.e., 'response to oxidative stress' (GO:0006979), 'respiratory tube development' (GO:0030323), 'respiratory system development' (GO:0006979), 'regulation of response to DNA damage stimulus' (GO:2001020), 'proteasome-mediated ubiquitin-dependent protein catabolic process' (GO:0043161), and 'cellular response to chemical stress' (GO:0062197) (Figures 6I-K, Table S9). The 'respiratory system development' GO term included the genes *FGF10, FOXF1,* and *COL3A1*, which are known to be crucial for lung fibroblast identity and function (Bellusci et al., 1997; Melboucy-Belkhir et al., 2014; Zoppi et al., 2004). To ensure that the pathways we had identified to be under negative selection were not false positives, we repeated the pathway analysis but employed bootstrapping to randomly shuffle genes amongst pathways before calculating the Log_2(O/E) ratios (Figures S7A-C, Methods). These results had no significant overlap with the pathways we identified to be under significant negative selection pressure (Figures S7D-E).

In summary, these results suggest that pathways supporting cell growth and survival in proliferating cells *in vitro*, are protected from the accumulation of SNVs by negative selection. Importantly, these pathways are biologically significant, as they appear to be relevant to the conditions of *in vitro* cell culture and to the unique characteristics of lung fibroblasts, highlighting the functional importance of preventing somatic mutation accumulation in functional genomic regions. These analyses show how negative selection allows a population of dividing cells to survive extremely high mutation burden and underscore the impact of random somatic SNVs on essential cellular functions.

Discussion

Although somatic mutations accumulate with age in normal tissues, their collective functional impact outside clonal expansion remains unclear (Franco et al., 2022; Ren et al., 2022). Here, we tested that question directly, showing that repeated low-dose ENU treatment increased somatic mutation burden in cultured fibroblasts from ~2,000 to ~56,000 SNVs per cell, more than 30-fold above controls, with only modest effects on growth and death rates. One possible interpretation is that fibroblasts tolerate large numbers of point mutations, which would be supported by results from a recent study in which as many as ~30,000 SNVs were observed in intestinal crypts of patients with genetic defects in DNA mismatch repair (Robinson et al., 2021).

An alternative explanation for the high tolerance of actively proliferating cells to mutation accumulation is negative selection against mutations in genomic regions important for cell survival. Evidence for negative selection against somatic mutations has indeed been found in tumors (Bányai et al., 2021; Ilya E. Vorontsov et al., 2016; Zapata et al., 2018), although positive

selection of driver mutations is the dominant force in cancer (Martincorena et al., 2017). Evidence for negative selection of somatic mutations in normal cells has so far been lacking, largely because the smaller numbers of mutations restrict statistical power (Ren et al., 2022). Our untreated control cells were similarly underpowered and did not show clear evidence of negative selection (Figure S4D, Table S5), whereas ENU-treatment increased power sufficiently to gain meaningful results (Figure S2H). Indeed, based on the high numbers of SNVs found after repeated treatment with ENU we found evidence for negative selection acting on potentially deleterious coding and non-coding variants. Although both synonymous and nonsynonymous variants increased linearly with ENU treatment (Figure 5A), O/E analyses showed significant depletion of high-impact nonsynonymous variants, especially missense variants, relative to the null model (Figures 5B-C). This suggests that proliferating fibroblasts can selectively purge the most harmful coding mutations, consistent with the selective constraint seen in germline datasets (Gudkov et al., 2024; Karczewski et al., 2020; Lek et al., 2016; X. Zhang et al., 2024). Our findings are in agreement with those from an *Msh2*$^{-/-}$ mouse model, which showed negative selection against somatic stop lost variants (Zhang et al., 2025). Additionally, our data showed negative selection acting on variants in non-coding regions, such as 3' UTR and splice sites (Figures 5G-I), which is in agreement with recent literature that has also analyzed selective pressures in these regions (Findlay et al., 2024).

Our gene-set and pathway analyses further indicated negative selection at a broader functional level, i.e., in expressed, essential, and haploinsufficient genes, which is consistent with a recent population-scale study of somatic mutations in oral epithelium (Figures 6A-F) (Lawson et al., 2025). We also identified signals of negative selection in actively expressed pathways, including oxidative stress, respiratory system development, and response to DNA damage (Figures 6I-K),

all of which are relevant to our experimental conditions and cell type. These findings are consistent with the idea that random somatic mutations affecting pathways important for cellular survival are detrimental for lung fibroblasts proliferating *in vitro*.

Our work provides strong evidence that high somatic mutation burdens can affect normal cells *independent* of clonally amplified mutations. Although negative selection was detected only at extremely high mutation burdens, it should be noted that the number of SNVs directly observed was no more than ~12,000, well below the ~56,000 estimate based on genome coverage and sensitivity (Figures 3A and S2H). This is only moderately above the ~2,000-5,000 SNVs per cell estimated at old age, depending on cell type (Ren et al., 2022). We therefore interpret the extreme mutation burden in our system as providing the statistical power needed to detect negative selection, rather than as evidence that this process operates only at extremely high mutation loads. Consistent with this, we observed no association between mutation burden and the strength of negative selection (Figure 5B), suggesting that negative selection may represent a general mechanism for eliminating somatic mutations that impair cell function even at physiological mutation burdens.

Another important consideration is how negative selection in our *in vitro* model compares with the *in vivo* context. *In vivo*, additional mechanisms such as immune surveillance may contribute to the elimination of damaged fibroblasts. More broadly, cell division rates are low in most adult tissues, with notable exceptions including the lymphoid and intestinal systems (Sender & Milo, 2021). In proliferating cells such as human B lymphocytes, we previously found that SNVs accumulate more slowly with age in functional genomic regions than genome-wide, consistent with negative selection during B-cell aging (Zhang et al., 2019). By contrast, somatic mutations in non-dividing cells such as neurons and cardiomyocytes cannot be removed through a growth disadvantage.

Nevertheless, these cells show robust accumulation of damaging mutations (Choudhury et al., 2022; Ganz et al., 2024). This suggests that age-related somatic mutation accumulation may have its greatest impact in non-proliferative cells, as proposed by the disposable soma theory (Kirkwood, 1977).

A key limitation of our study is that we mainly analyzed SNVs and the relatively small number of INDELs. These variant classes are generally less likely than structural variants (SVs) (e.g. large deletions, translocations, and copy number variants) to have major functional effects. Consistent with this, INDELs did not accumulate linearly during ENU treatment (Figures S2I-J), and in a recent study of *Msh2*$^{-/-}$ fibroblasts we observed an upper limit of ~16,000 INDELs per cell (Zhang et al., 2025). By contrast, somatic SVs remain poorly characterized because they are difficult to detect accurately with current single-cell and single-molecule methods (Hård et al., 2023). Although less frequent than SNVs, SVs can have major phenotypic consequences, as exemplified by germline SVs in human disease (Yang, 2020). SVs may therefore have a stronger impact on cell function than SNVs, but how they are selected against in non-dividing cells during aging remains unclear.

Future efforts should extend the framework presented in this study to the *in vivo* context, especially post-mitotic populations, and incorporate methods that capture SVs. Integrating mutational landscapes with functional readouts will be crucial for determining how the balance between information loss and selection shapes organismal aging and disease.

Methods

Additional methods can be found in the Supplemental Methods section in the Supplementary Materials.

Cell culture and treatment

At the beginning of each cycle, 1 x 10^6 human primary lung fibroblasts (IMR-90) cells were counted using a Countess 3 (Fisher Scientific), plated, and allowed to expand for 24 hours. Cells were then treated with 50 µg/mL *N*-ethyl-*N*-nitrosourea (ENU) and allowed to recover for 3 days. Cells were then collected and assayed for cell number and viability for the 3-day time point. Of the collected cells, 1 x 10^6 cells were re-plated and then allowed another 4 days to recover and expand until they were collected to assess cell number and viability for the 7-day time point. This procedure was repeated for 9 consecutive cycles over the course of ~10 weeks. This experiment was repeated 3 separate times to create 3 independent biological replicates. Cells were isolated at the 7-day time point at cycles 1, 3, 6, and 9 for somatic mutation analysis and at cycles 1 and 9 for transcriptome analysis.

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

Single-cells were isolated using the CellRaft AIR system (Cell Microsystems). The CellRaft array was prepared following the manufacturer's instructions and washed three times for 3 minutes with warm PBS before addition of cell suspension. 1,000 cells were seeded in 1 mL for one CellRaft

array. Cells were allowed to settle for 3-4 hours at 37°C before isolation. Then, the raft was carefully washed with warm medium to remove potential debris and dead cells. Individual rafts containing one cell were deposited in PCR tubes with 2.5 µL of PBS and stored at -80°C until further processing. Single-cell whole-genome amplification and library preparation for WGS was performed as previously described (L. Zhang et al., 2024). In brief, single-cell whole genomes were amplified using SCMDA and subject to quality control using a locus dropout test (Gundry et al., 2012). Libraries were prepared using a NEBNext Ultra II FS kit (NEB). Library quantity and size were assessed with a Qubit kit and TapeStation, respectively. Libraries were sequenced using the 150 paired-end mode on an Illumina NovoSeq platform by Novogene Corp Inc., CA.

Selection analysis of annotated variants

Selection analysis was performed by first simulating 10,000 instances of each single-cell in our dataset to be used as the null expectation (i.e. neutral selection) where mutations are randomly re-distributed throughout the genome. This was done using the SigProfilerSimulator program (Bergstrom et al., 2020), where for each single-cell, observed mutations were randomly re-distributed across the genome while accounting for cell-specific sequencing coverage, mutational signatures, and nucleotide contexts.

Log_2(observed/expected) ratios (written hereafter as Log_2(O/E)) were calculated for each single-cell or by group (mutations from all cells within a group are pooled), by first summing the frequency of annotated consequences for observed and expected mutations, respectively. The Log_2(O/E) ratio was then calculated with respect to each simulation instance (resulting in 10,000 Log_2(O/E) ratios per cell or group) using the following equation:

$$Log_2(O/E) = Log_2 \frac{(observed\ consequence\ freq. + 1)}{(simulated\ consequence\ freq. + 1)}$$

A permutation test was then used to determine the *p*-value of the observed consequence frequency where the simulated consequence frequency served as the null distribution and is as follows:

$$p = \frac{1}{n} \sum_{i=1}^{n} I\,(simulated\ consequence\ freq._i \geq observed\ consequence\ freq.)$$

Where *p* is the permutation test *p*-value, *n* is the number of permutations (i.e. 10,000), *i* is the *i*-th permutation, and *I* is the indicator function, which equals 1 if the condition inside is true, and 0 otherwise.

We performed a power analysis to determine how many observed variants are required to obtain reliable statistical results at varying Log_2(O/E) ratios. To do this, we used the mutation data from ENU-treated cells to obtain the relationship between the standard deviation in the number of expected variants given the number of observed variants for common types of variants that were identified. We fit a gamma GLM with a log link to model standard deviation as a function of log-transformed observed variant count. The model showed excellent fit (residual deviance = 3.19 on 144 df vs. null deviance = 147.37 on 145 df; dispersion = 0.020), with log-transformed observed variant count as a highly significant predictor irrespective of the variant type (*p-value* < 2 x 10^{-16}). With this, we then calculated Log_2(O/E) ratios by simulating 10,000 instances of expected variant counts while varying the observed variant counts from 1-1,000. We then varied the Log_2(O/E) ratios from 0-1 by multiplying the observed variant count by a scaling factor before calculating the Log_2(O/E) ratio. Finally, we calculated *p*-values for the Log_2(O/E) ratios using a permutation test, as described above.

The power analysis demonstrated how varying the observed variant count affects the *p*-value of the Log_2(O/E) ratio at different levels. Importantly, this analysis provided us with estimates of how many observed variants were required to achieve sufficient power (*p*-value<0.05) (i.e. to obtain robust statistical results) given a specific Log_2(O/E) (Figure S4B, solid line). This threshold was then applied to the Log_2(O/E) results, at both the levels of variant types and pathways, to determine which of these were amenable to further statistical testing where the results of this can be seen in the figure legends (e.g. Figure 5B; Table S5). For the Log_2(O/E) nonsynonymous/synonymous (N/S) calculation, we determined if these were amenable to further statistical testing if either the nonsynonymous or synonymous Log_2(O/E) results in each group reached the established observed variant threshold (Table S5).

When assessing for selection, it is also necessary to account for the cell-intrinsic bias in mutation frequency due to DNA repair mechanisms (i.e. transcription coupled repair), which is known to reduce mutation frequency in transcribed regions and is agnostic to damage that would result in synonymous or nonsynonymous variants (Selby et al., 2023). This bias in mutation frequency was indeed observed in our data and consistent with a depletion of mutations due to DNA repair (Figure 5H). Thus, we normalized the Log_2(O/E) ratio of coding variants (e.g. synonymous, missense, etc.) and non-coding variants (5'/3' UTR and splice region) by the Log_2(O/E) ratio of mutations which occurred within intronic regions and not in splice regions. We chose to use this normalization because the occurrence of both exonic and intronic mutations is dependent on transcription coupled repair (Heilbrun et al., 2021) however, since intronic mutations do not alter the translated protein (with the exception of splice site variants), they are not subject to selective mechanisms. In addition, the majority of mutations in our data occur within introns (~46.7%, Figure 5G), allowing us to make more confident Log_2(O/E) ratio for these mutations as compared to using the Log_2(O/E)

ratio of synonymous mutations which only made up only ~0.23% of all mutations. A comparison of unnormalized and normalized Log_2(O/E) ratio for nonsynonymous and synonymous variants is shown for example by comparing Figures 5B to S4C, where the effect of normalization can be seen to slightly increase the Log_2(O/E) ratio.

To estimate the number of variants that are protected against resulting from negative selection, we subtracted the average number of variants from simulated samples from the number of variants in the respective observed sample.

Selection analysis of gene sets and pathways

Nonexpressed and expressed genes were determined using bulk mRNA transcriptome data from cycle 9 control and ENU-treated groups using the zPKM program (Figure S2A; Table S7) (Hart et al., 2013). Nonessential and essential control gene sets were obtained from the DepMap database (depmap.org) (Broad, 2024), where the nonessential gene set was the negative controls from Hart *et al* (2015) and the essential gene set was the intersection of the essential gene sets resulting from the Hart *et al* (2015) and Blomen *et al* (2014) studies (Blomen et al., 2015; Hart et al., 2015). These gene sets were additionally filtered for expressed genes by overlapping them with the list of expressed genes we determined previously (Table S7). Haplosufficient and haploinsufficient gene sets were obtained from the ExAC database where haplosufficient genes were defined as pLI < 0.1 and haploinsufficient genes were defined as pLI > 0.9 (Lek et al., 2016). These gene sets were also filtered for expressed genes in the same manner as essential genes (Table S7). Pathway level gene sets were obtained by performing over representation analysis for Gene Ontology biological processes using the clusterProfiler program (Wu et al., 2021) on the list of expressed genes that

we determined previously (Table S8). For all gene sets, genomic coordinates of exons, introns, promoters (5KB upstream of transcription start site), and annotated enhancers (Boix et al., 2021) were retrieved using the GenomicRanges software (version 1.60.0) (Lawrence et al., 2013).

For each gene set, two Log_2(O/E) ratios were calculated. One for the version of the gene set where the intronic regions were included, and one for the version of the gene set intronic regions were excluded. To account for DNA repair (as described above for Log_2(O/E) calculation of coding and non-coding variants), the Log_2(O/E) ratios of each gene set were normalized by subtracting Log_2(O/E) ratio of the version that included introns from the version that excluded introns. *P*-values were estimated, as described above, and were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method.

To serve as a negative control for the pathway analysis, for each enriched biological process gene set, we randomly shuffled the genes between all other identified biological processes gene sets and created 10,000 randomly permuted gene sets. Average Log_2(O/E) ratios were then calculated for the 10,000 permutations of each gene set followed by *p*-value estimation, as described above.

Data Access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE271867. The WGS and RNA-seq data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number

PRJNA1129131. Scripts used for processing and analyzing data can be found in the Supplemental Material and have been deposited at:

figshare.com/articles/software/Negative_selection_allows_human_primary_fibroblasts_to_tolerate_high_somatic_mutation_loads_induced_by_N-ethyl-N-nitrosourea/26110453

Competing Interest Statement

M.L., A.Y.M., X.D., and J.V. are co-founders and shareholders of SingulOmics Corp. A.Y.M., and J.V. are co-founders of MutagenTech Inc. All other authors declare no conflict of interest.

Acknowledgements

We would like to acknowledge the members of the Vijg and Sidoli labs for helpful discussions during the preparation of the manuscript.

Funding

This study was supported by National Institutes of Health grants: T32GM007491 (R.C.) T32AG023475 (R.C.), U19AG056278 (J.V.), U01HL145560 (J.V.), U01ES029519 (J.V.), P01AG017242 (J.V.), P01AG047200 (J.V.), RF1AG068908 (J.V.), and P30AG038072 (J.V.). Other funding sources: US Department of Defense grant BC180689P1 (J.V.), Glenn Foundation for Medical Research Postdoctoral Fellowships in Aging Research (J.H.), The Michael J. Fox

Foundation (J.V.), and The Paul F. Glenn Center for the Biology of Human Aging at the Albert Einstein College of Medicine (J.V.).

Author Contributions

A.Y.M., and J.V. conceived this study and designed the experiments. J.H., M.S., and A.Y.M performed the experiments. R.C., J.H., and S.S. analyzed the data. X.D. and L.Z. provided guidance on the data analysis. R.C., J.H., S.S. and J.V. wrote the manuscript.

References

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Getz, G.,…Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94-101. https://doi.org/10.1038/s41586-020-1943-3

Bányai, L., Trexler, M., Kerekes, K., Csuka, O., & Patthy, L. (2021). Use of signals of positive and negative selection to distinguish cancer genes and passenger genes. *Elife*, *10*. https://doi.org/10.7554/eLife.59629

Barbaric, I., Wells, S., Russ, A., & Dear, T. N. (2007). Spectrum of ENU-induced mutations in phenotype-driven and gene-driven screens in the mouse. *Environ Mol Mutagen*, *48*(2), 124-142. https://doi.org/10.1002/em.20286

Bellusci, S., Grindley, J., Emoto, H., Itoh, N., & Hogan, B. L. (1997). Fibroblast growth factor 10 (FGF10) and branching morphogenesis in the embryonic mouse lung. *Development*, *124*(23), 4867-4878. https://doi.org/10.1242/dev.124.23.4867

Bergstrom, E. N., Barnes, M., Martincorena, I., & Alexandrov, L. B. (2020). Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics*, *21*(1), 438. https://doi.org/10.1186/s12859-020-03772-3

Blomen, V. A., Májek, P., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F. R., Olk, N., Stukalov, A., Marceau, C., Janssen, H., Carette, J. E., Bennett, K. L., Colinge, J., Superti-Furga, G., & Brummelkamp, T. R. (2015). Gene essentiality and synthetic lethality in haploid human cells. *Science*, *350*(6264), 1092-1096. https://doi.org/10.1126/science.aac7557

Boix, C. A. (2023). *EpiMap Repository*. https://compbio.mit.edu/epimap/

Boix, C. A., James, B. T., Park, Y. P., Meuleman, W., & Kellis, M. (2021). Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, *590*(7845), 300-307. https://doi.org/10.1038/s41586-020-03145-z

Broad, D. (2024). *DepMap 24Q4 Public* (Figshare+. https://doi.org/10.25452/figshare.plus.27993248.v1

Bronstein, S. M., Skopek, T. R., & Swenberg, J. A. (1992). Efficient Repair of O6-Ethylguanine, but not O4-Ethylthymine or O2-Ethylthymine, Is Dependent upon O6-Alkylguanine-DNA Alkyltransferase and Nucleotide Excision Repair Activities in Human Cells. *Cancer Research*, *52*(7), 2008-2011.

Carlson M, M. B. (2015). *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s)*. https://bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.UCSC.hg19.knownGene.html

Chen, Y., Yee, D., Dains, K., Chatterjee, A., Cavalcoli, J., Schneider, E., Om, J., Woychik, R. P., & Magnuson, T. (2000). Genotype-based screen for ENU-induced mutations in mouse embryonic stem cells. *Nature Genetics*, *24*(3), 314-317. https://doi.org/10.1038/73557

Choudhury, S., Huang, A. Y., Kim, J., Zhou, Z., Morillo, K., Maury, E. A., Tsai, J. W., Miller, M. B., Lodato, M. A., Araten, S., Hilal, N., Lee, E. A., Chen, M. H., & Walsh, C. A. (2022). Somatic mutations in single human cardiomyocytes reveal age-associated DNA damage and widespread oxidative genotoxicity. *Nature Aging*, *2*(8), 714-725. https://doi.org/10.1038/s43587-022-00261-5

Denisov, S. V., Bazykin, G. A., Sutormin, R., Favorov, A. V., Mironov, A. A., Gelfand, M. S., & Kondrashov, A. S. (2014). Weak negative and positive selection and the drift load at splice sites. *Genome Biol Evol*, *6*(6), 1437-1447. https://doi.org/10.1093/gbe/evu100

Dogliotti, E., Vitelli, A., Terlizzese, M., Muccio, A. D., Calcagnile, A., Saffiotti, U., & Bignami, M. (1987). Induction kinetics of mutations at two genetic loci, DNA damage and repair in CHO cells after different exposure times to N -ethyl- N -nitrosourea. *Carcinogenesis*, *8*(1), 25-31. https://doi.org/10.1093/carcin/8.1.25

Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A. Y., Wang, T., & Vijg, J. (2017). Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods*, *14*(5), 491-493. https://doi.org/10.1038/nmeth.4227

Dukler, N., Mughal, M. R., Ramani, R., Huang, Y. F., & Siepel, A. (2022). Extreme purifying selection against point mutations in the human genome. *Nat Commun*, *13*(1), 4312. https://doi.org/10.1038/s41467-022-31872-6

Erickson, R. P. (2010). Somatic gene mutation and human disease other than cancer: an update [Research Support, Non-U.S. Gov't

Review]. *Mutat Res*, *705*(2), 96-106. https://doi.org/10.1016/j.mrrev.2010.04.002

Failla, G. (1958). The aging process and cancerogenesis. *Ann N Y Acad Sci*, *71*(6), 1124-1140. https://doi.org/10.1111/j.1749-6632.1958.tb46828.x

Findlay, S. D., Romo, L., & Burge, C. B. (2024). Quantifying negative selection in human 3' UTRs uncovers constrained targets of RNA-binding proteins. *Nat Commun*, *15*(1), 85. https://doi.org/10.1038/s41467-023-44456-9

Franco, I., & Eriksson, M. (2022). Reverting to old theories of ageing with new evidence for the role of somatic mutations. *Nat Rev Genet*, *23*(11), 645-646. https://doi.org/10.1038/s41576-022-00513-5

Franco, I., Helgadottir, H. T., Moggio, A., Larsson, M., Vrtačnik, P., Johansson, A., Norgren, N., Lundin, P., Mas-Ponte, D., Nordström, J., Lundgren, T., Stenvinkel, P., Wennberg, L., Supek, F., & Eriksson, M. (2019). Whole genome DNA sequencing provides an atlas of

somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biol*, *20*(1), 285. https://doi.org/10.1186/s13059-019-1892-z

Franco, I., Johansson, A., Olsson, K., Vrtačnik, P., Lundin, P., Helgadottir, H. T., Larsson, M., Revêchon, G., Bosia, C., Pagnani, A., Provero, P., Gustafsson, T., Fischer, H., & Eriksson, M. (2018). Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. *Nat Commun*, *9*(1), 800. https://doi.org/10.1038/s41467-018-03244-6

Franco, I., Revêchon, G., & Eriksson, M. (2022). Challenges of proving a causal role of somatic mutations in the aging process. *Aging Cell*, *21*(5), e13613. https://doi.org/10.1111/acel.13613

Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., & López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. *Nat Genet*, *49*(12), 1684-1692. https://doi.org/10.1038/ng.3991

Ganz, J., Luquette, L. J., Bizzotto, S., Miller, M. B., Zhou, Z., Bohrson, C. L., Jin, H., Tran, A. V., Viswanadham, V. V., McDonough, G., Brown, K., Chahine, Y., Chhouk, B., Galor, A., Park, P. J., & Walsh, C. A. (2024). Contrasting somatic mutation patterns in aging human neurons and oligodendrocytes. *Cell*, *187*(8), 1955-1970.e1923. https://doi.org/10.1016/j.cell.2024.02.025

Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R., & Easton, D. F. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, *173*(4), 2187-2198. https://doi.org/10.1534/genetics.105.044677

Gudkov, M., Thibaut, L., & Giannoulatou, E. (2024). Context-adjusted proportion of singletons (CAPS): a novel metric for assessing negative selection in the human genome. *NAR Genom Bioinform*, *6*(3), lqae111. https://doi.org/10.1093/nargab/lqae111

Gundry, M., Li, W., Maqbool, S. B., & Vijg, J. (2012). Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic Acids Res*, *40*(5), 2032-2040. https://doi.org/10.1093/nar/gkr949

Hård, J., Mold, J. E., Eisfeldt, J., Tellgren-Roth, C., Häggqvist, S., Bunikis, I., Contreras-Lopez, O., Chin, C.-S., Nordlund, J., Rubin, C.-J., Feuk, L., Michaëlsson, J., & Ameur, A. (2023). Long-read whole-genome analysis of human single cells. *Nature Communications*, *14*(1), 5164. https://doi.org/10.1038/s41467-023-40898-3

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., & Moffat, J. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, *163*(6), 1515-1526. https://doi.org/10.1016/j.cell.2015.11.015

Hart, T., Komori, H. K., LaMere, S., Podshivalova, K., & Salomon, D. R. (2013). Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics*, *14*, 778. https://doi.org/10.1186/1471-2164-14-778

Heilbrun, E. E., Merav, M., & Adar, S. (2021). Exons and introns exhibit transcriptional strand asymmetry of dinucleotide distribution, damage formation and DNA repair. *NAR Genom Bioinform*, *3*(1), lqab020. https://doi.org/10.1093/nargab/lqab020

Huang, Z., Sun, S., Lee, M., Maslov, A. Y., Shi, M., Waldman, S., Marsh, A., Siddiqui, T., Dong, X., Peter, Y., Sadoughi, A., Shah, C., Ye, K., Spivack, S. D., & Vijg, J. (2022). Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nat Genet*, *54*(4), 492-498. https://doi.org/10.1038/s41588-022-01035-w

Hwang, T., Sitko, L. K., Khoirunnisa, R., Navarro-Aguad, F., Samuel, D. M., Park, H., Cheon, B., Mutsnaini, L., Lee, J., Otlu, B., Takeda, S., Lee, S., Ivanov, D., & Gartner, A. (2025). Comprehensive whole-genome sequencing reveals origins of mutational signatures associated with aging, mismatch repair deficiency and temozolomide chemotherapy. *Nucleic Acids Res*, *53*(1). https://doi.org/10.1093/nar/gkae1122

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A.,…Genome Aggregation Database, C. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443. https://doi.org/10.1038/s41586-020-2308-7

Kirkwood, T. B. L. (1977). Evolution of ageing. *Nature*, *270*(5635), 301-304. https://doi.org/10.1038/270301a0

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S. P., Arlt, V. M., Phillips, D. H., & Nik-Zainal, S. (2019). A Compendium of Mutational Signatures of Environmental Agents. *Cell*, *177*(4), 821-836 e816. https://doi.org/10.1016/j.cell.2019.03.001

Lackner, D. H., Hayashi, M. T., Cesare, A. J., & Karlseder, J. (2014). A genomics approach identifies senescence-specific gene expression regulation. *Aging Cell*, *13*(5), 946-950. https://doi.org/10.1111/acel.12234

Laconi, E., Marongiu, F., & DeGregori, J. (2020). Cancer as a disease of old age: changing mutational and microenvironmental landscapes. *Br J Cancer*, *122*(7), 943-952. https://doi.org/10.1038/s41416-019-0721-1

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Lawson, A. R. J., Abascal, F., Nicola, P. A., Lensing, S. V., Roberts, A. L., Kalantzis, G., Baez-Ortega, A., Brzozowska, N., El-Sayed Moustafa, J. S., Vaitkute, D., Jakupovic, B., Nessa, A., Wadge, S., Österdahl, M. F., Paterson, A. L., Rassl, D. M., Alcantara, R. E., O'Neill, L., Widaa, S.,…Martincorena, I. (2025). Somatic mutation and selection at population scale. *Nature*, *647*(8089), 411-420. https://doi.org/10.1038/s41586-025-09584-w

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J.,…MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285-291. https://doi.org/10.1038/nature19057

Li, C., & Williams, S. M. (2013). Human Somatic Variation: It's Not Just for Cancer Anymore. *Current Genetic Medicine Reports*, *1*(4), 212-218. https://doi.org/10.1007/s40142-013-0029-z

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., Sherman, M. A., Vitzthum, C. M., Luquette, L. J., Yandava, C. N., Yang, P., Chittenden, T. W., Hatem, N. E., Ryu, S. C., Woodworth, M. B., Park, P. J., & Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, *359*(6375), 555-559. https://doi.org/10.1126/science.aao4426

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2023). Hallmarks of aging: An expanding universe. *Cell*, *186*(2), 243-278. https://doi.org/10.1016/j.cell.2022.11.001

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029-1041.e1021. https://doi.org/10.1016/j.cell.2017.09.042

Maslov, A. Y., Makhortov, S., Sun, S., Heid, J., Dong, X., Lee, M., & Vijg, J. (2022). Single-molecule, quantitative detection of low-abundance somatic mutations by high-throughput sequencing. *Sci Adv*, *8*(14), eabm3259. https://doi.org/10.1126/sciadv.abm3259

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Melboucy-Belkhir, S., Pradère, P., Tadbiri, S., Habib, S., Bacrot, A., Brayer, S., Mari, B., Besnard, V., Mailleux, A., Guenther, A., Castier, Y., Mal, H., Crestani, B., & Plantier, L. (2014). Forkhead Box F1 represses cell growth and inhibits COL1 and ARPC2 expression in lung fibroblasts in vitro. *Am J Physiol Lung Cell Mol Physiol*, *307*(11), L838-847. https://doi.org/10.1152/ajplung.00012.2014

Miller, M. B., Huang, A. Y., Kim, J., Zhou, Z., Kirkham, S. L., Maury, E. A., Ziegenfuss, J. S., Reed, H. C., Neil, J. E., Rento, L., Ryu, S. C., Ma, C. C., Luquette, L. J., Ames, H. M., Oakley, D. H., Frosch, M. P., Hyman, B. T., Lodato, M. A., Lee, E. A., & Walsh, C. A. (2022). Somatic genomic changes in single Alzheimer's disease neurons. *Nature*, *604*(7907), 714-722. https://doi.org/10.1038/s41586-022-04640-1

Ren, P., Dong, X., & Vijg, J. (2022). Age-related somatic mutation burden in human tissues. *Front Aging*, *3*, 1018119. https://doi.org/10.3389/fragi.2022.1018119

Robinson, P. S., Coorens, T. H. H., Palles, C., Mitchell, E., Abascal, F., Olafsson, S., Lee, B. C. H., Lawson, A. R. J., Lee-Six, H., Moore, L., Sanders, M. A., Hewinson, J., Martin, L., Pinna, C. M. A., Galavotti, S., Rahbari, R., Campbell, P. J., Martincorena, I., Tomlinson, I., & Stratton, M. R. (2021). Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat Genet*, *53*(10), 1434-1442. https://doi.org/10.1038/s41588-021-00930-y

Saul, D., Kosinsky, R. L., Atkinson, E. J., Doolittle, M. L., Zhang, X., LeBrasseur, N. K., Pignolo, R. J., Robbins, P. D., Niedernhofer, L. J., Ikeno, Y., Jurk, D., Passos, J. F., Hickson, L. J., Xue, A., Monroe, D. G., Tchkonia, T., Kirkland, J. L., Farr, J. N., & Khosla, S. (2022). A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. *Nature Communications*, *13*(1), 4827. https://doi.org/10.1038/s41467-022-32552-1

Schumacher, B., Pothof, J., Vijg, J., & Hoeijmakers, J. H. J. (2021). The central role of DNA damage in the ageing process. *Nature*, *592*(7856), 695-703. https://doi.org/10.1038/s41586-021-03307-7

Selby, C. P., Lindsey-Boltz, L. A., Li, W., & Sancar, A. (2023). Molecular Mechanisms of Transcription-Coupled Repair. *Annu Rev Biochem*, *92*, 115-144. https://doi.org/10.1146/annurev-biochem-041522-034232

Sender, R., & Milo, R. (2021). The distribution of cellular turnover in the human body. *Nat Med*, *27*(1), 45-48. https://doi.org/10.1038/s41591-020-01182-9

Szilard, L. (1959). ON THE NATURE OF THE AGING PROCESS. *Proc Natl Acad Sci U S A*, *45*(1), 30-45. https://doi.org/10.1073/pnas.45.1.30

Vijg, J., & Dong, X. (2020). Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell*, *182*(1), 12-23. https://doi.org/10.1016/j.cell.2020.06.024

Vorontsov, I. E., Khimulya, G., Lukianova, E. N., Nikolaeva, D. D., Eliseeva, I. A., Kulakovskiy, I. V., & Makeev, V. J. (2016). Negative selection maintains transcription factor binding motifs in human cancer. *BMC Genomics*, *17 Suppl 2*(Suppl 2), 395. https://doi.org/10.1186/s12864-016-2728-9

Vorontsov, I. E., Khimulya, G., Lukianova, E. N., Nikolaeva, D. D., Eliseeva, I. A., Kulakovskiy, I. V., & Makeev, V. J. (2016). Negative selection maintains transcription factor binding motifs in human cancer. *BMC Genomics*, *17*(2), 395. https://doi.org/10.1186/s12864-016-2728-9

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., & Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, *350*(6264), 1096-1101. https://doi.org/10.1126/science.aac7041

Watson, D. E., Cunningham, M. L., & Tindall, K. R. (1998). Spontaneous and ENU-induced mutation spectra at the cII locus in Big Blue Rat2 embryonic fibroblasts. *Mutagenesis*, *13*(5), 487-497. https://doi.org/10.1093/mutage/13.5.487

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Yang, J.-L., Lee, P.-C., Lin, S.-R., & Lin, J.-G. (1994). Comparison of mutation spectra induced by N-ethyl-N-nitrosourea in the hprt gene of Mer+ and Mer− diploid human fibroblasts. *Carcinogenesis*, *15*(5), 939-945. https://doi.org/10.1093/carcin/15.5.939

Yang, L. (2020). A Practical Guide for Structural Variation Detection in the Human Genome. *Curr Protoc Hum Genet*, *107*(1), e103. https://doi.org/10.1002/cphg.103

Yurchenko, A. A., Rajabi, F., Braz-Petta, T., Fassihi, H., Lehmann, A., Nishigori, C., Wang, J., Padioleau, I., Gunbin, K., Panunzi, L., Morice-Picard, F., Laplante, P., Robert, C., Kannouche, P. L., Menck, C. F. M., Sarasin, A., & Nikolaev, S. I. (2023). Genomic mutation landscape of skin cancers from DNA repair-deficient xeroderma pigmentosum patients. *Nat Commun*, *14*(1), 2561. https://doi.org/10.1038/s41467-023-38311-0

Zapata, L., Pich, O., Serrano, L., Kondrashov, F. A., Ossowski, S., & Schaefer, M. H. (2018). Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol*, *19*(1), 67. https://doi.org/10.1186/s13059-018-1434-0

Zhang, L., Dong, X., Lee, M., Maslov, A. Y., Wang, T., & Vijg, J. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc Natl Acad Sci U S A*, *116*(18), 9014-9019. https://doi.org/10.1073/pnas.1902510116

Zhang, L., Lee, M., Hao, X., Ma, X., Xia, C., Zhao, Y., Ehlert, J., Chi, Z., Jin, B., Cutler, R., Maslov, A. Y., Barabási, A. L., Hoeijmakers, J. H. J., Edelmann, W., Vijg, J., & Dong, X. (2025). Divergent accumulation patterns of SNVs and INDELs reveal negative selection in noncancerous cells. *Innovation (Camb)*, *6*(10), 101008. https://doi.org/10.1016/j.xinn.2025.101008

Zhang, L., Lee, M., Maslov, A. Y., Montagna, C., Vijg, J., & Dong, X. (2024). Analyzing somatic mutations by single-cell whole-genome sequencing. *Nat Protoc*, *19*(2), 487-516. https://doi.org/10.1038/s41596-023-00914-8

Zhang, X., Theotokis, P. I., Li, N., Wright, C. F., Samocha, K. E., Whiffin, N., & Ware, J. S. (2024). Genetic constraint at single amino acid resolution in protein domains improves missense variant prioritisation and gene discovery. *Genome Med*, *16*(1), 88. https://doi.org/10.1186/s13073-024-01358-9

Zoppi, N., Gardella, R., De Paepe, A., Barlati, S., & Colombi, M. (2004). Human Fibroblasts with Mutations in COL5A1 and COL3A1 Genes Do Not Organize Collagens and Fibronectin in the Extracellular Matrix, Down-regulate α2β1 Integrin, and Recruit αvβ3 Instead of α5β1 Integrin*. *Journal of Biological Chemistry*, *279*(18), 18157-18168. https://doi.org/https://doi.org/10.1074/jbc.M312609200

Figure 1. Repeated Mutagen Treatment of Normal Cells Only Slightly Affects Cell Growth and Death

(A) Schematic depiction of the experimental design. Human fetal lung fibroblasts (IMR-90) were treated with a sublethal dose of 50 µg/mL of *N*-ethyl-*N*-nitrosourea (ENU), which was followed by a recovery period of 7 days. At 3-day and 7-day time points, 1 million cells were passaged, and a sample was taken for further analysis. This process was repeated for 9 cycles (Methods). **(B)** Cumulative population doublings (PDs) of the control (CTRL) and ENU-treated (ENU) groups throughout the experiment. Statistics: n=3; Data represent mean ± S.D.; P-value estimated using a two-way repeated measures ANOVA. **(C)** Cumulative PDs of the CTRL and ENU groups assessed at the 3-day time point. Statistics: Same as in (B). **(D)** Same as in (C), but at 7-day time point. Statistics: Same as in (B). **(E)** Percentage of cells detected in early apoptosis at the 3-day time point at each cycle (left) or with all cycles pooled (right). Statistics: n=2; Data represent mean ± S.D.; P-value of data from each cycle was estimated using a two-way repeated measures ANOVA. P-value of data from all cycles pooled was estimated using a paired t-test; **: p <= 0.005; ****: p<= 0.0001. **(F)** Same as in (E), but for late apoptosis. Statistics: Same as in (E). **(G)** Same as in (E), but for 7-day time point. Statistics: Same as in (E). **(H)** Same as in F, but for 7-day time point. Statistics: Same as in (E).

Figure 2. RNA-sequencing Reveals Increased Cell Cycle Inhibition and Apoptosis in ENU-treated Cells

(A) Differential gene expression (DGE) of control cycle 9 vs. control cycle 1 groups. Red points indicate Log_2 fold change > 2 and adjusted *p-value* < 0.05. See Table S1 for sample details and Table S2 for DGE results. Statistics: n=3; DGE p-values estimated using a Wald test followed by Benjamini-Hochberg correction for multiple hypothesis testing. **(B-D)** Gene set enrichment analysis (GSEA) of DGE results from the control cycle 9 vs. control cycle 1 comparison for KEGG cell cycle (B), KEGG apoptosis (C), and SenMayo senescence (D) gene sets. NES: normalized enrichment score. Statistics: n=3; p-values estimated using a permutation test. **(E)** Same as in (A), but for the ENU-treated cycle 9 vs. control cycle 9 comparison. See Supplemental Note 1 for batch

correction note. Statistics: Same as in (A). **(F-H)** Same as in (B-D), but for the ENU-treated cycle 9 vs. control cycle 9 comparison. Statistics: Same as in (B-D).

Figure 3. Repeated Mutagen Treatment of Normal Cells Results in Extremely High Mutation Burden.

(A) Estimated load of somatic single nucleotide variants (SNVs) per cell after correction for genome coverage, sensitivity, and technical artifacts (Methods). See Figure S2H for observed SNV load per cell and Figures S2K-L for ENU-specific mutation burden. See Table S3 for sample details and Table S10 for mutation catalogue. Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D. p-values of comparisons between cycles within the same condition were estimated with a two-sided linear model with estimated marginal means, Tukey's HSD. P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA and shown at very top of plot. p-value legend: ns: $p > 0.05$, ****: $p <= 0.0001$. **(B)** Fit of linear model to estimated SNV load per cell shown in (A) vs. the treatment cycle cells were collected at. Note that cells from control cycle 1 are included and control cycle 9 is omitted. Shown at top is the fit of a linear equation. Statistics: data represent mean ± S.D.; Pearson R correlation coefficient and p-value of the correlation was estimated using a t-test.

Figure 4. Mutational Spectra and Signatures Resulting from Repeated Mutagen Treatment

(A) Mutational spectra of the relative contribution of SNV types per cell grouped by condition. See Figure S3A for groups stratified by cycle. Note that cells from the control cycle 1 group were excluded from the mutational spectra and signatures analyses due to abnormally high C>T frequency suggestive of a sample preparation artifact (Figures S3A-D; Methods). Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; Data represent mean ± S.D. (**B)** Relative contribution of de novo mutational signatures SBSA and SBSB to individual cells (shown as rows) grouped by condition and cycle. See Figure S3F for determination of number of signatures to extract and Figure S3G for absolute contribution. **(C)** Comparison of de novo mutational signature SBSA to the ENU signature from the SIGNAL database. Statistics: cosine similarity 0.858, residual sum of squares=$7.83*10^{-3}$. **(D)** Comparison of de novo mutational signature SBSB to most similar fitted signature SBS5 from the COSMIC database. Statistics: cosine similarity=0.955, residual sum of squares= $1.79*10^{-3}$.

Figure 5. Negative Selection Against Damaging Coding and Non-Coding Variants in ENU-treated Cells.

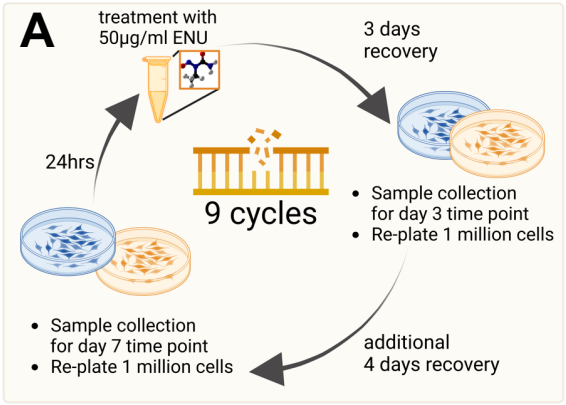
(A) Observed frequency of synonymous and nonsynonymous variants per cell. See Table S3 for sample details and Table S5 for variant annotation. Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA and shown at very top of plot; p-value legend: ***: $p <= 0.001$. **(B)** $Log_2(O/E)$ ratio of synonymous and nonsynonymous variants for ENU-treated cells. See figures S4C, S4D, and S4E for results of without normalization, results of control

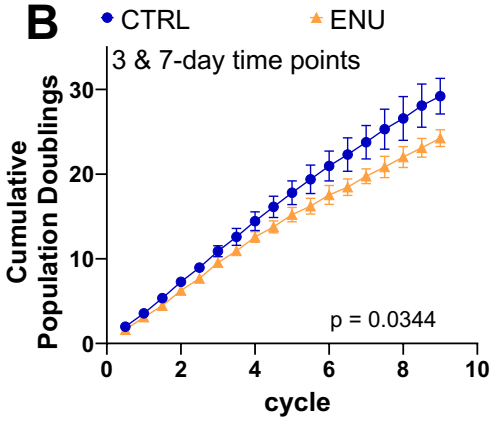
cells, and results per cell, respectively. See Table S5 for detailed results. Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values to assess if the Log_2(O/E) ratio significantly deviates from 0 were estimated using a permutation test (n=10,000) but were only estimated if the frequency of observed variants allowed for sufficiency statistical power (Figure S4B; Table S5); p-value legend: **: p <= 0.01. **(C)** Estimated percent of nonsynonymous variants out of total coding variants that are reduced in the ENU-treated group due to negative selection. Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; data represent mean ± S.D. **(D)** Log2O/E of the N/S ratio for ENU-treated cells. See figures S4I and S4J for results of control cells and per cell, respectively. Statistics: Same as in (B). p-value legend: *: p <= 0.05; **: p <= 0.01. **(E)** Observed frequency of various types of nonsynonymous variants per cell. Statistics: Same as in (A). p-value legend: ****: p <= 0.0001. **(F)** Log2O/E of various types of nonsynonymous variants for ENU-treated cells. See figures S4F, S4G, and S4H for results of without normalization, results of control cells, and results per cell, respectively. Statistics: Same as in (B). p-value legend: *: p <= 0.05; **: p <= 0.01. **(G)** Observed frequency of various types of non-coding variants per cell. Statistics: Same as in (A). ****: p <= 0.0001. **(H)** Log_2O/E of various types of non-coding variants for ENU-treated cells. See figures S4K, S4L, and S4M for results of control cells, without normalization, and per cell, respectively. See Supplemental Note 5 for discussion on results of the intergenic region. Statistics: Same as in (B). ****: p <= 0.0001; **: p <= 0.01. **(I)** Estimated percent reduction of variants in 3' UTR and splice region variants that are protected against in the ENU-treated group due to negative selection. Statistics: Same as in (C).

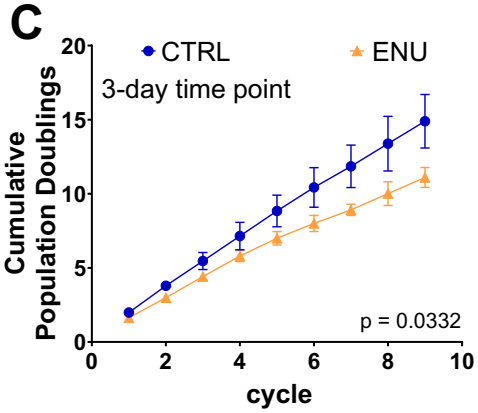
Figure 6. Pathways Supporting Fibroblast Growth and Cell Identity are Protected from Mutation Accumulation

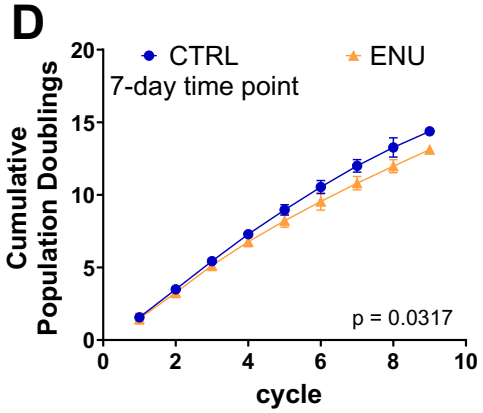
(A) Observed frequency of SNVs per cell in functional regions associated with genes that were determined to be non-expressed or expressed from bulk mRNA transcriptome data (Table S7). Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA; p-value legend: ****: p <= 0.0001. **(B)** Log_2(O/E) ratio of SNVs in functional regions associated with genes determined to be non-expressed or expressed in bulk mRNA transcriptome data (Table S7). See figures S6A, S6B, and S6C for results without normalization, results of control cells, and results per cell, respectively. Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; data represent mean ± S.D; P-values to assess if the Log_2(O/E) ratio significantly deviates from 0 were estimated using a permutation test (n=10,000) but were only estimated if the frequency of observed variants allowed for sufficiency statistical power (Figure S4B; Table S5); p-value legend: ****: p <= 0.0001. **(C)** Observed frequency of SNVs per cell in functional regions associated with nonessential or essential gene sets expressed in bulk mRNA transcriptome data (Table S7). Statistics: Same in in (A); p-value legend: ***: p <= 0.001. **(D)** Log_2(O/E) ratio of SNVs in functional regions associated with nonessential or essential gene sets expressed in bulk mRNA transcriptome data (Table S7). See figures S6D, S6E, and S6F for results without normalization, results of control cells, and results per cell, respectively. Statistics: Same as in (B); p-value legend: ****: p <= 0.0001. **(E)** Observed frequency of SNVs per cell in functional regions associated with haplosufficient or haploinsufficient gene sets expressed in bulk mRNA transcriptome data (Table S7). Statistics: Same in in (A); p-value legend: ***: p <= 0.001. **(F)** Log_2(O/E) ratio of SNVs in functional regions associated with haplosufficient or haploinsufficient gene sets expressed in bulk mRNA transcriptome data (Table S7). See figures S6G, S6H, and S6I

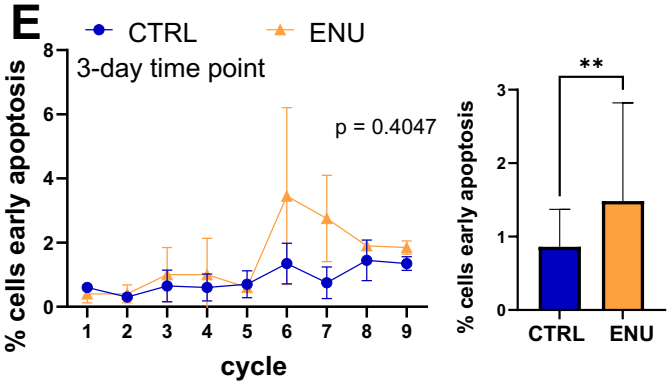
results without normalization, results of control cells, and results per cell, respectively. Statistics: Same as in (B); p-value legend: **: p <= 0.01; ****: p <= 0.0001. **(G)** Mean observed frequency of SNVs within the functional regions of pathways expressed in bulk mRNA transcriptome data (Table S8). See figure S6J for top 20 most enriched pathways. Statistics: n=399 pathways; P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA; p-value legend: ****: p <= 0.0001. **(H)** Mean Log_2(O/E) ratio of SNVs within the functional regions of pathways expressed in bulk mRNA transcriptome data (Table S8). See figures S6K and S6L for results of control cell and results without normalization, respectively. See table S9 for detailed results. Statistics: n=399 pathways; Otherwise, same as in (B), with the addition of Benjamini-Hochberg correction for multiple hypothesis testing of pathways. **(I-K)** Mean Log_2(O/E) ratio (I), estimated percent reduction in SNVs (J), and estimated reduction in SNV frequency (K) within functional regions of pathways found to be under significant negative selection (p <= 0.05) at cycle 9 in the ENU-treated group. Statistics: Same as in (B); p-value legend: *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001.

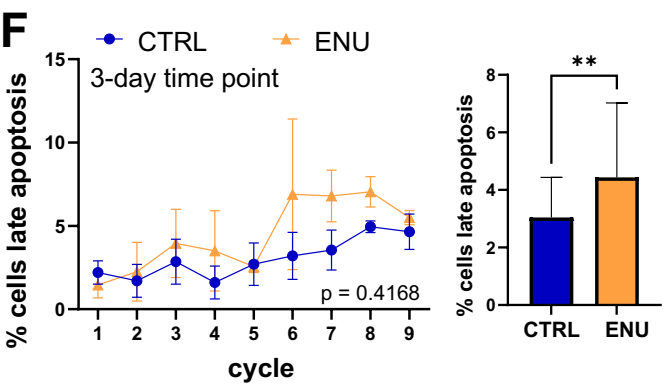

A
treatment with
50µg/ml ENU
3 days
recovery
24hrs
9 cycles
• Sample collection
for day 3 time point
• Re-plate 1 million cells
• Sample collection
for day 7 time point
• Re-plate 1 million cells
additional
4 days recovery

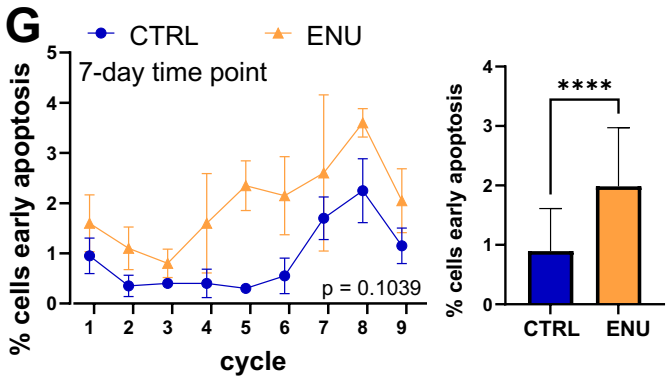

B
CTRL
ENU
3 & 7-day time points
Cumulative
Population Doublings
p = 0.0344
cycle

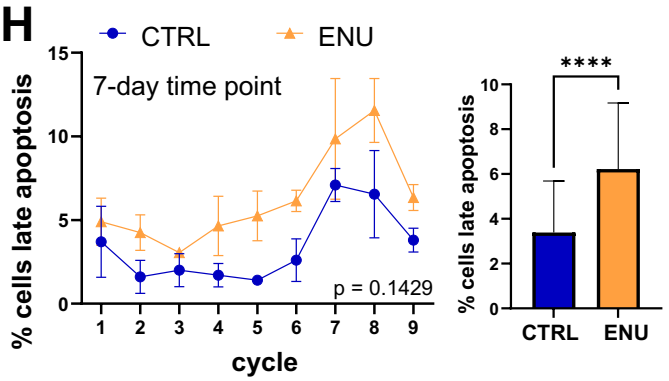

C
CTRL
ENU
3-day time point
Cumulative
Population Doublings
p = 0.0332
cycle

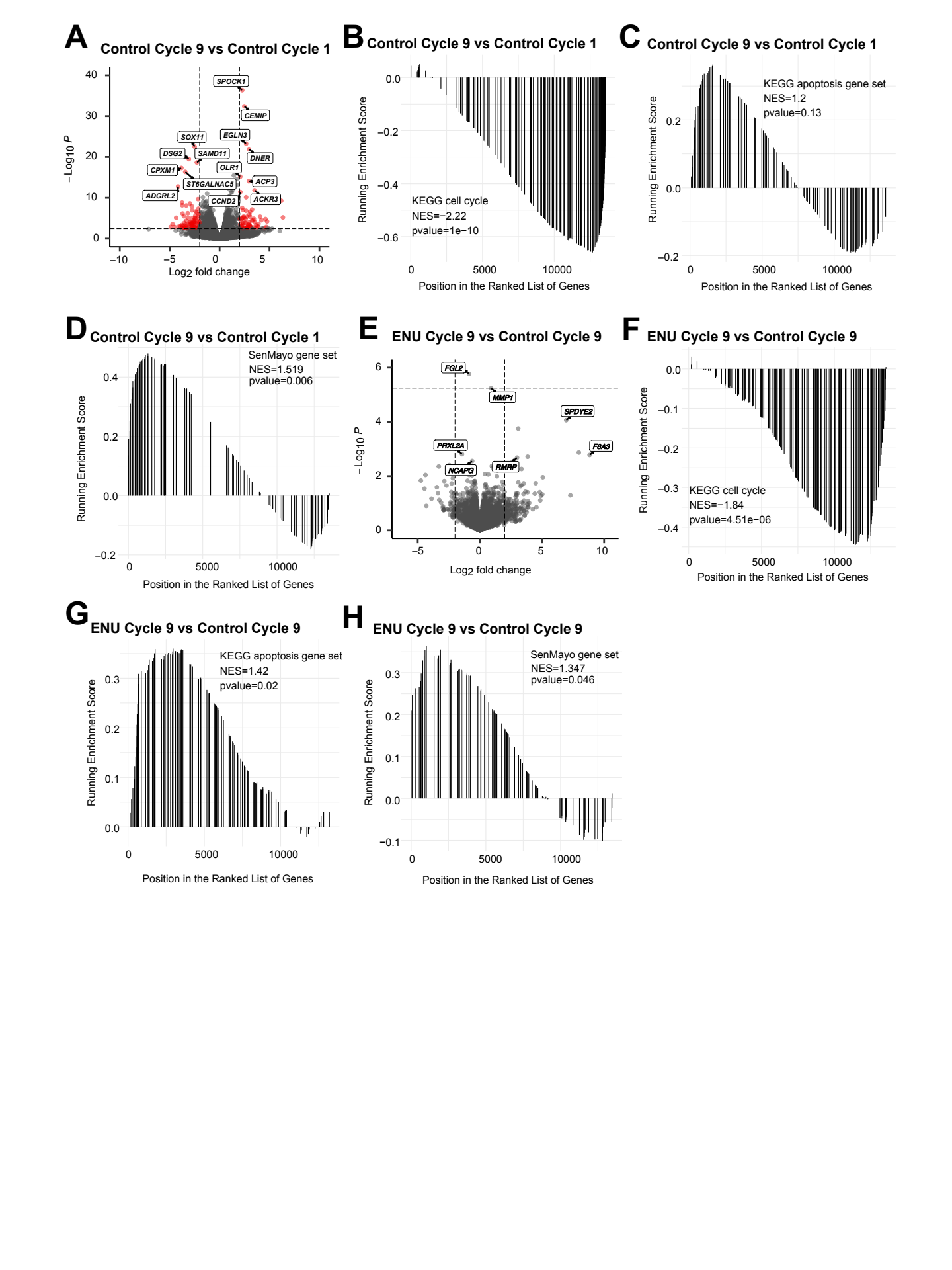

D
CTRL
ENU
7-day time point
Cumulative
Population Doublings
p = 0.0317
cycle

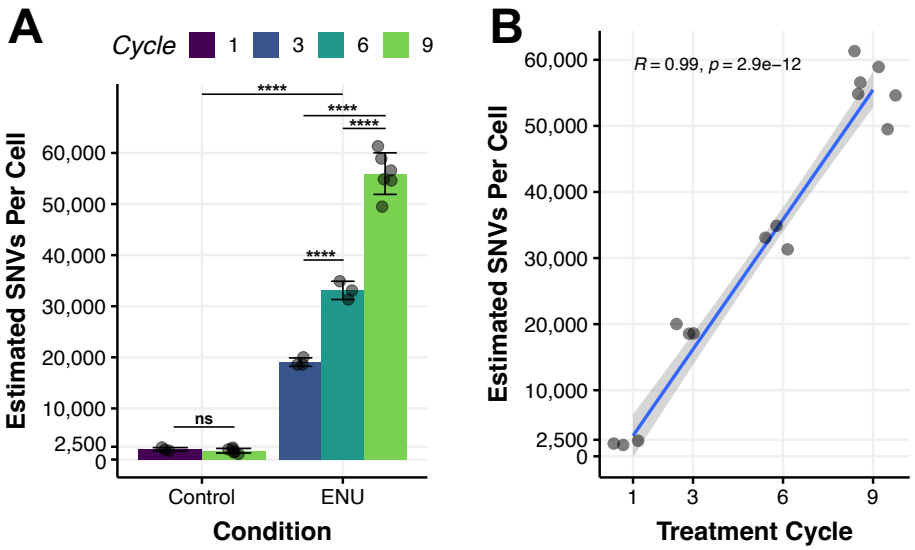

E
CTRL
ENU
3-day time point
p = 0.4047
% cells early apoptosis
cycle
**
CTRL
ENU

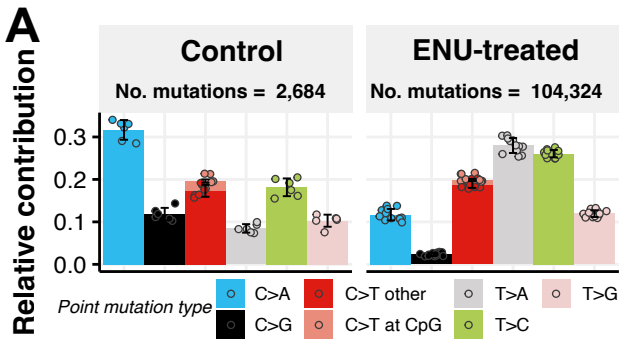

F
CTRL
ENU
3-day time point
p = 0.4168
% cells late apoptosis
cycle
**
CTRL
ENU

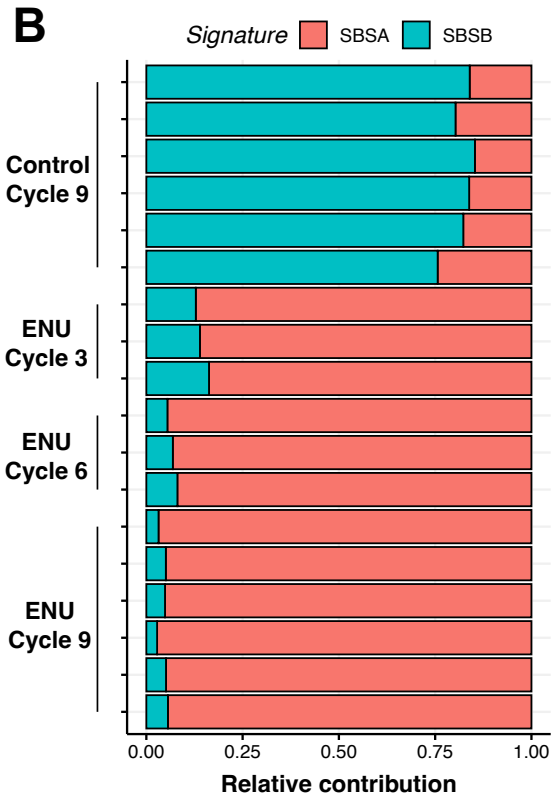

G
CTRL
ENU
7-day time point
p = 0.1039
% cells early apoptosis
cycle
****
CTRL
ENU

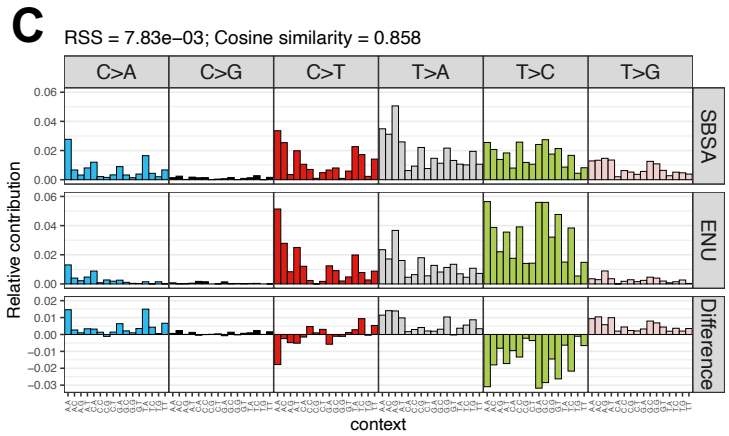

H
CTRL
ENU
7-day time point
p = 0.1429
% cells late apoptosis
cycle
****
CTRL
ENU

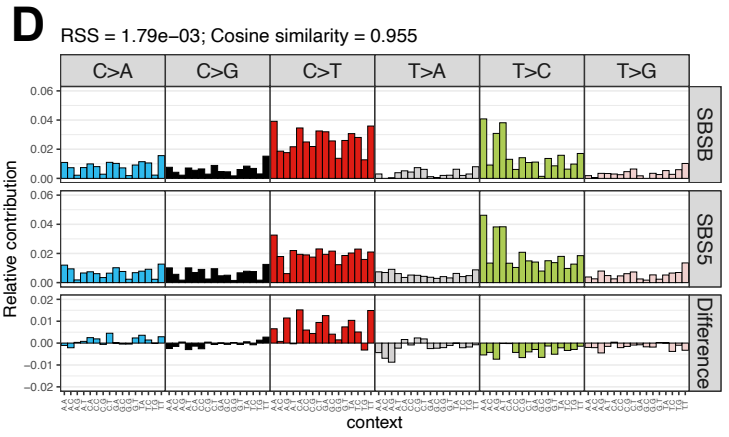

A
Control Cycle 9 vs Control Cycle 1
SPOCK1
CEMIP
SOX11
EGLN3
SAMD11
DSG2
DNER
CPXM1
OLR1
ST6GALNAC5
ACP3
ADGRL2
CCND2
ACKR3
−Log10 P
Log2 fold change
B
Control Cycle 9 vs Control Cycle 1
KEGG cell cycle
NES=−2.22
pvalue=1e−10
Running Enrichment Score
Position in the Ranked List of Genes
C
Control Cycle 9 vs Control Cycle 1
KEGG apoptosis gene set
NES=1.2
pvalue=0.13
Running Enrichment Score
Position in the Ranked List of Genes
D
Control Cycle 9 vs Control Cycle 1
SenMayo gene set
NES=1.519
pvalue=0.006
Running Enrichment Score
Position in the Ranked List of Genes
E
ENU Cycle 9 vs Control Cycle 9
FGL2
MMP1
SPDYE2
PRXL2A
FBA3
NCAPG
RMRP
−Log10 P
Log2 fold change
F
ENU Cycle 9 vs Control Cycle 9
KEGG cell cycle
NES=−1.84
pvalue=4.51e−06
Running Enrichment Score
Position in the Ranked List of Genes
G
ENU Cycle 9 vs Control Cycle 9
KEGG apoptosis gene set
NES=1.42
pvalue=0.02
Running Enrichment Score
Position in the Ranked List of Genes
H
ENU Cycle 9 vs Control Cycle 9
SenMayo gene set
NES=1.347
pvalue=0.046
Running Enrichment Score
Position in the Ranked List of Genes

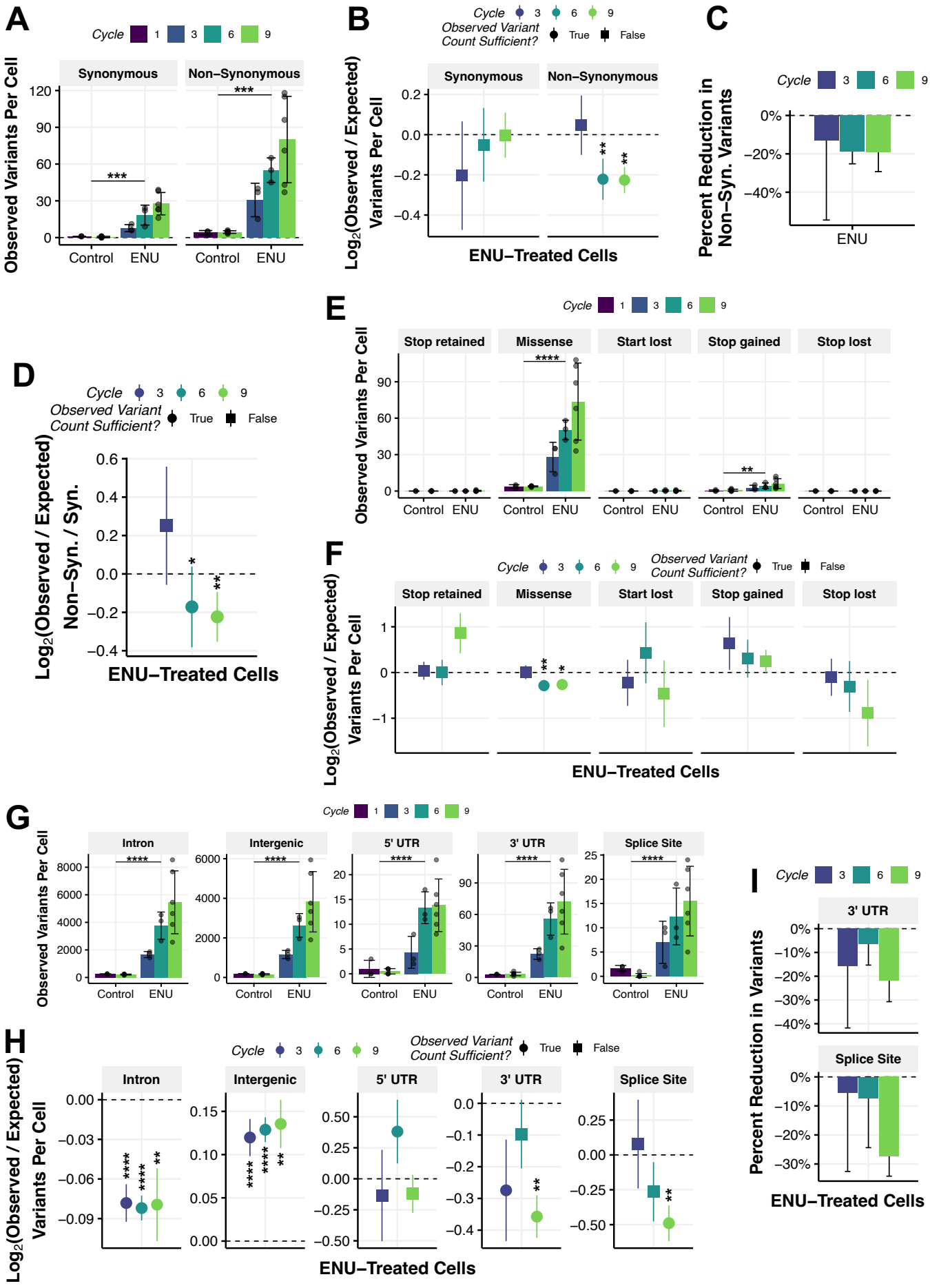

A
Cycle 1 3 6 9
Estimated SNVs Per Cell
60,000
50,000
40,000
30,000
20,000
10,000
2,500
0
****
****
****
****
ns
Control
ENU
Condition
B
R = 0.99, p = 2.9e−12
Estimated SNVs Per Cell
60,000
50,000
40,000
30,000
20,000
10,000
2,500
0
1
3
6
9
Treatment Cycle

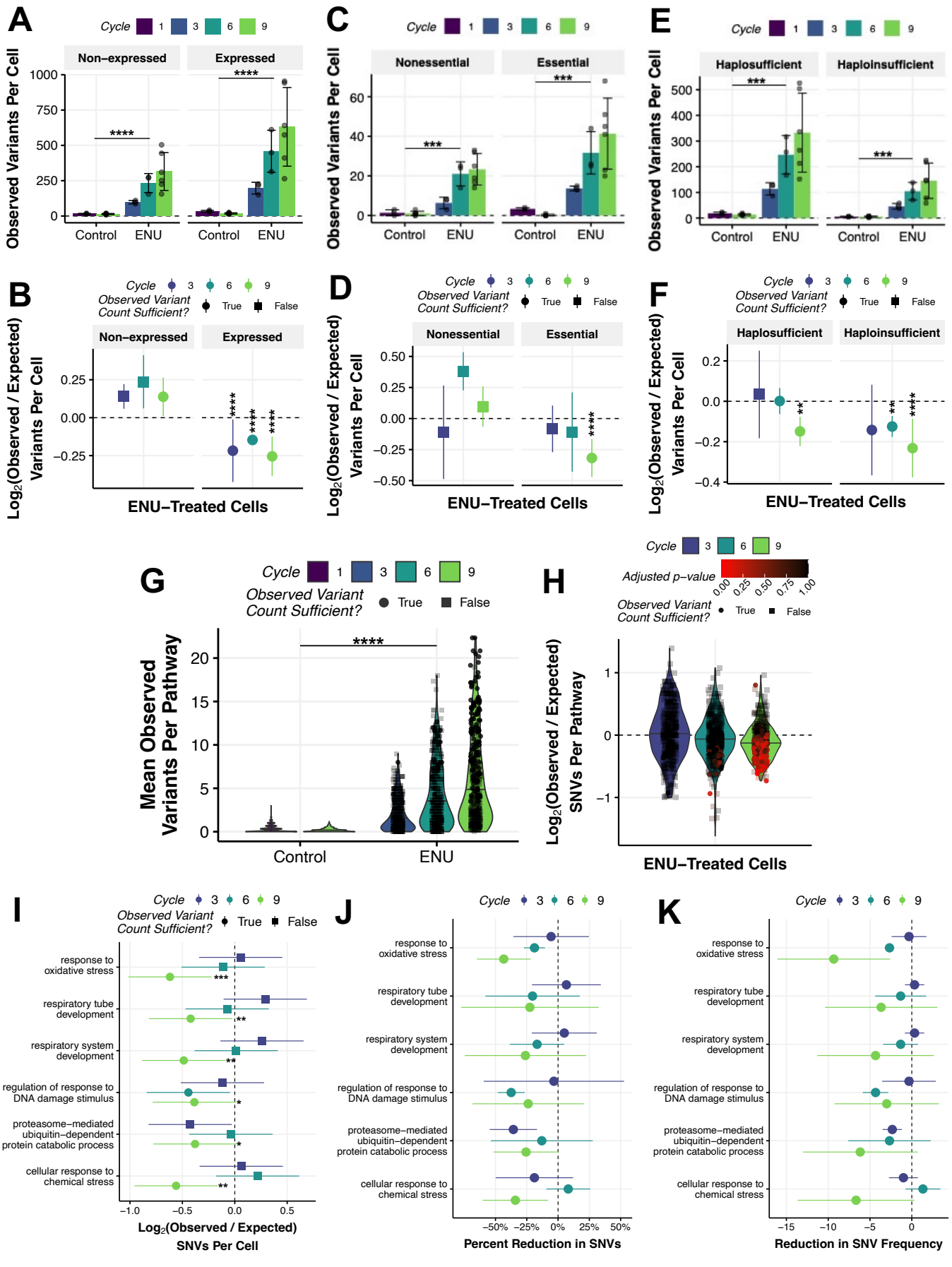

A
Control
No. mutations = 2,684
ENU-treated
No. mutations = 104,324
Relative contribution
0.3
0.2
0.1
0.0
Point mutation type
C>A
C>G
C>T other
C>T at CpG
T>A
T>C
T>G


B
Signature
SBSA
SBSB
Control Cycle 9
ENU Cycle 3
ENU Cycle 6
ENU Cycle 9
0.00
0.25
0.50
0.75
1.00
Relative contribution


C
RSS = 7.83e−03; Cosine similarity = 0.858
C>A
C>G
C>T
T>A
T>C
T>G
SBSA
ENU
Difference
Relative contribution
context


D
RSS = 1.79e−03; Cosine similarity = 0.955
C>A
C>G
C>T
T>A
T>C
T>G
SBSB
SBS5
Difference
Relative contribution
context


A
Cycle 1 3 6 9
Observed Variants Per Cell
Synonymous
Non–Synonymous
Control ENU
B
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU–Treated Cells
C
Cycle 3 6 9
Percent Reduction in Non–Syn. Variants
ENU
D
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Non–Syn. / Syn.
ENU–Treated Cells
E
Cycle 1 3 6 9
Observed Variants Per Cell
Stop retained
Missense
Start lost
Stop gained
Stop lost
F
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU–Treated Cells
G
Cycle 1 3 6 9
Observed Variants Per Cell
Intron
Intergenic
5' UTR
3' UTR
Splice Site
H
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU–Treated Cells
I
Cycle 3 6 9
3' UTR
Splice Site
Percent Reduction in Variants
ENU–Treated Cells


A
Cycle 1 3 6 9
Non−expressed
Expressed
Observed Variants Per Cell
Control
ENU
B
Cycle 3 6 9
Observed Variant Count Sufficient? True False
Log2(Observed / Expected) Variants Per Cell
ENU−Treated Cells
C
Nonessential
Essential
D
E
Haplosufficient
Haploinsufficient
F
G
Mean Observed Variants Per Pathway
H
Adjusted p−value
Log2(Observed / Expected) SNVs Per Pathway
I
response to oxidative stress
respiratory tube development
respiratory system development
regulation of response to DNA damage stimulus
proteasome−mediated ubiquitin−dependent protein catabolic process
cellular response to chemical stress
Log2(Observed / Expected) SNVs Per Cell
J
Percent Reduction in SNVs
K
Reduction in SNV Frequency