Error-prone bypass of DNA lesions during lagging strand replication is a common source of germline and cancer mutations

Vladimir B. Seplyarskiy[1,2,3], Maria A. Andrianova[3,4], Sergey I. Nikolaev[5,6], Georgii A. Bazykin[3,4] and Shamil R. Sunyaev[1,2*]

[1]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[3]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
[4]Skolkovo Institute of Science and Technology, Skolkovo, Russia
[5]Department. of Genetic Medicine and Development, University of Geneva, Switzerland.
[6]Service of Genetic Medicine, Geneva University Hospitals (HUG), Switzerland.
*Correspondence should be addressed to ssunyaev@rics.bwh.harvard.edu

Spontaneously occurring mutations are of great relevance in diverse fields including biochemistry, oncology, evolutionary biology, and human genetics. Studies in experimental systems have identified a multitude of mutational mechanisms including DNA replication infidelity as well as many forms of DNA damage followed by inefficient repair or replicative bypass[1–4]. However, the relative contributions of these mechanisms to human germline mutations remain completely unknown. Here, based on the mutational asymmetry with respect to the direction of replication and transcription, we suggest that error-prone damage bypass on the lagging strand plays a major role in human mutagenesis. Asymmetry with respect to transcription is believed to be mediated by the action of transcription-coupled DNA repair (TC-NER). TC-NER selectively repairs DNA lesions on the transcribed strand; as a result, lesions on the non-transcribed strand are preferentially converted into mutations. In human polymorphism we detect a striking similarity between transcriptional asymmetry and asymmetry with respect to replication fork direction. This parallels the observation that damage-induced mutations in human cancers accumulate asymmetrically with respect to the direction of replication, suggesting that DNA lesions are asymmetrically resolved during replication. Data from XR-seq experiments[5] and the analysis of cancers with defective NER corroborate the preferential error-prone bypass of DNA lesions on the lagging strand. The analysis of Damage-seq[6] data suggests that DNA damage on the lagging strand persists longer than damage on the leading strand in the population of dividing cells. We estimate that at least 9% of human germline mutations arise due to DNA damage rather than replication infidelity. Counterintuitively, the number of these damage-induced mutations is expected to scale with the number of replications and, consequently, paternal age.

Experiments in well-controlled genetic systems and *in vitro* experiments have uncovered that DNA polymerases make errors that are not repaired by the end of the cell cycle[7]. An alternative mechanism of mutagenesis due to misrepaired DNA damage or DNA damage bypassed by

translesion (TLS) polymerases has been extensively studied in experimental systems exposed to exogenous mutagens[8,9]. Although these studies do shed light on the mechanistic details of mutagenesis, well-controlled experimental systems provide little information on the relative contributions of these mechanisms in naturally occurring human mutations. More recently, computational genomics approaches have revealed statistical properties of mutations occurring in the germline[10–13], in tumors[14] and in embryo during early stages of development[15–17]. In cancer, many types of mutations have been successfully attributed to the action of specific mutagenic forces[18]. A number of studies have explored how cancer mutations scale with age at diagnosis[19,20] and how human germline mutations scale with paternal age[21–23]. It was hypothesized that the dependency of the number of accumulated mutations on the number of cell divisions may also reflect the replicative origin of mutations[24,25]. However, a quantitative model suggests that accumulation of both damage-induced and co-replicative mutations may scale with the number of cell divisions[26]. Therefore, we still do not know whether damage-induced mutations substantially contribute to heritable human mutations or whether natural mutagenesis in humans is mostly due to errors in replication.

To discriminate between co-replicative mutations and damage-induced mutations, we rely on statistical properties of mutations unequivocally associated with DNA damage. Both germline and cancer mutations leave footprints in the form of mutational asymmetry with respect to the direction of transcription (T-asymmetry). T-asymmetry reflects the prevalence of mutations that originate from lesions on the non-transcribed strand that could not be repaired by TC-NER[27,28]. Thus, the analysis of T-asymmetry may be used to quantify the prevalence of mutations arising from DNA lesions. Genomic data on cancers in which most mutations are caused by the action of specific, well-understood, DNA-damage-inducing agents provide an additional perspective on properties of damage-induced mutations. Notably, the level of T-asymmetry is exceptionally high in these cancers.

In contrast, the most obvious statistical feature associated with replication is asymmetry with respect to the direction of the replication fork (R-asymmetry). R-asymmetry may reflect differential fidelity of replication between the leading and lagging strands. Alternatively, R-asymmetry may be caused by the strand-specific bypass of DNA damage. DNA lesions not repaired prior to replication can either lead to fork regression followed by error-free repair or be bypassed by TLS polymerases[1,29]. TLS synthesis is error-prone and does not remove the lesion, which commonly introduces mutations on the newly synthesized strand. It has been asserted that the error-prone bypass process has different properties on leading and lagging strands[1,4] that would lead to R-asymmetry.

As a starting point, we compare R-asymmetry with T-asymmetry. To avoid the interference of statistical signals between the two types of asymmetries, R-asymmetry is estimated exclusively in intergenic regions and T-asymmetry only for genic regions (the estimates of T-asymmetry would remain the same if the analysis is confined to regions with no preferential direction of replication; Extended Figure 1). We calculate R-asymmetries for the 92 types of single-nucleotide mutations (excluding NpCpG>NpTpG mutations) in each trinucleotide context. C>T mutations in the NpCpG context are excluded because cytosine deamination in this context

usually results in conversion into the canonical nucleotide thymine[30] and does not relate to damage-based mutagenesis (Supplementary note 1). Figure 1 shows data for rare SNVs from the 1000 Genomes Project[31]. Extended Figure 2 shows that R-asymmetries across different contexts are concordant between rare SNPs and *de novo* mutations.

Strikingly, there is very high concordance between mutation types with substantial T-asymmetry and R-asymmetry (Figure 1). Mutation types that are predominant on the lagging strand are also more common on the non-transcribed strand (Figure 1a; R^2=0.80; p-value=$4.1*10^{-30}$). Moreover, this association holds even when six basic mutation classes are considered separately (Figure 1b).

As noted above, T-asymmetry arises from DNA damage on the non-transcribed strand that is invisible to TC-NER repair[12]. The unrepaired DNA lesions are occasionally converted into mutations. As a result, mutation types commonly induced by damage are biased towards the non-transcribed strand, and the level of T-asymmetry scales with the proportion of damage-induced mutations. Figure 1 suggests that R-asymmetry may be due to similarly differential resolution of DNA damage between leading and lagging strands. DNA lesions on the lagging strand would be more frequently converted into mutations, probably due to asymmetric damage bypass.

To follow up on this hypothesis, we analyze R-asymmetry in cancer genomes that have been influenced by specific mutagens. Four cancer types in the TCGA and ICGC datasets contain samples with high levels of T-asymmetry in specific mutation contexts: melanoma, predominated by UV-induced C>T mutations (signature 7)[14]; two lung cancers (LUAD and LUSC), predominated by smoking-induced G>A mutations (signature 4); and liver cancer, with a high prevalence of A>G mutations (signatures 12 and 16). All of these processes reflect the action of DNA-damaging mutagens rather than replication fidelity. We find that about 95% of these samples demonstrate a weak excess of mutations on the lagging strand during replication. These three mutagens that damage DNA primarily outside of replication also cause R-asymmetry, strongly suggesting that error-prone bypass on the lagging strand happens frequently (Figure 2a and Extended Figure 3). These findings also suggest that there is a common mechanism responsible for the R-asymmetry in human cancers in which specific mutagens are major sources of mutations.

Using the same cancer genome datasets, we test whether the direction of the replication fork is the primary contributor to regional, mutational, strand-specific biases in intergenic regions. We partition the genome into 500-kb genomic regions and sort these regions with respect to mutational imbalance (absolute value of the difference in mutation rates on the reference and non-reference strands calculated for the specific mutation type). Figure 2c shows that, in three of the cancer types, the overwhelming majority of damage-induced mutations in regions of high imbalance occur on the lagging strand (we exclude LUSC because of the insufficient number of samples). The same effect is also observed in rare human SNPs (Figure 2c).

If DNA lesions are more frequently bypassed by TLS on the lagging strand directly across the lesion, they will persist on this strand through replication. Therefore, mutational asymmetry caused by bypass in turn cause the asymmetry of unrepaired DNA damage. We utilize time series XR-seq data[5] to test whether the activity of the NER system is biased with respect to the replication fork direction. In agreement with the differential bypass hypothesis, repair is more frequently observed on the lagging strand (Figure 3a). Moreover, the difference between leading and lagging strands sharply increases with time after UV irradiation as more and more cells complete a round of replication.

To test whether the differential activity of the NER system reflects the preferential bypass of DNA damage, we analyze the Damage-seq dataset[6]. Damage-seq detects DNA damage (cyclobutane pyrimidine dimers), and we have data over a series of time points following the exposure of human fibroblasts to UV radiation. The data show a clear dependency on transcription and preferential retention of damage on the non-transcribed strand (Extended Figure 4). We observe a lagging strand bias of DNA damage that progressively increases with time, mirroring the trend in XR-seq data.

A different perspective on the effects of DNA damage on R-asymmetry can be gleaned from genomes of tumors lacking global genome NER (GG-NER). Patients with congenital loss of function of XPC, a key player in the GG-NER pathway, frequently develop skin cancers due to increased susceptibility to UV radiation. In these DNA-repair-deficient tumors, a larger fraction of UV-induced lesions are expected to remain unrepaired until replication. If these lesions are preferentially bypassed by TLS on the lagging strand, XPC deficiency is expected to enhance R-asymmetry. We analyze five squamous carcinoma genomes from patients with congenital XPC deficiency (Xeroderma Pigmentosa)[32]. These genomes have a distinct mutational spectrum that is dominated by Tp<u>Cp</u>T>TpTpT mutations (Extended Figure 5). These mutations indeed show elevated levels of R-asymmetry (Figure 3b).

Collectively, the above observations support the differential replication bypass hypothesis. However, a possible alternative explanation for the similarity between R-asymmetry and T-asymmetry in the human germline involves the exposure of DNA to a single-stranded conformation (ssDNA): The lagging strand stays in the single-stranded state during replication for a longer period, while the non-transcribed strand may occasionally adopt the single-stranded state because of R-loop formation between the transcribed strand and RNA[33,34]. We have tested the effect of R-loops on T-asymmetry and found that, in the germline, asymmetry does not increase in regions prone to R-loops compared with flanking regions within the same transcript (Extended Figure 6a). Additional clues to the role of ssDNA may be provided by APOBEC-induced mutations because APOBEC mutations have a strong affinity for ssDNA[35,36]. Again, we do not find that R-loops substantially affect the distribution of APOBEC-induced mutations in cancers (Extended Figure 6b). These analyses suggest that it is unlikely that the association between T-asymmetry and R-asymmetry is mediated by ssDNA.

Taken together, the observed mutation patterns in the germline and in cancer, and data from XR-seq and Damage-seq experiments point to differential damage bypass rather than

replication infidelity as a likely source of R-asymmetry in cancer genomes and in human germline mutations. Broadly, this suggests that DNA damage substantially contributes to spontaneous mutations. Although it is currently impossible to determine the precise proportion of damage-induced mutations, T-asymmetry allows us to quantify their contribution with a conservative estimate. Assuming that the DNA damage is uniform and that TC-NER is completely error-free and the only cause of the T-asymmetry, we compute the minimal fraction of damage-induced mutations in highly transcribed genes. Extrapolation of this estimate to the whole genome suggests that 9.4% of human germline mutations, 50% of mutations in melanoma, 40% of mutations in lung cancer, and 25% of mutations in liver cancer are due to DNA damage rather than replication infidelity. As expected, this estimate is much higher for cancers affected by known environmental mutagens. Still, the estimated fraction of damage-induced mutations in cancers obtained by our approach is much lower than previous estimates based on mutational spectra[14], attesting to the conservative nature of our analysis.

From the biochemical perspective, a higher conversion rate of damage due to mutations on the lagging strand is unsurprising, as replication of the leading strand is less tolerant to damage. Helicase is attached to the leading strand and is therefore more sensitive to damage on this strand[1,4]. Furthermore, damage on the leading strand blocks polymerase epsilon, which may cause fork uncoupling and stalling. This, in turn, may cause fork regression with lesion repair, template switch or homologous repair[1] – all these processes are error-free. Fork stalling may also lead to break-induced replication resulting in highly complex mutations not analyzed here. With the exception of break-induced replication, fork stalling is usually resolved by error-free mechanisms. Meanwhile, lesions on the lagging strand are unlikely to cause fork stalling and instead often only result in a short gap downstream from the lesion[1,4]. Consequently, damage on the lagging strand is rarely removed during replication and is instead simply bypassed by error-prone mechanisms (TLS) after replication.

Our analysis shows that mutations that are statistically associated with replication do not necessarily arise as a result of replication errors alone. A number of studies have demonstrated the dependency of the number of accumulated mutations on the number of cell divisions. This includes dependency on paternal age for germline mutations[21–23,37], the correlation of mutation burden in tumors with age at diagnosis[19], and properties of the molecular clock[38]. In line with theoretical models, we note that observations showing that mutation rate scales with the number of replications do not establish the mechanistic origins of mutations[26]. Instead of being responsible for generating mutations, DNA replication may simply convert pre-existing lesions, accumulated outside of S-phase, into mutations.

Acknowledgments

We thank Sergei Mirkin, Dmitry Gordenin, Cristopher Cassa and Donate Weghorn for useful comments on the manuscript, Lionel Sanz and Frédéric Chédin for help with R-loop data and Blake Boulerice for proofreading.

Authors Contributions

V.B.S., G.A.B. and S.R.S designed the study. V.S.B performed the data analyses. M.A.A. performed data preprocessing and the helped with results presentation. S.I.N. retrieved genomic data for squamous cell carcinoma. V.B.S. and S.R.S. drafted the manuscript. All authors contributed to the final version of the paper.

References

1. Yeeles, J. T. P., Poli, J., Marians, K. J. & Pasero, P. Rescuing stalled or damaged replication forks. *Cold Spring Harb. Perspect. Biol.* **5,** a012815 (2013).
2. Kunkel, T. A. & Erie, D. A. Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annu. Rev. Genet.* **49,** 291–313 (2015).
3. Hedglin, M., Pandey, B. & Benkovic, S. J. Characterization of human translesion DNA synthesis across a UV-induced DNA lesion. *eLife* **5,** (2016).
4. Hedglin, M. & Benkovic, S. J. Eukaryotic Translesion DNA Synthesis on the Leading and Lagging Strands: Unique Detours around the Same Obstacle. *Chem. Rev.* (2017). doi:10.1021/acs.chemrev.7b00046
5. Adar, S., Hu, J., Lieb, J. D. & Sancar, A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E2124-2133 (2016).
6. Hu, J., Adebali, O., Adar, S. & Sancar, A. Dynamic maps of UV damage formation and repair for the human genome. *Proc. Natl. Acad. Sci. U. S. A.* (2017). doi:10.1073/pnas.1706522114
7. Lujan, S. A. *et al.* Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res.* **24,** 1751–1764 (2014).
8. Boiteux, S. & Jinks-Robertson, S. DNA Repair Mechanisms and the Bypass of DNA Damage in Saccharomyces cerevisiae. *Genetics* **193,** 1025–1064 (2013).
9. Cohen, I. S. *et al.* DNA lesion identity drives choice of damage tolerance pathway in murine cell chromosomes. *Nucleic Acids Res.* **43,** 1637–1645 (2015).

10. Baker, A. *et al.* Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS Comput. Biol.* **8,** e1002443 (2012).

11. Chen, C.-L. *et al.* Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.* **28,** 2327–2337 (2011).

12. Polak, P. & Arndt, P. F. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res.* **18,** 1216–1223 (2008).

13. Seplyarskiy, V. B., Andrianova, M. A. & Bazykin, G. A. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. *Genome Res.* **27,** 175–184 (2017).

14. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

15. Harland, C. *et al.* Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. *bioRxiv* 79863 (2016). doi:10.1101/079863

16. Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T. & Hurles, M. Striking differences in patterns of germline mutation between mice and humans. *bioRxiv* 82297 (2016). doi:10.1101/082297

17. Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543,** 714–718 (2017).

18. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15,** 585–598 (2014).

19. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47,** 1402–1407 (2015).

20. Podolskiy, D. I., Lobanov, A. V., Kryukov, G. V. & Gladyshev, V. N. Analysis of cancer genomes reveals basic features of human aging and its role in cancer development. *Nat. Commun.* **7,** (2016).

21. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).

22. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).

23. Wong, W. S. W. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7,** 10486 (2016).

24. Moorjani, P., Gao, Z. & Przeworski, M. Human Germline Mutation and the Erratic Evolutionary Clock. *PLoS Biol.* **14,** e2000744 (2016).

25. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355,** 1330–1334 (2017).

26. Gao, Z., Wyman, M. J., Sella, G. & Przeworski, M. Interpreting the Dependence of Mutation Rates on Age and Time. *PLoS Biol.* **14,** e1002355 (2016).

27. Fousteri, M. & Mullenders, L. H. F. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res.* **18,** 73–84 (2008).

28. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15,** 465–481 (2014).

29. Roberts, S. A. & Gordenin, D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* **14,** 786–800 (2014).

30. Shen, J. C., Rideout, W. M. & Jones, P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* **22,** 972–976 (1994).

31. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

32. Zheng, C. L. *et al.* Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell Rep.* **9,** 1228–1234 (2014).

33. Sanz, L. A. *et al.* Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63,** 167–178 (2016).

34. Skourti-Stathaki, K. & Proudfoot, N. J. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev.* **28,** 1384–1396 (2014).

35. 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).

36. Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494,** 366–370 (2013).

37. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48,** 935–939 (2016).

38. Moorjani, P., Amorim, C. E. G., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 10607–10612 (2016).

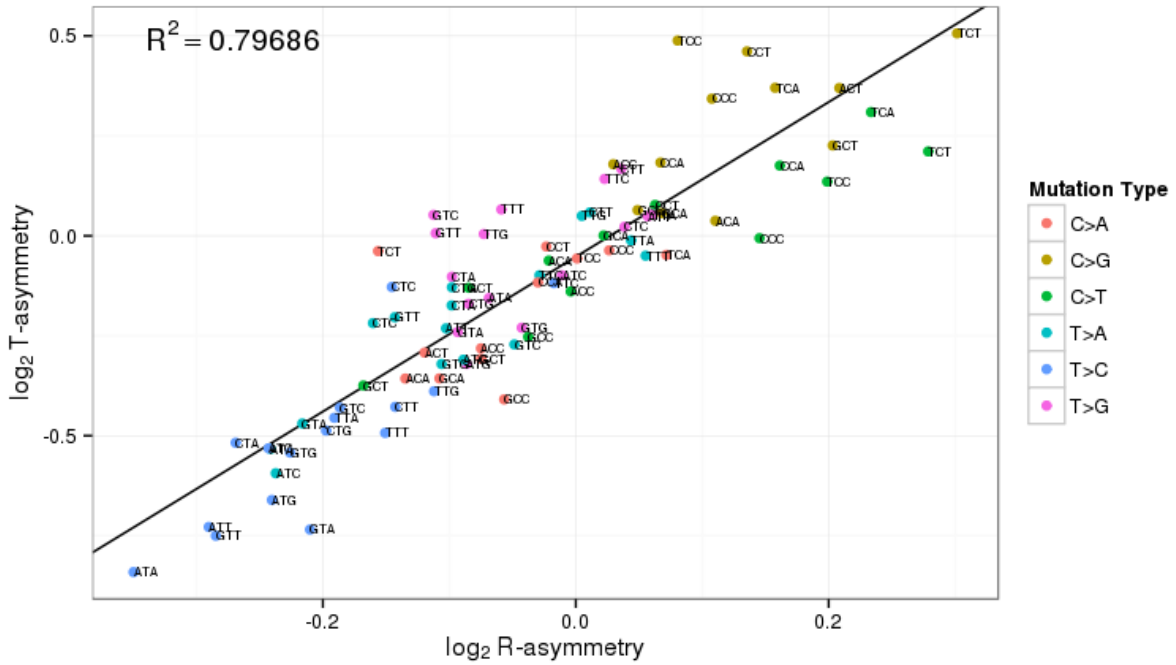
39. Seplyarskiy, V. B. *et al.* APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* **26,** 174–182 (2016).

40. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7,** 11383 (2016).

41. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164,** 538–549 (2016).

42. Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).

a



b

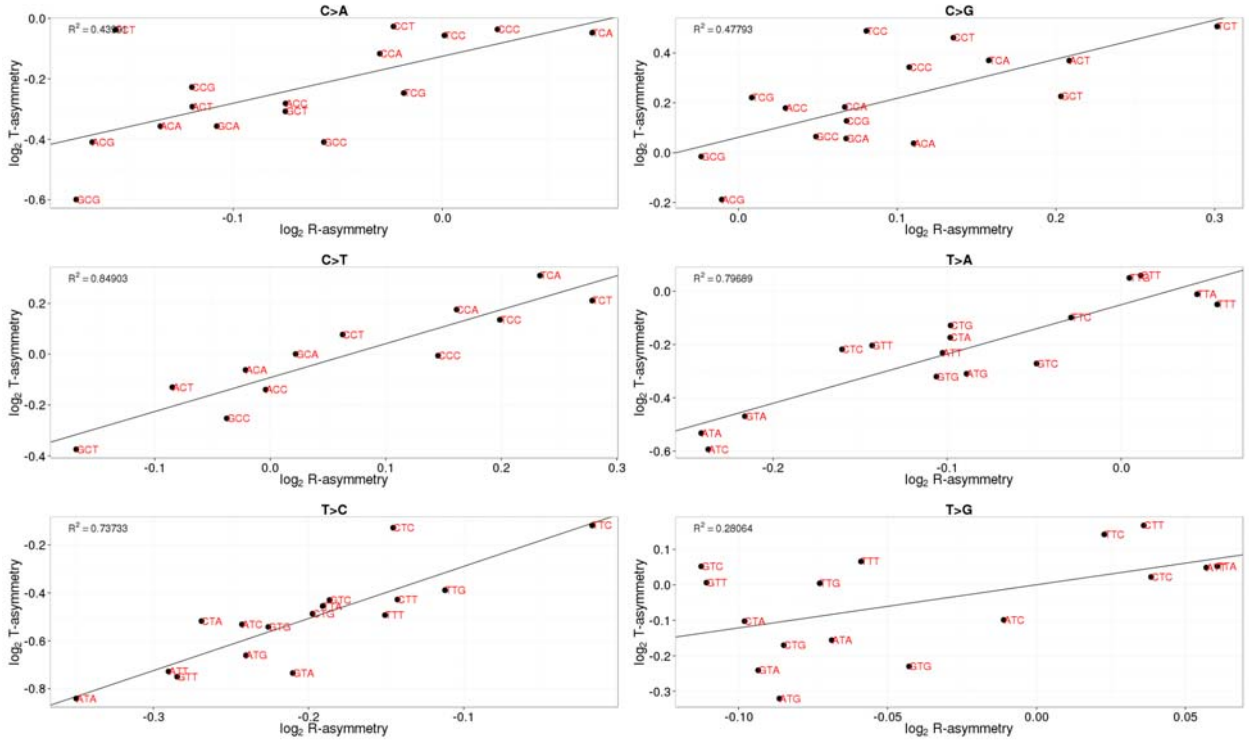
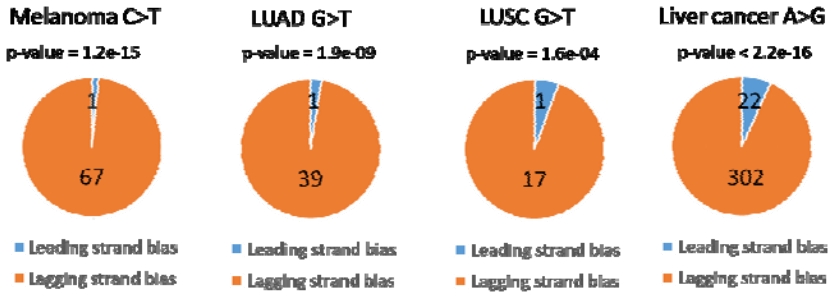
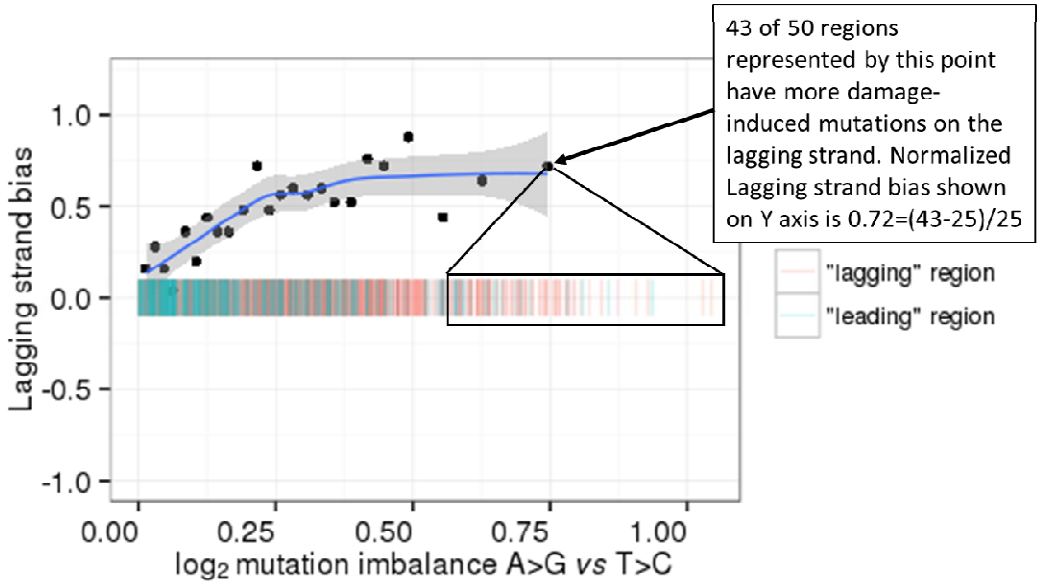


Figure 1. **R-asymmetry and T-asymmetry patterns in human polymorphism. a**, Relationship between R-asymmetry and T-asymmetry for 92 mutation types (NpCpG>T mutations excluded). **b**, Relationship between R-asymmetry and T-asymmetry shown separately for th six types of single-nucleotide mutations to highlight the effects of adjacent nucleotides.

a


Melanoma C>T
p-value = 1.2e-15
1
67
LUAD G>T
p-value = 1.9e-09
1
39
LUSC G>T
p-value = 1.6e-04
1
17
Liver cancer A>G
p-value < 2.2e-16
22
302
Leading strand bias
Lagging strand bias

b


43 of 50 regions represented by this point have more damage-induced mutations on the lagging strand. Normalized Lagging strand bias shown on Y axis is 0.72=(43-25)/25
"lagging" region
"leading" region
Lagging strand bias
1.0
0.5
0.0
-0.5
-1.0
0.00
0.25
0.50
0.75
1.00
log2 mutation imbalance A>G vs T>C

c

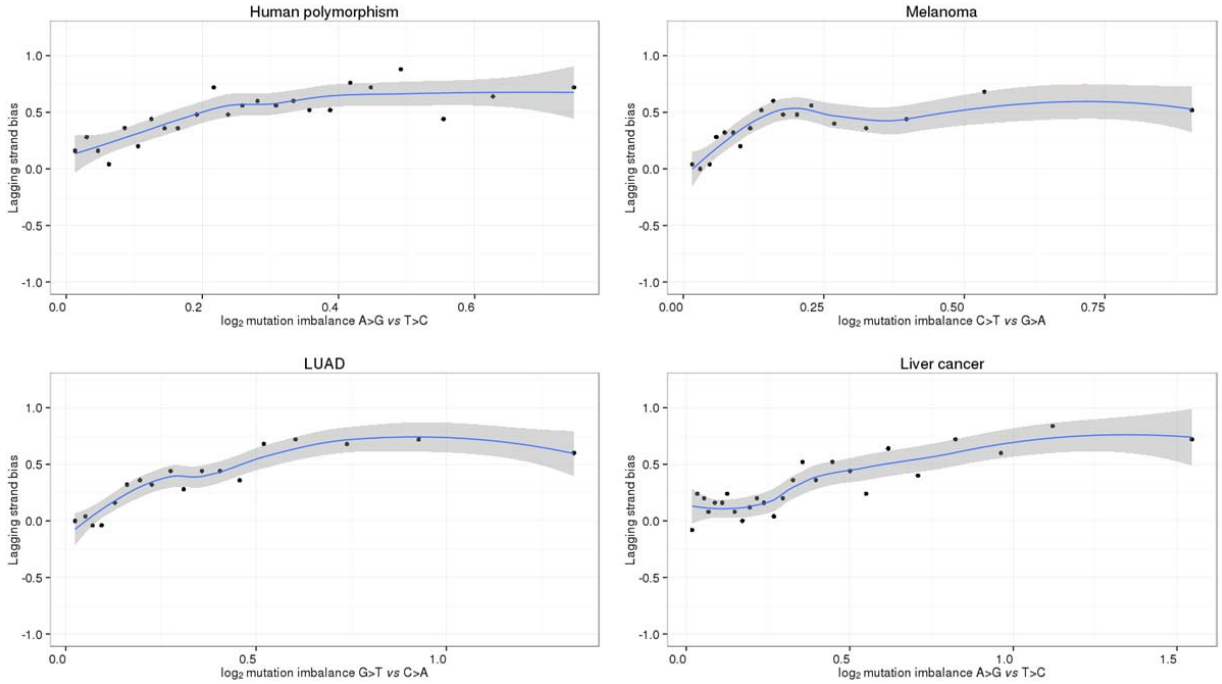

Human polymorphism
log2 mutation imbalance A>G vs T>C
Melanoma
log2 mutation imbalance C>T vs G>A
LUAD
log2 mutation imbalance G>T vs C>A
Liver cancer
log2 mutation imbalance A>G vs T>C
Lagging strand bias

Figure 2. **Damage-induced mutations preferentially reside on the lagging strand. a**, Number of tumor samples among melanomas, lung adeno carcinomas (LUAD), lung squamous carcinomas (LUSC), and liver cancers that have more damage-induced mutations on the leading or on the lagging strand (p-values shown for the goodness-of-fit chi-square test). **b.** Estimation of regional lagging strand bias. All genomic regions of 500-kb (short vertical lines) are sorted by the imbalance between complementary mutations. Each region is categorized as either "lagging" (red) or "leading" (turquoise) according to the strand on which damage-induced mutation is more prevalent (human polymorphism, A>G; melanoma, C>T; LUAD, G>T; liver cancer, A>G). Lagging strand bias is estimated from the fraction of "lagging" regions among the 50 regions with similar levels of mutation imbalance. For example, if all 50 regions are "lagging", the lagging strand bias=1; if 25 regions are "lagging" and 25 are "leading", the lagging strand bias=0; and if all 50 regions are "leading", the lagging strand bias=-1. **c**, Lagging strand bias for human polymorphism and cancers.

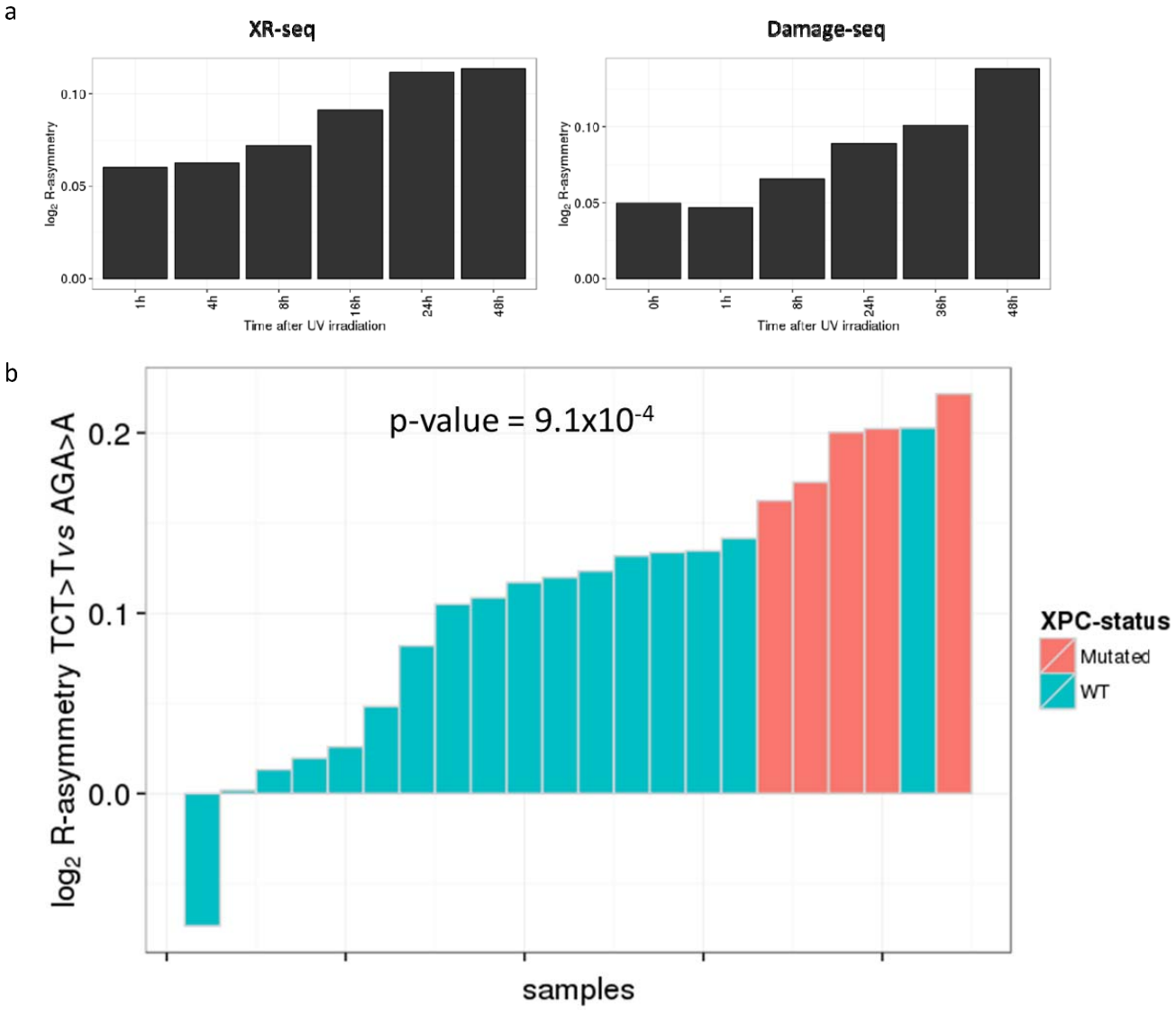


Figure 3. **R-asymmetry in UV-irradiated cells and in tumors in patients with the congenital GG-NER defect. a**, R-asymmetry of repaired CPD damage (left) and CPD damage remaining in DNA (right) as a function of time since UV irradiation. **b**, R-asymmetry for skin cancers. Bars corresponding to Xeroderma Pigmentosum patients due to loss of function mutation in *XPC* gene are colored red; bars corresponding to patients with intact *XPC* are colored turquoise. P-value for two-sided Mann–Whitney U test is shown.

Materials and methods

Human polymorphism and cancer mutation data

To analyze mutational patterns reflected in human DNA polymorphism, we extracted non-singleton SNPs with derived allele frequency <1% from 1000 Genomes Project[31] data. Cancer somatic mutations were assembled from data provided by the ICGC and TCGA projects and a dataset reported by Alexandrov *et al.*[14]. The resulting combined dataset included whole genomes of 87 melanomas, 64 lung adenocarcinomas, 20 lung squamous cell carcinomas, and 381 liver cancers. Cancer somatic mutations identified in XPC_{wt} and $XPC_{-/-}$ skin SCC samples were downloaded from dbGap (phs000830).

Experimental data on DNA damage and repair

The XR-seq dataset for cyclobutane pyrimidine dimers (CPD) reported in Adar *et al.*[5] allowed us to estimate the amount of DNA damage actively repaired by NER following UV irradiation. To directly assess the presence of unrepaired DNA damage, we used the Damage-seq data for CPDs provided by Hu *et al.*[6]. We did not use XR-seq and Damage-seq data for pyrimidinepyrimidone (6-4) photoproducts because these lesions are repaired too quickly to permit an accurate analysis of the effect of damage bypass over successive rounds of replication.

R-asymmetry

As described previously[39], the "derivative" (normalized rate of change) of replication timing may serve as a predictor of the preferential replication fork direction, which depends monotonically on it. This approach was proposed by Chen *et al.*[11] and has been used in recent cancer genomics studies[40,41].

We focused on genomic regions showing a strong preference for a specific fork direction as evident from the replication timing "derivative". For the analysis of SNPs, XR-seq, and Damage-seq (Figures 1a and b, Figures 3a and b), we used a conservative threshold corresponding to the 10% of genomic regions with the highest absolute values of the replication timing "derivative". However, this threshold appeared too restrictive for the cancer genome analyses because many individual tumors have insufficient numbers of mutations within the 10% of the genome, so we relaxed the threshold to 40% for these analyses. Both of these thresholds have been used in previous studies[13,41], and the results have generally been robust with respect to the threshold chosen.

For each individual analysis, we selected the most relevant available replication timing dataset: IMR-90 for lung cancers, HepG2 for liver cancer, and NHEK for melanoma and squamous carcinoma. For germline mutations, there is no relevant cell type with known replication timing. R-asymmetry of the SNPs from the 1000 Genomes appears stable irrespective of the tissue for which replication timing was obtained[13]. As in the previous study[11], we used replication timing data for the MCF-7 cell line. We also used replication timing data obtained from NHEK to predict the preferential fork direction in the analysis of XR-seq and Damage-seq data (matching the tissue but not the exact cell type).

For each mutation type, we calculated R-asymmetry as the ratio of mutation density on the lagging strand to the mutation density on the leading strand. Samples with fewer than 100 mutations on each strand were excluded from the analysis to reduce sampling noise.

In the analysis of XPC knockouts shown in Figure 2b, we focused on the magnitude of the effect in each tumor rather than on the presence of the effect. We therefore excluded samples with fewer than 500 mutations on each strand. The relaxation of the threshold to 100 mutations does not change the conclusions (data not shown).

In order to exclude the impact of T-asymmetry on the R-asymmetry estimation, we restricted the analysis of R-asymmetry to intergenic regions.

T-asymmetry

For each mutation type, we calculated T-asymmetry as a ratio of mutation density on the transcribed strand to mutation density on the non-transcribed strand. Gene annotations and transcription direction were determined according to the knownGene track of the UCSC genome browser. Tumors with T-asymmetry >1.2 (for any of the six major mutation classes) were considered to have high level of T-asymmetry. Even with this lenient criterion, only four cancer types (melanoma, LUAD, LUSC, and liver cancer) had more than 20 tumor samples in this category. To order the genes by their expression levels, we selected the most relevant tissues from Gtex[42]: testis for SNPs from 1000 Genomes, sun-exposed skin for melanoma, liver for liver cancer, and lung for lung cancers.

Regions with high mutational imbalance

Figure 2c shows the imbalance of rare human SNPs and cancer mutations between DNA strands and its relationship with the preferred fork direction. For this analysis, cancer data were aggregated by tumor type. The data on individual tumors are too sparse for the investigation of genomic regions.

The genome was subdivided into 500kb-long, non-overlapping regions. We considered only those regions that were informative with respect to the preferential fork direction (at the 40% threshold used). For all four datasets depicted in Figure 2c, we excluded regions that had fewer than 100 mutations on each DNA strand (only mutations in intergenic regions were considered). Mutational imbalance (MI) for each region was calculated as ratio of mutation densities for a mutation type and its reverse complement, e.g., for melanoma:

$$MI = \mu_{C>T_ref} / \mu_{G>A_ref}$$

where $\mu_{C>T_ref}$ and $\mu_{G>A_ref}$ are the mutation rates for C>T and G>A mutations with respect to the reference strand. Mutation is considered C>T if the unmutated reference strand has a C in this position and G>A if the unmutated reference strand has a G in this position. To have only positive values of $\log_2(MI)$ we reverse MI if MI was lower than 1, thus MI would be the same if mutations were considered with respect to the non-reference strand.

We assigned a label of “lagging” or “leading” to each 500-kb region on the basis of the relative prevalence of putatively damage-induced mutations on either the lagging or the leading strand.

For the visualization in Figure 2c, we arranged the 500-kb regions into groups of 50 regions according to their MI values. Next, we calculated the number of “leading” (N_leading) and “lagging” (N_lagging) windows for each group. By chance, we would expect an equal number of “lagging” and “leading” regions in the same group (N_leading = N_lagging = 25). We then calculated the observed excess of “lagging” windows in each group as (N_lagging-25)/25. The values of this excess are depicted on the Y-axis of Figure 2c. Negative values correspond to an excess of “leading” windows. Mutational imbalance for each group was calculated as the median MI value of windows in this group.

Exclusion of replica B2 at 48h from Damage-seq

T-asymmetry and the difference between genic and non-genic regions are the main results of the Damage-seq experiments[6] that support the utility of the data for the genome-wide analysis of bulky DNA damage and repair by the NER system. Thus, for quality control of the Damage-seq data, we calculated T-asymmetry and the ratio of reads in intergenic and genic regions separately for all replicas. T-asymmetry and the ratio of reads in intergenic and genic regions were normalized using the corresponding values for naked DNA. We found that the replicates were generally concordant at each time point with the exception of the 48h point, where we found substantial T-asymmetry and prevalence of mutations in intergenic regions in replica A but essentially no signal in replica B2 (Supplementary Figure 2). At other time points, we observed a clear, time-dependent increase in T-asymmetry and decrease in the fraction of damages in genic regions, as expected. Based on these observations, we argue that the absence of the signal in replica B2 at 48h is an artifact. Therefore, this data point was excluded. As shown in Supplementary Figure 2c, this replica is also a clear outlier in the analysis of R-asymmetry.

Estimate of the proportion of mutations arising due to DNA damage in human cancers and the germline

To conservatively estimate the proportion of damage-induced mutations, we capitalized on the statistical signal of T-asymmetry that is associated with DNA damage. The T-asymmetry introduced by co-transcriptional processes cannot be a consequence of replication infidelity. Therefore, mutations responsible for the T-asymmetry must be damage-induced. Since transcribed and non-transcribed regions can have different susceptibilities to DNA damage, we conservatively compared the levels of mutations between transcribed strand and immediately adjacent flanking sequences rather than between transcribed and non-transcribed strands:

$$t = \frac{\mu_{intergenic}}{\mu_{transcribed_strand}},$$

where $\mu_{transcribed_strand}$ is the mutation density on the transcribed strand and $\mu_{intergenic}$ is the mutation density in flanking intergenic regions.

To estimate *t*, we used the 10% of genes with the highest expression levels. We conservatively assumed that all damage on transcribed strands is efficiently repaired. Thus, the fraction of damage-induced mutations in transcribed regions and in intergenic regions is expressed as:

$$f_{genic} = \frac{t-1}{t+1}$$

$$f_{intergenic} = \frac{t-1}{t}$$

If *a* denotes the fraction of mutations in genic regions, and *b* is the fraction of mutations in intergenic regions, the fraction of damage-induced mutations for the whole genome (f_{genome}) is expressed as:

$$f_{genome} = a f_{genic} + b f_{intergenic}$$

The conservative nature of this estimate is evident in the cancer data. Although nearly all mutations in melanoma are caused by UV irradiation, our estimate attributes only 50% of mutations to DNA damage. Our procedure is shown schematically in Supplementary Figure 3.

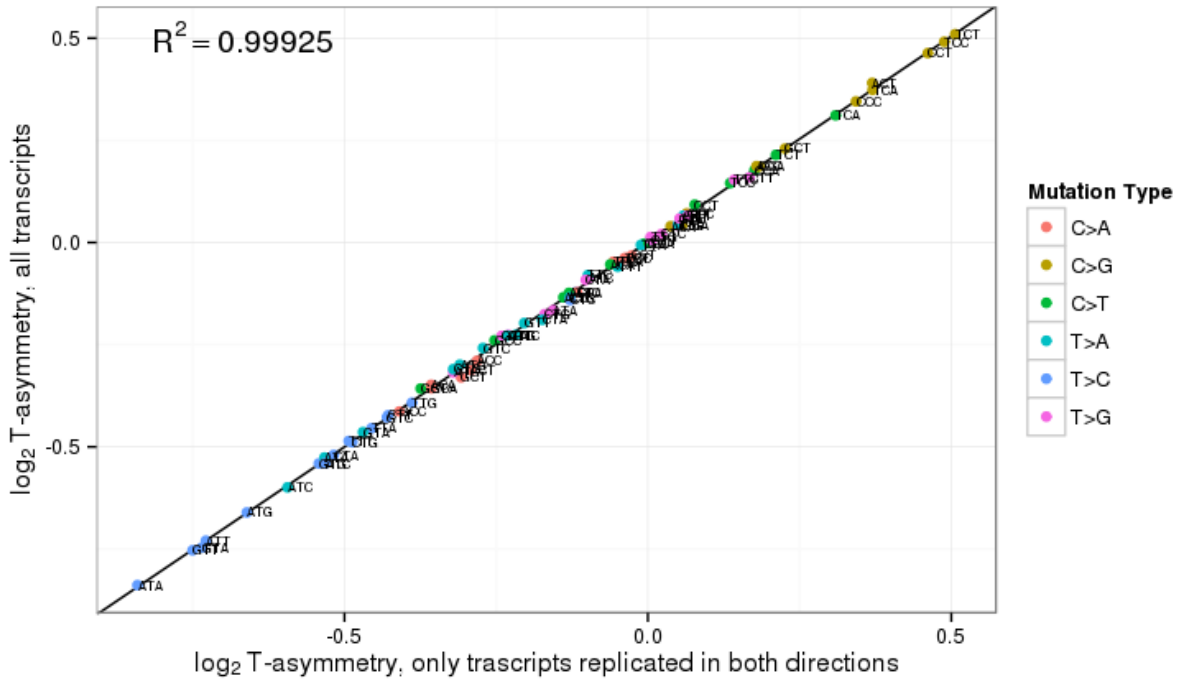
R-loops

We used data on strand-specific R-loops from Sanz *et al.* [33]. Most R-loops were on the template strand, and we considered only such R-loops. For control regions, we used intronic regions within the same gene that were 500 nucleotides apart from the R-loop peak and 500 nucleotides long.

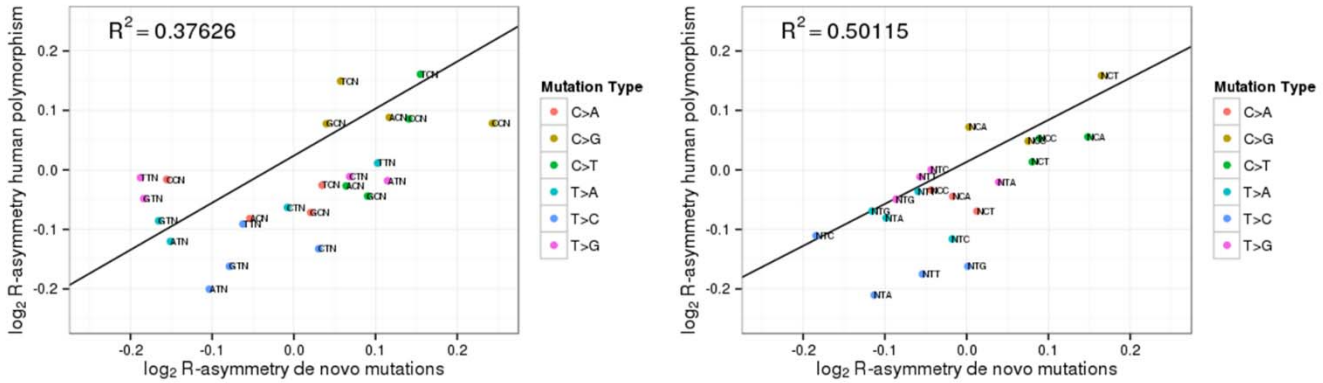
CpG islands

Annotation of CpG islands was downloaded from the UCSC genome browser (cpgIslandExt).

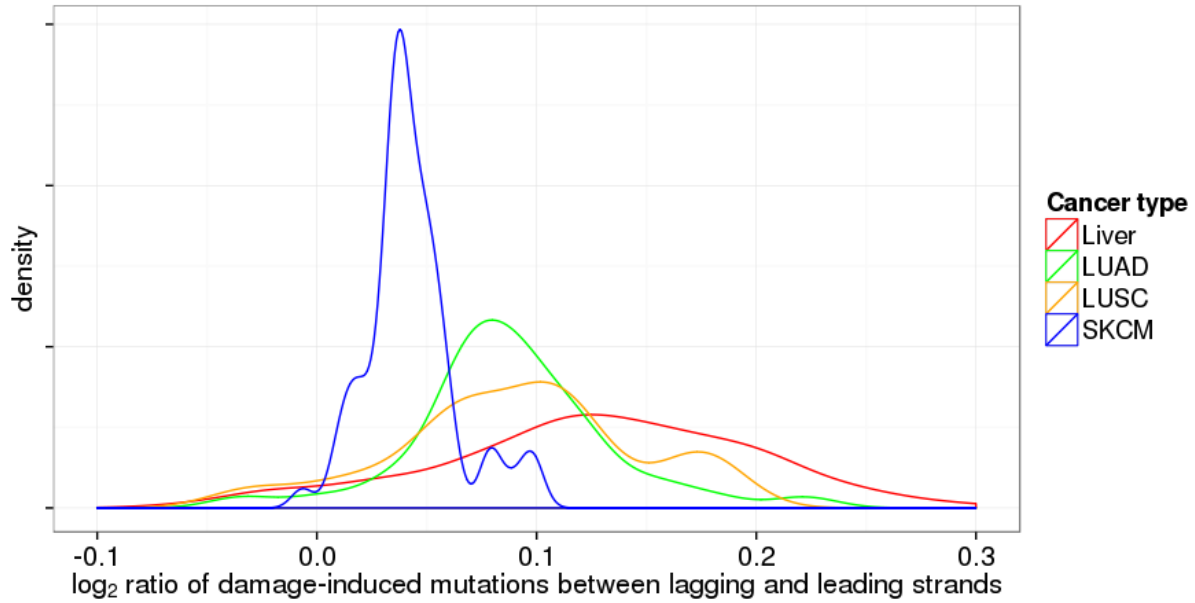
Extended Figures



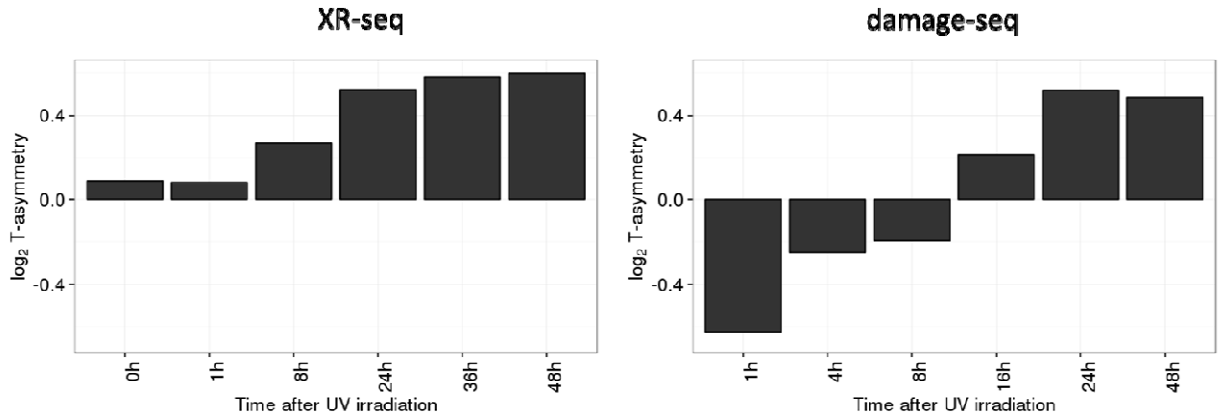
Extended Figure 1. **T-asymmetry for human polymorphism calculated for all genes vs. T-asymmetry for genes with no preferential fork direction.**



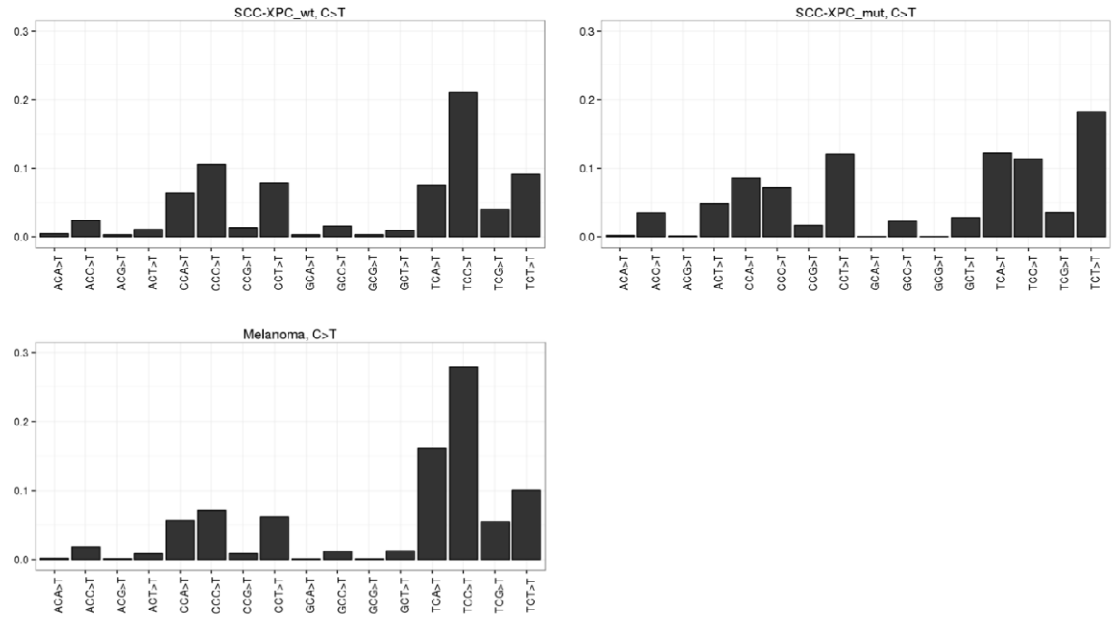
Extended Figure 2. **Concordance of R-asymmetries across different contexts between SNPs and *de novo* mutations.** The sparsity of currently available data on *de novo* mutations forces us to rely on the human polymorphism data as a proxy. R-asymmetry was calculated for six mutation types and one adjacent nucleotide (5' in the left panel and 3' in the right panel). We used one adjacent nucleotide rather than two because of insufficient data. CpG>TpG mutations were excluded; in the left panel, CpG>GpG and CpG>ApG mutations were also excluded because of the low numbers of *de novo* mutations of these types. *De novo* mutations were obtained from Francioli *et al.*[22] and Wong *et al.*[23]. The P-values for the correlations are $1.4*10^{-3}$ for the left panel and $3.3*10^{-4}$ for the right panel.



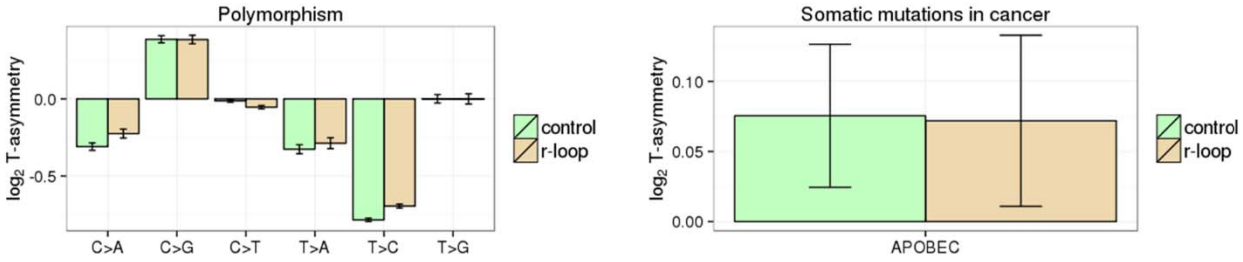
Extended Figure 3. **Distribution of R-asymmetry values for tumor samples with a high prevalence of damage-induced mutations.** Liver **–** liver cancer, LUAD – lung adeno carcinoma, LUSC – lung squamous carcinoma, SKCM **–** melanoma.



Extended Figure 4. **T-asymmetry of UV-induced damage.** T-asymmetry of repaired CPD damage (left) and CPD damage remaining in DNA (right) as a function of time since UV irradiation.



Extended Figure 5. **XPC deficiency associated with alterations in mutational spectra with prevalence of TCT>T mutations**. Fraction of C>T mutations in all trinucleotide contexts among all mutations. SCC – squamous carcinoma.



Extended Figure 6. **T-asymmetry is not elevated in regions prone to R-loops compared with flanking regions within the same transcript.**