DNA lesion bypass and the stochastic dynamics of transcription coupled repair

DNA lesion bypass and the stochastic dynamics of transcription coupled repair

Michael D. Nicholson[1,*], Craig J. Anderson[2], Duncan T. Odom[3,4], Sarah J. Aitken[4,5,6] & Martin S. Taylor[2,*]

[1] CRUK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, UK. EH4 2XU

[2] Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. EH4 2XU

[3] German Cancer Research Center (DKFZ), Heidelberg, Germany

[4] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

[5] Medical Research Council Toxicology Unit, University of Cambridge, Cambridge, CB2 1QR, UK

[6] Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK

* Correspondence to Michael D. Nicholson or Martin S. Taylor.

Email: michael.nicholson@ed.ac.uk or martin.taylor@ed.ac.uk

Author Contributions: M.S.T. conceived the project. M.D.N. and M.S.T. designed and implemented the analyses. C.J.A. provided supporting analysis on gene expression. D.T.O. and S.J.A. provided data resources and biological insight. M.D.N. and M.S.T. wrote the manuscript with input and editing from all authors. All authors approved the final manuscript.

Competing Interest Statement: The authors declare no competing interests.

Classification: Biological sciences > Genetics

Keywords: Mutation, damage, transcription, repair, mathematical modelling

Abstract

DNA base damage is a major source of oncogenic mutations (Alexandrov et al. 2020) and disruption to gene expression (Chiou et al. 2018). The stalling of RNA polymerase II (RNAP) at sites of DNA damage and the subsequent triggering of repair processes has major roles in shaping the genome wide distribution of mutations, clearing barriers to transcription and minimising the production of mis-coded gene products. Despite its importance for genetic integrity, key mechanistic features of this transcription coupled repair (TCR) process are controversial or unknown. Here we exploited a well-powered *in vivo* mammalian model system to explore the mechanistic properties and parameters of TCR for alkylation damage at fine spatial resolution and with discrimination of the damaged DNA strand. For rigorous interpretation, a generalisable mathematical model of DNA damage and TCR was developed. Fitting experimental data to the model and simulation revealed that RNA-polymerases frequently bypass lesions without triggering repair, indicating that small alkylation adducts are unlikely to be an efficient barrier to gene expression. Following a burst of damage, the efficiency of transcription coupled repair gradually decays through gene bodies with implications for the occurrence and accurate inference of driver mutations in cancer. The observed data is inconsistent with RNAP always reinitiating after repair, but is well explained by a model in which no reinitiation occurs, suggesting that RNAP reinitiation is not a general feature of transcription coupled repair. Collectively these results reveal how the directional, but stochastic activity of TCR shapes the distribution of mutations following DNA damage.

Significance

Damage to DNA can interfere with crucial cellular processes such as the transcription of genes into RNA and can ultimately lead to mutations, DNA sequence changes, that are inherited by subsequent generations of cells and organisms. Transcription coupled repair (TCR) works to ensure genes that are being used by a given cell are cleared of damage so they can continue to be utilised. We reveal mechanistic details of how TCR works, its efficiency and how that changes through the length of a gene. This helps understand how cells deal with a burst of DNA damage, for example from sunburn or chemotherapeutic treatment, and where the resulting genetic damage is likely to occur, with implications for cancer risk and treatment.

Introduction

Accurate and efficient DNA replication and DNA transcription are essential for life. However, cellular DNA is continuously assaulted with damage arising from both endogeneous and exogeneous sources. With hundreds of thousands of DNA adducts forming per genome per day, crucial molecular processes can be severely inhibited (Yousefzadeh et al. 2021). Damage falling within transcribed regions poses particularly acute challenges, potentially interfering with accurate and efficient transcription, as well as risking the formation of heritable, protein-altering mutations. Transcription coupled repair (TCR), a highly conserved branch of the nucleotide excision repair pathway (Gregersen and Svejstrup 2018; Sarsam et al. 2024), assists in minimising the risk of such aberrant outcomes (Fig 1.a). Triggered by the stalling of actively transcribing RNA polymerase II (RNAP), TCR excises the stalling-lesion and, by using the non-transcribed strand as a template for synthesis, results in repaired, lesion-free DNA.

Frequent RNAP stalling potentiates dysregulation of homeostatic expression and increased transcription-replication complex collisions (Lans et al. 2019). On the other hand, uncleared damage risks transcriptional mutagenesis (Brégeon and Doetsch 2011) and incorrect base-pairing at replication. Thus, a balance between damage tolerance and clearance must be struck. Central to understanding this balance, and our ability to quantitatively map damage to cellular outcome, is the measurement of how the transcriptional machinery interacts with damage. In this study we focus on two key elements of this interaction: the sensitivity with which RNAPs detect damage and trigger TCR, and how frequently RNAPs reinitiate transcription following repair (Fig 1.b).

The efficiency of TCR initiation is expected to be influenced by lesion type (Saxowsky and Doetsch 2006; Lans et al. 2019). Smaller adducts, such as the oxidative stress induced 8-oxoguanine, are bypassed with relative ease by RNAP (Tornaletti et al. 2004), while more bulky, helix-distorting lesions, e.g. UV-caused pyrimidine-dimers, provide a more stringent roadblock to transcribing RNAP, which may only rarely be bypassed (Marietta and Brooks 2007; Walmacq et al. 2012). When RNAP stalling and repair does occur, transcription must be rapidly resumed to maintain cellular function. It was commonly thought that stalled RNAPs resumed transcription from the damaged site (Geijer and Marteijn 2018), however recent work has demonstrated disassociation of RNAP following TCR at UV induced pyrimidine-dimers (Chiou et al. 2018). Without RNAP restart, further RNAP transcription initiations at a given gene's promoter are required, potentially necessitating numerous transcription initiations to clear a gene-body of multiple lesions and to generate a complete RNA transcript. While the bypass efficiency for varied lesions can be quantified *in vitro* (You et al. 2012), an integrative picture summarising the outcomes of transcriptional machinery encountering adducts *in vivo* is lacking.

For TCR-inducing lesions, we reasoned that analysing mutation burden as a function of both gene-expression and genic-position would provide insight into TCR mechanics. DNA damage that avoids repair and persists to replication can result in incorrect base-pairing, thus generating heritable mutations that are detectable in the damaged cell's progeny. Supposing that template strand lesions consistently stall RNAP, triggering lesion excision and repair and subsequent RNAP disassociation, then any downstream lesions will require a second RNAP for detection and clearance. Under this model the 5' end of moderately expressed genes would be cleared of lesions but the 3' end would remain unrepaired (Fig 1.c). If this positional bias in lesions persists through to DNA replication then a sigmoidal mutational pattern through the gene bodies would be expected, with the curve progressively moving towards the 3' end as transcription increases. Alternatively, if RNAPs consistently reinitiate transcription following lesion detection and repair, then no positional bias in lesion clearance should be expected, and hence a more uniform mutation burden through the gene body is predicted (Fig 1.c). Therefore, observing mutational patterns caused by template strand lesions as a function of genomic position and gene expression potentially offers a window into the mechanics of TCR.

As RNAP is only expected to trigger the repair of damage on the transcriptional template strand, a prerequisite for using mutation patterns to accurately infer the activity of TCR is the ability to resolve the lesion containing strand. Prior studies (Haradhvala et al. 2016; Seplyarskiy et al. 2019) have relied on inferences from the biochemistry of mutagenesis for lesion strand resolution, for example assuming that C->T mutations from UV photoadducts involve the C nucleotide rather than the G of the complementary strand. Such inferences can be confounded by atypical adducts (Vandenberg et al. 2023) and the spectrum of adducts produced by other mutagens is generally less well understood. An alternative strategy is to *ab initio* phase the stand of DNA damage. Following a burst of mutagenic damage in a single cell cycle, most mutations arise through replication using a damaged base as a template (Aitken et al. 2020). Through the semi-conservative replication of DNA, the two complementary strands of a DNA duplex will template the new synthesis of two sister chromatids that, through mitosis, segregate into separate daughter cells (Fig 1.a). Each daughter cell lineage receives the DNA lesions, and ultimately mutations, from just one of the parental DNA strands. This DNA lesion segregation (Aitken et al. 2020) results in chromosome scale, strand asymmetric mutation patterns that can be used to confidently discriminate the DNA lesion strand (Aitken et al. 2020) and through comparison to gene annotation, resolve it as either the transcriptional template or non-template strand (Fig 1.a; (Anderson et al. 2022)).

To explore the mechanism and efficiency of TCR *in vivo*, with spatial precision and lesion strand resolution, we have exploited an established mouse model of diethylnitrosamine (DEN) induced liver cancer (Verna, Whysner, and Williams 1996; Connor et al. 2018) (Fig 1.d). DEN is bioactivated into a potent but short-lived mutagen by the hepatocyte expressed enzyme Cyp2e1. This generates a range of DNA alkylation adducts, including the principal mutagenic lesion O^4-ethyldeoxythymidine (Verna, Whysner, and Williams 1996). Tumours reliably develop within 24 weeks of a single acute exposure to DEN; each of these represents a clonal expansion of one post-mutagenesis cell whose genome typically contains 60,000 base substitution mutations, and exhibits the pronounced mutation asymmetry of lesion segregation (Aitken et al. 2020).

Here, we examine strand-phased mutational patterns as a function of gene-expression and lesion-position to quantify the mechanics of TCR. We present a probabilistic mathematical model, incorporating the key mechanistic features of the TCR process, which is able to recapitulate the mutation patterns of DEN-induced tumour genomes. Analysing the murine liver data through the mathematical model we show that, for alkylation DNA adducts such as those created via DEN exposure, the initiation of TCR is stochastic, with frequent transcription occurring over mutagenic lesions. Overall our modelling approach provides a framework for translating strand-phased mutation data to the mechanics of TCR.

Results

TCR shapes mutation patterns through the gene-body in DEN-induced tumour genomes

We aimed to identify the speculated mutational patterns in the genomes of DEN-induced murine liver tumours. As previously described (Aitken et al. 2020), using lesion segregation we were able to call approximately 1.7 million high confidence, strand-resolved mutations within transcribed regions from 237 tumour genomes. Matching gene expression measures were generated contemporaneously by total cellular RNA sequencing on healthy liver tissue from untreated litter-mates (Aitken et al. 2020), and nascent transcription rates estimated from intron mapping reads (Anderson et al. 2022).

We first assessed the relationship between strand-specific mutation burden and gene expression. Consistent with TCR playing a dominant role in DEN-induced lesion repair, the mutation rate due to template strand lesions (hereafter, template mutation rate) markedly decreased with increasing transcription (Fig 2.a). We also observed that the mutation rate due to non-template strands lesions (hereafter, non-template mutation rate) was modestly reduced (Fig 2.a), which may occur due to greater chromatin accessibility in highly expressed genes (Anderson et al. 2022).

To isolate the signal of only TCR, we use the non-template mutation rate as the *expected* mutation rate (TCR absent), and compare with the *observed* mutation rate (TCR present) on the template strand. The observed:expected mutation rate quantifies the reduction in mutation burden due to template strand repair; observed:expected values of 1 imply equal lesion burden on both the template and non-template strand at DNA replication, suggesting a lack of TCR. In contrast an observed:expected value of 0 implies the complete removal of template strand lesions. This resulted in dose-response type patterns in each of the 237 tumour genomes (Fig 2.b). Mutation rates from different tumours may be expected to depend on the state of the tumour's ancestral cell at mutagenesis, for example the cell cycle phase at DEN exposure. However, by fitting log-logistic functions (Ritz et al. 2015) - commonly used to quantify dose-response relationships - the shape of the mutation rate decay was found to be remarkably homogeneous (Extended Data Fig 1.a,b). As described previously (Anderson et al. 2022) at high transcription levels the mutation rate plateaued, suggesting that the remaining mutagenic lesions were largely invisible to TCR. Invisible lesions potentially reflect subsets of lesions that are less efficient at stalling RNAPs or lesions in less recognisable genomic contexts; prior analysis of this data supports that lesions in certain trinucleotide contexts are less permissive to repair (Anderson et al. 2022). Given the consistency of the TCR pattern over individual genomes, henceforth we analysed the aggregated data across all genomes.

In order to jointly examine the effect of both expression and the genic position of lesions, the gene expression distribution was binned into six expression strata (Fig 2.b, top panel; Extended Data Fig 1.c). Strata boundaries were chosen to balance accurately reflecting the variation over expression, and to diminish noise by ensuring a sufficient number of genes per stratum. For each stratum, we measured the mutation rate aggregated over all genes in that stratum in consecutive 5 kb windows from the transcription start site (TSS). This demonstrated subtly (approximately 3.5%) lower mutation rates for both template and non-template strand lesions at the 5' end of non-expressed genes (Fig 2.c). This trend was also seen for the non-template strand at all expression strata (Fig 2.d).

We extended our analyses of observed:expected mutation rates (defined above) to focus on positional biases in mutation burden specifically due to TCR, negating potential confounding factors such as 5' end effects and enhanced non-TCR surveillance. We also recognised that as transcription is a processive and directional process, the probability of an upstream lesion on the same template strand could influence the TCR efficiency at a given gene-position. Consequently, both the upstream sequence composition and per tumour burden of lesions (inferred from mutations) could influence the repair efficiency of a focal analysis window.

Addressing these concerns, we created a normalised gene-position measure based on the expected number of upstream lesions that was calculated for each analysis window of each gene, in each tumour, prior to aggregated analysis (Methods) (Fig 2.e).

Comparison of the observed:expected mutation rates to the expected upstream lesion number (Fig 2.f; Extended Data Fig 1.d-k) leads to several immediate conclusions. First, the observed:expected mutation rate is approximately 1 for the lowest expressed genes (stratum 1), which indicates that, as expected, there is no TCR in the absence of detected transcription. Second, for intermediately expressed genes (strata 2-5) we see a linear increase in the mutation rate through the gene body - consistently found when considering only short, or only long genes (Extended Data Fig 1.i-j); suggesting that TCR efficiency decays approximately linearly with the upstream lesion number. Finally, the highly expressed genes, with >10 nascent transcripts per millions (nTPM), show negligible decay in TCR efficiency through the gene body, indicating that all detectable lesions have been removed. By comparing the observed linear decay in TCR efficiency through gene bodies to the hypothetical mutation pattern scenarios (Fig 1.c), these data support a model in which RNAP repairs 5' lesions before downstream 3' lesions, with regular disassociation of RNAP following repair. To robustly quantify the mechanistic origins of these effects we developed a mathematical model of TCR.

Mathematical model for transcription coupled repair dynamics

We defined a Markov chain model (Fig 3.a) characterising the dynamics of transcribing RNAPs in the interim period between DNA damage and replication. To model the initial damage distribution, we selected random positions through gene bodies. Following damage RNAPs sequentially initiate transcription and, upon encountering a lesion, the lesion is detected and repaired with probability *Pd.* Following repair, the RNAPs reinitiate transcription at the site of the damage with restart probability *Pr,* else they disassociate from the strand. Since the efficiency of repair appears to saturate at high levels of transcription without complete lesion removal (Fig 2.b), we assumed two types of lesions exist: lesions that are visible to TCR and so can be detected with probability *Pd*, and TCR-invisible lesions which will not be detected. As mentioned above, TCR-invisible lesions could have altered biochemistry or lie in less recognisable genomic contexts (Anderson et al. 2022); agnostic to mechanism, we include a parameter *Pv* in the mathematical model for the proportion of lesions that are visible.

To match the experimental analysis we consider 6 expression strata in the model such that the *k*th strata has an associated average expression level, e_k, measured in units of nascent transcripts per million (nTPM). We fixed the numerical values of (e_1,..,e_6) as the median nTPM for each strata in the experimentally defined expression data. For genes in a given stratum, we assumed that an average of n_k RNAPs initiated transcription between damage and replication. To relate the RNAP initiations in the model to the RNA sequencing measures, we included an expression multiplication factor (*m*) and specify that $n_k = m*e_k.$ As the per-strata expression values are fixed, the number of RNAP initiations per gene is controlled only through their associated stratum and *m*. Under mild assumptions, such as each produced RNA transcript having equal chance of being sampled in the RNA sequencing, *m* has the further interpretation as the total number of RNA transcription initiations between damage and replication, in units of transcription initiations (x10^6) (Methods).

Using techniques from Markov process theory (Supplementary File 1), we numerically determined the mathematical expectation of the template strand lesion count in the model, as a function of genic-position and the expression multiplier, *m*. The coding strand lesion burden is obtained by suppressing transcription in the model. Dividing the modelled template lesion count by the coding lesion count gives the proportion of unrepaired lesions obs:exp$_{\text{theory}}$, which is directly analogous to the experimentally measured observed:expected mutation rates. Matching the hypothesised lesion patterns (Fig 1.c), if RNAPs always restart following repair (*Pr*=1), then obs:exp$_{\text{theory}}$ is constant over gene position (Fig 3.b). With no RNAP restart and high RNAP sensitivity, obs:exp$_{\text{theory}}$ adopts a sigmoidal shape; while linear gradients emerge for low to medium values of RNAP sensitivity, similar to the experimental observed:expected mutation rates (Fig 2.f).

To examine the utility of the model to infer the mechanistic parameters of TCR, DNA damage followed by TCR was simulated at scales mimicking the murine liver data (Methods). A wide grid of parameter values was used, with *Pd* and *Pr* ranging between 0 and 1, while the expression multiplier *m* was constrained within a literature-informed plausible regime. As ~20% of lesions remain unrepaired even in highly expressed genes (Fig 2.f), we fixed the proportion of TCR-visible lesions, *Pv*, to be 0.8. For a given parameter combination, damage and repair was simulated for ~1.95 million genes (Methods), with genes stratified into 6 expression strata as in the experimental data. Each expression strata was associated with the same nascent expression values e_k measured for the murine liver. Thus, for a given *m* and a gene in strata *k*, an average of $m{*}e_k$ transcription initiations occurred per gene. For a given parameter combination, we aggregated over all simulated genes to construct the simulated observed:expected mutation rates as a function of expected upstream lesions (Fig 3.b). The Manhattan distance between the simulated data and the analytically determined obs:exp$_{\text{theory}}$ was minimised to estimate the underlying parameters (Fig 3.c).

Intuitively, certain parameter combinations could be challenging to uniquely identify, for example the same amount of damage may be cleared by many polymerases with low detection sensitivity, or a few polymerases with high lesion detection rates. Indeed, correlations in parameter estimates were observed in two dimensional heat maps illustrating plausible parameter fits (Fig 3.c), defined as those parameters such that the distance from obs:exp$_{\text{theory}}$ to the simulated data is less than the distance between the original data and bootstrapped original data. For example, overestimation of detection sensitivity often co-occurred with an underestimate of the expression multiplier. Despite this, as model outputs were required to match simulated data over both spatial (position in gene body) and transcriptomic (expression strata) dimensions, we broadly found the true parameters were identifiable in simulated data, with median percent errors of 10%, 22%, and 16% when estimating *Pd*, *Pr*, and *m*, respectively (Fig 3.d).

The results above indicate that we can accurately infer model parameters. However, the expression strata thresholds used for the simulated datasets were the same as those that were constructed to be highly informative on the experimental murine data. As a result the inference accuracy was dependent on the expression multiplier *m*, with an eightfold increase in the median percent error for *Pd* inference between *m*=0.5 and *m*=8.5. Consequently our simulation work likely underestimates the true accuracy of the inference workflow.

TCR is stochastic and RNAP frequently does not restart

We analysed the DEN-induced murine liver tumour mutation data using our mathematical model of TCR, fitting the data as described for the simulations. Despite its simplicity, the model is able to capture the key features of the experimental data (R^2 = 0.99), including linear decays in the efficiency of TCR for intermediate expression levels (Fig 4.a). For lesions visible to TCR, the lesion detection sensitivity, *Pd*, was estimated to be 0.42, with the 95% confidence interval of (CI95: 0.24, 0.74) (Fig 4.b,c). As the proportion of visible lesions, *Pv*, was estimated to be 0.8 (CI95: 0.79, 0.81), we infer that RNAP frequently transcribes over damage, failing to stall and trigger repair in 66% of lesion encounters (Fig 4.d).

The principal mutagenic adduct from DEN exposure is thought to be O^4-ethyldeoxythymidine (O^4-EtdT) (Verna, Whysner, and Williams 1996) and the relative bypass efficiency of O^4-EtdT by mammalian RNAP *in vitro* is ~60% (You et al. 2014), in close agreement with our inference from *in vivo* data. For those lesions accessible to TCR, our estimate suggests that each lesion will be transcribed over ~1.5 times before stalling an RNAP and initiating TCR. Transcription over template strand O^4-EtdT by mammalian Pol II misincorporates ribonucleotides in RNA at a rate of ~50% (You et al. 2014), suggesting wide-spread transcriptional mutagenesis occurred post-damage in the murine experiments.

The expression multiplier *m* was estimated as 1.59 (CI95: 0.79, 3.18), implying that in the mouse liver cells exposed to DEN, 1.59 million RNAPs initiated transcription between damage and replication. For highly expressed (stratum 6) genes with median expression of 11.15 nTPM, ~18 polymerases are expected to initiate transcription. To assess the validity of this inference, an orthogonal estimate of *m* was determined using estimates of transcription parameters obtained through analysis of single-molecule fluorescence *in situ* hybridisation imaging (Methods). Briefly, Bahar Halpern *et al.* (Bahar Halpern et al. 2015) measured the transcription rate and proportion of promoters actively transcribing for 7 genes, for which nascent RNA sequencing estimates (*e*) are available in the murine liver experimental data. Combining these values with literature estimates of the time between damage and replication, provides estimates of the transcript number produced for each gene (*n*) (Extended Data Fig 2.a). By the relation *n*=*m***e*, this suggests 2.77 million RNAP initiations occur between damage and replication. As plausible bounds for *m* range over nearly 2 orders of magnitude (Extended Data Fig 2.b) (Methods), the concordance between the orthogonal estimate to our inferred estimate of 1.59 confirms the robustness of our analytical approach despite the simplifications made.

RNAPs were estimated to restart transcription after 65% (CI95: 24%, 89%) of repair events. As the 95% confidence interval excludes 100%, the null hypothesis that RNAP always restarts from the damaged site after repair is not consistent with these data. Further, parameter combinations that include *Pr=0*, denoting the complete absence of polymerase restart, are within the plausible regions as defined above for simulations (Fig 4.c). When we considered a reduced model without RNAP restart (*Pr*=0), the optimal fit provided a near identical fit to the model with restart (Extended Data Fig 2.c) and model selection analysis, assuming normally distributed errors, indicated that the model without RNAP restart is marginally preferred (Akaike information criterion (AIC) with restart = -997.57, AIC without restart = -997.76). In the model without restart, lesion detection sensitivity is estimated as 0.19 (CI95: 0.11, 0.25), compared to that of 0.42 for the alternative model. Given that consistent RNAP restart is incompatible with the data, we conclude that transcription restart from the site of stalling is not an obligate feature of TCR. Application of Occam's razor favours the conclusion that RNAP restart is not a feature of TCR, though the present data does not allow us to exclude the possibility that restart occurs following some TCR events.

Discussion

In this study, we quantified the interactions between DNA damage and RNAP following exposure of murine hepatocytes to an alkylating agent (DEN) *in vivo*. DNA lesions that persist to replication are the templates for mutational changes inherited by daughter lineages, which are clonally expanded during tumorigenesis. The resulting mutational readout provides an integrated picture of the repair processes that occur between damage and replication; this offers a complimentary approach to the measurements of repair maps, which provide snapshots of repair at specific timespoints (Hu et al. 2015, 2017). By combining strand-phased whole genome sequencing data from 237 mouse liver tumours with RNA sequencing, we showed that transcription coupled repair leaves a highly reproducible and mechanistically informative footprint when comparing mutation burden to both gene expression and mutation position. To translate the mutation patterns into quantitative estimates of the mechanisms of TCR, we developed a mathematical model of damage and repair able to recapitulate the key features of the data. By analysing the mouse data through our model we demonstrated that (i) lesion bypass of small alkyl adducts is a common feature of transcription, and (ii) when lesions do stall RNAPs and elicit TCR, it is common for transcription not to restart from that damaged site (Fig 4.d).

Our finding that RNAP frequently bypasses DEN-induced lesions *in vivo*, extends previous *in vitro* studies (You et al. 2014; Xu et al. 2017) that have considered RNAP bypass of O^4-EtdT, the principle mutagenic adduct of DEN, and complements findings for other non-bulky adducts (Saxowsky and Doetsch 2006; You et al. 2012). However, the exact molecular mechanisms that lead to lesion bypass versus stalling and repair are presently unclear. For alkyl adducts, both nucleotide insertion and RNAP extension past damage can cause prolonged pausing, potentially facilitating damage recognition (Xu et al. 2017). Thus, contributing factors to the stochasticity of TCR upon lesion encounter may include the sequence of the DNA-RNA hybrid and/or local nucleotide concentrations. Regardless of the mechanism of lesion bypass, combining our estimates of lesion bypass frequency with the lack of fidelity of RNAP over alkyl adducts (You et al. 2014), suggests that alkylating agents can induce considerable transcriptional mutagenesis.

Following completion of TCR, it has been widely thought that RNAP restarts transcription from the site of damage (Geijer and Marteijn 2018). However, recent work on bulky UV-induced cyclobutane pyrimidine dimers (Chiou et al. 2018) challenges the universality of this model, reporting that RNAP dissociates from DNA at the damaged site and subsequent transcription initiation at the genic promoter is required for transcript synthesis. Our results corroborate these latter findings and extend them to the alkylation damage induced by DEN. The observed 5' bias of repair coupled with mathematical modelling indicates that RNAP does not always restart following repair. Furthermore, through analysing parameter regimes within bootstrap uncertainty (Fig. 4c) and model selection analysis (Fig. 4d), we conclude that our data are entirely consistent with RNAP always disassociating after repair. The 5' repair bias echoes the enhanced 5' repair found in the damage and repair maps generated from pyrimidine dimers (Hu et al. 2017) and agrees with the finding that TCR efficiency corresponds to gene length (Zeitler et al. 2022). Our finding that transcription does not consistently restart from the stall site following repair is particularly relevant when multiple lesions exist per gene, suggesting that damage-induced expression repression will disproportionately affect long (Stoeger et al. 2022), and lowly expressed genes. Supporting this hypothesis, *in vitro* damage experiments show that the degree of expression reduction was correlated with gene length following exposure to UV, the chemotherapeutic cisplatin, and the cigarette smoke component benzo(a)pyrene (Merav et al. 2024).

The gradient of mutation density we observe through gene bodies has implications for the accurate modelling of mutation patterns (Alexandrov et al. 2020; Vöhringer et al. 2021), necessary for the prediction of oncogenic selection (Muiños et al. 2021). Our model provides sufficient damage for this gradient to manifest, arising due to inefficient repair at downstream positions caused by the dissociation of RNAP. The co-dependency of damage burden and expression level enriches the developing mechanistic understanding of mutation patterns over the genome (Alexandrov et al. 2020; Seplyarskiy and Sunyaev 2021). Mutation patterns resulting from a high damage burden are not simply an amplification of the patterns expected from a lower dose of damage.

Quantitatively mapping the consequences of endogenous and exogenous DNA damage is necessary to understand mutagenesis, gene expression dysregulation, and the impact of environmental and therapeutic agents. Here, we have developed an integrative view of TCR following alkyl damage, complementing existing experimental assays that measure individual aspects of this fundamental repair process. Our results exemplify how mechanistic quantitative modelling can be used to bridge the molecular processes of damage and repair through to their presentation in large-scale genomics data.

Methods

DNA sequencing variant calling

The C3H/HeJ mouse strain reference genome assembly C3H_HeJ_v1 (Lilue et al. 2018) was used for read mapping, annotation and analysis. Mutation calling and quality filtering was performed using whole genome sequencing of 371 DEN induced liver tumours from n=104 male C3H mice, as previously reported (Aitken et al. 2020). A minimum variant allele frequency (VAF) threshold of 10% was applied to remove mutation calls from contaminating non-clonal cells. All mutation data was derived from sequence data in the European Nucleotide Archive (ENA) under accession PRJEB37808 and processed files directly used as input for this work are publicly available https://doi.org/10.1038/s41586-020-2435-1. Gene annotation in C3H_HeJ_v1 coordinates was obtained from Ensembl v.91 (Howe et al. 2021).

Mutation phasing

Genomic segmentation on mutational asymmetry was performed as previously reported (Aitken et al. 2020). In brief, mutational strand asymmetry was scored for each genomic segment using the relative difference metric S=(F-R)/(F+R) where F is the rate of mutations from T on the forward (plus) strand of the reference genome and R the rate of mutations from T on the minus strand (mutations from A on the plus strand). The phasing of mutation asymmetry is agnostic to which base harbours the mutagenic lesion, orthogonal data is required to resolve which asymmetry indicates the lesion containing strand. In the case of A versus T asymmetry from DEN damage prior studies have established T rather than A modification as the principal mutagenic lesion (Singer 1985; Mientjes et al. 1998; Aitken et al. 2020). A mutational asymmetry score of S >0.33 was used to identify the inheritance of forward strand lesions and S <-0.33 as the inheritance of reverse strand lesions. Analyses were confined to n=237, clonally distinct DEN induced tumours that met the combined criteria of: (i) not labelled as symmetric (mutationally symmetric tumours defined as >99% of autosomal mutations in genomic segments with abs(S) <0.2, see (Anderson et al. 2022)), (ii) tumour cellularity >50%, and (iii) >80% of substitution mutations attributed to the DEN1 signature (Aitken et al. 2020) by sigFit (v.2.0) (Gori and Baez-Ortega 2018).

Relative to the reference genome sequence, a plus (P) strand gene is transcribed using the reverse (R) strand as a template. So a P strand gene in a genomic segment with R strand lesions (denoted RP orientation) is expected to be subject to transcription coupled repair. A minus strand (M) gene with forward (F) strand lesions (FM orientation) is also expected to be subject to transcription coupled repair, as the retained lesions are on the transcription template strand. Conversely FP and RM orientation combinations will have lesions on the non-template strand for transcription and are therefore not expected to be subject to transcription coupled repair.

Gene expression

Paired-end, stranded total RNA-seq from C3H male mouse livers not exposed to DEN (n=4, matching the developmental time of mutagenesis, postnatal day 15, P15) was previously generated and is available from Array Express under accession E-MTAB-8518. RNA-seq was aligned to the reference genome C3H_HeJ_v1 using the splice aware aligner Star (v2.7.6a). A C3H liver specific splice junction database was generated from an initial round of RNA-seq read alignment to the C3H_HeJJ_v1 reference genome guided by Ensembl (v.91) genomic annotation. Using the sex, strain, and tissue matched splice junction database, a second iteration of Star alignment produced a final RNA to genome alignment with output attribute flags set to preserve read orientation information (outSAMattributes: NH HI AS nM). The transcription strand of RNA-seq reads was resolved using read-end and mapping orientation extracted by Samtools view (v.1.7.0) and read-pairs exclusively mapping within annotated exons were identified using Bedtools intersect (v.2.29.2). Intronic read-pairs were defined as those mapping within a genic span, derived from a sense-strand transcript, and not in the exonic set. Only read-pairs with a mapping quality (MAPQ) >10 were used to quantify gene expression. Nascent transcription was quantified by counting read-pairs in the intronic set using Bedtools multicov (v.2.29.2). The read count was normalised to reads per kilobase of analysed intron for each gene in each sequence library, and then normalised to nascent transcripts per million (nTPM) for each library. The final nascent transcript expression estimate per gene was taken as the mean of nascent TPM over replicate libraries. Nascent transcription estimates could be generated for 85% (n=17,304) of protein coding genes. Overlapping genes, defined by primary transcript coordinates, were hierarchically excluded from analysis: Starting with the most expressed gene, any overlapping less-expressed genes were excluded. Code for this analysis is available at: https://github.com/CraigJAnderson/lce-si_nascent.

Genes with similar estimates of nascent expression were aggregated for analysis of transcription coupled repair. The sigmoidal distribution relating nascent transcription rate to mutation rate (Fig 2.b) was segmented using linear regression models in the R package Segmented (v.1.3-3) (Muggeo 2003). This defined n=4,649 genes with zero or low detected nascent expression (<0.287 nTPM) in which reduced mutation rates associated with transcription coupled repair are essentially undetectable; subsequently stratum 1 genes (light blue in plots). Genes expressed at a greater rate than segmentation threshold >3.73 nTPM do not show a further decrease in mutation rate with increased expression; these n=7,176 highly expressed genes were defined as stratum 6 (bright red in plots). The n=4,005 genes with intermediate expression (0.287-3.73 nTPM) exhibited a log-linear relationship between expression and mutation rate. These were quantile split into strata 2 to 5, containing approximately 1,000 genes each. The median nascent expression for the six expression strata were (0, 0.49, 1.16, 2.07, 3.14, 11.15 nTPM).

Mutation rates

Strand resolved mutation rates were calculated as previously described (Aitken et al. 2020; Anderson et al. 2022). Vectors of 192 categories representing every possible single-nucleotide substitution conditioned on the identity of both 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 fraction of each trinucleotide in a reference sequence, for example the composition of the whole genome. Strand-specific mutation rates were calculated with respect to the lesion containing strand, with both mutation calls and sequence composition reverse complemented for reverse strand lesions. Autosomal chromosomes were considered diploid and the X chromosome haploid (all mice were male) for the purposes of calculating mutation rates and sequence composition.

Mutation rate versus expression

For those genes with measured nascent expression, genes with mean nTPM <0.01 were grouped (n=1757), as were genes with mean nTPM>100 (n=587). The remaining genes were equally split amongst 15 bins, resulting in a total of 17 expression bins. For each tumour, for each expression bin, the mutation rate due to template strand and non-template strand lesions was calculated as detailed above (proportion of mutated bases for given trinucleotide context). The average mutation rate for each strand was calculated similarly but without grouping genes by expression. Observed:expected as a function of expression (Fig 2.b, lower panel) was calculated as the ratio of template strand mutation rate to the non-template strand mutation rate. For each tumour, the expression-dependent observed:expected was fit to a four-parameter log-logistic model using the R package drc (Ritz et al. 2015) (Extended Data Fig 1.a,b).

Modelling transcription coupled repair

We defined a probabilistic model of lesion detection by RNAP (variable parameter *Pd*), and its subsequent re-initiation (*Pr*) or disassociation (1-*Pr*). The model also incorporated variables for the fraction of lesions that are visible to TCR (*Pv*) and a multiplier parameter (*m*) to translate experimental measurements of nascent TPM (nTPM) to the number of transcription initiations between mutagenesis and DNA replication. The model is illustrated in

Fig 3.a, and a detailed description is given in Supplementary File 1. The model was analysed both by stochastic simulations (details below) and analytic methods (details in Supplementary File 1). The analytic methods were used for parameter inference, which were assessed by simulation. The experimental nascent expression values determined for each strata (see ‘Gene Expression’, above) were used both for simulated data and for analysis of the tumour data.

Simulated mutagenesis and transcription coupled repair

For a given parameter set (*Pd*, *Pr*, *m, Pv*), we simulated damage and TCR on 1,940,237 phaseable genes, which is the cumulative number of phaseable genes from the mouse liver experiment. For each phaseable gene, the gene length was sampled from the length distribution of the filtered C3H gene list (see above, ‘Gene Expression’). The gene length was multiplied by the median per base mutation rate (13 x 10^{-6}/bp (Aitken et al. 2020)) resulting in the expected lesion number for that gene. The realised lesion number was obtained by sampling a Poisson distribution with mean given by the expected lesion number. Each lesion was placed on the gene at a location determined by sampling from a uniform distribution over [0, gene length]. Each gene was assigned to 1 of 6 expression strata with probabilities given by the strata proportions in the murine data. Each stratum is associated with a measured nascent transcription value *e*, and of the genes in a given stratum we assume a proportion *c* have floor(*e*.*m*) RNAPs that initiate transcription, while the other 1-*c* fraction of genes have floor(*e*.*m*) +1 RNAPs that initiate transcription. For given (*m, e*), *c* is uniquely given by 1-(*e.m* - floor(*e.m*)) (see Supplementary File 1). Thus, for our simulated gene in stratum *e*, we assign either floor(*e.m*) or floor(*e.m*) +1 RNAPs to initiation transcription with probabilities (*c*, 1-*c*). The RNAPs sequentially initiate transcription, and lesion detection and restart of the polymerases follow the rules illustrated in Fig 3.a, potentially resulting in lesion clearance. After all RNAPs have initiated and terminated transcription (potentially even bore the TES in the case of non-restart), the remaining lesion locations were recorded.

Lesion locations were converted to their position in units of ‘expected upstream lesions’ (base-pair location times 13 x 10^{-6}) and a spatial grid of 40 windows of width 0.1 expected lesions was applied (only few genes are long enough for >4 expected upstream lesions, thus further spatial grids would harbour substantial noise). Aggregating over all simulated genes, the summed number of lesions with positions within each spatial window was determined, resulting in the ‘observed’ lesion count. In the absence of TCR, for a given spatial bin, the aggregated lesion number is 0.1 multiplied by the number of phaseable genes with upstream lesion length not exceeding the right boundary of the spatial bin, resulting in the ‘expected’ lesion count for that bin. For each bin, the ratio of the ‘observed’ to the ‘expected’ resulted in the simulated observed:expected mutation rates.

Parameter inference on simulated or murine liver tumour data

With input as observed:expected mutation rates with 6 expression strata and 40 spatial windows through the gene in units of expected upstream lesions, parameter inference was performed as follows. Using the numerical output from the obs:exp$_{\text{theory}}$ expressions, the Manhattan distance (L_1 norm) between those 6x40 measures and the equivalent input data was minimised. Parameter space was initially explored as a grid-search. Probabilities *Pd*, *Pr*, and *Pv* were bounded at min=0, max=1 with steps of 0.01.

For both simulation and fitting of real data, the parameter range for the expression multiplier m was bounded at min=0.25, max=10 with steps of 0.25. This range was defined following initial grid search exploration with m=50/i for i=1, ..., 200, the rationale for the parameter bounds is given below in the paragraph ‘Plausible expression multiplier parameter ranges’. The optimal parameters obtained from the grid search were provided as the starting point for optimisation implemented in the R optim function (R Core Team 2020) with default parameters to return the final optimised parameter values.

To calculate confidence intervals, the observed:expected mutation rates for the six expression strata were re-calculated from the bootstrap sampling of genes (sampling with replacement to original gene list size, n=1,000 replicates for murine data, n=100 for simulated data). The inference procedure outlined above was performed for each bootstrapped dataset and reported 95% confidence intervals were calculated as the 0.025 and 0.975 quantiles of bootstrapped parameter estimates.

For AIC-based model selection on the murine data, the measured obs:exp values were assumed to be drawn from a normal distribution with mean obs:exp$_{\text{theory}}$ computed as detailed in Supplementary File 1, with a common variance *v*. Optimal fits were found by maximising the likelihood using the ‘L-BFGS-B’ method using the mle2 function from the R package bbmle2 (Ben Bolker and R Development Core Team 2022). Maximum likelihood estimates for parameters allowing restart were *Pd*=0.42, *Pr*=0.66, *m*=1.59, *Pv* =0.8, v=8.8*10^{-4}; maximum likelihood estimates for parameter without restart were *Pd*=0.18, *Pr*=0, *m*=4.14, *Pv* =0.8, v=8.9*10^{-4}.

Interpretation of expression multiplier *m*

For each expression stratum *k* we assume that, for each gene in that stratum, the average number of transcription initiation events between damage and replication, n_k, is related to the average expression (nTPM) over all genes in that stratum, e_k, by

$$n_k=m*e_k.$$

The variable *m* can be viewed solely as part of our statistical model, however it can be given a biological interpretation under some assumptions. Let the number of genes in stratum *k* be g_k. We assume that the gene expression for a given stratum is constant over time and that the RNA sequencing is reflective of this stable expression in the mutagenised cell. If RNA pol II can fail to restart transcription after repair (*Pr*<1) then not every transcription initiation will result in a transcript, hence let s_k be the probability a transcription initiation of a stratum *k* gene results in a transcript. Further, assume that a proportion p_k of these transcripts are detected in the RNA sequencing. Then the number of transcripts from stratum *k* detected in the RNA seq would be g_k* n_k*s_k*p_k .

Recall that by using units of nTPM, the interpretation of the expression level is that for every million nascent transcripts measured, e_k transcripts are apportioned to each gene in stratum *k*. Therefore, a total of g_k*e_k transcripts would be apportioned to stratum *k* for every million transcripts.

Hence

$$g_k * e_k = 10^6 * g_k * n_k * s_k * p_k / \sum_{\square=1}^{6} g_k * n_k * s_k * p_k ,$$

where the right hand side of the equation arises from multiplying 1 million with the proportion of transcripts produced and detected from stratum *k* genes.

So, as by definition $n_k = m * e_k$,

$$m = \sum_{\square=1}^{6} (g_k * n_k * s_k * p_k) / (10^6 * s_k * p_k).$$

Assuming that the s_k and p_k remain constant over each stratum,

$$m = 10^{-6} \sum_{\square=1}^{6} g_k * n_k .$$

Hence *m* is the number of transcription initiation events (measured in units of million initiations) between damage and replication.

Plausible expression multiplier *m* parameter ranges

We draw on prior literature for plausible parameter values for *m*, which, as discussed above, is the number of transcription initiations (x10^6) in a cell between DNA damage and replication. Note that when modelling the DEN mutagenesis murine experiment, the number of transcription initiations may not be directly equal to the number of transcripts produced as polymerases may not restart after lesion detection (in the most extreme case with *Pd=1, Pr=0* and *i* initial lesions, then the number of transcripts produced is equal to the transcription initiations - *i*). However, when comparing to non-mutagenesis experiments, where lesion numbers are expected to be greatly reduced, we equate transcript number and the number of transcription initiations.

For a lower bound on *m*, the number of transcription initiations (x10^6) between damage and replication, we note that an average time of 2,280 minutes between damage and DNA replication was estimated from the cell-cycle times of DEN mutagenised rat hepatocytes (Rotstein et al. 1984). As the the median mRNA half-life has been estimated as 139 minutes (Rabani et al. 2014), the transcript number measured at any moment can serve as a lower bound for the transcript initiation number; as the typical range estimated is 200-300k transcripts per mammalian cell (Velculescu et al. 1999; Marinov et al. 2014; Shapiro, Biezuner, and Linnarsson 2013), we adopt a lower bound of *m*=0.25. For a generous upper bound, we assume: 180,000 chromatin associated RNA Pol II complexes exist per cell (Kimura et al. 1999); all polymerases are continuously actively transcribing and only transcribing annotated genes; an average transcription rate of 2 kb min^{-1} in mouse liver (Bahar Halpern et al. 2015); a median gene length of 60 kb; and again 2,280 minutes between damage and replication. This implies 13.68 million transcripts are produced, hence *m*=13.68, and thus *m*=50 is a further upper bound for the parameter space used in inference. For a reduced upper bound, we note that of the 180,000 chromatin associated RNA Pol II complexes per cell measured in Kimura et al, only 110,000 were of the hyperphosphorylated form IIO - implying active elongation. Assuming only 110,000 RNA Pol II complexes actively transcribe between damage and replication implies that 8.36 million transcripts are produced; for this reason our simulated datasets were generated over a grid with an upper bound of *m*=8.5.

Orthogonal estimate of expression multiplier *m*

Bahar Halpern et al. (Bahar Halpern et al. 2015) estimated the transcription rate and proportion of time a gene is being transcribed in mouse hepatocytes using single molecule transcript counting; we focus on their periportal samples from mice in the “fed” condition. Taking the product of the estimated transcription parameters, and multiplying by the time between damage and replication (again assumed to be 2,280 minutes), provides an estimate for the number of transcripts produced by these genes before replication, a per gene estimate of *n*. Seven genes were both measured by single molecule transcript counting (Bahar Halpern et al. 2015) and quantified as nTPM from our RNA-seq data. Throughout we have assumed that for each set of genes that are associated to an expression stratum *k*, that $n_k=m*e_k$. If now, we assume this holds on a per-gene basis, that is for each gene $n=m*e$, then as both *n* and *e* are estimated per gene, we can readily infer *m*. The optimal least square fit for $\log_{10}(n)= \log_{10}(e)+\log_{10}(m)$ resulted in an *m* estimate of 2.77 (Extended Data Fig 2.a). Note that as the experiments of Bahar Halpern et al. occurred outside of a mutagenesis setting, we have again equated the number of transcripts with the number of transcription initiations *n*.

Acknowledgements

We thank the Liver Cancer Evolution (LCE) Consortium for informative discussions. This work was supported by: the MRC Human Genetics Unit core funding programme grants (MC_UU_00007/11, and MC_UU_00035/2), MRC Toxicology Unit core funding (RG94521), Cancer Research UK Cambridge Institute core funding (20412), Cancer Research UK (grant A22398), the Wellcome Trust (WT202878/B/16/Z), and European Research Council (615584, 788937). M.D.N is a cross-disciplinary post-doctoral fellow supported by funding from CRUK Brain Tumour Centre of Excellence Award (C157/A27589). This work made use of resources provided by the Edinburgh Compute and Data Facility (ECDF), including the IGC_Eddie3 high performance storage arrays (MC_PC-MR/X013677/1)

List of supplementary files

Extended Data Figures 1-2.

Supplementary File 1 | Mathematical model for DNA damage and transcription coupled repair (PDF).

References

Aitken, Sarah J., Craig J. Anderson, Frances Connor, Oriol Pich, Vasavi Sundaram, Christine Feig, Tim F. Rayner, et al. 2020. “Pervasive Lesion Segregation Shapes Cancer Genome Evolution.” *Nature* 583 (7815): 265–70.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human Cancer.” *Nature* 578 (7793): 94–101.

Anderson, Craig J., Lana Talmane, Juliet Luft, Michael D. Nicholson, John Connelly, Oriol Pich, Susan Campbell, et al. 2022. “Strand-Resolved Mutagenicity of DNA Damage and Repair.” *bioRxiv*, January, 2022.06.10.495644.

Bahar Halpern, Keren, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. 2015. “Bursty Gene Expression in the Intact Mammalian Liver.” *Molecular Cell* 58 (1): 147–56.

Ben Bolker and R Development Core Team. 2022. *Tools for General Maximum Likelihood Estimation [R Package Bbmle Version 1.0.25]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=bbmle.

Brégeon, Damien, and Paul W. Doetsch. 2011. “Transcriptional Mutagenesis: Causes and Involvement in Tumour Development.” *Nature Reviews. Cancer* 11 (3): 218–27.

Chiou, Yi-Ying, Jinchuan Hu, Aziz Sancar, and Christopher P. Selby. 2018. “RNA Polymerase II Is Released from the DNA Template during Transcription-Coupled Repair in Mammalian Cells.” *The Journal of Biological Chemistry* 293 (7): 2476–86.

Connor, Frances, Tim F. Rayner, Sarah J. Aitken, Christine Feig, Margus Lukk, Javier Santoyo-Lopez, and Duncan T. Odom. 2018. “Mutational Landscape of a Chemically-Induced Mouse Model of Liver Cancer.” *Journal of Hepatology* 69 (4): 840–50.

Geijer, Marit E., and Jurgen A. Marteijn. 2018. “What Happens at the Lesion Does Not Stay at the Lesion: Transcription-Coupled Nucleotide Excision Repair and the Effects of DNA Damage on Transcription in Cis and Trans.” *DNA Repair* 71 (November): 56–68.

Gori, Kevin, and Adrian Baez-Ortega. 2018. “Sigfit: Flexible Bayesian Inference of Mutational Signatures.” *bioRxiv*. bioRxiv. https://doi.org/10.1101/372896.

Gregersen, Lea H., and Jesper Q. Svejstrup. 2018. “The Cellular Response to Transcription-Blocking DNA Damage.” *Trends in Biochemical Sciences* 43 (5): 327–41.

Haradhvala, Nicholas J., Paz Polak, Petar Stojanov, Kyle R. Covington, Eve Shinbrot, Julian M. Hess, Esther Rheinbay, et al. 2016. “Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair.” *Cell* 164 (3): 538–49.

Howe, Kevin L., Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, et al. 2021. “Ensembl 2021.” *Nucleic Acids Research* 49 (D1): D884–91.

Hu, Jinchuan, Sheera Adar, Christopher P. Selby, Jason D. Lieb, and Aziz Sancar. 2015. “Genome-Wide Analysis of Human Global and Transcription-Coupled Excision Repair of UV Damage at Single-Nucleotide Resolution.” *Genes & Development* 29 (9): 948–60.

Hu, Jinchuan, Ogun Adebali, Sheera Adar, and Aziz Sancar. 2017. “Dynamic Maps of UV Damage Formation and Repair for the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (26): 6758–63.

Kimura, H., Y. Tao, R. G. Roeder, and P. R. Cook. 1999. “Quantitation of RNA Polymerase II and Its Transcription Factors in an HeLa Cell: Little Soluble Holoenzyme but Significant Amounts of Polymerases Attached to the Nuclear Substructure.” *Molecular and Cellular Biology* 19 (8): 5383–92.

Lans, Hannes, Jan H. J. Hoeijmakers, Wim Vermeulen, and Jurgen A. Marteijn. 2019. “The DNA Damage Response to Transcription Stress.” *Nature Reviews. Molecular Cell Biology* 20 (12): 766–84.

Lilue, Jingtao, Anthony G. Doran, Ian T. Fiddes, Monica Abrudan, Joel Armstrong, Ruth Bennett, William Chow, et al. 2018. “Sixteen Diverse Laboratory Mouse Reference Genomes Define Strain-Specific Haplotypes and Novel Functional Loci.” *Nature Genetics* 50 (11): 1574–83.

Marietta, Cheryl, and Philip J. Brooks. 2007. “Transcriptional Bypass of Bulky DNA Lesions Causes New Mutant RNA Transcripts in Human Cells.” *EMBO Reports* 8 (4): 388–93.

Marinov, Georgi K., Brian A. Williams, Ken McCue, Gary P. Schroth, Jason Gertz, Richard M. Myers, and Barbara J. Wold. 2014. “From Single-Cell to Cell-Pool Transcriptomes: Stochasticity in Gene Expression and RNA Splicing.” *Genome Research* 24 (3): 496–510.

Merav, May, Elnatan M. Bitensky, Elisheva E. Heilbrun, Tamar Hacohen, Ayala Kirshenbaum, Hadar Golan-Berman, Yuval Cohen, and Sheera Adar. 2024. “Gene Architecture Is a Determinant of the Transcriptional Response to Bulky DNA Damages.” *Life Science Alliance* 7 (3). https://doi.org/10.26508/lsa.202302328.

Mientjes, E. J., A. Luiten-Schuite, E. van der Wolf, Y. Borsboom, A. Bergmans, F. Berends, P. H. Lohman, R. A. Baan, and J. H. van Delft. 1998. “DNA Adducts, Mutant Frequencies, and Mutation Spectra in Various Organs of Lambda lacZ Mice Exposed to Ethylating Agents.” *Environmental and Molecular Mutagenesis* 31 (1): 18–31.

Muggeo, Vito M. R. 2003. “Estimating Regression Models with Unknown Break-Points.” *Statistics in Medicine* 22 (19): 3055–71.

Muiños, Ferran, Francisco Martínez-Jiménez, Oriol Pich, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. 2021. “In Silico Saturation Mutagenesis of Cancer Genes.” *Nature* 596 (7872): 428–32.

Rabani, Michal, Raktima Raychowdhury, Marko Jovanovic, Michael Rooney, Deborah J. Stumpo, Andrea Pauli, Nir Hacohen, et al. 2014. “High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies.” *Cell* 159 (7): 1698–1710.

R Core Team. 2020. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ritz, Christian, Florent Baty, Jens C. Streibig, and Daniel Gerhard. 2015. “Dose-Response Analysis Using R.” *PloS One* 10 (12): e0146021.

Rotstein, J., P. D. Macdonald, H. M. Rabes, and E. Farber. 1984. “Cell Cycle Kinetics of Rat Hepatocytes in Early Putative Preneoplastic Lesions in Hepatocarcinogenesis.” *Cancer Research* 44 (7): 2913–17.

Sarsam, Reta D., Jun Xu, Indrajit Lahiri, Wenzhi Gong, Qingrong Li, Juntaek Oh, Zhen Zhou, et al. 2024. “Elf1 Promotes Rad26’s Interaction with Lesion-Arrested Pol II for Transcription-Coupled Repair.” *Proceedings of the National Academy of Sciences* 121 (3): e2314245121.

Saxowsky, Tina T., and Paul W. Doetsch. 2006. “RNA Polymerase Encounters with DNA Damage: Transcription-Coupled Repair or Transcriptional Mutagenesis?” *Chemical Reviews* 106 (2): 474–88.

Seplyarskiy, Vladimir B., Evgeny E. Akkuratov, Natalia Akkuratova, Maria A. Andrianova, Sergey I. Nikolaev, Georgii A. Bazykin, Igor Adameyko, and Shamil R. Sunyaev. 2019. “Error-Prone Bypass of DNA Lesions during Lagging-Strand Replication Is a Common Source of Germline and Cancer Mutations.” *Nature Genetics* 51 (1): 36–41.

Seplyarskiy, Vladimir B., and Shamil Sunyaev. 2021. “The Origin of Human Mutation in Light of Genomic Data.” *Nature Reviews. Genetics* 22 (10): 672–86.

Shapiro, Ehud, Tamir Biezuner, and Sten Linnarsson. 2013. “Single-Cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science.” *Nature Reviews. Genetics* 14 (9): 618–30.

Singer, B. 1985. “In Vivo Formation and Persistence of Modified Nucleosides Resulting from Alkylating Agents.” *Environmental Health Perspectives* 62 (October): 41–48.

Stoeger, Thomas, Rogan A. Grant, Alexandra C. McQuattie-Pimentel, Kishore R. Anekalla, Sophia S. Liu, Heliodoro Tejedor-Navarro, Benjamin D. Singer, et al. 2022. “Aging Is Associated with a Systemic Length-Associated Transcriptome Imbalance.” *Nature Aging* 2 (12): 1191–1206.

Tornaletti, Silvia, Lauren S. Maeda, Richard D. Kolodner, and Philip C. Hanawalt. 2004. “Effect of 8-Oxoguanine on Transcription Elongation by T7 RNA Polymerase and Mammalian RNA Polymerase II.” *DNA Repair* 3 (5): 483–94.

Vandenberg, Brittany N., Marian F. Laughery, Cameron Cordero, Dalton Plummer, Debra Mitchell, Jordan Kreyenhagen, Fatimah Albaqshi, et al. 2023. “Contributions of Replicative and Translesion DNA Polymerases to Mutagenic Bypass of Canonical and Atypical UV Photoproducts.” *Nature Communications* 14 (1): 2576.

Velculescu, V. E., S. L. Madden, L. Zhang, A. E. Lash, J. Yu, C. Rago, A. Lal, et al. 1999. “Analysis of Human Transcriptomes.” *Nature Genetics* 23 (4): 387–88.

Verna, Lynne, John Whysner, and Gary M. Williams. 1996. “N-Nitrosodiethylamine Mechanistic Data and Risk Assessment: Bioactivation, DNA-Adduct Formation, Mutagenicity, and Tumor Initiation.” *Pharmacology & Therapeutics* 71 (1): 57–81.

Vöhringer, Harald, Arne Van Hoeck, Edwin Cuppen, and Moritz Gerstung. 2021. “Learning Mutational Signatures and Their Multidimensional Genomic Properties with TensorSignatures.” *Nature Communications* 12 (1): 3628.

Walmacq, Celine, Alan C. M. Cheung, Maria L. Kireeva, Lucyna Lubkowska, Chengcheng Ye, Deanna Gotte, Jeffrey N. Strathern, Thomas Carell, Patrick Cramer, and Mikhail Kashlev. 2012. “Mechanism of Translesion Transcription by RNA Polymerase II and Its Role in Cellular Resistance to DNA Damage.” *Molecular Cell* 46 (1): 18–29.

Xu, Liang, Wei Wang, Jiabin Wu, Ji Hyun Shin, Pengcheng Wang, Ilona Christy Unarta, Jenny Chong, Yinsheng Wang, and Dong Wang. 2017. “Mechanism of DNA Alkylation-Induced Transcriptional Stalling, Lesion Bypass, and Mutagenesis.” *Proceedings of the*

National Academy of Sciences of the United States of America 114 (34): E7082–91.

You, Changjun, Xiaoxia Dai, Bifeng Yuan, Jin Wang, Jianshuang Wang, Philip J. Brooks, Laura J. Niedernhofer, and Yinsheng Wang. 2012. "A Quantitative Assay for Assessing the Effects of DNA Lesions on Transcription." *Nature Chemical Biology* 8 (10): 817–22.

You, Changjun, Pengcheng Wang, Xiaoxia Dai, and Yinsheng Wang. 2014. "Transcriptional Bypass of Regioisomeric Ethylated Thymidine Lesions by T7 RNA Polymerase and Human RNA Polymerase II." *Nucleic Acids Research* 42 (22): 13706–13.

Yousefzadeh, Matt, Chathurika Henpita, Rajesh Vyas, Carolina Soto-Palma, Paul Robbins, and Laura Niedernhofer. 2021. "DNA Damage—how and Why We Age?" *eLife* 10 (January): e62852.

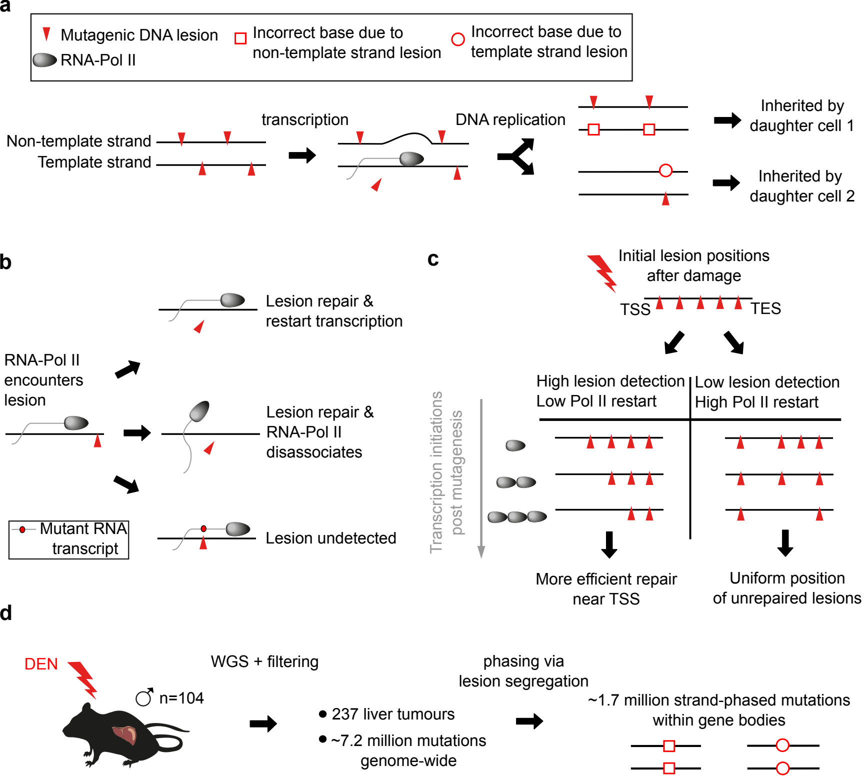
Zeitler, Leo, Cyril Denby Wilkes, Arach Goldar, and Julie Soutourina. 2022. "A Quantitative Modelling Approach for DNA Repair on a Population Scale." *PLoS Computational Biology* 18 (9): e1010488.

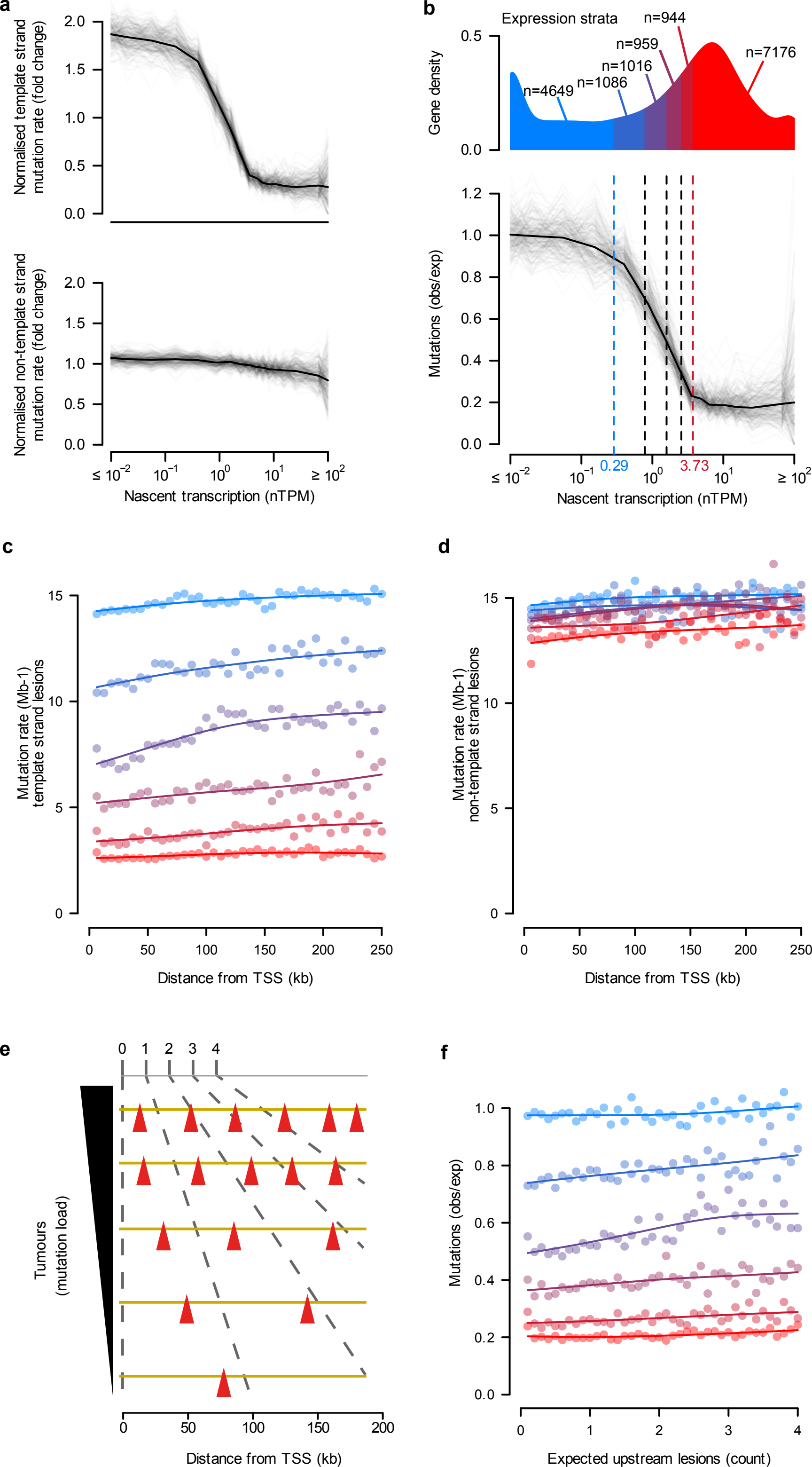
Figure 1 | Quantifying the dynamics of transcription coupled DNA repair with lesion-strand phased mutations and gene expression measures. a, Template strand DNA damage is alleviated during transcription by transcription coupled repair. Lesions that persist to replication can cause heritable mutations created through incorrect base-pairing. **b,** Alternate possible outcomes from transcription over a lesion-containing template DNA strand. **c,** Schematic of lesion clearance due to TCR following damage. The pattern of remaining lesions as a function of both expression and genic-position is dependent on the sensitivity of RNAP and whether the RNAP restarts following repair. **d,** We utilise strand-phased mutation data from 237 liver tumours induced by exposing male C3H mice to a single dose of DEN.

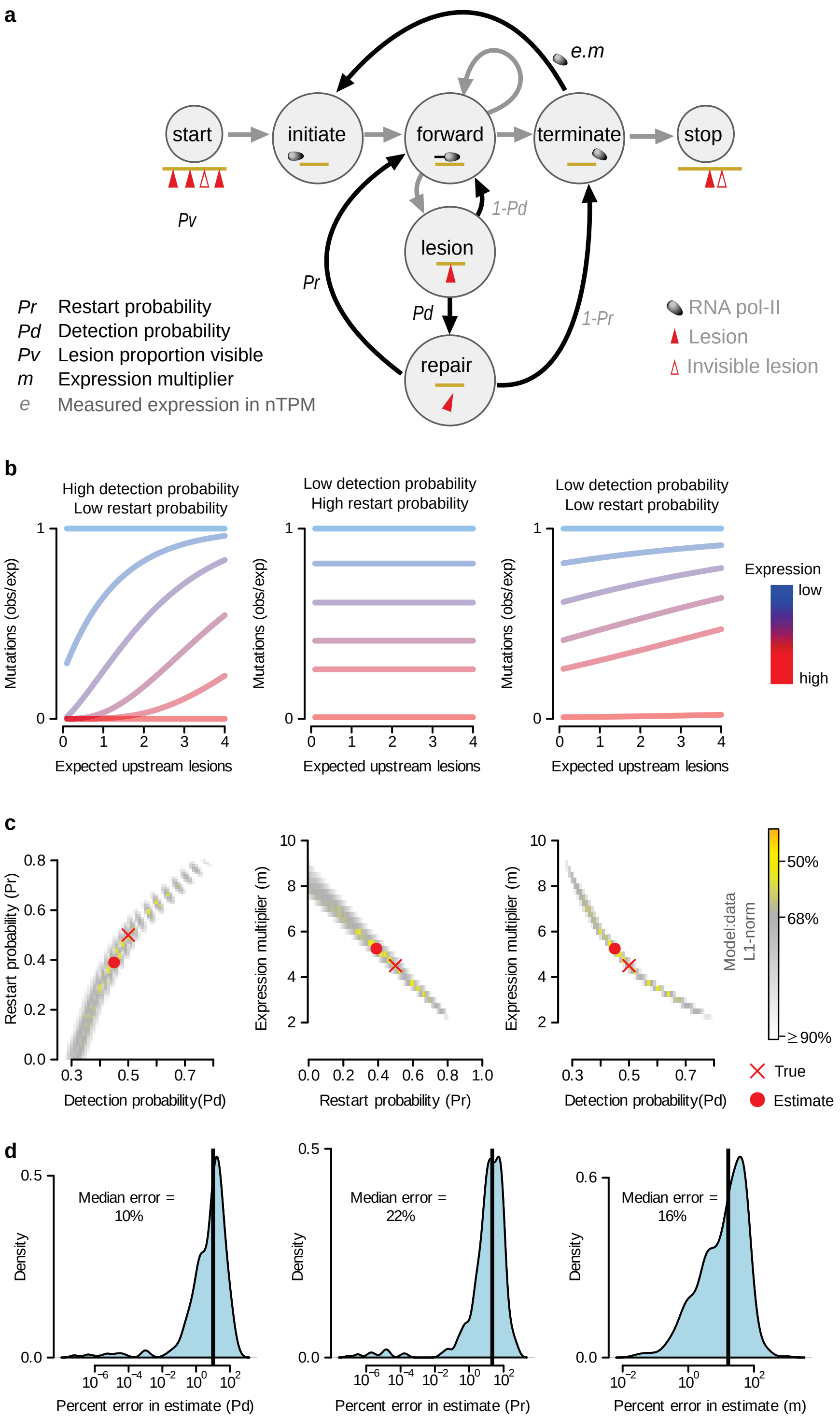
Figure 2 | Transcription coupled repair shapes the distribution of mutations through the body of expressed genes. a, Tumours (grey curves) consistently show the same normalised profile of transcription coupled repair: Increased expression (x-axis; plotted on log scale) corresponding to reduced mutation rate (y-axis) for lesions on the transcription template strand (upper panel). The mutation rate per tumour is normalised to the average for all genes in the tumour. For lesions on the non-template strand (lower panel), increased expression only subtly influences normalised mutation rate. Black line is the median of the per tumour rates. **b,** Lower panel shows observed versus expected mutations (y-axis) calculated as the ratio of template strand mutation rate to non-template strand mutation rate plotted against nascent transcription rate per tumour (x-axis). Expression >3.73 nascent transcripts per million (nTPM) does not further decrease the mutation rate. In subsequent analyses gene expression is binned into six strata of nascent gene expression (upper panel) blue→red denotes increasing expression, dashed lines demarcating strata boundaries (Methods). **c,** Mutation rates for genes with template strand lesions. Genes classified by expression strata and mutation rates calculated in 5 kb consecutive windows from the transcription start site (TSS). Points show observed data and curves show best-fit splines (3 degrees of freedom). **d,** As for c but considering genes with non-template strand lesions. **e,** Schematic of per-tumour normalisation to calculate the number of expected upstream lesions (red triangles) for each analysis window (Methods). **f,** Observed versus expected mutations (y-axis) calculated as the ratio of template to non-template strand. Expected upstream lesion count (x-axis) categories as per e. Points represent data while curves show best-fit splines (3 degrees of freedom). Genes with intermediate levels of expression (strata 2-5) exhibit a lower mutation rate at their 5’ end.

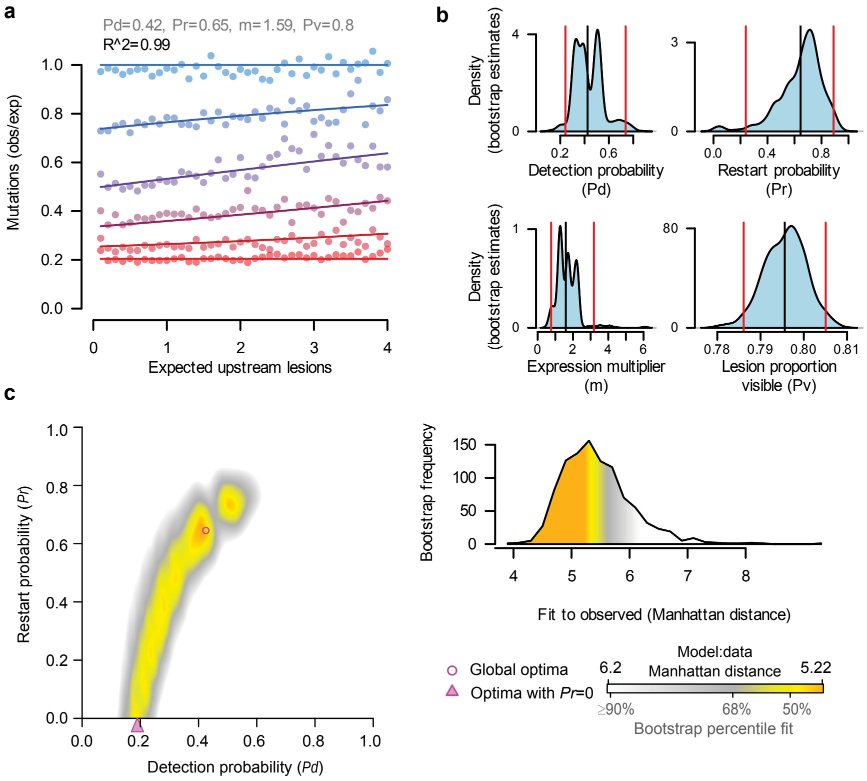
Figure 3 | Mathematical model of transcription coupled repair dynamics. a, Mathematical model of TCR dynamics. A string of nucleotides (yellow line) with DNA lesions (red triangles) is subject to transcription (grey arrows), and probabilistic TCR events (black arrows). On encountering a lesion, the probability of its detection (*Pd*) and of polymerase restart following lesion repair (*Pr*) are independent model variables. The fraction of lesions visible to TCR (*Pv*) and an expression multiplier parameter (*m*) are additional independent variables. **b,** Example mutation rate profiles generated analytically by the model under varied qualitative parameter regimes. Numerical parameters of *(Pd, Pr, m, Pv)* used were (left to right): (1,0.25,1.5,1); (0.25,1,1.5,1); (0.25,0.25,1.5,1). Expression level of gene sets denoted by colour with red to blue representing high to low expression, respectively (as per Fig 2.b). **c,** An analytic inference scheme was developed to infer model parameters. Heat map of the manhattan distance between obs:exp$_{\text{theory}}$ to simulated data is shown. Shading is determined by whether the obs:exp$_{\text{theory}}$ to simulation distance is smaller than the distance between bootstrapped simulated data and the original simulated data, at the displayed quantile levels. Yellow shading concentrated around true parameters illustrates that while errors in estimates are correlated, the true parameters are identifiable. **d,** Across a wide range of simulated datasets, true parameters can be recovered with small errors. Vertical black line denotes median percentage error.

Figure 4 | Stochastic dynamics of transcription coupled repair (TCR) in murine liver tumour genomes. a, Best fit between mathematical model (lines, model parameters in grey text) and data from murine liver genomes (points). Blue→red denotes increasing expression strata (as per Fig 2.b). **b,** Density of parameter estimates obtained from fitting the mathematical model to 1,000 bootstrap samples of mutation data. Red dashed lines indicate bootstrap confidence intervals, black vertical line denotes the estimate from original murine data. **c,** Heat map (left) showing optimal fits for all grid-search tested values of *Pd* and *Pr* ($8.4\text{x}10^8$ parameter combinations tested). Optimal fits (pink shapes; circle *Pr*≥0, triangle *Pr*=0) identified from gradient descent exploration initialised by high-quality grid-search fits. Landscape shading from the quantile distribution of fits between the observed data and bootstrap samples of it (right). **d,** Schematic summary of point estimates of interactions between RNAP and DNA lesions, for the full mathematical model including RNAP restart, and the reduced model without restart. Parameters values for the full model given as optimal in a, and for the reduced model as given in Extended Data Fig 2.c

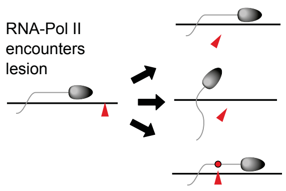

a
Mutagenic DNA lesion
RNA-Pol II
Incorrect base due to non-template strand lesion
Incorrect base due to template strand lesion
Non-template strand
Template strand
transcription
DNA replication
Inherited by daughter cell 1
Inherited by daughter cell 2
b
RNA-Pol II encounters lesion
Lesion repair & restart transcription
Lesion repair & RNA-Pol II disassociates
Mutant RNA transcript
Lesion undetected
c
Initial lesion positions after damage
TSS
TES
High lesion detection Low Pol II restart
Low lesion detection High Pol II restart
Transcription initiations post mutagenesis
More efficient repair near TSS
Uniform position of unrepaired lesions
d
DEN
♂ n=104
WGS + filtering
• 237 liver tumours
• ~7.2 million mutations genome-wide
phasing via lesion segregation
~1.7 million strand-phased mutations within gene bodies


a
Normalised template strand mutation rate (fold change)
2.0
1.5
1.0
0.5
0.0
Normalised non-template strand mutation rate (fold change)
2.0
1.5
1.0
0.5
0.0
≤ 10^{-2}
10^{-1}
10^{0}
10^{1}
≥ 10^{2}
Nascent transcription (nTPM)
b
Expression strata
Gene density
0.5
0.0
n=4649
n=1086
n=1016
n=959
n=944
n=7176
Mutations (obs/exp)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.29
3.73
≤ 10^{-2}
10^{-1}
10^{0}
10^{1}
≥ 10^{2}
Nascent transcription (nTPM)
c
Mutation rate (Mb-1) template strand lesions
15
10
5
0
0
50
100
150
200
250
Distance from TSS (kb)
d
Mutation rate (Mb-1) non-template strand lesions
15
10
5
0
0
50
100
150
200
250
Distance from TSS (kb)
e
0
1
2
3
4
Tumours (mutation load)
0
50
100
150
200
Distance from TSS (kb)
f
Mutations (obs/exp)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
Expected upstream lesions (count)


a
start
initiate
forward
terminate
stop
lesion
repair
e.m
1-Pd
Pd
Pr
1-Pr
Pv
Pr Restart probability
Pd Detection probability
Pv Lesion proportion visible
m Expression multiplier
e Measured expression in nTPM
RNA pol-II
Lesion
Invisible lesion
b
High detection probability
Low restart probability
Low detection probability
High restart probability
Low detection probability
Low restart probability
Mutations (obs/exp)
1
0
0 1 2 3 4
Expected upstream lesions
Expression
low
high
c
Restart probability (Pr)
0.8
0.6
0.4
0.2
0.0
0.3 0.5 0.7
Detection probability(Pd)
Expression multiplier (m)
10 8 6 4 2
0.0 0.2 0.4 0.6 0.8 1.0
Restart probability (Pr)
Model:data
L1-norm
50%
68%
≥90%
True
Estimate
d
Density
0.5
0.0
Median error =
10%
10^{-6} 10^{-4} 10^{-2} 10^{0} 10^{2}
Percent error in estimate (Pd)
Median error =
22%
Percent error in estimate (Pr)
0.6
Median error =
16%
10^{-2} 10^{0} 10^{2}
Percent error in estimate (m)



d



Event	Formula	Frequency under full model ○ (AIC=-997.57)	Frequency under reduced model △ (AIC=-997.76)
Lesion repair & restart transcription	*Pv*Pd*Pr*	22%	0%
Lesion repair & RNA-Pol II disassociates	*Pv*Pd**(1-*Pr*)	12%	15%
Lesion undetected	(1-*Pv*)+*Pv**(1-*Pd*)	66%	85%
Total		100%	100%