# Single-mitosis dissection of acute and chronic DNA mutagenesis and repair

# Single-mitosis dissection of acute and chronic DNA mutagenesis and repair

Paul Adrian Ginno[1], Helena Borgers[1], Sarah J. Aitken[2,3,4], Christina Ernst[3,6], Anja Schneider[1], Mikaela Behm[1], Martin S. Taylor[5,*], Duncan T. Odom[1,3,*]

1. German Cancer Research Center (DKFZ), Division of Regulatory Genomics and Cancer Evolution, Heidelberg, Germany

2. MRC Toxicology Unit, University of Cambridge, Cambridge, UK

3. Cancer Research UK - Cambridge Institute, University of Cambridge, Cambridge, UK

4. Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

5. MRC Human Genetics Unit, MRC Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

6. Present address: School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

* Corresponding authors

## Abstract

How chronic mutational processes and punctuated bursts of DNA damage drive evolution of the cancer genome is poorly understood. Here, we demonstrate a novel strategy to both disentangle and quantify distinct mechanisms underlying genome evolution in single cells, during single mitoses, and at single strand resolution both in vitro and in vivo. To distinguish mutations caused by chronic (ROS) and acute (UV) processes, we microfluidically separated individual pairs of sister cells resulting from the first mitosis following a burst of UV mutagenesis. Strikingly, ROS mutagenesis in transcribed regions is reduced in a strand-agnostic manner, while burst UV mutations manifest as sister-specific events, revealing mirror-image mutation phasing profiles across all chromosomes. We show that successive rounds of genome replication over persistent UV damage drives multi-allelic variation at tandem CC sites, and finally resolve phased mutational patterns to single-strands across the entire genome of liver tumours from F1 mice. The strategy we present here can be widely used to resolve the contributions of overlapping cancer relevant mutational processes.

## Introduction

Cancers are complex ecosystems of competing clones, each of which developed via highly idiosyncratic evolutionary changes. Indeed, tumours often develop after decades of both chronic mutational processes, such as oxidative damage[1], and punctuated bursts of damage, such as chromothripsis and mutagen exposure[2]. Thus, it has proven exceptionally difficult to retrospectively disentangle the interleaved mechanisms active in single cells that can trigger transformation. Methods to analyse the mutagenic processes in single cells are required as this is ultimately the origin of every tumour.

Two major sources that introduce variation into the somatic genome are chronic oxidative damage and acute mutagen exposure. Oxidation is one of the most common forms of recurring damage to DNA[3], whereas environmental genotoxins such as UV (sunburn)[4] or nitrosamines[5] can result in short bursts of mutagenesis. Mammals have complex repair systems to maintain DNA fidelity, including base excision repair (BER) and nucleotide excision repair (NER)[6]. Although these mechanisms can resolve most damage occurring in a cell[7,8], fixed mutations eventually accumulate in normal tissues, tumours, and cultured cells[2]. Several aspects regarding how DNA damage resolves into genomic mutations remain enigmatic.

DNA is chronically challenged by reactive oxygen species (ROS), often forming 8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxo-G). 8-oxo-G is usually repaired by BER[9], but if unrepaired can cause G > T transversions[9,10] introduced during DNA replication[1]. Deep whole genome sequencing (WGS) of cell culture models has revealed that ROS creates a background mutation landscape independent of any mutagenic exposure[10]. However, studying endogenous mutation patterns induced by oxidative damage is challenging given its extremely rare occurrence in the genome[11].

Acute pulses of DNA damage, such as exposure to UV light, can compromise genome integrity. The resulting bulky lesions include cyclobutane pyrimidine dimers (CPDs) or pyrimidine (6-4) pyrimidone photoproducts, which can be repaired by NER. The NER process itself is further distinguished by genomic context: global NER is responsible for bulky lesions across the entire genome, whereas transcription coupled NER (TCR) only resolves damage in transcribed regions[12]. Mutations in the NER protein XRCC1 result in the clinical phenotype xeroderma pigmentosum[13] where patients are sensitive to UV light, demonstrating how this form of DNA damage can trigger disease. Our work in mouse liver tumours suggests that lesions caused by acute genotoxic exposure are predominantly resolved as a function of DNA replication[14,15], resulting in strand-specific mutation stretches

79 across the entire genome. To directly test predictions of this model and mechanistically
80 dissect lesion segregation required a highly controlled system.

81 Here, we present both cell culture and *in vivo* strategies to dissect mutagenic processes
82 active in single cells. **Our adaptation of a microfluidics platform to precisely track single**
83 **mitotic events reveals mirror-image mutation phasing between mitotic sisters, a**
84 **previously-inferred result of the lesion segregation model.** These experiments have
85 afforded several key insights regarding how mutations are realised after genotoxic exposure.
86 More specifically, the resolution of our data have allowed us to revise the current model
87 regarding transcription associated ROS repair, confirm key predictions of lesion segregation,
88 reveal UV damage retention over multiple cell cycles and finally resolve mutations from
89 mouse tumours to single-strands of DNA. Finally, our approach of splitting cells after a
90 controlled number of divisions post genomic insult has the flexibility to address a myriad of
91 mutational processes relevant for transformation and cancer genome evolution.

92

93

## Results

### Mechanical separation of mammalian sister cells after a single mitosis

To disentangle acute and chronic mutagenesis and repair, we exploited an experimental system to physically separate two mammalian sister cells following a single mitosis. Such a system allows us to test several hypotheses about how DNA damage and repair together establish the resulting mutation landscape. For example, mutations acquired gradually in a single genome should be shared between mitotic sisters (Fig. 1a). In this way, the genome of each cell represents a unique evolutionary pathway of gradual mutations diverging from the reference origin to its current state in the growing population. In contrast, acute mutagenesis should not be shared between mitotic sisters unless damage is resolved into a mutation prior to genome replication.

The Berkeley Lights™ Lightning platform allows penning of individual cells and subsequent movement of these cells via light activation of a silicon membrane (Fig. 1b, Supp Video 1). Penning specifically refers to the process of moving a single cell into a single well, or pen, of the Berkeley Lights chip. This platform can image cells in regular intervals, physically separate sister cells after a single mitotic event and finally export the expanded populations. Subsequent whole-genome sequencing allowed us to determine the mutational landscape of these sister cell populations. Under optimised culture conditions (see methods), cells divide at a rate comparable to that measured in standard cell culture (Extended Data Fig. 1a) and the genome typically remains diploid (Extended Data Fig. 1b).

To control for ploidy and ensure each cell has gone through one cell cycle post acute mutagenesis, we first integrated the FastFUCCI construct[16] through lentiviral transduction into the nonadherent mouse cell line P388D1. FUCCI, or fluorescent ubiquitination-based cell cycle indicator, uses a combination of fluorophores degraded at particular times in the cell cycle. The clone selected from this line (PF1) revealed that fluorophore intensities were strongly correlated with DNA content, indicative of the cell cycle phase (red cells in G1, green cells in G2/M; Fig. 1b-c, Extended Data Fig. 1c-h). Since the Berkeley Lights™ system is equipped with lasers compatible with the FUCCI system (Fig. 1b) we could visualise the progression of PF1 cells through the cell cycle, where the first division points were clearly distinguishable from a G2/M (green) to G1 (red) switch in the penned cells (Extended Data Fig. 1c). In summary, we have established a controlled system to physically separate individual sister cells after a single mitotic cycle**.**

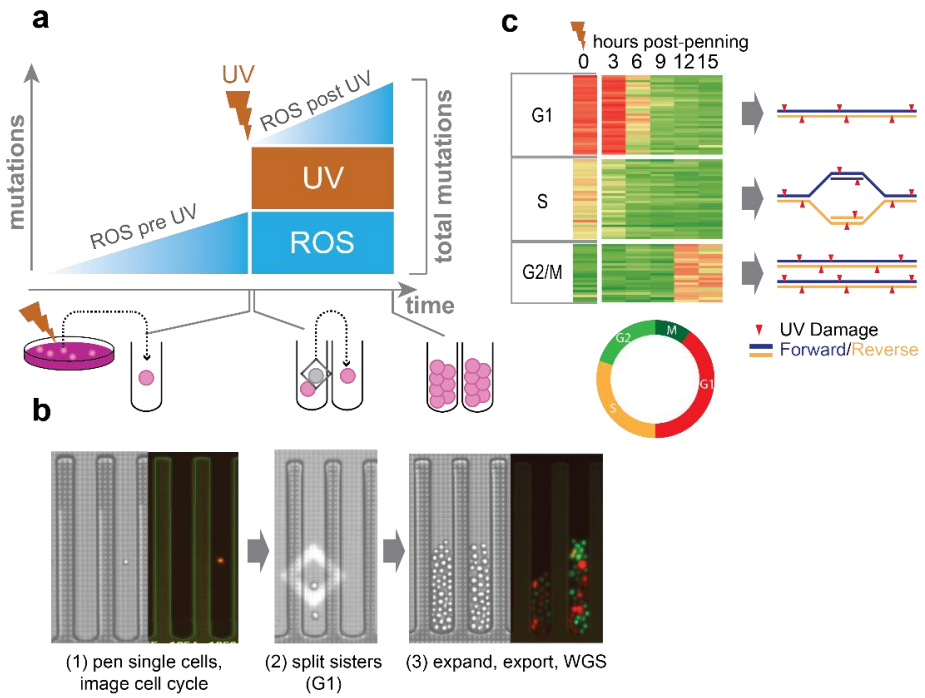



**Fig.1 | System to distinguish gradual and acute mutational processes in mammalian cells.** **a**, Model system to interrogate gradual (ROS, blue) and acute mutation pressure (UV light exposure, brown) on the mammalian genome. Triangles represent gradual accumulation of ROS mutations over several cell generations, while boxes represent fixed **mutations** in the genome after a specific cell division event. Schematic below depicts the experimental method from standard cell culture, exposure to UV, penning single cells, splitting two sister cells into separate pens after the first mitosis and finally proliferation/export. **b**, Images of single penned cells and FUCCI fluorophores imaged on the Berkeley Lights platform (left), splitting mitotic sisters with light cages (middle), as well as the expanded populations (right). WGS: Whole Genome Sequencing. **c**, A representative panel of cells (n=90) imaged to ascertain the intensity of red and green fluorophores in 3 hour intervals post-penning. Cell cycle designation to the left of each cluster was determined by the fluorophore intensity at the time of penning seen in column 1 and denoted by the grey border. Subsequent heatmap columns represent 3 hour imaging timepoints of the same cell. Cell cycle colour key circle is below. Models to the right implicate theoretical strand-specific (blue/yellow) distribution of DNA damage (red triangles) depending on the cell cycle phase at the time of UV treatment (time point 0).

**Sequencing of mitotic sisters distinguishes gradual and acute mutagenic processes**

We reasoned that UV and ROS mutations were distinguishable for two reasons. First, both mutation sources have clearly identifiable signatures for both the mutated base and trinucleotide context in which it arises. ROS predominantly causes G > T transversions[17–19] while UV damage results in C > T transitions, especially when adjacent to a pyrimidine on the same strand[10,20]. Second, ROS exposure is expected to be chronic in cell culture while UV treatment was a single acute pulse lasting 3 seconds. This has direct implications for the resulting variant allele frequency (VAF) in the growing population.

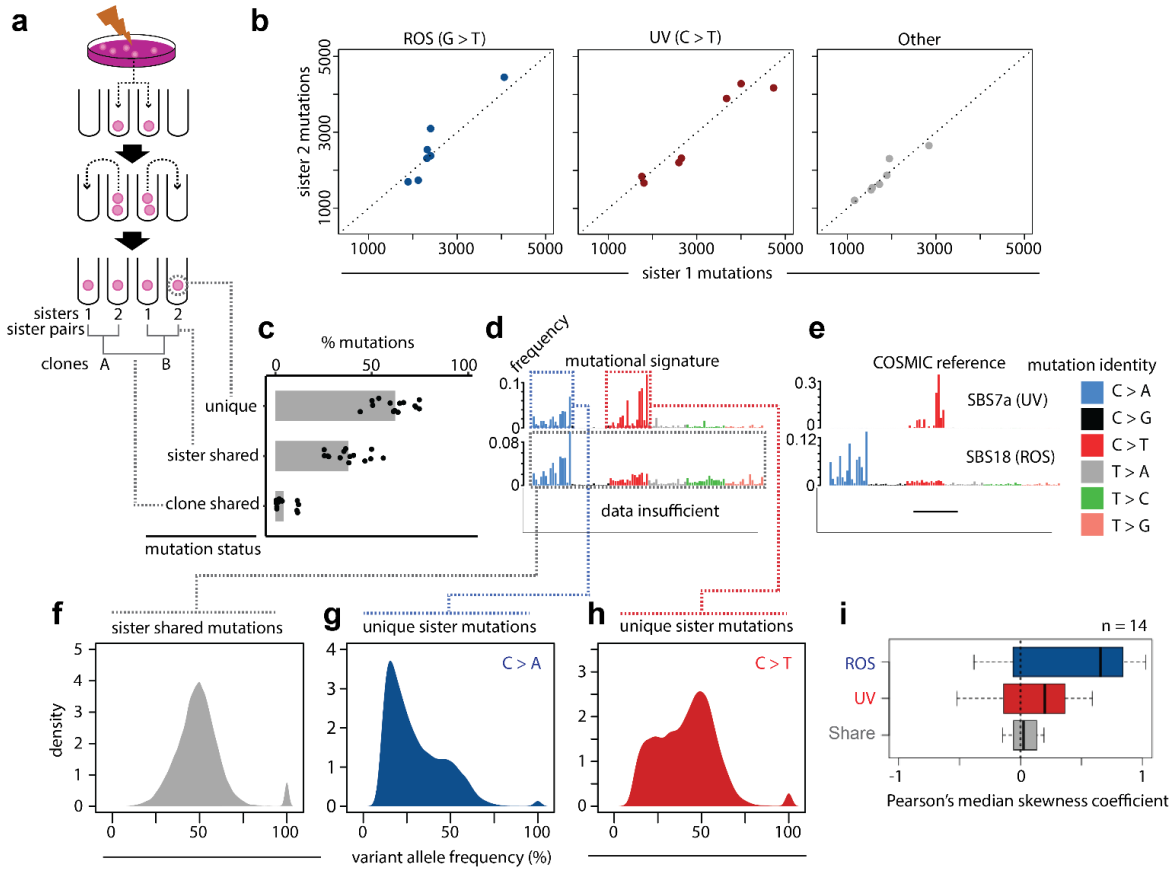
We maximised UV exposure by treating PF1 cells with an acute dose of UVC such that approximately 50% of the population proliferated post exposure (Extended Data Fig. 2a). Immediately after UV treatment, we penned individual cells on the Berkeley Lights™ platform and separated sisters after a single division. In further description of the results, each unique single penned cell is referred to as a clone, denoted by clone A, B, C etc (Fig. 2a), and the terms “sister 1” and “sister 2” represent the first mitotic sisters from a single penned cell. For example, clone A1 represents sister 1 from the singly penned clone A, while clone A2 represents sister 2 from the same singly penned clone A (Fig. 2a).

We expanded 14 daughters from 7 independent mitoses into clonal colonies large enough to perform whole genome sequencing (WGS). Each genome was sequenced to a minimum mean coverage of 20x, and mutations were called against libraries generated from untreated cells from the original culture. All clones contained 6000-9000 mutations per genome (Fig. 2b) and mutation signatures had high similarity to COSMIC[20] SBS7a/b and SBS18, patterns attributed to damage from UV light[20,21] and ROS, respectively (Fig. 2e, Extended Data Fig. 2b-c). We reasoned that sister cells should share a similar **number of mutations**, in line with a random distribution of damage from both UV and ROS sources. This was indeed the case, with near identical mutation frequency between mitotic sister genomes across mutation categories (Fig. 2b).

Given our observation of lesion segregation in mouse liver tumours[14], we surmised that mutations arising from acute, single dose UV damage will not usually be shared between mitotic sisters. This is because each sister inherits separate independently damaged strands. In agreement with this, roughly 90% of the C > T transitions characteristic of UV damage were unique to a single mitotic sister (Fig. 2c-d, upper). In contrast, mutations shared between sister cells (33% of aggregated mutations) were predominantly G > T transversions, **sharing a signature most similar to SBS18[20] and suggesting they are ROS induced mutations** (Fig. 2d middle, Extended Data Fig. 2c). **This is in agreement**

**with previous work profiling mutations influenced by high oxygen conditions in cell culture**[10] **and 8-oxo-G in human dermal fibroblasts**[22]. Finally, mutations shared between independent clones represent between 0.3 and 4.7% of the **total mutation count** (Fig. 2c), suggesting each clonal mutation landscape is predominantly the result of a unique evolutionary trajectory.

We hypothesised that G > T transversions shared between mitotic sisters represent the landscape of acquired ROS mutations present in the single cell, “clone”, when it was penned (Fig. 1a). If true, these mutations would overwhelmingly have a variant allele frequency of 50% at the population level, as both sister cells would inherit one mutated allele. In contrast, if mutations are accrued at later time points **a positive skew in the VAF distribution should be observed**. Indeed, in comparison to the VAF of all SNVs in these cells (Extended Data Fig. 2d), the shared mutations had a clear and distinct VAF of 50% (Fig. 2f). **In contrast, sister unique G > T transversions had a reduced VAF and more positive skewing in the VAF distribution (Fig. 2i), suggesting that more mutations tend to arise after sister clone separation (Fig. 2g)**. **This ancestral hypothesis is supported by the observation that the four sister genomes with higher clonal mutation sharing (Fig 2c, bottom) are derived from two singly penned clones on the same Berkeley lights chip (Extended Data Fig. 2e). This mutation sharing between clones is likely explained by a more recent common ancestor (Extended Data Fig. 2e) in agreement with previous findings in mammalian cell culture**[23]. Finally, removal of sister shared mutations increased overall signature similarity to that of SBS7a (Extended Data Fig. 2f). Overall, we show that mitotic sisters share similar overall mutational landscapes while acute UV mutations are unique to each mitotic sister. In contrast, ROS mutations can be subdivided into sister shared and sister specific accrued mutations.




202

203 **Fig. 2 | Resolving acute and chronic mutational processes.**
204 **a,** Schematic of separation experiment and nomenclature for mitotic sisters and clones. **b**, Shared
205 **mutations** between sister pairs for distinct mutational processes; each point is one sister pair. Scatter
206 plots represent the number of mutations between sisters for G > T (ROS, blue), C > T (UV, red) and
207 Other (grey). **c**, Barplot of mean overlap of mutations in percent between 7 mitotic sister pairs,
208 consisting of 14 genomes. Unique refers to mutations for each single sister cell, sister shared
209 represents mutations shared between mitotic sisters, and clone shared refers to mutations shared
210 with other clones. Points are overlaid for each individual genome. **d**, Mutation signatures for each
211 category of (c), each bar represents a specific type of mutation (identity key next to panel e) in a
212 specific trinucleotide context (96 total bars). **e**, COSMIC reference signature with highest similarity to
213 (d). **f**, Variant allele frequency (VAF) for mutations shared between mitotic sisters. (**g** and **h**) VAF as in
214 (f) but for ROS mutations (g) and UV mutations (h) unique to each sister cell, respectively. **i**,
215 Pearson's median skew of the VAF populations in f-h. Each box represents 14 measurements for the
216 specific mutation category from each sister genome.

## UV lesion retention causes multiallelic variation

While most UV mutations were resolved as mutations during the first cell division (VAF ~50%), thousands of UV mutations also had substantially suppressed VAFs, suggesting rounds of non-mutagenic replication over persistent lesions (Fig. 2h). Such lesion persistence could allow for the incorporation of distinct alternate bases from consecutive S-phases, a phenomenon termed multiallelism[14]. Succinctly, multiallelism is the observation of more than one alternative allele in sequencing reads at a single genomic position. We next sought to determine if multiallelism could be observed for mutations with a UV signature in our data.

Identifying multiallelism is challenging with the base substitution signature of UV damage because only one type of base change is commonly observed (C > T) at a single position. We therefore extended our observation to tandem CC mutations, because three possible C > T mutation combinations could occur at a dual site (CC > TT, CC > CT and CC > TC). While dinucleotide mutations represent only ~2.2% of all UV mutations across the 7 sister pairs sequenced, the well-described[24,25] CC > TT event (Extended Data Fig. 3a-c) is most common[24,25]. We identified reads fully overlapping CC > TT tandem mutation events (n = 373 sites), and discovered the presence of more than one alternative allele supported by read-level data in a single clone (Fig. 3a, left), indicating that an unrepaired UV lesion can result in the generation of both di-nucleotide and mononucleotide substitutions at the same site. **This observation is distinct from biallelic mutations seen in melanoma where both haplotypes are mutated**[26]. We confirmed and quantified this observation in a genome alignment independent manner (methods, Extended Data Fig. 3d-g).

We next asked whether both bases at each CC > TT dual mutation site have the same VAF. Differing VAFs between neighbouring cytosines at a dual mutation site would provide evidence that each mutation event was resolved in a different cell cycle (Fig. 3a, left). In contrast, the dinucleotide change occurring as a single mutational event would predict identical VAFs, resulting in a biallelic mutation (Fig. 3a, right). Identical VAFs (biallelic mutations) were observed for 70.2% of CC > TT dual mutations (Fig. 3b, black points, n=262 sites), while 29.8% of VAF pairs at tandem mutation sites were different, suggesting the presence of multi-allelism (Fig. 3b, red points). This analysis revealed that multiple alternative alleles were present at a subset of tandem mutation sites.

We next reasoned that the VAF for each alternate allele at multiallelic sites could reveal the order by which mutations were introduced. Indeed, the observed 2:1 ratio of the VAFs for the first and second most common alternate alleles was consistent with multiple error-prone replication events over an unrepaired pyrimidine dimer (Fig. 3c). The dominant sequence

mutated within these dimers is the 3’ cytosine, occurring in 85 of 111 multiallelic sites (Fig. 3d, Extended Data Fig. 3c-d). Situations where an initial 3’ cytosine mutation was followed by a double cytosine mutation in the next mitosis were approximately as common as the opposite order. In contrast, alleles with a single mutation at the 5’ cytosine occur in 22.2% of multiallelic sites. Almost no combination of single mutation events at these loci was observed, such as CC > CT followed by CC > TC. **To test whether a pyrimidine directly 3’ to the CC site might explain the 3’ bias, we measured base composition surrounding multiallelic sites and compared it to both biallelic and randomly sampled CC sites across the genome. On the contrary, pyrimidines were significantly enriched directly 5’ from CC dual mutations (Extended Data Fig. 3h), arguing that the 3’ bias is not caused by a dimer directly downstream of multiallelic sites.** Taken together, we show that multi-allelic variation at dual CC sites is observable for UV damage, and that the majority of mitotically consecutive mutation events are CC > CT and CC > TT, in either order.

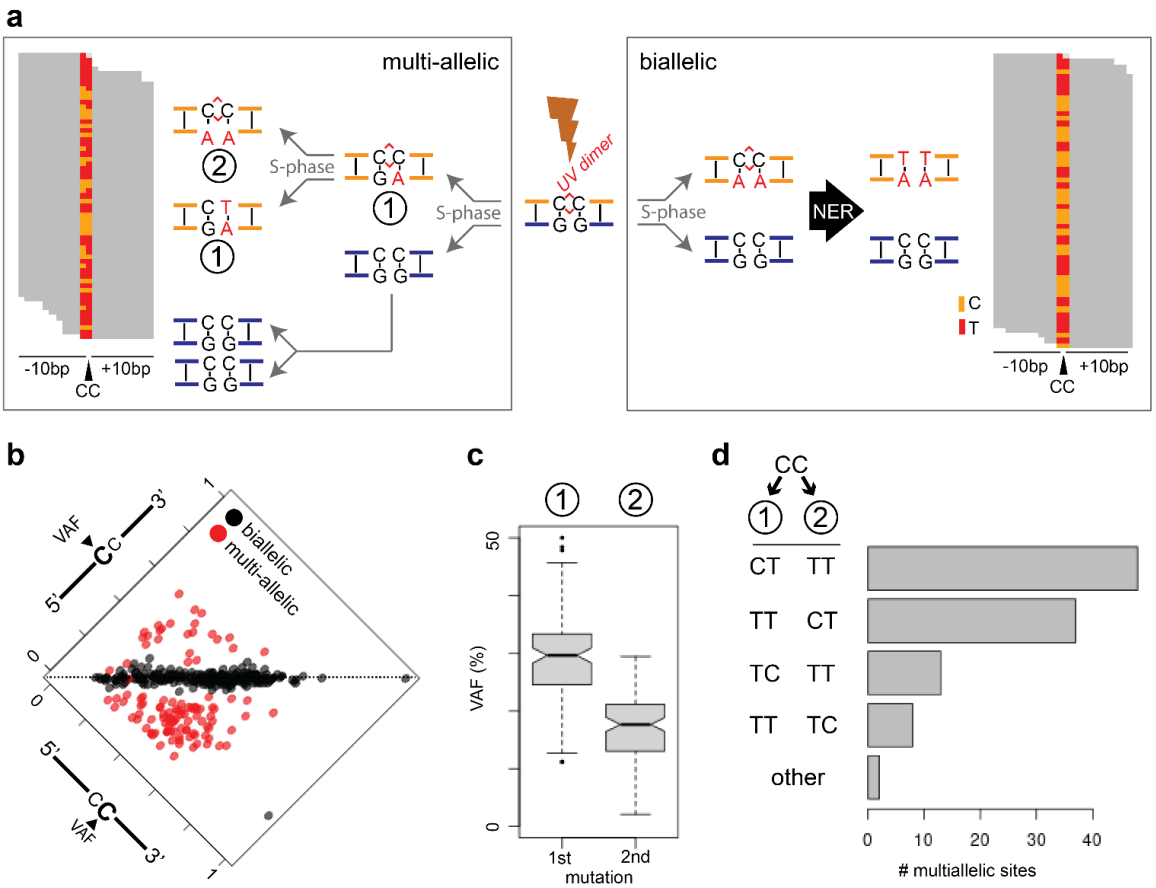



**Fig. 3 | Retained UV damage can drive multi-allelic variation.**
**a**, UV dimers (red lines/links, center) occur due to covalent linkages between consecutive bases along the DNA. Following S-phase replication, only one sister clone will inherit the lesion containing DNA duplex (gold). Left, lesion retention for two cell cycles can give rise to multi-allelic variation, in this case evidence for CC > CT mutations (1), as well as CC > TT (2). Right, schematic representation of a biallelic mutation occurring in one cell cycle, followed by NER of the lesion-containing strand. **b**, Scatter plot for VAF of the second C in a CC dual mutation (x-axis) and first C (y-axis). Black points represent instances where only one alternative allele is detected, while red points contain at least 2 alternative alleles with at least 3 unique reads supporting the presence of each allele. **c**, VAF for the most common, and second most common allele at multiallelic sites (n = 111). The second allele has half the VAF of the most common variant, suggesting it was added one cell cycle later. Boxplot elements: median is the thick middle line, quartile 1 to quartile 3 are represented by the grey box, 95% confidence interval is shown by notches, whiskers denote the minimum and maximum of non-outlier values, outlier values are shown as points. **d**, Frequency of mutation types for the first mutation event (1) and second mutation event (2). 85% of the cases are CC > CT and CC > TT.

## Transcription-associated repair dampens intragenic ROS mutation rate

We next asked how UV and ROS **mutation rates** are affected by transcriptional activity and chromatin accessibility. Transcription coupled repair (TCR) is a well-documented phenomenon[27] from prokaryotes[28] to humans[29] where DNA damage causes a transcribing RNA polymerase to stall, triggering nucleotide excision repair[30]. To this end, we first determined bulk RNA levels from the PF1 line (Fig. 4a and Extended Data Fig. 4a-b) and binned genes based on tags per million (TPM) (Fig. 4b) (bin 1 = 10.6k unexpressed genes, bins 2-4 = ~2.8k expressed genes). In agreement with previous work, active transcription significantly suppressed mutational rates ascribed to both UV[29–32] and ROS [33–353637] ($p= 1.7 \times 10^{-6}$, two tailed Mann-Whitney test; Fig. 4c), with UV showing a slightly stronger reduction. Both UV and ROS mutation rates were lowest within the most highly transcribed gene bodies where UV mutation rate was half the genomic rate, consistent with previous estimates for TCR[32] (Fig. 4c).

A hallmark of TCR is that lesions are repaired specifically on the transcript template strand[38,39]. Our data allow us to assign lesions to a specific strand as UV exposure causes a C > T transition due to pyrimidine dimer formation. More explicitly, C > T mutations in minus strand genes or G > A mutations overlapping plus strand genes would represent mutations resulting from damage on the 'template' strand (Fig. 4d). As expected, the template strand **mutation rate** in expressed genes is approximately 10% of the background level of UV-associated mutagenesis (Fig. 4e, upper), which is likely an underestimate of the repair rate given low background C > T mutations occurring in the absence of UV (Extended Data Fig. 5). Using the same rationale as for UV, with oxidative adducts of G identified as the main mutagenic damage of SBS18[22], we assigned ROS mutations to either the template or non-template strand. This analysis revealed that repair of ROS damage appeared to occur equally on both template and non-template strands (Fig. 4e, lower). This observation held whether ROS mutations occurred before or after the first division post-UV treatment (Extended Data Fig. 4f), and contrasts with prior work suggesting that strand-asymmetric repair of ROS damage may be occurring in highly transcribed genes[40]. Together, these data demonstrate that transcriptional activity correlates with reduced mutation rate from both UV and ROS, yet only UV repair is strand-asymmetric.

The reduction in ROS mutation rates within transcribed gene bodies was expected[33,35,36], but the symmetry in template and non-template mutation rates was unanticipated. This observation led us to test whether local accessibility as measured by ATAC-seq (Extended

Data Fig. 4c-e) may be partially responsible for this phenomenon. Globally, mutation rates within ATAC peak regions for UV and ROS were 50% and 32% of the genome average, respectively (Fig. 4f), in agreement with higher repair rates in open chromatin[41]. When profiling accessibility around all UV (37,756) and ROS (28,184) mutations, there was a clear decrease in open chromatin signal around both mutation types (Fig. 4g). Furthermore, ROS mutations revealed a stronger local depletion in ATAC signal, suggesting that BER is less efficient in detecting damaged bases in more highly compacted chromatin. **Furthermore, actively transcribed gene bodies tend to have greater accessibility than silent genes (Extended Data Fig. 4g) as well as increased OGG1 recruitment (Extended Data Fig 4h, previously published ChIP data[42])** which may contribute to the increased BER rate. In sum, our data suggest that the decrease in ROS mutation rates in transcribed gene bodies is not triggered by canonical TCR detection but rather, at least in part, due to increased local accessibility for repair proteins in actively transcribed regions.

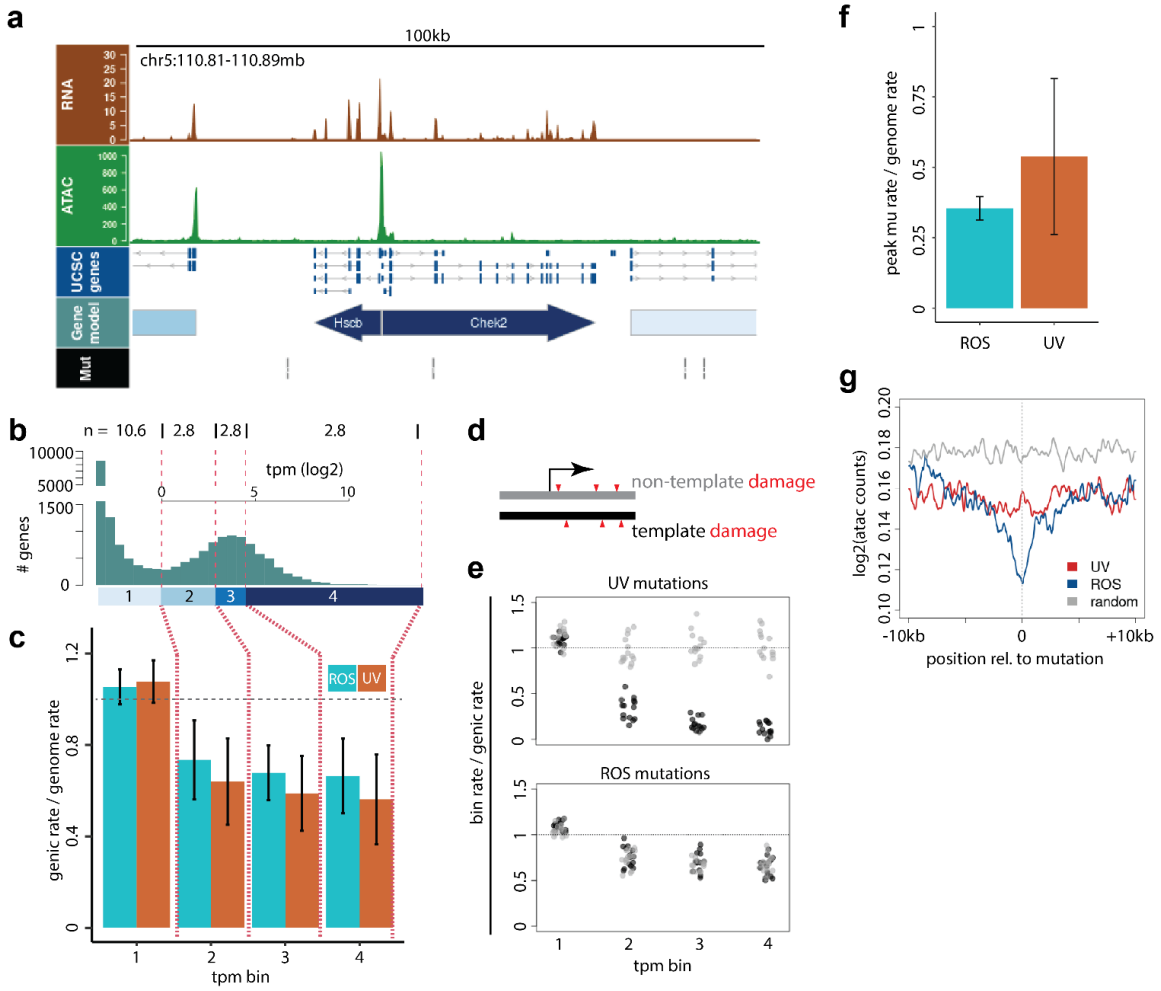
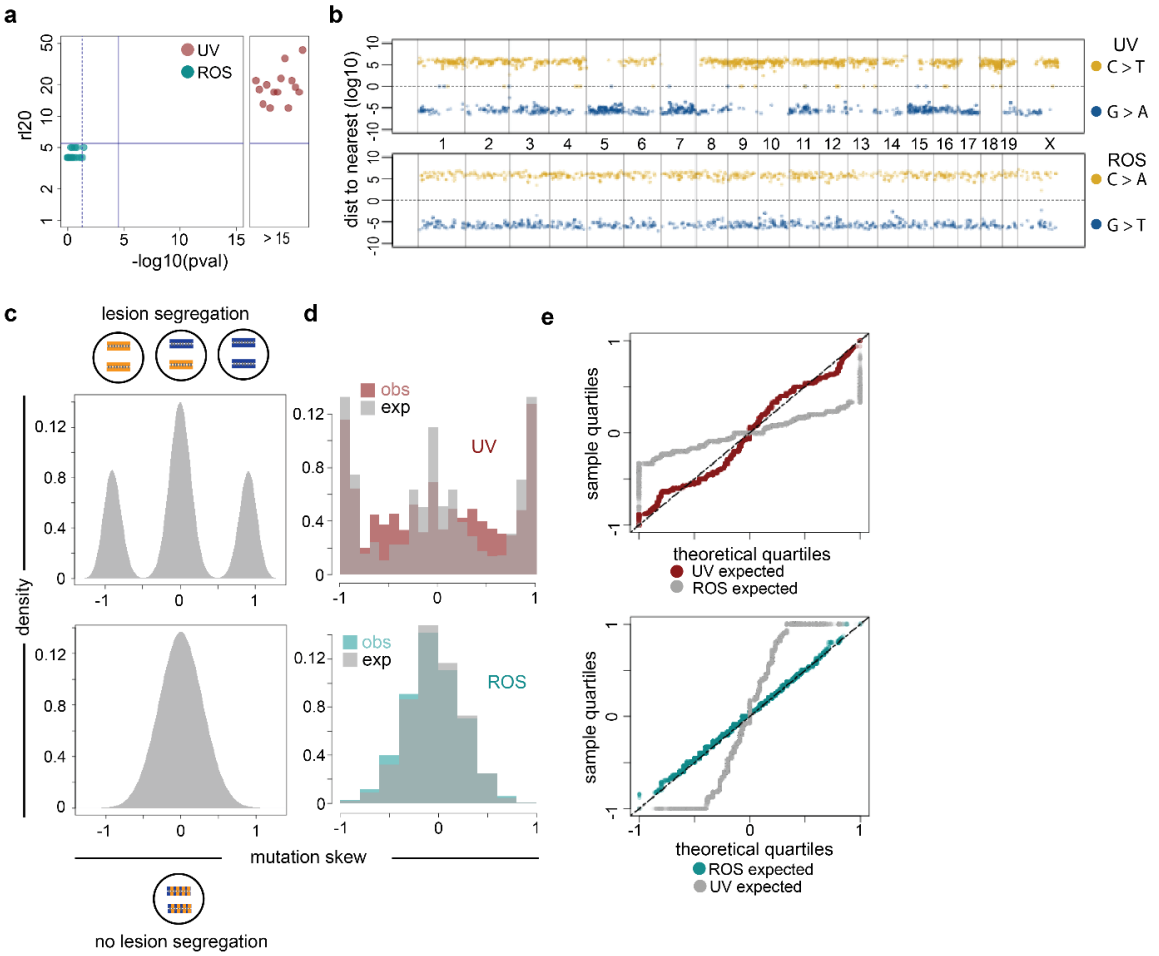



**Fig. 4 | Differential repair for UV and ROS in transcribed regions.**
**a**, Genomic screenshot of a 100 kb region of chromosome 5, from top to bottom showing RNA-seq reads (brown, coverage), ATAC-seq reads (green, coverage), UCSC gene annotation (blue), gene model based on ATAC and RNA data (light blue), and mutations in the region (black vertical lines, bottom track). Gene arrows are coloured based on their expression bin as noted in boxes under the histogram in (b). **b**, Histogram of log2(tpm) measurements for all mouse Refseq genes at least 1kb from the nearest gene and 1kb in length (19,091 total). Binning based on quantiles of tpm measurements are shown as vertical red dashed lines. Numbers above the dashes represent the number of genes (in thousands) in the respective bin. **c**, Mutation rates in gene bodies for all genes in bins outlined in (b). Y-axis represents the observed genic mutation rate for a bin divided by the expected mutation rate (calculated using all mutations). Horizontal grey dashed line represents genome average. Error bars represent 2 standard deviations from the mean of 14 genomes. **d**, Cartoon depicting stranded damage, which can be interrogated for gene bodies. **e**, Mutation rates for UV (upper) and ROS (lower), separated into template (black) and non-template (grey) mutation rates. Each point represents one sister genome and bins are as shown in (b) and (c). Y-axis is the mutation rate in the genic bin divided by the total genic mutation rate. **f**, Observed over expected mutation rates for ATAC peaks (accessible regions) across the genome. Error bars represent 2 standard deviations from the mean. Accessible regions show reduced **mutation rate**, especially for ROS. **g**, Accessibility metaplot around UV (red) and ROS (blue) mutations. Accessibility is shown as the log2 converted average number of reads at positions flanking the mutation. A 201bp smoothing window was applied to these averages.

**Mutational phasing is a unique characteristic of acute mutagenic exposure**

A prediction of the lesion segregation model is that a single, acute mutagenic exposure is required to cause mutation strand phasing across whole chromosomes. In contrast, although ROS damage is expected to be subject to lesion segregation during each cell division, its gradual accumulation over many cell generations should progressively mask the mutational asymmetry of any one generation. Our system provides an ideal landscape with which to test this hypothesis, where acute and chronic mutational processes occur in the same genome yet their signatures can be readily identified.

We separately considered the **mutation counts** and genomic distribution of SBS7 UV mutations (C > T / G > A) and SBS18 ROS mutations ( C > A / G > T) (Extended Data Fig. 5a). This confirmed the expectation that UV signature mutations exhibit pronounced chromosome-scale mutational asymmetry and significant tracts of strand-phased mutations in all UV exposed sister-clones (Fig. 5a-b; $p < 1e\text{-}15$; permutation based rl20 metric[14]). In contrast, the ROS mutations did not show mutational asymmetry in bulk analysis (Fig. 5a-b), or when partitioned into sister-shared and sister-unique mutations (Extended Data Fig. 5b). Direct comparison to Bernoulli models (Fig. 5c-d, Extended Data Fig. 5c-d) confirms the strand distribution of UV mutations is a good fit to a single burst of DNA damage followed by lesion segregation with random strand retention (Fig. 5e, upper), whereas the strand distribution of ROS mutations closely matches expectation for the random assignment of mutations to DNA strands (Fig. 5e, lower). Taken together, these results demonstrate that mutation phasing resulting from pulse UV mutagenesis is in agreement with the lesion segregation model, while chronic ROS exposure has a genomic distribution more closely fitting with a progressive accumulation over many cell generations.




379

380 **Fig. 5 | Mutational phasing via acute UV damage is established in a single cell cycle.**
381 **a**, rl20 metric used to determine if there is significant evidence for runs of a single mutation type. Light
382 blue points represent C > A or G > T mutations (ROS) analysed for each genome. Red points
383 represent C > T or G > A mutations (UV). **b**, Example of lesion segregation pattern due to UV induced
384 mutations (top) across all chromosomes of a single mitotic sister, and lack of phasing from ROS
385 induced mutagenesis (bottom) in the same sister. Reference cytosine mutations are shown as yellow
386 dots, while reference guanine residues are in blue. The y-axis represents log2(distance to nearest
387 neighbour), with G mutation distances converted to negative values to distinguish them from C
388 residues. Chromosomal boundaries are denoted by black vertical dashed lines and chromosomes
389 noted between the tracks. Horizontal dashed line represents a distance of 0. **c**, Theoretical distribution
390 of mutation phasing in lesion segregation (upper) and no phasing (lower) in 10mb genomic tiles.
391 **d,** Sampling of the distributions of C using the same number of tiles actually profiled in the data (grey),
392 and distribution of the skew from 10mb genomic tiles for all 14 genomes profiling acute UV (upper,
393 red) and chronic ROS (lower, light blue) mutagenesis. **e**, QQ plot comparing distributions of ROS, UV
394 and random sampling of the respective models in teal, red and grey respectively.

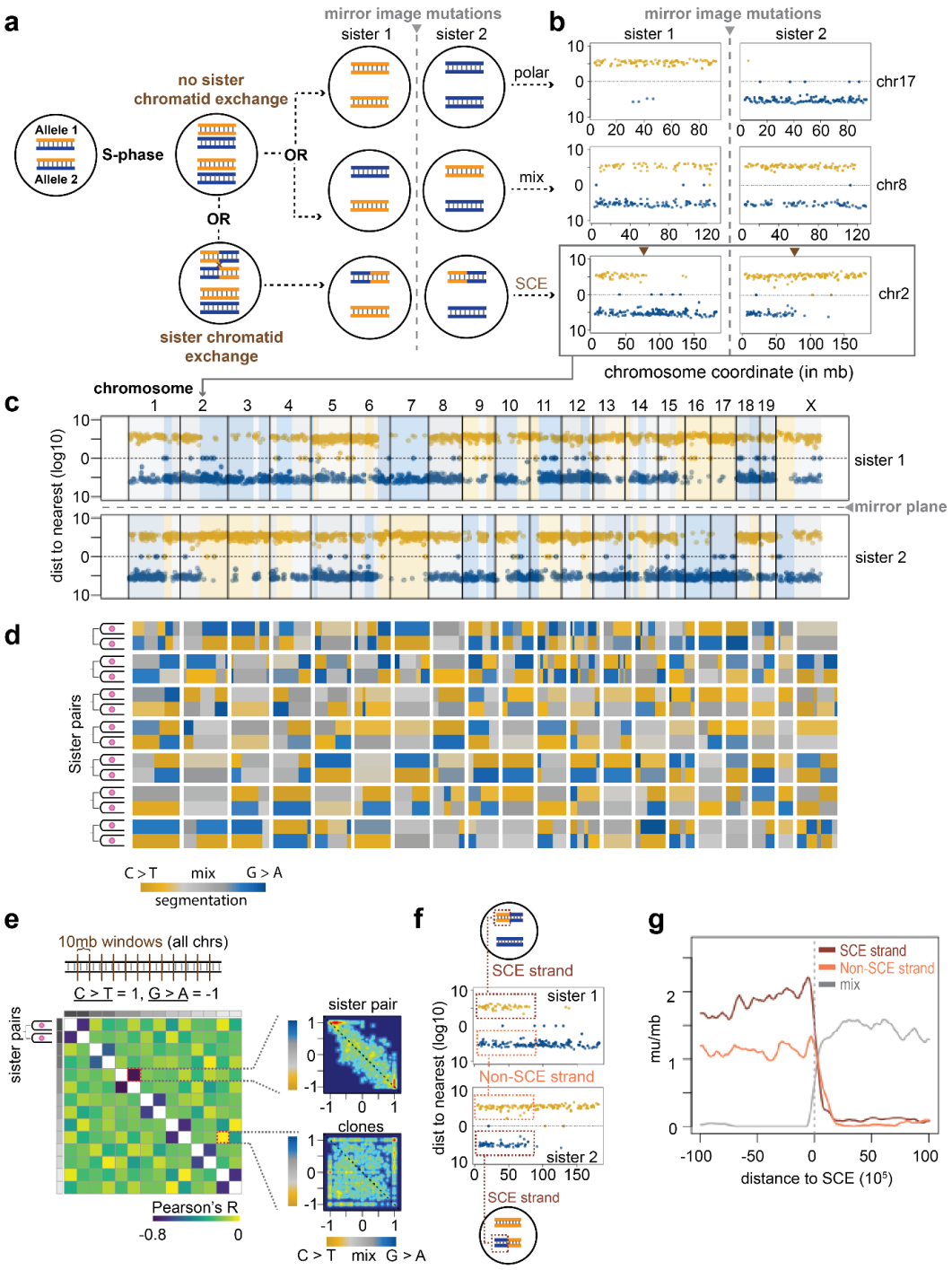
**Mitotic sisters have mirror-image mutation phasing across the genome**

In the first mitosis following a burst of mutagenic damage, the complementary lesion-containing DNA strands of each chromosome are expected to segregate into separate daughter cells (Fig. 6a). When both copies of a diploid chromosome inherit the C > T phase of lesions in mitotic sister 1, mitotic sister 2 will exhibit the complementary G > A asymmetry (Fig. 6b, chromosome 17). When mitotic sister 1 inherits a C > T and a G > A allele, the same should occur in mitotic sister 2, resulting in a mix of both mutation types across that chromosomal segment (Fig. 6b, chromosome 8). If there is a sister chromatid exchange event during that first mitosis, then approximately equivalent positions on the affected chromosome show a transition from mutation symmetry to asymmetry in both mitotic sister cells, but they will exhibit opposite mutational asymmetries (Fig. 6b, chromosome 2). Consequently, pairs of cellular clones derived from mitotic sister cells should have uniquely segmented “barcodes” of matched mutational symmetry and mirror image mutational asymmetry (Fig. 6c). All seven pairs of mitotic sisters exhibit such a matched genomic barcode with mirror image mutational asymmetry across their genomes (Fig. 6d).

We sought to quantify the significance of this mirror-image mutation phenotype by comparing mutation skew in 10mb genomic windows for all clones and sister pairs. For mitotic sisters, mirror-image phasing patterns should result in a linear negative relationship. For unrelated clones, the prediction would be a random relationship in their phasing across all windows because their UV mutations are independent damage events. This analysis revealed that on average, segment phasing patterns of sister 1 could explain 49% of the variance in phasing patterns for sister 2 from a single division event. In contrast, the variance in phasing patterns for clone A could explain less than 1% of the variance in phasing for clone B (Fig. 6e). In sum, these findings add strong support to a critical part of the lesion segregation model, namely that acute damage produces mirror-image mutation phasing patterns in mitotic sisters, independent of positive selection.

Our experimental system gave us a unique possibility to explore another specific mutation segment phenotype: switches in phasing within a single chromosome (Fig. 6a and b, bottom panels). If these switches reflect sister chromatid exchange (SCE) as previously suggested[14], reciprocity between two mitotic sisters should be observable. Put succinctly, a switch from a mixed segment to a pure C > T segment in mitotic sister one, should be reflected as a mixed segment switching to a pure G > A segment in mitotic sister 2 (Fig. 6a, bottom). This segment switching was evident at 130 total positions across all samples, with approximately one crossover event per chromosome (mean 0.92; Fig. 6b-d). As both normal lymphoblast and fibroblast cells from healthy human patients are estimated to have roughly

5 SCE events per mitosis[43], our 4-fold elevation suggests UV damage is responsible for this increase in SCE, consistent with previous studies[44–46]. There was a very modest increase in mutation density around SCE sites (Extended Data Fig. 6c), suggesting UV damage is locally correlated with and may trigger an SCE event. To probe this further, we separated mutations based on whether they originated from the recombined or non-recombined strand (Fig. 6f). By resolving mutations to the recombined strand, there was a clear trend for increased mutation density on the strand undergoing recombination (Fig. 6g). In conclusion, elevated local damage rates on one strand of DNA are more likely to trigger an SCE event during the first mitotic cycle.




(legend on next page)

**Fig. 6 | Mitotic sisters have mirror-image mutation distributions across the genome.** **a**, Model of lesion segregation creating mirror-image mutation patterns between mitotic sisters. Replication of damaged DNA creates stranded mutations, and sister cells from the subsequent division inherit either: only one type of strand (upper), one of each strand (middle), or can undergo SCE and switch from a mixed to a phased segment (lower). **b**, Representative examples for each strand inheritance type depicted in (a). **c**, Scatterplot of C (yellow) and G (blue) mutations across all chromosomes for one mitotic sister pair. Lightly coloured yellow/blue/grey background represents segmentation from changepoint analysis. **d**, Segmentation heatmap for all 7 pairs (14 total genomes). Colour legend of phasing noted below heatmap, mitotic sisters are adjacent to each other with a white gap between clones. **e**, Test for mirror image mutation patterns. Heatmap represents the correlation coefficient (Pearson's R) for mutation skew between mitotic sisters and clones. Smooth scatterplots to the right represent skew correlations between an example of mitotic sisters (top) and clones (bottom). **f**, Diagram demonstrating how recombined strand mutations are determined. **g**, Metaplot of mutations/megabase (mu/mb) for recombined and non-recombinant strands (brown and gold respectively), as well as mutation density for regions where one strand isn't uniquely assignable. Window represents 20mb flanking the SCE site shown at 0 (vertical grey line). Individual points represent smoothed 1mb sliding windows with a 100kb step size.

**F1 mice tumourigenesis resolves DNA damage to single haplotypes**

In DEN-induced liver tumours from inbred male mice, the vast majority (~95%) of X chromosome mutations are phased[14]. We therefore reasoned that if we could assign mutations to a single haplotype for autosomes, similar mutation phasing could be reproduced across the entire genome. An ideal model for this is the F1 progeny from a mouse cross where the parents have substantial germline single nucleotide polymorphisms (SNPs), which would allow us to assign mutations to a single allele. The C3H and *Mus castaneus* (CAST) mouse sub-species produce viable offspring, and their genomes differ by 20 million SNPs[47]. Given the SNP genomic distribution, paired-end 100bp sequencing allows 60.2% of all unique reads to be assigned to one haplotype on average (Extended Data Fig. 7a-b, methods).

We injected two F1 mice at P15 from a CAST and C3H cross with DEN and then isolated 6 tumours 30 weeks later (Fig. 7a, methods). We carried out WGS (>= 20x mean, Extended Data Fig. 7c) of these tumours and used N-masking[48] to map reads in an unbiased manner to a modified C57BL/6 reference genome. Acute DEN treatment predominantly induced mutations by damaging T bases (38.5% T, 37.3% A, 12.4% C, 11.8% G) with **total mutations**, signature, and stable chromosome copy number as documented previously[14] (Fig. 7b, Extended Data Fig. 7d-e). Haplotype agnostic mutations reproduced a lesion segregation phenotype identical to that found in the inbred parental strains (Fig. 7c), including mixed mutational phases (red-framed grey boxes in Fig. 7c).

By exploiting the co-occurrence of germline SNPs with mutations on the same read, we could phase 91.9% of mutations to either the C3H or CAST genomes (Fig. 7d-e, Extended Data Fig. 7d). We used the read mapping accuracy to the single copy of the X chromosome to estimate our false assignment rate, which is approximately 0.27% (sd = 0.25%). Once mutations were accurately assigned by haplotype, genomic segments with mixed phasing become resolved into two perfectly phased component alleles (Fig. 7f). For example, in the absence of haplotype assignment, Chromosome 2 of a representative tumour is a mixture of T > N and A > N mutations. Resolving the mutations by haplotype reveals that the CAST allele contains the vast majority of T > N mutations while the C3H allele has predominantly A > N mutations. Switches in mutational phasing are now clearly resolved to a single, recombined allele (Fig. 7f, chr3), compatible with fixation of stranded mutations.

We next segmented the genomes using changepoint analysis as done for UV above. Once the mutations were accurately assigned to a single allele, the mutational landscape of each autosome was virtually identical to the X chromosome (Fig. 7g). In total, 95.6% (sd = 0.92%) of haplotype assigned autosomal mutations are in agreement with their phase, in close

agreement with the single copy X chromosome at 94.9% (Extended Data Fig. 7f). This relationship held true for 5 additional tumours sequenced from two independent mice (Fig. 7h). Taken together, resolving mutations to single alleles demonstrates complete mutational phasing across the genome to a level nearly identical to the single copy X chromosome.

Across the genome we would predict that mutation phasing in one tumour is an independent event. We tested this independence by asking whether the phasing of one tumour can predict the phasing of another by comparing T <> A phasing in 10mb genome tiles across tumour genomes. In agreement with a completely independent origin, bin phases revealed an average Pearson's correlation of 0.07 when compared between tumours (Extended Data Fig. 7g right, h-I). While C > N and G > N mutations make up only ~20% of the data, these mutations are also phased in a manner analogous to T <> A phasing (Extended Data Fig. 7g, left). More specifically, genomic segments containing more T than A mutations also contain more C than G. As T bases have been shown to be the predominant mutation type caused by DEN, this suggests cytosines are the second major source of mutagenic base damage. This is in agreement with most damage occurring on pyrimidine bases and provides an independent representation of lesion segregation at the haplotype level. These data reveal that both species of DEN mutations are completely phased across the genome. Additionally, it demonstrates that intrachromosomal switches are indeed sister chromatid recombination events, as the mutation events from the opposite strand are always (140/140) juxtaposed within a single haplotype.

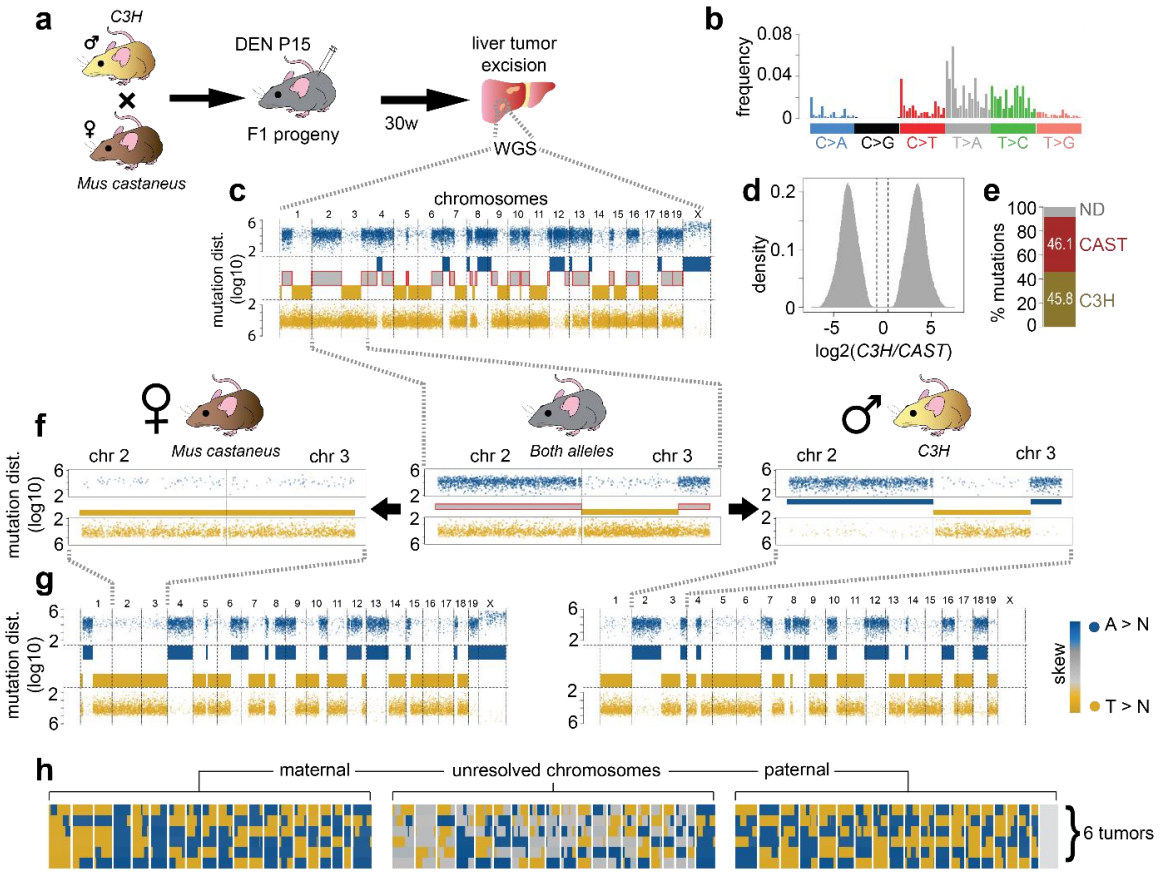



**Fig. 7 | Allelic resolution of lesion segregation *in vivo*.**
**a**, Experimental design: a *Mus musculus castaneus* femalewas crossed with a C3H male. Two 15 day old (P15) progeny were injected with DEN and 3 tumours from each mouse were isolated 30 weeks later for whole genome sequencing. **b**, Mutation signature of tumour mutations. **c**, Lesion segregation plot of a representative tumour with haplotype agnostic data. Segmentation boxes in the middle reflect if the chromosomal segment has mixed (grey with red outline) or phased mutations. Reference A mutations are shown in blue and reference T mutations are shown in gold. **d**, Frequency distribution of species specific delta read counts for all mutations in the sequenced population with at least 2 unique haplotype assignable reads. Grey vertical lines denote the cutoff for species specificity. **e**, Proportion of all mutations assigned to the C3H haplotype (gold), *Mus castaneus* (brown) or undetermined (grey). Actual percentages for haplotype specific reads are shown in the respective bars. **f**, Chromosomes 2 and 3 from the tumour depicted in (c). Middle panel shows mutation distribution (points) and segmentation (bars in the middle) for haplotype agnostic data. *Mus castaneus* (left) and *C3H* (right) show the same plots but after resolving haplotype. Note that the mixed segments become completely resolved into phased mutation stretches. **g**, Plotting the mutations after assignment to one haplotype. **h**, Segmentation before (middle) and after (left and right) haplotype resolution of mutations. Heatmap represents tumour segmentation phasing for the tumour in panel (c) and 5 additional tumours (6 in total).

## Discussion

Cancer genomes are formed by the long-term accumulation of changes arising from multiple mechanisms, often operating in parallel among many competing clones[49–51]. Here, we describe a methodology to identify and disentangle two distinct genomic mutation processes: acute UV and chronic ROS, active in single cells. We interrogated mutations in mammalian mitotic sister cell populations following a single division post UV exposure. Our data show that UV lesions can be retained for more than one cell cycle, driving multi-allelic variation in the context of CC dinucleotides. We furthermore demonstrate that mitotic sisters inherit stranded mutations after a single mitosis following acute damage, and present mirror-image mutation phasing. Mutation phasing follows a random inheritance of strands between sisters, analogous to a set of Bernoulli trials as predicted by the lesion segregation model[14]. Mitotic sisters have reciprocal intra-chromosomal switches in phasing, thus providing the first direct sequence-level evidence of individual sister chromatid exchanges. In contrast to acute base-specific damage, chronic ROS damage does not create phased mutation profiles and is subject to transcription-associated repair. Lastly, we demonstrate at a single chromatid level that all mutations are phased, by mapping the polarisation changes occurring in tumours from an F1 mouse cross.

Our repurposing of a microfluidics system provides several unique advantages. Automated imaging and physically separating mammalian cells allows us to account for several variables, including cell cycle state, cell division number, and expansion rate. Furthermore, cell lines can be genetically modified to offer flexibility in experimental design in combination with a diverse array of mutagens. Similarly, the use of F1 animals from inbred mouse subspecies minimised genetic heterogeneity, while still providing a robust means of haplotype-specific interrogation. Two aspects of our model systems limit their direct applicability. First, the Berkeley Lights platform works exclusively with non-adherent cells. Second, inbred mice and cell lines do not perfectly reflect many aspects of human tumours, including cell cycle time, population demographics, and mutagenic exposure. Despite these limitations, **our mechanistic insights are relevant to human cancers because: (1) similar phased mutational profiles are seen in patient data[14,52], (2) the mutation signatures in our system are incredibly similar to relevant tumour tissues (e.g. skin and UV), (3) we recapitulate reciprocal mutation phasing of the 'lost sister' predicted at transformation and clonal expansion. In addition, we also demonstrate that multi-allelic variation seen in hundreds of mouse tumours is observable in a non-cancerous setting. These findings further underscore the major role lesion segregation plays in establishing genomic mutational patterns observed in cancer.**

Evolution of the cancer genome is widely thought to be a Darwinian process[53] wherein random mutagenesis followed by selective pressure determines clonal expansion. Our approach exploits microfluidics to mechanically separate sister cells from a single division to decouple mutational pressure from positive selection, thus enabling the fate tracking of the forward and reverse strands of both alleles. We establish that acute UV mutagenesis in our system is predominantly resolved into mutations during DNA synthesis over damaged bases. The resulting phased mutations can provide information regarding when a cell transformed in relation to the acute damage event[14]. This mutation pattern parallels the Big Bang theory[54] of tumour inception, where clonal expansion of a primordial cell affords a shared landscape of ancestral mutations. In our case, subsequent populations from the initial sister cells continue to accumulate ROS damage, resulting in a heterogeneous mixture of mutations in the sequenced population, akin to a punctuated equilibrium concept[55]. Future studies of cancer genome phylogenies could take mutational phasing into account as a fingerprint of the originally transformed cell.

Our data suggests that the repair rate of ROS is primarily influenced by local chromatin accessibility, which is elevated in highly transcribed regions. This reinforces prior reports implicating the FACT chromatin remodelling complex[56], transcriptional activity[37,57] and damage position in relation to the nucleosome[58] as contributing factors to 8-oxoG repair. It has been hotly debated whether TCR repairs non-bulky lesions in transcribed regions[59–61]. Our system has unique features that give novel resolution on how ROS damage is repaired in transcribed gene bodies. First, we can assign base damage to a specific strand, as the dominant C > A mutation likely arises from oxidised G bases. Second, both TCR repair of UV damage and transcription associated repair for ROS are seen to be active in the same cell in thousands of gene bodies. Third, we can distinguish between ROS occurring before and after UV treatment, revealing that transcription-associated ROS repair is unperturbed by UV exposure. In summary, our data support a model where oxidative damage to guanosines is repaired better in transcribed gene bodies, but strand-agnostically. Although stranded BER repair has been described in yeast[35] and human[40], and for other oxidised bases[36], this strandedness does not appear to shape the repair of 8-oxo-G, which is one of the most common oxidative lesion occurring in mammalian genomes[8].

**The lesion segregation mechanism is defined by persistent DNA damage[14]. While eukaryotic cells clearly possess repair mechanisms to resolve these lesions[6], replication over altered bases seems to be driving most of the mutational process for acute genotoxins[14,15], regardless of transformation. Resulting genomic signatures of these mutations are ultimately the result of damaged base identities and several other**

**factors[62]. While the precise reasons for why a particular damage event may persist remain to be fully understood, several key factors have been implicated in this process. These include chromatin accessibility[63], repair pathway efficiency, total damage events at any one time, transcriptional state[64] and sequence context. One key missing piece of information is assigning an accurate quantitative estimate to the number of specific DNA lesions following a genotoxic event. Future studies will be required to help bridge the gap in our understanding between the total extent of damage following mutagenic exposure and the resulting mutational landscape.**

Our work establishes a novel approach, decoupling mutagenesis from selection, to study how DNA damage shapes the mammalian genome. The framework we present can be flexibly applied to separate otherwise-confounding mutational processes co-occurring in cells for mechanistic analysis.

## Data Availability

Fastq files for the WGS, RNA and ATAC-seq produced for this manuscript can be downloaded from Sequence Read Archive (SRA) under the accession number PRJNA934746. Processed files including mutation calls, TPM counts and ATAC peaks used in the analysis have been deposited in GEO under the accession GSE230579. HEK293 Flag-OGG1 ChIP-seq data was downloaded from the GEO accession GSE89017 while HEK293 Ribo-Zero total RNA was obtained from GEO accession GSE76496. Custom R scripts used for analysis are available upon request. Reviewer access:
SRA: https://dataview.ncbi.nlm.nih.gov/object/PRJNA934746?reviewer=2g34dbljblta1gcgr19obutt76
GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE230579
GEO token: odopmccwjlgplir

## Acknowledgements

We thank Frances Connor from the University of Cambridge, UK for assistance in the F1 mouse DEN experiments and Angelika Feldmann from the Deutsches Krebsforschungszentrum for critical input on the manuscript.



## Author contributions

P.A.G., M.S.T and D.T.O. conceived the study and designed experiments. P.A.G., H.B. and A.S. performed UV mutagenesis, cell-splitting experiments, WGS library preparation, ATAC and RNA-seq experiments. S.J.A. and C.E. performed mouse experiments. M.B. performed supporting experiments. P.A.G. and M.S.T designed and implemented computational analysis with D.T.O. support. D.T.O. and M.S.T. supervised the work. P.A.G., S.J.A., D.T.O. and M.S.T. wrote the manuscript with input from all other authors.

## Funding

This work was supported by the Deutsches Krebsforschungszentrum (DKFZ), the European Research Council (Grant ID:788937) to D.T.O., with additional support from the MRC Human Genetics Unit core funding programme grant MC_UU_00007/11 to M.S.T.. S.J.A. is supported by core funding from the MRC Toxicology Unit (RG94521).

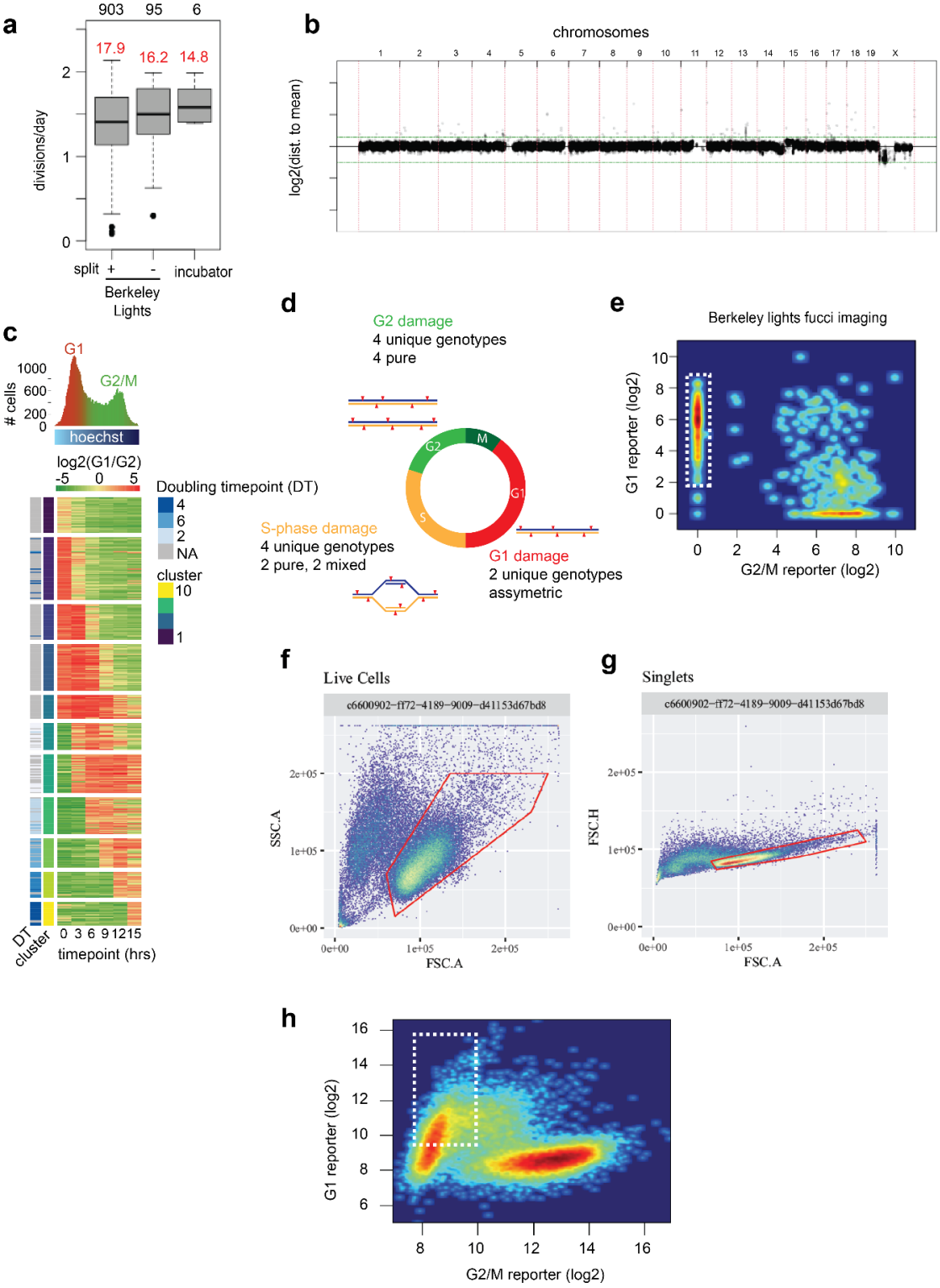
## Competing interests

The authors declare no competing interests.

## Materials & Correspondence

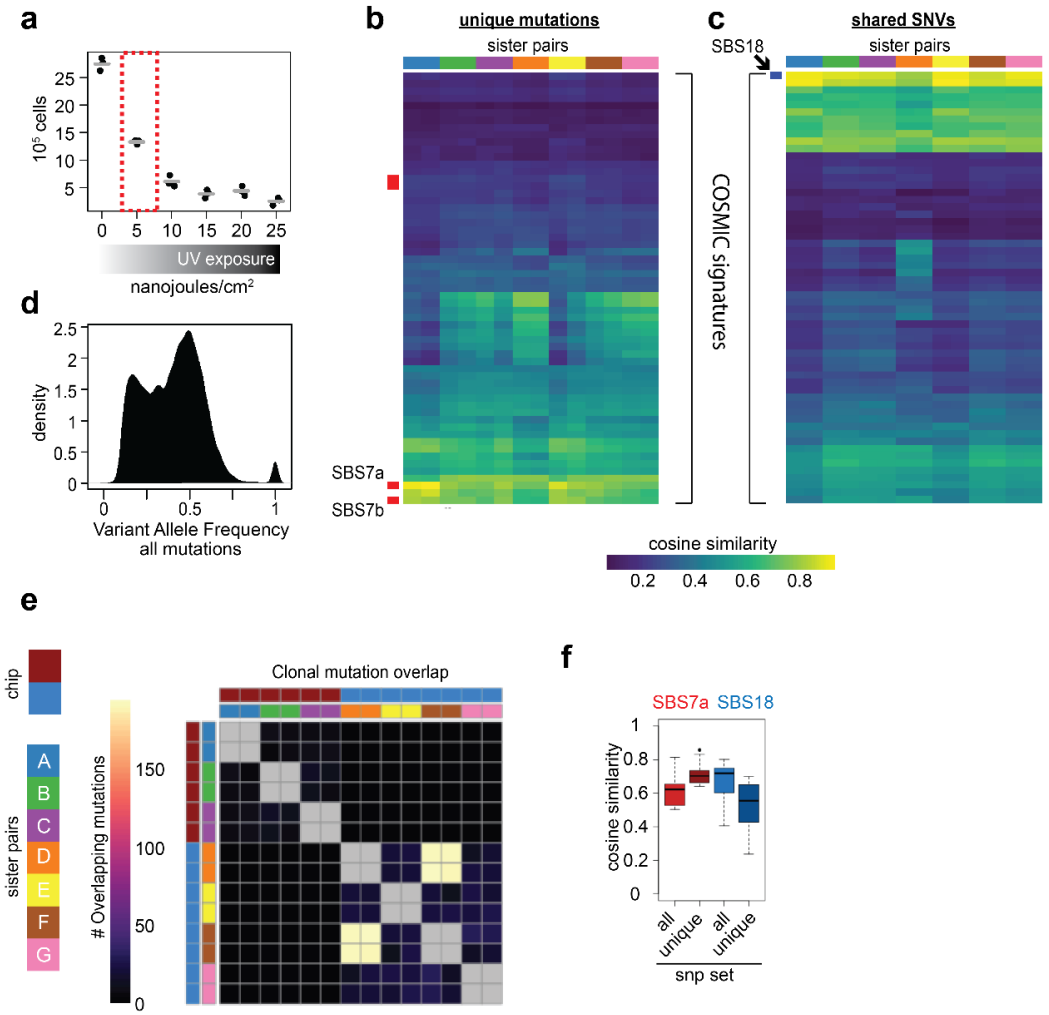
Correspondence and materials requests should be addressed to Martin S. Taylor martin.taylor@ed.ac.uk and Duncan T. Odom d.odom@dkfz-heidelberg.de

## Supplemental Figures

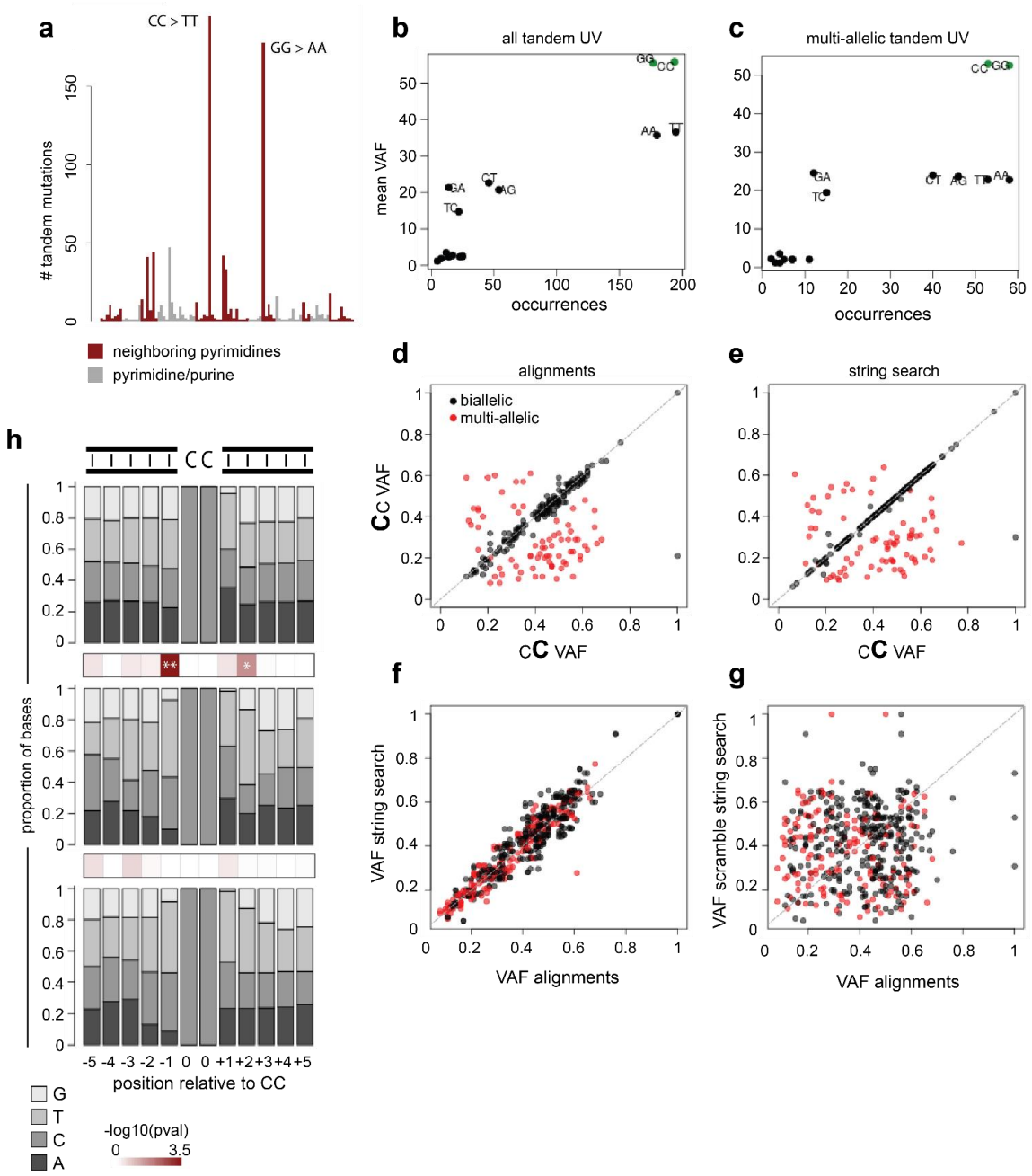



(legend on next page)

**Supplementary Fig. 1 | Genomic stability and cell cycle determination in the PF1 line.** **a**, Divisions per day on the Berkeley Lights instrument and in cell culture. Rates for cells split after a single mitosis are noted. Numbers above boxes represent the total number of rates measured, for Berkeley Lights this reflects pens, for incubator cells it is individual wells. Red numbers represent the mean doubling time in hours. Boxplot elements are as described in figure 3c, albeit without notches. Proliferation measurements from the Berkeley lights platform were taken on cells that proliferated post UV treatment. **b**, Copy Number analysis, showing diploid content for most of the genome. Reads were counted in 10kb bins, and the y-axis represents log2(distance to mean across all bins). Red vertical lines demarcate chromosome boundaries, and green horizontal lines represent counts expected for a single copy number gain or loss. **c**, Top: DNA content (x-axis = hoechst intensity) as a function of green and red fluorescence. Histogram bins have been coloured by scaled log2(red/green) for each cell. Bottom: FUCCI fluorophores imaging over time. 357 cells on one chip were hierarchically clustered based on FUCCI across all 6 timepoints. Colour scale is noted above and is identical to the hoechst histogram colour scheme. Doubling time point is indicated by the second annotation column scaling from early replicating timepoints (grey) to later replication timepoints (dark blue). Timepoints are 3 hour intervals and noted below each column. **d**, Fluorophore signal per cell cycle and theoretical effect of ploidy on mutation patterns for pulse mutagenesis (UV). Cells in S-phase would have intermittent lesion segregation patterns, while cells with duplicated DNA (G2/M) would not show lesion segregation patterns after a single mitosis. **e**, Scatter of scaled G1 (red) and G2/M (green) signal for 1120 cells measured on the Berkeley Lights platform. Cells to split are indicated by the white dashed box. **f**, Gating Live cells with FSC-area by SSC-height. **g**, Singlet determination by FSC-area by FSC-height. **h**, Fluorophore intensity for G1 fluorophore (FITC) and G2/M fluorophore (yellow-green laser). White box denotes FITC positive cells that were single-cell sorted to establish the PF1 line.

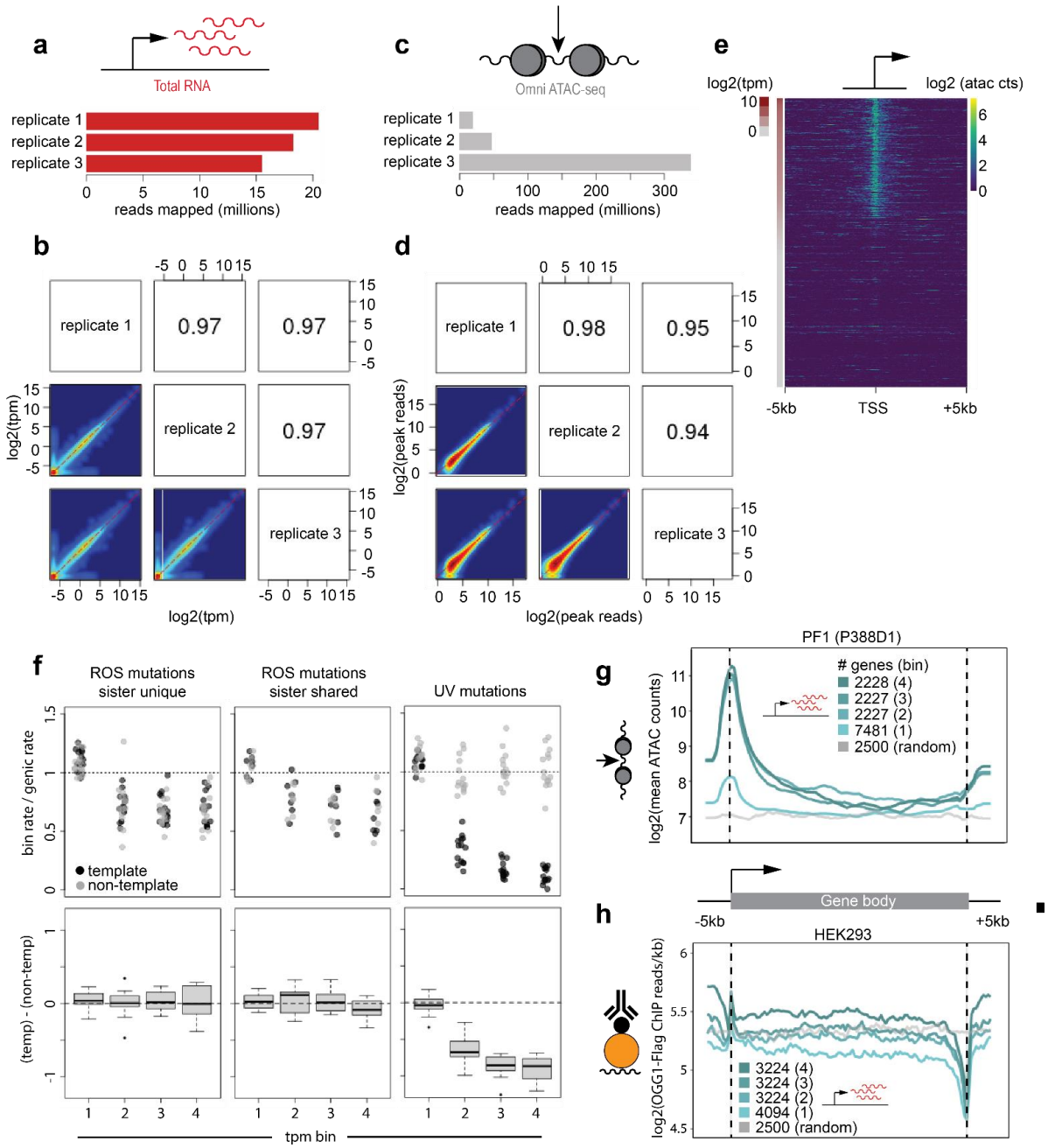



**Supplementary Fig. 2 | UV treatment, mutation signature determination and QC of WGS data.** **a**, Determination of UV treatment intensity and effect on viability 3 days post exposure. Intensity used to induce UV mutations is denoted by the red dashed box. Y-axis represents viable cells, X-axis represents UV intensity in nanojoules/cm². **b**, Cosine similarity between sister unique mutations and the 60 annotated SBS signatures in the COSMIC database. The four signatures attributed to UV damage are noted by the row annotation in red. Mitotic sister pairs are noted by the column annotation at the top of the heatmap. **c**, Same as in C but for SNVs shared between mitotic sisters. ROS signature is noted by the annotation row in blue. **d**, Distribution of VAF for all mutations across all sisters. **e**, Heatmap displaying the number of overlapping mutations between clones (see Fig 2c, bottom), sister shared mutations are greyed out to demonstrate clonal sharing. Chip annotation bar depicts two independent Berkeley lights chips (independent splitting experiments). Sister pairs are as shown in **b** and **c**. Sisters with higher clonal mutation overlap stem from two individual clones from a single splitting experiment. **f**, Similarity of mutation signatures to UV (SBS7a, red) and ROS (SBS18). Each pair of box plots represents the similarity of all 14 individual sisters to the respective signature when all mutations are considered (light red and light blue) or only mutations unique to that sister (dark red and dark blue). Boxplot elements are as described in figure 3c, albeit without notches.




(legend on next page)

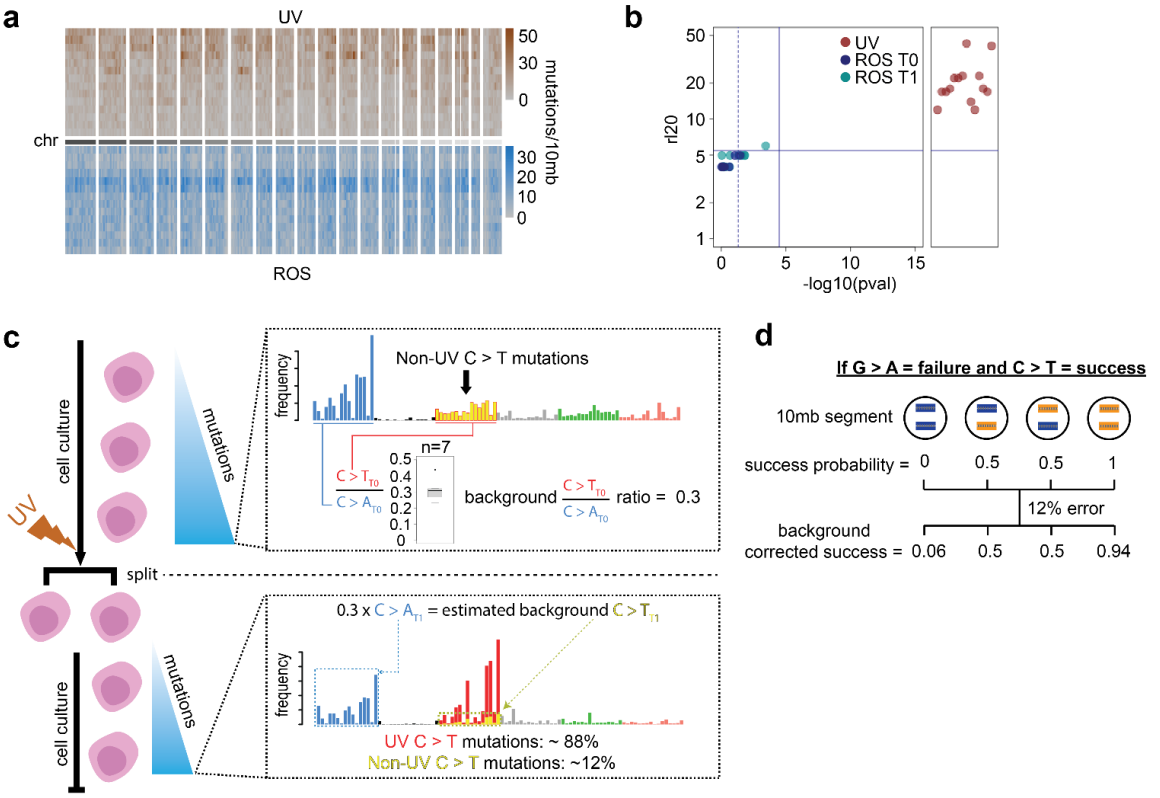
**Supplementary Fig. 3 | Unrepaired UV lesions can create multi-allelic variation.**
**a**, Bar plot of counts for all tandem mutation identities (929 total dual mutations, 94 total categories). Red bars represent reference alleles where two pyrimidines are adjacent to each other (ie, CC, CT, etc.) while grey bars represent purine/pyrimidine hybrids (ex CG). Canonical CC > TT UV mutations, and its reverse complement are noted (373 in total). **b**, Scatter plot depicting number of times an allele was seen at a tandem mutation, and the average VAF for that allele when it is detected. **c**, Same as in (b) but only for multi-allelic tandem mutations. **d**, Comparison of VAFs for each cytosine in a tandem mutation, calculated from reads where information for both bases is contained. Note how the second CC in is more likely to be a mutation seen in multi-allelic UV sites. 237 of the 373 sites are represented after filtering for dual mutations that can be interrogated using the string searching approach (see methods). **e**, The same as in (d), but instead of using alignment information directly, sequences representing each UV mutation allele were detected by string searching raw sequences in the corresponding genome (see methods). **f**, Scatter plot of VAFs calculated from alignments (d) with VAFs calculated from string searching sequences from raw data (e). **g**, Same as in (f), except the identity of sequences for VAFs calculated from string searching (Y axis in f) have been scrambled to depict no relationship. **h**, Sequence context surrounding CC dual mutations. Surrounding base identities have been converted to proportions. The identity of each base is shown in the key at bottom. The heatmap between any two bar plots displays the -log10(p-value) from a chi-square test of proportions between the two base distributions at that position. Positions with a p-value below 0.005 are noted by '**', while positional p-values below 0.05 are represented with '*'.(*top*) The average of 100 random sampling events of 111 CC dinucleotides in the mouse genome. (*middle*) Identities of surrounding bases for all multi-allelic sites (n = 111). (*bottom*) Identities of surrounding bases for all biallelic sites (n = 272).



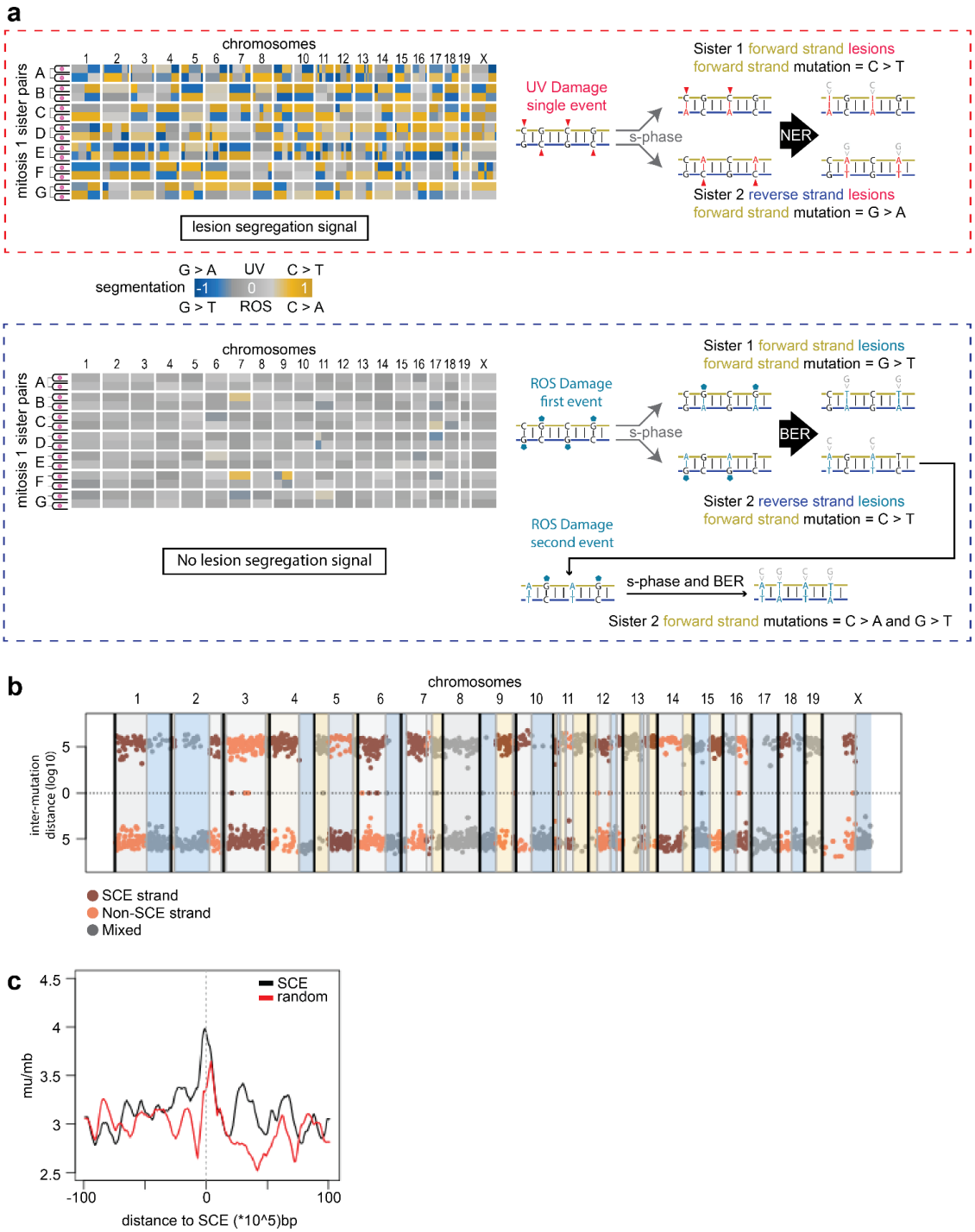

(legend on next page)

**Supplementary Fig. 4 | Transcriptome and accessibility profiling in the PF1 cell line.**

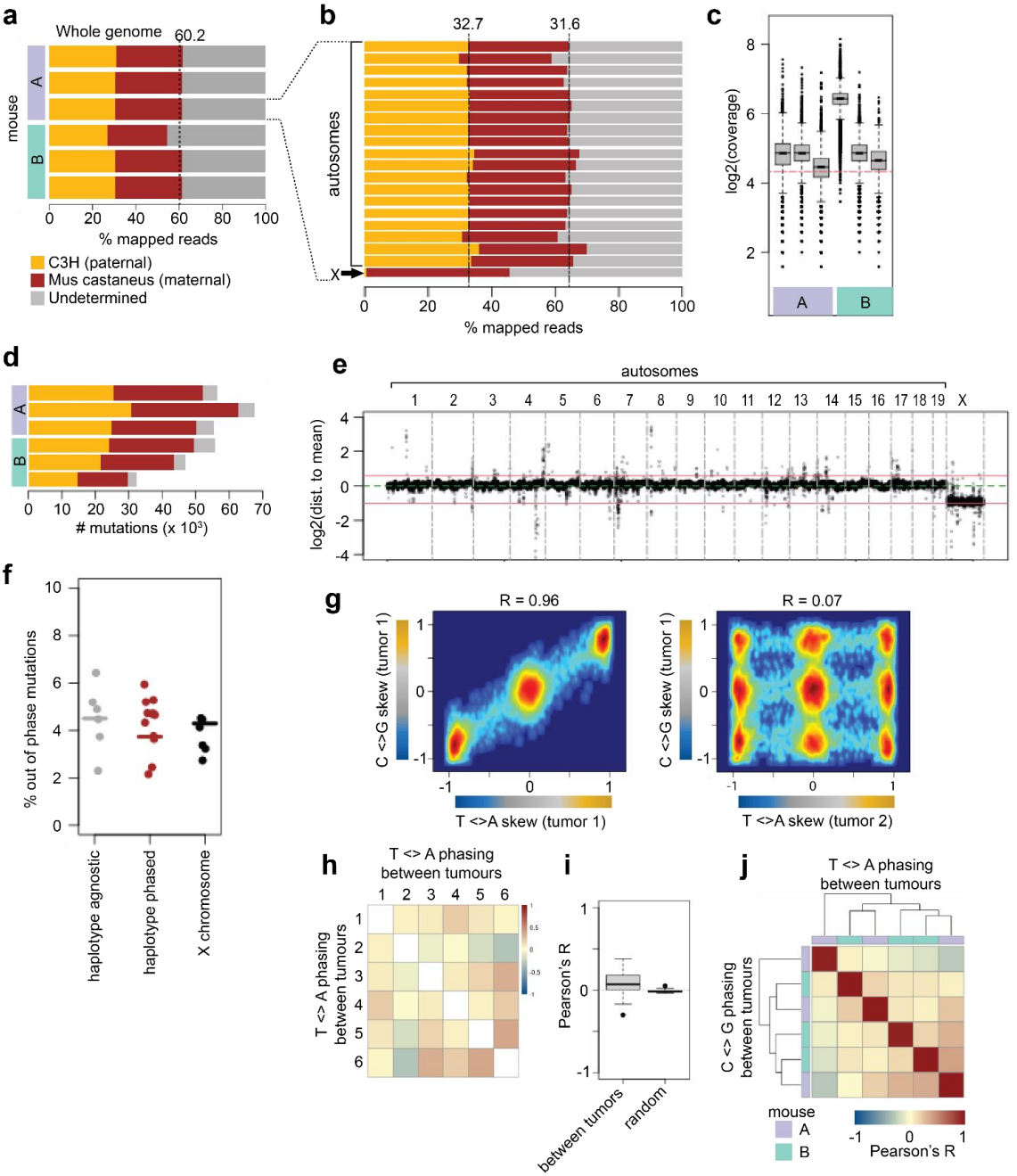
**a**, Number of unique reads mapped (in millions) per library for triplicate total RNA-seq replicates. **b**, Pairwise scatters of RNA measurements for all annotated mouse genes. Values are shown as log2(tpm + 0.01). Upper panels represent the Pearson correlation coefficient for the respective scatter. **c**, Same as in A but for unique reads in triplicate ATAC-seq samples. **d**, Pairwise scatter of reads in merged peaks across 3 ATAC-seq replicates. Axes represent log2(reads per kb + 1). **e**, Heatmap of ATAC-seq counts in a 10kb window surrounding transcription start sites. Rows are ordered by tpm from RNAseq data in A, and represented as the annotation column to the left of the heatmap. **f**, Upper: stripchart of template (black points) and non-template (grey points) mutation rates divided by the total genic mutation rate for all 14 genomes. Point clusters represent genic bins as described in figure 4. From left to right, ROS mutations unique to each sister cell (14 points per bin), ROS mutations shared between sisters (7 points per bin), and UV mutations. Lower: Boxplot of template - non-template rate for all 14 genomes, considering the mutations as for the stripchart panels above. **g**, Average ATAC signal over gene bodies. Genes at least 5kb in length were first binned based on TPM from low (1, light blue) to high (4, dark blue), and additionally 2500 coordinate shuffled gene positions (grey) were taken as a negative control. Gene bodies were divided into 100 tiles. Additionally, a window of 5kb was added flanking the TSS and TTS. Reads were counted in all genic tiles, summed by genic bin, and scaled to reads per kb of genomic representation. **h**, Genic signal for Flag-OGG1 ChIP data in HEK293 cells[42]. Transcriptional binning and gen body tiling were performed as in panel g, and numbers of genes per bin are shown as in g.




**Supplementary Fig. 5 | Mutational phasing for UV damage in mitotic sisters.**
**a**, **Mutation density** for UV (upper) and ROS (lower) across all chromosomes. Heatmaps represent 10mb sliding genomic windows with a 1 megabase step. **Mutation density** in windows is represented as mutations per 10mb window. Each row is a single sister genome and rows are sorted by **total UV mutation counts** from highest to lowest. **b**, rl20 analysis as in Fig. 4a, but distinguishing between ROS mutations shared between mitotic sisters at time point 0 (dark blue) and ROS mutations unique to each individual sister, acquired after the first division (light blue). Red dots represent UV mutations. **c**, Schematic depicting determination of background C > T mutation when modelling phasing for UV damage using Bernoulli trials. Upper box: Cells accumulate Non-UV C > T mutations in culture (yellow bars with red border) before UV damage. The ratio of C > T/ C > A for all 7 sister pairs is shown in the boxplot inset (Boxplot elements are as described in figure 3c, albeit without notches), and the average is ~ 0.3. Lower box: Total C > A **mutation counts** unique to each sister is multiplied by 0.3 to estimate the amount of total background C > T mutations (overlaid yellow bars). This background C > T estimate is then divided by the total C > T **mutation counts** to estimate the error adjustment for phased Bernoulli trials, which has a mean of 11.8% and ranges from 5% to 16% depending on UV **total mutations**. **d**, Error rate is used to adjust success or failure probability for completely phased segments. An example error rate of 12% is shown, whereby each phased segment probability is adjusted by 6%, as half of these background mutations will be randomly out of phase.




**Supplementary Fig. 6 | Acute, single pulse damage reveals asymmetric lesion segregation patterns between mitotic sisters.**
**a**, Segmented heatmap for UV (upper) and ROS (lower) induced mutations for 7 pairs of mitotic sisters. Model of mutation phasing for a single burst event (UV, upper right) that results in a lesion segregation phenotype, as opposed to chronic, low mutation rate (ROS, lower right). **b**, Plot of UV mutations with reference C bases in the upper and reference G bases in the lower halves respectively. Lightly coloured background (yellow/blue/grey) represents segmentation of the genome based on phasing. Switches in segmentation from a mixed segment to a phased segment represent sister chromatid exchange events. The SCE (brown) and non-SCE (salmon) strand mutations can be inferred in mixed regions neighbouring SCE sites. **c**, Metaplot of UV mutation density surrounding SCE sites. Shown are smoothed mutation rates/megabase for 1mb sliding windows with a 100kb step size. Actual UV mutation density around SCE sites shown in black, while random selection of an equal number of mutations from other clones shown in red.




792

793 (legend on next page)

794

795

796

**Supplementary Fig. 7 | Haplotype resolved lesion segregation in F1 mice.**
**a**, Percentage of mapped reads to a specific genome. Each bar represents a tumour where brown depicts *Mus castaneus* specific reads and gold C3H. Mean haplotype specific reads for all 6 libraries is denoted by the vertical dashed line. Coloured boxes on the left denote which mouse the tumour was isolated from. **b**, Chromosome specific mapping rate for one of the tumours in (a). Note *Mus castaneus* specific mapping to the X. **c**, Read coverage per mutation in each library. Horizontal red dashed line represents 20x coverage. Coloured boxes below denote mouse of origin as in (a). Boxplot elements are as described in figure 3c. **d**, Number of mutations per tumour. Colours and symbols as in panel (a). **e**, Genomic stability of F1 tumours. Read counts in 10kb windows for a representative tumour. Y-axis represents log2 counts subtracted from the mean. Horizontal green line represents no difference, while horizontal red lines depict duplication or haploid content. **f**, Percent of out-of-phase mutations for phased regions in haplotype agnostic mutations (grey), haplotype resolved whole genome mutations (red), and X chromosome localised mutations (black). **g**, Left: Correlation between T <> A phasing and C <> G phasing in a single tumour within 10mb windows. C and T mutations (as well as G and A) share phase within a single tumour. Right: comparison between T <> A phasing in tumour 2 to C <> G phasing in tumour 1. **h**, Comparison of T <> A phasing across all 6 tumours, as measured by Pearson correlation for phasing in 10mb bins. **i**, Boxplot of correlation of T <> A phasing between all non-self correlations as in (h), as well as random expectation of phasing correlation by mixing bins for a respective tumour. **j**, Heatmap of T <> A phasing (x-axis) with C <> G phasing between and within tumours. Mouse of origin are coloured as in (a).

## Online Methods

### Tumour induction in CASTxC3H F1 mice

Animal experimentation was carried out in accordance with the Animals (Scientific Procedures) Act 1986 (United Kingdom) and with the approval of the Cancer Research UK Cambridge Institute Animal Welfare and Ethical Review Body (AWERB). Inbred female *Mus musculus castaneus* (CAST/EiJ) mice were crossed with inbred male C3H/HeOuJ (C3H) mice. The F1 offspring were treated with a single intraperitoneal dose of N-Nitrosodiethylamine (DEN; Sigma-Aldrich N0258; 20 mg/kg body weight) as described previously[14]. Liver tumours were isolated 30 weeks after treatment, flash frozen in liquid nitrogen and stored at -80°C for DNA extraction and sequencing. Liver tissue from an untreated P15 litter mate was sampled for control experiments. Control samples (liver tissue) were also collected from untreated, age-matched littermates.

### F1 liver tumour genomic DNA extraction

Liver DNA was extracted using the Qiagen© AllPrep DNA/RNA Mini Kit. In brief, approximately 30 mg of tumour was placed in a 2mL Eppendorf tube, and 600ul buffer RLT supplemented with 10 µl ß-mercaptoethanol was added. A 5mM stainless steel bead from Qiagen© (#69989) was added to the tube and the sample was allowed to shake for 2 x 20s at 15 Hz on the Qiagen© TissueLyser II. The lysate was subsequently transferred to an AllPrep DNA spin column and centrifuged for 30 seconds at 8,000g. The column was washed with 500 µl buffer AW1 and subsequently 500 µl AW2, and eluted in 100 µl buffer AE.

### Cell culture and splitting on the Berkeley Lights Lightning platform

P388D1 cells were obtained from ATCC® and cultured before and during incubation on the Berkeley Lights™ Lightning system using 5% $C0_2$ at 37°C in media consisting of DMEM supplemented with 10% Fetal Bovine serum, 2.5mM L-glutamine, 1x Pen-Strep and 5x B27. The 5x excess of B27 was added to the medium as recommended by Berkeley Lights™ to reduce free radicals in the media. The proliferation assay for incubator cells was carried out by plating 2.5 x $10^5$ cells in 6 well plates in triplicate. At 24 hour timepoints, cells were stained with trypan blue and counted on a Countess 3™, recording the number of live cells at each timepoint. Doubling rates for incubator cells were calculated as for Berkeley Lights™ (below). For Berkeley Lights™ **cell cycle Fucci measurements (Extended Data Fig 1c)**, cells were imaged in 3 hour windows over the course of 15 hours.

Before penning on the Berkeley Lights™ platform, a full clean was carried out on the system as per the manufacturer's instructions. Shortly after UV treatment (< 30 minutes), PF1 cells at a concentration of 2 x $10^6$/mL were penned as singlets and images acquired in ambient light, FITC and Texas Red excitation, and cell number/pen counted using the Jurkat CNN algorithm. One day post penning, cells were assayed for doublings using the cell analysis suite of Berkeley Lights™ and singlet cells that were in G1 at time of penning (Texas Red, no FITC) were split into two pens. Default settings for manual OEP of the instrument were used to move cells except wavelength voltage was increased to 6. Cells were allowed to proliferate for 4-6 days and exported to 96 well plates. Clones were expanded to approximately 2 x $10^6$ cells for DNA extraction and WGS library preparation.

**For calculating doubling rates for UV treated PF1 cells in Berkeley Lights™ pens, cell numbers were obtained using the automated Jurkat Convolution Neural Network (CNN) algorithm developed by Berkeley Lights™ over a 64 hour period using a minimum of 4 measurements. Pens that did not proliferate were first filtered, and a linear model was fit whereby the log2(cell counts) of each time point was regressed on the time in hours of culture on the chip. The mean adjusted R squared for all fits was 0.905 with a standard deviation of 0.166 (n = 998). The slope of the fit thus represented the number of doublings per hour, and was multiplied by 24 to represent divisions per day in Extended Data Panel 1a.**

### Introduction of the FastFucci system in P388D1

The lentiviral vector pBOB-EF1-FastFucci-Puro was obtained by AddGene, transformed into chemically competent DH5α E.coli and midi-preps carried out using the Qiagen MidiKit. Lentiviral packaging vectors VSVG and R8.91 were obtained as a kind gift from the lab of Michaela Frye. 5x$10^6$ Lentix HEK293T cells from Takara were transfected with 12, 5 and 12 µg of the FastFucci, VSVG and R8.91 vectors using Lipofectamine 3000. On day two and three post transfection, media from the transfected Lentix cells was sterile filtered (0.45 µM), pooled and spun in an ultracentrifuge @ 25,000 rpm @ 4ºC for 90 minutes. The pelleted virus was resuspended in 100 µl Opti-mem media. 1 x $10^6$ P388D1 cells were resuspended in 1 mL media containing 25 µl of the concentrated virus. New media and virus were added 24 hours later. 24 hours post transfection, fresh media with 2 µg/mL puromycin was added and selection carried out for 48 hours. Selected cells were expanded and passaged 4 times and subsequently single cells with GFP signal were FACs sorted into a 96 well plate. The PF1 clone was selected from this line after subsequent FACs analysis where both GFP and Kusabira orange 2 signal was analysed.

**Hoechst and FACS analysis**

To analyse G1 and G2/M fluorophores in conjunction with cellular DNA content, 2 drops of Hoechst 33342 Ready Flow™ Reagent from Invitrogen™ was added to 2 x $10^6$ cells and placed in the incubator for 15 minutes. Cells were spun down, resuspended in Miltenyi Biotec FACs buffer and assayed using the BD FACSAria™ Fusion 3 system. Green fluorophores were ascertained with excitation at 488 nM and emission at 530 nM, while the orange fluorophore of G1 cells was excited at 561 nM and emission recorded at 586 nM. For the Berkeley Lights™ Lightning system, fluorophore detection was carried out using the FITC and TRED excitation/emission filters in the Cell Analysis Suite software. Hoechst staining intensity was ascertained using a UV laser at 375 nM with emission at 450 nM. To determine overlap for measurements of all FACs signals, tables with measurements of 10000 cells were read into an R environment, variables were scaled from 1 to 100 and intensities compared.

**UV treatment conditions**

To determine UV intensity, 2.5 x $10^5$ PF1 cells were resuspended in 500 µl media and plated in 6 well dishes. UVC treatment was carried out using the Analytik Jena crosslinker (model CL-1000, 254 nm Wavelength) at exposures of 5,000-30,000 microjoules/$cm^2$ (see Extended Data Fig. 2a), and 2 mL of fresh media was added post treatment. Cells were cultured for 3 days post exposure and live cell numbers were measured in triplicate using the Countess™ 3 system from Thermo Fisher©. The exposure of 5,000 microjoules/$cm^2$ was used in subsequent experiments as roughly half of the cell count of unexposed cells were dividing post treatment.

**P388D1 Genomic DNA extraction**

Cells were spun at 500 x g for 5 minutes and supernatant was removed. The cells were then washed with 1 mL PBS, spun again at 500 x g for 5 minutes, supernatant removed, and resuspended in 200 µl PBS. DNA extraction was performed using the DNeasy kit from Qiagen© as per the manufacturer's instructions. In brief, after resuspension in 200 µl PBS, 20 µl proteinase K was added to the reaction as well as 200 µl buffer AL, briefly vortexed and then incubated for 30 minutes at 56°C with rotation of 400 rpm. 200 µl of 100% EtOH was added and the lysate was spun through a DNeasy mini spin column at 8,000 x g for 1 minute. Column was then washed with 500 µl AW1, spun as above, washed with 500 µl AW2 and spun at 18,000g for 3 minutes. To elute DNA, 100 µl buffer AE was added to the column, the column was incubated for 5 minutes at 37° and spun for 1 minute at 8,000g.

**Whole Genome Library Construction and Sequencing**

DNA size and quality post extraction was assayed using the NanoDrop™ and Agilent 4200 Tapestation® systems. Approximately 100-500 µg of genomic DNA was fragmented and libraries prepared using the NEBNext® Ultra II kit and the Unique Dual Index primers for Illumina® as per the manufacturer's instructions, with the exception that the enzymatic fragmentation step was carried out for 15 minutes instead of 5. Libraries were amplified between 4-6 cycles using the NEBNext® UDI primers (article E6440), depending on the amount of starting material. Library size and molarity was determined using the TapeStation system and libraries were pooled at a concentration of 2 µM. Paired-end 100 bp sequencing was performed using the NextSeq 2000 and NovaSeq platforms.

**Total RNA extraction and library preparation**

Total RNA was extracted from 3 replicates of $10^6$ PF1 cells using the RNeasy® Plus Mini kit from Qiagen as per the manufacturer's instructions and total RNA quality and quantity assayed using the RNA ScreenTape on the Agilent TapeStation system. 100 ng of total RNA from each replicate was then processed with the TruSeq Stranded total RNA with Illumina Ribo-zero Plus RNA depletion protocol as per the manufacturer's instructions. Libraries were quantified using the TapeStation High Sensitivity D5000 ScreenTape and QuBit dsDNA High sensitivity kit. Libraries were Sequenced on the NextSeq 2000 with 50 bp paired-end reads (100 cycle P2 chemistry).

**Omni-ATAC library preparation**

ATAC-seq libraries of the PF1 line were carried out as described [65] with slight modifications[66]. Three replicates of 50,000 cells were spun down, washed in 100 µl PBS and resuspended in 50 µl lysis buffer (10mM Tris-HCl pH 7.5, 10mM NaCl, 3mM $MgCl_2$, 0.1% NP-40, 0.1% Tween-20 and 0.01% Digitonin). After 3 minutes of incubation on ice, 1 ml of wash buffer was added (10mM Tris-HCl pH 7.5, 10mM NaCl, 3mM $MgCl_2$, 0.1% Tween-20) and subsequently centrifuged for 10 min at 500 x g at 4ºC. Supernatant was removed from the nuclear pellet, and nuclei were resuspended in 50 µl transposition mix (25 µl 2X Tagment DNA buffer (Illumina), 16.5 µl PBS, 0.5 µl 10% Tween-20, 0.5 µl 1% Digitonin, 2.5 µl Tn5 Tagment DNA enzyme (Illumina) and 5 µl $H_20$). Nuclei were incubated 30 min at 37ºC and subsequently DNA was cleaned up using the Qiagen MinElute Reaction Cleanup kit and eluted in 10ul of elution buffer. Libraries were amplified for five cycles as described[67] using the universal primers and barcoded primers A2.1, A2.2 and A2.3 for replicates 1, 2 and 3 respectively. After the initial amplification, additional cycles were determined as described[67] and an additional 8 amplification cycles were done. Libraries were sequenced on the NextSeq 2000 platform using paired-end 50 bp reads (100 cycle kit, P2).

## Sequencing Analysis

### Whole genome and ATAC alignment and filtering

Raw reads from all sequencing data were trimmed with the TrimGalore[68] software using the –stringency 3 flag. Mapping was carried out using bowtie2[69] with the --end-to-end and --maxins 1000 flags for all WGS and ATAC data. For PF1 samples, the mm10 reference genome was used while a dual hybrid N-masked reference genome (see below) was used for F1 tumour samples. Resulting bam files were then processed with samtools v1.10[70] matefix and markdup tools, and subsequently filtered using the 0x2 flag to retrieve only reads mapping antiparallel to each other on the same contig.

### Dual-hybrid N-masked reference for F1 tumour haplotype discrimination

For genomic DNA from F1 mouse tumours, an N-based mapping approach was used with the SNPsplit[48] program. In brief a dual-hybrid reference was first created where germline SNPs from both C3H and *Mus castaneus* were replaced with “Ns” in the mm10 reference using the command:

*SNPsplit_genome_preparation --vcf_file mgp.v5.merged.snps_all.dbSNP142_UCSC.vcf --strain C3H_HeJ --reference_genome BSGenome_mm10.fasta --full_sequence --nmasking --dual_hybrid --strain2 CAST_EiJ*

The VCF file containing germline SNPs in reference to the mm10 build was downloaded from the Mouse Genomes Project[47,71,72]. Trimmed WGS reads were mapped to this reference, matefixed and duplicates marked and filtered as above. Reads were then split using the SNPsplit command on the bam files with the –conflicted and –paired flags, using the SNP file created from the genome preparation file created above.

### Mutation calling and filtering

A Strelka2[73] pipeline was used to call mutations for mapped whole genome reads in both the PF1 cell line and F1 tumours. For F1 tumours, the genome prepared for mapping above was used as the reference, while for PF1 cells the standard mm10 reference was used. To determine larger structural variants, Manta was run on both tumour/normal and cell line/UV treated cells with default parameters. The input from the Manta run was used subsequently with Strelka2 and otherwise default parameters to call mutations in the whole genome sets. Mutations were originally processed with bcftools for the PASS flag, and then the GATK[74] CalculateSNVMetrics walker was used to further filter mutations. Mutations were removed if:

*VariantAlleleCount < 4, VariantAlleleCountControl > 1, VariantMapQualMedian < 40, MapQualDiffMedian < -5.0, MapQualDiffMedian > 5.0, LowMapQual > 0.05, VariantBaseQualMedian < 30, (VariantAlleleCount >= 7 & VariantStrandBias < 0.05 & ReferenceStrandBias >= 0.2), DistanceToAlignmentEndMedian < 10, DistanceToAlignmentEndMAD < 3*

CNVkit[75] was used to calculate read depth in 1kb windows across the genome, and mutations overlapping regions with coverage greater than 2 standard deviations from the mean (~11% of the genome) were filtered out. For F1 tumours, mutations overlapping germline SNPs were also removed from the analysis.

**PF1 ATAC-seq data processing**

Aligned ATAC reads from 3 replicates were used as input for MACS2[76] to call peaks with the flags -f BAMPE -g mm –nomodel –nolambda –keep-dup-all –call-summits -B -q 0.01. Peaks from individual replicates were converted to GRanges objects in the R environment and merged using the reduce function of the GenomicRanges[77] package. Read counts within merged peaks were calculated using the qCount function of the QuasR[78] package.

**PF1 RNA-seq data processing and analysis**

Transcript abundances were quantified using Kallisto[79] (v0.46.0) with the –bias and –rf-stranded flags and the Gencode M25 transcript release. To assign tags per million to a single gene instance, transcripts were split based on shared Entrez gene ID. Gene IDs with transcripts on more than one chromosome or transcripts on both strands were removed. Additionally, transcripts with no Entrez gene ID or genes that had the same alias and were overlapping were removed. To determine a gene model, ATAC signal was calculated within a 1.5kb window around each annotated TSS (-1000, +500). Gene starts were selected from the transcript with maximum ATAC signal, or in the case of no-expression, the longest isoform was used. All transcripts from an Entrez gene ID with unique signal from quantification with Kallisto were summed, and the end of the gene model determined by the longest transcript with quantified reads. To bin genes based on transcriptional level, genes were first filtered to be at least 1kb long and 1kb away from the nearest neighbour. This was done to prevent confounding signals from repetitive small genes and genes in the direct vicinity of each other, resulting in 19091 genes total. A pseudocount of 0.1 was added to the mean TPM of each gene and then log2 converted. Genes with a value of 0 or below were designated as ‘unexpressed’, or bin 1 and represented 10657 genes. The remaining 8434 ‘expressed’ genes were binned into 3 quantiles, resulting in 2812, 2811 and 2811 genes in bins 2, 3 and 4 respectively.

**HEK293 Flag-OGG1 ChIP and total RNA-seq processing and analysis**

Previously published Flag-OGG1 ChIP[42] and Ribo-Zero total RNA[80] data were downloaded from GEO using the accession numbers GSE89017 and GSE76496, respectively. Reads were trimmed using trim-galore version 0.6.6. RNA was mapped to the T2T-CHM13 version 2.0[81] reference genome using the STAR[82] aligner version 2.7.10b while Flag-OGG1 data was mapped using bowtie2[69] version 2.3.5.1 and unique reads were retained. Gene annotation in the form of a GFF3 RefSeq file was downloaded from the UCSC genome browser, and genes were filtered to have gene_biotype = 'protein coding' and extra_copy_number='0', resulting in 19,776 genes for subsequent analysis. To avoid neighbouring interference and noise from very small gene bodies, genes were further filtered to be at least 5kb in genomic length and at least 5kb away from the nearest neighbouring gene (5kb profile flanks +1kb buffer), resulting in 13,766 filtered genes. Tags Per Million (TPM) were tallied for each gene, and genes with greater than 0 TPM were quantile binned, resulting in 4 bins of 4094 for non/very low transcribed genes, and 3224 genes each for low, medium and high categories. Each gene body was divided into 100 equal tiles using the tileGenome function of the GenomicRanges package, and each 5kb flank was divided into 10 tiles of 500bp each. Reads per kilobase were tallied for each tile. Mean log2(counts) for each tile in a respective genomic bin was calculated to produce metaplots of signal (Extended Data Fig 4h).

**Multi-allelic analysis**

Tandem mutations were identified by using the distanceToNearest function in the GenomicRanges R package and selecting those with an intermutation distance of 0. To calculate VAFs for each base in a tandem mutation, reads were extracted from the relevant PF1 genome using Rsamtools[83], whereby each read in the calculation had sequence information for both bases. Mutli-allelic sites were selected on the criteria that two alternative alleles with at least 3 unique reads with information at both bases were present. To determine mutation order, alleles at each multi-allelic site were ordered based on VAF, with the higher VAF allele assumed to be the first mutation at that site. To confirm VAF bias for the second C in tandem CC > TT mutations, we used a grepping approach. For each multiallelic site, we used 10bp of sequence on each side of the tandem site (22 base long sequence as query) for all possible alleles and their reverse complement (8 in total per multi-allelic site). To compare with the alignment data, we filtered sites where one of the alternate alleles had an exact match to another 22 bp sequence in the genome. We further filtered for regions with 0 matches that occurred because of single base pair changes in the genome

that precluded an exact string matching event. In total this resulted in 237 tandem CC > TT mutations to compare VAFs for each base in the tandem mutation (Extended Data Fig. 3).

**RL20 metric**

The rl20 metric was carried out as previously described[14]. In short, run lengths of relevant mutations (ex. C > T or G > A for UV) were calculated on a chromosomal basis using the rle (run length encoding) function in R and runs were ordered by decreasing size. The smallest run length in the top 20% of this list was set as the run length for that particular genome and mutation type. The significance of seeing such a run length given equal probability of either mutation orientation was calculated using the Wald-Wolfowitz runs test with the runs.tests function in the R package randtests[84].

**Mutational phasing**

To compare mutational phasing between samples, the genome was first tiled into sliding 10mb bins with a 100kb step using the slidingWindows function of the GenomicRanges package. Bins were retained with 95% mappability, as determined using the function mappabilityCalc in the Repitools[85] R package. Overlapping mutations of either stranded-orientation, for example C > T and G > A with UV, were assigned a 1 or -1 respectively. The average of this number for each bin represented the phasing of that bin. Thus if there were equal numbers of both UV mutation types in a bin, the phasing of UV mutations in that bin would be 0. In contrast, if all mutations were C > T, the phasing would be 1, while all G > A mutations would be -1.

To simulate a mutation phasing distribution assuming a lesion segregation phenotype with this bin based approach, we created 3 sets of bernoulli trials in a 1:2:1 proportion. Using UV as an example, these sets represented C > T phased segments, mixed segments and G > A phased segments, respectively (see Extended Data Fig. 5). Each mutation within the bin was thus a trial, while bins represented sets of trials. The 1:2:1 proportion was used as the expected ratio under the Hardy-Weinberg[86] assumption, given each mitotic sister inherits the mutation result of two strands non-selectively in mitosis. We next set C > T mutations as a success while G > A mutations were a failure. The probability of success in mixed bins of the genome was thus set to 50%, resulting in an equal representation of both mutation types. For phased bins, under perfect circumstances in the framework of this model it would be assumed the probability of success is 100% in a C > T phased segment, and 0 in a G > A phased segment. This accuracy was not reflected in the genome, as even in phased segments of F1 resolved genomes and the singular X chromosome, roughly 5% mutations are out of phase (see Extended Data Fig. 6).

To calculate the out of phase rate in our PF1 cell line data, we focused on the mutations shared between mitotic sister cells at the time of penning. Our reasoning was that while C > A mutations made up the majority of the population, on average 31% of this number were C > T or G > A and thus indistinguishable from true UV mutations. This meant that given the number of sister specific ROS mutations, between 5 and 16% of this number are false positive UV mutations (see Extended Data Fig. 5). Probability of success was adjusted to reflect this fact. More specifically, if 12% of C > T mutations in a particular genome are assumed to be background, the probability of success was shifted by 6% as these 'false positive' UV mutations would be assumed to be incorrectly phased half the time.

For a 10mb bin to be considered in the model, at least 10 mutations needed to be present in that bin, which equated to 13,971 bins for UV and 16,337 bins for ROS. To create an exhaustive population of these segments and establish an ideal distribution, we carried out 100 fold more Bernoulli trials than were present in all 14 genomes, which equated to 139k for UV and 163K for ROS. This number was chosen as it is two orders of magnitude larger than the actual population. We also carried out a set of trials with the same amount of data points actually represented in the data (13.9k and 16.3k), to directly compare the distributions qualitatively and subsequently with a qq-plot of the resulting distributions.

**Sister chromatid exchange**

Our previous work describing lesion segregation suggested that switches in mutation phasing were the result of sister chromatid exchange[14]. These switches were exclusively noted to be composed of a mixed segment directly adjacent to a polarised segment (with the exception of genome build errors). This behaviour is precisely what would be predicted if mutations are fixed in a stranded manner during the first mitosis following DNA damage, followed by a recombination event between homologues. The mirror-image mutation phasing for mitotic sisters we describe here provided direct evidence that switches in mutation phasing were the result of sister chromatid exchange, as we could now account for both products of the recombination event.

A direct consequence of this observation is that we could identify mutations on the recombined and non-recombined strand residing in mixed segments. While it is impossible to query recombination status directly in phased regions, mixed regions inherently contain mutation information from both orientations, and thus both strands. More specifically, mutations in mixed segments with the opposite orientation of the adjacent phased segment would come predominantly from the recombined strand (see Fig. 5f).

Thus to delineate mutations specific to the recombined strand, we first identified the phased and unphased segments on each side of an SCE site. This was done by taking the absolute value of the skew, which scaled from -1 to 1. The mixed segment was determined as the smaller absolute value of skew. After identifying the skewed segment, we identified the polarity of the adjacent segment by asking if the skew was greater or less than 0, in the case of UV meaning it was either a C > T or G > A phased segment respectively. Finally, mutations in the mixed segment with the opposite orientation of the adjacent skewed segment were identified as recombined strand mutations. In contrast, mutations in the mixed segment with the same polarity as the adjacent skewed segment were noted as non-recombined strand mutations.

To profile mutation density for both strands and the mixed regions around SCE sites, a 1mb sliding window approach with 100kb step size was used as above, covering in total 20mb centered on the SCE site. Mutation rate was reported as the number of mutations/megabase (mu/mb).

**Mutation rate calculation**

Mutation rates were calculated as previously reported[14]. In brief, to account for genomic representation in mutation rate, each mutation type (ie C > T) in addition to bases immediately adjacent to the mutation site were first summed, creating a trinucleotide vector of 192 unique mutation instances. This was folded into a vector of 64 unique mutations by combining identical trinucleotide contexts, where only a different alternate base was observed. For instance a C > T mutation in the ACG context was treated identically to a C > A mutation in the ACG context. The number of trinucleotide mutation instances was then divided by the total number of possible trinucleotides in that window. The weighted mean of this number for all trinucleotides was then calculated, with the weights being the relative representation of that specific trinucleotide in either the window of interest or the whole genome, depending on the comparison. This number was then multiplied by $10^6$ to represent mutations/megabase.

**Mutation signatures**

To identify mutation signatures, trinucleotide sequences centered on each mutation were first reverse complemented if the reference base was either A or G. This created a vector of length 96, representing all mutations in the context of either a C or T reference base. The number of that mutation type was divided by the total mutations for that sample to depict a frequency of each mutation identity in the population. To compare our observed mutation signatures to previously identified ones, we downloaded SBS signatures from the COSMIC

database version 3.2 for the mouse genome reference GRCm38. This consisted of 79 total signatures, 19 of which were filtered out given evidence of possible sequencing artefacts. Using our frequency scaled signatures defined above, we compared the cosine similarity of the 96 length vectors for each sequenced genome to the 60 filtered COSMIC signatures.

**Transcription coupled repair analysis**

To compare mutation rates to transcription output we used the gene models and bins defined above. In the case of UV mutations, template strand damage was determined as a C > T mutation in minus strand genes and a G > A mutation in plus strand genes. Conversely, C > T mutations in plus strand genes and G > A mutations in minus strand genes were defined as non-template mutations. The same logic was applied to ROS mutations, meaning G > T mutations on minus strand genes and C > A mutations on plus strand genes were determined to be template mutations. The inverse again were designated as non-template mutations. Mutation rates were calculated as described above, with mean weights calculated using the trinucleotide representation of the whole genome. Stranded mutation rates were multiplied by 2 and then divided by the genome average, to represent rate relative to the genome (Fig. 3c-e).

## References


1. Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J.* **17**, 1195–1214 (2003).
2. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
3. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* vol. 362 709–715 Preprint at https://doi.org/10.1038/362709a0 (1993).
4. Noonan, F. P. *et al.* Neonatal sunburn and melanoma in mice. *Nature* **413**, 271–272 (2001).
5. Connor, F. *et al.* Mutational landscape of a chemically-induced mouse model of liver cancer. *J. Hepatol.* **69**, 840–850 (2018).
6. Huang, R. & Zhou, P.-K. DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy. *Signal Transduct Target Ther* **6**, 254 (2021).
7. Lindahl, T. & Barnes, D. E. Repair of endogenous DNA damage. *Cold Spring Harb. Symp. Quant. Biol.* **65**, 127–133 (2000).
8. Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644–656 (2017).
9. Whitaker, A. M., Schaich, M. A., Smith, M. R., Flynn, T. S. & Freudenthal, B. D. Base excision repair of oxidative DNA damage: from mechanism to disease. *Front. Biosci.* **22**, 1493–1522 (2017).
10. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821–836.e16 (2019).
11. Gedik, C. M., Collins, A. & ESCODD (European Standards Committee on Oxidative DNA Damage). Establishing the background level of base oxidation in human

lymphocyte DNA: results of an interlaboratory validation study. *FASEB J.* **19**, 82–84 (2005).

12. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15**, 465–481 (2014).
13. Leung, A. K., Barankin, B., Lam, J. M., Leong, K. F. & Hon, K. L. Xeroderma pigmentosum: an updated review. *Drugs Context* **11**, (2022).
14. Aitken, S. J. *et al.* Pervasive lesion segregation shapes cancer genome evolution. *Nature* **583**, 265–270 (2020).
15. Anderson, C. J. *et al.* Strand-resolved mutagenicity of DNA damage and repair. *bioRxiv* 2022.06.10.495644 (2022) doi:10.1101/2022.06.10.495644.
16. Koh, S.-B. *et al.* A quantitative FastFUCCI assay defines cell cycle dynamics at a single-cell level. *J. Cell Sci.* **130**, 512–520 (2017).
17. Klein, J. C. *et al.* Repair and replication of plasmids with site-specific 8-oxodG and 8-AAFdG residues in normal and repair-deficient human cells. *Nucleic Acids Res.* **20**, 4437–4443 (1992).
18. Kamiya, H. *et al.* c-Ha-ras containing 8-hydroxyguanine at codon 12 induces point mutations at the modified and adjacent positions. *Cancer Res.* **52**, 3483–3485 (1992).
19. Moriya, M. Single-stranded shuttle phagemid for mutagenesis studies in mammalian cells: 8-oxoguanine in DNA induces targeted G.C-->T.A transversions in simian kidney cells. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 1122–1126 (1993).
20. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
21. Lopes, F. C. P. S. *et al.* UV Exposure and the Risk of Cutaneous Melanoma in Skin of Color: A Systematic Review. *JAMA Dermatol.* **157**, 213–219 (2021).
22. Jin, S.-G., Meng, Y., Johnson, J., Szabó, P. E. & Pfeifer, G. P. Concordance of hydrogen peroxide–induced 8-oxo-guanine patterns with two cancer mutation signatures of upper GI tract tumors. *Science Advances* **8**, eabn3815 (2022).

23. Brody, Y. *et al.* Quantification of somatic mutation flow across individual cell division events by lineage sequencing. *Genome Res.* **28**, 1901–1918 (2018).
24. Brash, D. E. UV signature mutations. *Photochem. Photobiol.* **91**, 15–26 (2015).
25. Laughery, M. F. *et al.* Atypical UV Photoproducts Induce Non-canonical Mutation Classes Associated with Driver Mutations in Melanoma. *Cell Rep.* **33**, 108401 (2020).
26. Demeulemeester, J., Dentro, S. C., Gerstung, M. & Van Loo, P. Biallelic mutations in cancer genomes reveal local mutational determinants. *Nat. Genet.* **54**, 128–133 (2022).
27. Jiang, G. & Sancar, A. Recruitment of DNA damage checkpoint proteins to damage in transcribed and nontranscribed sequences. *Mol. Cell. Biol.* **26**, 39–49 (2006).
28. Ganesan, A., Spivak, G. & Hanawalt, P. C. Transcription-coupled DNA repair in prokaryotes. *Prog. Mol. Biol. Transl. Sci.* **110**, 25–40 (2012).
29. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* **9**, 958–970 (2008).
30. Bohr, V. A., Smith, C. A., Okumoto, D. S. & Hanawalt, P. C. DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. *Cell* **40**, 359–369 (1985).
31. Mellon, I., Spivak, G. & Hanawalt, P. C. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. *Cell* **51**, 241–249 (1987).
32. van Hoffen, A., Venema, J., Meschini, R., van Zeeland, A. A. & Mullenders, L. H. Transcription-coupled repair removes both cyclobutane pyrimidine dimers and 6-4 photoproducts with equal efficiency and in a sequential way from transcribed DNA in xeroderma pigmentosum group C fibroblasts. *EMBO J.* **14**, 360–367 (1995).
33. Spivak, G. & Hanawalt, P. C. Host cell reactivation of plasmids containing oxidative DNA lesions is defective in Cockayne syndrome but normal in UV-sensitive syndrome fibroblasts. *DNA Repair* **5**, 13–22 (2006).
34. Wilson, I. D. M. *Base Excision Repair Pathway, The: Molecular Mechanisms And Role In Disease Development And Therapeutic Design*. (World Scientific, 2016).

35. Reis, A. M. C. *et al.* Targeted detection of in vivo endogenous DNA base damage reveals preferential base excision repair in the transcribed strand. *Nucleic Acids Res.* **40**, 206–219 (2012).
36. Banerjee, D. *et al.* Preferential repair of oxidized base damage in the transcribed genes of mammalian cells. *J. Biol. Chem.* **286**, 6006–6016 (2011).
37. Menoni, H., Hoeijmakers, J. H. J. & Vermeulen, W. Nucleotide excision repair-initiating proteins bind to oxidative DNA lesions in vivo. *J. Cell Biol.* **199**, 1037–1046 (2012).
38. Sancar, A. DNA excision repair. *Annu. Rev. Biochem.* **65**, 43–81 (1996).
39. Wood, R. D. Nucleotide excision repair in mammalian cells. *J. Biol. Chem.* **272**, 23465–23468 (1997).
40. Guo, J., Hanawalt, P. C. & Spivak, G. Comet-FISH with strand-specific probes reveals transcription-coupled repair of 8-oxoGuanine in human cells. *Nucleic Acids Res.* **41**, 7700–7712 (2013).
41. Poetsch, A. R. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Comput. Struct. Biotechnol. J.* **18**, 207–219 (2020).
42. Hao, W. *et al.* Effects of the stimuli-dependent enrichment of 8-oxoguanine DNA glycosylase1 on chromatinized DNA. *Redox Biol* **18**, 43–53 (2018).
43. van Wietmarschen, N. & Lansdorp, P. M. Bromodeoxyuridine does not contribute to sister chromatid exchange events in normal or Bloom syndrome cells. *Nucleic Acids Res.* **44**, 6787–6793 (2016).
44. Latt, S. A. Sister chromatid exchange formation. *Annu. Rev. Genet.* **15**, 11–55 (1981).
45. St Charles, J. *et al.* High-resolution genome-wide analysis of irradiated (UV and γ-rays) diploid yeast cells reveals a high frequency of genomic loss of heterozygosity (LOH) events. *Genetics* **190**, 1267–1284 (2012).
46. Eppink, B. *et al.* The response of mammalian cells to UV-light reveals Rad54-dependent and independent pathways of homologous recombination. *DNA Repair* **10**, 1095–1105 (2011).
47. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene

regulation. *Nature* **477**, 289–294 (2011).

48. Krueger, F. & Andrews, S. R. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res.* **5**, 1479 (2016).
49. Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. *Nat. Rev. Genet.* **23**, 298–314 (2022).
50. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat. Rev. Cancer* **21**, 379–392 (2021).
51. Elliott, K. & Larsson, E. Non-coding driver mutations in human cancer. *Nat. Rev. Cancer* **21**, 500–509 (2021).
52. de Kanter, J. K. *et al.* Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. *Cell Stem Cell* **28**, 1726–1739.e6 (2021).
53. Greaves, M. Evolutionary determinants of cancer. *Cancer Discov.* **5**, 806–820 (2015).
54. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
55. Cross, W. C., Graham, T. A. & Wright, N. A. New paradigms in clonal evolution: punctuated equilibrium in cancer. *J. Pathol.* **240**, 126–136 (2016).
56. Charles Richard, J. L. *et al.* FACT Assists Base Excision Repair by Boosting the Remodeling Activity of RSC. *PLoS Genet.* **12**, e1006221 (2016).
57. Campalans, A. *et al.* Distinct spatiotemporal patterns and PARP dependence of XRCC1 recruitment to single-strand break and base excision repair. *Nucleic Acids Res.* **41**, 3115–3129 (2013).
58. Olmon, E. D. & Delaney, S. Differential Ability of Five DNA Glycosylases to Recognize and Repair Damage on Nucleosomal DNA. *ACS Chem. Biol.* **12**, 692–701 (2017).
59. Cozzarelli, N. R. Editorial expression of concern. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11816 (2003).
60. Cooper, P. K., Nouspikel, T. & Clarkson, S. G. Retraction. *Science* **308**, 1740 (2005).
61. Le Page, F. *et al.* Transcription-coupled repair of 8-oxoguanine: requirement for XPG,

TFIIH, and CSB and implications for Cockayne syndrome. *Cell* **123**, 711 (2005).

62. Hauer, M. H. & Gasser, S. M. Chromatin and nucleosome dynamics in DNA damage and repair. *Genes Dev.* **31**, 2204–2221 (2017).
63. Bhakat, K. K., Mokkapati, S. K., Boldogh, I., Hazra, T. K. & Mitra, S. Acetylation of human 8-oxoguanine-DNA glycosylase by p300 and its role in 8-oxoguanine repair in vivo. *Mol. Cell. Biol.* **26**, 1654–1665 (2006).
64. Menoni, H., Di Mascio, P., Cadet, J., Dimitrov, S. & Angelov, D. Chromatin associated mechanisms in base excision repair - nucleosome remodeling and DNA transcription, two key players. *Free Radic. Biol. Med.* **107**, 159–169 (2017).
65. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
66. [No title]. https://www.med.upenn.edu/kaestnerlab/assets/user-content/documents/ATAC-seq-Protocol-(Omni)-Kaestner-Lab.pdf.
67. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
68. GitHub - FelixKrueger/TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. *GitHub* https://github.com/FelixKrueger/TrimGalore.
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* vol. 9 357–359 Preprint at https://doi.org/10.1038/nmeth.1923 (2012).
70. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
71. Doran, A. G. *et al.* Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biol.* **17**, 167 (2016).
72. Lilue, J. *et al.* Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **50**, 1574–1583 (2018).
73. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat.*

*Methods* **15**, 591–594 (2018).

74. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
75. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
76. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
77. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
78. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130–1132 (2015).
79. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-Seq quantification with kallisto. *Nat. Biotechnol.*
80. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26**, 1742–1752 (2016).
81. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
82. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
83. Morgan, M., Pages, H., Obenchain, V. & Hayden, N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. *R package version*.
84. Caeiro, F. & Mateus, A. randtests: Testing randomness in R. 2014. *Available in: https://cran. r-project. org/package*.
85. Statham, A. L. *et al.* Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics* **26**, 1662–1663 (2010).
86. Hardy, G. H. MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science* **28**, 49–50 (1908).

1389