Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair

Paz Polak[1–3,8], Michael S Lawrence[3,8], Eric Haugen[4], Nina Stoletzki[1–3], Petar Stojanov[3], Robert E Thurman[4], Levi A Garraway[2,3,5,6], Sergei Mirkin[7], Gad Getz[3], John A Stamatoyannopoulos[4] & Shamil R Sunyaev[1–3]

Carcinogenesis and neoplastic progression are mediated by the accumulation of somatic mutations. Here we report that the local density of somatic mutations in cancer genomes is highly reduced specifically in accessible regulatory DNA defined by DNase I hypersensitive sites. This reduction is independent of any known factors influencing somatic mutation density and is observed in diverse cancer types, suggesting a general mechanism. By analyzing individual cancer genomes[1], we show that the reduced local mutation density within regulatory DNA is linked to intact global genome repair machinery, with nearly complete abrogation of the hypomutation phenomenon in individual cancers that possess mutations in components of the nucleotide excision repair system. Together, our results connect chromatin structure, gene regulation and cancer-associated somatic mutation.

Somatic mutations are a major contributor to cancer development and progression. In cancer cells, the density of somatic mutations is highly heterogeneous throughout the genome[2,3]. However, mechanisms governing the genomic distribution of somatic mutations are poorly understood. Recently, cancer genomics efforts have accumulated data on somatic mutations in tumors[4], revealing that the relative density of somatic mutations in protein coding genes (including both introns and exons) is lower than the genome average[5]. This effect has been posited to result from transcription-coupled DNA repair (TCR)[2,3], which is mediated by the recruitment of the nucleotide excision repair (NER) system through RNA polymerase II (RNA Pol II) stalled at premutation lesions[6,7]. The existence of such an effect raises the question of whether other similarly specialized repair mechanisms operate on other functionally important genomic regions.

Regulatory DNA (e.g., promoters, enhancers, insulators) active within a given cell type is characterized by hypersensitivity to DNase I[8], resulting in DNase I hypersensitive sites (DHSs) that quantitatively reflect regulatory factor binding in place of canonical nucleosomes[9,10]. It has long been posited that the accessibility of DNA within regulatory regions may render such regions more susceptible to DNA damage–induced mutation[11]. Evolutionary rates of sequence divergence within DHSs found in cancer genomes and primitive cells are higher than within normal differentiated cells[8], and the density of somatic variants detected in a cultured cancer cell sample was shown to be reduced in DHSs more than the density of common single-nucleotide polymorphisms[12]. However, particularly with regard to the variability in somatic mutation rates in cancer genomes, a quantitative understanding of mutation within regulatory DNA has not been achieved, and the underlying biological mechanisms have not been explored.

RESULTS

Reduced local density of somatic mutations in DHSs

To examine mutation frequencies in regulatory DNA, we mapped DHSs genome-wide in 12 cancer cell lines, as well as normal cellular counterparts of major malignancies (Online Methods). We then analyzed whole-genome sequencing data from 34 tumor-normal pairs from seven distinct data sets: small-cell lung cancer[3], melanoma[2], 23 multiple myeloma[5] samples and 9 colon cancers[13]. We used published mutation data for small-cell lung cancer[3] and melanoma cell lines[2] (http://icgc.org/) and reanalyzed primary tumor data on multiple myeloma and colon cancer using MuTect[14] (http://www.broadinstitute.org/cancer/cga/mutect). These 34 cancer genomes contained 364,226 somatic point mutations in about 2.6 Gbp of sequence that could be uniquely mapped in the DHSs assay, i.e., density of 0.000139 per base pair (bp).

We observed a substantial reduction in the frequency of somatic nucleotide substitutions in DHSs compared to the genome average (**Fig. 1** and **Supplementary Fig. 1**). This reduction is highly significant and consistent across all tumors ($P < 10^{-36}$, χ^2 test). The reduction was most prominent in the core transcription factor binding regions of DHSs marked by the maxima of DNase I cleavage intensity (**Fig. 1**).

We next confirmed that the reduction of frequency of somatic mutations in DHSs was neither the result of confounding factors influencing local variation in cancer mutation density, nor the result of sequencing and mapping biases[15]. Confounding factors may include differences between intergenic regions and genes (including both exons and introns), distance from transcription start

[1]Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Boston, Massachusetts, USA. [2]Harvard Medical School, Boston, Massachusetts, USA. [3]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]Departments of Genome Sciences and Medicine (Oncology), University of Washington, Seattle, Washington, USA. [5]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [6]Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [7]Department of Biology, Tufts University, Medford, Massachusetts, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to S.R.S. (ssunyaev@rics.bwh.harvard.edu) or J.A.S. (jstam@u.washington.edu).

Received 12 December 2012; accepted 22 November 2013; published online 15 December 2013; doi:10.1038/nbt.2778

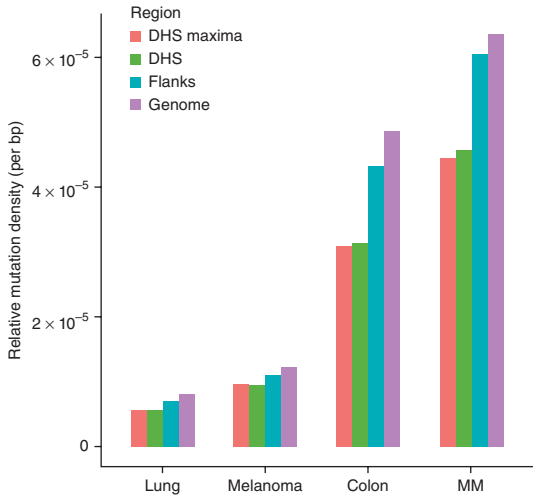


Figure 1 Relative density of somatic mutations is reduced in DHSs of all analyzed cancer genomes (lung[3], melanoma[2], colon[13], multiple myeloma (MM)[5]). Mutation density per (uniquely mappable) bp is shown for (red) DHS maxima defined as ±75 bp around the peak of DNase I hypersensitivity (marked as DHS peaks), (green) DHSs, (cyan) 1,000-bp flanking regions and (purple) overall genome. Mutation density in DHSs is substantially lower in comparison with immediate flanking regions and genome average. The effect is stronger for DHS maxima compared to overall DHSs.

sites[2] (**Supplementary Fig. 2**), time of DNA replication during the S-phase[16], distances to telomeres and centromeres, and local G+C content[15]. Relative density of somatic mutations also depends on sequence context, especially flanking nucleotides; different tumors exhibit different context dependencies[2,3,13] (**Supplementary Fig. 3**). The relative density of mutations expected from the sequence context is higher in DHSs, magnifying our observation ($P < 5 \times 10^{-181}$).

We separately analyzed the mutation density in DHSs located in intronic regions, intergenic regions and in regions proximal (<1 kb) and distal with respect to transcription start sites. In all cases, mutation density within DHSs was lower than in corresponding regions outside DHSs ($P < 10^{-4}$) (**Supplementary Fig. 4**). More notably, the reduction was evident in comparison to immediately flanking 1-kb regions ($P < 2.2 \times 10^{-16}$, χ^2 test in lung, multiple myeloma and colon; $P = 7.21 \times 10^{-13}$ in melanoma). Consequently, the observed reduction in the density of somatic mutations cannot be explained by a regional factor acting over long ranges[17], such as transcription or DNA replication timing.

To rule out biases related to sequencing and mapping, for two of the cancer types (colon[13] and multiple myeloma[5]) with available raw sequencing data, we repeated the analysis restricting it to nucleotide positions with >80% detection power based on sequencing coverage. This analysis confirmed that the density of somatic sequence alterations is significantly reduced ($P < 2.2 \times 10^{-16}$) in DHSs compared to 1-kb flanking sequences.

To account collectively for all of the above potential confounding factors, we applied a Poisson regression model[18]. DHSs remained a significant ($P < 10^{-24}$) contributor to the local somatic mutation frequency on top of other factors, including DNA replication timing[19], distance from transcription start sites, distance from the DHS itself, CpG islands status, G+C content and region type (exonic, intergenic, intronic) (**Supplementary Tables 1** and **2**). Because our regression analysis included neighboring windows, short-range regional dependencies could potentially inflate statistical significance. We repeated the analysis using a small subset (20%) of spatially separated windows and confirmed that DHSs remain a highly significant ($P < 10^{-5}$) contributor, even if only 20% of data are used (**Supplementary Tables 1** and **2**).

Notably, the effect of chromatin accessibility is monotonic and continuous and thus does not depend on the specific thresholds used to define DHSs (**Supplementary Fig. 5**). Finally, DHSs mapped in potential cells or tissues of origin (e.g., lung tissue for lung carcinoma) substantially contribute to the regression model that already includes pooled DHSs from multiple cell types (**Supplementary Tables 4–7**). This demonstrates that cell-selective chromatin architecture and not simply genomic location is the driving feature.

The observed reduction in the frequency of somatic sequence changes within DHSs might be explained by either reduced occurrence of somatic mutation or by the action of purifying selection (selective removal of alleles with deleterious effect). At present, purifying selection in cancers has not been carefully studied, so we lack information that would support or contradict the action of purifying selection. In general, population genetics and comparative genomics studies in a variety of organisms suggest that purifying selection is usually stronger in coding regions than in regulatory regions[20,21]. To investigate the possible action of purifying selection, we compared relative mutation densities in regulatory and protein-coding regions. The average reduction (across cancers) in frequency of somatic mutations in coding sequences relative to flanking sequences is smaller than analogous reduction in DHSs compared to their flanks (**Supplementary Fig. 6**). Furthermore, the observed reduction of mutation frequency in exons may not necessarily represent the action of purifying selection. The frequency of missense mutations is not lower than the frequency of synonymous changes (**Supplementary Fig. 6**). Thus, although it is possible that cancers may differ from evolving populations and we cannot rule out the action of purifying selection in regulatory regions, mutation attenuation seems to play a more important role.

Association with nucleotide excision repair

Relative mutation density depends on replication fidelity, levels of DNA damage or efficiency of DNA repair. The fact that the observed relative reduction of mutation density was almost entirely limited to DHSs makes it difficult to explain the reduction by an increase in global replication fidelity. It is also unlikely that more accessible DNA at DHSs would be less prone to damage than less accessible DNA elsewhere. In fact, mutation frequencies observed in model organisms are reduced by positioned nucleosomes[22,23], whereas the effects of nucleosome positioning on somatic mutations in cancer are relatively small and differ in whether they are positive or negative between various cancer types[15].

Chromatin accessibility plays a major role in targeting nuclear proteins to regulatory DNA, and may provide a mechanism for preferential access by the repair machinery. Preferential activity of DNA repair proteins in accessible regulatory DNA may thus offer an explanation for the observed effect, analogous to the action of TCR in protein coding genes. We hypothesized specifically that potentiation of nucleotide excision repair (NER) and base excision repair (BER) by chromatin accessibility could be responsible for the observed relative reduction of mutation density at DHSs. The level of oxidative stress and the subsequent accumulation of lesions targeted by BER[24] is higher in malignant than in normal cells[25]. BER is an evolutionarily conserved DNA repair pathway, which starts from the recognition and excision of various base lesions by specific DNA glycosylases, followed by the processing of the resulting apurinic/apyrimidinic sites, DNA repair synthesis and ligation. Direct access of glycosylases to DNA lesions is pivotal for this repair process[26]. Not surprisingly, therefore, BER complexes preferentially assemble in nonnucleosomal regions in response to oxidative stress[27], which naturally targets them to DHSs.

The NER pathway consists of two converging branches: global genome repair and TCR. DNA damage is first recognized by

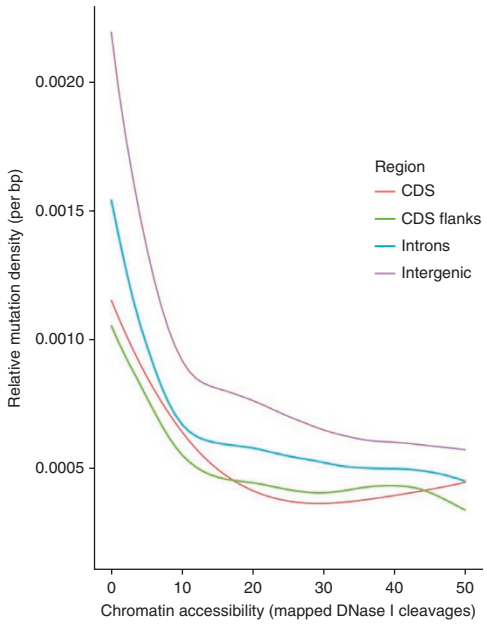
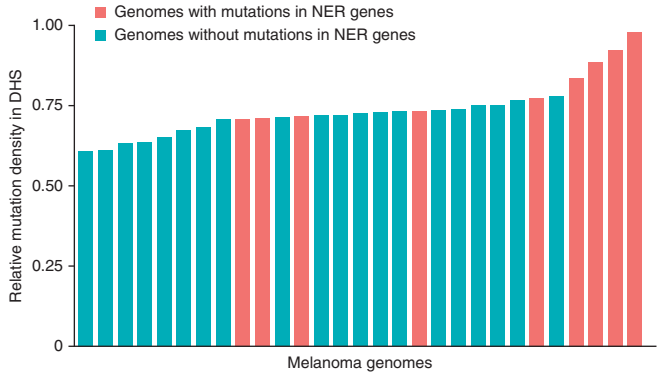


Figure 2 Density of somatic C:G→T:A transition mutations in melanoma samples strongly depends on chromatin accessibility in a monotonic and continuous fashion. Density of C:G→T:A transitions per C:G base-pair in 400-bp genomic intervals is shown as a function of chromatin accessibility in melanocytes measured by the density DNase I cleavages. The dependence is presented separately for introns and intergenic regions, and is equally present in both. Mutation densities are parametrically fitted to a spline function using a generalized additive model Poisson regression model[18].

the Xeroderma pigmentosum C (XPC) complex; the DNA duplex is opened by XPD and XPB helicases, followed by the excision of the damaged strand by XPF-ERCC1 and XPG nucleases, then gap filling by the replication polymerases using the intact strand as a template and finally DNA ligation[28]. A priori, NER machinery should be able to correct damaged DNA regardless of its chromatin state. A fully assembled NER complex, however, has a footprint of ~100 bp in DNA, which is significantly longer than the length of an internucleosomal linker. As a result, chromatin structure inhibits functional NER complex assembly and function[29–31]. In the case of TCR, this problem is circumvented by the fact that DNA damage is first sensed by the RNA polymerase followed by NER recruitment to an already unraveled chromatin required by the CSB and CSA proteins (reviewed in ref. 6). This is not the case for global genome repair, and the problem is additionally exacerbated by the fact that the damage sensor, the XPC complex, cannot bind to DNA adducts embedded in nucleosomes[32]. Thus, both lesion recognition and repair-complex assembly may be potentiated in accessible chromatin[29,33]. The fact that roughly half of DHSs lie in intergenic regions suggests global genome repair as a more likely candidate than TCR. Notably, nucleosomal chromatin appears to inhibit global genome repair function[29–31]. Moreover, the XPC complex involved in damage recognition is inhibited from binding to lesions in nucleosomal DNA[32]. Thus, both lesion recognition and repair complex assembly may be potentiated in accessible chromatin[29,33].

Failure of NER predisposes the body to cancer. This is best illustrated by the extreme frequency of cancers caused by exposure to sunlight in Xeroderma pigmentosum patients, evidently due to their inability to repair ultraviolet (UV) photoproducts in DNA. Mutations in NER are commonly detected in melanomas. Similar to what is observed in Xeroderma pigmentosum, most somatic mutations in melanoma cells are C:G→T:A transitions caused by UV[34] damage, which are primarily repaired by NER.

BER and NER are high-fidelity pathways, as suggested by studies showing that deactivation of these pathways leads to an increased chance of mutation from cryptic lesions or exogenous DNA damage[34–37]. As was shown recently, NER defects lead to increases in the density of C:G→T:A mutations under chronic low-dose UV radiation, conditions specifically relevant to melanoma[38].



We therefore reasoned that the relative reduction of mutation density in DHSs in individual melanoma genomes should parallel the integrity of NER pathway components. To test this, we analyzed 29 individual melanoma genomes sequenced at high coverage[1]. **Figure 2** and **Supplementary Table 3** show the continuous dependence of C:G→T:A mutation densities on chromatin accessibility in melanocytes in both introns and exons. We note that the observed effect is inconsistent with the recently discovered action of apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) proteins[39–41]. First, APOBEC proteins act on single-stranded DNA and it is unlikely that inaccessible and untranscribed DNA would more readily adopt a single-stranded conformation than DNA in DHSs and transcribed regions. Second, action of APOBEC proteins would preferentially increase rates of C→T transitions and C→G transversions within a TCA/TCT motif. Our observation is not confined to this motif, and we did not detect a parallel effect for C→G changes (**Supplementary Fig. 7**). Quantitatively, dependence of mutation density within a TCA/TCT motif on the number of DNase I cleavages was slightly lower (rather than higher) than the dependence of mutations within the TCG/TCC motif (*P* value for the interaction term in the regression analysis in **Supplementary Fig. 7** is 2.11×10^{-5}). This is inconsistent with the hypothesis that action of APOBEC proteins induces the observed dependency on chromatin accessibility.

This effect was far more pronounced in the chromatin of melanocytes than in that of other cells. Furthermore, the signature of TCR activity is also observed by the higher density of C→T over G→A in the nontemplate strand[2] of exons and introns (**Supplementary Fig. 8**). This demonstrates that NER provides multilayer protection against UV-light damage in genes because of activity of TCR and the accessibility of chromatin to global genome repair.

Overall, 9 out of 29 melanoma genomes harbored nonsynonymous mutations in NER genes. Notably, four melanoma genomes with the

Figure 3 Normalization of DHS hypomutation in melanoma genomes with mutated nucleotide excision repair pathway genes. Relative mutation density in DHSs of melanoma genomes is shown for samples with an intact NER system (blue) and samples with nonsynonymous mutations in NER pathway genes (red). Nonsynonymous changes in NER pathway genes significantly track relative mutation reduction in DHSs ($P < 0.0237$, Wilcoxon-Mann-Whitney test).

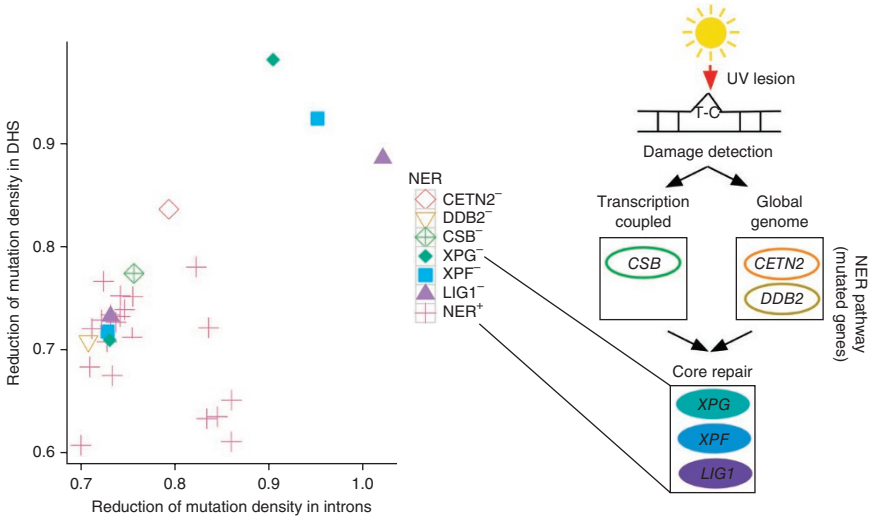


Figure 4 Reduction of mutation density in DHSs and in transcribed regions. Shown for individual melanoma samples (scatter plot) are nonsynonymous mutations in genes involved in NER (marked by shape and color corresponding to each gene). Roles of these genes within the NER pathway are shown by the diagram on the right. *XPG*, *XPF* and *LIG1* are core repair components; *CETN2* and *DDB2* are specific to global genome repair and are involved in lesion recognition. *CSB* is specific to TCR and is involved in recruiting NER to the stalled RNA Pol II. Samples with low (or no) reduction of somatic mutations in DHSs and carrying nonsynonymous changes in genes of core NER components also show low (or no) reduction of mutation frequency in transcribed regions, suggesting that core NER genes are responsible for both effects. NER, samples with mutations in NER genes; NER⁺, samples with intact NER genes.

lowest levels of mutation reduction at DHSs all harbor mutations in NER genes[42,43] (http://sciencepark.mdanderson.org/labs/wood/dna_repair_genes.html; $P < 0.0237$, Wilcoxon-Mann-Whitney test; **Fig. 3**). In three of these samples, mutation frequencies in DHSs returned close to the genomic baseline. The presence of genomes with mutations in NER genes and reduced mutation frequencies in DHSs is not surprising because NER mutations may appear late in cancer development, and some of the missense mutations may be functionally benign. This result implicates NER in observed reduction of mutation frequency in DHSs. It also provides an additional argument against selection as an explanation because purifying selection would not be expected to differentially affect melanoma samples.

Eight out of the nine samples with mutations in NER pathway genes harbored mutations in major components of the NER machinery (*XPG*/*ERCC5*, *XPF*/*ERCC4* and *LIG1*). These lesions would be expected to compromise both NER and TCR and therefore should affect mutation density in both DHSs and transcribed regions. Intriguingly, one of these samples also harbored a mutation in *CETN2*, which recognizes DNA distortions and therefore preferentially affects global genome repair function over TCR[44]. In agreement with this reasoning, three genomes carrying mutations in core subunits had a markedly smaller or negligible reduction of mutation density in transcribed regions compatible with defective TCR (**Fig. 4**). Concordantly, in the genome carrying a mutation in the *CETN2* gene, strong suppression of mutations in transcribed regions remained, implying that TCR function was not significantly compromised.

DISCUSSION

Taken together, our results suggest that the relative density of somatic mutations in cancer genomes is substantially suppressed in regulatory DNA, and that mutation frequency closely tracks chromatin accessibility. The hypomutational effect is highly localized and is statistically associated with intact global genome repair. The analysis of individual melanoma samples suggests that the relationship between relative mutation density and chromatin accessibility may be mediated by DNA repair. Our analysis could not completely rule out alternative explanations, such as selection for regulatory function or increased C→T deamination through enzymatic activities getting abnormal access to DNA. However, these alternatives would require an as-yet-unknown mechanism[45] to explain the association with NER in melanoma.

Our results link fine-scale chromatin accessibility with cancer mutation accumulation. Given the growing interest in the role of regulatory sequences in cancer progression[46], these results will help provide a necessary baseline for cancer genomics projects targeting noncoding regions, similar to computational approaches used in the analysis of protein-coding genes[47].

With the increasing amount of whole genome sequencing data, our approach can also be formalized and extended to associate mutational patterns with specific pathways, including DNA repair, DNA replication and chromatin remodeling. Mutational patterns can be treated as traits of individual tumor samples. With the large number of tumor samples available, these mutational traits can be associated with recurrent mutations in specific genes controlling mutagenesis, potentially identifying important players shaping somatic mutational landscapes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. These data have been deposited in GEO under accession numbers GSE29692 and GSE18927.

Note: Any Extended Data Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health (NIH) grants U54CA143874 and RO1MH101244 to S.R.S. and U54HG004592 and HG007010 to J.A.S. We thank D. Gordenin for his valuable feedback.

AUTHOR CONTRIBUTIONS

P.P., M.S.L., N.S., G.G., J.A.S. and S.R.S. conceived the study. P.P. led the analysis of sequencing data. M.S.L. and P.S. contributed to the analysis of sequencing data. R.E.T. and E.H. analyzed DHS data. L.A.G. provided melanoma sequencing data. S.M., G.G., J.A.S. and S.R.S. supervised the analysis. P.P., S.M., J.A.S. and S.R.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502–506 (2012).
2. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
3. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).

4. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
5. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
6. Hanawalt, P.C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* **9**, 958–970 (2008).
7. Lainé, J. & Egly, J. Initiation of DNA repair mediated by a stalled RNA polymerase IIO. *EMBO J.* **25**, 387–397 (2006).
8. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
9. Gross, D.S. & Garrard, W.T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
10. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
11. Legault, J., Tremblay, A., Ramotar, D. & Mirault, M.E. Clusters of S1 nuclease-hypersensitive sites induced *in vivo* by DNA damage. *Mol. Cell Biol.* **17**, 5437–5452 (1997).
12. Parker, S.C. *et al.* Mutational signatures of de-differentiation in functional non-coding regions of melanoma genomes. *PLoS Genet.* **8**, e1002871 (2012).
13. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A–TCF7L2 fusion. *Nat. Genet.* **43**, 964–968 (2011).
14. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
15. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33**, 136–143 (2012).
16. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
17. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
18. Faraway, J.J. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (Chapman & Hall/CRC, Boca Raton, 2006).
19. Hansen, R.S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* **107**, 139–144 (2010).
20. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
21. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
22. Chen, X. *et al.* Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* **335**, 1235–1238 (2012).
23. Sasaki, S. *et al.* Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323**, 401–404 (2009).
24. Cheng, K.C., Cahill, D.S., Kasai, H., Nishimura, S. & Loeb, L.A. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G—T and A—C substitutions. *J. Biol. Chem.* **267**, 166–172 (1992).
25. Kawanishi, S., Hiraku, Y., Pinlaor, S. & Ma, N. Oxidative and nitrative DNA damage in animals and patients with inflammatory diseases in relation to inflammation-related carcinogenesis. *Biol. Chem.* **387**, 365–372 (2006).
26. Hitomi, K., Iwai, S. & Tainer, J.A. The intricate structural chemistry of base excision repair machinery: implications for DNA damage recognition, removal, and repair. *DNA Repair* (*Amst.*) **6**, 410–428 (2007).
27. Amouroux, R., Campalans, A., Epe, B. & Radicella, J.P. Oxidative stress triggers the preferential assembly of base excision repair complexes on open chromatin regions. *Nucleic Acids Res.* **38**, 2878–2890 (2010).
28. Friedberg, E.C. *et al. DNA Repair and Mutagenesis* (ASM Press, 2006).
29. Bell, O., Tiwari, V.K., Thomä, N.H. & Schübeler, D. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* **12**, 554–564 (2011).
30. Thoma, F. Repair of UV lesions in nucleosomes–intrinsic properties and remodeling. *DNA Repair (Amst.)* **4**, 855–869 (2005).
31. Aboussekhra, A. *et al.* Mammalian DNA nucleotide excision repair reconstituted with purified protein components. *Cell* **80**, 859–868 (1995).
32. Yasuda, T. *et al.* Nucleosomal structure of undamaged DNA regions suppresses the non-specific DNA binding of the XPC complex. *DNA Repair (Amst.)* **4**, 389–395 (2005).
33. Fei, J. *et al.* Regulation of nucleotide excision repair by UV-DDB: prioritization of damage recognition to internucleosomal DNA. *PLoS Biol.* **9**, e1001183 (2011).
34. Shuck, S.C., Short, E.A. & Turchi, J.J. Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. *Cell Res.* **18**, 64–72 (2008).
35. Sugasawa, K. Xeroderma pigmentosum genes: functions inside and outside DNA repair. *Carcinogenesis* **29**, 455–465 (2008).
36. Hanawalt, P.C., Ford, J.M. & Lloyd, D.R. Functional characterization of global genomic DNA repair and its implications for cancer. *Mutat. Res.* **544**, 107–114 (2003).
37. Girard, P.M. & Boiteux, S. Repair of oxidized DNA bases in the yeast *Saccharomyces cerevisiae*. *Biochimie* **79**, 559–566 (1997).
38. Haruta, N., Kubota, Y. & Hishida, T. Chronic low-dose ultraviolet-induced mutagenesis in nucleotide excision repair-deficient cells. *Nucleic Acids Res.* **40**, 8406–8415 (2012).
39. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
40. 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).
41. Burns, M.B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
42. Lange, S.S., Takata, K. & Wood, R.D. DNA polymerases and cancer. *Nat. Rev. Cancer* **11**, 96–110 (2011).
43. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
44. Palomera-Sanchez, Z. & Zurita, M. Open, repair and close again: chromatin dynamics and the response to UV-induced DNA damage. *DNA Repair (Amst.)* **10**, 119–125 (2011).
45. Iyer, L.M., Zhang, D., Rogozin, I.B. & Aravind, L. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Res.* **39**, 9473–9497 (2011).
46. Huang, F.W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
47. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
48. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).

ONLINE METHODS

DNaseI hypersensitivity mapping. DNaseI mapping was conducted on cultured cancer cell lines, primary *ex vivo* hematopoietic cells, cultured primary cells and isolated fetal tissues using appropriate nuclei isolation protocols (below), followed by a standard processing pipeline. Data from lines A549, HepG2, LNCap, CACO2, PANC1, CLL, K562, CMK, NB4 and MCF7 are derived from reference 8. Generation of new data from M059J (glioblastoma), RPMI_7951 (melanoma), CD19 (B cell), CD20 (B-cell), melanocytes, fetal lung and fetal intestine are described below.

Isolation of nuclei from cultured cancer cell lines. Cells were cultured as described[8] and in accordance with the detailed protocols provided at http://www.uwencode.org/protocols. To prepare nuclei, freshly grown cells were centrifuged at 500*g* for 5 min (4 °C) in an Eppendorf Centrifuge 5810R, and washed in cold PBS (Cellgro/Mediatech Inc.). Cell pellets were resuspended in Buffer A (15 mM Tris-Cl pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA (Ambion/Life Technologies Corp)) pH 8.0, 0.5 mM EGTA (Boston BioProducts) pH 8.0, 0.5 mM spermidine (MP Biomedicals, LLC) and 0.15 mM spermine (MP Biomedicals, LLC) to a final concentration of 2 × 106 cells/ml. Nuclei were obtained by dropwise addition of an equal volume of Buffer A containing 0.04% IGEPAL CA-630 (Sigma-Aldrich) to the cells, followed by incubation on ice for 10 min. Nuclei were centrifuged at 1,000*g* for 5 min and then resuspended and washed with 25 ml of cold Buffer A. Nuclei were resuspended in 2 ml of Buffer A at a final concentration of 1 × 107 nuclei/ml.

Isolation of nuclei from hematopoietic cells. $CD19^+$ and $CD20^+$ cells (separately) were isolated by immunomagnetic separation by the Large Scale Cell Processing Facility at the Fred Hutchinson Cancer Research Center from healthy volunteer donors under an institutional review board (IRB)-approved protocol. Cells were pelleted by centrifugation for 5 min at 500*g* at 4 °C. Cells were washed in ice-cold PBS, then resuspended to 5 million cells per ml in Buffer A. An equal volume of ice-cold 2× IGEPAL CA-630 solution (0.02–0.06% range) was added and the tube was incubated for 5–6 min on ice to lyse the cells. Nuclei were pelleted by centrifugation for 5 min at 500*g* at 4 °C, resuspended in Buffer A and counted with a hemocytometer.

Isolation of nuclei from fetal tissues. Fetal lung and intestine tissues were obtained from morphologically normal fetuses by the Birth Defects Research Laboratory in the department of pediatrics at the University of Washington, collected under an IRB-approved protocol. Tissue was minced, resuspended in cold 250 mM sucrose, 1 mM $MgCl_2$, 10 mM Tris-Cl pH 7.5, with added EDTA Protease Inhibitor Cocktail (Roche Applied Science Corp.). Resuspended tissue from fetal brain, fetal lung, fetal kidney and fetal adrenal gland was dissociated by slowly homogenizing with a Dounce homogenizer. Resuspended tissue from fetal heart or fetal intestine was dissociated in a gentleMACS Dissociator (Miltenyi Biotech Inc.). After dissociation, all fetal tissues were filtered through a 100-μM filter, and nuclei pelleted by centrifugation 600*g* for 10 min. Pelleted nuclei were washed with Buffer A, resuspended in Buffer A and counted in a hemocytometer.

DNaseI mapping from isolated nuclei. Briefly, DNaseI digestion was done as described[8], with minor modifications. Isolated nuclei (2 × 106) from suspension cells or dissociated tissue were washed with 15 mM Tris-Cl pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.5 mM spermidine and 0.15 mM spermine then subjected to DNaseI digestion for 3 min at 37 °C in 13.5 mM Tris-HCl pH 8.0, 87 mM NaCl, 54 mM KCl, 6 mM $CaCl_2$, 0.9 mM EDTA, 0.45 mM EGTA, 0.45 mM spermidine. Digestion was stopped by addition of 50 mM Tris-HCl pH 8.0, 100 mM NaCl, 0.1% SDS, 100 mM EDTA pH 8.0, 1 mM spermidine, 0.3 mM spermine. A range of DNaseI (Sigma-Aldrich), 10–80 U/ml) concentrations was used for each preparation of nuclei and the sample with the optimum difference between treatment and no treatment with DNaseI was used for sequencing library construction. DNaseI double-hit fragments were collected by ultra-centrifugation and gel-purified. Adaptors were ligated to the ends of purified fragments, and the resulting libraries sequenced on an Illumina Genome Analyzer IIx according to a standard protocol. *Processing of DNaseI-seq data.* 36-base reads with up to two mismatches were mapped to the human genome (GRCh37/hg19) using the sequence aligner BOWTIE. DHSs were identified using the Hotspot algorithm (8) at a false-discovery rate (FDR) threshold of 5%. Genomic feature overlaps and distance calculations were performed using the BEDOPS[48] suite of software tools available at https://github.com/bedops/bedops.

Data availability. All DNaseI data used in this study have been released to the ENCODE Project repository or to the Roadmap Epigenomics Mapping Consortium data coordination center. These data have been deposited in GEO under accession numbers GSE29692 and GSE18927. Data are also available for download through http://www.uwencode.org/data and through the data links at http://www.epigenomebrowser.org/.

Cancer data sets. We restricted our study to cancer genomes sequenced by the Sanger Institute and by Broad Institute. Whole genomes of COLO-829 and NCI-H209 cell lines have been sequenced by the Sanger Institute[2,3]. The COLO-829 cell line was derived from metastatic tissue. NCI-H209, an immortal cell line of a small-cell lung cancer, was derived from a bone marrow metastasis. The mutation lists that we used in this study can be found in http://dcc.icgc.org/download/legacy_data_releases/version_07/ under the folders Malignant_Melanoma-WTSI-UK (COLO-829) and Small_Cell_Lung_Carcinoma-WTSI-UK (NCI-H209). Nine colon cancer genomes[13], 23 multiple myelomas[5] and 29 melanoma genomes have been sequenced by the Broad Institute. For the four Broad Institute data sets we identified sites in the genome where we have over 80% power to detect mutations (this is defined by at least 14 reads the covered this position in the tumor sample and 8 reads that cover this position in normal). About 81% of the bases in colon and multiple myeloma genomes are sufficiently covered, and about 86% of the melanoma genomes are sufficiently covered.

Annotations. Gene and exon coordinates were retrieved for hg19 from UCSC genome browser. For the flank analysis we took 1,000-bp windows around the DHSs, as some of flank regions overlapped with DHSs. We removed the DHS defined sites from the original set of flanks.

Density of mutations. We calculated the number of mutations per bp that can be mapped uniquely by the DHS essays.

Poisson regression. To test whether DHS regions have an additional impact on rates in addition to other cofounding factors, we used multivariate Poisson regression[18]. We divided the genomes into nonoverlapping 400-bp windows. Every bin was classified as intergenic, intronic and exonic regions. The windows were also classified as DHS regions when overlap at least 80 bp of any DHSs. For each window we calculated the following quantities: GC content, CpG content, distance from the nearest transcription start site, distance from nearest CpG islands, distance from nearest DHSs, and mutation counts and coverage (i.e., the number of bases for which we have 80% power to detect mutation) for colon and multiple myeloma, and similarly the number of GC bases in a window that are covered in the Broad Institute melanoma samples. Then using the glm function in R we calculated the estimated rates[18]. We repeated the analysis for spatially separated windows comprising 20% of the data to ensure that possible short-range interdependence between neighboring windows does not artificially inflate statistical significance.

doi:10.1038/nbt.2778