nature structural & molecular biology

Nuclear topology modulates the mutational landscapes of cancer genomes

Kyle S Smith[1,2,5], Lin L Liu[3–5], Shridar Ganesan[1], Franziska Michor[3,4] & Subhajyoti De[1]

Nuclear organization of genomic DNA affects processes of DNA damage and repair, yet its effects on mutational landscapes in cancer genomes remain unclear. Here we analyzed genome-wide somatic mutations from 366 samples of six cancer types. We found that lamina-associated regions, which are typically localized at the nuclear periphery, displayed higher somatic mutation frequencies than did the interlamina regions at the nuclear core. This effect was observed even after adjustment for features such as GC percentage, chromatin, and replication timing. Furthermore, mutational signatures differed between the nuclear core and periphery, thus indicating differences in the patterns of DNA-damage or DNA-repair processes. For instance, smoking and UV-related signatures, as well as substitutions at certain motifs, were more enriched in the nuclear periphery. Thus, the nuclear architecture may influence mutational landscapes in cancer genomes beyond the previously described effects of chromatin structure and replication timing.

Emerging evidence indicates that somatic mutations in cancer genomes are nonrandomly distributed and are influenced by factors such as genomic context and DNA secondary structures, chromatin organization, transcriptional activity, and replication timing[1–11]. Local variation in the mutation burden stems from variability in processes of DNA damage and/or repair[3,5,12,13], and has implications for the identification of potential cancer-driver genes[14] and the clinical management of cancer patients, for example, radiosensitivity and immunotherapy[15]. However, the factors identified to date do not explain the entire extent of regional variation of the mutational burden in cancer genomes, thus suggesting that other factors remain to be identified.

Genomic DNA is folded into higher-order domains, which occupy different territories in the three-dimensional architecture of the nucleus[16–18], and nuclear-lamina-binding regions are usually at the nuclear periphery[16,19,20]. Nuclear organization of genetic material plays an important role in DNA replication[21] as well as the processes of DNA damage and repair[22–24]. For instance, the nuclear-lamina-associated regions are refractory to homologous-recombination-mediated repair and use an error-prone alternative end-joining mechanism to repair DNA double-strand breaks[25]. Oct-1- and p53-dependent pathways link lamin functions to the oxidative-stress response[26]. Indeed, a previous multivariate analysis has suggested that nuclear-lamina association significantly contributes to variations in germline mutation rates[27]. Furthermore, it has recently been reported that regulatory-domain boundaries are frequently disrupted in cancer[28], and in some cases, such boundaries and the chromatin loops that underlie them are associated with unusual mutational spectra[29]. Here, we hypothesized that the nuclear organization of genomic DNA might modulate the somatic mutational landscape of cancer genomes and that its effects might surpass the variations due to known covariates such as chromatin domains and DNA-replication timing[4,6].

RESULTS

Integration of mutation data from multiple cancer types

To test these hypotheses, we obtained somatic point-mutation data from 366 completely sequenced genomes of six different cancer types: melanoma (SKCA, 25 samples)[30], lung squamous cell carcinoma (LUSC, 31 samples)[31], gastric cancer (STAD, 100 samples)[32], diffuse large B cell lymphoma (DLBCL, 40 samples)[33], chronic lymphocytic leukemia (CLL, 150 samples)[34], and prostate cancer (PRAD, 20 samples)[35,36]. The somatic mutation frequencies for these cancer cohorts were comparable to published estimates of the mutation burden for the respective cancer type[14] (**Supplementary Fig. 1**). We chose these cancer types because they have distinct etiologies, different patterns of DNA damage and repair, and a difference of several orders of magnitude in somatic mutation frequencies[14,37], thus enabling us to identify the effects of nuclear localization on somatic mutational patterns across diverse cancer types. We focused on the noncoding, nonrepetitive, nonconserved regions of the genome and analyzed somatic mutations therein to minimize biases due to selection during clonal evolution as well as sequencing and mapping artifacts (details in Online Methods). We denote the mutation detection frequency per base pair in these regions, normalized to the mutation detection frequency per base pair in the genome, as the adjusted mutation rate (AMR).

[1]Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey, USA. [2]Department of Pharmacology, University of Colorado Denver, Aurora, Colorado, USA. [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [4]Department of Stem Cells and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to F.M. (michor@jimmy.harvard.edu) or S.D. (subhajyoti.de@rutgers.edu).

Received 6 February; accepted 28 August; published online 2 October 2017; doi:10.1038/nsmb.3474

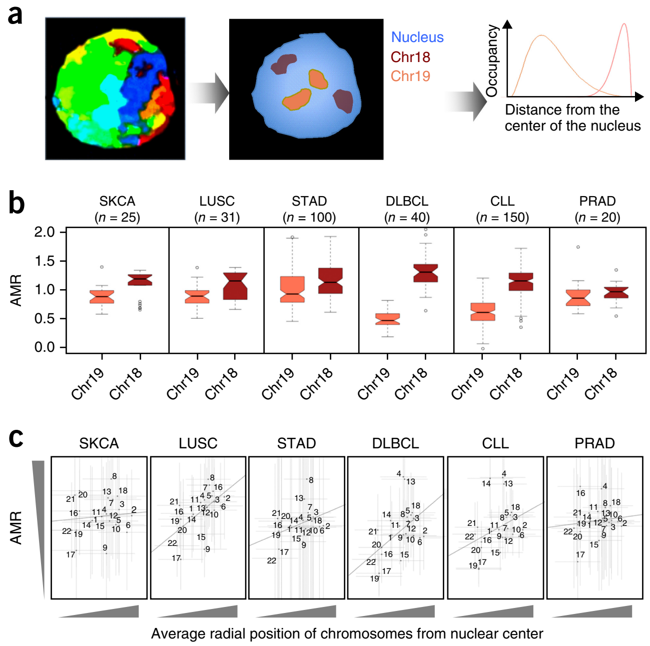

Figure 1 Somatic mutation frequencies differ between chromosomes located at the nuclear core versus the periphery. (**a**) Eukaryotic chromosomes occupy different radial positions from the center of the nucleus. Classic examples are human chr18 and chr19, which are located at the nuclear periphery and core, respectively. (**b**) The adjusted mutation rate (AMR) tends to be significantly higher for chr18 relative to chr19. Mann–Whitney *U*-test $P < 1 \times 10^{-2}$ for all cohorts. In the box plots, the upper whisker is 1.5× the interquartile range (IQR) more than the third quartile or the maximal value of the adjusted mutation rate (depending on which value is greater); the lower whisker is 1.5× IQR lower than the first quartile or the minimum value of the adjusted mutation rate (depending on which value is smaller), where the IQR is the difference between the third quartile and the first quartile, i.e., the box length. Center line, median; points, data outside the upper and lower whiskers. (**c**) AMR values for chr1 to chr22, plotted against the average normalized radial distances from the center of the nucleus. Average and s.d. of normalized radial distances of chromosomes from the center of the nucleus were estimated from 54 measurements, as described in ref. 38. The numbers of samples used for AMR estimation are as in **b**. The s.d. values of AMR and radial positions are shown with vertical and horizontal error bars, respectively. The coefficient of determination was <0.1 in all cohorts.

Mutation patterns for chromosomes 18 and 19

First, we investigated whether nuclear localization of chromosomes correlated with the AMR. We used chromosome (chr) 18 and chr19 as classic examples, because human chr18 preferentially localizes close to the nuclear periphery, whereas chr19 primarily localizes to the nuclear core[38] (**Fig. 1a**). Indeed, the AMR for chr18 was significantly higher than that for chr19 across all six cancer types analyzed (**Fig. 1b**; Mann–Whitney *U*-test $P < 1 \times 10^{-2}$ for all cohorts). Integrating paired copy-number data when available (for example, LUSC; **Supplementary Fig. 2**), we established that the difference was not due to proportionally more copy-number-deletion events on chr19. Extending this investigation to all other autosomes, whose nuclear positioning was determined through 3D fluorescence *in situ* hybridization, we observed a similar association between the overall nuclear positioning of chromosomes and their AMR: those predominantly in the nuclear periphery had a higher AMR than those in the core (**Fig. 1c**). The coefficient of determination was weak (<0.1) in all cohorts, at least partly because chromosomes are large nuclear entities that typically span multiple nuclear domains; i.e., some parts of the same chromosome may be localized at the periphery while other parts are localized at the relative interior of the nucleus[38]. Therefore, we investigated whether more precise measures of the nuclear localization of genomic regions within and across chromosomes might be able to explain the observed differences in chromosome-level variations in AMR.

Genome-wide mutation patterns are associated with nuclear localization

We obtained chromatin immunoprecipitation (ChIP) data for the lamin family proteins lamin A and lamin B1 (**Fig. 2a**), and classified a region as being constitutively present in the nuclear periphery if the region was associated with lamins in all cell types examined; in contrast, a region was categorized as being constitutively present in the nuclear core if it did not overlap with lamin-associated domains in any of the cell types analyzed (**Fig. 2b**). As before, we prioritized noncoding, nonrepetitive, nonconserved segments of genomic regions that were constitutively present at the periphery (constitutive lamina-associated domains; cLADs) and core (interlamina-associated domains; iLADs), respectively. We then in tegrated somatic mutation data from each cancer cohort and calculated the AMRs for these two types of regions for each sample. We found that the AMR for cLADs was significantly higher than that for iLADs, and once again, this observation was consistent across all six cancer cohorts (**Fig. 2c**; Mann–Whitney *U*-test $P < 1 \times 10^{-5}$ for all cohorts). Within respective chromosomes, cLAD and iLAD regions displayed a systematic difference in their AMRs, regardless of the average nuclear localization of the chromosomes. A minor subset of lamins accumulates away from the nuclear periphery, usually in nucleoli-associated domains (NADs)[39], and we found consistent results after excluding NADs (**Supplementary Fig. 3**). We also repeated the experiments more conservatively by analyzing only the cLADs and iLADs with evolutionarily conserved patterns of nuclear localization, after integrating data on lamina-associated regions from multiple cell types in mice, and we found similar results (**Supplementary Fig. 3**). Therefore, our findings were not sensitive to our choice of definition for cLADs and iLADs, and indicated that lamina-associated regions localized at the periphery have higher somatic mutation frequencies than do the interlamina regions at the nuclear core.

We next focused on mutational-signature differences between the nuclear core and periphery. In the SKCA cohort, UV-induced C-to-T substitutions, including those in the pi-pyrimidine context, were proportionally more common in the cLADs than the iLADs (Mann–Whitney *U*-test $P < 1 \times 10^{-8}$) (**Fig. 2d** and **Supplementary Fig. 4**). Similarly, C-to-A substitutions displayed a higher enrichment in the cLADs in the LUSC cohort, thus indicating that smoking-associated oxidative DNA damage was greater in the nuclear periphery than the nuclear core (Mann–Whitney *U*-test $P < 1 \times 10^{-2}$, **Fig. 2d** and **Supplementary Fig. 4**). For LUSC patients, data on their smoking history and the number of pack-years were available. We calculated the AMR for cLAD and iLAD, considering only C-G to A-T mutations, and plotted the AMR(cLAD)/AMR(iLAD) ratio against the number of pack-years. Indeed, we found that the number of pack-years was weakly correlated with the AMR(cLAD)/AMR(iLAD) ratio (Spearman correlation coefficient of 0.29; **Supplementary Fig. 4**), thus suggesting that the relative strength of the signature of oxidative damage induced by smoking in the nuclear periphery was higher for heavy smokers than light smokers. Therefore, in cancer types driven by external carcinogens, the nuclear periphery had a proportionally higher burden of corresponding mutation signatures.

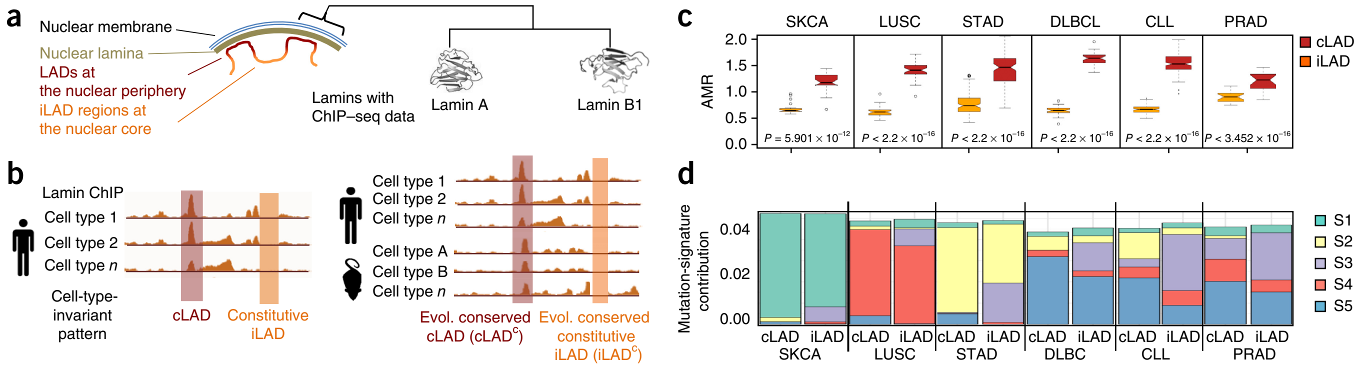

Figure 2 Somatic mutation patterns differ between genomic regions located at the nuclear core versus the periphery. (**a**) Genomic regions interacting with lamina proteins such as lamins A and B1 are predominantly localized at the nuclear periphery (with some exceptions). (**b**) Identification of genomic regions that are predominantly at the nuclear core (iLAD) and periphery (cLAD), respectively, in a cell-type-invariant manner. Lamin ChIP was used to identify genomic regions interacting with lamins in individual cell types. We classified a region as being constitutively present in the nuclear periphery if the region was associated with lamins in all cell types examined; in contrast, a region was categorized as being constitutively present in the nuclear core if it did not overlap with lamin-associated domains in any of the indicated cell types analyzed. A subset of these regions also showed preferential positioning at the nuclear core (iLAD^c) and periphery (cLAD^c) in an evolutionarily (evol.) conserved manner (**Supplementary Fig. 3**). (**c**) cLADs tend to have a significantly higher AMR than do iLADs. Box plots and number of samples in each cohort are as in **Figure 1b**. Mann–Whitney *U*-test $P < 1 \times 10^{-5}$ for all cohorts. Similar results were observed when mutations in iLAD^c and cLAD^c regions were considered (**Supplementary Fig. 3**). (**d**) Mutational signatures differ between the nuclear core and periphery across different cancer types. Somatic mutational signatures were identified on the basis of non-negative matrix factorization and principal component analysis in the somaticSignature[40] R package. Details of the mutation signatures S1–S5 are provided in **Supplementary Figure 4**.

Even though the patterns of DNA damage and response in the cancer cohorts were dominated by disease etiology, there were some other differences in mutational signatures between cLAD and iLADs, which were tissue-type invariant (**Fig. 2d** and **Supplementary Fig. 4**). For instance, when we summarized the trinucleotide-substitution patterns into mutational signatures by using nonnegative matrix factorization[40], mutation signatures 3 and 5, compared with the other signatures, had a proportionally larger contribution in the iLAD and cLADs, respectively, in most cancer types. Translation of the mutational signatures into substitution patterns clearly indicated that most of the cancer types had a proportional increase in the contribution of mutations in the WNW context (in which W is A or T, and N is A, G, C, or T) in the cLADs at the periphery compared with the iLADs in the core. Different cancer types, however, showed subtle variations in the preference for specific submotifs; for instance, in the DLBCL and CLL cohorts, W[T-to-G]W and also W[T-to-C]W mutations were relatively more common in the cLADs than in the iLADs (Mann–Whitney *U*-test $P < 1 \times 10^{-10}$). There were other differences in mutational signatures that were dominated by the biology of the cancer type. For instance, in the SKCA cohort, T[C-to-T]W substitutions were more common in the cLADs than the iLADs (Mann–Whitney *U*-test $P < 1 \times 10^{-8}$; **Supplementary Fig. 4**).

Mutation patterns are associated with nuclear localization after adjustment for covariates

Nuclear localization of genomic DNA is coupled with many genomic and epigenomic features: regions in the nuclear periphery compared with the core tend to be, on average, more AT rich, gene poor, and heterochromatic, and to have later replication timing[16,18–20]. Features such as replication timing and chromatin influence processes of DNA damage and repair, thus affecting mutational frequencies and signatures[4,6,41–43] (**Fig. 3a**). However, not all point mutations arise during replication, and nuclear lamins play a key role in DNA double-strand-break repair, such that the preference for repair mechanisms in the nuclear periphery is different from that in the nuclear interior[25]. We thus assessed whether nuclear localization influences the mutational landscape in cancer genomes beyond what is already captured by chromatin and replication timing. Using a multiple linear regression including chromatin, replication timing, gene density, and GC content as covariates, we observed that the cLAD density was significantly associated with somatic mutation frequency, even after adjustment for other features, in all cancer types tested (**Supplementary Fig. 5** and **Supplementary Tables 1** and **2**). After normalizing all features to zero mean and unitary variance, we also computed the variable importance metrics by using random forest regression (**Fig. 3b**) and determined the effect sizes by using multiple linear regression (**Supplementary Fig. 5**) for all features including the cLAD density in each 1-Mb bin. In general, the variable importance metrics of the cLAD density computed from the random forest regression were of similar magnitudes to those for the trimethyl histone H3 K9 (H3K9me3) signal and replication timing. We also computed the approximate conditional-variable importance metrics to address the multicollinearities among the features (Online Methods). We found that the cLAD density had a similar metric magnitude to that of H3K9me3 and replication timing in most cases. We further ascertained that the influence of the sample size in the collected cohort on the results was not significant, on the basis of a subsampling analysis of the lymphoma cohort (Online Methods).

Key differences between the nuclear core and periphery in the detected mutational signatures also persisted even when we adjusted for both chromatin (**Fig. 3c**) and replication timing (**Supplementary Note**). In the SKCA cohort, we found a proportionally higher burden of UV-mediated DNA damage and translesion-synthesis errors in the pyrimidine-dimer context in the nuclear periphery relative to that in the core, even when controlling for replication timing and chromatin. We also found that cLADs had a larger contribution of the mutational signature $S_{SKCA}1$, dominated by T[C-to-T]W substitutions, whereas iLADs had a relative enrichment in mutational signature $S_{SKCA}2$, representing C[C-to-T]Y; these preferences were observed even after adjustment for both chromatin and replication timing. Indeed, there is evidence that nuclear lamin B1 is critical for the nucleotide-excision-repair pathway for effective repair of the DNA-damage response to UV irradiation[44]. The preference for C[C-to-T]N (where

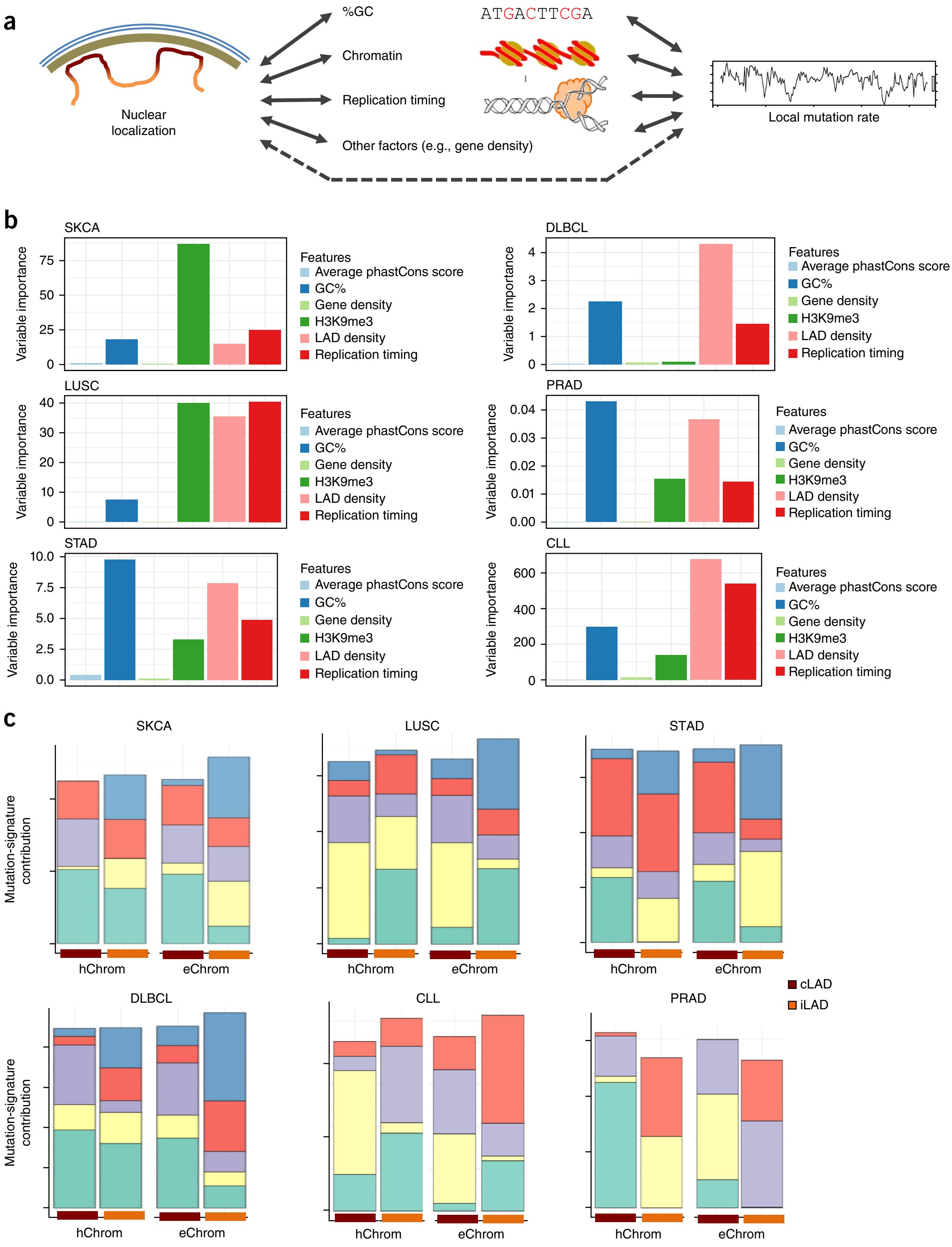

Figure 3 Differences in somatic mutation patterns between the nuclear core and periphery are not solely due to chromatin. (**a**) Nuclear localization of genomic DNA is at least partly associated with chromatin. Other features, such as GC content, replication timing and gene density, also modulate local mutation rates. Effects of nuclear localization beyond those explained by these known covariates are examined in this study. (**b**) Marginal variable importance metrics for the indicated genomic features were computed from random forest regression and compared in six cancer types. (**c**) Mutational signatures differ between genomic material localized at the nuclear core and periphery even when assessed within similar euchromatic (eChrom) or heterochromatic (hChrom) contexts. Color codes of mutation signatures and their relative contributions are comparable only within respective cancer cohorts. Additional comparative assessments are shown in **Supplementary Figure 5**. Details of the mutational signature analysis are presented in the **Supplementary Note**.

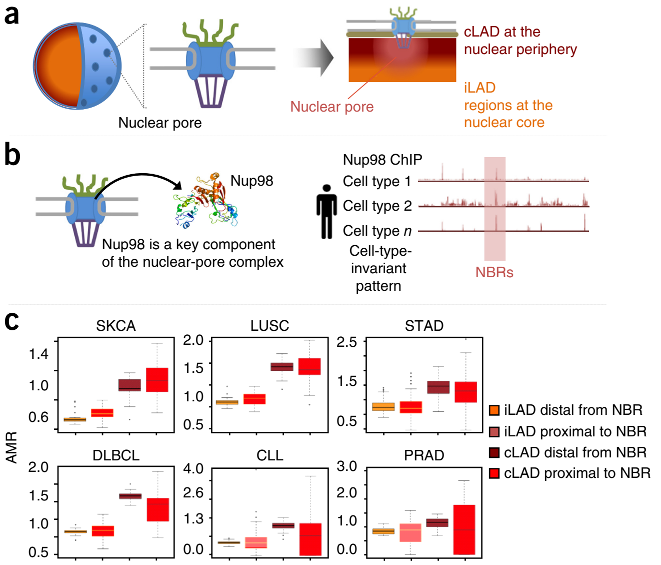

Figure 4 Nuclear-pore-proximal genomic regions have characteristic somatic mutation patterns. (**a**) Schematic representation of nuclear pores as large multiprotein complexes on the nuclear envelope that regulate nuclear transport of biomolecules, including some mutagens and DNA-repair factors. (**b**) Nup98 is a key component of the nuclear-pore complex, and NBRs were identified from Nup98 ChIP data through an approach similar to that in **Figure 1b**. We classified genomic regions as nuclear-pore proximal if they were within 50 kb of Nup58 ChIP peaks in all cell types examined. In contrast, genomic regions that were at least 50 kb from Nup58 ChIP peaks in all cell types were considered distal to nuclear pores. (**c**) AMR of cLAD and iLADs that are proximal to and distal from nuclear-pore regions were compared. False-discovery-rate-adjusted Mann–Whitney *U*-test *P* values were $< 5 \times 10^{-2}$ in the STAD, lymphoma, and CLL cohorts. The box plots are as in **Figure 1b**.

N is A, T, G, or C) in iLADs over cLADs was detectable in other cancer types including LUSC (signature $S_{\mathrm{LUSC}}1$). Moreover, in the LUSC cohort, the signature of oxidative DNA damage marked by C-to-A substitutions, especially W[C-to-A]W, was more common in the cLADs even after adjustment for chromatin and replication timing (Mann–Whitney *U*-test $P < 1 \times 10^{-10}$). Therefore, a higher burden of mutation signatures arising because of external mutagens in the nuclear periphery was at least partly attributable to nuclear localization even with adjustment for replication timing and chromatin context. The increased incidence of somatic mutations in the WNW context was also detected across most cancer types regardless of replication timing and chromatin context. In the DLBCL and CLL cohorts, we observed an increase in C-to-T transitions in iLADs and an increase in T-to- G transversions in the WTN trinucleotide context in cLADs (Mann–Whitney *U*-test $P < 1 \times 10^{-5}$) (**Supplementary Note**). The former signature is similar to COSMIC signature 2 and therefore might have been due to cytosine deamination mediated by off-target effects of AICDA/APOBEC family enzymes[37,45,46]. This hypothesis is also consistent with the observation that AICDA is predominantly localized in nucleoli and Cajal bodies in the nuclear core[47]. The latter signature is similar to COSMIC signature 9, and a variant of this signature, N[T-to-G]T, was also observed in cLADs in the STAD cohort (Mann–Whitney *U*-test *P* value $< 1\times 10^{-7}$; **Supplementary Note**). On the basis of the interpretation of COSMIC signature 9, we suspect that the signature arises primarily because of mutations attributed to DNA polymerase η (ref. 37), but other factors may also play a role.

Nuclear-pore-associated regions have distinct mutation patterns

Nuclear pores are large multiprotein channels that are conduits for the nuclear transport of many small molecules and proteins, including DNA-damage-response and DNA-damage-repair factors, and nuclear pores play a key role in DNA repair[24,48]. To further extend our analysis, we investigated whether nuclear-pore-proximal regions (**Fig. 4a**) displayed mutational patterns different from those observed for the nuclear core and periphery regions. Nup98 is a component of the nuclear-pore complex (**Fig. 4b**); it is predominantly localized in the nuclear periphery, but it is also present in the nuclear interior, and its dynamics of interaction with genomic regions depends on the developmental trajectory of the cell[49]. Using Nup98 ChIP–seq data from multiple cell types[49], we identified genomic regions that bound to Nup98 in one or more cell types. Accordingly we identified cLAD and iLADs that were localized in the neighborhood of, or distal from, Nup98-bound regions (NBRs) in a cell-type-invariant manner (Online Methods). cLADs at the nuclear periphery that were also close to NBRs in a cell-type-invariant manner were likely to be nuclear-pore proximal. Unfortunately, the number of mutations in these subregions was small; nonetheless, cLADs that were nuclear-pore proximal had a lower AMR than those that were distal (**Fig. 4c**) in the STAD, lymphoma, and CLL cohorts (false-discovery-rate-adjusted Mann–Whitney *U*-test $P < 5 \times 10^{-2}$). The trinucleotide contexts of the substitution patterns in NBRs did not show any prominent cancer-type-invariant mutational signatures (**Supplementary Note**). Interaction of genomic DNA with the nuclear pore is dynamic, and DNA breaks are shunted to nuclear pores for a repair pathway controlled by a conserved SUMO-dependent E3 ligase[50]. Therefore, the effects of nuclear-pore-assisted repair may not be restricted to NBRs. Nonetheless, DNA lesions within NBRs may be relocated to the nuclear-pore complex more quickly for repair, and this process may play a role in lowering the AMR in NBRs. Further evidence is required to conclusively establish this conjecture.

DISCUSSION

Together, our mutational signatures and multivariate analyses indicated that the nuclear localization of genomic DNA may potentially modulate somatic mutational patterns of cancer genomes, and that the effect attributed to nuclear localization on mutational landscapes in cancer is of similar magnitude to those of previously determined features such as chromatin and replication timing. This finding probably arose because a subset of mutations do not emerge during replication, and the nuclear lamina plays a role in DNA-damage recognition and repair[21–24]. Our observations are consistent with the reported effects of the nuclear lamina on variations in the germline mutation rate[27]. Even benign somatic tissue samples, although having considerably fewer somatic mutations, also showed similar patterns (**Supplementary Fig. 6**; $P > 5 \times 10^{-2}$). However, our results should be interpreted with caution, because: (i) The LAD information used in our study does not correspond to the (potentially unknown) cell type of origin of the six cancer types examined in this paper. To identify the effects of cell-type-specific LADs on mutation frequencies requires matched data, which are not yet available. (ii) The multicollinearities among features such as replication timing, chromatin, and nuclear localization pose a statistical challenge to dissecting their individual effects. Here, we performed our analyses from multiple angles, looking only at 'neutrally' evolving genomic regions and investigating the data by using different multivariate models (**Supplementary Figs. 7** and **8** and **Supplementary Table 3**). Although the results of these different analyses are generally consistent with one another, further investigation is still needed to confirm the effects of nuclear localization on somatic mutations in somatic tissues.

There are multiple biological processes that might contribute to the observed differences in the mutation burden between the nuclear core and periphery. In 1975, Hsu proposed the 'bodyguard hypothesis', suggesting that constitutive heterochromatin is used by the cell as a bodyguard to protect the vital euchromatin by forming a layer of dispensable shield on the outer surface of the nucleus[51]. In agreement with this hypothesis, in the melanoma and lung squamous cell carcinoma cohorts, we found that the nuclear periphery, compared with the core, had a larger mutation burden and also displayed mutation signatures consistent with greater exposure to external mutagens. In addition, some of the processes of DNA-damage recognition and repair depend on lamina association or nuclear localization. For instance, lamin B1 controls oxidative-stress responses through sequestration of Oct-1 at the nuclear periphery[52], thus also leading to slow repair of DNA lesions. Furthermore, competing DNA-repair mechanisms may recruit different DNA polymerases or their cofactors with variable fidelity and signature error profiles[53], depending on nuclear localization. For instance, XPC and XPA are two damage-recognition proteins associated with the nucleotide-excision-repair pathway, and after UV radiation, both XPC and XPA quickly accumulate in the border region of condensed chromatin called perichromatin at the nuclear core, but in condensed heterochromatin domains, accumulation of only XPC has been observed[54]. Another possibility could be that competing DNA-repair mechanisms recruit different DNA polymerases or their cofactors with variable fidelity and signature error profiles[53], depending on nuclear localization and cancer type. Furthermore, there is substantial evidence that DNA double-strand-break repair is nuclear localization-dependent repair in the nuclear interior or at the nuclear pores that occurs through the repair pathways of classical homologous recombination and nonhomologous end-joining; however, the nuclear-lamina-proximal regions tend to be refractory to homologous recombination and instead allow repair primarily through the error-prone mechanisms of alternative end-joining[25], which may be a source of point mutations in the nuclear periphery. In any case, our findings support analyzing somatic mutations in tumor and benign tissues in the context of their 3D nuclear architecture.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

ACKNOWLEDGMENTS

The authors acknowledge financial support from T15LM009451 (K.S.S.), U54CA193461 (F.M.), P30CA072720, the American Cancer Society, and the Boettcher Foundation (S.D.). The authors thank other members of the laboratories of F.M. and S.D. for helpful discussions. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

AUTHOR CONTRIBUTIONS

S.D. conceived the project with F.M.; K.S.S., L.L.L., F.M., and S.D. designed the experiments. K.S.S., L.L.L., and S.D. performed the experiments. K.S.S., L.L.L., S.G., F.M., and S.D. interpreted the results. F.M. and S.D. wrote the manuscript with input from other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. De, S. & Michor, F. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat. Struct. Mol. Biol.* **18**, 950–955 (2011).
2. De, S. & Michor, F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.* **29**, 1103–1108 (2011).
3. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
4. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* **4**, 1502 (2013).
5. Roberts, S.A. & Gordenin, D.A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* **14**, 786–800 (2014).
6. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
7. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
8. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
9. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
10. Smith, K.S. *et al.* Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. *Nucleic Acids Res.* **43**, 5307–5317 (2015).
11. Pedersen, B.S. & De, S. Loss of heterozygosity preferentially occurs in early replicating regions in cancer genomes. *Nucleic Acids Res.* **41**, 7615–7624 (2013).
12. Watson, I.R., Takahashi, K., Futreal, P.A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14**, 703–718 (2013).
13. Alexandrov, L.B. & Stratton, M.R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).
14. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
15. De, S. & Ganesan, S. Looking beyond drivers and passengers in cancer genome sequencing data. *Ann. Oncol.* **28**, 938–945 (2016).
16. Bickmore, W.A. The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* **14**, 67–84 (2013).
17. Gibcus, J.H. & Dekker, J. The hierarchy of the 3D genome. *Mol. Cell* **49**, 773–782 (2013).
18. Cavalli, G. & Misteli, T. Functional implications of genome topology. *Nat. Struct. Mol. Biol.* **20**, 290–299 (2013).
19. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
20. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603–613 (2010).
21. Meister, P., Taddei, A. & Gasser, S.M. In and out of the replication factory. *Cell* **125**, 1233–1235 (2006).
22. Ball, A.R. Jr. & Yokomori, K. Damage site chromatin: open or closed? *Curr. Opin. Cell Biol.* **23**, 277–283 (2011).
23. Bell, O., Tiwari, V.K., Thomä, N.H. & Schübeler, D. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* **12**, 554–564 (2011).
24. Lemaître, C. & Bickmore, W.A. Chromatin at the nuclear periphery and the regulation of genome functions. *Histochem. Cell Biol.* **144**, 111–122 (2015).
25. Lemaître, C. *et al.* Nuclear position dictates DNA repair pathway choice. *Genes Dev.* **28**, 2450–2463 (2014).
26. Shimi, T. & Goldman, R.D. Nuclear lamins and oxidative stress in cell proliferation and longevity. *Adv. Exp. Med. Biol.* **773**, 415–430 (2014).
27. Ananda, G., Chiaromonte, F. & Makova, K.D. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol.* **12**, R27 (2011).
28. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
29. Kaiser, V.B., Taylor, M.S. & Semple, C.A. Mutational biases drive elevated rates of substitution at regulatory sites across cancer types. *PLoS Genet.* **12**, e1006207 (2016).
30. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent *PREX2* mutations. *Nature* **485**, 502–506 (2012).
31. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
32. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* **46**, 573–582 (2014).
33. Morin, R.D. *et al.* Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265 (2013).
34. Puente, X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
35. Abeshouse, A. *et al.* The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
36. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
37. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
38. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157 (2005).

39. Németh, A. *et al.* Initial genomics of the human nucleolus. *PLoS Genet.* **6**, e1000889 (2010).
40. Gehring, J.S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
41. Kazanov, M.D. *et al.* APOBEC-induced cancer mutations are uniquely enriched in early-replicating, gene-dense, and active chromatin regions. *Cell Rep.* **13**, 1103–1109 (2015).
42. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
43. Woo, Y.H. & Li, W.H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.* **3**, 1004 (2012).
44. Butin-Israeli, V., Adam, S.A. & Goldman, R.D. Regulation of nucleotide excision repair by nuclear lamin b1. *PLoS One* **8**, e69169 (2013).
45. Di Noia, J.M. & Neuberger, M.S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).
46. Puente, X.S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).
47. Hu, Y. *et al.* Activation-induced cytidine deaminase (AID) is localized to subnuclear domains enriched in splicing factors. *Exp. Cell Res.* **322**, 178–192 (2014).
48. Misteli, T. & Soutoglou, E. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat. Rev. Mol. Cell Biol.* **10**, 243–254 (2009).
49. Liang, Y., Franks, T.M., Marchetto, M.C., Gage, F.H. & Hetzer, M.W. Dynamic association of NUP98 with the human genome. *PLoS Genet.* **9**, e1003308 (2013).
50. Nagai, S. *et al.* Functional targeting of DNA damage to a nuclear pore-associated SUMO-dependent ubiquitin ligase. *Science* **322**, 597–602 (2008).
51. Hsu, T.C. A possible function of constitutive heterochromatin: the bodyguard hypothesis. *Genetics* **79** (Suppl.), 137–150 (1975).
52. Malhas, A.N., Lee, C.F. & Vaux, D.J. Lamin B1 controls oxidative stress responses via Oct-1. *J. Cell Biol.* **184**, 45–55 (2009).
53. Lange, S.S., Takata, K. & Wood, R.D. DNA polymerases and cancer. *Nat. Rev. Cancer* **11**, 96–110 (2011).
54. Solimando, L. *et al.* Spatial organization of nucleotide excision repair proteins after UV-induced DNA damage in the human cell nucleus. *J. Cell Sci.* **122**, 83–91 (2009).

ONLINE METHODS

Somatic mutation data. We obtained somatic point-mutation data from 366 completely sequenced genomes from melanoma (SKCA, 25 samples)[30], lung squamous cell carcinoma (LUSC, 31 samples)[31], gastric cancer (STAD, 100 samples)[32], diffuse large B cell lymphoma (DLBCL, 40 samples)[33], chronic lymphocytic leukemia (CLL, 150 samples)[34], and prostate cancer (PRAD, 20 samples)[35,36]. Somatic mutation data and other data types were mapped to the human reference genome (hg19). Mutation frequencies for the samples in these cohorts were comparable to those published in the literature[14], and there were no outlier subsets of samples with excessive mutations or skewed mutational signatures that dominated the overall patterns observed in our analyses.

Annotation of noncoding, nonrepetitive, nonconserved regions. Because the mutational landscape of cancer genomes is shaped by the incidence of mutations as well as natural selection during clonal evolution acting on the variability thus generated[55,56], and because variant calling is technically challenging in some genomic regions (for example, centromeres, telomeres, and repetitive regions), we focused only on the noncoding, nonrepetitive, nonconserved regions (tier III annotation obtained from Mardis *et al.*[57]). In brief, such regions were identified after exclusion of repeat-masked regions, coding regions of annotated exons, canonical splice sites, and RNA-encoding genes, conserved genomic elements (cutoff: conservation score ≥500 on the basis of either the phastConsElements28way table or the phastConsElements17way table from UCSC genome browser), and regions with regulatory potential. (Regulatory annotations included were targetScanS, ORegAnno, tfbsConsSites, vistaEnhancers, eponine, firstEF, L1 TAF1 Valid, poly(A), switchDbTss, encodeUViennaRnaz, and cpgIslandExt[57].) Such regions are generally expected to evolve in the absence of strong (positive or negative) selective pressure[58] and should have no major issues with next-generation sequencing or mappability.

Annotation of nuclear-core and nuclear-periphery regions. Data on nuclear localization of human chromosomes were obtained from Bolzer *et al.*[38]. We obtained genome-wide data on LADs for multiple human and mouse cell types[19,20]. In these data sets, LADs were identified by using DamID treatment with a chimeric protein consisting of DNA adenine methyltransferase fused to lamin A or B1. DamID maps of (i) lamin B1 in mouse embryonic stem cells, astrocytes, neuronal precursor cells, and mouse embryonic fibroblasts were obtained from Peric-Hupkes *et al.*[20]; (ii) lamin B1 in human Tig3 fibroblasts were obtained from Guelen *et al.*[19]; and (iii) lamin B1 in human embryonic stem cells and HT1080 cells, and in mouse *Pou2f1*$^{-/-}$ and matching wild-type mouse embryonic fibroblasts, as well as lamin A in human HT1080 cells and in mouse neuronal precursor cells and astrocytes, were obtained from Meuleman *et al.*[59] Genomic regions associated with lamins are predominantly at the nuclear periphery, although some nucleoplasmic LADs accumulate around nucleoli in the interior[19,20,39], whereas those at the core were distinguished by the absence of interactions with nuclear lamina. Genome-wide distributions of lamina-associated regions are largely similar (73–87%) between different cell types in higher eukaryotes[20].

Overlaying lamin A and B1 data, we identified the regions that overlapped lamin-associated regions in (i) all the human cell lines tested and (ii) none of the human cell line tested, and denoted them as being constitutively present at the nuclear periphery (denoted cLADs) and core (denoted constitutive iLADs), respectively, in a cell-type-invariant manner (**Fig. 2b**). Genomic regions in the nuclear core and periphery have differences in gene density, repetitive elements, and evolutionarily conserved elements, and those features can influence selection on the somatic mutations (for example, gene region) and mutation calling (for example, repetitive regions). Therefore, to minimize biases in our analysis, for all analyses presented in **Figures 1**, **2**, and **4**, we considered only tier III segments[57] (i.e., noncoding, nonrepetitive, nonconserved genomic segments) of the cLAD and iLAD regions. In the multivariate analysis presented in **Figure 3b**, we used gene density, repetitive elements, evolutionary conservation, and other features as covariates.

As an even more conservative approach, by integrating human and mouse LA data in a similar manner, we also identified tier III segments of cLAD and iLADs with evolutionarily conserved patterns of localization in the nuclear periphery (denoted conserved and constitutive cLAD regions, cLAD^c) and nuclear core (denoted conserved and constitutive iLAD regions, iLAD^c; **Fig. 2b**), respectively, and compared the AMR between them (**Supplementary Fig. 3**).

Annotation of nuclear-pore-proximal regions. Nucleoporins are key components of nuclear-pore complexes that control nucleocytoplasmic trafficking. Liang *et al.* have examined genomic regions bound to Nup98, a nucleoporin family nuclear-pore protein, through ChIP using multiple antibodies to Nup98 in four cell types, three of which are related by direct lineage[49]. In tissue stem- and progenitor-cell populations, NBRs are predominantly at the nuclear periphery, but some NBRs also exist at the nuclear core, and Nup98 binding dynamically changes between cell types and during development[49]. We classified the cLAD and iLAD genomic regions as nuclear pore proximal if they were within 50 kb of Nup58 ChIP peaks in all cell types examined. We observed similar results by using 20 kb and 100 kb windows.

Annotation of replication timing, chromatin, and other covariates. Repli-Seq signals were downloaded for multiple tissue types[60] from the ENCODE data portal (**Supplementary Table 1**) and, following the approach used in a previous study[61], we kept only one GM12878 cell line data set to decrease the bias toward blood. Similarly, H3K9me3 histone modification marks across different tissue types were obtained from the Epigenomic Roadmap project[62], including tissues such as liver and lung (**Supplementary Table 3**). The transcripts, GC percentage information and phastCons conservation scores for the human genome (hg19), calculated from multiple alignments with other 99 vertebrates, were extracted from the UCSC genome browser database[63]. For each 1-Mb bin, the GC percentage, the number of genes overlapping with the bin, the proportion of nucleotides located in gene region, and the average phastCons conservation score were computed. For replication timing and H3K9me3 signals, we first calculated the average signal for each 1 Mb within each cell type, then averaged the values across different cell types. Because, in general, the cell of origin of different cancer types is unknown, the average signal across different cell types can be used as a more robust measure of such signals, with the trade-off of loss of cell-type-specific information.

Statistical analysis. We conducted both random forest regression and multiple linear regression to analyze the effects of LADs on the average mutation frequency over different tumors within a certain cancer type, adjusting for conservation score, GC percentage, gene density, average replication timing signal (with a higher signal indicating more enrichment with early replication timing on average), and the average signal of the heterochromatin mark H3K9me3 across multiple cell lines (**Supplementary Figs. 7** and **8** and **Supplementary Tables 1–3**). The adjusted R^2 for the linear model and the variance explained by the features of the random forest regression are shown in **Supplementary Table 3**. The use of linear regression was justified by using the residual plots and central limit theorem when averaging the mutation frequencies of each 1-Mb bin over different tumors (**Supplementary Fig. 7**). To account for potential correlation among 1-Mb bins, we calculated the robust sandwich standard error[64] in all regression analyses. When analyzing the mutation frequency, averaging across different tumors within the same cancer type, the appropriateness of a linear model with additive effects of different genomic features can be justified by using residual plots (**Supplementary Fig. 7**). To make the scale of coefficients of different features comparable, we normalized all the features to zero mean and unitary variance.

For the random forest regression, the function *cforest()* in the R package ‘party’ was used. The variable importance metrics for the genomic features were computed on the basis of permutation methods by using the *varimp()* function in the same package (**Fig. 3b**). The same set of features was included during random forest regression, again with average mutation frequencies in 1-Mb windows across samples as the dependent variable. The goodness of fit of random forest regression was again justified by using the residual plots (**Supplementary Fig. 7**). Because the genomic features analyzed were generally correlated, we also computed the conditional variable importance metric[65], which aims to remove some of the bias due to multicollinearities among the features (**Supplementary Fig. 8**). Because of the computational complexity, we were not able to compute the genome-wide metrics. In an alternative approach, we randomly divided the genome into ten groups 50 times, computed the metrics within each group, calculated the median metrics across groups, and finally plotted the distribution of these median scores across 50 randomizations. However, as outlined in Strobl *et al.*[65], such an attempt cannot guarantee the complete removal of the multicollinearity bias. Therefore, even though **Supplementary Figure 8** shows that LAD has a conditional variable importance metric similar to and sometimes even stronger than those of H3K9me3 and replication timing, such results cannot necessarily be interpreted to indicate that LAD is a more important factor than H3K9me3 in DLBCL.

Finally, because the different cancer cohorts had different sample sizes, we explored how the sample size might have influenced our key results. To test the robustness of our findings over different sample sizes, we computed the variable importance metrics for the genomic features on the basis of sample sizes equal to 10, 20, 30, and 40 in the lymphoma cohort, and we found that the patterns were highly similar across different sample sizes (**Supplementary Fig. 8**).

Mutational signatures are patterns in the occurrence of somatic single-nucleotide variants that can reflect underlying mutational and/or repair processes. We applied non-negative matrix factorization and principal component analysis to define mutation signatures, then evaluated their contributions to each sample's mutational spectrum by using the somaticSignature R package[40]. To examine the significance of nuclear localization in mutagenic and repair processes, we partitioned the genome according to chromatin or replication-timing context, and then analyzed differences in mutation signatures between cLAD and iLAD regions within the respective context. Two-tailed *P* values for respective cohorts were calculated with Mann–Whitney *U* tests. COSMIC mutational signatures were obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC; http://cancer.sanger.ac.uk/cosmic/) and were based on a previously published report[37].

A **Life Sciences Reporting Summary** for this paper is available online.

Data availability. Publicly available data sets were used for this study. Nonetheless, all data are available upon request.

55. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
56. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).
57. Mardis, E.R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
58. Ohta, T. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286 (1992).
59. Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res.* **23**, 270–280 (2013).
60. 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).
61. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
62. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
63. Speir, M.L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* **44** D1, D717–D725 (2016).
64. Freedman, D.A. On the so-called 'Huber sandwich estimator' and 'robust standarderrors'. *Am. Stat.* **60**, 299–302 (2006).
65. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics* **9**, 307 (2008).

doi:10.1038/nsmb.3474

Corresponding Author Name: Michor

Manuscript Number: NSMB-AN37603C

Reporting Checklist For Life Sciences Articles

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more information, please read Reporting Life Sciences Research. List items are standard for all Nature journal articles but may not apply to all disciplines or manuscripts.

Figure legends

☑ Check here to confirm that the following information is available in all relevant figure legends (or Methods section if too long):

- the **exact sample size (*n*)** for each experimental group/condition, given as a number, not a range;
- a **description of the sample collection** allowing the reader to understand whether the samples represent **technical or biological replicates** (including how many animals, litters, culture, etc.);
- a **statement of how many times the experiment shown was replicated in the laboratory**;
- **definitions of statistical methods and measures**: (For small sample sizes (n<5) descriptive statistics are not appropriate, instead plot individual data points)
  - very common tests, such as *t*-test, simple χ^2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - **statistical test results**, e.g., ***P* values**;
  - definition of '**center values**' as **median** or **mean**;
  - definition of **error bars** as **s.d.** or **s.e.m.** or **c.i.**

This checklist will not be published. Please ensure that the answers to the following questions are reported in the manuscript itself. We encourage you to include a specific subsection in the Methods section for statistics, reagents and animal models. Below, provide the page number or section and paragraph number (e.g. "Page 5" or "Methods, 'reagents' subsection, paragraph 2").

Statistics and general methods

	Reported in section/paragraph or page #:
1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? (Give section/paragraph or page #)	pp 5, Method summary. ⊞
For animal studies, include a statement about sample size estimate even if no statistical methods were used.	Animals were not used for this study.
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established? (Give section/paragraph or page #)	pp1, last para
3. If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it. (Give section/paragraph or page #)	NA
For animal studies, include a statement about randomization even if no randomization was used.	Animals were not used for this study.
4. If the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, state the extent of blinding. (Give section/paragraph or page #)	NA
For animal studies, include a statement about blinding even if no blinding was done.	NA
5. For every figure, are statistical tests justified as appropriate?	Yes
Do the data meet the assumptions of the tests (e.g., normal distribution)?	Yes. In most cases non-parametric statistics was used.
Is there an estimate of variation within each group of data?	Yes
Is the variance similar between the groups that are being statistically compared? (Give section/paragraph or page #)	Not necessarily, but appropriate statistics was used. pp 6, para 5: Multivariate analysis ⊞

(Continues on following page)

Reagents	Reported in section/paragraph or page #:
6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile (e.g., Antibodypedia, 1DegreeBio).	NA
7. Cell line identity:	
a. Are any cell lines used in this paper listed in the database of commonly misidentified cell lines maintained by ICLAC (also available in NCBI Biosample)?	NA
b. If yes, include in the Methods section a scientific justification of their use – indicate here on which page (or section and paragraph) the justification can be found.	NA
c. For each cell line, include in the Methods section a statement that specifies:	
- the source of the cell lines	NA
- have the cell lines been authenticated? If so, by which method?	NA
- have the cell lines been tested for mycoplasma contamination? In this checklist, indicate on which page (or section and paragraph) the information can be found.	NA

Animal Models	Reported in section/paragraph or page #:
8. Report species, strain, sex and age of animals	NA
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	NA

10. We recommend consulting the ARRIVE guidelines (PLoS Biol. 8(6), e1000412,2010) to ensure that other relevant aspects of animal studies are adequately reported.

Human Subjects	Reported in section/paragraph or page #:
11. Identify the committee(s) approving the study protocol.	NA
12. Include a statement confirming that informed consent was obtained from all subjects.	NA
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	NA
14. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent).	NA
15. For phase II and III randomized controlled trials, please refer to the CONSORT statement and submit the CONSORT checklist with your submission.	NA
16. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines.	NA

(Continues on following page)

Data Availability

Reported in section/paragraph or page

17. Please provide a Data Availability statement in the Methods section under “Data Availability”. Data availability statements should include, where applicable, accession codes, other unique identifiers and associated web links for publicly available datasets, and any conditions for access of non-publicly available datasets. Where figure source data are provided, statements confirming this should be included in data availability statements. Please refer our data availability and data citations policy for detailed guidance on information that must be provided in this statement.

Data deposition in a public repository is mandatory for:

a. Protein, DNA and RNA sequences
b. Macromolecular structures
c. Crystallographic data for small molecules
d. Microarray data

We used publicly available DNA mutation data from published papers and publicly available databases. Citations of the papers are provided.

Deposition is strongly recommended for many other datasets for which structured public repositories exist; more details on our data policy are available here. We encourage the provision of other source data in supplementary information or in unstructured repositories such as Figshare and Dryad. We encourage publication of Data Descriptors (see Scientific Data) to maximize data reuse

18. If computer code was used to generate results that are central to the paper’s conclusions, include a statement in the Methods section under “**Code availability**” to indicate whether and how the code can be accessed. Include version information as necessary and any restrictions on availability.

Standard statistical tests were used on publicy available data. Nevertheless, computer codes for the study will be available upon request.