Pervasive lesion segregation shapes cancer genome evolution

Sarah J. Aitken[1,2,3,13], Craig J. Anderson[4,13,#], Frances Connor[1,13,#], Oriol Pich[5,13], Vasavi Sundaram[1,6,13], Christine Feig[1,13], Tim F. Rayner[1,13], Margus Lukk[1,13], Stuart Aitken[4,13], Juliet Luft[4,13] , Elissavet Kentepozidou[6,13], Claudia Arnedo-Pac[5,13], Sjoerd V. Beentjes[7], Susan E. Davies[3], Ruben M. Drews[1,13], Ailith Ewing[4,13], Vera B. Kaiser[4,13], Ava Khamseh[4,8], Erika López-Arribillaga[5,13], Aisling M. Redmond[1], Javier Santoyo-Lopez[9], Inés Sentís[5,13], Lana Talmane[4,13], Andrew D. Yates[6], Colin A. Semple[4,13], Núria López-Bigas[5,10,11,13], Paul Flicek[1,6,13], Duncan T. Odom[1,12,13,*], Martin S. Taylor[4,13,*].

[1] Cancer Research UK - Cambridge Institute, University of Cambridge, Cambridge, UK

[2] Department of Pathology, University of Cambridge, Cambridge, UK

[3] Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

[4] MRC Human Genetics Unit, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, UK

[5] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain

[6] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

[7] School of Mathematics and Maxwell Institute, University of Edinburgh, Edinburgh, UK

[8] Higgs Centre for Theoretical Physics, University of Edinburgh, Edinburgh, UK

[9] Edinburgh Genomics (Clinical), The University of Edinburgh, Edinburgh, UK

[10] Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain

[11] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[12] German Cancer Research Center (DKFZ), Division of Regulatory Genomics and Cancer Evolution, Heidelberg, Germany

[13] Liver Cancer Evolution Consortium

# Equal contribution

* Corresponding authors

Cancers arise through the acquisition of oncogenic mutations and grow through clonal expansion[1,2]. Here we reveal that most mutagenic DNA lesions are not resolved as mutations within a single cell-cycle. Instead, DNA lesions segregate unrepaired into daughter cells for multiple cell generations, resulting in the chromosome-scale phasing of subsequent mutations. We characterise this process in mutagen-induced mouse liver tumours and show that DNA replication across persisting lesions can produce multiple alternative alleles in successive cell divisions, thereby generating both multi-allelic and combinatorial genetic diversity. The phasing of lesions enables the accurate measurement of strand biased repair processes, quantification of oncogenic selection, and fine mapping of sister chromatid exchange events. Finally, we demonstrate that lesion segregation is a unifying property of exogenous mutagens, including UV light and chemotherapy agents in human cells and tumours, which has profound implications for the evolution and adaptation of cancer genomes.

Analysis of cancer genomes has identified a wealth of driver mutations and mutation signatures[1,3], illustrating how environmental mutagens cause genetic damage and elevate cancer risk[4,5]. The numerous mutation patterns identified in cancer genomes is testament to the temporal and spatial heterogeneity of exogenous and endogenous exposures, mutational processes, and germline variation amongst patients. A study of diverse human cancers identified 49 distinct single base substitution signatures, with almost all tumours demonstrating evidence of at least three signatures[3].

Such intrinsic heterogeneity leads to overlapping mutation signatures that confound our ability to accurately disentangle the biases of DNA damage and repair, or to interpret the dynamics of clonal evolution. We reasoned that a more controlled and genetically uniform cancer model system would overcome some of these limitations. By effectively re-running cancer evolution hundreds of times, we aimed to explore oncogenesis and mutation patterns at high resolution and with good statistical power.

We chemically induced liver tumours in fifteen-day-old (P15) male C3H/HeOuJ inbred mice (**Fig. 1a**; subsequently C3H, n=104) using a single dose of diethylnitrosamine (DEN)[6]. For comparison and validation, we replicated the study in the divergent mouse strain CAST/EiJ[7] (subsequently CAST, n=54) (**Extended Data Fig. 1**).

Whole genome sequencing (WGS) of 371 independently-evolved tumours from 104 C3H mice (**Supplementary Table 1**) revealed that each genome had ~60,000 (~13 per Mb) somatic point mutations (**Extended Data Fig. 1a**), comparable to human cancers caused by exogenous mutagens such as tobacco[8] and UV[9]. Insertion-deletion mutations and larger segmental changes were rare (**Extended Data Fig. 1a-f**). Point mutations were dominated (76%) by T→N/A→N changes (where N represents any alternate nucleotide; **Fig. 1b; Extended Data Fig. 1g-j**), consistent with the long-lived thymine adduct O^4-ethyl-deoxythymidine as the principal mutagenic lesion[10]. Known driver mutations were in the EGFR/RAS/RAF pathway[6,11,12] (**Fig. 1c**) and usually mutually exclusive. Similar results were replicated in CAST mice (**Extended Data Fig. 1j**).

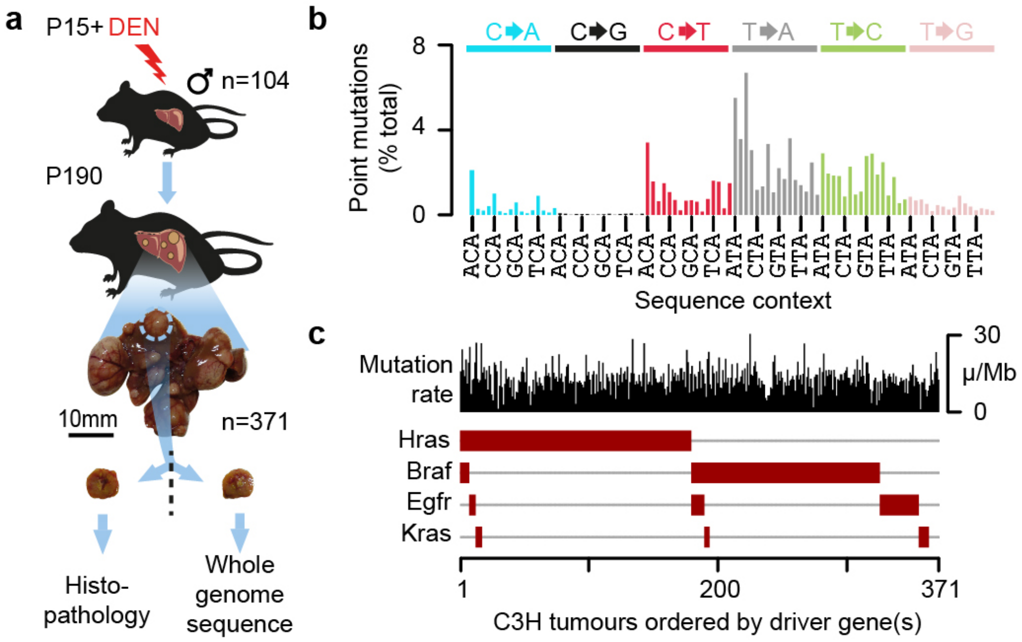


Fig.1 | DEN-initiated tumours have a high burden of T->N/A->N mutations and driver changes in the EGFR/RAS/RAF pathway. **a**, Fifteen-day-old (P15) male C3H/HeOuJ mice received a single dose of diethylnitrosamine (DEN); 371 tumours were isolated 25 weeks later (P190), histologically analysed, and whole-genome sequenced. **b**, The aggregated mutations showing the distribution of nucleotide substitutions; every fourth trinucleotide context is displayed (x-axis). **c**, Each tumour is shown as a column with its mutation rate (μ/Mb, black) and driver mutations (brown boxes).

Chromosome-scale segregation of lesions

In each tumour we observed multi-megabase genomic segments with pronounced Watson versus Crick strand asymmetry of mutations, often encompassing entire chromosomes (**Fig. 2**). We define Watson strand bias as an excess of T→N over A→N mutations when called on the forward strand of the reference genome, and the converse as Crick strand bias. With a median span of 55 MB (**Fig. 2a-d**), these asymmetrically mutated segments are orders of magnitude longer than asymmetries generated by transcription coupled repair (TCR)[13], APOBEC mutagenesis[14,15], or replication biases[13,16]. Total mutation load and DNA copy-number remains uniform across the genome (**Fig. 2e,f**).

Pervasive, strand-asymmetric mutagenesis can be explained by DEN-induced lesions remaining unrepaired prior to genome replication. The first round of replication after DEN results in two sister chromatids with independent lesions on each parent strand, and daughter strands containing misincorporation errors complementary to the lesions (**Fig. 2i**). The sister chromatids segregate into separate daughter cells during mitosis, and lesion:mutation duplexes are resolved by later replication cycles. Asymmetric regions show a 23-fold (median) excess of their preferred mutation over its reverse complement, thus >95% of lesions that generate a mutation segregate for at least one mitotic division. We subsequently refer to this phenomenon as “lesion segregation”.

The haploid X chromosome always contains segments with a strong strand bias (**Fig. 2g**). On autosomal chromosomes, when both allelic copies have the same bias, the genome shows that bias (e.g. Watson bias, chromosome 15 in **Fig. 2a-d**); when one copy has Watson and the other has Crick bias, the chromosome appears unbiased (e.g. chromosome 19 in **Fig. 2a-d**). A model based on random retention of Watson or Crick biased chromosomes accurately predicts that (1) ~50% of the autosomal genome and (2) ~100% of the haploid X chromosome show mutational asymmetry (**Fig. 2g; Extended Data Fig. 2**). A few tumours (3.5%) have absent or muted asymmetry; cellularity estimates indicate they are polyclonal or polyploid (**Supplementary Table 1**).

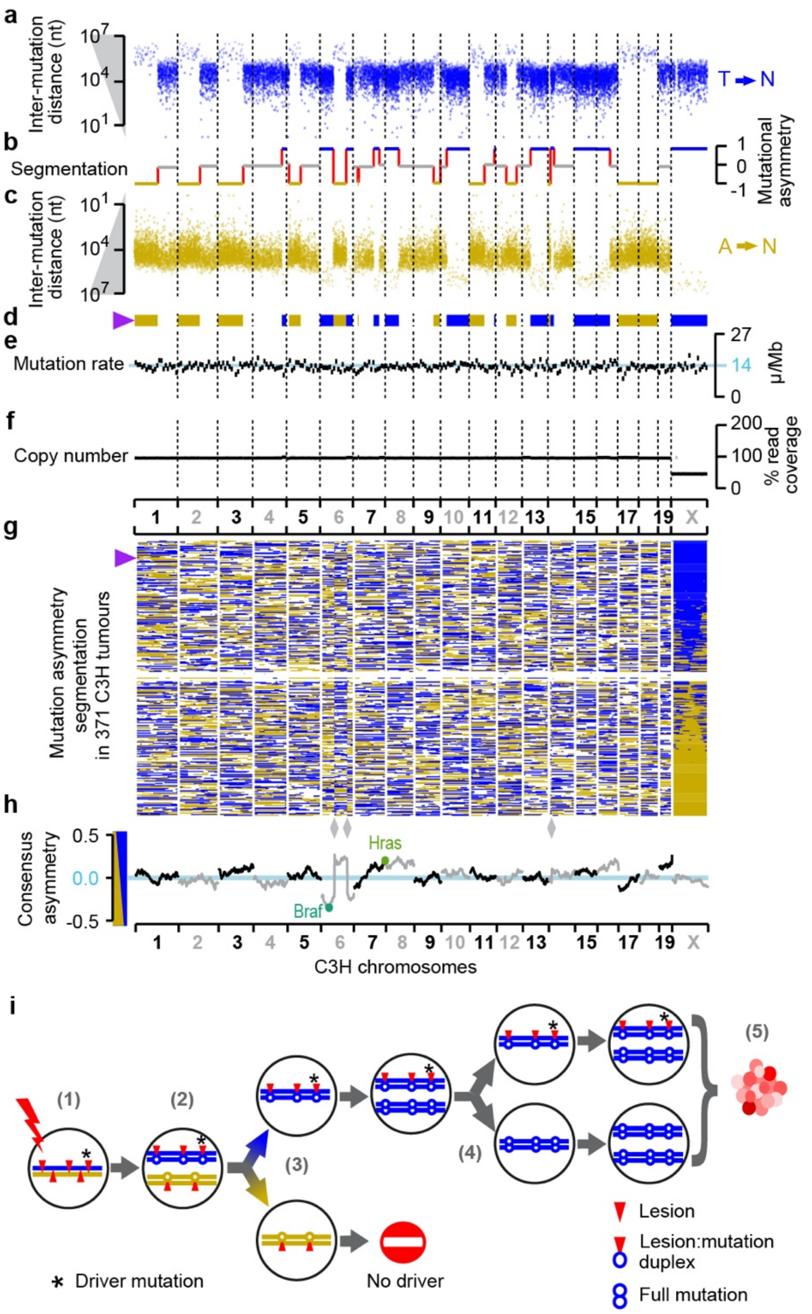


Fig.2 | Chromosome-scale and strand asymmetric segregation of DNA lesions. **a-f**, An example DEN-induced C3H tumour (identifier: 94315_N8). **a-c**, Mutational asymmetry. Individual T→N mutations shown as blue points if Watson strand (**a**) and gold if Crick (**c**); y-axis plots distance to nearest same-strand T→N mutation. **b**, Segmentation of mutation strand asymmetry patterns. Y-axis shows degree of asymmetry (no bias: grey); symmetry switches are red lines. **d**, Asymmetric segments shown as ribbon-plot. **e**, Mutation rate (μ/Mb) in 10Mb windows, blue line shows genome-wide average. **f**, DNA copy number in 10Mb windows (grey) and for each asymmetric segment (black). **g**, Ribbon-plots (as in **d**) for 371 C3H tumours ranked by chromosome X asymmetry. Purple triangle indicates example tumour (**a-f**). Grey diamonds mark reference genome mis-assemblies. **h**, Driver genes distort the balance of Watson and Crick asymmetries (Methods). **i**, Mechanistic model of lesion segregation. A mutagen generates lesions (red triangles) on both DNA strands *(1)*. If not removed *(2),* lesions will segregate into sister chromatids: one carrying only Watson (blue) strand lesions and the second only Crick (gold) strand lesions. Post-mitotic daughter cells will have independent lesions and resulting replication errors *(3)*, resolved into full mutations in later replication *(4)*. Only lineages containing driver changes (* in *step* (*1)*) will expand into substantial populations *(5)*.

Resolving sister chromatid exchange

The lesion segregation model predicts that mutational asymmetries should span whole chromosomes. However, we observe symmetry switches between multi-megabase segments of Watson and Crick bias within chromosomes (**Fig. 2a-d,g**). These likely represent sister chromatid exchanges (SCEs) from homologous recombination (HR) mediated DNA repair[17] (**Extended Data Fig. 4a**). SCEs are typically invisible to sequencing technologies because HR between sister chromatids is thought to be error-free[18].

SCE frequency per-tumour positively correlates with point mutation rate (**Extended Data Fig. 3a,b**). With ~27 (median) SCEs in each tumour genome (n=371), we were well-powered to detect recurrent exchange sites and biases in genomic context (**Extended Data Fig. 3c,d**). After removing three reference genome mis-assemblies (**Fig. 2g**; **Extended Data Fig. 3e,f**), we found that SCEs occur with modest enrichment in transcriptionally inactive, late-replicating regions (**Extended Data Fig. 4b**). The fine mapping (~20kb resolution) of SCEs allowed us to test the fidelity of HR. The mutation rate appears locally elevated at SCEs, but the mutational spectrum matches the rest of the genome (**Extended Data Fig. 4c-f**). A model of Holliday intermediate branch migration could explain these observations (**Extended Data Fig. 4g**).

Lesion segregation reveals selection

Cumulatively the tumours have equal Watson and Crick lesion strand retention across most of the genome (**Fig. 2h**). However, we observe striking deviations at loci spanning known driver genes (**Fig. 2h**). The driver *Braf* T→A mutation at codon 584[6] is observed in 153/371 C3H tumours, and we would expect the surrounding chromosomal segment to retain T-lesions on the same strand. This is the case in 94% (144/153) of tumours (Fisher's exact test $p=3.6x10^{-19}$). In contrast, tumours lacking the *Braf* mutation do not show a retention bias (47% Crick, 53% Watson, p=0.88, not rejecting 50:50 null expectation). We applied this novel test for oncogenic selection at sites with sufficient recurrent mutations to have statistical power, which confirmed significant oncogenic selection in *Hras*, *Braf,* and *Egfr* (**Fig. 1c; Extended Data Table 1**).

DNA repair with lesion strand resolution

Resolving DNA lesions to specific strands within a single mutagenised cell cycle presents a unique opportunity to investigate strand-specific DNA damage and repair *in vivo*. For example, transcription-coupled nucleotide excision repair (TCR, **Fig. 3a**) specifically removes DNA lesions from the RNA template strand[19,20].

We generated transcriptomes from the tissue of origin at the developmental time of DEN mutagenesis. Mutation rates were calculated for each gene in each tumour, stratified by both expression level and the strand containing lesions (**Fig. 3b**). As expected, TCR was highly specific to the template strand and correlated closely with gene expression. The mutation rate in non-expressed genes had no observable transcription strand-bias. In contrast, mutations in highly expressed genes were reduced 79.8±1.0%, if the tumour had template-strand lesions.

To evaluate the specificity of TCR, we compared mutation rates for each trinucleotide context between template and non-template strands, stratified by expression level (**Fig. 3c**). For highly expressed genes, thymines have an 82% (s.d. 6.8% across sequence contexts) lower mutation rate on the template strand; the non-template mutation rate is indifferent to expression (**Fig. 3c**, dark-blue lines are close to vertical), as expected[19]. Mutations from C and G show highly efficient TCR on the template strand; 70% (s.d. 7.8%) and 34% (s.d. 21%) respectively. In contrast to T mutations, they also show an expression-dependent reduction in mutation rate on the non-template strand, suggesting non-TCR repair processes are active. Rare mutations from adenine on the lesion containing strand increase with transcription, possibly due to error-prone trans-lesion DNA polymerase Pol-η[21].

The ability to resolve the lesion strand unmasks the striking contribution of bidirectional transcription from active promoters[22] in shaping mutation patterns (**Fig. 3d-f; Extended Data Fig. 5**). Genic transcription is associated with a sharp, sustained reduction in mutation rate from template strand lesions. A local increase in the mutation rate over the ~200 nucleotides upstream of the TSS (**Fig. 3d**) is revealed to result from genic and upstream bidirectional transcription emerging from opposite edges of the core promoter[23] leading to a local depletion of within-promoter TCR activity (**Fig. 3e,f**).

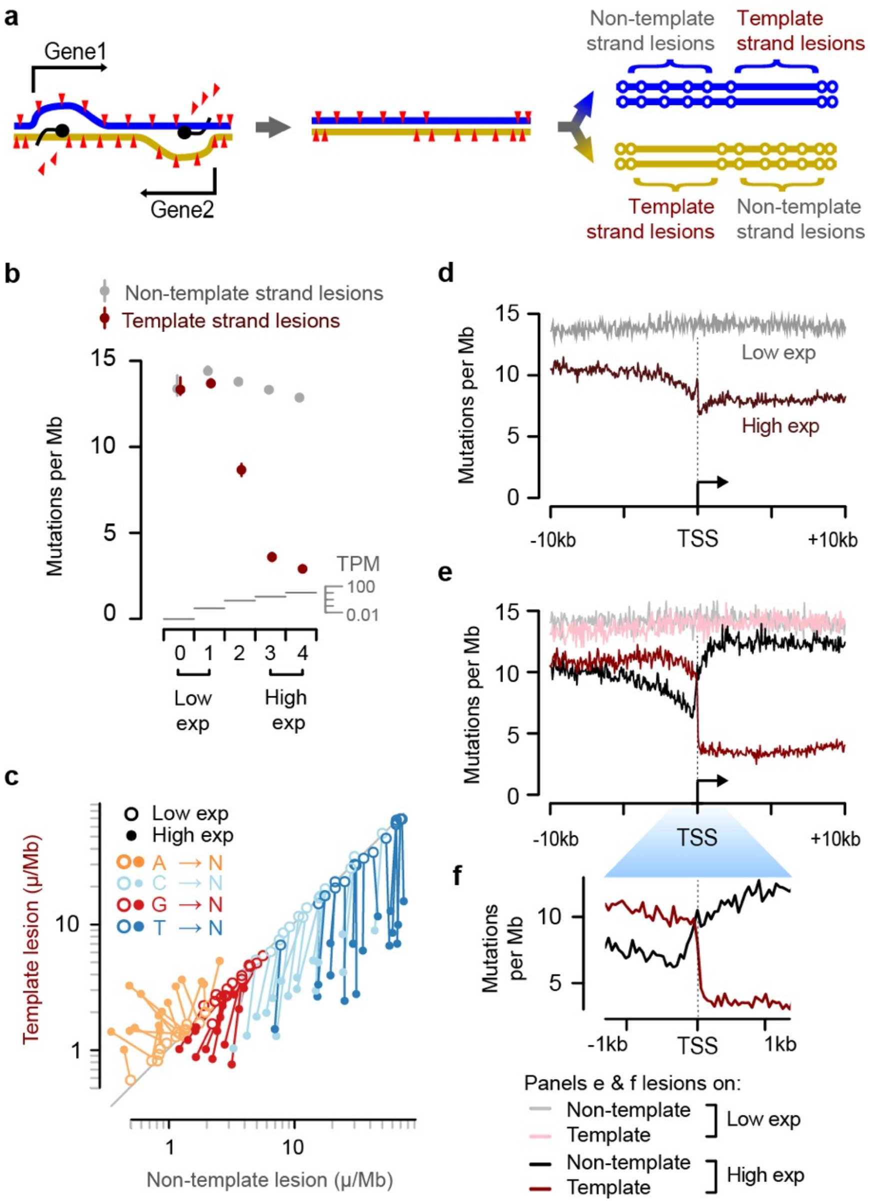


Fig.3 | Identification of the lesion containing DNA strand allows transcription coupled repair (TCR) to be quantified with strand specificity. a, TCR of DNA lesions is only expected to reduce the mutation rate when lesions are on the template strand of an expressed gene. **b**, TCR of template strand lesions is dependent on transcription level (P15 liver, median transcripts per million (TPM)). Mutation rate estimates (circles) are the aggregate rates for expression level binned genes across C3H tumours (n=371). Expression level bin 0 contains n=2,835 genes, all subsequent bins contain n=4,351±1 genes (inclusion criteria, see Methods); empiric confidence intervals (99%) were calculated through bootstrap sampling (n=100 replicates) of genes within each bin. **c**, Comparison of template versus non-template mutation rates for the 64 trinucleotide contexts: each context has a high and a low expression point linked by a line. **d**, Sequence composition normalised profiles of mutation rate around transcription start sites (TSS). **e**, Stratifying by lesion strand reveals how bidirectional transcription initiation shapes the observed mutation patterns. **f**, Higher resolution of TSS region, from panel above (**e**).

An engine for genetic diversity

A segregating lesion may template multiple rounds of replication in successive cell cycles (**Fig. 2i**). Each replication could incorporate different incorrectly - or even correctly - paired nucleotides opposite a persistent lesion, resulting in multiple alleles at the same position. Consistent with this notion, multi-allelic mutations have been reported in human cancers[24] and a cell lineage tracking system[25].

We evaluated multi-allelic variation by identifying sites with multiple high-confidence, but conflicting, mutation calls. On average, 8% of mutated sites in DEN-induced tumours have multi-allelic variants (n=1.8 million sites in C3H tumours); per tumour, this value ranges from <1% to 26% (**Fig. 4a**). As a control, only 0.098% (95% CI: 0.043-0.25%) of sites permuted between tumours show evidence of non-reference nucleotides. We further validated WGS multi-allelic variant calls using independently performed exome sequencing[6] (**Fig. 4b**).

The generation of multi-allelic variation produces combinatorial genetic diversity not expected under purely clonal expansion. This can be directly visualised in pairs of mutations spanned by individual sequencing reads (**Fig. 4c,d**). The observed combinations of biallelic sites require replication over lesions without the generation of mutations in some cell divisions (**Fig. 4d**). This directly demonstrates that non-mutagenic synthesis over DNA lesions occurs, and allele frequency analysis indicates it is common (**Extended Data Fig. 6**). The per-tumour rates of combinatorial diversity and multi-allelic sites correlate closely and highlight the wide variation between tumours (**Fig. 4e**).

The explanation for such inter-tumour variance becomes evident when plotting the distribution of multi-allelic sites along each genome (**Fig. 4f-i**). Tumours with high rates of genetic diversity have consistently high rates of multi-allelism throughout their genome (**Fig. 4g**). They likely expanded from a first-generation daughter of the original DEN mutagenised cell, in which all DNA is a duplex of a lesion containing and non-lesion containing strand. Therefore, replication over lesion containing strands in subsequent generations produces multi-allelic variation uniformly across the genome. Tumours with lower total levels of genetic diversity exhibit discrete genomic segments of high and low multi-allelism (**Fig. 4h,i**). These tumours likely developed from a cell some generations subsequent to DEN treatment. Each mitosis following DEN exposure is expected to dilute the number of lesion containing strands in each daughter cell by approximately 50%. Only lesion-retaining fractions of the genome generate multi-allelic and combinatorial genetic diversity in the daughter lineages; consistent with this, the multi-allelic segments mirror the mutational asymmetry segmentation pattern.

By estimating the fraction of multi-allelic chromosomal segments, we can infer the cell generation post-DEN exposure that the tumour expanded from (**Fig. 4j**). In 67% of C3H and 21% of CAST tumours the initial burst of mutations was instantly transformative. For the remainder, the observed fractions of multi-allelic segments cluster around expectations for subsequent cell generations, suggesting that transformation required a specific mutation allele combination, an additional mutation, or an external trigger. Intriguingly, *Egfr* driven tumours appear to transform significantly later, suggesting that driver gene identity influences the timing of tumour inception (**Fig. 4k**).

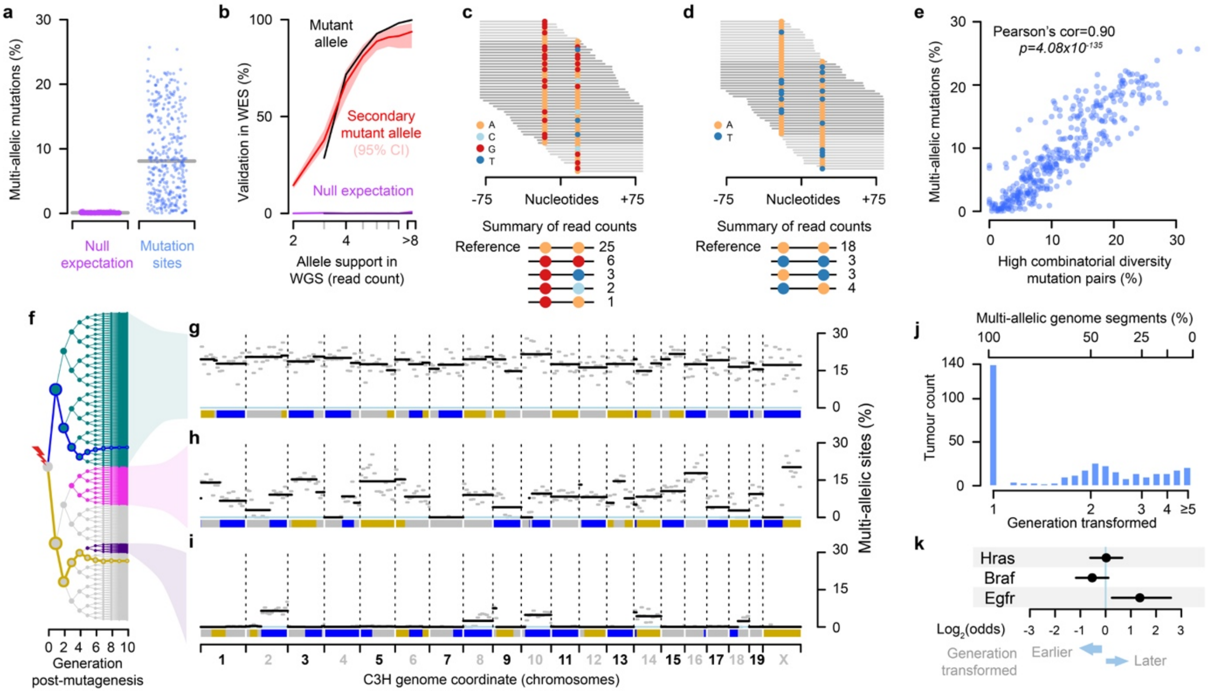


Fig.4 | Lesion segregation generates multi-allelic and combinatorial genetic diversity. **a,** Mutation sites per tumour with robust support for multi-allelic variation; grey line is median, null expectation from permutation between tumours. **b**, Validation rate for whole-genome (WGS) mutations in independent exome sequencing (WES); n=15 tumours, collectively with n=20,683 WGS mutations meeting inclusion criteria (Methods). Curves show validation rate stratified by WGS read support. Empiric confidence interval (95%) from 100 bootstrap samplings of the aggregated WGS mutations. The null expectation permuted tumour identity between WGS and WES. **c,** Sequence reads spanning proximal mutated sites. **d**, As (**c**), showing combinatorial diversity between a pair of biallelic sites. **e**, Correlation between per-tumour multi-allelic rate and combinatorially diverse mutation pairs (as in **c**,**d**), one point per tumour. **f**, Tree of all possible progeny of a DEN mutagenised cell for 10 generations. Blue and gold lines trace simulated segregation of lesion-containing strands from a single haploid chromosome. Coloured nodes show hypothetical transformed daughter lineages with their multi-allelic patterns shown, right. **g-i**, Mutation asymmetry summary ribbons for example C3H tumours that show high (**g**), variable (**h**), or low (**i**) rates of genetic diversity. The percent of mutation sites with robust support for multi-allelic variation calculated in 10Mb windows (grey) and for each asymmetric segment (black). **j**, Histogram of the estimated cell generation post-DEN exposure from which tumours developed based on the proportion of multi-allelic segments. **k**, Enrichment of specific driver gene mutations in earlier (generation 1) and later (generation >1) transforming tumours. Log2(odds) ratios (circles) from Fisher's exact test with 95% confidence intervals (whiskers) calculated from the hypergeometric distribution. All n=371 tumours were included in the analysis for each gene.

Lesion segregation is ubiquitous

Lesion segregation is a feature of DEN mutagenesis in mice. This raises two critical questions: Do other DNA damaging agents induce lesion segregation? Does lesion segregation occur in human cells and cancers? Recently, human induced pluripotent stem cells (iPSCs) exposed to 79 environmental mutagens revealed 41 of them produced excess nucleotide substitutions[5]. Although not previously noted in these *in vitro* data, many of the exposures generated chromosome-scale lesion segregation patterns (**Extended Data Fig. 7**) similar to our *in vivo* DEN model. Applying runs based tests (**Fig. 5a,b; Extended Data Fig. 8**) we detect significant mutational asymmetry in every sample with >1,000 informative mutations (**Fig. 5b, Extended Data Fig. 8b**), including clinically relevant insults: sunlight (simulated solar radiation, SSR), tobacco smoke (BPDE) and chemotherapeutics (temozolomide). In contrast, mutations induced by perturbation of replication and repair pathways[26] independent of DNA lesions, showed no detectable asymmetry, as expected (**Extended Data Fig. 8c**). We conclude that the chromosome-scale segregation of lesions, and the resulting strand asymmetry of mutation patterns, is a general feature of all tested DNA damaging mutagens.

The striking mutation asymmetry observed in both DEN-induced tumours and mutagen exposed human iPSCs[5] occurs after a single mutagenic insult. By contrast, most human cancers accumulate mutations as a result of multiple damaging events over their history. Lesion segregation predicts that such tumours will acquire new waves of segregating lesions after each exposure, thus progressively masking their asymmetry patterns. Therefore, even though UV exposure does cause striking lesion segregation in human cells (**Fig. 5a,b; Extended Data Fig. 7a, 8b**), it is unlikely that skin cancers would show mutational asymmetry following repeated UV exposure.

Nevertheless, exploring human cancer genomes[27] (n=18,850 tumours, 22 primary sites) identified multiple cancers with the characteristic mutational asymmetry of lesion segregation (**Fig. 5c,d**). The majority of these tumours are renal, hepatic or biliary in origin, and show a high mutation rate and strand asymmetry of T→A/A→T mutations, consistent with known aristolochic acid exposure[3] (**Supplementary Table 2**). Though visualised most clearly in tumours subjected to a single dose of a mutagen, lesion segregation has likely shaped all genomes subjected to DNA damage, with important implications for tumour evolution and heterogeneity.

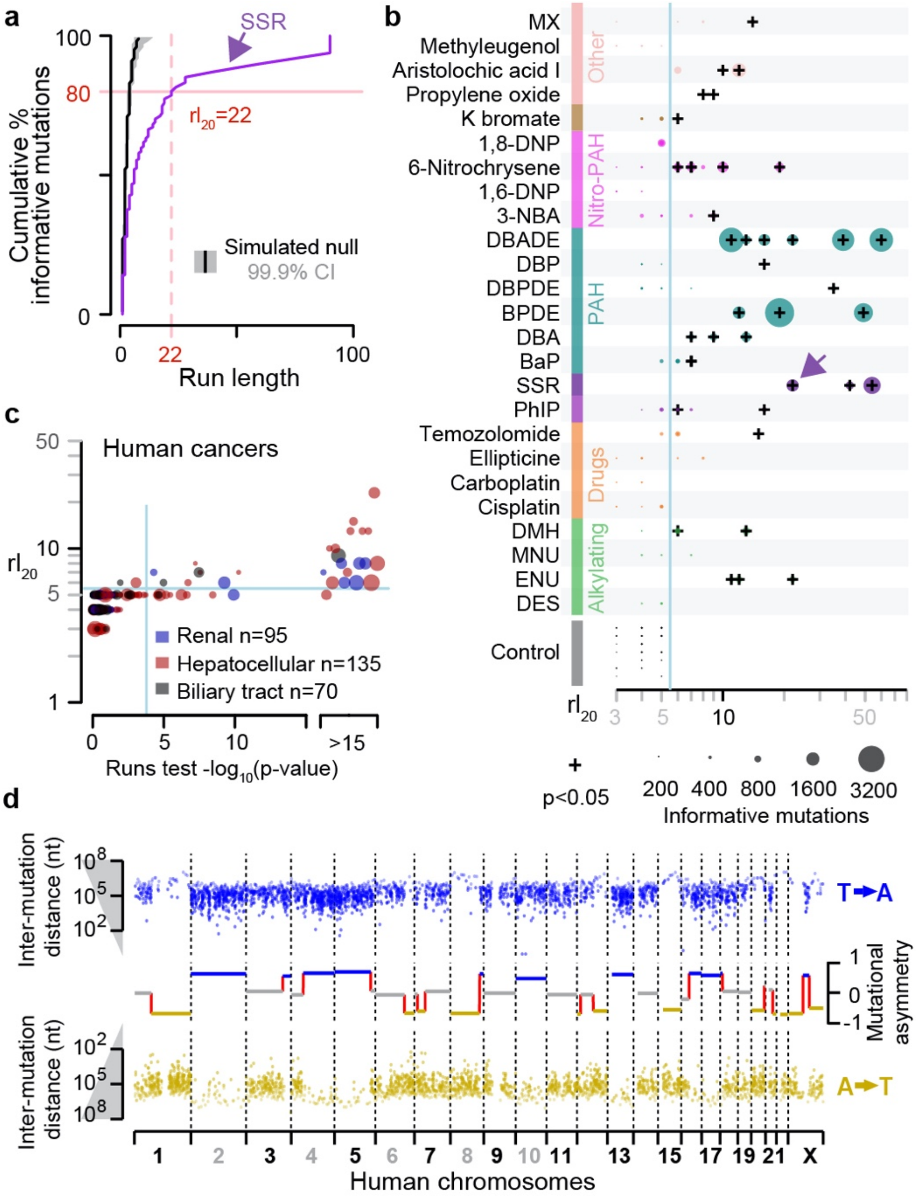


Fig.5 | Lesion segregation is a pervasive feature of exogenous mutagens and is evident in human cancers. **a**, The runs-based rl$_{20}$ metric, calculated for an example simulated solar radiation (SSR) clone (**Extended Data Fig. 7a**); 20% of informative mutations (C→T/G→A) are in strand asymmetric runs of ≥22 consecutive mutations (e.g. ≥22 C→T without an intervening G→A). Simulated null based on 100,000 permutations of 1,000 mutations, black curve shows median. **b**, All robust mutagens in human iPSCs[5] (**Supplementary Table 2**). The rl$_{20}$ metric (x-axis) is plotted for each clone (n=325), including multiple replicates per exposure, data point size quantifies informative mutations. "+" Bonferroni corrected (two-sided) p-value <0.05. **c**, The rl$_{20}$ metric and runs tests for human cancers[27], n=18,850 cancers screened, three cohorts plotted. Blue lines show Bonferroni adjusted p=0.05 threshold for the runs test (two-sided) and an empiric threshold for rl$_{20}$ (Methods). X-axis p-values < 1x10^{-15} are rank-ordered. **d**, Mutational asymmetry (plotted as per **Fig. 2a-c**) in a human hepatocellular carcinoma (donor DO231953) with a dominant mutation signature for aristolochic acid exposure.

Discussion

Here we have shown that most mutation-causing DNA lesions are not resolved as mutations within a single cell-cycle. Instead, lesions segregate unrepaired into daughter cells for multiple cellular generations, resulting in chromosome-scale strand asymmetry of subsequent mutations. This suggests that lesion removal prior to replication is high fidelity, rarely resulting in mutations. Initially discovered in a well-powered *in vivo* mouse model of oncogenesis, we demonstrate that lesion segregation is ubiquitous to all tested mutagens, occurs in human cells, and is evident in human cancers. Similar patterns of asymmetry in bacterial mutagenesis posit that the underlying mechanisms are deeply conserved[28,29].

Our discovery of lesion segregation challenges long-standing assumptions of cancer evolution[30]. For example, the widely-used infinite sites model[31] does not allow for recurrent mutation at the same site. Our findings also provide new perspectives to understand cancer evolution, using mutational asymmetry and multi-allelism patterns to track events during oncogenesis and to quantify selection. Perhaps most importantly, lesion segregation is a previously unrecognised mechanism for a cancer to sample the fitness effects of mutation combinations, thus evading Muller's ratchet[32] and Hill-Robertson interference, which assumes low selection efficiency due to the inability to separate mutations of opposing fitness[33,34]. Consequently, DNA damaging chemotherapeutics, particularly large or closely spaced doses generating persistent lesions, could inadvertently provide an opportunity for cancer to efficiently select resulting mutations. This insight may guide the development of more effective chemotherapeutic regimens.

Once identified, lesion segregation is a deeply intuitive concept. Its practical applications provide new vistas for the exploration of genome maintenance and fundamental molecular biology. The discovery of pervasive lesion segregation profoundly revises our understanding of how the architecture of DNA repair and clonal proliferation can conspire to shape the cancer genome.

References

1. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
3. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
4. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
5. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821–836 (2019).
6. Connor, F. *et al.* Mutational landscape of a chemically-induced mouse model of liver cancer. *J. Hepatol.* **69**, 840–850 (2018).
7. Maronpot, R. R. Biological Basis of Differential Susceptibility to Hepatocarcinogenesis among Mouse Strains. *J. Toxicol. Pathol.* **22**, 11–33 (2009).
8. Wang, C. *et al.* Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. *Nat. Commun.* **9**, 2054 (2018).
9. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
10. Verna, L., Whysner, J. & Williams, G. M. N-nitrosodiethylamine mechanistic data and risk assessment: bioactivation, DNA-adduct formation, mutagenicity, and tumor initiation. *Pharmacol. Ther.* **71**, 57–81 (1996).
11. Maronpot, R. R., Fox, T., Malarkey, D. E. & Goldsworthy, T. L. Mutations in the ras proto-oncogene: clues to etiology and molecular pathogenesis of mouse liver tumors. *Toxicology* **101**, 125–156 (1995).
12. Buchmann, A., Karcier, Z., Schmid, B., Strathmann, J. & Schwarz, M. Differential selection for B-raf and Ha-ras mutated liver tumors in mice with high and low susceptibility to hepatocarcinogenesis. *Mutat. Res.* **638**, 66–74 (2008).
13. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538–549 (2016).
14. Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
15. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
16. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
17. Perry, P. & Evans, H. J. Cytological detection of mutagen-carcinogen exposure by sister chromatid exchange. *Nature* **258**, 121–125 (1975).
18. Guirouilh-Barbat, J., Lambert, S., Bertrand, P. & Lopez, B. S. Is homologous recombination really an error-free process? *Front. Genet.* **5**, 175 (2014).
19. Strick, T. R. & Portman, J. R. Transcription-Coupled Repair: From Cells to Single Molecules and Back Again. *J. Mol. Biol.* **431**, 4093–4102 (2019).
20. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human

global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29**, 948–960 (2015).

21. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170**, 534–547.e23 (2017).
22. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
23. Preker, P. *et al.* PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res.* **39**, 7179–7193 (2011).
24. Kuipers, J., Jahn, K., Raphael, B. J. & Beerenwinkel, N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* **27**, 1885–1894 (2017).
25. Brody, Y. *et al.* Quantification of somatic mutation flow across individual cell division events by lineage sequencing. *Genome Res.* **28**, 1901–1918 (2018).
26. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
27. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
28. Parkhomchuk, D., Amstislavskiy, V., Soldatov, A. & Ogryzko, V. Use of high throughput sequencing to observe genome dynamics at a single cell level. *Proceedings of the National Academy of Sciences* vol. 106 20830–20835 (2009).
29. Chan, K. & Gordenin, D. A. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu. Rev. Genet.* **49**, 243–267 (2015).
30. Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**, 213–229 (2017).
31. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903 (1969).
32. Zhang, Y. *et al.* Genetic Load and Potential Mutational Meltdown in Cancer Cell Populations. *Mol. Biol. Evol.* **36**, 541–552 (2019).
33. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
34. Tilk, S., Curtis, C., Petrov, D. & McFarland, C. D. Most cancers carry a substantial deleterious load due to Hill-Robertson interference. *bioRxiv* 764340 (2019) doi:10.1101/764340.

Methods

Mouse colony management

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). Animals were maintained using standard husbandry: mice were group housed in Tecniplast GM500 IVC cages with a 12-hour light / 12-hour dark cycle and *ad libitum* access to water, food (LabDiet 5058), and environmental enrichments.

Chemical model of hepatocarcinogenesis

15-day-old (P15) male C3H and CAST mice were treated with a single intraperitoneal (IP) injection of N-Nitrosodiethylamine (DEN; Sigma-Aldrich N0258; 20 mg/kg body weight) diluted in 0.85% saline. Liver tumour samples were collected from DEN-treated mice 25 weeks (C3H) or 38 weeks (CAST) after treatment. All macroscopically identified tumours were isolated and processed in parallel for DNA extraction and histopathological examination. Non-tumour tissue from untreated P15 mice (ear, tail, and background liver) was sampled for control experiments.

Tissue collection and processing

Liver tumours of sufficient size (≥2 mm diameter) were bisected; one half was flash frozen in liquid nitrogen and stored at -80°C for DNA extraction, and the other half was processed for histology. Tissue samples for histology were fixed in 10% neutral buffered formalin for 24 h, transferred to 70% ethanol, machine processed (Leica ASP300 Tissue Processor; Leica, Wetzlar, Germany), and paraffin embedded. All formalin-fixed paraffin-embedded (FFPE) sections were 3 μm in thickness.

Histochemical staining

FFPE tissue sections were haematoxylin and eosin (H&E) stained using standard laboratory techniques. Histochemical staining was performed using the automated Leica ST5020; mounting was performed on the Leica CV5030.

Imaging

Tissue sections were digitised using the Aperio XT system (Leica Biosystems) at 20x resolution; all H&E images are available in the BioStudies archive at EMBL-EBI under accession S-BSST383.

Tumour histopathology

H&E sections of liver tumours were blinded and assessed twice by a pathologist (S.J.A); discordant results were reviewed by an independent hepatobiliary pathologist (S.E.D). Tumours were classified according to the International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice (INHAND) guidelines[35]. In addition, tumour grade, size, morphological subtype, nature of steatosis, and mitotic index were assessed (**Supplementary Table 1**), as well as the presence of cystic change, haemorrhage, necrosis, or vascular invasion.

Sample selection for WGS

Tumours which met the following histological criteria were selected for whole genome sequencing (C3H n=371, CAST n=84): (i) diagnosis of either dysplastic nodule (DN) or hepatocellular carcinoma (HCC), (ii) homogenous tumour morphology, (iii) tumour cell percentage >70%, and (iv) adequate tissue for DNA extraction. Neoplasms with extensive necrosis, mixed tumour types, a nodule-in-nodule appearance (indicative of an HCC arising within a DN), or contamination by normal liver tissue were excluded. Since carcinogen-induced tumours arising in the same liver are independent[6], multiple tumours were selected from each mouse to minimise the number of animals used. A subset of normal (non-tumour) samples from untreated mice were also sequenced (C3H n=13, CAST n=7).

Whole genome sequencing

Genomic DNA was isolated from liver tissue and liver tumours using the AllPrep 96 DNA/RNA Kit (Qiagen, 80311) according to the manufacturer's instructions. DNA quality was assessed on a 1% agarose gel and quantified using the Quant-IT dsDNA Broad Range Kit (Thermo Fisher Scientific). Genomic DNA was sheared using a Covaris LE220 focused-ultrasonicator to a 450 bp mean insert size.

WGS libraries were generated from 1 µg of 50 ng/ul high molecular weight gDNA using the TruSeq PCR-free Library Prep Kit (Illumina), according to the manufacturer's instructions. Library fragment size was determined using a Caliper GX Touch with a HT DNA 1k/12K/Hi Sensitivity LabChip and HT DNA Hi Sensitivity Reagent Kit to ensure 300-800 bp (target ~450 bp).

Libraries were quantified by real-time PCR using the Kapa library quantification kit (Kapa Biosystems) on a Roche LightCycler 480. 0.75 nM libraries were pooled in 6-plex and sequenced on a HiSeq X Ten (Illumina) to produce paired-end 150 bp reads. Each pool of 6 libraries was sequenced over eight lanes (minimum of 40x coverage).

Variant calling and somatic mutation filtering

Sequencing reads were aligned to respective genome assemblies (C3H = C3H_HeJ_v1; CAST = CAST_EiJ_v1)[36] with bwa-mem (v.0.7.12)[37] using default parameters. Reads were annotated to read groups using the picard (v.1.124)[38] tool AddOrReplaceReadGroups, and minor annotation inconsistencies corrected using the picard CleanSam and FixMateInformation tools. Bam files were merged as necessary, and duplicate reads were annotated using the picard tool MarkDuplicates.

Single nucleotide variants were called using Strelka2 (v.2.8.4)[39] implementing default parameters. Initial variant annotation was performed with the GATK (v.3.8.0)[40] walker CalculateSNVMetrics[41]. Genotype calls with a variant allele frequency < 0.025 were removed. Although inbred strains were used, fixed genetic differences between the colonies and the reference genome, as well as small numbers of germline variants segregating within the colonies were identified. For each strain, fixed differences identified as homozygous changes present in 100% of genotyped samples were filtered out. Segregating variants were filtered based on the excess clustering of mutations to animals with shared mothers. To generate a null expectation taking into account the family structure of the colonies, the parent-offspring relationships were randomly permuted 1,000 times. For each count of recurrent mutation (range 5 to 371 inclusive), we determined the null distribution of expected distinct mothers. Comparing this to the observed count of distinct mothers for each recurrent (n>4) mutation, those with a low probability ($p<1x10^{-4}$, pnorm function from R (v.3.5.1)[42]) under the null were excluded from analyses.

Copy number variation between tumours within strains was called using CNVkit (v.0.9.6)[43]. Non-tumour reference coverage was provided from non-tumour control WGS data (C3H n=11, CAST n=7) and per tumour cellularity estimates (see below) were provided.

RNA-sequencing

Total RNA was extracted from P15 liver tissue (n=4 biological replicates per strain) using QIAzol Lysis Reagent (Qiagen), according to manufacturer's instructions. DNase treatment and removal were performed using the TURBO DNA-freeTM Kit (Ambion, Life Technologies), according to manufacturer's instructions. RNA concentration was measured using a NanoDrop spectrophotometer (Thermo Fisher); RNA integrity was assessed on a Total RNA Nano Chip Bioanalyzer (Agilent).

Total RNA (1 µg) was used to generate sequencing libraries using the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina), according to manufacturer's instructions. Library fragment size was determined using a 2100 Bioanalyzer (Agilent). Libraries were quantified by qPCR (Kapa Biosystems). Pooled libraries were sequenced on a HiSeq4000 to produce ≥40 million paired-end 150 bp reads per library.

RNA-seq data processing and analysis

Transcript abundances were quantified with Kallisto (v.0.43.1)[44] (using the flag --bias) and a transcriptome index compiled from coding and non-coding cDNA sequences defined in Ensembl v91[45]. Transcripts per million (TPM) estimates were generated for each annotated transcript and summed across alternate transcripts of the same gene for gene-level analysis. Transcription start sites (TSS) for each gene were annotated with Ensembl v91 and based upon the most abundantly expressed transcript. RNA-seq data are available at Array Express at EMBL-EBI under accession E-MTAB-8518.

Genomic annotation data

Mouse liver proximity ligation sequencing (HiC) data were downloaded from GEO (GSE65126)[46], replicates were combined, then aligned to GRCm38[47] and processed using the Juicebox (v.7.5) and Juicer scripts[48] to obtain the HiC matrix. Eigenvectors were obtained for 500kb consecutive genomic windows over each chromosome from the HiC matrix using Juicebox and subsequently oriented (to distinguish compartment A from B) using GC content per 500kb bin. We used progressiveCactus[49] to project the 500kb windows into the C3H reference genome and Bedtools (v.2.28.0) to merge syntenic loci between 450 and 550 kb in size, removing the second instance where we observed overlaps.

Genic annotation was obtained from Ensembl v91[45] for the corresponding C3H and CAST reference genome assemblies (C3H_HeJ_v1, CAST_EiJ_v1). Genomic repeat elements were annotated using RepeatMasker (v.20170127)[50] with the default parameters and libraries for mouse annotation.

The analysable fraction of the genome

Analysis and sequence composition calculations were confined to the main chromosome assemblies of the reference genome (chromosomes 1-19 and X). Using WGS of non-tumour liver, ear and tail samples (C3H n=11, CAST n=7) collected and sequenced contemporaneously with tumour samples, genome sequencing coverage was calculated for 1kb windows using multicov in Bedtools (v.2.28.0)[51]. Windows with read coverage >2 s.d. from the autosomal mean were flagged as suspect in each tumour. Read coverage over the X chromosome was doubled in these calculations to account for the expected hemizygosity in these male mice. Any 1kb window identified as suspect in >90% of these non-tumour samples was flagged as "abnormal read coverage" (ARC) and masked from subsequent analysis. This masked 12.7% of the C3H and 11.5% of the CAST reference genomes yielding analysable haploid genome sizes of C3H = 2,333,783,789 nt and CAST = 2,331,370,397 nt.

Mutation rate calculations

Mutation rates were calculated as 192 category vectors representing every possible single nucleotide substitution conditioned on the identity of the upstream and downstream nucleotides. Each rate being the observed count of a mutation category divided by the count of the trinucleotide context in the analysed sequence. To report a single aggregate mutation rate, the three rates for each trinucleotide context were summed to give a 64 category vector and the weighted mean of that vector reported as the mutation rate. The vector of weights being the trinucleotide sequence frequency of a reference sequence, for example the composition of the whole genome. In the case of whole genome analysis, the same trinucleotide counts are used in (1) the individual category rates calculation and (2) the weighted mean of the rates, cancelling out. For windowed comparisons of mutation rates, the weighted mean is calculated using the genome wide composition of trinucleotides rather than the local sequence composition, providing a compositionally adjusted mutation rate estimate. For mutation rates in TCR analysis, the same compositional adjustment was carried out but using the trinucleotide composition of the aggregate genic spans of genome (minus ARC regions) for normalisation.

Mutation signatures

The 96 category "folded" mutation counts for each of the 371 C3H tumours were deconvolved into the best fitting number (*K*) of component signatures using sigFit (v.2.0)[52] with 1,000 iterations and *K* set to integers 2 to 8 inclusive. A heuristic goodness of fit score based on cosine similarity favoured instances where *K*=2. The DEN1 and DEN2 signatures reported were obtained by running sigFit with 30,000 iterations for *K*=2. Analysis of CAST tumours gave less distinct separation of signatures so the C3H derived DEN1 and DEN2 were used for both strains. To fit signatures to each tumour we used sigFit provided with the DEN signatures and additional SPONT1 and SPONT2 signatures that were derived from equivalent WGS analysis of spontaneous (non-DEN induced) C3H tumours.

Driver mutation identification

Candidate cancer driver genes were identified by applying oncodriveFML (v.2.2.0 using the SIFT scoring scheme)[53] and oncodriveCLUSTL (v.1.1.1)[54] to mutations identified in C3H tumours. The only genes convincingly identified as significantly enriched for functionally impactful or clustered mutations were *Braf*, *Egfr* and *Hras*. *Kras* appeared as marginally significant. These four genes were identified for C3H[6]. Protein altering mutations in those genes were annotated as driver mutations in C3H and CAST tumours.

Mutational asymmetry segmentation and scoring

For each tumour a focal subset of "informative" mutation types were defined, T→N/A→N mutations, in the case of DEN-induced tumours. The order of focal mutations along each chromosome was represented as a binary vector (e.g. 0 for T→N, 1 for A→N). Vectors corresponding to each chromosome of each tumour were processed with the cpt.mean function of the R Changepoint (v.2.2.2)[55] package run with an Akaike information criterion (AIC) penalty function, maximum number of changepoints set to 12 (Q=12), and implementing the PELT algorithm for optimal changepoint detection. Following segmentation, the defined segments were scored for strand asymmetry, taking into account the sequence composition of the segment. For example in tumours with T→N/A→N informative mutations the number of Ts on the forward strand is the count of Watson sites G_W and the number of T→N mutations is μ_W which together give the Watson strand rate $R_W=\mu_W/G_W$. The forward strand count of As and mutations from A likewise give the Crick strand rate $R_C=\mu_C/G_C$. From these two rates we calculate a relative difference metric, the mutational asymmetry score $S=(R_W-R_C)/(R_W+R_C)$.

The parameter S scales from 1 all Watson (e.g. DEN T→N mutations) through 0 (50:50 T→N:A→N) to -1 for all Crick (e.g. DEN A→N). For the categorical assignment, $S \geq 0.3$ is Watson strand asymmetric, $S \leq -0.3$ Crick strand asymmetric and in the range $-0.3 < S < 0.3$ symmetric, though more stringent filtering was applied where noted. Segments containing <20 informative mutations were discarded from subsequent analyses.

To test for oncogenic selection at sites with recurrent mutations, mutational asymmetry segments overlapping the focal mutation were categorised based on their asymmetry score S, as above. The test was implemented as a Fisher's exact test with the 2x2 contingency table comprising the counts chromosomes (two autosomes per cell) stratified by Watson versus Crick asymmetry and the presence of the focal mutation in the tumour. Tumours containing another known driver gene or recurrent mutation within the focal asymmetry segment were discarded from the analysis. We estimated the minimum recurrence of a mutation necessary to reliably detect oncogenic selection through simulation. Biased segregation of chromosomes containing drivers was modelled using the observed median excess of T→N over A→N lesions (23 fold), and random segregation of non-driver containing strands (1:1 ratio). Our model predicted >33 C3H recurrences or >41 CAST recurrences would give 80% power to detect oncogenic selection if present.

Tumour cellularity estimates

We calculated tumour cellularity as a function of the non-reference read count in autosomal chromosomes $(1-R/d)*2$ where R is the reference read count at a mutated site and d is the total read depth at the site. For each tumour these values were binned in percentiles and the midpoint of the most populated (modal) percentile taken as the estimated cellularity of the tumour. Given the low rate of copy number variation across the DEN induced tumours, no correction was made for copy-number distortion. Skew in the variant allele frequency (VAF=$(1-R/d)$) distribution was calculated using Pearson's median skewness coefficient implemented in R as (3(mean-median))/sd of the VAF distribution.

Identifying and filtering reference genome mis-assemblies

Since lesion segregation, mutation asymmetry patterns allow the long-range phasing of chromosome strands, they can detect discrepancies in sequence order and orientation between the sequenced genomes and the reference. We identified autosomal asymmetry segments that immediately transitioned from Watson bias ($S > 0.3$) to Crick ($S < -0.3$) or vice versa without occupying the intermediate unbiased state ($-0.3 > S < 0.3$); such discordant segments are unexpected. Allowing for ±100kb uncertainty in the position of each exchange site we produced the discordant segment coverage metric. At sites with discordant segment coverage >1 we calculated percentage consensus for mis-assembly $M=ds/(ds+cs)$ where ds is the number of discordant segments over the exchange site and cs the number of concordant: where either Watson or Crick mutational asymmetry extends at least $1x10^6$ nucleotides on both sides of the exchange site. The approximate genomic coordinates for a C3H strain specific inversion on chromosome 6 were previously reported[56].

Sister chromatid exchange site analysis

Identified SCE sites were aggregated across tumours from each strain. Exchange sites within $1x10^6$ nt of known and proposed reference genome mis-assembly sites were excluded from analysis. The mid-point between the flanking informative mutations was taken as the reference genome position of the exchange event, and the distance between those flanking mutations as the positional uncertainty of the estimate. To generate null expectations for mutation rate measures, the coordinate of an exchange was projected into the genome of a proxy tumour and the mutation rates and patterns measured from that proxy tumour (repeated 100 times). The permutation of tumour identifiers for the selection of proxy tumours was a shuffle without replacement that preserved the total number of exchange sites measured in each tumour.

The comparison of mutation spectra between windows was calculated as the cosine distance between the 96 category trinucleotide context mutation spectra for the whole genome and that calculated for the aggregated 5kb window. The 96 categories were equally weighted for this comparison.

Exchange site enrichment analysis used Bedtools[51] shuffle to permute the genomic positions of exchange sites into the analysable fraction of the genome (defined above). Observed rates of annotation overlap were compared to the distribution of values from 1,000 permuted exchange sites. For genic overlaps we used Ensembl v91[45] coordinates for genic spans; gene expression status was based on the summed expression over all annotated transcripts for the gene from P15 liver from the matched mouse strain. Expression thresholds were defined as >50th centile for active and <50th centile for inactive genes.

A higher count of informative mutations provides greater power to identify shorter mutational asymmetry segments. To fairly test for correlation between nucleotide substitution rate and SCE rate we randomly down-sampled informative mutations to 10,000 per tumour genome and recomputed the mutational asymmetry segmentation patterns from the sampled data. Tumours with <10,000 informative mutations were excluded. We then correlated the total (not down sampled) nucleotide substitution load to the count of SCE events inferred from the down-sampled data.

Transcription coupled repair calculations

For each protein coding gene, the maximally expressed transcript isoform was identified from P15 liver in the matched strain (TPM expression), subsequently the primary transcripts. In the case of ties, transcript selection was arbitrary. Genes were partitioned into five categories based on the expression of the primary transcript: expression level 0 (<0.0001 TPM) and four quartiles of detected expression.

Using the segmental asymmetry patterns of each tumour and the annotated coordinates (Ensembl v91) of the selected transcripts, we identified transcripts completely contained in a single Watson or Crick asymmetric segment and located at least 200kb from the segment boundary at both ends. We also applied strict asymmetry criteria of mutational asymmetry scores $S > 0.8$ for Watson and $S < -0.8$ for Crick asymmetry segments, though analysis with the standard asymmetry thresholds and no segment boundary margin give similar results and identical conclusions. For each transcript in each tumour we then used both the transcriptional orientation of the gene and the mutational asymmetry of the segment containing it to resolve the segregated lesions to either the template (anti-sense) or non-template (sense) strand of the gene. Transcripts contained in mutationally symmetric regions or not meeting the strict filtering criteria were excluded from analysis.

We then analysed mutation rates stratifying by gene expression level and the template/non-template strand of the lesions but aggregating between tumours within the same strain. The transcription start site coordinates used correspond to the annotated 5' end of the primary transcripts.

Multi-allelic variation

Aligned reads spanning genomic positions of somatic mutations were re-genotyped using Samtools mpileup (v.1.9)[57]. Genotypes supported by ≥2 reads with a nucleotide quality score of ≥20 were reported, considering sites with two alleles as biallelic, those with three or four alleles as multi-allelic. The fraction of called mutations exhibiting multi-allelic variation was calculated for the analysable fraction of the genome, across 10Mb consecutive windows and also for each of the mutational asymmetry segments calculated for each tumour.

A null expectation for the multi-allelic rate estimate was generated per C3H tumour; genomic positions identified as mutated across the other 370 tumours were down-sampled to match the mutation count in the focal tumour. Any of these proxy mutation sites with a non-reference genotype supported by ≥2 reads and nucleotide quality score ≥20 at the focal site were referred to as "multi-allelic" for the purposes of defining a background expectation for the calling of multi-allelic variation. For each tumour, this was repeated 100 times and the mean reported.

We used whole exome sequencing (WES) of fifteen C3H tumours from prior work[6] that have subsequently been used to generate WGS data in this study as a basis for validating multi-allelic calls. Multi-allelic variant positions derived from WGS were genotyped in WES using Samtools mpileup, as described above. Only sites with ≥30x WES coverage were considered and alleles were found to be concordant if a WGS genotype was supported by ≥1 read in the WES data. To provide a null expectation, the analysis was repeated using WES data from a different tumour and validation rates reported for all versus all combinations of mismatched WGS-WES pairs (n=15^2-15=210).

To quantify combinatorial genetic diversity for each tumour, pairs of mutations located between 3-150nt apart were phased using sequencing reads that traversed both mutation sites. Distinct allelic combinations were counted after extraction with Samtools mpileup using only reads with nucleotide quality score ≥20 over both mutation sites.

Estimating the cell generation of transformation

Knowing the fraction of lesion segregation segments that generated multi-allelic variation across a tumour genome allows the inference of the generation time post-mutagenesis of the cell from which the tumour developed, because each successive cell generation is expected to retain only 50% of the lesion containing segments. We estimate this fraction as follows. Let p denote the fraction of multi-allelic segments and let q be its complement, i.e. the fraction of non-multi-allelic segments, for each tumour genome. Segment boundaries being SCE sites or chromosome boundaries. In order to determine p, we re-purpose the quadratic Hardy-Weinberg equation: $p+q=p^2+2pq+q^2=1$, which holds since the two possible fractions need to sum to unity. Given an asymmetric segment of interest in the diploid genome, there are 3 distinct scenarios: (i) both chromosomes are multi-allelic (p^2), (ii) One of the chromosomes is multi-allelic and the other is not ($pq+qp$) and (iii) both chromosomes are non-multi-allelic (q^2). The first two scenarios are not distinguishable from the data as both appear multi-allelic (m). However, in the third scenario, for a segment to be non-multi-allelic (biallelic, b), both chromosomal copies have to be non-multi-allelic. As described below, q^2 can be estimated directly from the data and is subsequently used to estimate $p=1\text{-}sqrt(q^2)$ and hence the cell generation number of transformation post-mutagenesis.

The estimation of q^2 requires computing the ratio $q^2=b/(b+m)$. We can directly observe the counts of b as non-multi-allelic segments. The number of autosomal chromosome pairs (n=19) and count of sister chromatid exchange events (x) give the total number of segments in the genome $b+m=n+x$. Exchange events are not expected to align between allelic chromosomes which will result in the partial overlap of segments between allelic copies. Although this increases the number of observed segments (b and m) relative to actual segments, assuming the independent behaviour of allelic chromosomes and that segment length is independent of multi-allelic state, this partial overlap does not systematically distort the quantification of b or the estimation of q^2.

To call a non-multi-allelic segment (b) we require less than 0.04% multi-allelic sites. The threshold is based on the tri-modal frequency distribution of multi-allelic rates per-segment, aggregated over all 371 C3H tumours. The 0.04% threshold separates the lower distribution of multi-allelic rates from the mid and higher distributions.

To test for the enrichment of specific driver gene mutations in early generation versus late generation transformation post-DEN treatment, we applied Fisher's exact test (fisher.test function in R) to compare the generation 1 ratio of tumours with, versus those without a focal mutation, to the same ratio for tumours inferred to have transformed in a later generation. We additionally report the same odds ratios, but requiring that the "with focal mutation" tumours had a driver mutation in only one of the driver genes: *Hras, Braf*, or *Egfr*.

Cell-line and human cancer mutation analysis

Somatic mutation calls were obtained from DNA maintenance and repair pathway perturbed human cells[26]. Of the 128,054 reported single nucleotide variants, 6,587 unique mutations (genomic site and specific change) were shared between two or more sister clones, so likely represent mutations present but not detected in the parental clone. All occurrences of the shared mutations were filtered out leaving 106,688 mutations for analysis, although the inclusion of these filtered mutations does not alter any conclusions drawn. Somatic mutation calls from mutagen exposed cells[5] were obtained, no additional filtering was applied to these sub-clone mutations.

Somatic mutation calls from the International Cancer Genome Consortium (ICGC)[58] were obtained as simple_somatic_mutation.open.* files from release 28 of the consortium, one file for each project. These somatic mutations have been called from a mixture of whole genome and whole exome sequencing. Of the 18,965 patients represented (and not embargoed in the release 28 dataset), 116 were excluded from analysis; these represent a distinct whole exome sequenced subset of the LICA-CN project that appear to show a processing artefact in the distribution of specific mutation subsets. ICGC mutations were filtered to remove insertion and deletion mutations and also filtered for redundancy so that each mutation was only reported once for each patient. Mutation signatures deconvolution was performed using the R MutationPatterns (v.1.4.2)[59] package and COSMIC signature 22 was interpreted as aristolochic acid[3].

The rl_{20} metric and runs tests

Amongst only the informative mutations (e.g. T→N/A→N in DEN) three consecutive T→N without an intervening A→N is a run of three. The R function rle was used to encode the run-lengths for binary vectors of informative mutations along the genome of a focal tumour. Ranking them from the longest to the shortest run, we find the set of longest runs that encompass 20% of all informative mutations in the tumour. The run-length of the shortest of those is reported as the rl_{20} metric. The threshold percent of mutations was defined as having to be less than 50%, as on average only 50% of the autosomal genomes are expected to show mutational asymmetry patterns. On testing with randomised data, the value of 20% gave a stable null expectation (maximum observed value of a run of five when simulating 10,000 informative mutations) and still encompassed a large fraction of the informative mutations. All rl_{20} results reported were implemented so that runs were broken when crossing chromosome boundaries. To define an empirical significance threshold for genomes with fewer mutations, we simulated 1,000 random informative mutations 100,000 times, >99.995% simulations had $rl_{20} \leq 5$ and 100% $rl_{20} \leq 6$.

The Wald-Wolfowitz runs test was performed using the runs.test function of the R randtests (v.1.0)[60] library. It was applied to binary vectors of informative changes as described above, with threshold=0.5.

The Wald-Wolfowitz runs test significance is inflated by coordinated dinucleotide changes, such as those produced by UV light exposure and also other local mutational asymmetries such as replication asymmetry[13] and kataegis events[14,61]. The rl_{20} metric appears robust to most such distortions but we find it efficiently detects kataegis events that are in an otherwise mutationally quiet background, as is often the case for breast cancer. For this reason we also indicate the total genomic span of mutations in the rl_{20} subset of mutation runs: kataegis events typically span a tiny (<5%) fraction of the whole genome.

Computational analysis environment

Primary data processing was performed in shell-scripted environments calling the software indicated. Except where otherwise noted, analysis processing post-variant calling was performed in a Conda environment and choreographed with Snakemake running in an LSF batch control system (**Supplementary Table 3**). The analysis pipeline including Conda and Snakemake configuration files can be obtained from the repository https://git.ecdf.ed.ac.uk/taylor-lab/lce-ls.

Data availability

The WGS BAM files are available from the European Nucleotide Archive (ENA) under accession: PRJEB37808. RNA-seq files are available from Array Express E-MTAB-8518. Digitised histology images are available from Biostudies under accession S-BSST383.

Key resources

The key reagents and resources required to replicate our study are listed in **Supplementary Table 3**. For externally sourced data, where applicable, URLs that we used can be found in the Git repository https://git.ecdf.ed.ac.uk/taylor-lab/lce-ls.

Methods references

35. Thoolen, B. *et al.* Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. *Toxicol. Pathol.* **38**, 5S–81S (2010).
36. Lilue, J. *et al.* Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **50**, 1574–1583 (2018).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. Broad Institute. Picard Tools. *Broad Institute, GitHub Repository* http://broadinstitute.github.io/picard (2019).
39. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
40. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
41. Eldridge, M. gatk-tools: Utilities for processing sequencing data and genomic variants using GATK. https://github.com/crukci-bioinformatics/gatk-tools.
42. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
43. 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).
44. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-Seq quantification with kallisto. *Nat. Biotechnol.* **34**, 525–527 (2016).
45. Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2019).
46. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
47. Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of

the mouse. *PLoS Biol.* **7**, e1000112 (2009).

48. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
49. Armstrong, J. *et al.* Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era. *bioRxiv* 730531 (2019) doi:10.1101/730531.
50. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0., http://www.repeatmasker.org (2013-2015).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
52. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv* 372896 (2018) doi:10.1101/372896.
53. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
54. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 5396 (2019).
55. Killick, R. & Eckley, I. A. changepoint: An R Package for Changepoint Analysis. *J. Stat. Softw.* **58**, (2014).
56. Akeson, E. C. *et al.* Chromosomal Inversion Discovered in C3H/HeJ Mice. *Genomics* **87**, 311–313 (2006).
57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
59. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
60. Caeiro, F. & Mateus, A. randtests: Testing randomness in R. (2014).
61. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
62. Singer, B. In vivo formation and persistence of modified nucleosides resulting from alkylating agents. *Environ. Health Perspect*. **62**, 41–48 (1985).

Acknowledgements and funding

We gratefully acknowledge Maša Roller and Florian Markowetz for supervision, and Loris Mularoni and Graham Ritchie for software support. Also, the valuable contributions by the CRUK Cambridge Institute Core facilities: CRUK Biological Resources (Angela Mowbray), Preclinical Genome Editing (Lisa Young, Steven Kupczak, Maureen Cronshaw, Paul Mackin, Yi Cheng, Lena Hughes-Hallett), Genomics (James Hadfield, Fatimah Bowater), Bioinformatics (Gord Brown, Matthew Eldridge, Richard Bowers), Histopathology and ISH (Leigh-Anne McDuffus, Cara Brodie, Jo Arnold), and Research Instrumentation (Jane Gray); Edinburgh Genomics (Clinical) Facility; and the EMBL-EBI technical services cluster (Zander Mears, Andrea Cristofori, Tomasz Nowak, Sundeep Nanuwa, Vahit Tabak, Alessio Checcucci). We thank Wendy Bickmore and Chris Ponting for comments on the manuscript.

This work was supported by: Cancer Research UK (20412, 22398), the European Research Council (615584, 682398), the Wellcome Trust (WT108749/Z/15/Z, WT106563/Z/14/A, WT202878/B/16/Z), the European Molecular Biology Laboratory, the MRC Human Genetics Unit core funding programme grants (MC_UU_00007/11, MC_UU_00007/16), and the ERDF/Spanish Ministry of Science, Innovation and Universities-Spanish State Research Agency/DamReMap Project (RTI2018-094095-B-I00). S.J.A. received a Wellcome Trust PhD Training Fellowship for Clinicians (WT106563/Z/14/Z) and is now funded by a National Institute for Health Research (NIHR) Clinical Lectureship. O.P. is funded by a BIST PhD fellowship supported by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the Barcelona Institute of Science and Technology. V.S. is supported by an EMBL Interdisciplinary Postdoc (EIPOD) fellowship under Marie Skłodowska-Curie actions COFUND (664726). E.K. is supported by the EMBL International PhD Programme. C.A-P. is supported by La Caixa Foundation fellowship (ID 100010434; LCF/BQ/ES18/11670011). S.V.B. is supported by ERC Starter Grant 759967. A.E. is supported by a UKRI Innovation Fellowship (MR/RO26017/1). A.K. is a cross-disciplinary postdoctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (core grant to the MRC Institute of Genetics and Molecular Medicine).

Author contributions

S.J.A., F.C., C.F., D.T.O. conceived the project and designed the experiments. S.J.A., F.C., C.F., performed the mutagenesis experiments and sequencing experiments. E.L-A, A.M.R. performed supporting experiments. J.S-L provided contract sequencing. S.J.A. performed the histopathological analyses with advice from S.E.D.. C.J.A., M.S.T. designed and implemented computational analysis. M.S.T. discovered lesion segregation. O.P., V.S., T.F.R., M.L., S.A., E.K., J.L. performed supporting computational analysis. C.A-P., S.V.B., R.M.D., A.E., V.B.K., A.K., I.S., L.T. contributed to the computational analyses. T.F.R., M.L., S.A., A.D.Y. curated data. S.J.A., C.A.S., N.L.B., P.F., D.T.O., M.S.T. supervised the work. S.J.A., C.A.S., N.L.B., P.F., D.T.O., M.S.T. lead the Liver Cancer Evolution Consortium. S.J.A. and P.F. provided scientific and administrative organisation. S.J.A., C.A.S., N.L.B., P.F., D.T.O., M.S.T. funded the work. S.J.A., D.T.O., M.S.T. wrote the manuscript. All authors had the opportunity to edit the manuscript. All authors approved the final manuscript.

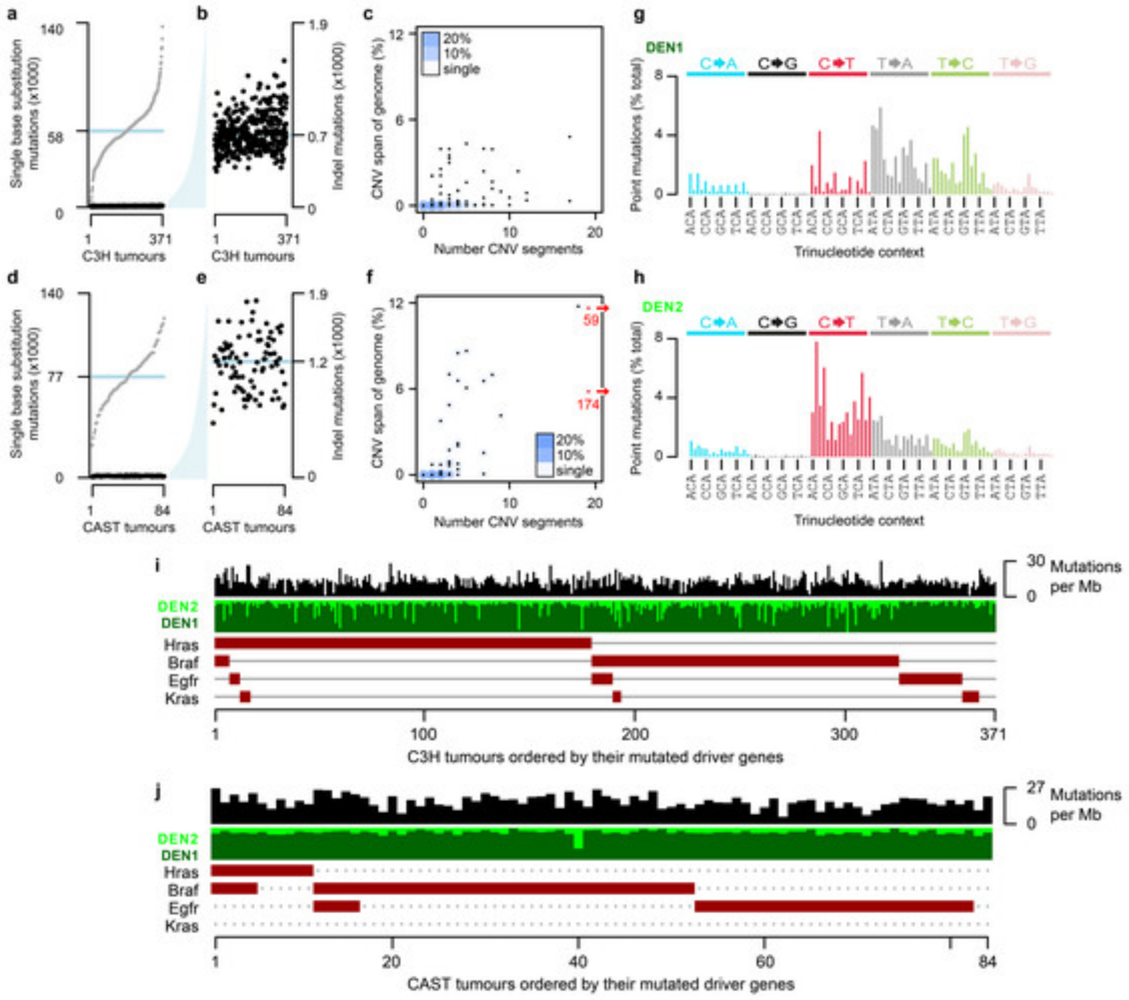
Competing interests

P.F. is a member of the Scientific Advisory Boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd.

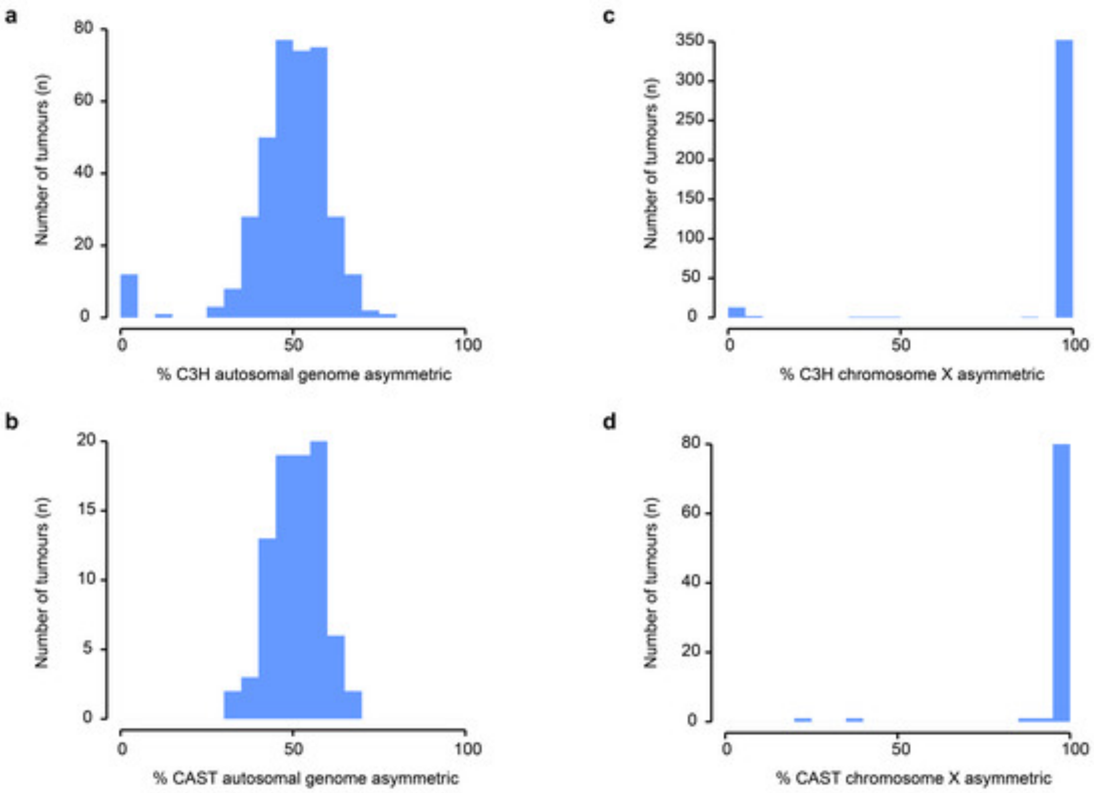
Correspondence

Correspondence and materials requests should be addressed to Duncan T. Odom Duncan.Odom@cruk.cam.ac.uk and Martin S. Taylor martin.taylor@igmm.ed.ac.uk

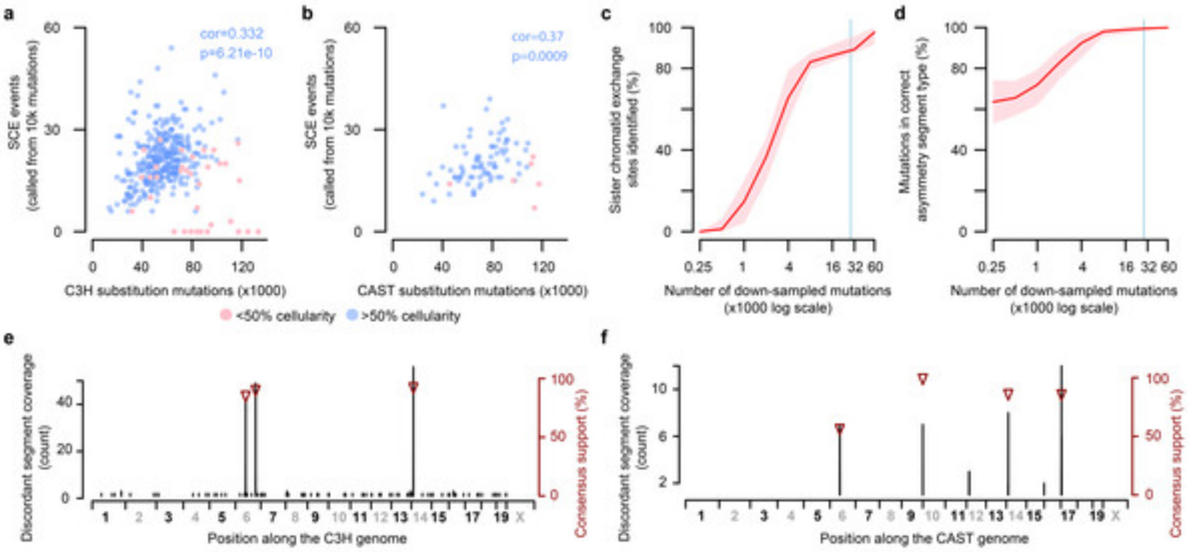
Extended data



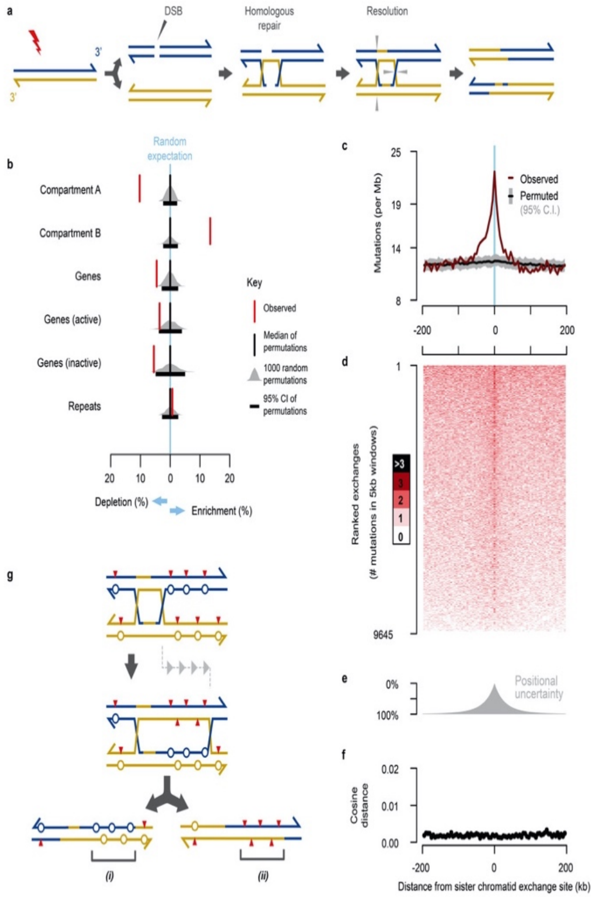
Extended Data Fig.1 | Summary mutation metrics for both C3H and CAST tumours. **a**, Single nucleotide substitution rates per C3H tumour, rank ordered over x-axis (grey points, median blue line). Insertion/deletion (indel, <11 nt) rates show as black. **b**, Y-axis from **a,** expanded to show distribution of indel rates with preserved tumour order. **c**, Number of C3H copy number variant (CNV) segments and their total span as a percent of the haploid genome. Blue shading shows intensity of overlapping points as a percent of all tumours in the plot. **d-f**, Corresponding plots for CAST derived tumours; **f**, two extreme x-axis outliers relocated (red) and x-axis value shown. **g-h**, Mutation spectra deconvolved from the aggregate spectra of 371 C3H tumours, subsequently referred to as the DEN1 and DEN2 signatures. DEN1 is dominated by T→N/A→N changes thought to arise from the O^4-ethyl-deoxythymidine adduct of T[10]. DEN2 substitutions are primarily C→T/G→A changes likely from O^6-ethyl-2-deoxyguanosine lesions of G[10]. **i**, Oncoplot summarising mutation load, mutation signature composition, and driver gene mutation complement of C3H tumours. **j**, Oncoplot of CAST derived tumours as (**i**). The DEN2 signature is a minor component of most tumours but prominent in a minority (**i,j**).



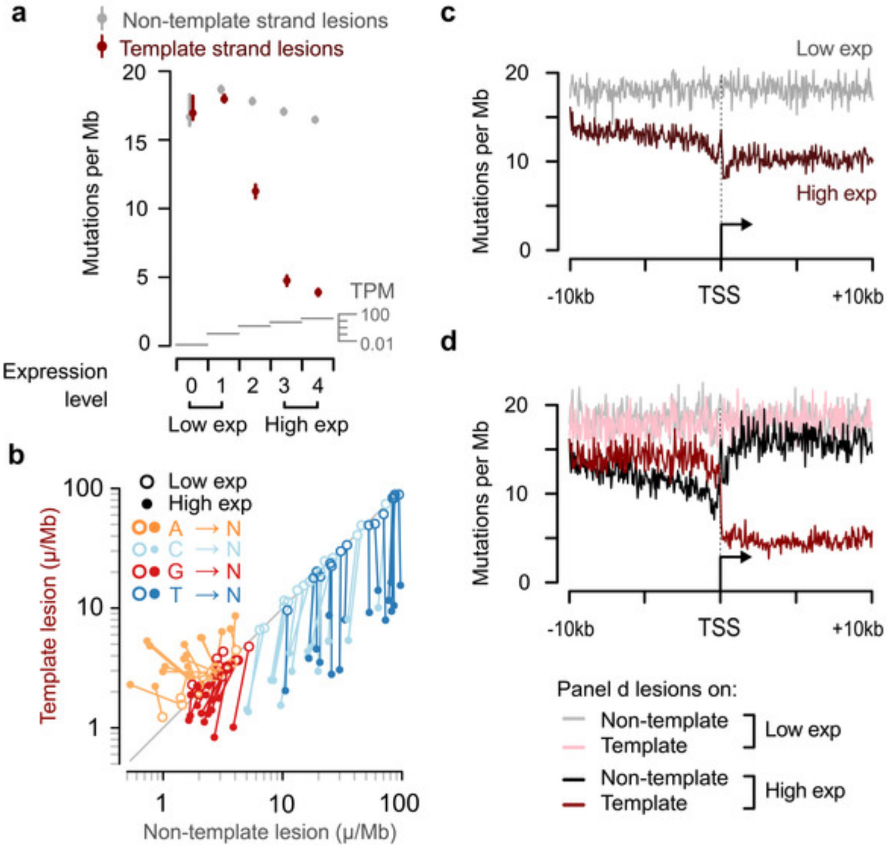
Extended Data Fig.2 | Mutational asymmetry across 50% of the autosomal genome and 100% of the haploid X chromosomes. **a,b**, Typically 50% of the autosomal genomic span (percent of nucleotides) in tumours is contained in segments with either Watson or Crick strand mutational asymmetry. **a**, C3H tumours, n=371. **b**, CAST tumours, n=84. **c**,**d**, Typically 100% of the haploid X chromosome shows Watson or Crick strand mutational asymmetry. **c**, C3H tumours (n=371). **d**, CAST tumours (n=84).



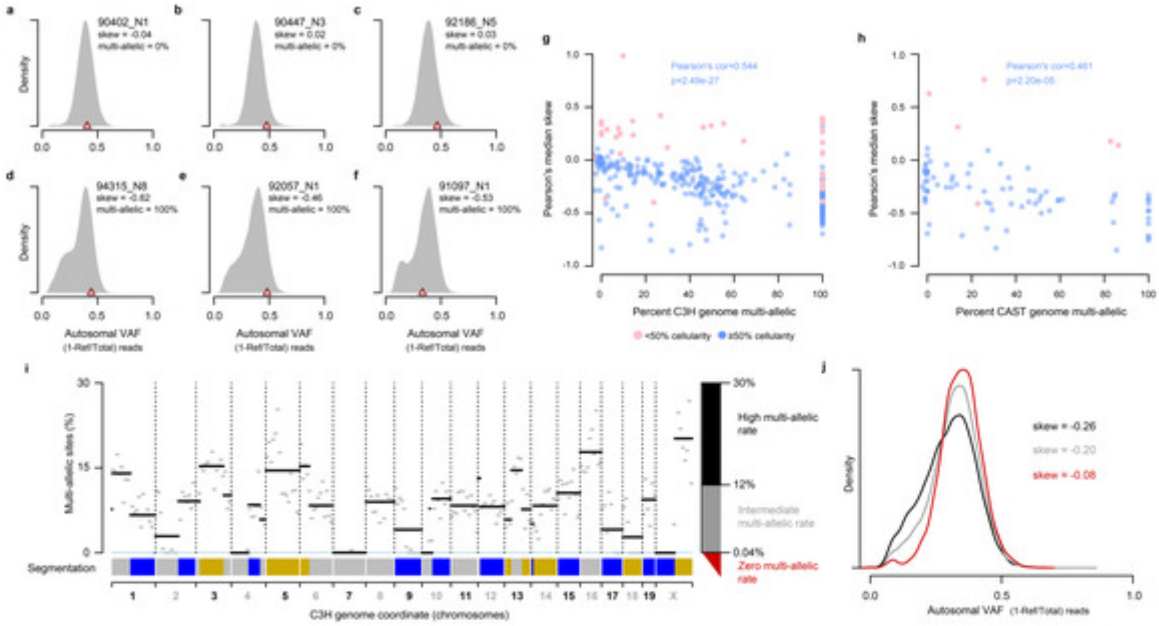
Extended Data Fig.3 | The frequency of sister chromatid exchanges correlates with mutation rate, and localising reference genome assembly errors. a, The relationship between single nucleotide substitution mutation load and detected sister chromatid exchange (SCE) events in C3H tumours. DEN is known to produce ethyl aducts on the sugar-phosphate backbone of DNA as well as mutation-inducing modifications to the bases[10] which could lead to strand breaks[62] triggering SCE. The frequent observation and correlation between rates of SCE and point mutation supports this view. Counts of SCE (y-axis) are based on down-sampling to 10,000 informative mutations per tumour to ensure equal power to detect SCE in each tumour. Tumours with <50% cellularity (pink) have high mutation load and form a sub-group with few detected sister chromatid exchange events; these are suspected to be polyclonal tumours and were excluded from the Pearson's correlation reported (n=335 independent tumour samples, implemented in a two-sided test, significance from Fisher's transform). **b,** As for (**a**) but showing CAST derived tumours (n=84, after cellularity exclusions n=77). **c,** Evaluation of the relationship between mutation load and ability to detect sister chromatid exchange events. Mutations from C3H tumour 94315_N8 (shown in **Fig. 2**) randomly down-sampled and segmentation analysis applied. Y-axis shows the percentage of sister chromatid exchange events detected (100 replicates, mean red, 95% C.I. pink). X-axis is on a log-scale: 95% of C3H and >95% of CAST tumours have mutation counts to the right of the blue vertical line. Down-sampling other tumours gave comparable results. **d**, The same down-sampling data as shown in panel **c** but the y-axis shows the percent of mutations with the correct (same as full data) mutational asymmetry assignment (mean red, 95% C.I. pink). **e**, Candidate C3H reference genome assembly errors. Genome coordinates shown on the x-axis. Immediate switches between Watson and Crick asymmetry are not expected on autosomes unless both copies of the chromosome have a SCE event at equivalent sites. However, inversions and translocations between the sequenced genomes and the reference assembly are expected to produce immediate asymmetry switches. The discordant segment coverage count (black y-axis) shows the number of informative tumours (those with either Watson or Crick strand asymmetry at the corresponding genome position) that suggest a tumour genome to reference genome discrepancy. Consensus support (brown y-axis) plotted as triangles shows the percentage of informative tumours that support a genomic discrepancy at the indicated position (only shown for values >50% support). The two sites on chromosome 6 in C3H correspond to a previously identified C3H strain specific inversion that is known to be incorrectly oriented in the C3H reference assembly[56]. **f**, Candidate CAST reference genome assembly errors, plotted as per (**e**). The candidate mis-assembly on chromosome 14 in both strains occurs at an approximately orthologous position, suggesting a rearrangement shared between strains or a misassembly in the BL6 GRCm38 reference assembly against which other mouse reference genome assemblies have been scaffolded.



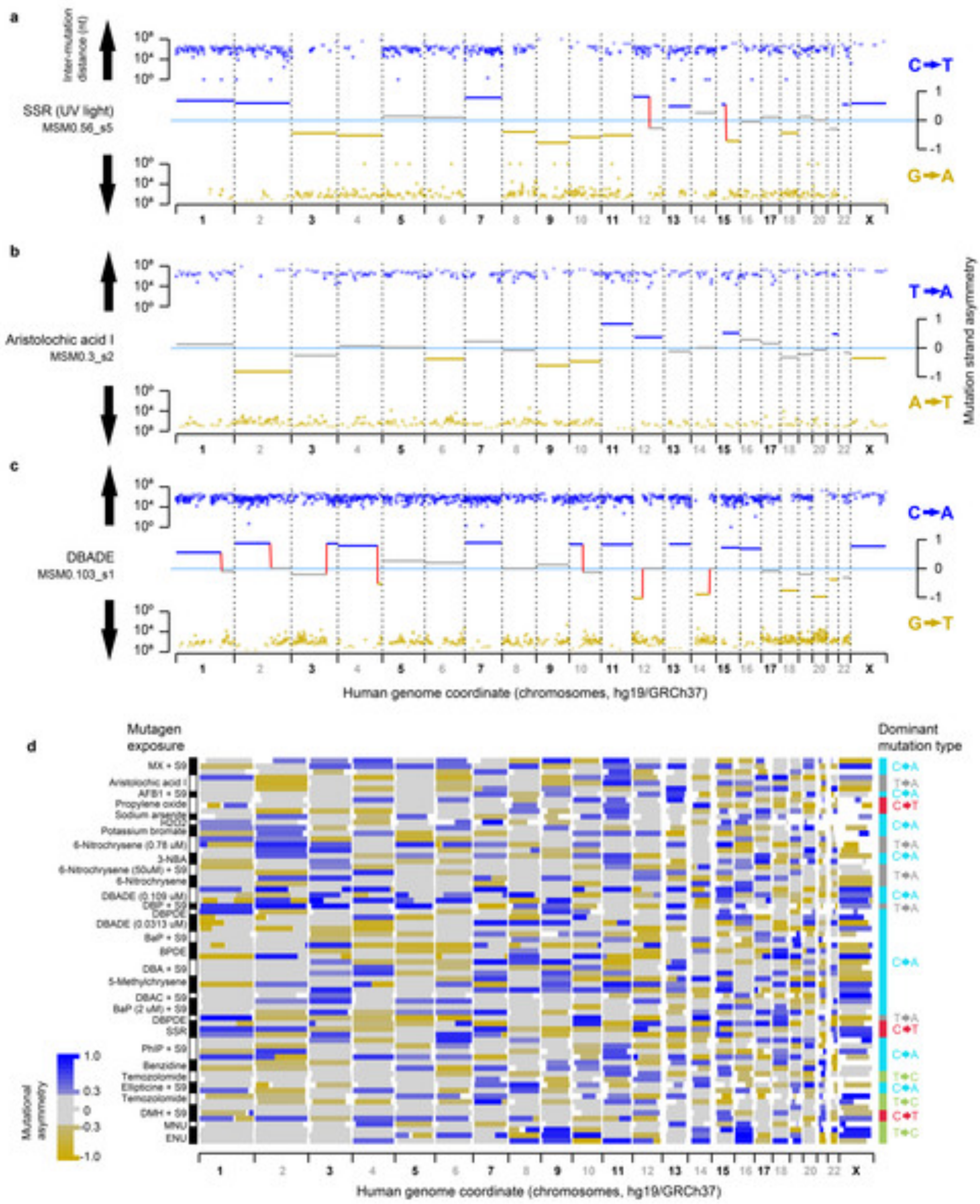
Extended Data Fig.4 | Locally elevated mutation load is driven by sister-chromatid exchange. a, Double strand breaks (DSBs) and other DNA damage can trigger homologous recombination (HR) mediated DNA repair between sister chromatids. The repair intermediate resolves into separate chromatids through cleavage and ligation; grey triangles denote cleavage sites for one of the possible resolutions that would result in a large-scale sister-chromatid exchange event. Although illustrated for double-ended DNA breaks, single ended breaks from collapsed replication forks can be repaired through HR and could similarly lead to the formation of repair intermediate structures that can be resolved as SCEs. **b**, Enrichment analysis of sister chromatid exchanges sites (red) compared with null expectations from randomly permuting locations into the analysable fraction of the genome (grey distributions), the black boxes denote 95% of 1,000 permutations. Sister chromatid exchange events are enriched in later replicating and transcriptionally less active genomic regions (Hi-C defined compartment B), and correspondingly depleted from early replicating active regions. **c**, Aggregating across n=9,645 sister chromatid exchange sites, the observed mutation rate approximately doubles at the inferred site of exchange (x=0). Aggregate mutation rates (brown) were calculated in consecutive 5kb windows. Compositionally matched null expectation was generated by permuting each exchange site into 100 proxy tumours and calculating median (black) and 95% confidence intervals (grey) while preserving the total number of projected sites per proxy tumour. **d**, The elevated mutation count is not the result of a high mutation density in a subset of exchange sites, rather it is a subtle increase in mutations across most exchange sites. Heatmap showing mutation counts calculated in consecutive 5kb windows across each exchange site. Rows represent each exchange site, rank-ordered by total mutation count across each 400kb interval. **e**, The distribution of positional uncertainty in exchange site location approximately mirrors the decay profile of elevated mutation frequency. **f**, Divergence of mutation rate spectra is shown as cosine distance between the analysed window and the genome wide mutation rate spectrum aggregated over all C3H tumours. Despite the elevated mutation frequency, there is no detected distortion of the mutation spectrum. **g**, A model based on HR repair intermediate, branch migration that produces heteroduplex segments of *(i)* mismatch:mismatch (circles) and *(ii)* lesion:lesion (red triangles) strands. Subsequent strand segregation would increase the mutational diversity of a descendant cell population but not the mutation count per cell (key as per **Fig. 2**).



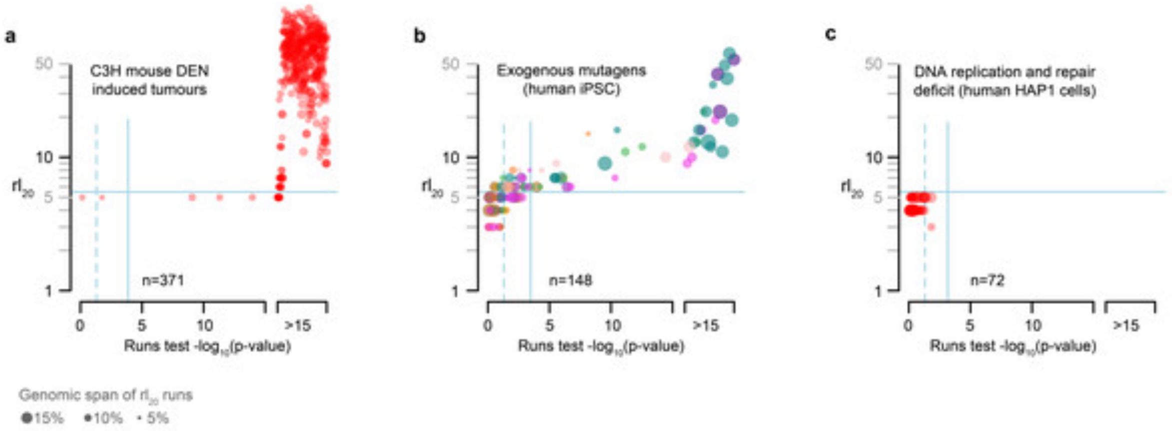
Extended Data Fig.5 | Replication of transcription coupled repair with lesion strand resolution in *Mus musculus castaneus*. a, Transcription coupled repair of template strand lesions is dependent on transcription level (P15 liver, median transcripts per million (TPM)). Mutation rate estimates (circles) are the aggregate rates for expression level binned genes across CAST tumours (n=84). Expression level bin 0 contains n=2,645 genes, all subsequent bins contain n=4,323 genes. See methods for per-gene, per-tumour inclusion criteria. Empiric confidence intervals (99%) were calculated through bootstrap sampling (n=100 replicates) of genes within the expression level bin. **b,** Comparison of mutation rates for the 64 trinucleotide contexts: each context has a high and a low expression point linked by a line. **c,** Sequence composition normalised profiles of mutation rate around transcription start sites (TSS). **d,** Stratifying the data plotted in (**c**) by lesion strand reveals greater detail on the observed mutation patterns, including the pronounced influence of bidirectional transcription initiation.



Extended Data Fig.6 | Variant allele frequency distributions demonstrate high rates of non-mutagenic replication over segregating lesions. a-f, Variant allele frequency (VAF) distributions shown as probability density functions (total area under curve=1) for six example tumours, calculated taking into account observed multi-allelic variation. The VAF for identified driver mutations is indicated (brown triangle). Tumour identifiers are shown top right along with the percent of genomic segments (based on mutation asymmetry segmentation) that are multi-allelic. Skew shows Pearson's median skewness coefficient for the VAF distributions. Panels **a-c** show tumours with no multi-allelic segments and exhibit a symmetric VAF distribution showing minimal sub-clonal structure; **d-f** show tumours with all segments multi-allelic, illustrating the sub-clonal structure generated by segregating lesions. **g**, Tumours with a high proportion of multi-allelic segments have a left-skewed VAF distribution indicating frequent non-mutagenic replication over segregating lesions. Percent of genome segments that are multi-allelic (x-axis) plotted against VAF distribution skew for 371 C3H tumours. Tumours with low estimated cellularity indicated in pink and excluded from correlation analysis (n=335 independent tumour samples in Pearson's correlation, two-sided significance from Fisher's transform). **h**, As for (**g**) but showing 84 CAST tumours (n=77 independent tumours included in Pearson's correlation). **i**, Mutation asymmetry summary ribbon for example C3H tumour 90797_N2; C3H genome on the x-axis. The percent of mutation sites with robust support for multi-allelic variation (y-axis) calculated in 10Mb windows (grey) and for each asymmetric segment (black). Thresholds for high (black), intermediate (grey) and zero (red) rates of multi-allelic sites shown on the right axis. **j**, VAF density plots for the example tumour 90797_N2 (shown in (**i**)) mutations in asymmetry segments stratified by the multi-allelic rate thresholds defined in (**i**). As with individual tumour based analysis (**a-h**), high multi-allelic rates correspond to a leftward skew of the VAF (black, grey) whereas segments without multi-allelic variation (red) show a minimally skewed distribution.



Extended Data Fig.7 | Examples of mutation patterns generated by lesion segregation from a diverse range of clinically relevant mutagens. **a-c**, Genome wide mutation asymmetry plots (shown as per **Fig. 2a-c**) for mutagen exposed human induced pluripotent stem cells (iPSCs)[5]. Cells exposed to simulated solar radiation (SSR) illustrate lesion segregation for ultraviolet damage (**a**). Immediately adjacent mutations (inter-mutation distance 10^0) indicate CC->TT dinucleotide changes. Despite a low total mutation load (1,308 nucleotide substitutions, 842 informative T→A changes), the mutational asymmetry of lesion segregation is evident for the aristolochic acid exposed clone[5] (**b**), and the polycyclic aromatic hydrocarbon DBADE (**c**) that is found in tobacco smoke. **d**, Summary mutation asymmetry ribbons (as per **Fig. 2d**) for all mutagen exposed clones with rl_{20} >5, which illustrates the independence of asymmetry pattern between replicate clones, almost universal asymmetry on chromosome X, and approximately 50% of the autosomal genome with asymmetry over autosomal chromosomes. The dominant mutation type is indicated for each mutagen. In those clones with low mutation rates, some sister exchange sites are likely to have been missed leading to reduced asymmetry signal (e.g. on the X chromosome). Segments with <20 informative mutations are shown in white.



Extended Data Fig.8 | Lesion segregation is evident for multiple DNA damaging agents but not for damage independent mutational processes. **a**, DEN induced C3H tumour genomes (n=371) typically show significant mutational asymmetry across their genome. Wald-Wolfowitz runs test (x-axis) p-values calculated using a normal approximation (two-sided). Nominal p=0.05 significance threshold indicated by dashed blue line, Bonferroni corrected threshold shown as solid vertical blue line. P-values < $1x10^{-15}$ are rank-ordered. The rl_{20} metric (**Fig.5a**; Methods) is shown on the y-axis, horizontal blue line gives empric significance threshold of $rl_{20}>5$. **b**, Many human induced pluripotent stem cells (iPSCs) grown from single cells after exogenous mutagen exposure[5] show significant mutation asymmetry (n=148 whole genome sequenced, mutagen exposed cell lines). Statistical calculations and plotting as in panel (**a**) with adjustment of Bonferroni correction. Diverse categories of mutagen, denoted by point colour (see **Fig. 5b**), show asymmetry indicative of lesion segregation. **c**, Cell-lines with genetically perturbed genome replication and maintenance machinery[26] and similar mutation load to those in panel (**b**) do not show significant mutation asymmetry (n=72 whole genome sequenced, genetically perturbed cell-lines). Statistical calculations and plotting as in panel (**a**) with adjustment of Bonferroni correction.

Extended Data Table 1 | A lesion segregation based test for oncogenic selection.

Strain	Gene	Mutation	Mutation count	Odds ratio	P-value	Known driver
C3H	Braf	6:37548568_A/T	151	2.13	5.77×10^{-6}	Yes
C3H	Hras	7:145859242_T/C	81	2.67	6.88×10^{-6}	Yes
C3H	Hras	7:145859242_T/A	65	1.02	1	Yes
C3H	Intronic Fmnl1	11:105081902_A/C	44	1.03	1	No
C3H	Intergenic	9:73125689_G/C	42	1.13	1	No
C3H	Egfr	11:14185624_T/A	34	3.87	1.23×10^{-4}	Yes
CAST	Braf	6:37451282_A/T	42	1.41	0.338	Yes

Recurrently mutated sites in C3H and CAST tumours with sufficient estimated power to detect oncogenic selection through biased strand retention analysis (required >33 C3H recurrences or >41 CAST recurrences). Odds ratio values >1 indicate the predicted correlation of driver mutation and Watson/Crick strand retention in tumours with the candidate driver mutation, but not for those without the mutation. The Fisher's exact test was performed on counts of chromosomes with Watson and Crick strand asymmetries (Methods). Each tested site was autosomal, thus total sample sizes were: n=2x371=742 for C3H, and n=2x84=168 for CAST. P-values (two-sided) are shown after Bonferroni correction (7 tests performed). Known driver indicates the mutation or its orthologous change has previously been implicated as a driver of hepatocellular carcinoma[6]. The CAST 6:37451282_A/T mutation is orthologous to the C3H 6:37548568_A/T mutation.

Supplementary Table 1 | Table of tumours sequenced containing key parameters and mutation spectra signature matrices (Excel file).

Supplementary Table 2 | Table of exogenous mutagen and ICGC scan results (Excel file).

Supplementary Table 3 | Table of key resources and software (Excel file).